Attachment 1 (1)

AFOSR-TR- 91 0982

THESIS BY:

JOHN S. GIERKE

MICHIGAN TECHNOLOGICAL UNIVERSITY

Subcontract No.# S-789-000-007

ATTACHMENT 6 IS BEST QUALITY AVIABLE AT THIS TIME. ATTACHMENT WILL BE SUBMITTED AT A LATER DATE AS AN ERRATA.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	30 April 1991	Annual Report 1 Aug 89-28 Feb 90

4. TITLE AND SUBTITLE
Thesis with 6 attachments
Annual Report
For 1990
Laboratory Graduate Fellowship Program

5. FUNDING NUMBERS
F49620-86-C-0127

6. AUTHOR(S)
Mr Rodney Darrah

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Universal Energy Systems, Inc. (UES)
4401 Dayton-Xenia Road
Dayton OH 45432

8. PERFORMING ORGANIZATION REPORT NUMBER
AFOSR-TR- [illegible] 2

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
AFOSR/NI
Bldg 410
Bolling AFB DC 20332-6448

Lt Col V. Claude Cavender

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION/AVAILABILITY STATEMENT
UNLIMITED

12b. DISTRIBUTION CODE

13. ABSTRACT (Maximum 200 words)

Critical to the success of the Air Force Office of Scientific Research (AFOSR) mission is the ability of AFOSR to draw upon the research community in the United States to respond to its needs. In recent years, however, the number of U. S. citizens seeking advanced degrees in the areas of Air Force research interests has been decreasing. This refers specifically to the number of U. S. citizens obtaining Ph.D. degrees in areas of mathematics, science, and engineering that are of interest to the Air Force. This situation points toward the potential problem of a future shortage of qualified researchers in areas critical to the nation's security interest.

To address this problem, the United States Air Force Laboratory Graduate Fellowship Program (USAF/LGFP) was established. The program annually provides three-year fellowships for at least 25 Ph.D. students in research areas of interest to the Air Force.

This report includes information on the following topics: Volatile Organic Materials in soil and Their Removal, Seismological Studies of Earth Structure, Cerebral Configurations of Parents and Siblings of Language Diordered Boys, Thermo Inelasticity, Theorems of Linear Systems, Feedback Stablization in Deformable Tokomak Plasmas Magnetisphere Ionosphere Coupling Measurements, and Analysis of Autonomic Activity During Motion Sickness.

15. NUMBER OF PAGES

CODE

TION OF ABSTRACT

298 (Rev. 2-89)

Vacuum extraction of VOCs from contaminated soils is an effective weapon against ground-water contamination.

VAPORIZING VOCs

NEIL J. HUTZLER
JOHN S. GIERKE
BLAINE E. MURPHY

When soils are contaminated with volatile organic chemicals (VOCs), the potential for ground-water contamination persists. Even after discharges have been stopped, the unsaturated zone above an aquifer retains a portion or all of a chemical discharge. Soil contaminants are flushed to ground water with infiltrating rain and diffuse through the air-filled pores. Soil vapor extraction is a cost-effective technique for VOC removal, and its use is becoming widespread.

A recent study for EPA of a representative sampling of installations offers some useful information. Basically, the method involves removing air that contains volatile chemicals from unsaturated soil. Fresh air is injected or flows into the subsurface at locations around a spill site, and the vapor-laden air is withdrawn under vacuum from recovery vents.

The method, also known as subsurface venting, vacuum extraction, in situ soil air-stripping and soil venting, has several advantages. It is a relatively simple concept and can be used in conjunction with other soil decontamination procedures—biological degradation, washing, on-site or off-site treatment or disposal. Air extraction processes cause minimal disturbance of contaminated soil and they use standard equipment. There is demonstrated experience at pilot and field scale up to several years duration, and the process can be used to treat larger volumes of soil than can be practically excavated. There is also the potential for product recovery.

SYSTEM VARIABLES

According to our findings for the EPA study, the design and operation of a vapor extraction system depends on several factors.

- Contamination volume. The extent to which the contaminants are dispersed in the soil, both vertically and horizontally, is important in deciding if vapor extraction is cost-effective. Soil excavation and treatment is probably less costly when only a few hundred cubic yards of near-surface soils are contaminated.
- Ground-water depth. Soil vapor extraction systems have been used in shallow as well as deep unsaturated zones. However, where ground water is more than 40 ft deep and contamination extends to the water table, a soil vapor extraction system may be the only way to remove VOCs from the unsaturated zone. In some cases, ground-water depth can be lowered to increase the volume of the unsaturated zone.
- Soil heterogeneity. Heterogeneities influence air movement aswell as the location of chemicals, making it more difficult to position extraction and inlet vents. There are significant differences in the air conductivity of the various soil horizons. A horizontally-stratified soil may be favorable for vapor extraction. The relatively impervious strata will limit the rate of vertical inflow from the ground surface and make the vacuum more effective horizontally from the extraction point. Some soil layering can make it easier to extract VOCs where horizontal air channeling occurs through sand layers with subsequent diffusion from less permeable layers.
- Contaminant location and area development. If the contamination extends across property lines, beneath a building or beneath an extensive utility trench network, vapor extraction should be considered instead a more intrusive decontaminating technique. This is especially so in a highly developed area.
- Site soil characteristics. Air conductivity controls the rate at which air can be drawn from soil by an applied vacuum. The soil moisture content or degree of saturation is also important because it is easier to draw air through drier soils. At one site in the Southwest where

0885-7024/90-0004-0057/$01.00 + 15¢ per page

the soil was relatively dry (2-5% moisture content), it took only 7 months to remove over 6 tons of dichloropropene using moderate air-flow rates of between 85 and 250 cu ft per min. Some designers think soil vapor extraction systems should be used only in highly permeable soils. However, these systems have been installed in soils with a wide range of permeabilities. Even clayey or silty soils can be effectively ventilated by the usual levels of vacuum developed in a soil vapor extraction system. The success of the soil vapor extraction in these soils depends on the presence of more conductive strata, as would be expected in alluvial settings or with relatively low moisture contents.

- Chemical properties. In conjunction with site conditions and soil properties, chemical properties determine the feasibility of a soil vapor extraction system. The system most effectively removes compounds that exhibit significant volatility at ambient subsurface temperatures. Common screening tools are the air-water partitioning coefficient and vapor pressure. Compounds with values of Henry's Law constants greater than 0.01 or vapor pressures over 1 in. of mercury are removable by vapor extraction. Trichloroethylene, 1,1,1-trichloroethane, methylene chloride, carbon tetrachloride, dichloroethylene, tetrachloroethylene, 1,3-dichloropropene and gasoline constituents (benzene, toluene, ethyl benzene and xylene) have all been successfully removed by vapor extraction. Compounds less easily removed include trichlorobenzene, acetone and heavier petroleum fuels.
- Operating variables. Higher air-flow rates tend to increase vapor removal because the zone of influence is increased and air is forced through more of the air-filled pores. More vents allow better control of air flow but also increase construction and operation costs. The water infiltration rate can be controlled by placing an impermeable cap over the site. Intermittent extraction from different vents allows time for chemicals to diffuse from immobile water and air zones, resulting in the removal of higher concentrations.
- Response variables. System performance parameters include air pressure gradients, VOC concentrations and power usage. Vapor removal rate is affected by the chemical's volatility, its sorptive capacity into soil, the air-flow rate, the distribution of air flow, the initial distribution of the chemical, soil stratification or aggregation and the soil moisture content.

DESIGNING THE SYSTEM

A typical system consists of (1) extraction vents, (2) air inlets or injection vents (optional), (3) air headers, (4) vacuum pumps or air blowers, (5) flow meters and controllers, (6) vacuum gauges, (7) sampling ports, (8) air-water separator (optional), (9) vapor treatment (optional) and (10) an impermeable cap (optional). Designing both pilot and field scale systems involves sizing and locating these components.

Vent design and placement. Extraction vents are typically designed to penetrate fully the unsaturated soil zone or the geologic stratum to be cleaned. An extraction vent is usually constructed of slotted plastic pipe placed in a permeable packing, and can be either vertical or horizontal. Vertical alignment is typical for deeper contamination zones and results in radial, subsurface air-flow patterns. If the depth of the contaminated soil or the depth to the ground-water table is less than 15 ft, it may be more practical to dig trenches across the area of contamination and install perforated piping horizontally. It is important to look at the depth of the vents relative to the water table elevation. Not only does the water table elevation vary seasonally—maybe even daily—but clay lenses can lead to perched water.

DESIGN AND OPERATION ARE FLEXIBLE AND EFFECTIVE.

In addition, reduced air pressures resulting from the applied vacuum at an extraction vent leads to mounding of the water table. Larger hole borings for the vents promote water vapor condensation within the bore packing.

The first few feet of augured columns for vertical vents or trenches for horizontal vents are usually grouted to prevent the direct inflow of air. Usually, several vents are installed. In stratified systems,

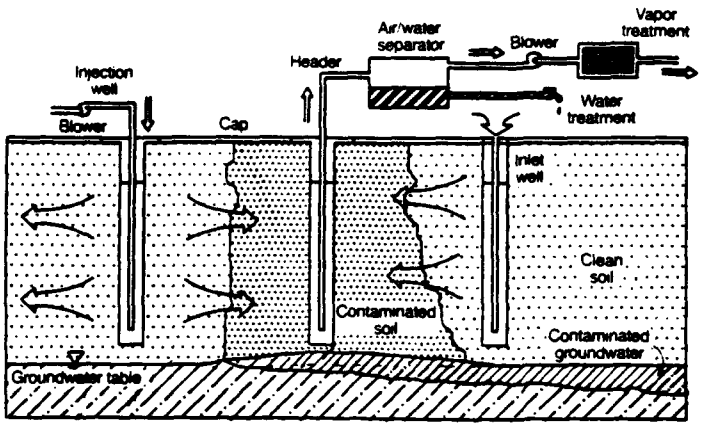
TABLE 1.
CONDITIONS AT TYPICAL VAPOR EXTRACTION SITES

	Site 1	Site 2	Site 3	Site 4
Soil type	sand/silt/gravel	sand	clayey sand	sand
Depth to ground water (ft)	240	40-50	48-53	>30
Contaminated volume (cu yd)	>4,000	>400,000	unknown	>33,000
Extraction vent diameter (in.)	2 (PVC)	(galvanized steel)	4 (PVC)	2 (PVC)
Screen depth (ft)	15-25	unknown	10-15	6-25
Number of vents	79	>20	6	7
Vent spacing (ft)	variable	unknown	14-50	40-90
Fresh-air source	inlet vents	inlet vents	surface	surface
Blowers	3	8	one vacuum pump	1
Total flow (cu ft/min)	86-250	unknown	unknown	210
Cap	none	clay and concrete	existing pavement	none
Vapor treatment	none	combustion	none	none
Contaminants	dichloropropene	acetone, ketone, toluene, xylene	gasoline	organic solvents
Initial mass (lb)	50,000-90,000	unknown	unknown	unknown
Amount extracted (lb)	90,000	>78,000	22,000	240
Duration of operation (months)	7	>5	7	unknown

more than one vent can be installed at the same location, each venting a given stratum. Extraction vents can be installed incrementally starting with areas of highest contamination. This allows the system to be brought online as soon as possible.

Spacing is usually based on an estimate of the radius of influence of an individual extraction vent.

FIGURE 1.
SYSTEM SCHEMATIC



A typical soil vapor extraction system configuration.

Vent spacing typically ranges from 15 to 100 ft. Screening depth probably has an effect only at low air-flow rates. At such flow rates, higher recovery rates result when the screen is placed near the water table rather than screening the full depth of the vadose zone. Vents should be constructed with solid pipe between the top of the screen and the soil surface to prevent the short circuiting of air and to extract deep contamination.

Air inlets. In the simplest soil vapor extraction systems, air flows to an extraction vent from the ground surface. To enhance air flow through zones of maximum contamination, air inlet or injection vents can be part of the installation. Air vents are passive, and usually only a fraction of extracted air comes from air inlets. This indicates that air drawn from the surface is the predominant source of clean air. Injection vents force air into the ground and must be installed at the edge of a site so as not to force contamination away from the extraction vents. They are also placed between adjacent extraction points to ensure pressure gradients in the direction of the extraction vents. Typically, injection and inlet vents are similar in construction to extraction vents. Extraction vents can be designed to be used as air inlets.

Piping and blower systems. Piping materials and headers are usually made of plastic or steel. However, headers should be constructed of steel for durability, especially in colder climates. Both pipes and headers are usually buried or wrapped with heat tape and insulated in northern climates to prevent condensate freezing. Headers can be configured in a grid, but manifold construction appears to be more common. A control valve is usually installed at each vent head and at other critical locations, such as lateral-header connections, to provide operational flexibility. Typically, ball or butterfly valves are used because they provide better flow control.

The vacuum for extracting soil air is developed by an ordinary positive displacement industrial blower, a rotary blower, vacuum or aspirator pump, or a turbine. Most commonly used blowers have ratings ranging from 100 to 6,000 cu ft per min at vacuums up to 30 in. of mercury gauge. Electric drive motor ratings are usually 10 hp or less. The pressure from the outlet side of the pumps or blowers is usually used to push the exit gas through a treatment system and can be used to force air back into the ground if injection vents are used. It is more common, however, to use a separate blower for injection. Vapor treatment efficiency can be improved by installing the blower between the moisture separator and the vapor treatment system to take advantage of the heat generated by the blower.

A flow meter should be installed to monitor the volume of extracted air. This measurement is used in conjunction with gas analysis to determine the total mass of contaminants extracted from the soil. Flow measurements from individual vents are useful for optimizing extraction system operation. A flow meter consisting of an orifice plate and manometer, together with the appropriate rating curve, will yield the system discharge air flow rate.

Impermeable caps. Capping the entire site with plastic sheeting, clay, concrete or asphalt enhances horizontal movement toward the extraction vent. A cap controls the air-flow pathway so that clean air is more likely to come from air vents or injection vents. Without a cap, air enters the soil from the surface nearest the vent. Impermeable caps extend the radius of influence around the extraction vent.

The use of a ground surface cover will also prevent or minimize infiltration, which, in turn, reduces moisture content and further chemical migration. With little or no infiltration, water is less likely to be extracted from the system. In very dry climates, a reduction of moisture content to where partial drying of the soil occurs reduces system efficiency due to increased adsorption capacity of the dry soil.

Air-water separator. If water is pulled from the extraction vents, an air-water separator is required to protect the blowers or pumps and to increase the efficiency of vapor treatment systems. The condensate may need treatment as a hazardous waste depending on the types and concentrations of contaminants. The need for a separator can be eliminated by covering the treatment area with an impermeable cap. In some cases, gasoline can be recovered by using a gasoline-water separator combined with a vapor extraction/pumping system.

Vapor treatment. Air emission problems should not be created while remediating soil contamina-

tion. However, vapor treatment may not be required for systems that produce easily degradable chemicals at low emission rates. There are several treatment systems available that limit or control air emissions. These include liquid/vapor condensers, incinerators, catalytic converters and granular activated carbon (GAC).

If air-emissions control or vapor treatment is required for an installation, a vapor-phase activated carbon adsorber system will probably be the most practical system, depending on chemical emission rates and VOC levels, although catalytic oxidation units have produced favorable results. Gas-phase GAC may require heating the extracted air. Heating controls the relative humidity, which, in turn, optimizes the carbon usage rate. As the fraction of water vapor increases, the capacity for the target chemical decreases and the carbon replacement rate increases. The spent carbon may be considered a hazardous waste due to the type of contaminants. Incineration can be self-sustained combustion if the vapor contains high concentrations of hydrocarbons or combustible volatile chemicals. Usually there is a lag time to achieve a high concentration of combustibles. Concentrations of volatiles in the air stream might be increased by intermittent blower operation or by intermittently operating different extraction vents. Some systems have auxiliary fuels to maintain a desired exhaust temperature.

There may be other uses for extraction system off-gas. On one project, where the initial extraction rate of volatiles was over 2,000 lb per day, the extracted gas was piped to the combustion air intake zone of a nearby industrial boiler in continuous operation.

Monitoring systems. Various monitoring devices such as sampling ports, vacuum gauges and pilot tubes are required for estimating vapor discharges. Pressure gauges are used to monitor the pressure losses in the overall system and to optimize air flow. Vapor and pressure monitoring probes can be placed in the soil surrounding the extraction system to measure vapor concentrations and the radius of influence. Monitoring wells are usually necessary to assess final site cleanup.

Sampling ports are usually installed at each vent head, at the blower and after gas treatment. The basic measurements required to assess system performance are the system air-flow rate and the concentration of volatile organic chemicals in the extracted flow. The system VOC concentration data is checked using a gas chromatograph with a detector appropriate for the compounds expected in the exhaust gas.

At most sites, the initial VOC recovery rates are relatively high and then decrease asymptotically to zero with time. Vapor extraction is more effective at sites where the more volatile chemicals are still present—when the spill is relatively recent. Several studies show that intermittent venting from individual vents is probably more efficient in terms of mass of VOC extracted per unit of energy expended. This is especially true when extracting from soils where mass transfer is limited by the rate at which chemicals diffuse out of immobile air zones and impermeable layers. Optimal system operation may involve taking individual vents in and out of service to allow time for liquid and gas diffusion and to change air-flow patterns in the region being vented.

The design and operation of soil vapor extraction systems can be quite flexible, allowing for changes during the course of operation in vent placement, blower size and air flows from individual vents. If the system is not operating effectively, changes in vent placement or capping the surface may improve it. At one site, the blowers were housed in modules with quick disconnect attachments, allowing for portability.

A major problem with soil vapor extraction systems is determining when the site is clean enough to cease operation. Mass balances using initial and final soil borings have not been particularly successful in predicting amounts of chemical actually removed. Soil vapor measurements in conjunction with soil boring and ground-water monitoring can better determine the remaining chemicals. Some designers suggest intermittent operation near the end of cleanup. If vapor concentration shows no significant increase on restart, one can assume the site is decontaminated.

OTHER APPLICATIONS

Soil air extraction is used in conjunction with ground-water pumping and treatment as a low-cost alternative for the cleanup of petroleum and solvent spills. Large quantities of organic chemicals can be retained in the vadose zone by capillary forces, dissolution in soil water, volatilization and sorption. If this product can be removed before it reaches the ground water then the problem is mitigated. Some studies show that vapor extraction is effective in removing organic chemical vapor, sorbed chemical and free product at the water table. This suggests that the soil should be decontaminated by vapor extraction before groundwater cleanup can be completed.

FOR DEVELOPED AREAS, SOIL VAPOR SYSTEMS ARE LESS INTRUSIVE THAN OTHER METHODS.

Researchers have also observed that inducing subsurface air-flow enhances microbial degradation of organic compounds. Ideally, then, vapor extraction could remove the lighter molecular weight, more volatile compounds while creating aerobic conditions to promote the degradation of the heavier, less volatile chemicals.

Soil vapor extraction is effective for removing volatile chemicals over a wide range of conditions. The design and operation of these systems is flexible enough to allow for rapid changes in operation, optimizing chemical removal.

Neil J. Hutzler, M.ASCE, is an associate professor of civil and environmental engineering at Michigan Technological University, Houghton. John S. Gierke, A.M.ASCE, is completing his doctoral work under Professor Hutzler. Blaine E. Murphy, A.M.ASCE, a graduate student when this study was performed, is an environmental engineer for Bechtel Environmental Inc., Houston. *The complete study is EPA 600/2-89-024, for the agency's Risk Reduction Engineering Laboratory, Cincinnati. The findings, opinions and conclusions are the authors' and do not necessarily reflect those of EPA.*

JRP/RL

WATER RESOURCES RESEARCH, VOL. ?, NO. ?, PAGES ?-?, MONTH ?, 19??

Modeling the Movement of Volatile Organic Chemicals in Columns of Unsaturated Soil

JOHN S. GIERKE, NEIL J. HUTZLER, AND JOHN C. CRITTENDEN

Department of Civil Engineering, Michigan Technological University, Houghton

Mechanisms affecting the fate of nondegradable volatile organic chemicals in soils include (1) advection in air and water, (2) dispersion in air and water, (3) air-water mass transfer and equilibrium, (4) diffusion in immobile water, (5) mass transfer between mobile and immobile water, and (6) sorption. A deterministic model was developed to account for these processes in laboratory columns of unsaturated soil. The general form of the model was solved numerically. The numerical solution was verified with analytic solutions for simplified conditions. Column experiments were conducted to validate the model and to determine the relative importance of each mechanism in two soil types. The movement of trichloroethene was measured in a column packed with a uniform sand and one packed with uniformly sized aggregates that were made from clay. Parameter values for the model predictions were independently determined from direct measurements and literature correlations. Bromide tracer studies were performed to determine parameter values that could not be measured directly or were not estimated accurately by literature correlations. For the sand column the amount of immobile water, the rate of liquid diffusion, and the liquid dispersion coefficient were measured in a tracer study. A batch rate study was used to measure the rate of intraaggregate diffusion in the clay aggregates. The liquid dispersion coefficient for the column containing aggregates was measured in a tracer study. These parameter values were used in the model to predict the breakthrough and elution of trichloroethene in the two columns. To describe the column data, however, Henry's constant was increased from a literature value of 0.4 to 0.7, and the predicted gas dispersion coefficient was reduced by a factor of 10.

INTRODUCTION

Subsurface contaminant transport and attenuation is governed by a number of spreading, retardation, and transformation mechanisms such as advection, dispersion, diffusion, and interfacial mass transfer; adsorption and volatilization; and biological and chemical reactions. These mechanisms and their impacts on chemical fate are discussed in detail by *MacKay et al.* [1985] and *Nielsen et al.* [1986]. As a chemical travels through soil with fluid flow, the shape of its concentration profile is affected by dispersing or spreading mechanisms, the profile's position is slowed by retardation mechanisms, and the concentration may also decrease due to biological and chemical transformations.

Previous work on modeling unsaturated transport has focused in three areas: (1) vapor transport in the upper soil layer for predicting pesticide movement [*Rolston et al.*, 1969; *Mayer et al.*, 1974] and for assessing the behavior of organic chemicals [*Jury et al.*, 1980, 1983], (2) tracer and nonvolatile chemical transport for simulating the one-dimensional movement of salts and heavy metals [*van Genuchten and Wierenga*, 1976; *Jury*, 1982], and (3) three-dimensional subsurface movement of liquids and vapors for estimating the travel time of organic solvents and petroleum products to groundwater [*Abriola and Pinder*, 1985a; *Lindstrom and Piver*, 1986; *Corapcioglu and Baehr*, 1987]. Although the three-dimensional models are conceptually closer to field conditions, these models have not considered all of the mechanisms. Moreover, the three-dimensional models have not been validated experimentally. To determine a model's predictive capability it is important to do so [*Abriola and Weber*, 1986].

Paper number 89WR03634.
0043-1397/90/89WR-03634$05.00

Mechanisms of nonequilibrium [*Nielsen et al.*, 1986; *Brusseau and Rao*, 1989], specifically those associated with physical reactions, have typically been ignored in the derivation of three-dimensional transport models. Models that describe the one-dimensional movement of vapors or aqueous solutes have studied some nonequilibrium effects. For example, *DeSmedt et al.* [1986] used a model developed by *van Genuchten and Wierenga* [1976] to study the impact of diffusion in immobile water on tracer breakthrough curves in columns of unsaturated sand. They propose that the importance of diffusion in immobile water increases with decreasing water content. The three-dimensional model derivations ignore this mechanism by assuming that all of the water is mobile.

There continues to be a need for a better understanding of subsurface fate processes [*Nielsen et al.*, 1986; *Abriola and Weber*, 1986], including those associated with physical nonequilibrium. To date, many studies aimed at gaining a better understanding of chemical fate processes in the vadose zone have been segmented and disciplinary [*Nielsen et al.*, 1986]. Especially lacking are integrated modeling-experimental approaches [*Abriola and Weber*, 1986]; approaches that not only formulate hypotheses of fate mechanisms but test the hypotheses experimentally as well as numerically.

This paper presents a one-dimensional model that describes the movement of dilute solutions of nondegradable volatile organic chemicals (VOCs) in unsaturated soil. Verification of the model's numerical solution and experimental results for determining the validity of the model as a tool for understanding subsurface transport are also presented. Verification of the numerical solution is obtained by comparing numerical calculations to analytic solutions for simplified conditions. Experimental results are used to validate the model and, in conjunction with numerical calculations, to improve the level of understanding of several important

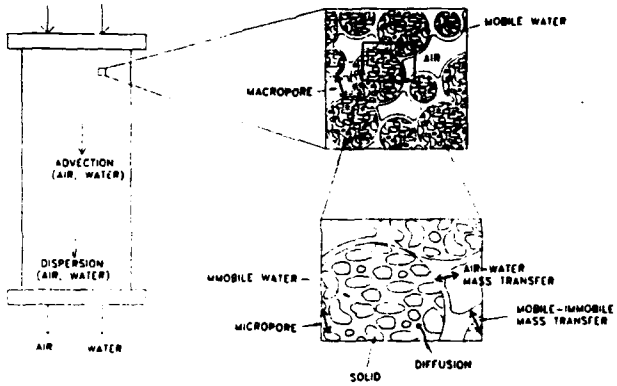


Fig. 1. Conceptual picture of unsaturated soil column used for model development.

mechanisms of subsurface transport. This paper shows that the model can be used to examine the contributions of advection, dispersion, mass transfer resistance, diffusion in immobile water, sorption, and volatilization on the spreading and retardation of VOC breakthrough curves in columns packed with homogeneous soil materials.

Model Development

Figure 1 is a conceptual picture of a soil column that is used for the model development that follows. The mechanisms being considered are (1) air and water advection, (2) liquid and gas dispersion in the direction of flow, (3) liquid diffusion in pores filled with immobile water, (4) mass transfer resistance at the air-water and the mobile-immobile water interfaces, (5) partitioning between the air-water phase, and (6) sorption to soil organic matter from aqueous solution.

Descriptions of unsaturated water flow are complicated by hysteresis and the variation of soil properties with location. In addition, dynamic environmental conditions such as atmospheric pressure, ambient temperature, and rainfall events alter the flow of water. Air flow also affects liquid permeability. Since one of the primary objectives of this study is to compare the relative impact of liquid advection to dispersion, diffusion, and mass transfer resistance, it is advantageous to assume constant pressure, temperature, and wetting conditions and assume that soil properties and moisture profiles are uniform. *Wierenga* [1977] showed that when analyzing breakthrough curves of noninteracting solutes on the basis of cumulative drainage, transient and steady flow calculations give similar results.

Air flow in unsaturated soil is usually small in comparison to water flow, but it is becoming common practice to induce air flow for removing VOCs from soil [*Hutzler et al.*, 1989*a*]. Steady, one-dimensional air flow in the same direction (cocurrent) as water flow was included in this model development so that the model could be used to determine when mass transfer resistance at the air-water interface is an important mechanism in unsaturated soil. The model development assumes that the air is saturated with water so that moisture content remains constant.

Liquid dispersion in the direction of flow is much greater in unsaturated soil than in saturated soil [*DeSmedt et al.*, 1986]. Liquid dispersion is assumed constant and is described using a Fickian approach because of the assumption of steady water flow and uniform moisture content [*Wierenga*, 1977]. Because of these simplifying conditions, gas dispersion is described in the same manner as liquid dispersion.

Diffusion in immobile water has been shown to be important in saturated soil systems [*Crittenden et al.*, 1986; *Hutzler et al.*, 1986; *Roberts et al.*, 1987]. The significance of intraaggregate diffusion in unsaturated soil is thought to be greater than in saturated systems [*van Genuchten and Wierenga*, 1977; *van Genuchten et al.*, 1977; *DeSmedt et al.*, 1986]. Diffusion in immobile water zones is described here as Fickian diffusion in saturated micropores contained within uniformly sized, spherical aggregates [*Rao et al.*, 1982].

Mass transfer resistance at the air-water interface could be an important mechanism in fate modeling of volatile pollutants in unsaturated soil. Others have ignored its impact in subsurface transport by assuming rapid air-water equilibrium [*Abriola and Pinder*, 1985*b*; *Lindstrom and Piver*, 1986; *Corapcioglu and Baehr*, 1987]. Air-water mass transfer is included in the model to test the impact it has on chemical movement. Although mass transfer resistance at the mobile-immobile water interface (film transfer) has been found by others to be unimportant in saturated soils [*Crittenden et al.*, 1986; *Hutzler et al.*, 1986; *Roberts et al.*, 1987], it is included here to study its importance in unsaturated systems.

Retardation of the rate of movement of a chemical profile relative to rate of water or air movement is a function of the solute phase distribution at equilibrium. The model development assumes VOC solutions are dilute. Hence air-water equilibrium can be described by Henry's law. Freundlich adsorption equilibrium [*Freundlich*, 1922] is used to describe chemical equilibrium between water and soil organic matter. Adsorption of vapors onto soil was assumed to be unimportant [*Roy and Griffin*, 1987] because soil surfaces are usually covered by water.

Derivation of Dimensioned Equations

The model was developed to describe the movement of a single VOC in laboratory columns of unsaturated soil where the mechanisms described above are operative. The soil system is divided into the three zones shown in Figure 1: mobile or immobile air, mobile water, and aggregates composed of immobile water and solid soil particles. Mass balances on these zones result in three partial differential equations in terms of dimensioned variables and parameters.

The air and mobile water, as depicted in Figure 1, are assumed to be continuous phases. The smallest pores, such as those contained in aggregates, contain water that is immobile. If the water content exceeds the soil's field capacity, water will flow through larger pores while the largest pore spaces still contain air. Since there is no theoretical method for determining the fraction of interphase contact between air and immobile water and between air and mobile water, the rate of air-water mass transfer is assumed to be lumped into a single transfer rate between air and mobile water. Therefore it is assumed that chemical phase transfer occurs between the air and mobile water and the mobile and immobile water. It is also assumed that the water in contact with soil surfaces, where sorption can occur, is immobile.

A mass balance on the air zone results in the following equation:

$$\frac{\partial C_v(Z, T)}{\partial T} = E_v \frac{\partial^2 C_v(Z, T)}{\partial Z^2} - u \frac{\partial C_v(Z, T)}{\partial Z} - \frac{K_L a}{\epsilon(1 - S)}\left[\frac{C_v(Z, T)}{H} - C_b(Z, T)\right] \quad (1)$$

Equation (1) describes the change in VOC vapor concentration ($C_v(Z, T)$) with respect to time. The terms on the right side represent gas dispersion or diffusion, gas advection, and air-water mass transfer.

A mass balance on the mobile water zone results in a similar expression for the change in VOC concentration in mobile water ($C_b(Z, T)$):

$$\frac{\partial C_b(Z, T)}{\partial T} = E_z \frac{\partial^2 C_b(Z, T)}{\partial Z^2} - v \frac{\partial C_b(Z, T)}{\partial Z} + \frac{K_L a}{\epsilon(S - S_i)} \cdot \left[\frac{C_v(Z, T)}{H} - C_b(Z, T)\right] + \frac{3k_f[1 - \epsilon(1 - S_i)]}{\epsilon(S - S_i)R_a} \cdot [C_p(R = R_a, Z, T) - C_b(Z, T)] \quad (2)$$

The terms on the right side of (2) represent liquid dispersion, liquid advection, mass transfer between air and mobile water, and film transfer. The mathematical representation of mass transfer between the mobile and immobile water in (2) assumes that an aggregate whose center is at axial position Z absorbs chemical at a rate proportional to the deficit between the immobile water concentration at the aggregate surface ($C_p(R = R_a, Z, T)$) and the mobile water concentration. This representation of film transfer is appropriate when the axial water concentration gradient ($\partial C_b/\partial Z$) is small across an aggregate diameter.

Figure 1 shows that soil aggregates are represented by uniformly porous spheres [*Rao et al.*, 1982] within which the aqueous chemical concentration in the micropores is at equilibrium with the sorbed-phase concentration on the adjacent soil surface. The change in the total intraaggregate concentration ($Y(R, Z, T)$) with respect to time is equal to the rate at which chemical diffuses through the internal pores:

$$\frac{\partial Y(R, Z, T)}{\partial T} = \frac{1}{R^2}\frac{\partial}{\partial R}\left[\frac{D_p \epsilon S_i R^2}{\rho_s(1 - \epsilon)}\frac{\partial C_p(R, Z, T)}{\partial R}\right] \quad (3)$$

Like (2), (3) assumes that changes in axial water concentration are small across an aggregate diameter. The variable $Y(R, Z, T)$ is the total chemical concentration per unit mass of soil at a specific radial position within an aggregate:

$$Y(R, Z, T) = \frac{\epsilon S_i}{\rho_s(1 - \epsilon)} C_p(R, Z, T) + Q(R, Z, T) \quad (4)$$

The equilibrium between the aqueous and sorbed phases within the micropores is described with the *Freundlich* [1922] isotherm equation:

$$Q(R, Z, T) = KC_p(R, Z, T)^{1/n} \quad (5)$$

where $Q(R, Z, T)$ is the mass of chemical sorbed per unit mass of soil.

The initial condition for solving the above equations can be any specified concentration profile within the column. Typically, for soil columns the initial concentrations are zero:

$$C_v(0 \leq Z \leq L, T = 0) = C_b(0 \leq Z \leq L, T = 0) = Y(0 \leq R \leq R_a, 0 \leq Z \leq L, T = 0) = 0 \quad (6)$$

Boundary conditions for (1) and (2) are derived from the fact that soil columns are closed reactors [*Levenspiel*, 1962]. It will be shown later that analytical solutions of the model exist for simplified conditions; however, for the general case shown a numerical solution method is necessary. In an attempt to force the numerical method to conserve chemical mass, overall mass balances on the mobile water and air phases are used for two boundary conditions. The difference between the mass of VOC entering and leaving the column by advection in air must equal the mass accumulating in the air minus the mass transferred to the air from the water:

$$u[C_{in}(T) - C_v(Z = L, T)] = \frac{\partial}{\partial T}\int_0^L C_v(Z, T)\,\partial Z - \int_0^L \frac{K_L a}{\epsilon(1 - S)}\left[C_b(Z, T) - \frac{C_v(Z, T)}{H}\right]\partial Z \quad (7)$$

$C_{in}(T)$ is the time-varying influent concentration in the air.

The difference in chemical mass entering and leaving the column in water must equal the accumulation in the water and in the aggregates plus the mass transferred to the air from the water:

$$v[C_{in}(T) - C_b(Z = L, T)] = \frac{\partial}{\partial T} \cdot \int_0^L \left[C_b(Z, T) + \frac{3\rho_s(1 - \epsilon)}{\epsilon(S - S_i)R_a^3}\int_0^{R_a} Y(R, Z, T)R^2\,dR\right]\partial Z$$

TABLE 1. Definitions of Dimensionless Groups

Group	Definition	Equation
	Mass Transfer Groups	
Ar	rate of advection in air / rate of advection in mobile water	$\frac{u}{v} Dg_a$
Ed_p	rate of diffusion in immobile water / rate of advection in mobile water	$\frac{D_p Dg_p L}{vR_a^2}$
Pe_b	rate of advection in mobile water / rate of axial dispersion in mobile water	$\frac{vL}{E_z}$
Pe_a	rate of advection in mobile water / rate of axial dispersion in air	$\frac{vL}{E_z Dg_a}$
Pe	total advective flux in air and in water / total dispersive flux in air and in water	$\frac{1 + Ar}{Pe_b^{-1} + Pe_a^{-1}}$
St_b	rate of transport across mobile–immobile water interface / rate of advection in mobile water	$\frac{k_f[1 - \varepsilon(1 - S_i)]L}{v\varepsilon(S - S_i)R_a}$
St_a	rate of transport across air–mobile water interface / rate of advection in mobile water	$\frac{K_L a L}{3v\varepsilon(S - S_i)}$
	Chemical Distribution Groups	
Dg_p	mass of chemical in immobile water / mass of chemical in mobile water	$\frac{S_i}{(S - S_i)}$
Dg_s	mass of chemical adsorbed to soil / mass of chemical in mobile water	$\frac{\rho_s(1 - \varepsilon)KC_{bn}^{(1/n - 1)}}{\varepsilon(S - S_i)}$
Dg_a	mass of chemical in air / mass of chemical in mobile water	$\frac{(1 - S)H}{(S - S_i)}$
Dg	mass of chemical in air, in immobile water, and on soil / mass of chemical in mobile water	$Dg_p + Dg_s + Dg_a$
R_d	velocity of chemical front / velocity of water	$\frac{1 + Dg}{(1 + Ar)(1 + Dg_p)}$
$1/n$	isotherm intensity	

$$+ \int_0^L \frac{K_L a}{\varepsilon(S - S_i)} \left[C_b(Z, T) - \frac{C_a(Z, T)}{H} \right] \partial Z \qquad (8)$$

$C_{bn}(T)$ is the time varying influent concentration in water.

Exit boundary conditions for (1) and (2) are obtained by solving the following equations for the given initial conditions:

$$\frac{\partial^2 C_a(Z = L, T)}{\partial Z\, \partial T} = 0 \qquad \frac{\partial C_a(Z = L, T = 0)}{\partial Z} = 0 \qquad (9)$$

$$\frac{\partial^2 C_b(Z = L, T)}{\partial Z\, \partial T} = 0 \qquad \frac{\partial C_b(Z = L, T = 0)}{\partial Z} = 0 \qquad (10)$$

One boundary condition for (3) results from symmetry; that is, no concentration gradient exists at the center of an aggregate:

$$\frac{\partial C_p(R = 0, Z, T)}{\partial R} = 0 \qquad (11)$$

Symmetry is based on the assumptions of spherical aggregates and that axial changes in chemical concentration are small across an aggregate diameter.

The other boundary condition for (3) is derived by performing a mass balance on an aggregate. A change in mass of chemical in an aggregate is equal to the mass transferred to the aggregate from the mobile water:

$$\frac{\partial}{\partial T} \int_0^{R_a} Y(R, Z, T) R^2\, dR = \frac{k_f R_a^2}{\rho_s(1 - \varepsilon)}$$

$$\cdot [1 - \varepsilon(1 - S_i)][C_b(Z, T) - C_p(R = R_a, Z, T)] \qquad (12)$$

Equation (12) is consistent with the assumption used to represent film transfer in (2). Like (7) and (8), (12) attempts to conserve mass during the numerical solution of the model equations.

Conversion to Dimensionless Form

To reduce the complexity of the equations, so that model solutions could be based on and characterized by fewer parameters, the dimensioned equations derived above were converted to a dimensionless form. Soil column model predictions in terms of relative (dimensionless) concentration as a function of dimensionless time (throughput) can be characterized by the groups defined in Table 1. These groups

represent mass transfer mechanisms and chemical distributions at equilibrium. Because water is the most common transport medium in soil, the mass transfer groups are based on the rate of mass transport by water advection, and the chemical distribution groups are based on chemical mass in mobile water. The magnitudes of the five mass transfer groups (air Peclet (Pe_v), mobile water Peclet (Pe_b), immobile water diffusion modulus (Ed_p), mobile-immobile water Stanton (St_b), air-water Stanton (St_v)) represent the degree of spreading exhibited by a breakthrough curve [*Roberts et al.*, 1987]. A large value of any of these groups indicates a small contribution from the corresponding mechanism toward the observed spreading. For example, a large value of Pe_b means that transport by liquid dispersion is slow in comparison to that by water advection and therefore not important. An increase in the air-water advective flux ratio (Ar) has the same effect as increasing Pe_b and Pe_v and decreasing St_b, St_v, and Ed_p. The chemical distribution groups (immobile water (Dg_p), sorbed (Dg_s), vapor (Dg_v)) and the isotherm intensity ($1/n$) impact spreading because they determine the amount of chemical in a given phase. Only Ar, Dg_s, Dg_v, and $1/n$ affect the magnitude of the retardation coefficient (R_d). If $1/n$ is not equal to 1, then R_d is also dependent on concentration.

Throughput (t) is defined by assuming that a soil column is initially free of chemical and that the influent concentrations are constant and in equilibrium ($C_{bi}(T) = C_{bm}$, $C_{vi}(T) = HC_{bm}$). Throughput is equal to the ratio of chemical mass fed to the mass contained in the column at equilibrium with C_{bm}. Dimensionless concentrations are derived by dividing a particular phase concentration by its concentration in equilibrium with C_{bm}. Axial position (Z) is normalized by the column length (L) and radial position (R) by the aggregate radius (R_a).

The dimensionless forms of the air mass balance (equation (1)) and its boundary conditions (equations (7) and (9)) are

$$\frac{\partial c_v(z,t)}{\partial t} = \frac{[1+Dg]}{Dg_v[1+Ar]}\left[\frac{1}{Pe_v}\frac{\partial^2 c_v(z,t)}{\partial z^2} - Ar\frac{\partial c_v(z,t)}{\partial z} - 3St_v[c_v(z,t) - c_b(z,t)]\right] \quad (13)$$

$$c_m(t) - c_v(z=1,t) = \frac{Dg_v[1+Ar]}{Ar[1+Dg]}\frac{\partial}{\partial t}\int_0^1 c_v(z,t)\,dz - \int_0^1 \frac{3St_v}{Ar}[c_b(z,t) - c_v(z,t)]\,dz \quad (14)$$

$$\frac{\partial^2 c_v(z=1,t)}{\partial z\,\partial t} = 0 \qquad \frac{\partial c_v(z=1,t=0)}{\partial z} = 0 \quad (15)$$

In dimensionless form the mobile water mass balance (equation (2)) and its boundary conditions (equations (8) and (10)) become

$$\frac{\partial c_b(z,t)}{\partial t} = \frac{[1+Dg]}{[1+Ar]}\left[\frac{1}{Pe_b}\frac{\partial^2 c_b(z,t)}{\partial z^2} - \frac{\partial c_b(z,t)}{\partial z} + 3St_v[c_v(z,t) - c_b(z,t)] + 3St_b[c_p(r=1,z,t) - c_b(z,t)]\right] \quad (16)$$

$$c_{bi}(t) - c_b(z=1,t) = \frac{[1+Ar]}{[1+Dg]}\frac{\partial}{\partial t}\cdot\int_0^1 [c_b(z,t) + 3[Dg_p + Dg_s]\int_0^1 y(r,z,t)r^2\,dr]\,\partial z + \int_0^1 3St_v[c_b(z,t) - c_v(z,t)]\partial z \quad (17)$$

$$\frac{\partial^2 c_b(z=1,t)}{\partial z\,\partial t} = 0 \qquad \frac{\partial c_b(z=1,t=0)}{\partial z} = 0 \quad (18)$$

The dimensionless form of the intraaggregate mass balance (equation (3)) is

$$\frac{\partial y(r,z,t)}{\partial t} = \frac{Ed_p[1+Dg]}{[Dg_p + Dg_s][1+Ar]}\frac{1}{r^2}\frac{\partial}{\partial r}\left[r^2\frac{\partial c_p(r,z,t)}{\partial r}\right] \quad (19)$$

The dimensionless total intraaggregate concentration must satisfy

$$y(r,z,t) = \frac{Dg_p c_p(r,z,t) + Dg_s c_p(r,z,t)^{1/n}}{Dg_p + Dg_s} \quad (20)$$

Equation (20) was obtained by substituting (5) into (4) and dividing the result by the total intraaggregate concentration in equilibrium with C_{bm}.

The dimensionless form of the boundary conditions (equations (11) and (12)) for the intraaggregate mass balance are

$$\frac{\partial c_p(r=0,z,t)}{\partial r} = 0 \quad (21)$$

$$St_b[c_b(z,t) - c_p(r=1,z,t)] = \frac{[Dg_p + Dg_s][1+Ar]}{[1+Dg]}\frac{\partial}{\partial t}\int_0^1 y(r,z,t)\,r^2\,dr \quad (22)$$

The dimensionless initial condition for solving (13), (16), and (19) is

$$c_{vi}(0 \leqq z \leqq 1, t=0) = c_{bi}(0 \leqq z \leqq 1, t=0) = y(0 \leqq r \leqq 1, 0 \leqq z \leqq 1, t=0) = 0 \quad (23)$$

Converting the model into a dimensionless form reduces the number of parameters that characterize a solution from 17 (a, D_p, E_v, E_z, H, K, k_f, K_L, L, $1/n$, R_a, S, S_i, u, v, ϵ, ρ_s) to 10 (Ar, Dg_p, Dg_s, Dg_v, Ed_p, $1/n$, Pe_b, Pe_v, St_b, St_v). It is also easier to characterize a solution in terms of these groups. Five groups (Ed_p, Pe_b, Pe_v, St_b, St_v) affect only the shape of a breakthrough curve. Nine dimensioned parameters (H, K, L, $1/n$, S, u, v, ϵ, ρ_s) impact both shape and position; the other eight affect only the shape.

Model Solutions

The general form of the model given by (13)–(23) is solved numerically. Exact solutions can be obtained, however, when $1/n$ is equal to 1 and certain mass transfer mechanisms are unimportant and when simpler boundary conditions than those given by (14), (15), (17), and (18) are used. Several analytic solutions are given below in increasing order of complexity, and the numerical solution is presented after them.

In the results and discussion section, model verification and validation are reported. The numerical solution method is verified by comparing its calculations to the analytic solutions for a series of special cases. Validation of the model is achieved by comparing model predictions to experimental results.

Local equilibrium solutions. When the mechanisms affecting the rates to chemical equilibrium between phases are fast in comparison to chemical movement in the direction of air and water flow, then a condition of local equilibrium is said to exist [*Brusseau and Rao*, 1989]. It was assumed in the model development that three mechanisms affect the time to equilibrium between the air, water, and soil phases: air-water mass transfer, mobile-immobile water mass transfer, and intraaggregate diffusion. Therefore when the magnitudes of St_v, St_h, and Ed_p are large in comparison to Pe_h or Pe_v the model will simulate local equilibrium. Chemical concentrations in the air and on the soil can be determined from the water concentration when local equilibrium exists by the following equilibrium relationships:

$$C_v(Z, T)/H = C_h(Z, T) = C_p(0 \leqq R \leqq R_a, Z, T)$$
$$Q(0 \leqq R \leqq R_a, Z, T) = KC_h(Z, T)^{1/n} \quad (24)$$

Throughout the remainder of this paper, when local equilibrium is assumed then it is also assumed that (24) is satisfied and that $1/n$ is equal to 1.

A mass balance across all phases in a column where local equilibrium exists results in the following expression in terms of dimensionless water concentration:

$$\frac{\partial c_h(z, t)}{\partial t} = \frac{1}{Pe}\frac{\partial^2 c_h(z, t)}{\partial t} = \frac{1}{Pe}\frac{\partial^2 c_h(z, t)}{\partial z^2} - \frac{\partial c_h(z, t)}{\partial z} \quad (25)$$

The three mass balances in the general model reduce to one (equation (25)) because $C_v(Z, T)$, $C_p(R, Z, T)$, and $Q(R, Z, T)$ are determined from (24). The three mass transfer groups that are important are combined into one Peclet number (Pe):

$$Pe = [1 + Ar]\left[\frac{1}{Pe_v} + \frac{1}{Pe_h}\right]^{-1} \quad (26)$$

Pe represents the ratio of mass transport by air and water advection to mass transport by dispersion in both fluids.

Equation (25) is the classical convection-dispersion equation most commonly used in subsurface transport modeling [*van Genuchten and Jury*, 1987]. The assumption of local equilibrium is used here to verify the numerical representation of gas and liquid dispersion in the general model.

Various solutions of (25) can be obtained by changing the boundary conditions. The conditions that are used for the general model assume that the column acts as a closed reactor [*Levenspiel*, 1962] and are used to simplify the numerical solution. Closed reactor boundary conditions for (25) that enable an analytic solution to be obtained are those proposed by *Danckwerts* [1953]:

$$1 - c_b(z = 0^+, t) = -\frac{1}{Pe}\frac{\partial c_b(z = 0^+, t)}{\partial z} \quad (27)$$

$$\frac{\partial c_b(z = 1, t)}{\partial z} = 0 \quad (28)$$

Equation (27) assumes that the influent chemical concentration is constant ($c_{bi}(t) = 1$). The initial condition is (23).

For Pe less than 2 a soil column acts as a completely mixed reactor [*Levenspiel*, 1962], and the solution of (25) approaches

$$c_b(0 \leqq z \leqq 1, t) = 1 - \exp[-t] \quad (29)$$

A breakthrough curve described by (29) represents the maximum observed spreading caused by dispersion.

For Pe greater than 40 the following asymptotic solution is valid [*Hashimoto et al.*, 1964]:

$$c_b(z = 1, t) = \frac{1}{2}\operatorname{erfc}\left[\frac{Pe^{1/2}(1 - t)}{2t^{1/2}}\right] + \left[\frac{t}{\pi Pe}\right]^{1/2}\frac{(t^2 + 4t - 1)}{(t + 1)^3}\exp\left[\frac{-Pe(1 - t)^2}{4t}\right] \quad (30)$$

In general, the solution of (25) constrained by conditions (27) and (28) is [*Hashimoto et al.*, 1964]

$$\begin{aligned} c_b(z = 1, t) = {} & \frac{1}{2}\operatorname{erfc}\left[\frac{Pe^{1/2}(1 - t)}{2t^{1/2}}\right] \\ & - \frac{1}{2}\exp(Pe)\operatorname{erfc}\left[\frac{Pe^{1/2}(1 + t)}{2t^{1/2}}\right] \\ & + 3(Pe\,t)^{1/2}\exp(Pe)\,i\operatorname{erfc}\left[\frac{Pe^{1/2}(1 + t)}{2t^{1/2}}\right] \\ & - 2(Pe\,t)\exp(Pe)\,i^2\operatorname{erfc}\left[\frac{Pe^{1/2}(1 + t)}{2t^{1/2}}\right] \\ & + 4(Pe\,t)^{1/2}\exp(3Pe/2)\,i\operatorname{erfc}\left[\frac{Pe^{1/2}(2 + t)}{2t^{1/2}}\right] \\ & - 16(Pe\,t)\exp(3Pe/2)\,i^2\operatorname{erfc}\left[\frac{Pe^{1/2}(2 + t)}{2t^{1/2}}\right] \\ & + 20(Pe\,t)^{3/2}\exp(3Pe/2)\,i^3\operatorname{erfc}\left[\frac{Pe^{1/2}(2 + t)}{2t^{1/2}}\right] \\ & - 8(Pe\,t)^2\exp(3Pe/2)\,i^4\operatorname{erfc}\left[\frac{Pe^{1/2}(2 + t)}{2t^{1/2}}\right] + \cdots \end{aligned} \quad (31)$$

Danckwerts [1953] solved (25) for open boundary conditions which correspond to a column of infinite length:

$$c_b(z = -\infty, t) = 1 \quad (32)$$

$$c_b(z = \infty, t) = 0 \quad (33)$$

The solution of (25) for these boundary conditions is [*Danckwerts*, 1953]

$$c_b(z = 1, t) = \frac{1}{2}\operatorname{erfc}\left[\frac{Pe^{1/2}(1 - t)}{2t^{1/2}}\right] \quad (34)$$

Plug flow solution. In structured or aggregated soils where the air or water is flowing fast, the spreading that is caused by axial dispersion in air and water could be negligible compared to the spreading caused by other mechanisms. Plug flow is assumed here so that the numerical approximation of intraaggregate diffusion and film transfer could be

tested against an analytic solution in the same manner as dispersion is verified with the local equilibrium solutions.

Rosen [1952] derived an analytic solution for single-phase plug flow through a packed bed. The general model reduces to the equations solved by *Rosen* [1952] if $1/n$ is equal to 1, the air-water mass transfer rate is fast (large St_a), and axial dispersion is slow in comparison to advection (large Pe). For this condition, (13) and (16) combine to give

$$\frac{\partial c_b(z, t)}{\partial t} = \frac{[1 + Dg]}{[1 + Dg_a]}\left[-\frac{\partial c_b(z, t)}{\partial z} + \frac{3St_b}{[1 + Ar]}[c_p(r = 1, z, t) - c_b(z, t)]\right] \quad (35)$$

The boundary condition for (35) is

$$c_b(z = 0, t > 0) = 1 \quad (36)$$

Equations (19)–(22) are used to represent film transfer and intraaggregate diffusion. The initial condition is (23).

The exact solution of (35) is an integral, however, *Rosen* [1954] developed the following asymptotic solution:

$$c_b(z = 1, t) = \frac{1}{2}\operatorname{erfc}\left\{\frac{(1 - t)[1 + Dg]}{2[Dg_p + Dg_s]}\left[\frac{1 + Ar}{15Ed_p} + \frac{1 + Ar}{3St_b}\right]^{-1/2}\right\} \quad (37)$$

Equation (37) is valid for [*Rosen*, 1954]

$$Ed_p/(1 + Ar) \geq 13.33 \quad (38)$$

Numerical solution. The general form of the model (equations (13)–(23)) is solved numerically by converting the partial differential equations (PDEs) to a system of ordinary differential equations (ODEs). Orthogonal collocation (OC), a method of weighted residuals, lends itself well to converting similar types of PDEs to systems of ODEs [*Raghavan and Ruthven*, 1983; *Crittenden et al.*, 1986]. The resulting set of ODEs can then be solved by a number of standard techniques.

Raghavan and Ruthven [1983] used OC to solve equations similar to those comprising the general model except that an inlet condition similar to (27) was used in their formulation. The inlet boundary condition they imposed required an iterative solution method which involved guessing the inlet concentration. The boundary conditions employed here (equations (14), (17), and (22)) avoid the iteratve step. More important is the fact that using (14), (17), and (22) will help conserve chemical mass during the process of numerically solving the model.

Weighted residual methods allow separation of the time and spatial dependency of a PDE by approximating the exact solution with a series of products of time-varying coefficients and spatial basis or trial functions. The collocation method requires that the residual between the numerical approximation of the PDE and its exact value be orthogonal to the Dirac delta function at specified collocation points. This results in the residuals being zero at the collocation points [*Finlayson*, 1980].

Orthogonal collocation uses orthogonal polynomials as basis functions and specifies that the collocation points be located at the basis function roots. The polynomials are constructed orthogonal to each other with respect to a weight function. The weight functions used in the construction of the polynomials for the different equations were chosen to make the numerical solution stable.

Application of OC to the air and mobile water mass balances and their boundary conditions (equations (13)–(18)) yields $2J$ ODEs, where J is the number of axial collocation points. Additional ODEs ($J \times I$, where I is the number of radial collocation points) are produced by the application of OC to the intraaggregate mass balance and its boundary conditions (equations (19), (21), and (22)). Figure 2 is a schematic of the OC descretization of the solution domain and shows the coupling of the ODEs. This system of ODEs is solved using an algorithm called GEAR which can be found in the International Mathematics and Scientific Library (IMSL). The application of OC is shown below in the order in which GEAR receives the derivatives.

The application of OC to (19) results in

$$\frac{dy(i, j, t)}{dt} = \frac{Ed_p[1 + Dg]}{[Dg_p + Dg_s][1 + Ar]}\sum_{n=1}^{I} B_{i,n}^{I}c_p(n, j, t) \quad (39)$$

Figure 2 shows that (39) is evaluated at $I - 1$ radial collocation points at each axial collocation point ($j = 1$ to J). $B_{i,n}^{I}$ is a member of an OC coefficient matrix for spherical geometry that is used to approximate the Laplacian of $c_p(r, z, t)$. This matrix is constructed from a set of symmetric Jacobi polynomials that represent the radial dependence of $c_p(r, z, t)$. The radial orthogonal polynomials are constructed with only even powers of r up to degree $2I$ using a weight function of $1 - r^2$ over the interval of r from 0 to 1. The internal radial collocation locations shown in Figure 2 are the positive roots of the $2(I - 1)$ degree polynomial and lie between 0 and 1. Because the matrix is symmetrical, (21) is satisfied by the application of OC to (19) [*Finlayson*, 1980].

Applying OC to (22) and solving for the change in the total intraaggregate concentration at r equal to 1 leads to the following condition at the aggregate surface:

$$\frac{dy(I, j, t)}{dt} = \frac{1}{W_I^I}\left[\frac{[1 + Dg]St_b}{[Dg_p + Dg_s][1 + Ar]}\cdot[c_b(j, t) - c_p(I, j, t)] - \sum_{p=1}^{I-1} W_p^I\frac{dy(p, j, t)}{dt}\right] \quad (40)$$

Equation (40) is evaluated at all axial collocation locations. W_p^I is a member of a coefficient vector that is used in the quadrature approximation of the radial integrals. W_I^I is nonzero because a weight factor of $1 - r^2$ is used for the construction of the radial basis functions [*Finlayson*, 1980].

The application of OC to the air and mobile water mass balances (equations (13) and (16)) gives the following equations:

$$\frac{dc_a(j, t)}{dt} = \frac{[1 + Dg]}{[1 + Ar]\,Dg_a}\left\{\sum_{m=1}^{I}\left[\frac{B_{j,m}^{I}}{Pe_a} - Ar\,A_{j,m}^{I}\right]c_a(m, t) - 3St_a[c_a(j, t) - c_b(j, t)]\right\} \quad (41)$$

$$\frac{dc_b(j, t)}{dt} = \frac{[1 + Dg]}{[1 + Ar]}\left\{\sum_{m=1}^{I}\left[\frac{B_{j,m}^{I}}{Pe_b} - A_{j,m}^{I}\right]c_b(m, t) + 3St_a[c_a(j, t) - c_b(j, t)] + 3St_b[c_b(j, t) - c_p(I, j, t)]\right\} \quad (42)$$

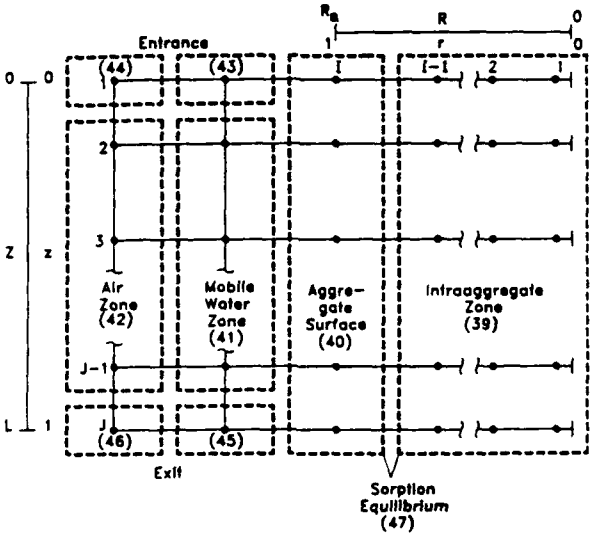


Fig. 2. Coupling of the ordinary differential equations (equation numbers indicated in parentheses) resulting from the orthogonal collocation application to the dimensionless model equations. The locations of the axial collocation points are the roots of $J - 2$ degree polynomial, and the locations of the radial collocation points are the positive roots of the $2(I - 1)$ degree polynomial.

Equations (41) and (42) are evaluated at the $J - 2$ internal axial collocation locations shown in Figure 2 ($j = 2$ to $J - 1$). $A^z_{j,m}$ and $B^z_{j,m}$ are members of OC coefficient matrices for planar geometry that are used to approximate the first and second spatial derivatives, respectively, of $c_b(z, t)$ and $c_v(z, t)$. These matrices are obtained from a set of asymmetric Jacobi polynomials that represent the axial dependence of $c_b(z, t)$ and $c_v(z, t)$. The axial orthogonal polynomials are constructed up to degree $J - 1$ using a weight factor of $z(1 - z)$ over the interval 0 to 1. The locations of the internal axial collocation points shown in Figure 2 are the roots of the $J - 2$ degree orthogonal polynomial. The boundary conditions for (13) and (16) are located at j equal to 1 and J.

Entrance conditions are obtained by using OC to convert (14), (15), (17), and (18) to ODEs and solving for the derivatives at j equal to 1:

$$\frac{dc_v(1, t)}{dt} = \left[W^z_1 - W^z_J \frac{A^z_{J,1}}{A^z_{J,J}}\right]^{-1} \left\{\frac{Ar[1 + Dg]}{Dg_v[1 + Ar]} \cdot \left[[c_w(t) - c_v(J, t)] + \sum_{m=1}^{J} \frac{St_v W^z_m}{Ar}[c_b(m, t) - c_v(m, t)]\right] - \sum_{m=2}^{J-1}\left[W^z_m - W^z_J \frac{A^z_{J,m}}{A^z_{J,J}}\right]\frac{dc_v(m, t)}{dt}\right\} \quad (43)$$

$$\frac{dc_b(1, t)}{dt} = \left[W^z_1 - W^z_J \frac{A^z_{J,1}}{A^z_{J,J}}\right]^{-1} \left\{\frac{[1 + Dg]}{[1 + Ar]} \cdot \left[[c_{bi}(t) - c_b(J, t)] - \sum_{m=1}^{J} St_v W^z_m[c_b(m, t) - c_v(m, t)]\right] - \sum_{m=2}^{J-1}\left[W^z_m - W^z_J \frac{A^z_{J,m}}{A^z_{J,J}}\right]\frac{dc_b(m, t)}{dt} - 3[Dg_p + Dg_s] \sum_{m=1}^{J} W^z_m \sum_{p=1}^{I} W^r_p \gamma(m, p, t)\right\} \quad (44)$$

W^z_m is a member of a coefficient vector that is used in the quadrature approximation of the axial integrals. W^z_1 and W^z_J are nonzero because a weight factor of $z(1 - z)$ is used in the generation of the axial basis functions [*Finlayson*, 1980]. The exit ($j = J$) conditions are then

$$\frac{dc_v(J, t)}{dt} = -\sum_{m=1}^{J-1} \frac{A^z_{J,m}}{A^z_{J,J}} \frac{dc_v(m, t)}{dt} \quad (45)$$

$$\frac{dc_b(J, t)}{dt} = -\sum_{m=1}^{J-1} \frac{A^z_{J,m}}{A^z_{J,J}} \frac{dc_b(m, t)}{dt} \quad (46)$$

TABLE 2. Experimental Conditions for Column Runs With Bromide (Br^-) and Trichloroethene (TCE) in Ottawa Sand (OS) and Verilite (VE)

Column Length, cm	Soil	Chemical	Influent Concentration, $\mu g\ L^{-1}$	Water Flow Rate, $cm^3\ s^{-1}$	Degree of Saturation	Pulse Time, hours	Total Time, hours
29.9	OS	Br^-	45 600	0.075	0.33	2.65	6.47
29.9	OS	TCE	650	0.076	0.33	11.00	23.50
20.1	VE	Br^-	97 600	0.084	0.64	9.68	18.86
20.1	VE	TCE	1 160	0.084	0.64	12.18	22.36

Cross-sectional area of columns is 91.6 cm^2, there is no air flow, and temperature is 22°C.

Evaluations of (39)–(46) are made after solving (20) for $c_p(r, z, t)$ at each of the radial collocation points:

$$y(i, j, t) = \frac{Dg_p c_p(i, j, t) + Dg_z c_p(i, j, t)^{1/n}}{Dg_p + Dg_z} \quad (47)$$

For $1/n$ not equal to 1, values of $c_p(i, j, t)$ are determined with a root finding subroutine called ZBRENT, which also is an IMSL algorithm.

Initially ($t = 0$), (47) is ignored, and (23) is used for concentration values at all of the collocation points:

$$c_r(j, t = 0) = c_h(j, t = 0) = y(i, j, t = 0) = 0 \quad (48)$$

Equation (48) is used in (39)–(46) to calculate the initial derivatives; the derivatives are sent to GEAR, and GEAR returns values of $y(i, j, t)$, $c_h(j, t)$, and $c_r(j, t)$; (47) is solved for $c_p(i, j, t)$; and the algorithm is repeated until the desired throughput is reached.

MATERIALS AND METHODS

Laboratory columns and experimental procedures were designed to measure the breakthrough and elution of trichloroethene (TCE) and bromide (Br^-) from a cohesionless soil and a structured soil. Table 2 lists the column conditions for each experiment. Details of the column design and experimental procedures are reported elsewhere [*Krause*, 1987; *Hutzler et al.*, 1989*b*].

TABLE 3. Properties of Water, Trichloroethene, and Bromide at 22°C Used for Parameter Estimation

	Value
Water	
Viscosity, μ_l ($g\ cm^{-1}\ s^{-1}$)	0.00955[a]
Density, ρ_l ($g\ cm^{-3}$)	0.988[a]
Trichloroethene, TCE (C_2HCl_3)	
Molecular weight, M_A ($g\ mol^{-1}$)	131.3[a,a]
Molar volume, V_A ($cm^3\ mol^{-1}$)	98.1[b]
Boiling point, T_b (K)	360[a]
Henry's constant, H (dimensionless)	0.4[c]
Bromide, Br^- (made from KBr)	
Valence, n_-, n_+	1
Limiting ionic conductance in water at 25°C	
Anion, λ_- (A V g-equiv cm^{-6})	78.3[d]
Cation, λ_+ (A V g-equiv cm^{-6})	73.5[d]

[a] [From *Weast*, 1981].
[b] [*LeBas*, 1915].
[c] [*Ashworth et al.*, 1988].
[d] [*Reid et al.*, 1977]; temperature correction factor, $0.00299T_c/\mu_l$.

Trichloroethene was chosen because it is a common groundwater contaminant of intermediate volatility. Saturated and unsaturated column runs were performed with a bromide tracer to characterize the columns and for estimating certain parameters. Bromide was chosen as the tracer because it is nonadsorbing and nonvolatile, and it can be measured in low concentrations. Chemical properties of Br^- and TCE corresponding to the conditions of the column experiments are given in Table 3.

Ottawa sand (Ottawa, Illinois) was chosen to simulate cohesionless soils, and SCR Veri-lite (Mapleton Development, Incorporated, Minerva, Ohio) was chosen to simulate aggregated soils. Table 4 summarizes the characteristics of each material as packed in the columns. Ottawa sand is a uniform, silica sand containing little or no organic material and thus does not adsorb most organic compounds from aqueous solution. A saturated TCE column run in the sand showed no adsorption of TCE [*Hutzler et al.*, 1989*b*]. SCR Veri-lite (Verilite) is a lightweight, fired clay used mostly in industry as an insulator for steel and iron ladles. The particles are porous and more angular than Ottawa sand. Verilite was chosen because of its availability and low cost. An aqueous isotherm experiment with Verilite showed no adsorption of TCE [*Krause*, 1987].

Hydrodynamic measurements were made with both

TABLE 4. Properties of Porous Media

	Ottawa Sand	Verilite
Properties Measured Directly		
Bulk density, ρ_b, $g\ cm^{-3}$	1.78[a]	0.45[a]
Total porosity, ϵ	0.33[a]	0.70[a]
Microporosity, ϵ_a	0[b], 0.043[c]	0.50[a]
Particle radius[d], cm	0.035[a]	0.035[e]
Hydraulic conductivity, K_s, cm s^{-1}	0.26[f]	0.22[f]
Derived Parameter values		
Solid density, ρ_s, $g\ cm^{-3}$	2.65	1.51
Particle density, ρ_a, $g\ cm^{-3}$	2.65	0.75
Macroporosity, ϵ_m	0.33	0.40
Immobile saturation, S_i	0, 0.10	0.42

[a] Measured gravimetrically [*Black et al.*, 1965].
[b] Assumed.
[c] Value fit to tracer study and close to field capacity measurement.
[d] Aggregate radius (R_a) was assumed to be equal to the particle radius.
[e] Half of geometric mean particle size contained in U.S. standard 20–30 sieves (0.085–0.055 cm).
[f] Determined from slope of specific discharge (v_p) versus headloss per unit length of column.

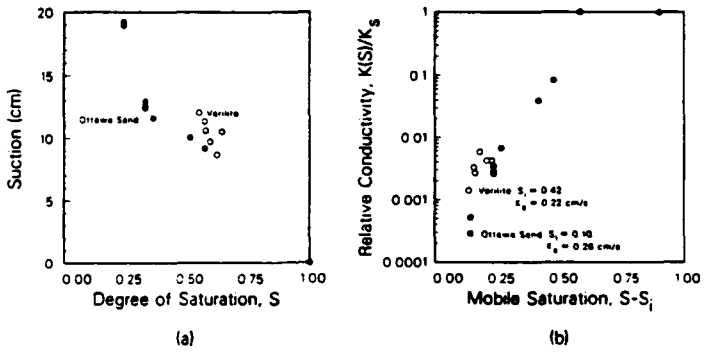


Fig. 3. (*a*) Soil water suction in Ottawa sand and in Verilite as a function of the degree of saturation measured for wetting conditions. (*b*) Relative hydraulic conductivity at 25°C in Ottawa sand and in Verilite as a function of mobile saturation. S, of Ottawa sand was taken as the field capacity; S_i of Verilite was measured gravimetrically.

packed columns. The saturated conductivities (K_s) are given in Table 4. Figure 3*a* shows the relationship between suction and degree of saturation for both media. Figure 3*b* shows the relationship between unsaturated hydraulic conductivity ($K(S)$) relative to K_s and the mobile saturation ($S - S_i$) for both media. The value of S_i used for calculating the mobile saturation in the sand was determined from a tracer study while S_i for the Verilite was measured gravimetrically. The particle sizes of the sand and of the Verilite used in this work were equal, and the interparticle (macro) porosity (ε_m) of the materials when packed in the columns were nearly the same. Therefore the saturated conductivities of the sand and Verilite columns were almost the same. Figure 3*b* shows that the hydrodynamic properties of the two materials under unsaturated flow conditions are also similar. Since both media have similar flow properties, then axial dispersion in both columns should also be similar. Therefore gas and liquid dispersion coefficients measured in one column could be used to predict axial dispersion in the other.

Prior to performing the unsaturated experiments, saturated column experiments were performed with Br^- in both columns. The results were predicted using independently derived dispersion coefficients so it was assumed that apparatus-induced dispersion was negligible [*Hutzler et al.*, 1989*b*].

The experiments were performed in such a manner that the simplifying assumptions made in the model development were satisfied. For example, the model derivation assumed steady state flow and uniform moisture content; therefore influent water was supplied to the tops of the columns at steady rates, and the suction inside the column was monitored along the depth and adjusted by applying a suction at the bottom of the column to achieve a uniform degree of saturation [*Krause*, 1987; *Hutzler et al.*, 1989*b*]. The columns were packed so that no stratification was visible. In addition, ambient temperature was held constant.

PARAMETER ESTIMATION

Parameter values for transport models can be obtained from direct measurement, literature correlations, laboratory experiments, and by fitting model solutions to concentration data that are obtained from a soil column. For a mathematical model to be predictive, however, the parameters must be determined independently of the system being modeled and not by fitting simulations to data. Accordingly, the model parameters should be based on physical properties of the soil and chemical being studied. Tables 2, 3, and 4 list the measured parameter values and the basic chemical and soil properties that were used in this study. The remaining model parameters that were required by the model were determined from these values with the correlations given below and from tracer studies. Table 5 is a list of the parameter values used for the model input.

Gas Dispersion

For low gas velocities (u), gas dispersion in soil will be primarily due to gas diffusion through tortuous air-filled pores. Thus a free-air diffusion coefficient (D_G) can be corrected for the tortuosity of the air-filled pores (τ_a) to obtain the gas dispersion coefficient ($E_v = D_G\tau_a^{-1}$). A tortuosity correction can be used when u is less than about $[1 - \varepsilon(1 - S)]D_G[2R_a\varepsilon(1 - s)]^{-1}$ [*Miyauchi and Kikuchi*, 1975]. In this work the Wilke-Lee modification [*Wilke and Lee*, 1955] of the Hirschfelder-Bird-Spotz method for calculating diffusivities of nonpolar organics is used to calculate D_G in air:

$$D_G = \frac{[4.336 - (0.0345 + M_A^{-1})^{0.5}]T_c^{1.5}(0.0345 + M_A^{-1})^{0.5}}{P_t(0.118V_A^{0.33} + 0.371)^2 f(0.1025T_cT_b^{-0.5})} \tag{49}$$

The value of the collision function for diffusion ($f(0.1025T_cT_b^{-0.5})$) was obtained from a polynomial fit to a graph found by *Treybal* [1980].

Many correlations exist for determining the tortuosity of the air-filled pores [see *Roy and Griffin*, 1987]. A relationship adapted from *Millington* [1959] was used in this work:

$$\tau_a = \varepsilon^2[\varepsilon(1 - S)]^{-7/3} \tag{50}$$

TABLE 5. Parameter Values for Model Calculations in Figures 5–8

	Chemical/Soil				
	Br^-/OS	TCE/OS	Br^-/VE	TCE/VE	Estimation Method[a]
			Independently Determined Values		
S_i	0[b]	0.10[c]	0.42	0.42	measured gravimetrically
S	0.33	0.34	0.64	0.64	measured gravimetrically
v, cm s^{-1}	0.0075	0.010	0.0060	0.0060	calculated from definition
v_a, cm s^{-1}	0.0075	0.0074	0.0020	0.0020	calculated from definition
H	NA	0.4	NA	0.4	*Ashworth et al.* [1988]
k_f, cm s^{-1}	NA	0.0056	0.0030	0.0018	equation (54)
$K_L a$, s^{-1}	NA	6.4 (10^{-4})	NA	6.6 (10^{-4})	equation (53)
D_l cm^2 s^{-1}	2.0 (10^{-5})	9.4 (10^{-6})	2.0 (10^{-5})	9.4 (10^{-6})	equation (52) for Br^-; (51) for TCE
τ_p	NA	NA	90	90	measured in batch experiment
D_p, cm^2 s^{-1}	NA	NA	2.2 (10^{-7})	1.0 (10^{-7})	$D_l \tau_p^{-1}$
D_G, cm^2 s^{-1}	NA	0.088	NA	0.088	equation (49)
τ_a	NA	3.7	NA	4.1[d]	equation (50)
E_v, cm^2 s^{-1}	NA	0.024	NA	0.021	$D_G \tau_a^{-1}$
E_z cm^2 s^{-1}	1.7 (10^{-3})		5.0 (10^{-4})		equation (55)
	1.7 (10^{-4})		1.4 (10^{-4})		equation (57)
	3.7 (10^{-1})	0.10[c]	0.060[e]	0.020[f]	equation (56)
			Values Resulting From Model Calibration		
S_i	0.10	NC	NC	NC	fit
v, cm s^{-1}	0.011	NC	NC	NC	calculated from definition
H	NC	0.7	NC	0.7	fit
k_f, cm s^{-1}	0.0056	NC	NC	NC	equation (56)
$D_p R_a^{-2}$ s^{-1}	1.6 (10^{-6})	8.0 (10^{-7})	NC	NC	fit Br^-, divided value for Br^- by 2
E_v, cm^2 s^{-1}	NC	*<0.0024*	NC	*<0.0028*	adjusted to make unimportant
E_z, cm^2 s^{-1}	0.10	NC	0.020	NC	fit

NA, not applicable; NC, no change. The notation 6.4 (10^{-4}) means 6.4 × 10^{-4}.
[a]Unless otherwise noted.
[b]Assumed.
[c]Used value fit to Br^- in OS experiment.
[d]ϵ_m^2 was substituted for ϵ^2.
[e]Multiplied E_z fit for Br^- in OS by ratio of v_{VE} to v_{OS}.
[f]Used value fit from Br^- in VE experiment.

The air-filled porosity term in (50) is raised to the −7/3 power instead of −10/3, as reported by *Millington* [1959], because Millington included the area available for gas diffusion in the determination of the effective diffusion coefficient, while in this work it is separated from E_v. Because *Millington* [1959] studied diffusion through cohesionless soils, mobile porosity (ϵ_m) is substituted for total porosity in ϵ^2 in (50) for estimating τ_a in the Verilite.

Diffusion in Pores Containing Immobile Water

The description of liquid diffusion in intraaggregate pores must account for the tortuous paths that molecules travel around soil particles that form an aggregate ($D_p = D_l \tau_p^{-1}$). Many correlations exist for estimating liquid diffusion coefficients (D_l). For TCE the Hayduk-Laudie correlation given by *Sherwood et al.* [1975] was used:

$$D_l = 6.96(10^{-7})\mu_l^{-1.14} V_a^{-0.589} \quad (51)$$

For Br^- the Nernst-Haskell equation [*Reid et al.*, 1977] was used to calculate D_l:

$$D_l = 8.931(10^{-10}) T_e \frac{1/n_+ + 1/n_-}{1/\lambda_+ + 1/\lambda_-} \quad (52)$$

Internal pore tortuosity (τ_p) is a function of the pore shape and the amount of immobile water. The immobile degree of saturation in the sand ($S_i = 0.10$) and the specific intraaggregate diffusion rate ($D_p R_a^{-2} = 1.6(10^{-6})$ s^{-1}) was determined from an unsaturated bromide column run because these parameters could not be measured directly. The values fit to the unsaturated bromide data were used to predict the movement of TCE in the sand. The amount of immobile water inside the Verilite particles ($S_i = 0.42$) was measured gravimetrically, and the tortuosity of the internal pores ($\tau_p = 90$) was measured in a batch study [*Hutzler and Gierke*, 1988]. The batch experiment was similar to those performed by *Rao et al.* [1982]. The τ_p of 90 used in this work is large compared to values between 2 and 10 used by most researchers [*Roberts et al.*, 1987], however, when the Verilite was viewed under electron scanning microscopy (Air Force Engineering and Services Center, Tyndall Air Force Base, Florida) it was observed that the particle surface was impervious except at a relatively few number of locations. Electron microscopic photographs of the internal pores showed that many were not connected. For these reasons the measured τ_p is reasonable. In addition, a saturated Br^- column run in Verilite could be predicted using the value of τ_p measured in the batch experiment [*Hutzler et al.*, 1989b].

Air-Water Mass Transfer

The air-water mass transfer coefficient (K_L) and the specific air-water interfacial area (a) have not been studied in unsaturated soils. Many correlations do exist, however, for these parameters in the analysis of packed-tower operation such as air stripping and the performance of trickle bed

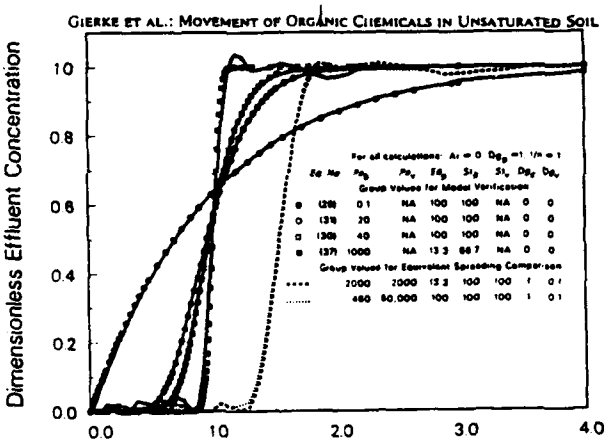


Fig. 4. Verification of ... with 4 analytical solutions (symbols) for advection-dispersion dominant and advection-intraaggregate diffusion film transfer dominant conditions. Comparison of equivalent spreading relationship (equation (58)) for intraaggregate diffusion dominant (dashed curve) and axial dispersion dominant (dotted curve) conditions.

reactors. *Turek and Lange* [1981] developed a correlation for $K_L a$ in low-velocity trickle bed reactors:

$$K_L a \approx 16.8 D_l \left[\frac{8R_g^3 g \rho_l^2}{\mu_l}\right]^{-0.22}\left[\frac{2R_g v \varepsilon(S-S_r)\rho_l}{\mu_l}\right]^{0.25}\left[\frac{\mu_l}{\rho_l D_l}\right]^{0.5} \quad (53)$$

Equation (53) is valid for values of R_g between 0.028 and 0.15 cm, $\varepsilon(S - S_r)$ between 0.05 and 0.3, and Re between 0.1 and 5. The experiments reported in this work satisfy all but the Reynolds number requirements. For the experiments reported herein Re was less than 0.007.

Mobile-Immobile Water Mass Transfer

Mass transfer rates across the mobile-immobile water interface were found to be fast in saturated soil systems [*Crittenden et al.*, 1986; *Hutzler et al.*, 1986; *Roberts et al.*, 1987] and are expected to be fast in unsaturated systems. It was included in this modeling effort to test this hypothesis. A correlation by *Wilson and Geankoplis* [1966] for saturated systems was adapted to estimate k_f:

$$k_f = 1.09\varepsilon\{2R_g v \varepsilon(S - S_r)/D_l\}^{-2/3} \quad (54)$$

Equation (54) is valid for values of Re between 0.0016 and 55 and $\varepsilon(S - S_r)$ between 0.35 and 0.75.

Liquid Dispersion

Liquid dispersion has been studied in soils more often than any mass transport mechanism other than advection, at least in saturated media. However, there is still a lack of accurate correlations for predicting liquid dispersion coefficients (E_z) in unsaturated media. *Yule and Gardner* [1978] fit the following relationship for E_z in unsaturated sand columns:

$$E_z = 5.33(10^{-5}) + 0.216 v_p \quad (55)$$

Equation (55) was fit to data obtained for average pore velocities (v_p) between $1.7(10^{-4})$ and 0.0043 cm s^{-1} and degrees of saturation between 0.34 and 0.76.

DeSmedt and Wierenga [1984] propose the following relationship for observed dispersion in unsaturated columns of glass beads:

TABLE 6. Dimensionless Group Values Resulting From Parameter Values Listed in Table 5

	Chemical/Soil			
	Br^-/OS	TCE/OS	Br^-/VE	TCE/VE
	Independently Determined Values			
Dg_p	0	0.42	1.9	1.9
Dg_v	NA	1.1	NA	0.66
R_d	1	1.8	1	1.2
Ed_p	NA	$9.0\ (10^{-4})$	1.1	0.52
St_h	NA	260	1100	660
St_v	NA	0.66	NA	4.8
Pe_v	NA	12	NA	8.7
Pe_h	61[a]	3.1	860[b]	6.0
Pe	Pe_h	2.4	Pe_h	3.6
	Values Resulting From Model Calibration			
Dg_p	0.44[c]	NC	NC	NC
Dg_v	NC	1.9[c]	NC	1.2[c]
R_d	NC	2.3[d]	NC	1.4[d]
Ed_p	$1.9\ (10^{-3})$[c]	NC	NC	NC
St_h	3900[d]	NC	NC	NC
Pe_v	NC	>67[e]	NC	>50[e]
Pe_h	3.2[c]	NC	6.0[c]	NC

$Ar = 0$, $Dg_s = 0$, and $1/n = 1$ for all calculations.
NA, not applicable; NC, no change.
[a]Determined using E_z from (56).
[b]Determined using E_z from (57).
[c]Fit.
[d]Calculated from fit parameters.
[e]Increased to decrease impact.

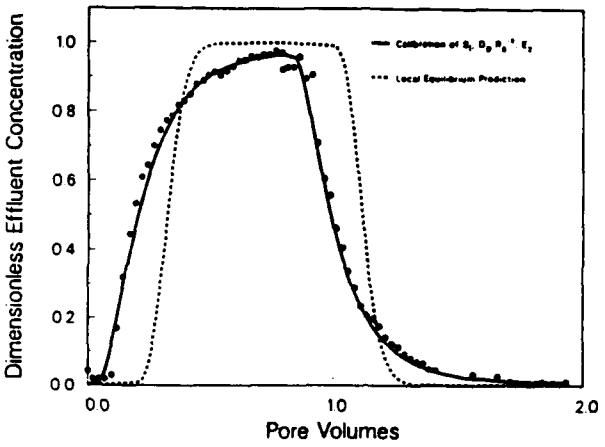


Fig. 5. Comparison of local equilibrium model prediction based on parameter values obtained from the literature (dashed curve) and numerical model calibration of S_r, $D_p R_d^{-2}$, and E_z (solid curve) to experimentally measured aqueous concentrations of bromide in effluent from unsaturated Ottawa sand column (dots).

$$E_z = \frac{1.18(10^{-5}) + 0.021v_p + 1900v_p^2}{1.88 + 3630v_p} \quad (56)$$

Values of E_z given by (56) include contributions from axial dispersion and diffusion in immobile water. *DeSmedt and Wierenga* [1984] developed (56) by assuming that the contribution of axial dispersion is given by

$$E_z = 1.4(10^{-5}) + 0.021v \quad (57)$$

DeSmedt and Wierenga [1984] used a first-order exchange model to describe the transfer of solute between mobile and immobile water, while in this work the rate of transfer through immobile water by diffusion is also considered.

Results and Discussion

A complete model study consists of verification and validation steps. Verification of the numerical solution was performed by comparing model calculations to the analytical solutions presented above for simplified conditions. Unsaturated miscible displacement experiments were performed in order to validate the model and to determine its ability to describe chemical transport in cohesionless soils, such as sands, and in structured or aggregated soils.

Model Verification of the Numerical Solution

The general model numerically approximates advection, axial dispersion, diffusion in immobile water, film transfer, and air-water mass transfer. To verify the numerical method, model calculations were compared to analytical solutions that account for one or more of these mechanisms. The numerical approximation of advection and axial dispersion was compared to the local equilibrium solutions. The numerical approximation of advection, air-water mass transfer, film transfer, and diffusion in pores containing immobile water was compared to the plug flow solution.

Model input parameters for the verification step were chosen to satisfy assumptions that were made to obtain a particular analytical solution. To simulate local equilibrium, large values of Ed_p, St_b, and St_v were input to the numerical solution. To simulate plug flow, large values of Pe_b and Pe_v were input instead.

Figure 4 compares the breakthrough curves that were calculated with the numerical model to analytical solutions for different values of the dimensionless groups. These breakthrough curves are plots of the column effluent water concentration relative to the influent versus the number of pore volumes of water fed. For the calculations shown in Figure 4 it was assumed that $1/n$ was equal to 1 and influent concentrations were constant. Numerical solutions, shown as solid curves, using J (axial) equal to 10 and I (radial) equal to 3 collocation points agree with the analytic solutions (equations (29), (30), (31), and (37)), shown as symbols, for the dimensionless group values listed on Figure 4. Ten axial and 3 radial collocation points were also used for the model validation calculations.

Additional simulations were performed with the model to observe the impact of Ed_p, Pe, St_b, and St_v on the numerical solution. Dispersion calculations by the model are accurate for values of Pe from 0.1 to 40 and greater; however, values of Pe greater than 1000 cause significant numerical error. For low values of Pe the observed spreading causes the breakthrough curve to become asymmetric, and the model can simulate this. As the value of Ed_p or St_b or St_v decrease below 1, the early portion of the breakthrough curve sharpens and the latter part tails.

By examining the shapes of the breakthrough curves depicted in Figure 4 it is seen that the mass transfer mechanisms can have similar impacts on chemical transport. A relationship for equivalent spreading between axial dispersion, diffusion in immobile water, and film transfer can be developed by equating the arguments of (34) and (37).

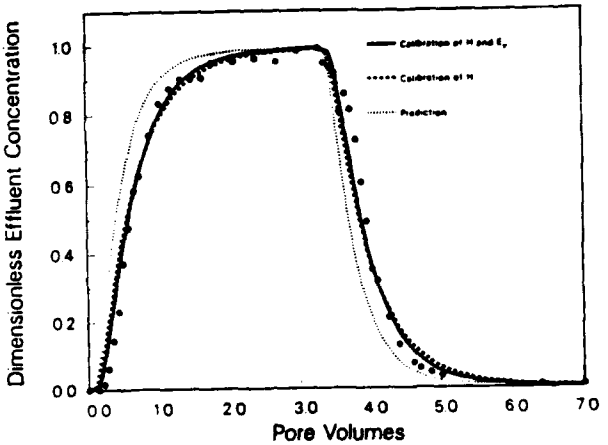


Fig. 6. Comparison of numerical model prediction based on parameter values obtained from the literature and S_t, $D_p R_d^{-1}$, and E_z from the Br^- in Ottawa sand experiment (dotted curve), calibration of H (dashed curve), and calibration of H and E_z (solid curve) to experimentally measured aqueous concentrations of TCE in effluent from unsaturated Ottawa sand column (dots).

Equations (30), (31), and (34) give comparable results for large Pe. If only the central portion of the breakthrough curve is considered (i.e., t is near 1), the following relationship can be derived:

$$Pe = \frac{15Ed_p[1 + Dg]^2}{[1 + Ar][Dg_p + Dg_v]^2} = \frac{3St_h[1 + Dg]^2}{[1 + Ar][Dg_p + Dg_v]^2} \quad (58)$$

Equation (58) is similar to the equivalent spreading relationship for saturated flow developed by *Crittenden et al.* [1986]. When Ar is equal to 0, then (58) is also equivalent to $3St_v[1 + Dg]^2 Dg_v^{-2}$. The model was used to calculate the dotted and dashed curves shown in Figure 4 to simulate conditions where axial dispersion and diffusion in immobile water have equivalent impacts on a breakthrough curve. These curves were calculated for dimensionless group values that were obtained according to (58). Identical results were obtained with the numerical model for film transfer predominant and for air-water mass transfer predominant spreading. These conditions corresponded to a value of St_h equal to 66.7 (Ed_p = 100, Pe = 1000, St_v = 100) and a value of St_v equal to 0.16 (Ed_p = 100, Pe = 1000, St_h = 100), respectively. It is evident that it is not always possible to distinguish between the impacts of different mechanisms by fitting model solutions to data [*Roberts et al.*, 1987; *Brusseau and Rao*, 1989]. Equation (58) can be used to compare the relative contributions of axial dispersion, intraaggregate diffusion, and film transfer on the observed spreading of a chemical front and to determine which mechanisms are important for different conditions.

Model Validation

Column experiments were performed to show that the model described above is able to predict the breakthrough and elution of volatile organic chemicals from unsaturated soil columns under controlled conditions. The experimental procedure was designed to show that the model is versatile enough to simulate chemical movement in different types of porous media as well as under different flow and moisture conditions. Table 4 summarizes the column conditions for each run. Tables 2–5 list parameters values used for the model calculations. Table 6 is a list of the magnitudes of the corresponding dimensionless groups. Model predictions were obtained with the parameter and group values in the upper portion of Tables 5 and 6. These values were independently determined. Values used in model simulations and fits were taken from the lower portion of Tables 5 and 6.

The model developed herein was intended for describing one-dimensional movement of VOCs in unsaturated columns of soil. The numerical solution of the general model is used to simulate movement of TCE with unsaturated water flow. The general model was altered [*Hutzler et al.*, 1989*b*] to ignore vapor movement for simulating the Br^- column experiments.

Column experiments with Ottawa sand. Bromide and TCE experiments were run on a column packed with unsaturated Ottawa sand. A 45.6 mg L^{-1} Br^- solution was fed to the top of the column at a rate of 0.075 $cm^3\ s^{-1}$ (v_p = 0.0075 $cm\ s^{-1}$) for 2.65 hours. Clean water was then applied at the same rate to elute the bromide from the column. The degree of saturation was 0.33. Figure 5 compares the data to model calculations. Because the sand particles are solid and uniformly sized, it was first assumed that the amount of immobile water was negligible, such as in the saturated runs. Accordingly, (31) was used to calculate the dashed curve shown in Figure 5 (Pe = 61) by using an E_z of $3.7(10^{-3})\ cm^2\ s^{-1}$ which was estimated with (56). This value was larger than that which was estimated by (55).

The breakthrough data is shifted to the left of the dispersion equation prediction, and this is attributed to the presence of immobile water. This shift could not be simulated with (31). Five parameter values were not known with

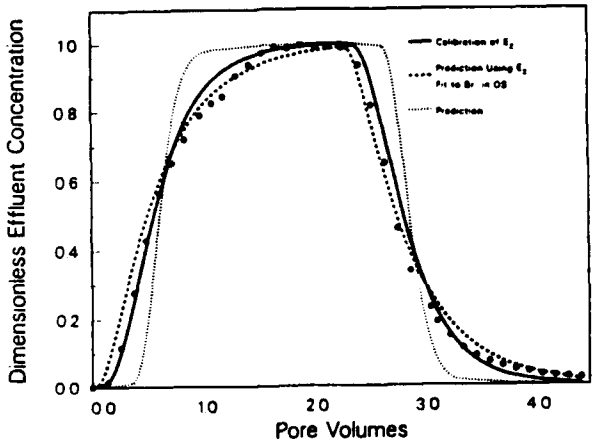


Fig. 7. Comparison of numerical model prediction based on parameter values obtained from the literature and a batch rate study (dotted curve), prediction based on E_z fit from the Br^- in Ottawa sand experiment (dashed curve), and calibration of E_z (solid curve) to experimentally measured aqueous concentrations of bromide in effluent from unsaturated Verilite column (dots).

certainty (S_i, D_p, R_a, k_f, E_z). The numerical model simulated the data by adjusting only three dimensionless groups (Dg_p, Ed_p, Pe_b). Either Ed_p or St_b could have been reduced so that the breakthrough curve would shift to the left. Since film transfer was unimportant in saturated soils [*Hutzler et al.*, 1986; *Roberts et al.*, 1987], Ed_p was fit. When either Ed_p, St_b, or St_v is small, the early portion of the breakthrough curve will be sharp and appear sooner than R_dS pore volumes. For this Br^- run ($R_d = 1$), the early portion of the breakthrough will be located approximately at $S(1 + Dg_p)^{-1}$ pore volumes. Hence Dg_p was increased to simulate the first sharp increase in Br^- concentration. Reducing Ed_p did not result in enough spreading to simulate the data, therefore Pe_b had to be reduced too. A value of 0.44 for Dg_p, 0.0019 for Ed_p, and 3.2 for Pe_b best describe the data. These group values correspond to an S_i of 0.10, $D_pR_a^{-2}$ equal to $1.6(10^{-6})$ s^{-1}, and an E_z of 0.10 $cm^2\ s^{-1}$. The fit value of S_i is close to the degree of saturation of the Ottawa sand column after it has drained freely by gravity but is about twice as much as that observed by *DeSmedt and Wierenga* [1984] in unsaturated columns of glass beads. Intraaggregate diffusion was fit as the ratio $D_pR_a^{-2}$ to keep the uncertainty in one term. The low value of Ed_p is possibly due to channeling inside the column or an increased moisture content at the bottom of the column. The E_z fit to this data is larger than that observed in other unsaturated studies [*Yule and Gardner*, 1978; *DeSmedt and Wierenga*, 1984]. The large amount of dispersion could also be due to channeling or an increase in moisture content at the bottom of the column. A rapid rate of film transfer was indicated by the large value of St_b (3900).

A TCE solution of 650 $\mu g\ L^{-1}$ was then fed to the sand column at a degree of saturation of 0.34 over a period of 11 hours at a rate of 0.076 $cm^3\ s^{-1}$ ($v_p = 0.0074\ cm\ s^{-1}$). Figure 6 shows a comparison of the TCE data to calculations of the general model. Even after using the results of the unsaturated Br^- run, there are still two parameters (H, E_v), which are needed for predicting the TCE results, that have uncertain values. For the calculations shown the E_z and S_i fit in the unsaturated Br^- experiment were used. The ratio of $D_pR_a^{-2}$ was reduced by a factor of two because (52) predicts that Br^- diffuses approximately twice as fast as TCE (equation (51)). The model prediction is shown as a dotted curve, and the corresponding magnitudes of the dimensionless groups are listed in Table 6. The breakthrough curve is shifted to the right of the prediction, indicating that the description of TCE equilibrium in the sand was incorrect. This could be due to either H for TCE being higher than predicted or sorption of TCE vapors onto the drier particle or column surfaces. TCE did not adsorb to the sand or the column in saturated experiments. Sorption capacity (K) could be adjusted in an attempt to account for vapor adsorption, but because Ed_p was small, the model calculations would exhibit more asymmetry as K is increased. Adjusting H, however, produced a curve that simulated the data. A slightly better fit of the data was obtained by reducing the gas dispersion coefficient (E_v) by a factor of 10 ($Pe_v = 67$), that is, making gas dispersion unimportant. It was concluded from a comparison of the magnitudes of the dimensionless groups that the rates of volatilization and film transfer had little impact on the observed spreading.

A slower water flow ($S = 0.30$, $v_p = 0.00028\ cm\ s^{-1}$) TCE experiment in the Ottawa sand column used here is reported by *Hutzler et al.* [1989b]. Because gas diffusion was predominant for the slow flow conditions, their experimental results could be simulated with (29). However, H was increased to 1.0 to fit the data.

Column Experiments with Verilite. Most natural soils are not as uniform as Ottawa sand. Instead, they exhibit some structure and a pore size distribution. To simulate a structured or aggregated system, uniformly sized Verilite was used as a packing material. The particle size, hydraulic conductivity, and macroporosity of Verilite and Ottawa sand

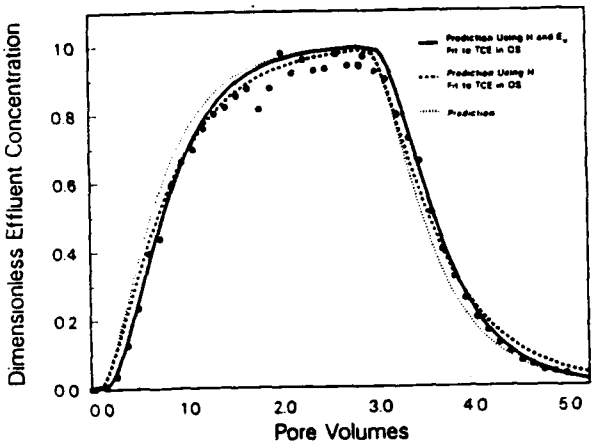


Fig. 8. Comparison of numerical model prediction based on parameter values obtained from the literature, a batch rate study, and E_z fit from the Br^- in Verilite experiment (dotted curve), prediction based on H fit from the TCE in Ottawa sand experiment (dashed curve), and prediction based on H and E_z fit from the TCE in Ottawa sand experiment (solid curve) to experimentally measured aqueous concentrations of TCE in effluent from unsaturated Verilite column (dots).

are similar. Verilite particles contain a microporosity or internal porosity inside which the flow of water is negligible in comparison to the flow around the particles.

The Verilite column was drained to a degree of saturation of 0.64, which gave about the same air-filled porosity as the Ottawa sand column. A 97.6 mg L^{-1} Br^- solution was then applied to the top of the column at a rate of 0.084 $cm^3 s^{-1}$ (v_p = 0.0020 cm s^{-1}) for 9.68 hours. Figure 7 shows the effluent data and three model calculations. Because S_i and D_p were measured independently, only the magnitude of E_z was uncertain for this experiment. The dotted curve is a prediction using the D_p measured in the batch experiment and E_z predicted by (57) (Pe_b = 860). The dashed curve was obtained by multiplying the value of E_z fit to the unsaturated run of Br^- in the sand by the ratio of the interstitial water velocity in the Verilite to that in the sand (Pe_b = 2.0). Because this prediction is close, it adds confidence to the fit E_z for the unsaturated sand run. The solid curve was obtained by adjusting E_z to fit the data (Pe_b = 6.0).

After eluting the Br^- from the unsaturated Verilite a 1160 $\mu g\ L^{-1}$ solution of TCE was fed at the same rate (v_p = 0.0020 cm s^{-1}) for 12.18 hours. The breakthrough and elution of TCE from the unsaturated Verilite column is displayed in Figure 8. As was the case for estimating the parameters for the TCE run in the sand, H and E_z were not known with certainty for this experiment. A model prediction using the E_z fit for the unsaturated Br^- run in Verilite (Pe_b = 6.0) and an H of 0.7 (Dg_v = 1.2; R_d = 1.4), which was used to described the TCE movement through the unsaturated sand, is shown in Figure 8 as a dashed curve. A slightly better fit of the data was obtained by reducing the estimated gas dispersion coefficient by a factor of 10 (Pe_v = 50), as was done in the description of TCE diffusion through the sand. The air-filled porosity of the Verilite was equal to 0.25, which was a little larger than that of the sand (0.21). Gas diffusion was not an important transport mechanism in the Verilite (Pe_v = 50) in comparison to liquid dispersion (Pe_b = 2.0) and intraaggregate diffusion (Ed_p = 0.52). Again, film transfer (St_b = 660) and the air-water mass transfer resistances (St_v = 4.8) were unimportant. The dotted curve was calculated using a literature value of H of 0.4 (Dg_v = 0.68; R_d = 1.1). It is not known at this time whether TCE vapors are sorbing to the column or soil materials or whether Henry's constant is, indeed, higher in unsaturated soils.

CONCLUSIONS

The development of a mathematical model that simulates chemical fate in soil requires a fundamental understanding of transport and attenuation processes. The current level of understanding of the processes affecting nondegradable chemical movement in unsaturated soil is enhanced with a combination of theoretical model development and a series of laboratory column experiments. A one-dimensional model accounting for air and water dispersion in the direction of flow and diffusion in immobile water is capable of simulating the movement of aqueous solutions of trichloroethene in laboratory columns packed with homogeneous soils. Laboratory results and numerical calculations indicate the following about the mechanisms affecting the transport of nondegradable VOCs in homogeneous, unsaturated soils: (1) both liquid dispersion and diffusion in immobile water are important, (2) vapor diffusion is not an important transport mechanism for trichloroethene in sands or structured soils when the average pore water velocities are greater than about 0.07 and 0.02 cm s^{-1}, respectively, but is important in sand for v_p less than about 0.0003 cm s^{-1}, and (3) the rates of mass transfer across the air-water and the mobile-immobile water interfaces are fast. In addition, it was also found for unsaturated conditions that Henry's partitioning for trichloro-

ethene was almost twice as large as reported in the literature, or there was adsorption of vapors onto the packing or column materials. If vapor sorption does occur in unsaturated soil, then a model would have to include air-soil partitioning. The rate of internal diffusion and the amount of immobile water in the sand could not be independently determined. A batch rate study was successful at measuring the intraaggregate diffusion rate in Verilite. These results indicate a need for more studies of liquid dispersion with unsaturated flow and diffusion through immobile water filled pores. In addition, measurements of chemical partitioning between phases in an unsaturated soil system are needed.

NOTATION

a specific interfacial area between the air and water (L^{-1}).

Ar advective flux ratio, ratio of chemical mass transport rate by advection in air to that by advection in water, equal to $uDg_v v^{-1}$ (dimensionless).

A^x_{ij} i, j member of the OC coefficient matrix that is used for approximating the first axial derivative of $c_b(z, t)$ and $c_v(z, t)$ (dimensionless).

B^r_{ij} i, j member of the OC coefficient matrix that is used for approximating the laplacian of $c_p(r, z, t)$ (dimensionless).

B^z_{ij} i, j member of the OC coefficient matrix that is used for approximating the second axial derivative of $c_b(z, t)$ and $c_v(z, t)$ (dimensionless).

$C_b(Z, T)$ mobile water phase chemical concentration as a function of axial position and time (M L^{-3}).

$c_b(z, t) = C_b(z, t)C_{bn}^{-1}$ (dimensionless).

$C_{bn}(T)$ influent chemical concentration in water as a function of time (M L^{-3}).

$c_{bn}(t) = C_{bn}(t)C_{\sim}^{-1}$ (dimensionless).

C_{bn} normalizing water concentration, equal to average influent concentration measured during breakthrough (M L^{-3}).

$C_p(R, Z, T)$ immobile water phase chemical concentration as a function of radial position, axial position, and time (M L^{-3}).

$c_p(r, z, t)$ $C_p(r, z, t)C_{bn}^{-1}$ (dimensionless).

$C_v(Z, T)$ air-phase chemical concentration as a function of axial position and time (M L^{-3}).

$c_v(z, t) = C_v(z, t)(HC_{bn})^{-1}$ (dimensionless).

$C_{vn}(T)$ influent chemical concentration in air as a function of time (M L^{-3}).

$c_{vn}(t) = C_{vn}(t)(HC_{bn})^{-1}$ (dimensionless).

Dg total solute distribution parameter, ratio of chemical mass contained in immobile water, on soil, and in air to that in mobile water at equilibrium with C_{bn}, equal to $Dg_v + Dg_p + Dg_s$ (dimensionless).

D_G gas diffusion coefficient ($L^2 T^{-1}$).

Dg_p immobile water solute distribution parameter, ratio of chemical mass contained in immobile water to that in mobile water at equilibrium with C_{bn}, equal to $S_i(S - S_i)^{-1}$ (dimensionless).

Dg_s adsorbed solute distribution parameter, ratio of chemical mass adsorbed onto soil particle surfaces to that contained in mobile water at equilibrium with C_{bn}, equal to $\rho_s(1 - \epsilon)KC_{bn}^{1/n-1}[\epsilon(S - S_i)]^{-1}$ (dimensionless).

Dg_v vapor solute distribution parameter, ratio of chemical mass contained in air-filled pores to that in mobile water at equilibrium C_{bn}, equal to $(1 - S)H(S - S_i)^{-1}$ (dimensionless).

D_l liquid diffusion coefficient ($L^2 T^{-1}$).

D_p intraaggregate liquid diffusion coefficient ($L^2 T^{-1}$).

Ed_p intraaggregate diffusion modulus, ratio of chemical mass transfer rate by diffusion in immobile water to that by advection in mobile water, equal to $LD_pDg_p(vR_a^2)^{-1}$ (dimensionless).

E_v axial dispersion or diffusion coefficient in air ($L^2 T^{-1}$).

E_z combined axial dispersion and diffusion coefficient in mobile water ($L^2 T^{-1}$).

$f(\ )$ collision function for diffusion as a function of temperature and energy of molecular attraction (dimensionless).

g gravitational acceleration ($L T^{-2}$).

H Henry's air-water partitioning coefficient (dimensionless).

I number of radial collocation points.

J number of axial collocation points.

K soil sorption capacity ($[L^3/M]^{1/n} M M^{-1}$).

$K(S)$ unsaturated hydraulic conductivity as a function of degree of saturation ($L T^{-1}$).

k_f film transfer coefficient ($L T^{-1}$).

K_L overall mass transfer coefficient between the air and the mobile water ($L T^{-1}$).

K_s saturated hydraulic conductivity ($L T^{-1}$).

L column length (L).

M_A molecular weight ($M mol^{-1}$).

n_+ valence of cation (dimensionless).

n_- valence of anion (dimensionless).

$1/n$ adsorption intensity (dimensionless).

$Pe = (1 + Ar)\{Pe_b^{-1} + Pe_v^{-1}\}^{-1}$ (dimensionless).

Pe_b Peclet number of mobile water, ratio of chemical mass transfer rate by advection in mobile water to that by dispersion in mobile water, equal to vL/E_z (dimensionless).

Pe_v Peclet number for air, ratio of chemical mass transfer rate by advection in mobile water to that by diffusion in air, equal to $vL[E_vDg_v]^{-1}$ (dimensionless).

P_t atmospheric pressure ($M L^{-1} T^{-2}$).

$Q(R, Z, T)$ adsorbed-phase chemical concentration as a function of radial position, axial position, and time (M/M).

Q_G volumetric air flow rate ($L^3 T^{-1}$).

Q_L volumetric water flow rate ($L^3 T^{-1}$).

$r = RR_a^{-1}$ (dimensionless).

R radial coordinate (L).

R_a radius of aggregated particle (L).

R_d retardation coefficient, equal to $(1 + Dg)\{(1 + Dg_p)(1 + Ar)\}^{-1}$ (dimensionless).

Re liquid Reynolds number based on approach

$\tau^{-1/N}$ velocity, equal to $2R_a\rho_1 v\varepsilon(S - S_i)\mu_1^{-1}$ (dimensionless).
S degree of saturation, relative volume of the pores filled with water (dimensionless).
S_i immobile degree of saturation, relative volume of the pores filled with immobile water (dimensionless).
St_b Stanton number for film transfer, ratio of chemical mass transfer rate by film transfer to the transport rate by advection in mobile water, equal to $k_f[1 - \varepsilon(1 - S_i)]L[v\varepsilon(S - S_i)R_a]^{-1}$ (dimensionless).
St_v Stanton number for volatilization, ratio of chemical mass transfer rate across air-water interface to the transport rate by advection in mobile water, equal to $K_L aL[3v\varepsilon(S - S_i)]^{-1}$ (dimensionless).
t throughput, ratio of chemical mass fed to a column to the mass of chemical in the column at equilibrium with the influent, proportional to time by $(1 + Ar)v[(1 + Dg)L]^{-1}$ (dimensionless).
T time (T).
T_e temperature (°K).
T_b boiling point (°K).
u interstitial air velocity, equal to $Q_G[A\varepsilon(1 - S)]^{-1}$ (L T^{-1}).
v interstitial water velocity, equal to $Q_L[A\varepsilon(S - S_i)]^{-1}$ (L T^{-1}).
V_a molar volume (L^3 mol^{-1}).
v_p average pore water velocity, equal to $Q_L(A\varepsilon S)^{-1}$ (L T^{-1}).
W_i^r ith component of the weight vector used in the quadrature approximation of the radial integrals (dimensionless).
W_i^z ith component of the weight vector used in the quadrature approximation of the axial integrals (dimensionless).
$Y(R, Z, T)$ total intraaggregate concentration per mass of soil as a function of radial position, axial position, and time, equal to $\varepsilon S_i C_p(R, Z, T)[\rho_s(1 - \varepsilon)]^{-1} + Q(R, Z, T)$ (M/M).
$y(r, z, t) = Y(r, z, t)\{\varepsilon S_i C_{hm}[\rho_s(1 - \varepsilon)]^{-1} + KC_{hm}^{1/n}\}^{-1}$ (dimensionless).
$z = Z/L$ (dimensionless).
Z axial position.
ε total porosity, void fraction of column (dimensionless).
ε_a microporosity, void fraction containing immobile water, equal to $\varepsilon S_i[1 - \varepsilon(1 - S_i)]^{-1}$ (dimensionless).
ε_m macroporosity, void fraction containing air and mobile water, equal to $(\varepsilon - \varepsilon_a)(1 - \varepsilon_a)^{-1}$ (dimensionless).
λ_+ ionic conductance of cation in water (A V g-equiv L^6).
λ_- ionic conductance of anion in water (A V g-equiv L^6).
μ_1 water viscosity (M L^{-1} T^{-1}).
ρ_a particle density, equal to $\rho_s(1 - \varepsilon_a)$ (M L^{-3}).
ρ_b bulk density (M L^{-3}).
ρ_1 water density (M L^{-3}).
ρ_s soil particle density, equal to $\rho_b(1 - \varepsilon)^{-1}$ (M L^{-3}).
τ_a tortuosity of the air-filled pores (dimensionless).
τ_p tortuosity of the micropores (dimensionless).

Acknowledgments. This material is based upon work supported by the National Science Foundation under grant ECE-8501395 and by the Environmental Engineering Center for Water and Waste Management, Michigan Technological University. Any opinions, findings, and conclusions or recommendations expressed in this publication are those of the authors and do not necessarily reflect the views of the National Science Foundation or Michigan Technological University. The authors would also like to acknowledge the laboratory efforts of Lawrence C. Krause and David B. McKenzie, both graduate student research assistants at Michigan Technological University at the time of this study.

REFERENCES

Abriola, L. M., and G. F. Pinder, A multiphase approach to the modeling of porous media contamination by organic compounds, 1, equation development, *Water Resour. Res.*, *21*(1), 11–18, 1985*a*.
Abriola, L. M., and G. F. Pinder, A multiphase approach to the modeling of porous media contamination by organic compounds, 2, numerical simulation, *Water Resour. Res.*, *21*(1), 19–26, 1985*b*.
Abriola, L. M., and W. J. Weber, Jr., Summary report: Forum on NSF research activities in subsurface systems, University of Michigan, Ann Arbor, July, 1986.
Ashworth, R. A., G. B. Howe, M. E. Mullins, and T. N. Rogers, Air-water partitioning coefficients of organics in dilute aqueous solutions, *J. Haz. Mater.*, *18*, 25–36, 1988.
Bachr, A. L., and M. Y. Corapcioglu, A compositional multiphase model for groundwater contamination by petroleum products, 2, numerical solution, *Water Resour. Res.*, *23*(1), 201–213, 1987.
Black, C. A., D. D. Evans, J. L. White, L. E. Ensminger, and F. E. Clark, *Methods of Soil Analysis*, part 1, *Physical and Mineralogical Properties, Including Statistics of Measurement and Sampling*, 700 pp., American Society of Agronomy, Madison, Wis., 1965.
Brusseau, M. L., and P. S. C. Rao, Sorption nonideality during organic contaminant transport in porous media, *CRC Crit. Rev. Environ. Control*, *19*(1), 33–99, 1989.
Corapcioglu, Y. M., and A. L. Bachr, A compositional multiphase model for groundwater contamination by petroleum products, 1, theoretical considerations, *Water Resour. Res.*, *23*(1), 191–200, 1987.
Crittenden, J. C., N. J. Hutzler, D. G. Geyer, J. L. Oravitz, and G. Friedman, Transport of organic compounds with saturated groundwater flow: Model development and parameter sensitivity, *Water Resour. Res.*, *22*(3), 271–284, 1986.
Danckwerts, P. V., Continuous flow systems, *Chem. Eng. Sci.*, *2*(1), 1–13, 1953.
DeSmedt, F., and P. J. Wierenga, Solute transfer through columns of glass beads, *Water Resour. Res.*, *20*(2), 225–232, 1984.
DeSmedt, F., F. Wauters, and J. Sevilla, Study of tracer movement through unsaturated soil, *J. Hydrol.*, *85*, 169–181, 1986.
Finlayson, B. A., *Nonlinear Analysis in Chemical Engineering*, 366 pp., McGraw-Hill, New York, 1980.
Freundlich, H., *Colloid and Capillary Chemistry*, translated from German by J. S. Hatfield, 883 pp., E. P. Dutton, New York, 1922.
Hashimoto, I., K. B. Deshpande, and H. C. Thomas, Peclet numbers and retardation factors for ion exchange columns, *Ind. Eng. Chem. Fundam.*, *3*(3), 213–218, 1964.
Hutzler, N. J., and J. S. Gierke, Validation of an unsaturated VOC transport model, paper presented at the Proceedings of the 1988 Joint Canadian Society of Civil Engineering and American Society of Civil Engineering Specialty Conference, Vancouver, Br. Columbia, Canada, July 13–15, 1988.
Hutzler, N. J., J. C. Crittenden, J. S. Gierke, and A. M. Johnson, Transport of organic compounds with saturated groundwater

flow: Experimental results, *Water Resour. Res.*, *22*(3), 285–295, 1986.

Hutzler, N. J., B. E. Murphy, and J. S. Gierke, State of technology review: Vapor extraction systems, *EPA 600/2-89-024*, 87 pp., Environ. Prot. Agency, Washington, D. C., 1989*a*.

Hutzler, N. J., J. S. Gierke, and L. C. Krause, Movement of volatile organic chemicals in soils, in *Reactions and Movement of Organic Chemicals in Soils*, *Spec. Publ. 22*, edited by B. L. Sawhney and K. Brown, pp. 373–403, Soil Science Society of America and American Society of Agronomy, Madison, Wis., 1989*b*.

Jury, W. A., Simulation of solute transport using a transfer function model, *Water Resour. Res.*, *18*(2), 363–368, 1982.

Jury, W. A., R. Grover, W. F. Spencer, and W. J. Farmer, Modeling vapor losses of soil-incorporated triallate, *Soil Sci. Soc Am. J.*, *44*, 445–450, 1980.

Jury, W. A., W. F. Spencer, and W. J. Farmer, Behavior assessment model for trace organics in soil, I, Model description, *J. Environ. Qual.*, *12*, 558–563, 1983.

Krause, L. C., Modeling the transport of volatile organic chemicals in unsaturated media: Experimental results, M.S. thesis, Mich. Tech. Univ., Houghton (available from Univ. Microfilms, Ann Arbor, Mich.), 1987.

LeBas, *The Molecular Volumes of Liquid Chemical Compounds*, Longmans Green, London, 1915.

Levenspiel, O., *Chemical Reaction Engineering*, 501 pp., John Wiley and Sons, New York, 1962.

Lindstrom, F. T., and W. T. Piver, Vertical-horizontal transport and fate of low water solubility chemicals in unsaturated soils, *J. Hydrol.*, *86*, 93–131, 1986.

MacKay, D. M., P. V. Roberts, and J. A. Cherry, Transport of organic contaminants in groundwater, *Environ. Sci. Technol.*, *19*(5), 384–392, 1985.

Mayer, R. J. Letey, and W. J. Farmer, Models for predicting volatilization of soil-incorporated pesticides, *Soil Sci. Soc. Am. J.*, *38*, 563–568, 1974.

Millington, R. J., Gas diffusion in porous media, *Science*, *130*, 100–102, 1959.

Miyauchi, T., and T. Kikuchi, Axial dispersion in packed beds, *Chem. Eng. Sci.*, *30*, 343–348, 1975.

Nielsen, D. R., M. Th. van Genuchten, and J. W. Biggar, Water flow and solute transport in the unsaturated zone, *Water Resour. Res.*, *22*(9), 89S–108S, 1986.

Raghavan, N. S., and D. M. Ruthven, Numerical simulation of a fixed-bed adsorption column by the method of orthogonal collocation, *AIChE J.*, *29*(6), 922–925, 1983.

Rao, P. S. C., R. E. Jessup, and T. M. Addiscott, Experimental and theoretical aspects of solute diffusion in spherical and nonspherical aggregates, *Soil Sci.*, *133*, 342–349, 1982.

Reid, R. C., J. M. Prausnitz, and T. K. Sherwood, *The Properties of Gases and Liquids*, 3rd ed., 688 pp., McGraw-Hill, New York, 1977.

Roberts, P. V., M. N. Goltz, R. S. Summers, J. C. Crittenden, and F. Nkedi-Kizza, The influence of mass transfer on solute transport in column experiments with an aggregated soil, *J. Contaminant Hydrol.*, *1*, 375–393, 1987.

Rolston, D. E., D. Kirkham, and D. R. Nielsen, Miscible displacement of gases through soil columns, *Soil Sci. Soc. Am. J.*, *33*, 488–492, 1969.

Rosen, J. B., Kinetics of a fixed bed system for solid diffusion into spherical particles, *J. Chem. Phys.*, *20*(3), 387–394, 1952.

Rosen, J. B., General numerical solutions for solid diffusion in fixed beds, *Ind. Eng. Chem.*, *46*(8), 1590–1594, 1954.

Roy, W. R., and R. A. Griffin, Vapor-phase movement of organic solvents in the unsaturated zone, *Open File Rep. 16*, 37 pp., Environ. Inst. of Waste Manage. Studies, Univ. of Alabama, Tuscaloosa, June 1987.

Sherwood, T. K., R. L. Pigford, and C. R. Wilke, *Mass Transfer*, McGraw-Hill, New York, 1975.

Treybal, R. E., *Mass Transfer Operations*, 784 pp., McGraw-Hill, New York, 1980.

Turek, F., and R. Lange, Mass transfer in trickle bed reactors at low Reynolds number, *Chem. Eng. Sci.*, *36*(3), 569–579, 1981.

van Genuchten, M. Th., and W. A. Jury, Progress in unsaturated flow and transport modeling, *Rev. Geophys.*, *25*(2), 135–140, 1987.

van Genuchten, M. Th., and P. J. Wierenga, Mass transfer studies in sorbing porous media, I, Analytical solutions, *Soil Sci. Soc. Am. J.*, *40*(4), 473–480, 1976.

van Genuchten, M. Th., and P. J. Wierenga, Mass transfer studies in sorbing porous media, II, Experimental evaluation with tritium (^3H_2O), *Soil Sci. Soc. Am. J.*, *41*, 272–278, 1977.

van Genuchten, M. Th., P. J. Wierenga, and G. A. O'Connor, Mass transfer studies in sorbing porous media, III, Experimental evaluation with 2,4,5-T, *Soil Sci. Soc. Am. J.*, *41*, 278–285, 1977.

Weast, R. C., *CRC Handbook of Chemistry and Physics*, 61st ed., CRC Press, Boca Raton, Florida, 1981.

Wierenga, P. J., Solute distribution profiles computed with steady-state and transient water movement models, *Soil Sci. Soc. Am. J.*, *41*, 1050–1055, 1977.

Wilke, C. R., and C. Y. Lee, Estimation of diffusion coefficients for gases and vapors, *Ind. Eng. Chem.*, *47*(6), 1253–1257, 1955.

Wilson, E. J., and C. J. Geankoplis, Liquid mass transfer at very low Reynolds numbers in packed beds, *Ind. Eng. Chem. Fundam.*, *5*(1), 9–14, 1966.

Yule, D. F., and W. R. Gardner, Longitudinal and transverse dispersion coefficients in unsaturated Plainfield sand, *Water Resour. Res.*, *14*(4), 582–588, 1978.

J. C. Crittenden, J. S. Gierke, and N. J. Hutzler, Department of Civil Engineering, Michigan Technological University, Houghton, MI, 49931.

(Received March 31, 1989;
revised November 17, 1989;
accepted December 6, 1989.)

LIST OF PRESENTATIONS COAUTHORED BY JOHN S. GIERKE DURING HIS LABORATORY GRADUATE FELLOWSHIP TENURE

"Validation of an Unsaturated VOC Transport Model," with N.J. Hutzler, 1988 International Conference on Environmental Engineering, American Society of Civil Engineers/Canadian Society of Civil Engineers, Vancouver, British Columbia, Canada, July 13-15, 1988.

"Review of Soil Vapor Extraction System Technology," with N.J. Hutzler and B.E. Murphy, Soil Vapor Extraction Technology for Underground Storage Tank Sites Workshop, USEPA Risk Reduction Engineering Laboratory, Edison, New Jersey, June 27-28, 1989.

"Vapor Extraction of Volatile Organic Chemicals from Unsaturated Soil," with N.J. Hutzler and D.B. McKenzie, International Symposium on Processes Governing the Movement and Fate of Contaminants in the Subsurface Environment, International Association on Water Pollution Research and Control, Stanford, California, July 24-26, 1989.

"Soil Vapor Extraction Systems: State of Technology Review," with N.J. Hutzler and B.E. Murphy, Twelfth Annual Madison Waste Conference: Municipal and Industrial Waste, Department of Engineering Professional Development, University of Wisconsin-Madison, Madison, Wisconsin, September 20-21, 1989.

"Review of Soil Vapor Extraction System Technology," with N.J. Hutzler and B.E. Murphy, HazMat West 1989 Conference, Long Beach, California, November 7-9, 1989.

United States
Environmental Protection
Agency

Risk Reduction
Engineering Laboratory
Cincinnati, OH 45268

Research and Development

EPA/600/S2-89/024 Jan. 1990

EPA

Project Summary

State of Technology Review: Soil Vapor Extraction Systems

Neil J. Hutzler, Blane E. Murphy, and John S. Gierke

Extracting vapor from soil is a cost-effective technique for the removal of volatile organic chemicals (VOCs) from contaminated soils. Among the advantages of the soil vapor extraction process are that it minimally disturbs the contaminated soil, it can be constructed from standard equipment, it has been demonstrated at pilot- and field-scale, it can be used to treat larger volumes of soil than can be practically excavated, and it has potential for product recovery.

Unfortunately, there are few guidelines for the optimal design, installation, and operation of soil vapor extraction systems. A large number of pilot- and full-scale soil vapor extraction systems have been constructed and studied under a wide range of conditions. The major objectives of the Report summarized here are to critically review available documents that describe current practices and to summarize this information as concisely as possible. A typical vapor extraction system is briefly described, the experience with existing extraction systems has been reviewed, and information about each system is briefly summarized.

Soil vapor extraction can be effectively used for removing a wide range of volatile chemicals over a wide range of conditions. The design and operation of this system are flexible enough to allow for rapid changes in operation, thus optimizing the removal of chemicals. Although a number of variables intuitively affect the rate of chemical extraction, no extensive study to correlate variables to extraction rates has been identified.

This Project Summary was developed by EPA's Risk Reduction Engineering Laboratory, Cincinnati, OH, to announce key findings of the research project that is fully documented in a separate report of the same title (see Project Report ordering information at the back).

Introduction

Soils may become contaminated in a number of ways with such volatile organic chemicals as industrial solvents and gasoline components. The sources of contamination at or near the earth's surface include intentional disposal, leaking underground storage tanks, and accidental spills. Contamination of groundwater from these sources can continue even after discharge has stopped because the unsaturated zone above a groundwater aquifer can retain a portion or all of the contaminant discharge. As rain infiltrates, chemicals elute from the contaminated soil and migrate toward groundwater.

Alternatives for decontaminating unsaturated soil include excavation with on-site or off-site treatment or disposal, biological degradation, and soil flushing. Soil vapor extraction is also an accepted, cost-effective technique to remove volatile organic chemicals from contaminated soils. Among the advantages of the soil vapor extraction process are that it minimally disturbs the contaminated soil, it can be constructed from standard equipment, it has been demonstrated at pilot- and field-scale, it can be used to treat larger volumes of soil than are practical for excavation, and

it has a potential for product recovery. With vapor extraction, spills can be cleaned up before the chemicals reach the groundwater table. Soil vapor extraction technology is often used with other clean up technologies to provide complete restoration of contaminated sites.

Unfortunately, there are few quidelines for the optimal design, installation, and operation of soil vapor extraction system. Theoretically based design equations defining the limits of this technology are especially lacking. Because of this, the design of these systems is mostly empirical. Alternative designs can only be compared by the actual construction, operation, and monitoring of each design.

A large number of pilot- and full-scale soil vapor extraction systems have been constructed and studied under a wide range of conditions. The information gathered from these experiences can be used to deduce the effectiveness of this technology. One of the major objectives of the Report is to review available reports describing current practices critically and to summarize this information as concisely as possible. A brief description of a typical vapor extraction system is presented. The experience with existing extraction systems has been reviewed, and information about each system is briefly summarized in a standard form. The information is further summarized in several tables, which form the bases for a discussion of the design, installation, and operation of these systems. Because soil vapor extraction is a relatively new soil remediation technology, this Technology Review document will evolve as more information becomes available.

Process Description

A soil vapor extraction, forced air venting, or in situ air stripping system (Figure 1) revolves around the extraction of air containing volatile chemicals from unsaturated soil. Fresh air is injected or flows into the subsurface at locations around a spill site, and the vapor-laden air is withdrawn under vacuum from recovery or extraction wells.

System Components

Extraction wells are typically designed to fully penetrate the unsaturated zone to the capillary fringe. Extraction wells usually consist of slotted elastic pipe placed in permeable packing.

System Operations

During remediation, the blower is turned on and the air flow through the soil comes to an equilibrium. The flows that are finally established are a function of the equipment, the flow control devices, the geometry of well layout, the site characteristics, and the air permeability of the soil. At the end of operation, the final distribution of VOCs in the soil can be measured to ensure decontamination of the site. Wells may be aligned vertically or horizontally. Vertical alignment is typical for deeper contamination zones and for residue in radial flow patterns. If the depth of the contaminated soil or the depth to the groundwater table is less than 10 to 15 ft, it may be more practical to dig a trench across the area of contamination and install horizontal perforated piping in the trench bottom rather than to install vertical extraction wells. Usually several wells are installed at a site.

System Variables

A number of variables characterize the successful design and operation of a vapor extraction system: site conditions, soil properties, control variables, response variables and chemical properties. The specific variables belonging to these groups include:

Site Conditions: distribution of VOCs, depth to groundwater, infiltration rate, location of heterogeneities, temperature, atmospheric pressure.

Soil Properties: permeability, porosity, organic carbon content, soil structure, soil moisture characteristics, particle size distribution.

Control Variables: air withdrawal rate, well configuration, extraction well spacing, vent well spacing, ground surface covering, inlet air VOC concentration and moisture content, pumping duration.

Response Variables: pressure gradients, final distribution of VOCs, final moisture content, extracted air concentration, extracted air temperature, extracted air moisture, power usage.

Chemical Properties: Henry's constant, solubility, adsorption equilibrium, diffusivity (air and water), density, viscosity.

Well Design and Placement

Well spacing is usually based on some estimate of the radius of influence of an individual extraction well. In the studies reviewed, well spacing has ranged from 15 to 100 ft. Well spacing should be decreased as soil bulk density increases or the porosity of the soil decreases. One of the major differences noted between systems was the soil boring diameter. Larger borings are preferred to minimize extracting liquid water from the soil.

In the simplest soil vapor extraction systems, air flows to an extraction well from the ground surface. To enhance air flow through zones of maximum contamination, it may be desirable to include air inlet wells in the installation. These injection wells or air vents, whose function is to control the flow of air into a contaminated zone, may be located at numerous places around the site. Typically, injection wells and air vents are constructed similarly to extraction wells. In some installations, extraction wells have been designed so they can also be used as air inlets. Usually, only a fraction of extracted air comes from air inlets. This indicates that air drawn from the surface is the predominant source of clean air.

One study investigated the effects of air-flow rate and the configuration of the inlet and extraction wells on gasoline recovery from an artificial aquifer. It was determined that screening geometry only had an effect at the low air-flow rates. At low flow rates, higher recovery rates resulted when the screen was placed near the water table rather than when the well was screened the full depth of the aquifer.

Woodward-Clyde made a simila. assessment at the Time Oil Company site. Their engineers suggested that wells should be constructed with approximately 20 ft of blank casings between the top of the screen and the soil surface to prevent the short circuiting of air and to aid in the extraction of deep contamination. At most sites, the initial VOC recovery rates were relatively high then decreased asymptotically to zero with time. Several studies have indicated that intermittent venting from individual wells is probably more efficient in terms of mass of VOC extracted per unit of energy expended. This is especially true when extracting from soils where mass transfer is limited by diffusion out of immobile water. Optimal operation of a soil vapor extraction system may involve taking individual wells in and out of service to allow time for liquid diffusion and to change air flow patterns in the region being vented. Little work has been done to study this.

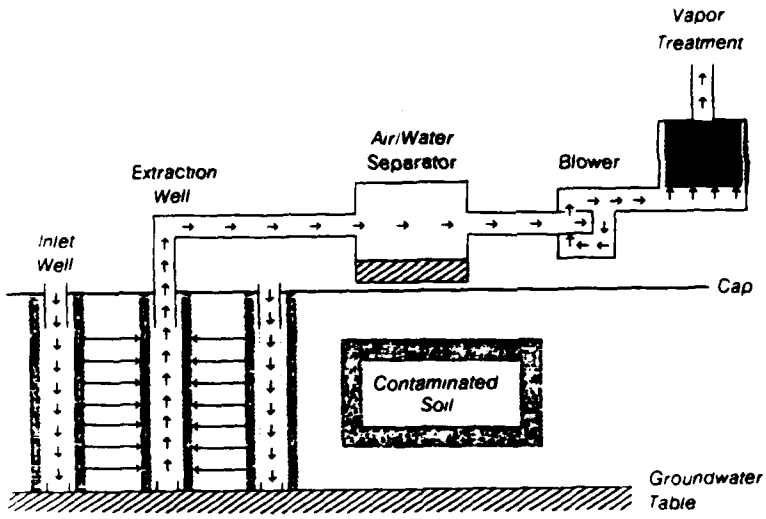


Figure 1 Soil vapor extraction system.

One of the major problems in the operation of a soil vapor extraction system is determining when the site is sufficiently clean to cease operation.

The design and operation of soil vapor extraction systems can be quite flexible; changes can be made during the course of operation with regard to well placement, or blower size, or air flows from individual wells. If the system is not operating effectively, changes in the well placement or capping the surface may improve it.

Conclusions

Based on the current state of the technology of soil vapor extraction systems, the following conclusions can be made:

1. Soil vapor extraction can be effectively used for removing a wide range of volatile chemicals in a wide range of conditions.
2. The design and operation of these systems is flexible enough to allow for rapid changes in operation, thus, optimizing the removal of chemicals.
3. Intermittent blower operation is probably more efficient in terms of removing the most chemical with the least energy.
4. Volatile chemicals can be extracted from clays and silts but at a slowerrate. Intermittent operation is certainly more efficient under these conditions.
5. Air injection has the advantage of controlling air movement, but injection systems need to be carefully designed.
6. Extraction wells are usually screened from a depth of from 5 to 10 ft below the surface to the groundwater table. For thick zones of unsaturated soil, maximum screen lengths of 20 to 30 ft are specified.
7. Air-water separators are simple to construct and should probably be installed in every system.
8. Installation of a cap over the area to be vented reduces the chance of extracting water and extends the path that air follows from the ground surface, thereby increasing the volume of soil treated.
9. Incremental installation of wells, although probably more expensive, allows for a greater degree of freedom in design. Modular construction where the most contaminated zones are vented first is preferable.
10. Use of soil vapor probes in conjunction with soil borings to assess final clean up is less expensive than use of soil borings alone. Usually a complete materials balance on a given site is impossible because most sites have an unknown amount of VOC in the soil and in the groundwater.
11. Soil vapor extraction systems are usually only part of a site remediation system.
12. Although a number of variables intuitively affect the rate of chemical extraction, no extensive study to correlate variables to extraction rates has been identified.

The full report was submitted in partial fulfillment of Cooperative Agreement No. CR-814319-01-1 by Michigan Technological University under the sponsorship of the U.S. Environmental Protection Agency.

Neil J. Hutzler, Blane E. Murphy, and John S. Gierke are with Michigan Technological University, Houghton, MI 49931.
Paul R. de Percin *is the EPA Project Officer (see below).*
The complete report, entitled "State of Technology Review: Soil Vapor Extraction Systems," (Order No. PB 89-195 184/AS; Cost: $15.95, subject to change) will be available only from:
National Technical Information Service
5285 Port Royal Road
Springfield, VA 22161
Telephone: 703-487-4650
The EPA Project Officer can be contacted at:
Risk Reduction Engineering Laboratory
U.S. Environmental Protection Agency
Cincinnati, OH 45268

United States
Environmental Protection
Agency

Center for Environmental Research
Information
Cincinnati OH 45268

BULK RATE
POSTAGE & FEES PAID
EPA
PERMIT No. G-35

Official Business
Penalty for Private Use $300

EPA/600 S2-89/024

MODELING THE TRANSPORT OF VOLATILE ORGANIC CHEMICALS IN UNSATURATED SOIL AND THEIR REMOVAL BY VAPOR EXTRACTION

By

JOHN S. GIERKE

A DISSERTATION

Submitted in partial fulfillment of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

(Environmental Engineering)

MICHIGAN TECHNOLOGICAL UNIVERSITY

1990

This dissertation, "Modeling the Transport of Volatile Organic Chemicals in Unsaturated Soil and Their Removal by Vapor Extraction," is hereby approved in partial fulfillment of the requirements for the degree of DOCTOR OF PHILOSOPHY in the field of Environmental Engineering.

PROGRAM: Environmental Engineering

Neil J. Hutzler

Dissertation Advisor: Dr. Neil J. Hutzler

C. Robert Baillod

Program Chair: Dr. C. Robert Baillod

June 15, 1990

Date

ABSTRACT

The approach used here to develop models for predicting subsurface chemical transport and vapor extraction performance is to: (1) develop a conceptual picture, (2) derive and solve equations that describe the conceptual picture, (3) simplify the governing equations to examine the impacts of different transport mechanisms and obtain analytic solutions, (4) verify the numerical solution of the model against analytic solutions, and (5) validate the model against experimental results. One-dimensional model calculations and the results of laboratory experiments were used to study chemical transport in unsaturated soil columns. From the results of these column studies, it was shown that: (1) liquid dispersion and immobile water diffusion affect transport when water is flowing; (2) if only air is flowing, diffusion in immobile water does not affect transport in uniform sands; (3) vapor diffusion is important when the pore water velocity is less than 0.001 cm s^{-1} and when the air is mobile; (4) rates of mass transfer between phases are fast; and (5) sorption of vapors is important in dry soils and adsorption equilibrium may be nonlinear for low relative humidities. Model sensitivity calculations showed that intraaggregate diffusion could be simulated as first-order, air-water mass transfer. Vapor extraction system models were derived for two system configurations (radial flow and planar flow). For a wide range of air flow rates, both configurations remove chemical at the same rate; however, a planar system can be operated at about one-half of the power consumption. For a given soil and system configuration, a range of air flow rates exist where the volume of air extracted per unit of chemical removed is a minimum. At low air flow rates, more air is required because of mixing due to gas diffusion, and at the high flow rates, removal is limited by the rate of phase mass transfer. This range is narrower for a radial configuration and for aggregated soils at high moisture contents. Models that assume complete mixing overestimate the time for cleanup unless the air velocity is low. Ignoring the rate to equilibrium between phases can lead to an underestimate of the cleanup time for high air flow rates.

ACKNOWLEDGEMENTS

Financial support for this work has been provided by the following: United States Air Force, Faculty Research Program/Graduate Student Research Program (Contract No. F49620-88-C-0053), Research Initiation Program (Contract No. F49620-88-C-0053/SB5881-0378), and Laboratory Graduate Fellowship Program (Subcontract No. S-789-000-007) (all of which are administered by Universal Energy Systems, Dayton, Ohio); National Science Foundation (Grant No. ECE-8501395); U.S. Environmental Protection Agency (Contract No. CR-814319-01-0); and Michigan Technological University in conjunction with the Department of Civil and Environmental Engineering and the Environmental Engineering Center for Water and Waste Management.

I am grateful to Larry Krause and Dave McKenzie for their laboratory work, Mitzi Johnson for her secretarial support, Dave Perram for his analytical and experimental assistance, Dave Hand for his modeling advice, Mike Bednar for his help with checking some of the equation derivations, and Leif Hauge for creating some of the graphics. I appreciate the help with the administration of the USAF Laboratory Graduate Fellowship Program from Tom Staufer, Dr. Jimmy Cornette, and Capt. Ed Heyse (all from Tyndall AFB, Florida)and Mr. Rodney Darrah and Ms. Judy Conover (Universal Energy Systems). I am grateful to Dave DePaoli from Oak Ridge National Laboratory for allowing me to visit their vapor extraction site and making his results available to me. A special thanks goes out to the committee members for reviewing this document: Dr. Gilbert Lewis, Mathematics Department; Dr. Joel W. Massmann, Department of Geological Engineering; and Drs. John C. Crittenden, Neil J. Hutzler, and James R. Mihelcic, Department of Civil

and Environmental Engineering. I also appreciate the professional advice from Drs. Crittenden, Hutzler, and Massmann.

During my graduate school "tenure" I have been helped either professionally, academically, technically, athletically, or socially by many others in the Department of Civil and Environmental Engineering: Drs. Marty Auer, Bob Baillod, Bill Bulleit, Bogue Sandberg, and Vern Watwood, Jim Yates, and many graduate and undergraduate students too numerous to mention. I also appreciate the encouragement from their families.

This work and the requirements for this degree would not have been completed without the advice, encouragement, and friendship of my advisor, Neil Hutzler. I owe all of my professional career training to him and thank him for allowing me this opportunity to study under him. His family has also been like my own, and I appreciate all their instructions, especially those regarding child rearing.

I also appreciate the support of my parents, family, and friends and those I have come to know in the Copper Country. My wife, Lynn, has been a tremendous help and I love her for it. In addition, I appreciate the encouragement of my daughter, Genevieve, who kept reminding me it was time to get out of bed and go to school.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

SECTION 1. INTRODUCTION

Large areas of soil have become contaminated with volatile organic chemicals such as industrial solvents and gasoline components. The sources of contamination at or near the earth's surface include intentional disposal, leaking underground storage tanks, and accidental spills. Contamination of ground water from these sources continues even after discharge is stopped because the vadose or unsaturated zone retains a portion or all of a contaminant discharge due to capillary attraction, dissolution, volatilization, and sorption. As rain infiltrates, chemicals elute from the vadose zone and migrate towards ground water.

There are a number of alternatives for decontaminating the unsaturated zone including excavation with onsite or offsite treatment and/or disposal, biological degradation, and soil washing. Use of vapor extraction is also becoming an accepted, cost-effective technique for the removal of volatile organic chemicals from contaminated soils (cf. Hutzler *et al.* [1989a]). Among the advantages of the soil air extraction process are minimal disturbance of the contaminated soil, use of standard components, experience with process at field- and pilot-scale, treatment of larger volumes than are practical for excavation, and a potential for product recovery. With vapor extraction, it is possible to clean up spills before the chemicals reach the ground water table.

Johnson *et al.* [1990a] have reported some important practical considerations for the design and installation of vapor extraction systems. Unfortunately there are few guidelines for optimal design, installation, and operation [Bennedsen, 1987; Wilson *et al.*, 1987].

Theoretically-based design equations which define the limits of this technology are especially lacking. Because of this, the design of these systems is mostly empirical. Different designs are compared by construction, operation, and monitoring of each alternative. For example, comparison of removal rates using different pumping strategies and vent placement geometries requires that such alternatives be implemented and studied either in the field or by pilot- or laboratory-scale experiments. Such analyses are generally expensive in terms of time and resources. Predictive transport models that account for the chemical fate mechanisms important in unsaturated soil would enhance the design and operation of vapor extraction systems.

Subsurface transport models are needed for understanding basic phenomena affecting the movement of contaminants, for predicting chemical fate, and for designing soil treatment systems. The design of vapor removal systems and effective ground water protection devices can be enhanced by using mechanistic models that accurately predict chemical fate in the unsaturated zone. The development of models for engineering applications requires a fundamental understanding of transport and reaction mechanisms that affect the movement and persistence of organic compounds in soil.

Subsurface contaminant transport and attenuation is governed by a number of spreading, retardation, and transformation mechanisms such as: advection, dispersion, diffusion, and interfacial mass transfer; adsorption and volatilization; and biological and chemical reactions. These mechanisms and their impacts on chemical fate are discussed in detail by MacKay *et al.* [1985] and Nielsen *et al.* [1986]. In general, as a chemical travels through soil with fluid flow, the shape of its

concentration profile is affected by dispersing or spreading mechanisms, the profile's position is slowed by retardation mechanisms, and the mass may decrease due to biological and chemical transformations.

The remainder of this section briefly describes the vapor extraction process and previous modeling efforts that have focused on chemical movement in unsaturated soils and on describing vapor extraction performance. The next section states the objectives and scope of this work and is followed by a step-by-step description of the approach used here to develop subsurface fate models. The models and results are described in detail in Section 4. This section begins with descriptions of the important mechanisms that are considered and the approach used to model each process and includes details and justification of the assumptions used in deriving the models.

Section 4 is divided into two main sections, one describes the column (one-dimensional) model derivations and validation experiments, and the other, the two-dimensional model (those describing the vapor extraction process) derivations and numerical results. The one-dimensional model section includes a set of models for describing transport of volatile chemicals with water flow and either with or without air flow and a set for air transport and no water flow. The subsections that describe the equation derivations, solution, simplification, verification, and validation for the column models are complete and contain the most detail about the importance of fate mechanisms and transport. The two-dimensional model sections are subdivided into descriptions of models for a single vertical vent system and a model for one cell of a trench system. Equation derivations, solution, simplification, and verification are performed for the models

describing both configurations, and the numerical results sections for these two sets of models report some considerations as to which mechanisms are most important and the optimal vent configuration (trench or vertical) and range of air the extraction rates that should be used.

1.1 PROCESS DESCRIPTION

A vapor extraction, forced air venting, or in-situ air stripping system, such as the one shown in Figure 1.1, involves extracting air from unsaturated soil. Fresh air is injected or flows into the subsurface at locations around a spill site, and the vapor-laden air is withdrawn under vacuum from recovery or extraction vents.

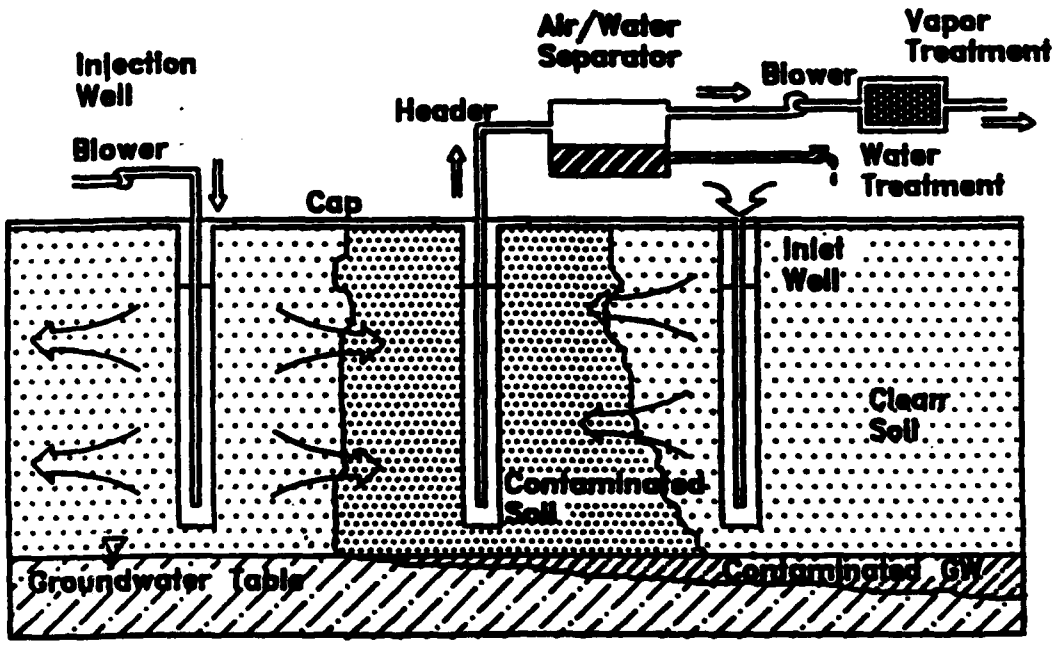


Figure 1.1. Vapor extraction system schematic [Hutzler *et al.*, 1989a].

1.1.1 System Components

A composite vapor extraction system is shown in Figure 1.1, and it is comprised of: (1) extraction vents, (2) air inlet or injection vents (optional), (3) piping or air headers, (4) vacuum pumps or air blowers, (5) flow meters and controllers, (6) vacuum gages, (7) sampling ports,

(8) air-water separator (optional), (9) vapor treatment (optional), and (10) a cap (optional). Details regarding the specifications of these components are summarized in a report by Hutzler *et al.* [1989a].

From the list of components given above, only details about the vent configuration, the blower rate, and the installation of a cap are important for developing a system model. Extraction vents are typically designed to fully penetrate the unsaturated zone to the capillary fringe. If the ground water is at a shallow depth (10-15 ft) or if the contamination is confined to a depth of less than 15 ft, then the extraction vents may be placed horizontally in trenches [Hutzler *et al.*, 1989a]. Extraction vents are slotted pipe which are placed in a permeable packing. The surface of an augered column for vertical vents or a trench for horizontal vents is usually grouted to prevent short circuiting of air from the surface. It may be desirable to also install air inlet or injection vents to enhance air flow through zones of maximum contamination. They are constructed in the same way as extraction vents. Inlet vents are passive and allow air to be drawn into the ground at a particular location. Injection vents force air into the ground and are used in closed-loop systems. The function of inlet and injection vents is to enhance air movement in strategic locations and promote horizontal air flow. Pumps or blowers reduce gas pressure in the extraction vents to induce air flow towards the extraction vents. The pressure from the outlet side of the pumps or blowers can be used to push the exit gas through a treatment system, such as activated carbon or incineration, and back into the ground if injection vents are used. An impermeable cap may be used to cover the treatment site to minimize infiltration and prevent short circuiting of

inlet air.

1.1.2 System Operation

During operation, the blower is turned on, and the air flow comes to steady-state. The steady-state flow rate is a function of the equipment, the flow control devices, the geometry of the vent layout, the site characteristics, and the air-permeability of the soil. Exhaust air is sampled on a routine basis and used along with flow measurements to determine the rate of chemical extraction. Typically, the mass removal rate is high at first, and subsequently decreases [Dynamac Corp., 1986]. At the end of operation, the final distribution of contaminants should be measured to ensure decontamination of the site.

1.1.3 System Variables

A number of variables characterize the successful design and operation of a vapor extraction system. They may be classified as site conditions, soil properties, chemical characteristics, control variables, and response variables [Anastos *et al.*, 1985; Dynamac Corp., 1986]. Table 1.1 is a list of specific variables that belong to these groups.

Site conditions, soil properties, and chemical properties are important for determining the feasibility of a vapor extraction system [Hutzler *et al.*, 1989a] as well as predicting the subsurface fate of contaminants. A predictive model could be used in aiding the feasibility determination and in designing the control variables, especially the vent geometry and extraction rate. The response variables should be used to compare different designs and manage the system operation. For example, a model could be used to predict power usage for different designs or off-gas vapor concentrations for various

air extraction rates.

Table 1.1. Important Variables for Assessing the Performance of Vapor Extraction [Hutzler *et al.*, 1989a].

Site Conditions	Control Variables
Contaminant distribution	Air withdrawal rate
Depth to ground water	Vent configuration
Infiltration rate	Extraction vent spacing
Location of heterogeneities	Vent spacing
Temperature	Ground surface covering
Atmospheric pressure	Pumping duration
	Inlet air concentration and moisture content
Soil Properties	
Permeability (air and water)	
Porosity	Response Variables
Organic carbon content	Pressure gradients
Soil structure	Contaminant distribution
Soil moisture characteristics	Moisture profile
Particle size distribution	Extracted air concentration
	Extracted air moisture
Chemical Properties	Extracted air temperature
Henry's constant	Power usage
Solubility	
Adsorption equilibrium	
Diffusivity (air and water)	
Density	
Viscosity	

1.2 PREVIOUS MODELING WORK

Research efforts towards developing a better understanding of chemical fate processes in the vadose zone have been to date segmented and disciplinary [Nielsen *et al.*, 1986]. Most models that describe unsaturated chemical transport use Fickian-based convection-diffusion expressions [van Genuchten and Jury, 1987]. Few mechanistic models include air flow so they can not be used to simulate vapor extraction, and models that include air flow, usually ignore other mass transfer mechanisms.

1.2.1 Unsaturated Transport Models

Previous work on modeling chemical transport in unsaturated soils

has focused in three areas: (1) vapor transport in the upper soil layer for predicting pesticide movement [Rolston *et al.*, 1969; Mayer *et al.*, 1974] and for assessing the behavior of organic chemicals [Jury *et al.*, 1980 and 1983; Falta *et al.*, 1989]; (2) tracer and nonvolatile chemical transport for simulating the one-dimensional movement of salts and heavy metals [van Genuchten and Wierenga, 1976; Jury, 1982]; and (3) multi-dimensional subsurface movement of liquids and vapors for estimating the travel time of organic solvents and petroleum products to ground water [Abriola and Pinder, 1985a; Lindstrom and Piver, 1986; Corapcioglu and Baehr, 1987; Sleep and Sykes, 1989; Mendoza and Frind, 1990]. Although the multi-dimensional models are conceptually closer to field conditions, these models have ignored phase mass transfer mechanisms. Moreover, the three-dimensional models have not been validated against experimental data, and validation tests are important for determining the predictive capabilities of a model [Abriola and Weber, 1986].

Mechanisms of nonequilibrium [Nielsen *et al.*, 1986; Brusseau and Rao, 1989], specifically those associated with phase-transfer reactions, have typically been ignored in the derivation of three-dimensional transport models. Models that describe the one-dimensional movement of vapors or aqueous solutes have studied some nonequilibrium effects. For example, DeSmedt *et al.* [1986] used a model developed by van Genuchten and Wierenga [1976] to study the impact of diffusion in immobile water on tracer breakthrough curves in columns of unsaturated sand. They propose that the relative importance of diffusion in immobile water increases with decreasing water content.

Abriola and Pinder [1985a] developed equations to describe the three-dimensional movement of organics in air, water, and pure-liquid

phases. They did not include advective air transport nor did they attempt to account for the rate of mass transfer between phases. They solved their equations for a one-dimensional flow problem which included axial dispersion in both fluid phases and instantaneous equilibrium between phases [Abriola and Pinder, 1985b]. Model calculations by Abriola and Pinder [1985b] concentrate primarily on convergence and numerical sensitivity with respect to time step and mesh refinement.

Lindstrom and Piver [1986] considered heat transport in their formulations describing the movement of moisture and dissolved organic contaminants. They did not consider advective vapor movement or phase transfer processes.

Corapcioglu and Baehr [1987] derive mathematical equations for the three-dimensional, subsurface movement of petroleum product as vapors, aqueous solutes, and pure liquid. Like the models of Abriola and Pinder [1985a] and Lindstrom and Piver[1986], Coracioglu and Baehr [1987] did not consider air flow. Abriola and Pinder [1985a] derived their equations based on two contaminants, nonreactive and reactive components; Corapcioglu and Baehr [1987] accounted for many components, specifically, those which comprise petroleum. Furthermore, Corapcioglu and Baehr [1987] account for sorption and biological reactions where Abriola and Pinder [1985a] ignore these reactions. Abriola and Pinder [1985a] include soil compressibility while Corapcioglu and Baehr [1987] do not. The importance of this difference in vadose zone modeling is probably negligible. Baehr and Corapcioglu [1987] solved the equations of Corapcioglu and Baehr [1987] for a vertical, one-dimensional contamination scheme of hydrocarbon constituents. Their solution is restricted to homogeneous soils where the moisture content is uniform.

They discovered that model simulations were most sensitive to vapor diffusion. Neither Abriola and Pinder [1985b] nor Baehr and Corapcioglu [1987] test their model simulations against experimental data.

The model by Sleep and Sykes [1989] was used to examine density-driven gas flow due to organic chemical volatilization in conjunction with the movement of organic and aqueous phases. Unlike other multi-phase models, they account for volatilization and dissolution rates in their descriptions of chemical movement. A similar model was developed by Mendoza and Frind [1990], but it did not consider liquid movement or an interphase mass transfer step. Mendoza and Frind [1990] were able to fit model calculations to measured trichloroethene concentrations up to a distance of 1 m from an immobile source. In their experiments, density-driven gas flow was only important to consider at the locations furthest from the contamination source. Falta *et al.* [1989] developed a model similar to Mendoza and Frind [1990] and arrived at similar conclusions. These density-driven flow models do not consider pressure-induced flow, such as would result from a vapor extraction system. Externally induced vacuums may negate the affects of density variations.

In addition to the above mentioned research efforts, similar work has been performed in chemical engineering analysis of trickle-bed reactors [Hofmann, 1983; Crine and L'Homme, 1983]. The mathematical description of a trickle-bed reactor is analogous to chemical transport through columns of unsaturated soil. Trickle-bed reactors operate under conditions much different than soil columns. Gas and liquid flows are higher in trickle-bed reactors, and packing materials are usually synthetic and of more uniform size and shape. Reactions which take

place in a trickle-bed reactor are often catalytic and occur between pure vapors and pure liquids. For these reasons the mechanisms of interest in trickle-bed reactors may not be important in soil. For example, gas-liquid mass transfer is an important consideration in a trickle-bed reactor while gas and liquid dispersion are of more importance in soil.

There continues to be a need for a better understanding of subsurface fate processes [Nielsen *et al.*, 1986; Abriola and Weber, 1986], including those associated with physical nonequilibrium [Brusseau and Rao, 1989]. Especially lacking are integrated modeling-experimental approaches [Abriola and Weber, 1986]; approaches that not only formulate hypotheses of fate mechanisms but test the hypotheses experimentally as well as numerically.

1.2.2 Vapor Extraction System Models

The mathematical models that have been developed to date have focused on the description of air velocity and chemical phase-equilibrium [Wilson *et al.*, 1987; Hochmuth *et al.*, 1988; Wilson *et al.*, 1988; Baehr *et al.*, 1989; Roy and Griffin, 1989; Johnson *et al.*, 1990b]. Vapor diffusion is only considered in a few instances: diffusion from layers of low permeability [Hochmuth *et al.*, 1988] and diffusion from a layer of floating product [Wilson *et al.*, 1987; Johnson *et al.*, 1990b]. Only one study has included diffusion in the direction of flow [Wilson *et al.*, 1987] while others have considered diffusion as a predominant mechanism by assuming complete mixing [Hochmuth *et al.*, 1988; Roy and Griffin, 1989; Johnson *et al.*, 1990b]. In the draft conceptual plan for the in situ air-stripping of volatile contaminants at the Seymour Recycling Corporation Hazardous Waste Site [CH_2M-Hill, 1987], a

numerical, two-dimensional, multilayered gas flow model called GASMAIN was used to compute air flows, and a complete mixing model coupled with local equilibrium (retardation) calculations was used to estimate removals. They also calculated mass removals using the one-dimensional convection-dispersion equation [van Genuchten and Jury, 1987], even though the system they were trying to model is two-dimensional. For the air flows they selected, the complete mixing model and the convection-dispersion model gave similar results.

Wilson *et al.* [1987] developed a model to describe the removal of gasoline from a pilot-scale vapor extraction system study by Wootan and Voynick [1984]. Simplifying assumptions of ideal gases, isothermal processes, isotropic soils, nondegradable constituents represented as a lumped-single component, one-dimensional darcian flow of vapors, and no liquid flow in the unsaturated zone were imposed. They also assumed that a pure organic phase existed only at the capillary fringe. Relationships are reported for determining the importance of gas diffusion and gas density stratification. Model calculations were successfully compared to an analytical solution for constant velocity. Model sensitivities were performed on gas velocity, exit pressure, compound molecular weight, temperature, porosity, and outlet boundary condition. Their model calculations appeared to simulate experimental pilot data of Wootan and Voynick [1984].

Hochmuth *et al.* [1988] developed a model for simulating removals of solvent vapors from a site in Santa Clara County, California. They modeled the site as a cylinder consisting of intermittent layers of permeable and impermeable soil materials. They considered diffusion of chemical from the impermeable lenses into completely mixed permeable

zones. Calibrations of initial concentrations and retardation coefficients provided good fits of observed extraction results.

Roy and Griffin [1989] derived a method for estimating the time for removing dissolved volatile organics. Because their formulations assume complete mixing, their method may overestimate the cleanup time.

Wilson *et al.* [1988] developed a laboratory vapor extraction column model that accounted for the compressibility of air. They also derived a single-vent extraction model for estimating field-scale performance. The field scale model assumes vapor withdrawal at a single point. In their models they ignore diffusional transport, mass transfer resistances, and sorption, nor did they attempt to even validate their column models.

Baehr *et al.* [1989] developed a multicomponent, local equilibrium model for describing gasoline extraction in laboratory columns. They concluded that mass transfer resistances and diffusional transport were negligible in comparison to advective vapor movement. They used flow rates that were much higher (15 - 30 times) than the normal operating rates of a vapor extraction, so it may not be surprising that diffusion effects were not observed. In addition, it did not appear that they attempted to simulate the latter part of the experiments where air-water mass transfer resistances may become more important [McClellan and Gillham, 1990]. They concluded that Henry's partitioning and soil sorption were unimportant considerations. Again, this is probably due to the fact that they modeled only the part of the extraction where the volatilization of gasoline from the organic-liquid phase was important.

Johnson *et al.* [1990b] developed screening models for estimating the extraction of volatile mixtures from soils. One model can be used

to predict the pressure-flow field around an extraction vent in either a homogeneous or a layered system. A separate model is used to calculate the equilibrium concentrations of the volatile mixture components by assuming complete mixing in each layer.

To date, there has been no integrated modeling-experimental studies that have examined the impacts of vapor diffusion in the direction of flow, air-water mass transfer resistance, and diffusion from immobile zones. Models that have attempted to describe vapor extraction performance have ignored vertical transport with water infiltration. Most laboratory experiments have focused on hydrocarbon mixtures where multicomponent interactions can not be distinguished from other effects. The models and laboratory studies reported here will examine the impacts of air-water mass transfer, vapor diffusion, liquid diffusion, and air and water advection and dispersion on the transport and removal of volatile organic chemicals from unsaturated soil. Model calculations will be used to examine the impact of air flow rate and vent configuration on the removal efficiency of vapor extraction.

1.2.3 Summary of Previous Modeling Work

The previous modeling work described above is summarized in Table 1.2 in terms of the mechanisms considered in the model developments. Only the models that have actually been solved are included in this chronological summary. For example, even though Abriola and Pinder [1985a] and Corapcioglu and Baehr [1987] derived equations for three dimensional transport, they did not solve these general conditions.

Table 1.2. Summary of the Mechanisms Included in Models of Unsaturated Transport and Vapor Extraction.

Mechanism:	Advection		Dispersion	Diffusion		Interfacial Mass Transfer				Sorption			Heat	Bio.	Chem.	Reference
Phase:	Air	Water	Organic	Water	Air	Water	A-W	MW-IW	O-W	O-A	A-S	O-S	W-S				
	1C				X						X						Rolston et al. [1969]
	I				X						X						Mayer et al. [1974]
		1C		X				X					X				van Genuchten and Wierenga [1976]
	I	1C		X	X		X						X				Jury et al. [1980]
	I	1U															Jury [1982]
	I	1C		X	X								X		X or X		Jury et al. [1983]
	I	1U	1U	X	X												Abriola and Pinder [1985b]
	I	2U		X	X						X		X	X	X	X	Lindstrom and Piver [1986]
	I	1C	I	X	X						X	X	X		X	X	Baehr and Corapcioglu [1987]
	2U				X												Wilson et al. [1987]
	1C				P						X						Hochmuth et al. [1988]
	2U	I									X		X				Wilson et al. [1988]
	1C	I	I										X				Baehr et al. [1989]
	2D	I			X								X				Falta et al. [1989]
	1C	I			P								X				Roy and Griffin [1989]
	2D	2U		X	X		X		X	X							Sleep and Sykes [1989]
	1C	I			P								X		X or X		Johnson et al. [1990b]
	2D	I			X								X				Mendoza and Frind [1990]
	1C	1C		X	X	L	X	X					X				This Work

Abbreviation Key: A=air phase, Bio.=biological reactions, Chem.=chemical reactions, C=constant flow, D=density-driven flow, I=immobile, L=intraaggregate diffusion, M=mobile, O=organic phase, P=perfect mixing (infinite diffusion rate), S=soil phase, U=unsteady flow, W=water phase, X=included, 1=one dimensional, 2=two dimensional

Included as the last row of Table 1.2 is a summary of the mechanisms that are included in the model developments reported in this work. The distinguishing features of this study are the inclusion of two nonequilibrium mechanisms in the models, specifically air-water mass transfer and intraaggregate diffusion, and the modeling of simultaneous movement of chemicals with air and water flow. In addition, model calculations are compared to laboratory experiments to determine the validity of the models.

SECTION 2. OBJECTIVES AND SCOPE

The results of this study are complementary to the goals of a project where the focus is the assessment of vapor extraction system performance [Hutzler *et al.*, 1989a]. This research has focused on laboratory- and pilot-scale model development.

The primary objectives of this research are (1) to demonstrate a rational modeling approach for describing subsurface chemical transport and the removal of volatile organic chemicals from unsaturated soil and (2) to enhance the current level of understanding of subsurface chemical transport. The first objective is met by developing and testing a modeling approach. The approach reported in this work is used to derive and test models that describe the movement of volatile organic chemicals with air or water flow in laboratory columns of unsaturated soil. Models are then developed to account for the movement of chemicals with water flow and removal with air flow to simulate the vapor extraction process. A logical set of experiments is combined with numerical calculations to enhance the current level of understanding of subsurface transport and give guidance to the design of extraction rates for simple vapor extraction configurations. The modeling approach is given in Section 3.

A series of subobjectives are defined to satisfy the two primary objectives of this research:

(a) Develop a set of models to describe the movement of aqueous and gaseous concentrations of volatile chemicals in laboratory columns of unsaturated soil.

(b) Design a set of column experiments to test the water-phase and air-phase transport aspects of the column models. These experiments are performed in an order that allows the impacts of different mechanisms to be observed separately and in combination. Two different porous materials are used

to compare chemical movement in cohesionless soils to movement in aggregated or structured soils.

(c) Based on the column models, derive and solve equations that describe the removal of volatile chemicals from soil by vapor extraction with either a trench or vertical vent system.

(d) The vapor extraction system models are used to examine the effect of air withdrawal rate on removal performance. In addition, the numerical calculations are used to give guidance for design and operation, and provide insight into the important removal mechanisms.

(e) Determine the applicability of using methods for developing column models to derive models that describe larger-scale systems. Examine the feasibility of this step in terms of equation derivations, simplifications, and solution methods.

This research focuses on studying simple, laboratory column systems and single-extraction vent configurations. The results are used to obtain a better understanding of the importance of physical transport (advection, dispersion, diffusion), phase transfer (liquid diffusion, volatilization/dissolution), and equilibrium (air-water, water-soil, air-soil) mechanisms and to provide a set of mathematical tools for interpreting laboratory and pilot-scale vapor extraction system results.

The complexity of the model developments is different for the various systems. Figure 2.1 shows the systems that are studied by model simulations herein. The one-dimensional, column models are used to examine the relative impacts of various mechanisms on the transport of volatile organic chemicals in unsaturated soils with air and water flow. Experimental results are used to validate the column models and observe the impacts of liquid dispersion, vapor diffusion, diffusion in immobile water, sorption, phase mass transfer, and Henry's partitioning in uniform and structured soils. The axially symmetric and two-dimensional models are used for numerical examination of the importance of

advection, dispersion, diffusion, volatilization, air-water (Henry's) partitioning, and sorption.

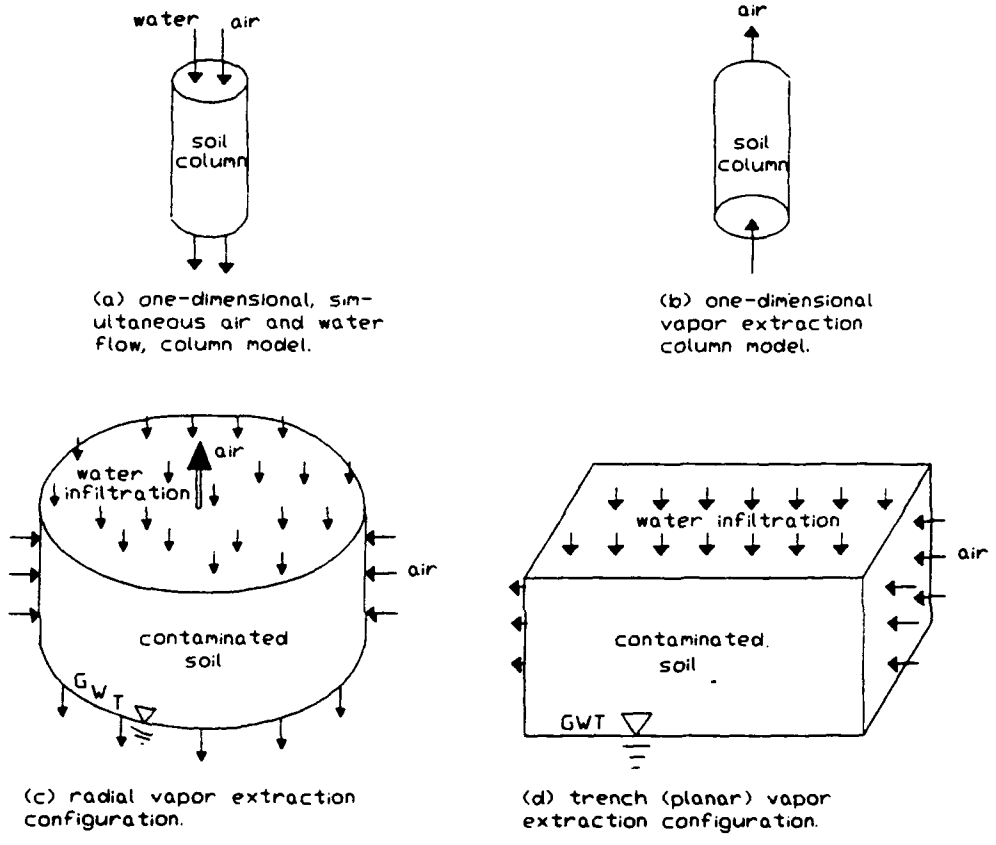


Figure 2.1. Schematics of the soil system configurations for which models are developed in this research.

A set of models is developed for each system studied. Each model within a set differs by the number of mechanisms that are included. The one-dimensional, column models include more mechanisms than the two-dimensional models because of the additional numerical complexity involved with solving the coupled equations for two-dimensional transport.

SECTION 3. MODELING APPROACH

This section describes an approach for constructing subsurface fate models. The approach is summarized in Table 3.1. The model development begins with construction of a conceptual picture of the system. Mathematical equations are derived to describe chemical transport. Dimensioned variables are normalized to obtain dimensionless equations. The system dimensionless of equations is then solved analytically when possible and numerically when not. When a numerical solution is necessary, the dimensionless equations are simplified to obtain analytic solutions for verification of the numerical code and for performing numerical sensitivity of the simplifying assumptions. The models are used to determine model parameter sensitivity, the results of which, in turn, are used to modify the conceptual picture, if necessary, and design experiments. Data from laboratory- and pilot-scale experiments are compared to model output for validation of the models. The validation results can then be used to modify the conceptual picture as required to best simulate the vapor extraction process.

Table 3.1. Approach Used for Constructing Subsurface Fate Models.

1. Construct a conceptual picture of the system.
2. Derive equations to describe the conceptual picture.
3. Convert equations to a dimensionless form.
4. Solve dimensionless equations.
5. Simplify equations and solve.
6. Verify the numerical solution by comparing to solutions of simplified conditions.
7. Model Sensitivity: Use the general model and its simplifications to determine which mechanisms, parameters, and simplifying assumptions are most important.
8. Validate the model and model sensitivity by comparing model predictions to experimental results.

The approach given in Table 3.1 is an integrated model

development-experimental study procedure that can enhance the current level of understanding of some of the subsurface chemical transport processes. It is similar to a generic approach detailed by Hern *et al.* [1986]. The modeling approach used here has been successfully applied to laboratory studies of organic chemical transport in saturated soil columns [Crittenden *et al.*, 1986; Hutzler *et al.*, 1986]. Their studies encompass model development, parameter sensitivity, and laboratory validation experiments. The column model developments reported here are complete in that all of the steps listed in Table 3.1 are performed. The two-dimensional model developments address, at least in part, steps 1-7. Even though the two-dimensional vapor extraction models developed here are not validated, they are useful for design and operation guidance and will be especially helpful for understanding removal processes. These vapor extraction model calculations should not be used in lieu of field- or pilot-scale experience.

3.1 CONCEPTUAL PICTURE

The first step in the development of a mathematical model is the conceptualizing of a system as a geometrically simpler picture of the interaction of transport and fate processes. There are numerous processes which could affect the fate of pollutants in unsaturated soil [cf. Donigian and Rao, 1986]. Table 1.1 showed that there are many potential variables to be included in a field-scale vapor extraction model. It is sometimes frustrating and inefficient to attempt to develop and validate a model that could describe the removal of all contaminants under a wide variety of conditions. The models developed in this research are constructed in a stepwise procedure. That is, the conceptual pictures begin simple, such as shown in Figure 2.1.a. The

next level of model sophistication is then based on the development of simpler models which have been satisfactorily solved and tested. By adding new complexities one at a time, their impact is observed separately. Moreover, increased confidence in the model assumptions and verification of the numerical solution is gained by comparing numerical calculations to other models.

3.2 DERIVATION OF THE DIMENSIONED EQUATIONS

A set of mathematical equations describing flow conditions and solute transport is developed from the conceptual picture. Mathematical equations are derived from physical and chemical principles, and the governing equations are solved with appropriate initial and boundary conditions.

3.3 NORMALIZATION OF THE DIMENSIONED VARIABLES

The governing equations are transformed into a dimensionless form by normalizing the dependent and independent variables. Two advantages result: (1) the equations are of a simpler form and the solutions are more general, and (2) the solutions can be characterized by the values of parameter groupings which are dimensionless and fewer in number than the total number of parameters. The normalization is a process by which linear substitutions of the normalizing coefficients and the dimensionless variables are made for the corresponding dimensioned quantities.

3.4 SOLUTION OF THE DIMENSIONLESS EQUATIONS

Typically, subsurface transport models are too complicated to be solved analytically, so numerical solution methods are employed. The numerical method employed in this work is a finite element technique, belonging to the family of weighted residual methods, called orthogonal

collocation [Finlayson, 1980]. Orthogonal collocation has been used often in the solution of column models [Raghavan and Ruthven, 1983; Crittenden *et al.*, 1986] and has only recently been applied to two-dimensional systems [Im, 1988]. This method is chosen here because of prior success at solving column equations and so its attributes can be evaluated for field scale equation applications. More details of the application of orthogonal collocation are given in the sections where it is applied to the equations. A brief mathematical justification for its use and an application example is given in Appendix B of a thesis by Gierke [1986]. After solving the general form of the different system models, the dimensionless equations are subsequently simplified by making assumptions about the importance of various mechanisms so that analytic solutions can then be used to verify the numerical solution method.

3.5 SIMPLIFICATION OF THE GOVERNING DIMENSIONLESS EQUATIONS

The general form of the dimensionless equations that describe a specific system is simplified by making assumptions about the impacts of different mechanisms. All of the analytic solutions assume linear sorption equilibrium. Three simplifying assumptions are used in this work to obtain analytic solutions: (1) local equilibrium (cf. Brusseau and Rao [1989]), (2) perfect or complete mixing, and (3) plug flow. The local equilibrium assumption is applied by assuming that volatilization and intraaggregate diffusion rates are fast in comparison to mass transport by advection and dispersion. Perfect mixing is employed in conjunction with the local equilibrium assumption, and it implies that dispersion and diffusion are fast enough to consider the contents of the system as completely mixed. Plug flow can be assumed either with

regards to water flow or air flow by ignoring liquid dispersion or gas diffusion, respectively. The local equilibrium assumption reduces the number of partial differential equations that describes a system, and given simplified boundary and initial conditions, it can lead to analytic solutions. Likewise, the assumption of complete mixing leads to a single mass balance and analytic solution for each system. The plug flow assumption allows an analytic solution for the column models. Different levels of complexity of the numerical solution codes result from the implementation of various combinations of these assumptions.

3.6 VERIFICATION OF THE NUMERICAL SOLUTION

Verification of a numerical solution is attained by comparing its calculations of a simplified condition to an analytic solution matching this condition. For example, by including fast rates of mass transfer the numerical approximation of advection (some call this "convection") and dispersion can be compared to analytic solutions of the convection-dispersion equation. Similarly, the numerical approximation of interfacial mass transfer or intraaggregate diffusion can be compared to a plug flow solution by inputting low values of the dispersion coefficients. All of the mechanisms included in the one-dimensional models were tested with at least one analytic solution. The two-dimensional numerical models could be compared to the perfect mixing solution or the convection-dispersion solution in one-dimension by ignoring either air or water flow.

The complexity of analytical solutions varies. A perfect mixing model with linear equilibrium relationships can be computed on a hand-held calculator. A one-dimensional, advection-dispersion model for finite boundary conditions involves the computation of an infinite

series [Hashimoto *et al.*, 1964]. Analytical solutions exist for some more complex models, but numerical solutions are sometimes faster [Crittenden *et al.*, 1986]. Analytic solutions are most useful for verifying the numerical solution of a more complex model.

3.7 VALIDATION OF THE MODEL

The validity of a model is determined by comparing calculations with independently determined parameters to experimental data. The validation step is divided into three sections: (1) experimental design, (2) parameter estimation, and (3) model predictions and fits of the experimental results. The experimental design involved constructing an apparatus, selecting porous media and chemicals, and determining the system conditions such as flow rates and degrees of saturation. Independently determined model parameters come from basic soil and chemical properties, direct measurements, literature correlations, and independent experiments such as batch rate studies. Model predictions are calculations using independently determined parameters. System specific characteristics or a lack of literature correlations for some parameters, such as dispersion coefficients, requires some fitting of model solutions to data. When model fits are performed then it is necessary to perform more experiments to validate the fit parameters. Column studies are performed to validate the column models. Experimental validation of the two-dimensional models is beyond the scope of this research.

SECTION 4. MODEL DEVELOPMENTS AND RESULTS

The model development approach described in Section 3 is applied to the one- and two-dimensional systems shown in Figure 2.1. For the one-dimensional systems, equations are derived where flow proceeds axially, such as depicted in Figures 2.1.a and 2.1.b (laboratory columns). The column models consider transport in either uniform or aggregated-porous materials. The two-dimensional systems assume horizontal air flow and vertical water flow. One set of two-dimensional models is derived for the cylinder of soil depicted in Figure 2.1.c where air is withdrawn uniformly along the axial center of the system (axially-symmetric, radially-converging flow), the other for the block shown in Figure 2.1.d.

A complete model development approach, such as described in Section 3, consists of conceptualizing the geometry of the system and the interaction of important process mechanisms, equation derivations, variable normalization, solution, simplification of the equations, solution verification, model sensitivity, and validation. This approach is performed in entirety for the column models. The two-dimensional models are not validated experimentally. The two-dimensional models consider fewer mechanisms than the column models due, in part, to computational convenience. Nevertheless, the two-dimensional model sensitivity results indicate important considerations for design of air flow rates and vent configuration, for further vapor extraction model developments, and for laboratory- and pilot-scale validation experiments.

The model developments reported here focus on describing chemical transport, and less attention is given to descriptions of air and water

flow. Subsurface heterogeneities, such as layering, are not included in these model studies. In addition, chemical and biological reactions that alter the form of a contaminant are ignored. Only dilute solutions are considered so that density effects could be ignored. Controlled experiments by McClellan and Gillham [1990] indicate that the removal of the organic phase of trichloroethene is advection dominant and that mass transfer limitations become important in the latter stages of cleanup where the dissolved component is being removed. The paragraphs that follow contain brief descriptions of the mechanisms that influence the migration of dilute solutions of nondegradable volatile organic chemicals in homogeneous, unsaturated soil systems and how the models consider these mechanisms.

Natural air flow in unsaturated soil is usually slower than water flow, but it is becoming common practice to induce higher air flows for removing volatile contaminants from soil [Bennedsen, 1987; Hutzler *et al.*, 1989a]. Because pressure drops associated with vapor extraction are small, air is treated as an incompressible fluid [Massmann, 1989; Croise *et al.*, 1989]. Massmann [1989] determined under what conditions it was appropriate to use ground-water flow models to predict air-pressure flow fields that result from the operation of a vapor removal system. Coupling air flow calculations with chemical transport and fate was beyond the scope of his study. Massmann [1989] determined that it was valid to assume incompressible flow of air for total pressure drops of less than 50%. This conclusion is supported with numerical studies by Croise *et al.* [1989].

Because gas diffusion is faster than liquid diffusion and gases are less viscous than liquids, changes in vapor concentrations can

develop pressure gradients that induce gas flow. Falta *et al.* [1989], Sleep and Sykes [1989], and Mendoza and Frind [1990] modeled the movement of gases due to density differences resulting from the evaporation of volatile contaminants. Their formulations showed that the impact of density-driven gas flow is greatest in high permeability soils (*i.e.* permeabilities greater than 0.1 darcies -- a sand and gravel mix has a permeability between 1 and 100 darcies [Massmann, 1989]) for contaminants having air-saturated densities greater than 150% that of from dry air (*e.g.* air saturated with carbon tetrachloride). Toluene vapor, which is only 10% heavier than dry air, shows no density affects unless the permeability is greater than 30 darcys. Their calculations are also based on situations where there is no external pressure-induced flow such as would occur with vapor extraction.

In the vapor extraction model derivations, air flow is assumed to occur horizontally either at a uniform rate (trench system) or converging radially to a vertical axis (vertical vent). The model developments also assume that the air is saturated with water so that moisture content remains constant.

Descriptions of unsaturated water flow are complicated by hysteresis and the variation of soil properties with location. In addition, dynamic environmental conditions such as atmospheric pressure, ambient temperature, and rainfall events alter the flow of water. Air flow also affects liquid permeability. Since one of the primary objectives of this study is to compare the relative impacts of mechanisms, it is advantageous to assume constant pressure, temperature, and wetting conditions and assume that soil properties and moisture profiles are uniform. Moreover, water and air flow are assumed to be

steady. Wierenga [1977] showed that when analyzing breakthrough curves of noninteracting solutes on the basis of cumulative drainage, transient and steady flow calculations give similar results.

Dispersion in porous media is attributed to the combination of two effects: molecular diffusion and fluid mixing. In water-saturated systems, both effects may contribute to the total dispersion flux. Gas diffusion is about 10,000 times faster than liquid diffusion and, thus, molecular diffusion is more important than the fluid mixing component in air at low flow (for volatile organic chemicals in air, low flow is defined as air velocities less than about 3 cm s^{-1}). The terms gas dispersion and gas diffusion will be used synonymously throughout this report, and gas dispersion will be considered to be independent of gas flow rate in the radially-converging extraction models.

Fickian-based diffusion is the most common mathematical description of liquid dispersion [van Genuchten and Jury, 1987]. This description is where the dispersive flux is the product of a concentration-independent dispersion coefficient and the concentration gradient. Liquid dispersion is a function of the moisture content and water velocity, and its impact in unsaturated soil is much greater than in saturated [DeSmedt *et al.*, 1986].

A common approach for describing gas diffusion is to use Fick's first law [Falta *et al.*, 1989; Thortenson and Pollock, 1989; Wilson *et al.*, 1988]. In unsaturated soil, gas dispersion is a function of the air-filled void tortuosity, which varies with the moisture content [Millington, 1959]. Thortenson and Pollock [1989] imply that subsurface gas diffusion may deviate from Fick's law. They describe calculations that include the effects of wall collisions and viscous and viscous-slip

flow for multicomponent mixtures. The processes they consider are important when describing gas transport in soils where a pure gas is mixing with a mixture or another pure gas. When the mole fraction of the diffusing gas is less than about 20%, mixing is primarily due to molecular diffusion, which can be described by Fick's law. This modeling effort assumes that concentrations are dilute.

Diffusion in immobile water has been shown to be important in saturated soil systems [Nkedi-Kizzi *et al.*, 1982; Crittenden *et al.*, 1986; Hutzler *et al.*, 1986; Roberts *et al.*, 1987]. The significance of intraaggregate diffusion in unsaturated soil is thought to be greater than in saturated systems [van Genuchten and Wierenga, 1977; van Genuchten *et al.*, 1977; DeSmedt *et al.*, 1986]. Diffusion in immobile water zones is described here as Fickian diffusion in saturated micropores contained within uniformly-sized, spherical aggregates [Rao *et al.*, 1982].

Mass transfer resistances exist at the boundaries of phases such as the air-water and water-soil interfaces and at the boundary of fluid layers traveling at different rates. Mass transfer resistance at the air-water interface could be an important mechanism in fate modeling of volatile pollutants in unsaturated soil [Sleep and Sykes, 1989]. Others have ignored its impact in subsurface transport by assuming rapid air-water equilibrium [Abriola and Pinder, 1985b; Lindstrom and Piver, 1986; Corapcioglu and Baehr, 1987; Mendoza and Frind, 1990]. Mass transfer resistance at the mobile-immobile water interface (film transfer) has been found by others to be unimportant in saturated soils [Crittenden *et al.*, 1986; Hutzler *et al.*, 1986; Roberts *et al.*, 1987]. Others have just simply dispensed with it [Abriola and Pinder, 1985b; Lindstrom and

Piver, 1986; Corapcioglu and Baehr, 1987; Sleep and Sykes, 1989] or combined it with intraaggregate diffusion [van Genuchten and Wierenga, 1977]. Nicoud and Schweih [1989] claim that it may be the predominant mechanism if the description of mobile-immobile mass transfer is performed correctly by accounting for the distribution of particle sizes instead of using the mean diameter as was done in work by others [Crittenden *et al.*, 1986; Hutzler *et al.*, 1986; Roberts *et al.*, 1987]. Air-water and mobile-immobile water mass transfer are included in the column models to test their impact on chemical movement in laboratory columns. Mathematically they are treated following film theory [*cf.* Finlayson, 1980]. The resistance to mass transfer at the soil-water interface is considered here to be small compared to the diffusive resistance of chemicals in the liquid phase.

Retardation or the rate of movement of a chemical profile relative to the rate of water or air movement is a function of the solute phase distribution at equilibrium. Sorption tends to retard chemical movement. Organic chemicals can sorb to soil either by mineral adsorption or partitioning (dissolution) into soil organic material [Burchill *et al.*, 1981; Hamaker and Thompson, 1972]. Mineral adsorption in water-saturated situations is usually negligible for nonpolar species because of the competition of polar water molecules [Hassett *et al.*, 1983]. The equilibrium relationship is linear for chemicals that dissolve from aqueous solution into soil organic material [Burchill *et al.*, 1981]. That is, water-phase concentrations are proportional to the sorbed-phase concentrations. In saturated soil, most sorption of nonpolar organics onto soil is linear [Hassett *et al.*, 1983; Karickhoff *et al.*, 1979; Hutzler *et al.*, 1986]. Freundlich adsorption equilibrium

[Freundlich, 1922] is used in the column models to allow for nonlinear adsorption equilibrium. Nonlinear adsorption of organic vapors has been observed for dry soils [Chiou and Shoup, 1985]. The column models account for nonlinear adsorption, and the vapor extraction models assume linear partitioning to reduce computation times and because sorption is probably linear in moist natural soils. For moist conditions, it is assumed that the soil particles are completely covered by water, even if only by several molecular layers, so that adsorption is described by an aqueous-phase isotherm equation. When liquid water is present, adsorption of vapors onto soil is probably unimportant [Roy and Griffin, 1987] because soil surfaces are usually covered by water. Others have measured moisture impacts on sorption [Chiou and Shoup, 1985; Peterson *et al.*, 1988]. Peterson *et al.* [1988] measured the sorption of trichloroethene vapors on a laboratory-derived aluminum silicate material with a humic acid coating. They measured moisture effects up to moisture contents of 12%; however, the material that they used had a surface area (200 $m^2\ g^{-1}$), which is comparable to granular activated carbon (800 $m^2\ g^{-1}$). The moisture contents they achieved probably did not coat the entire surface of the porous material and so vapor sorption would occur. Granular activated carbon also shows moisture effects for moisture contents lower than 30% [Okazaki *et al.*, 1978]. Granular soils have much lower surface areas (30 $m^2\ g^{-1}$) and, thus, will probably be covered entirely by water at moisture contents where water is present as a liquid. Liquid water was not present in the studies of Chiou and Shoup [1985], and it is not apparent whether liquid water was present in the experiments reported by Peterson *et al.* [1988].

Because these model developments assume that solution

concentrations are dilute, air-water equilibrium is described by Henry's law. That is, a linear relationship exists between the water-phase and air-phase concentrations which is analogous to the linear sorption equilibrium. Most volatile organic chemicals of environmental concern are only slightly soluble in water.

The remainder of this section incorporates the mechanisms described above into models for describing chemical transport in laboratory columns and models for predicting the performance of single-extraction vent systems. The column models and assumptions regarding the mechanisms as discussed above are tested against experimental results. The vapor extraction models are used to examine the applicability of extending the approach used for developing column models to field-scale model development and for comparing the impacts of the transport mechanisms, the air withdrawal rate, and vent configuration on the performance of vapor extraction.

4.1 ONE-DIMENSIONAL TRANSPORT MODELS

The important fate mechanisms are studied numerically with one-dimensional transport models and by experimental observation with laboratory columns. The column models account for water as the only flowing medium or for air and water flow and for air flow alone. The simultaneous flow model is derived and tested first and followed by a similar description of the air flow or vapor extraction column model. The column model sections parallel the outline given in the approach section.

4.1.1 Column Model for Air and Water Flow

4.1.1.1 Conceptual Picture. Figure 4.1.1.1 is a conceptual picture of a soil column that is used for the model development that

follows. The mechanisms being considered here are: (a) air and water advection, (b) liquid and gas dispersion in the direction of flow, (c) liquid diffusion in pores filled with immobile water, (d) mass transfer resistance at the air-water and the mobile-immobile water interfaces, (e) partitioning among the air and water, and (f) nonlinear sorption from aqueous solution.

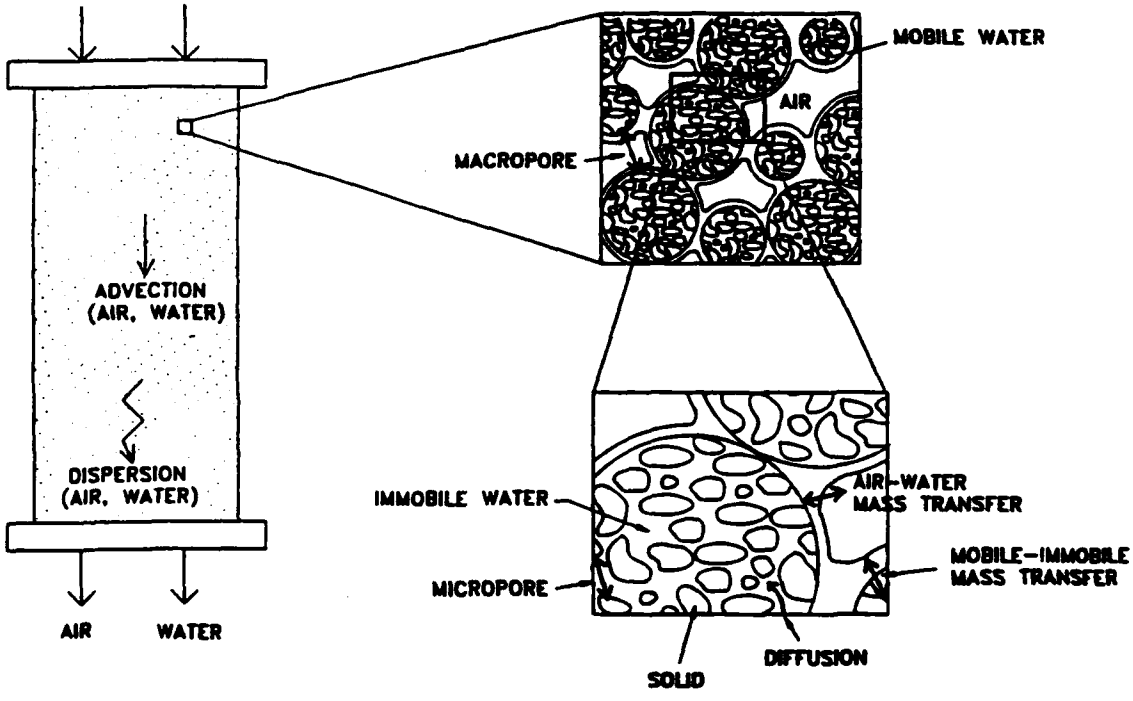


Figure 4.1.1.1. Conceptual picture for modeling transport in soil columns with air and water flow.

The model is developed to describe the movement of a volatile organic contaminant in laboratory columns of unsaturated soil where the mechanisms described above are operative. The soil system is divided into the three zones shown in Figure 4.1.1.1: mobile or immobile air, mobile water, and aggregates comprised of immobile water and solid soil particles. Mass balances on these zones result in three partial differential equations.

The air and mobile-water phases depicted in Figure 4.1.1.1 are assumed to be continuous. The smallest pores, such as those inside aggregates, contain water that is immobile. If the water content exceeds the field capacity of the soil, water will flow around the aggregates through the larger pores while the largest pore spaces are filled with air. Since there is no theoretical method for predicting the fraction of interphase contact between air and immobile water and between air and mobile water, the rate of air-water mass transfer is assumed to be lumped into a single transfer rate between air and mobile water. Therefore it is assumed that chemical phase transfer occurs between the air and mobile water and the mobile and immobile water. It is also assumed that water in contact with soil surfaces inside an aggregate is immobile.

4.1.1.2 Derivation of the Dimensioned Equations. A mass balance on the air zone results in the following equation:

$$\frac{\partial C_v(Z,T)}{\partial T} = E_v \frac{\partial^2 C_v(Z,T)}{\partial Z^2} - u \frac{\partial C_v(Z,T)}{\partial Z} - \frac{K_L a}{\epsilon(1-S)} \left[\frac{C_v(Z,T)}{H} - C_b(Z,T) \right] \qquad (4.1.1.1)$$

Equation (4.1.1.1) accounts for the change in vapor concentration ($C_v(Z,T)$) with respect to time. The terms on the right side represent: gas dispersion or diffusion, gas advection, and air-water mass transfer. The notation used in the model equations is defined in Section 7.

A mass balance on the mobile-water zone results in a similar expression for the change in chemical concentration in mobile water:

$$\frac{\partial C_b(Z,T)}{\partial T} = E_z \frac{\partial^2 C_b(Z,T)}{\partial Z^2} - v \frac{\partial C_b(Z,T)}{\partial Z} + \frac{K_L a}{\epsilon(S-S_i)} \left[\frac{C_v(Z,T)}{H} - C_b(Z,T) \right]$$

$$+ \frac{3k_f[1-\epsilon(1-S_i)]}{\epsilon(S-S_i)R_a} [C_p(R=R_a,Z,T) - C_b(Z,T)] \qquad (4.1.1.2)$$

The terms on the right side of (4.1.1.2) represent: liquid dispersion, liquid advection, mass transfer between air and mobile water, and film transfer. The mathematical representation of mass transfer between the mobile and immobile water in (4.1.1.2) assumes that an aggregate centered at axial position Z absorbs chemical at a rate proportional to the deficit between the immobile water concentration at the aggregate surface ($C_p(R=R_a,Z,T)$) and the mobile-water concentration ($C_b(Z,T)$). This representation of film transfer is appropriate when the axial water concentration gradient ($\partial C_b/\partial Z$) is small across an aggregate diameter.

Figure 4.1.1.1 shows that soil aggregates are represented by uniformly-porous spheres [Rao *et al.*, 1982] within which the aqueous chemical concentration in the micropores is at equilibrium with the sorbed-phase concentration on the adjacent soil surface. The change in the total intraaggregate concentration ($Y(R,Z,T)$) with respect to time is equal to the rate at which chemical diffuses through the internal pores:

$$\frac{\partial Y(R,Z,T)}{\partial T} = \frac{1}{R^2} \frac{\partial}{\partial R} \left[\frac{D_p \epsilon S_i R^2}{\rho_s(1-\epsilon)} \frac{\partial C_p(R,Z,T)}{\partial R} \right] \qquad (4.1.1.3)$$

where: $$Y(R,Z,T) = \frac{\epsilon S_i}{\rho_s(1-\epsilon)} C_p(R,Z,T) + Q(R,Z,T)$$

Like (4.1.1.2), (4.1.1.3) assumes that changes in axial water concentration are negligible over an aggregate diameter. The variable

Y(R,Z,T) is the total chemical concentration at a specific radial position within an aggregate per unit mass of soil. Q(R,Z,T) is the mass of chemical sorbed per unit mass of soil. The equilibrium between the aqueous and sorbed phases within the micropores is described with the Freundlich [1922] isotherm equation:

$$Q(R,Z,T) = KC_p(R,Z,T)^{1/n} \qquad (4.1.1.4)$$

The initial condition for solving the above equations can be any specified concentration profile within the column. Typically for soil columns the initial concentrations are zero:

$C_v(Z,T=0) = C_{vi}(Z)$, typically $C_{vi}(Z) = 0$;

$C_b(Z,T=0) = C_{bi}(Z)$, typically $C_{bi}(Z) = 0$;

$C_p(R,Z,T=0) = C_{pi}(R,Z)$, typically $C_{pi}(R,Z) = 0$;

and

$Q(R,Z,T=0) = Q_i(R,Z) = KC_{pi}(R,Z)^{1/n}$,

thus

$$Y(R,Z,T=0) = \frac{\epsilon S_i}{\rho_s(1-\epsilon)} C_{pi}(R,Z) + Q_i(R,Z) \qquad (4.1.1.5)$$

Boundary conditions for (4.1.1.1) and (4.1.1.2) are derived from the fact that soil columns are closed reactors [Levenspiel, 1962]. It will be shown later that analytical solutions of the model exist for simplified conditions; however, for the general case shown, a numerical solution method is necessary. In an attempt to force the numerical method to conserve chemical mass, overall mass balances on the mobile-water and air phases are used for two of the boundary conditions. The difference between the mass of chemical entering and leaving the column

by advection in air must equal the mass accumulating in the air minus the mass transferred to the air from the water:

$$u[C_{vo}(T) - C_v(Z=L,T)] = \frac{\partial}{\partial T}\int_0^L C_v(Z,T)\partial Z - \int_0^L \frac{K_L a}{\epsilon(1-S)}\left[C_b(Z,T) - \frac{C_v(Z,T)}{H} \right] \partial Z \qquad (4.1.1.6)$$

The difference in chemical mass entering and leaving the column in water must equal the accumulation in the water and in the aggregates plus the mass transferred to the air from the water:

$$v[C_{bo}(T) - C_b(Z=L,T)] = \frac{\partial}{\partial T}\int_0^L \left[C_b(Z,T) + \frac{3\rho_s(1-\epsilon)}{\epsilon(S-S_i)R_a^3}\int_0^{R_a} Y(R,Z,T)R^2 dR \right] \partial Z + \int_0^L \frac{K_L a}{\epsilon(S-S_i)}\left[C_b(Z,T) - \frac{C_v(Z,T)}{H} \right] \partial Z \qquad (4.1.1.7)$$

The influent concentration in the water ($C_{bo}(T)$) is also allowed to vary with time.

Exit boundary conditions for (4.1.1.1) and (4.1.1.2) are the following zero-gradient conditions:

$$\frac{\partial^2 C_v(Z=L,T)}{\partial Z \partial T} = 0 \qquad (4.1.1.8)$$

$$\frac{\partial^2 C_b(Z=L,T)}{\partial Z \partial T} = 0 \qquad (4.1.1.9)$$

One boundary condition for (4.1.1.3) results from symmetry; that is, no concentration gradient exists at the center of an aggregate:

$$\frac{\partial C_p(R=0,Z,T)}{\partial R} = 0 \qquad (4.1.1.10)$$

Symmetry is based on the assumptions of spherical aggregates and that axial changes in chemical concentration are small across an aggregate diameter.

The other boundary condition for (4.1.1.3) is derived by performing a mass balance on an aggregate. A change in mass of chemical in an aggregate is equal to the mass transferred to the aggregate from the mobile water:

$$\frac{\partial}{\partial T}\int_0^{R_a} Y(R,Z,T)R^2dR = \frac{k_f R_a^2[1-\epsilon(1-S_i)]}{\rho_s(1-\epsilon)}[C_b(Z,T) - C_p(R=R_a,Z,T)] \qquad (4.1.1.11)$$

Equation (4.1.1.11) is consistent with the assumption used to represent film transfer in (4.1.1.2). Like (4.1.1.6) and (4.1.1.7), (4.1.1.11) attempts to conserve mass during the numerical solution of the model equations.

4.1.1.3 Conversion to Dimensionless Form. To reduce the complexity of the equations, so that model solutions could be based on and characterized by fewer parameters, the dimensioned equations derived above are converted to a dimensionless form. Dimensionless equations result from substitutions of the terms in the middle column of Table 4.1.1.1 into (4.1.1.1-11). Dimensionless time or throughput (t) is defined by assuming that a soil column is initially free of chemical and that the influent concentrations are constant and in equilibrium ($C_{bo}(T) = C_{bn}$, $C_{vo}(T) = HC_{bn}$). Throughput is equal to the ratio of chemical mass fed to the mass contained in the column at equilibrium with C_{bn}.

Dimensionless concentrations are derived by dividing a particular phase concentration by its concentration in equilibrium with C_{bn}. Axial position (Z) is normalized by the column length (L), and radial position (R), by the aggregate radius (R_a).

Table 4.1.1.1. Variable Substitutions to Convert Dimensioned Equations into a Dimensionless Form.

Dimensioned Variable	Substitution	Dimensionless Variable
$C_b(Z,T)$	$C_{bn}\ c_b(Z,T)$	$c_b(Z,T)$
$C_p(R,Z,T)$	$C_{bn}\ c_p(R,Z,T)$	$c_p(R,Z,T)$
$C_v(Z,T)$	$C_{vn}\ c_v(Z,T)$	$c_v(Z,T)$
$Y(R,Z,T)$	$Y_n\ y(R,Z,T)$	$y(R,Z,T)$
$C_{bi}(Z)$	$C_{bn}\ c_{bi}(Z)$	$c_{bi}(Z)$
$C_{pi}(R,Z)$	$C_{bn}\ c_{pi}(R,Z)$	$c_{pi}(R,Z)$
$C_{vi}(Z)$	$C_{vn}\ c_{vi}(Z)$	$c_{vi}(Z)$
$C_{bo}(T)$	$C_{bn}\ c_{bo}(T)$	$c_{bo}(T)$
$C_{vo}(T)$	$C_{vn}\ c_{vo}(T)$	$c_{vo}(T)$
Z	$L\ z$	z
R	$R_a\ r$	r
T	$\frac{v}{L}\left[1+\frac{u(1-S)H}{v(S-S_i)}\right]\left[1+\frac{S}{S-S_i}+\frac{(1-S)H}{S-S_i}+\frac{\rho_s(1-\epsilon)KC_{bn}^{1/n-1}}{\epsilon(S-S_i)}\right]^{-1} t$	t

Note: $C_{vn} = HC_{bn}$ and $Y_n = \epsilon S_i C_{bn}[\rho_s(1-\epsilon)]^{-1} + KC_{bn}^{1/n}$, where $C_{bn} = \max(C_{bo}(T), C_{bi}(Z))$.

Column model predictions in terms of relative (dimensionless)

concentration as a function of dimensionless time (throughput) are characterized by the groups defined in Table 4.1.1.2. These groups represent mass transfer mechanisms and chemical distributions at equilibrium. Because water is the most common transport medium in soil, the mass transfer groups are based on the rate of mass transport by water advection, and the chemical distribution groups are based on chemical mass in mobile water. In the design and operation of vapor extraction, water flow is minimized while air flow is induced. Therefore in the model development sections following these sections on simultaneous air and water transport, mass transfer and transport rates and equilibrium are normalized by air advective rates and mass of chemical in air. The magnitudes of five of the mass transfer groups (air Peclet (Pe_v), mobile-water Peclet (Pe_b), immobile-water diffusion modulus (Ed_p), mobile-immobile water Stanton (St_b), air-water Stanton (St_v)) represent the degree of spreading exhibited by a breakthrough curve [Crittenden *et al.*, 1986]. A large value of any of these groups indicates a small contribution from the corresponding mechanism towards the observed spreading. For example, a large value of Pe_b means that transport by liquid dispersion is slow in comparison to that by water advection, and therefore liquid dispersion is not important. An increase in the air-water advective flux ratio (Ar) has the same effect as increasing Pe_b and Pe_v and decreasing St_b, St_v, and Ed_p. The chemical distribution groups (immobile water (Dg_p), sorbed (Dg_s), vapor (Dg_v)) and the isotherm intensity (1/n) impact spreading because they determine the amount of chemical in a given phase. Only Ar, Dg_s, Dg_v, and 1/n affect the magnitude of the retardation coefficient (R_d). If 1/n is not equal to 1, then R_d is also dependent on concentration.

Table 4.1.1.2. Definitions of Dimensionless Groups.

Group	Definition	Equation
Mass Transfer Groups:		
Ar	$\frac{\text{rate of advection in air}}{\text{rate of advection in mobile water}}$	$\frac{u}{v} - Dg_v$
Ed_p	$\frac{\text{rate of diffusion in immobile water}}{\text{rate of advection in mobile water}}$	$\frac{D_p Dg_p L}{vR_a^2}$
Pe_b	$\frac{\text{rate of advection in mobile water}}{\text{rate of axial dispersion in mobile water}}$	$\frac{vL}{E_z}$
Pe_v	$\frac{\text{rate of advection in mobile water}}{\text{rate of axial dispersion in air}}$	$\frac{vL}{E_v Dg_v}$
Pe	$\frac{\text{total advective flux in air and in water}}{\text{total dispersive flux in air and in water}}$	$\frac{1+Ar}{Pe_b^{-1} + Pe_v^{-1}}$
St_b	$\frac{\text{rate of transport across mobile-immobile water interface}}{\text{rate of advection in mobile water}}$	$\frac{k_f[1-\epsilon(1-S_i)]L}{v\epsilon(S-S_i)R_a}$
St_v	$\frac{\text{rate of transport across air-mobile water interface}}{\text{rate of advection in mobile water}}$	$\frac{K_L aL}{3v\epsilon(S-S_i)}$
Chemical Distribution Groups:		
Dg_p	$\frac{\text{mass of chemical in immobile water}}{\text{mass of chemical in mobile water}}$	$\frac{S_i}{S-S_i}$
Dg_s	$\frac{\text{mass of chemical adsorbed to soil}}{\text{mass of chemical in mobile water}}$	$\frac{\rho_s(1-\epsilon)KC_{bn}^{(1/n-1)}}{\epsilon(S-S_i)}$
Dg_v	$\frac{\text{mass of chemical in air}}{\text{mass of chemical in mobile water}}$	$\frac{(1-S)H}{S-S_i}$
Dg	$\frac{\text{mass of chemical in air, in immobile water, and on soil}}{\text{mass of chemical in mobile water}}$	$Dg_p + Dg_s + Dg_v$
R_d	$\frac{\text{velocity of chemical front}}{\text{velocity of water}}$	$\frac{1+Dg}{(1+Ar)(1+Dg_p)}$
1/n	isotherm intensity	

The dimensionless forms of the air mass balance (equation (4.1.1.1)) and its boundary conditions (equations (4.1.1.6) and (4.1.1.8)) are

$$\frac{\partial c_v(z,t)}{\partial t} = \frac{1 + Dg}{Dg_v[1 + Ar]} \left[\frac{1}{Pe_v} \frac{\partial^2 c_v(z,t)}{\partial z^2} - Ar \frac{\partial c_v(z,t)}{\partial z} - 3St_v[c_v(z,t) - c_b(z,t)] \right] \quad (4.1.1.12)$$

$$c_{vo}(t) - c_v(z=1,t) = \frac{Dg_v[1 + Ar]}{Ar[1 + Dg]} \frac{\partial}{\partial t} \int_0^1 c_v(z,t)\partial z - \int_0^1 \frac{3St_v}{Ar} [c_b(z,t) - c_v(z,t)]\partial z \quad (4.1.1.13)$$

$$\frac{\partial^2 c_v(z=1,t)}{\partial z \partial t} = 0 \quad (4.1.1.14)$$

In dimensionless form, the mobile-water mass balance (equation (4.1.1.2)) and its boundary conditions (equations (4.1.1.7) and (4.1.1.9)) become

$$\frac{\partial c_b(z,t)}{\partial t} = \frac{1 + Dg}{1 + Ar} \left[\frac{1}{Pe_b} \frac{\partial^2 c_b(z,t)}{\partial z^2} - \frac{\partial c_b(z,t)}{\partial z} + 3St_v[c_v(z,t) - c_b(z,t)] + 3St_b[c_p(r=1,z,t) - c_b(z,t)] \right] \quad (4.1.1.15)$$

$$c_{bo}(t) - c_b(z=1,t) = \frac{1 + Ar}{1 + Dg} \frac{\partial}{\partial t} \int_0^1 \left[c_b(z,t) + 3[Dg_p + Dg_s] \int_0^1 y(r,z,t) r^2 dr \right] \partial z + \int_0^1 3St_v[c_b(z,t) - c_v(z,t)]\partial z \quad (4.1.1.16)$$

$$\frac{\partial^2 c_b(z=1,t)}{\partial z \partial t} = 0 \quad (4.1.1.17)$$

The dimensionless form of the intraaggregate mass balance

(equation (4.1.1.3)) is

$$\frac{\partial y(r,z,t)}{\partial t} = \frac{Ed_p[1 + Dg]}{[Dg_p + Dg_s][1 + Ar]} \frac{1}{r^2} \frac{\partial}{\partial r} \left[r^2 \frac{\partial c_p(r,z,t)}{\partial r} \right] \qquad (4.1.1.18)$$

The dimensionless total intraaggregate concentration must satisfy

$$y(r,z,t) = \frac{Dg_p c_p(r,z,t) + Dg_s c_p(r,z,t)^{1/n}}{Dg_p + Dg_s} \qquad (4.1.1.19)$$

Equation (4.1.1.18) is obtained by substituting (4.1.1.4) into the definition of Y(R,Z,T) and dividing the result by the total intraaggregate concentration in equilibrium with C_{bn}. For 1/n equal to 1, (4.1.1.19) reduces to $y(r,z,t) = c_p(r,z,t)$. The dimensionless form of the boundary conditions (equations (4.1.1.10) and (4.1.1.11)) for the intraaggregate mass balance are

$$\frac{\partial c_p(r=0,z,t)}{\partial r} = 0 \qquad (4.1.1.20)$$

$$St_b[c_b(z,t) - c_p(r=1,z,t)] = \frac{[Dg_p + Dg_s][1 + Ar]}{[1 + Dg]} \frac{\partial}{\partial t} \int_0^1 y(r,z,t) r^2 dr \qquad (4.1.1.21)$$

The dimensionless initial condition for solving (4.1.1.12), (4.1.1.15), and (4.1.1.18) is

$$c_v(z,t=0) = c_{vi}(z);\ c_b(z,t=0) = c_{bi}(z);\ \text{and}$$
$$c_p(r,z,t=0) = c_{pi}(r,z) \qquad (4.1.1.22)$$

The initial condition is set by specifying $c_{vi}(z)$, $c_{bi}(z)$, and $c_{pi}(r,z)$.

Equation (4.1.1.19) is used to determine $y(r,z,t=0)$. Typically, a continuous concentration profile is specified, that is $c_{vi}(z) = c_{bi}(z) = c_{pi}(r=1,z)$; and for an initial uniform concentration throughout, $c_{vi}(0\leq z\leq 1) = c_{bi}(0\leq z\leq 1) = c_{pi}(0\leq r\leq 1, 0\leq z\leq 1) = 0$ (for breakthrough or contamination) or 1 (for elution or cleanup).

Converting the model into a dimensionless form reduces the number of parameters that characterize a solution from 17 (a, D_p, E_v, E_z, H, K, k_f, K_L, L, 1/n, R_a, S, S_i, u, v, ϵ, ρ_s) to 10 (Ar, Dg_p, Dg_s, Dg_v, Ed_p, 1/n, Pe_b, Pe_v, St_b, St_v). It is also easier to characterize a solution in terms of these groups. Five groups (Ed_p, Pe_b, Pe_v, St_b, St_v) affect only the shape of a breakthrough curve. The impact of Ar, Dg_p, Dg_s, and Dg_v is primarily on position. Nine dimensioned parameters (H, K, L, 1/n, S, u, v, ϵ, ρ_s) impact both shape and position; the other eight affect only the shape.

4.1.1.4 Numerical Solution. The dimensionless form of the column model (equations (4.1.1.12-22)) is solved numerically by converting the partial differential equations (PDEs) to a system of ordinary differential equations (ODEs). Orthogonal collocation (OC), a method of weighted residuals, lends itself well to converting similar types of PDEs to systems of ODEs [Raghavan and Ruthven, 1983; Crittenden et al., 1986]. The resulting set of ODEs can then be solved by a number of standard techniques.

Raghavan and Ruthven [1983] used OC to solve equations similar to those comprising the general model except that a different inlet condition was used in their formulation. The inlet boundary condition they imposed required an iterative solution method which involved guessing the inlet concentration. The boundary conditions employed here

(equations (4.1.1.13), (4.1.1.16), and (4.1.1.21)) avoid the iterative step, and they help conserve chemical mass during the process of numerically solving the model.

Weighted residual methods allow separation of the time and spatial dependency of a PDE by approximating the exact solution with a series of products of time-varying coefficients and spatial basis or trial functions. The collocation method requires that the residual between the numerical approximation of the PDE and its exact value be orthogonal to the Dirac delta function at specified collocation points. This results in the residuals being zero at the collocation points [Finlayson, 1980].

Orthogonal collocation employs orthogonal polynomials as basis functions and specifies that the collocation points be located at the roots of a basis function. The polynomials are constructed orthogonal to each other with respect to a weight function. The weight functions used in the construction of the polynomials for the different equations are chosen to satisfy boundary conditions and/or to make the numerical solution stable.

Application of OC to the air and mobile-water mass balances and the corresponding boundary conditions (equations (4.1.12)-(4.1.1.17)) yields 2J ODEs, where J is the number of axial collocation points. Additional ODEs (JxI, where I is the number of radial collocation points) are produced by the application of OC to the intraaggregate mass balance and its boundary conditions (equations (4.1.1.18), (4.1.1.20), and (4.1.1.21)). Figure 4.1.1.2 is a general schematic of the OC descretization of the solution domain and shows the coupling of the ODEs. The resulting system of ODEs is solved using an algorithm called

GEAR, which can be found in the International Mathematics and Scientific Library (IMSL). The application of OC is shown below in the order in which GEAR receives the derivatives.

The application of OC to (4.1.1.18) results in:

$$\frac{dy(i,j,t)}{dt} = \frac{Ed_p[1 + Dg]}{[Dg_p + Dg_s][1 + Ar]} \sum_{n=1}^{I} B^r{}_{i,n}\, c_p(n,j,t) \qquad (4.1.1.23)$$

Figure 4.1.1.2 shows that (4.1.1.23) is evaluated at I-1 radial collocation points at each axial collocation point (j = 1 to J). $B^r{}_{i,n}$ is a member of an OC coefficient matrix for spherical geometry that is used to approximate the Laplacian of $c_p(r,z,t)$. This matrix is constructed from a set of symmetric Jacobi polynomials that represent the radial dependence of $c_p(r,z,t)$. The radial orthogonal polynomials are constructed with only even powers of r up to degree 2I using a weight function of $1-r^2$ over the interval of r from 0 to 1. The internal radial collocation locations shown in Figure 4.1.1.2 are the positive roots of the 2(I-1) degree polynomial and lie between 0 and 1. Because the matrix is symmetrical, (4.1.1.20) is satisfied by the application of OC to (4.1.1.18) [Finlayson, 1980].

Applying OC to (4.1.1.21) and solving for the change in the total intraaggregate concentration at r equal to 1 leads to the following condition at the aggregate surface:

$$\frac{dy(I,j,t)}{dt} = \frac{1}{W^r{}_I}\left[\frac{[1 + Dg]St_b}{[Dg_p + Dg_s][1 + Ar]}[c_b(j,t) - c_p(I,j,t)] - \sum_{p=1}^{I-1} W^r{}_p \frac{dy(p,j,t)}{dt}\right] \qquad (4.1.1.24)$$

Equation (4.1.1.24) is evaluated at all axial collocation locations. W^r_p is a member of a coefficient vector that is used in the quadrature approximation of the radial integrals. W^r_I is nonzero because a weight factor of $1-r^2$ is used for the construction of the radial basis functions [Finlayson, 1980].

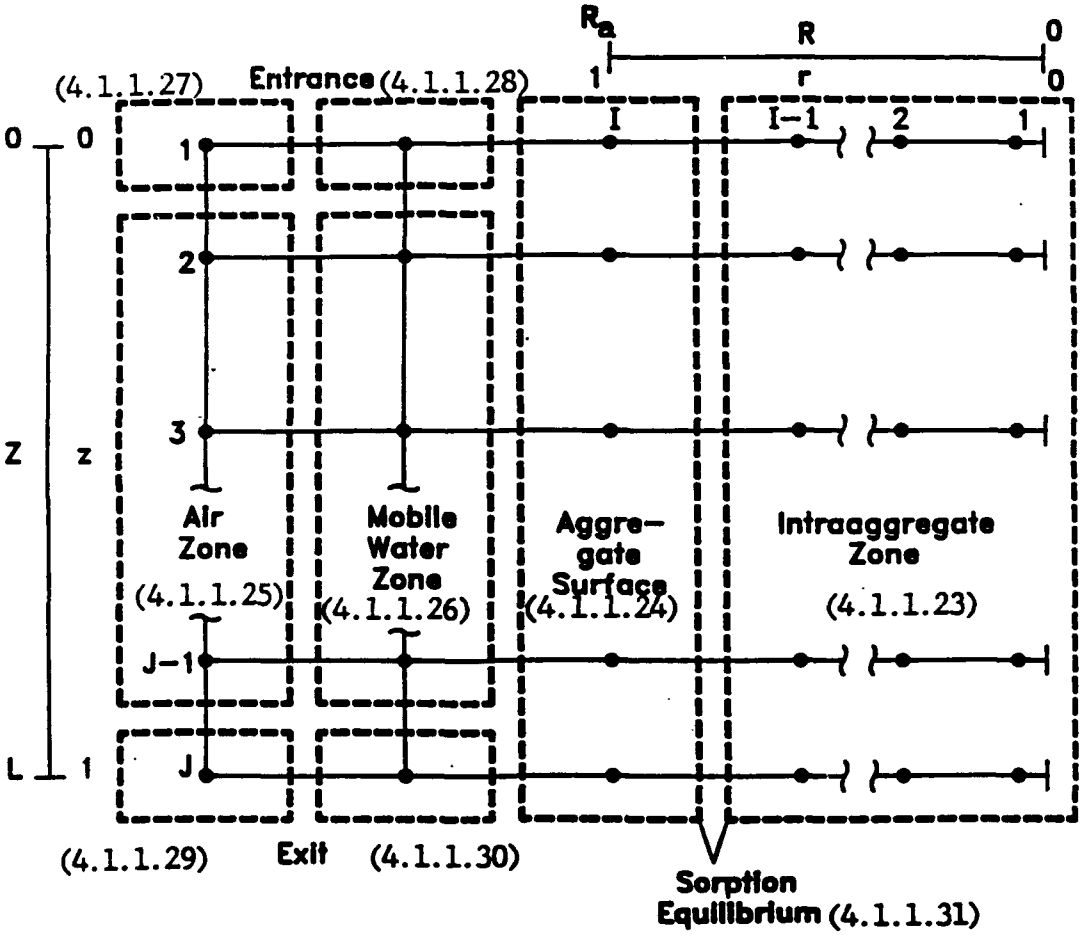


Figure 4.1.1.2. General schematic showing the coupling of the ordinary differential equations resulting from the application of orthogonal collocation to the partial differential equations comprising the column model. The number of radial points is I and number of axial is J. Equation numbering is in parentheses.

The application of OC to the air and mobile-water mass balances (equations (4.1.1.12) and (4.1.1.15)) gives the following equations:

$$\frac{dc_v(j,t)}{dt} = \frac{1 + Dg}{[1 + Ar]Dg_v} \left\{ \sum_{m=1}^{J} \left[\frac{B^z{}_{j,m}}{Pe_v} - Ar\ A^z{}_{j,m} \right] c_v(m,t) \right.$$

$$\left. - 3St_v[c_v(j,t) - c_b(j,t)] \right\} \qquad (4.1.1.25)$$

$$\frac{dc_b(j,t)}{dt} = \frac{1 + Dg}{1 + Ar} \left\{ \sum_{m=1}^{J} \left[\frac{B^z{}_{j,m}}{Pe_b} - A^z{}_{j,m} \right] c_b(m,t) \right.$$

$$\left. + 3St_v[c_v(j,t) - c_b(j,t)] + 3St_b[c_b(j,t) - c_p(I,j,t)] \right\} \qquad (4.1.1.26)$$

Equations (4.1.1.25) and (4.1.1.26) are evaluated at the J-2 internal axial collocation locations shown in Figure 4.1.1.2 (j = 2 to J-1). $A^z{}_{j,m}$ and $B^z{}_{j,m}$ are members of OC coefficient matrices for planar geometry that are used to approximate the first and second spatial derivatives, respectively, of $c_b(z,t)$ and $c_v(z,t)$. These matrices are obtained from a set of asymmetric Jacobi polynomials that represent the axial dependence of $c_b(z,t)$ and $c_v(z,t)$. The axial orthogonal polynomials are constructed up to degree J-1 using a weight factor of z(1-z) over the interval 0 to 1. The locations of the internal axial collocation points shown in Figure 4.1.1.2 are the roots of the J-2 degree orthogonal polynomial. The boundary conditions for (4.1.1.12) and (4.1.1.15) are located at j equal to 1 and J.

Entrance conditions are obtained by using OC to convert (4.1.1.13), (4.1.1.14), (4.1.1.16), and (4.1.1.17) to ODEs and solving for the derivatives at j equal to 1:

$$\frac{dc_v(1,t)}{dt} = \left[W^z_1 - W^z_J \frac{A_{J,1}{}^z}{A_{J,J}{}^z} \right]^{-1} \left\{ \frac{Ar[1 + Dg]}{Dg_v[1 + Ar]} \left[[c_{vo}(t) - c_v(J,t)] + \sum_{m=1}^{J} \frac{St_v\ W^z_m}{Ar} [c_b(m,t) - c_v(m,t)] \right] - \sum_{m=2}^{J-1} \left[W^z_m - W^z_J \frac{A^z_{J,m}}{A^z_{J,J}} \right] \frac{dc_v(m,t)}{dt} \right\} \quad (4.1.1.27)$$

$$\frac{dc_b(1,t)}{dt} = \left[W^z_1 - W^z_J \frac{A^z_{J,1}}{A^z_{J,J}} \right]^{-1} \left\{ \frac{1 + Dg}{1 + Ar} \left[[c_{bi}(t) - c_b(J,t)] - \sum_{m=1}^{J} St_v\ W^z_m\ [c_b(m,t) - c_v(m,t)] \right] - \sum_{m=2}^{J-1} \left[W^z_m - W^z_J \frac{A^z_{J,m}}{A^z_{J,J}} \right] \frac{dc_b(m,t)}{dt} - 3[Dg_p + Dg_s] \sum_{m=1}^{J} W^z_m \sum_{p=1}^{I} W^r_p \frac{dy(p,m,t)}{dt} \right\} \quad (4.1.1.28)$$

W^z_m is a member of a coefficient vector that is used in the quadrature approximation of the axial integrals. W^z_1 and W^z_J are nonzero because a weight factor of $z(1-z)$ is used in the generation of the axial basis functions [Finlayson, 1980]. The exit ($j = J$) conditions are then

$$\frac{dc_v(J,t)}{dt} = - \sum_{m=1}^{J-1} \frac{A^z_{J,m}}{A^z_{J,J}} \frac{dc_v(m,t)}{dt} \quad (4.1.1.29)$$

$$\frac{dc_b(J,t)}{dt} = - \sum_{m=1}^{J-1} \frac{A^z_{J,m}}{A^z_{J,J}} \frac{dc_b(m,t)}{dt} \quad (4.1.1.30)$$

Evaluations of (4.1.1.23-30) are made after solving (4.1.1.19) for $c_p(r,z,t)$ at each of the radial collocation points:

$$y(i,j,t) = \frac{Dg_p c_p(i,j,t) + Dg_s c_p(i,j,t)^{1/n}}{Dg_p + Dg_s} \quad (4.1.1.31)$$

For $1/n$ not equal to 1, values of $c_p(i,j,t)$ are determined with a root

finding subroutine called DZBREN, which also is an IMSL algorithm, and for 1/n = 1, $y(i,j,t) = c_p(i,j,t)$.

Initially (t=0), (4.1.1.22) is used for concentration values at all of the collocation points:

$c_v(j,t=0) = c_{vi}(j)$,

$c_b(j,t=0) = c_{bi}(j)$,

$c_p(i,j,t=0) = c_{pi}(i,j) \Rightarrow y_i(i,j)$ from (4.1.1.31),

or, typically,

$$c_v(j,t=0) = c_b(j,t=0) = y(i,j,t=0) = \begin{cases} 0 \text{ (breakthrough)} \\ 1 \text{ (elution)} \end{cases} \qquad (4.1.1.32)$$

Condition (4.1.1.32) is used in (4.1.1.23-30) to calculate the initial derivatives; the derivatives are sent to GEAR, and GEAR returns values of $y(i,j,t)$, $c_b(j,t)$, and $c_v(j,t)$; (4.1.1.31) is solved for $c_p(i,j,t)$; and the algorithm is repeated until the desired throughput is reached.

4.1.1.5 Model Simplifications and Corresponding Solutions. Even though the general form of the column model is solved numerically, exact solutions can be obtained when 1/n is equal to 1, certain mass transfer mechanisms are unimportant, and different boundary conditions are used. In the following section, verification of the numerical solution of the column model is reported. The numerical solution method is verified by comparing its calculations to the analytic solutions given below for a series of special cases. There are two assumptions that are commonly employed in column models that simplify the equations derived above. The local equilibrium assumption is the most common in ground water quality modeling, and the plug flow assumption is used more often in trickle-bed reactor analysis.

Local Equilibrium Assumption. When the mechanisms affecting the rates to chemical equilibrium between phases are fast in comparison to chemical movement in the direction of air and water flow, then a condition of local equilibrium is said to exist [Brusseau and Rao, 1989]. It was assumed in the model development that three mechanisms affect the time to equilibrium between the air, water, and soil phases: air-water mass transfer, mobile-immobile water mass transfer, and intraaggregate diffusion. Therefore when the magnitudes of St_v, St_b, and Ed_p are large in comparison to Pe_b or Pe_v the model will simulate local equilibrium. Chemical concentrations in the air and on the soil can be determined from the water concentration when local equilibrium exists by the following equilibrium relationships:

$$C_v(Z,T) = HC_b(Z,T)$$
$$C_p(0 \le R \le R_a, Z, T) = C_b(Z,T)$$
$$Q(0 \le R \le R_a, Z, T) = KC_b(Z,T)^{1/n} \qquad (4.1.1.33)$$

Throughout the remainder of this report, when local equilibrium is assumed then it is also assumed that (4.1.1.33) is satisfied and that 1/n is equal to 1.

A mass balance across all phases in a column where local equilibrium exists results in the following expression in terms of dimensionless water concentration:

$$\frac{\partial c_b(z,t)}{\partial t} = \frac{1}{Pe}\frac{\partial^2 c_b(z,t)}{\partial z^2} - \frac{\partial c_b(z,t)}{\partial z} \qquad (4.1.1.34)$$

The three mass balances in the general model reduce to one (equation (4.1.1.34)) because $C_v(Z,T)$, $C_p(R,Z,T)$, and $Q(R,Z,T)$ are determined

according to (4.1.1.33). The three mass transfer groups that are important are combined into one Peclet number (Pe). As listed in Table 4.1.1.2, Pe represents the ratio of mass transport by air and water advection to mass transport by dispersion in both fluids. Equation (4.1.1.34) is the classical convection-dispersion equation most commonly used in subsurface transport modeling [van Genuchten and Jury, 1987]. The assumption of local equilibrium is used here to verify the numerical approximation of gas and liquid dispersion in the general model.

Various solutions of (4.1.1.34) can be obtained by changing the boundary conditions. The conditions that are used for the general model assume that the column acts as a closed reactor [Levenspiel, 1962] and are used to simplify the numerical solution. Closed reactor boundary conditions for (4.1.1.34) that enable an analytic solution to be obtained are those proposed by Danckwerts [1953]:

$$1 - c_b(z=0^+,t) = -\frac{1}{Pe}\frac{\partial c_b(z=0^+,t)}{\partial z} \qquad (4.1.1.35)$$

$$\frac{\partial c_b(z=1,t)}{\partial z} = 0 \qquad (4.1.1.36)$$

Equation (4.1.1.35) assumes that the influent chemical concentration is constant ($c_{bi}(t) = 1$). The initial condition to be used for obtaining an analytic solution is

$$c_b(z,t=0) = 0 \qquad (4.1.1.37)$$

For Pe less than 2, the solution of the convection-dispersion equation approaches a completely mixed reactor solution [Brenner, 1962],

and so the solution of (4.1.1.34) approaches

$$c_b(0 \le z \le 1, t) = 1 - \exp[-t] \qquad (4.1.1.38)$$

A breakthrough curve described by (4.1.1.38) represents the maximum observed spreading caused by dispersion.

In general, the solution of (4.1.1.34) constrained by conditions (4.1.1.35-37) is [Hashimoto *et al.*, 1964]:

$$\begin{aligned} c_b(z=1,t) = {} & \frac{1}{2}\,\mathrm{erfc}\left[\frac{Pe^{1/2}(1-t)}{2t^{1/2}}\right] \\ & - \frac{1}{2}\exp(Pe)\,\mathrm{erfc}\left[\frac{Pe^{1/2}(1+t)}{2t^{1/2}}\right] \\ & + 3(Pe\ t)^{1/2}\exp(Pe)\ \mathrm{ierfc}\left[\frac{Pe^{1/2}(1+t)}{2t^{1/2}}\right] \\ & - 2(Pe\ t)\exp(Pe)\ i^2\mathrm{erfc}\left[\frac{Pe^{1/2}(1+t)}{2t^{1/2}}\right] \\ & + 4(Pe\ t)^{1/2}\exp(3Pe/2)\ \mathrm{ierfc}\left[\frac{Pe^{1/2}(2+t)}{2t^{1/2}}\right] \\ & - 16(Pe\ t)\exp(3Pe/2)\ i^2\mathrm{erfc}\left[\frac{Pe^{1/2}(2+t)}{2t^{1/2}}\right] \\ & + 20(Pe\ t)^{3/2}\exp(3Pe/2)\ i^3\mathrm{erfc}\left[\frac{Pe^{1/2}(2+t)}{2t^{1/2}}\right] \\ & - 8(Pe\ t)^2\exp(3Pe/2)\ i^4\mathrm{erfc}\left[\frac{Pe^{1/2}(2+t)}{2t^{1/2}}\right] + \ldots \end{aligned} \qquad (4.1.1.39)$$

For values of Pe greater than 40, the following asymptotic solution is valid [Hashimoto *et al.*, 1964]:

$$c_b(z=1,t) = \frac{1}{2}\,\mathrm{erfc}\left[\frac{Pe^{1/2}(1-t)}{2t^{1/2}}\right] + \left[\frac{t}{\pi Pe}\right]^{1/2} \cdot \left[\frac{t^2+4t-1}{(t+1)^3}\right] \exp\left[\frac{-Pe(1-t)^2}{4t}\right] \qquad (4.1.1.40)$$

Danckwerts [1953] solved (4.1.1.34) for open reactor boundary conditions, which correspond to a column of infinite length:

$$c_b(z=-\infty,t) = 1 \qquad (4.1.1.41)$$

$$c_b(z=\infty,t) = 0 \qquad (4.1.1.42)$$

The solution of (4.1.1.34) for these boundary conditions and the initial condition given by (4.1.1.37) is [Danckwerts, 1953]:

$$c_b(z=1,t) = \frac{1}{2}\,\mathrm{erfc}\left[\frac{Pe^{1/2}(1-t)}{2t^{1/2}}\right] \qquad (4.1.1.43)$$

Plug Flow Assumption. In structured or aggregated soils where the air or water is flowing fast, the spreading that is caused by axial dispersion in air and water could be negligible compared to the spreading caused by other mechanisms. Plug flow is assumed here so that the numerical approximation of intraaggregate diffusion and film transfer could be tested against an analytic solution in the same manner as dispersion is verified with the local equilibrium solutions.

Rosen [1952] derived an analytic solution for single-phase plug flow through a packed bed. The general model reduces to the equations solved by Rosen [1952] if 1/n is equal to 1, the air-water mass transfer rate is fast (large St_v), and axial dispersion is slow in comparison to advection (large Pe). For this condition, (4.1.1.12) and (4.1.1.13)

combine to give

$$\frac{\partial c_b(z,t)}{\partial t} = \frac{1 + Dg}{1 + Dg_v}\left[- \frac{\partial c_b(z,t)}{\partial z} + \frac{3St_b}{1 + Ar}\left[c_p(r=1,z,t) - c_b(z,t)\right]\right] \qquad (4.1.1.44)$$

The boundary condition for (4.1.1.44) is

$$c_b(z=0,t>0) = 1 \qquad (4.1.1.45)$$

Equations (4.1.1.18-21) are used to account for film transfer and intraaggregate diffusion. The initial condition is the breakthrough scenario (zero concentrations at t = 0) for (4.1.1.22).

The exact solution of (4.1.1.44) is an integral, but Rosen [1954] developed the following asymptotic solution for $Ed_p/[1 + Ar] > 13.33$:

$$c_b(z=1,t) = \frac{1}{2}\,\mathrm{erfc}\left[\frac{[1 - t][1 + Dg]}{2[Dg_p + Dg_s]}\left[\frac{1 + Ar}{15Ed_p} + \frac{1 + Ar}{3St_b}\right]^{-1/2}\right] \qquad (4.1.1.46)$$

4.1.1.6 Model Verification and Sensitivity. Verification of a numerical solution is performed by comparing model calculations to other numerical or analytical solutions. The column model numerically approximates air and water advection, water dispersion and gas diffusion in the direction of flow (axial), diffusion in immobile water, film transfer, and air-water mass transfer. To verify the numerical method, calculations of this model are compared to analytical solutions for simplified situations where one or more of the mechanisms are important.

The numerical approximation of advection and axial dispersion is compared to the local equilibrium solutions (equations (4.1.1.38-40)). The numerical approximation of advection, air-water mass transfer, film transfer, and diffusion in pores containing immobile water is compared to the plug flow solution (equation (4.1.1.46)).

Model input parameters for the verification step are chosen to satisfy assumptions that were made to obtain a particular analytical solution. To simulate local equilibrium, large values of Ed_p, St_b, and St_v are used as input. To simulate plug flow, large values of Pe_b and Pe_v are used instead.

Figure 4.1.1.3 compares the breakthrough curves that are calculated with the column model to analytical solutions for different values of the dimensionless groups. These breakthrough curves are plots of the column effluent water concentration relative to the influent versus the number of pore volumes of water fed. For the calculations shown in Figure 4.1.1.3 it is assumed that 1/n is equal to 1 and influent concentrations are constant. Numerical solutions, shown as solid lines, using J (axial) equal to 10 and I (radial) equal to 3 collocation points compare closely to the analytic solutions, shown as symbols, for the dimensionless group values listed on Figure 4.1.1.3. Oscillations appear in the numerical solution when simulating plug flow conditions, and the numerical error increases as Pe increases. These errors appear to dampen as time increases. Because overall mass balances are used as boundary conditions, it is possible that even though numerical error occurs initially for certain situations, the error does not propogate except for very large values of Pe. Ten axial and 3 radial collocation points are also used for the model validation

calculations.

Additional simulations are performed with the column model to observe the impact of Ed_p, Pe, St_b, and St_v on the numerical solution. Dispersion calculations by the model are accurate for values of Pe from 0.1 to 40 and greater. Values of Pe greater than 1000, however, cause unacceptable amounts of numerical error. For low values of Pe, the observed spreading causes the breakthrough curve to become asymmetric, and the model can simulate this. As the value of Ed_p, St_b, or St_v decrease below 1, the early portion of the breakthrough curve sharpens and the latter part tails.

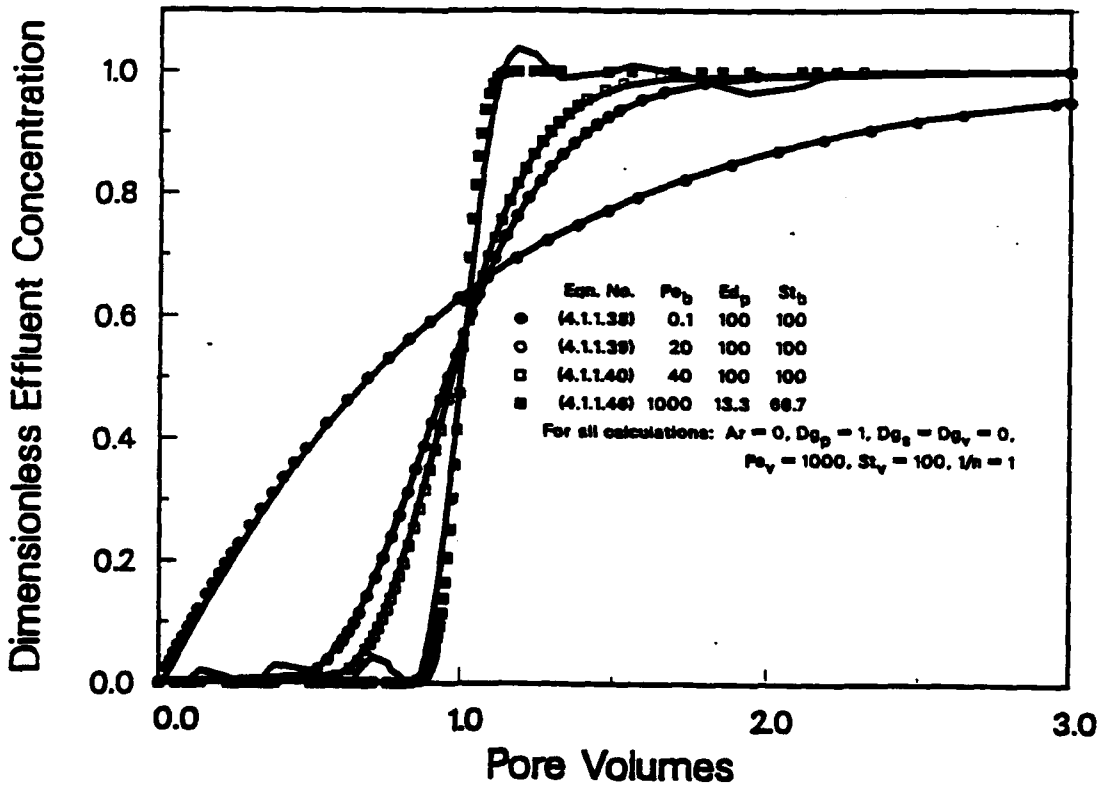


Figure 4.1.1.3. Comparisons of the numerical solution (solid curves) for 10 axial and 3 radial collocation points to analytic solutions (symbols) for verification of the orthogonal collocation approximation.

Mass transfer mechanisms can have similar impacts on chemical

transport [Crittenden *et al.*, 1986; Roberts *et al.*, 1987]. A relationship for equivalent spreading between axial dispersion, diffusion in immobile water, and film transfer is developed by equating the arguments of (4.1.1.43) and (4.1.1.46). Equations (4.1.1.39), (4.1.1.40), and (4.1.1.43) give comparable results for large Pe [Crittenden *et al.*, 1986]. If only the central portion of the breakthrough curve is considered (*i.e.* t is near 1), the following relationship is derived:

$$Pe = \frac{15Ed_p[1 + Dg]^2}{[1 + Ar][Dg_p + Dg_s]^2} = \frac{3St_b[1 + Dg]^2}{[1 + Ar][Dg_p + Dg_s]^2} \qquad (4.1.1.47)$$

Equation (4.1.1.47) is similar to the equivalent spreading relationship for saturated flow developed by Crittenden *et al.* [1986]. When Ar is equal to 0 then (4.1.1.47) is also equivalent to $3St_v[1 + Dg]^2\ Dg_v^{-2}$.

The column model is used to calculate the curves shown in Figure 4.1.1.4 to simulate conditions where axial dispersion, diffusion in immobile water, air-water mass transfer, and film transfer have equivalent impacts on a breakthrough curve as determined from (4.1.1.47). These conditions correspond to a values of: Pe = 480 (Ed_p = 100, St_b = 100, St_v = 100); Ed_p = 13.3 (Pe = 1000, St_b = 100, St_v = 100); St_b = 66.7 (Ed_p = 100, Pe = 1000, St_v = 100); and St_v = 0.16 (Ed_p = 100, Pe = 1000, St_b = 100), respectively. The values of the dimensionless groups are obtained according to (4.1.1.47). It is evident that it is not always possible to distinguish between the impacts of different mechanisms by fitting model solutions to data [Roberts *et al.*, 1987; Brusseau and Rao, 1989]. Equation (4.1.1.47) can be used to compare the relative contributions of axial dispersion,

intraaggregate diffusion, and film transfer on the observed spreading of a chemical front and to determine which mechanisms are important.

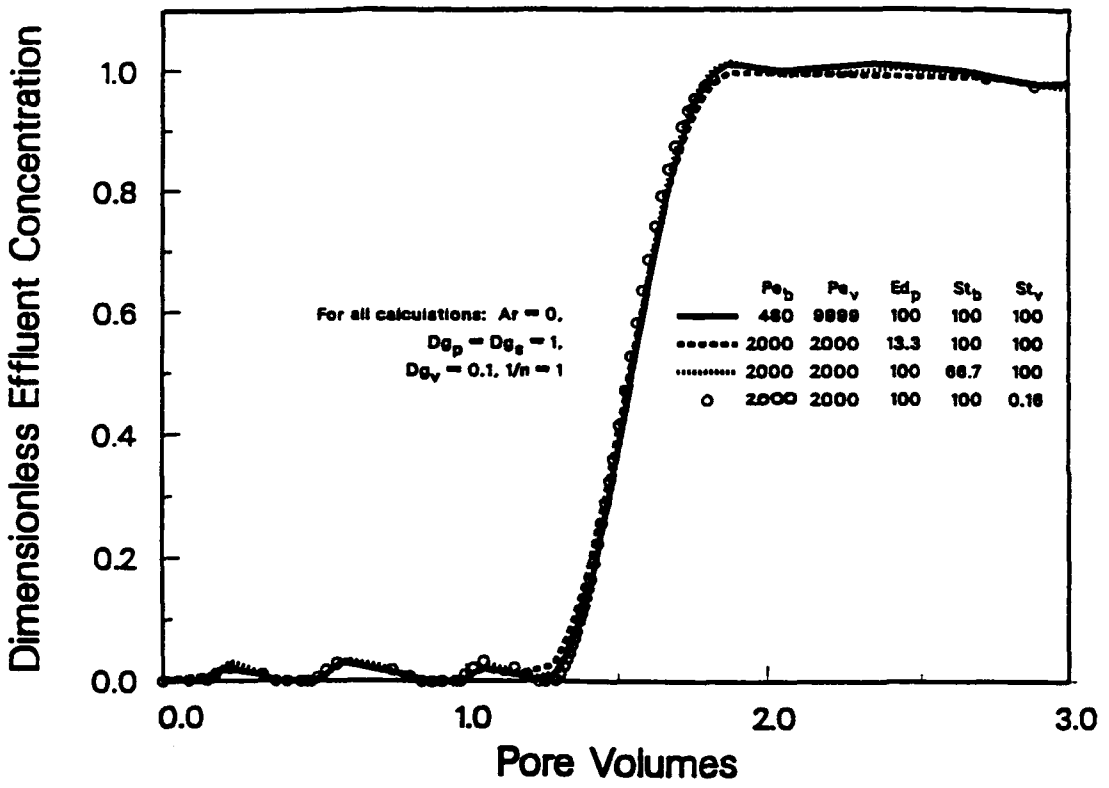


Figure 4.1.1.4. Comparisons of the numerical solution for different spreading conditions: axial dispersion dominant (solid line), intraaggregate diffusion dominant (dashed line), film transfer dominant (dotted line), and air-water mass transfer dominant (open circles).

4.1.1.6 Validation of the Column Model. A complete model study includes independent validation tests. Validation of a model is achieved by predicting experimental results. Unsaturated miscible displacement experiments were performed in order to validate the one-dimensional column model and to determine its ability to describe chemical transport with water flow in cohesionless soils, such as sands, and in structured or aggregated soils.

Materials and Methods. Laboratory columns and experimental procedures were designed to measure the breakthrough and elution of trichloroethene (TCE) and bromide (Br^-) from a cohesionless soil and a aggregated soil. Table 4.1.1.3 lists the column conditions for each experiment. Details of the column design and experimental procedures are reported elsewhere [Krause, 1987; Hutzler *et al.*, 1989b]. Most of the experiments were performed by Krause [1987] and are noted as such in the figure captions. Those performed as a part of this work have no such designation, and the procedures that were followed were the same as reported by Krause [1987].

Table 4.1.1.3. Conditions for Column Runs With Bromide (Br^-) and Trichloroethene (TCE) in Ottawa Sand (OS) and in Aggregates (APM).

Column Length (cm)	Soil	Chemical	Influent Concentration ($\mu g\ L^{-1}$)	Water Flow Rate ($cm^3\ s^{-1}$)	Degree of Saturation	Pulse Time (hours)	Total Time (hours)
29.9	OS	Br^-	45,600	0.075	0.33	2.65	6.47
29.9	OS	TCE	650	0.076	0.33	11.00	23.50
20.1	APM	Br^-	97,600	0.084	0.64	9.68	18.86
20.1	APM	TCE	1,160	0.084	0.64	12.18	22.36

Cross-sectional area of columns is 91.6 cm^2, there is no air flow, and the temperature is 22 ^oC.

Trichloroethene was chosen because it is a common ground water contaminant of intermediate volatility. Saturated and unsaturated column runs were performed with a bromide tracer to characterize the columns and for estimating certain parameters [Krause, 1987; Hutzler *et al.*, 1989b]. Bromide was chosen as the tracer because it is nonadsorbing and nonvolatile, and it can be measured in low concentrations. Chemical properties of Br^- and TCE corresponding to the

conditions of the column experiments are given in Table 4.1.1.4.

Table 4.1.1.4. Properties of Water, Trichloroethene, and Bromide at 22 °C that are Used for Parameter Estimation.

	Value
Water	
Viscosity, μ_l (g cm^{-1} s^{-1})	0.00955[a]
Density, ρ_l (g cm^{-3})	0.998[a]
Trichloroethene, TCE (C_2HCl_3)	
Molecular weight, M_A (g mol^{-1})	131.3[a]
Molar volume, V_A (cm^3 mol^{-1})	98.1[b]
Boiling point, T_b (K)	360[a]
Henry's constant, H (dimensionless)	0.4[c]
Bromide, Br^- (made from KBr)	
Valence, n_-, n_+	1
Limiting ionic conductance in water at 25 °C	
Anion, λ_- (A V g-equiv cm^{-6})	78.3[d]
Cation, λ_+ (A V g-equiv cm^{-6})	73.5[d]

[a]From Weast [1981].
[b]LeBas [1915].
[c]Ashworth *et al.* [1988].
[d]Reid *et al.* [1977], temp. correction factor: $0.00299T_e\ \mu_l^{-1}$.

Ottawa sand (Ottawa, Illinois) was chosen to simulate cohesionless soils, and SCR Veri-lite (Mapleton Development, Inc., Minerva, Ohio) was chosen to simulate aggregated soils. Table 4.1.1.5 summarizes the characteristics of each material as packed in the columns. Ottawa sand is a uniform, silica sand containing little or no organic material and, thus, does not adsorb most organic compounds from aqueous solution. A

saturated TCE column run in the sand showed no adsorption of TCE [Hutzler *et al.*, 1989b]. The aggregated porous material (APM) is a lightweight, fired clay used mostly in industry as an insulator for steel and iron ladles. The particles are porous and more angular than Ottawa sand. An aqueous isotherm experiment with the APM showed no adsorption of TCE [Krause, 1987].

Table 4.1.1.5. Properties of Porous Media.

	Ottawa Sand	APM
Properties Measured Directly:		
Total porosity, ϵ	0.33[a]	0.70[a]
Microporosity, ϵ_a	0[b], 0.043[c]	0.50[a]
Bulk density, ρ_b (g cm^{-3})	1.78[a]	0.45[a]
Hydraulic Cond., K_s (cm s^{-1})	0.26[e]	0.22[e]
Particle Radius[d] (cm)	0.035[f]	0.035[f]
Derived Parameter Values:		
Solid density, ρ_s (g cm^{-3})	2.65	1.51
Particle density, ρ_a (g cm^{-3})	2.65	0.75
Macroporosity, ϵ_m	0.33	0.40
Immobile saturation, S_i	0, 0.10	0.42

[a]Measured gravimetrically [Black *et al.*, 1965].
[b]Assumed.
[c]Value fit to tracer study and close to field capacity measurement.
[d]Aggregrate radius (R_a) was assumed to be equal to the particle radius.
[e]Determined from slope of specific discharge (v_p) versus headloss per unit length of column.
[f]Half of geometric mean particle size contained in U.S. Std. no. 20-30 sieves (0.085-0.055 cm).

Hydrodynamic measurements were made with both packed columns. The saturated conductivities (K_s) are given in Table 4.1.1.5. Figure 4.1.1.5a shows the relationship between suction and degree of saturation

for both media. Figure 4.1.1.5b shows the relationship between unsaturated hydraulic conductivity (K(S)) relative to K_s and the mobile saturation ($S-S_i$) for both media. The value of S_i used for calculating the mobile saturation in the sand was determined from a tracer study while S_i for the aggregated material was measured gravimetrically. The particle sizes of the sand and of the aggregates used in this work are equal, and the interparticle (macro) porosity (ϵ_m) of the materials when packed in the columns are nearly the same. Therefore the saturated conductivities of the sand and of the APM columns are almost the same. The hydrodynamic properties of the two materials under unsaturated flow conditions are also similar. Since both media have similar flow properties then axial dispersion in both columns should also be similar. Gas and liquid dispersion coefficients measured in one column are used to predict axial dispersion in the other.

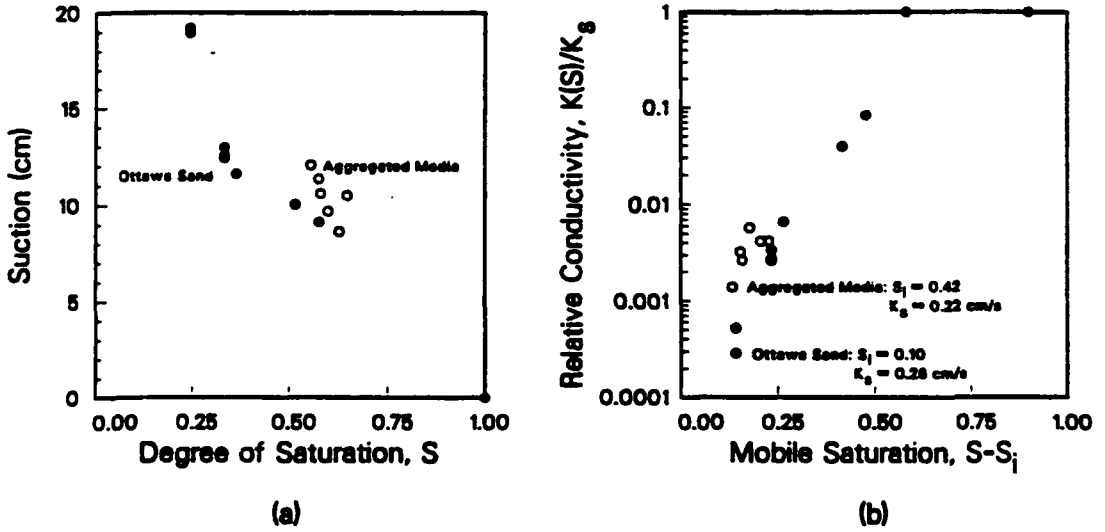


Figure 4.1.1.5. Moisture characteristics of Ottawa sand and aggregates.

Prior to performing the unsaturated experiments, saturated column

experiments were performed with Br^-. The results were predicted using independently derived dispersion coefficients so it was assumed that apparatus-induced dispersion was negligible [Hutzler *et al.*, 1989b].

Experiments were performed in such a manner that the simplifying assumptions made in the model development are satisfied. For example, the model derivation assumed steady flow and uniform moisture content; therefore influent water was supplied to the tops of the columns at steady rates, and water tension inside the column was monitored along the depth and adjusted by applying a suction at the bottom of the column to achieve a uniform degree of saturation [Krause, 1987; Hutzler *et al.*, 1989b]. The columns were packed so that no stratification was visible. In addition, ambient temperature was held constant.

Model Parameter Estimation. Parameter values for transport models can be obtained from direct measurement, literature correlations, laboratory experiments, and by fitting model solutions to column data. For a mathematical model to be predictive, however, the parameters must be determined independently of the system being modeled and not by fitting simulations to data. Accordingly, the model parameters should be based on physical properties of the soil and chemical being studied. Tables 4.1.1.3, 4.1.1.4, and 4.1.1.5 list the measured parameter values and the basic chemical and soil properties that were used in this study. The remaining model parameters that are required by the model are determined from these values with the correlations given below and from tracer studies. All of the units are grams-centimeters-seconds (cgs) unless otherwise noted.

Gas Dispersion

For low gas velocities (u), gas dispersion in soil will be

primarily due to gas diffusion through tortuous air-filled pores. Thus a free-air diffusion coefficient (D_G) can be corrected for the tortuosity of the air-filled pores (τ_a) to obtain the effective gas diffusion coefficient ($E_v = D_G\ \tau_a^{-1}$). A tortuosity correction can be used when u is less than about $[1-\epsilon(1-S)]D_G[2R_a\epsilon(1-S)]^{-1}$ [Miyauchi and Kikuchi, 1975]. In this work the Wilke-Lee modification [Wilke and Lee, 1955] of the Hirschfelder-Bird-Spotz method for calculating diffusivities of nonpolar organics is used to calculate D_G in air:

$$D_G = \frac{[4.353 - (0.0345 + M_A^{-1})^{0.5}]T_e^{1.5}(0.0345 + M_A^{-1})^{0.5}}{P_t(0.118V_A^{0.33} + 0.371)^2 f(0.1025T_e\ T_b^{-0.5})} \qquad (4.1.1.48)$$

Ambient pressure is in Pascals and the value of the collision function for diffusion ($f(0.1025T_e\ T_b^{-0.5})$) is obtained from a polynomial fit to a graph found in Treybal [1980].

Many correlations exist for determining the tortuosity of the air-filled pores (see Roy and Griffin [1987]). A relationship adapted from Millington [1959] is used in this work:

$$\tau_a = \epsilon^2[\epsilon(1-S)]^{-7/3} \qquad (4.1.1.49)$$

The air-filled porosity term in (4.1.1.49) is raised to the -7/3 power instead of -10/3 as reported by Millington [1959] because he included the area available for gas diffusion in the determination of the effective diffusion coefficient, while in this work it is separated from E_v [Baehr, 1987]. Because Millington [1959] studied diffusion through cohesionless soils, mobile porosity (ϵ_m) is substituted for total porosity in ϵ^2 in (4.1.1.49) when estimating τ_a in the columns containing APM.

Diffusion in Pores Containing Immobile Water

The description of liquid diffusion in intraaggregate pores must account for the tortuous paths that molecules travel around soil particles that form an aggregate ($D_p = D_l \tau_p^{-1}$). Many correlations exist for estimating liquid diffusion coefficients (D_l). For TCE, the Hayduk-Laudie correlation given by Sherwood *et al.* [1975] is used:

$$D_l = 6.96(10^{-7})\ \mu_l^{-1.14}\ V_a^{-0.589} \qquad (4.1.1.50)$$

For Br^-, the Nernst-Haskell equation [Reid *et al.*, 1977] is used to calculate D_l:

$$D_l = 8.931(10^{-10})\ T_e \frac{1/n_+ + 1/n_-}{1/\lambda_+ + 1/\lambda_-} \qquad (4.1.1.51)$$

Internal pore tortuosity (τ_p) is a function of the pore shape and the amount of immobile water. The immobile degree of saturation in the sand ($S_i = 0.10$) and the specific intraaggregate diffusion rate ($D_p\ R_a^{-2} = 1.6(10^{-6})\ s^{-1}$) was determined from an unsaturated bromide column run because these parameters could not be measured directly. The values fit to the unsaturated bromide data are used to predict the movement of TCE in the sand. The amount of immobile water inside the APM particles ($S_i = 0.42$) was measured gravimetrically, and the tortuosity of the internal pores ($\tau_p = 90$) was measured in a batch study [McKenzie, 1990]. The batch experiment was similar to those performed by Rao *et al.* [1982]. The τ_p of 90 used in this work is large compared to values between 2 and 10 used by most researchers [Roberts *et al.*, 1987]. After viewing the APM under electron scanning microscopy at the Air Force Engineering and Services Center, Tyndall AFB, Florida, it was observed that the particle

surface is impervious except at a relatively few number of locations [McKenzie, 1990]. Electron microscopic photographs of the internal pores showed that many are not connected. For these reasons, the measured τ_p is reasonable. In addition, a saturated Br^- column run in the APM could be predicted using the value of τ_p measured in the batch experiment [Hutzler *et al.*, 1989b].

Air-Water Mass Transfer

The air-water mass transfer coefficient (K_L) and the specific air-water interfacial area (a) have not been studied in unsaturated soils. Many correlations do exist, however, for these parameters in the analysis of packed-tower operation, such as air stripping, and the performance of trickle-bed reactors. Turek and Lange [1981] developed a correlation for K_La in low velocity trickle-bed reactors:

$$K_La = 16.8D_l \left[\frac{8R_a^3 g\rho_l^2}{\mu_l} \right]^{-0.22} \left[\frac{2R_a v\epsilon(S-S_i)\rho_l}{\mu_l} \right]^{0.25} \cdot \left[\frac{\mu_l}{\rho_l D_l} \right]^{0.5} \qquad (4.1.1.52)$$

Equation (4.1.1.52) is valid for values of R_a between 0.028 and 0.15 cm, $\epsilon(S-S_i)$ between 0.05 and 0.3, and Re_l between 0.1 and 5. The experiments reported in this work satisfy all but the Reynolds number requirements. For the experiments reported herein Re_l is less than 0.007. This correlation, like most others for estimating air-water mass transfer rates, estimates a value of K_La of zero for a water velocity of zero. There has been no work to date that measures air-water mass transfer rates in soil columns.

Mobile-Immobile Water Mass Transfer

Mass transfer rates across the mobile-immobile water interface were found to be fast in saturated soil systems [Crittenden *et al.*, 1986; Hutzler *et al.*, 1986; Roberts *et al.*, 1987] and are expected to be fast in unsaturated systems. It is included in this modeling effort to test this hypothesis. A correlation by Wilson and Geankoplis [1966] for saturated systems was adapted to estimate k_f:

$$k_f = 1.09v[2R_a v\epsilon(S-S_i)/D_l]^{-2/3} \qquad (4.1.1.53)$$

Equation (4.1.1.53) is valid for values of Re_l between 0.0016 and 55 and $\epsilon(S-S_i)$ between 0.35 and 0.75.

Liquid Dispersion

Liquid dispersion has been studied in soils more often than any mass transport mechanism other than advection, at least in saturated media. There is still a lack of accurate correlations for predicting liquid dispersion coefficients (E_z) in unsaturated media. Yule and Gardner [1978] fit the following relationship for E_z in unsaturated sand columns:

$$E_z = 5.33(10^{-5}) + 0.216v_p \qquad (4.1.1.54)$$

Equation (4.1.1.54) was fit to data obtained for average pore velocities (v_p) between $1.7(10^{-4})$ and 0.0043 cm s^{-1} and degrees of saturation between 0.34 and 0.76.

DeSmedt and Wierenga [1984] propose the following relationship for observed dispersion in unsaturated columns of glass beads:

$$E_z = \frac{1.18(10^{-5}) + 0.021v_p + 1900v_p^2}{1.88 + 3630v_p} \qquad (4.1.1.55)$$

Values of E_z given by (4.1.1.55) include contributions from axial dispersion and diffusion in immobile water. DeSmedt and Wierenga [1984] developed (4.1.1.55) by assuming that the contribution of axial dispersion is given by

$$E_z = 1.4(10^{-5}) + 0.021v \qquad (4.1.1.56)$$

DeSmedt and Wierenga [1984] used a first-order exchange model to describe the transfer of solute between mobile and immobile water, while in this work the rate of transfer through immobile water by diffusion is also considered.

Experimental Results. The column experiments show that the column model is able to predict the breakthrough and elution of volatile organic chemicals from unsaturated soil columns under controlled conditions. The model is versatile in that it can simulate chemical movement in different types of porous media as well as under different flow and moisture conditions. The results from sand column experiments are given first and followed by the aggregated porous media results. For each soil material, unsaturated bromide transport was measured first and then followed by an experiment with trichloroethene.

Table 4.1.1.3 summarizes the column conditions for each run. The length of time for the column runs were short enough that biodegradation of TCE was not observed. Integration of the effluent concentration histories reported here indicate that at least 95% of the chemical mass retained in a soil column for a given breakthrough experiment is removed during elution. Tables 4.1.1.6 list parameters values used for the model calculations. These values were calculated from the correlations,

taken from direct measurements, and fit to experimental data. They are used in conjunction with the column conditions and soil and chemical properties listed in Tables 4.1.1.3-5 for model input.

Table 4.1.1.6. Parameter Values for the Column Model Calculations of the Validation Experiments.

Chemical/Soil:	Br^-/OS	TCE/OS	Br^-/APM	TCE/APM	Estimation Method[a]
Independently Determined Values:					
S_i	0[b]	0.10[c]	0.42	0.42	Measured gravimetrically
S	0.33	0.34	0.64	0.64	Measured gravimetrically
v ($cm\ s^{-1}$)	0.0075	0.010	0.0060	0.0060	Calculated from definition
v_p ($cm\ s^{-1}$)	0.0075	0.0074	0.0020	0.0020	Calculated from definition
H	NA	0.4	NA	0.4	Ashworth et al. [1988]
k_f ($cm\ s^{-1}$)	NA	0.0056	0.0030	0.0018	Equation (4.1.1.53)
$K_L a$ (s^{-1})	NA	$6.4(10^{-4})$	NA	$6.6(10^{-4})$	Equation (4.1.1.52)
D_l ($cm^2\ s^{-1}$)	$2.0(10^{-5})$	$9.4(10^{-6})$	$2.0(10^{-5})$	$9.4(10^{-6})$	Equation (4.1.1.51) for Br^-; (4.1.1.50) for TCE
τ_p	NA	NA	90	90	[McKenzie, 1990]
D_p ($cm^2\ s^{-1}$)	NA	NA	$2.2(10^{-7})$	$1.0(10^{-7})$	$D_l\ \tau_p^{-1}$
D_G ($cm^2\ s^{-1}$)	NA	0.088	NA	0.088	Equation (4.1.1.48)
τ_a	NA	3.7	NA	4.1[d]	Equation (4.1.1.49)
E_v ($cm^2\ s^{-1}$)	NA	0.024	NA	0.021	$D_G\ \tau_a^{-1}$
E_z ($cm^2\ s^{-1}$)	$1.7(10^{-3})$		$5.0(10^{-4})$		Equation (4.1.1.54)
	$1.7(10^{-4})$		$1.4(10^{-4})$		Equation (4.1.1.56)
	$3.7(10^{-3})$	0.10[c]	0.060[e]	0.020[f]	Equation (4.1.1.55)
Values Resulting From Model Calibration:					
S_i	0.10	NC	NC	NC	Fit
v ($cm\ s^{-1}$)	0.011	NC	NC	NC	Calculated from definition
H	NC	0.7	NC	0.7	Fit
k_f ($cm\ s^{-1}$)	0.0056	NC	NC	NC	Equation (4.1.1.55)
$D_p\ R_a^{-2}$ (s^{-1})	$1.6(10^{-6})$	$8.0(10^{-7})$	NC	NC	Fit Br^-, Divided value for Br^- by 2
E_v ($cm^2\ s^{-1}$)	NC	<0.0024	NC	<0.0028	Adjusted to make unimportant
E_z ($cm^2\ s^{-1}$)	0.10	NC	0.020	NC	Fit

[a] unless otherwise noted.
[b] assumed.
[c] used value fit to Br^- in OS experiment.
[d] ϵ_m^2 was substituted for ϵ^2.
[e] multiplied E_z fit for Br^- in OS by ratio of v_{APM} to v_{OS}.
[f] used value fit from Br^- in APM experiment.
NA - not applicable.
NC - no change.

Table 4.1.1.7 is a list of the magnitudes of the corresponding

dimensionless groups that characterize the model solutions in terms of which mechanisms have the most impact on the chemical front. Model predictions correspond to the group values in the upper part of Table 4.1.1.7. The magnitudes of these groups were calculated from independently determined parameter values. Group values that were fit are listed in the lower part of Table 4.1.1.7.

Table 4.1.1.7. Dimensionless Group Values Resulting from Parameter Values Listed in Table 4.1.1.6 ($Ar = 0$, $Dg_s = 0$, and $1/n = 1$ for all calculations).

Chemical/Soil:	Br^-/OS	TCE/OS	Br^-/APM	TCE/APM
Independently Determined Values:				
Dg_p	0	0.42	1.9	1.9
Dg_v	NA	1.1	NA	0.66
R_d	1	1.8	1	1.2
Ed_p	NA	$9.0(10^{-4})$	1.1	0.52
St_b	NA	260	1100	660
St_v	NA	0.66	NA	4.8
Pe_v	NA	12	NA	8.7
Pe_b	61[a]	3.1	860[b]	6.0
Pe	Pe_b	2.4	Pe_b	3.6
Values Resulting From Model Calibration:				
Dg_p	0.44[c]	NC	NC	NC
Dg_v	NC	1.9[c]	NC	1.2[c]
R_d	NC	2.3[d]	NC	1.4[d]
Ed_p	$1.9(10^{-3})$[c]	NC	NC	NC
St_b	3900[d]	NC	NC	NC
Pe_v	NC	>67[e]	NC	>50[e]
Pe_b	3.2[c]	NC	6.0[c]	NC

[a]determined using E_z from (4.1.1.55).
[b]determined using E_z from (4.1.1.56).
[c]fit.
[d]calculated from fit parameters.
[e]increased to decrease impact.
NA - not applicable.
NC - no change.

The air and water flow column model is intended for describing one-dimensional movement of volatile organic chemicals in unsaturated columns of soil. The numerical solution of the general model is used to simulate movement of TCE with unsaturated water flow. The numerical code for solving the general model was altered by Hutzler *et al.* [1989b] to ignore vapor movement for simulating the Br^- column experiments.

Column Experiments with Ottawa Sand

Bromide and TCE experiments were run on a column packed with unsaturated Ottawa sand. A 45.6 mg L^{-1} Br^- solution was fed to the top of the column at a rate of 0.075 cm^3 s^{-1} (v_p = 0.0075 cm s^{-1}) for 2.65 hours. Clean water was then applied at the same rate to elute the bromide from the column. The degree of saturation was 0.33. Figure 4.1.1.6 compares the data to model calculations. Because the sand particles are solid and uniformly sized it is first assumed that the amount of immobile water is negligible. This was a valid assumption for the saturated flow experiments [Hutzler *et al.*, 1989b]. Equation (4.1.1.39) is used to calculate the dashed line shown in Figure 4.1.1.6 (Pe = 61) by using an E_z of 3.7(10^{-3}) cm^2 s^{-1}, which was estimated with (4.1.1.55). This value is larger than that which was estimated by (4.1.1.54).

The breakthrough data is shifted to the left of the dispersion equation prediction, and this is attributed to the presence of immobile water. This shift could not be simulated with a model assuming local equilibrium. Five parameter values are not known with certainty (S_i, D_p, R_a, k_f, E_z). The numerical model simulated the data by adjusting only three dimensionless groups (Dg_p, Ed_p, Pe_b). Either Ed_p or St_b could have been reduced so that the model calculation of the

breakthrough curve would shift to the left. Since film transfer is unimportant in saturated soils [Hutzler *et al.*, 1986; Roberts *et al.*, 1987], Ed_p is fit for this experiment. When either Ed_p, St_b, or St_v is small, the early portion of the breakthrough curve will be sharp and appear sooner than R_dS pore volumes.

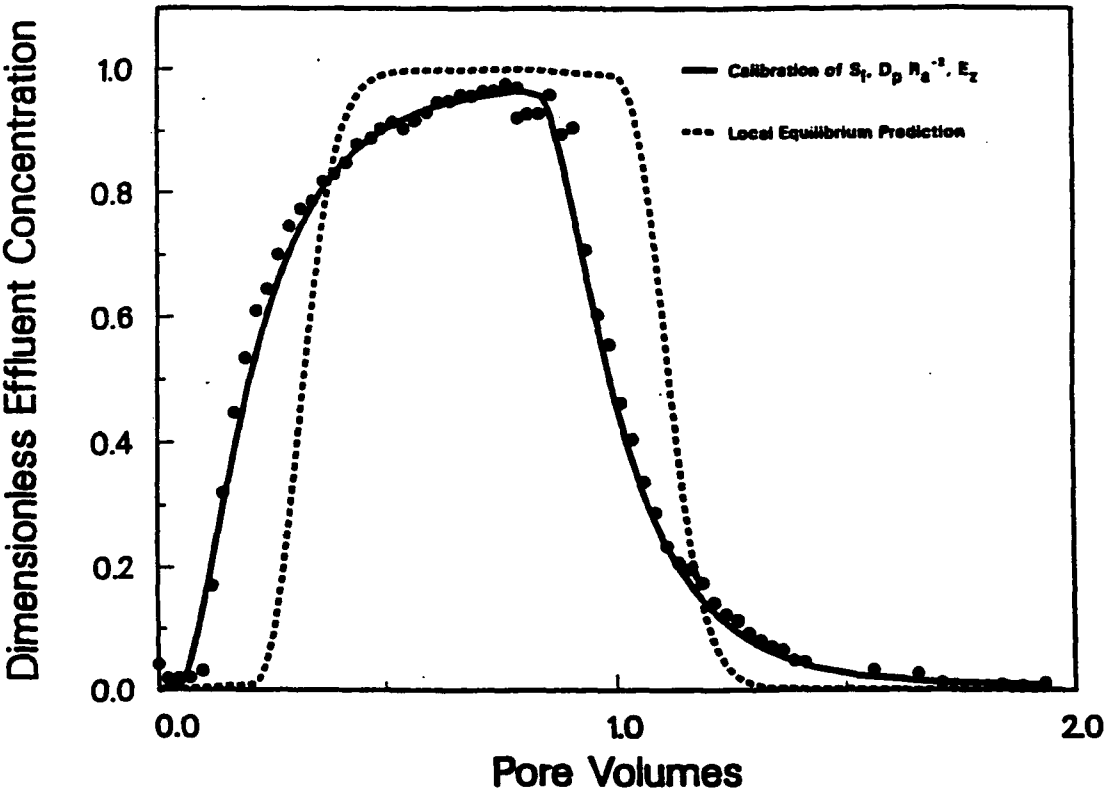


Figure 4.1.1.6. Model prediction and fit of bromide movement in a column of Ottawa sand at a degree of saturation of 0.33.

For this Br^- run ($R_d = 1$), the early portion of the breakthrough will be located approximately at $S\ (1+Dg_p)^{-1}$ pore volumes. Hence Dg_p is increased to simulate the first sharp increase in Br^- concentration. Reducing Ed_p does not result in enough spreading to simulate the data, so Pe_b had to be reduced, too. A value of 0.44 for Dg_p, 0.0019 for Ed_p, and 3.2 for Pe_b best describe the data. These group values correspond

to an S_i of 0.10, $D_p R_a^{-2}$ equal to $1.6(10^{-6})$ s^{-1}, and an E_z of 0.10 cm^2 s^{-1}. The fit value of S_i is close to the degree of saturation of the Ottawa sand column after it has drained freely by gravity but is about twice as much as that observed by DeSmedt and Wierenga [1984] in unsaturated columns of glass beads. Intraaggregate diffusion is fit as the ratio $D_p R_a^{-2}$ to keep the uncertainty in one term. The low value of Ed_p is possibly due to channeling inside the column or an increased moisture content at the bottom of the column. The E_z fit to this data is larger than that observed in other unsaturated studies [Yule and Gardner, 1978; DeSmedt and Wierenga, 1984; Wierenga and van Genuchten, 1989]. Wierenga and van Genuchten [1989] measured dispersion coefficients for chloride in a similar sized column of sandy soil. Their values are a factor of 10 less than measured here (accounting linearly for differences in pore water velocity). The moisture content was about two times greater in their experiments. The large amount of dispersion could also be due to channeling or an increase in moisture content at the bottom of the column. A rapid rate of film transfer is indicated by the large value of St_b (3900).

A TCE solution of 650 μg L^{-1} was then fed to the sand column at a degree of saturation of 0.34 over a period of 11 hours at a rate of 0.076 cm^3 s^{-1} (v_p = 0.0074 cm s^{-1}). Figure 4.1.1.7 shows a comparison of the TCE data to calculations of the column model. Even after using the parameters fit in the unsaturated Br^- run, there are still two parameters (H, E_v) that are needed for predicting the TCE results and, yet, the values are uncertain. For the calculations shown, the E_z and S_i fit in the unsaturated Br^- experiment are used. The ratio of $D_p R_a^{-2}$ is reduced by a factor of two because (4.1.1.51) predicts that Br^-

diffuses approximately twice as fast as TCE (equation (4.1.1.50)). The model prediction is shown as a dotted line, and the corresponding magnitudes of the dimensionless groups are listed in Table 4.1.1.7.

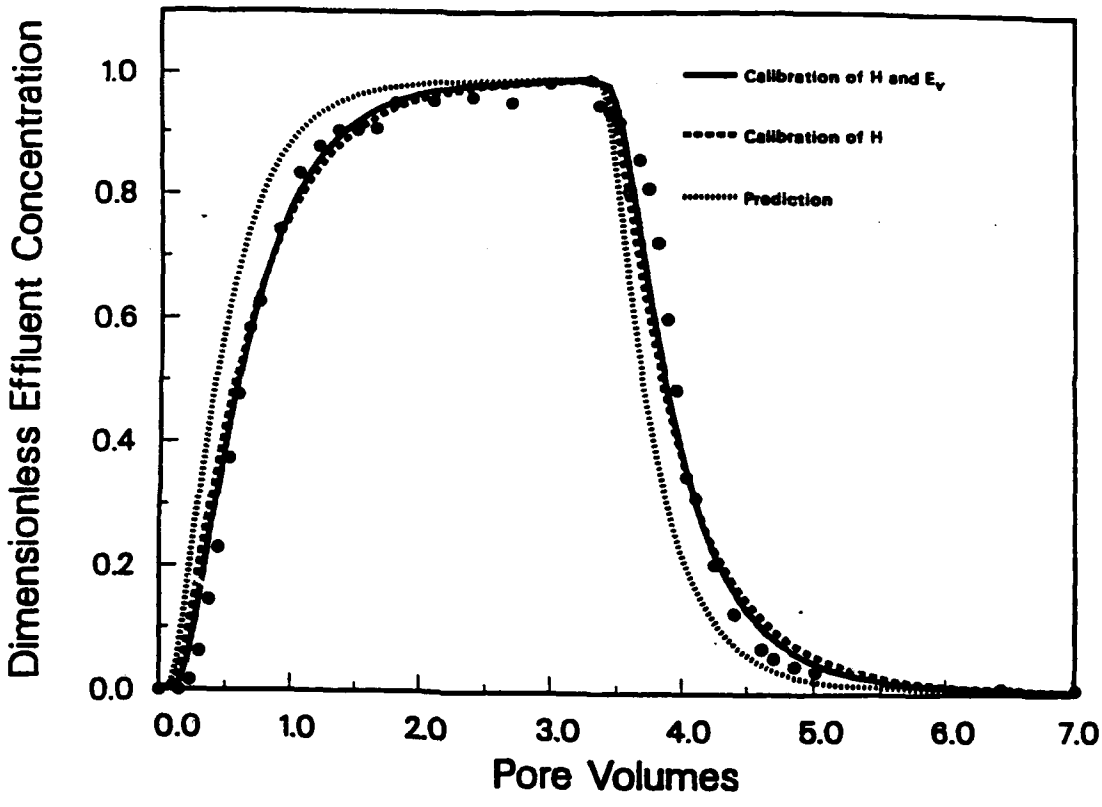


Figure 4.1.1.7. Model prediction and fits of trichloroethene movement in a column of Ottawa sand at a degree of saturation of 0.34.

The breakthrough curve is shifted to the right of the prediction indicating that the description of TCE equilibrium in the sand is incorrect. This could be due to either H for TCE being higher than predicted or sorption of TCE vapors onto the drier particle or column surfaces. TCE did not adsorb to the sand or the column in saturated experiments [Hutzler *et al.*, 1989b]. Sorption capacity (K) could be adjusted to account for vapor adsorption, but because Ed_p is small, the

model calculations would exhibit more asymmetry as K is increased. Increasing H produced a curve that simulated the data. A better fit of the data is obtained by increasing the air-filled tortuousity by a factor of 10 (Pe_v = 67), i.e. making gas diffusion unimportant. It is concluded from a comparison of the magnitudes of the dimensionless groups that the rates of volatilization and film transfer have little impact on the observed spreading.

A similar experiment at a slower water flow rate (S = 0.30, v_p = 0.00028 cm s^{-1}) was reported by Hutzler *et al.* [1989b]. Because gas diffusion is predominant for low-velocity conditions, their experimental results could be simulated with (4.1.1.38) as shown in Figure 4.1.1.8. Henry's constant was increased to 1.0 to fit the data.

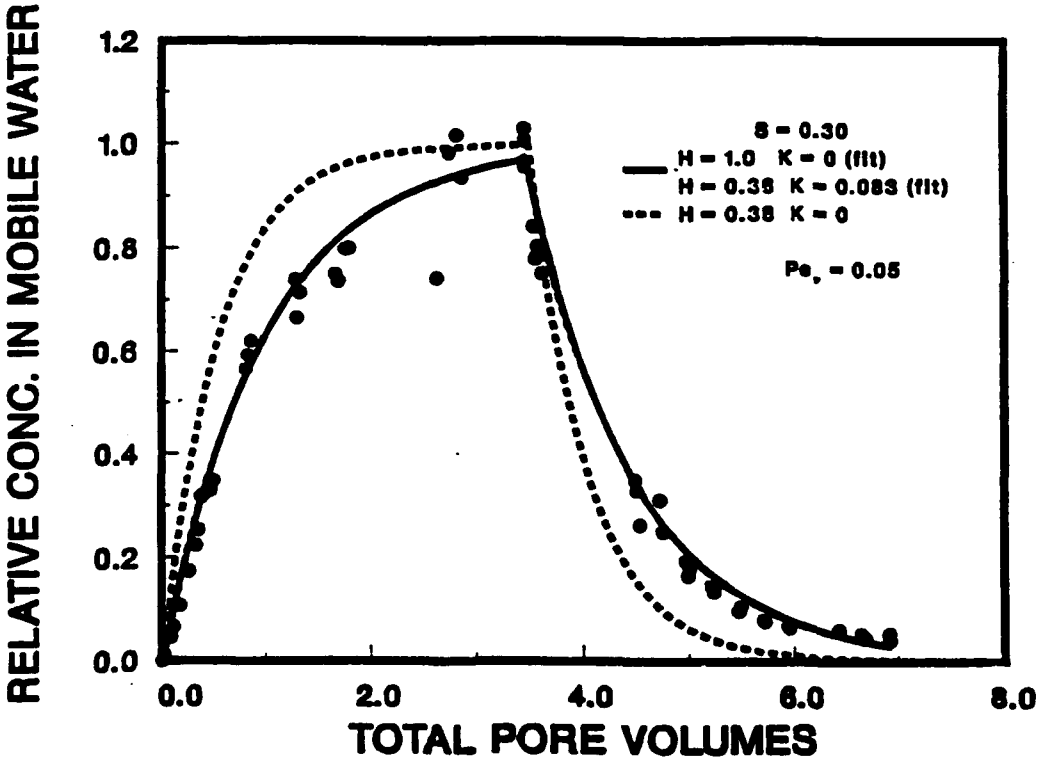


Figure 4.1.1.8. Complete mixing calculations for describing trichloroethene movement under a slow water velocity condition in a column of Ottawa sand at a degree of saturation of 0.30 (figure from Hutzler *et al.* [1989b], data from Krause [1987]).

Column Experiments with Aggregated Material

Natural soils are not as uniform as Ottawa sand. Instead they exhibit some structure and a pore size distribution. To simulate a structured or aggregated system, a uniformly-sized, aggregated porous material (APM) was used as a packing material. The particle size, hydraulic conductivity, and macroporosity of APM and Ottawa sand are similar. The APM particles contain a micro- or internal porosity inside which the flow of water is negligible in comparison to the flow around the particles.

The column containing APM was drained to a degree of saturation of 0.64 which gave about the same air-filled porosity as the Ottawa sand column. A 97.6 mg L^{-1} Br^- solution was then applied to the top of the column at a rate of 0.084 cm^3 s^{-1} (v_p = 0.0020 cm s^{-1}) for 9.68 hours. Figure 4.1.1.9 shows the effluent data and three model calculations. Because S_i and D_p were measured independently [McKenzie, 1990], only the magnitude of E_z is uncertain for this experiment.

The dotted line in Figure 4.1.1.9 is a prediction using the D_p measured in the batch experiment and E_z predicted by (4.1.1.56) (Pe_b = 860). The dashed line is obtained by multiplying the value of E_z fit to the unsaturated run of Br^- in the sand by the ratio of the interstitial water velocity in the APM to that in the sand (Pe_b = 2.0). Because this prediction is close, it adds confidence to the fit E_z for the unsaturated sand run. The solid line is obtained by adjusting E_z for a better fit to the data (Pe_b = 6.0).

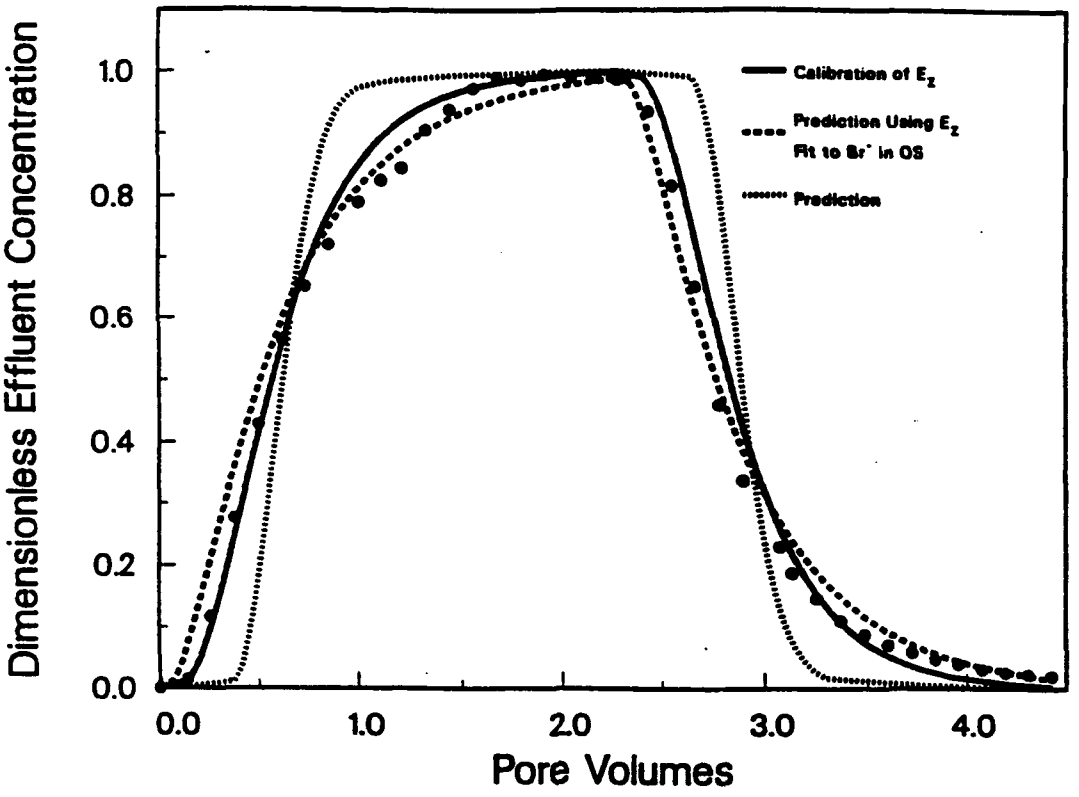


Figure 4.1.1.9. Model predictions and fit of bromide movement in a column of aggregated porous material at a degree of saturation of 0.64 (data from Krause [1987]).

After eluting the Br^- from the unsaturated APM a 1160 $\mu g\ L^{-1}$ solution of TCE was fed at the same rate (v_p = 0.0020 cm s^{-1}) for 12.18 hours. The breakthrough and elution of TCE from the unsaturated APM column is displayed in Figure 4.1.1.10. As was the case for estimating the parameters for the TCE run in the sand, H and E_v are not known with certainty for this experiment. A model prediction using the E_z fit for the unsaturated Br^- run in APM (Pe_b = 6.0) and an H of 0.7 (Dg_v = 1.2; R_d = 1.4), which was used to described the TCE movement through the unsaturated sand, is shown in Figure 4.1.1.10 as a dashed line. A better fit of the data is obtained by increasing the air-filled tortuousity by a factor of 10 (Pe_v = 50), as was done in the description

of TCE diffusion through the sand. The air-filled porosity of the APM is equal to 0.25, which is almost the same as the sand (0.21). Gas diffusion is not an important transport mechanism in the APM (Pe_v = 50) in comparison to liquid dispersion (Pe_b = 2.0) and intraaggregate diffusion (Ed_p = 0.52). Again, film transfer (St_b = 660) and air-water mass transfer resistance (St_v = 4.8) are unimportant. The dotted line is calculated using a literature value of H of 0.4 (Dg_v = 0.68; R_d = 1.1). It is not known at this time whether TCE vapors are sorbing to the column or soil materials or whether Henry's constant is, indeed, higher in unsaturated soil.

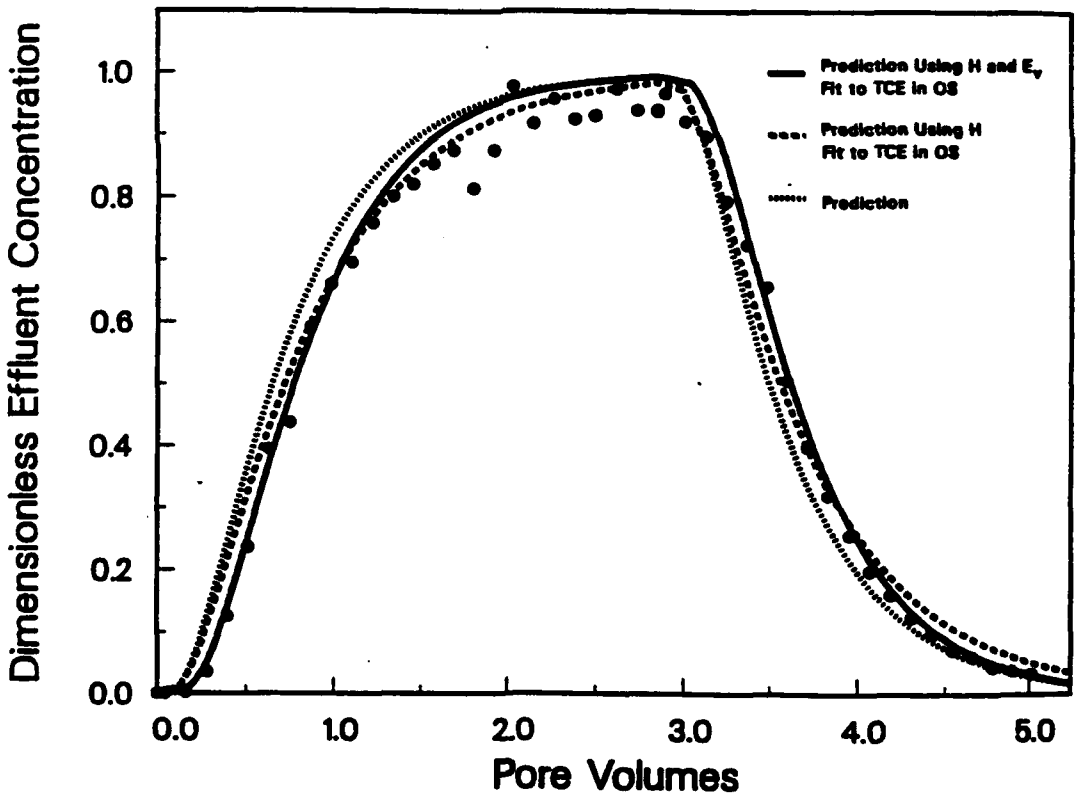


Figure 4.1.1.10. Model predictions of trichloroethene movement in a column of aggregated porous material at a degree of saturation of 0.64 (data from Krause [1987]).

4.1.1.8 Summary of Unsaturated Water Flow Column Results. An integrated model development, numerical verification, and experimental validation approach leads to a better understanding of subsurface chemical transport. The model developed here accounted for the transport of dissolved, nondegradable organic chemicals in unsaturated soil columns. The following mechanisms were included in the model development: advection and dispersion in air and in water, air-water mass transfer, mobile-immobile water mass transfer, intraaggregate diffusion, and sorption. The model was solved numerically using orthogonal collocation. The numerical approximation was verified by comparing calculations to analytic solutions for simplified conditions. Validation experiments were performed to assess the predictive capabilities of the column model. Two soil materials were used: a cohesionless sand and an aggregated soil. Trichloroethene was used for the volatile chemical and bromide was used as a tracer. Experiments were performed for different water flow conditions.

Model sensitivity calculations showed that it is not always possible to distinguish the impacts of various mechanisms by observing the shape of breakthrough curves. Furthermore, experimental results alone do not always provide sufficient information to ascertain which mechanisms are important. Laboratory experiments and numerical calculations indicate the following about the mechanisms affecting the transport of nondegradable organic compounds with water flow in unsaturated soil columns: (1) liquid dispersion and intraaggregate diffusion are important for cohesionless and aggregated soils, (2) vapor diffusion is not important in comparison to liquid advection and dispersion except for low water velocities (*e.g.* for Ottawa sand, low

velocity is defined as average pore water velocities less than 0.0003 cm s^{-1}), and (3) interfacial mass transfer rates (air-water and mobile-immobile water) are fast. These results were observed for average pore water velocities less than 0.07 cm s^{-1} in Ottawa sand and 0.02 cm s^{-1} in the aggregated material. Nonvolatile tracer experiments are sometimes needed to measure transport rates such as liquid dispersion and intraaggregate diffusion. Intraaggregate diffusion rates were successfully measured for the aggregated material in independent batch experiments, however, diffusion rates in immobile water and the amount of immobile water could not be predicted independently for the sand column. Discrepancies occurred between predicted air-water-soil chemical equilibrium for both soil materials, however, it is uncertain whether this was due to chemical interactions with the column apparatus.

Even though this model development included advective air transport, laboratory experiments were not performed it in conjunction with water-phase movement. In addition, the air diffusion representation in this model was not satisfactorily tested because air diffusion was unimportant for the conditions studied above. This model could not be used for no-water-flow conditions because time was normalized by advective water movement. A model and set of laboratory experiments are developed below to examine vapor-phase movement with no water flowing.

4.1.2 Vapor Extraction Column Model

4.1.2.1 Conceptual Picture. In vapor extraction, it is desirable to minimize the infiltration of water so to simulate this, the air and water flow column model is modified to accommodate situations where there is no flow of water. A situation where neither air nor water are flowing was studied by Bednar [1990]. The mechanisms being considered are: (a) air advection, (b) gas diffusion, (c) liquid diffusion in pores filled with immobile water, (d) mass transfer resistance at the air-water interface, (e) partitioning between the air-water phases, and (f) sorption to soil organic matter from aqueous solution. Even though the following model is conceptually simpler than the previous one, it is developed afterward and many of the steps that were applied in the simplification and verification parallel those taken for the air and water flow column (see Sections 4.1.1.5 and 4.1.1.6).

The model is developed to describe the movement of an organic vapor in laboratory columns of unsaturated soil where the mechanisms described above are operative. Figure 4.1.2.1 is an adaption of Figure 4.1.1.1 which shows a conceptual picture of a soil column that is used to develop a column model for vapor extraction. The soil system is divided into the two zones: mobile air and aggregates comprised of immobile water and solid soil particles. Mass balances on these zones result in two partial differential equations.

The air depicted in Figure 4.1.2.1 is assumed to be a continuous phase. The smallest pores, such as those contained in aggregates, contain water that is immobile. It is assumed that the water surrounds the soil surfaces.

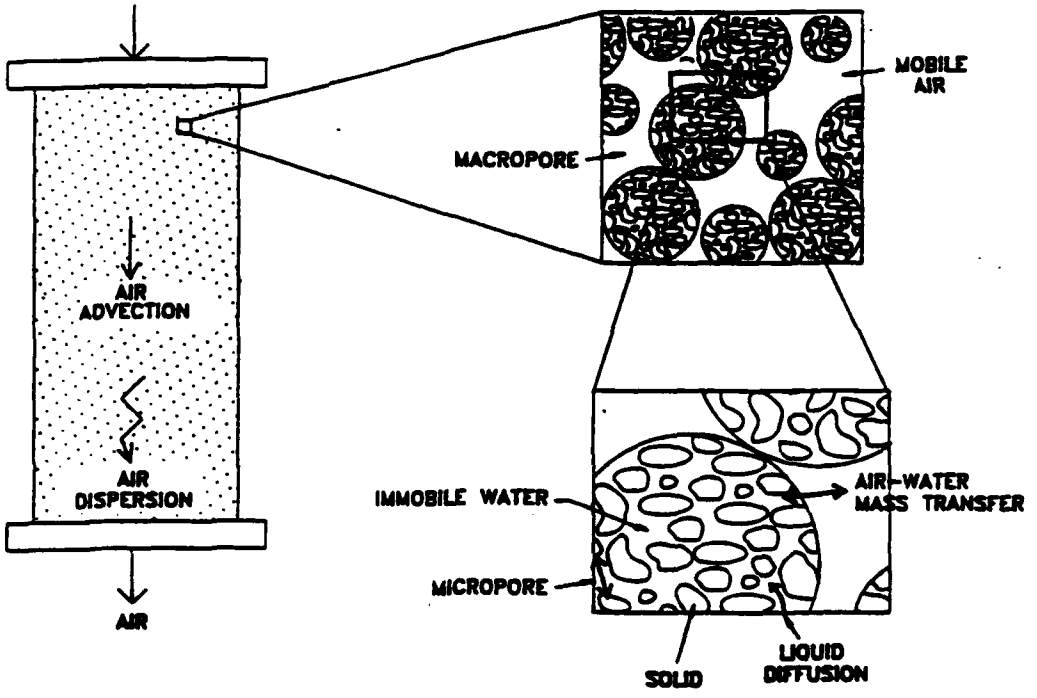


Figure 4.1.2.1. Conceptual picture for developing the vapor extraction column model (this figure is an adaption of Figure 4.1.1.1).

4.1.2.2 Derivation of the Dimensioned Equations. A mass balance on the air zone results in the following equation:

$$\frac{\partial C_v(Z,T)}{\partial T} = E_v \frac{\partial^2 C_v(Z,T)}{\partial Z^2} - u \frac{\partial C_v(Z,T)}{\partial Z} - K_L a \left[\frac{C_v(Z,T)}{H} - C_p(R=R_a,Z,T) \right] \qquad (4.1.2.1)$$

Equation (4.1.2.1) describes the change in vapor concentration ($C_v(Z,T)$) with respect to time. The terms on the right side represent: gas diffusion, gas advection, and air-water mass transfer. It is assumed that the vapor concentration gradient ($\partial C_v(Z,T)/\partial T$) in the axial direction is small over the diameter of an aggregate.

Boundary conditions for (4.1.2.1) are derived from the fact that soil columns are closed reactors [Levenspiel, 1962]. In an attempt to

force the numerical method to conserve chemical mass, an overall mass balance is used for one boundary condition. The difference between the mass of chemical entering and leaving the column by advection in air must equal the mass accumulating in the air and in the aggregates:

$$u[C_{vo}(T) - C_v(Z=L,T)] = \frac{\partial}{\partial T}\int_0^L \left[C_v(Z,T) + \frac{3\rho_s(1-\epsilon)}{\epsilon(1-S)}\int_0^{R_a} Y(R,Z,T)R^2dR \right] \partial Z \qquad (4.1.2.2)$$

The influent concentration in the air $C_{vo}(T)$ is allowed to vary, and, typically, the influent concentration is zero. The exit boundary condition is still (4.1.1.8).

Intraaggregate diffusion is still represented by (4.1.1.3) along with the symmetry boundary condition (equation (4.1.1.10)). The other boundary condition for (4.1.1.3) is derived by performing a mass balance on an aggregate. A change in mass of chemical in an aggregate is equal to the mass transferred to the aggregate from the air:

$$\frac{\partial}{\partial T}\int_0^{R_a} Y(R,Z,T)R^2dR = \frac{K_L a\epsilon(1-S)}{\rho_s(1-\epsilon)}\left[\frac{C_v(Z,T)}{H} - C_p(R=R_a,Z,T) \right] \qquad (4.1.2.3)$$

The definition of $Y(R,Z,T)$ is the same as shown in (4.1.1.3), and it is assumed that (4.1.1.4) describes the sorption equilibrium. Equation (4.1.2.3) is consistent with the assumption that axial gradients in $C_v(Z,T)$ are small over the diameter of an aggregate. Like (4.1.2.2), (4.1.2.3) attempts to conserve mass during the numerical solution of the model equations.

In summary, the column model equations for conditions where there

is air flow but no water flow consist of (4.1.2.1), (4.1.2.2), (4.1.1.8), (4.1.1.3), (4.1.1.10), and (4.1.2.3) and the initial condition given by (4.1.1.5).

4.1.2.3 Conversion to Dimensionless Form. As was done in Section 4.1.1.3, the complexity of the above equations is reduced, by converting them to a dimensionless form. Concentration, space, and time variables are normalized by the variables listed in Table 4.1.2.1. Throughput (t) is defined here by assuming that a soil column is initially free of chemical and that the influent concentration is constant ($C_{vo}(T) = C_{vn}$). Throughput is equal to the ratio of chemical mass fed or removed in air to the mass contained in the column at equilibrium with C_{vn}.

Table 4.1.2.1. Variable Substitutions to Convert Dimensioned Equations into a Dimensionless Form.

Dimensioned Variable	Substitution	Dimensionless Variable
$C_p(R,Z,T)$	$C_{bn}\ c_p(R,Z,T)$	$c_p(R,Z,T)$
$C_v(Z,T)$	$C_{vn}\ c_v(Z,T)$	$c_v(Z,T)$
$Y(R,Z,T)$	$Y_n\ y(R,Z,T)$	$y(R,Z,T)$
$C_{pi}(R,Z)$	$C_{bn}\ c_{pi}(R,Z)$	$c_{pi}(R,Z)$
$C_{vi}(Z)$	$C_{vn}\ c_{vi}(Z)$	$c_{vi}(Z)$
$C_{vo}(T)$	$C_{vn}\ c_{vo}(T)$	$c_{vo}(T)$
Z	$L\ z$	z
R	$R_a\ r$	r
T	$\frac{u}{L}\left[1 + \frac{S}{(1-S)H} + \frac{\rho_s(1-\epsilon)KC_{bn}^{1/n-1}}{\epsilon(1-S)H}\right]^{-1} t$	t

$C_{bn}=C_{vn}/H$ and $Y_n=\epsilon S_i C_{bn}[\rho_s(1-\epsilon)]^{-1}+KC_{bn}^{1/n}$, where $C_{vn}=\max(C_{vo},C_{vi})$.

Soil column model predictions in terms of relative (dimensionless) concentration as a function of dimensionless time (throughput) are characterized by the groups defined in Table 4.1.2.2. Like those listed in Table 4.1.1.2, these groups represent mass transfer mechanisms and chemical distributions at equilibrium. Because in this case, air is the only fluid for axial transport, the mass transfer groups are based on the rate of mass transport by air advection, and the chemical distribution groups are based on chemical mass in air. The magnitudes of the three mass transfer groups (air Peclet (Pe_v), immobile-water diffusion modulus (Ed_p), air-water Stanton (St_v)) represent the degree of spreading exhibited by a breakthrough curve [Roberts *et al.*, 1987]. A large value of any of these groups indicates a small contribution from the corresponding mechanism towards the observed spreading.

Table 4.1.2.2. Definitions of Dimensionless Groups.

Group	Definition	Equation
Mass Transfer Groups:		
Ed_p	$\frac{\text{rate of diffusion in water}}{\text{rate of advection in air}}$	$\frac{D_p Dg_p L}{uR_a^2}$
Pe_v	$\frac{\text{rate of advection in air}}{\text{rate of diffusion in air}}$	$\frac{uL}{E_v}$
St_v	$\frac{\text{rate of transport across air-water interface}}{\text{rate of advection in air}}$	$\frac{K_L aLS}{3u(1-S)H}$
Chemical Distribution Groups:		
Dg_p	$\frac{\text{mass of chemical in water}}{\text{mass of chemical in air}}$	$\frac{S}{(1-S)H}$
Dg_s	$\frac{\text{mass of chemical adsorbed to soil}}{\text{mass of chemical in air}}$	$\frac{\rho_s(1-\epsilon)KC_{bn}^{(1/n-1)}}{\epsilon(1-S)H}$
Dg	$\frac{\text{mass of chemical in water and on soil}}{\text{mass of chemical in air}}$	$Dg_p + Dg_s$
R_d	$\frac{\text{velocity of chemical front}}{\text{velocity of air}}$	$1 + Dg$
$1/n$	isotherm intensity	

The dimensionless forms of the air mass balance (equation (4.1.2.1)) and the entrance boundary condition (equation (4.1.2.2)) are

$$\frac{\partial c_v(z,t)}{\partial t} = [1 + Dg]\left[\frac{1}{Pe_v}\frac{\partial^2 c_v(z,t)}{\partial z^2} - \frac{\partial c_v(z,t)}{\partial z} - 3St_v[c_v(z,t) - c_p(r=1,z,t)]\right] \qquad (4.1.2.4)$$

$$c_{vo}(t) - c_v(z=1,t) = \frac{1}{[1 + Dg]}\frac{\partial}{\partial t}\int_0^1 \left[c_v(z,t) - 3Dg\int_0^1 y(r,z,t)r^2 dr\right]\partial z \qquad (4.1.2.5)$$

The dimensionless form of the exit boundary condition (equation (4.1.1.8)) is (4.1.1.14).

The dimensionless form of the intraaggregate mass balance (equation (4.1.1.3)) is

$$\frac{\partial y(r,z,t)}{\partial t} = \frac{Ed_p[1 + Dg]}{Dg}\frac{1}{r^2}\frac{\partial}{\partial r}\left[r^2\frac{\partial c_p(r,z,t)}{\partial r}\right] \qquad (4.1.2.6)$$

The dimensionless total intraaggregate concentration must satisfy

$$y(r,z,t) = \frac{Dg_p c_p(r,z,t) + Dg_s c_p(r,z,t)^{1/n}}{Dg} \qquad (4.1.2.7)$$

Equation (4.1.2.7) reduces to $y(r,z,t) = c_p(r,z,t)$ for $1/n = 1$.

The dimensionless form of (4.1.1.10) is unchanged from (4.1.1.20). The dimensionless form of (4.1.2.3) is

$$St_v[c_v(z,t) - c_p(r=1,z,t)] = \frac{Dg}{[1 + Dg]} \frac{\partial}{\partial t} \int_0^1 y(r,z,t)r^2dr \qquad (4.1.2.8)$$

The dimensionless initial condition for solving these equations is (4.1.1.22).

Converting the vapor extraction column model into a dimensionless form reduces the number of parameters that characterize a solution from 13 (a, D_p, E_v, H, K, K_L, L, 1/n, R_a, S, u, ϵ, ρ_s) to 6 (Dg_p, Dg_s, Ed_p, 1/n, Pe_v, St_v). It is also easier to characterize a solution in terms of these groups. Three groups (Ed_p, Pe_v, St_v) affect only the shape of a breakthrough curve. Eight parameters (H, K, L, 1/n, S, u, ϵ, ρ_s) impact both shape and position; the other five affect only the shape.

4.1.1.4 Numerical Solution. The dimensionless form of the above column model (equations (4.1.1.14), (4.1.1.20), (4.1.1.22), and (4.1.2.4-8)) is solved numerically in the same manner as in Section 4.1.1.4.

Application of orthogonal collocation (OC) to the air mass balance and its boundary conditions (equations (4.1.2.4), (4.1.2.5), and (4.1.1.14)) yields J ordinary differential equations (ODEs), where J is the number of axial collocation points. Additional ODEs (JxI, where I is the number of radial collocation points) are produced by the application of OC to the intraaggregate mass balance and its boundary conditions (equations (4.1.2.6), (4.1.2.7), and (4.1.1.20)). Figure 4.1.2.2 is a schematic of the OC discretization of the solution domain and shows the coupling of the ODEs. This system of ODEs is again solved using an algorithm called GEAR, which is found in the International Mathematics and Scientific Library (IMSL). The application of OC is shown below in the order in which GEAR receives the derivatives.

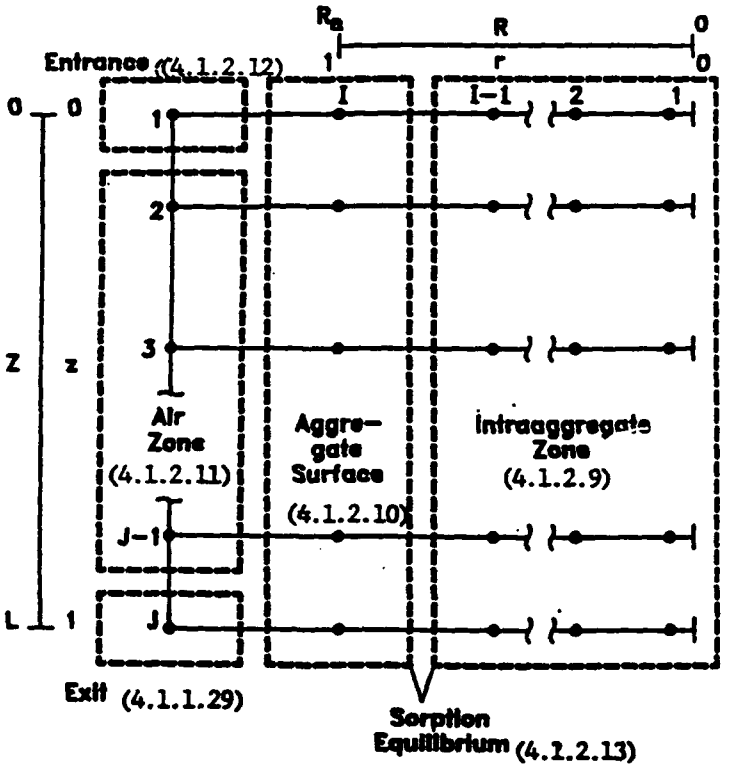


Figure 4.1.2.2. General schematic of the coupling of the ordinary differential equations resulting from the application of orthogonal collocation to the partial differential equations comprising the vapor extraction column model (this figure is an adaption of Figure 4.1.1.2). Equation numbering is in parentheses.

The application of OC to (4.1.2.6) results in:

$$\frac{dy(i,j,t)}{dt} = \frac{Ed_p[1 + Dg]}{Dg} \sum_{n=1}^{I} B^r_{i,n}\, c_p(n,j,t) \qquad (4.1.2.9)$$

Figure 4.1.2.2, which is an adaption of Figure 4.1.1.2, shows that (4.1.2.9) is evaluated at I-1 radial collocation points at each axial collocation point (j = 1 to J). The descriptions of the OC coefficient matrices is given in Section 4.1.1.4. Because the matrix is symmetrical, (4.1.1.20) is satisfied by the application of OC to (4.1.2.9) [Finlayson, 1980].

Applying OC to (4.1.2.7) and solving for the change in the total intraaggregate concentration at r equal to 1 leads to the following condition at the aggregate surface:

$$\frac{dy(I,j,t)}{dt} = \frac{1}{W^r_I}\left[\frac{[1 + Dg]St_v}{Dg}[c_v(j,t) - c_p(I,j,t)] - \sum_{p=1}^{I-1} W^r_p \frac{dy(p,j,t)}{dt}\right] \qquad (4.1.2.10)$$

Equation (4.1.2.10) is evaluated at all axial collocation locations.

The application of OC to the air mass balance (equation (4.1.2.4)) gives the following equation:

$$\frac{dc_v(j,t)}{dt} = [1 + Dg]\left[\sum_{m=1}^{J}\left[\frac{B^z_{j,m}}{Pe_v} - A^z_{j,m}\right] c_v(m,t) - 3St_v[c_v(j,t) - c_p(I,j,t)]\right] \qquad (4.1.2.11)$$

Equation (4.1.2.11) is evaluated at the J-2 internal axial collocation locations shown in Figure 4.1.2.2 (j = 2 to J-1).

An entrance condition is obtained by using OC to convert (4.1.1.14) and (4.1.2.5) to ODEs and solving for the derivative at j equal to 1:

$$\frac{dc_v(1,t)}{dt} = \left[W^z_1 - W^z_J \frac{A^z_{J,1}}{A^z_{J,J}}\right]^{-1}\left[[1 + Dg][c_{vo}(t) - c_v(J,t)] - \sum_{m=2}^{J-1}\left[W^z_m - W^z_J \frac{A^z_{J,m}}{A^z_{J,J}}\right]\frac{dc_v(m,t)}{dt}\right] - 3Dg \sum_{m=1}^{J} W^z_m \sum_{p=1}^{I} W^r_p \frac{dy(p,m,t)}{dt} \qquad (4.1.2.12)$$

The exit (j = J) condition is then (4.1.1.29)

Evaluations of (4.1.2.9-12) and (4.1.1.29) are made after solving (4.1.2.7) for $c_p(r,z,t)$ at each of the radial collocation points:

$$y(i,j,t) = \frac{Dg_p c_p(i,j,t) + Dg_s c_p(i,j,t)^{1/n}}{Dg} \qquad (4.1.2.13)$$

For 1/n not equal to 1, values of $c_p(i,j,t)$ are determined with a root finding subroutine called DZBREN, which also is an IMSL algorithm, and for 1/n = 1, $y(i,j,t) = c_p(i,j,t)$.

Initially (t=0), (4.1.1.22) is used for concentration values at all of the collocation points. Condition (4.1.2.13) is used with (4.1.2.9-12) and (4.1.1.29) to calculate the initial derivatives, the derivatives are sent to GEAR, and GEAR returns values of $y(i,j,t)$, $c_b(j,t)$, and $c_v(j,t)$. Equation (4.1.2.13) is solved for $c_p(i,j,t)$ and the algorithm is repeated until the desired throughput is reached.

4.1.2.5 Model Simplifications and Verification. The model derived above simplifies in the same manner as the model described in Section 4.1.1.5. For example, if Ed_p and St_v are large, then the local equilibrium solutions apply here. These substitutions in (4.1.1.34) give identical results: $Pe = Pe_v$, t as defined in Table 4.1.2.1, and $c_b(z,t) = c_v(z,t)$. The plug flow solution in Section 4.1.1.5 applies here if Pe_v is large. The appropriate changes in (4.1.1.46) are: t as defined in Table 4.1.2.1; Ar = 0; Dg, Dg_p, Dg_s, and Ed_p as defined in Table 4.1.2.2; $St_b = St_v$; and $c_b(z,t) = c_v(z,t)$. Because the mechanisms included in this model are approximated in the same way as the model described in Section 4.1.1, the simplifications derived in Section 4.1.1.5 apply here. Therefore the numerical solution verification given in Section 4.1.1.6 applies to the numerical approximation above as well.

4.1.2.6 Model Validation. The experimental validation approach used in Section 4.1.1.7 started with simple conditions where the fewest mechanisms were important and proceeded to more complicated situations where the impacts of several mechanisms were observed. This same approach is followed in this section.

Materials and Methods. Experimental procedures for validating the air-flow column model are similar to those followed in Section 4.1.1.7 except here the mobile fluid is air. Given below are the experimental results and model simulations of the laboratory scale vapor extraction experiments. The column apparatus and experimental procedures used for these experiments are reported by McKenzie [1990]. A modification to the final version of the apparatus used by McKenzie [1990] was made to redo two of his experiments where the results could not be explained. The modification involved replacing the direct connection from the soil column to an FID with a 6-port switching valve and 2-ml sample loop (Valco Instruments, Inc., Houston, Texas) connected to a 5880 Hewlett-Packard gas chromatograph (GC) (Hewlett-Packard Co., Mt. View, California) with an FID. This modification made it possible to sample effluent gas periodically instead of continuously and, thus, integrated areas from the GC/FID analysis are used to determine concentrations instead of peak height, which was used by McKenzie [1990]. The peak height method is not as sensitive, precise, or accurate as the area integration technique. Experimental results that are taken from McKenzie [1990] are noted as such in the figure captions.

The experimental conditions are summarized in Table 4.1.2.3. The porous materials used in these vapor extraction column model validation experiments are the same as used in the validation of the column model

described in Section 4.1.1. The media characteristics are listed in Table 4.1.1.5. Two chemicals were used in these experiments: methane was used as a gas tracer for observing the impact of gas diffusion on subsurface vapor transport, and (2) toluene was used as the volatile organic contaminant. The chemical properties that are used for the model parameter estimation are listed in Table 4.1.2.4. Water viscosity and density are listed in Table 4.1.1.4. Moist soil experiments were performed by saturating the columns with water and then allowing the columns to drain by gravity. Air flow was induced either by a vacuum pump or a regulated tank.

Table 4.1.2.3. Experimental Conditions for Column Runs With Methane (CH_4) and Toluene (TOL) in Ottawa Sand (OS) and in Aggregates (APM).

Soil	Chemical	Influent Concentration (mg L^{-1})	Gas Flow Rate[a] (cm^3 s^{-1})	Degree of Saturation	Pulse Time (hours)	Total Time (hours)
OS	CH_4[b]	35.7	0.10 /0.10	0	3.70	10.65
OS	TOL[c]	60	0.097/0.097	0	2.56	4.25
OS	TOL[d]	1	0.076/0.062	0	27.75	181.22
OS	TOL[c]	60	0.094/0.094	0.26	5.41	9.19
APM	CH_4[b]	35.7	0.37 /0.33	0	1.64	3.36
APM	TOL[c]	60	0.084/0.084	0	6.39	10.92
APM	TOL[d]	1	0.23 /0.30	0.64	15.95	39.92

[a]Breakthrough/Elution
[b]Breakthrough gas is 95% argon/5% methane, elution gas is nitrogen.
[c]Breakthrough gas is toluene in air, elution gas is air.
[d]Breakthrough gas is toluene in nitrogen, elution gas is nitrogen, and relative humidity of the gases is 0% (relative humidity is greater than 0% for the other experiments).
Cross-sectional area of columns is 19.6 cm^2, column length is 30.13 cm, the temperature was 23 ^oC (±2), and atmospheric pressure was 101,300 Pa (±2000).

Table 4.1.2.4. Properties of Air, Methane, and Toluene at 23 °C that are Used for Model Parameter Estimation.

	Value
Air	
Viscosity, μ_g (g cm^{-1} s^{-1})	0.000184[a]
Density, ρ_g (g cm^{-3})	0.00130[a]
Toluene, TOL ($C_5H_5CH_3$)	
Molecular weight, M_A (g mol^{-1})	92.2[a]
Molar volume, V_A (cm^3 mol^{-1})	132.5[b]
Boiling point, T_b (K)	374[a]
Vapor Pressure, P_v (mm Hg)	28[d]
Solubility, C_s (mg L^{-1})	515[d]
Henry's constant, H (dimensionless)	0.27[c]
Methane, CH_4	
Molecular weight, M_A (g mol^{-1})	16.0[a]
Molar volume, V_A (cm^3 mol^{-1})	29.6[b]
Boiling point, T_b (K)	109[a]
Vapor Pressure, P_v (mm Hg)	205,000[d]
Solubility, C_s (mg L^{-1})	24[d]
Henry's constant, H (dimensionless)	27[d]

[a]From Weast [1981].
[b]LeBas [1915].
[c]Ashworth *et al.* [1988].
[d]MacKay and Shiu [1981].

Model Parameter Estimation. Parameter values for the model calculations are given in Table 4.1.2.5. These values were obtained either from the literature correlations given in Section 4.1.1.7, values fit in the water flow experiments (Table 4.1.1.6), direct measurements and by assumption, or, when necessary, by calibrating the model. The parameter for which the greatest uncertainty exists is $K_L a$, and, because for these experiments there is no water flow, the value of $K_L a$ that is assumed fast such that it has no impact on the results.

Table 4.1.2.5. Parameter Values for Column Model Calculations of the Validation Experiments.

Chem./Soil:	CH_4/OS	TOL/OS	TOL/OS	TOL/OS	CH_4/APM	TOL/APM	TOL/APM	Estimation Method[a]
Independently Determined Values:								
S	0	0	0[b]	0.26	0	0	0.64	Measured gravimetrically
u (cm s^{-1}) (BT)	0.016	0.015	0.012	0.014	0.045[c]	0.010[c]	0.042	Calculated from definition
u (cm s^{-1}) (EL)	0.016	0.015	0.0096	0.014	0.040[c]	0.010[c]	0.055	Calculated from definition
H	27	0.27	0.27	0.27	27	0.27	0.27	CH_4 [MacKay and Shiu, 1981]; TOL [Ashworth et al., 1988]
D_G (cm^2 s^{-1})	0.25	0.078	0.078	0.078	0.25	0.078	0.078	Equation (4.1.1.48)
D_p (cm^2 s^{-1})	NA	NA	NA	NA	$2.8(10^{-3})$	$8.7(10^{-4})$	$8.4(10^{-8})$	Dry: $D_G \tau_p^{-1}$; Moist: $D_l \tau_p^{-1}$
$D_p R_a^{-2}$ (s^{-1})	NA	NA	NA	$1.0(10^{-7})$	NA	NA	NA	Used value fit water flow exp. adjusted for diff. coef.
τ_a	1.4	1.4	1.4	2.9	4.1	4.1	4.1	Equation (4.1.1.49)
E_v (cm^2 s^{-1})	0.18	0.056	0.056	0.027	0.18	0.056	0.020	$D_G \tau_a^{-1}$
K $((cm^3/mg)^{1/n}mg/g)$	0	0	0.08	0, 0.02	0	0	0, 0.2	Assumed, Value fit from dry experiment adjusted for H.
1/n	1	1	1	1	1	1	1	Assumed
Values Resulting From Model Calibration:								
K $((cm^3/mg)^{1/n}mg/g)$	NC	0.08	2.10	NC	NC	0.8	NC	Fit
1/n	NC	NC	0.4	NC	NC	NC	NC	Fit

[a]Unless otherwise noted.
[b]Relative humidity (RH) is 0%, while RH is greater than 0% for the other experiments.
[c]Assumed air inside APM is immobile.
BT = breakthrough, EL = elution, NA = not applicable, and NC = no change.

Experimental Results. No biodegradation was observed over the duration of an experiment. Mass balance determinations showed that 95% or more of the chemical retained in a column during breakthrough was removed during extraction. The results that follow are summarized by the magnitudes of the dimensionless groups given in Table 4.1.2.6. When a group value is given in Table 4.1.2.6 as "NA" the mechanism for which it represents is unimportant so an arbitrary value is selected so that it has no impact on the model calculation. Diffusion in immobile water and liquid dispersion were the most important mechanisms in the water flow experiments; while gas diffusion was predominant in the dry vapor extraction column studies, and intraaggregate diffusion was significant only for moist conditions in the column containing APM.

Table 4.1.2.6. Dimensionless Group Values Resulting from Parameter Values Listed in Table 4.1.2.5.

Chemical/Soil:	CH_4/OS	TOL/OS	TOL/OS	TOL/OS	CH_4/APM	TOL/APM	TOL/APM
Independently Determined Values:							
Dg_p	0	0	0	1.3	0	0	6.6
Dg_s	0	0	0.43	0	0	0	0
R_d	1	1	1.43	2.3	1	1	7.6
Ed_p	NA	NA	NA	NA	1100	4700	0.3/0.25
St_v	100	100	100	100	100	100	100/100
Pe_v	2.7	7.9	6.2/5.1	22	7.5/4.0	5.5	69/83
1/n	1	1	1	1	1	1	1
Values Resulting From Model Calibration:							
Dg_s	NC	0.43	11.2	NC	NC	0.84	NC
1/n	NC	NC	0.4	NC	NC	NC	NC
R_d	NC	1.43	12.2	NC	NC	1.8	NC

NA - not applicable, input a large value to make unimportant. NC - not changed. Two values are given when the breakthrough and elution flow rates are different, the first is for breakthrough and the second for elution.

Extraction Experiments with Ottawa Sand

A gas tracer (95% argon/5% methane) experiment was performed first to measure the air-filled pore tortuosity (τ_a) of the sand column. Measured effluent concentrations of methane (CH_4) relative to an influent concentration of 35.7 mg L^{-1} are shown in Figure 4.1.2.3 as a function of the total pore volumes of air fed to the column. The breakthrough portion of the experiment lasted for 3.70 hours (this corresponds to 7.18 pore volumes for a gas flow rate of 0.10 cm^3 s^{-1}), after which the methane was eluted with clean air at the same rate for almost nine hours. The model prediction shown in Figure 4.1.2.3 is based on parameter values given in Table 4.1.2.5 and it agrees with the observed effluent concentrations.

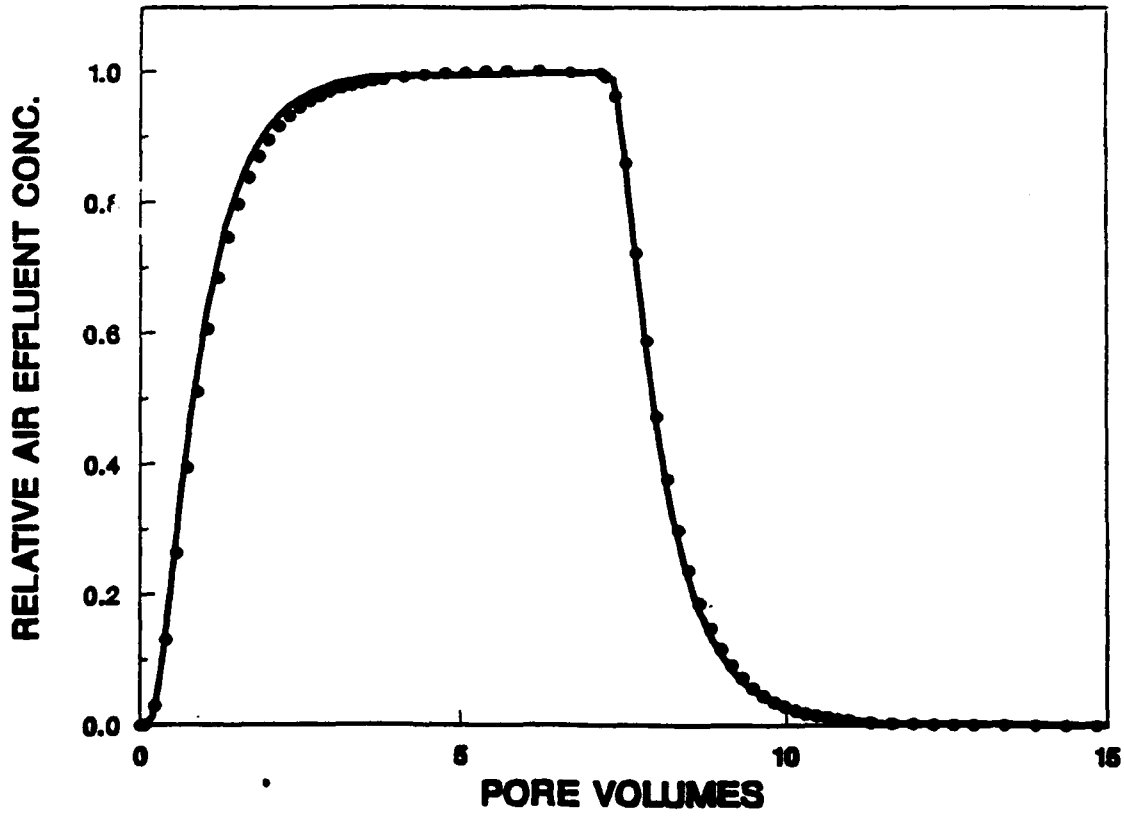


Figure 4.1.2.3. Model prediction of methane gas movement in a column of dry Ottawa sand.

Magnitudes of the dimensionless group values given in Table 4.1.2.6 indicate that the predominant spreading mechanism for the experiment shown in Figure 4.1.2.3 is gas diffusion (Pe_v = 2.6). In the unsaturated water flow experiments in sand reported in Section 4.1.1.7, the transport of bromide (Br^-) and of trichloroethene (TCE) could only be described by accounting for diffusion in immobile water (see Figure 4.1.1.6). For the experimental results shown in Figure 4.1.2.3, a model calculation that accounts for diffusion in an immobile air fraction (10% of the void volume, which is equivalent to the fraction of immobile water that was measured in the unsaturated water flow experiments with sand) gives identical results. Therefore if a fraction of the air in the sand column is immobile, the diffusion into these zones is fast enough for the existing advective rates that it can be ignored.

Toluene (TOL) vapor at a concentration of 60 mg L^{-1} was then drawn through the dry sand column at a rate of 0.097 cm^3 s^{-1} for 2.56 hours (4.59 pore volumes). The results are shown in Figure 4.1.2.4. The model prediction is shown as a solid line, and this calculation assumes no sorption of the toluene vapors. Although the relative humidity (RH) of the influent gases was not measured, the RH was above zero because the gases were in contact with liquid water prior to being drawn through the column. Sorption of TCE onto the Ottawa sand was not observed in saturated flow experiments [Hutzler *et al.*, 1989b] because TCE is nonpolar and no organic material is associated with the Ottawa sand [Hasset *et al.*, 1983]. Because toluene is also nonpolar, it is reasonable to assume no sorption of the toluene onto the sand. The observed effluent data is fit with the model by adjusting K from zero to 0.08 cm^3 g^{-1}. The observed retardation can not be attributed to

partitioning in condensed or adsorbed water vapor because the degree of saturation would have to be at least 0.10, which would have been measured in the gravimetric analysis at the completion of the experiment. The calibrated model calculation is shown as a dashed line. Again, the predominant spreading mechanism is gas diffusion (Pe_v = 8.1) and the air-filled tortousity used in the methane experiment is used to describe the effective gas diffusion rate here. Another experiment was performed by McKenzie [1990] at a higher air flow rate, and the model predicted the data using the K fit for this experiment.

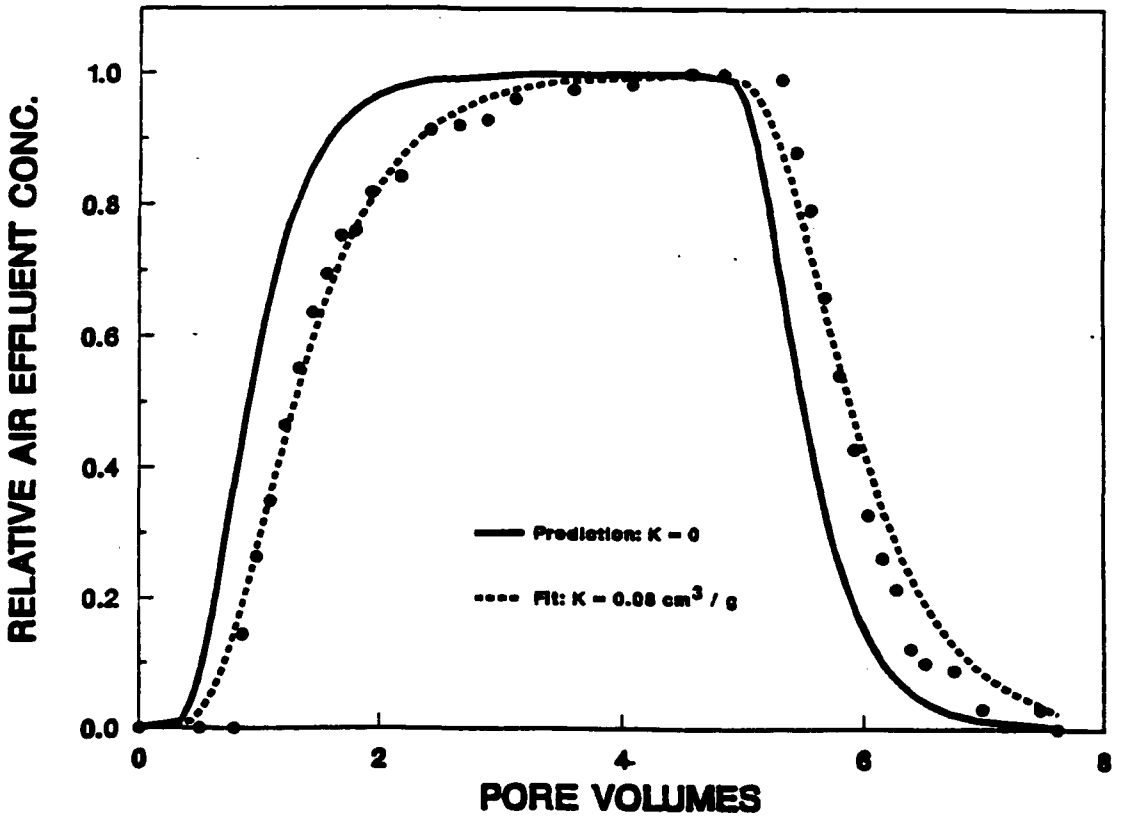


Figure 4.1.2.4. Model prediction and fit of toluene vapor movement in a column of dry Ottawa sand (data from McKenzie [1990]).

Since the influent concentration of toluene (60 mg TOL L^{-1} = 15 mm Hg) was over 50% of the vapor pressure, it is also possible that the

observed retardation was due instead to condensation of the toluene vapors. The dry experiment was repeated at a lower toluene influent concentration (1 mg L^{-1}) to determine if condensation occurred. The relative humidity of the influent gases was reduced to zero to observe the impact of water vapor on the sorption of organic vapors. The results of this 0% RH experiment are shown in Figure 4.1.2.5. The dashed line is a calculation based on the K fit in the previous experiment where water vapor was present.

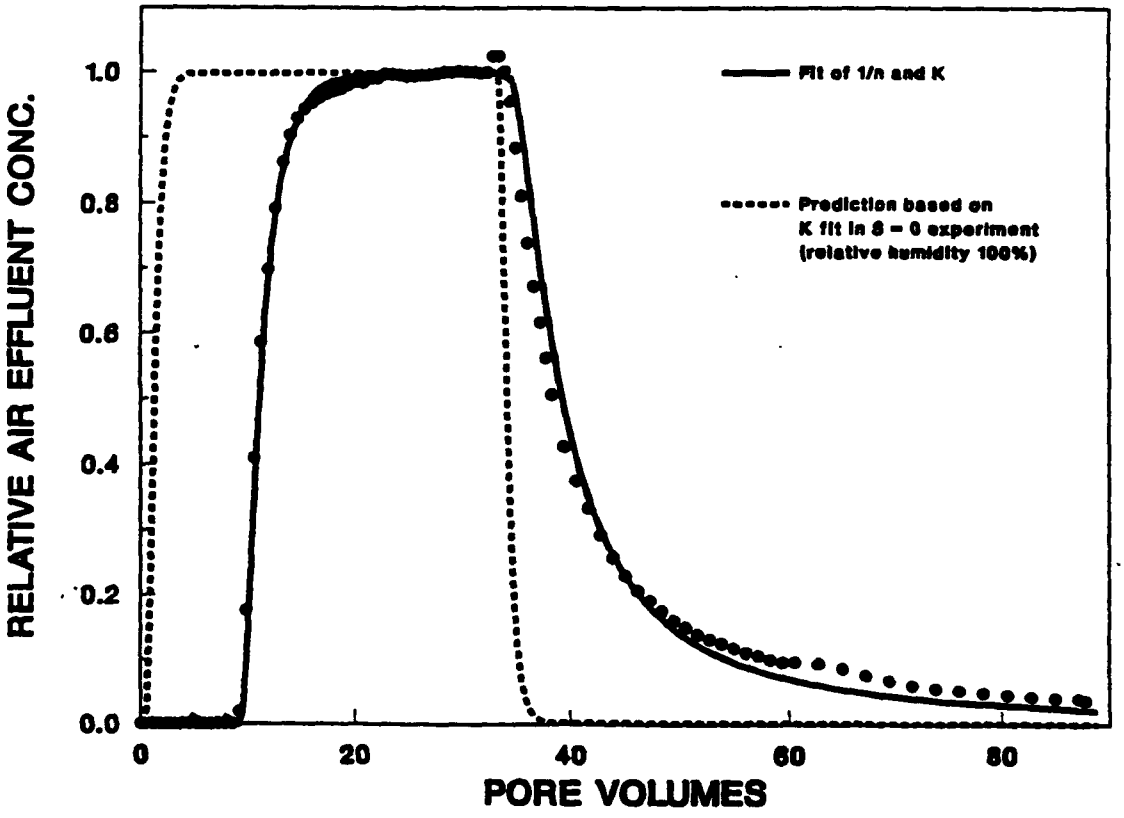


Figure 4.1.2.5. Model prediction and fit of toluene vapor movement in a column of dry Ottawa sand at 0% relative humidity.

The retardation observed during this experiment is about 10 times more in the 0% RH experiment. The breakthrough part of the experiment lasted 27.75 hours (38.94 pore volumes at an extraction rate of 0.076

$cm^3\ s^{-1}$) and the elution took 154 hours (176 pore volumes at an extraction rate of 0.062 $cm^3\ s^{-1}$). Retardation would be less if condensation was occurring because the influent concentration was less. Moreover, the breakthrough is sharper than the elution, indicating nonlinear ($1/n < 1$, cf. Crittenden *et al.* [1986]) sorption of organic vapors in the absence of water vapor. Nonlinear sorption equilibrium was also observed by Chiou and Shoup [1985] under dry conditions. The model is fit to the data resulting in a best fit 1/n equal to 0.4 and K of 2.1 $(cm^3\ mg^{-1})^{0.4}\ mg\ g^{-1}$. In this case, nonlinear equilibrium tended to diminish the impact of gas diffusion. Since water is present in most soils, further study of relative humidity impacts is not pursued here.

The Ottawa sand column was then saturated with water and allowed to drain. Remaining free water was removed by blowing water out of the bottom of the column with a pipette bulb. Toluene at a concentration of 60 mg L^{-1} was drawn through the column at a rate of 0.094 $cm^3\ s^{-1}$ for 5.41 hours (9.39 pore volumes). A comparison of a model prediction, assuming no adsorption, with the observed effluent concentrations is shown in Figure 4.1.2.6. The model calculation assumes all of the spreading is due to gas diffusion ($Pe_v = 22$, $Ed_p = 100$, $St_v = 100$). The dashed line is a model calculation using the sorption capacity fit in the dry experiment where water vapor was present, corrected for equilibrium between water and soil. The retardation due to dissolution of toluene vapors is much larger than a contribution due to sorption so it is not possible to determine if vapor sorption did occur. In the unsaturated water flow experiments in Ottawa sand, the amount of immobile water and the specific rate of intraaggregate diffusion ($D_p R_a^{-2}$) had to be calibrated to the tracer experiment because it was not

possible to independently determine the size of the immobile water regions. A model calculation using the value of $D_p R_a^{-2}$ fit in the water flow experiments is shown in Figure 4.1.2.5 as a dotted line (Pe_v = 22, Ed_p = 0.0012, St_v = 100). It is apparent that the immobile water regions present in situations where water flow occurs are different than those where there is no water flow. The model calculations in Figure 4.1.2.6 assume that air-water mass transfer rates (St_v = 100) are fast so the value of $K_L a$ is not important.

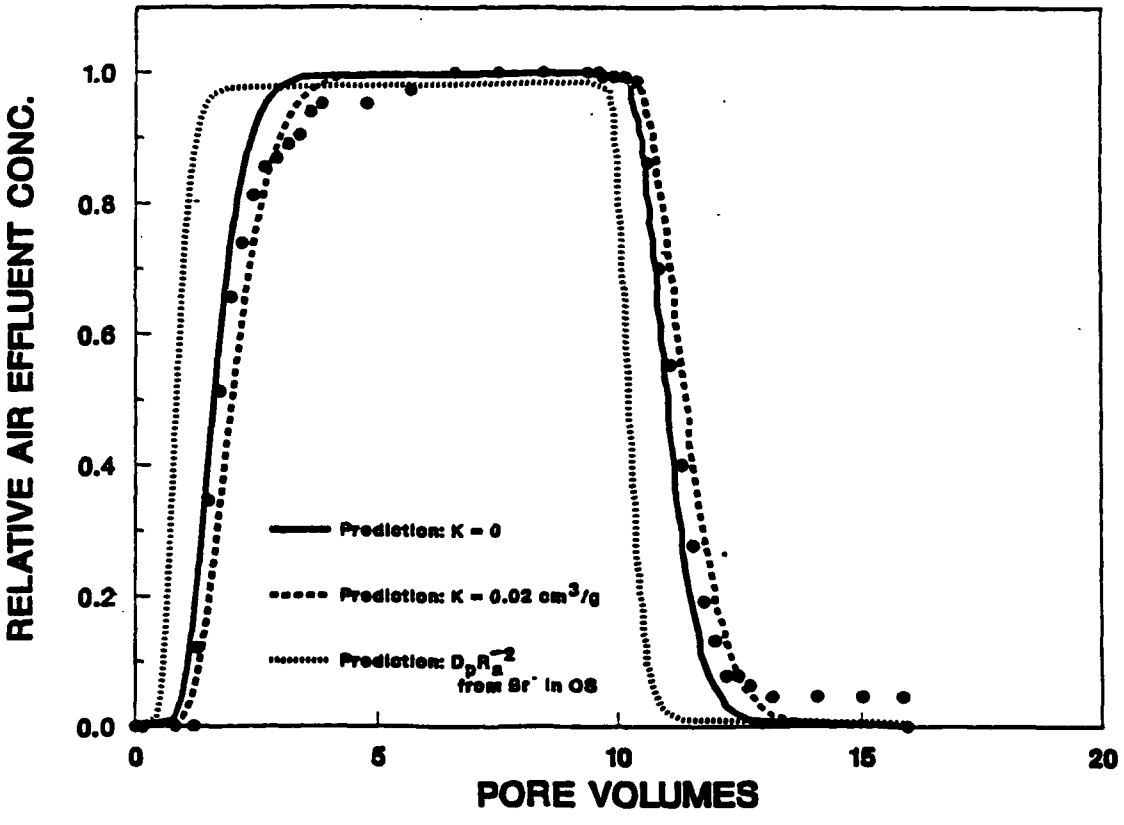


Figure 4.1.2.6. Model predictions of toluene vapor movement in a column of moist Ottawa sand (data from McKenzie [1990]).

Extraction Experiments with Aggregates

These experiments with an aggregated porous material (APM) were performed under dry conditions first and then a moist experiment to

observe the impact of soil aggregation on vapor extraction. The APM was used because its soil structure characteristics (size and tortuosity) were previously measured in the unsaturated water flow study described in Section 4.1.1.7.

Figure 4.1.2.7 shows the observed movement of methane in a dry column containing APM and a model prediction. The model calculation assumes that the air inside the aggregate particles is immobile and chemical transport inside the particles is due to gas diffusion. The intraaggregate diffusion rate in the dry material is fast enough (Ed_p = 1060) to be unimportant when compared to axial diffusion (Pe_v = 7.6).

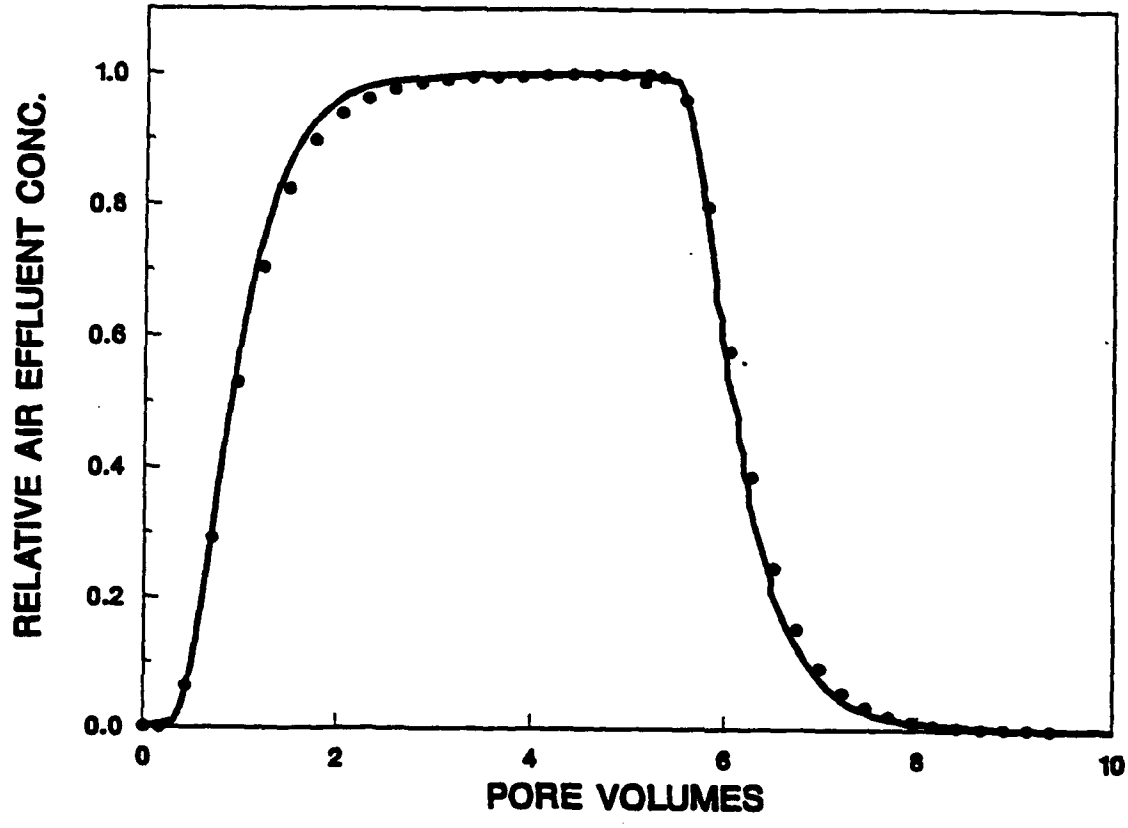


Figure 4.1.2.7. Model prediction of methane gas movement in a column of dry aggregated porous media.

Transport of toluene in the dry APM column is shown in Figure

4.1.2.8. A model prediction assuming no adsorption of the toluene vapors is shown as a solid line. The model is fit by adjusting K to 0.8 $cm^3\ g^{-1}$. Relative humidity was not measured but it was greater than zero. As in the dry methane experiment, axial gas diffusion with air flow is the predominant spreading mechanism (Pe_v = 5.5, Ed_p = 4700, St_v = 10,000), and as was the case in the dry sand experiment, toluene sorbed to the dry aggregates.

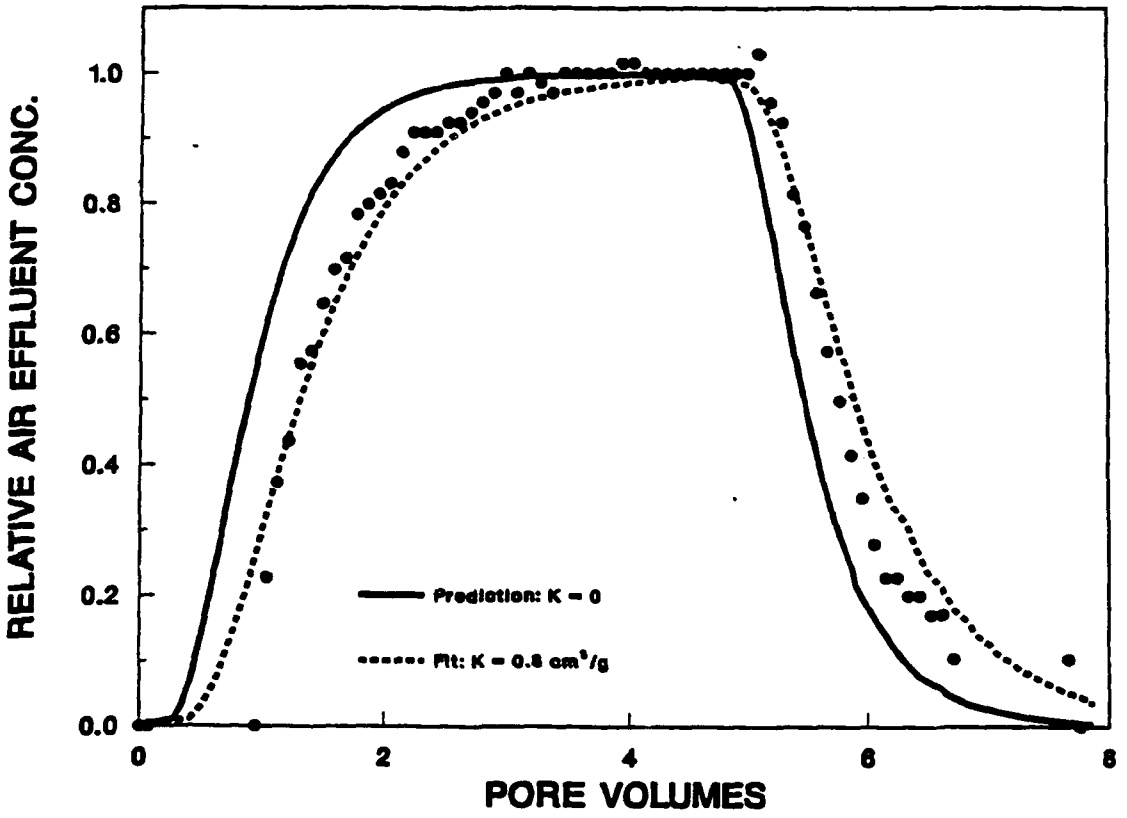


Figure 4.1.2.8. Model prediction and fit of toluene vapor movement in a column of dry aggregated porous media (data from McKenzie [1990]).

The aggregate particles were then saturated with water and packed in a column. Toluene vapor breakthrough and elution from the moist APM column are shown in Figure 4.1.2.9. A model prediction assuming no adsorption is shown as a solid line. The dashed line shows the impact

of including the amount of adsorption that was observed in the dry case. The dotted curved assumes that intraaggregate diffusion is fast. All three model calculations assume that air-water mass transfer rates are fast. In this moist APM experiment, intraaggregate diffusion could not be ignored, and the impact of intraaggregate diffusion would increase as the air flow rate is increased.

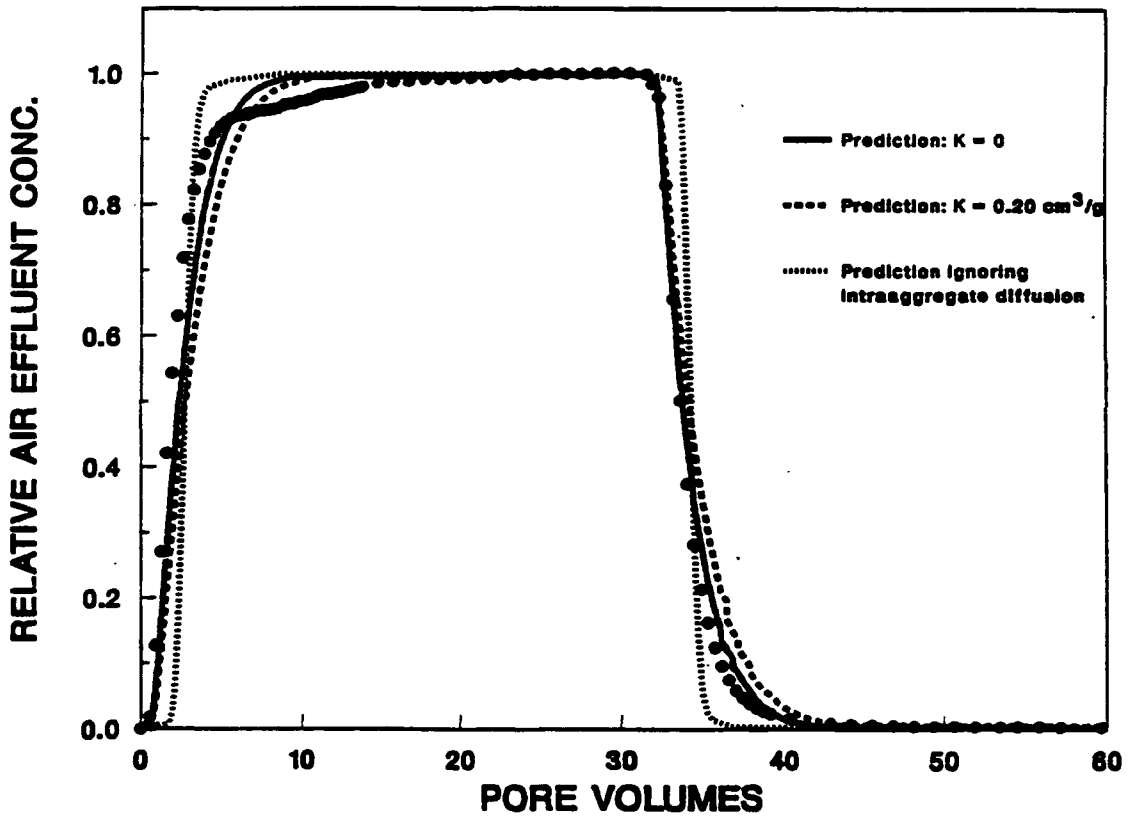


Figure 4.1.2.9. Model predictions of toluene vapor movement in a column of moist aggregated porous media.

4.1.2.7 Summary of Column Extraction Results. The approach demonstrated in the previous section is appropriate for developing a vapor extraction column model that simulates the removal of dissolved volatile organic chemicals from cohesionless and aggregated soils. This model was also solved with orthogonal collocation and the numerical

approximation verified. Validation experiments were performed with the same two porous materials (sand and an aggregated soil) used in the water flow column experiments. Toluene was used in the vapor extraction experiments for the volatile contaminant and methane was used as an insoluble gas tracer.

The results reported here indicate two important differences between extraction of chemicals from cohesionless soils and from aggregated soils. First, the moisture content of an aggregated soil is usually higher and this results in an increase in retardation due to partitioning between the air and water. Second, the presence of moisture in an aggregated soil increases extraction time because of intraaggregate diffusion rate limitations. For dry conditions, vapor transport in the aggregated material and the sand are similar. Vapor transport in either the aggregated soil or the sand is advection-diffusion dominant for low air velocities (0.04 and 1 cm s^{-1}, respectively) or dry conditions. Air-water mass transfer rates appear to be fast for the air velocities used here. Sorption of organic vapors can occur for dry conditions even when no organic material is present, and the sorption equilibrium was nonlinear for zero relative humidity.

There are differences between the important mechanisms for the vapor extraction experiments reported in this section and the unsaturated water flow results reported in Section 4.1.1.7. For the situations where water was flowing, vapor diffusion could be ignored, but gas diffusion could not be neglected when air was the only mobile fluid. Diffusion out of water was not important for the moist sand vapor extraction experiment; however, it was important in the water flow sand experiment. This discrepancy occurred even though the degrees of

saturation were about the same for the two experiments. The distribution of liquid water may be different between the water-flow and the no-water-flow conditions. The impact of intraaggregate diffusion that was observed in the unsaturated sand column for flowing water could not be used to predict the vapor extraction column results for moist, no-water-flow conditions.

Even though intraaggregate diffusion rates are difficult to predict, the impact of diffusion out of water should be tested in a vapor extraction model. Since the model sensitivity results showed that air-water mass transfer and intraaggregate diffusion can have similar impacts by appropriate manipulations of the mass transfer parameters, air-water mass transfer is used in lieu of intraaggregate diffusion in the two dimensional models derivations that follow. Liquid dispersion is important when water is flowing and vapor diffusion is important for air flow conditions, so both of these spreading mechanisms are included along with the primary direction of fluid flow in the following model developments.

4.2 TWO-DIMENSIONAL VAPOR EXTRACTION MODELS

This section describes the development of models that simulate the removal of volatile organic chemicals from soils using vapor extraction. The models developed here describe the vertical movement of aqueous solutions of nondegradable volatile organic chemicals and simultaneous removal with horizontal air flow. All of the assumptions regarding the mathematical descriptions of the important mechanisms discussed previously apply here with the following exceptions. The solution of coupled two-dimensional equations is more complicated than the solutions of the one-dimensional systems. To reduce the complexity, soil aggregation and nonlinear sorption are not considered in the two-dimensional model developments.

The numerical method used in the previous sections to solve the one-dimensional column models, called orthogonal collocation (OC), has been successfully used in fixed-bed applications, however, the solution of two-dimensional system equations using OC has been less successful. Im [1988] applied OC to transport equations that included advection in two parallel layers and transverse dispersion in each layer, but he was unable to verify the solution when the advective rates in both layers were of the same order of magnitude (*i.e.* when the vertical transport in each layer is significant). In this work, it is deemed necessary to demonstrate whether OC is appropriate for solving two-dimensional equations before including the effects of aggregation. It has already been proven that OC is appropriate for describing intraaggregate diffusion [Crittenden *et al.*, 1986; Gierke, 1986], and so it would be a simple task to include intraaggregate diffusion in the two-dimensional models. Doing so would significantly increase the already large number

(200) of equations by at least a factor of three.

Nonlinear sorption was shown to be important only in very dry conditions (see Figure 4.1.2.4). Moisture will probably be present in most vapor extraction applications, and because calculations for nonlinear sorption isotherms take at least ten- and often one hundred-times longer, 1/n is assumed to be 1 for the remaining model developments.

Many types of vapor extraction systems are currently in use throughout the country [Hutzler *et al.*, 1989a]. Two system geometries are studied here. The most common configuration is a grid of vertical vents, so the first set of models are derived to study the performance of a single vertical vent in a homogeneous soil system. Trench or planar extraction systems are modeled next using the same approach. The radial model development consists of the construction of a conceptual picture, derivation of dimensioned equations, conversion of dimensioned equations to dimensionless form, application of the numerical method to the dimensionless equations, simplification of the dimensionless equations and subsequent solution, and model sensitivity. Model calculations are used to determine the important removal mechanisms for field-scale systems, to give additional considerations for specifying air withdrawal rates and vent configuration, and to demonstrate the appropriateness of using a laboratory-scale modeling approach to develop field-scale models.

4.2.1 Radial Geometry

Vapor extraction systems are typically installed as a grid of vertical vents. Air flow towards a vent is then primarily in a horizontal, radial direction when the ground surface is sufficiently

covered to minimize short circuiting of air flow. For systems where the vents are installed on equally-spaced centers and the soil system is homogeneous, the flow could be considered to be axially symmetric. A two-dimensional, axially symmetric system as shown in Figure 2.1c corresponds to a situation where there is a small spill and one extraction vent is installed in the center of the contamination and several inlet vents are completed at the edge of the contamination. Water infiltration could occur while still not allowing air short circuiting if a compacted soil cover that is saturated is used as a cap.

The volume of an extraction vent is small in comparison to the volume of contamination; hence, the derivation of the continuum transport equations does not consider the transport within the vent packing or pipe. The assumption of vapor removal at R = 0 instead of the radius of the vent pipe or hole simplifies the numerical method used to solve the equations. In addition, this simplification allows an analytic solution to be obtained for radial-dispersed flow. Pressure drop calculations requires a value of the vent pipe radius, so a typical vent packing diameter is chosen for determining the system pressure drop.

4.2.1.1 Conceptual Picture. The equation derivations for radial geometry are for the cylindrical configuration shown in Figure 4.2.1.1. Equations are derived for downward (Z) water flow and horizontal, axially-symmetric (R) air flow. The radial extraction system assumes that air enters the soil at $R = R_i$ at a uniform rate. Vapor concentrations in the influent air can range from zero to the vapor pressure (typically, influent vapor concentrations are zero or equal to the effluent concentration from an off-gas treatment system used in a

closed-loop configuration). Subsurface vapors are removed at R = 0 at a flow rate equal to the inlet rate (*i.e.* air is assumed to behave as an incompressible fluid).

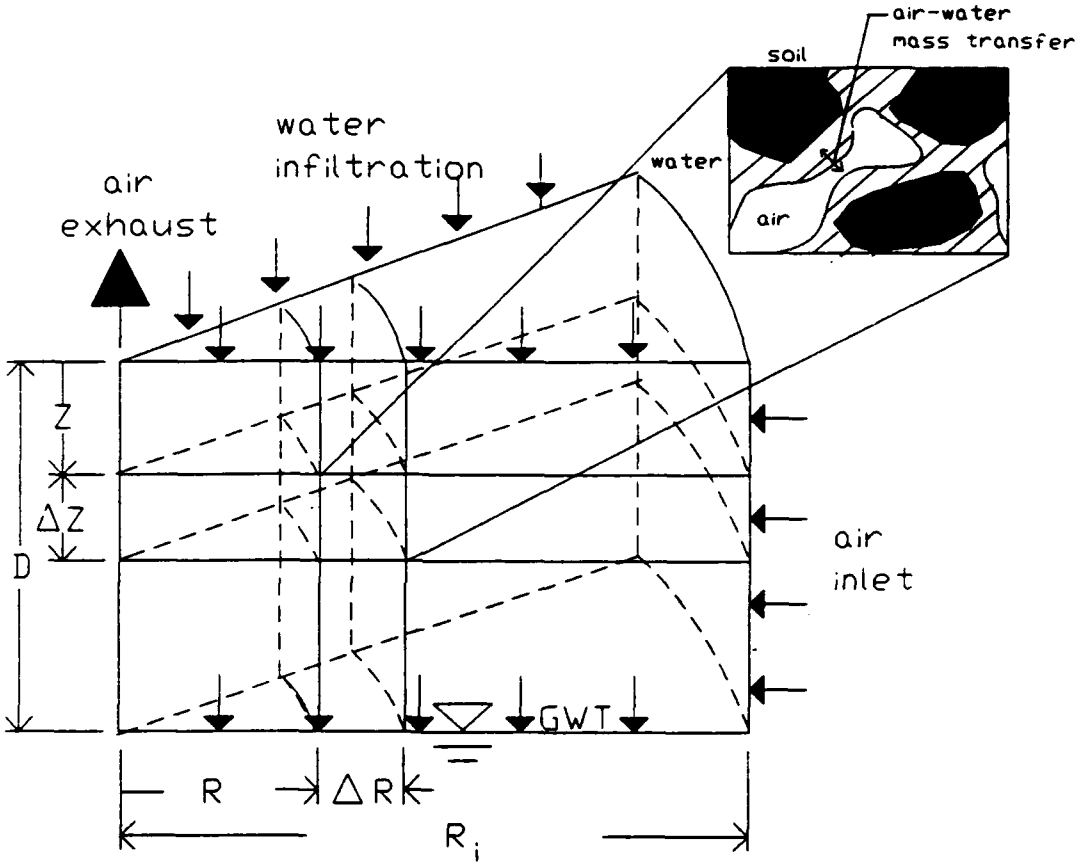


Figure 4.2.1.1. Conceptual picture of a vapor extraction system for radial flow.

Water containing a volatile chemical at a constant concentration ranging from zero to its solubility limit infiltrates the soil system at a uniform rate at Z = 0. The water flows downward through the soil towards the ground water table at Z = D. The ground water table is assumed to be horizontal and at a constant elevation. Reduced air pressures around extraction vents result in a concomitant rise of the ground water table [Johnson *et al.*, 1990b]. Although ground water table

fluctuations are important considerations in the design of vapor extraction systems, the models do not consider these effects because when necessary the fluctuations can and should be controlled. The soil system is assumed to be homogeneous so that porosity and bulk density can be considered constant and the capillary fringe is assumed to be thin in comparison to the ground water depth so that moisture content is also considered constant. For the following equation derivations it is also assumed that all of the water is mobile or that soil aggregation is insignificant. The impact of aggregation will be deduced from the effects of air-water mass transfer.

4.2.1.2 Equation Derivations. Dimensioned equations are obtained by performing mass balances on the differential soil volumes shown in Figure 4.2.1.1. Water flow occurs in the positive Z direction (downward); air flows in the negative R direction. Concentrations are assumed to be uniform at every radial distance (R) from the center at any depth (Z). In addition to mass transport by advection, chemical movement by axial dispersion in water and vertical and horizontal diffusion in air are considered. Mass transfer between the air and water is accounted for in the equation derivations, and chemical equilibrium between the water and soil is assumed to be instantaneous.

A mass balance on the air phase gives:

$$\frac{\partial C_v(R,Z,T)}{\partial T} = \frac{Q_G}{D\epsilon(1-S)2\pi R}\frac{\partial C_v(R,Z,T)}{\partial R} + \frac{D_e}{R}\frac{\partial}{\partial R}\left[R\frac{\partial C_v(R,Z,T)}{\partial R}\right] + D_e\frac{\partial^2 C_v(R,Z,T)}{\partial Z^2} + K_L a\left[C_b(R,Z,T) - \frac{C_v(R,Z,T)}{H}\right] \quad (4.2.1.1)$$

Equation (4.2.1.1) relates accumulation of chemical at given depth for a

given distance from the extraction vent to the advective rate in air, the radial gas dispersive rate, the vertical gas diffusive rate, and the air-water mass transfer rate, respectively. Horizontal gas dispersion is described using air diffusion because it is assumed that air velocity is slow enough that molecular diffusion predominates.

The initial vapor concentrations in the soil system shown in Figure 4.2.1.1 are specified from measurements or by assumption:

$$C_v(R,Z,T=0) = C_{vi}(R,Z) \qquad (4.2.1.2)$$

The boundary condition for (4.2.1.1) at the extraction vent ($R=0, 0 \leq Z \leq D$) is derived by assuming that the system is a closed reactor [Levenspiel, 1962]:

$$\frac{\partial^2 C_v(R=0, 0 \leq Z \leq D, T)}{\partial R \partial T} = 0 \qquad (4.2.1.3)$$

The boundary conditions for (4.2.1.1) at the ground surface ($0 \leq R \leq R_i, Z=0$) and at the water table ($0 \leq R \leq R_i, Z=D$) are obtained by assuming that vapors could not diffuse out the system at either boundary:

$$\frac{\partial C_v(0 \leq R \leq R_i, Z=0, T)}{\partial Z} = 0 \qquad (4.2.1.4)$$

$$\frac{\partial^2 C_v(0 \leq R \leq R_i, Z=D, T)}{\partial Z \partial T} = 0 \qquad (4.2.1.5)$$

The ground surface boundary condition ($Z=0$, equation (4.2.1.4)) is valid for a system that is covered by an impermeable cap or thin, saturated

soil cover. Jury *et al.* [1990] calculate minimum soil cover thicknesses for inhibiting diffusion of various volatile organic chemicals to the atmosphere. For example, less than 1% of toluene contamination beneath a 15 cm clay layer at a degree of saturation of 0.75 would volatilize to the atmosphere in a year. The ground water table boundary condition (Z = D, equation (4.2.1.5)) is based on the assumption that dissolution of organic vapors and subsequent transport with ground water flow is negligible. This is probably a valid assumption when the velocity of the ground water (assuming a horizontal ground water table) is two or more orders of magnitude slower than the air velocity.

The boundary condition for (4.2.1.1) at the air inlet ($R=R_i, 0 \leq Z \leq D$) is obtained by performing a mass balance on the air phase in the disk-shaped volume shown in Figure 4.2.1.1:

$$\frac{\partial}{\partial T}\int_0^{Ri} C_v(R,Z,T)R\partial R = \int_0^{Ri}\left(K_L a\left[C_b(R,Z,T) - \frac{C_v(R,Z,T)}{H}\right] + D_e\frac{\partial^2 C_v(R,Z,T)}{\partial^2 Z}\right) R\partial R + \frac{Q_G}{D\epsilon(1-S)2\pi}\left[C_{vo}(T) - C_v(R=0,Z,T)\right] \qquad (4.2.1.6)$$

A mass balance on the water-soil phase is given by:

$$\left[1 + \frac{\rho_s(1-\epsilon)K}{\epsilon S}\right]\frac{\partial C_b(R,Z,T)}{\partial T} = E_z\frac{\partial^2 C_b(R,Z,T)}{\partial Z^2} - v\frac{\partial C_b(R,Z,T)}{\partial Z} - \frac{K_L a(1-S)}{S}\left[C_b(R,Z,T) - \frac{C_v(R,Z,T)}{H}\right] \qquad (4.2.1.7)$$

The accumulation term in (4.2.1.7) is set equal to the sum of the terms representing liquid dispersion, water advection, and air-water mass transfer, respectively.

For this development it is assumed that the aqueous- and sorbed-phase concentrations are in equilibrium. Furthermore, this equilibrium is assumed to be linear:

$$Q(R,Z,T) = K\ C_b(R,Z,T) \qquad (4.2.1.8)$$

Because in most soils, moisture will probably cover the soil particles [Roy and Griffin, 1987], sorption of vapors by soil is not considered here. Descriptions of vapor sorption such as that given by Shoemaker *et al.* [1990] could be performed with this model, if local equilibrium exists, by modifying the value of K accordingly.

Initially the soil system shown in Figure 4.2.1.1 is at chemical equilibrium:

$$C_b(R,Z,T=0) = C_{bi}(R,Z) = C_{vi}(R,Z) \ / \ H \qquad (4.2.1.9)$$

Boundary conditions for (4.2.1.7) are obtained by assuming that the soil system is a closed reactor [Levenspiel, 1962]. The Danckwerts [1953] exit condition is:

$$\frac{\partial^2 C_b(0 \le R \le R_i, Z=D, T)}{\partial Z \partial T} = 0 \qquad (4.2.1.10)$$

The description of aqueous-phase transport to ground water assumes that water velocities are slow enough that changes in the depth to ground water are negligible. In addition, the mass flux of chemical away from the water table is equal to the mass flux to the aquifer.

An entrance boundary condition for (4.2.1.7) is obtained by performing a mass balance on the water-soil phases in the cylindrical volume shown in Figure 4.2.1.1:

$$\left[1 + \frac{\rho_s(1-\epsilon)K}{\epsilon S} \right] \frac{\partial}{\partial T} \int_0^D C_b(R,Z,T)\partial Z = v\,[C_{bo}(T) - C_b(R,Z=D,T)]$$
$$- \int_0^D \frac{K_L a(1-S)}{S} \left[C_b(R,Z,T) - \frac{C_v(R,Z,T)}{H} \right] \partial Z \qquad (4.2.1.11)$$

4.2.1.3 Conversion to Dimensionless Form. The dimensioned equations derived above were converted to a dimensionless form by making substitutions of the middle column in Table 4.2.1.1.

Table 4.2.1.1. Variable Substitutions to Convert Dimensioned Equations into a Dimensionless Form.

Dimensioned Variable	Substitution	Dimensionless Variable
$C_b(R,Z,T)$	$C_{bn}\ c_b(R,Z,T)$	$c_b(R,Z,T)$
$C_v(R,Z,T)$	$C_{vn}\ c_v(R,Z,T)$	$c_v(R,Z,T)$
$C_{bi}(R,Z)$	$C_{bn}\ c_{bi}(R,Z)$	$c_{bi}(R,Z)$
$C_{vi}(R,Z)$	$C_{vn}\ c_{vi}(R,Z)$	$c_{vi}(R,Z)$
$C_{bo}(T)$	$C_{bn}\ c_{bo}(T)$	$c_{bo}(T)$
$C_{vo}(T)$	$C_{vn}\ c_{vo}(T)$	$c_{vo}(T)$
R	$R_i\ r$	r
Z	$D\ z$	z
T	$\frac{Q_G H + \epsilon S \pi R_i^2 v}{\epsilon(1-S)D\pi R_i^2 H} \left[1 + \frac{S}{(1-S)H} + \frac{\rho_s(1-\epsilon)K}{\epsilon(1-S)H} \right]^{-1} t$	t

Note: $C_{bn} = C_{vn} / H$ where $C_{vn} = \max(C_{vo}(T), C_{vi}(R,Z))$.

These substitutions resulted in natural groupings of the dimensioned parameters. The resulting groups are dimensionless and are similar to those derived for the one-dimensional system models. Mass

transfer rates are compared to the advective rates in air and chemical equilibrium distributions are made relative to the mass of chemical in air. The dimensionless groups that characterize the solutions of radial models are given in Table 4.2.1.2.

Table 4.2.1.2. Definitions of Dimensionless Groups.

Dimensionless Group		Equation	Ratio of	Ratio to
Ar	=	$\frac{v\epsilon S\pi R_i^2}{Q_G H}$	Advective rate in water	Advective rate in air
Dg	=	$Dg_S + Dg_V$	Mass in water and on soil	Mass in air
Dg_S	=	$\frac{\rho_S(1-\epsilon)K}{\epsilon(1-S)H}$	Mass adsorbed to soil	Mass in air
Dg_V	=	$\frac{S}{(1-S)H}$	Mass in water	Mass in air
Ed_V	=	$\frac{D_e\epsilon(1-S)\pi R_i^2}{Q_G D}$	Vapor diffusive rate in vertical direction	Advective rate in air
Pe_b	=	$\frac{Q_G HD}{E_Z\epsilon S\pi R_i^2}$	Advective rate in air	Dispersive rate in water
Pe_V	=	$\frac{Q_G}{D_e\epsilon(1-S)2\pi R_i}$	Advective rate in air	Horizontal dispersive rate in air
St_V	=	$\frac{K_L a\epsilon(1-S)\pi R_i^2}{Q_G H}$	Volatilization rate	Advective rate in air

Equations (4.2.1.1-11) become the following after substituting the variables defined in Tables 4.2.1.1 and 4.2.1.2:

$$\frac{\partial c_v(r,z,t)}{\partial t} = \frac{1 + Dg}{2[1 + Ar]} \left[\frac{1}{r} \frac{\partial c_v(r,z,t)}{\partial r} + \frac{1}{Pe_v r} \frac{\partial}{\partial r} \left[r \frac{\partial c_v(r,z,t)}{\partial r} \right] \right.$$

$$\left. + 2Ed_v \frac{\partial^2 c_v(r,z,t)}{\partial z^2} + 2St_v \left[c_b(r,z,t) - c_v(r,z,t) \right] \right] \qquad (4.2.1.12)$$

$$c_v(r,z,t=0) = c_{vi}(r,z) \qquad (4.2.1.13)$$

$$\frac{\partial^2 c_v(r=0,0\leq z\leq 1,t)}{\partial r \partial t} = 0 \qquad (4.2.1.14)$$

$$\frac{\partial c_v(0\leq r\leq 1,z=0,t)}{\partial z} = 0 \qquad (4.2.1.15)$$

$$\frac{\partial^2 c_v(0\leq r\leq 1,z=1,t)}{\partial z \partial t} = 0 \qquad (4.2.1.16)$$

$$\frac{\partial}{\partial t} \int_0^1 c_v(r,z,t) r \partial r = \frac{1 + Dg}{2[1 + Ar]} \left[\int_0^1 \left[2St_v \left[c_b(r,z,t) - c_v(r,z,t) \right] \right. \right.$$

$$\left. \left. + 2Ed_v \frac{\partial^2 c_v(r,z,t)}{\partial z^2} \right] r \partial r + \left[c_{vo}(t) - c_v(r=0,z,t) \right] \right] \qquad (4.2.1.17)$$

$$\frac{\partial c_b(r,z,t)}{\partial t} = \frac{1 + Dg}{Dg[1 + Ar]} \left[\frac{1}{Pe_b} \frac{\partial^2 c_b(r,z,t)}{\partial z^2} - Ar \frac{\partial c_b(r,z,t)}{\partial z} \right.$$

$$\left. - St_v[c_b(r,z,t) - c_v(r,z,t)] \right] \qquad (4.2.1.18)$$

$$c_b(r,z,t=0) = c_{bi}(r,z) = c_{vi}(r,z) \qquad (4.2.1.19)$$

$$\frac{\partial^2 c_b(0\leq r\leq 1,z=1,t)}{\partial z \partial t} = 0 \qquad (4.2.1.20)$$

$$\frac{\partial}{\partial t}\int_0^1 c_b(r,z,t)\partial z = \frac{1 + Dg}{Dg[1 + Ar]}\left[Ar\,[c_{bo}(t) - c_b(r,z=1,t)] - \int_0^1 St_v[c_b(r,z,t) - c_v(r,z,t)]\partial z \right] \qquad (4.2.1.21)$$

4.2.1.4 Numerical Solution. As before, orthogonal collocation (OC) is the numerical method used to convert the partial differential equations (PDEs) given above to a system of ordinary differential equations (ODEs) in time. Details of the method are given elsewhere [Finlayson, 1980]. Steps for converting PDEs to ODEs are documented in Appendix B of Gierke [1986].

Orthogonal collocation constants that are used for differential approximations in the column models (see Section 4.1.1.4) are used in the approximation of the radial derivatives by making a variable substitution of the independent variable r. Equations (4.2.1.12-17) become the following after substituting $x = r^2$:

$$\frac{\partial c_v(x,z,t)}{\partial t} = \frac{1 + Dg}{1 + Ar}\left[\frac{1}{Pe_v}\frac{\partial^2 c_v(x,z,t)}{\partial x^2} + \left[1 + \frac{1}{Pe_v} \right]\frac{\partial c_v(x,z,t)}{\partial x} + Ed_v \frac{\partial^2 c_v(x,z,t)}{\partial z^2} + St_v[c_b(x,z,t) - c_v(x,z,t)] \right] \qquad (4.2.1.22)$$

$$c_v(x,z,t=0) = c_{vi}(x,z) \qquad (4.2.1.23)$$

$$\frac{\partial^2 c_v(x=0,0\le z\le 1,t)}{\partial x \partial t} = 0 \qquad (4.2.1.24)$$

$$\frac{\partial c_v(0\le x\le 1,z=0,t)}{\partial z} = 0 \qquad (4.2.1.25)$$

$$\frac{\partial^2 c_v(0\le x\le 1, z=1, t)}{\partial z \partial t} = 0 \qquad (4.2.1.26)$$

$$\frac{\partial}{\partial t}\int_0^1 c_v(x,z,t)\partial x = \frac{1+Dg}{1+Ar}\left[\int_0^1 \left[St_v[c_b(x,z,t) - c_v(x,z,t)] + Ed_v \frac{\partial^2 c_v(x,z,t)}{\partial z^2}\right]\partial x + [c_{vo}(t) - c_v(x=0,z,t)]\right] \qquad (4.2.1.27)$$

Because the water-soil-phase equations (equations (4.2.1.18-21)) do not contain derivatives with respect to r, then the variable r in the arguments of the dependent variables can just simply be replaced with x:

$$\frac{\partial c_b(x,z,t)}{\partial t} = \frac{1+Dg}{Dg(1+Ar)}\left[\frac{1}{Pe_b}\frac{\partial^2 c_b(x,z,t)}{\partial z^2} - Ar\frac{\partial c_b(x,z,t)}{\partial z} - St_v[c_b(x,z,t) - c_v(x,z,t)]\right] \qquad (4.2.1.28)$$

$$c_b(x,z,t=0) = c_{bi}(x,z) = c_{vi}(x,z) \qquad (4.2.1.29)$$

$$\frac{\partial^2 c_b(0\le x\le 1, z=1, t)}{\partial z \partial t} = 0 \qquad (4.2.1.30)$$

$$\frac{\partial}{\partial t}\int_0^1 c_b(x,z,t)\partial z = \frac{1+Dg}{Dg[1+Ar]}\left[Ar\,[c_{bo}(t) - c_b(x,z=1,t)] - \int_0^1 St_v[c_b(x,z,t) - c_v(x,z,t)]\partial z\right] \qquad (4.2.1.31)$$

The initial condition for the transformed equations are given by (4.2.1.23) and (4.2.1.29) and the OC form of the initial condition is

$$c_v(i,k,t=0) = c_{vi}(x_i, z_k) \text{ for } i = 1 \text{ to NR and } k = 1 \text{ to NDV} \qquad (4.2.1.32)$$

$$c_b(i,j,t=0) = c_{bi}(x_i,z_j) \text{ for } i = 1 \text{ to NR and } j = 1 \text{ to NDB} \qquad (4.2.1.33)$$

The parameter NR is the number of collocation locations for the radial approximations, NDB is the number of vertical collocation points for approximating water transport, and NDV is the number of vertical collocation points for approximating gas diffusion in the z direction.

The application of OC to the transformed air-phase mass balance (equation (4.2.1.22)) results in:

$$\frac{dc_v(i,k,t)}{dt} = \frac{1 + Dg}{1 + Ar}\left\{ \sum_{n=1}^{NR} \left[\frac{B^x{}_{i,n}}{Pe_v} + A^x{}_{i,n}\left[1 + \frac{1}{Pe_v} \right] \right] c_v(n,k,t) + \sum_{m=1}^{NDV} Ed_v B^{zv}{}_{k,m}\, c_v(i,m,t) + St_v[c_b(i,k,t) - c_v(i,k,t)] \right\}$$

$$\text{for } i = 2 \text{ to NR-1 and } k = 1 \text{ to NDV-1} \qquad (4.2.1.34)$$

The matrices A^x and B^x are the same asymmetric, planar coefficient matrices used for approximating the spatial derivatives in the axial mass transport equations of the column models and are used here to approximate the first and second partial derivatives in the x coordinate, respectively, in (4.2.1.22) and (4.2.1.24) (see Section 4.1.1.4). The matrix B^{zv} is a symmetric, planar OC coefficient matrix for approximating the second partial of $c_v(x,z,t)$ with respect to z. B^{zv} is analogous to the spherical matrix used in the column models. It is a function only of even powers of z and is generated with respect to a weighting function of 1 over the interval of z from 0 to 1. The locations of the vertical collocation points used in the approximation of $\partial^2 c_v/\partial z^2$ are obtained from the positive roots of the 2(NDV-1) degree polynomial and lie between 0 and 1. Because B^{zv} is symmetrical, its use

satisfies (4.2.1.25).

The air-phase concentrations at the extraction vent come from the OC transformation of (4.2.1.24):

$$\frac{dc_v(1,k,t)}{dt} = -\sum_{n=2}^{NR} \frac{A^x_{1,n}}{A^x_{1,1}} \frac{dc_v(n,k,t)}{dt} \quad \text{for } k = 1 \text{ to NDV} \qquad (4.2.1.35)$$

The boundary condition at the ground water table (equation (4.2.1.26)) is given by:

$$\frac{dc_v(i,NDV,t)}{dt} = -\sum_{m=1}^{NDV-1} \frac{A^{zv}_{NDV,m}}{A^{zv}_{NDV,NDV}} \frac{dc_v(i,m,t)}{dt} \quad \text{for } i = 2 \text{ to NR-1} \qquad (4.2.1.36)$$

Changes in the air-phase concentrations at the inlet vents are described by:

$$\frac{dc_v(NR,k,t)}{dt} = \left\{ \frac{1 + Dg}{1 + Ar} \left(\sum_{n=1}^{NR} w^x_n \left[St_v[c_b(n,k,t) - c_v(n,k,t)] - \sum_{m=1}^{NDV} Ed_v B^{zv}_{k,m} c_v(n,m,t) \right] + [c_{vo}(t) - c_v(1,k,t)] \right) - \sum_{n=2}^{NR-1} \left[w^x_n - \frac{A^x_{1,n}}{A^x_{1,1}} w^x_1 \right] \frac{dc_v(n,k,t)}{dt} \right\} \left[w^x_{NR} - \frac{A^x_{1,NR}}{A^x_{1,1}} w^x_1 \right]^{-1}$$

$$\text{for } k = 1 \text{ to NDV} \qquad (4.2.1.37)$$

Equation (4.2.1.37) is obtained by applying OC to (4.2.1.27), substituting (4.2.1.35) into the result, and solving for the time-derivative of $c_v(NR,j,t)$. The construction of the quadrature vector for approximating integrals with respect to x (w^x) is described in Section 4.1.1.4. The influence of the boundary conditions are redundant at the

intersections of the vents and the ground water table (i,k = 1,NDV and NR,NDV). Combinations of the different boundary conditions were tried and the combination given above gave the most stable results.

The water-soil mass balance in transformed coordinates (equation (4.2.1.28)) is the following after approximating with OC:

$$\frac{dc_b(i,j,t)}{dt} = \frac{1 + Dg}{Dg[1 + Ar]} \left\{ \sum_{m=1}^{NDB} \left[\frac{B^{zb}{}_{j,m}}{Pe_b} - Ar\, A^{zb}{}_{j,m} \right] c_b(i,m,t) - St_v[c_b(i,j,t) - c_v(i,j,t)] \right\}$$

$$\text{for } i = 1 \text{ to } NR \text{ and } j = 2 \text{ to } NDB-1 \qquad (4.2.1.38)$$

The matrices A^{zb} and B^{zb} are constructed in the same manner as A^x and B^x. They are given different notation because they will not be the same if NR ≠ NDB.

The boundary condition at the ground water table (equation (4.2.1.30)) is given by:

$$\frac{dc_b(i,NDB,t)}{dt} = - \sum_{m=1}^{NDB-1} \frac{A^{zb}{}_{NDB,m}}{A^{zb}{}_{NDB,NDB}} \frac{dc_b(i,m,t)}{dt}$$

$$\text{for } i = 1 \text{ to } NR \qquad (4.2.1.39)$$

Application of OC to (4.2.1.31) becomes the following after substitution of (4.2.1.39), and rearranging to obtain the time-derivative of $c_b(i,1,t)$:

$$\frac{dc_b(i,1,t)}{dt} = \left\{ \frac{1 + Dg}{Dg[1 + Ar]} \left[Ar[c_{bo}(t) - c_b(i,NDB,t)] - \sum_{m=1}^{NDB} W^{zb}{}_m St_v[c_b(i,m,t) - c_v(i,m,t)] \right] - \sum_{m=2}^{NDB-1} \left[W^{zb}{}_m - \frac{A^{zb}{}_{NDB,m}}{A^{zb}{}_{NDB,NDB}} W^{zb}{}_{NDB} \right] \frac{dc_b(i,m,t)}{dt} \right\} \left[W^{zb}{}_1 - \frac{A^{zb}{}_{NDB,1}}{A^{zb}{}_{NDB,NDB}} W^{zb}{}_{NDB} \right]^{-1}$$

for i = 1 to NR (4.2.1.40)

The coupling of the NR(NDB+NDV) ODEs given above are shown graphically in Figure 4.2.1.2. This schematic shows that the derivatives of the air-phase concentrations at a particular horizontal (x_i) and vertical (z_k) position are functions of the air-phase concentrations along an x-z axis centered at (i,k), and that the water concentration derivatives at a particular vertical collocation point (z_j) for a given axial location (x_i) are functions of the water concentrations only at the other vertical points at the same radial position. To properly determine the mass of chemical transfered between the air and water, air-phase concentrations need to be determined for the z_j positions from the values given at the z_k locations and water-phases concentrations at the z_k positions from the z_j locations. The locations of the vertical collocation locations for the air (z_k) and the water (z_j) are different because (4.2.1.22) is symmetrical with respect to z and (4.2.1.28) is not. Symmetric and asymmetric equations employ different weighting functions in the generation of the spatial polynomials for the OC approximations so the roots (*i.e.* the collocation locations) are different. Therefore the air and water concentrations must be determined for the collocation locations for the other phase as shown in Figure 4.2.1.2. Concentrations at locations between the

collocation points are obtained by linear interpolation and beyond the collocation points for the air phase by linear extrapolation. If linear extrapolation determines values of the dimensionless air-phase concentrations greater than 1 or less than 0, the value is set equal to 1 or 0, respectively.

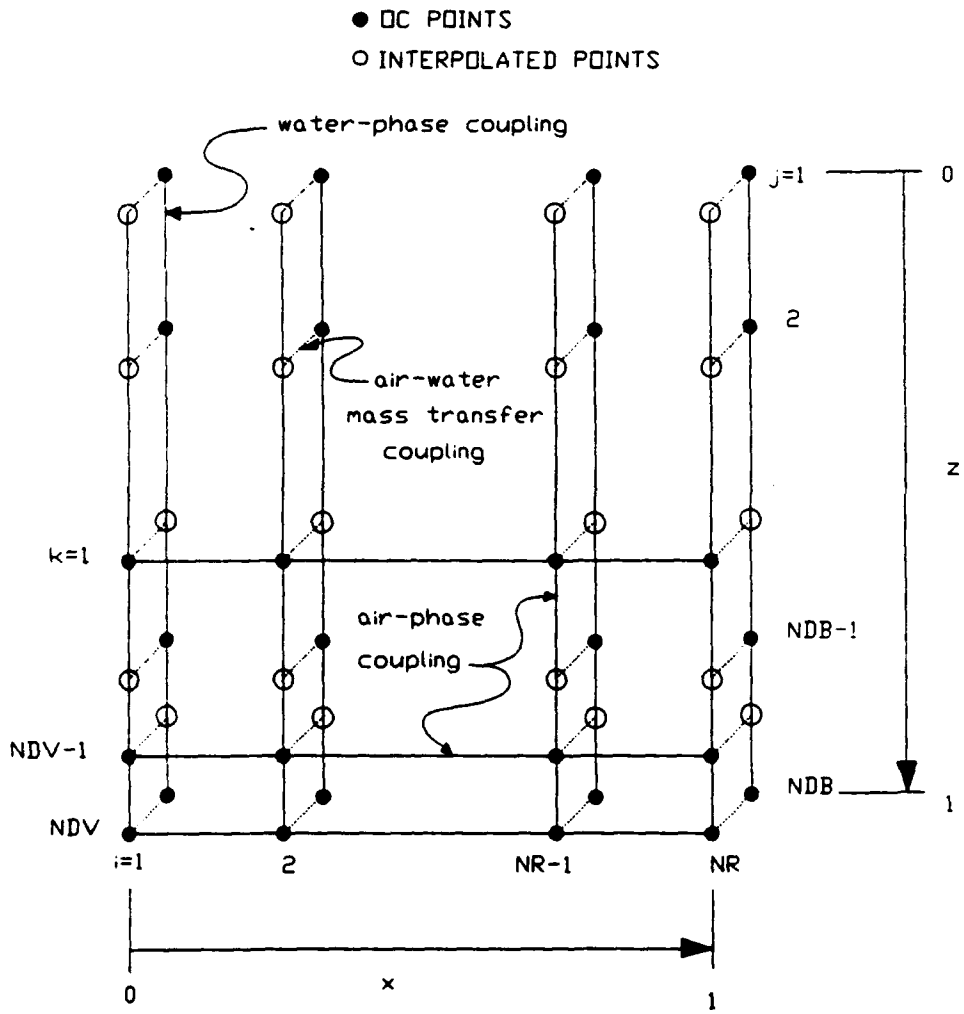


Figure 4.2.1.2. General schematic of the coupling of the ordinary differential equations resulting the application of orthogonal collocation to the transformed partial differential equations comprising the general form of the vapor extraction system model for radial configuration.

Early versions of the model used a nonlinear interpolating matrix

derived from the definition of OC to determine concentrations at locations different than the OC points. Bednar [1990] found that this type of interpolation exhibited unacceptable numerical error for steep concentration profiles, but that a linear interpolation routine did not. For chemical fronts that are spread out, the OC and linear interpolation methods give similar results [Bednar, 1990].

The system of ODEs given above (equations (4.2.1.32-40)) comprise the most general form of the radial-configuration vapor extraction models. The derivatives are evaluated in the following order: (4.2.1.38), (4.2.1.40), (4.2.1.39), (4.2.1.34), (4.2.1.38), (4.2.1.37), and (4.2.1.35). Initially, (4.2.1.32) and (4.2.1.33) are used in the determination of the derivatives. An International Mathematics and Statistics Library package for solving sets of ODEs, called GEAR, is used to determine concentration values at times greater than 0.

4.2.1.5 Model Simplifications and Corresponding Solutions. Variations of the model given above are derived by making assumptions about the relative importance of the various transport mechanisms. For each of the following simplified cases, the initial conditions (equations (4.2.1.23 and 29)) do not change. The boundary conditions are different for the various simplifications.

Radial Model without Vertical Diffusion. When vertical vapor diffusion is ignored, (4.2.1.12) reduces to:

$$\frac{\partial c_v(r,z,t)}{\partial t} = \frac{1 + Dg}{2[1 + Ar]} \left[\frac{1}{r} \frac{\partial c_v(r,z,t)}{\partial r} + \frac{1}{Pe_v r} \frac{\partial}{\partial r} \left[r \frac{\partial c_v(r,z,t)}{\partial r} \right] + 2St_v \left[c_b(r,z,t) - c_v(r,z,t) \right] \right] \qquad (4.2.1.41)$$

In addition, (4.2.1.15 and 16) are not needed, and (4.2.1.17) reduces to:

$$\frac{\partial}{\partial t}\int_0^1 c_v(r,z,t)r\partial r = \frac{1+Dg}{2[1+Ar]}\left[\int_0^1 2St_v[c_b(r,z,t) - c_v(r,z,t)]\, r\partial r + [c_{vo}(t) - c_v(r=0,z,t)]\right] \qquad (4.2.1.42)$$

The other boundary condition is given by (4.2.1.14). The water-phase transport equations (equations (4.2.1.18-21)) do not change.

In the transformed coordinate $x = r^2$, (4.2.1.41) and (4.2.1.42) become:

$$\frac{\partial c_v(x,z,t)}{\partial t} = \frac{1+Dg}{1+Ar}\left[\frac{1}{Pe_v}\frac{\partial^2 c_v(x,z,t)}{\partial x^2} + \left[1 + \frac{1}{Pe_v}\right]\frac{\partial c_v(x,z,t)}{\partial x} + St_v[c_b(x,z,t) - c_v(x,z,t)]\right] \qquad (4.2.1.43)$$

$$\frac{\partial}{\partial t}\int_0^1 c_v(x,z,t)\partial x = \frac{1+Dg}{1+Ar}\left[\int_0^1 St_v[c_b(x,z,t) - c_v(x,z,t)]\, \partial x + [c_{vo}(t) - c_v(x=0,z,t)]\right] \qquad (4.2.1.44)$$

The transformed (4.2.1.14) is given by (4.2.1.24) and water-phase transport equations by (4.2.1.28-31).

Because there are no vertical derivatives in the air-phase mass balance for no vertical diffusion, the air-phase transport equations can be evaluated at the collocation locations for the water-phase transport equations and concentration interpolation is not necessary (NDB = NDV = ND). The application of OC to (4.2.1.43 and 44) results in:

$$\frac{dc_v(i,k,t)}{dt} = \frac{1 + Dg}{1 + Ar}\left\{ \sum_{n=1}^{NR} \left[\frac{B^x_{i,n}}{Pe_v} + A^x_{i,n}\left[1 + \frac{1}{Pe_v} \right] \right] c_v(n,k,t) + St_v[c_b(i,k,t) - c_v(i,k,t)] \right\}$$

for i = 2 to NR-1 and k = 1 to ND (4.2.1.45)

$$\frac{dc_v(NR,k,t)}{dt} = \left\{ \frac{1 + Dg}{1 + Ar}\left[\sum_{n=1}^{NR} W^x_n St_v[c_b(n,k,t) - c_v(n,k,t)] + [c_{vo}(t) - c_v(1,k,t)] \right] - \sum_{n=2}^{NR-1}\left[W^x_n - \frac{A^x_{1,n}}{A^x_{1,1}} W^x_1 \right] \frac{dc_v(n,k,t)}{dt} \right\} \cdot \left[W^x_{NR} - \frac{A^x_{1,NR}}{A^x_{1,1}} W^x_1 \right]^{-1}$$

for k = 1 to ND (4.2.1.46)

The remaining OC equations are (4.2.1.35, 38-40) where NDB = NDV = ND, $A^{zb} = A^z$, and $B^{zb} = B^z$. The initial condition is

$$c_v(i,j,t=0) = c_b(i,j,t=0) = c_{vi}(i,j)$$

for i = 1 to NR and j = 1 to ND (4.2.1.47)

The total number of equations to solve is 2(NRxND). Equation (4.2.1.47) is used initially to calculate the derivatives of air and water concentrations in the order: (4.2.1.38), (4.2.1.40), (4.2.1.39), (4.2.1.45), (4.2.1.46), and (4.2.1.35). The values of the derivatives are input to GEAR and new values of the concentrations are returned. This cycle is repeated until the desired output time is reached.

Radial Plug Flow, Air-Water Mass Transfer Model. By ignoring dispersion and diffusion in air and water, the transport equations for air and water become, respectively:

$$\frac{\partial c_v(r,z,t)}{\partial t} = \frac{1 + Dg}{2[1 + Ar]} \left[\frac{1}{r} \frac{\partial c_v(r,z,t)}{\partial r} + 2St_v[c_b(r,z,t) - c_v(r,z,t)] \right] \quad (4.2.1.48)$$

$$\frac{\partial c_b(r,z,t)}{\partial t} = - \frac{1 + Dg}{Dg[1 + Ar]} \left[Ar \frac{\partial c_b(r,z,t)}{\partial z} + St_v[c_b(r,z,t) - c_v(r,z,t)] \right] \quad (4.2.1.49)$$

For plug air and water flow, the boundary conditions are replaced with a single boundary conditions at r equal to 1 for (4.2.1.48) and z equal to 0 for (4.2.1.49):

$$c_v(r=1, 0 \leq z \leq 1, t) = c_{vo}(t) \quad (4.2.1.50)$$

$$c_b(0 \leq r \leq 1, z=0, t) = c_{bo}(t) \quad (4.2.1.51)$$

Again, before applying the numerical method, the r variable in (4.2.1.48-51) is transformed to $x = r^2$:

$$\frac{\partial c_v(x,z,t)}{\partial t} = \frac{1 + Dg}{1 + Ar} \left[\frac{\partial c_v(x,z,t)}{\partial x} + St_v[c_b(x,z,t) - c_v(x,z,t)] \right] \quad (4.2.1.52)$$

$$\frac{\partial c_b(x,z,t)}{\partial t} = - \frac{1 + Dg}{Dg[1 + Ar]} \left[Ar \frac{\partial c_b(x,z,t)}{\partial z} + St_v[c_b(x,z,t) - c_v(x,z,t)] \right] \quad (4.2.1.53)$$

$$c_v(x=1,0\le z\le 1,t) = c_{vo}(t) \qquad (4.2.1.54)$$

$$c_b(0\le x\le 1,z=0,t) = c_{bo}(t) \qquad (4.2.1.55)$$

The application of OC to (4.2.1.52-55) gives:

$$\frac{dc_v(i,k,t)}{dt} = \frac{1+Dg}{1+Ar}\left[\sum_{n=1}^{NR} A^x_{i,n}\, c_v(n,k,t) + St_v[c_b(i,k,t) - c_v(i,k,t)]\right] \text{ for } i = 1 \text{ to } NR-1 \text{ and } k = 1 \text{ to } ND \qquad (4.2.1.56)$$

$$\frac{dc_b(i,j,t)}{dt} = -\frac{1+Dg}{Dg[1+Ar]}\left[\sum_{m=1}^{ND} A^z_{j,m}\, Ar\, c_b(i,m,t) + St_v[c_b(i,j,t) - c_v(i,j,t)]\right]$$

$$\text{for } i = 1 \text{ to } NR \text{ and } j = 2 \text{ to } ND \qquad (4.2.1.57)$$

$$c_v(NR,j,t) = c_{vo}(t) \text{ for } j = 1 \text{ to } ND \qquad (4.2.1.58)$$

$$c_b(i,1,t) = c_{bo}(t) \text{ for } i = 1 \text{ to } NR \qquad (4.2.1.59)$$

The total number of equations to solve is NR(ND-1)+ND(NR-1). For plug flow approximations, the matrices A^x and A^z are constructed orthogonal with respect to a weighting function of 1. Equation (4.2.1.58) defines the air-phase concentration at the inlet vent (typically it is zero) and (4.2.1.59) defines the water-phase concentration at the ground surface. Equation (4.2.1.47) is used initially to calculate the time-derivatives of air and water concentrations in the order: (4.2.1.56) and (4.2.1.55). The values of the derivatives are input to GEAR and new values of the concentrations are returned. This cycle is repeated until the desired output time is

reached.

An attempt was made to develop an analytic solution for steady state conditions to test the numerical approximation of air-water mass transfer. At steady state the time derivatives become zero so that (4.2.1.48) and (4.2.1.49) reduce to:

$$\frac{1}{r}\frac{\partial c_v(r,z)}{\partial r} = -2St_v[c_b(r,z) - c_v(r,z)] \qquad (4.2.1.60)$$

$$\frac{\partial c_b(r,z)}{\partial z} = -\frac{St_v}{Ar}[c_b(r,z) - c_v(r,z)] \qquad (4.2.1.61)$$

The boundary conditions for (4.2.1.60) and (4.2.1.61) are (4.2.1.50) and (4.2.1.51), respectively.

Again, the radial coordinate is transformed by $x = r^2$ to get:

$$\frac{\partial c_v(x,z)}{\partial x} = -St_v[c_b(x,z) - c_v(x,z)] \qquad (4.2.1.62)$$

$$\frac{\partial c_b(x,z)}{\partial z} = -\frac{St_v}{Ar}[c_b(x,z) - c_v(x,z)] \qquad (4.2.1.63)$$

and (4.2.1.54) and (4.2.1.55).

This system is singular and can not be solved by the conventional method of characteristics. An Euler integration routine is used as another method for solving the steady-state problem. Equations (4.2.1.62), (4.2.1.63), (4.2.1.54), and (4.2.1.55) can be approximated with the following finite difference equations for a uniform mesh size:

$$c_v(x_i,z_j) = c_v(x_{i+1},z_j)[1 - \Delta x2St_v] + \Delta x2St_v c_b(x_{i+1},z_j)$$
$$\text{for } i \text{ from } I\text{-}1 \text{ to } 1 \text{ and } j \text{ from } 1 \text{ to } J \qquad (4.2.1.64)$$

$$c_b(x_i,z_{j+1}) = c_b(x_i,z_j)[1 - \Delta z2St_v/Ar] + \Delta z2St_v c_v(x_i,z_j)/\,Ar$$
$$\text{for } i \text{ from } I \text{ to } 1 \text{ and } j \text{ from } 2 \text{ to } J \qquad (4.2.1.65)$$

$$c_v(x_I,z_j) = c_{vo} \quad \text{for } j \text{ from } 1 \text{ to } J \qquad (4.2.1.66)$$

$$c_b(x_i,z_1) = c_{bo} \quad \text{for } i \text{ from } I \text{ to } 1 \qquad (4.2.1.67)$$

Here, the parameters I and J are the total number of horizontal and vertical mesh points, respectively. This approximation is performed by first defining the boundary concentrations at $x_I = 1$ and $z_1 = 0$ with (4.2.1.66) and (4.2.1.67). Equation (4.2.1.64) is used to determine $c_v(x_i,z_1)$ starting at i=I-1 and continuing to i=1, then (4.2.1.65) is used to evaluate $c_b(x_i,z_2)$ from i = I to 1, and evaluations of (4.2.1.64) are repeated for z_2. This cycle continues to z_J. The result is the steady-state concentration profile for a given set of values of St_v, Ar, c_{vo}, and c_{bo}. If $c_{vo} = c_{bo}$, Ar = 0, or $St_v = 0$, the solution is trivial. A common situation is where there is cleanup of continuous infiltration of contaminant such as beneath a leaking underground storage tank, where $c_{vo} = 0$ and $c_{bo} = 1$.

This finite difference approach could be applied to the unsteady-state equations (equations (4.2.1.52) and (4.2.1.53)), however, it is beyond the scope of this work to study numerical solving techniques in detail, and the comparisons to the steady-state problem should be sufficient for numerical verification.

Dispersed Flow, Local Equilibrium Model. When local equilibrium (see Section 4.1.1.5) is assumed, (4.2.1.12) and (4.2.1.18) combine to

give:

$$\frac{\partial c_v(r,z,t)}{\partial t} = \frac{1}{2[1 + Ar]} \left[\frac{1}{r} \frac{\partial c_v(r,z,t)}{\partial r} + \frac{1}{Pe_v r} \frac{\partial}{\partial r} \left[r \frac{\partial c_v(r,z,t)}{\partial r} \right] \right.$$
$$\left. - Ar \frac{\partial c_v(r,z,t)}{\partial z} + \left[2Ed_v + \frac{1}{Pe_b} \right] \frac{\partial^2 c_v(r,z,t)}{\partial z^2} \right] \qquad (4.2.1.68)$$

The initial condition is (4.2.1.13). The boundary condition at the extraction vent is (4.2.1.14) and at the ground water table it is (4.2.1.16). Because of limitations of the OC method, mass balance boundary conditions could not be employed. The boundary condition given by (4.2.1.50) is used at the air inlet and a similar condition is used at the ground surface:

$$c_v(0 \le r \le 1, z=0, t) = c_{bo}(t) \qquad (4.2.1.69)$$

In transformed coordinates, this model consists of (4.2.1.26), (4.2.1.27), (4.2.1.54), and

$$\frac{\partial c_v(x,z,t)}{\partial t} = \frac{1}{1 + Ar} \left[\left[1 + \frac{1}{Pe_v} \right] \frac{\partial c_v(x,z,t)}{\partial x} - Ar \frac{\partial c_v(x,z,t)}{\partial z} \right.$$
$$\left. + \frac{1}{Pe_v} \frac{\partial^2 c_v(x,z,t)}{\partial x^2} + \left[2Ed_v + \frac{1}{Pe_b} \right] \frac{\partial^2 c_v(x,z,t)}{\partial z^2} \right] \qquad (4.2.1.70)$$

$$c_v(x, z=0, t) = c_{bo}(t) \qquad (4.2.1.71)$$

The OC equations are (again, ND = NDB = NDV) (4.2.1.35), (4.2.1.36), (4.2.1.58), and

$$\frac{dc_v(i,k,t)}{dt} = \frac{1}{1 + Ar} \left\{ \sum_{n=1}^{NR} \left[\frac{B^x{}_{i,n}}{Pe_v} + A^x{}_{i,n} \left[1 + \frac{1}{Pe_v} \right] \right] c_v(n,k,t) \right.$$

$$\left. + \sum_{m=1}^{ND} \left[B^z{}_{k,m} \left[Ed_v + \frac{1}{Pe_b} \right] - A^z{}_{k,m}\, Ar \right] c_v(i,m,t) \right\}$$

for i = 2 to NR-1 and k = 2 to ND-1 (4.2.1.72)

$$c_v(i,1,t) = c_{bo}(t) \text{ for } i = 1 \text{ to NR} \quad (4.2.1.73)$$

The matrix A^z used in (4.2.1.72) is equivalent to A^{zb}, which is used for (4.2.1.38-40), and the coefficients in A^{zv} for (4.2.1.36) are replaced with A^z.

The total number of equations to solve is (NR-1)(ND-1). Equation (4.2.1.58) defines the concentrations at the inlet vent (typically they are zero) and (4.2.1.73) defines the concentrations at the ground surface. Equation (4.2.1.47) is used initially to calculate the time-derivatives in the order: (4.2.1.72), (4.2.1.36), and (4.2.1.58). The values of the derivatives are input to GEAR and new values of the concentrations are returned. This cycle is repeated until the desired output time is reached.

No Water Flow, Local Equilibrium Model. If the water velocity is zero, then the condition described by (4.2.1.68) reduces to a one-dimensional, axially-symmetric problem:

$$\frac{\partial c_v(r,t)}{\partial t} = \frac{1}{2} \left[\frac{1}{r} \frac{\partial c_v(r,t)}{\partial r} + \frac{1}{Pe_v r} \frac{\partial}{\partial r} \left[r \frac{\partial c_v(r,t)}{\partial r} \right] \right] \quad (4.2.1.74)$$

Equation (4.2.1.14) reduces to

$$\frac{\partial c_v(r=0,t)}{\partial r} = 0 \quad (4.2.1.75)$$

Equation (4.2.1.17) reduces to:

$$c_{vo}(t) - c_v(r=0,t) = 2\int_0^1 \frac{dc_v(r,t)}{dt}\, r dr \qquad (4.2.1.76)$$

The dimensionless initial condition is

$$c_v(r,t=0) = c_{vi}(r) \qquad (4.2.1.77)$$

As before, a variable substitution of $x = r^2$ is made in (4.2.1.74-77) prior to applying the numerical method:

$$\frac{\partial c_v(x,t)}{\partial t} = \left[1 + \frac{1}{Pe_v}\right] \frac{\partial c_v(x,t)}{\partial x} + \frac{1}{Pe_v} \frac{\partial^2 c_v(x,t)}{\partial x^2} \qquad (4.2.1.78)$$

$$\frac{\partial c_v(x=0,t)}{\partial x} = 0 \qquad (4.2.1.79)$$

$$c_{vo}(t) - c_v(x=0,t) = \int_0^1 \frac{\partial c_v(x,t)}{\partial t}\, \partial x \qquad (4.2.1.80)$$

$$c_v(x,t=0) = c_{vi}(x) \qquad (4.2.1.81)$$

Orthogonal collocation is now applied to these equations in the same manner as before. The resulting set of NR ODEs is

$$\frac{dc_v(j,t)}{dt} = \sum_{m=1}^{J} \left(\left[1 + \frac{1}{Pe_v}\right] A^x_{j,m} + \frac{B^x_{j,m}}{Pe_v} \right) c_v(m,t)$$

$$\text{for } j = 2 \text{ to } NR-1 \qquad (4.2.1.82)$$

$$\frac{dc_v(NR,t)}{dt} = \left[W^x_{NR} - W^x_1 \frac{A^x_{1,NR}}{A^x_{1,1}} \right]^{-1} \left\{ [c_{vo}(t) - c_v(1,t)] - \sum_{m=2}^{NR-1} \left[W^x_m - W^x_1 \frac{A^x_{1,m}}{A^x_{1,1}} \right] \frac{dc_v(m,t)}{dt} \right\} \quad (4.2.1.83)$$

$$\frac{dc_v(1,t)}{dt} = - \sum_{m=2}^{NR} \frac{A^x_{1,m}}{A^x_{1,1}} \frac{dc_v(m,t)}{dt} \quad (4.2.1.84)$$

$$c_v(j,t=0) = c_{vi}(j) \quad \text{for } j = 1 \text{ to NR} \quad (4.2.1.85)$$

Equation (4.2.1.85) is used initially to calculate the time-derivatives in the order: (4.2.1.82), (4.2.1.83), and (4.2.1.84). The values of the derivatives are input to GEAR and new values of the concentrations are returned. This cycle is repeated until the desired output time is reached.

Analytic solutions also exist for (4.2.1.74). Al-Niami and Rushton [1978] used (4.2.1.74) to describe the dispersed flow of a tracer to an abstraction well operating under steady-flow conditions. They solved (4.2.1.74) subject to (4.2.1.75) and:

$$c_v(0 \le r \le 1, t=0) = 0 \quad (4.1.3.86)$$

$$c_v(r=1, t>0) = 1 \quad (4.1.3.87)$$

The resulting solution is in terms of modified Bessel functions, which for this problem are only expandable for specific values of Pe_v. The solutions of (4.2.1.74) developed by Al-Niami and Rushton [1978] for values of Pe_v of 1, 5, and 11 are presented below.

For $Pe_v = 1$, $r = 0$:

$$c_v(r=0,t) = 1 + 2\sum_{n=1}^{\infty}(-1)^n \exp[-(n\pi)^2 t/2] \qquad (4.2.1.88)$$

For $Pe_v = 5$, $r = 0$:

$$c_v(r=0,t) = 1 - \frac{2}{15}\sum_{n=2}^{\infty}\frac{\alpha_n{}^3\exp(-\alpha_n{}^2 t/10)}{\alpha_n\cos(\alpha_n) - \sin(\alpha_n)} \qquad (4.2.1.89)$$

Where α_n, n=2,3,..., are the roots of:

$$\alpha_n\cot(\alpha_n) = [3 - \alpha_n{}^2]/3 \qquad (4.2.1.90)$$

For $Pe_v = 11$, $r = 0$:

$$c_v(r=0,t) = 1 + \frac{2}{10395} \cdot \sum_{n=2}^{\infty}\frac{\alpha_n{}^9\exp(-\alpha_n{}^2 t/22)}{[\alpha_n{}^4-45\alpha_n{}^2+105]\sin(\alpha_n) + [10\alpha_n{}^3-105\alpha_n]\cos(\alpha_n)} \qquad (4.2.1.91)$$

Where α_n, n=2,3,..., are the roots of:

$$\alpha_n\cot(\alpha_n) = [15\alpha_n{}^4-420\alpha_n{}^2+945]/[\alpha_n{}^4-105\alpha_n{}^2+945] \qquad (4.2.1.92)$$

Complete Mixing. A solution is obtained by assuming complete mixing within the soil cylinder. The solution of a complete mixing model for constant influent concentrations and uniform influent concentrations is

$$c_v(t) = c_{vi}e^{-t} + \left[\frac{c_{vo} + Ar\ c_{bo}}{1 + Ar}\right][1 - e^{-t}] \qquad (4.2.1.93)$$

Typically, Ar = 0, $c_{vo} = 0$, and $c_{vi} = 1$, thus, (4.2.1.93) reduces to:

$$c_V(t) = e^{-t} \qquad (4.2.1.94)$$

Table 4.2.1.3 is a summary of the equations that comprise each model. Most vapor extraction models reported in the literature assume complete mixing, and the models account for finite rates of diffusion, usually assume local equilibrium (see Table 1.2). To date, no one has considered rate limitations, such as air-water mass transfer or intraaggregate diffusion in vapor extraction models. It is probably valid to assume local equilibrium for removing the pure organic phase [Baehr, 1989; McClellan and Gillham, 1990]. The removal of the dissolved phase may become limited by phase equilibrium [McClellan and Gillham, 1990]. In addition, no one has accounted for removal with air flow and simultaneous leaching in water. For impermeable capped systems (plastic liners or pavement or concrete structures), leaching will not be important, however, leaching could occur beneath a soil cover.

Table 4.2.1.3. Summary of Equations Comprising the Radial Vapor Extraction Models.

	Included Mechanisms								
Model	Air Adv.	H. Air Disp.	V. Air Diff.	Air-Water Trans.	Water Adv.	Water Disp.	Dimensionless PDEs	OC ODEs	Verifcation Equations
General Form	X	X	X	X	X	X	(4.2.1.12-21)	(4.2.1.32-40)	none
No Vertical Diffusion	X	X		X	X	X	(4.2.1.13,14, 18-21,41,42)	(4.2.1.35,38-40,45-47)	none
Plug Flow	X			X	X		(4.2.1.13,14, 48-51)	(4.2.1.47,56-59)	(4.2.1.64-67)[a]
Local Equilibrium	X	X	X		X	X	(4.2.1.13,14, 16,50,68,69)	(4.2.1.35,36, 47,58,72,73)	none
No Water Flow L.E.	X	X					(4.2.1.74-77)	(4.2.1.82-85)	(4.2.1.88[b] 89[c],90[d])
Complete Mixing[e]	X	X	X		X	X			(4.2.1.93)

[a] Finite difference approximation for steady-state conditions.
[b] Analytic solution for $Pe_v = 1$.
[c] Analytic solution for $Pe_v = 5$.
[d] Analytic solution for $Pe_v = 11$.
[e] Analytic model that assumes dispersion is fast enough to establish perfect mixing throughout.

4.2.1.6 Model Verification. Verification of the numerical method used to solved the radial system models is not straightforward because analytic solutions for the coupled two-dimensional system of equations do not exist. Nevertheless, the description of vertical transport with water flow was verified in the one-dimensional development. The OC approximation of radial air advection and diffusion is compared to the analytic solutions for the no water flow, local equilibrium model. The coupling aspect of the horizontal and the vertical equations and the vertical vapor diffusion approximation have not yet been addressed in this study. Numerical approximation of the vertical vapor diffusion component of the air-phase transport equation was verified by Bednar [1990] in his description of vapor diffusion through columns under no-flow conditions. The verification of the coupling by the air-water mass transfer description (volatilization/dissolution) is performed by comparing to a finite difference approximation of the plug flow model for steady-state conditions.

Table 4.2.1.4 shows the calculations of the plug flow model for 6 horizontal and 6 vertical collocation points at t = 500 to calculations of the finite difference approximation of steady-state, plug-flow conditions (4.2.1.64-67). The dimensionless group values used for these calculations were chosen arbitrarily, but similar results were obtained with a wide range of values. The agreement is good and increases the confidence in the OC approximation of air-water mass transfer rates. In addition, this exercise has shown that verification of a numerical solution can be obtained by comparing calculations to either analytic solutions or approximations with other numerical methods.

Table 4.2.1.4. Comparison of a Finite Difference (FD) Approximation of the Steady-State Plug Flow Model for a 100 by 100 mesh to the Orthogonal Collocation (OC) Approximation of the Plug Flow Model at t = 500 for 6 Horizontal and 6 Vertical Collocation Points (25 Nodes). ($Ar = 0.5$, $c_{bo} = 1$, $c_{vo} = 0$, $St_v = 1$)

Num. Meth.	x:	0.00		0.14		0.42		0.72		0.94		1.00
	z	c_b	c_v	c_b	c_v	c_b	c_v	c_b	c_v	c_b	c_v	c_b
OC	0.00		0.63		0.58		0.44		0.24		0.06	
FD			0.63		0.58		0.44		0.24		0.06	
OC	0.14	0.90	0.54	0.88	0.48	0.85	0.36	0.81	0.19	0.77	0.04	0.76
FD		0.90	0.53	0.88	0.48	0.85	0.36	0.81	0.19	0.77	0.04	0.75
OC	0.42	0.71	0.38	0.68	0.34	0.61	0.24	0.53	0.12	0.46	0.02	0.44
FD		0.71	0.38	0.68	0.34	0.62	0.24	0.53	0.12	0.46	0.03	0.44
OC	0.72	0.53	0.26	0.49	0.22	0.42	0.15	0.33	0.07	0.26	0.01	0.24
FD		0.52	0.26	0.49	0.22	0.42	0.15	0.33	0.07	0.25	0.01	0.23
OC	0.94	0.42	0.20	0.38	0.17	0.31	0.11	0.23	0.05	0.17	0.01	0.15
FD		0.42	0.20	0.39	0.17	0.31	0.11	0.23	0.05	0.17	0.01	0.15
OC	1.00	0.39	0.18	0.36	0.15	0.29	0.10	0.21	0.04	0.15	0.01	0.14
FD		0.40	0.18	0.36	0.15	0.29	0.10	0.21	0.04	0.15	0.01	0.14

Coefficient matrices used to approximate radial derivatives for symmetric partial differential equations (*e.g.* equations describing only radial diffusion) can be found in the literature [Finlayson, 1980]. Matrices for asymmetric radial PDEs (such as for radially-converging flow and diffusion) have not been reported. By transforming the radial coordinate, the resulting equations resemble planar geometry and asymmetric, planar coefficients can be used to approximate the transformed derivatives. To check the validity of this manipulation, the numerical model approximation of radial flow and diffusion (equations (4.2.1.82-85)) are compared to analytic solutions (equations (4.2.1.88-90)).

Figure 4.2.1.3 shows analytical-numerical comparisons for values of Pe_v of 1, 5, and 11. The solid lines in all of the graphs are calculations of the solution of (4.2.1.82-85) for 10 radial points, and the dashed lines in Figures 4.2.1.3a, 4.2.1.3b, and 4.2.1.3c are calculations of (4.2.1.88), (4.2.1.89), and (4.2.1.90), respectively. The difference between the numerical model solution and the analytic solutions increases as the value of Pe_v decreases. The center of breakthrough predicted by the numerical model is located at a dimensionless time of 1 while the center determined with analytic solution is almost 1 for Pe_v of 11 and appears earlier as the value of Pe_v decreases. The center of breakthrough is not suppose to be a function of Pe_v [Hashimoto *et al.*, 1964]. For a low value of Pe_v, it would be expected that a complete mixing model calculation could be used to describe the spreading [Brenner, 1962]. The dotted line in Figure 4.2.1.3d is a calculation of (4.2.1.94) and it agrees with the numerical solution.

There is an important difference between the boundary conditions for the numerical model and the analytic solutions of Al-Niami and Rushton [1978]. The numerical model employs an overall mass balance as a boundary condition while the analytic solutions are subject to a prescribed concentration. Parker and van Genuchten [1984] found for the one-dimensional convection-dispersion equation that using a specified concentration as a boundary condition (they call it a boundary condition of "first-type") corresponds to flux-averaged concentrations, and "third-type" boundary conditions, as used in the numerical model, give results corresponding to volume-averaged (resident) concentrations. The distinction between the two is more important as dispersion increases.

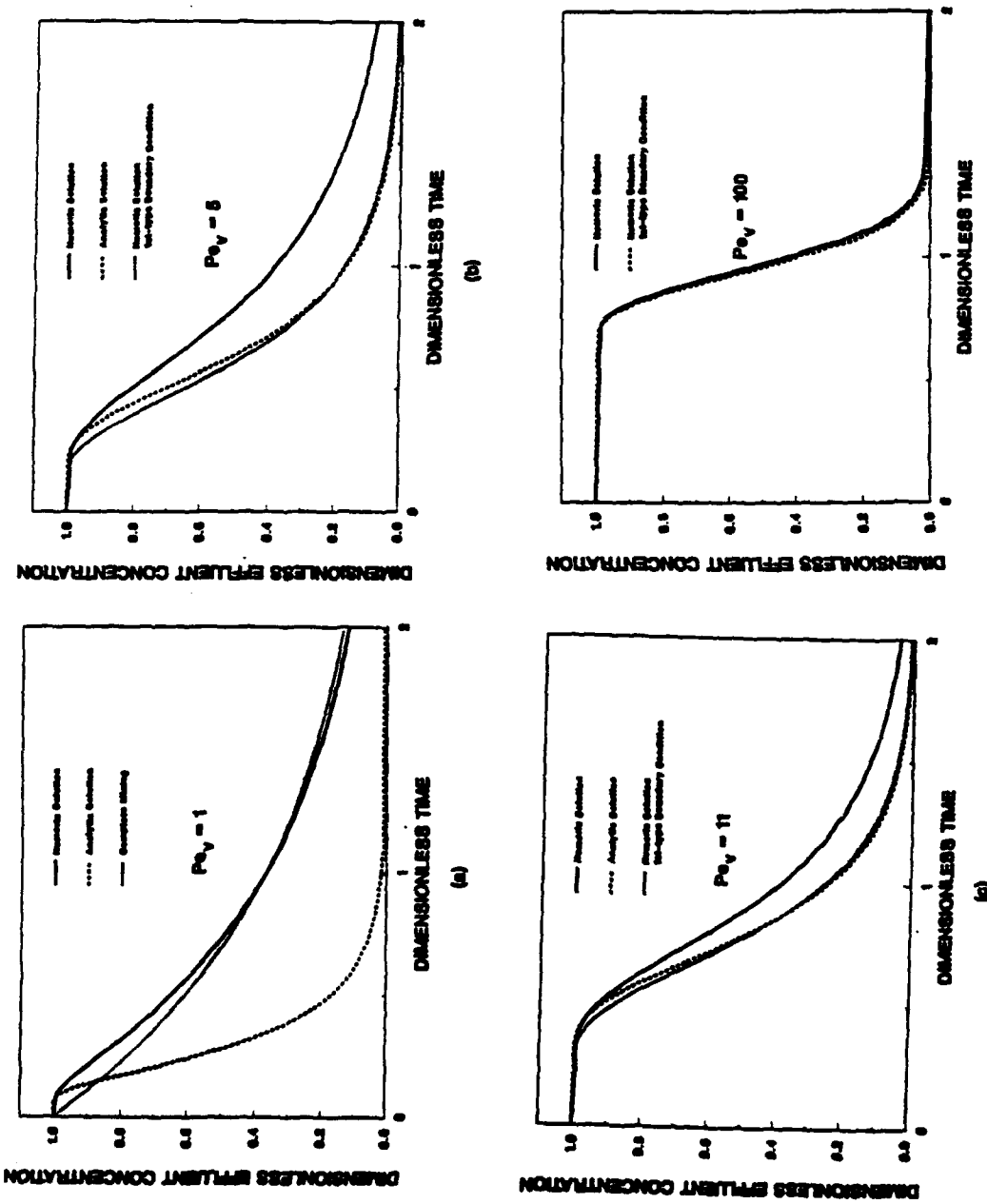


Figure 4.2.1.3. Comparisons of the numerical approximation for of the no water flow, local equilibrium model for first- and third-type boundary conditions and 10 collocation points to analytic solutions by Al-Niami and Rushton [1978] for Pe_y = 1, 5, and 11.

Flux-averaged concentrations have no physical meaning [Parker and van Genuchten, 1984], and this is indicated by the variation of the center of mass of the elution curve with Pe_v for the analytic solution. Chen [1987] solved (4.2.1.74) for radial dispersion from an injection well using Cauchy (third-type) conditions, however, he assumed that radial dispersion was proportional to velocity so his solutions could not be used here for comparing to the numerical solution. He was also able to show that flux-averaged and resident concentration formulations give the same results for low dispersion (high value of Pe_v). Batu and van Genuchten [1990] extended the work of Parker and van Genuchten [1984] to two dimensions and found similar results.

To show that the numerical approximation could simulate the incorrect analytic solutions, the mass balance boundary condition (equation (4.2.1.83)) was replaced with a first-type condition:

$$c_v(NR,t>0) = 1 \qquad (4.2.1.95)$$

The results of this modified numerical version are shown as dotted lines on Figures 4.2.1.3b and 4.2.1.3c. As the value of Pe_v increases the solutions for first- and third-type conditions coincide as shown in Figure 4.2.1.3d for Pe_v = 100 (analytic solutions for values of Pe_v greater than 11 are not reported). These results are identical to those reported by Parker and van Genuchten [1984].

Because the models are derived for resident concentrations, the mass balance boundary condition is used in the numerical solution of the general form of the model. The dispersed flow, local equilibrium model uses a mix of first- and third-type boundary conditions for numerical

convenience, but this model is used only for values of Pe_v greater than 64.

Verification of the numerical approximation of the general version of the radial vapor extraction model is not complete, but this piecewise approach where specific aspects of the model are verified separately gives confidence to the numerical results that follow.

4.2.1.7 Numerical Results (Model Sensitivity). A complete modeling approach, as followed in the preceding sections describing the column model developments, requires that at this stage the model be tested against laboratory or field results. The development of a set of laboratory experiments to validate the two-dimensional vapor extraction models is beyond the scope of this research. Many of the field and pilot scale results that are reported in the literature (*cf.* Hutzler *et al.* [1989a]) do not contain sufficient information to obtain accurate model comparisons or the system configurations are much different than the assumptions that are used to derive the models reported herein.

Models that are not validated can be used to determine which mechanisms are most important, and, in addition, give some guidance to system design. More importantly, models are additional tools which are available for enhancing the decision process of design engineers. The mechanisms that are examined with the models are vapor diffusion, air-water mass transfer (and with this, intraaggregate diffusion), air and water advection, and water dispersion. Calculations for giving design guidance of the optimum air withdrawal rate are also performed. Finally, the suitability of using a column model development approach to developing field-scale system models is discussed.

Parameter Estimation. Parameter values for the model calculations

are based on properties of the Ottawa sand and aggregated porous media used in the column studies. Values used in the calculations of the two-dimensional radial models are given in Table 4.2.1.5.

Table 4.2.1.5. Parameter Values for Model Calculations of Toluene Removal from Ottawa Sand (OS) and Aggregated Porous Material (APM).

Parameter Values	OS	APM
Radius of Influence, R_i (m):	4.5	4.5
Radius of Vent Hole, R_w (m):	0.12	0.12
Depth to ground water, D (m):	4.5	4.5
Infiltration rate, $v\epsilon S$ (cm s^{-1}):	0, 0.00026	0, 0.00026
Air withdrawal rate, Q_G (cm^3 s^{-1}):	21-2.1(10^6)	21-2.1(10^6)
Air conductivity, K_a (cm s^{-1}):	0.018	0.018
Degree of saturation, S:	0.30	0.67
Porosity, ϵ:	0.33	0.70
Soil density, ρ_s (g cm^{-3}):	2.65	1.51
Henry's constant, H:	0.27	0.27
Air diffusion coef., D_e (cm^2 s^{-1}):	0.050	0.050
Liquid dispersion coef., E_z (cm^2 s^{-1}):	0.0033	0.0033
Air-water mass transfer rate, K_La (s^{-1}):	0.010	0.0020
Sorption capacity, K (cm^3 g^{-1}):	0	0, 1.0
Initial contaminant concentration		
in air, C_{vi} (mg L^{-1}):	1.0	1.0
in water, C_{bi} (mg L^{-1}):	3.7	3.7
on soil, Q_i (mg kg^{-1}):	0	0, 3.7
Influent contaminant concentration		
in air, C_{vo} (mg L^{-1}):	0	0
in water, C_{bo} (mg L^{-1}):	0	0

The chemical of interest for these calculations is toluene. System size is chosen to be similar to the soil volume of one unit of the trench system at Hill AFB, Utah (see DePaoli *et al.* [1990]). Water velocity is chosen to compare with the measured moisture characteristics given in Figure 4.1.1.4. Degree of saturation is chosen so that the air-filled porosity of the sand and aggregated material are the same

(0.231). Air flowrate is varied to examine its effect on vapor extraction performance and because it is the most flexible variable with regard to system operation. The value of Henry's constant and of the gas diffusion coefficient of toluene is obtained from the literature as described in Sections 4.1.1.3 and 4.1.2.3, the liquid dispersion coefficient is taken as that measured in the column studies adjusted for water velocity.

The parameter with the most uncertainty is the specific air-water mass transfer rate ($K_L a$). Its value was shown to be unimportant in the column studies, in part, because it was assumed that the limiting step to equilibrium was intraaggregate diffusion. Intraaggregate diffusion is not accounted for in the two-dimensional models at this time and, thus, the limiting resistance is air-water mass transfer. Values of the air-water mass transfer coefficient that are used in the following sensitivity analyses are determined to simulate the impact of diffusion out of water because the experimental efforts thus far indicate air-water mass transfer is fast. The remainder of this subsection give a brief justification for this assumption. First, an approximation is derived to determine the minimum air-water mass transfer rate for subsurface systems. Second, model calculations are used to show that the impact of diffusion from immobile can be simulated with an appropriate value of $K_L a$.

There is a substantial lack of volatilization studies in the subsurface transport literature. Most of the work on gas-liquid mass transfer has been performed using columns of artificial packing with pure fluids at high velocities. Most of these correlations are based on liquid velocity and in such a way that the mass transfer rate decreases

with decreasing velocity -- often going to zero as water velocity tends to zero. For volatilization/dissolution between a gas and a liquid, the overall mass transfer coefficient (K_L) is considered to be a combination of a resistance to mass transfer in the gas and a resistance in the liquid (cf. Treybal [1980]):

$$\frac{1}{K_L} = \frac{1}{k_l} + \frac{1}{Hk_g} \qquad (4.2.1.96)$$

If the resistance in one fluid is much greater than in the other, the fluid with the greater resistance is said to "control." Correlations commonly used for estimating the relative contributions of gas and liquid resistances are of the form [Nicoud and Schweich, 1989]:

$$Sh_l = c_{l,1} + c_{l,2} (Re_l)^{c_{l,3}} (Sc_l)^{c_{l,4}} \qquad (4.2.1.97a)$$

$$Sh_g = c_{g,1} + c_{g,2} (Re_g)^{c_{g,3}} (Sc_g)^{c_{g,4}} \qquad (4.2.1.97b)$$

The Sherwood numbers (Sh_l or Sh_g) are ratios of the product of a local mass transfer coefficient (k_l or k_g) and a diffusion thickness (δ_l or δ_g) to a diffusion coefficient (D_l or D_G); the Reynolds numbers (Re_l or Re_g) are ratios of the product of a fluid velocity (v or u) and particle size (R_a) to a kinematic viscosity (μ_l/ρ_l or μ_g/ρ_g); the Schmidt numbers (Sc_l or Sc_g) are ratios of a kinematic viscosity to a diffusion coefficient; and the $c_{l,i}$'s and $c_{g,i}$'s are experimental constants. For no fluid flow in packed beds, (4.2.1.97) reduces to (*loc. cit.* Sontheimer *et al.* [1988])

$$Sh_l = Sh_g = c_{l,1} = c_{g,1} = c_l = 2 + 3[1 - \epsilon(1-S)] \qquad (4.2.1.98)$$

In soil columns where there is no air flow, such as in the water

flow experiments described in Section 4.1.1.7, (4.2.1.98) predicts a value of k_g of approximately 10 cm s^{-1} (this estimate assumes an air-filled porosity of 0.27, a gas diffusion coefficient of 0.08 cm^2 s^{-1}, and δ_g of $0.155R_a$, which is the radius of a geometric estimate of the air-filled pore size); while for extraction conditions where no flow of water occurs, (4.2.1.97) estimates a value of k_l of approximately 0.03 cm s^{-1} (assuming $\epsilon(1-S) = 0.27$, $D_l = 0.000008$ cm^2 s^{-1}, and $\delta_l = 0.001$ cm, which is a geometric estimate of the water film thickness surrounding a 0.07 cm diameter sand particle at a degree of saturation of 0.3). Hence chemicals having a low H in a quiescent air-water system, the limiting resistance is diffusion in water. Furthermore, values estimated by (4.2.1.98) should be a minimum mass transfer rate because (4.2.1.97) predicts higher rates as flow increases. A value of $K_L = k_l = 4D_l/\delta l$ could be used to determine the greatest impact of volatilization/dissolution rates in subsurface systems. This analysis contradicts (predicts values of the mass transfer coefficient higher by a factor of 10,000) estimates of K_La by the Turek and Lange [1981] correlation (equation (4.1.1.52)) that was used in estimating K_La in Section 4.1.1.7. This is probably due to the fact that the Reynolds number for fluid flow in soils is on the order of one thousand times less than those used in trickle-bed studies. This contradiction exists between the theory from which (4.2.1.98) is derived and all correlations developed for mass transfer coefficients in packed towers when they are extrapolated to air-water-soil systems.

Results from the one-dimensional studies indicate that diffusion inside aggregates could be more important in describing chemical fate, and this was as observed in the experimental results shown in Figures

4.1.1.6, 4.1.1.7, 4.1.1.9, 4.1.1.10, and 4.1.2.9. The impact of intraaggregate diffusion can be studied by adjusting the air-water mass transfer rate. Figure 4.1.1.5 shows that air-water mass transfer and intraaggregate diffusion have similar impacts on the spreading of chemical breakthrough curves. Equation (4.1.1.47) can be used to determine when intraaggregate diffusion and mass transfer across the aggregate surface are equivalent. The same approach that was used to obtain (4.1.1.47) is used to determine, from the vapor extraction column model (equations 4.1.1.14, 4.1.1.20, 4.1.1.22, and 4.1.2.4-8), when intraaggregate diffusion has the same impact as air-water mass transfer when air is flowing. The result is similar to (4.1.1.47) and to the equivalent spreading relationship derived by Crittenden *et al.* [1986] for saturated chemical transport:

$$Pe_v = 15Ed_p = 3St_v \qquad (4.2.1.99)$$

The difference between (4.2.1.99) and (4.1.1.47) is that the groups in the former are based on advection in air and the latter on advection in water.

Equation (4.2.1.99) is used with the air extraction column model results displayed in Figure 4.1.2.9 to determine a value of K_La for the aggregated porous media that would affect the model calculation in the same manner as intraaggregate diffusion. The results are shown in Figure 4.2.1.4. The solid curve shows a model calculation assuming fast air-water mass transfer as indicated by the magnitudes of the dimensionless groups shown (Ed_p = 0.23, St_v = 1000). The dashed curve assumes fast intraaggregate diffusion (Ed_p = 230) and the value of St_v (1.2) is obtained with (4.2.1.99). Therefore not including

intraaggregate diffusion in this model development does not preclude studying its impact. Moreover, because of the difficulties in predicting the air-water mass transfer rate and because this transfer rate has had relatively little influence on organic chemical transport in the columns studies reported here, it is probably better to use the average liquid diffusion flux for estimating $K_L a$.

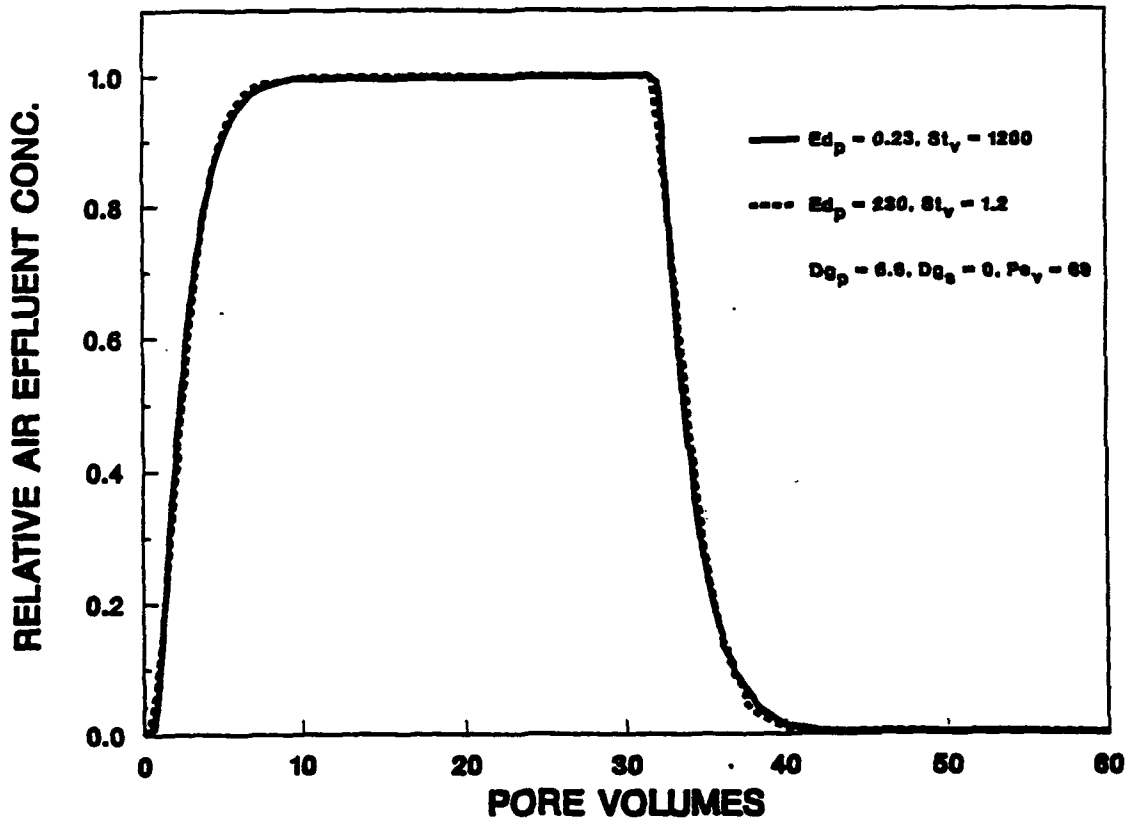


Figure 4.2.1.4. Vapor extraction column model comparisons for simulating the impact of intraaggregate diffusion with air-water mass transfer using the ratio $15Ed_p = 3St_v$. Parameters are from the model calculations that describe the data in Figure 4.1.2.9.

Equation (4.2.1.99) was derived for an assumption of no water flow. Equation (4.1.1.47) equates the spreading caused by axial dispersion, intraaggregate diffusion, and film transfer for simultaneous air and water flow, but no equivalent spreading relationship between

intraaggregate diffusion and air-water mass transfer exists for simultaneous air and water flow. The simultaneous flow model simulates the order of chemical transfer for elutuion as intraaggregate diffusion, film transfer, and volatilization. The influence of intraaggregate diffusion (or film transfer) on air-phase concentration could then be simulated using air-water mass transfer. As an approximation, the equivalent spreading relationship (4.2.1.99) is used to estimate a value of $K_L a$ for simulating the observed spreading of the air-phase concentration profile in a column where air and water are flowing simultaneously. Figure 4.2.1.5 shows calculations of the simultaneous air and water flow column model described in Section 4.1.1 for the conditions given in Figure 4.2.1.4, except in this case, water is also flowing at a rate such that the advective flux in water is about one-third the mass flux in air. The air-phase concentration history is shown in Figure 4.2.1.5b as the dashed line for intraaggregate diffusion limited conditions and the solid line is the approximation for equivalent impact by air-water mass transfer ($K_L a = 0.002\ s^{-1}$). The corresponding water-phase concentration histories are displayed in Figure 4.2.1.5a.

Figures 4.2.1.4 and 4.2.1.5 show that using a value of $K_L a$ that accounts for the impact of intraaggregate diffusion can be used in the two-dimensional extraction models to approximate the effect of the combined mass transfer rates (intraaggregate diffusion and volatilization).

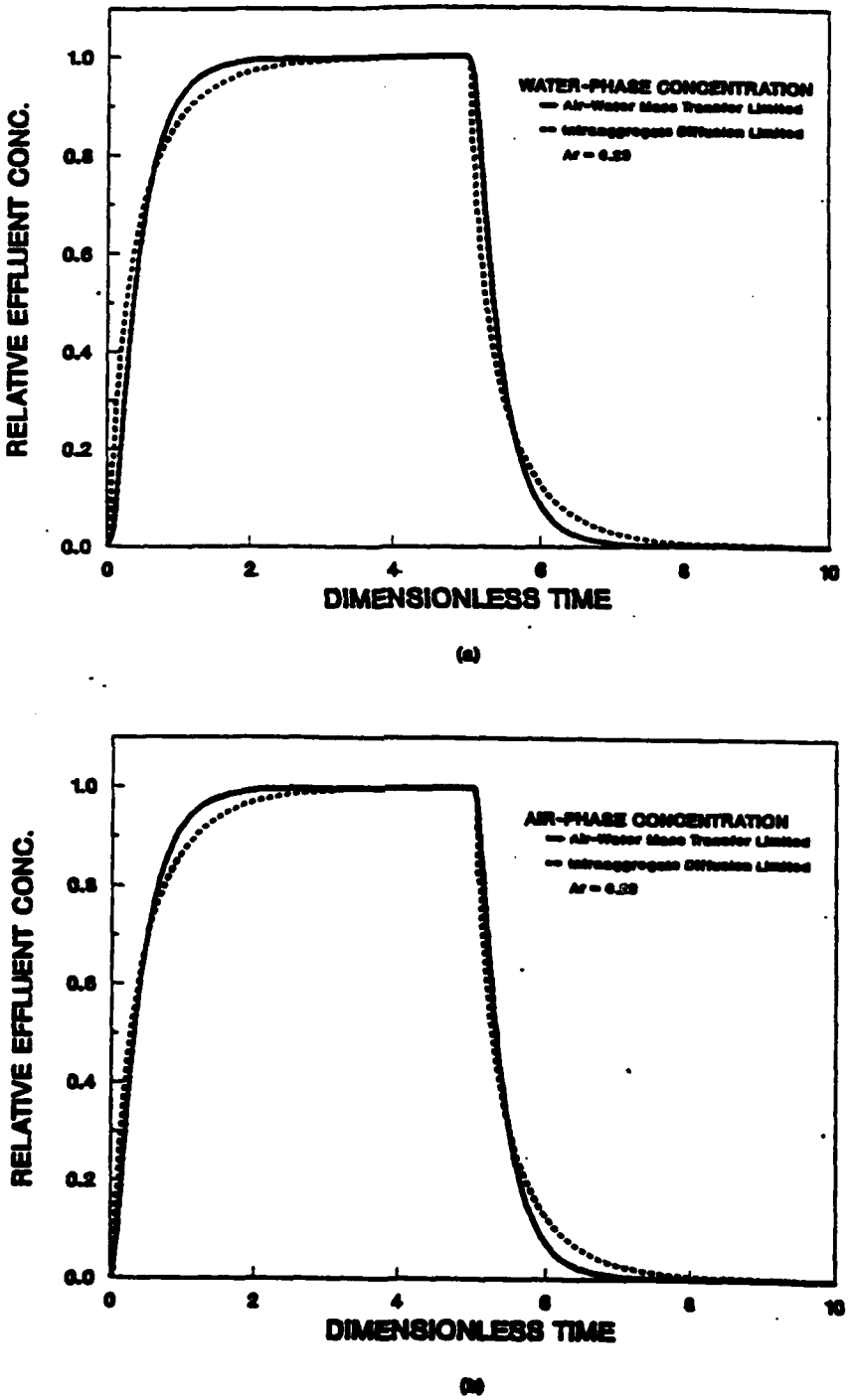


Figure 4.2.1.5. Simultaneous flow column model comparisons for simulating the impact of intraaggregate diffusion with air-water mass transfer using the ratio $15Ed_D = 3St_V$ that is derived from vapor extraction column model. Part (a) is the result for water-phase concentration and part (b) is the result for air-phase concentration.

As a result of the uncertainty involved with estimating $K_L a$, it was decided in this work to use two values and study their impact for various air velocities. Liquid diffusion and air-water mass transfer had no impact on extraction from Ottawa sand, so an arbitrary value of $K_L a$ of 0.01 s^{-1} is used for cohesionless soil cleanup simulations. The impact of diffusion out of saturated aggregates was observed and predicted in the aggregated porous material. This impact is shown in Figure 4.2.1.4 to be equivalent to a value of $K_L a$ of 0.002 s^{-1}, so this value is used to simulate aggregated soils.

The vapor extraction performance figures that follow show the dimensionless toluene concentration from the extraction vent (starting at 1 along the left ordinate) and the percent of mass removed from the system with air flow (starting at 0 and approaching 100% along the right ordinate) as a function of the number of air pore volumes extracted. Extraction calculations are performed for cohesionless soils (sands) first, and similar numerical experiments are performed with an aggregated porous material.

Model Calculations for Toluene Removal from Sand

The first model scenario simulations assume a capped system where no water is flowing. The initial condition is a uniform concentration distribution. For this initial condition and no water flow, the vertical diffusion component of the contaminant transport equations has no impact. The model could not solve nonuniform initial concentration profiles in general. Numerical instabilities arise when an initial profile with an abrupt change in concentration is used. Smooth profiles, for example an exponential decay, give results that are physically acceptable with regards to conservation of mass for short

times, but for longer times the boundary conditions at z equal to 0 and 1 for the water phase and at z = 1 for the air phase introduce instabilities in the numerical solution. The time derivatives for air-phase concentration at the boundaries are dependent on changes in concentration at all of the interior collocation points along the same vertical plane. This is a disadvantage of the OC method because in some instances the domain of dependence encompasses concentrations that should not impact a particular location. Bednar [1990] solved this problem by replacing a no gradient boundary condition at the column exit of his no-flow, diffusion model with a specified concentration that varied such that mass in the column is conserved. This type of approach could not be used here because there would be too many unknowns and not enough equations. This leads to another obvious shortcoming of OC in that it is not easy to change boundary conditions without substantial revisions to the code.

The base case air flow rate is 2100 $cm^3\ s^{-1}$, which corresponds to an air detention time of about nine hours. The pressure drop (ΔP) can be determined from

$$\Delta P = Q_G \ln(R_i/R_w)/(2K_a \pi D) \qquad (4.2.1.100)$$

Equation (4.2.1.100) assumes that air acts as an incompressible fluid and that flow obeys Darcy's law. System size (R_i and D) are given in Table 4.2.1.5, R_w is the radius of the hole for the vent (usually this is about 12 cm), and K_a is the air conductivity. Air conductivity is estimated from the saturated hydraulic conductivity (K_s) given in Table 4.1.1.5 [$K_a = K_s \rho_l \mu_g (\rho_g \mu_l)^{-1} = 0.018$ cm s^{-1}]. It is assumed for these pressure drop calculations that the air conductivity is not a function

of moisture content. Air conductivity decreases with increasing water content but this was not measured. A flow rate of 2100 cm^3 s^{-1} would result in a pressure drop of approximately 0.1 mm Hg (field systems typically operate at vacuums between 5 and 50 mm Hg [Hutzler *et al.*, 1989a]). In field situations, the value of R_i would vary with air flow rate and ΔP, however, here it is assumed that the system is vented to the atmosphere, such as by multiple inlet vents, at some distance R_i from the extraction vent. Therefore alterations of flow rate from the base case value will result in proportional changes in pressure drop.

A summary of system performance for various air flow rates is listed in Table 4.2.1.6. The calculations assume no water flow and no sorption (increases in removal times resulting from sorption are addressed in the calculations for the aggregated soil which follow). Increases in the air flow rate results in proportionally higher pressure drops, and power consumption rates increase with flow squared. Time to reach a level of cleanup decreases with increasing flow rate, however, the total energy consumed increases.

Table 4.2.1.6. Summary of Extraction Performance for Removing 99.9% of Toluene from a Cohesionless Soil and from an Aggregated Soil with Various Air Flow Rates in a Capped, Radial System. (Calculations assume no water flow and no sorption)

			Ottawa Sand				Aggregated Porous Material			
Q_G ($cm^3 s^{-1}$)	ΔP (cm air)	Power (Watts)	time (hours)	t	apv	Energy (kJoules)	time (hours)	t	apv	Energy (kJoules)
21	1.5	$4(10^{-6})$	16,000	6.9	18	0.23	47,000	7.0	54	0.68
210	15	$4(10^{-4})$	900	4.0	10	1.3	3,100	6.3	35	4.5
2,100	150	0.04	57	2.5	6	8.2	180	2.4	21	26
21,000	1500_4	4.0	6.8	3.0	8	98	21	2.9	24	300
210,000	$1.5(10^4_5)$	400_4	0.6	2.6	7	860	4.8	5.0	42	6900
2,100,000	$1.5(10^5)$	$4(10^4)$	0.1	5.3	14	14,000				

apv = air pore volumes, t = throughput.

Extraction performance for three air flow rates is shown in Figure 4.2.1.6. The base case is a solid line, and the dashed line depicts the

reduction in extraction efficiency for a 10-fold decrease in the air withdrawal rate and the dotted line corresponds to a 100-fold decrease in air flow rate. The decreases in extraction efficiency that result are due to diffusional mixing in the air phase as indicated by the decreasing value of Pe_v with air flow rate.

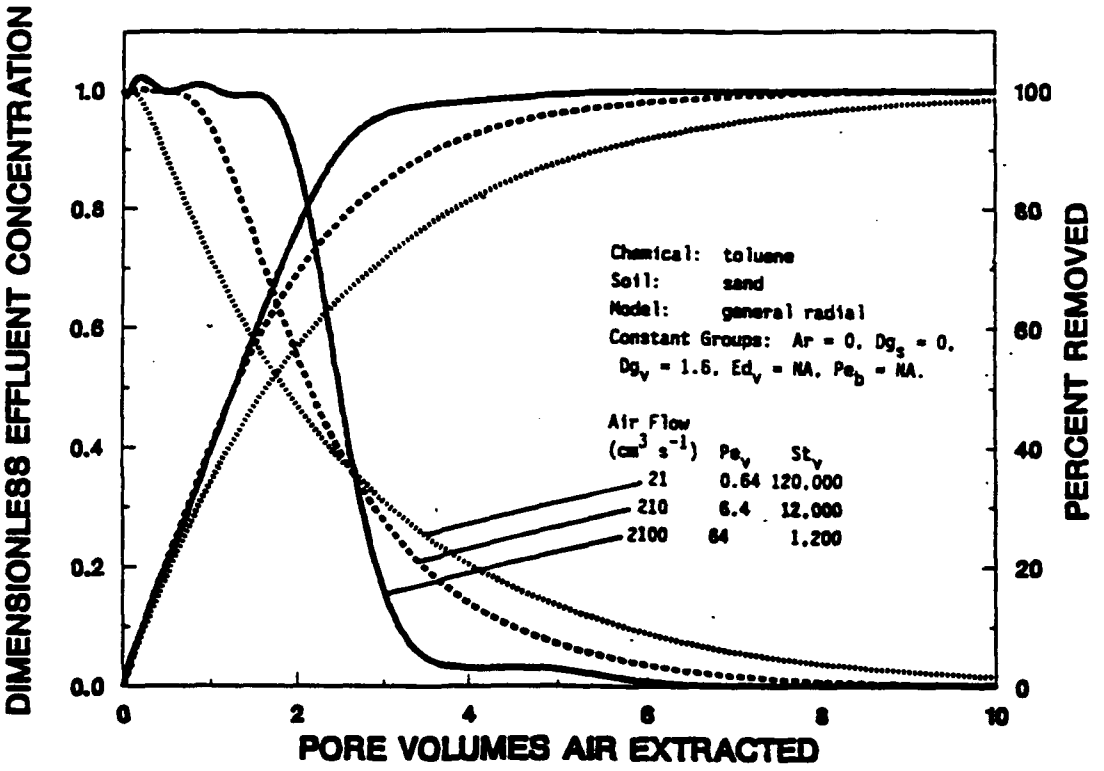


Figure 4.2.1.6. Toluene extraction performance as a function of air withdrawal rate (21, 210, 2100 $cm^3\ s^{-1}$) for a vertical vent configuration in Ottawa sand at a degree of saturation of 0.3.

Figure 4.2.1.7 shows the change in extraction performance for order-of-magnitude increases in the air extraction rate. The solid line is a simulation for an air withdrawal rate of 21,000 $cm^3\ s^{-1}$. One hundred- and 1000-fold increases in air flow from the base case are shown as dashed and dotted lines, respectively. Not only does power consumption increase but extraction efficiency in terms of air volume

extracted and total energy consumed decreases. Reductions in extraction efficiency with air flow rate are due to mass transfer rate limitations from the water-phase, as indicated by a decrease in St_v with increases in air flow rate. It is assumed that the air-water mass transfer rate does not change with air velocity. This is probably a good assumption if diffusion from water controls the mass transfer rate.

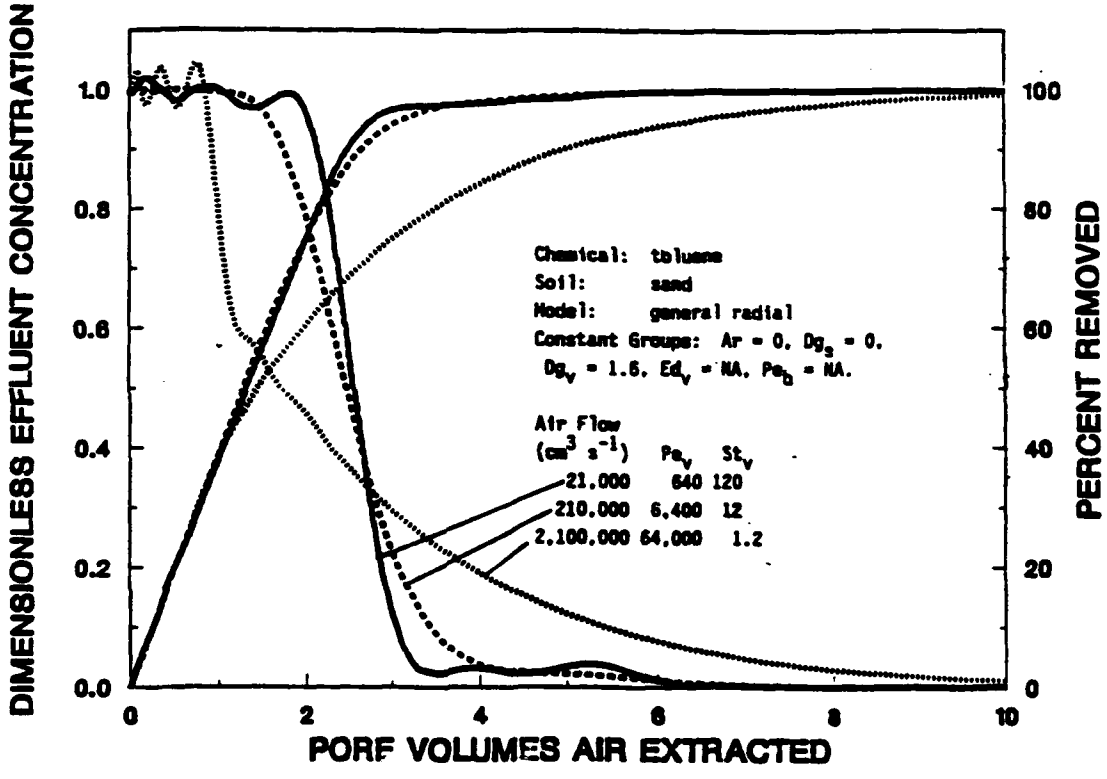


Figure 4.2.1.7. Toluene extraction performance as a function of air withdrawal rate (21,000, 210,000, 2,100,000 cm^3 s^{-1}) for a vertical vent configuration in Ottawa sand at a degree of saturation of 0.3.

Although flow rates has high as 2,100,000 cm^3 s^{-1} are typically not extracted from a single vent, these calculations show that there exists an optimum range of air flow rates. For these calculations the assumption of a constant air diffusion coefficient is not valid for air flow rates greater than 68,000 cm^3 s^{-1} (this value is determined from a

graph by Miyauchi and Kikuchi [1976] for the conditions given in Table 4.2.1.5) since fluid mixing becomes more important at higher air flow rates. Because for radial flow, air velocity increases with decreasing distance from the extraction vent, the model formulations in this section could not be used to study situations where air dispersion is greater than air diffusion.

Column model calculations have shown that it is not always possible to distinguish impacts of different mechanisms by just examining the shape of breakthrough curves. Figure 4.2.1.8 compares performance calculations for air flow rates of 2,100,000 $cm^3\ s^{-1}$ (dotted line) and 21 $cm^3\ s^{-1}$ (solid line) with a complete mixing calculation (dashed line). The equivalent spreading relationship given by (4.2.1.99) predicts that the two curves would match for a ratio of Pe_v to St_v of 3.

The following model calculations are used to observe the impact of water infiltration at a rate of 0.00026 cm s^{-1} (this corresponds to a degree of saturation of 0.30 for the sand and may represent vapor extraction during a period of heavy rain fall (9 inches per day) or snow melt). The introduction of water flow makes it possible to examine the impact of vertical gas diffusion and fully utilize the attributes of the two-dimensional model equations developed thus far. Although situations where water infiltration are rare, these calculations show that it is important from a mass transfer point of view to minimize water infiltration or if this can not be done, use higher air flow rates. Even when a portion of the contamination is leaching towards ground water, there is an upper limit of the air flow rate where there advantage in further increases in flow rate.

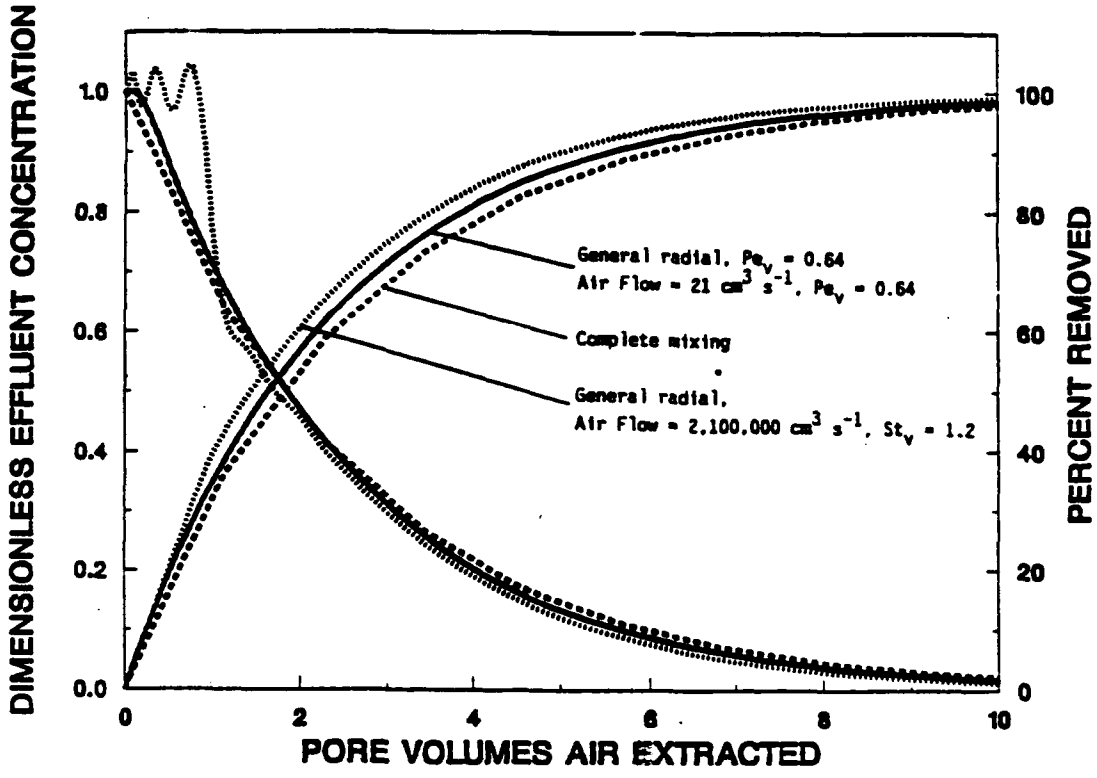


Figure 4.2.1.8. Comparison of toluene extraction performance for air withdrawal rates of 21 and 2,100,000 cm^3 s^{-1} to a completely mixed calculation.

During conditions where water infiltration occurs, the air flow rate should be fast enough such that mass removal rate in air is greater than the advective flux in water (Ar less than 1). If no air flow is induced, water flow column results in Section 4.1.1.7 indicate that it would take between approximately 60 and 120 hours to elute the toluene from the sand. Therefore an air flow rate of at least 2100 cm^3 s^{-1} (Ar = 0.29) is needed to reduce the amount of toluene leaching to ground water. Calculations for no water flow at this flow rate (see Figure 4.2.1.7) indicate that equilibrium between air and water could be assumed (St_v = 1200) so the general form of the model should compare with the dispersed flow, local equilibrium model. This comparison is

made in Figure 4.2.1.9 where the solid line is the calculation of the numerical solution to the general form of the model and the dashed line is the local equilibrium prediction. Even though the important mechanisms for these conditions are accounted for in both models, the calculations do not agree.

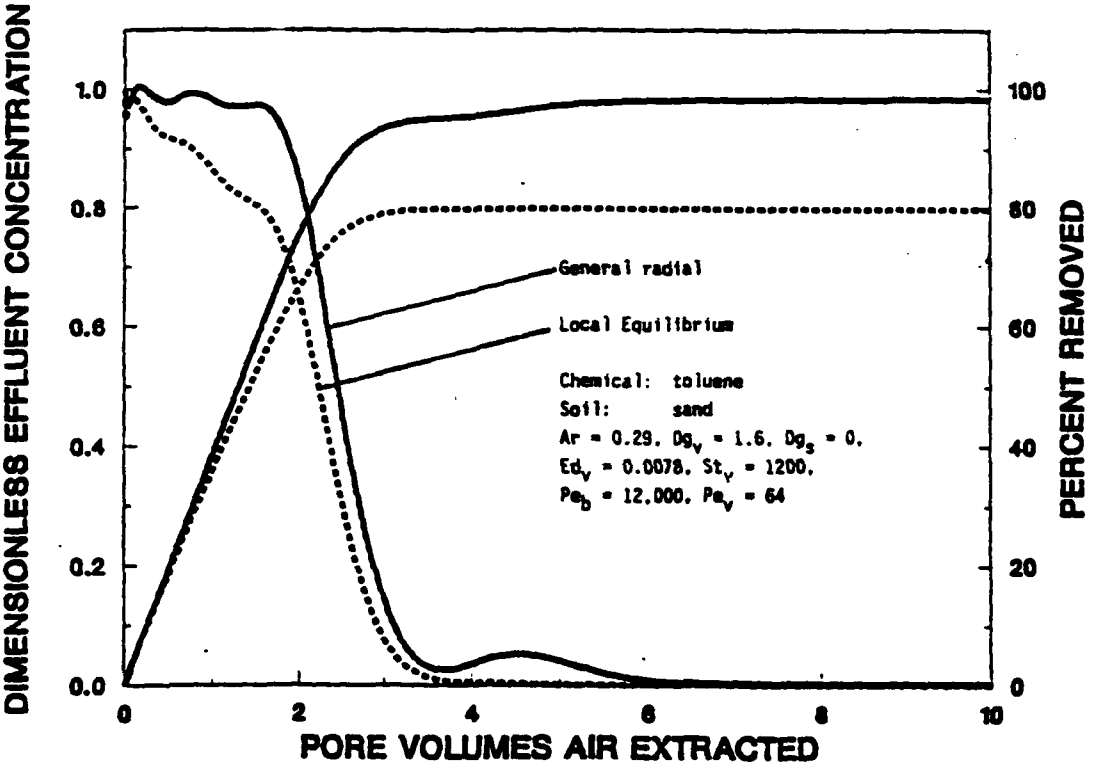


Figure 4.2.1.9. Comparison of the numerical solution of the general form of the radial configuration vapor extraction model for an air withdrawal rate of 2100 $cm^3\ s^{-1}$ and 0.00026 cm s^{-1} water infiltration rate to the numerical solution of the local equilibrium model.

The local equilibrium calculation is compared to the numerical solution for no vertical diffusion (dotted line), which is listed in Table 4.2.1.3, in Figure 4.2.1.10 and these agree for a value of D_e of 0 (dashed line) for the vertical diffusion component and differ slightly for a nonzero value of the vertical D_e (solid line).

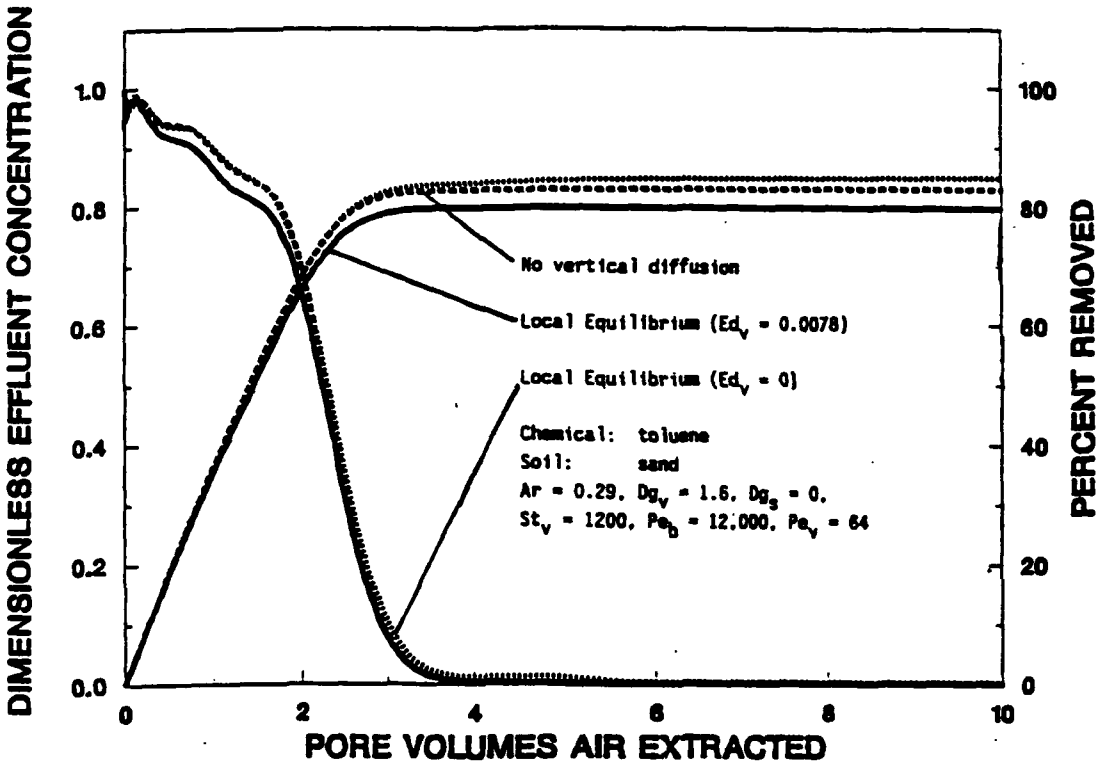


Figure 4.2.1.10. Comparison of the numerical solution of the no vertical diffusion model to the numerical solution of the local equilibrium model with and without vertical diffusion.

The general form of the model did not exhibit sensitivity to the value of D_e for vertical diffusion. Predictions by the general model and the model where vertical diffusion is ignored, as well as the local equilibrium model, differ more as the value of Ar increases. This discrepancy is attributed to the shortcomings of orthogonal collocation for approximating the air-phase boundary conditions at the water table in the general model, and, therefore, the numerical approximation of the general model is not satisfactory. Remaining calculations for simultaneous air and water flow will be made with the model that ignores vertical gas diffusion.

Multidimensional vapor diffusion may be important only where the

initial concentration profile is nonuniform or the air velocity varies with depth. If the air velocity is uniform with depth, then it is probably not important to consider vertical gas diffusion since the exhaust air-phase concentration at an extraction vent is a vertical average. Variations in air velocity with depth either due to changes in permeability or degree of saturation are probably more important to consider.

At an air flow rate of 21,000 $cm^3\ s^{-1}$, model calculations predict that nearly all (98%) of the toluene could be removed with air as shown by the dashed line in Figure 4.2.1.11. A ten-fold increase in air flow rate does not significantly improve the removal efficiency, while a ten-fold reduction reduces the percent removed in air to 85%.

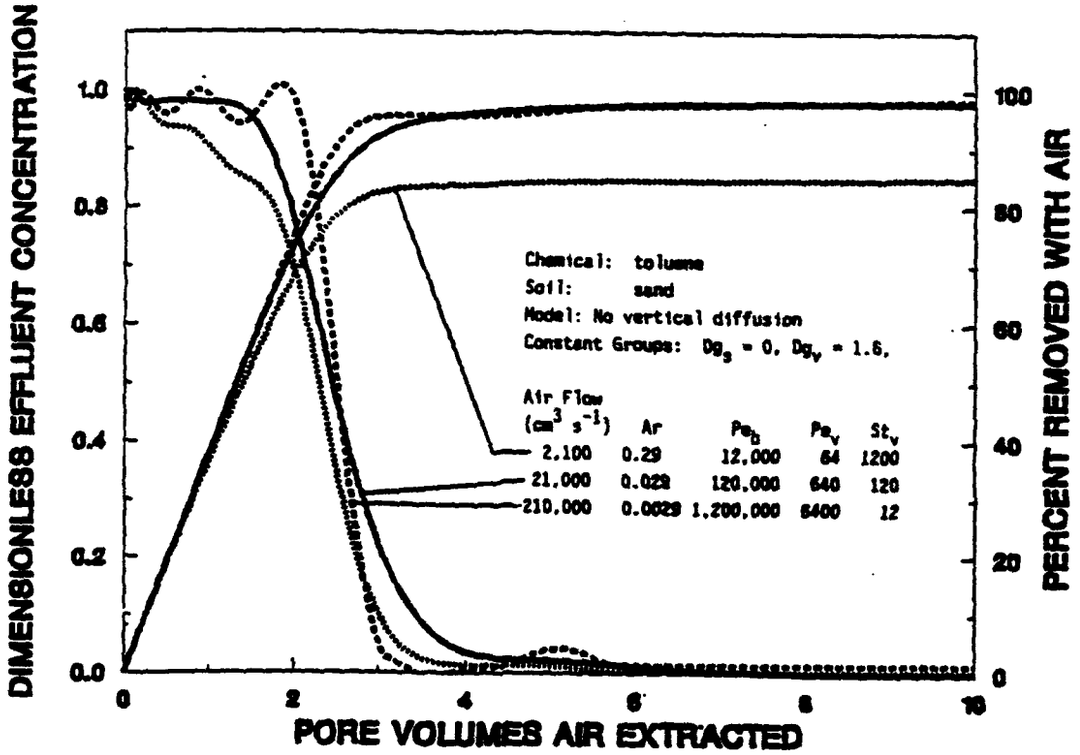


Figure 4.2.1.11. Comparison of toluene extraction performance for radial configuration at air withdrawal rates of 2100, 21,000, and 210,000 $cm^3\ s^{-1}$ and 0.00026 cm s^{-1} water infiltration rate.

The results described above indicate that the extraction rate of air should be between 21,000 and 210,000 cm^3 s^{-1} for either water infiltration situation. The slower flow rate could be used to attain the same level of cleanup for about 16% of the power consumption (see Table 4.2.1.6).

Model Calculations for Toluene Removal from APM

The conditions for determining the rate of toluene extraction from an aggregated porous material (APM) are similar to those reported above and are listed in Table 4.2.1.5. There are two primary differences in extraction from the sand and the APM. The degree of saturation in the APM is higher and the air-water mass transfer rate is slower ($K_L a$ = 0.002 s^{-1}) to account for diffusion of chemical from immobile water. The degree of saturation of this system is 0.67 and this results in an air porosity of 0.23, which is the same as that used in the calculations for the sand given above. An example is also given to show the impact of sorption on vapor extraction performance.

The impact of sorption is easy to examine. Column model calculations have shown that sorption has its greatest impact on retardation. Extraction performance calculations for removal of toluene from the APM at an air flow rate of 21,000 cm^3 s^{-1} assuming no sorption is compared to a calculation for a sorption coefficient (K) of 1 cm^3 g^{-1} (this corresponds to fraction of organic carbon of about 0.5% [Hassett *et al.*, 1983]) in Figure 4.2.1.12a for a dimensionless time scale (throughput). The throughput scale normalizes differences in retardation.

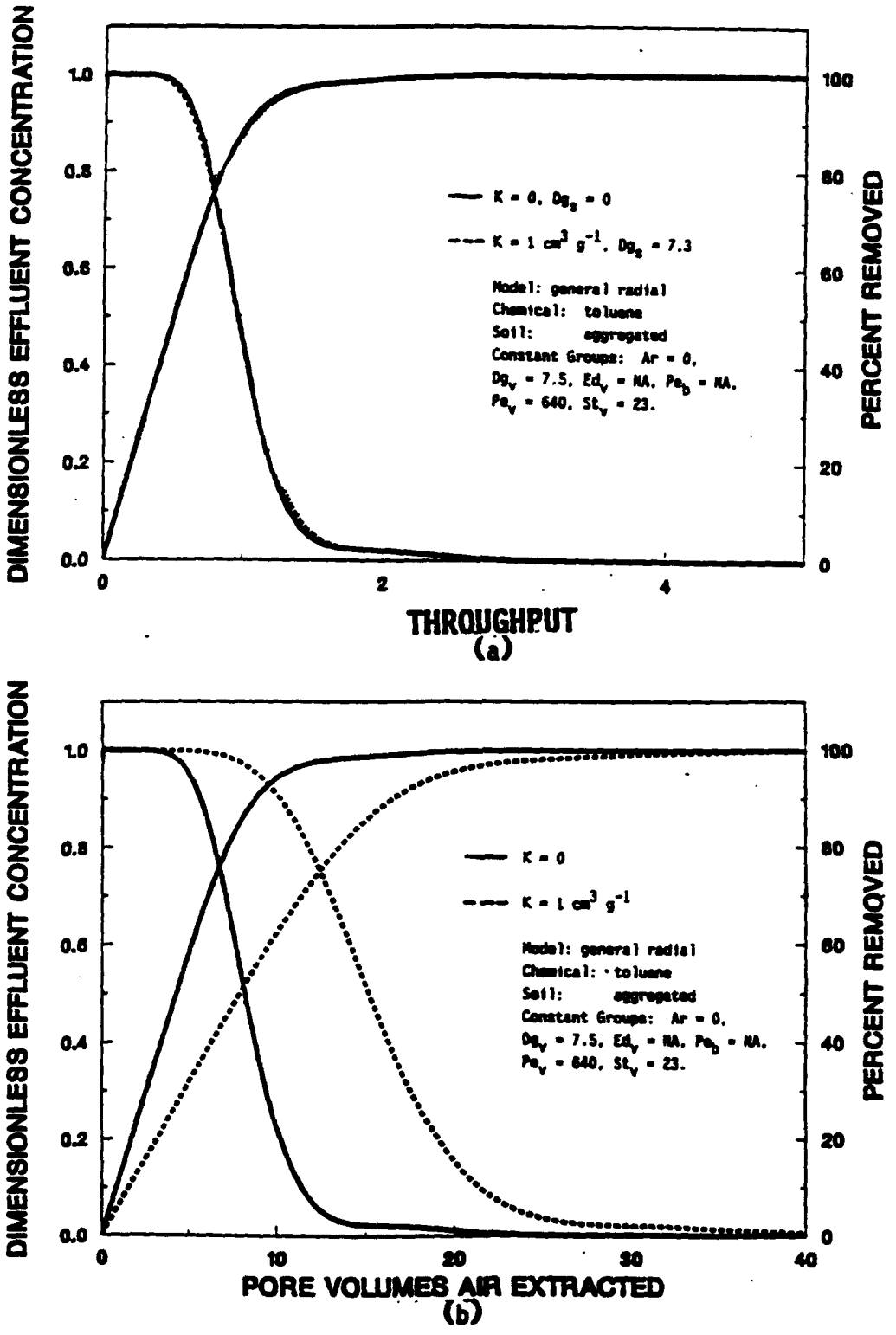


Figure 4.2.1.12. Comparison of toluene extraction performance for radial configuration from the aggregated porous media for K = 0 and 1 cm^3 g^{-1}: (a) throughput scale and (b) air pore volumes extracted.

Adsorption has little or no impact on chemical mixing so performance curves determined for no sorption can be adjusted for a given amount of sorption by increasing the time scale to match the increased retardation as shown in Figure 4.2.1.12b. By the same token, because the air-filled porosity of the sand and the APM are assumed to be the same, then when gas diffusion is the predominant spreading mechanism, it would be expected that performance curves from sand could be used to estimate those for the APM by adjusting for differences in retardation.

Calculations for no water flow and three air flow rates (21, 210, 2100 $cm^3\ s^{-1}$) are shown in Figure 4.2.1.13. The pore volume scale is a factor of four greater than for the sand because the increased water content in the APM results in an increase in retardation from 2.6 for the sand to 8.5 for the APM. The results shown in Figure 4.2.1.13 are similar to those shown in Figure 4.2.1.6 in that the extraction efficiency decreases with decreasing flow rate due to diffusional mixing.

Figure 4.2.1.14 displays calculations for air flow rates of 21,000, and 210,000 $cm^3\ s^{-1}$ and these can be compared to the solid and dashed lines in Figure 4.2.1.7. Because the value of K_La used for the APM calculations is a factor of 5 less than that used in the sand, the air velocity for which extraction becomes volatilization rate limited is less than for the sand. The air-filled porosity for the APM and the sand calculations are the same so the impact of diffusion is the same for air velocities where the air-water mass transfer is unimportant.

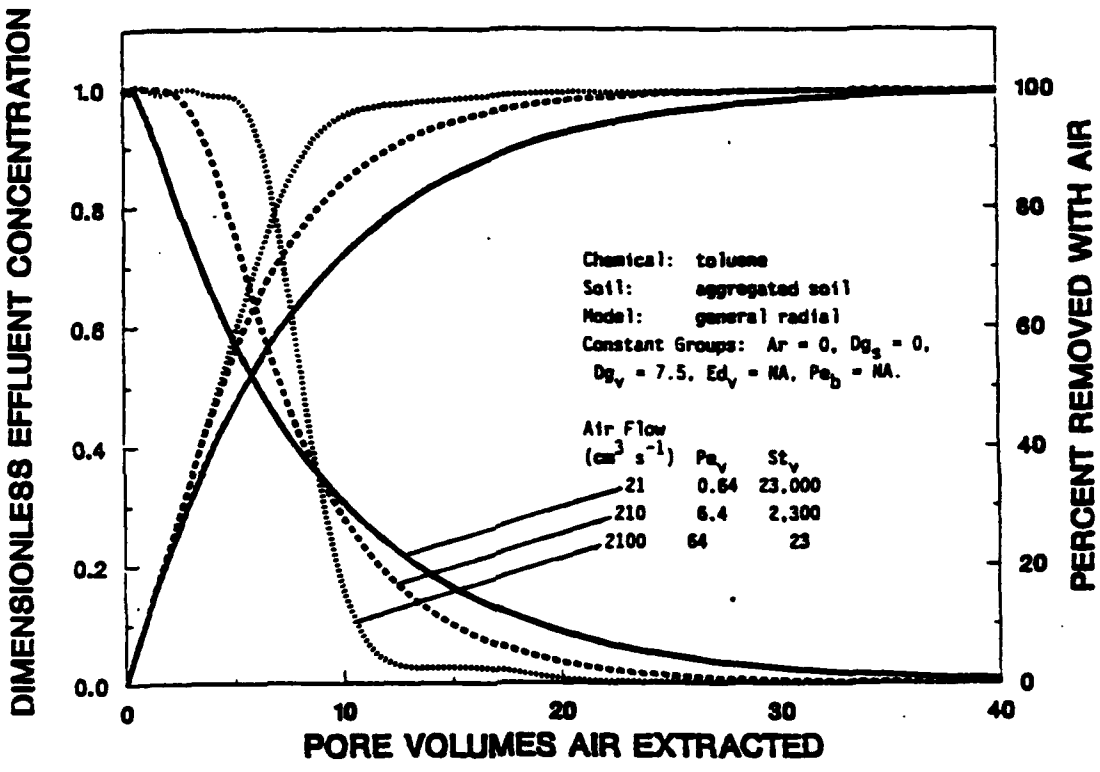


Figure 4.2.1.13. Toluene extraction performance as a function of air withdrawal rate (21, 210, 2100 $cm^3\ s^{-1}$) for a vertical vent configuration in aggregates at a degree of saturation of 0.67.

A summary of extraction performance is shown in Figure 4.2.1.15 as a function of air flow rate for both soil materials. This figure shows the number of air pore volumes required to extract 99.9% of the initial mass of toluene from both soils for six order-of-magnitude changes in air flow rate. These results are also summarized in Table 4.2.1.6 in terms of extraction time, pressure drop, power, and energy. Pressure drop increases proportionally with air flow rate, assuming a constant radius of influence, and power consumption rate increases with flow to the second power. Even though cleanup time decreases as air flow rate increases, total energy consumption increases with air flow rate. Total energy should not be the only basis for design of air flow rate because if gas diffusion flux is of the same order as advective air flux, then

organic vapors can also diffuse away from the site. There should be some balance between cleanup time and energy costs.

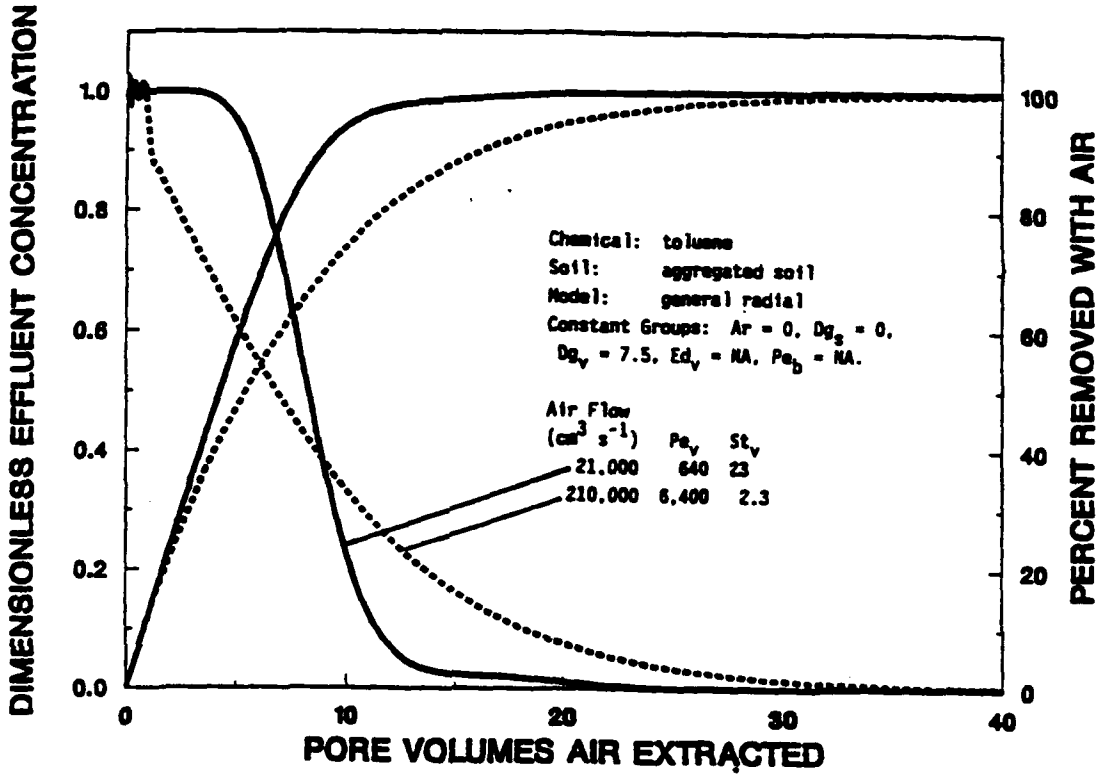


Figure 4.2.1.14. Toluene extraction performance as a function of air withdrawal rate (21,000 and 210,000 cm^3 s^{-1}) for a vertical vent configuration in aggregates at a degree of saturation of 0.67.

Given an air flow rate corresponding to a minimum volume of air required to achieve a specific level of cleanup, the extraction efficiency decreases for flow rates lower than the optimum because of diffusional mixing in the air phase and for higher flow rates because volatilization rates are limiting extraction. In terms of air volume this impact is more pronounced as the amount of retardation is increased. Also shown in Figure 4.2.1.15, as symbols connected by horizontal lines, are the minimum pore volumes of air required for complete cleanup if no diffusional mixing and mass transfer resistances

occur. At best, an extraction system will require at least twice this minimum. The shapes of these performance curves indicate that a range of optimum air velocities exist where the air velocity is fast enough to minimize diffusional mixing while slow enough to allow volatilization. Given some air diffusion rate, the size of the optimal range decreases with $K_L a$.

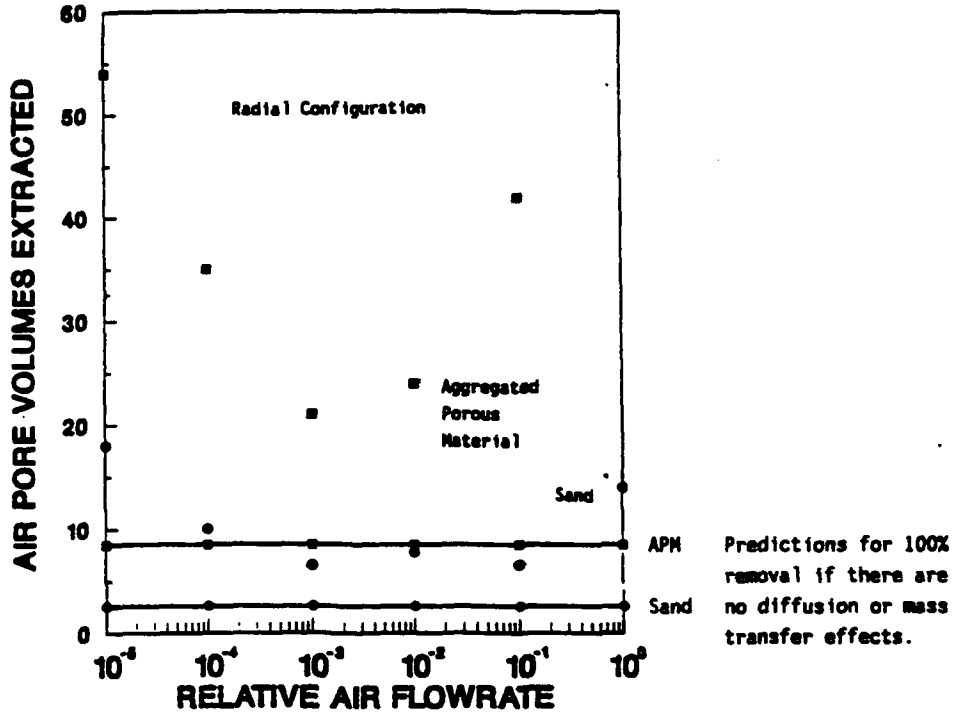


Figure 4.2.1.15. Air volumes required for various air withdrawal rates to remove 99.9% of dissolved toluene contamination from sand at an S of 0.30 and from aggregates at an S of 0.67 assuming no adsorption.

For advection-dispersion controlled removal, the performance for the sand could be used to predict performance for the APM by increasing the time scale of the sand performance curves by a factor of the relative amount of retardation. For example, at an air flow rate of 21,000 $cm^3\ s^{-1}$, 8 air pore volumes are needed to remove the toluene from the sand. The ratio of retardation $[(1 + Dg)\ (1 + Ar)^{-1}]$ in the APM to

that in the sand is 3.3, and so the removal from APM is estimated to be 26 air pore volumes. The model predicted 99.9% removal after 24 pore volumes. This method of estimation is not valid when the air withdrawal rate is fast enough that removal is influenced by the rate of air-water mass transfer. For example, at a flow rate of 210,000 $cm^3\ s^{-1}$, removal from the sand is not limited by volatilization, and, as a result, 99.9% of the toluene could be removed in 7 air pore volumes. Using the ratio of the retardation coefficients as before, the cleanup of the APM would be underestimated as 23 pore volumes where, in fact, due to volatilization rate limitations, it would take almost twice as long.

The infiltration of water in the system containing APM has the same effect as infiltration in the sand. Figure 4.2.1.16 shows that for air flow rates of 21,000 and 210,000 $cm^3\ s^{-1}$, 97% of the contamination in the APM is removed in the air phase, and for an air flow rate of 2100 $cm^3\ s^{-1}$, only 85% of the contamination is removed with air flow. These same removal contributions were also predicted for the sand, however, the difference in air volume required between flow rates of 21,000 and 210,000 $cm^3\ s^{-1}$ is greater in the APM.

For low air flow rates, advection and gas diffusion control the performance of vapor extraction. Extraction performance can diminish if the air withdrawal rate is faster than the phase equilibrium rate. The rate to equilibrium is slower in aggregated soils. If the extraction process is phase mass transfer limited (either by volatilization, diffusion from immobile regions, or desorption kinetics), then increases in air extraction rates will not increase efficiency. On the other hand, the extraction rate should be high enough to minimize the mixing caused by gas diffusion. These types of calculations can help an

engineer decide on a range of air withdrawal rates to specify in the design and operation of a vapor extraction system. In cases where mass transfer rates limit the air flow rate to value such that diffusional mixing occurs, intermittent operation of the extraction process might be better. This type of operation would allow for fast air velocities to sweep out the air-filled voids of the contaminant vapors at rate which minimizes diffusional mixing, and then during flow stoppage, time is allowed for the air and water to reach chemical equilibrium so that the cycle could be repeated (cf. Bednar [1990]).

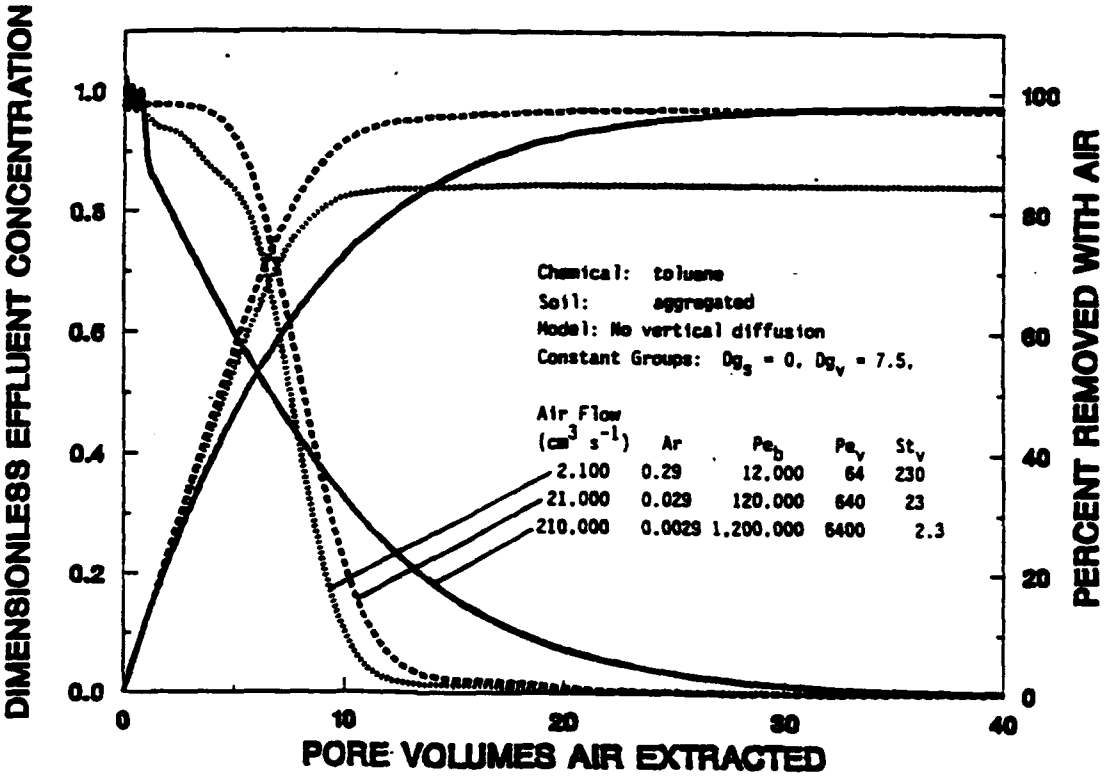


Figure 4.2.1.16. Comparison of toluene extraction performance for radial configuration at air withdrawal rates of 2100, 21,000, and 210,000 $cm^3\ s^{-1}$ and 0.00026 cm s^{-1} water infiltration rate.

4.2.1.8 Summary of the Radial Model Results. A set of models was developed to study the relative impacts of various combinations of

mechanisms on the performance of vapor extraction. The system geometry corresponds to a typical field configuration for a vertical extraction vent completed in a homogeneous soil. The mechanisms considered are believed to have an influence on the removal of dissolved, volatile organic chemicals over a time period short enough so that degradation can be ignored. The impact of these mechanisms were studied in laboratory column experiments, and the results are reported in earlier sections of this document. Aqueous-phase transport was modeled as steady, vertical advective and dispersive movement in water. Vapor-phase transport included horizontal air advection to a radially-converging, vertical axis and horizontal and vertical gas diffusion. Aqueous and vapor concentrations were coupled by an air-water mass transfer mechanism that could account either for volatilization rate limitations or intraaggregate diffusion. Linear sorption from aqueous solution was also considered. The models were used to study the impacts of these mechanisms for a six order-of-magnitude range of air withdrawal rates. Calculations were performed to examine the removal of toluene from a cohesionless sand and from an aggregated soil. Some of the models in this section were solved numerically by orthogonal collocation. Orthogonal collocation was acceptable for solving the column models described in previous sections. The numerical approximations used here were verified by comparing them to analytic solutions for simplified conditions. Two-dimensional air diffusion was not approximated satisfactorily in the general model. Orthogonal collocation was not appropriate for solving the general set of radial equations. The numerical solution of the coupled, one-dimensional transport equations (vertical aqueous-phase movement and horizontal

vapor-phase movement) was verified.

Model calculations indicate that a range of air flow rates exist, for a given soil treatment situation, where the least volume of air is required to reach a specified level of cleanup. More air is required than the minimum to remove chemical for low air flow rates because of diffusional mixing in the vapor phase and for high flow rates because of phase mass transfer limitations. The importance of diffusional mixing is a function of the air-filled porosity and the gas diffusion coefficient. Phase mass transfer impacts are affected by the amount and distribution of water or the degree of soil aggregation. Limitations due to phase mass transfer will be realized in aggregated soils at lower air flow rates than in sands. System operation should be designed for air flow rates large enough to minimize diffusional mixing and low enough to allow phase equilibrium to be reached. The model developed here is useful for estimating this range.

The volume of air required to treat a soil can be directly related to energy requirements given the system pressure drop. Pressure drop in the soil system increases with flow rate to the second power. Total energy consumption calculations also require an estimate of the pressure drop in the piping and the blower efficiency.

Even though vertical vent configurations are the most common in use today, there may be advantages to installing vents horizontally. A planar (trench) system model is developed in the following section. Vertical gas diffusion is ignored in the development of the planar model because it was not approximated satisfactorily in this section.

4.2.2 Planar Geometry

Vapor extraction vents are installed either horizontally or vertically. Horizontal vents are placed in trenches and backfilled with pea gravel and grouted at the surface. One cell of a trench (planar) system configuration, shown graphically in Figure 4.2.2.1, is studied in this section. A horizontal extraction vent is placed between two inlet vents. The ground surface is covered with a low permeability soil, such as compacted clay, or an impermeable cap. For systems with a soil cover, water infiltration at the ground surface is accounted for and assumed to be steady. Likewise, air flows and pressures are assumed to be steady and the pressure drop between vents is assumed to be small.

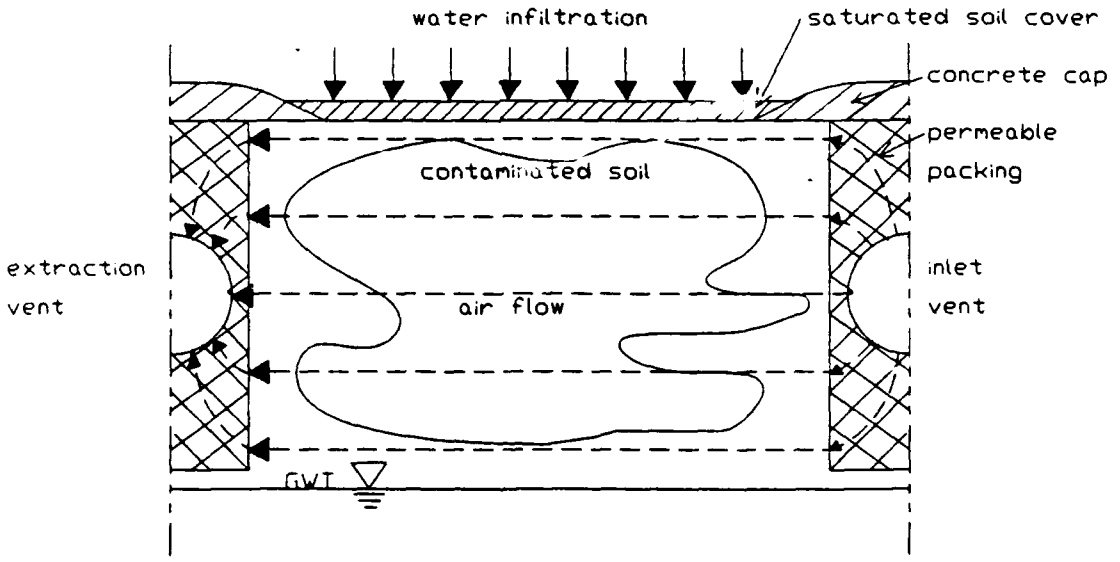


Figure 4.2.2.1. Schematic of one cell of a trench system where air flows horizontally from the inlet vent to the extraction vent.

4.2.2.1 Conceptual Picture. A conceptual picture is derived from the system schematic shown in Figure 4.2.2.1. The impact of adjacent cells is assumed to be additive. The equation derivations for planar geometry are based on the rectangular configuration of unit thickness

shown in Figure 4.2.2.2. Transport equations are derived for downward (Z) water flow and horizontal (X) air flow. Air enters this extraction system at the inlet (X = L) at a uniform rate (and vapor concentration for a closed loop system). The vapors are removed at the extraction vent (X = 0) at the same flowrate (*i.e.* it is assumed that air behaves as an incompressible fluid).

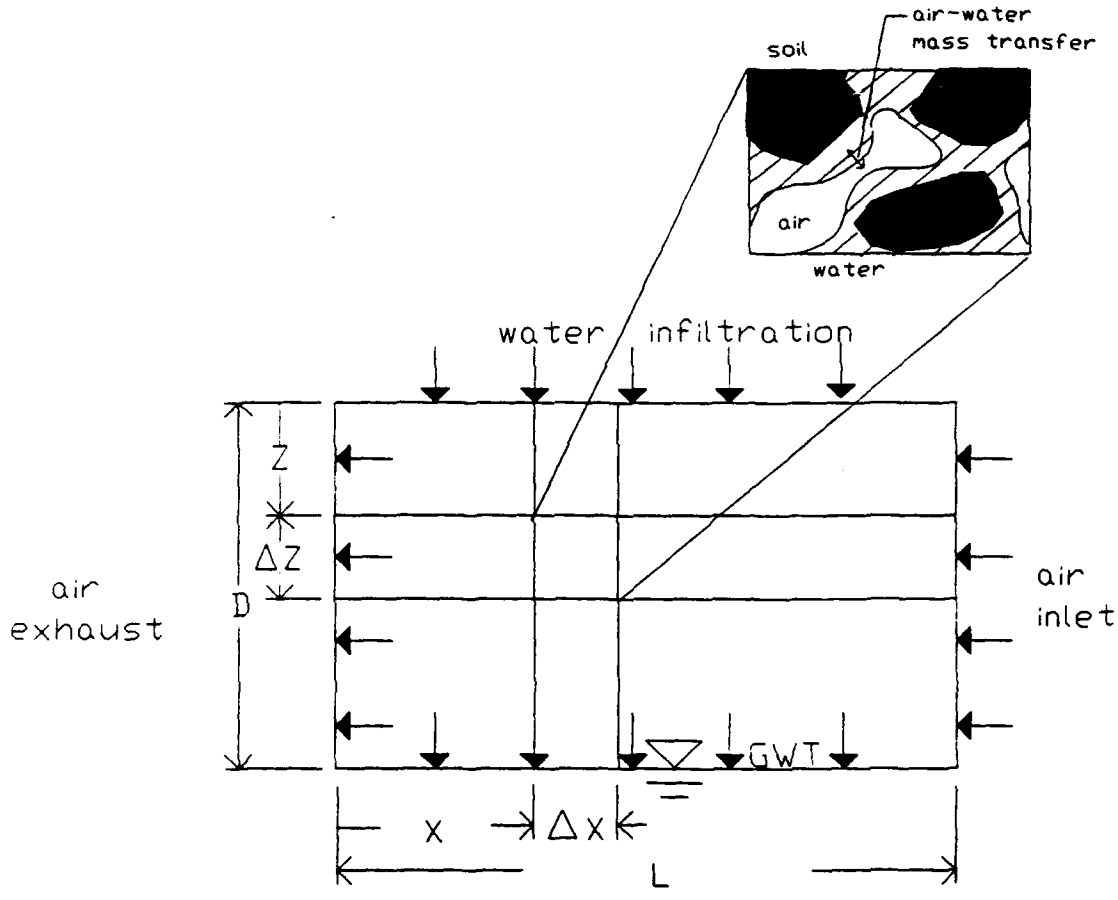


Figure 4.2.2.2. Conceptual picture for the trench or planar configuration, vapor extraction model derivation.

Water containing a volatile chemical at a constant concentration ranging from zero to its solubility limit infiltrates the soil system shown in Figure 4.2.2.2 at a uniform rate through a thin, saturated soil cover (Z = 0). Water flows downward through the soil towards a

horizontal ground water table (Z = D). It is assumed that the soil system is homogeneous, so that porosity and bulk density are constant, and that the capillary fringe is thin in comparison to the system depth, so that the moisture content can be approximated by a uniform profile. For the following equation derivations it is also assumed that all of the water is mobile or that soil aggregation is insignificant. The impact of soil aggregation will be determined from the effects of air-water mass transfer resistance as previously discussed.

4.2.2.2 Equation Derivations. Dimensioned equations are obtained by performing mass balances on the differential soil volumes shown in Figure 4.2.2.2. Water flow occurs in the positive Z direction (downward); air flows in the negative X direction. Flows and concentrations are assumed uniform in the direction perpendicular to the X-Z plane. In addition to mass transport by advection, chemical movement by axial dispersion in water and horizontal dispersion (diffusion) in air are considered. Because the numerical method employed in the solution of the radial models did not satisfactorily approximate vertical air diffusion, it is ignored in the following development. Mass transfer between the air and water is accounted for in the equation derivations, and chemical equilibrium between the water and soil is assumed to be instantaneous.

A mass balance on the air phase gives:

$$\frac{\partial C_v(X,Z,T)}{\partial T} = u \frac{\partial C_v(X,Z,T)}{\partial X} + E_v \frac{\partial^2 C_v(X,Z,T)}{\partial X^2} + K_L a \left[C_b(X,Z,T) - \frac{C_v(X,Z,T)}{H} \right] \qquad (4.2.2.1)$$

The terms in (4.2.2.1) represent: accumulation in the air phase, advection in air, air dispersion, and air-water mass transfer, respectively.

It is assumed that initially the soil system shown in Figure 4.2.2.2 is at equilibrium:

$$C_v(X,Z,T=0) = C_{vi}(X,Z) \qquad (4.2.2.2)$$

The boundary condition for (4.2.2.1) at the extraction vent ($X=0, 0 \leq Z \leq D$) is derived by assuming that the system approximates a closed reactor [Levenspiel, 1962]:

$$\frac{\partial^2 C_v(X=0,0\leq Z\leq D,T)}{\partial X \partial T} = 0 \qquad (4.2.2.3)$$

The boundary condition for (4.2.2.1) at the inlet ($X=L, 0 \leq Z \leq D$) is obtained by performing a mass balance on the air phase in the volume shown in Figure 4.2.2.2 transecting from X equal to 0 to L:

$$\frac{\partial}{\partial T}\int_0^L C_v(X,Z,T)\partial X = \int_0^L K_L a \left[C_b(X,Z,T) - \frac{C_v(X,Z,T)}{H} \right] \partial X + u[C_{vo}(T) - C_v(X=0,Z,T)] \qquad (4.2.2.4)$$

A mass balance on the water-soil phase is given by:

$$\left[1 + \frac{\rho_s(1-\epsilon)K}{\epsilon S} \right] \frac{\partial C_b(X,Z,T)}{\partial T} = E_z \frac{\partial^2 C_b(X,Z,T)}{\partial Z^2} - v \frac{\partial C_b(X,Z,T)}{\partial Z} - \frac{K_L a(1-S)}{S} \left[C_b(X,Z,T) - \frac{C_v(X,Z,T)}{H} \right] \qquad (4.2.2.5)$$

Like (4.2.2.1), the terms in (4.2.2.5) account for accumulation, liquid

dispersion, advection in water, and air-water mass transfer. For this development it is assumed that the aqueous- and sorbed-phase concentrations are in equilibrium. Furthermore, this equilibrium is assumed to be linear:

$$Q(X,Z,T) = K\ C_b(X,Z,T) \qquad (4.2.2.6)$$

Because in most soils, moisture will probably cover the soil particles, sorption of vapors by soil is not considered here. In situations where there is no water flow and air-water mass transfer is fast, then (4.2.2.6) can be used to determine the impact of linear vapor sorption.

It is assumed that initially the soil system shown in Figure 4.2.2.2 is at equilibrium:

$$C_b(X,Z,T=0) = C_{bi}(X,Z) = C_{vi}(X,Z) \ / \ H \qquad (4.2.2.7)$$

Boundary conditions for (4.2.2.5) are obtained by assuming that the soil system acts as a closed reactor [Levenspiel, 1962]. The Danckwerts [1953] exit condition is:

$$\frac{\partial^2 C_b(0 \le X \le L, Z=D, T)}{\partial Z \partial T} = 0 \qquad (4.2.2.8)$$

The description of aqueous-phase transport to ground water assumes that water velocities are slow enough that changes in the depth to ground water are negligible. In addition, the mass flux of chemical away from the water table is equal to the mass flux to the aquifer.

An entrance boundary condition for (4.2.2.5) is obtained by performing a mass balance on the water-soil phases in the volume shown in Figure 4.2.2.2 transecting from Z equal to 0 to D:

$$\left[1 + \frac{\rho_s(1-\epsilon)K}{\epsilon S} \right] \frac{\partial}{\partial T} \int_0^D C_b(X,Z,T)\partial Z = v[C_{bo}(T) - C_b(X,Z=D,T)]$$

$$- \int_0^D \frac{K_L a(1-S)}{S} \left[C_b(X,Z,T) - \frac{C_v(X,Z,T)}{H} \right] \partial Z \qquad (4.2.2.9)$$

4.2.2.3 Conversion to Dimensionless Form. The dimensioned equations derived above are converted to a dimensionless form by making substitutions of the middle column of Table 4.2.2.1. The normalization step is the same as the radial geometry derivation in Section 4.2.2.3.

Table 4.2.2.1. Variable Substitutions to Convert Dimensioned Equations into a Dimensionless Form.

Dimensioned Variable	Substitution	Dimensionless Variable
$C_b(X,Z,T)$	$C_{bn}\ c_b(X,Z,T)$	$c_b(X,Z,T)$
$C_v(X,Z,T)$	$C_{vn}\ c_v(X,Z,T)$	$c_v(X,Z,T)$
$C_{bi}(X,Z)$	$C_{bn}\ c_{bi}(X,Z)$	$c_{bi}(X,Z)$
$C_{vi}(X,Z)$	$C_{vn}\ c_{vi}(X,Z)$	$c_{vi}(X,Z)$
$C_{bo}(T)$	$C_{bn}\ c_{bo}(T)$	$c_{bo}(T)$
$C_{vo}(T)$	$C_{vn}\ c_{vo}(T)$	$c_{vo}(T)$
X	$L\ x$	x
Z	$D\ z$	z
T	$\frac{u}{L}\left[1 + \frac{LvS}{Du(1-S)H} \right]\left[1 + \frac{S}{(1-S)H} + \frac{\rho_s(1-\epsilon)K}{\epsilon(1-S)H} \right]^{-1} t$	t

Note: $C_{bn} = C_{vn} / H$ where $C_{vn} = \max(C_{vo}(T), C_{vi}(X,Z))$.

Substitutions of the dimensionless variables resulted in natural groupings of the dimensioned parameters. The resulting groups are

dimensionless and are similar to those derived for the one-dimensional column models and the radial configuration model. Mass transfer rates are compared to the advective rates in air, and chemical equilibrium distributions are normalized by the mass of chemical in air. The dimensionless groups that characterize the solutions of the planar models are given in Table 4.2.2.2.

Table 4.2.2.2. Dimensionless Groups Representing Mass Transfer Rates and Chemical Equilibrium Distributions.

Dimensionless Group		Equation	Ratio of	Ratio to
Ar	=	$\frac{LvDg_v}{Du}$	Advective rate in water	Advective rate in air
Pe_b	=	$\frac{D^2u}{E_zLDg_v}$	Advective rate in air	Dispersive rate in water
Pe_v	=	$\frac{uL}{E_v}$	Advective rate in air	Horizontal dispersive rate in air
St_v	=	$\frac{K_LaL}{uH}$	Volatilization rate	Advective rate in air
Dg	=	$Dg_S + Dg_v$	Mass in water and on soil	Mass in air
Dg_S	=	$\frac{\rho_S(1-\epsilon)K}{\epsilon(1-S)H}$	Mass adsorbed to soil	Mass in air
Dg_v	=	$\frac{S}{(1-S)H}$	Mass in water	Mass in air

The dimensionless forms of (4.2.2.1-9) are:

$$\frac{\partial c_v(x,z,t)}{\partial t} = \frac{1 + Dg}{1 + Ar}\left[\frac{\partial c_v(x,z,t)}{\partial x} + \frac{1}{Pe_v}\frac{\partial^2 c_v(x,z,t)}{\partial x^2} + St_v[c_b(x,z,t) - c_v(x,z,t)]\right] \qquad (4.2.2.10)$$

$$c_v(x,z,t=0) = c_{vi}(x,z) \qquad (4.2.2.11)$$

$$\frac{\partial^2 c_v(x=0,0\leq z\leq 1,t)}{\partial x \partial t} = 0 \qquad (4.2.2.12)$$

$$\frac{\partial}{\partial t}\int_0^1 c_v(x,z,t)\partial x = \frac{1 + Dg}{1 + Ar}\left[\int_0^1 St_v[c_b(x,z,t) - c_v(x,z,t)]\partial x + [c_{vo}(t) - c_v(x=0,z,t)]\right] \qquad (4.2.2.13)$$

$$\frac{\partial c_b(x,z,t)}{\partial t} = \frac{1 + Dg}{Dg[1 + Ar]}\left[\frac{1}{Pe_b}\frac{\partial^2 c_b(x,z,t)}{\partial z^2} - Ar\frac{\partial c_b(x,z,t)}{\partial z} - St_v[c_b(x,z,t) - c_v(x,z,t)]\right] \qquad (4.2.2.14)$$

$$c_b(x,z,t=0) = c_{bi}(x,z) = c_{vi}(x,z) \qquad (4.2.2.15)$$

$$\frac{\partial^2 c_b(0\leq x\leq 1,z=1,t)}{\partial z \partial t} = 0 \qquad (4.2.2.16)$$

$$\frac{\partial}{\partial t}\int_0^1 c_b(x,z,t)\partial z = \frac{1 + Dg}{Dg[1 + Ar]}\left[Ar[c_{bo}(t) - c_b(x,z=1,t)] - \int_0^1 St_v[c_b(x,z,t) - c_v(x,z,t)]\partial z\right] \qquad (4.2.2.17)$$

Compare the dimensionless equations given above to those in the

transformed coordinate that comprise the radial configuration model where vertical air diffusion is ignored (equations (4.2.1.43, 23, 12, 44, 28-31), respectively). The two models are identical with the exception of the coefficient of the first spatial derivative in (4.2.2.10) and (4.2.1.43). The coefficient of this term in (4.2.2.10) is 1 and it is $[1 + 1/Pe_v]$ in (4.2.1.43). Therefore the planar and radial models should give similar results for large Pe_v.

4.2.2.4 Numerical Solution. Orthogonal collocation (OC) is the numerical method used to convert the coupled partial differential equations (PDEs) given above to a system of ordinary differential equations (ODEs) in time. Details of the method are given elsewhere [Finlayson, 1980]. Steps for converting these types of PDEs to ODEs are documented in the Appendix B of Gierke [1986].

Initial conditions for the models are given by (4.2.2.11) and (4.2.2.15). They become the following as a result of the application of OC:

$$c_v(i,j,t=0) = c_{vi}(x_i,z_k) \quad \text{for } i = 1 \text{ to NL and } j = 1 \text{ to ND} \qquad (4.2.2.18)$$

$$c_b(i,j,t=0) = c_{bi}(x_i,z_j) \quad \text{for } i = 1 \text{ to NL and } j = 1 \text{ to ND} \qquad (4.2.2.19)$$

Because vertical air diffusion is not considered in the planar system model, only the aqueous-phase transport equations contain partial derivatives with respect to z, so the air-phase transport equations can be evaluated at the vertical OC collocation locations for the aqueous-phase equations. This coupling of the equations is shown in Figure 4.2.2.3 and it is identical to the coupling for the no vertical diffusion model for radial geometry. Hence for this model and its radial counterpart, no distinction between vertical OC locations for

air- and water-phase transport approximations is made. As a result, the OC coefficient matrices for approximating horizontal and vertical derivatives are constructed in the same manner and will be the identical if NL = ND.

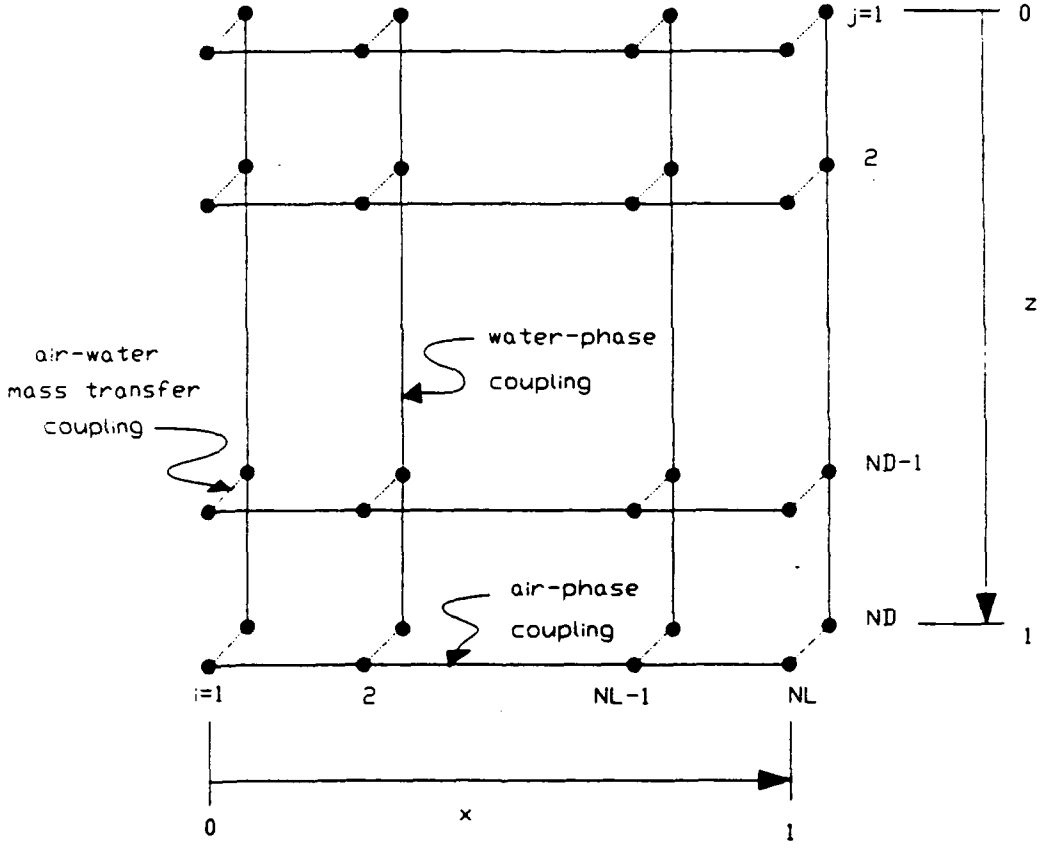


Figure 4.2.2.3. General schematic showing the coupling of the ordinary differential equations resulting from the application of orthogonal collocation to the partial differential equations comprising the trench-configuration vapor extraction model.

The application of OC to the air-phase mass balance (equation (4.2.2.10)) results in the following (NL-2)(ND-1) ODEs:

$$\frac{dc_v(i,j,t)}{dt} = \frac{1 + Dg}{1 + Ar}\left\{ \sum_{n=1}^{NL} \left[\frac{B^x_{i,n}}{Pe_v} + A^x_{i,n} \right] c_v(n,j,t) + St_v[c_b(i,j,t) - c_v(i,j,t)] \right\} \quad \text{for } i = 2 \text{ to } NL-1 \text{ and } j = 1 \text{ to } ND-1 \qquad (4.2.2.20)$$

The air-phase concentrations at the extraction vent come from the application of OC to (4.2.2.12):

$$\frac{dc_v(1,j,t)}{dt} = - \sum_{n=2}^{NL} \frac{A^x_{1,n}}{A^x_{1,1}} \frac{dc_v(n,j,t)}{dt} \quad \text{for } j = 1 \text{ to } ND \qquad (4.2.2.21)$$

Changes in the air-phase concentration at the inlet vent are described by

$$\frac{dc_v(NL,j,t)}{dt} = \left\{ \frac{1 + Dg}{1 + Ar} \left[\sum_{n=1}^{NL} W^x_n St_v[c_b(n,j,t) - c_v(n,j,t)] + [c_{vo}(t) - c_v(1,j,t)] \right] - \sum_{n=2}^{NL-1} \left[W^x_n - \frac{A^x_{1,n}}{A^x_{1,1}} W^x_1 \right] \frac{dc_v(n,j,t)}{dt} \right\} \left[W^x_{NL} - \frac{A^x_{1,NL}}{A^x_{1,1}} W^x_1 \right]^{-1} \quad \text{for } j = 1 \text{ to } ND \qquad (4.2.2.22)$$

Equation (4.2.2.22) is obtained by applying OC to (4.2.2.13) substituting (4.2.2.21) into the result and solving for the time-derivative of $c_v(x,z,t)$ at i equal to NL.

The water-soil mass balance (equation (4.2.2.14)) becomes the following after application of OC:

$$\frac{dc_b(i,j,t)}{dt} = \frac{1 + Dg}{Dg[1 + Ar]} \left\{ \sum_{m=1}^{ND} \left[\frac{B^z_{j,m}}{Pe_b} - A^z_{j,m} \right] Ar\ c_b(i,m,t) \right.$$

$$\left. - St_v[c_b(i,j,t) - c_v(i,j,t)] \right\} \text{ for } i = 1 \text{ to } NL$$

$$\text{and } j = 2 \text{ to } ND-1 \qquad (4.2.2.23)$$

The boundary condition at the ground water table for (4.2.2.14) (equation (4.2.2.16)) becomes the following after application of OC:

$$\frac{dc_b(i,ND,t)}{dt} = - \sum_{m=1}^{ND-1} \frac{A^z_{ND,m}}{A^z_{ND,ND}} \frac{dc_b(i,m,t)}{dt} \quad \text{for } i = 1 \text{ to } NL \qquad (4.2.2.24)$$

Application of OC to (4.2.2.17) creates the following after substitution of (4.2.2.24) and rearrangement to solve for the time-derivative of $c_b(x,z,t)$ at z equal to 0:

$$\frac{dc_b(i,1,t)}{dt} = \left\{ \frac{1 + Dg}{Dg[1 + Ar]} \left[Ar\ [c_{bo}(t) - c_b(i,ND,t)] \right.\right.$$

$$\left. - \sum_{m=1}^{ND} W^z_m St_v[c_b(i,m,t) - c_v(i,m,t)] \right] - \sum_{m=2}^{ND-1} \left[W^z_m \right.$$

$$\left.\left. - \frac{A^z_{ND,m}}{A^z_{ND,ND}} W^z_{ND} \right] \frac{dc_b(i,m,t)}{dt} \right\} \left[W^z_1 - \frac{A^z_{ND,1}}{A^z_{ND,ND}} W^z_{ND} \right]^{-1}$$

$$\text{for } i = 1 \text{ to } NL \qquad (4.2.2.25)$$

The system of ODEs given above (equations (4.2.2.18-25)) are evaluated in the order: (4.2.2.23), (4.2.2.25), (4.2.2.24), (4.2.2.20), (4.2.2.22), and (4.2.2.21). Initially, (4.2.2.18) and (4.2.2.19) are used in the determination of the derivatives. An International Mathematics and Statistics Library package for solving sets of ODEs, called GEAR, is used to determine concentration values at times greater

than 0 at the collocation points.

4.2.2.5 Model Verification. The numerical approximation of vertical water advection and dispersion was verified in Section 4.1.1.6 and air advection and dispersion in this planar model is approximated in the same manner. The numerical approximation of air-water mass transfer is the same that is used for the radial flow models, which was verified in Section 4.2.1.6. Like the radial flow models, no analytic solutions exist for the coupled differential equations so the previous piecewise verification approach reported in Section 4.2.1.6 is assumed to apply here.

4.2.2.6 Numerical Results (Model Sensitivity). The primary purpose of developing a planar model is to compare treatment efficiency of radial systems with the less common trench configuration. Even though comparing the governing equations indicates that the performance will be similar for large values of Pe_v, there is a significant difference in pressure drop between the two configurations. Pressure loss resulting from radial flow is approximated with (4.2.1.100), where for planar steady, incompressible air flow ΔP is given by

$$\Delta P = u\epsilon(1-S)L/K_a \qquad (4.2.2.26)$$

It will be shown later that (4.2.2.26) predicts pressure drops that are about one-half of those estimated by (4.2.1.100) for equivalent air detention times (air-filled porosity / air flow rate). Similar results are obtained when the compressibility of air is accounted for. The pressure relationship for steady, compressible, planar air flow is given by Wilson *et al.* [1988] and for radial by Johnson *et al.* [1990b].

Parameter Estimation. Parameter values for the model calculations are, again, based on properties of the Ottawa sand and aggregated porous media used in the column studies. Furthermore, system size is specified to be similar to the trench network at Hill AFB, Utah [DePaoli *et al.*, 1990], and match the radial configuration studied in Section 4.2.1. Values used in the calculations of the two-dimensional planar models are given in Table 4.2.2.3. Air velocities are chosen here to give air detention times equivalent to those for the air flow rates used in Section 4 2.1.7. The mass transfer parameters are identical to those in Table 4.2.1.5.

Table 4.2.2.3. Parameter Values for Model Calculations of Toluene Removal from Ottawa Sand (OS) and Aggregated Porous Material (APM).

Parameter Values	OS	APM
Dist. from extraction to inlet vent, L (m):	4.5	4.5
Depth to ground water, D (m):	4.5	4.5
System width (m):	14.1	14.1
Infiltration rate, $v\epsilon S$ (cm s^{-1}):	0, 0.00026	0, 0.00026
Air withdrawal rate, $u\epsilon(1-S)$ (cm s^{-1}):	0.000033-3.3	0.000033-0.33
Air conductivity, K_a (cm s^{-1}):	0.018	0.018
Degree of saturation, S:	0.30	0.67
Porosity, ϵ:	0.33	0.70
Soil density, ρ_s (g cm^{-3}):	2.65	1.51
Henry's constant, H:	0.27	0.27
Air dispersion coef., E_v (cm^2 s^{-1}):	0.050	0.050
Liquid dispersion coef., E_z (cm^2 s^{-1}):	0.0033	0.0033
Air-water mass transfer rate, K_La (s^{-1}):	0.010	0.0020
Sorption capacity, K (cm^3 g^{-1}):	0	0
Initial contaminant concentration		
in air, C_{vi} (mg L^{-1}):	1.0	1.0
in water, C_{bi} (mg L^{-1}):	3.7	3.7
on soil, Q_i (mg kg^{-1}):	0	0
Influent contaminant concentration		
in air, C_{vo} (mg L^{-1}):	0	0
in water, C_{bo} (mg L^{-1}):	0	0

Model Calculations for Toluene Removal from Sand

The first planar model simulations assume a capped system where no water is flowing. The vapor extraction column model could be used to describe a capped, planar system if the initial concentration profile does not vary with depth. Even though the following simulations assume a uniform initial concentration, the two-dimensional model is used to estimate performance because the results of these calculations will be compared to situations where water infiltration occurs. For the same reasons as in the radial case, the planar model could not solve nonuniform initial concentration profiles in general.

It was observed that for large values of Pe_v, the dimensionless equations describing the performance of a planar system (equations (4.2.2.10-17)) are identical to the dimensionless transport equations for a radial system in transformed coordinates (equations (4.2.1.43, 23, 12, 44, 28-31)). Planar model calculations for no water flow are almost identical to the radial model calculations for air flow rates greater than 2100 $cm^3\ s^{-1}$ ($u = 0.014\ cm\ s^{-1}$ for planar). Comparisons of planar and radial performance are shown in Figure 4.2.2.4 for flow rates of 2100, 21,000, 210,000, and 2,100,000 $cm^3\ s^{-1}$. A summary of the planar performance is given in Table 4.2.2.4 (radial performance is summarized in Table 4.2.1.6). Even though the volume of air required to treat the soil is equivalent for both configurations, the power and energy requirements for the planar system are about one-half of that for the radial system. The pressure drop calculations assume an air conductivity of 0.018 $cm\ s^{-1}$ and do not consider pressure losses in the trench or piping nor blower efficiency.

Table 4.2.2.4. Summary of Extraction Performance for Removing 99.9% of Toluene from a Cohesionless Soil and from an Aggregated Soil with Various Air Flow Rates in a Capped, Trench System. (Calculations Assume No Water Flow and No Sorption)

			Ottawa Sand				Aggregated Porous Material			
u ($cm\ s^{-1}$)	ΔP (cm air)	Power (Watts)	Time (hours)	t	apv	Energy (kJoules)	Time (hours)	t	apv	Energy (kJoules)
0.00014	0.8	$2(10^{-6})$	$14(10^3)$	6.1	16	0.11	$46(10^3)$	6.1	52	0.33
0.0014	8.0	$2(10^{-4})$	600	2.7	7	0.43	1,900	2.5	22	1.4
0.014	80	0.02	51	2.3	6	3.7	190	2.5	21	14
0.14	800	2	6.8	3.0	8	49	21	2.9	24	151
1.4	8000	200	0.6	2.6	7	430	3.8	5.0	43	2700
14	$8(10^4)$	$2(10^4)$	0.1	5.7	15	7200				

apv = air pore volumes, t = throughput.

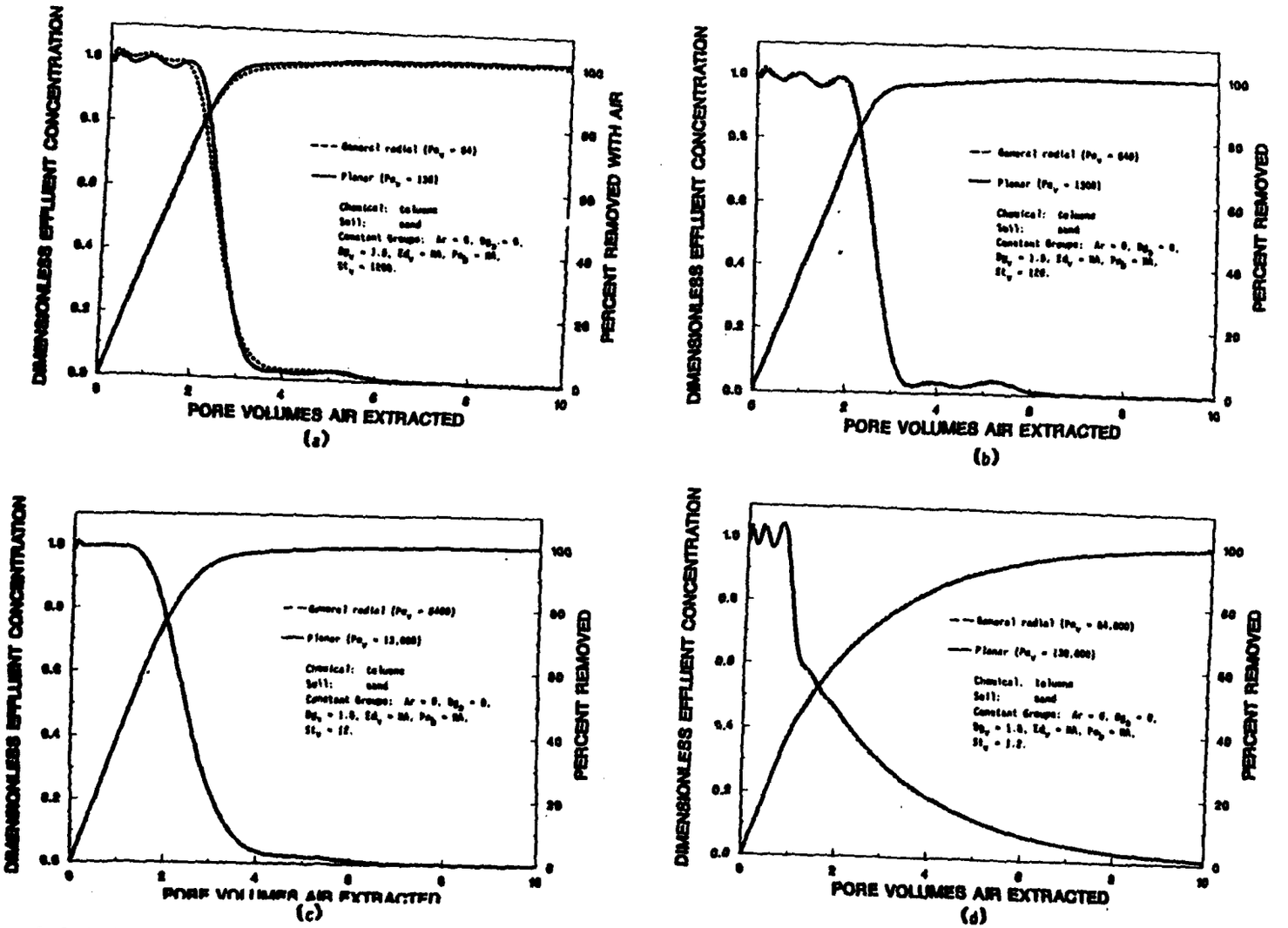


Figure 4.2.2.4. Comparison of radial and trench configuration on the toluene extraction performance in sand for air flow rates of: (a) 2100, (b) 21,000, (c) 210,000, and (d) 2,100,000 $cm^3\ s^{-1}$.

There is little difference in performance between radial and planar configurations for complete mixing conditions (low flow, Pe_v less than 2 [Brenner, 1962]) as shown in Figure 4.2.2.5.

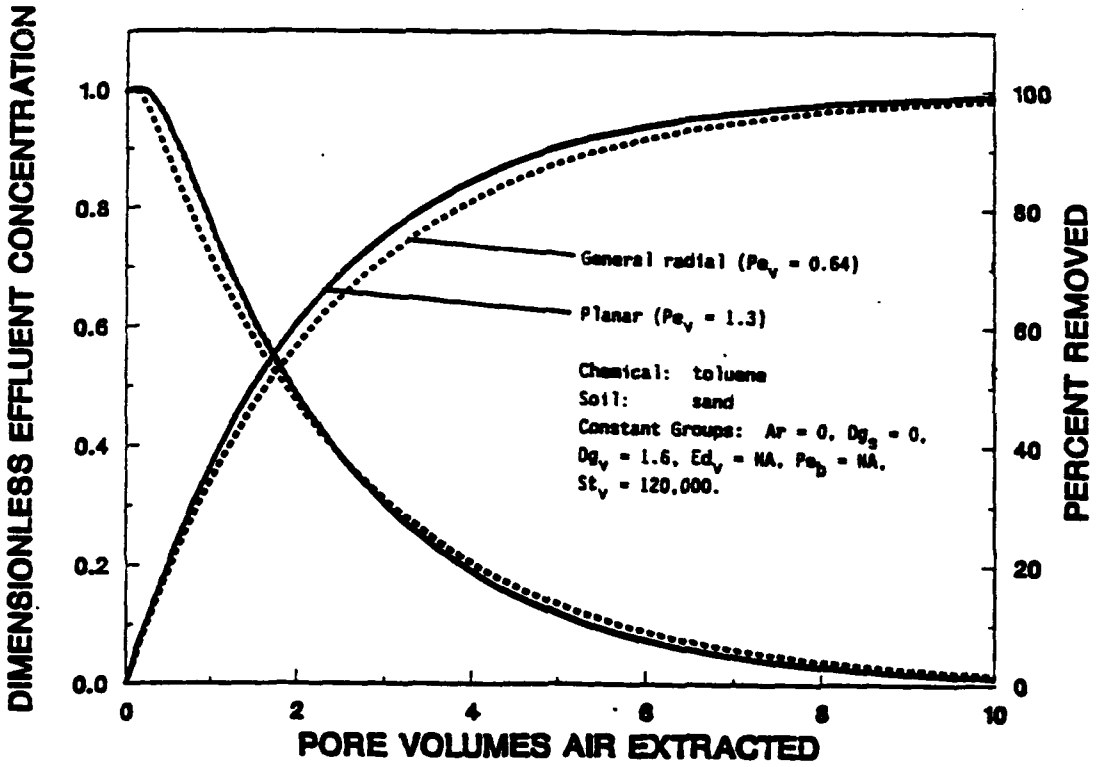


Figure 4.2.2.5. Comparison of radial and trench configuration on the toluene extraction performance in sand for diffusion dominant conditions (*i.e.* low air flow rate of 21 $cm^3\ s^{-1}$).

The largest difference occurs for values of the radial Pe_v between 2 and 10. For example, Figure 4.2.2.6 compares calculations for the two systems for a flow rate of 210 $cm^3\ s^{-1}$. The planar system can treat the soil with about one-third less air, which results in an energy savings of 67% even though the savings in power usage is only 50%. The radial flow system is less efficient in terms of air volume extracted because the air velocity near the air inlet ($r \rightarrow 1$) is slow enough to allow diffusional mixing. In a trench system, air velocity is uniform.

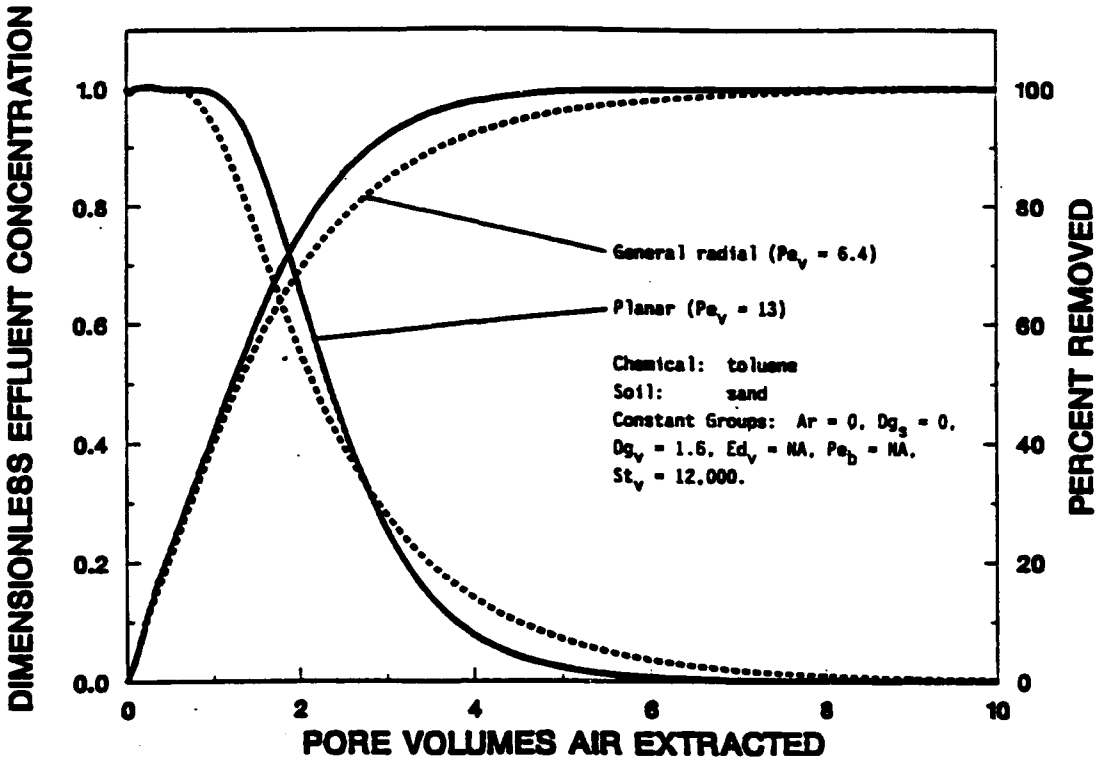


Figure 4.2.2.6. Comparison of radial and trench configuration on the toluene extraction performance in sand for an air flow of 210 cm^3 s^{-1}.

Because of the similarities between the radial and planar equations for values of Pe_v greater than 10, the results for simultaneous water infiltration and vapor extraction are identical for the conditions shown in Figure 4.2.2.7. The case where there would be a difference (an air flow rate of 210 cm^3 s^{-1}) is not interesting because the advective mass flux rate in water would be greater than that in air.

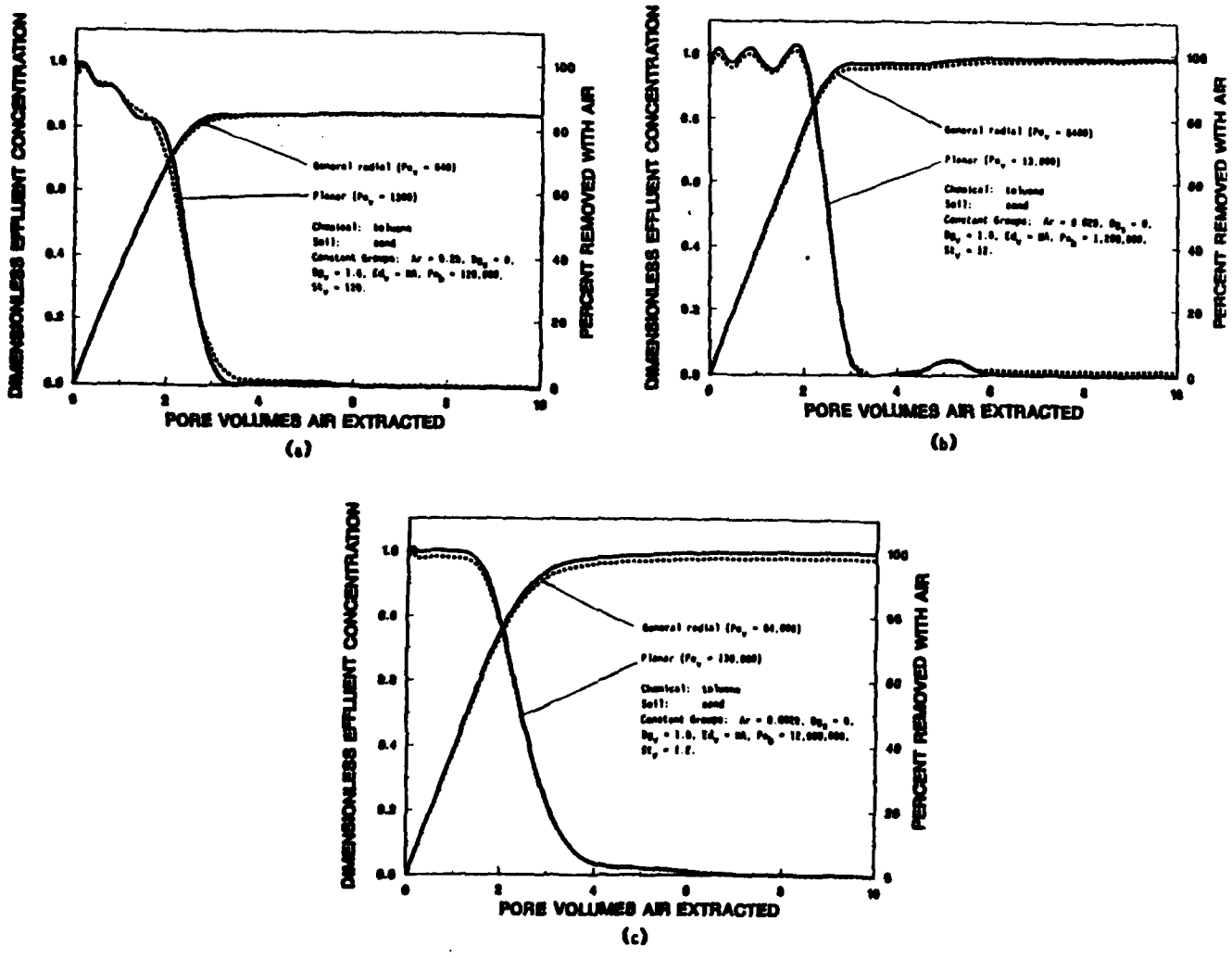


Figure 4.2.2.7. Comparison of radial and trench configuration on the toluene extraction performance in sand for water infiltration at a rate of 0.00026 cm s^{-1} and air flow rates of: (a) 2100, (b) 21,000, and (c) 210,000 cm^3 s^{-1}.

Performance calculations given in Table 4.2.2.4 show that for the value of $K_L a$ chosen for the sand, volatilization-limited removal is not important unless the air velocity is greater than 1.4 cm s^{-1}. In a controlled vapor extraction experiment with uniform, planar air flow through a 1 m^3 box, McClellan and Gillham [1990] observed an impact of volatilization rate on the removal of dissolved and sorbed trichloroethene (TCE) from Borden sand for air velocity as low as 0.2 cm s^{-1}. The impact was more pronounced when the air velocity was 0.7 cm s^{-1}, which was the highest velocity they used. Volatilization rate was not important during the stage of removal where pure liquid TCE was being extracted with air flow.

Model Calculations for Toluene Removal from APM

As in the previous results, any difference between radial and planar performance for toluene removal from the aggregated porous material (APM) is observed only for values of Pe_v between 2 and 20. For example, Figure 4.2.2.8 shows planar and radial performance for an air flow rate of 210 cm^3 s^{-1}. This is similar to the results shown in Figure 4.2.2.6. A summary of the performance for a trench system used to clean a contaminated aggregated soil is summarized in Table 4.2.2.4.

Figure 4.2.2.9 is a graphical representation of the performance summary comparing toluene removal from sand with removal from APM. The symbols connected with lines are the corresponding minimum removal volumes. Extraction limited by phase equilibrium in the APM occurs at air velocities greater than 1 cm s^{-1}, which is comparable to the observations of McClellan and Gillham [1990].

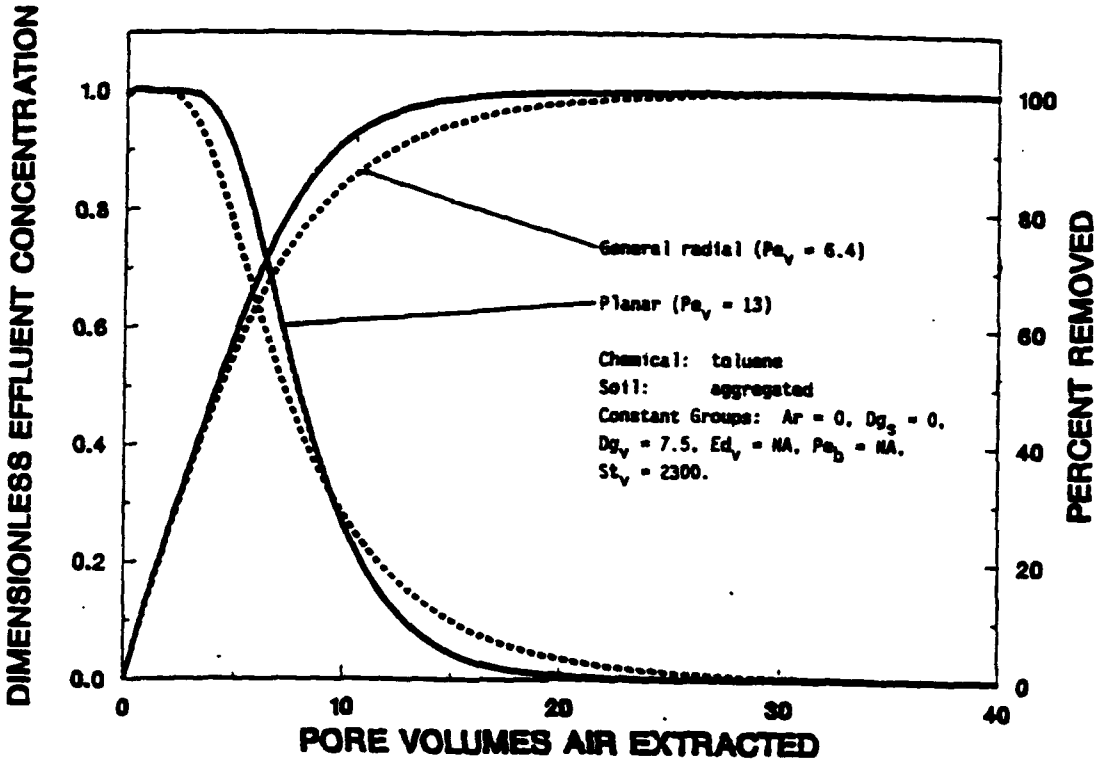


Figure 4.2.2.8. Comparison of radial and trench configuration on the toluene extraction performance in aggregates for an air flow of 210 cm^3 s^{-1}.

Comparing the planar-configuration performance summary in Figure 4.2.2.9 with the summary for radial configuration in Figure 4.2.1.15, it is observed that the range of air flow rates for optimal removal is larger for planar configuration. In terms of removal efficiency, planar systems are equal to or better than radial, and in terms of power and energy, planar requires half of that for radial. Therefore, if the site conditions are such that trenches can be used, a trench system would be more efficient to operate. Limitations for using trenches include depth of contamination, because they must be installed with a backhoe instead of a drill rig, and site surface conditions, such as the location of buildings or other barriers.

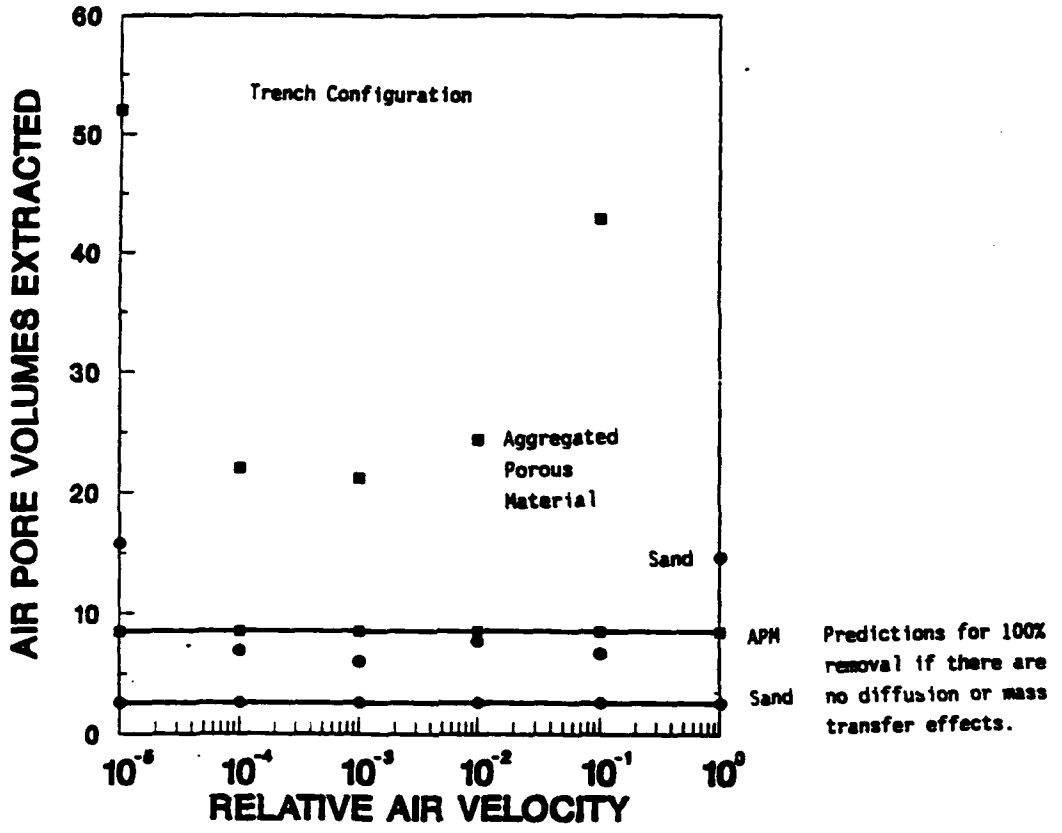


Figure 4.2.2.9. Volume of air required to remove 99.9% of toluene from sand and from aggregates for various air velocities in a trench system.

4.2.2.7 Summary of Planar Model Results. A model for planar configuration was developed to compare its performance to a radial configuration. For most air flow rates, planar and radial configurations treat soils at similar rates. There is a narrow range of air flow rates where planar geometry is more efficient, in terms of air volume extracted, than radial. This range occurs near the lower limit of the optimum range for planar geometry. The optimum operating range for planar geometry is larger than radial. Pressure drops associated with radially converging flow are at least twice that for uniform planar flow. Therefore a trench system is probably a better configuration in terms of operating flexibility and energy requirements.

SECTION 5. SUMMARY AND CONCLUSIONS

This modeling effort focused on mathematically describing organic chemical movement through unsaturated soil with air and water flow and volatile chemical removal from the vadose zone by vapor extraction. Column models were used to gain an understanding of the important mechanisms to consider for modeling chemical fate in the subsurface. It is not always possible to distinguish the impacts of various mechanisms from model sensitivity calculations. Furthermore, experimental results alone do not always provide sufficient information to ascertain which mechanisms are important. An integrated model development, numerical verification, and experimental validation approach leads to a better understanding of subsurface chemical transport. The understanding that was gained from the results of the laboratory column experiments in conjunction with numerical calculations was incorporated into models that could be used to simulate the performance of radial and trench vapor extraction configurations.

5.1 SUMMARY OF COLUMN RESULTS

The column models developed here accounted for the transport of dissolved, nondegradable organic chemicals in unsaturated soil columns. The following mechanisms were included in the model development: advection and dispersion in air and in water, air-water mass transfer, mobile-immobile water mass transfer, intraaggregate diffusion, and sorption. Two types of laboratory experiments were performed. Chemical movement with steady water flow was monitored in unsaturated soil columns to assess the validity of a column model that included all of the mechanisms described above. Then vapor movement experiments were performed in the same soils for both dry and moist conditions to assess

the validity of a vapor extraction column model.

Experimental and modeling results indicate the following about the mechanisms affecting the transport of organic compounds with water flow in unsaturated soil columns: (1) liquid dispersion and intraaggregate diffusion are important for cohesionless and aggregated soils, (2) vapor diffusion is not important in comparison to liquid advection and dispersion except for low water velocities (*e.g.* for Ottawa sand, low velocity is defined as average pore water velocities less than 0.0003 cm s^{-1}), and (3) interfacial mass transfer rates (air-water and mobile-immobile water) are fast. These results were observed for average pore water velocities less than 0.07 cm s^{-1} in Ottawa sand and 0.02 cm s^{-1} in an aggregated material. Nonvolatile tracer experiments were sometimes needed to measure transport rates such as liquid dispersion and intraaggregate diffusion. Intraaggregate diffusion rates were successfully measured for the aggregated material in independent batch experiments; however, diffusion rates in immobile water and the amount of immobile water could not be predicted independently for the sand column. Discrepancies occurred between predicted air-water-soil chemical equilibrium for both soil materials. It is uncertain whether this was due to chemical interactions with the column apparatus.

The vapor extraction column results indicate two important differences between the extraction of chemicals from cohesionless soils and from aggregated soils. First, the moisture content of an aggregated soil is usually higher and this results in an increase in retardation due to partitioning between the air and water. Second, the presence of moisture in an aggregated soil increases extraction time because of intraaggregate diffusion rate limitations. For dry conditions, vapor

transport in the aggregated material and the sand are similar. Vapor transport in either the aggregated soil column or the sand column is advection-diffusion dominant for low air velocities (0.04 and 1 cm s^{-1}, respectively) or dry conditions. Air-water mass transfer rates appear to be fast for the air velocities used here. Sorption of organic vapors can occur for dry conditions even when no organic material is present, and if the relative humidity of the vapor is zero, adsorption equilibrium is nonlinear.

There are two differences between the important mechanisms for the vapor extraction experiments and the unsaturated water flow results. For the situations where water was flowing, vapor diffusion could be ignored, but gas diffusion could not be neglected when air was the only mobile fluid. Diffusion out of water was not important for the moist sand vapor extraction experiment; however, it was important in the water flow sand experiment. This discrepancy occurred even though the degrees of saturation were about the same for the two experiments. The distribution of liquid water may be different between the water-flow and the no-water-flow conditions. The impact of intraaggregate diffusion that was observed in the unsaturated sand column for flowing water could not be used to predict the vapor extraction column results for moist, no-water-flow conditions.

5.2 SUMMARY OF VAPOR EXTRACTION MODEL RESULTS

A set of models was derived to examine the relative impacts of the mechanisms studied above on the performance of vapor extraction. Two extraction configurations were studied. The first system geometry corresponds to a typical field configuration of a vertical extraction vent completed in a homogeneous soil. The other system geometry

simulates a trench or planar flow configuration. The mechanisms considered in the models are believed to have an influence on the removal of dissolved, volatile organic chemicals. Aqueous-phase transport was modeled as steady, vertical advective and dispersive movement in water. Vapor-phase transport included horizontal air advection and horizontal and vertical gas diffusion. Aqueous and vapor concentrations were coupled by an air-water mass transfer mechanism that could account either for volatilization rate limitations or intraaggregate diffusion. Linear sorption from aqueous solution was also considered. The models were used to study the impacts of these mechanisms for a six order-of-magnitude range of air withdrawal rates. Calculations were performed for both configurations to examine the removal of toluene from a cohesionless sand and from an aggregated soil.

The extraction models were solved numerically using orthogonal collocation. Orthogonal collocation was successful for solving the column models. The numerical approximations used here were verified by comparing them to analytic solutions for simplified conditions. Two-dimensional air diffusion was not approximated satisfactorily in the general radial model. Orthogonal collocation was not appropriate for solving the general set of radial equations. The numerical method would not allow general initial conditions, and it was not very flexible with regard to using different boundary conditions. The numerical solution of the coupled, one-dimensional transport equations (vertical aqueous-phase movement and horizontal vapor-phase movement) was verified. Vertical gas diffusion was ignored in the development of the planar model because it was not approximated satisfactorily in the radial development. The extrapolation of the approach for developing column

models to describe field-scale processes is useful for aiding in the understanding of important processes and for giving additional guidance for design. The mathematical simplifications imposed for the column models reported here, however, are not adequate for predicting field performance in general.

Model calculations indicate that a range of air flow rates exist, for a given vapor extraction scheme, where a minimum volume of air is required to reach a specified level treatment level. More air is required than the minimum to remove a chemical for low air flow rates because of diffusional mixing in the vapor phase and for high flow rates because of phase mass transfer limitations. The importance of diffusional mixing is a function of the air-filled porosity and the gas diffusion coefficient. Phase mass transfer impacts are affected by the amount and distribution of water or the degree of soil aggregation. Limitations due to phase mass transfer will be realized in aggregated soils at lower air flow rates than in sands. Vapor extraction models that assume complete mixing overestimate the time for cleanup unless the air velocity is low. Ignoring the rate of mass transfer between the phases can lead to an underestimate of the cleanup time of dissolved contaminants being treated by high air flows. System operation should be designed for air flow rates large enough to minimize diffusional mixing and low enough to allow phase equilibrium to be reached. The models developed here are useful for estimating this range.

For most air flow rates, planar and radial configurations treat soils at similar rates. There is a narrow range of air flow rates where planar geometry is more efficient, in terms of air volume extracted, than radial. This range occurs near the lower limit of the optimum

range for planar geometry.

The optimum operating range for planar geometry is larger than radial. For similar system size, pressure drops associated with radially converging flow are at least twice that for uniform planar flow. Therefore a trench system is probably a better configuration in terms of operating flexibility and energy requirements. The volume of air that is required to treat a soil can be directly related to energy requirements given the system pressure drop. Pressure drop in the soil system increases with flow rate to the second power, hence it is best to use as low as air flow rate as time will allow.

SECTION 6. RECOMMENDATIONS FOR FUTURE WORK

Before mathematical models can be used for fate predictions and treatment system design, experimental validation must be performed. This work included validation experiments for the column models developed here. The results of the validation steps generated several aspects of future study. Chemical equilibrium in the multiphase systems is still not well understood, especially for contaminant mixtures. The impact of moisture content on vapor sorption is important for vapor extraction design as well as the effect, if any, of soil structure on air-water, organic-air, and organic-water equilibrium. Air-water, water-organic, and air-organic mass transfer rates have not been determined for multiphase flow in soils. These impacts could be studied in columns by measuring breakthrough curves for a range of fluid velocities and saturations. The impact of degree of saturation on liquid dispersion is probably not important to consider for designing vapor extraction systems, but will probably be important for predicting chemical migration through the vadose zone. Correlations exist for estimating dispersion coefficients for unsaturated conditions but they are often soil specific so a more fundamental approach that considers variations in moisture content and fluid velocity is necessary. The column models developed in this work have been validated for soils that are not found naturally, so applications of the column models to experiments using natural soils would be a logical extension.

The vapor extraction system models have not been validated but should be. Although these models are not yet attractive for engineering design applications, the models can be used to increase the level of understanding of the vapor extraction process. By the same token,

controlled large-scale experiments, that closely simulate field situations, in conjunction with column experiments for parameter estimation, could be used to validate some of the conclusions resulting from the numerical studies reported here. Controlled comparisons of radial and trench configurations would be especially useful. In addition, since air advection is the primary removal mechanism in vapor extraction, studies that address the suitability of assumptions such as steady and incompressible flow are necessary for field-scale design. Pressure distribution studies for radial and trench configurations are needed to examine air flow patterns and the impact of adjacent vents.

Validation experiments could enhance the understanding of subsurface chemical fate as well. For example, simultaneous vertical water flow and horizontal air flow experiments would allow measurements of the impact of moisture content on air permeability (an impact not well studied to date). The results of this fluid flow study would determine whether it is important to account for variations in moisture content with depth in the model development. It will probably be important to include nonuniform and time-varying moisture profiles as the system scale increases. Layering and heterogeneities are also important considerations when describing field situations. Models that account for layering may be used to approximate a nonuniform moisture profile.

Finally, the vapor extraction process provides fresh air to the soil system and this may promote aerobic biodegradation. This may be an important consideration in deciding on the feasibility of vapor for treating soils that are contaminated with mixtures of volatile and nonvolatile chemicals.

SECTION 7. LIST OF NOTATION

To reduce the complexity of the notation list, sometimes the same notation is used for describing similar groups, parameters, and variables in different models. The exact definitions are given in the sections where they are used. For definitions of the dimensionless variables and groups, refer to Tables 4.1.1.1-2, 4.1.2.1-2, 4.2.1.1-2, and 4.2.2.1-2.

a specific interfacial area between the air and water (L^{-1}).

apv air pore volumes (dimensionless).

Ar advective flux ratio, ratio of chemical mass transport rate by advection in air to that by advection in water or vice versa (dimensionless).

$A^{x}_{i,j}$ i, j member of the OC coefficient matrix that is used for approximating the first x-derivative of $c_v(z,t)$ (dimensionless).

$A^{z}_{i,j}$ i, j member of the OC coefficient matrix that is used for approximating the first z-derivative of $c_b(z,t)$ or $c_v(z,t)$ (dimensionless).

$A^{zb}_{i,j}$ i, j member of the OC coefficient matrix that is used for approximating the first z-derivative of $c_b(z,t)$ (dimensionless).

$A^{zv}_{i,k}$ i, k member of the OC coefficient matrix that is used for approximating the first z-derivative of $c_v(z,t)$ (dimensionless).

$B^{r}_{i,j}$ i, j member of the OC coefficient matrix that is used for approximating the laplacian of $c_p(r,z,t)$ (dimensionless).

$B^{z}_{i,j}$ i, j member of the OC coefficient matrix that is used for approximating the second z-derivative of $c_b(z,t)$ or $c_v(z,t)$ (dimensionless).

$B^{zv}_{i,k}$ i, k member of the OC coefficient matrix that is used for approximating the second z-derivative of $c_v(z,t)$ (dimensionless).

$c_{g,1}, c_{g,2}, c_{g,3}, c_{g,4}$ dimensionless, empirical constants.

$c_{l,1}, c_{l,2}, c_{l,3}, c_{l,4}$ dimensionless, empirical constants.

$C_b(R,Z,T)$, $C_b(X,Z,T)$, $C_b(Z,T)$ mobile-water-phase chemical concentration ($M\ L^{-3}$).

$c_b(r,z,t) = C_b(r,z,t)\ C_{bn}^{-1}$ (dimensionless).
$c_b(x,z,t) = C_b(x,z,t)\ C_{bn}^{-1}$ (dimensionless).
$c_b(z,t) = C_b(z,t)\ C_{bn}^{-1}$ (dimensionless).

$C_{bi}(R,Z)$, $C_{bi}(X,Z)$, $C_{bi}(Z)$ initial mobile-water-phase chemical concentration ($M\ L^{-3}$).

$c_{bi}(r,z) = C_{bi}(r,z)\ C_{bn}^{-1}$ (dimensionless).
$c_{bi}(x,z) = C_{bi}(x,z)\ C_{bn}^{-1}$ (dimensionless).
$c_{bi}(z) = C_{bi}(z)\ C_{bn}^{-1}$ (dimensionless).

$C_{bo}(T)$ influent chemical concentration in water ($M\ L^{-3}$).

$c_{bo}(t) = C_{bo}(t)\ C_{bn}^{-1}$ (dimensionless).

C_{bn} normalizing water-phase concentration ($M\ L^{-3}$).

$C_p(R,Z,T)$ immobile-water-phase chemical concentration ($M\ L^{-3}$).

$c_p(r,z,t) = C_p(r,z,t)\ C_{bn}^{-1}$ (dimensionless).

$C_{pi}(R,Z)$ initial immobile-water-phase chemical concentration ($M\ L^{-3}$).

$c_{pi}(r,z) = C_{pi}(r,z)\ C_{bn}^{-1}$ (dimensionless).

$C_v(R,Z,T)$, $C_v(X,Z,T)$, $C_v(Z,T)$ air-phase chemical concentration ($M\ L^{-3}$).

$c_v(r,z,t)$ = $C_v(r,z,t)\ C_{vn}^{-1}$ (dimensionless).
$c_v(x,z,t)$ = $C_v(x,z,t)\ C_{vn}^{-1}$ (dimensionless).
$c_v(z,t)$ = $C_v(z,t)\ C_{vn}^{-1}$ (dimensionless).

$C_{vi}(R,Z)$,
$C_{vi}(X,Z)$,
$C_{vi}(Z)$ initial air-phase chemical concentration ($M\ L^{-3}$).

$c_{vi}(r,z)$ = $C_{vi}(r,z)\ C_{vn}^{-1}$ (dimensionless).
$c_{vi}(x,z)$ = $C_{vi}(x,z)\ C_{vn}^{-1}$ (dimensionless).
$c_{vi}(z)$ = $C_{vi}(z)\ C_{vn}^{-1}$ (dimensionless).

C_{vn} normalizing air-phase concentration ($M\ L^{-3}$).

$C_{vo}(T)$ influent chemical concentration in air as a function of time ($M\ L^{-3}$).

$c_{vo}(t)$ = $C_{vo}(t)\ C_{vn}^{-1}$ (dimensionless).

D_e effective gas diffusion coefficient ($L^2\ T^{-1}$).

Dg total solute distribution parameter (dimensionless).

D_G gas diffusion coefficient ($L^2\ T^{-1}$).

Dg_p immobile-water solute distribution parameter, ratio of chemical mass contained in immobile water to that in mobile water or air (dimensionless).

Dg_s adsorbed solute distribution parameter, ratio of chemical mass adsorbed onto soil particle surfaces to that contained in mobile water or air at equilibrium (dimensionless).

Dg_v vapor solute distribution parameter, ratio of chemical mass contained in air-filled pores to that in mobile water or mass in water to that in air at equilibrium (dimensionless).

D_l liquid diffusion coefficient ($L^2\ T^{-1}$).

D_p intraaggregate liquid diffusion coefficient ($L^2\ T^{-1}$).

Ed_p intraaggregate diffusion modulus, ratio of chemical mass transfer rate by diffusion in immobile water to transport by advection in water or air (dimensionless).

Ed_v vertical diffusion modulus, ratio of chemical mass transfer rate by vertical diffusion in air to transport by advection in air (dimensionless).

E_v axial dispersion or diffusion coefficient in air ($L^2\ T^{-1}$).

E_z combined axial dispersion and diffusion coefficient in water ($L^2\ T^{-1}$).

f() collision function for diffusion as a function of temperature and energy of molecular attraction (dimensionless).

g gravitational acceleration ($L\ T^{-2}$).

H Henry's air-water partitioning coefficient (dimensionless).

I number of radial collocation points or horizontal mesh points.

J number of axial collocation points or vertical mesh points.

K soil sorption capacity ($[L^3/M]^{1/n_M}\ M^{-1}$).

K(S) unsaturated hydraulic conductivity as a function of degree of saturation ($L\ T^{-1}$).

K_a air conductivity ($L\ T^{-1}$).

k_f film transfer coefficient ($L\ T^{-1}$).

k_g gas-phase transfer coefficient ($L\ T^{-1}$).

k_l liquid-phase mass transfer coefficient ($L\ T^{-1}$).

K_L overall mass transfer coefficient between air and water ($L\ T^{-1}$).

K_s saturated hydraulic conductivity ($L\ T^{-1}$).

L column length or distance between extraction and inlet vents (L).

M_A molecular weight (M mol^{-1}).

n_+ valence of cation (dimensionless).

n_- valence of anion (dimensionless).

1/n adsorption intensity (dimensionless).

NR number of radial collocation points.

NX number of horizontal collocation points.

NZ number of vertical collocation points.

NZB number of vertical collocation points for water-phase equations.

NZV number of vertical collocation points for air-phase equations.

ΔP air pressure drop in soil system (L).

Pe = $(1+Ar)\ [Pe_b^{-1}+Pe_v^{-1}]^{-1}$ (dimensionless).

Pe_b Peclet number for water, ratio of chemical mass transfer rate by advection in water or air to transport by dispersion in water (dimensionless).

Pe_v Peclet number for air, ratio of chemical mass transfer rate by advection in water or air to transport by dispersion or diffusion in air, (dimensionless).

P_t atmospheric pressure (M L^{-1} T^{-2}).

Q(R,Z,T) adsorbed-phase chemical concentration (M M^{-1}).

Q_i(R,Z) initial adsorbed-phase chemical concentration (M M^{-1}).

Q_G volumetric air flow rate (L^3 T^{-1}).

Q_L volumetric water flow rate (L^3 T^{-1}).

r dimensionless R coordinate.

R radial coordinate (L).

R_a radius of aggregated particle (L).

R_d retardation coefficient (dimensionless).

R_i radius from extraction vent to inlet vent (L).

Re_g gas Reynolds number, equal to $2R_a\rho_g u\ \mu_g^{-1}$ (dimensionless).

Re_l liquid Reynolds number, equal to $2R_a\rho_l v\ \mu_l^{-1}$ (dimensionless).

R_w radius of extraction vent hole (L).

S degree of saturation, relative volume of the pores filled with water (dimensionless).

Sc_g gas Schmidt number, equal to $\mu_g\ (\rho_g D_G)^{-1}$ (dimensionless).

Sc_l liquid Schmidt number, equal to $\mu_l\ (\rho_l D_l)^{-1}$ (dimensionless).

Sh_g gas Sherwood number, equal to $k_g\delta_g\ D_G^{-1}$ (dimensionless).

Sh_l liquid Sherwood number, equal to $k_l\delta_l\ D_l^{-1}$ (dimensionless).

S_i immobile degree of saturation, relative volume of the pores filled with immobile water (dimensionless).

St_b Stanton number for film transfer, ratio of chemical mass transfer rate by film transfer to the transport rate by advection in water (dimensionless).

St_v Stanton number for volatilization, ratio of chemical mass transfer rate across air-water interface to the transport rate by advection in water or air (dimensionless).

t dimensionless time or throughput.

T time (T).

T_e temperature (^oK).

T_b boiling point (^oK).

u interstitial air velocity ($L\ T^{-1}$).

v interstitial water velocity ($L\ T^{-1}$).

V_a molar volume ($L^3\ mol^{-1}$).

v_p average pore water velocity, equal to $v\epsilon S$ ($L\ T^{-1}$).

W^r_i i-th component of the weight vector used in the quadrature approximation of r integrals (dimensionless).

W^x_i i-th component of the weight vector used in the quadrature approximation of x integrals (dimensionless).

W^z_j j-th component of the weight vector used in the quadrature approximation of z integrals of air- and water-phase concentrations (dimensionless).

W^{zb}_j j-th component of the weight vector used in the quadrature approximation of z integrals of water-phase concentration (dimensionless).

W^{zv}_k k-th component of the weight vector used in the quadrature approximation of z integrals of air-phase concentration (dimensionless).

$Y(R,Z,T)$ total intraaggregate concentration per mass of soil, equal to $\epsilon S_i C_p(R,Z,T)\ [\rho_s(1-\epsilon)]^{-1} + Q(R,Z,T)$ ($M\ M^{-1}$).

$y(r,z,t)$ = $Y(r,z,t)\ \{\epsilon S_i C_{bn}\ [\rho_s(1-\epsilon)]^{-1} + K C_{bn}^{1/n}\}^{-1}$ (dimensionless).

z dimensionless Z coordinate.

Z axial position in columns or depth in two-dimensional systems.

α_n n-th root.

δ_g gas diffusion thickness (L).

δ_l liquid diffusion thickness (L).

ϵ total porosity, void fraction of column (dimensionless).

ϵ_a microporosity, void fraction containing immobile water, equal to $\epsilon S_i\ [1-\epsilon(1-S_i)]^{-1}$ (dimensionless).

ϵ_m macroporosity, void fraction containing air and mobile water, equal to $(\epsilon-\epsilon_a)\ (1-\epsilon_a)^{-1}$ (dimensionless).

λ_+ ionic conductance of cation in water (A V g-equiv L^6).

λ_- ionic conductance of anion in water (A V g-equiv L^6).

μ_g air viscosity (M L^{-1} T^{-1}).

μ_l water viscosity (M L^{-1} T^{-1}).

ρ_a particle density, equal to $\rho_s(1-\epsilon_a)$ (M L^{-3}).

ρ_b bulk density (M L^{-3}).

ρ_g air density (M L^{-3}).

ρ_l water density (M L^{-3}).

ρ_s soil particle density, equal to $\rho_b\ (1-\epsilon)^{-1}$ (M L^{-3}).

τ_a tortuosity of the air-filled pores (dimensionless).

τ_p tortuosity of the micropores (dimensionless).

SECTION 8. REFERENCES

Abriola, L. M., and G. F. Pinder, A multiphase approach to the modeling of porous media contamination by organic compounds 1. equation development, Water Resources Research, 21(1), 11-18, 1985a.

Abriola, L. M., and G. F. Pinder, A multiphase approach to the modeling of porous media contamination by organic compounds 2. numerical simulation, Water Resources Research, 21(1), 19-26, 1985b.

Abriola, L. M., and W. J. Weber, Jr., Summary Report: Forum on NSF Research Activities in Subsurface Systems, University of Michigan, Ann Arbor, July 23-25, 1986.

Anastos, G. J., P. J. Marks, M. H. Corbin, and M. F. Coia, Task 11. In situ air stripping of soils, pilot study, final report, Report No. AMXTH-TE-TR-85026, U. S. Army Toxic & Hazardous Material Agency, Aberdeen Proving Grounds, Edgewood, Maryland, 88 pp., October 1985.

Ashworth, R. A., G. B. Howe, M. E. Mullins, and T. N. Rogers, Air-water partitioning coefficients of organics in dilute aqueous solutions, J. Haz. Materials, 18, 25-36, 1988.

Baehr, A. L., Selective transport of hydrocarbons in the unsaturated zone due to aqueous and vapor phase partitioning, Water Resources Research, 23(10), 1926-1938, 1987.

Baehr, A. L., and M. Y. Corapcioglu, A compositional multiphase model for groundwater contamination by petroleum products 2. numerical solution, Water Resources Research, 23(1), 201-213, 1987.

Baehr, A. L., G. E. Hoag, and M. C. Marley, Removing volatile contaminants from the unsaturated zone by inducing advective air-phase transport, J. Contaminant Hydrology, 4, 1-26, 1989

Bednar, M. L., Modeling intermittent operation of laboratory-scale vapor extraction systems, M.S. Thesis, Dept. of Chem. Eng., Michigan Technological University, Houghton, in preparation, 1990.

Bennedsen, M. B., Vacuum VOC's from soil, Pollution Engineering, 19(2), 66-68, 1987.

Black, C. A., D. D. Evans, J. L. White, L. E. Ensminger, and F. E. Clark, Methods of Soil Analysis, Part 1, Physical and Mineralogical Properties, Including Statistics of Measurement and Sampling, 700 pp., American Society of Agronomy, Madison, Wisconsin, 1965.

Brenner, H., The diffusion model for longitudinal mixing in beds of finite length. Numerical values, Chemical Engineering Science, 17(4), 229-243, 1962.

Brusseau, M. L., and P. S. C. Rao, Sorption nonideality during organic contaminant transport in porous media, CRC Critical Reviews in Environmental Control, 19(1), 33-99, 1989.

Burchill, S., M. H. B. Hayes, and D. J. Greenland, Adsorption, in The Chemistry of Soil Processes, D. J. Greenland and M. B. H. Hayes, eds., John Wiley and Sons, New York, pp. 221-400, 1981.

Corapcioglu, Y. M., and A. L. Baehr, A compositional multiphase model for groundwater contamination by petroleum products 1. theoretical considerations, Water Resources Research, 23(1), 191-200, 1987.

Chiou, C. T., and T. D. Shoup, Soil sorption of organic vapors and effects of humidity on sorptive mechanism and capacity, Environ. Sci. Technol., 19(12), 1196-1200, 1985.

CH_2M-Hill, Inc., Operable unit remedial action, soil vapor extraction at Thomas Solvents Raymond Road Facility, Battle Creek, Michigan, Quality Assurance Project Plan, U.S. Environmental Protection Agency, Chicago, August 1987.

Crine, M., and G. A. L'Homme, Recent trends in the modeling of catalytic trickle bed reactors, in Mass Transfer with Chemical Reaction in Multiphase Systems, vol II: Three-Phase Systems, edited by E. Alper, pp. 99-132, Martinus Nijhoff, Boston, Mass., 1983.

Crittenden, J. C., N. J. Hutzler, D. G. Geyer, J. L. Oravitz, and G. Friedman, Transport of organic compounds with saturated groundwater flow: model development and parameter sensitivity, Water Resources Research, 22(3), 271-284, 1986.

Croise, J., W. Kinzelbach, and J. Schmolke, Computation of air flows induced in the zone of aeration during in situ remediation of volatile hydrocarbon spills, in Contaminant Transport in Groundwater, Kobus and Kinzelbach, eds., Balkema, Rotterdam, pp. 437-444, 1989.

Danckwerts, P. V., Continuous flow systems, Chem. Eng. Sci., 2(1), 1-13, 1953.

DePaoli, D. W., S. E. Herbes, J. H. Wilson, D. K. Solomon, H. L. Jennings, J. E. Nyquist, R. M. Counce, and C. O. Thomas, In situ soil venting: guidance document, final report prepared for the Air Force Engineering and Services Center, Tyndall AFB, Florida, in review, 1990.

DeSmedt, F., and P. J. Wierenga, Solute transfer through columns of glass beads, Water Resources Research, 20(2), 225-232, 1984.

DeSmedt, F., F. Wauters, and J. Sevilla, Study of tracer movement through unsaturated soil, J. of Hydrology, 85, 169-181, 1986.

Donigian, A. S., and P. S. C. Roa, Overview of terrestrial processes and modeling, in Vadose Zone Modeling of Organic Pollutants, S. C. Hern and S. M. Melancon, eds., Chapter 1, 3-35, Lewis Publishers, Chelsea, Michigan, 295 pp., 1986.

Dynamac Corporation, Literature Review of Forced Air Venting to Remove Subsurface Organic Vapors from Aquifers and Soil. Subtask Statement No. 3. U.S. Air Force Engineering and Services Center, Tyndall AFB, Florida, 30 pp., July 28, 1986.

Falta, R. W., I. Javandel, K. Pruess, and P. A. Witherspoon, Density-driven flow of gas in the unsaturated zone due to the evaporation of volatile organic compounds, Water Resources Research, 25(10), 2159-2169, 1989.

Finlayson, B. A., Nonlinear Analysis in Chemical Engineering, 366 pp., McGraw-Hill, New York, 1980.

Freundlich, H., Colloid and Capillary Chemistry, English trans. of 3rd German ed. by J. S. Hatfield, 883 pp., E. P. Dutton, New York, 1922.

Gierke, J. S., Modeling the movement of volatile organic chemicals through homogeneous, isotropic, unsaturated soils with cocurrent air and water flow, M.S. Thesis, Dept. of Civil Eng., Mich. Tech. Univ., Houghton, (Available from University Microfilms, Ann Arbor, Michigan) 1986.

Hamaker, J. W. and J. M. Thompson, Adsorption, in Organic Chemicals in the Environment, vol. 1, C. A. I. Goring and J. W. Hamaker, eds., Marcel Dekker, New York, pp. 49-143, 1972.

Hashimoto, I., K. B. Deshpande, and H. C. Thomas, Peclet numbers and retardation factors for ion exchange columns, Ind. Eng. Chem. Fund., 3(3), 213-218, 1964.

Hassett, J. J., W. L. Banwart, and R. A. Griffin, Correlation of compound properties with sorption characteristics of nonpolar compounds by soils and sediments, in Environment and Solid Wastes, C. W. Francis, S. I. Auerbach, and V. A. Jacobs, eds., Butterworths, Boston, pp. 161-178, 1983.

Hern, S. C., S. M. Melancon, and J. E. Pollard, Generic steps in the field validation of vadose zone fate and transport models, in Vadose Zone Modeling of Organic Pollutants, S. C. Hern and S. M. Melancon, eds., Chapter 3, 61-80, Lewis Publishers, Chelsea, Michigan, 295 pp., 1986.

Hofmann, H., Fluiddynamics, mass transfer and chemical reaction in multiphase catalytic fixed bed reactors, in Mass Transfer with Chemical Reaction in Multiphase Systems, vol II: Three-Phase Systems, edited by E. Alper, pp. 73-97, Martinus Nijhoff, Boston, Mass., 1983.

Hutzler, N. J., J. C. Crittenden, J. S. Gierke, and A. M. Johnson, Transport of organic compounds with saturated groundwater flow: experimental results, Water Resources Research, 22(3), 285-295, 1986.

Hutzler, N. J., B. E. Murphy, and J. S. Gierke, State of technology review: vapor extraction systems, EPA 600/2-89-024, 87 pp., Environmental Protection Agency, Washington, D. C., 1989a.

Hutzler, N. J., J. S. Gierke, and L. C. Krause, Movement of volatile organic chemicals in soils, in Reactions and Movement of Organic Chemicals in Soils, Special Pub. no. 22, edited by B. L. Sawhney and K. Brown, pp. 373-403, Soil Sci. Soc. Amer. and Amer. Soc. Agron., Madison, Wisconsin, 1989b.

Im, J. S., Contaminant transport models in layered porous media: Application of the orthogonal collocation method, Ph.D. Dissertation, Michigan Technological University, Houghton (Available from University Microfilms, Ann Arbor, Michigan), 1988.

Johnson, P. C., C. C. Stanley, M. W. Kemblowski, D. L. Byers, and J. D. Colthart, A practical approach to the design, operation, and monitoring of in situ soil-venting systems, Ground Water Monitoring Review, 10(2), 159-178, 1990a.

Johnson, P. C., M. W. Kemblowski, and J. D. Colthart, Quantitative analysis for the cleanup of hydrocarbon-contaminated soils by in situ soil venting, Ground Water, 28(3), 413-429, 1990b.

Jury, W. A., Simulation of solute transport using a transfer function model, Water Resources Research, 18(2), 363-368, 1982.

Jury, W. A., R. Grover, W. F. Spencer, and W. J. Farmer, Modeling vapor losses of soil-incorporated triallate, Soil Sci. Soc. Am. J., 44, 445-450, 1980.

Jury, W. A., D. Russo, G. Streile, and H. El Abd, Evaluation of volatilization by organic chemicals residing below the soil surface, Water Resources Research, 26(1), 13-20, 1990.

Jury, W. A., W. F. Spencer, and W. J. Farmer, Behavior assessment model for trace organics in soil: I. Model description, J. Environ. Qual., 12, 558-563, 1983.

Karickhoff, S. W., D. S. Brown, and T. A. Scott, Sorption of hydrophobic pollutants on natural sediments, Water Research, 13, 1979.

Krause, L. C., Modeling the transport of volatile organic chemicals in unsaturated media: Experimental results, M.S. Civil Eng. thesis, Michigan Tech. Univ., Houghton, (Available from University Microfilms, Ann Arbor, Michigan) 1987.

Krishnayya, A. V., M. J. O'Conner, J. G. Agar, and R. D. King, Vapour extraction systems: factors affecting their design and performance, Proceedings of the NWWA/API Conference on Petroleum Hydrocarbons and Organic Chemicals in Ground Water, 547-569, 1988.

LeBas, G., The Molecular Volumes of Liquid Chemical Compounds, Longmans Green, London, 1915.

Levenspiel, O., Chemical Reaction Engineering, 501 pp., Wiley and Sons, New York, 1962.

Lindstrom, F. T., and W. T. Piver, Vertical-horizontal transport and fate of low water solubility chemicals in unsaturated soils, J. Hydrol., 86, 93-131, 1986.

Massmann, J. W., Applying groundwater flow models in vapor extraction system design, J. ASCE, 115(1), 129-149, 1989.

MacKay, D. M., P. V. Roberts, and J. A. Cherry, Transport of organic contaminants in groundwater, Environ. Sci. Technol., 19(5), 384-392, 1985.

MacKay, D. M., and W. Y. Shiu, A critical review of Henry's law constants for chemicals of environmental interest, J. Phys. Chem. Ref. Data, 10(4), 1175-1199, 1981.

Mayer, R., J. Letey, and W. J. Farmer, Models for predicting volatilization of soil-incorporated pesticides, Soil Sci. Soc. Amer. Proc., 38, 563-568, 1974.

McClellan, R. D., and R. W. Gillham, Vapour extraction of trichloroethylene under controlled field conditions, Proceedings of the IAH Conference on Subsurface Contamination by Immiscible Fluids, April 18-20, Calgary, Alberta, Canada, 1990.

McKenzie, D. B., Extraction of volatile organic chemicals from unsaturated soil: experimental results and model predictions, M. S. Thesis, Dept. of Civil and Environmental Eng., Mich. Tech. Univ., Houghton, (Available from University Microfilms, Ann Arbor, Mich.) 1990.

Mendoza, C. A., and E. O. Frind, Advective-dispersive transport of dense organic vapors in the unsaturated zone: 1. model development, Water Resources Research, 26(3), 379-387, 1990.

Millington, R. J., Gas diffusion in porous media, Science, 130, 100-102, 1959.

Miyauchi, T., and T. Kikuchi, Axial dispersion in packed beds, Chem. Eng. Sci., 30, 343-348, 1975.

Morsi, B. I., and J. C. Charpentier, Hydrodynamics and gas-liquid interfacial parameters with organic and aqueous liquids in catalytic and noncatalytic packings in trickle-bed reactors, in Mass Transfer with Chemical Reaction in Multiphase Systems, vol II: Three-Phase Systems, edited by E. Alper, pp. 133-159, Martinus Nijhoff, Boston, Mass., 1983.

Nicoud, R. M., and D. Schweich, Solute transport in porous media with solid-liquid mass transfer limitations: application to ion exchange, Water Resources Research, 25(6), 1071-1082, 1989.

Nielsen, D. R., J. W. Biggar, and J. M. Davidson, Experimental consideration of diffusion analysis in unsaturated flow problems, Soil Sci. Soc. Amer. J., 26(2), 107-111, 1962.

Nielsen, D. R., M. Th. van Genuchten, and J. W. Biggar, Water flow and solute transport in the unsaturated zone, Water Resources Research, 22(9), 89S-108S, 1986.

Nkedi-Kizzi, P., P. S. C. Rao, R. E. Jessup, and J. M. Davidson, Ion exchange and diffusive mass transfer during miscible displacement through an aggregated oxisol, Soil Sci. Soc. Amer. J., 46, 471-476, 1982.

Okazaki, M., H. Tamon, and R. Toei, Prediction of binary adsorption equilibria of solvent and water vapor on activated carbon, J. of Chem. Eng. of Japan, 11(3), 209-215, 1978.

Peterson, M. S., L. W. Lion, and C. A. Shoemaker, Influence of vapor-phase sorption and diffusion on the fate of trichloroethene in an unsaturated aquifer system, Environ. Sci. & Technol., 22(5), 571-578, 1988.

Raghavan, N. S., and D. M. Ruthven, Numerical simulation of a fixed-bed adsorption column by the method of orthogonal collocation, AIChE J., 29(6), 922-925, 1983.

Rao, P. S. C., R. E. Jessup, and T. M. Addiscott, Experimental and theoretical aspects of solute diffusion in spherical and nonspherical aggregates, Soil Sci., 133, 342-349, 1982.

Reid, R. C., J. M. Prausnitz, and T. K. Sherwood, The Properties of Gases and Liquids, 3rd ed., 688 pp., McGraw-Hill, New York, 1977.

Roberts, P. V., M. N. Goltz, R. S. Summers, J. C. Crittenden, and P. Nkedi-Kizzi, The influence of mass transfer on solute transport in column experiments with an aggregated soil, J. Contaminant Hydrol., 1, 375-393, 1987.

Rolston, D. E., D. Kirkham, and D. R. Nielsen, Miscible displacement of gases through soil columns, Soil Sci. Soc. Amer. Proc., 33, 488-492, 1969.

Rosen, J. B., Kinetics of a fixed bed system for solid diffusion into spherical particles, J. Chem. Physics, 20(3), 387-394, 1952.

Rosen, J. B., General numerical solutions for solid diffusion in fixed beds, Ind. Eng. Chem., 46(8), 1590-1594, 1954.

Roy, W. R., and R. A. Griffin, Vapor-phase movement of organic solvents in the unsaturated zone, Open File Report No. 16, 37 pp., Environ. Inst. for Waste Mgmt. Studies, Univ. of Alabama, Tuscaloosa, June 1987.

Roy, W. R., and R. A. Griffin, In-situ extraction of organic vapors from unsaturated porous media, Open File Report prepared for the Environmental Inst. for Waste Mgmt. Studies, Univ. of Alabama, Tuscaloosa, June 29, 1989.

Sherwood, T. K., R. L. Pigford, and C. R. Wilke, Mass Transfer, McGraw-Hill, New York, 1975.

Shoemaker, C. A., T. B. Culver, L. W. Lion, and M. G.Peterson, Analytical models of the impact of two-phase sorption on subsurface transport of volatile chemicals, Water Resources Research, 26(4), 745-758, 1990.

Sleep, B. E., and J. F. Sykes, Modeling the transport of volatile organics in variably saturated soil, Water Resources Research, 25(1), 81-92, 1989.

Sontheimer, H., J. C. Crittenden, and R. S. Summers, Activated Carbon for Water Treatment, 2nd ed., AWWA, New York, 722 pp., 1988.

Thortenson, D. C., and D. W. Pollock, Gas transport in unsaturated zones: multi-component systems and the adequacy of Fick's laws, Water Resources Research, 25(3), 477-507, 1989.

Treybal, R. E., Mass Transfer Operations, 784 pp., McGraw-Hill, New York, 1980.

Turek, F., and R. Lange, Mass transfer in trickle bed reactors at low Reynold's number, Chem. Eng. Sci., 36(3), 569-579, 1981.

van Genuchten, M. Th., and W. A. Jury, Progress in unsaturated flow and transport modeling, Reviews of Geophysics, 25(2), 135-140, 1987.

van Genuchten, M. Th., and P. J. Wierenga, Mass transfer studies in sorbing porous media: I. Analytical solutions, Soil Sci. Soc. Am. J., 40(4), 473-480, 1976.

van Genuchten, M. Th., and P. J. Wierenga, Mass transfer studies in sorbing porous media: II. Experimental evaluation with tritium (^3H_2O), Soil Sci. Soc. Am. J., 41, 272-278, 1977.

van Genuchten, M. Th., P. J. Wierenga, and G. A. O'Connor, Mass transfer studies in sorbing porous media: III. Experimental evaluation with 2,4,5,-T, Soil Sci. Soc. Am. J., 41, 278-285, 1977.

Weast, R. C., CRC Handbook of Chemistry and Physics, 61st ed., CRC Press, Boca Raton, Florida, 1981.

Wierenga, P. J., Solute distribution profiles computed with steady-state and transient water movement models, Soil Sci. Soc. Am. J., 41, 1050-1055, 1977.

Wierenga, P. J., and M. Th. van Genuchten, Solute transport through small and large unsaturated soil columns, Ground Water, 27(1), 35-42, 1989.

Wilke, C. R., and C. Y. Lee., Estimation of diffusion coefficients for gases and vapors, Ind. Eng. Chem., 47(6), 1253-1257, 1955.

Wilson, D. E., R. E. Montgomery, and M. R. Sheller, A mathematical model for removing volatile subsurface hydrocarbons by miscible displacement, Water, Air, and Soil Pollution, 33, 231-255, 1987.

Wilson, D. J., A. N. Clarke, and J. H. Clarke, Soil Clean Up by in-situ aeration 1. Mathematical modeling, Separation Science and Technology, 23(10-11), 991-1037, 1988.

Wilson, E. J., and C. J. Geankoplis, Liquid mass transfer at very low Reynolds numbers in packed beds, Ind. Eng. Chem. Fundam., 5(1), 9-14, 1966.

Wootan, W. L., and T. Voynick, Forced venting to remove gasoline vapor from a large-scale model aquifer. Texas Research Institute, Inc., Final Report to American Petroleum Institute, 1984.

Yule, D. F., and W. R. Gardner, Longitudinal and transverse dispersion coefficients in unsaturated Plainfield sand, Water Resources Research, 14(4), 582-588, 1978.

Attachment 2

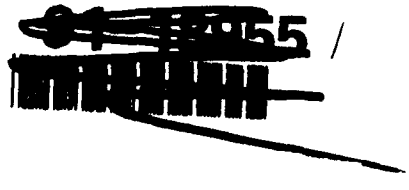
AFOSR-TR- 91 0902

THESIS BY:

LIND SHELMERDINE GEE

HARVARD UNIVERSITY

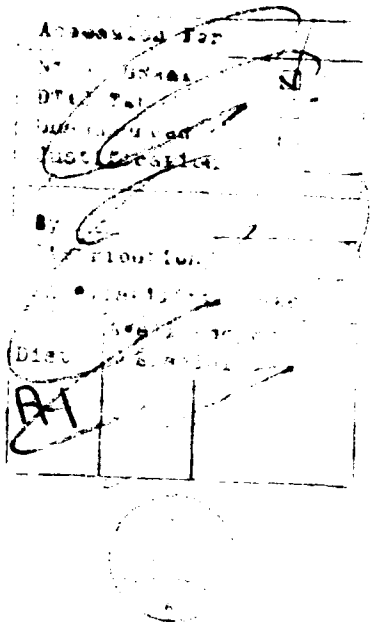
Subcontract No.# S-789-000-005



NEW TECHNIQUES FOR SEISMOLOGICAL STUDIES OF EARTH STRUCTURE

BY

LIND SHELMERDINE GEE
A.B., HARVARD UNIVERSITY 1982



SUBMITTED TO THE DEPARTMENT OF
EARTH, ATMOSPHERIC, AND PLANETARY SCIENCES
IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

AT THE

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
FEBRUARY, 1990

SIGNATURE OF AUTHOR _____

DEPARTMENT OF EARTH, ATMOSPHERIC, AND PLANETARY SCIENCES
FEBRUARY 9, 1990

CERTIFIED BY _____

THOMAS H. JORDAN
THESIS SUPERVISOR

ACCEPTED BY _____

THOMAS H. JORDAN
CHAIRMAN, DEPARTMENT COMMITTEE

NEW TECHNIQUES FOR SEISMOLOGICAL STUDIES OF EARTH STRUCTURE

By

Lind Shelmerdine Gee

Submitted to the Department of Earth, Atmospheric, and Planetary Sciences
in partial fulfillment of
the requirements for the degree of
Doctor of Philosophy

February 9, 1990

ABSTRACT

We have formulated a new waveform-analysis procedure to recover phase and amplitude information from observed seismograms that makes use of our ability to calculate realistic synthetic seismograms. The methodology is based on the representation of the seismogram $s(t)$ as a superposition of traveling wave groups and the construction of an isolation filter $\bar{f}(t)$ as a weighted sum of these wave groups synthesized from a referen e earth model, with the weights chosen to yield a desirable sampling of structural features. We compute the broad-band cross-correlation between the isolation filter and the observed seismogram, window the correlation function about its peak in the time domain, and apply a set of narrow-band filters at discrete center frequencies $\{\omega_i : i = 1, ..., N\}$. We have developed theoretical expressions for the filtered, windowed correlation function based on a Hermite-polynomial expansion of the autocorrelation of the isolation filter, the windowing operator, and the narrow-band filter. When the half-width of the applied filter σ_i is small compared to its center frequency ω_i, the filtered, windowed correlation function may be approximated as a harmonic carrier modulated by a Gaussian envelope: $F_i WC_{\bar{f}s}(t) = E(t) \cos \Phi(t)$, where $E(t) = \exp\{-\sigma_i'^2(t - \Delta\bar{\tau}_g - \delta\tau_g)^2/2\}$ and $\Phi(t) = \omega_i'(t - \Delta\bar{\tau}_p - \delta\tau_p) - \sigma_i'^2(\Delta\bar{\tau}_a + \delta\tau_a)(t - \Delta\bar{\tau}_g - \delta\tau_g)$. In these expressions, ω_i' and σ_i' represent the center frequency and half-width of the filtered, windowed autocorrelation of $\bar{f}(t)$, $\Delta\bar{\tau}_p$, $\Delta\bar{\tau}_g$, and $\Delta\bar{\tau}_a$ represent known phase, group, and attenuation time parameters associated with the reference model used to calculated the isolation filter, and $\delta\tau_p$, $\delta\tau_g$, and $\delta\tau_a$ represent the corresponding unknown time shifts owing to differences between the reference model and the actual Earth at ω_i'. We show how these latter quantities, which we call "generalized seismological data functionals," can be estimated by a waveform-fitting procedure.

The generalized seismological data functionals, $\delta\tau_p$, $\delta\tau_g$, and $\delta\tau_a$, are the phase, group, and attenuation time shifts estimated from the windowed and filtered cross-correlation between an isolation filter $\tilde{f}(t)$ and an observed seismogram $s(t)$ which measure the departure of a reference model $\mathbf{m}_0$ from the Earth. Examples of isolation filters include individual body-wave arrivals such as *S* waves, dispersed wavetrains such as the fundamental-mode Love and Rayleigh waves, as well as general sums over traveling modes constructed to represent complex wavegroups, such as *Sa*, or to sample specific structural features. We have developed a general methodology to synthesize seismologically useful isolation filters, to measure the associated time shifts, and to invert these data for earth structure.

In particular, we have derived expressions for the Fréchet kernels that relate a perturbation in the earth model $\delta\mathbf{m}$ to the first-order perturbations in the generalized data functionals $\delta\tau_p$, $\delta\tau_g$, and $\delta\tau_a$, thus allowing us to pose a linearized inverse problem. For example, the kernels for the phase-delay functional $\delta\tau_p$ are shown to be a sum of the ordinary functional derivatives for individual mode branches obtained from the variational principle. We apply this methodology to portions of the seismogram dominated by multiply reflected body waves, where the complex interferences among the various wave groups confound standard body-wave and surface-wave techniques. We also discuss how isolation filters can be constructed to sample features of earth structure not easily constrained using the standard taxonomy of seismological waveforms, such as the shear velocity of the inner core.

Thesis Committee:

Dr. Thomas H. Jordan, Professor of Geophysics — Thesis Supervisor

Dr. Sean C. Solomon, Professor of Geophysics

Dr. Bradford H. Hager, Professor of Geophysics

Dr. Adam M. Dziewonski, Professor of Geophysics, Harvard University

DEDICATION

This thesis is dedicated to my grandfather, who taught me about the important things in life: hiking, camping, biking, and fishing. He showed me how to feed snapdragons, how to open garage doors by crossing my arms like a genie, and how to flatten pennies on the train tracks. He helped me collect endless quantities of rocks and shells and showed me the fascinating life within the California tidepools. Thank you, Inga, you made my childhood rich with memories.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS

Tom Jordan has been an outstanding advisor. In addition to his skills at throwing parties, playing poker, inventing obscene charades, and mixing chilitinis, he is an excellent scientist. His ability to identify scientific problems of merit ("This is really fundamental") and proscribe innovative solutions ("This is going to revolutionize seismology") is remarkable. His intuition is astonishing ("I don't know what is wrong with this plot, but something isn't right") and often correct. He provided support when I needed it ("This is pretty good ... for a first draft") as well as ample motivation ("The trouble with women is ..."). Working with him on this thesis has been a pleasure.

I would like to thank my committee for participating in my thesis defense: Tom Jordan, Sean Solomon, Brad Hager, and Adam Dziewonski. I appreciate their patience in wading through pages of equations; their comments have improved this thesis greatly. Sean carefully read and commented all aspects of my thesis, from science to hyperbole. Brad made a number of insightful suggestions which have clarified my presentation. Adam contributed his tremendous intuition about seismology.

A number of people have contributed substantially to my interest in earth sciences: Arch, Bill Brace, Ken Creager, Adam Dziewonski, Tom Heaton, Kate Hutton, Ray Jeanloz, Luci Jones, Art Lerner-Lam, Karen McNally, Bernard Minster, and Jim Whitcomb. I owe a special thanks to Bill Brace. Bill convinced me that MIT was the best place to study the Earth and encouraged me to follow my interests, even when they led out of his lab.

The technical skills of Dave Krowitz have saved me from an open-ended career as a graduate student. His ability to diagnose hardware and software problems, install operating systems, write device drivers, advise on programming efficiency, locate and replace bad chips on aging equipment, and generally keep the Apple-Apollo-Alliant

network running smoothly is nothing short of remarkable He and Sandy have been very special friends.

The support staff in the Department have guided me through the maze of MIT procedures and requirements. In particular, I want to thank Debby Roecker, Donna Martel, Sharon Feldstein, Jan Sahlstrom, Laura Doughty, Marie Senat, and Katherine Ware. Katherine and Marie have exercised infinite gentleness and patience with me over the last few months.

I have enjoyed my years as a graduate student at MIT. The environment has been exceptionally stimulating and supportive. I have been fortunate to share offices with a number of talented and interesting people: Geoff Abers, Greg Beroza, Steve Bratt, Carol Bryan, John Goff, Dave Olgaard, Peter Puster, Anne Sheehan, and Jeanne Sauber.

A few people deserve particular recognition. Karen Fischer has been a very special friend and a helpful and entertaining roommate during many AGU meetings. Greg Beroza and Eva Huala have hosted many wonderful Halloween parties and Greg has been a willing conspirator in a multitude of hacks. Mark Murray has contributed significantly to my knowledge of movies, to my ability (and need) to shoot rubber bands, and to my surviving the last year. Bob Grimm was always willing to listen and to play games. Justin Revenaugh taught me all the dirty jokes I know. Dan Burns and Jim Mendelson are always willing to have lunch and shoot the breeze. Peter Puster has been very good natured about a lot of teasing and has acted as a friendly taxi-service. Robin Jordan has been extremely supportive throughout the years and I value our friendship highly.

Michael Baker has been my best friend for over six years. Our relationship has survived my generals, both of our theses, and his post-doctoral position in California. I am grateful for his patience and honored by his love.

My family has been very supportive of my career. My father has attended several AGU talks and my mother has baked enough cookies to feed an army. Next time, let's hide the baby pictures.

I want to thank the National Science Foundation for a NSF Graduate Fellowship and the Air Force for an Air Force Graduate Fellowship. I am grateful to Jim Lewkowicz and John Cipar for allowing me participate in the Air Force Graduate Fellowship program. My research has also been supported by NSF through contract #EAR-8904710 and by DARPA through contracts #F19628-87-K-0040 and #F19628-85-K-0024. The data used in my thesis was provided by the National Earthquake Information Center. John Woodhouse and Adam Dziewonski kindly allowed me to borrow tapes from their collection of NEIC Network Day Tapes as well as use their optical disk drive.

Biographical Note

My interest in earth sciences was awakened at 6:09 AM Feburary 9, 1971. Although our home was not badly damaged, the San Fernando earthquake made a lasting impression, particularly since my brother was born nine months later. Despite my initial confusion about the relationship between earthquakes and babies, I have been fascinated by seismology ever since.

GLOSSARY

$\otimes$	cross-correlation operator
$*$	convolution operator (as a superscript, the complex-conjugate operator)
$s(t)$	the observed seismogram
$\tilde{s}(t)$	the complete synthetic seismogram
$\tilde{f}(t)$	the isolation filter
$u_m(t)$	an observed wavegroup
$\tilde{u}_m(t)$	a synthetic wavegroup, generally assumed to a traveling-wave branch
$C_{\tilde{f}\tilde{f}}$	the autocorrelation of the isolation filter $\equiv \tilde{f}(t) \otimes \tilde{f}(t)$
$C_{\tilde{f}\tilde{s}}$	the cross-correlation between between $\tilde{f}(t)$ and $\tilde{s}(t) \equiv \tilde{f}(t) \otimes \tilde{s}(t)$
$C_{\tilde{f}s}$	the cross-correlation between between $\tilde{f}(t)$ and $s(t) \equiv \tilde{f}(t) \otimes s(t)$
$\mathrm{Ga}(x)$	Gaussian function $\equiv \exp(-x^2/2)$
$\mathrm{He}_k(x)$	the Hermite polynomial of degree k
$\mathrm{H}(x)$	the Heaviside step function
$F_i(\omega)$	the narrow-band filter, characterized by a center frequency ω_i and half-width σ_i
$F_iC_{\tilde{f}\tilde{f}}$	the filtered autocorrelation function of the isolation filter $\equiv F_i(t) * C_{\tilde{f}\tilde{f}}(t)$
$F_iC_{\tilde{f}\tilde{s}}$	the filtered cross-correlation between $\tilde{f}(t)$ and $\tilde{s}(t) \equiv F_i(t) * C_{\tilde{f}\tilde{s}}(t)$
$F_iC_{\tilde{f}s}$	the filtered cross-correlation between $\tilde{f}(t)$ and $s(t) \equiv F_i(t) * C_{\tilde{f}s}(t)$
$\Delta\tilde{\tau}_p^{nm}$	the differential phase delay between the nth and mth branches
$\Delta\tilde{\tau}_g^{nm}$	the averaged differential group delay between the nth and mth branches
$\Delta\tilde{\tau}_a^{nm}$	the averaged differential attenuation delay between the nth and mth branches
$\Delta\tilde{\tau}_p$	the averaged differential phase delay due to branch-branch interference
$\Delta\tilde{\tau}_g$	the averaged differential group delay due to branch-branch interference
$\Delta\tilde{\tau}_a$	the averaged differential attenuation delay due to branch-branch interference
δk_m	the complex-valued differential wavenumber of the mth branch

$\delta\tau_p^m$	the differential phase delay of the mth branch
$\delta\tau_g^m$	the differential group delay of the mth branch
$\delta\tau_a^m$	the differential attenuation delay of the mth branch
$\delta\tau_p$	the averaged differential phase delay
$\delta\tau_g$	the averaged differential group delay
$\delta\tau_a$	the averaged differential attenuation delay
$W(t)$	the windowing operator, characterized by a location time t_w and half-width σ_w^{-1}
$WC_{\tilde{f}\tilde{f}}$	the windowed autocorrelation function of $\tilde{f}(t) \equiv W(t)C_{\tilde{f}\tilde{f}}(t)$
$WC_{\tilde{f}\tilde{s}}$	the windowed cross-correlation between $\tilde{f}(t)$ and $\tilde{s}(t) \equiv W(t)C_{\tilde{f}\tilde{s}}(t)$
$WC_{\tilde{f}s}$	the windowed cross-correlation between $\tilde{f}(t)$ and $s(t) \equiv W(t)C_{\tilde{f}s}(t)$
$F_iWC_{\tilde{f}\tilde{f}}$	the filtered, windowed autocorrelation function of $\tilde{f}(t) \equiv F_i(t) * WC_{\tilde{f}\tilde{f}}(t)$
$F_iWC_{\tilde{f}\tilde{s}}$	the filtered, windowed cross-correlation between $\tilde{f}(t)$ and $\tilde{s}(t) \equiv F_i(t) * WC_{\tilde{f}\tilde{s}}(t)$
$F_iWC_{\tilde{f}s}$	the filtered, windowed cross-correlation between $\tilde{f}$ and $s(t) \equiv F_i(t) * WC_{\tilde{f}s}(t)$

CHAPTER 1

"A PLAYGROUND FOR MATHEMATICIANS"

> Of all regions of the earth none invites speculation more than that which lies beneath our feet, and in none is speculation more dangerous; yet, apart from speculation, it is little that we can say regarding the constitution of the interior of the earth. We know, with sufficient accuracy for most purposes, its size and shape: we know that its mean density is about 5 1/2 times that of water, that density must increase towards the centre, and that the temperature must be high, but beyond these facts little can be said to be known. Many theories of the earth have been propounded at different times: the central substance of the earth has been supposed to be fiery, fluid, solid, and gaseous in turn, till geologists have turned in despair from the subject and become inclined to confine their attention to the outermost crust of the earth, leaving its centre as a playground for mathematicians.
>
> R. D. Oldham, 1906

INTRODUCTION

In the eighty years which have passed since Oldham published "The constitution of the interior of the Earth as revealed by earthquakes," seismologists have made spectacular progress in illuminating the deep structure of the Earth. Our knowledge has grown from a rudimentary grasp of radial structure to detailed images of three-dimensional velocity variations. The last fifteen years have been particularly exciting as the structural perturbations associated with boundary layers and internal dynamics of the mantle and core have come to light. Only a small fraction of the information available from existing data has been used for this purpose, however, and many important geophysical questions related to the details of Earth structure remain unanswered.

For example, the structure of the continental upper mantle is not well understood, despite the accessibility of the continents and the intense scrutiny given to regional phases by generations of seismologists. In particular, the classical argument between Gutenberg and Jeffreys about the existence of a low-velocity zone in the uppermost mantle has not

been fully resolved [*e.g.*, Dziewonski and Anderson, 1981], and questions regarding the role of a low-velocity zone in decoupling continental plates from the the motions of the underlying mantle are the subject of active debate. Linked to this debate are questions about the degree and vertical extent of seismic anisotropy in the upper mantle, and the interplay between lids, low-velocity zones, and anisotropy. Consider the two recent models of continental shear velocity illustrated in Figure 1.1. EU2 (solid line) is a northern Eurasian platform model, derived from the inversion of fundamental and higher-mode Rayleigh waves [Lerner-Lam and Jordan, 1987], while SNA (dashed line) is a Canadian shield model derived from the forward modeling of SH-polarized waveforms and travel times [Grand and Helmberger, 1984]. While EU2 shows a simple, monotonic structure in the upper 200 km with an average velocity of 4.5 km/s, SNA has a thick lid with an average velocity of 4.8 km/s and a well-developed low-velocity zone. The differences between these two models of stable continental structure are as large as the observed variations between continents and oceans [Lerner-Lam and Jordan, 1987] and, if real, have implications for the mechanical, thermal, and chemical evolution of the continents, since the presence of a low-velocity zone is interpreted by some as marking the base of the lithosphere.

Figure 1.2 presents the ω–l diagrams, where ω is frequency and l is angular order, and Figures 1.3 and 1.4 compare the phase and group velocity for the fundamental and first few overtones of toroidal and spheroidal normal modes of these models. SNA (dashed line) predicts phase velocities which are everywhere higher than those of EU2 (solid line) (Figure 1.3). The nearly horizontal branch (which is not a single branch, but rather is composed of the tessellations of several branches) on the spheroidal-mode diagram is the Stoneley wave at the core-mantle boundary and is insensitive to the model differences. The picture is somewhat more complicated for group velocity (Figure 1.4). For example, the predicted group velocities for the fundamental-mode Love wave from SNA (dashed line) are lower than those from EU2 (solid line) in the frequency range from 25 to 50 mHz. The

branches which "bend" over in the group velocity diagram are those which are sensitive to the core-mantle boundary. The structure of the spheroidal-mode diagram is more complex; the core-mantle boundary Stoneley wave forms the cross-cutting horizontal branch. Figure 1.5 illustrates the travel times predicted by these models for the phases *S*, *SS*, and *SSS*. SNA (dashed line) predicts travel times which are substantially less than those of EU2 (solid line) - a difference which is exacerbated with the number of surface reflections. In addition, the triplication structure of the models is different, producing dissimilar waveforms (see Figure 2.1).

There are several possible explanations for the disagreement between these models. First, the differences may represent true path variations between the northern Eurasian platforms and the Canadian shield. SNA has been used to model propagation across the Russian platform and predicts the observed *SH*-polarized waveforms and travel times [Rial *et al.*, 1984; Grand and Helmberger, 1985]. While EU2 may have slightly lower velocities because the path across northern Eurasian from events in the western Pacific to stations in western Europe has a larger tectonic component than the path across the Russian platform from events in the Hindu Kush to the same stations (Figure 1.6), the disagreement between EU2 and SNA is too large to be explained by path differences alone. Second, the differences may be due to polarization anisotropy; EU2 was derived from a study of *PSV*-polarized waveforms whereas SNA was derived from a study of *SH*-polarized waveforms. Although the deviation between the models has the right sense for transverse isotropy with a vertical axis of symmetry, *i.e.*, *SH* advanced with respect to *PSV*, the discrepancy requires significant deep anisotropy. However, most observations of polarization anisotropy are limited to the crust and upper mantle [McEvilly, 1964; Forsyth, 1975; Cara *et al.*, 1980; Dziewonski and Anderson, 1981; Kirkwood and Crampin, 1981; Lévêque and Cara, 1983; Nataf *et al.*, 1984; Silver and Chan, 1988]. Finally, the differences between these models may be due to frequency-dependent propagation. The lid velocity in SNA was determined from observations of *Sn*, a phase which propagates at the top of the mantle

[Brune and Dorman, 1963; Båth, 1966, Heustis *et al.*, 1973]. The apparent velocities of high-frequency (~1 Hz) *Sn* waves on both horizontal and vertical components are typically 100–300 m/s higher than the average shear velocity v_S of the uppermost mantle derived from low-frequency (~.01 Hz) Rayleigh waves. Although not systematically treated in the literature, this problem is evident in the comparison of the *Sn* velocities of Heustis *et al.* [1973] with Rayleigh-wave models such as EU2.

Examples of observed seismograms with synthetic seismograms calculated from the models EU2 and SNA are presented in Figures 1.7 and 1.8. Figure 1.7 illustrates seismograms for the northern Eurasian path; the synthetic seismograms are calculated by normal-mode summation and are complete to 50 mHz. EU2 provides a good fit to the Rayleigh waves and *PSV*-polarized waveforms of multiply reflected *S* phases for the northern Eurasian corridor, but has velocities which are too low for the Love waves and *SH*-polarized phases with turning points in the upper mantle. Figure 1.8 compares seismograms for the Russian platform path, with the same conventions. SNA is generally consistent with the waveforms and travel times of *SH*-polarized body phases for paths to KONO across the Russian platform, but has velocities which are too high for the *PSV*-polarized shear waves with turning points in the upper mantle. In particular, comparison of the travel times of *SS* on the transverse and vertical components provides evidence of apparent shear-wave splitting. On the other hand, the path to GRFO along the southwestern margin of the Russian platform does not show this polarization dependence; the observed travel times exceed those predicted by SNA by about 15 s on both components, consistent with the *SH* observations of Rial *et al.* [1984].

From this cursory examination of observed and synthetic seismograms, we see that EU2 predicts the both *PSV*-polarized waveforms and *SH*-polarized, vertically propagating waveforms for the northern Eurasian corridor. There is a suggestion of shallow anisotropy along this path, based on the misfit between EU2 and the *SH*-polarized, horizontally propagating waves, but there is no evidence for deep-seated anisotropy. Second, there is

considerable lateral variation in structure, as demonstrated by the differences between the waveforms at KONO and GRFO for the Hindu Kush events. Third, there is strong evidence of shallow anisotropy along the Russian platform, based on the observation of apparent shear-wave splitting in *SS*. Thus, the EU2-SNA discrepancy may be explained in part by path variations and in part by intrinsic anisotropy. These observations are discussed in greater detail in Gee and Jordan [1988], which is reproduced in Appendix F.

The EU2-SNA discrepancy symbolizes many of the problems facing seismologists today, as researchers grapple with structural models derived from a variety of specialized techniques which depend on wavegroup, polarization, or frequency. In particular, it demonstrates the need for a self-consistent analysis of seismic data. Because the Earth is neither spherically symmetric nor isotropic, seismologists need techniques which may be applied to *SH* and *PSV*-polarized waveforms and to body waves as well as surface waves in order to reveal the three-dimensional variations in structure. In addition, these methodologies must include the complexity of the frequency-dependent propagation such as shear-coupled *PL* or triplicated arrivals. In this chapter, we discuss two broad categories of techniques for making inferences about earth structure: the "classical" approach, based on the recovery of kinematic properties such as travel time, and the more modern approach of waveform inversion, based on the difference between an observed seismogram and a synthetic seismogram calculated from a reference model. We then introduce the methodology which is the basis of this thesis.

THE CLASSICAL APPROACH

A seismogram $s(t)$ may be represented as a sum over traveling waveforms $\{u_n(t) : n = 1, 2, \ldots\}$:

$$s(t) = \sum_{n=0}^{\infty} u_n(t) \tag{1.1}$$

where a particular element $u_n(t)$ may be a body-wave pulse, a surface-wave group, or any other convenient representation of the seismic wavetrain. Each $u_n(t)$ may be described by several kinematic properties, such as travel time, amplitude, and dispersion. The classical approach in structural seismology is to separate the process of measuring waveform properties from the process of inverting for earth structure (Figure 1.9). Discrete body-wave pulses are identified and their travel times and amplitudes are determined; surface-wave groups are isolated and their dispersion and attenuation properties are measured. These data are then inverted for an earth model whose parameterization is sufficiently complete to explain the observed variations. If a good starting model is available, the latter step may be accomplished using a perturbation theory based on the variational principles of Fermat and Rayleigh. Examples of the classical approach include Jeffreys' application of the Herglotz-Weichert formula [1939], the tomographic inversion of travel times [*e.g.*, Dziewonski *et al.*, 1977; Dziewonski, 1984; Creager and Jordan, 1986; Morelli and Dziewonski, 1987], early studies of surface-wave dispersion [*e.g.*, Dorman *et al.*, 1960; Brune and Dorman, 1963], recent work on phase and group velocities [*e.g.*, Nakanishi and Anderson, 1982, 1983, 1984; Nataf *et al.*, 1984, 1986], measurements of normal-mode eigenfrequencies [*e.g.*, Gilbert and Dziewonski, 1975; Masters *et al.*, 1982; Giardini *et al.*; 1987], and estimates of attenuation [*e.g.*, Jordan and Sipkin, 1977; Sailor and Dziewonski, 1978; Masters and Gilbert, 1982].

A problem with this approach concerns the various wave effects that complicate the measurement of individual wave groups. In the case of body-wave travel times, these include caustic phase shifts, diffraction effects, and the physical dispersion associated with attenuation. In the case of surface-wave dispersion, they involve source-related phase shifts and the problems associated with isolating individual modes. Indeed, for portions of the seismogram where many wave groups arrive simultaneously, waveforms cannot generally be resolved into either individual body waves or surface waves, and the classical

measurement schemes that rely on waveform isolation may fail to produce reliable results. Techniques based on frequency-wavenumber filtering have been used to separate interfering surface waves [Nolet, 1975, 1977; Cara, 1979; Cara *et al.*, 1980], but they generally require large-aperture arrays of seismometers not common in global studies of earth structure.

WAVEFORM INVERSION

Many of these difficulties can be avoided by inverting the complete seismogram directly for earth structure. In the ideal situation, where the entire wavefield is recorded by a spatially dense set of receivers from a spatially dense set of sources, powerful nonlinear inversion techniques may be applied to recover an image of the three-dimensional structure [Tarantola, 1986]. Although the collection of these sorts of ideal data sets is feasible in exploration seismology, the data sets available to global seismology are limited by the distribution of large-magnitude sources, primarily earthquakes, and the sparse distribution of stations, especially those with high-quality, digitally recording seismometers. In this situation, fully nonlinear methods cannot be applied because the solution manifolds have multiple minima, and the problem must be linearized by assuming the solution to the waveform-inversion problem is in some sense close to a chosen reference earth structure.

Theoretical and computational advances over the last two decades now permit the routine calculation of synthetic seismograms $\tilde{s}(t)$ using a variety of waveform representations. A seismogram computed from a reference earth model $\mathbf{m}_0$ may be written as a sum over synthetic waveforms $\{\tilde{u}_n(t)\}$:

$$\tilde{s}(t) = \sum_{n=0}^{N} \tilde{u}_n(t) \tag{1.2}$$

although the number of elements in the sum, N, must necessarily be finite. Synthetic seismograms can model accurately source excitation effects, elastic and anelastic wave propagation, as well as instrument response. Most waveform-inversion algorithms [Mellman, 1980; Dziewonski and Steim, 1982; Lerner-Lam and Jordan, 1983; Shaw, 1983; Woodhouse and Dziewonski, 1984; Tanimoto, 1984, 1987; Nolet *et al.*, 1986] subtract synthetic seismograms computed for the reference structure from the observed time series to form differential seismograms that are then inverted for a structural perturbation using first-order perturbation theory (Figure 1.10). This linearized inverse problem thus takes the form

$$\mathbf{G}\,\delta\mathbf{m} = \delta\mathbf{s} \tag{1.3}$$

where $\delta\mathbf{s}$ is the vector containing the differential time series, $\delta\mathbf{m}$ is the model perturbation, and $\mathbf{G}$ is the matrix of partial derivatives.

The primary problem with waveform inversion is its "black-box" character: it is difficult to evaluate exactly what features on the seismograms determine particular characteristics of the earth model. Although waveform inversion allows more information on the seismogram to be used in constraining earth structure, it is of uneven quality. Amplitude and phase information is combined, so understanding the robustness of the solution to departures from the modeling assumptions (which are usually incomplete with respect to first-order amplitude perturbations) is often impossible. Moreover, the results may be very sensitive to how the data are weighted, and the resolving power of any particular data set is difficult to assess. For example, it is difficult to determine how much of the residual contained in a differential seismogram may be attributed to a one-dimensional path-averaged perturbation, as opposed to two- or three-dimensional along-path and off-path perturbations. Finally, waveform inversion is limited by the need for the synthetic seismogram to be "linearly" close to the observed seismogram. If the waveforms

of the synthetic and the data are too dissimilar, the inversion will not converge to the true minimum.

ISOLATION FILTERING

The application of isolation filtering techniques to isolate individual wave groups prior to inversion has facilitated the understanding of those features on the seismograms which are most significant in constraining the solution. An isolation filter $\tilde{f}(t)$ may generally be defined as a sum over wavegroups:

$$\tilde{f}(t) = \sum_{m=0}^{N} \tilde{\alpha}_m(t) * \tilde{u}_m(t) \tag{1.4}$$

where $\tilde{\alpha}_m$ is a linear filter and the * operator denotes convolution. This approach has been used by Lerner-Lam and Jordan [1983, 1987] to isolate higher-mode arrivals on *PSV* polarized waveforms. In their formulation, the isolation filter is defined as a single traveling wave branch synthetic: $\tilde{f}(t) = \tilde{u}_m(t)$. They define the observed branch cross-correlation function between $\tilde{u}_m(t)$ and $s(t)$ and the synthetic branch cross-correlation function between $\tilde{u}_m(t)$ and $\tilde{s}(t)$:

$$\begin{aligned} C_{\tilde{u}\tilde{s}}(t) &\equiv \tilde{u}_m(t) \otimes \tilde{s}(t) = \int_{-\infty}^{\infty} \tilde{u}_m(\tau)\, \tilde{s}(\tau+t)\, d\tau \\ C_{\tilde{u}s}(t) &\equiv \tilde{u}_m(t) \otimes s(t) = \int_{-\infty}^{\infty} \tilde{u}_m(\tau)\, s(\tau+t)\, d\tau \end{aligned} \tag{1.5}$$

where the $\otimes$ operator denotes cross-correlation. By forming the difference between the observed and synthetic branch cross-correlation functions, Lerner-Lam and Jordan [1983,

1987] invert for model perturbations via (1.3). While this application of isolation filtering eliminates some of the "black box" character of waveform inversion, the approach is still limited by the requirement that the model be linearly close to the Earth.

GENERALIZED DATA FUNCTIONALS

This thesis presents a new set of waveform-analysis procedures to recover estimates of phase and amplitude from seismograms. These procedures have a number of advantages over existing methods. Like waveform-inversion techniques, they make use of our ability to compute synthetic seismograms from realistic earth models, and they provide a uniform methodology for inverting body-wave, surface-wave, and other types of wave groups from three-component data. Unlike waveform-inversion techniques, they isolate from the seismogram time-like quantities that correspond to well-defined scalar-valued functions of earth structure: phase delays, group delays, amplitude factors, and their generalizations. An inversion of these quantities for earth structure (1-D, 2-D or 3-D) can thus be accomplished in a separate step using standard perturbation techniques. The separation of the measurement of data functionals from the inversion for earth structure facilitates the assessment of the significance and robustness of the measurements and allows a variety of model parameterizations and inversion schemes to be compared (Figure 1.11).

Overview of the method

The methodology is based on the representation of the seismogram as a superposition of traveling-wave branches and the construction of a match or isolation filter as a weighted sum of these wave branches synthesized from a reference earth model. We compute the broad-band cross-correlation between the isolation filter and the observed seismogram, window the correlation function about its peak in the time domain, and apply a set of narrow-band filters at discrete center frequencies $\{\omega_i : i = 1, ..., N\}$. We have

developed theoretical expressions for the filtered, windowed correlation function based on a Hermite-polynomial expansion of the autocorrelation of the isolation filter, the windowing operator, and the narrow-band filter. We will show that the correlation function may be approximated as a harmonic carrier modulated by a Gaussian envelope, when the bandwidth of the applied filter σ_i is small compared with its center frequency ω_i':

$$F_i WC_{fs}(t) \equiv E(t) \cos\Phi(t) \tag{1.6}$$

where $F_i WC_{fs}(t)$ is the filtered, windowed cross-correlation between the isolation filter and the observed seismogram and $E(t)$ and $\Phi(t)$ are defined by

$$E(t) \equiv \exp\{-\sigma_i'^2 (t - \Delta\tilde{\tau}_g - \delta\tau_g)^2/2\} \tag{1.7}$$

$$\Phi(t) \equiv \omega_i' (t - \Delta\tilde{\tau}_p - \delta\tau_p) - \sigma_i'^2 (\Delta\tilde{\tau}_a + \delta\tau_a)(t - \Delta\tilde{\tau}_g - \delta\tau_g) \tag{1.8}$$

In these expressions, ω_i' and σ_i' represent the center frequency and half-width of the filtered, windowed autocorrelation of $\tilde{f}(t)$, $\Delta\tilde{\tau}_p$, $\Delta\tilde{\tau}_g$, and $\Delta\tilde{\tau}_a$ represent known phase, group, and amplitude time parameters associated with the reference model used to calculated the isolation filter, and $\delta\tau_p$, $\delta\tau_g$, and $\delta\tau_a$ represent the corresponding unknown time shifts owing to differences between the reference model and the actual Earth at ω_i'. The generalized seismological data functionals, $\delta\tau_p$, $\delta\tau_g$, and $\delta\tau_a$, are the phase, group, and attenuation time shifts estimated from the windowed and filtered cross-correlation between an isolation filter $\tilde{f}(t)$ and an observed seismogram $s(t)$ which measure the departure of a reference model $\mathbf{m}_0$ from the Earth. Examples of isolation filters include individual body-wave arrivals such as *S* waves, dispersed wavetrains such as the fundamental-mode Love and Rayleigh waves, as well as general sums over traveling modes constructed to represent complex wavegroups, such as *Sa*, or to sample specific

structural features. We have developed a general methodology to synthesize seismologically useful isolation filters, to measure the associated time shifts, and to invert these data for earth structure.

In particular, we have derived expressions for the Fréchet kernels that relate a perturbation in the earth model δm to the first-order perturbations in the generalized data functionals $\delta\tau_p$, $\delta\tau_g$, and $\delta\tau_a$, thus allowing us to pose a linearized inverse problem. For example, the kernels for the phase-delay functional $\delta\tau_p$ are shown to be a sum of the ordinary functional derivatives for individual mode branches obtained from the variational principle. We apply this methodology to portions of the seismogram dominated by multiply reflected body waves, where the complex interferences among the various wave groups confound standard body-wave and surface-wave techniques. We also discuss how isolation filters can be constructed to sample features of earth structure not easily constrained using the standard taxonomy of seismological waveforms, such as the shear velocity of the inner core.

Overview of the thesis

In this chapter, we demonstrated the need for a uniform, self-consistent approach to the analysis of seismic data by comparing two recent radial models of the continental upper mantle. The disagreement between EU2 and SNA, which is as large as the observed variation between continents and oceans, emphasizes the importance of techniques which are independent of polarization, wavetype, and frequency band. We outlined the approach of current techniques and introduced the conceptual basis of our methodology. In particular, we illustrated how our procedure combines elements of the classical analysis of seismic phases with the construction of synthetic seismograms which is the cornerstone of waveform inversion.

In Chapter 2, we introduce the tools of our waveform-analysis procedure. We consider an isolation filter composed of a single waveform, such as the fundamental-mode

surface wave, and expand $C_{\bar{u}\bar{u}}(t)$ in a Gram-Charlier series. We show that this expansion may be truncated after the second-order term, provided that the signal is peaked in the spectral domain and may be described by a few of its low-order moments. Although not generally valid, this condition may be met by the application of a narrow-band filter. Using a first-order Taylor series expansion of differential wavenumber and neglecting differences between the actual and assumed source function and instrument response, the filtered cross-correlation between an isolation filter and the observed seismogram may be expressed in terms of time parameters due to differences between a reference model and the actual Earth at ω_i'. These time parameters, which we call "generalized seismological data functionals," represent phase, group, and amplitude time shifts and may be interpreted in terms of Earth structure by variational principles. We develop explicit expressions for the effect of interference from other wavegroups and derive an expression which relates the observed phase perturbation to a linear sum of the individual phase perturbations of each traveling-wave branch. Finally, we illustrate our methodology with the example of fundamental-mode surface waves.

In Chapter 3, we introduce an additional step in our waveform analysis procedure by windowing the broad-band correlation functions before filtering. This step reduces the contamination due to interfering arrivals, although it limits the application of the linear dispersion approximation. We formulate analytic expressions for the cross-correlation of the isolation filter with the complete synthetic and the observed seismogram which model the mode branch interference as a sum over Gaussian wavelets, as well as parameterizing the correlation functions as a single Gaussian wavelet.

In Chapter 4, we explore the general form of the isolation filter as a sum over traveling-wave branches convolved with the weight coefficients $\bar{\alpha}_m(t)$. We consider a method for constructing isolation filters for wavegroups based on the summation of traveling-wave branches with weight coefficients designed to "window" about a particular group arrival time and illustrate this approach with the phases *S* and *SSS*. We develop

expressions for $F_i WC_{\mathcal{T}\mathcal{T}}(t)$, $F_i WC_{\mathcal{T}\mathcal{S}}(t)$, and $F_i WC_{\mathcal{T}\mathcal{S}}(t)$ in terms of sums over traveling-wave branches. We illustrate the implementation with the *S* and *SSS* examples and discuss the Fréchet kernels obtained from our analysis. These kernels provide new insight into the way wavegroups average the Earth at finite frequencies, which is very different from the partial derivatives associated with the infinite frequency approximation of ray theory.

In the final chapter, we consider several outstanding problems in geophysics which may be resolved with these waveform-analysis procedures.

CONCLUSIONS

This thesis presents techniques for the analysis of three-component seismograms. The methodology may be divided into three steps: formulation of the measurement procedure, calculation of the Fréchet kernels, and construction of the isolation filters.

The measurement procedure is based on a cross-correlation formalism. Our approach differs from traditional cross-correlation techniques in several ways. First, we have developed an analytic expression for the cross-correlation between an isolation filter and the observed seismogram which depends on differential time parameters that measure the departure of the Earth from our reference model. Second, the use of complete synthetic seismograms permits the modeling of the source function, the instrument response, and interfering energy. Third, we use a waveform-fitting procedure in order to estimate the differential time parameters, rather than picking the peak of the correlation function or the maximum of the envelope. These differential time parameters recovered with this formalism may be interpreted using standard variation techniques. We derive formulae which express the measured differential phase delay as a weighted sum of the differential phase delays associated with the traveling-wave groups which compose the isolation filter. This construction allows us to employ the full power of inverse theory in interpreting our measurements through mechanisms such as hypothesis testing. While waveform inversion is an inverse procedure, the problem of determining which wiggles contribute to particular

aspects of the model make it difficult to use inversion as a diagnostic tool. Finally, we have developed the ability to design isolation filters for arbitrary wavegroups from normal-mode summation. To our knowledge, this represents the first time that seismologists have used the normal-mode formulation to calculate individual body-wave arrivals such as *S*. Our normal-mode approach allows us to include the effects of polarization, such as coupled and converted arrivals, explicitly. These phenomena are difficult to model using the traditional ray-theoretical approach to pulse propagation, such as WKBJ. Consequently, we may begin to utilize the information available in *PSV*-polarized seismograms; information which has been neglected previously. We may also design isolation filters which do not correspond to the standard taxonomy of seismic phases. For example, we may construct an isolation filter which is sensitive to particular regions within the Earth, such as the shear velocity of the inner core.

FIGURE CAPTIONS

FIGURE 1.1

Shear velocity as a function of depth for continental upper mantle models EU2 [Lerner-Lam and Jordan, 1987] and SNA [Grand and Helmberger, 1984]. SNA (dashed line) was derived from forward modeling of multiply reflected *SH*-polarized waveforms for stable North America paths; EU2 (solid line) was derived from waveform inversion of fundamental Rayleigh waves and *PSV*-polarized higher-modes for northern Eurasia. SNA has been shown to model propagation across the Russian platform [Rial *et al.*, 1984; Grand and Helmberger, 1985]. SNA has significantly higher velocities than EU2 in the upper 400 km; the discrepancy between these models is due in part to path differences and in part to strong polarization anisotropy [Lerner-Lam and Jordan, 1987; Gee and Jordan, 1988 (Appendix F)].

FIGURE 1.2

Frequency (ω) as a function of angular order (l) for the fundamental mode and first few overtones of the models EU2 (solid line) and SNA (dashed line). The lines connect normal modes of constant radial-order number (n). The branch structure of toroidal modes (left) is considerably less complicated that that of spheroidal modes (right). Modes in the left-hand corner of both diagrams are high-phase velocity modes (a straight line passing through the origin of these diagrams is a line of constant phase velocity). In the toroidal-mode diagram, these correspond to *ScS*-equivalent modes. In the spheroidal-mode diagram, these correspond to *PKP*, *PKIKP*, and *PKJKP* in addition to the *ScS*-equivalent modes. Properties of the toroidal modes, such as energy density, vary smoothly along a branch, while those of spheroidal modes do not. The spheroidal-mode diagram is composed of several families of modes which do not correspond to the conventional nomenclature [Okal, 1978]. The ω–l diagram is described in greater detail in Appendix C.

FIGURE 1.3

Phase velocity as a function of frequency for models EU2 and SNA. SNA (dashed line) predicts phase velocities which are everywhere greater than those of EU2 (solid line). The nearly horizontal branch (which is not a single branch, but rather is composed of the tessellations of several branches) on the spheroidal-mode diagram is the Stoneley wave at the core-mantle boundary and is insensitive to the model differences. At 25 mHz (40 s) and a source-receiver distance of 70°, the difference between EU2 and SNA in fundamental-mode phase velocity accumulates to more than 63.3 s differential time in the Love wave and 78.3 s in the Rayleigh wave.

FIGURE 1.4

Group velocity as a function of frequency for models EU2 and SNA. The predicted group velocities for the fundamental-mode Love wave from SNA (dashed line) are less than those from EU2 (solid line) in the frequency range from 25 to 50 mHz. The branches which "bend" over in the group velocity diagram are those which are sensitive to the core-mantle boundary. The structure of the spheroidal-mode diagram is more complex; the core-mantle boundary Stoneley wave forms the cross-cutting horizontal branch.

FIGURE 1.5

Reduced travel time (s) as a function of epicentral distance (degrees) for *S*, *SS*, and *SSS* from the models EU2 and SNA. SNA (dashed line) predicts travel times which are substantially less than those of EU2 (solid line). In addition, the triplication structure of the models is different. The reducing velocity is 18 km/s.

FIGURE 1.6

Azimuthal equidistant projection centered on KONO, illustrating the northern Eurasian and southwestern Eurasian corridors. The northern Eurasian corridor includes the marginal basins and active foldbelts east of the Verkhoyansk suture, as well as the stable cratons of the Siberian and Russian platforms. EU2 was derived from a study of fundamental and higher-mode Rayleigh waves from this path. The southwestern Eurasian corridor includes two paths, one crossing the central part of the Russian platform from Hindu Kush events to KONO, and one traversing the southwestern margin of the platform along the Alpine-Himalayan front to GRFO. SNA has been used to model the *SH*-polarized waveforms and travel times of wavegroups propagating across the Russian platform [Rial *et al.*, 1984; Grand and Helmberger, 1985]. The triangles are earthquake locations and octagons are receiver locations of events used in Gee and Jordan [1988]. Shields and stable platforms (shaded) are from Jordan [1981].

FIGURE 1.7

Comparison of observed and synthetic seismograms for propagation across the northern Eurasian corridor. This figure displays transverse (top) and vertical (bottom) component seismograms for four events in the western Pacific, recorded at Global Digital Seismic Network stations in western Europe. Within each triplet, the solid line is the observed seismogram, the upper trace is the complete synthetic seismogram calculated SNA and the lower trace is the complete synthetic seismogram calculated from EU2. The synthetic seismograms were computed by normal-mode summation and are complete to 50 mHz; all seismograms were filtered with a Hanning taper between 0 and 5 and between 40 and 50 mHz. EU2 provides a good fit to the Rayleigh waves and *PSV*-polarized waveforms of multiply reflected *S* phases for the northern Eurasian corridor, but has velocities which are too low for the Love waves and *SH*-polarized phases with turning points in the upper mantle. SNA predicts travel times which are substantially less than those observed on both the transverse and vertical-component wavegroups.

FIGURE 1.8

Comparison of observed and synthetic seismograms for propagation across the southwestern Eurasian corridor. This figure displays transverse (top) and vertical (bottom) component seismograms for two events in the Hindu Kush, recorded at Global Digital Seismic Network stations in western Europe, with the same conventions as Figure 1.7. SNA is generally consistent with the waveforms and travel times of *SH*-polarized body phases for paths to KONO across the Russian platform, but has velocities which are too high for the *PSV*-polarized *SS* phases with turning points in the upper mantle. On the other hand, the path to GRFO along the southwestern margin of the Russian platform does not show this polarization dependence; the observed travel times exceed those predicted by SNA by about 15 s on both components, consistent with the *SH* observations of Rial *et al.* [1984].

FIGURE 1.9

"Classical" seismological methodology breaks the problem of determining earth structure into two parts: (1) the measurement of well-defined data functionals--*e.g.*, the travel times of body phases; phase and group velocities of surface waves--and (2) the inversion of these data for earth models.

FIGURE 1.10.

The more modern methodology of waveform inversion computes the differences between the observed seismograms and synthetic seismograms and inverts these differential seismograms directly for earth structure.

FIGURE 1.11.

The methodology discussed in this thesis combines the advantages of the classical approach with those of waveform inversion. Measurements of time-like data functionals are made using synthetic seismograms to account for wave effects; *e.g.*, caustic phase shifts, dispersion, and diffraction. These data are subsequently inverted for earth models. By separating the measurement process from the inversion process, the significance of the data and the robustness of the modeling can be more easily assessed than in a one-step, "black-box" waveform inversion scheme. The technique provides a uniform methodology for inverting body-wave, surface-wave, and other types of wave groups from three-component data.

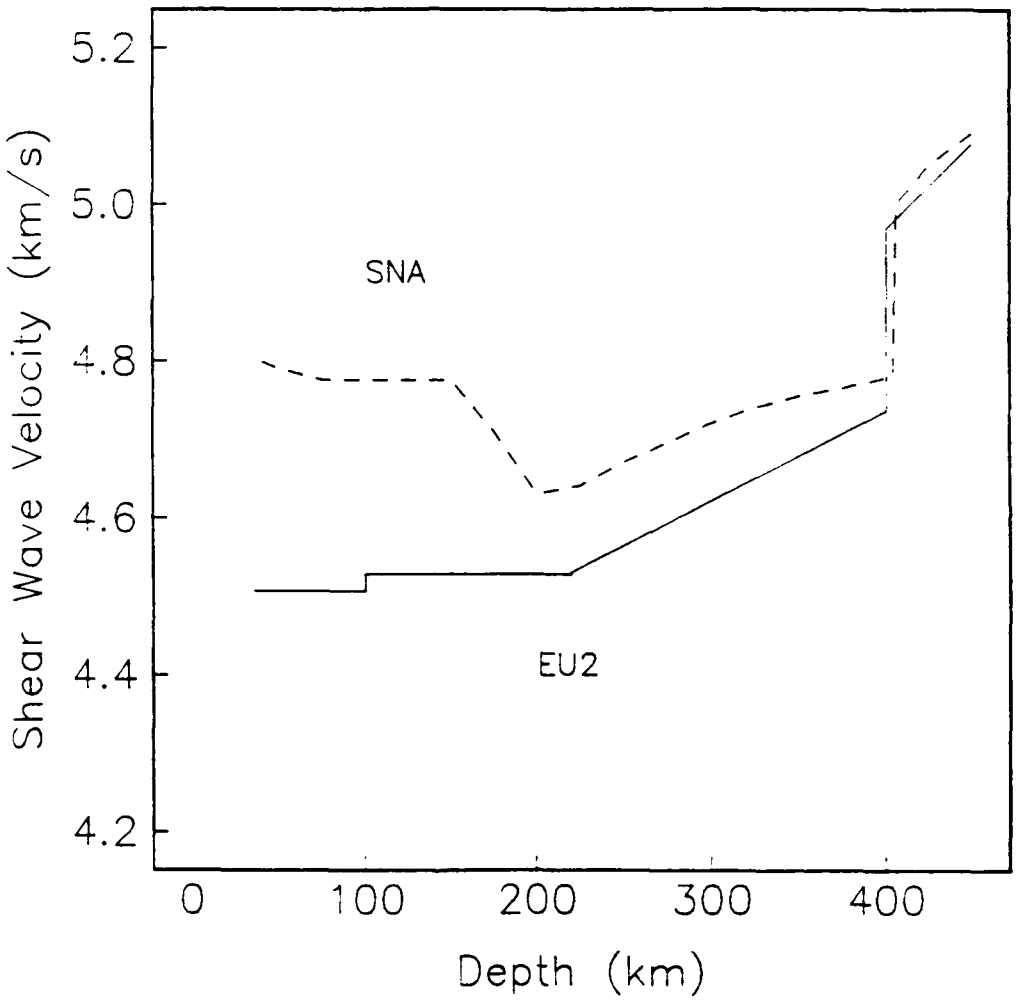

5.2
5.0
4.8
4.6
4.4
4.2
Shear Wave Velocity (km/s)
SNA
EU2
0
100
200
300
400
Depth (km)

Figure 1.1

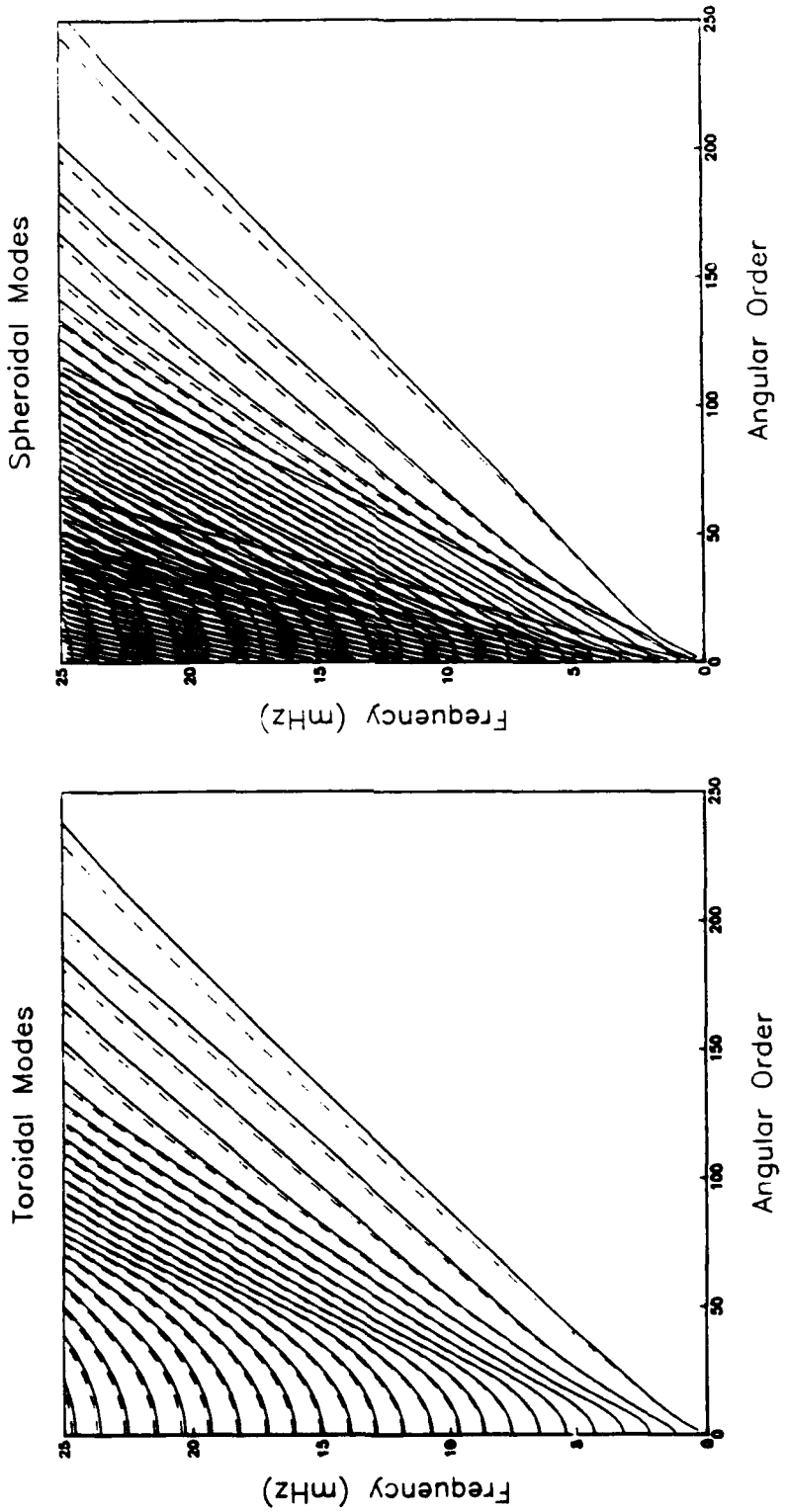

Spheroidal Modes
Angular Order
Frequency (mHz)
Toroidal Modes
Angular Order
Frequency (mHz)

Figure 1.2

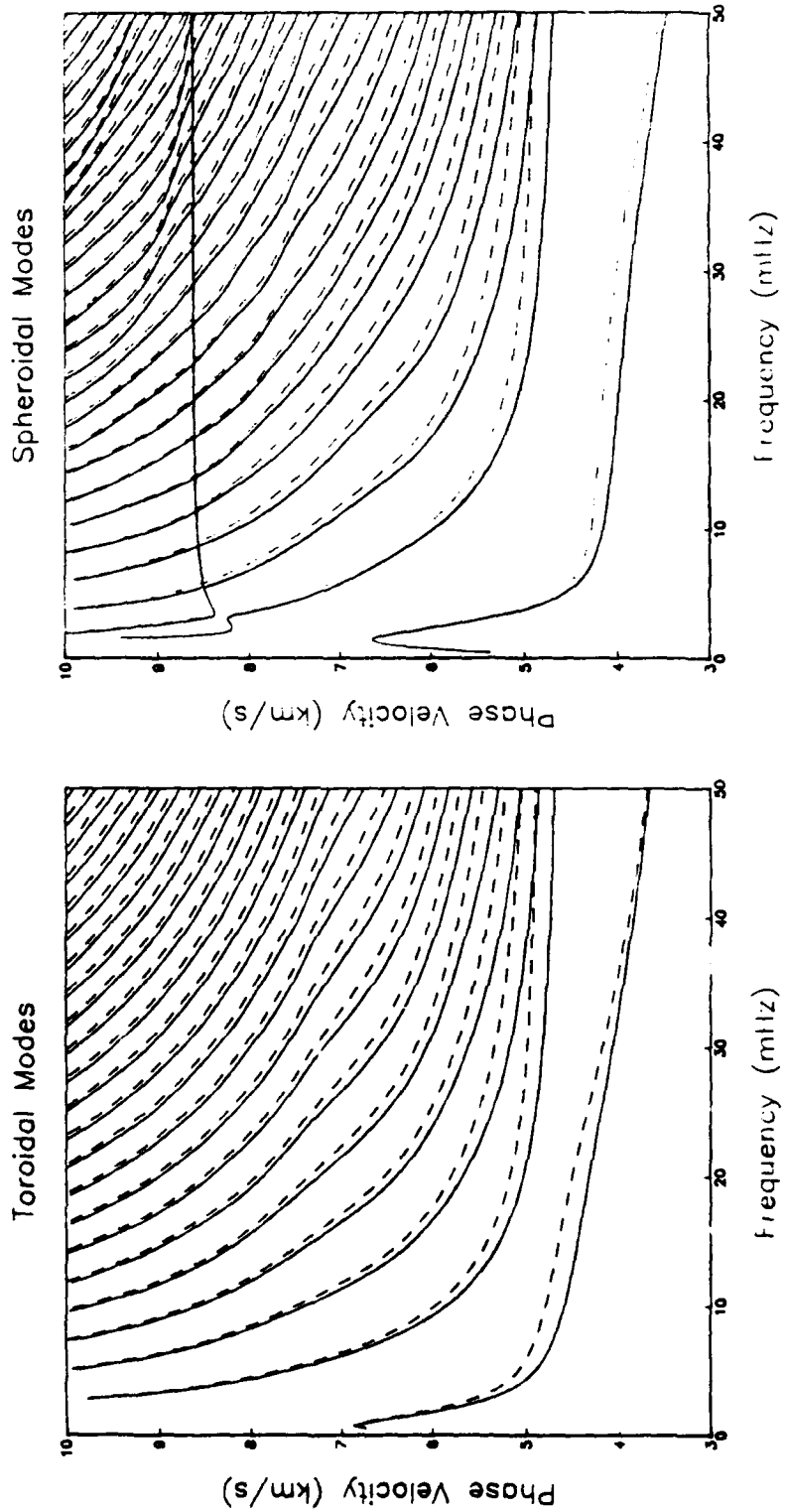

Spheroidal Modes
Phase Velocity (km/s)
Frequency (mHz)
Toroidal Modes
Phase Velocity (km/s)
Frequency (mHz)

Figure 1.3

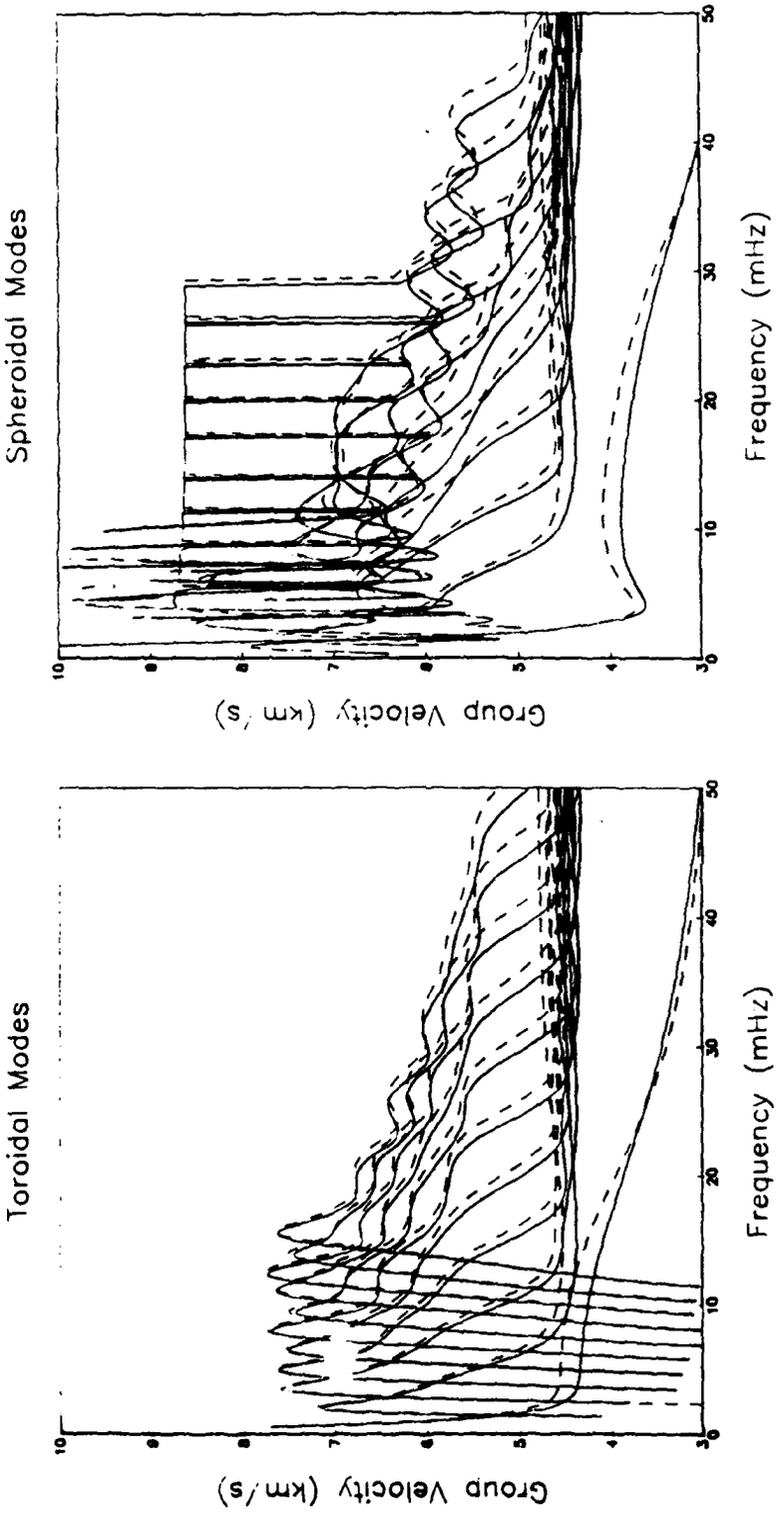

Spheroidal Modes
Frequency (mHz)
Group Velocity (km/s)
Toroidal Modes
Frequency (mHz)
Group Velocity (km/s)

Figure 1.4

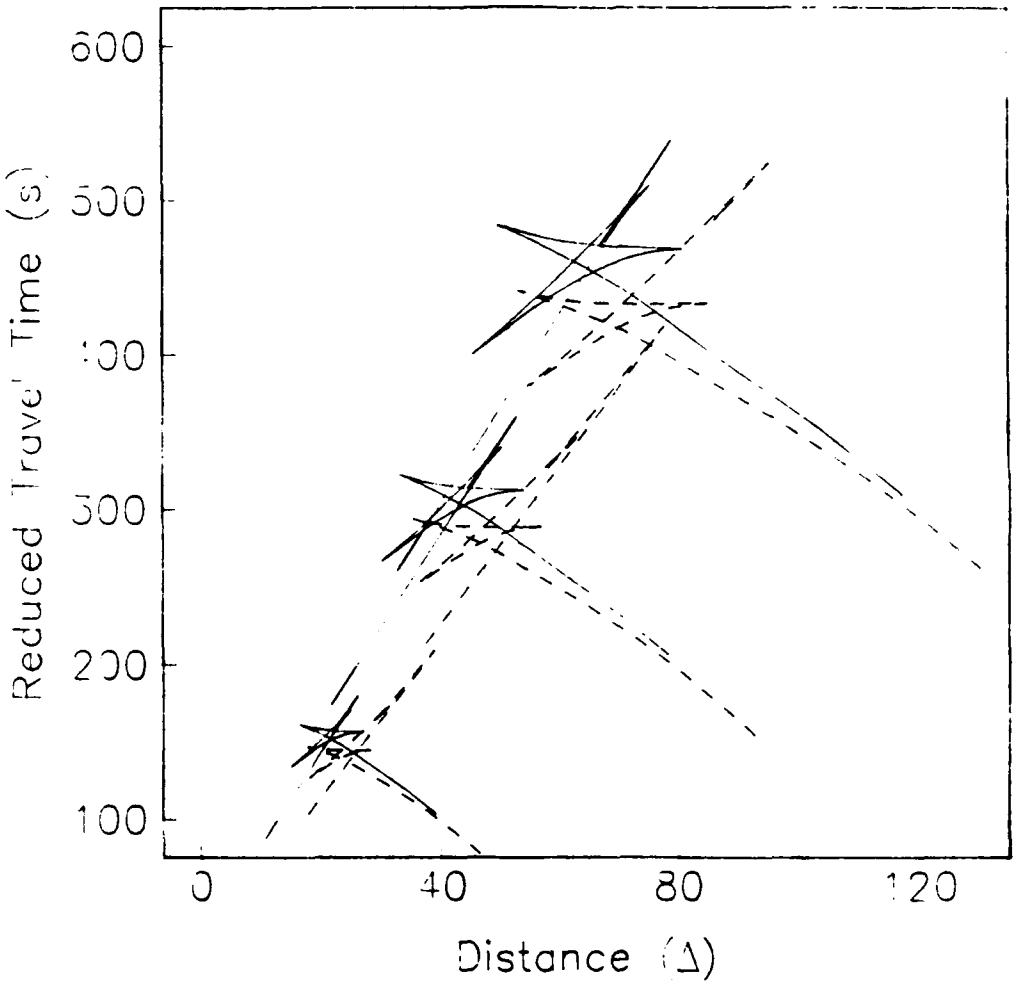

600
500
400
300
200
100
Reduced Travel Time (s)
0
40
80
120
Distance (Δ)

Figure 1.5

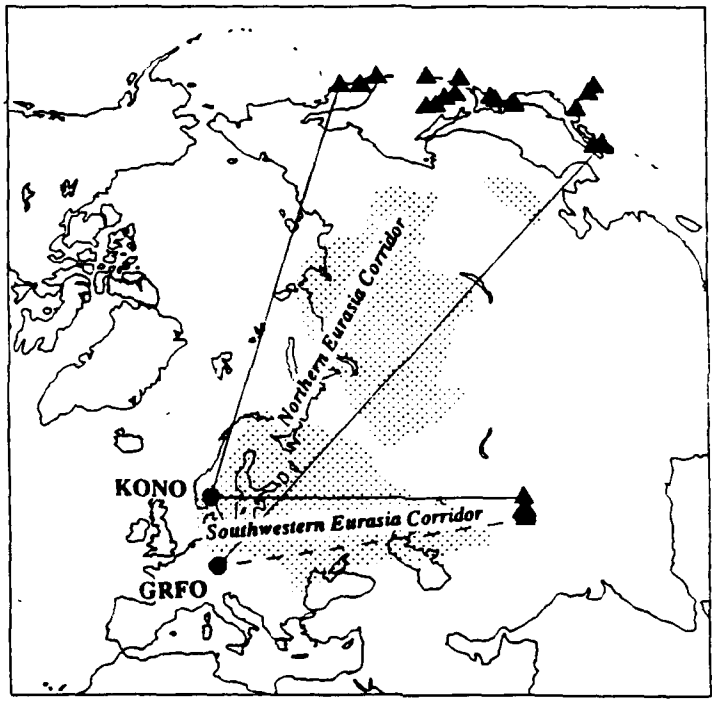

Northern Eurasia Corridor
KONO
Southwestern Eurasia Corridor
GRFO

Figure 1.6

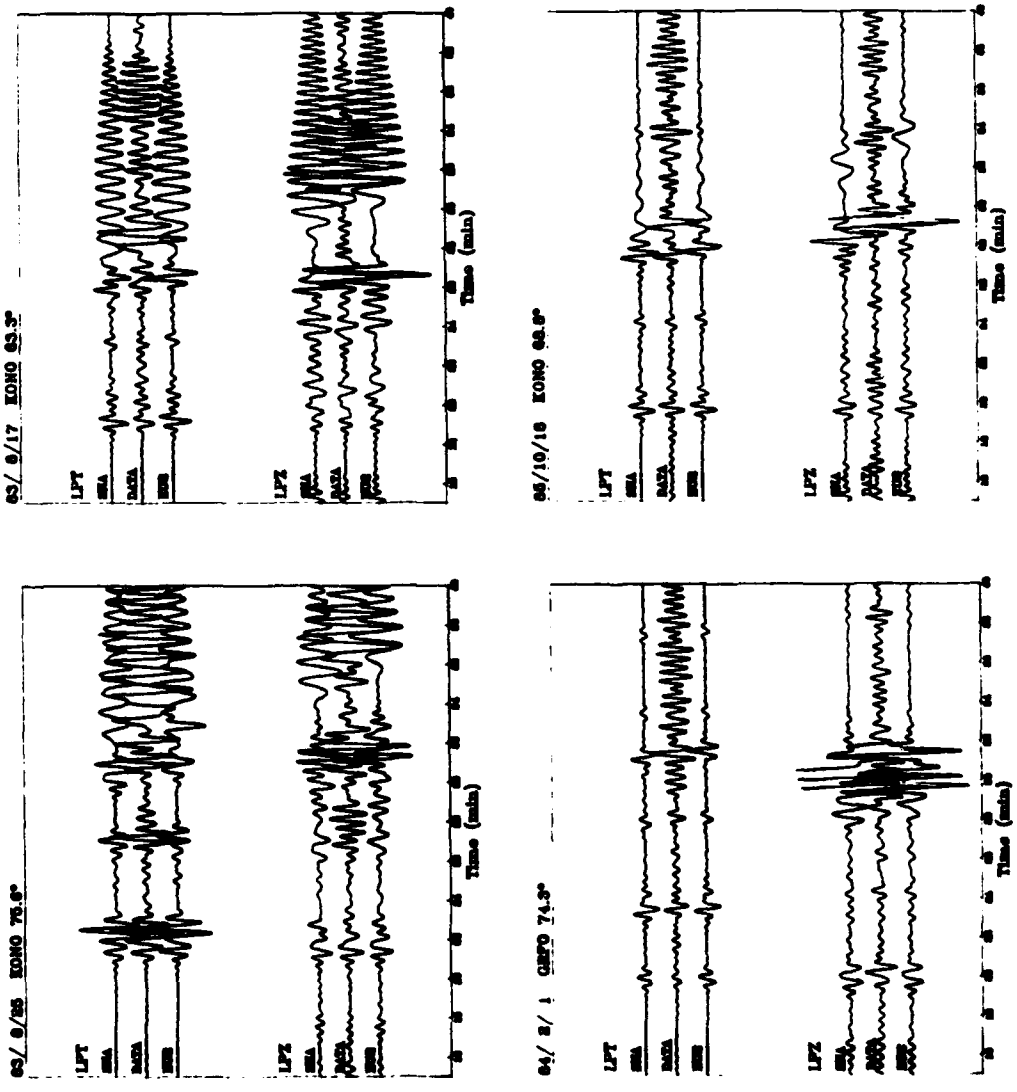


Figure 1.7

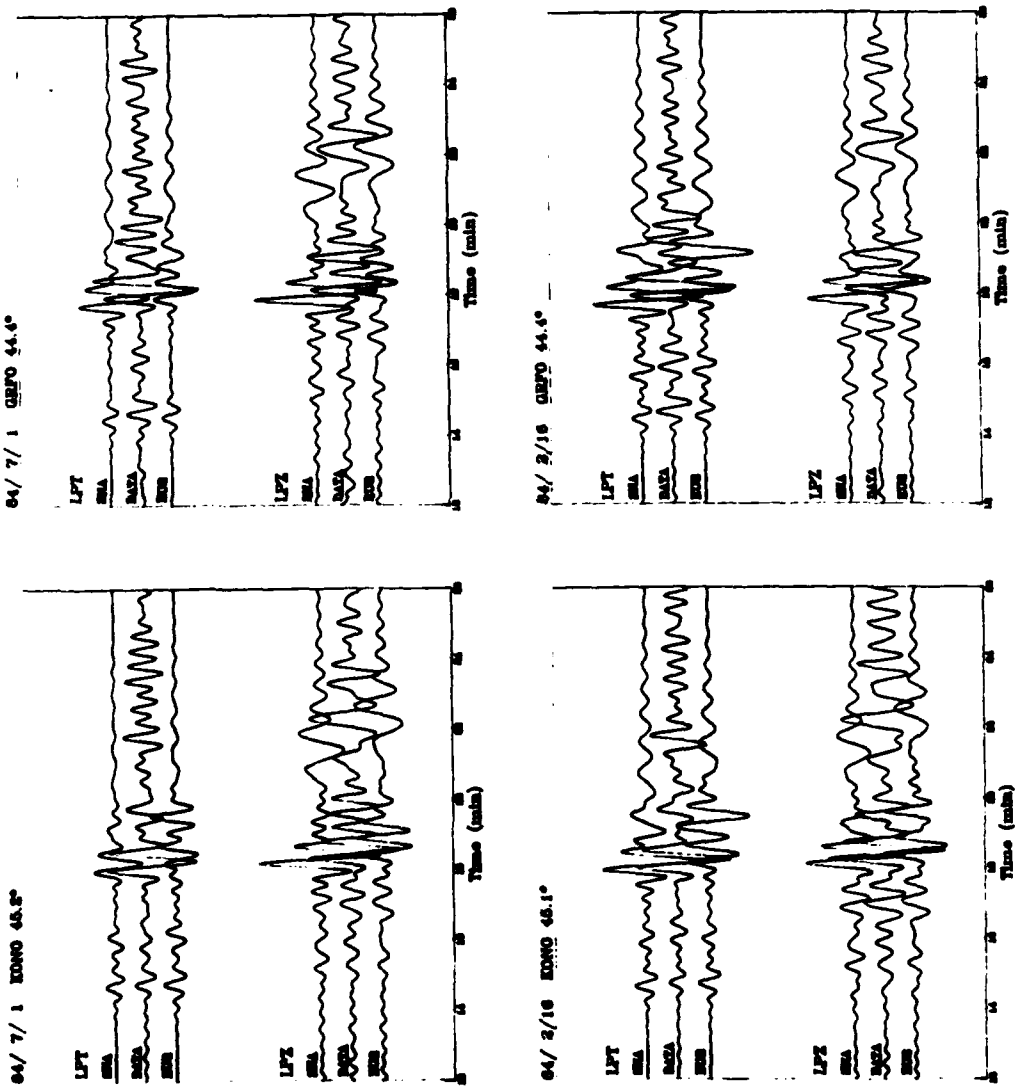

84/ 7/ 1 QUPO 44.6°
84/ 2/16 QUPO 44.6°
84/ 7/ 1 KONO 46.8°
84/ 2/16 KONO 46.1°
LPT
LPZ
Time (min)

Figure 1.8

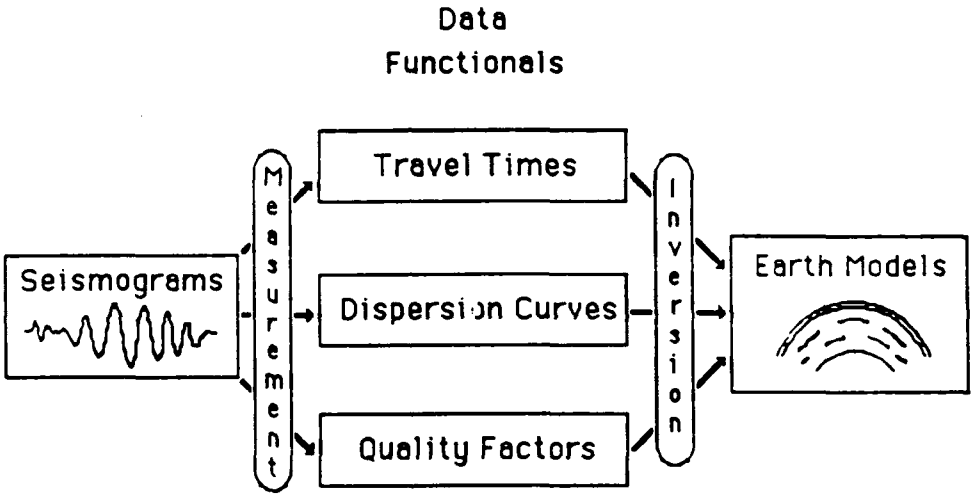

Data
Functionals
Seismograms
Measurement
Travel Times
Dispersion Curves
Quality Factors
Inversion
Earth Models

Figure 1.9

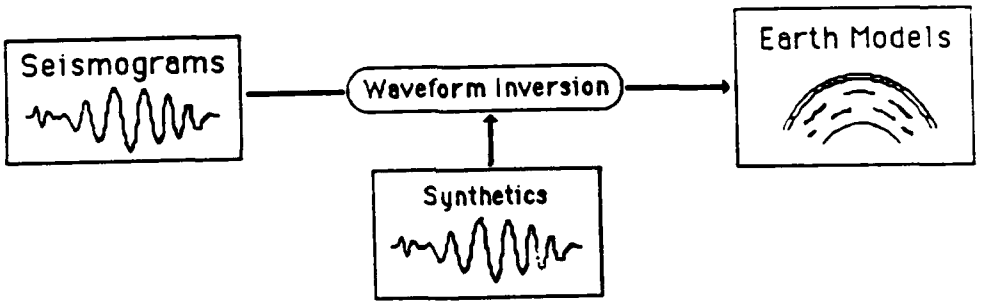


Figure 1.10

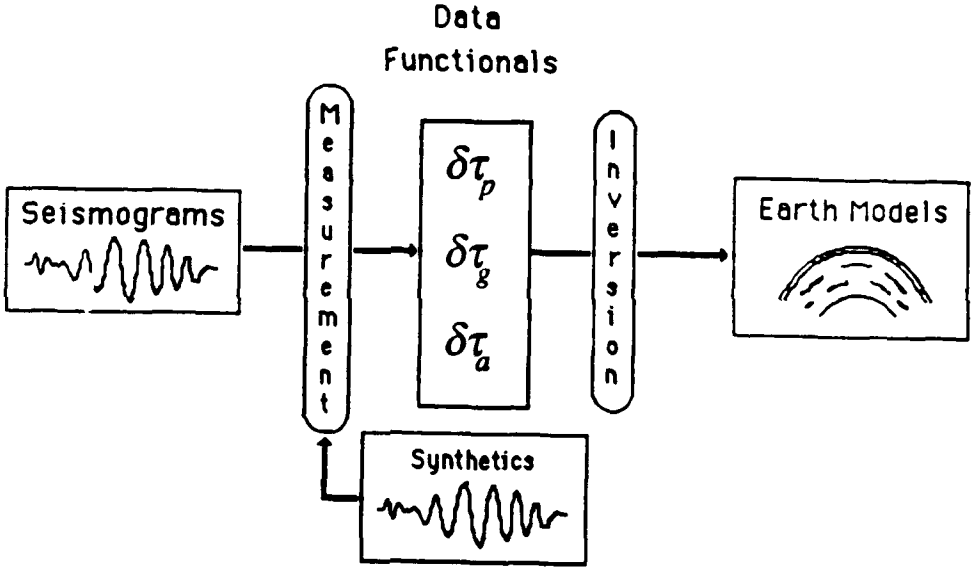

Data
Functionals
Seismograms
Measurement
$\delta\tau_p$
$\delta\tau_g$
$\delta\tau_a$
Inversion
Earth Models
Synthetics

Figure 1.11

CHAPTER 2

WAVEFORM ANALYSIS OF NARROW-BAND SEISMOGRAMS

INTRODUCTION

Cross-correlation has played an important role in structural seismology ever since advances in digital computing made the Fourier transform a standard time-series analysis tool. In particular, analysis of the cross-correlation function has two distinct advantages over processing of the individual signals. First, the Fourier phase spectrum of the cross-correlation between two real time series, $g(t)$ and $h(t)$, is simply the phase difference of the input time series:

$$\begin{aligned} g(t) \otimes h(t) &= \int_{-\infty}^{\infty} g(\tau)\, h(\tau + t) d\tau \\ &= \frac{1}{2\pi} \int_{-\infty}^{\infty} g^*(\omega)\, h(\omega)\, e^{-i\omega t}\, d\omega \\ &= \frac{1}{2\pi} \int_{-\infty}^{\infty} |g(\omega)|\,|h(\omega)| \exp[i(\varphi_h(\omega) - \varphi_g(\omega) - \omega t)] d\omega \end{aligned} \qquad (2.1)$$

where the superscript * denotes the complex-conjugate operator and we have assumed the Fourier convention $g(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} g(\omega) \exp(-i\omega t)\, d\omega$. φ_h is the phase spectrum of $h(\omega)$,

φ_g is the phase spectrum of $g(\omega)$, and $\varphi_h - \varphi_g$ is the phase spectrum of the cross-correlation function. Second, the cross-correlation function has a higher signal-to-noise ratio than the individual signals, due to cancellation of the uncorrelated noise. These properties have been exploited by a number of investigators in seismology. Aki [1964] cross-correlated transverse and radial component records at the same station in order to obtain the phase difference between Love and Rayleigh waves. Landisman *et al.* [1969] determined inter-station phase and group velocities of surface waves using cross-correlation. Dziewonski and Landisman [1970] computed the autocorrelation of a seismogram and used multiple-filter analysis and time-variable filtering to measure the differential phase and group delay between mantle waves such as R_5 and R_3. Cross-correlation has also been applied to body-wave studies for the estimation of differential travel-times, such as ScS_2-ScS [Okal and Anderson, 1975; Sipkin and Jordan, 1976; Stark and Forsyth, 1983] and SS-S [Stark and Forsyth, 1983; Kuo *et al.*, 1987; Woodward and Masters, 1989; Sheehan and Solomon, 1989].

Synthetic seismograms have expanded the application of cross-correlation. Aki [1960] defined a "phase equalization" technique where he cross-correlated a waveform which included the effects of propagation and instrument response in order to study the source function. Dziewonski *et al.* [1972] introduced the residual dispersion analysis of surface waves for the recovery of group velocity, based the cross-correlation of a synthetic mode and the observed seismogram. Herrin and Goforth [1977, 1986] developed the phase-matched filtering technique to recover group-velocity curves by the iterative cross-correlation of a filter computed from a trial group-velocity curve and an observed seismogram. Lerner-Lam and Jordan [1983] developed a waveform-inversion formalism for fundamental and higher mode data, using cross-correlation to isolate the traveling wave branch of interest. Cara and Lévêque [1987] parameterized the envelope of cross-correlation between a synthetic mode and the observed seismogram in terms of "secondary observables" such as attenuation and group delay. Cross-correlation of synthetic

seismograms with data have also been used to measure differential body-wave travel times such as *S* [Hart, 1975; Hart and Butler, 1978] and *SS* [Butler, 1979].

In this chapter, we present the tools which form the core of our waveform-analysis procedure. We introduce an isolation filter composed of a single waveform and expand its autocorrelation in terms of Hermite polynomials in the frequency domain. We show that the expansion may truncated at second order if the spectrum is strongly peaked and accurately described by a few of its low-order moments, a situation which may be enforced by the application of a narrow-band filter. Using a first-order Taylor series expansion of differential wavenumber, we derive a formula for the cross-correlation of the isolation filter with the observed seismogram. The formula is parameterized in terms of a differential phase delay, differential group delay, and differential amplitude factor. We develop this methodology first for the example of an isolated waveform, where the interaction or cross terms between the isolation filter and the synthetic and observed seismograms may be neglected. We illustrate the technique with an example and demonstrate how the recovered parameters may be interpreted using Rayleigh's principle. We then relax the isolated-waveform restriction and derive expressions for the contribution of the cross terms. We conclude the chapter with a discussion on the implementation of this methodology and the interpretation of the observable parameters.

The technique is best described in the context of a simple numerical experiment. Consider the two continental structures discussed in Chapter 1: SNA, a model of stable North America derived by forward modeling of *SH*-polarized waveforms of multiply reflected shear waves [Grand and Helmberger, 1984], and EU2, a model of northern Eurasia derived by waveform inversion of fundamental Rayleigh waves and *PSV*-polarized higher modes [Lerner-Lam and Jordan, 1987]. The average upper-mantle shear velocity in SNA is significantly greater than in EU2, evidently due in part to genuine path differences and in part to strong polarization anisotropy in the upper mantle [Lerner-Lam and Jordan, 1987; Gee and Jordan, 1988]. These hypotheses are discussed in Appendix F; here we

simply consider the two models to be representative of differences in regional upper-mantle structure that a seismologist might seek to determine by waveform measurements. The specific problem we pose is to estimate the Love and Rayleigh-wave dispersion of SNA using EU2 as a reference model. This is a rigorous test of our methodology, since the differences between seismograms predicted by these models are greater than those between observed seismograms and the appropriate synthetic seismograms (recall Figures 1.7 and 1.8).

METHODOLOGY

Isolation filter

The basis of the waveform-analysis procedure described in this the is is the construction of an isolation filter $\tilde{f}(t)$, defined in (1.4) as a weighted sum over synthetic waveforms:

$$\tilde{f}(t) = \sum_{m=0} \tilde{\alpha}_m(t) * \tilde{u}_m(t) \tag{2.2}$$

In our formulation, we consider the $\tilde{u}_m(t)$ to be individual traveling-wave branches [Gilbert, 1976a; Lerner-Lam and Jordan, 1983], calculated for a chosen reference model $\mathbf{m}_0$, and the $\tilde{\alpha}_m$ are frequency-dependent weight coefficients. In this chapter, we focus on the special case of an isolation filter that corresponds to a single waveform, such as an individual surface wave, and take the coefficients $\tilde{\alpha}_m$ to be zero for all but one value of m, where it is assumed to be unity:

$$\tilde{f}(t) = \tilde{u}_m(t) \tag{2.3}$$

For the problem posed above, therefore, $\tilde{f}(t)$ is simply the fundamental-mode surface wave, calculated from the reference model EU2. Examples of transverse and vertical-component synthetic seismograms calculated from EU2 and SNA corresponding to an earthquake in Kamchatka (83/08/17, h = 77 km) recorded at the Global Digital Seismic Network (GDSN) station KONO in Norway ($\Delta = 63°$) are plotted in Figure 2.1.

Isolated-waveform approximation

The cross-correlation of the single-waveform isolation filter with the complete synthetic seismogram may be written as the sum of two terms:

$$C_{\tilde{u}\tilde{s}}(t) \equiv \tilde{u}_m(t) \otimes \tilde{s}(t) = \int_{-\infty}^{\infty} \tilde{u}_m(\tau)\, \tilde{s}(\tau+t)\, d\tau$$

$$= \int_{-\infty}^{\infty} \tilde{u}_m(\tau)\, \tilde{u}_m(\tau+t)\, d\tau + \sum_{\substack{n=0 \\ (n \neq m)}}^{N} \int_{-\infty}^{\infty} \tilde{u}_m(\tau)\, \tilde{u}_n(\tau+t)\, d\tau$$

$$= \frac{1}{2\pi}\int_{-\infty}^{\infty} |\tilde{u}_m(\omega)|^2\, e^{-i\omega t}\, d\omega + \frac{1}{2\pi} \sum_{\substack{n=0 \\ (n \neq m)}}^{N} \int_{-\infty}^{\infty} \tilde{u}_m^*(\omega)\, \tilde{u}_n(\omega)\, e^{-i\omega t}\, d\omega \quad (2.4)$$

The first term in (2.4) is the autocorrelation term of $C_{\tilde{u}\tilde{s}}(t)$ while the second term represents interference or cross-term contributions. If $\tilde{u}_m(t)$ is isolated on the seismogram from interfering arrivals, then the autocorrelation term will dominate the sum near the peak of the cross-correlation, and we may make an "isolated-waveform approximation" by neglecting the interaction terms:

$$C_{\tilde{u}\tilde{s}}(t) \approx C_{\tilde{u}\tilde{u}}(t) \quad (2.5)$$

where $C_{\tilde{u}\tilde{u}}(t)$ is the autocorrelation of $\tilde{u}_m(t)$. This approximation is only valid for t near $t = 0$, as the cross-term contributions will be significant away from the peak of the correlation function. We shall abandon this approximation later in this chapter and discuss the contribution of the interference terms.

Autocorrelation function

$C_{\tilde{u}\tilde{u}}(t)$ is a symmetric function peaked at zero lag (Figure 2.2) whose real-valued Fourier spectrum $C_{\tilde{u}\tilde{u}}(\omega)$ is the squared modulus of the complex Fourier spectrum $\tilde{u}_m(\omega)$ ($C_{\tilde{u}\tilde{u}}(\omega) \equiv \tilde{u}_m^*(\omega)\,\tilde{u}_m(\omega) = |\tilde{u}_m(\omega)|^2$). Before proceeding further, it will be useful to develop the properties of this autocorrelation function in terms of Hermite polynomials. We shall introduce them briefly here, deferring an extended discussion of their properties to Appendices A and B.

We expand the spectrum $C_{\tilde{u}\tilde{u}}(\omega)$ on the positive ω-axis in terms of an unnormalized Gaussian, $\mathrm{Ga}(x) \equiv \exp(-x^2/2)$, multiplied by a sum over Hermite polynomials, $\mathrm{He}_k(x)$:

$$C_{\tilde{u}\tilde{u}}(\omega)\,\mathrm{H}(\omega) = \frac{1}{\sqrt{2\pi}\,\tau_f}\,\mathrm{Ga}\left(\frac{\omega-\omega_f}{\sigma_f}\right)\sum_{k=0}^{\infty} a_k\,\mathrm{He}_k\left(\frac{\omega-\omega_f}{\sigma_f}\right) \tag{2.6}$$

where $\mathrm{H}(\omega)$ is the Heaviside step function, ω_f is a location parameter, and σ_f is a scale parameter. This expansion is based on the approximation that $C_{\tilde{u}\tilde{u}}(\omega)$ is small near $\omega = 0$. The Hermite polynomial of degree k is defined by the expression

$$\mathrm{He}_k(x) = k!\sum_{j=0}^{[k/2]} \frac{(-1/2)^j\,x^{k-2j}}{j!\,(k-2j)!} \tag{2.7}$$

where $[k/2]$ is the largest integer $\leq k/2$. Sansone [1959], Szegö [1959], and Lebedev [1965] give further definitions and properties; in particular, on the interval $(-\infty,\infty)$ the functions $\{He_k(x) : k = 0, 1, 2, ...\}$ are complete and orthogonal with respect to the Gaussian weight $Ga(x)$. The coefficients in (2.6) are real and given by

$$a_k = \frac{1}{k!} \int_0^{\infty} C_{\tilde{u}\tilde{u}}(\omega)\ He_k\left(\frac{\omega-\omega_f}{\sigma_f}\right) d\omega \tag{2.8}$$

and may be written in terms of spectral moments of $C_{\tilde{u}\tilde{u}}(\omega)$ [Rietz, 1971]:

$$a_k = \sum_{m=0}^{[k/2]} \frac{(-1/2)^m\ \hat{\mu}_{k-2m}(\omega_f)}{m!\,(k-2m)!} \tag{2.9}$$

where $\hat{\mu}_p(\omega_f)$ is the pth normalized one-sided moment of $C_{\tilde{u}\tilde{u}}(\omega)$ about $\omega = \omega_f$

$$\hat{\mu}_p(\omega_f) = \frac{1}{\sigma_f^p} \int_0^{\infty} C_{\tilde{u}\tilde{u}}(\omega)\ (\omega-\omega_f)^p\ d\omega \tag{2.10}$$

If $C_{\tilde{u}\tilde{u}}(\omega)$ is normalized such that $\hat{\mu}_0(0) = 1$, then $a_0 = 1$. If the location parameter ω_f is chosen to be the center frequency of $C_{\tilde{u}\tilde{u}}(\omega)$ (*i.e.*, $\omega_f = \sigma_f \hat{\mu}_1(0)$), then $a_1 = 0$. Finally, if the scale parameter σ_f is the half-width of $C_{\tilde{u}\tilde{u}}(\omega)$ (*i.e.*, $1 = \hat{\mu}_2(\omega_f)$), then $a_2 = 0$. With these choices, the zeroth-order term of the Hermite-polynomial expansion represents the Gaussian approximation to $C_{\tilde{u}\tilde{u}}(\omega)$, and the first correction term is third-order. We shall assume this representation throughout the thesis.

The expansion described by (2.6), (2.7), and (2.8) is known as a Gram-Charlier series in statistics [Jackson, 1961; Rietz, 1971] and is used to represent arbitrary

distributions in terms of the derivatives of the normal distribution. Further details about the representation of real functions using Gram-Charlier series may be found in Appendix A, which catalogs theorems for the manipulation of Hermite polynomials and discusses the higher-order terms of the expansion.

This representation of $C_{\tilde{u}\tilde{u}}(\omega)$ has two advantages. First, the convenient Fourier-transform properties of Hermite polynomials [Hille, 1926] yield a simple series expression for its time-domain image:

$$
\begin{aligned}
C_{\tilde{u}\tilde{u}}(t) &= \frac{1}{(2\pi)^{3/2}} \frac{1}{\sigma_f} \mathrm{Re}\left\{ \int_{-\infty}^{\infty} \mathrm{Ga}\left(\frac{\omega-\omega_f}{\sigma_f}\right) \sum_{k=0}^{\infty} a_k \mathrm{He}_k\left(\frac{\omega-\omega_f}{\sigma_f}\right) e^{-i\omega t}\, d\omega \right\} \\
&= \frac{1}{2\pi} \mathrm{Ga}(\sigma_f t) \sum_{k=0}^{\infty} a_k\, \gamma_f^k\, (\omega_f t)^k \cos(\omega_f t + k\tfrac{\pi}{2}) \\
&= \frac{1}{2\pi} \mathrm{Ga}(\sigma_f t)\ \{\cos(\omega_f t)\,[1 + a_4\, \gamma_f^4\, (\omega_f t)^4 - a_6\, \gamma_f^6\, (\omega_f t)^6 + \cdots] \\
&\qquad + \sin(\omega_f t)\,[a_3\, \gamma_f^3\, (\omega_f t)^3 - a_5\, \gamma_f^5\, (\omega_f t)^5 + \cdots]\ \}
\end{aligned}
\tag{2.11}
$$

where $\gamma_f \equiv \sigma_f/\omega_f$ is a measure of the relative bandwidth of the signal. This expression has the form of a Gaussian envelope with half-width σ_f^{-1} multiplied by harmonic carrier function and a power series in t. Second, because $C_{\tilde{u}\tilde{u}}(\omega)$ is expected to be peaked in the pass-band of the instrument response, (2.11) can usually be approximated by its first few terms. If terms of third and higher order may be neglected, then we obtain the standard "wave-packet" approximation [*e.g.*, Bracewell, 1978] by neglecting terms of order γ_f^3:

$$
C_{\tilde{u}\tilde{u}}(t) = E(t) \cos\Phi(t) \tag{2.12}
$$

where $E(t)$ and $\Phi(t)$ are defined by

$$E(t) \equiv \frac{1}{2\pi}\,\mathrm{Ga}(\sigma_f t) \tag{2.12a}$$

$$\Phi(t) \equiv \omega_f t \tag{2.12b}$$

In other words, the autocorrelation function of the isolation filter can be represented as a cosinusoidal "carrier" modulated by a Gaussian "envelope." The carrier frequency is the center frequency of $C_{\tilde{u}\tilde{u}}(\omega)$, ω_f, and the half-width of the envelope is the inverse of the spectral half-width, σ_f. Because equation (2.12) corresponds to taking the spectrum $C_{\tilde{u}\tilde{u}}(\omega)$ to be Gaussian, we refer to it as the "Gaussian-wavelet approximation."

Figure 2.2 compares the actual autocorrelation function computed from the fundamental-mode Love and Rayleigh waves (solid line) with that obtained from the Gaussian-wavelet approximation (dashed line). The fit is quite good near zero lag, degrading at the shoulders of the correlation functions due to the non-Gaussian character of the spectra. Since seismic signals do not have narrow-band Gaussian spectra in general, we enforce the Gaussian approximation by convolving a series of narrow-band filters $\{F_i: i = 1, \ldots, I\}$ with varying center frequencies ω_i and half-widths σ_i to the broad-band correlation $C_{\tilde{u}\tilde{u}}(t)$:

$$F_i C_{\tilde{u}\tilde{u}}(t) \equiv F_i(t) * C_{\tilde{u}\tilde{u}}(t) \tag{2.13}$$

where the $F_i C_{\tilde{u}\tilde{u}}(t)$ is the filtered autocorrelation function. We may represent an arbitrary filter by expanding $F_i(\omega)$ along the positive axis in terms of Hermite polynomials:

$$F_i(\omega)\,\mathrm{H}(\omega) = \frac{1}{\sqrt{2\pi}\,\sigma_i}\,\mathrm{Ga}\left(\frac{\omega-\omega_i}{\sigma_i}\right)\sum_{m=0}^{\infty} f_m\,\mathrm{He}_m\left(\frac{\omega-\omega_i}{\sigma_i}\right) \tag{2.14}$$

where the coefficients f_m are linear combinations of the spectral moments of $F_i(\omega)$. We have worked out the details of filtering with Gram-Charlier expansions in Appendix B. making use of the theorems developed in Appendix A. In the present application, we take $F_i(\omega)\mathrm{H}(\omega)$ to be exactly Gaussian

$$F_i(\omega)\ \mathrm{H}(\omega) = \frac{1}{\sqrt{2\pi}\ \sigma_i}\ \mathrm{Ga}\left(\frac{\omega-\omega_i}{\sigma_i}\right) \tag{2.15}$$

and assume that $F_i(\omega)$ is small near $\omega = 0$. The expression for the filtered autocorrelation resembles that of the broad-band correlation function, but is written as a power series in γ_i' $\equiv \sigma_i' / \omega_i'$:

$$F_i C_{\tilde{u}\tilde{u}}(t) = \frac{1}{(2\pi)^{3/2}} \frac{\sigma_i'}{\sigma_i\ \sigma_f}\ \mathrm{Ga}\left(\frac{\omega_i-\omega_f}{\sqrt{\sigma_i^2+\sigma_f^2}}\right) \mathrm{Ga}\,(\sigma_i' t) \sum_{l=0}^{\infty} b_l\ \gamma_i'^{\,l}\ (\omega_i' t)^l \cos\,(\omega_i' t + l\,\frac{\pi}{2}) \tag{2.16}$$

where ω_i' is the weighted center frequency

$$\omega_i' = \frac{\sigma_i^2\omega_f + \sigma_f^2\omega_i}{\sigma_i^2 + \sigma_f^2} \tag{2.17}$$

and σ_i' is the weighted half-width

$$\sigma_i'^2 = \frac{\sigma_i^2\sigma_f^2}{\sigma_i^2 + \sigma_f^2} \tag{2.18}$$

Using the scale and shift theorems developed in Appendix A, we can manipulate the Hermite polynomials and compute an expression for the b_l coefficients. The details of this

formulation may be found in Appendix B; the coefficients depend on the parameters of $C_{\tilde{u}\tilde{u}}$ and the filter:

$$b_l = \left(\frac{\sigma_i}{\sqrt{\sigma_i^2+\sigma_f^2}}\right)^l \sum_{j=0}^{\infty} \binom{j+l}{j} \left(\frac{\sigma_f(\omega_i-\omega_f)}{\sigma_i^2+\sigma_f^2}\right)^j \sum_{m=0}^{\infty} a_{2m+j+l} \frac{(2m+j+l)!\,(-1/2)^m}{(j+l)!\,m!} \left(\frac{\sigma_f^2}{\sigma_i^2+\sigma_f^2}\right)^m \quad (2.19)$$

The sum over m arises from the scale theorem (A.11) and converges for all values of σ_f and σ_i. The sum over j proceeds from the shift theorem (A.7), and is well-behaved as long as $\sigma_f(\omega_f-\omega_i) < (\sigma_f^2+\sigma_i^2)$, which is simply a requirement that the filter center frequency lie within the bandwidth of the signal. Finally, the b_l depend on the ratio $\sigma_i^2 / (\sigma_f^2+\sigma_i^2)$, which is always less than 1.

If $\gamma_i' << 1$, we may make the Gaussian-wavelet approximation by neglecting terms of order $\gamma_i'^3$:

$$F_i C_{\tilde{u}\tilde{u}}(t) = E(t)\cos\Phi(t) \quad (2.20)$$

where $E(t)$ and $\Phi(t)$ are defined by

$$E(t) = \frac{1}{(2\pi)^{3/2}} \frac{\sigma_i'}{\sigma_i\,\sigma_f} \mathrm{Ga}\left(\frac{\omega_i-\omega_f}{\sqrt{\sigma_i^2+\sigma_f^2}}\right) \mathrm{Ga}(\sigma_i' t) \quad (2.20a)$$

$$\Phi(t) = \omega_i' t \quad (2.20b)$$

Figure 2.3 illustrates the comparison between (2.20) (dashed line) and the filtered autocorrelation of the fundamental-mode Love and Rayleigh waves (solid line) for several filters.

γ_i' is the fundamental control parameter in the Gaussian-wavelet approximation (2.20). It measures the narrowness of the spectrum relative to its center frequency. In the case of the unfiltered autocorrelation function, $\gamma_i' = \gamma_f$, which is of order 0.25 for our fundamental-mode example. From Figure 2.2, we see that the Gaussian-wavelet approximation is quite good for one cycle about the peak of the $C_{\tilde{u}\tilde{u}}(t)$, but breaks down at greater lag. We generally apply filters with a fixed ratio of σ_i / ω_i of 0.1. The resulting values of γ_i' are near 0.1, which improves the time-domain fit of the Gaussian-wavelet approximation considerably (Figure 2.3). Implicit in this approximation is the assumption that the applied filter may be described by a few of its low-order moments. The application of an extremely narrow boxcar filter would not be well-described by the Gaussian-wavelet approximation.

Cross-correlation function

Now we consider the cross-correlation between the isolation filter and the observed seismogram, $C_{\tilde{u}s}(t) = \tilde{u}_m(t) \otimes s(t)$. If $u_m(t)$ is reasonably well-separated from other energy on the observed seismogram, then we may neglect the interaction terms near the peak of the cross-correlation function, as we did in the case of the complete synthetic seismogram:

$$C_{\tilde{u}s}(t) \approx C_{\tilde{u}u}(t) \tag{2.21}$$

$C_{\tilde{u}u}(t)$ will not be the same as $C_{\tilde{u}\tilde{u}}(t)$ (Figure 2.4) due to differences in propagation, which we hope to measure, as well as differences between the actual and assumed source function and instrument response, which we ignore, at least for the present discussion. We suppose the actual waveform u_m has a spectrum that is related to the synthetic $\tilde{u}_m$ by a differential response operator D_m,

$$u_m(\omega) = D_m(\omega)\,\tilde{u}_m(\omega) \tag{2.22}$$

With this parameterization, the expression for the filtered cross-correlation function $F_iC_{\tilde{u}u}(t)$ has the following form:

$$F_iC_{\tilde{u}u}(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} C_{\tilde{u}\tilde{u}}(\omega)\,D_m(\omega)\,F_i(\omega)\,e^{-i\omega t}\,d\omega \tag{2.23}$$

At this point that we adopt a traveling wave, rather than standing wave, representation of the seismogram and write the differential response in the form $D_m(\omega) = D_0^m \exp[i\delta k_m(\omega)x]$, where D_0^m is a real-valued constant, x is the propagation distance, and $\delta k_m(\omega) \equiv k_m(\omega) - \tilde{k}_m(\omega)$ is the complex-valued differential wavenumber between the mth observed and synthetic branches. Since we are interested in measuring the differential dispersion as a function of frequency, we expand the latter in a Taylor series about the center frequency of the filtered autocorrelation:

$$\begin{aligned}\delta k_m(\omega)\,x &\approx [\delta k_m(\omega_i') + (\omega-\omega_i')\,\dot{\delta k}_m(\omega_i') + \frac{1}{2}(\omega-\omega_i')^2\,\ddot{\delta k}_m(\omega_i') + \cdots]\,x \\ &= \omega_i'\delta\tau_p^m(\omega_i') + (\omega-\omega_i')[\delta\tau_g^m(\omega_i') + i\,\delta\tau_a^m(\omega_i')] + \frac{1}{2}(\omega-\omega_i')^2\,\delta\tau_s^{m\,2}(\omega_i') + \cdots\end{aligned} \tag{2.24}$$

where the "superdot" denotes the derivative with respect to ω. $\delta\tau_p^m(\omega_i') \equiv x\ \mathrm{Re}\{\delta k_m(\omega_i')\} / \omega_i'$ is the differential phase delay at the center frequency ω_i'; $\delta\tau_g^m(\omega_i') \equiv x\ \mathrm{Re}\{\dot{\delta k}_m(\omega_i')\}$ is the differential group delay; $\delta\tau_a^m(\omega_i') \equiv x\ \mathrm{Im}\{\dot{\delta k}_m(\omega_i')\}$ is a differential amplitude factor; and $\delta\tau_s^{m\,2}(\omega_i') \equiv x\ \ddot{\delta k}_m(\omega_i')$ measures the differential curvature of the dispersion function and is, in general, a complex constant.

We shall consider two separate approximations to (2.24). The first, when we neglect terms greater than the second, is a quadratic-dispersion approximation. The

second, when we neglect terms greater than the first, is a linear-dispersion approximation. Figure 2.5 displays the real part of the differential dispersion between the models EU2 and SNA (solid line) with the linear (long dashes) and quadratic approximations (short dashes) for the fundamental Love and Rayleigh wave. In this example, we shall show that the linear-dispersion approximation is generally sufficient for most applications, as long as the signal is sufficiently narrow-band. However, we will return to the quadratic-dispersion approximation later in Chapter 3.

Quadratic-dispersion approximation. By neglecting terms of order higher than the second in (2.24), we make a quadratic-dispersion approximation. With this parameterization of differential wavenumber, we solve for $F_i C_{\tilde{u}_u}(t)$:

$$F_i C_{\tilde{u}_u}(t) = \mathrm{Re}\Big\{ \frac{1}{(2\pi)^{3/2}} \frac{\sigma_z}{\sigma_i \sigma_f} D_0^m \, \mathrm{Ga}\left(\frac{\omega_i - \omega_f}{\sqrt{\sigma_i^2 + \sigma_f^2}}\right) \exp[(\sigma_z \delta\tau_a^m)^2/2] \, \mathrm{Ga}\,(\sigma_z(t - \delta\tau_g^m)) \times \sum_{j=0}^{\infty} c_j \,(\sigma_z(t-\delta\tau_g^m))^j \, \exp[-i(\omega_i(t-\delta\tau_p^m) - \sigma_z^2 \delta\tau_a^m(t-\delta\tau_g^m) + j\tfrac{\pi}{2})] \Big\} \quad (2.25)$$

where the c_j coefficients are complex and depend on the coefficients of the filtered autocorrelation, the differential attenuation, and differential curvature:

$$c_j = \sum_{l=0}^{\infty} \binom{l+j}{l} b_{l+j} (-\sigma_z \delta\tau_a^m)^l \quad (2.26)$$

In this expression for the filtered cross-correlation function, the quadratic-dispersion term enters only through the effective half-width σ_z:

$$\sigma_z^2 = \frac{\sigma_i'^2 \, \sigma_s^{m\,2}}{\sigma_i'^2 + \sigma_s^{m\,2}} \quad (2.27)$$

where $\sigma_s^{m^2} \equiv i/\delta\tau_s^{m^2}$. If $\sigma_i'^2 << \sigma_s^{m^2}$, then $\sigma_z \approx \sigma_i'$, allowing us to control the effect of quadratic dispersion through the parameters of the narrow-band filter. Of course, $\delta\tau_s^{m^2}$ is proportional to the epicentral distance, which means that the effect of quadratic dispersion will increase with distance. The maximum value of $\mathrm{Re}\{\delta\tau_s^{m^2}\}$ from EU2-SNA at 70° for the fundamental Love wave is 1489 s^2 at 22 mHz; the maximum value for the fundamental Rayleigh wave is 1640 s^2 at 38 mHz. These values imply that a γ_i' of 0.1 or so (which is typical of the values we employ) is sufficient to control the contribution of quadratic dispersion in this extreme case. A more detailed discussion of the effect of quadratic dispersion is included in Appendix D.

Linear-dispersion approximation. Since the effect of quadratic dispersion can be controlled by filtering, we may approximate (2.24) by truncating the Taylor series expansion after the first term, a common assumption in seismology [Dziewonski *et al.*, 1972]. Under the linear-dispersion approximation, the complete Gram-Charlier expansion of $F_i C_{\tilde{u}\tilde{u}}(t)$ has the following form:

$$F_i C_{\tilde{u}\tilde{u}}(t) = \frac{1}{(2\pi)^{3/2}} \frac{\sigma_i'}{\sigma_i \sigma_f} D_0^m \exp[(\sigma_i' \delta\tau_a^m)^2/2] \; \mathrm{Ga}\left(\frac{\omega_i - \omega_f}{\sqrt{\sigma_i^2 + \sigma_f^2}}\right) \mathrm{Ga}(\sigma_i'(t - \delta\tau_g^m))$$

$$\times \sum_{j=0}^{\infty} d_j \; \gamma_i'^j \; (\omega_i'(t - \delta\tau_g^m))^j \cos[\omega_i'(t - \delta\tau_p^m) - \sigma_i'^2 \delta\tau_a^m (t - \delta\tau_g^m) + j\frac{\pi}{2}] \quad (2.28)$$

The coefficients d_j depend on the expansion coefficients of the filtered autocorrelation and on the differential attenuation parameter:

$$d_j = \sum_{l=0}^{\infty} \binom{l+j}{l} b_{l+j} (-\sigma_i' \delta\tau_a^m)^l \quad (2.29)$$

Figure 2.6 compares the analytic linear and quadratic approximations (equations (2.25) and (2.28)) with the filtered cross-correlation functions in the case of the EU2-SNA comparison (fundamental mode) for various values of γ_i', demonstrating that the effect of quadratic dispersion may be controlled through the appropriate choice of filter parameters.

In the situation where $\gamma_i' << 1$, we may make the Gaussian narrow-band approximation by neglecting terms of third and higher order:

$$F_i C_{\tilde{u}u}(t) = E(t)\cos\Phi(t) \tag{2.30}$$

where $E(t)$ and $\Phi(t)$ are defined by

$$E(t) = \frac{1}{(2\pi)^{3/2}} \frac{\sigma_i'}{\sigma_i \sigma_f} D_0^m \exp[(\sigma_i' \delta\tau_a^m)^2/2] \; \mathrm{Ga}\left(\frac{\omega_i - \omega_f}{\sqrt{\sigma_i^2+\sigma_f^2}}\right) \mathrm{Ga}\,(\sigma_i'(t-\delta\tau_g^m)) \tag{2.30a}$$

$$\Phi(t) = \omega_i'(t-\delta\tau_p^m) - \sigma_i'^2 \delta\tau_a^m (t-\delta\tau_g^m) \tag{2.30b}$$

As with the autocorrelation, $F_i C_{\tilde{u}u}(t)$ may be represented as a cosinusoidal "carrier" modulated by a Gaussian "envelope." The envelope function reaches a maximum at the time of the differential group delay ($t_E = \delta\tau_g^m$), a result which is the basis of cross-correlation techniques for the recovery of surface-wave dispersion, such as the residual dispersion method of Dziewonski *et al.* [1972] and the phase-matched filtering technique of Herrin and Goforth [1977]. The maximum of the carrier function, however, occurs at the time of the differential phase delay, with a correction term for differential amplitude ($t_\Phi = \delta\tau_p^m + \gamma_i'^2 \omega_i' \, \delta\tau_a^m (\delta\tau_p^m - \delta\tau_g^m) + O(\gamma_i'^4)$, with a $2n\pi/\omega_i'$ ambiguity). The peak of the cross-correlation function is determined by the equation:

$$\dot{E}(t)/E(t) = \dot{\Phi}(t) \tan\Phi(t) \qquad (2.31)$$

By linearizing the tangent near the peak of the phase function, we find an expression for δt_c, the peak of the cross-correlation function:

$$\delta t_c \approx \frac{\omega_i'^2 \delta\tau_p^m + \sigma_i'^2 \delta\tau_g^m - \omega_i' \sigma_i'^2 \delta\tau_a^m (\delta\tau_p^m - \delta\tau_g^m) + 2\pi n (\omega_i' - \sigma_i'^2 \delta\tau_a^m)}{\omega_i'^2 - 2\omega_i' \sigma_i'^2 \delta\tau_a^m + \sigma_i'^2} \qquad (2.32)$$

where we have neglected terms of order $\gamma_i'^3$. Thus, to order $\gamma_i'^2$, the peak of the cross-correlation function occurs at the time of the differential phase delay at ω_i' [Sipkin and Jordan, 1980; Jordan, 1980]. However, if the difference between $\delta\tau_p^m$ and $\delta\tau_g^m$ is greater than $2\pi/\omega_i'$, a cycle-skipping problem will arise. Figure 2.7 illustrates the dependence of the cross-correlation function upon these parameters.

Cross-correlation travel times. There has been considerable discussion in the literature [Sipkin and Jordan, 1976; Butler, 1977; Sipkin and Jordan, 1980; Jordan, 1980; Stark and Forsyth, 1983] about the meaning of cross-correlation travel times, *i.e.*, differential travel times determined by the peak of the cross-correlation function. At issue is the approach of using cross-correlation to measure the differential travel time between two phases with distinct attenuation histories. For example, Okal and Anderson [1975] and Sipkin and Jordan [1976] measured the differential travel times of multiple *ScS* pulses by cross-correlating ScS_m with ScS_n. While Sipkin and Jordan [1976] discussed the frequency dependence of travel time introduced by attenuation and noted that their values represented measurements at a specific reference frequency, Butler [1977] contended that the travel times measured by this technique are biased by attenuation with respect to those values determined by visual inspection of the seismogram.

The phase delay time of a body wave is a well-known function of frequency in the presence of attenuation [Aki and Richards, 1980]. To first order in Q^{-1}:

$$\tau_p(\omega) = \tau_p(\omega_0)\,[1 - \frac{1}{\pi Q} \ln \frac{\omega}{\omega_0}] \tag{2.33}$$

where ω_0 is the reference frequency. Using this relation, Sipkin and Jordan [1980] and Jordan [1980] showed that the peak of the cross-correlation function is a measure of the differential phase delay at the center frequency of the first-arriving pulse to first order, and thus did not require a correction for the differential attenuation. Their asymptotic result is confirmed by (2.32). The "bias" discussed by Butler [1977] is merely a measure of the frequency-dependence of travel time, since pulses with different attenuation histories will have different center frequencies. The approach of creating a "synthetic" waveform by windowing an arrival and applying an appropriate attenuation operator before cross-correlation [Stark and Forsyth, 1983; Kuo *et al.*, 1987; Woodward and Masters, 1989; Sheehan and Solomon, 1989] removes the ambiguity of the reference frequency, but is not necessary in order to interpret cross-correlation travel times. In our work, estimating the reference frequency is part of the processing operation, and we shall always refer to a travel time at a particular reference frequency.

EXAMPLE

In order to test our approach, we have applied this methodology to a synthetic example of an isolated fundamental mode (Figure 2.8). In the first step, we compute the autocorrelation of $\tilde{u}_m(t)$, apply a narrow-band filter, and fit expression (2.20) to the filtered autocorrelation function in order to estimate ω_i' and σ_i' (Figure 2.3). In the second step, we compute the cross-correlation between $\tilde{u}_m(t)$ and $u_m(t)$, apply the narrow-band filter, and fit equation (2.30) to the filtered cross-correlation function with ω_i' and σ_i' fixed in

order to estimate $\delta\tau_p^m(\omega_i')$, $\delta\tau_g^m(\omega_i')$, and $\delta\tau_a^m(\omega_i')$ (Figure 2.9). Steps one and two are performed iteratively for a series of narrow-band filters. In the final step, we correct $\delta\tau_p^m$ for cycle-skipping, a problem alluded to in the previous section. If the difference between the differential group and phase delays is more than a cycle, our estimate of $\delta\tau_p^m$ will contain a $2\pi n/\omega_i'$ ambiguity. This will introduce discontinuities in the differential phase function, which we expect to be a smooth function of frequency (Figure 2.10*a*). However, we may use our estimate of $\delta\tau_g^m$ to predict the variation of differential phase delay with frequency and thus identify cycle-skipping offsets:

$$\frac{d}{d\omega}\delta\tau_p^m = \frac{1}{\omega}(\delta\tau_g^m - \delta\tau_p^m) \tag{2.34}$$

where we have used the fact that phase velocity is ω/k and group velocity is $d\omega/dk$. Thus, if we assume that some value of $\delta\tau_p^m$ is not contaminated by cycle-skipping, we may revise the other est nates by checking the numerical derivative of phase delay with the estimate of group delay. In practice, we assume that the lowest frequency estimate, typically 10 mHz, is unaffected by cycle-skipping. This is a reasonable assumption; even in the extreme example of EU2-SNA, $\delta\tau_p^m - \delta\tau_g^m$ does not exceed 100 s.

Figure 2.10*b* illustrates our estimates of $\delta\tau_p^m$, $\delta\tau_g^m$, and $\delta\tau_a^m$ (squares) as a function of frequency in the isolated waveform example with the actual EU2-SNA dispersion and attenuation (solid line) after correction for cycle skipping. The estimation of the phase delay is excellent, to within one second of the predicted values. The estimation of group delay is also quite good. There is greater error in the estimation of $\delta\tau_a^m$ than in either the phase or group delay, which reflects the inherent difficulty of amplitude measurements.

INTERPRETATION

We refer to the parameters $\delta\tau_p^m$, $\delta\tau_g^m$, and $\delta\tau_a^m$ as generalized data functionals. They are functions of model parameters (α, β, ρ), which are in turn functions of position (r, θ,

φ). The advantage of techniques such as the one described in this thesis over waveform inversion is that the recovered parameters may be interpreted using variational principles. In particular, we make use of Rayleigh's principle [Backus and Gilbert, 1967; Woodhouse, 1976; Woodhouse and Dahlen, 1978], which states that on average the total kinetic energy of the νth normal mode is equal to its total potential energy:

$$\omega_\nu^2 \int_0^a [U_\nu^2(r) + V_\nu^2(r) + W_\nu^2(r)]\, \rho(r)\, r^2 dr = \int_0^a [K_\nu(r)\kappa(r) + M_\nu(r)\mu(r) + \Gamma_\nu(r)]\, dr \qquad (2.35)$$

where ω_ν is the eigenfrequency; $U_\nu(r)$, $V_\nu(r)$, and $W_\nu(r)$ are the radial scalars; $K_\nu(r)\kappa(r)$ is the compressional energy density; $M_\nu(r)\mu(r)$ is the shear energy density; $\int_0^a \Gamma_\nu(r)\, r^2\, dr$ is the gravitational potential energy; and a is the radius of the Earth. The mode index ν is a function of the radial order number, the angular order number, and the azimuthal order number. Rayleigh's principle is a statement that the first variation in eigenfrequency of a normal mode due to a variation in the eigenfunction is zero. This stationarity may be exploited to calculate the eigenfrequency shift due to small changes in the model parameters:

$$\frac{\delta\omega_\nu}{\omega_\nu} = \frac{1}{2} \int_0^a \left[K_\nu(r)\kappa(r) \frac{\delta\kappa(r)}{\kappa(r)} + M_\nu(r)\mu(r) \frac{\delta\mu(r)}{\mu(r)} + R_\nu(r)\rho(r) \frac{\delta\rho(r)}{\rho(r)} \right] dr \qquad (2.36)$$

where we only consider perturbations to the radial distribution of isotropic elastic parameters. $\delta\omega_\nu$ is the perturbation to eigenfrequency of the νth normal mode at constant wavenumber, $K_\nu(r)\kappa(r)$ is the Fréchet kernel for the normalized perturbation to bulk modulus $\delta\kappa/\kappa$, $M_\nu(r)\mu(r)$ is the Fréchet kernel for the normalized perturbation to shear modulus $\delta\mu/\mu$, and $R_\nu(r)\rho(r)$ is the Fréchet kernel for the normalized perturbation to

density $\delta\rho/\rho$. Explicit expressions for the Fréchet kernels may be found in Backus and Gilbert [1967], Woodhouse [1976], and Woodhouse and Dahlen [1978] and are reproduced in Appendix C for completeness. We may rewrite this formulation of Rayleigh's principle in order to make the dependence on compressional and shear velocity more explicit:

$$\frac{\delta\omega_\nu}{\omega_\nu} = \frac{1}{2}\int_0^a [K_\alpha^\nu(r)\frac{\delta\alpha(r)}{\alpha(r)} + K_\beta^\nu(r)\frac{\delta\beta(r)}{\beta(r)} + K_\rho^\nu(r)\frac{\delta\rho(r)}{\rho(r)}]\,dr \qquad (2.37)$$

where $K_\alpha^\nu(r) = \rho(r)\,\alpha^2(r)\,K_\nu(r)$ is the Fréchet kernel for compressional velocity, $K_\beta^\nu(r) = \rho(r)\,\beta^2(r)\,(M_\nu(r) - 4/3\,K_\nu(r))$ is the Fréchet kernel for shear velocity, and $K_\rho^\nu(r) = \rho(r)[R_\nu(r) + \rho(r)^{-1}\,(\mu(r)\,M_\nu(r) + \kappa(r)\,K_\nu(r))]$ is the Fréchet kernel for density. (2.36) and (2.37) have two important applications. First, these expressions may be used to predict the eigenfrequencies of a nearby model, from the eigenfunctions and eigenfrequencies of the reference model, provided that $\delta\kappa$, $\delta\mu$, and $\delta\rho$ are small. Second, observations of eigenfrequency shifts may be used to calculate deviations from the reference model using a linear inversion.

We may also use Rayleigh's principle to calculate the effect of attenuation. If we allow the elastic parameters in (2.35) to have a small complex component, we find:

$$q_\nu = \int_0^a [K_\nu(r)\kappa(r)\,q_\kappa(r) + M_\nu(r)\mu(r)\,q_\mu(r)]\,dr \qquad (2.38)$$

where the attenuation $q_\nu = Q_\nu^{-1}$ of vth normal mode depends on the bulk attenuation $q_\kappa = Q_\kappa^{-1}$ and shear attenuation $q_\mu = Q_\mu^{-1}$ and we have neglected the frequency-dependent variation of attenuation [Backus and Gilbert, 1968; Sailor and Dziewonski, 1978; Masters

and Gilbert, 1983]. If we consider a perturbation to the attenuation due to small changes in the bulk and shear attenuations, we find:

$$\delta q_v = \int_0^a [K_v(r)\kappa(r)q_\kappa(r)\frac{\delta q_\kappa(r)}{q_\kappa(r)} + M_v(r)\mu(r)q_\mu(r)\frac{\delta q_\mu(r)}{q_\mu(r)}]\,dr \qquad (2.39)$$

This expression, similar to (2.36), poses a linear relationship between perturbations to the attenuation and variations in the anelastic model parameters. (2.39) may be used to update a reference model by calculating new values of q for small changes in δq_κ and δq_μ, or it may be used to invert observations of δq_v for perturbations to the model parameters. In practice, the values of q_κ and q_μ are so poorly constrained that (2.38) is generally used to invert attenuation measurements.

Phase delay

The differential phase delay at ω_i' arises from the differential phase velocity between $\tilde{u}_m(t)$ and $u_m(t)$:

$$\delta\tau_p^m(\omega_i') \equiv x\left(\frac{1}{c_m(\omega_i')} - \frac{1}{\tilde{c}_m(\omega_i')}\right)$$

$$\approx -\frac{x}{\tilde{c}_m(\omega_i')}\frac{\delta c_m(\omega_i')}{\tilde{c}_m(\omega_i')} \qquad (2.40)$$

where $\tilde{c}_m(\omega_i')$ is the phase velocity of $\tilde{u}_m(t)$ at ω_i', $c_m(\omega_i')$ is the phase velocity of $u_m(t)$ at ω_i', and $\delta c_m = c(\omega_i') - \tilde{c}(\omega_i')$. In order to apply Rayleigh's principle, we make use of a result which converts our observation of differential phase velocity at fixed frequency to an observation at fixed wavenumber [Dahlen, 1975]:

$$(\delta c_m)_\omega = \frac{\tilde{c}_m}{\tilde{U}_m}(\delta c_m)_k \tag{2.41}$$

where $\tilde{U}_m$ is the group velocity. We may now write:

$$\begin{aligned}\delta\tau_p^m(\omega_i') &\approx -\frac{x}{\tilde{U}_v}\frac{\delta\omega_v}{\omega_v} \\ &\approx -\frac{1}{2}\frac{x}{\tilde{U}_v}\int_0^a [K_\alpha^v(r)\frac{\delta\alpha(r)}{\alpha(r)} + K_\beta^v(r)\frac{\delta\beta(r)}{\beta(r)} + K_\rho^v(r)\frac{\delta\rho(r)}{\rho(r)}]\,dr\end{aligned} \tag{2.42}$$

where the vth normal mode corresponds to the mth traveling-wave branch at ω_i'. Figure 2.11 displays the velocity and density kernels for the example of the fundamental-mode Love (a) and Rayleigh (b) at four sample frequencies. In each diagram, the 400 km discontinuity and the base of the crust are marked by horizontal lines. At 10 mHz, the Love-wave partials sample the upper mantle to a depth of 300 km, but become concentrated in the crust with increasing frequency. The Rayleigh-wave partials also become localized at shallow depth with frequency, but remain more sensitive to upper-mantle structure than the Love-wave partials. In addition, the Rayleigh-wave kernels display some sensitivity to compressional velocity in the crust, a feature which becomes more pronounced with frequency.

We may use these kernels to predict the differential phase delay, employing the known SNA-EU2 velocity perturbations in (2.42). Figure 2.12 compares the measured estimates of $\delta\tau_p^m$ (squares) with the values predicted by integration of the kernels (filled circles). This test of first-order perturbation theory is quite successful, particularly considering the large difference between the models. The systematic overestimation of $\delta\tau_p^m$ is due to the linearization of (2.40), not any limitation of Rayleigh's principle.

Group delay

If the measurements of $\delta\tau_p^m$ were perfect and available at all frequencies, then estimates of $\delta\tau_g^m$ would be superfluous since the measurements of $\delta\tau_p^m$ are sufficient to reconstruct the entire dispersion curve. However, this is seldom the case, and we wish to make use of the values of differential group delay. It is possible to differentiate (2.35) with respect to wavenumber and construct formulae for group-velocity kernels [Gilbert, 1976b]. However, the resulting expressions are extremely complicated and we shall not reproduce them here. We shall use the estimates of differential group delay to correct $\delta\tau_p^m$ for cycle skipping.

Amplitude delay

The differential amplitude time $\delta\tau_a^m(\omega_i')$ arises from both the elastic and anelastic contributions which give rise to changes in amplitude. For example, $\delta\tau_a^m(\omega_i')$ will measure the differential amplitude effects associated with the complications of triplicated waveforms. In general, however, differential attenuation will dominate most elastic effects. In this situation, $\delta\tau_a^m$ represents the differential t^* operator ($t^* \equiv \tau_p^m / 2Q_m$) between $\tilde{u}_m(t)$ and $u_m(t)$:

$$\delta\tau_a^m(\omega_i') \equiv \frac{1}{2}\left[\tau_p^m(\omega_i')\, q_m(\omega_i') - \tilde{\tau}_p(\omega_i')\, \tilde{q}_m(\omega_i')\right]$$

$$\approx \frac{1}{2}\,\tilde{\tau}_p^m(\omega_i')\, \delta q_m(\omega_i') \qquad (2.43)$$

where $\tilde{q}_m(\omega_i') = \tilde{Q}_m^{-1}(\omega_i')$ is the attenuation of $\tilde{u}_m(t)$ at ω_i', $q_m = Q_m^{-1}(\omega_i')$ is the attenuation of $u_m(t)$ at ω_i', and $\delta q_m = q_m(\omega_i') - \tilde{q}_m(\omega_i')$.

SUMMARY OF THE ISOLATED-WAVEFORM CASE

Thus far, we have considered the case of an isolation filter composed of a single waveform and assumed that the interference due to energy with similar group velocities was negligible. We derived an expression for the autocorrelation of the isolation filter based on a Hermite polynomial expansion of $C_{\tilde{u}\tilde{u}}(\omega)$, where we assumed that $C_{\tilde{u}\tilde{u}}(\omega)$ was small near $\omega = 0$. The time-domain expression for the autocorrelation has a simple form and may be approximated by its first few terms. In particular, if $C_{\tilde{u}\tilde{u}}(\omega)$ is Gaussian, then $C_{\tilde{u}\tilde{u}}(t)$ may be described by a cosinusoidal carrier modulated by a Gaussian envelope. In general, we wish to consider signals which will not satisfy this constraint. We have shown that the application of a Gaussian narrow-band filter to the broad-band correlation function enforces this approximation, allowing us to neglect terms of order greater than γ_i^3 in the Gram-Charlier series expansion. We considered the cross-correlation between $\tilde{u}_m$ and $u_m(t)$ and derived an expression for $F_i C_{\tilde{u}u}(t)$ assuming a Taylor series expansion of differential wavenumber and neglecting any differences between the actual and assumed source function and instrument response. We investigated both the quadratic and linear approximations for $\delta k_m(\omega)$ and concluded that the effect of quadratic-dispersion could be controlled by the application of a narrow-band filter. We demonstrated that the estimates of $\delta\tau_p^m$, $\delta\tau_g^m$, and $\delta\tau_a^m$ recovered by an implementation of this methodology are quite good. Finally, we showed how $\delta\tau_p^m$ and $\delta\tau_a^m$ may be linearly related to model parameters by Rayleigh's principle. While it is possible to develop similar expressions for $\delta\tau_g^m$ [Gilbert, 1976b], we have not made use of them here. Instead, we use our estimates of the differential group delay to correct the measurements of differential phase delay for cycle skipping. Now we shall relax the assumption of an isolated waveform and consider the effect of interfering wavegroups on the estimation of $\delta\tau_p^m$, $\delta\tau_g^m$, and $\delta\tau_a^m$.

NON-ISOLATED WAVEFORMS

In general, one cannot neglect the interaction terms in the computation of $F_i C_{\tilde{u}\tilde{s}}(t)$ and $F_i C_{\tilde{u}s}(t)$. Interference is particularly acute at low frequencies, where the application of a narrow-band filter spreads out the energy of the time series and increases the interaction with other waveforms. In this section, we develop explicit expressions for the interference terms in $F_i C_{\tilde{u}\tilde{s}}(t)$ and $F_i C_{\tilde{u}s}(t)$ by carrying through the cross terms in equation (2.4) and approximate the resulting sums over traveling-wave branches as single Gaussian wavelets characterized by average perturbations due to interfering arrivals ($\Delta\tilde{\tau}_p$, $\Delta\tilde{\tau}_g$, $\Delta\tilde{\tau}_a$) and average perturbations due to differences between the reference model and the Earth ($\delta\tau_p$, $\delta\tau_g$, $\delta\tau_a$). We will formulate expressions for the average perturbations in terms of the branch contributions; in particular, we seek expressions for the Fréchet kernels which describe the variation of the generalized data functionals with respect to changes in the model parameters.

The complete synthetic cross-correlation

In the case where the isolation filter is composed of a single waveform, $\tilde{f}(t) = \tilde{u}_m(t)$, the filtered cross-correlation between the isolation filter and the synthetic seismogram may be written as a sum over the cross-correlations between $\tilde{u}_m(t)$ and the individual traveling-wave branches which comprise the synthetic seismogram, convolved with the narrow-band filter:

$$F_i C_{\tilde{u}\tilde{s}}(t) = F_i(t) * \tilde{u}_m(t) \otimes \sum_n \tilde{u}_n(t)$$

$$= \frac{1}{2\pi} \sum_n \int_{-\infty}^{\infty} \tilde{u}_m^*(\omega)\, \tilde{u}_n(\omega)\, F_i(\omega)\, e^{-i\omega t}\, d\omega \qquad (2.44)$$

In the isolated-waveform approximation, we assumed that the autocorrelation term would dominate (2.44) near zero lag. In this section, we retain the complete sum.

In order to solve equation (2.44), we need an expression which relates $\tilde{u}_m(\omega)$ and $\tilde{u}_n(\omega)$, similar to the differential response operator in (2.22). We can formulate the mth branch as the product of the excitation function $\tilde{A}_m(\omega)$, the dispersion function $\exp[i\tilde{k}_m(\omega)x]$, and the instrument response $\tilde{I}(\omega)$:

$$\begin{aligned}\tilde{u}_m(\omega) &= \tilde{A}_m(\omega)\ \exp[i\tilde{k}_m(\omega)x]\ \tilde{I}(\omega) \\ &= |\tilde{A}_m(\omega)|\ |\tilde{I}(\omega)|\ \exp[i(\tilde{k}_m(\omega)x + \tilde{\varphi}_m(\omega) + \tilde{\varphi}_I(\omega))] \end{aligned} \tag{2.45}$$

where $\tilde{\varphi}_m(\omega)$ is the real-valued excitation phase function, $\tilde{\varphi}_I(\omega)$ is the real-valued instrument phase function, and $\tilde{k}_m(\omega)$ is the complex-valued wavenumber. With this parameterization, the integral for $F_i C_{\tilde{u}\tilde{s}}(t)$ assumes the form:

$$F_i C_{\tilde{u}\tilde{s}}(t) = \frac{1}{2\pi} \sum_n \int_{-\infty}^{\infty} |\tilde{A}_m(\omega)||\tilde{A}_n(\omega)||\tilde{I}(\omega)|^2 F_i(\omega)\ e^{i(\Delta\tilde{\psi}^{nm} - \omega t)}\, d\omega \tag{2.46}$$

where $\Delta\tilde{\psi}^{nm}$ measures the differential dispersion and excitation phase between the nth and mth branches:

$$\Delta\tilde{\psi}^{nm}(\omega) = [\tilde{k}_n(\omega) - \tilde{k}_m(\omega)]\, x + [\tilde{\varphi}_n(\omega) - \tilde{\varphi}_m(\omega)] \tag{2.47}$$

and the phase contribution from the instrument cancels, since it is not a function of branch number. We expect the source amplitude spectrum to be a slowly varying function of frequency and branch number and assume that $|\tilde{A}_n(\omega)|\ |\tilde{A}_m(\omega)|\ |\tilde{I}(\omega)|^2$ is a peaked spectrum which is completely described by a few of its low-order moments. Near $\omega = \omega_i'$, we describe this product as a scaled version of the autocorrelation function:

$$|\tilde{A}_n(\omega)|\,|\tilde{A}_m(\omega)|\,|\tilde{I}(\omega)|^2 \approx |\tilde{A}_n(\omega_i')|\,|\tilde{A}_m(\omega_i')|\,C_{\tilde{u}\tilde{u}}(\omega) \tag{2.48}$$

This statement is an approximation that the amplitude spectrum changes scale, but not shape, as a function of branch number. For the cross-correlation of a high phase-velocity branch with a low phase-velocity branch, (2.48) is not a good approximation of the amplitude spectrum. However, most significant contributions to the travel-wave sum will be from "nearest-neighbor" branches, which will satisfy this assumption.

The phase spectrum of $F_i C_{\tilde{u}\tilde{s}}(\omega)$, on the other hand, is a rapidly varying function of frequency and branch number, primarily due to the source-phase contribution. We expand $\Delta\tilde{\psi}^{nm}$ in a Taylor series expansion about $\omega = \omega_i'$, neglecting terms of order higher than the first:

$$\Delta\tilde{\psi}^{nm}(\omega) \approx \omega_i' \Delta\tilde{\tau}_p^{nm} + (\omega - \omega_i')(\Delta\tilde{\tau}_g^{nm} + i\Delta\tilde{\tau}_a^{nm}) \tag{2.49}$$

$\Delta\tilde{\tau}_p^{nm}$ is the differential phase delay, $\Delta\tilde{\tau}_g^{nm}$ is the differential group delay, and $\Delta\tilde{\tau}_a^{nm}$ is the differential amplitude delay between the nth and mth branches due to propagation and source excitation:

$$\Delta\tilde{\tau}_p^{nm}(\omega_i') \equiv \mathrm{Re}[\tilde{k}_n(\omega_i') - \tilde{k}_m(\omega_i')]x / \omega_i' + [\tilde{\varphi}_n(\omega_i') - \tilde{\varphi}_m(\omega_i')] / \omega_i' \tag{2.49a}$$

$$\Delta\tilde{\tau}_g^{nm}(\omega_i') \equiv \mathrm{Re}[\tilde{\dot{k}}_n(\omega_i') - \tilde{\dot{k}}_m(\omega_i')]x + [\tilde{\dot{\varphi}}_n(\omega_i') - \tilde{\dot{\varphi}}_m(\omega_i')] \tag{2.49b}$$

$$\Delta\tilde{\tau}_a^{nm}(\omega_i') \equiv \mathrm{Im}[\tilde{\dot{k}}_n(\omega_i') - \tilde{\dot{k}}_m(\omega_i')]x \tag{2.49c}$$

The case of n equal to m corresponds to the autocorrelation of an individual branch and $\Delta\tilde{\tau}_p^{nm} = \Delta\tilde{\tau}_g^{nm} = \Delta\tilde{\tau}_a^{nm} = 0$.

With this parameterization of the amplitude and phase spectrum, we may solve (2.44) for the filtered cross-correlation between the isolation filter and the complete synthetic seismogram:

$$F_i C_{\tilde{u}\tilde{s}}(t) = \frac{1}{(2\pi)^{3/2}} \frac{\sigma_i'}{\sigma_i\ \sigma_f} \mathrm{Ga}\left(\frac{\omega_i - \omega_f}{\sqrt{\sigma_i^2+\omega_f^2}}\right) \sum_n \tilde{A}_{nm} \exp[(\sigma_i' \Delta\tilde{\tau}_a^{nm})^2/2] \mathrm{Ga}\,(\sigma_i'(t-\Delta\tilde{\tau}_g^{nm}))$$

$$\times \cos[\omega_i'(t-\Delta\tilde{\tau}_p^{nm}) - \sigma_i'^2 \Delta\tilde{\tau}_a^{nm}(t-\Delta\tilde{\tau}_g^{nm})] \quad (2.50)$$

where $\tilde{A}_{nm} \equiv |\tilde{A}_n(\omega_i')|\ |\tilde{A}_m(\omega_i')|$. Using the Hermite-polynomial formalism developed in this Chapter and in Appendices A and B, we may carry the correction terms through our treatment. However, for the sake of brevity, we have not done so.

Single wavelet approximation. Equation (2.50) expresses $F_i C_{\tilde{u}\tilde{s}}(t)$ as a sum over Gaussian wavelets; each one parameterized by a differential phase, group, and amplitude time delay describing the interaction between two traveling-wave branches. Only branches with nearly the same group velocity at ω_i' will contribute significantly to the sum, and we expect that the correlation function will be dominated by the autocorrelation term. Therefore, we approximate (2.50) by a single Gaussian wavelet, characterized by an average perturbation of amplitude as well as phase, group, and attenuation delays due to the cross-term contributions:

$$F_i C_{\tilde{u}\tilde{s}}(t) \approx E(t) \cos\Phi(t) \quad (2.51)$$

where $E(t)$ and $\Phi(t)$ are defined by

$$E(t) = \frac{\tilde{A}}{(2\pi)^{3/2}} \frac{\sigma_i'}{\sigma_i\,\sigma_f} \exp(\sigma_i'^2 \Delta\tilde{\tau}_a^2/2)\ \mathrm{Ga}\left(\frac{\omega_i - \omega_f}{\sqrt{\sigma_i^2+\sigma_f^2}}\right) \mathrm{Ga}\,(\sigma_i'(t-\Delta\tilde{\tau}_g)) \quad (2.51a)$$

$$\Phi(t) = \omega_i'(t-\Delta\tilde{\tau}_p) - \sigma_i'^2\Delta\tilde{\tau}_a(t-\Delta\tilde{\tau}_g) \tag{2.51b}$$

$\Delta\tilde{\tau}_p(\omega_i')$ is an estimate of the average differential phase shift at ω_i', $\Delta\tilde{\tau}_g(\omega_i')$ is an estimate of the average differential group delay, $\Delta\tilde{\tau}_a(\omega_i')$ is a measure of the average differential amplitude, and $\tilde{A}$ is a scale parameter introduced by the interfering waveforms on the synthetic seismogram.

We can develop expressions for these average perturbations due to interfering arrivals in terms of the individual contributions of the traveling-wave branches. We illustrate the approach in the simple case without differential attenuation in order to fix ideas; extending the methodology to include differential attenuation is straightforward. Neglecting the effect of differential attenuation, we compare the single Gaussian-wavelet approximation (2.51) to the traveling-wave sum (2.50):

$$\tilde{A}\,\mathrm{Ga}(\sigma_i'(t-\Delta\tilde{\tau}_g))\cos[\omega_i'(t-\Delta\tilde{\tau}_p)] = \sum_n \tilde{A}_{nm}\,\mathrm{Ga}(\sigma_i'(t-\Delta\tilde{\tau}_g^{nm}))\cos[\omega_i'(t-\Delta\tilde{\tau}_p^{nm})] \tag{2.52}$$

If $\sigma_i' << \omega_i'$, then the Gaussian envelopes will be slowly varying, compared to the carrier functions, near their peaks. We expand both envelopes about $t = \Delta\tilde{\tau}_g$ and compare expressions on a term-by-term basis:

0th:
$$\tilde{A}\cos[\omega_i'(t-\Delta\tilde{\tau}_p)] = \sum_n \tilde{B}_{nm}\cos[\omega_i'(t-\Delta\tilde{\tau}_p^{nm})] \tag{2.53}$$

1st:
$$-\sigma_i'^2\sum_n \tilde{B}_{nm}\,(\Delta\tilde{\tau}_g - \Delta\tilde{\tau}_g^{nm})\cos[\omega_i'(t-\Delta\tilde{\tau}_p^{nm})] = 0 \tag{2.54}$$

where $\tilde{B}_{nm} = \tilde{A}_{nm}\ \mathrm{Ga}(\sigma_i'(\Delta\tilde{\tau}_g - \Delta\tilde{\tau}_g^{nm}))$

The zeroth-order term is a statement of the cosine averaging theorem: a sum of cosines oscillating at the same frequency but with different amplitude and phase may be

represented as a single cosine with an average amplitude and phase. By rearranging (2.53), we obtain expressions for $\tilde{A}$ and $\Delta\tilde{\tau}_p$:

$$\tilde{A}^2 = \left[\sum_n \tilde{B}_{nm}\cos(\omega_i'\Delta\tilde{\tau}_p^{nm})\right]^2 + \left[\sum_n \tilde{B}_{nm}\sin(\omega_i'\Delta\tilde{\tau}_p^{nm})\right]^2 \tag{2.55}$$

$$\Delta\tilde{\tau}_p(\omega_i') = \frac{1}{\omega_i'}\tan^{-1}\left(\frac{\sum_n \tilde{B}_{nm}\sin(\omega_i'\Delta\tilde{\tau}_p^{nm})}{\sum_n \tilde{B}_{nm}\cos(\omega_i'\Delta\tilde{\tau}_p^{nm})}\right) \tag{2.56}$$

which formulate these parameters as weighted sums of sines and cosines of the differential phase delay between the *m*th and *n*th branches.

We may also use the cosine average theorem to obtain an expression for the average differential group delay. We represent (2.54) as a single cosine characterized by an average amplitude $\tilde{B}$ and phase $\Delta\tilde{\tau}_b$:

$$-\sigma_i'^2\sum_n \tilde{B}_{nm}\,(\Delta\tilde{\tau}_g - \Delta\tilde{\tau}_g^{nm})\,\cos[\omega_i'(t-\Delta\tilde{\tau}_p^{nm})] = \tilde{B}\cos[\omega_i'(t-\tilde{\tau}_b)] \tag{2.57}$$

where $\tilde{B}^2$ and $\Delta\tilde{\tau}_b$ are defined by

$$\tilde{B}^2 = \left[\sum_n \tilde{B}_{nm}\,(\Delta\tilde{\tau}_g-\Delta\tilde{\tau}_g^{nm})\cos(\omega_i'\Delta\tilde{\tau}_p^{nm})\right]^2 + \left[\sum_n \tilde{B}_{nm}\,(\Delta\tilde{\tau}_g-\Delta\tilde{\tau}_g^{nm})\sin(\omega_i'\Delta\tilde{\tau}_p^{nm})\right]^2 \tag{2.58}$$

$$\tilde{\tau}_b(\omega_i') = \frac{1}{\omega_i'}\tan^{-1}\left(\frac{\sum_n \tilde{B}_{nm}\,(\Delta\tilde{\tau}_g-\Delta\tilde{\tau}_g^{nm})\sin(\omega_i'\Delta\tilde{\tau}_p^{nm})}{\sum_n \tilde{B}_{nm}\,(\Delta\tilde{\tau}_g-\Delta\tilde{\tau}_g^{nm})\cos(\omega_i'\Delta\tilde{\tau}_p^{nm})}\right) \tag{2.59}$$

The first-order term of the expansion requires that $\tilde{B}\cos[\omega_i'(t - \Delta\tilde{\tau}_b)]$ be zero for all time. Since there is no single value of $\Delta\tilde{\tau}_b$ which will satisfy this constraint for all values of time, we seek an approximate solution to (2.57). Rather than expanding the cosine about some particular time, we determine the value of $\Delta\tilde{\tau}_g$ which minimizes $\tilde{B}$. Assuming that the $\tilde{B}_{nm}$ are slowly-varying functions of $\Delta\tilde{\tau}_g$, then we may derive a simple expression for $\Delta\tilde{\tau}_g$ by taking the derivative of the expression (2.58) with respect to $\Delta\tilde{\tau}_g$:

$$\Delta\tilde{\tau}_g(\omega_i') \approx \frac{c_0 c_1 + s_0 s_1}{c_0^2 + s_0^2} \tag{2.60}$$

where the coefficients c_0, c_1, s_0, and s_1 are defined in terms of sums over the differential branch phase and group delays:

$$\begin{aligned} c_0 &= \sum_n \tilde{B}_{nm}\cos(\omega_i'\Delta\tilde{\tau}_p^{nm}) \\ c_1 &= \sum_n \tilde{B}_{nm}\cos(\omega_i'\Delta\tilde{\tau}_p^{nm})\,\Delta\tilde{\tau}_g^{nm} \\ s_0 &= \sum_n \tilde{B}_{nm}\sin(\omega_i'\Delta\tilde{\tau}_p^{nm}) \\ s_1 &= \sum_n \tilde{B}_{nm}\sin(\omega_i'\Delta\tilde{\tau}_p^{nm})\,\Delta\tilde{\tau}_g^{nm} \end{aligned} \tag{2.61}$$

Our expression for $\Delta\tilde{\tau}_g(\omega_i')$ is depends of sums of sines and cosines of the differential phase delay between the *m*th and *n*th branches, multiplied by the differential group delay and appropriately weighted.

We have seen that the cross-correlation between a single waveform isolation filter and the complete synthetic seismogram may be described as a sum over Gaussian wavelets (equation 2.50) which describe the individual branch correlation functions. This formula was derived by assuming that the amplitude spectrum of the branch correlation functions is a scaled form of the autocorrelation spectrum and by approximating the differential

dispersion between two branches with a first-order Taylor series expansion. We compared (2.50) to a single Gaussian wavelet which is characterized by an average perturbation to amplitude and phase (2.51). In the case where differential attenuation may be neglected, we determined expressions for the average perturbation to phase and group delay due to the cross-term contributions, assuming that the envelope functions are slowly varying compared to the carrier function. $\Delta\tilde{\tau}_p$ and $\Delta\tilde{\tau}_g$ are simple weighted sums of sines and cosines of the differential phase delay between the *m*th and *n*th branches.

The observed cross-correlation

Proceeding as we did in the case of the synthetic cross-correlation function, we derive an expression for the filtered cross-correlation between $\tilde{u}_m(t)$ and $s(t)$. $F_i C_{\tilde{u}s}$ may be written as a sum over the cross-correlations between $\tilde{u}_m(t)$ and the individual traveling-wave branches which comprise the observed seismogram, convolved with the narrow-band filter:

$$\begin{aligned} F_i C_{\tilde{u}s}(t) &= F_i(t) * \tilde{u}_m(t) \otimes \sum_n u_n(t) \\ &= \frac{1}{2\pi} \sum_n \int_{-\infty}^{\infty} \tilde{u}_m^*(\omega)\, u_n(\omega)\, F_i(\omega)\, e^{-i\omega t}\, d\omega \\ &= \frac{1}{2\pi} \sum_n \int_{-\infty}^{\infty} |\tilde{A}_m(\omega)||A_n(\omega)|\, I(\omega)\, \tilde{I}^*(\omega)\, F_i(\omega)\, e^{i(\Delta\psi^{nm} - \omega t)}\, d\omega \end{aligned} \quad (2.62)$$

where $|A_n(\omega)|$ is the source excitation amplitude of the *n*th observed branch, $I(\omega)$ is the observed instrument response, and $\Delta\psi^{nm}$ is the differential dispersion and the source phase function between the *n*th and *m*th branches:

$$\Delta\psi^{nm}(\omega) = [k_n(\omega) - \tilde{k}_m(\omega)]\, x + [\varphi_n(\omega) - \tilde{\varphi}_m(\omega)] \quad (2.63)$$

and φ_n is the real-valued source phase of the nth observed branch. As before, we assume that the differences between the observed and synthetic waveforms are due to elastic and anelastic propagation alone, *i.e.*, we neglect any differences in the source mechanism or instrument response: $|A_n(\omega)| = |\tilde{A}_n(\omega)|$; $\varphi_n \approx \tilde{\varphi}_n$; $I(\omega) \approx \tilde{I}(\omega)$. Under this assumption, we may rewrite the expression for $\Delta\psi^{nm}$ in terms of a perturbation to $\Delta\tilde{\psi}^{nm}$:

$$\Delta\psi^{nm}(\omega) \approx \Delta\tilde{\psi}^{nm}(\omega) + \delta k_n(\omega)\, x \tag{2.64}$$

and $\delta k_n(\omega) = k_n(\omega) - \tilde{k}_n(\omega)$ is the differential wavenumber between the nth observed and synthetic branches. This formulation explicitly represents $\Delta\psi^{nm}$ in terms of the differential dispersion and source phase between the nth and mth branches of the synthetic seismogram, plus a perturbation which accounts for the differences between the reference model and the Earth for the nth branch. We expand $\Delta\psi^{nm}$ in a Taylor series about $\omega = \omega_i'$, neglecting terms of order greater than the first:

$$\Delta\psi^{nm}(\omega) \approx \omega_i' (\delta\tau_p^n + \Delta\tilde{\tau}_p^{nm}) + (\omega - \omega_i')((\delta\tau_g^n + \Delta\tilde{\tau}_g^{nm}) + i(\delta\tau_a^n + \Delta\tilde{\tau}_a^{nm})) \tag{2.65}$$

where $\delta\tau_p^n(\omega_i')$ is the differential phase delay, $\delta\tau_g^n(\omega_i')$ is the differential group delay, and $\delta\tau_a^n(\omega_i')$ is the differential attenuation delay between the nth observed and synthetic branch at ω_i' due to differences in elastic and anelastic propagation, as discussed in the isolated waveform approximation.

With this parameterization of the amplitude and phase spectrum, we may solve (2.62) for the filtered cross-correlation between the isolation filter and the observed seismogram:

$$F_i C_{\tilde{u}\,s}(t) = \frac{1}{(2\pi)^{3/2}} \frac{\sigma_i'}{\sigma_i \sigma_f} \mathrm{Ga}\left(\frac{\omega_i - \omega_f}{\sqrt{\sigma_i^2+\sigma_f^2}}\right) \sum_n \tilde{A}_{nm} \exp[(\sigma_i'(\delta\tau_a^n + \Delta\tilde{\tau}_a^{nm}))^2/2]$$

$$\times \mathrm{Ga}\,(\sigma_i'(t-\delta\tau_g^n-\Delta\tilde{\tau}_g^{nm})) \cos[\omega_i'(t-\delta\tau_p^n-\Delta\tilde{\tau}_p^{nm}) - \sigma_i'^2(\delta\tau_a^n+\Delta\tilde{\tau}_a^{nm})(t-\delta\tau_g^n-\Delta\tilde{\tau}_g^{nm})] \tag{2.66}$$

Equation (2.66) differs from (2.50), the expression for $F_i C_{\tilde{u}\tilde{s}}$, through the phase, group and amplitude perturbations: $\delta\tau_p^n$, $\delta\tau_g^n$, and $\delta\tau_a^n$, which parameterize the differences between the nth observed and synthetic waveforms. If we allow these perturbations to go to zero, then we recover (2.50).

Single wavelet approximation. Although equation (2.66) expresses $F_i C_{\tilde{u}\,s}(t)$ as a sum over Gaussian wavelets, we expect that it will be dominated by the autocorrelation term. Therefore, we approximate (2.66) by a single Gaussian wavelet, characterized by average differential phase, group, and attenuation delays which depend on differences between the model and the Earth as well as interference effects:

$$F_i C_{\tilde{u}\,s}(t) \approx E(t) \cos\Phi(t) \tag{2.67}$$

where $E(t)$ and $\Phi(t)$ are defined by

$$E(t) = \frac{A}{(2\pi)^{3/2}} \frac{\sigma_i'}{\sigma_i \sigma_f} \exp[\sigma_i'^2(\delta\tau_a+\Delta\tilde{\tau}_a)^2/2] \mathrm{Ga}\left(\frac{\omega_i-\omega_f}{\sqrt{\sigma_i^2+\sigma_f^2}}\right) \mathrm{Ga}\,(\sigma_i'(t-\delta\tau_g-\Delta\tilde{\tau}_g)) \tag{2.67a}$$

$$\Phi(t) = \omega_i'(t-\delta\tau_p-\Delta\tilde{\tau}_p) - \sigma_i'^2(\delta\tau_a+\Delta\tilde{\tau}_a)(t-\delta\tau_g-\Delta\tilde{\tau}_g) \tag{2.67b}$$

As before, $\Delta\tilde{\tau}_p(\omega_i')$, $\Delta\tilde{\tau}_g(\omega_i')$, and $\Delta\tilde{\tau}_a(\omega_i')$ represent a measure of the interference. $\delta\tau_p(\omega_i')$ is an estimate of the average differential phase shift due to the difference between the Earth and reference model at ω_i', $\delta\tau_g(\omega_i')$ is an estimate of the average differential

group delay, $\delta\tau_a(\omega_i')$ is a measure of the average differential amplitude, and A is a scale parameter.

Fréchet kernels. We can develop expressions for the generalized data functionals, $\delta\tau_p$, $\delta\tau_g$, and $\delta\tau_a$, in terms of the individual contributions of the traveling-wave branches, which will allow us to formulate expressions for the Fréchet kernels as linear sums of the branch parameters. To demonstrate the approach, we shall derive a formula for the Fréchet kernel for $\delta\tau_p(\omega_i')$ in the case of no differential attenuation. Expanding the theory to derive expressions for the group and attenuation kernels is a direct extension of this methodology.

We compare the expression of the averaged Gaussian wavelet (2.67) with the traveling-wave sum (2.66), where we have neglected the effect of differential attenuation:

$$\begin{aligned} & A\,\mathrm{Ga}(\sigma_i'(t-\delta\tau_g-\Delta\tilde{\tau}_g))\cos[\omega_i'(t-\delta\tau_p-\Delta\tilde{\tau}_p)] \\ &= \sum_n \tilde{A}_{nm}\,\mathrm{Ga}(\sigma_i'(t-\delta\tau_g^n-\Delta\tilde{\tau}_g^{nm}))\cos[\omega_i'(t-\delta\tau_p^n-\Delta\tilde{\tau}_p^{nm})] \end{aligned} \tag{2.68}$$

If $\sigma_i' << \omega_i'$, then the Gaussian envelopes will be slowly varying compared to the carrier function near their peaks. We expand the envelopes about $t = \Delta\tilde{\tau}_g + \delta\tau_g$ and equate terms:

$$\text{0th:} \qquad \tilde{A}\,\cos[\omega_i'(t-\delta\tau_p-\Delta\tilde{\tau}_p)] = \sum_n \tilde{B}_{nm}\cos[\omega_i'(t-\delta\tau_p^n-\Delta\tilde{\tau}_p^{nm})] \tag{2.69}$$

where $\tilde{B}_{nm} = \tilde{A}_{nm}\ \mathrm{Ga}(\sigma_i'(\Delta\tilde{\tau}_g - \Delta\tilde{\tau}_g{}^{nm}))$ as before. We have neglected the amplitude perturbations in (2.69), since they will be second-order with respect to the phase perturbations. By rearranging the terms, we find:

$$\begin{aligned} & \sin[\omega_i'(\delta\tau_p+\Delta\tilde{\tau}_p)]\sum_n \tilde{B}_{nm}\cos[\omega_i'(\delta\tau_p^n+\Delta\tilde{\tau}_p^{nm})] \\ & \qquad = \cos[\omega_i'(\delta\tau_p+\Delta\tilde{\tau}_p)]\sum_n \tilde{B}_{nm}\sin[\omega_i'(\delta\tau_p^n+\Delta\tilde{\tau}_p^{nm})] \end{aligned} \tag{2.70}$$

If we assume that the first-order differential phase shifts are small (*i.e.*, $\cos\omega_i'\delta\tau_p \approx 1$, $\cos\omega_i'\delta\tau_p^n \approx 1$, $\sin\omega_i'\delta\tau_p \approx \omega_i'\delta\tau_p$, $\sin\omega_i'\delta\tau_p^n \approx \omega_i'\delta\tau_p^n$), then we obtain:

$$\delta\tau_p(\omega_i') = \frac{\sum_n \tilde{P}_n \, \delta\tau_p^n(\omega_i')}{\sum_n \tilde{P}_n} \tag{2.71}$$

where $\tilde{P}_n$ is defined by

$$\tilde{P}_n = \tilde{B}_{nm}\cos\omega_i'(\Delta\tilde{\tau}_p - \Delta\tilde{\tau}_p^{nm}) \tag{2.72}$$

This equation expresses the average phase perturbation as a linear sum of appropriately scaled branch phase perturbations; that is, the average phase perturbation due to the cross-correlation of a single waveform $\tilde{u}_m(t)$ with the observed seismogram $s(t)$ depends on all the individual branch perturbations $\delta\tau_p^n$. This remarkably simple formula includes the contribution of the cross-terms explicitly through the $\Delta\tilde{\tau}_p^{nm}$.

IMPLEMENTATION

We have developed software to fit the average Gaussian wavelet expressions (2.51 and 2.67) to the observed functions $F_iC_{\tilde{u}\tilde{s}}(t)$ and $F_iC_{\tilde{u}s}(t)$ and thereby determine the values of $\delta\tau_p(\omega_i')$, $\delta\tau_g(\omega_i')$, and $\delta\tau_a(\omega_i')$. Our algorithm is outlined in Figure 2.13. Here we shall briefly discuss the implementation of this methodology.

Data selection

The data presented in this thesis are from the GDSN. This network, operated by the U.S. Geological Survey, has been active for over ten years and consists of over 50

contributing stations [Peterson *et al.*, 1976, 1980; Peterson, 1980; Peterson and Hutt, 1982]. We identify events located in the Western Pacific and Eurasian continent with moments greater than 1.0×10^{24} dyne-cm for which centroid moment tensors (CMT) [Dziewonski *et al.*, 1981] are determined. Two hours of data from selected events are retrieved from the National Earthquake Information Center (NEIC) Network Day Tapes and examined for glitches, gaps, and non-linearities. A raw seismogram is demeaned, edited, and rotated into transverse and radial components based on the CMT location. After rotation, a Hanning taper is applied in the frequency domain between 0 and 5 and between 40 and 50 mHz.

Calculation of synthetic seismograms

A simple recipe for the calculation of synthetic seismograms is based on the convolution of three operators:

$$\tilde{s}(t) = \tilde{S}(t) * \tilde{P}(t) * \tilde{R}(t) \tag{2.73}$$

where $\tilde{S}$ is source operator, $\tilde{P}$ is the elastic and anelastic propagation operator, and $\tilde{R}$ is the receiver operator. We use normal-mode summation [Gilbert, 1971; Gilbert and Dziewonski, 1975] to calculate our synthetic seismograms. We have calculated normal-mode catalogs for the models EU2 and SNA which are complete to 50 mHz. For *SH*, this encompasses 8,000 modes; for *PSV*, more than 16,000. Although time-intensive, these computations are only made once for each model. The attenuation model of Masters and Gilbert [1983] was used to calculate the quality factor of each mode. Once the eigenfunctions and eigenfrequencies are computed, any number of synthetic seismograms may be calculated. We use the parameters of the Harvard CMT solutions as the source operator with a step function time dependence. A typical toroidal-mode seismogram, one hour in length with one second sampling, may be computed in under two minutes on a

four-processor Alliant FX/40; a spheroidal-mode seismogram in under four minutes. After summation, the synthetic seismograms are convolved with the appropriate instrument response and filtered with a Hanning taper between 0 and 5 and between 40 and 50 mHz. Additional details about the procedure for calculating synthetic seismograms may be found in Appendix C.

Waveform fitting

We have referred to the technique of waveform fitting several times in this chapter, and we shall now describe the procedure. In the first step, we normalize the amplitude of the correlation function, eliminating any information which may be explained by a scale parameter. We then estimate the relevant parameters (ω_i' and σ_i' for $F_iC_{\tilde{u}\tilde{u}}$; $\Delta\tilde{\tau}_p$, $\Delta\tilde{\tau}_g$, and $\Delta\tilde{\tau}_a$ for $F_iC_{\tilde{u}\tilde{s}}$; and $\delta\tau_p$, $\delta\tau_g$, and $\delta\tau_a$ for $F_iC_{\tilde{u}s}$) by a peak-and-trough technique. For example, for $F_iC_{\tilde{u}\tilde{u}}$, ω_i' is determined from the location of the minima on either side of the peak and σ_i' is determined by their amplitude (Figure 2.14). These initial guesses are used in the appropriate expression (equations 2.20, 2.51, or 2.67) to calculate an analytic correlation function. We compute the difference between the observed and analytic correlation functions and weight the residual with a cosine taper centered at the peak of the correlation function. This weighting scheme reinforces our approximation that the expansion of the correlation functions is valid in the neighborhood about the peak of the correlation function. The weighted residual is inverted for perturbation to the parameters and the algorithm is repeated iteratively, until the χ^2 minimization criteria is satisfied.

Isolation filter and the correlation functions

In this chapter, we have assumed that $\tilde{f}(t)$ is composed of a single waveform, such as the fundamental-mode surface wave. Having calculated $\tilde{u}_m(t)$, we compute its autocorrelation function $C_{\tilde{u}\tilde{u}}(t)$, transform to the frequency domain, and apply a narrow-band filter. We estimate ω_i' and σ_i' in the time domain by fitting (2.20) to the correlation

function. The filtered autocorrelation of the fundamental-mode surface waves and resulting fit are illustrated in Figure 2.3 for several filter center frequencies. The high quality of our fit is demonstrated by the inability to distinguish the solid line from the dashed line in Figure 2.3.

We then compute the cross-correlation between $\tilde{u}_m(t)$ and $\tilde{s}(t)$ and apply the narrow band filter. With ω_i' and σ_i' fixed, we use equation (2.51) to estimate $\Delta\tilde{\tau}_p(\omega_i')$, $\Delta\tilde{\tau}_g(\omega_i')$, and $\Delta\tilde{\tau}_a(\omega_i')$ and correct the estimates of phase delay for cycle-skipping as described earlier. If $\tilde{f}(t)$ is isolated from other arrivals on the seismogram, then these values will be near zero. If, however, the interference is large, then these values will be non-zero. Figure 2.15 illustrates the situation by comparing $C_{\tilde{u}\tilde{s}}(t)$ and $F_i C_{\tilde{u}\tilde{s}}(t)$ for three filters. Although $C_{\tilde{u}\tilde{s}}(t)$ is not strongly influenced by interference near zero lag, the arrivals visible between -400 and -200 s affect $F_i C_{\tilde{u}\tilde{s}}$ at low frequencies. For example, the best-fitting model in the 15 mHz example is biased by the additional energy, as indicated by the misfit between $F_i C_{\tilde{u}\tilde{s}}$ and the model. Figure 2.16 presents the values of the apparent phase, group, and attenuation time shifts determined from the waveform-fitting. The estimation of phase delay is quite robust with respect to interference, while that of group delay and attenuation are not. In particular, the estimate of differential group delay is extremely sensitive to the presence of interfering energy; this may be observed in Figure 2.15. For comparison, Figure 2.17 displays the measured values of $\Delta\tilde{\tau}_p$ and $\Delta\tilde{\tau}_g$ (squares) with the values predicted from our expressions (2.56) and (2.60) (filled circles). The prediction of the average phase delay is excellent across the entire frequency range. The prediction of the average group delay is quite good for the Love wave, but manifests some problems for the Rayleigh wave.

Finally, we compute the filtered cross-correlation between the isolation filter and the observed seismogram. With ω_i' and σ_i' fixed, we use waveform fitting to estimate the total phase shift $\delta\tau_p(\omega_i') + \Delta\tilde{\tau}_p$, the total group shift $\delta\tau_g(\omega_i') + \Delta\tilde{\tau}_g$, and the total attenuation shift $\delta\tau_a(\omega_i') + \Delta\tilde{\tau}_a$, from which we may determine $\delta\tau_p(\omega_i')$, $\delta\tau_g(\omega_i')$, and

$\delta\tau_a(\omega_i')$, and correct the estimates of phase delay for cycle skipping. Figure 2.18 compares the numerical cross-correlation function with the model obtained by waveform-fitting equation (2.67). These filtered correlation functions also show the influence of interference at low frequencies.

EXAMPLES

Complete synthetic seismogram

Figure 2.19 displays the estimated parameter (squares) for the fundamental-mode surface wave example. The measured values of differential phase delay are very stable, despite the effects of interference at low frequencies, and the fit to the exact values (solid line) is excellent. However, the estimates of group and attenuation delay are considerably less robust. As we saw in the example of the synthetic cross-correlation function, the estimates of group delay are strongly biased at low frequencies due to interference. Although we are primarily interested in phase information, this bias may be a problem since we use the estimates of group delay to correct for cycle-skipping.

We may use equation (2.71) to calculate expressions for the Fréchet kernels. Figure 2.20 displays the fundamental-mode kernels for Love (*a*) and Rayleigh (*b*) waves at four frequencies. These kernels are not very different from those in Figure 2.11 because the estimates of differential phase delay are not strongly influenced by interference in this example. Figure 2.21 compares the measured estimates of $\delta\tau_p$, $\delta\tau_g$, and $\delta\tau_a$ (squares) with the values predicted by first-order perturbation theory via (2.71) (filled circles). The comparison between the two is quite good.

Observed seismogram

Figure 2.22 compares the observed seismograms at KONO for the 83/08/17 event with synthetic seismograms calculated from EU2. The comparison between the observed

and synthetic seismograms is good, although not excellent, on the vertical component. There is a greater discrepancy between the data and synthetic on the transverse component. Both the fundamental Rayleigh and Love waves show the effect of unmodeled scattering. Figure 2.23 presents the results of the parameter retrieval algorithm for the generalized data functionals. It is interesting to note that the Rayleigh differential phase delay for this seismogram is near zero across the frequency range examined, suggesting EU2 is an appropriate model for *PSV* propagation along this path. This is not surprising since EU2 was derived from a study of fundamental and higher Rayleigh modes along this path. However, the Love differential phase delay indicates that the SH propagation is faster than that predicted by EU2. This observation of apparent polarization anisotropy is consistent with the results of Cara *et al.* [1980], Lévêque and Cara [1983], and Gee and Jordan [1988].

SUMMARY

In this chapter, we have used a Hermite polynomial expansion to express the autocorrelation of the isolation filter. We have shown that this expansion may be truncated after the second-order term, provided that the signal is peaked in the spectral domain and may be described by a few of its low-order moments. Although not generally valid, this condition may be met by the application of a narrow-band filter. Using a first-order Taylor series expansion of differential wavenumber and neglecting differences between the actual and assumed source function and instrument response, the filtered cross-correlation between an isolation filter and the observed seismogram may be expressed in terms of time parameters due to differences between a reference model and the actual Earth at ω_i'. These time parameters, which we call "generalized seismological data functionals," represent phase, group, and attenuation time shifts. We have developed explicit expressions for the effect of interference from other wavegroups and derived an expression which relates the

observed phase perturbation to a linear sum of the individual phase perturbations of each traveling-wave branch.

We illustrated our methodology with the example of fundamental-mode surface waves. In these examples, the effect of interference was manifested primarily in the estimation of group delay. Although we are primarily interested in the measurement of the differential phase, we use $\delta\tau_g$ to correct $\delta\tau_p$ for cycle skipping. Errors in the estimates of $\delta\tau_g$ may lead to problems in resolving the $2\pi n$ uncertainty. In the next chapter, we expand this methodology to include a windowing operator. The application of a windowing operator in the time-domain, prior to filtering, dramatically reduces the effect of interference.

FIGURE CAPTIONS

FIGURE 2.1

Transverse (left) and vertical (right) component synthetic seismograms calculated from the models EU2 and SNA for a Kamchatka event (83/08/17, $h = 77$ km) recorded at the GDSN station KONO. $\tilde{s}(t)$ is the seismogram calculated from EU2; $s(t)$ is the seismogram calculated from SNA. $\tilde{f}(t)$ is the isolation filter for the fundamental-mode surface wave, calculated from EU2. Notice the differences in the waveforms predicted by the models, particularly in *SSS*. All six synthetic seismograms were calculated by normal-mode summation and are complete to 50 mHz.

FIGURE 2.2

$C_{\tilde{u}\tilde{u}}(t)$ calculated from the fundamental-mode Love (left) and Rayleigh (right) waves. $C_{\tilde{u}\tilde{u}}(t)$ is a symmetric function, peaked at zero lag. The dashed line indicates the Gaussian-wavelet approximation (2.72) to $C_{\tilde{u}\tilde{u}}(t)$, which is quite good. The center frequency of the Love wave is 23 mHz with half-width 6 mHz; the center frequency of the Rayleigh wave is 27 mHz with half-width 7 mHz.

FIGURE 2.3

$F_i C_{\tilde{u}\tilde{u}}$ (solid line) calculated from the fundamental-mode Love (left) and Rayleigh (right) waves. The top waveform in each box is the unfiltered correlation function; the bottom three illustrate filters with varying center frequencies (1 = 35, 2 = 25, and 3 = 15 mHz). The relative half-width of the applied filter (γ_i) was 0.1 in these examples. The dashed line, which is indistinguishable from the solid line, indicates the best-fitting model of $F_i C_{\tilde{u}\tilde{u}}(t)$ determined from waveform fitting of (2.20).

FIGURE 2.4

Comparison between $C_{\tilde{u}u}(t)$ (solid line) and $C_{\tilde{u}\tilde{u}}(t)$ (dashed line) for the fundamental-mode Love and Rayleigh waves. While $C_{\tilde{u}\tilde{u}}(t)$ is a symmetric function about zero lag, $C_{\tilde{u}u}(t)$ is not. Neglecting errors due to variations in the actual and assumed source function and instrument response, we model the differences between these functions through the differential response operator $D_m(\omega)$, where $D_m(\omega) = D_0^m \exp[i\delta k_m(\omega)x]$.

FIGURE 2.5

Differential wavenumber from the models EU2 and SNA ($\delta k_{SNA} - \delta k_{EU2}$) as a function of frequency (solid line) for toroidal (left) and spheroidal (right) fundamental mode. The long dashed line is the linear approximation to differential wavenumber, centered at 30 mHz; the short dashed line is the quadratic approximation. Over the entire band, the quadratic approximation is superior to the

linear; however, the linear approximation is quite good in a small region about the frequency of interest.

FIGURE 2.6

Comparison between $F_i C_{\tilde{u}u}$ and the linear and quadratic approximations as a function of the γ_i' at 25 mHz center frequency. Figure 2.6*a* illustrates the linear approximation (Love wave - left; Rayleigh wave - right); Figure 2.6*b* illustrates the quadratic approximation. In each box, the numerical correlation functions are designated by a solid line; the analytic correlation functions, calculated using the actual values from the EU2/SNA differential wavenumber, are designated by dashed lines (the dashed line *is not* a model derived by waveform fitting of the analytic expression). The quadratic approximation provides a good fit to the numerical correlation function for the range of γ_i' investigated (0.01 – 0.3). The fit between the linear approximation and the numerical correlation function improves as the filter becomes increasingly narrow.

FIGURE 2.7

Illustration of the effect of the differential time parameters on $F_i C_{\tilde{u}u}(t)$. The top three figures demonstrate the effect of the individual time parameters on the filtered cross-correlation function (solid line) compared to the filtered autocorrelation function (dashed line). $\delta\tau_p^m$ shifts the peak of the correlation function; $\delta\tau_g^m$ shifts the peak of the envelope; $\delta\tau_a^m$ changes the apparent center frequency. If $\delta\tau_g^m$ shifts the envelope more than one cycle, $\delta\tau_p^m$ will contain a $2n\pi/\omega_i'$ phase ambiguity. The combination of the three parameters is presented in the bottom figure.

FIGURE 2.8

The fundamental-mode surface waves used in the synthetic, isolated waveform example. The left panel presents the Love waves calculated from EU2 (top) and SNA (bottom); the right panel presents the Rayleigh wave calculated from EU2 (top) and SNA (bottom).

FIGURE 2.9

$F_i C_{\tilde{u}u}(t)$ (solid line) calculated from the fundamental-mode Love (left) and Rayleigh (right) waves. The top waveform in each box is the unfiltered correlation function; the bottom three illustrate filters with varying center frequencies (1 = 35, 2 = 25, and 3 = 15 mHz). The relative half-width of the applied filter (γ_i) was 0.1 in these examples. The dashed line indicates the best-fitting model of $F_i C_{\tilde{u}u}(t)$ determined by waveform fitting of (2.30).

FIGURE 2.10

Results of the application of the narrow-band technique to the isolated waveforms in Figure 2.8. The top three panels present the measured values of the "generalized seismological data functionals" ($\delta\tau_p^m$, $\delta\tau_g^m$, and $\delta\tau_a^m$) (squares) with the actual values (solid line) for the fundamental Love wave; the bottom three panels present the

Rayleigh wave results. Figure 2.10*a* illustrates the measurements prior to the cycle-skipping correction. Large discontinuities are visible in the estimates of $\delta\tau_p^m$, indicating that cycle skipping is a problem. Figure 2.10*b* displays the measurements after applying the correction. The estimates of $\delta\tau_p^m$ are excellent, with errors around 1 s or so. The estimates of $\delta\tau_g^m$ are also quite good, with somewhat greater scatter. The estimates of $\delta\tau_a^m$ show the greatest error, which may be attributed to the inherent difficulty of estimating attenuation.

FIGURE 2.11

The Fréchet kernels for differential phase delay (2.42). In the case of an isolated waveform, the Fréchet kernel for $\delta\tau_p^m$ at fixed frequency is simply related to the Fréchet kernel for $\delta\omega$ at fixed wavenumber. Figure 2.11*a* presents the Fréchet kernels for shear-wave velocity (solid line) and density (dashed line) for the fundamental Love wave as a function of depth. Figure 2.11*b* presents the Fréchet kernels for shear-wave velocity (solid line) and density (dashed line), as well as the kernel for compressional velocity (short dashed line) for the fundamental Rayleigh wave as a function of depth. In both *a* and *b*, the kernels are presented at four frequencies. The two horizontal dashed lines indicate the 400 km discontinuity and the base of the crust. While the kernels for both the Love and Rayleigh waves become localized at shallow depth with increasing frequency, the Rayleigh-wave kernels are consistently more sensitive to deeper structure than the Love-wave kernels.

FIGURE 2.12

Comparison between $\delta\tau_p^m$ estimated by waveform fitting (squares) and the predictions from first-order perturbation theory (filled circles), with the actual values (solid line). The values predicted from first-order perturbation theory were obtained by integrating the kernels with the known velocity perturbation between EU2 and SNA. The comparison is excellent, considering the difference between the models.

FIGURE 2.13

Flow chart illustrating the implementation of the narrow-band technique. (*a*) illustrates the processing procedure; (b) indicates the hierarchy of parameters recovered from the waveform-analysis procedure.

FIGURE 2.14

Demonstration of the peak-and-trough technique for estimating parameters of the filtered correlation functions. The left panel illustrates the determination of ω_i' and σ_i' from $F_i C_{\bar{u}\bar{u}}$ and the right panel illustrates the determination of the phase shift $\delta\tau_p(\omega_i')$, the group shift $\delta\tau_g(\omega_i')$, and the total attenuation shift $\delta\tau_a(\omega_i')$ from $F_i C_{\bar{u}\bar{u}}$. In the first step, we use the amplitudes, a_1 and a_2, to estimate the relative bandwidth parameter: $\gamma_i'^2 \approx -(2/\pi^2) \ln(-a_1)$. In the second step, we use the location of the peaks to estimate the center frequency $\omega_i' \approx 2\pi (t_2 - t_1)^{-1} (1 - \gamma_i'^2)^{-1}$. From consideration of the cross-correlation function we estimate $\delta\tau_a$, $\delta\tau_g$, and $\delta\tau_p$. The estimate of $\delta\tau_a$ is derived from the location of the secondary peaks, t_1'

and t_2', $\delta\tau_a \approx (1/2\pi\gamma_i'^2)(t_2' - t_1') - (1/\omega_i'\gamma_i'^2)(1 - \gamma_i'^2)$. The estimate of $\delta\tau_g$ is determined from the amplitude of the secondary peaks, a_1' and a_2', $\delta\tau_a \approx [(t_2'^2 - t_1'^2) - (2/\omega_i'^2\gamma_i'^2)\ln(a_2'/a_1')]/[2(t_2' - t_1')]$. Finally, the estimate of $\delta\tau_p$ depends on the peak time of the cross-correlation function, $\delta\tau_c$, $\delta\tau_p \approx \delta\tau_c - \gamma_i'^2(1 - \omega_i'\delta\tau_a)(\delta\tau_g - \delta\tau_c)$. These initial guesses are used in the appropriate Gaussian-wavelet expression to calculate the analytic correlation function for waveform-fitting procedure.

FIGURE 2.15

$F_i C_{\tilde{u}\tilde{s}}$ (solid line) calculated from the fundamental-mode Love (left) and Rayleigh (right) waves. The top waveform in each box is the unfiltered correlation function; the bottom three illustrate filters with center frequencies of 35, 25, and 25 mHz. The relative half-width of the applied filter (γ_i) was 0.1 in these examples. The dashed line indicates the best-fitting model of $F_i C_{\tilde{u}\tilde{s}}(t)$ determined by waveform fitting of equation (2.51). The unfiltered correlation function is not strongly influenced by interference at zero lag, but the energy appearing between -400 and -200 s influences the filtered correlation functions at low frequencies. This is reflected as bias in the estimates of the generalized data functionals, particularly in the differential group delay. The estimates of $\Delta\tilde{\tau}_p$, $\Delta\tilde{\tau}_g$, and $\Delta\tilde{\tau}_a$ recovered from waveform fitting are displayed in Figure 2.16.

FIGURE 2.16

Estimates of $\Delta\tilde{\tau}_p$, $\Delta\tilde{\tau}_g$, and $\Delta\tilde{\tau}_a$ from the waveform fitting of (2.51) to $F_i C_{\tilde{u}\tilde{s}}$. If the effect of interference is small, then these values will be near zero. The top three panels present the measured values for the fundamental Love wave; the bottom three panels present the fundamental Rayleigh wave results. The estimates of $\Delta\tilde{\tau}_p$ are near zero for both the Love and Rayleigh wave, indicating that the phase measurements are not strongly biased. On the other hand, the estimates of $\Delta\tilde{\tau}_g$ at frequencies less than 20 mHz are non-zero, particularly for the Rayleigh wave. The estimates of $\Delta\tilde{\tau}_a$ show some variance from the isolated waveform example as well.

FIGURE 2.17

Comparison of the measured values of $\Delta\tilde{\tau}_p$ and $\Delta\tilde{\tau}_g$ (squares) with the values predicted from the expressions (2.56) and (2.60) (filled circles). The prediction of the average phase delay is excellent across the entire frequency range. The prediction of the average group delay is quite good for the Love wave, but manifests some problems for the Rayleigh wave.

FIGURE 2.18

$F_i C_{\tilde{u}\tilde{s}}(t)$ (solid line) calculated from the fundamental-mode Love (left) and Rayleigh (right) waves. The top waveform in each box is the unfiltered correlation function; the bottom three illustrate filters with center frequencies of 35, 25, and 15 mHz. The relative half-width of the applied filter (γ_i) was 0.1 in these examples. The symbols indicate the best-fitting model of $F_i C_{\tilde{u}\tilde{s}}(t)$ determined by waveform fitting of equation (2.67). The influence of interfering arrivals may be observed in the 15

mHz example. The estimates of $\delta\tau_p$, $\delta\tau_g$, and $\delta\tau_a$ obtained from waveform fitting are displayed in Figure 2.19.

FIGURE 2.19

Estimates of the differential phase, group, and attenuation delays (squares) derived from waveform fitting of (2.67) to $F_i C_{\tilde{u}s}(t)$ in the EU2/SNA fundamental-mode example, with the actual values (solid line). The phase delay measurements are very stable with respect to interference, while the group and attenuation delay are biased at low frequencies.

FIGURE 2.20

The Fréchet kernels for differential phase delay (2.71). Figure 2.20*a* presents the Fréchet kernels for shear-wave velocity (solid line) and density (dashed line) for the fundamental Love wave as a function of depth. Figure 2.20*b* presents the Fréchet kernels for shear-wave velocity (solid line) and density (dashed line), as well as the kernel for compressional velocity (short dashed line) for the fundamental Rayleigh wave as a function of depth. In both *a* and *b*, the kernels are presented at four frequencies. The two horizontal dashed lines indicate the 400 km discontinuity and the base of the crust. These kernels are not very different from those in Figure 2.11 as the estimates of differential phase delay are not strongly influenced by interference in this example.

FIGURE 2.21

Comparison between $\delta\tau_p$ estimated by waveform fitting (squares) and the predictions from first-order perturbation theory (filled circles), with the actual values (solid line). The values predicted from first-order perturbation theory were obtained by integrating the kernels (2.71) with the known velocity perturbation between EU2 and SNA. The comparison is excellent, considering the differences between the models.

FIGURE 2.22

Comparison of the isolation filter $\tilde{f}(t)$ for the fundamental-mode surface waves of the Kamchatka event with the complete synthetic seismograms $\tilde{s}(t)$ (top) and the observed seismograms $s(t)$ (bottom). The comparison between $\tilde{f}(t)$ and $s(t)$ on the vertical component (right) is more favorable than the transverse component (left). The observed surface waves on both components show the effect of unmodeled interference, probably due to scattering.

FIGURE 2.23

Estimates of $\delta\tau_p$, $\delta\tau_g$, and $\delta\tau_a$ derived from waveform fitting of (2.67) to $F_i C_{\tilde{u}s}(t)$ for the Kamchatka event. The Rayleigh differential phase delay for this seismogram is near zero across the frequency range examined, suggesting EU2 is an appropriate model for *PSV* propagation along this path. This is not surprising as EU2 was derived from a study of fundamental and higher Rayleigh modes along

this path. However, the Love differential phase delay indicates that the SH propagation is faster than that predicted by EU2. This observation of apparent polarization anisotropy is consistent with the results of Aki and Kiminuna [1963], McEvilly [1964], Cara *et al.* [1980], Lévêque and Cara [1983], and Gee and Jordan [1988].

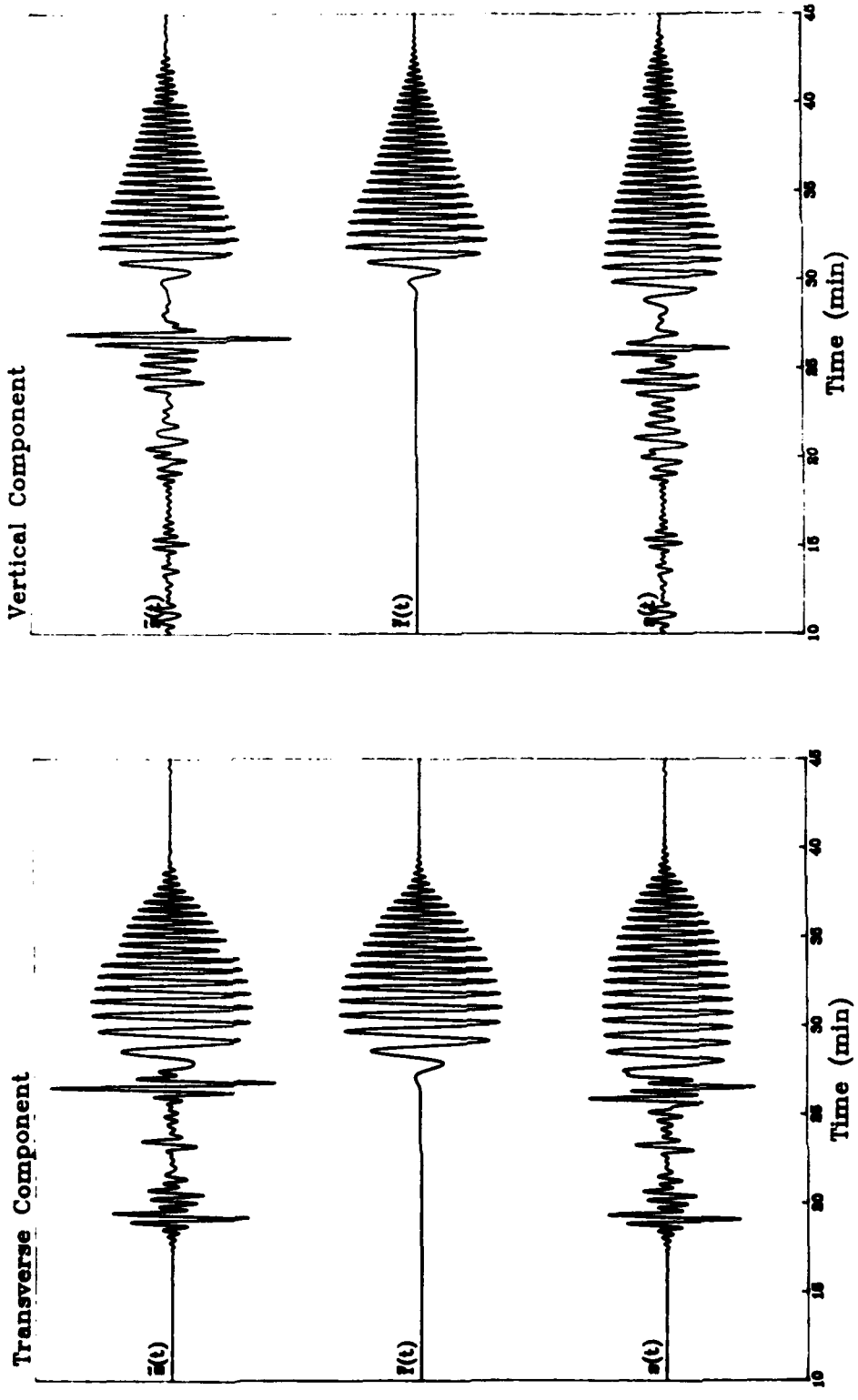

Vertical Component
Transverse Component
Time (min)
Time (min)

Figure 2.1

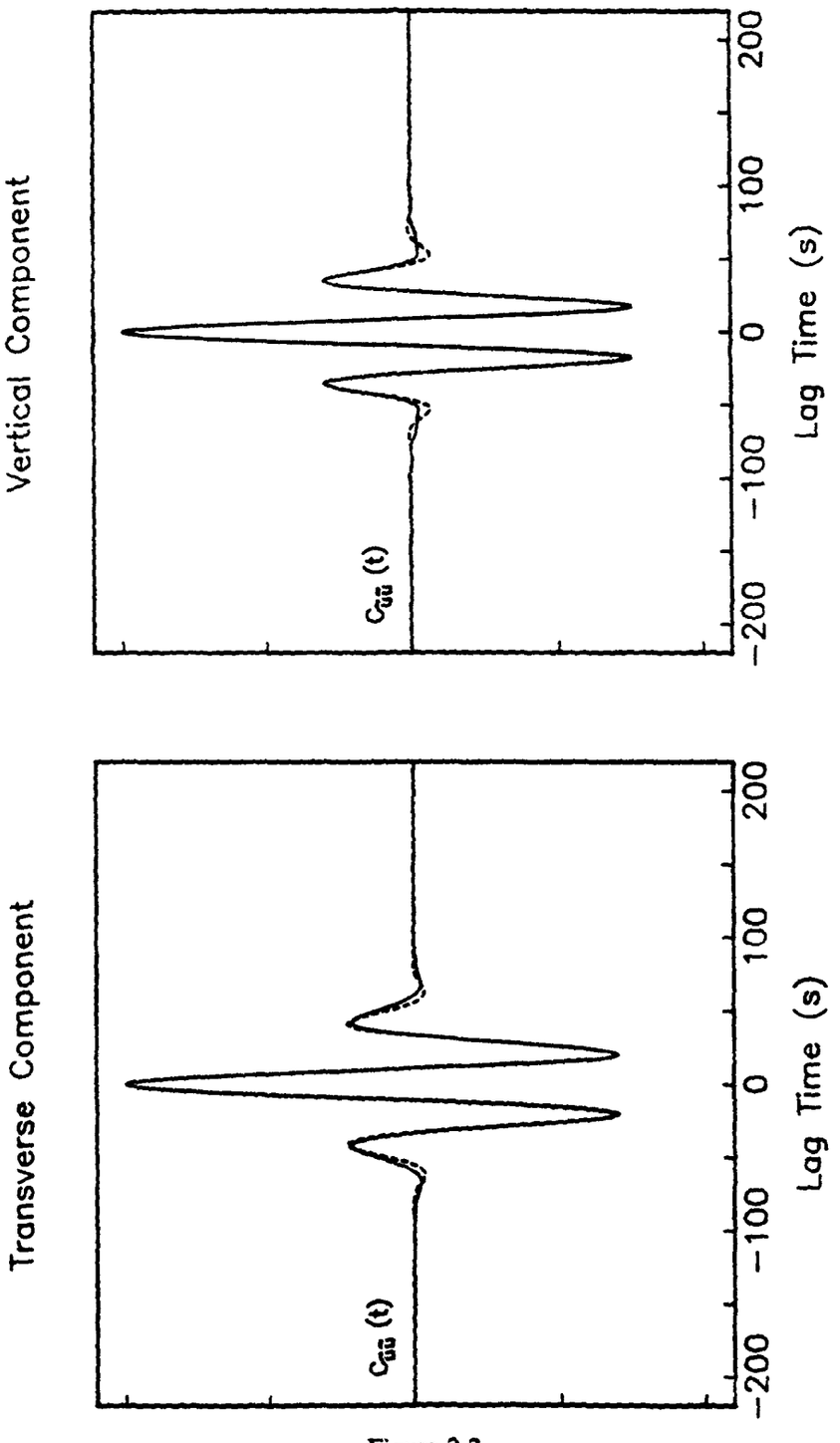

Vertical Component
Transverse Component
$C_{\tilde{u}\tilde{u}}$ (t)
$C_{\tilde{u}\tilde{u}}$ (t)
−200
−100
0
100
200
Lag Time (s)
Lag Time (s)

Figure 2.2

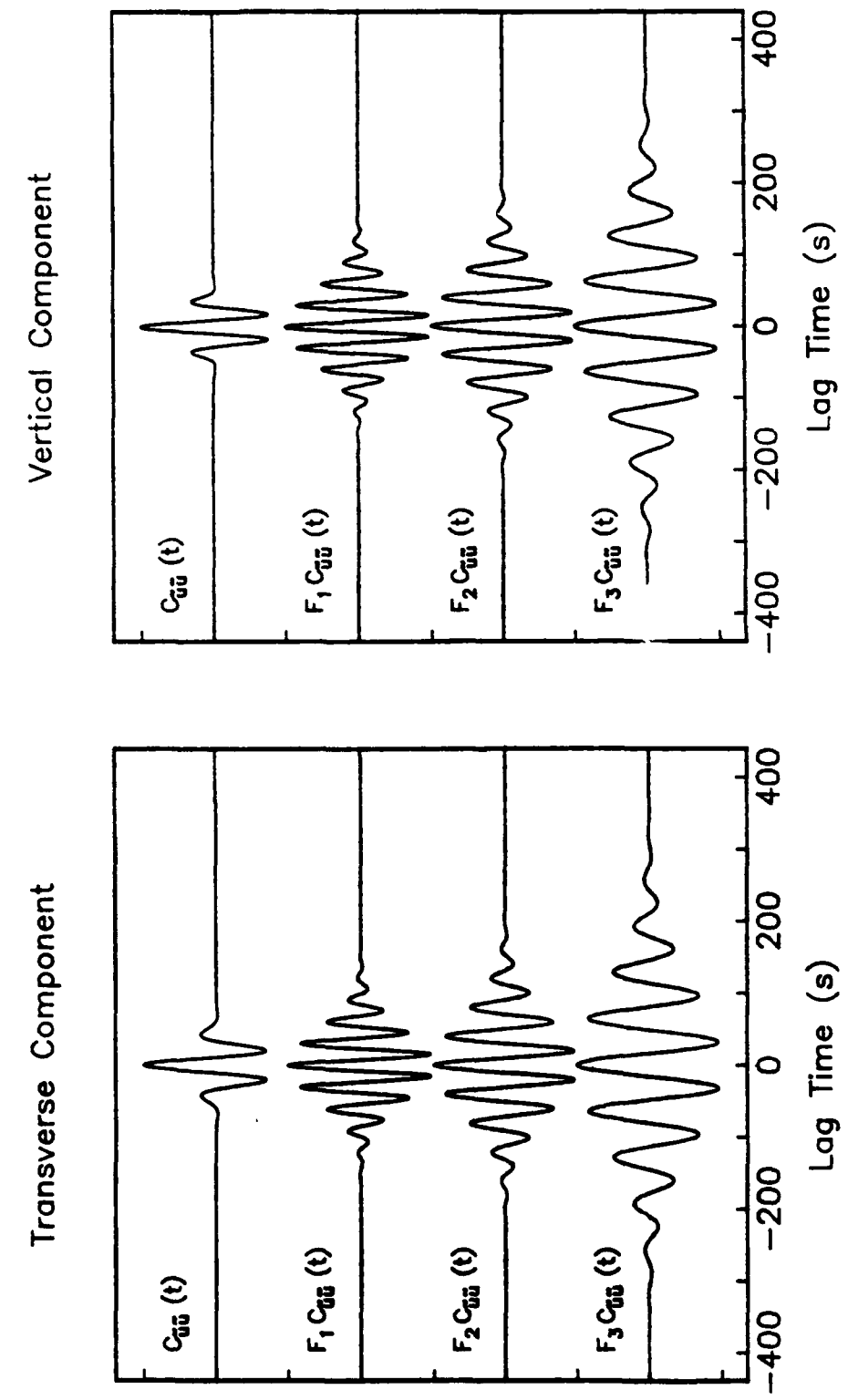

Vertical Component
Transverse Component
$C_{\bar{u}\bar{u}}$ (t)
$F_1 C_{\bar{u}\bar{u}}$ (t)
$F_2 C_{\bar{u}\bar{u}}$ (t)
$F_3 C_{\bar{u}\bar{u}}$ (t)
−400
−200
0
200
400
Lag Time (s)

Figure 2.3

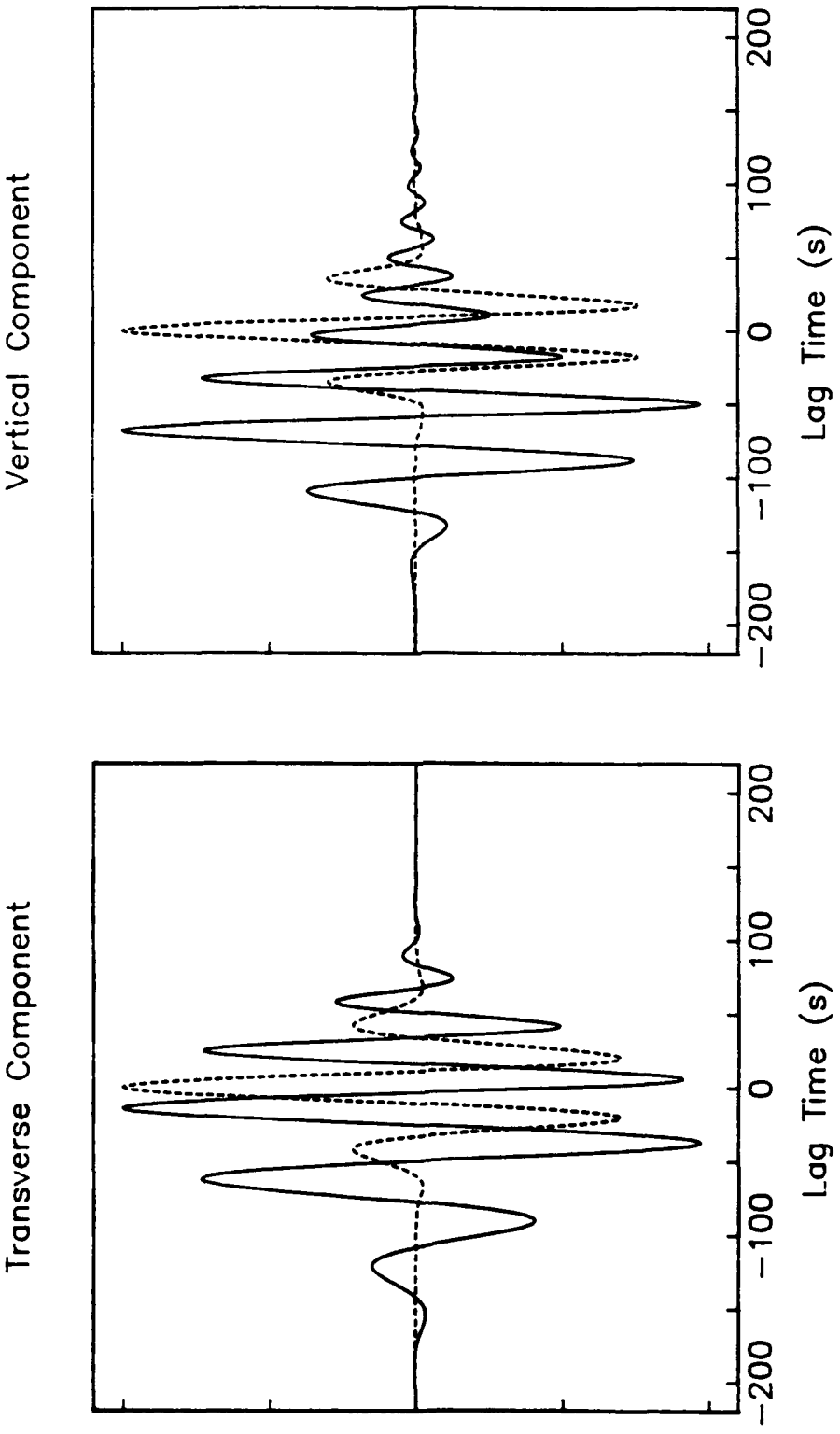

Vertical Component
Transverse Component
Lag Time (s)
Lag Time (s)
200
100
0
−100
−200
200
100
0
−100
−200

Figure 2.4

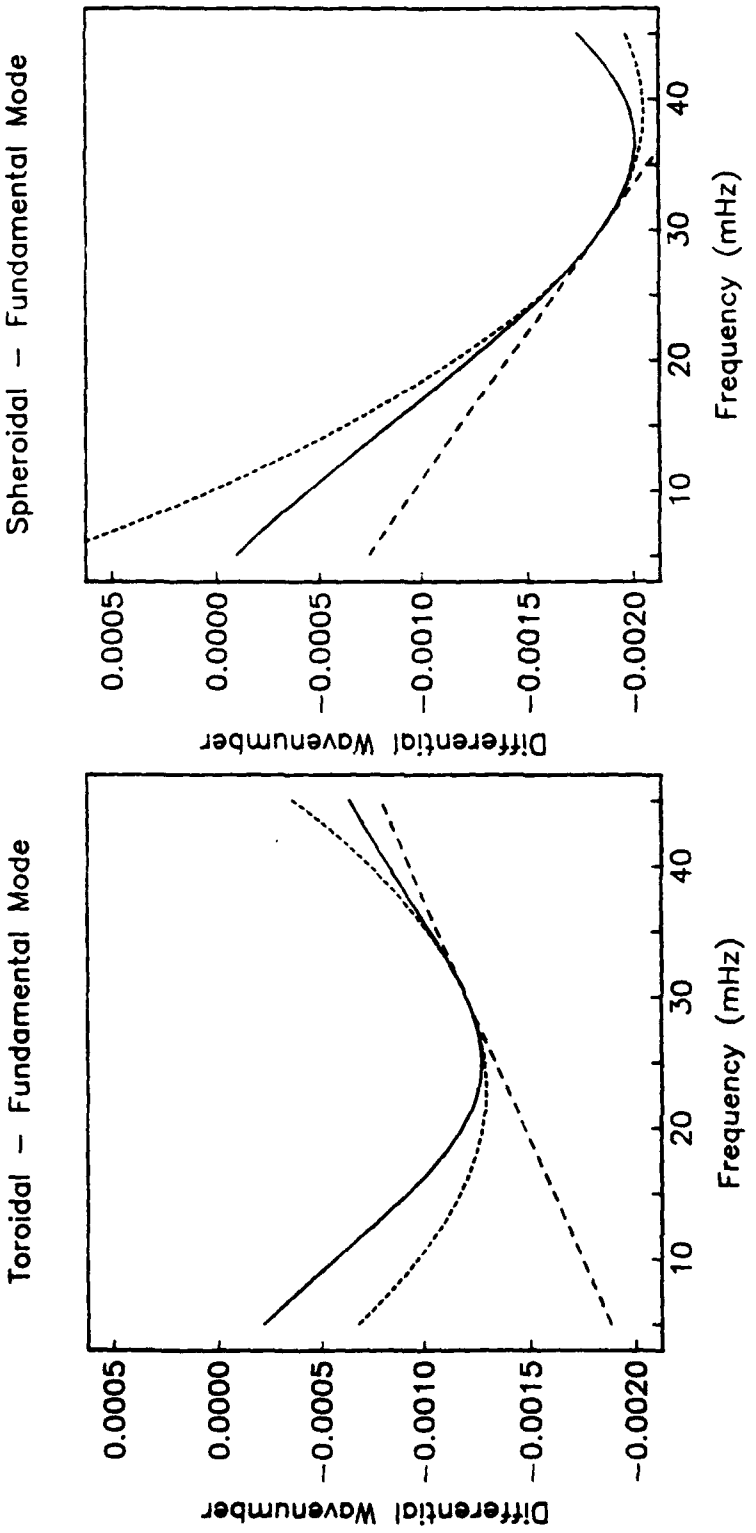

Spheroidal – Fundamental Mode
Toroidal – Fundamental Mode
Differential Wavenumber
Frequency (mHz)
0.0005
0.0000
−0.0005
−0.0010
−0.0015
−0.0020
10
20
30
40

Figure 2.5

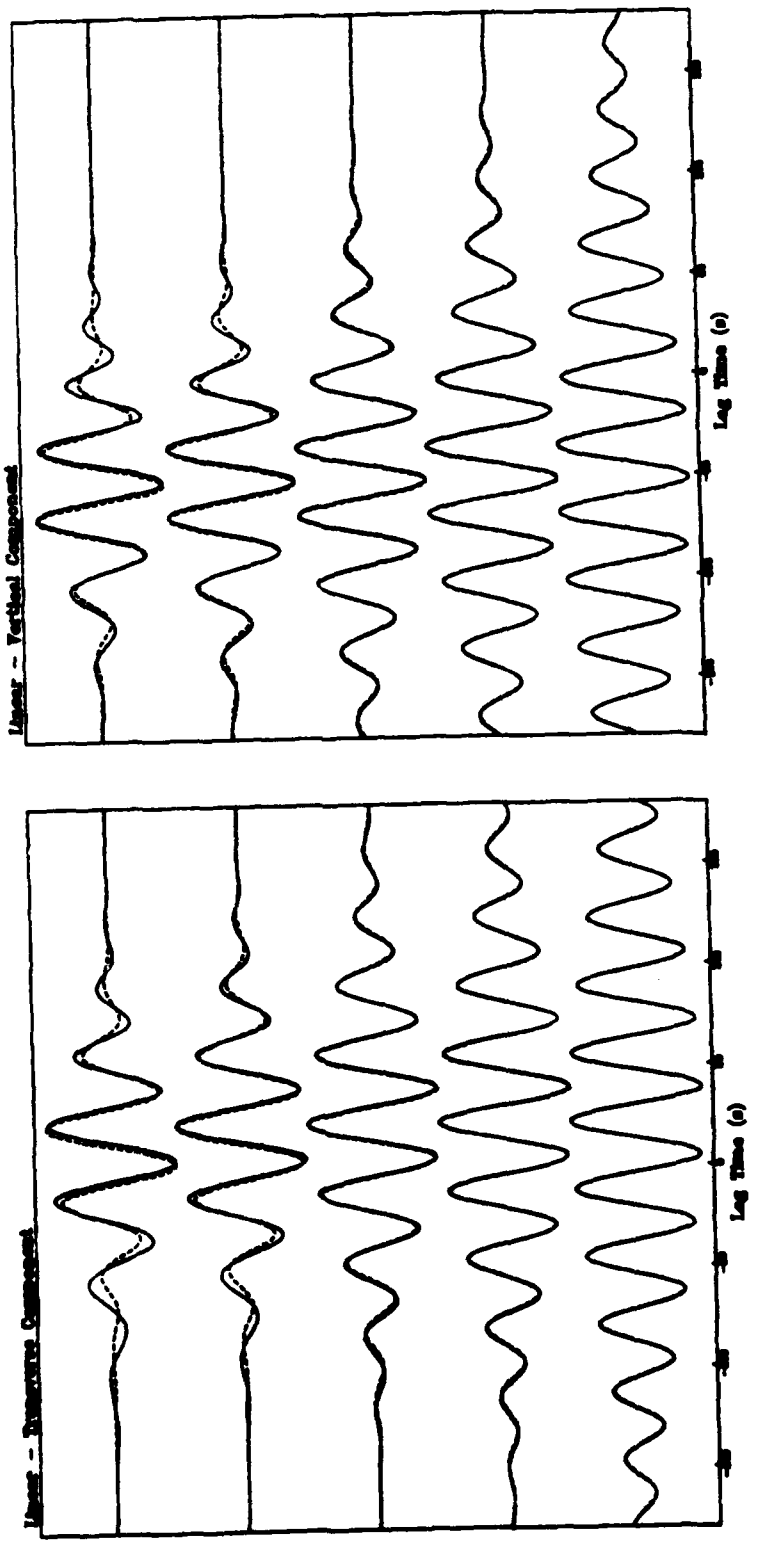


Figure 2.6a

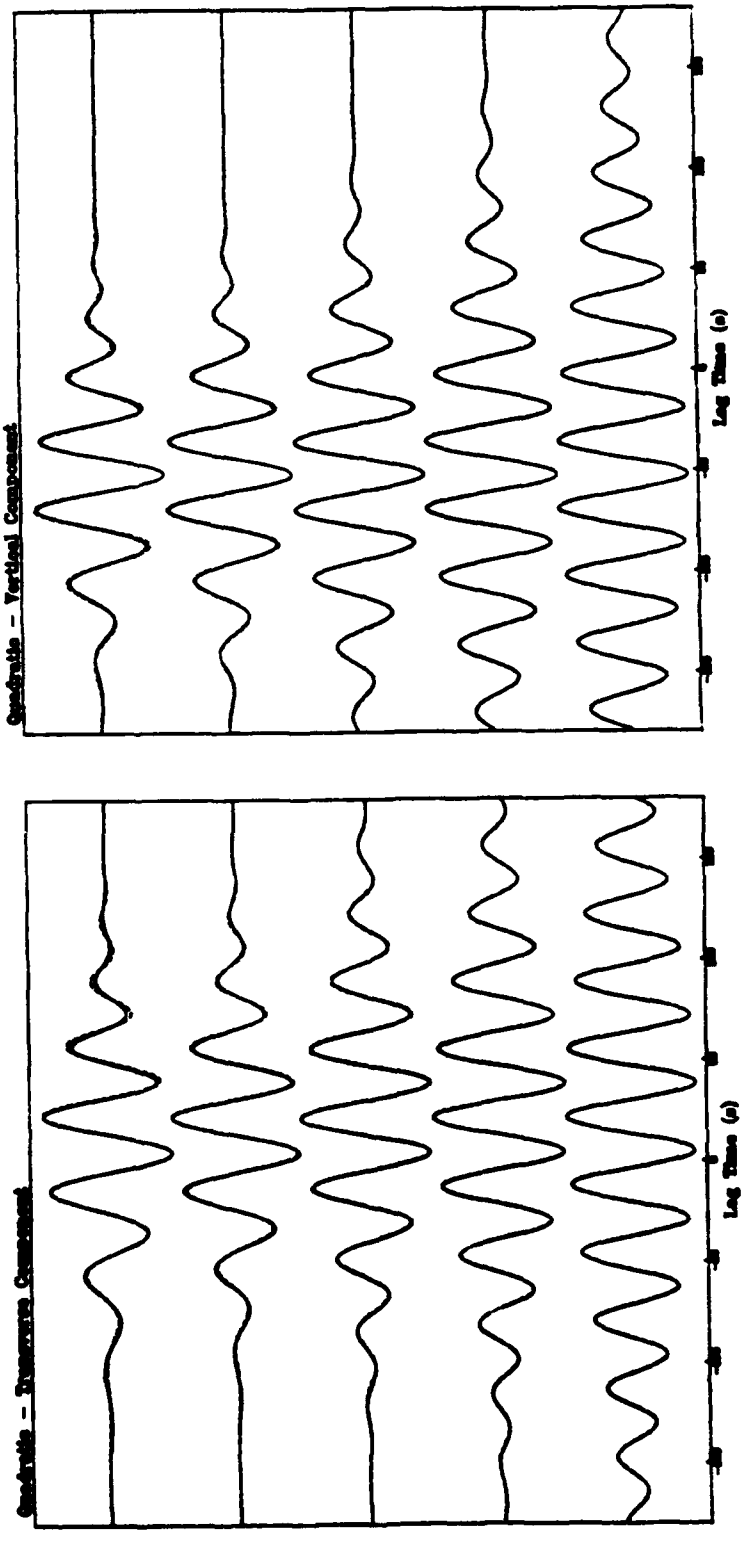

Quadrature - Vertical Component
Lag Time (s)
Quadrature - Transverse Component
Lag Time (s)

Figure 2.6b

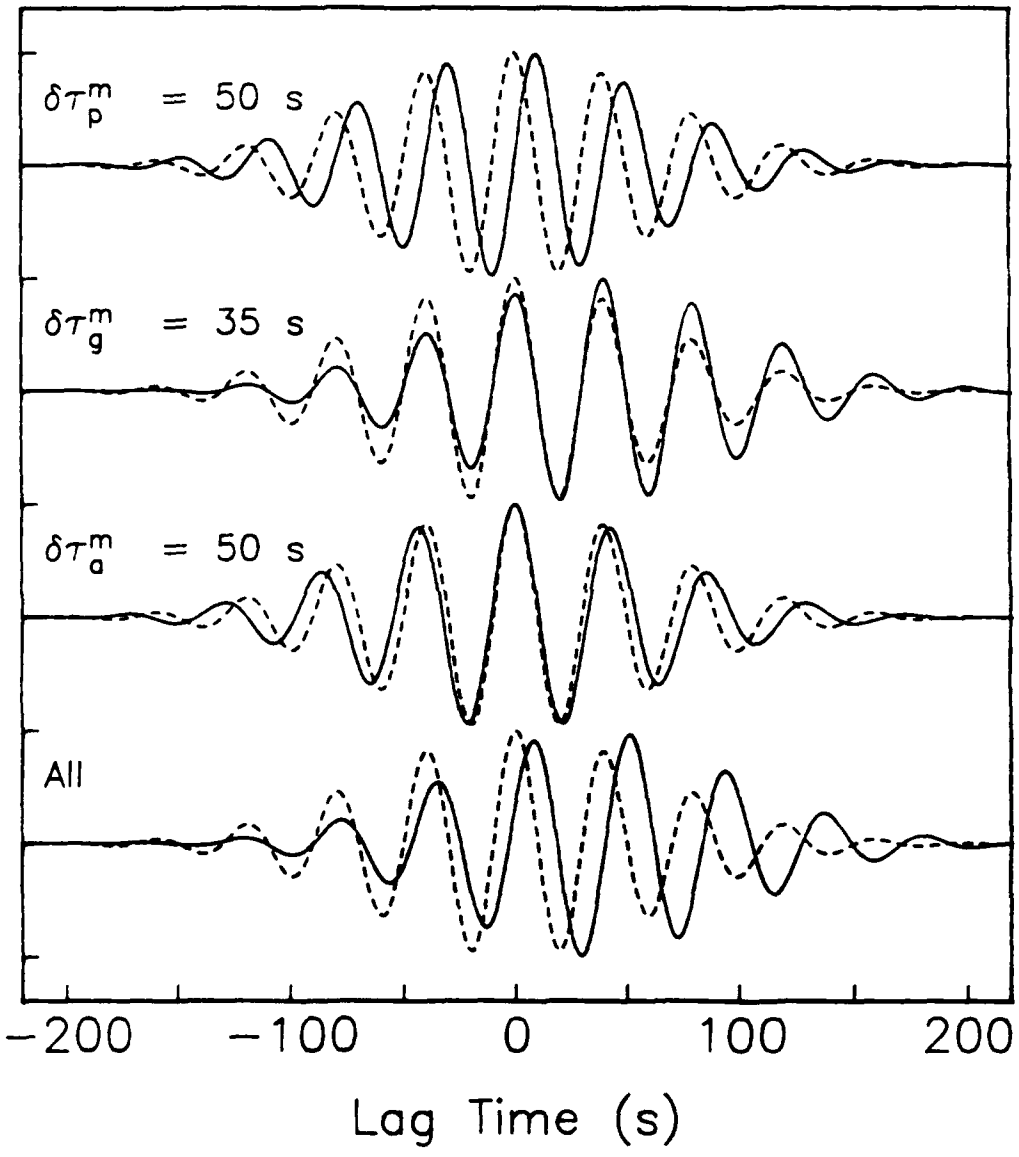

$\delta\tau_p^m$ = 50 s
$\delta\tau_g^m$ = 35 s
$\delta\tau_a^m$ = 50 s
All
−200
−100
0
100
200
Lag Time (s)

Figure 2.7

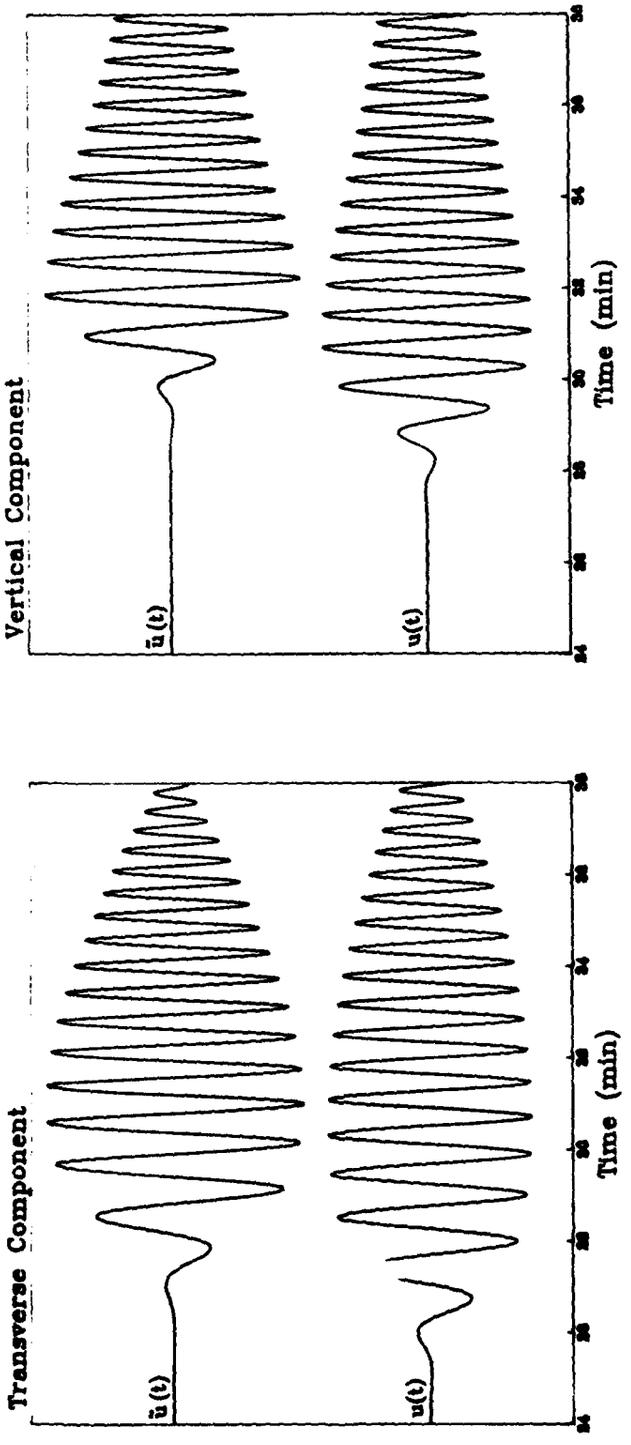

Vertical Component
Transverse Component
Time (min)
Time (min)

Figure 2.8

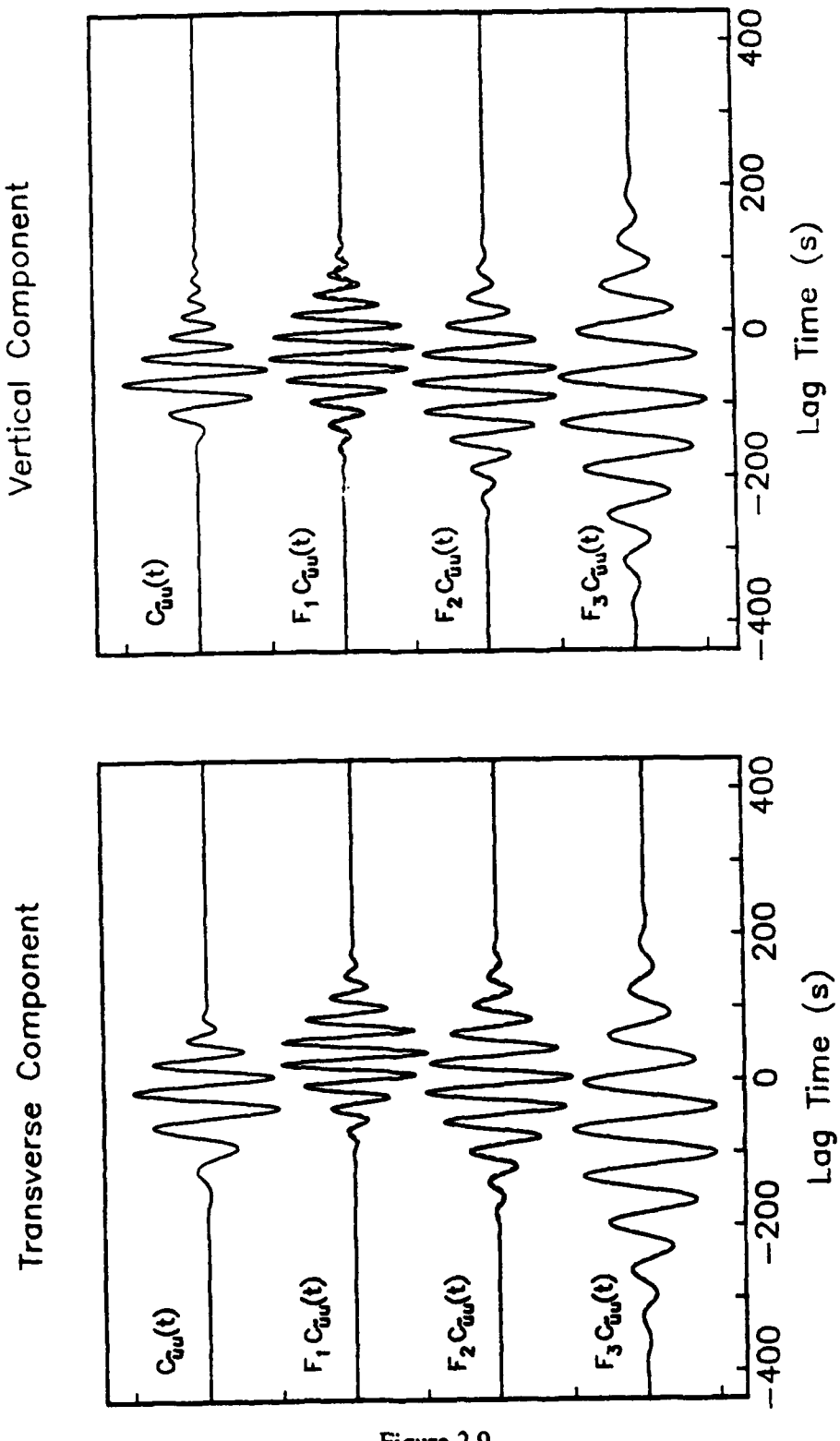

Vertical Component
Transverse Component
$C_{\tilde{u}\tilde{u}}(t)$
$F_1 C_{\tilde{u}\tilde{u}}(t)$
$F_2 C_{\tilde{u}\tilde{u}}(t)$
$F_3 C_{\tilde{u}\tilde{u}}(t)$
400
200
0
−200
−400
Lag Time (s)

Figure 2.9

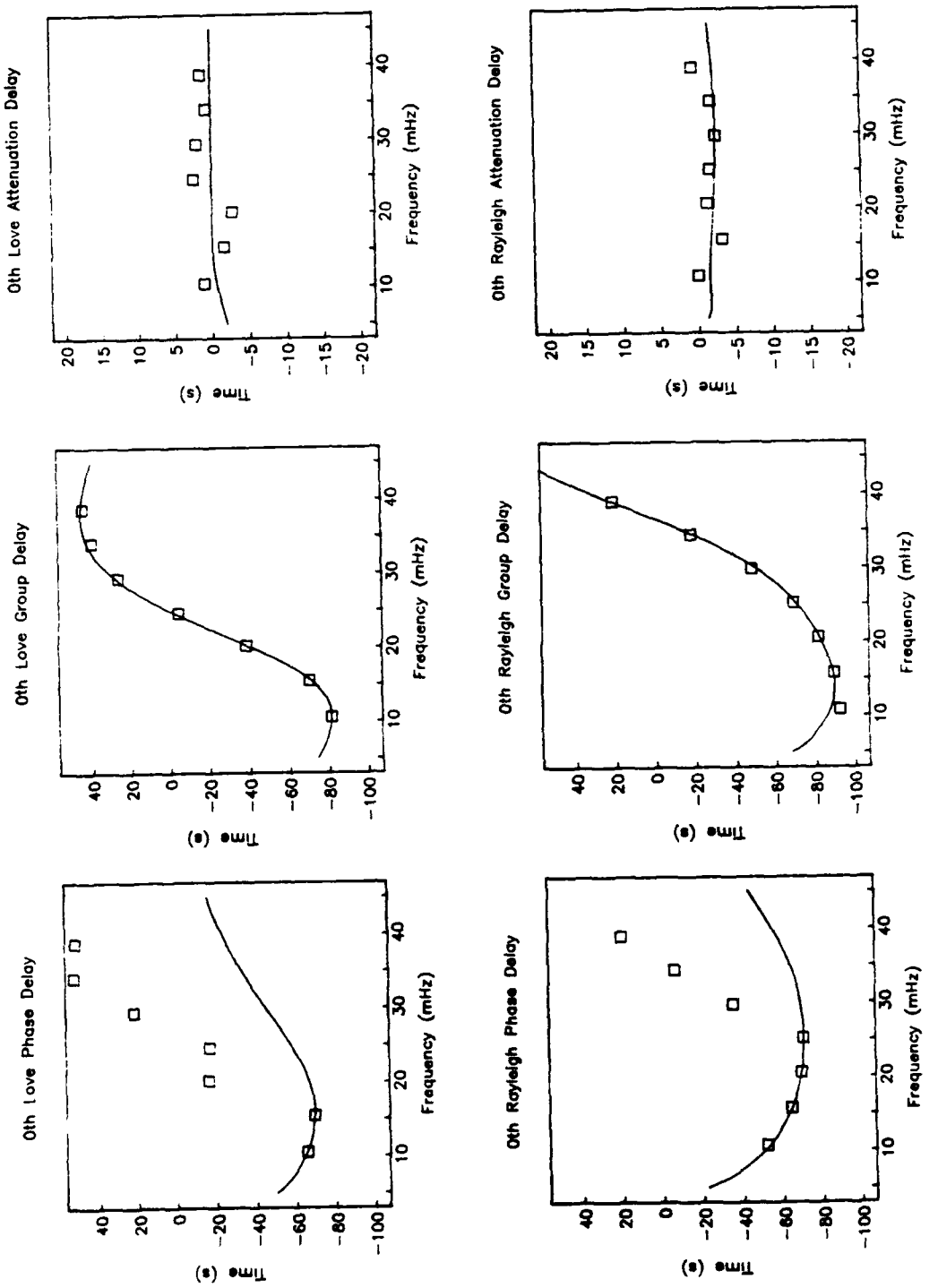

0th Love Attenuation Delay
0th Rayleigh Attenuation Delay
0th Love Group Delay
0th Rayleigh Group Delay
0th Love Phase Delay
0th Rayleigh Phase Delay
Time (s)
Frequency (mHz)

Figure 2.10a

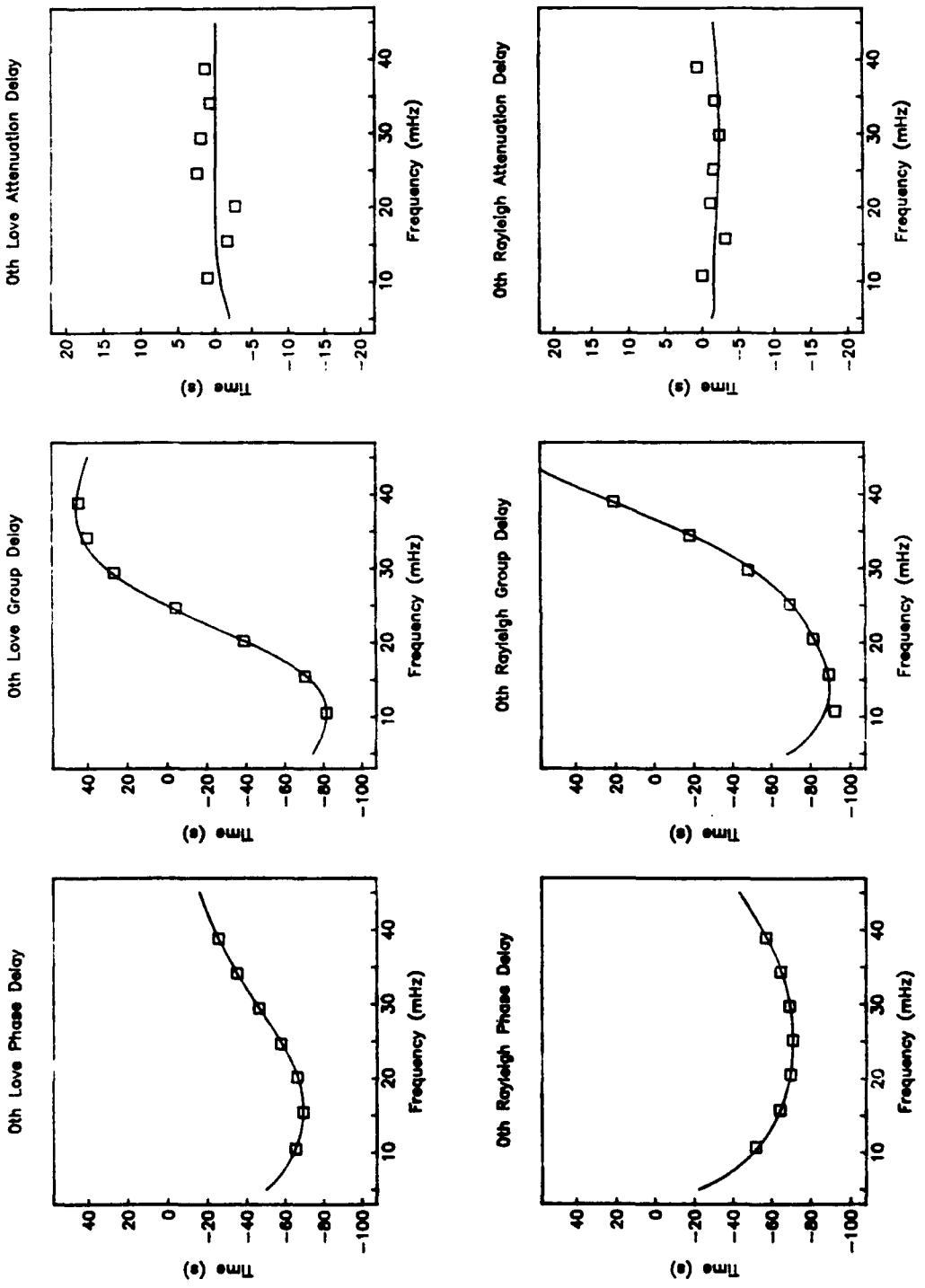

0th Love Phase Delay
0th Love Group Delay
0th Love Attenuation Delay
0th Rayleigh Phase Delay
0th Rayleigh Group Delay
0th Rayleigh Attenuation Delay
Time (s)
Frequency (mHz)

Figure 2.10b

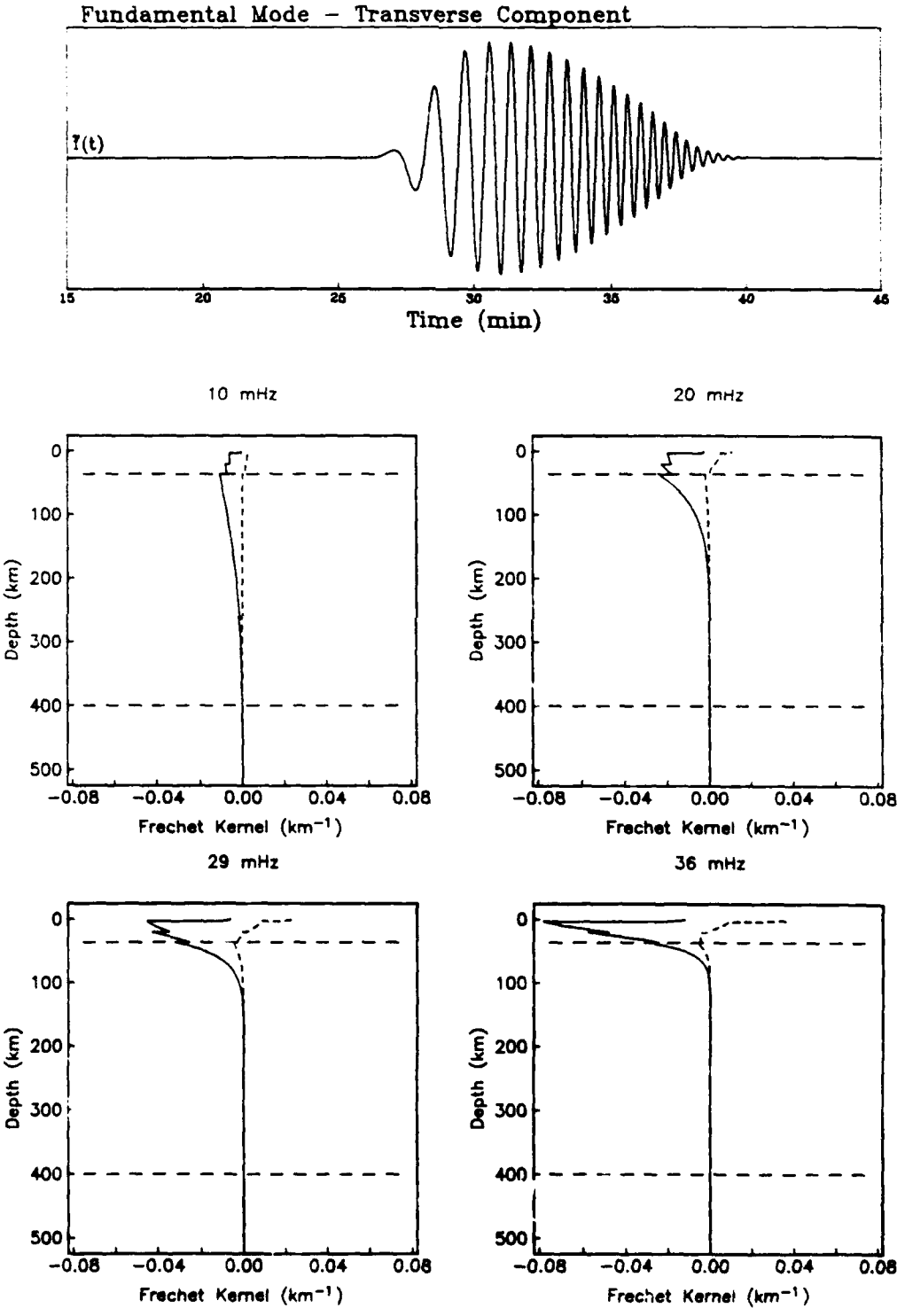

Fundamental Mode – Transverse Component
T(t)
15
20
25
30
35
40
45
Time (min)
10 mHz
20 mHz
29 mHz
36 mHz
Depth (km)
0
100
200
300
400
500
−0.08
−0.04
0.00
0.04
0.08
Frechet Kernel (km⁻¹)

Figure 2.11a

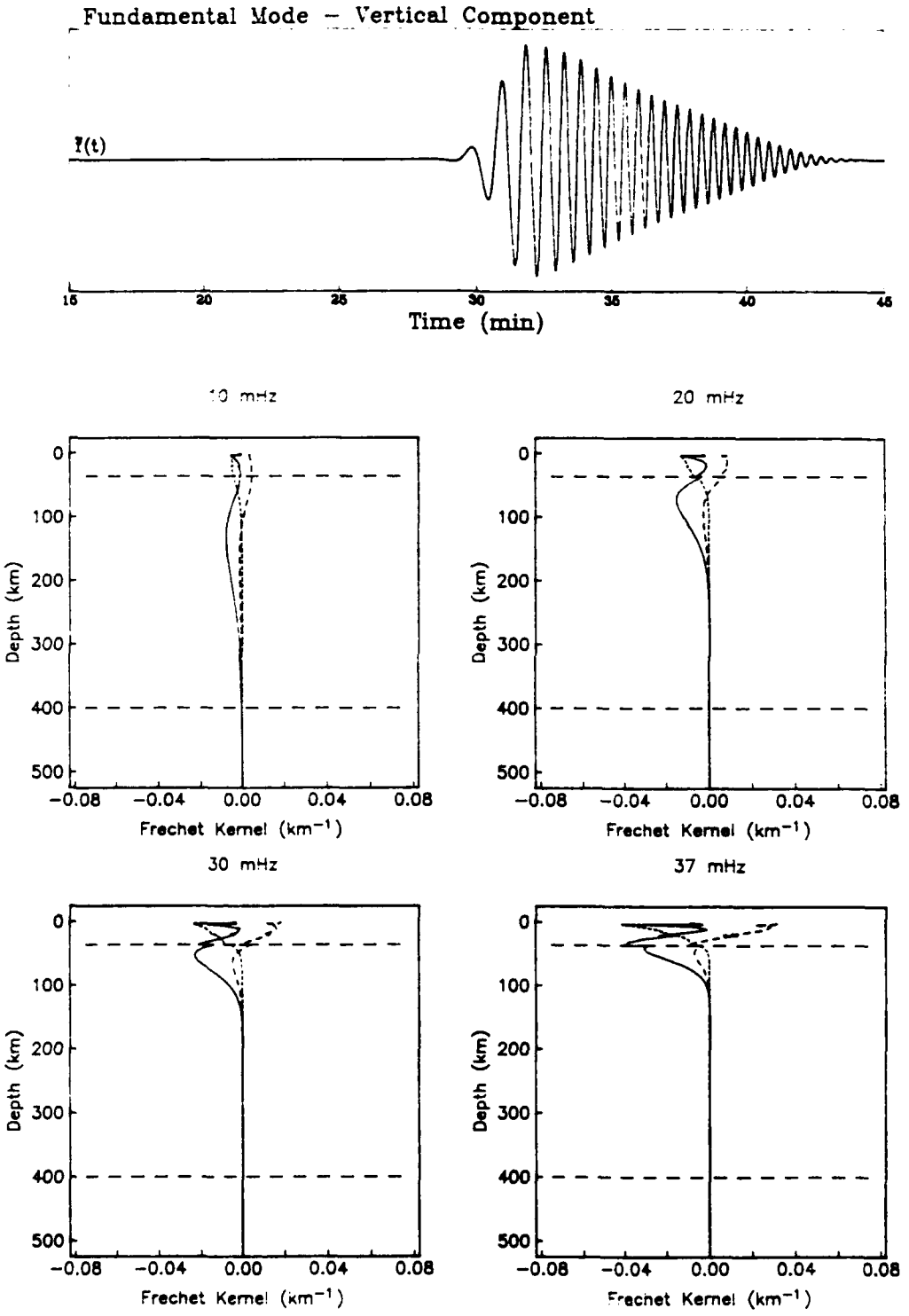

Fundamental Mode – Vertical Component
Y(t)
15
20
25
30
35
40
45
Time (min)
10 mHz
20 mHz
30 mHz
37 mHz
Depth (km)
0
100
200
300
400
500
−0.08
−0.04
0.00
0.04
0.08
Frechet Kernel (km⁻¹)

Figure 2.11b

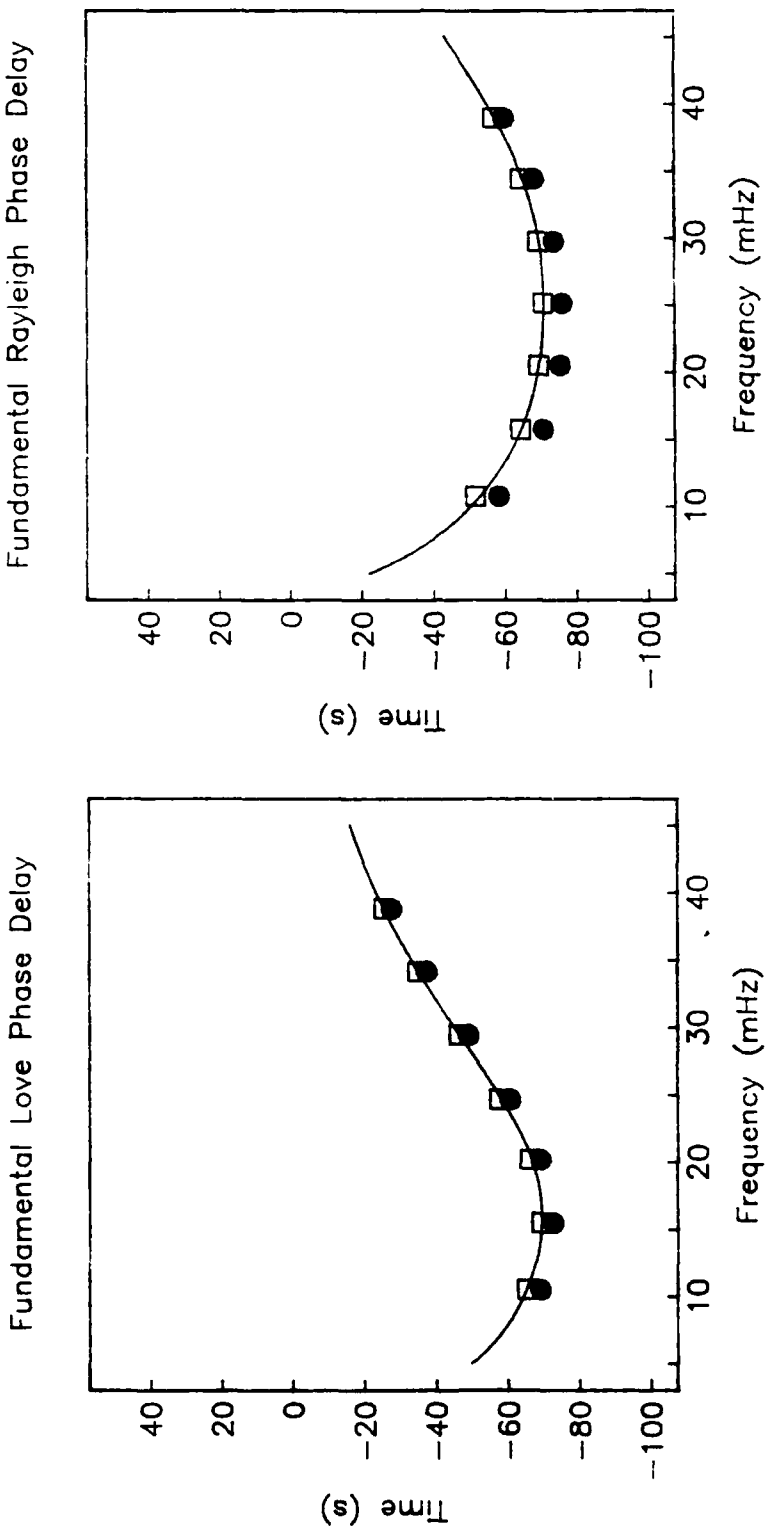

Fundamental Rayleigh Phase Delay
Time (s)
40
20
0
-20
-40
-60
-80
-100
10
20
30
40
Frequency (mHz)
Fundamental Love Phase Delay
Time (s)
40
20
0
-20
-40
-60
-80
-100
10
20
30
40
Frequency (mHz)

Figure 2.12

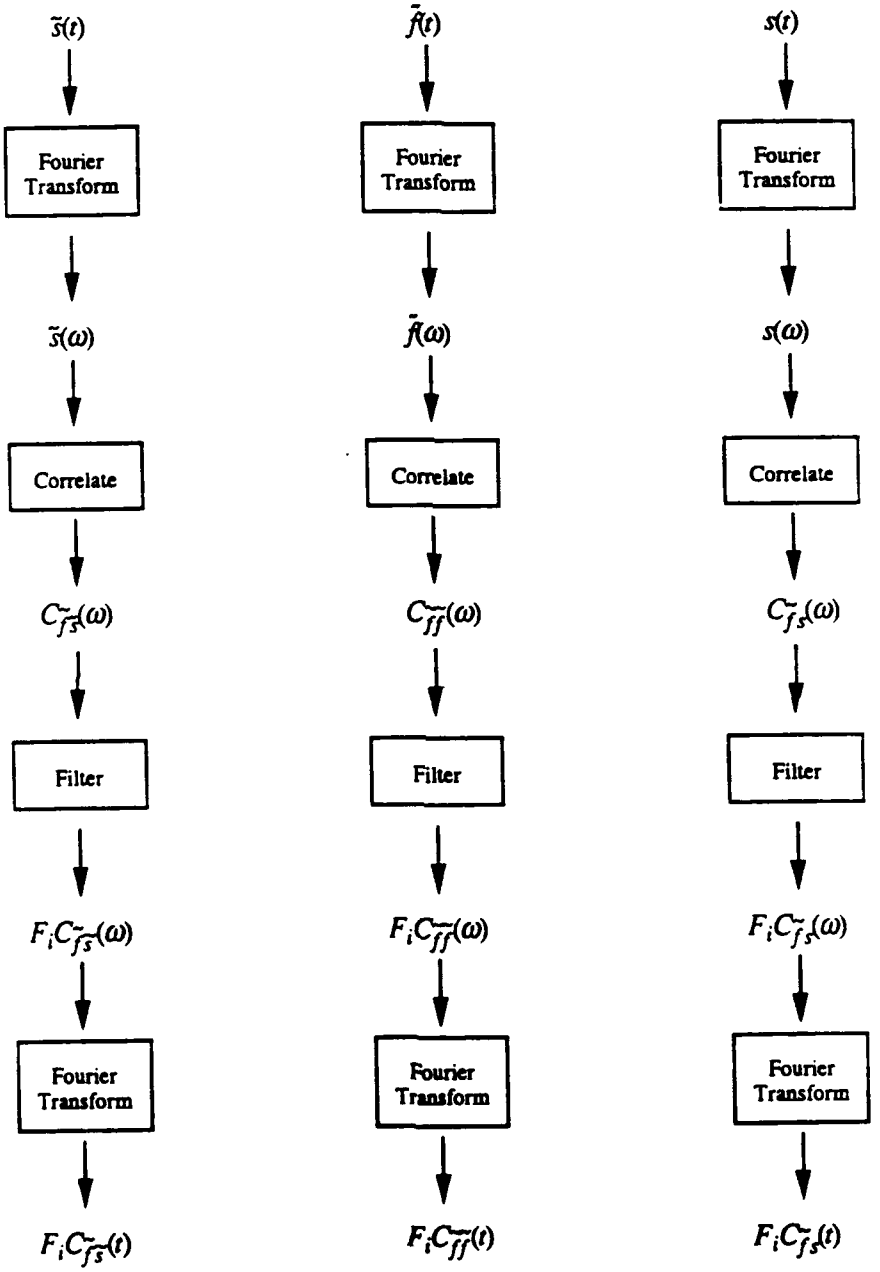

$\tilde{s}(t)$
Fourier Transform
$\tilde{s}(\omega)$
Correlate
$C_{\tilde{f}\tilde{s}}(\omega)$
Filter
$F_i C_{\tilde{f}\tilde{s}}(\omega)$
Fourier Transform
$F_i C_{\tilde{f}\tilde{s}}(t)$
$\tilde{f}(t)$
Fourier Transform
$\tilde{f}(\omega)$
Correlate
$C_{\tilde{f}\tilde{f}}(\omega)$
Filter
$F_i C_{\tilde{f}\tilde{f}}(\omega)$
Fourier Transform
$F_i C_{\tilde{f}\tilde{f}}(t)$
$s(t)$
Fourier Transform
$s(\omega)$
Correlate
$C_{\tilde{f}s}(\omega)$
Filter
$F_i C_{\tilde{f}s}(\omega)$
Fourier Transform
$F_i C_{\tilde{f}s}(t)$

Figure 2.13a

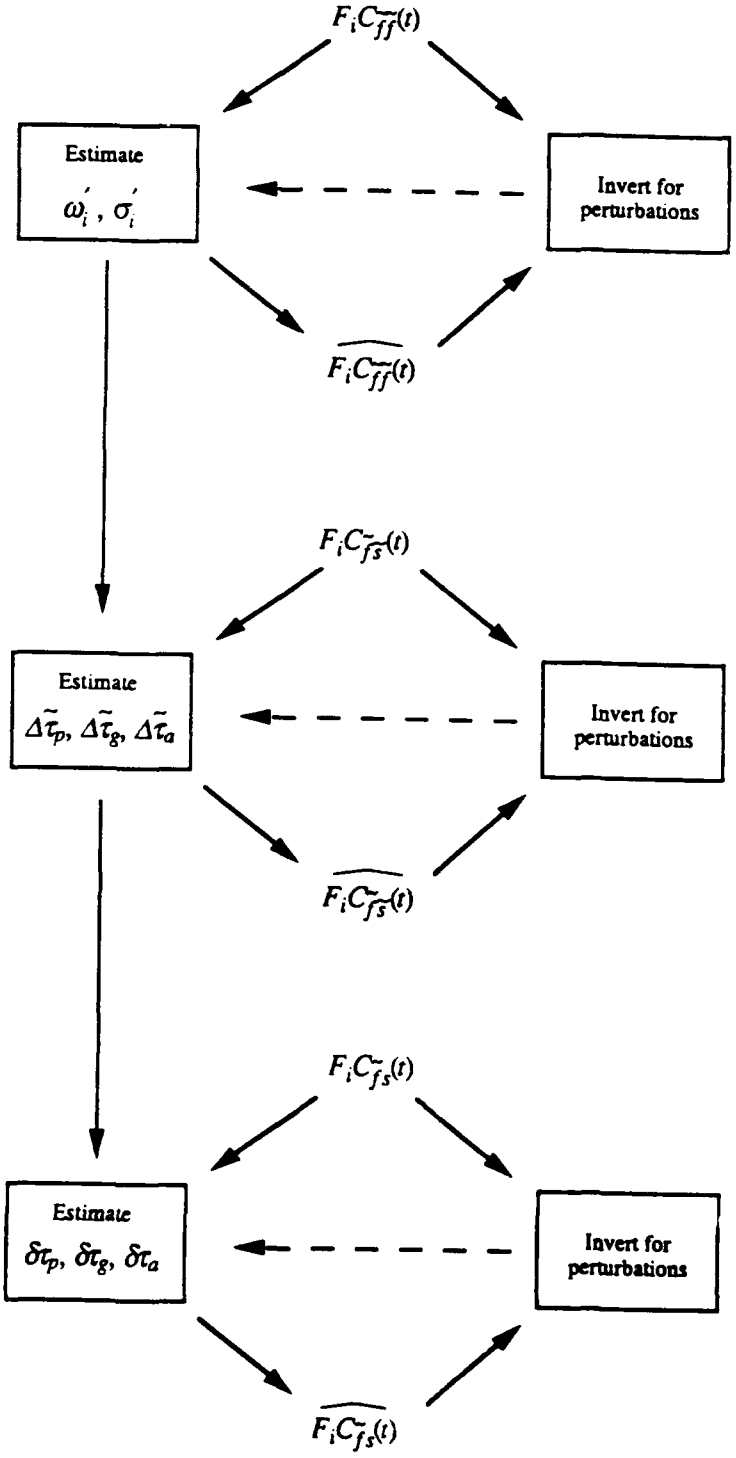

$F_i C_{ff}(t)$
Estimate
ω_i' , σ_i'
Invert for perturbations
$\widehat{F_i C_{ff}(t)}$
$F_i C_{fs}(t)$
Estimate
$\Delta\tilde{\tau}_p$, $\Delta\tilde{\tau}_g$, $\Delta\tilde{\tau}_a$
Invert for perturbations
$\widehat{F_i C_{fs}(t)}$
$F_i C_{fs}(t)$
Estimate
$\delta\tau_p$, $\delta\tau_g$, $\delta\tau_a$
Invert for perturbations
$\widehat{F_i C_{fs}(t)}$

Figure 2.13b

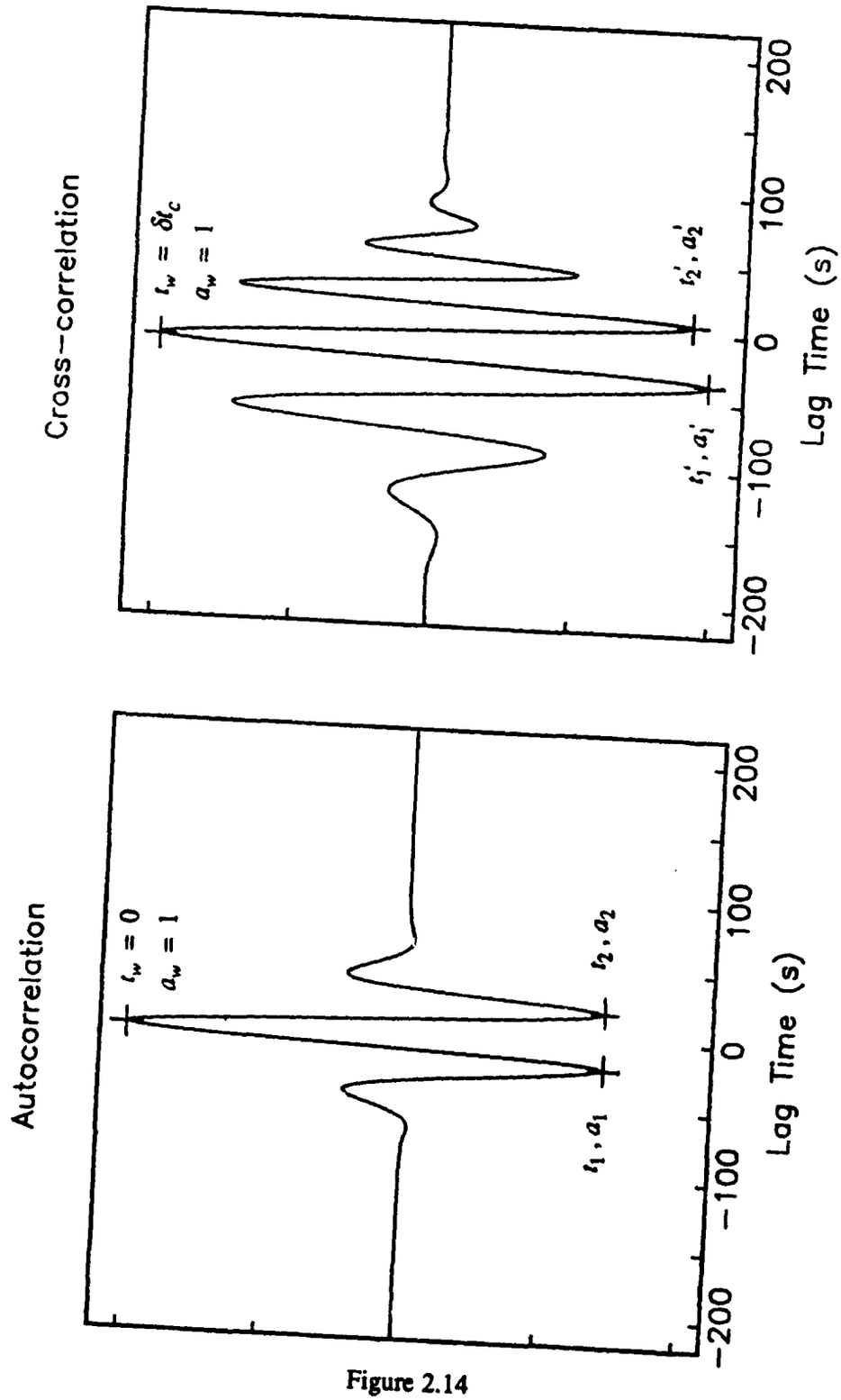

Cross–correlation
$t_w = \delta t_c$
$a_w = 1$
t_2', a_2'
t_1', a_1'
Lag Time (s)
200
100
0
−100
−200
Autocorrelation
$t_w = 0$
$a_w = 1$
t_2, a_2
t_1, a_1
Lag Time (s)
200
100
0
−100
−200

Figure 2.14

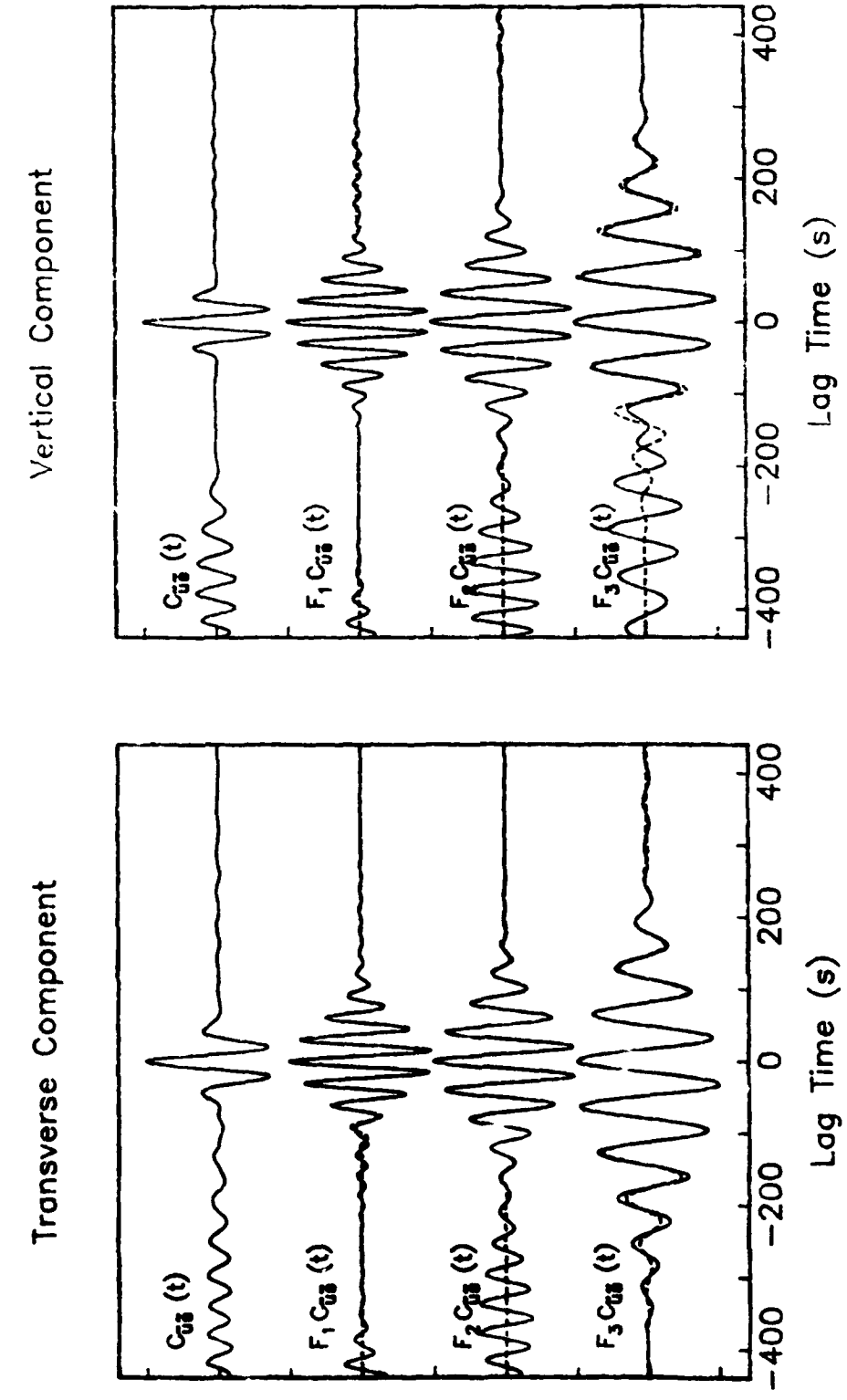

Vertical Component
Transverse Component
$C_{\bar{u}\bar{s}}$ (t)
$F_1 C_{\bar{u}\bar{s}}$ (t)
$F_2 C_{\bar{u}\bar{s}}$ (t)
$F_3 C_{\bar{u}\bar{s}}$ (t)
-400
-200
0
200
400
Lag Time (s)

Figure 2.15

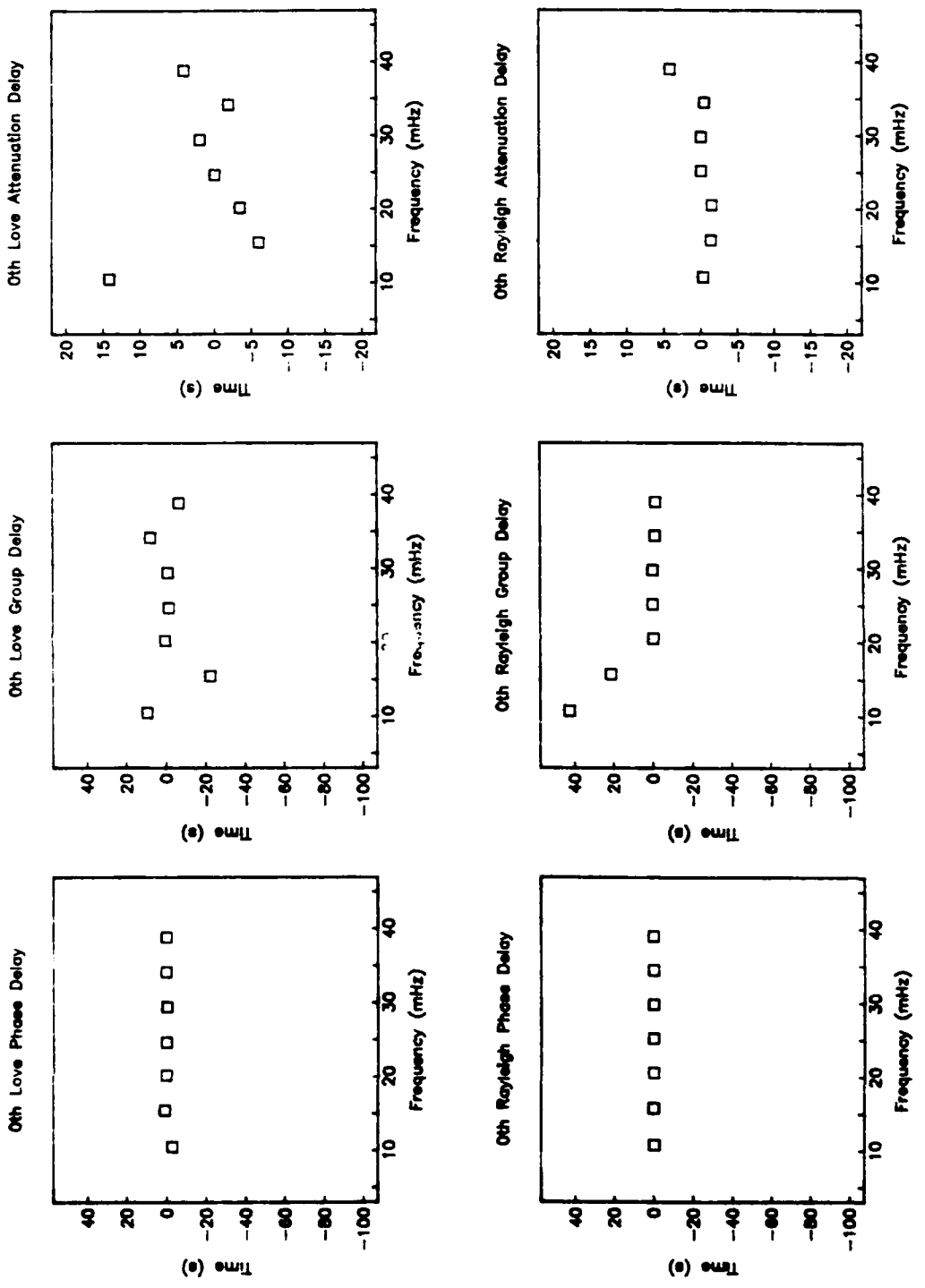

0th Love Attenuation Delay
0th Rayleigh Attenuation Delay
0th Love Group Delay
0th Rayleigh Group Delay
0th Love Phase Delay
0th Rayleigh Phase Delay
Time (s)
Frequency (mHz)

Figure 2.16

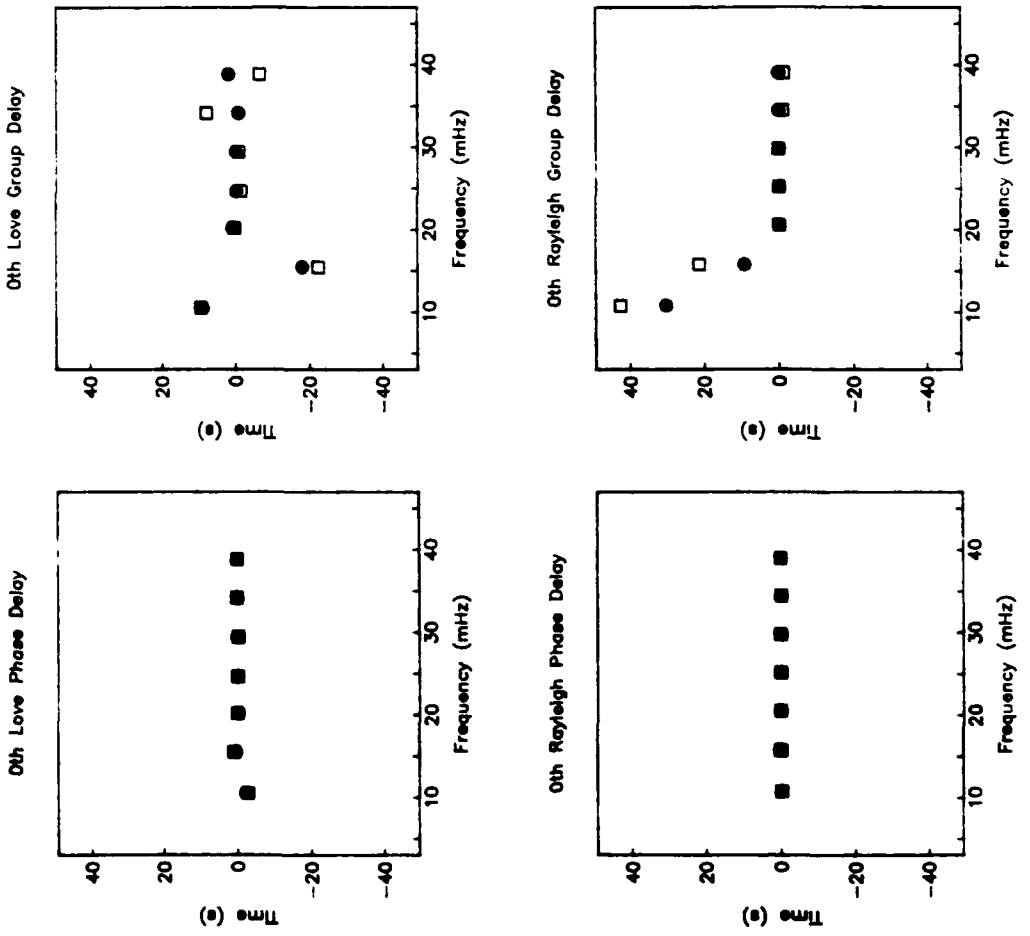

0th Love Group Delay
0th Rayleigh Group Delay
0th Love Phase Delay
0th Rayleigh Phase Delay
Time (s)
Frequency (mHz)
40
20
0
-20
-40
10
20
30
40

Figure 2.17

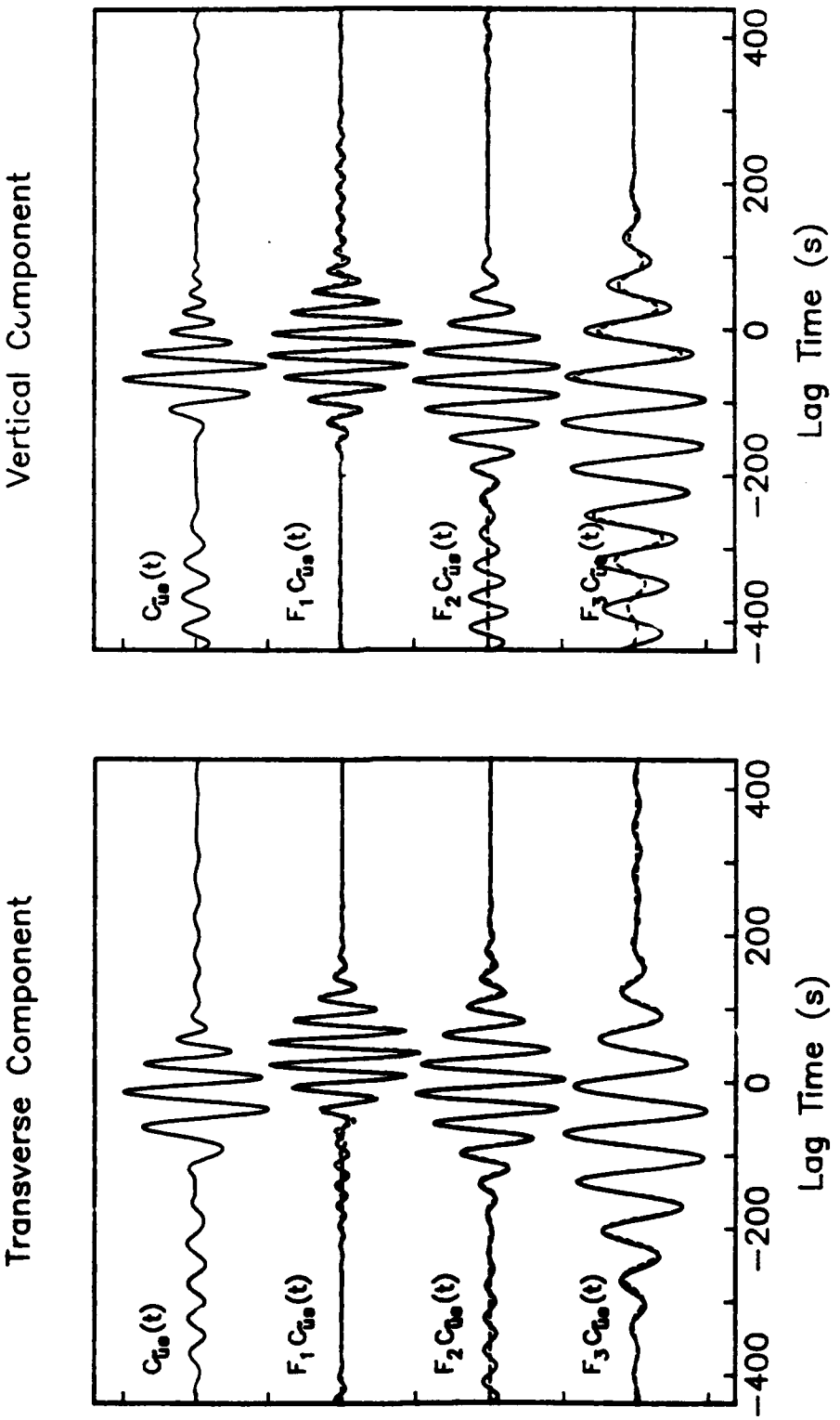

Vertical Component
Transverse Component
$C_{\bar{u}\bar{u}}(t)$
$F_1 C_{\bar{u}\bar{u}}(t)$
$F_2 C_{\bar{u}\bar{u}}(t)$
$F_3 C_{\bar{u}\bar{u}}(t)$
$C_{\bar{u}\bar{u}}(t)$
$F_1 C_{\bar{u}\bar{u}}(t)$
$F_2 C_{\bar{u}\bar{u}}(t)$
$F_3 C_{\bar{u}\bar{u}}(t)$
-400
-200
0
200
400
Lag Time (s)
Lag Time (s)

Figure 2.18

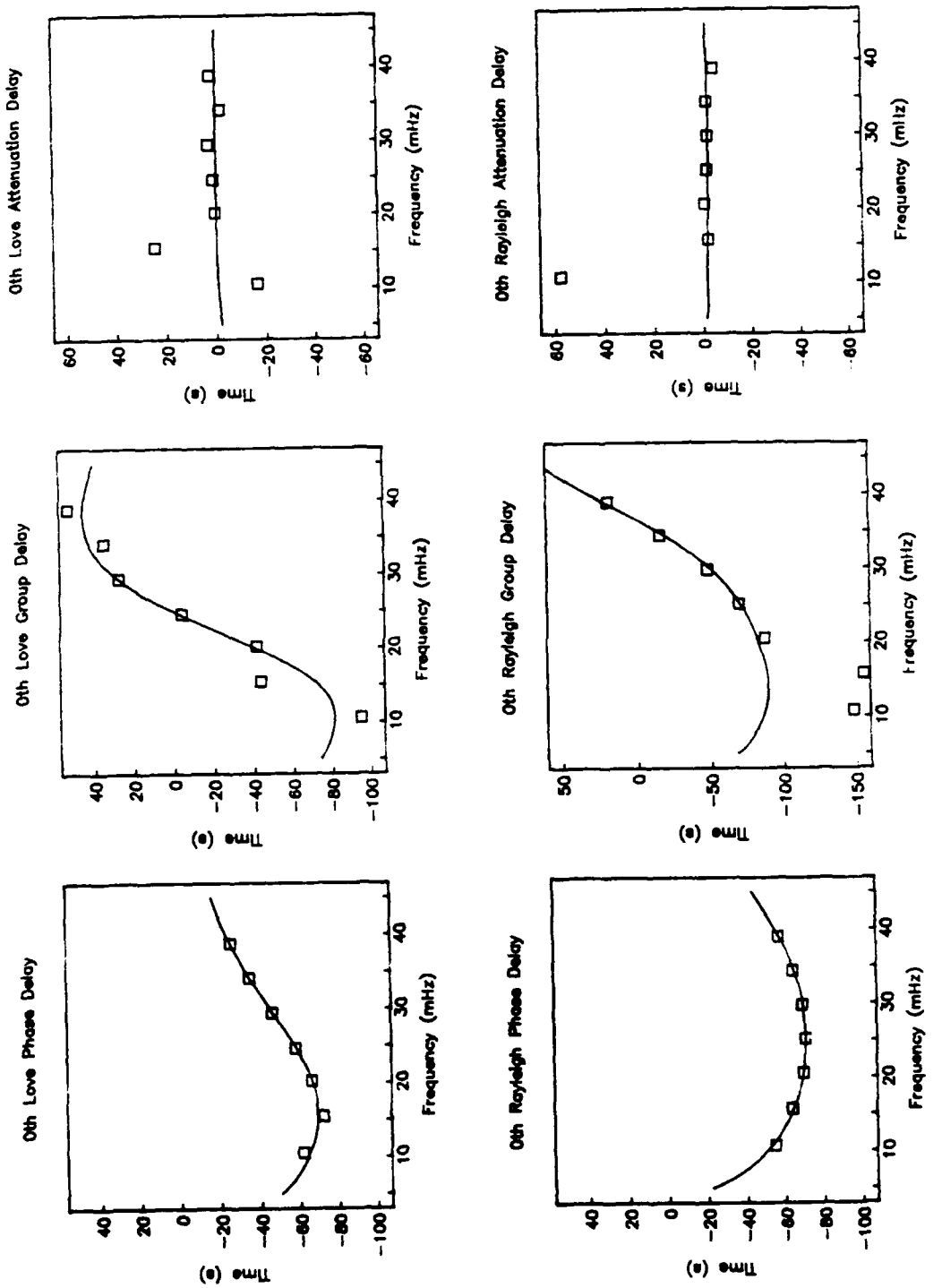

0th Love Attenuation Delay
0th Rayleigh Attenuation Delay
0th Love Group Delay
0th Rayleigh Group Delay
0th Love Phase Delay
0th Rayleigh Phase Delay
Time (s)
Frequency (mHz)

Figure 2.19

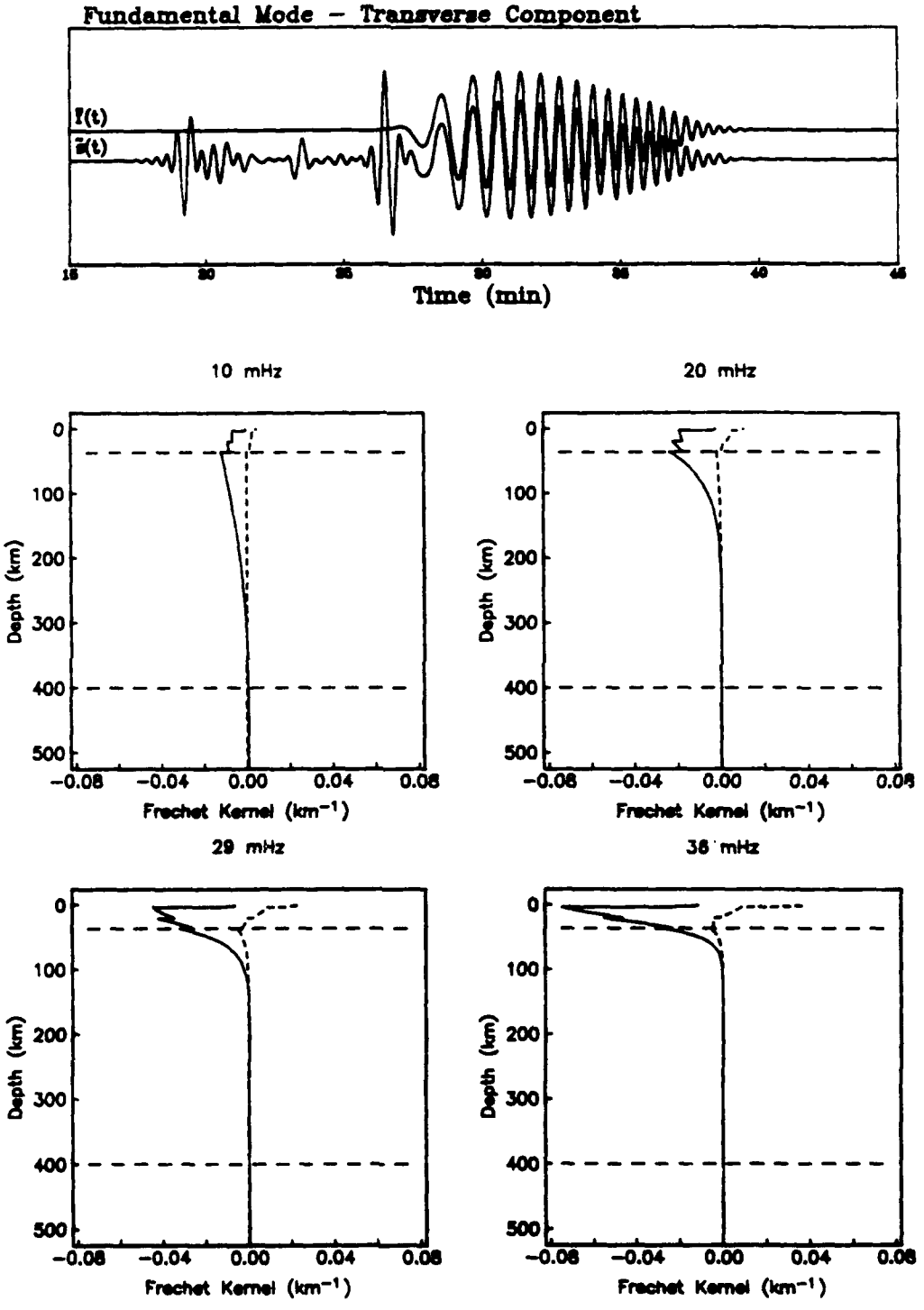

Fundamental Mode – Transverse Component
Time (min)
10 mHz
20 mHz
29 mHz
36 mHz
Depth (km)
Frechet Kernel (km⁻¹)

Figure 2.20a

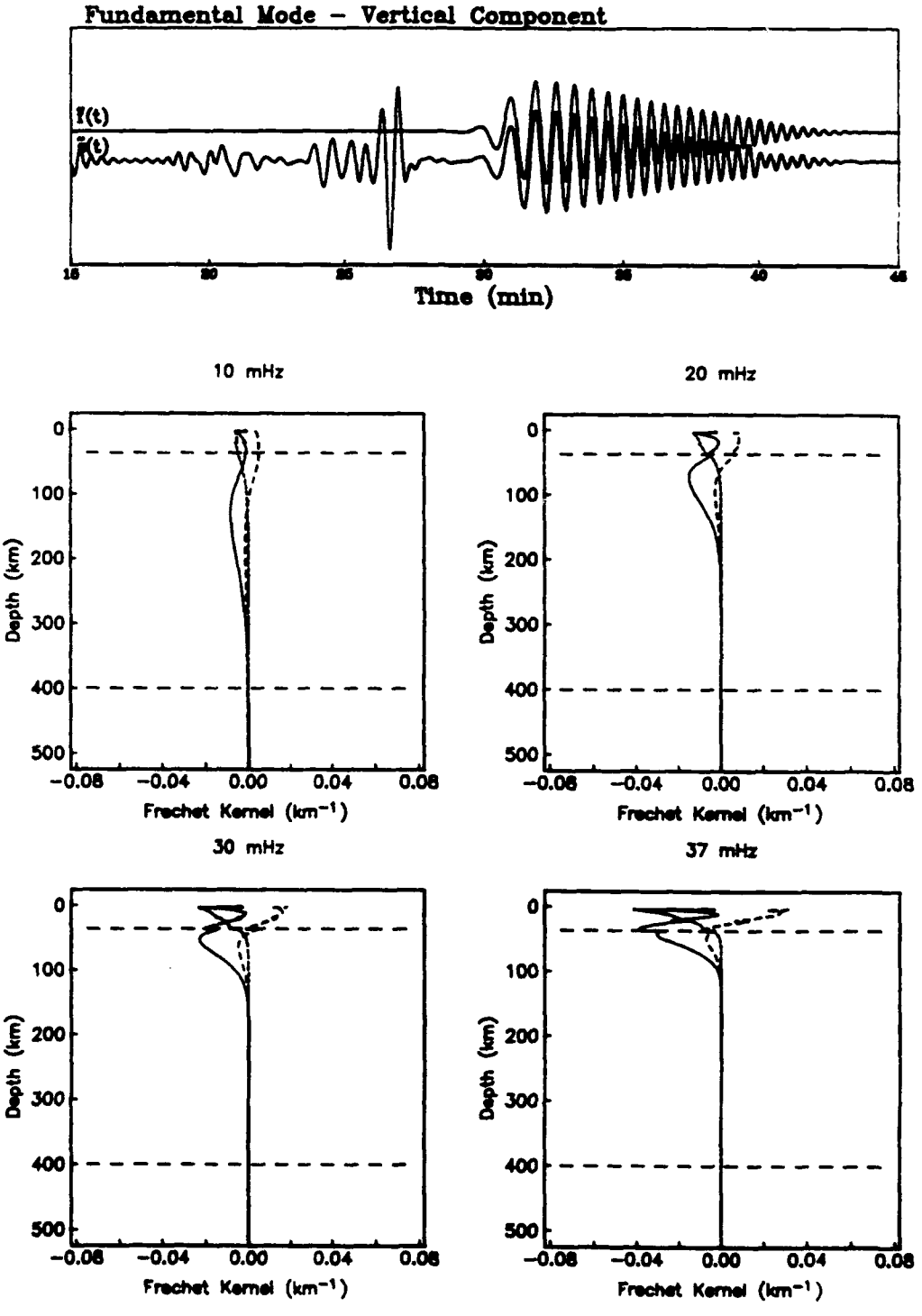

Fundamental Mode – Vertical Component
f(t)
Time (min)
10 mHz
20 mHz
30 mHz
37 mHz
Depth (km)
Frechet Kernel (km⁻¹)

Figure 2.20b

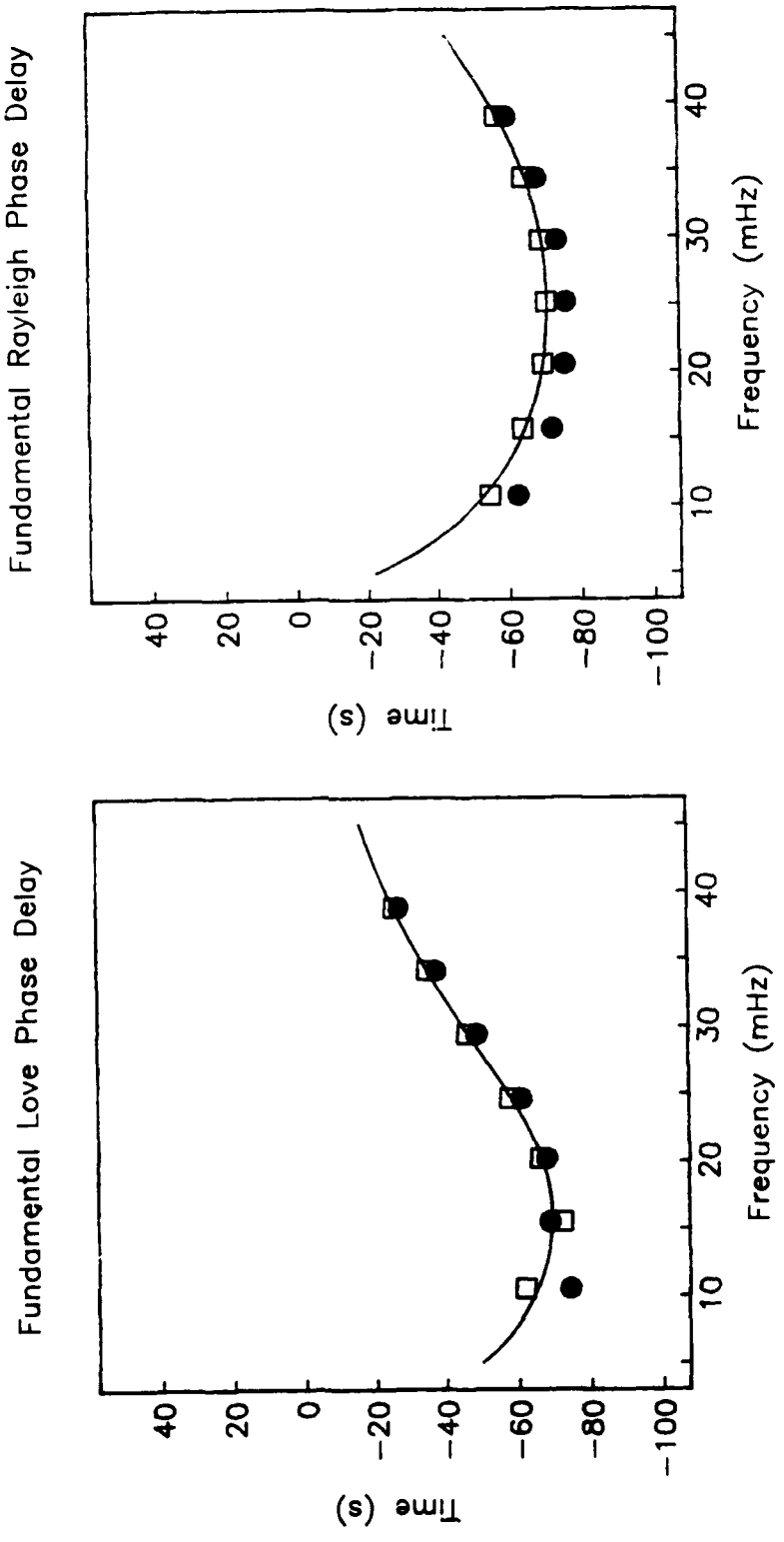

Fundamental Rayleigh Phase Delay
Time (s)
40
20
0
-20
-40
-60
-80
-100
10
20
30
40
Frequency (mHz)
Fundamental Love Phase Delay
Time (s)
40
20
0
-20
-40
-60
-80
-100
10
20
30
40
Frequency (mHz)

Figure 2.21

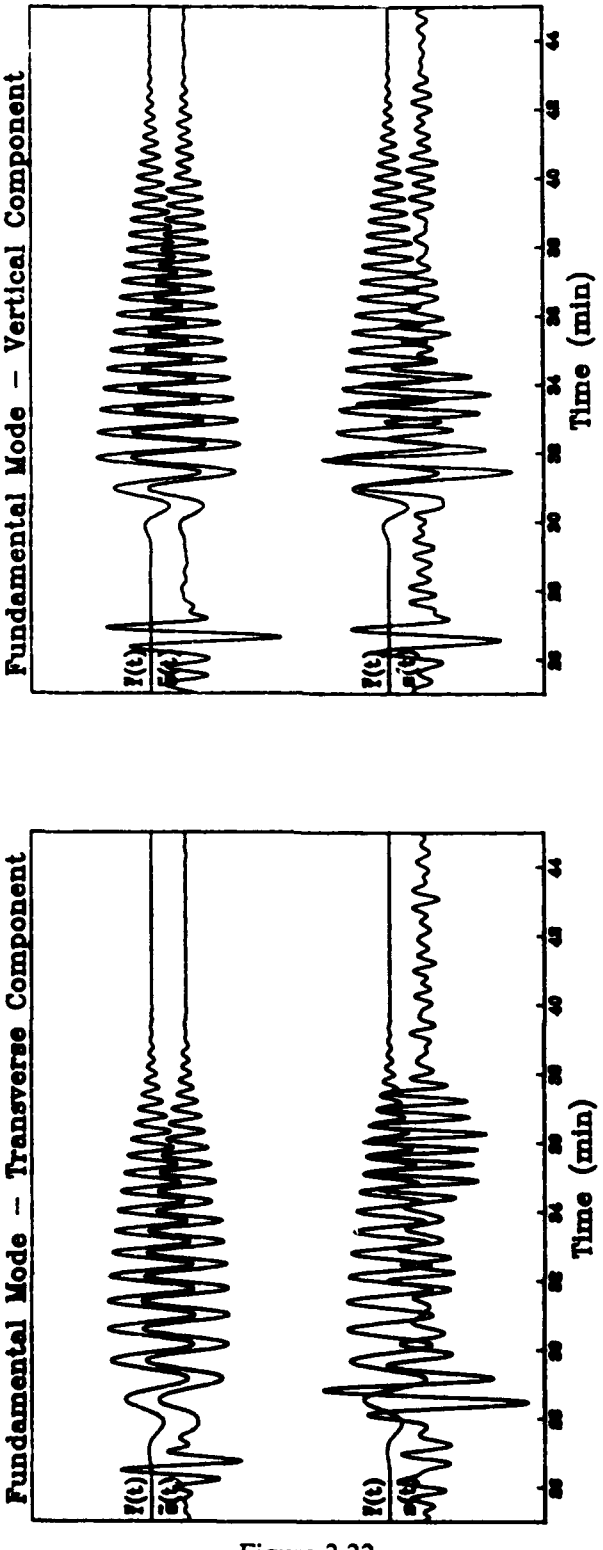

Fundamental Mode – Vertical Component
Time (min)
Fundamental Mode – Transverse Component
Time (min)

Figure 2.22

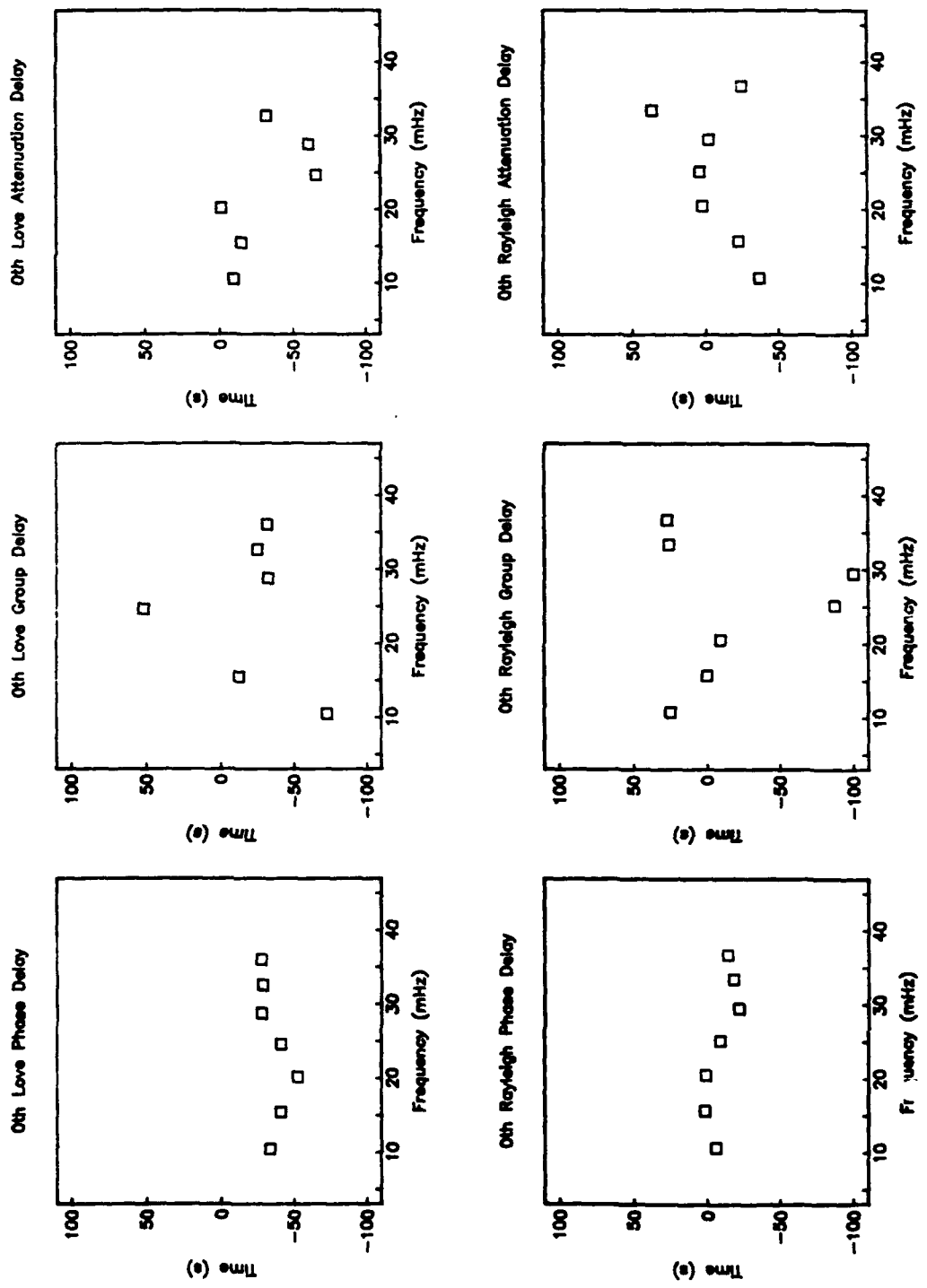

0th Love Attenuation Delay
0th Love Group Delay
0th Love Phase Delay
0th Rayleigh Attenuation Delay
0th Rayleigh Group Delay
0th Rayleigh Phase Delay
Time (s)
Frequency (mHz)
100
50
0
-50
-100
10
20
30
40

Figure 2.23

CHAPTER 3

WAVEFORM ANALYSIS OF BROAD-BAND SEISMOGRAMS

INTRODUCTION

The problem of interference, whether due to wavegroups from the same event or seismic noise, has plagued seismologists for decades. For example, Aki and Kaminuma [1963] and McEvilly [1964] reported observations of unexpectedly high Love-wave phase velocities. Their estimates of fundamental-mode Love and Rayleigh phase velocities could not be explained by a simple radial model of smooth variations in isotropic parameters. McEvilly [1964] and Kaminuma [1966] proposed smooth anisotropic models to resolve their observations, although Aki [1968] and Hales and Bloch [1969] demonstrated that isotropic models containing thin low-velocity zones could satisfy the measured variation. However, Thatcher and Brune [1969] and Boore [1969] postulated that the anomalous Love-wave phase velocities were caused by higher-mode interference, rather than intrinsic anisotropy or laminated Earth structure, and James [1971] provided additional evidence for this hypothesis. Although the body of evidence for some anisotropy in upper mantle is quite compelling [Forsyth, 1975; Schule and Knopoff, 1977; Yu and Mitchell, 1979; Cara *et al.*, 1980; Dziewonski and Anderson, 1981; Kirkwood and Crampin, 1981; Lévêque and Cara, 1983; Nataf *et al.*, 1984; Silver and Chan, 1988; Gee and Jordan, 1988], the significance of the discrepancy between Love and Rayleigh phase velocities is still contested.

A number of approaches to the problem of interference have been investigated over the years. In situations where an array of seismometers is available, this has included a

diverse group of sophisticated techniques for frequency-wavenumber filtering [Capon, 1969; La Cross *et al.*, 1969; Mendiguren, 1973; Gilbert and Dziewonski, 1975; Nolet, 1975, 1977; Cara, 1979; Cara *et al.*, 1980]. When an array of seismometers is not available or stacking is not feasible, the common approach has been the application of a windowing operator. For example, the "time-variable filtration" of Pilant and Knopoff [1964] and the "moving window analysis" of Landisman *et al.* [1969] are based on a window centered at a particular group arrival time on the observed seismogram, whose width varies with the frequency component of interest, and the analysis is repeated for a series of group arrival times. A different approach was taken by Herrin and Goforth [1977, 1986], who combined a cross-correlation methodology with a windowing operator centered on the peak of the correlation function.

In this chapter, we add a step to our data analysis procedure in order to reduce the effect of interfering wavegroups on $C_{\bar{u}\bar{s}}(t)$ and $C_{\bar{u}\bar{s}}(t)$. Following the strategy of Herrin and Goforth [1977, 1986], we apply a window of half-width σ_w^{-1} centered at a reference time t_w first, yielding the windowed correlation functions $WC_{\bar{u}\bar{s}}(t)$ and $WC_{\bar{u}s}(t)$. Windowing the broad-band correlation functions about their peaks prior to narrow-band filtering allows the full bandwidth to be used to separate the waveform of interest from interfering arrivals, improving the approximation $WC_{\bar{u}\bar{s}}(t) \approx WC_{\bar{u}\bar{u}}(t)$ and $WC_{\bar{u}s}(t) \approx WC_{\bar{u}\bar{u}}(t)$. We then apply a narrow-band filter $F_i(\omega)$ in the frequency domain, producing the filtered, windowed correlation functions $F_iWC_{\bar{u}\bar{s}}(t)$ and $F_iWC_{\bar{u}s}(t)$. We derive expressions for the filtered, windowed correlation functions, first for the case of an isolated waveform and then for the complete seismograms, and develop a formulation for the Fréchet kernels of the recovered parameters. Finally, we discuss the implementation of this methodology, continuing with our example of fundamental-mode dispersion.

WINDOWING OPERATOR

Many windowing operators $W(t)$ have been used in the analysis of seismic data, including the boxcar, the Hanning taper, and the Gaussian window. The Hanning taper is particularly popular, as it eliminates the large spectral sidelobes of the boxcar window, while falling to zero within some specified width α:

$$W(t) = \cos^2\alpha(t - t_w) \tag{3.1}$$

where t_w is the center of the window. Rather than develop expressions for a specific window, we represent a general windowing operator as a Gram-Charlier series:

$$W(t) = \frac{1}{2\pi}\,\mathrm{Ga}\,(\sigma_w(t-t_w))\;\left[w_0 - w_2\,(\sigma_w\,(t-t_w))^2 + w_4\,(\sigma_w\,(t-t_w))^4 - \cdots\right] \tag{3.2}$$

where $W(t)$ is real and even and the coefficients w are real, even, and depend on the spectral moments of $W(\omega)$. The window is centered at $t = t_w$, which we generally choose to be the peak of the correlation function, with temporal half-width σ_w^{-1}. As we shall see, the influence of the window on the correlation function depends on the relative size of σ_w and σ_f. If σ_w is small compared to σ_f (corresponding to a large window), the effect of the window on the correlation function will be negligible. If σ_w is large compared to σ_f (corresponding to a narrow window), then the effect of the window will be considerable. We have carried through the Hermite-polynomial expansions for each of the expressions below. In the interest of brevity, we do not detail them here; instead, we present the Gaussian-wavelet approximations. The complete expressions and coefficients are presented in Appendix E; the mathematical details of windowing with Gram-Charlier series are discussed in Appendix B.

ISOLATED WAVEFORM APPROXIMATION

The autocorrelation function

We recall that our expression for the broad-band autocorrelation function was of the form:

$$C_{\tilde{u}\tilde{u}}(t) = \frac{1}{2\pi} \mathrm{Ga}(\sigma_f t) \cos(\omega_f t) \tag{3.3}$$

where ω_f is the center frequency and σ_f is the spectral half-width of $C_{\tilde{u}\tilde{u}}(\omega)$. We multiply $C_{\tilde{u}\tilde{u}}(t)$ by the windowing operator, assuming that the window is centered at the peak of the autocorrelation function ($t_w \equiv 0$):

$$\begin{aligned} WC_{\tilde{u}\tilde{u}}(t) &\equiv W(t)\, C_{\tilde{u}\tilde{u}}(t) \\ &= \frac{1}{2\pi} \mathrm{Ga}(\sigma_g t) \cos(\omega_f t) \end{aligned} \tag{3.4}$$

where σ_g^{-1}, the effective half-width of the Gaussian envelope, depends on σ_f and σ_w:

$$\sigma_g^2 = \sigma_w^2 + \sigma_f^2 \tag{3.5}$$

While the application of windowing operator narrows or limits the correlation function in the time domain by decreasing the effective half-width, the "uncertainty principle" of Fourier analysis [*e.g.*, Bracewell, 1978] results in the broadening or spreading of the signal in the frequency domain.

After windowing, we calculate the Fourier transform of $WC_{\tilde{u}\tilde{u}}(t)$ and apply the Gaussian filter (2.15) in the frequency domain. Assuming that $WC_{\tilde{u}\tilde{u}}(\omega)$ is small near $\omega =$

0, we return to the time domain with an expression for the filtered, windowed autocorrelation function:

$$F_i WC_{\tilde{u}\tilde{u}}(t) \equiv F_i(t) * WC_{\tilde{u}\tilde{u}}(t)$$
$$= E(t)\cos\Phi(t) \tag{3.6}$$

where $E(t)$ and $\Phi(t)$ are defined by:

$$E(t) = \frac{1}{(2\pi)^{5/2}} \frac{\sigma_i'}{\sigma_i \sigma_g} \mathrm{Ga}\left(\frac{\omega_i - \omega_f}{\sqrt{\sigma_i^2+\sigma_g^2}}\right) \mathrm{Ga}(\sigma_i' t) \tag{3.6a}$$

$$\Phi(t) = \omega_i' t \tag{3.6b}$$

and the weighted center frequency ω_i' and spectral half-width σ_i' depend on the window half-width through σ_g:

$$\omega_i' = \frac{\sigma_i^2 \omega_f + \sigma_g^2 \omega_i}{\sigma_i^2 + \sigma_g^2} \tag{3.7}$$

$$\sigma_i'^2 = \frac{\sigma_i^2 \sigma_g^2}{\sigma_i^2 + \sigma_g^2} \tag{3.8}$$

Thus, the windowing operator affects the filtered, windowed autocorrelation function only through the effective center frequency and half-width.

The cross-correlation function

In Chapter 2, we considered both a quadratic and linear approximation to differential wavenumber and showed that the effect of quadratic dispersion could be neglected as long as the signal was sufficiently narrow-band. Here, we begin with the linear approximation and examine the constraints introduced by the windowing operator.

In the case of linear dispersion, $C_{\tilde{u}u}$ depends on the differential phase, group, and amplitude time parameters:

$$C_{\tilde{u}u}(t) = \frac{1}{2\pi} D_0^m \exp[(\sigma_f \delta\tau_a^m)^2/2 + (\omega_i' - \omega_f)\delta\tau_a^m] \; \mathrm{Ga}\,(\sigma_f(t-\delta\tau_g^m)) \times \cos[\omega_i'(t-\delta\tau_p^m) - (\omega_i' - \omega_f + \sigma_f^2\delta\tau_a^m)(t-\delta\tau_g^m)] \tag{3.9}$$

where we have expanded the differential wavenumber about $\omega = \omega_i'$ and $\delta\tau_p^m \equiv \delta\tau_p^m(\omega_i')$, $\delta\tau_g^m \equiv \delta\tau_g^m(\omega_i')$, and $\delta\tau_a^m \equiv \delta\tau_a^m(\omega_i')$. This expression is more complicated than the equation for $F_i C_{\tilde{u}u}(t)$ in Chapter 2, as the center frequency of the broad-band correlation function is, in general, different from that of the applied filter, and (3.9) contains terms that account for the frequency shift. $C_{\tilde{u}u}(t)$ reaches its maximum, δt_c, when:

$$\delta t_c = \frac{\omega_i' (\omega_f - \sigma_f^2\delta\tau_a^m)(\delta\tau_p^m + 2\pi n/\omega_i' - \delta\tau_g^m) + (\omega_f^2 - 2\,\omega_f\sigma_f^2\delta\tau_a^m + \sigma_f^2)\,\delta\tau_g^m}{\omega_f^2 - 2\,\omega_f\sigma_f^2\delta\tau_a^m + \sigma_f^2} \tag{3.10}$$

and the $2\pi n/\omega_i'$ accounts for the cycle skipping in the peak of the cross-correlation function. We apply the windowing operator in the time domain, centering the window on the peak of the cross-correlation function ($t_w \equiv \delta t_c$) :

$$WC_{\tilde{u}u}(t) = \frac{1}{(2\pi)^2} D_0^m \exp[(\sigma_f \delta\tau_a^m)^2/2 + (\omega_i' - \omega_f)\delta\tau_a^m] \; \mathrm{Ga}\left(\frac{\sigma_f \sigma_w}{\sigma_g}(\delta t_c - \delta\tau_g^m)\right)$$
$$\times \mathrm{Ga}(\sigma_g(t-\delta\tau_g'^m)) \cos[\omega_i'(t-\delta\tau_p^m) - (\omega_i' - \omega_f + \sigma_f^2 \delta\tau_a^m)(t-\delta\tau_g'^m)] \quad (3.11)$$

Similar to the expression for the windowed autocorrelation, the Gaussian envelope of $WC_{\tilde{u}u}(t)$ is characterized by a new half-width, σ_g^{-1}, which depends on the spectral half-width of the broad-band autocorrelation function and the window. However, the envelope is now centered at an effective group delay, $\delta\tau_g'^m$, which is the weighted average of the window center time and the differential group delay at ω_i':

$$\delta\tau_g'^m = \frac{\sigma_w^2 \delta t_c + \sigma_f^2 \delta\tau_g^m}{\sigma_w^2 + \sigma_f^2} \quad (3.12)$$

We calculate the Fourier transform of the windowed function, apply the narrow-band filter, and return to the time domain. The filtered, windowed cross-correlation has the form of a Gaussian window multiplied by a harmonic carrier function:

$$F_i WC_{\tilde{u}u}(t) = E(t) \cos\Phi(t) \quad (3.13)$$

where $E(t)$ and $\Phi(t)$ are defined by

$$E(t) = \frac{E}{(2\pi)^{5/2}} \frac{\sigma_i'}{\sigma_i \sigma_g} \mathrm{Ga}(\sigma_i'(t-\delta\tau_g'^m)) \quad (3.13a)$$

$$\Phi(t) = \omega_i'(t-\delta\tau_p'^m) - \sigma_i'^2 \delta\tau_a'^m (t-\delta\tau_g'^m) \quad (3.13b)$$

and we have introduced E as notation for the amplitude of the envelope function:

$$E = D_0^m \exp[(\sigma_f \delta\tau_a^m)^2/2 + (\omega_i' - \omega_f)\delta\tau_a^m] \, \mathrm{Ga}\left(\frac{\sigma_f \sigma_w}{\sigma_g}(\delta t_c - \delta\tau_g^m)\right) \, \mathrm{Ga}\left(\frac{\omega_i - \omega_f + \sigma_f^2 \delta\tau_a^m}{\sqrt{\sigma_i^2 + \sigma_g^2}}\right) \quad (3.13c)$$

$F_i WC_{\tilde{u}\tilde{u}}(t)$ depends on the effective phase, group, and amplitude parameters ($\delta\tau_p^{m'}$, $\delta\tau_g^{m'}$, $\delta\tau_a^{m'}$):

$$\begin{aligned}\delta\tau_p'^m &= \delta\tau_p^m + \frac{(\omega_i' - (\omega_f - \sigma_f^2 \delta\tau_a^m))}{\omega_i'} \frac{\sigma_w^2}{\sigma_w^2 + \sigma_f^2} (\delta\tau_c - \delta\tau_g^m) \\ &= \delta\tau_p^m + \frac{(\omega_i' - (\omega_f - \sigma_f^2 \delta\tau_a^m))(\omega_f - \sigma_f^2 \delta\tau_a^m)}{\omega_f^2 - 2\omega_f \sigma_f^2 \delta\tau_a^m + \sigma_f^2} \frac{\sigma_w^2}{\sigma_w^2 + \sigma_f^2} (\delta\tau_p^m + 2\pi n/\omega_i' - \delta\tau_g^m) \quad (3.14a)\end{aligned}$$

$$\begin{aligned}\delta\tau_g'^m &= \frac{\sigma_w^2 \delta t_c + \sigma_f^2 \delta\tau_g^m}{\sigma_w^2 + \sigma_f^2} \\ &= \delta\tau_g^m + \frac{\omega_i' (\omega_f - \sigma_f^2 \delta\tau_a^m)}{\omega_f^2 - 2\omega_f \sigma_f^2 \delta\tau_a^m + \sigma_f^2} \frac{\sigma_w^2}{\sigma_w^2 + \sigma_f^2} (\delta\tau_p^m + 2\pi n/\omega_i' - \delta\tau_g^m) \quad (3.14b)\end{aligned}$$

$$\delta\tau_a'^m = \delta\tau_a^m \left[1 + \frac{\sigma_w^2}{\sigma_w^2 + \sigma_f^2}\right] \quad (3.14c)$$

which represent the actual generalized data functionals with correction terms which accounts for the effect of the windowing operator. Common to all three expressions is the bandwidth parameter $[\sigma_w^2/(\sigma_w^2 + \sigma_f^2)]$, which measures the narrowness the window relative to the signal width of the correlation function. $\delta\tau_p^{m'}$ and $\delta\tau_g^{m'}$ contain two additional parameters. The dispersion parameter $[\delta\tau_p^m + 2\pi n/\omega_i' - \delta\tau_g^m]$ measures the asymmetry of the correlation function. The frequency parameter $[(\omega_i' - (\omega_f - \sigma_f^2 \delta\tau_a^m))(\omega_f -$

$\sigma_f^2 \delta\tau_a^m)/(\omega_f^2 - 2\omega_f\sigma_f^2\delta\tau_a^m + \sigma_f^2)]$ in the case of $\delta\tau_p^{m\prime}$ and $[\omega_i' (\omega_f - \sigma_f^2 \delta\tau_a^m)/(\omega_f^2 - 2\omega_f\sigma_f^2\delta\tau_a^m + \sigma_f^2)]$ in the case of $\delta\tau_g^{m\prime}$ are a measure of how far away the filter is from the center of the band.

Wide-window approximation. The critical parameters in determining the size of the correction term are the measures of bandwidth and dispersion. In general, the frequency parameter will be of order 0.5 or so; in the particular example of the fundamental-mode Rayleigh wave, a filter centered at 10 mHz implies a frequency parameter of 0.3. If the cross-correlation function is symmetric, either because we are considering an undispersed arrival or because the differential group delay is fortuitously equal to the differential phase delay plus an integral number of cycles, then the correction term will be zero. If the differential dispersion is large, the cross-correlation function will be asymmetrical, and the correction term depends on the product of the bandwidth and dispersion parameters. Thus, if the differential dispersion is not large, we may apply a narrow windowing operator. However, if the differential dispersion is large, then we must increase the window width in order to neglect the correction term. Because the $2\pi n/\omega_i'$ term ensures that the difference between $\delta\tau_p^m$ and $\delta\tau_g^m$ will never be greater than one cycle, the dispersion parameter at 10 mHz would be 100 s at worst. In the actual case of EU2-SNA, the differential-dispersion parameter is 40 s at 10 mHz. We typically use a window of fixed width (which is a function of ω_f) and the resulting bandwidth parameter is of order 0.005. Consequently, numerical analysis suggests that we may neglect the correction terms:

$$\delta\tau_p^{'m} \approx \delta\tau_p^m \tag{3.15a}$$

$$\delta\tau_g^{'m} \approx \delta\tau_g^m \tag{3.15b}$$

$$\delta\tau_a^{'m} \approx \delta\tau_a^m \tag{3.15c}$$

which we shall refer to as the wide-window approximation.

Comparing equation (3.13) with (2.30), we see that the addition of the windowing operator does not modify the cross-correlation function considerably, as long we make a judicious choice of window parameters. In particular, the choice of window width depends on the amount of dispersion. If the wavegroup is not very dispersed, then we may apply a narrow window; however, large dispersion requires that we open the window in order to make the wide-window approximation. If the wide-window approximation does not apply, that is if the product $[\sigma_w^2/(\sigma_w^2 + \sigma_f^2)]\,[\delta\tau_p^m + 2\pi n/\omega_i' - \delta\tau_g^m]$ is large, the estimates of differential phase, group, and amplitude will contain correction terms.

Example

In order to explore the influence of windowing on the measurement procedure, we have applied this methodology to the synthetic example of an isolated fundamental-mode. First, we compute the autocorrelation of $\tilde{u}_m(t)$ and determine ω_f and σ_f by fitting equation (3.3) to $C_{\tilde{u}\tilde{u}}$. In the second step, we apply the windowing operator, typically a Hanning taper, transform the windowed autocorrelation function into the frequency domain, and apply the narrow-band filter. We fit (3.6) to $F_i W C_{\tilde{u}\tilde{u}}(t)$ in order to estimate ω_i' and σ_i' (Figure 3.1). In the third step, we compute the cross-correlation between $\tilde{u}_m(t)$ and $u_m(t)$, apply the window centered at the peak of $C_{\tilde{u}u}(t)$, transform the windowed autocorrelation function into the frequency domain and apply the narrow-band filter. With ω_i' and σ_i' fixed, we fit (3.13) to $F_i W C_{\tilde{u}u}(t)$ and estimate $\delta\tau_p^m(\omega_i')$, $\delta\tau_g^m(\omega_i')$, and $\delta\tau_a^m(\omega_i')$ (Figure 3.2). Once the windowed correlation functions are calculated, a series of narrow-band filters may be applied. Finally, the estimates of $\delta\tau_p^m$ are corrected for cycle skipping.

Figure 3.3 compares the estimated values of $\delta\tau_p^m$, $\delta\tau_g^m$, and $\delta\tau_a^m$ retrieved with this technique (squares) with the actual values in the EU2-SNA isolated fundamental-mode example (solid line). The estimates of differential phase delay recovered with this technique are quite good. However, the differential group delay measurements display

systematic errors and the amplitude measurements are scattered. These results are disappointing, especially when compared to the isolated waveform results from Chapter 2 (crosses).

We attribute this misfit to the breakdown of the linear-dispersion approximation in (3.9). In the technique described in Chapter 2, we apply the narrow-band Gaussian filter directly to the broad-band correlation function. This allows us to use the linear-dispersion approximation as a local approximation at ω_i'. In this approach, the window is applied to the correlation function before filtering, introducing the requirement that the differential dispersion approximation to be valid from ω_f to ω_i', rather than in a small neighborhood about ω_i'. In the extreme example of EU2 and SNA, the linear-dispersion approximation is not valid over a large frequency range (recall Figure 2.5, which compares the linear and quadratic-dispersion approximations of the fundamental-mode differential wavenumber) and the estimates of $\delta\tau_p^n$, $\delta\tau_g^n$, and $\delta\tau_a^n$ are most accurate near $\omega = \omega_f$ (ω_f for the Love wave is 23 mHz; 27 mHz for the Rayleigh wave). Thus, the application of a windowing operator limits the approximation of differential wavenumber.

Quadratic dispersion

Returning to the expansion of differential wavenumber (2.24), we derive an expression for $F_i WC_{\tilde{u}u}(t)$ in the case of quadratic dispersion. This requires carrying the quadratic term through the operations of windowing and filtering. Because the quadratic term is complex-valued, the resulting expression contains frequency and bandwidth parameters which are complex-valued. We write this formula as a Gaussian envelope times a harmonic function, where we take the real value of their product:

$$F_i WC_{\tilde{u}u}(t) = \mathrm{Re}\,\{E(t)\exp[-i\Phi(t)]\} \tag{3.16}$$

where $E(t)$ and $\Phi(t)$ are defined by:

$$E(t) = E\ \mathrm{Ga}\left(\sigma_z(t-\delta\tau_g^m)\right) \tag{3.16a}$$

$$\Phi(t) = \omega_i'(t-\delta\tau_p^m) - (\omega_i' - (\omega_z - \sigma_z^2\delta\tau_a^m))(t-\delta\tau_g^m) \tag{3.16b}$$

and we have made the wide-window approximation. Since we are primarily concerned with the time-dependent properties of (3.16), we defer discussion of the envelope coefficient to Appendix D. The constants ω_z and σ_z are complex:

$$\omega_z = \frac{\sigma_i^2\omega_x + (\sigma_w^2 + \sigma_x^2)\omega_i}{\sigma_i^2 + \sigma_w^2 + \sigma_x^2} \tag{3.17}$$

$$\sigma_z^2 = \frac{\sigma_i^2(\sigma_w^2 + \sigma_x^2)}{\sigma_i^2 + \sigma_w^2 + \sigma_x^2} \tag{3.18}$$

The influence of quadratic dispersion appears through the parameters ω_x and σ_x, which depend on the differential-quadratic contribution:

$$\begin{aligned}\omega_x &= \frac{\sigma_s^{m2}\omega_f + \sigma_f^2\omega_i'}{\sigma_s^{m2} + \sigma_f^2} \\ &= \frac{\omega_f - i\gamma_d^2\omega_i'}{1 - i\gamma_d^2}\end{aligned} \tag{3.19}$$

$$\begin{aligned}\sigma_x^2 &= \frac{\sigma_s^{m2}\sigma_f^2}{\sigma_s^{m2} + \sigma_f^2} \\ &= \frac{\sigma_f^2}{1 - i\gamma_d^2}\end{aligned} \tag{3.20}$$

AD-A-243 935

PAGE
132
missing
FROM ORIGINAL
DOCUMENT
AS SENT FROM
THE ORIGINATOR

displays the results (circles) obtained with this technique. We have nearly eliminated the bias in differential group delay due to quadratic dispersion.

Interpretation

The addition of the windowing operator does not affect our interpretation of the generalized data functionals recovered by this technique in the case of an isolated waveform. In particular, the estimate of differential phase delay is linearly related to perturbations in the model parameters, as discussed in Chapter 2 (equation (2.42)). Figure 3.5*a* and *b* display the Fréchet kernels for $\delta\tau_p^m$ at four center frequencies. Figure 3.6 compares the estimates of $\delta\tau_p^m$ obtained by waveform fitting (open circles) with those predicted by integration of the kernels (closed circles).

Summary of the isolated waveform example

We introduced a windowing operator into our analysis procedure to reduce the effect of interference on the estimation of the generalized data functionals. By windowing the broad-band correlation functions before filtering, we allow the full bandwidth of the signal to separate the waveform of interest. We saw that the windowing operator changes the differential-dispersion approximation from a local approximation about ω_i' to a global approximation from ω_f to ω_i'. In the example of EU2-SNA, we found that the linear-dispersion approximation was not valid at the edges of the band. We demonstrated a bootstrapping technique which allows us to estimate the quadratic contribution at each point along the dispersion curve and thus obtain unbiased estimates of the differential phase, group, and amplitude delays. In general, we suspect that quadratic dispersion will not be a significant problem because our example of the recovery of SNA dispersion from EU2 is an extreme application of this methodology. In general, we expect that our reference model will be "closer" to the actual Earth. In the following section, we shall examine the effectiveness of the windowing operator in reducing interference.

AD-A-243935

PAGE
134

MISSING
FROM ORIGINAL
DOCUMENT
AS SENT TO
DTIC
FROM THE
ORIGINATOR

$$\tilde{C}_{nm} = \tilde{A}_{nm} \exp[(\sigma_f \Delta\tilde{\tau}_a^{nm})^2/2 + (\omega_i' - \omega_f)\Delta\tilde{\tau}_a^{nm}] \; \mathrm{Ga}\left(\frac{\sigma_f \sigma_w}{\sigma_g}(t_c - \Delta\tilde{\tau}_g^{nm})\right) \mathrm{Ga}\left(\frac{\omega_i - \omega_f + \sigma_f^2 \Delta\tilde{\tau}_a^{nm}}{\sqrt{\sigma_i^2 + \sigma_g^2}}\right) \tag{3.22a}$$

and the effective differential time parameters contain correction terms for the window:

$$\Delta\tilde{\tau}_p^{'nm} = \Delta\tilde{\tau}_p^{nm} + \frac{(\omega_i' - (\omega_f - \sigma_f^2 \Delta\tilde{\tau}_a^{nm}))}{\omega_i'} \frac{\sigma_w^2}{\sigma_w^2 + \sigma_f^2} (t_c - \Delta\tilde{\tau}_g^{nm}) \tag{3.23a}$$

$$\Delta\tilde{\tau}_g^{'nm} = \frac{\sigma_w^2 t_c + \sigma_f^2 \Delta\tilde{\tau}_g^{nm}}{\sigma_w^2 + \sigma_f^2} \tag{3.23b}$$

$$\Delta\tilde{\tau}_a^{'nm} = \Delta\tilde{\tau}_a^{nm}[1 + \frac{\sigma_w^2}{\sigma_w^2 + \sigma_f^2}] \tag{3.23c}$$

Equation (3.22) describes the filtered, windowed cross-correlation between the isolation filter and the complete synthetic seismogram as a sum over Gaussian wavelets; each one describing the interaction between two traveling-wave branches. If we make the wide-window approximation, then the correction terms can be neglected, and the primary effect of window is the addition of the $\mathrm{Ga}[(\sigma_w \sigma_f/\sigma_g)(t_c - \Delta\tilde{\tau}_g^{nm})]$ coefficients in the $\tilde{C}_{nm}$, which penalize any cross-terms which have differential group arrivals outside the window.

Single-wavelet approximation. Since the windowing operator will screen out most interfering arrivals, we may more accurately model the sum as a single Gaussian wavelet, characterized by an average perturbation of amplitude as well as phase, group, and amplitude delays:

$$F_i W C_{\tilde{u}\tilde{s}}(t) = E(t) \cos\Phi(t) \tag{3.24}$$

AD-A-243935

PAGE
#136

MISSING
FROM ORIGINAL
DOCUMENT
AS SENT TO
DTIC
FROM THE
ORIGINATOR

seismogram and the windowing operator. In the case of the wide-window approximation, the correction terms in (3.25a-c) can be neglected.

We can develop expressions for these average perturbations in terms of the individual contributions of the traveling-wave branches. We illustrate the approach in the simple case where there is no differential attenuation; extending the methodology to include differential attenuation is straightforward. We compare the Gaussian-wavelet approximation (3.24) to the traveling-wave sum (3.22):

$$\tilde{C}\ \mathrm{Ga}\left(\frac{\sigma_f \sigma_w}{\sigma_g}(t_c-\Delta\tilde{\tau}_g)\right)\ \mathrm{Ga}\left(\sigma_i'(t-\Delta\tilde{\tau}_g)\right)\ \cos[\omega_i'(t-\Delta\tilde{\tau}_p)]$$

$$= \sum_n \tilde{A}_{nm}\,\mathrm{Ga}\left(\frac{\sigma_f \sigma_w}{\sigma_g}(t_c-\Delta\tilde{\tau}_g^{nm})\right)\ \mathrm{Ga}\left(\sigma_i'(t-\Delta\tilde{\tau}_g^{nm})\right)\ \cos[\omega_i'(t-\Delta\tilde{\tau}_p^{nm})] \qquad (3.26)$$

where we have made the wide-window approximation. If $\sigma_i' << \omega_i'$, then the Gaussian envelopes will be slowly varying, compared to the carrier functions, near their peaks. We expand the envelopes about $t = \Delta\tilde{\tau}_g$ and compare expressions on a term-by-term basis. From the zeroth-order term, we find:

$$\tilde{C}^2 \approx \left[\sum_n \tilde{D}_{nm}\cos(\omega_i'\Delta\tilde{\tau}_p^{nm})\right]^2 + \left[\sum_n \tilde{D}_{nm}\sin(\omega_i'\Delta\tilde{\tau}_p^{nm})\right]^2 \qquad (3.27)$$

$$\Delta\tilde{\tau}_p(\omega_i') \approx \frac{1}{\omega_i'}\tan^{-1}\left(\frac{\sum_n \tilde{D}_{nm}\sin(\omega_i'\Delta\tilde{\tau}_p^{nm})}{\sum_n \tilde{D}_{nm}\cos(\omega_i'\Delta\tilde{\tau}_p^{nm})}\right) \qquad (3.28)$$

From the first-order term, we recover

$$\Delta\tilde{\tau}_g(\omega_i') \approx \frac{c_0 c_1 + s_0 s_1}{c_0^2 + s_0^2} \qquad (3.29)$$

AD-A-243935

PAGE
#138

MISSING
FROM ORIGINAL
DOCUMENT
AS SENT TO
DTIC
FROM THE
ORIGINATOR

$$F_i WC_{\tilde{u}\,s}(t) = \frac{1}{(2\pi)^{5/2}} \frac{\sigma_i'}{\sigma_i \sigma_g} \sum_n C_{nm}\ \mathrm{Ga}\left(\sigma_i'(t-\delta\tau_g'^{n}-\Delta\tilde{\tau}_g'^{nm})\right)$$

$$\times \cos\left[\omega_i'(t-\delta\tau_p'^{n}-\Delta\tilde{\tau}_p'^{nm}) - \sigma_i'^{2}\,(\delta\tau_a'^{n}+\Delta\tilde{\tau}_a'^{nm})(t-\delta\tau_g'^{n}-\Delta\tilde{\tau}_g'^{nm})\right] \quad (3.32)$$

where the C_{nm} are defined:

$$C_{nm} = \tilde{A}_{nm} \exp\left[(\sigma_f(\delta\tau_a^n+\Delta\tilde{\tau}_a^{nm}))^2/2 + (\omega_i'-\omega_f)(\delta\tau_a^n+\Delta\tilde{\tau}_a^{nm})\right]$$

$$\times \mathrm{Ga}\left(\frac{\sigma_f\sigma_w}{\sigma_g}((\delta t_c-\delta\tau_g^n)+(t_c-\Delta\tilde{\tau}_g^{nm}))\right) \mathrm{Ga}\left(\frac{\omega_i-\omega_f+\sigma_f^2(\delta\tau_a^n+\Delta\tilde{\tau}_a^{nm})}{\sqrt{\sigma_i^2+\sigma_g^2}}\right) \quad (3.32a)$$

and the effective differential time parameters contain correction terms for the windowing operator:

$$\delta\tau_p'^{n} = \delta\tau_p^n + \frac{(\omega_i'-(\omega_f-\sigma_f^2\delta\tau_a'^{n}))}{\omega_i'} \frac{\sigma_w^2}{\sigma_w^2+\sigma_f^2} (\delta t_c-\delta\tau_g^n) \quad (3.33a)$$

$$\delta\tau_g'^{n} = \frac{\sigma_w^2\,\delta t_c + \sigma_f^2\,\delta\tau_g^n}{\sigma_w^2+\sigma_f^2} \quad (3.33b)$$

$$\delta\tau_a'^{n} = \delta\tau_a^n\left[1 + \frac{\sigma_w^2}{\sigma_w^2+\sigma_f^2}\right] \quad (3.33c)$$

In the case of the wide-window approximation, the correction terms can be neglected.

Single-wavelet approximation. Although (3.32) represents the filtered, windowed cross-correlation between $\tilde{u}_m(t)$ and $s(t)$ as a sum over many Gaussian wavelets, only a few will

AD-A-243935

PAGE
#140

MISSING
FROM ORIGINAL
DOCUMENT
AS SENT TO
DTIC
FROM THE
ORIGINATOR

$$\delta\tau_g' = \frac{\sigma_w^2\,\delta t_c + \sigma_f^2\,\delta\tau_g}{\sigma_w^2 + \sigma_f^2} \tag{3.35b}$$

$$\delta\tau_a' = \delta\tau_a\,[1 + \frac{\sigma_w^2}{\sigma_w^2 + \sigma_f^2}] \tag{3.35c}$$

As before, the correction terms can be neglected in the wide-window approximation.

Fréchet kernels. We can develop expressions for the generalized data functionals, $\delta\tau_p$, $\delta\tau_g$, and $\delta\tau_a$, in terms of the individual contributions of the traveling-wave branches, which will allow us to formulate expressions for the Fréchet kernels as linear sums of the branch parameters. To demonstrate the approach, we shall derive a formula for the Fréchet kernel for $\delta\tau_p(\omega_i')$ in the case of no differential attenuation. Expanding the theory to derive expressions for the group and attenuation kernels is a direct extension of this methodology. We compare the expression of the averaged Gaussian wavelet (3.34) with the traveling-wave sum (3.32):

$$\begin{aligned} C\ \mathrm{Ga}\left(\frac{\sigma_f\sigma_w}{\sigma_g}((\delta t_c-\delta\tau_g)+(t_c-\Delta\tilde{\tau}_g))\right)\ \mathrm{Ga}\,(\sigma_i'(t-\delta\tau_g-\Delta\tilde{\tau}_g))\cos[\omega_i'(t-\delta\tau_p-\Delta\tilde{\tau}_p)] \\ = \sum_n \tilde{A}_{nm}\ \mathrm{Ga}\left(\frac{\sigma_f\sigma_w}{\sigma_g}((\delta t_c-\delta\tau_g^n)+(t_c-\Delta\tilde{\tau}_g^{nm}))\right) \\ \mathrm{Ga}\,(\sigma_i'(t-\delta\tau_g^n-\Delta\tilde{\tau}_g^{nm}))\ \cos[\omega_i'(t-\delta\tau_p^n-\Delta\tilde{\tau}_p^{nm})] \end{aligned} \tag{3.36}$$

where we have made the wide-window approximation. If $\sigma_i' << \omega_i'$, then the Gaussian envelopes will be slowly varying compared to the carrier function near their peaks. We expand the envelopes about $t = \Delta\tilde{\tau}_g + \delta\tau_g$ and neglect the amplitude perturbations, as they will be second-order compared to the phase perturbation. From consideration of the first-order term, we find:

AD-A-243935

PAGE
#142

MISSING
FROM ORIGINAL
DOCUMENT
AS SENT TO
DTIC
FROM THE
ORIGINATOR

correlation function. The total length of the window is generally chosen to be five times the dominant period of $C_{\tilde{u}\tilde{u}}$, which is sufficiently broad that we may assume the wide-window approximation. This window width is still adequate to eliminate most of the interfering arrivals, such as the first-higher Love mode. We transform the windowed correlation functions into the frequency domain and apply the narrow-band filter. Returning to the time domain, we fit (3.6) to $F_i WC_{\tilde{u}\tilde{u}}$ to determine ω_i' and σ_i' (Figure 3.1). With center frequency and half-width fixed, we fit (3.24) to $F_i WC_{\tilde{u}\tilde{s}}(t)$ to estimate $\Delta\tilde{\tau}_p$, $\Delta\tilde{\tau}_g$, and $\Delta\tilde{\tau}_a$ (Figures 3.7 and 3.8). Finally, we use equation (3.34) to determine $\delta\tau_p$, $\delta\tau_g$, and $\delta\tau_a$ from $F_i WC_{\tilde{u}\tilde{s}}(t)$ (Figures 3.10 and 3.11). In the last step, the estimates of $\delta\tau_p$ are corrected for cycle skipping.

Example – complete synthetic seismogram

Figure 3.4 illustrates our ability to recover estimates of the generalized data functionals with the broad-band technique, in the case of an isolated waveform. As discussed then, the errors in the determination of some these parameters are due to the inadequacy of the linear-dispersion approximation across the frequency band of interest. We have demonstrated a procedure for estimating $\delta\tau_p$, $\delta\tau_g$, and $\delta\tau_a$ which reduces the bias introduced by quadratic dispersion. We now shall apply this methodology to the complete seismograms in order to evaluate the effect of windowing in terms of reducing interference.

Figure 3.7 illustrates the effect of the windowing operator graphically, by comparing $C_{\tilde{u}\tilde{s}}$, $WC_{\tilde{u}\tilde{s}}$, and $F_i WC_{\tilde{u}\tilde{s}}$ for three filters. Clearly, the window removes all the interfering energy, improving the approximation $WC_{\tilde{u}\tilde{s}} \approx WC_{\tilde{u}\tilde{u}}$. Comparison of this figure with Figure 2.15 is quite striking, which displays the filtered correlations without windowing. Figure 3.8 displays the values of $\Delta\tilde{\tau}_p$, $\Delta\tilde{\tau}_g$, and $\Delta\tilde{\tau}_a$ determined from $F_i WC_{\tilde{u}\tilde{s}}$ (squares) with those determined from $F_i C_{\tilde{u}\tilde{s}}$ (crosses). The windowing operator has greatly reduced the bias introduced by interfering arrivals, particularly for the differential group delay. Figure 3.9 compares the values of $\Delta\tilde{\tau}_p$ and $\Delta\tilde{\tau}_g$ recovered by

AD-A-243935

PAGE
#144

MISSING
FROM ORIGINAL
DOCUMENT
AS SENT TO
DTIC
FROM THE
ORIGINATOR

APPROXIMATION OF DIFFERENTIAL DISPERSION

We have seen that the application of the windowing operator changes the linear-dispersion approximation from a local approximation near ω_i' to a global approximation from ω_f to ω_i'. We found that the EU2-SNA differential dispersion does not satisfy this approximation over a broad-range of frequencies. One solution to this problem is to use better starting models. EU2-SNA is an extreme example with which to test this technique. An alternative solution is to use the quadratic form of differential wavenumber. While we have solved this problem (Appendix D) for the filtered, windowed cross-correlation function, we do not believe that it is necessary to fit the correlation functions for a four-parameter model. Instead, as discussed above, we have developed a correction scheme which allows the values of $\delta\tau_p$, $\delta\tau_g$, and $\delta\tau_a$ estimated with the linear-dispersion approximation to be corrected for quadratic dispersion.

SUMMARY

We have introduced an additional step in our waveform analysis procedure by windowing the broad-band correlation functions before filtering. This step reduces the contamination due to interfering arrivals, as illustrated by the complete synthetic case, although it limits the application of the linear-dispersion approximation. We formulated analytic expressions for the cross-correlation of the isolation filter with the complete synthetic and the observed seismogram which model the mode-branch interference as a sum over Gaussian wavelets, as well as parameterizing the correlation functions as a single Gaussian wavelet. We have developed explicit expressions for the effect of interference from other wavegroups and derived an expression which relates the observed phase perturbation to a linear sum of the individual phase perturbations of each traveling-wave branch.

AD-A-243935

PAGE
#146

MISSING
FROM ORIGINAL
DOCUMENT
AS SENT TO
DTIC
FROM THE
ORIGINATOR

effect of quadratic dispersion by an iterative-fitting process. In this example, the estimation of $\delta\tau_s^m$ has been greatly improved by the correction procedure.

FIGURE 3.5

The Fréchet kernels for the first-order perturbation in phase delay. Figure 3.5*a* presents the Fréchet kernels for shear-wave velocity (solid line) and density (dashed line) for the fundamental Love wave as a function of depth. Figure 3.5*b* presents the Fréchet kernels for shear-wave velocity (solid line) and density (dashed line), as well as the kernel for compressional velocity (short dashed line) for the fundamental Rayleigh wave as a function of depth. In both *a* and *b*, the kernels are presented at four frequencies. The two horizontal dashed lines indicate the 400 km discontinuity and the base of the crust. As observed in Chapter 2, Love and Rayleigh waves of the same frequency average the upper mantle in different ways.

FIGURE 3.6

Comparison between the measured values of $\delta\tau_p^m$ (open circles) and the values predicted from integration of the Fréchet kernels (filled circles). The predicted values compare favorably to the actual EU2-SNA fundamental-mode dispersion (solid line) of the Love wave (left) and Rayleigh wave (right).

FIGURE 3.7

$F_i WC_{\bar{u}\bar{s}}(t)$ for fundamental Love (left) and Rayleigh (right) waves. In each box, the uppermost trace is the broad-band cross-correlation function, $C_{\bar{u}\bar{s}}(t)$; the trace immediately below is the windowed function $WC_{\bar{u}\bar{s}}(t)$. Some interfering energy is apparent in $C_{\bar{u}\bar{s}}(t)$ and is eliminated by the window. The remaining traces (solid line) illustrate three filters with varying center frequencies (1 = 35, 2 = 25, 3 = 15 mHz) and a fixed relative bandwidth ($\gamma_i = 0.1$) applied to $WC_{\bar{u}\bar{s}}(t)$. The dashed line indicates the model derived from waveform fitting of equation (3.24). Comparing these correlations to Figure 2.15 demonstrates the effectiveness of the windowing operator in reducing interference. The values of $\Delta\tilde{\tau}_p$, $\Delta\tilde{\tau}_g$, and $\Delta\tilde{\tau}_a$ estimated from the waveform-fitting procedure are displayed in Figure 3.8.

FIGURE 3.8

Estimates of $\Delta\tilde{\tau}_p$, $\Delta\tilde{\tau}_g$, and $\Delta\tilde{\tau}_a$ (squares) recovered from waveform fitting of equation (3.24) to $F_i WC_{\bar{u}\bar{s}}(t)$. The top panels present the values for the Love wave; the bottom panels present the values for the Rayleigh wave. These estimates are near zero, indicating that interference has been significantly reduced by the windowing operator. The values estimated from $F_i C_{\bar{u}\bar{s}}(t)$ (crosses) in Chapter 2 are included for comparison.

FIGURE 3.9

Comparison between the values of $\Delta\tilde{\tau}_p$ and $\Delta\tilde{\tau}_g$ recovered by waveform fitting (squares) with those values predicted by (3.28) and (3.29) (filled circles). The match between the measured and predicted values is excellent.

AD-A-243935

PAGE
#148

MISSING
FROM ORIGINAL
DOCUMENT
AS SENT TO
DTIC
FROM THE
ORIGINATOR

does not alter the values of $\delta\tau_p$ greatly, although it does modify the estimates of $\delta\tau_g$ and $\delta\tau_a$. In the case of the fundamental-mode Rayleigh wave, however, the new estimates of $\delta\tau_g$ correct a cycle-skipping problem in the differential phase delay.

AD-A-243935

PAGE
#150

MISSING
FROM ORIGINAL
DOCUMENT
AS SENT TO
DTIC
FROM THE
ORIGINATOR

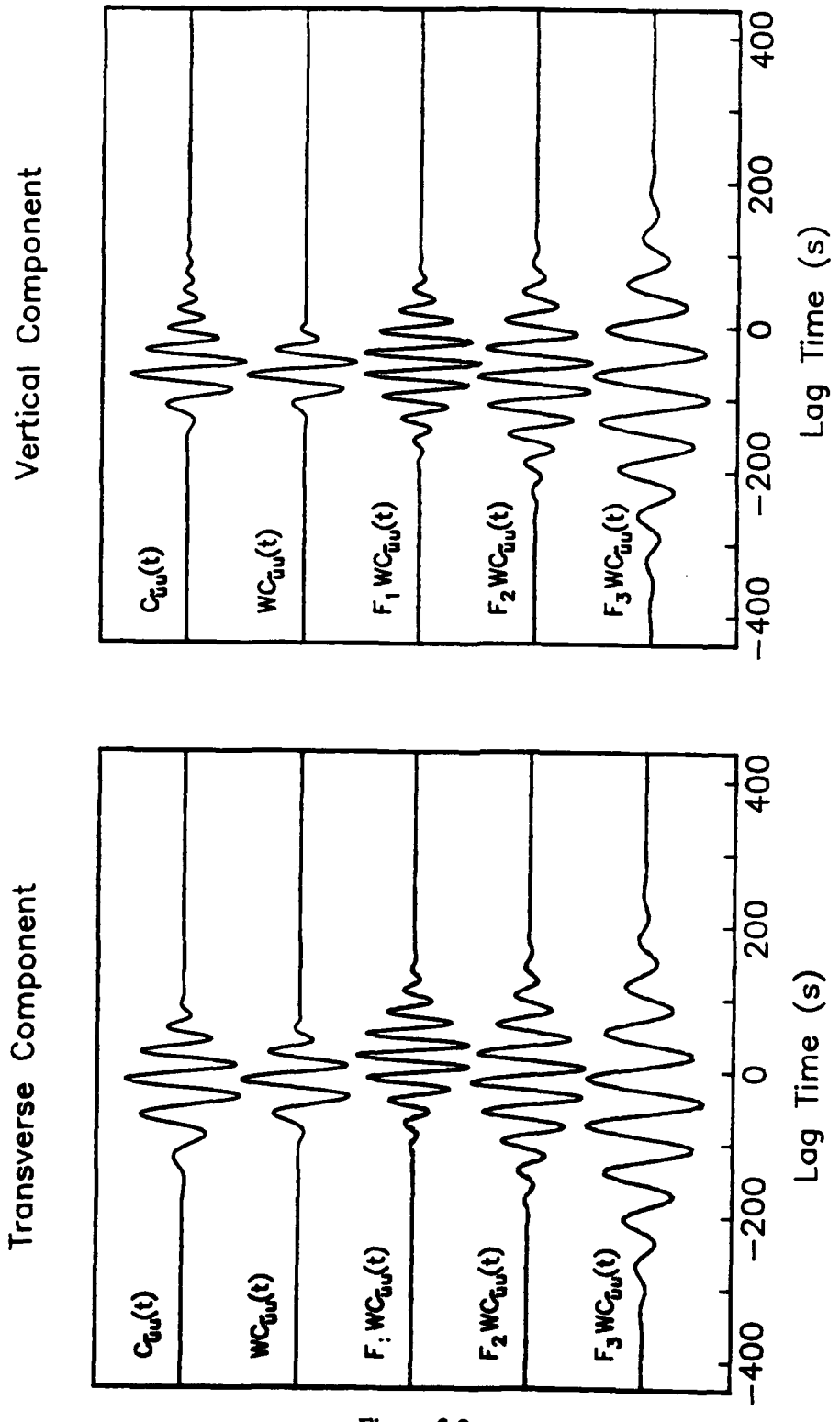

Vertical Component
Transverse Component
$C_{uu}(t)$
$WC_{uu}(t)$
$F_1 WC_{uu}(t)$
$F_2 WC_{uu}(t)$
$F_3 WC_{uu}(t)$
-400
-200
0
200
400
Lag Time (s)

Figure 3.2

AD-A-243935

PAGE
#152

MISSING
FROM ORIGINAL
DOCUMENT
AS SENT TO
DTIC
FROM THE
ORIGINATOR

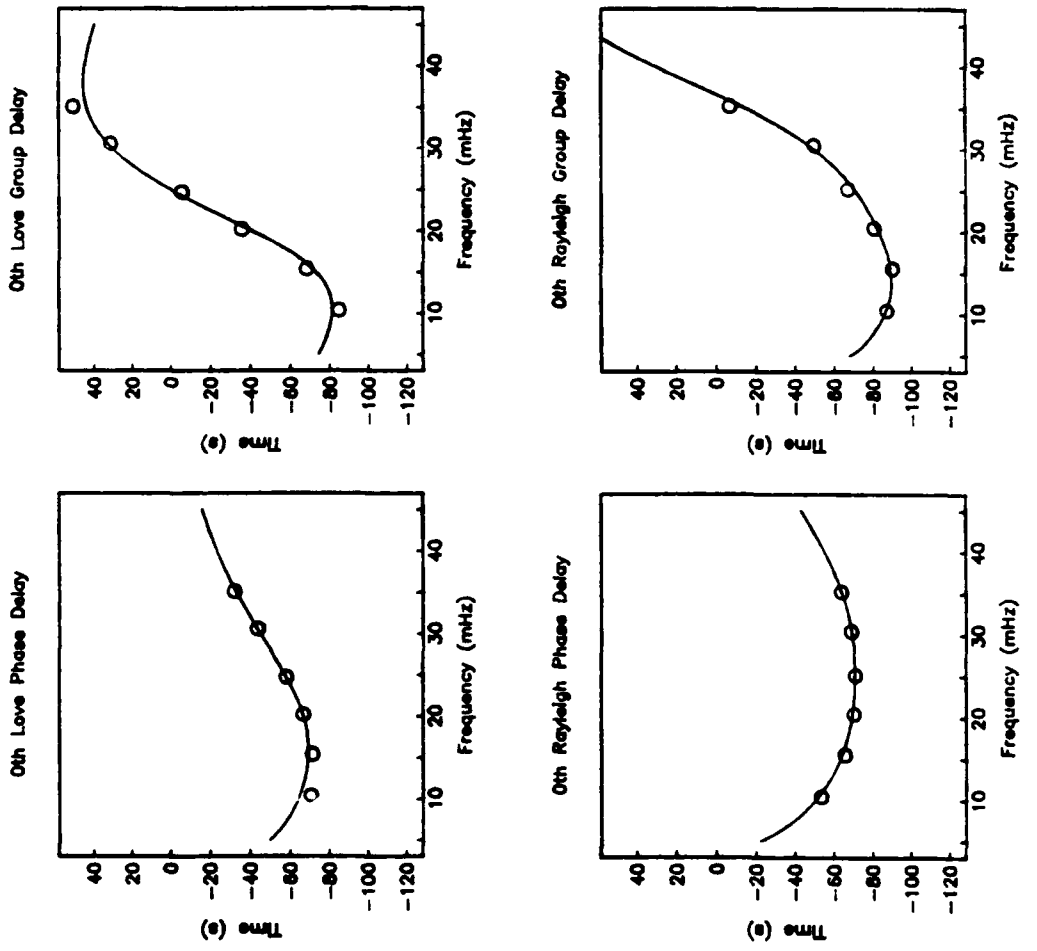

0th Love Group Delay
0th Rayleigh Group Delay
0th Love Phase Delay
0th Rayleigh Phase Delay
Time (s)
Frequency (mHz)
40
20
0
-20
-40
-60
-80
-100
-120
10
20
30
40

Figure 3.4

AD-A-243935

PAGE

#154

MISSING

FROM ORIGINAL

DOCUMENT

AS SENT TO

DTIC

FROM THE

ORIGINATOR

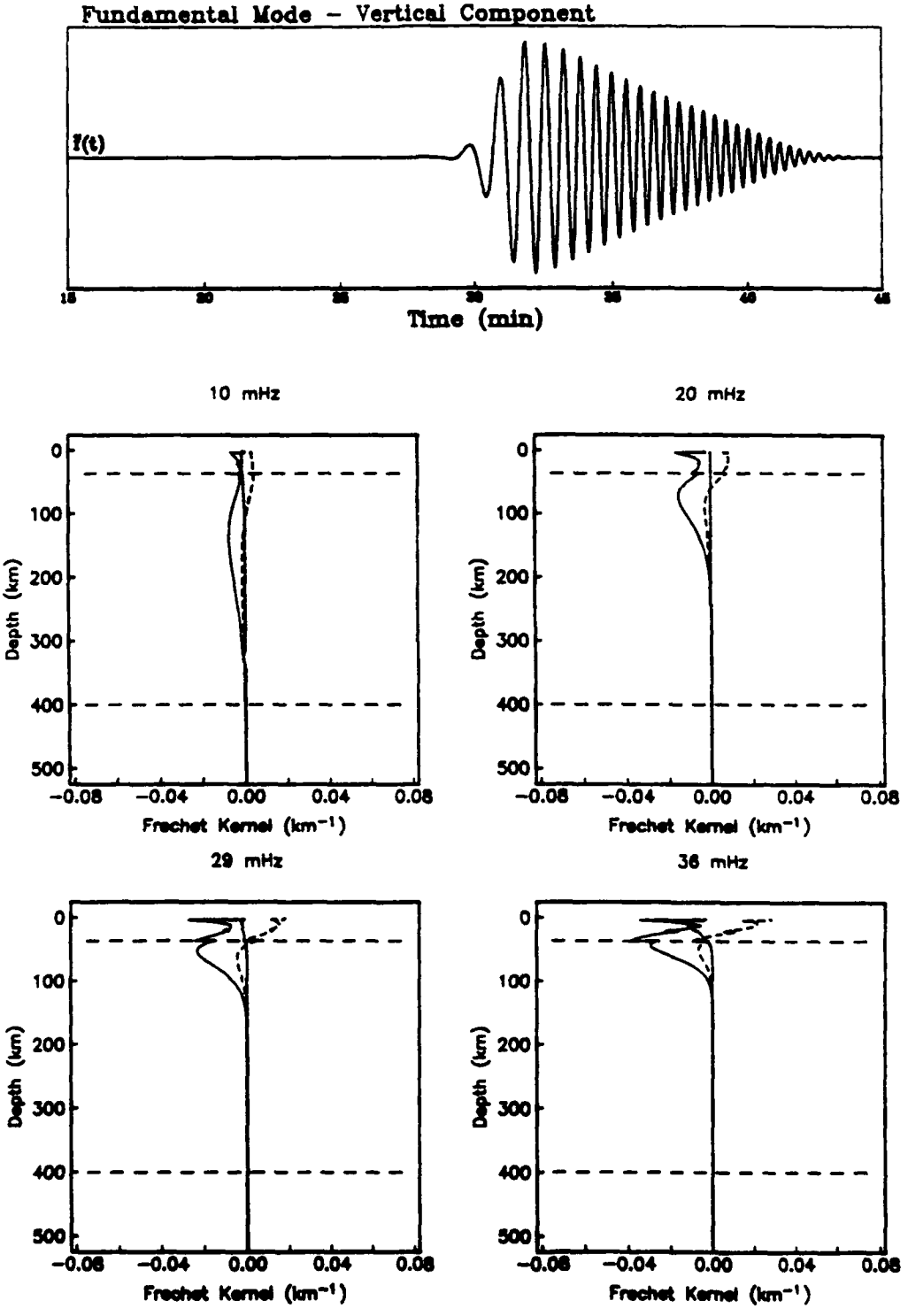

Fundamental Mode – Vertical Component
Y(t)
Time (min)
10 mHz
20 mHz
29 mHz
36 mHz
Depth (km)
Frechet Kernel (km⁻¹)
0
100
200
300
400
500
−0.08
−0.04
0.00
0.04
0.08

Figure 3.5b

AD-A-243935

PAGE

#156

MISSING

FROM ORIGINAL

DOCUMENT

AS SENT TO

DTIC

FROM THE

ORIGINATOR

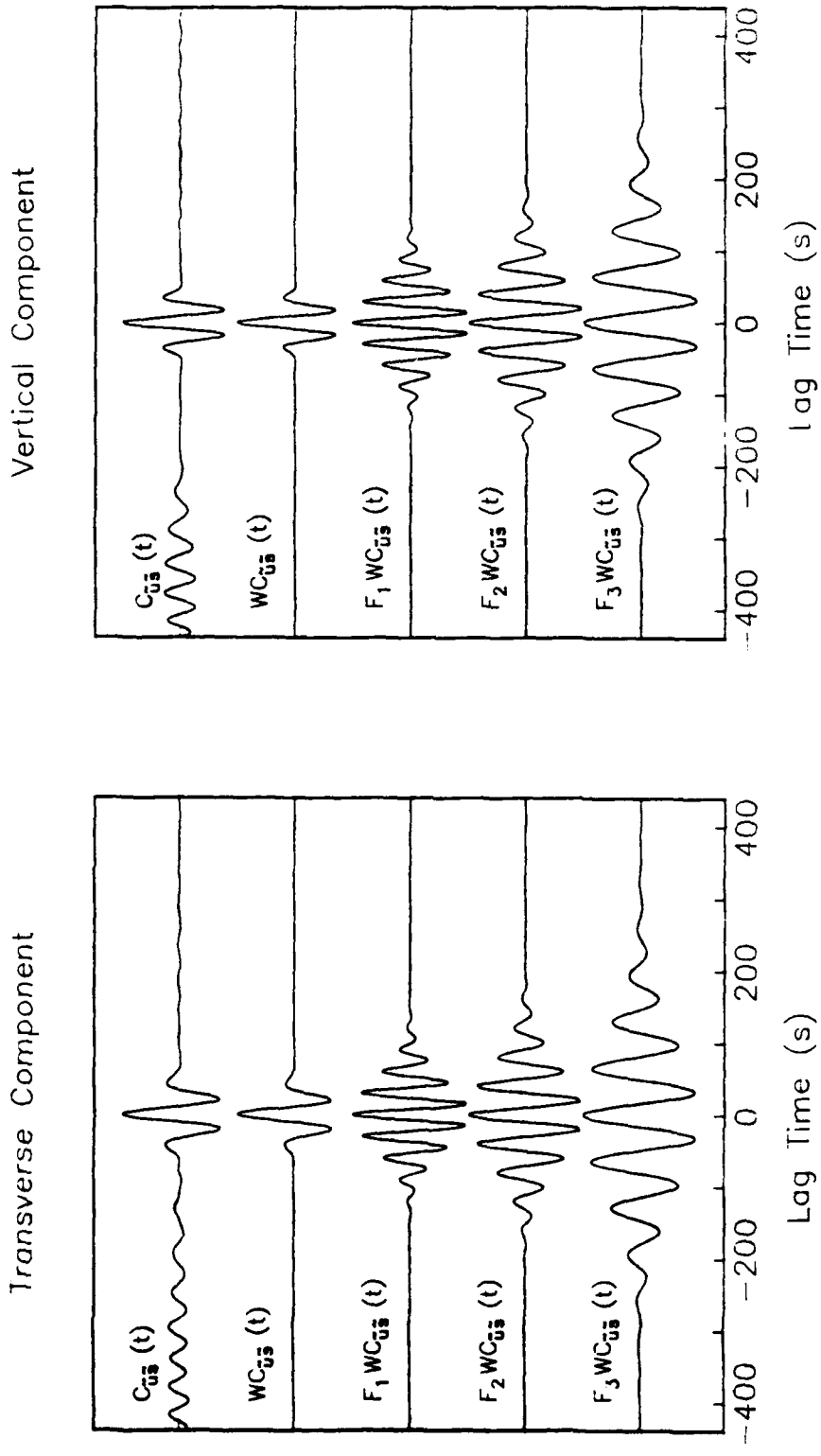

Vertical Component
Transverse Component
$C_{\bar{u}\bar{s}}$ (t)
$WC_{\bar{u}\bar{s}}$ (t)
$F_1WC_{\bar{u}\bar{s}}$ (t)
$F_2WC_{\bar{u}\bar{s}}$ (t)
$F_3WC_{\bar{u}\bar{s}}$ (t)
-400
-200
0
200
400
Lag Time (s)

Figure 3.7

AD-A-243935

PAGE

#158

MISSING

FROM ORIGINAL

DOCUMENT

AS SENT TO

DTIC

FROM THE

ORIGINATOR

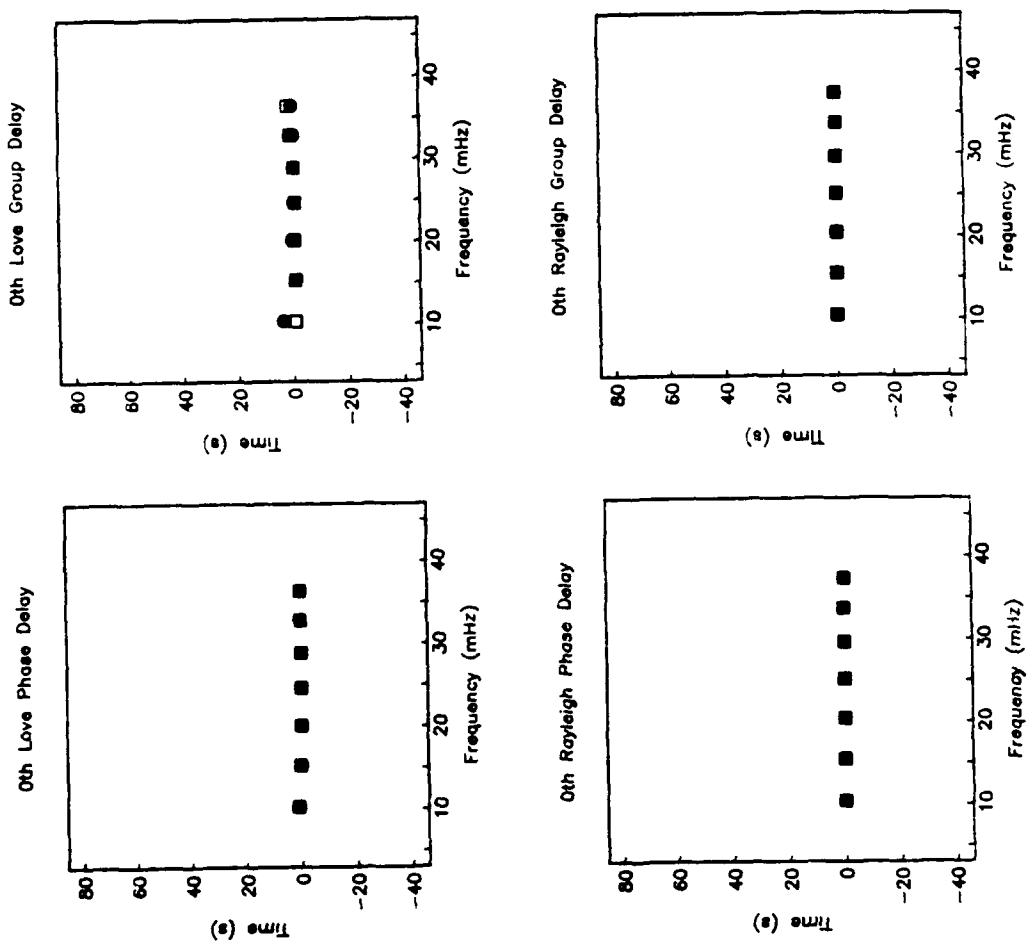

0th Love Group Delay
0th Love Phase Delay
0th Rayleigh Group Delay
0th Rayleigh Phase Delay
Time (s)
Frequency (mHz)
80
60
40
20
0
-20
-40
10
20
30
40

Figure 3.9

AD-A-243935

PAGE
#160

MISSING
FROM ORIGINAL
DOCUMENT
AS SENT TO
DTIC
FROM THE
ORIGINATOR

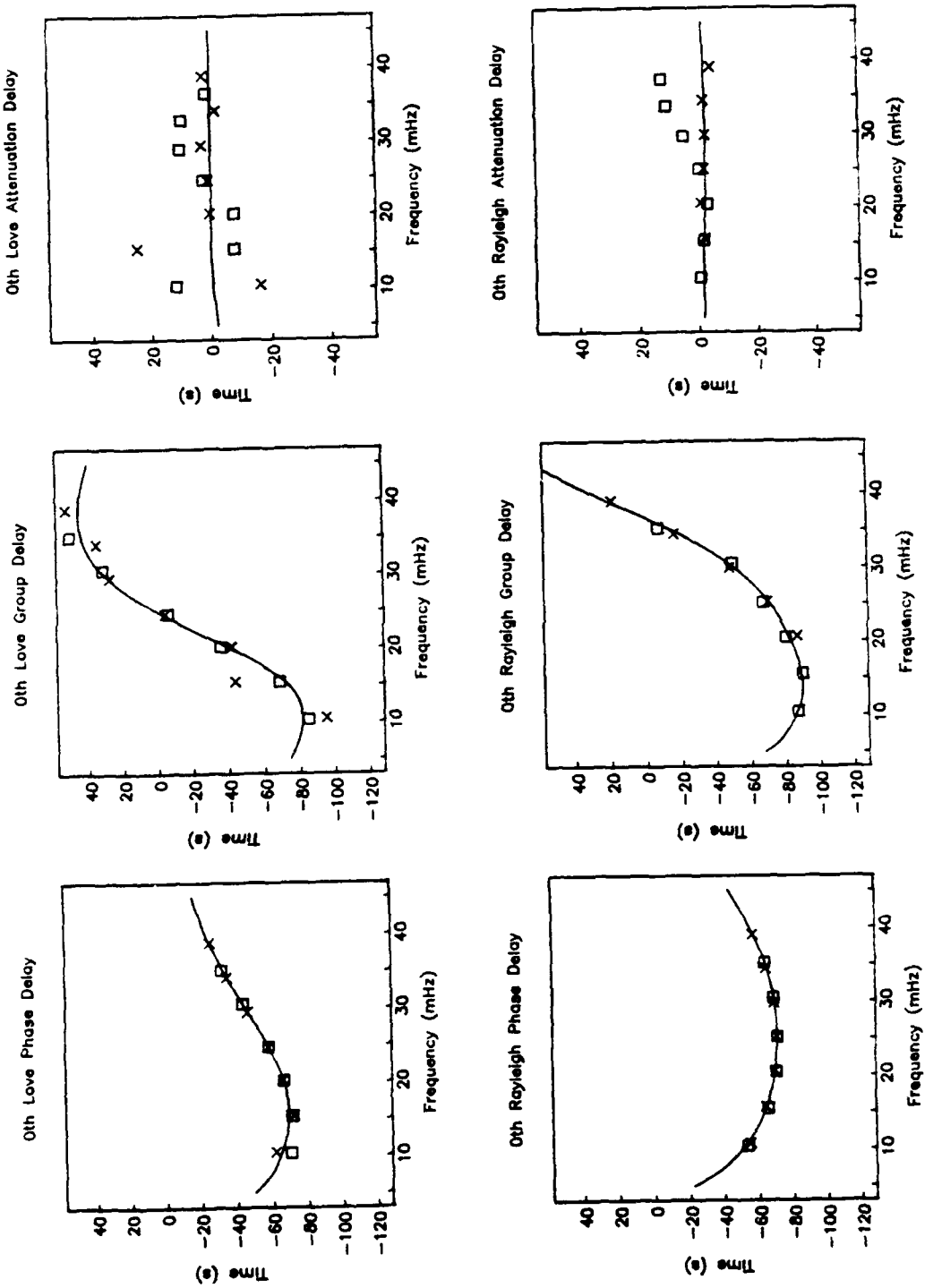

0th Love Attenuation Delay
0th Love Group Delay
0th Love Phase Delay
0th Rayleigh Attenuation Delay
0th Rayleigh Group Delay
0th Rayleigh Phase Delay
Time (s)
Frequency (mHz)

Figure 3.11

AD-A-243935

PAGE

#162

MISSING

FROM ORIGINAL

DOCUMENT

AS SENT TO

DTIC

FROM THE

ORIGINATOR

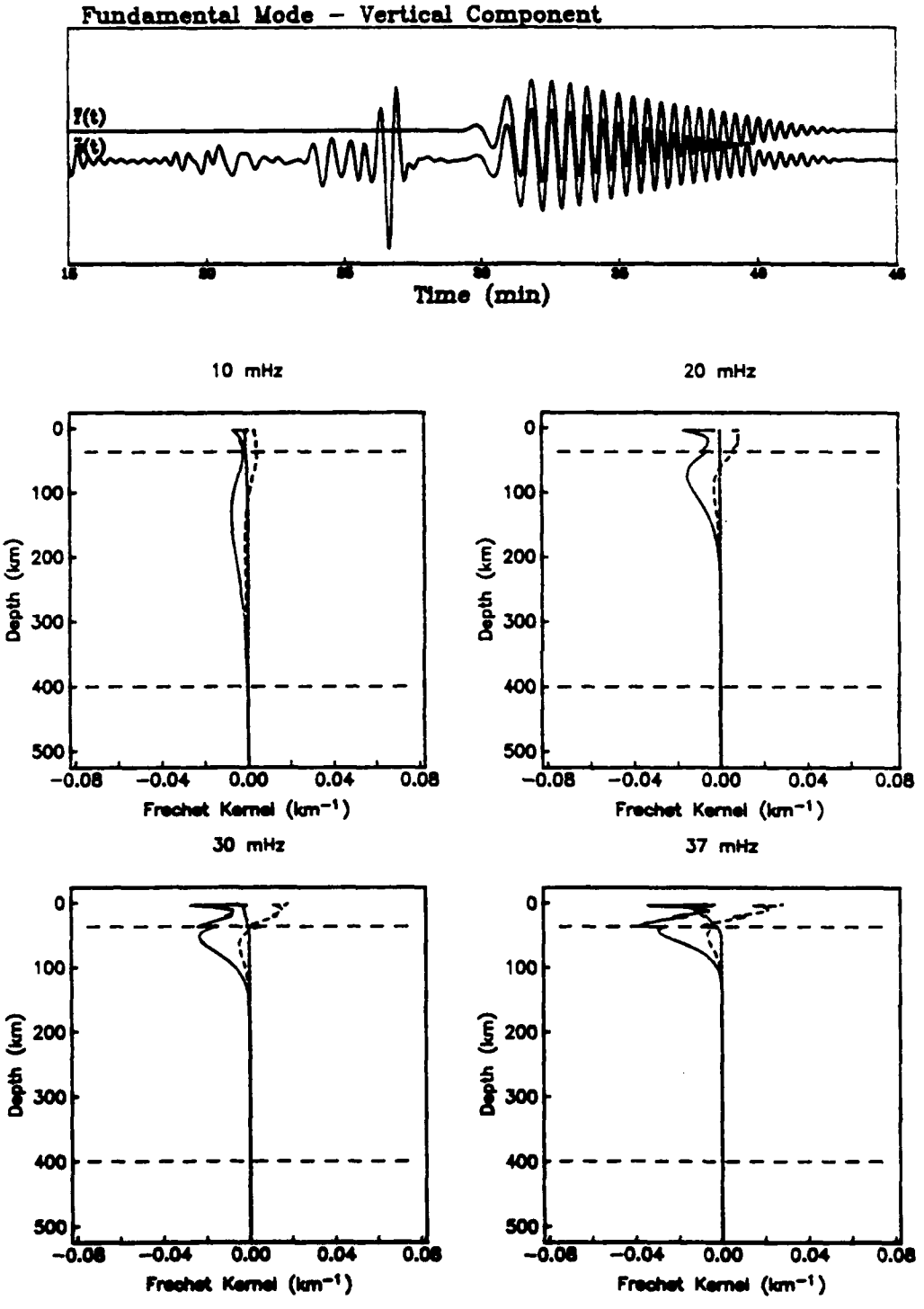

Fundamental Mode – Vertical Component
Time (min)
10 mHz
20 mHz
30 mHz
37 mHz
Depth (km)
Frechet Kernel (km⁻¹)

Figure 3.12b

AD-A-243935

PAGE
#164

MISSING
FROM ORIGINAL
DOCUMENT
AS SENT TO
DTIC
FROM THE
ORIGINATOR

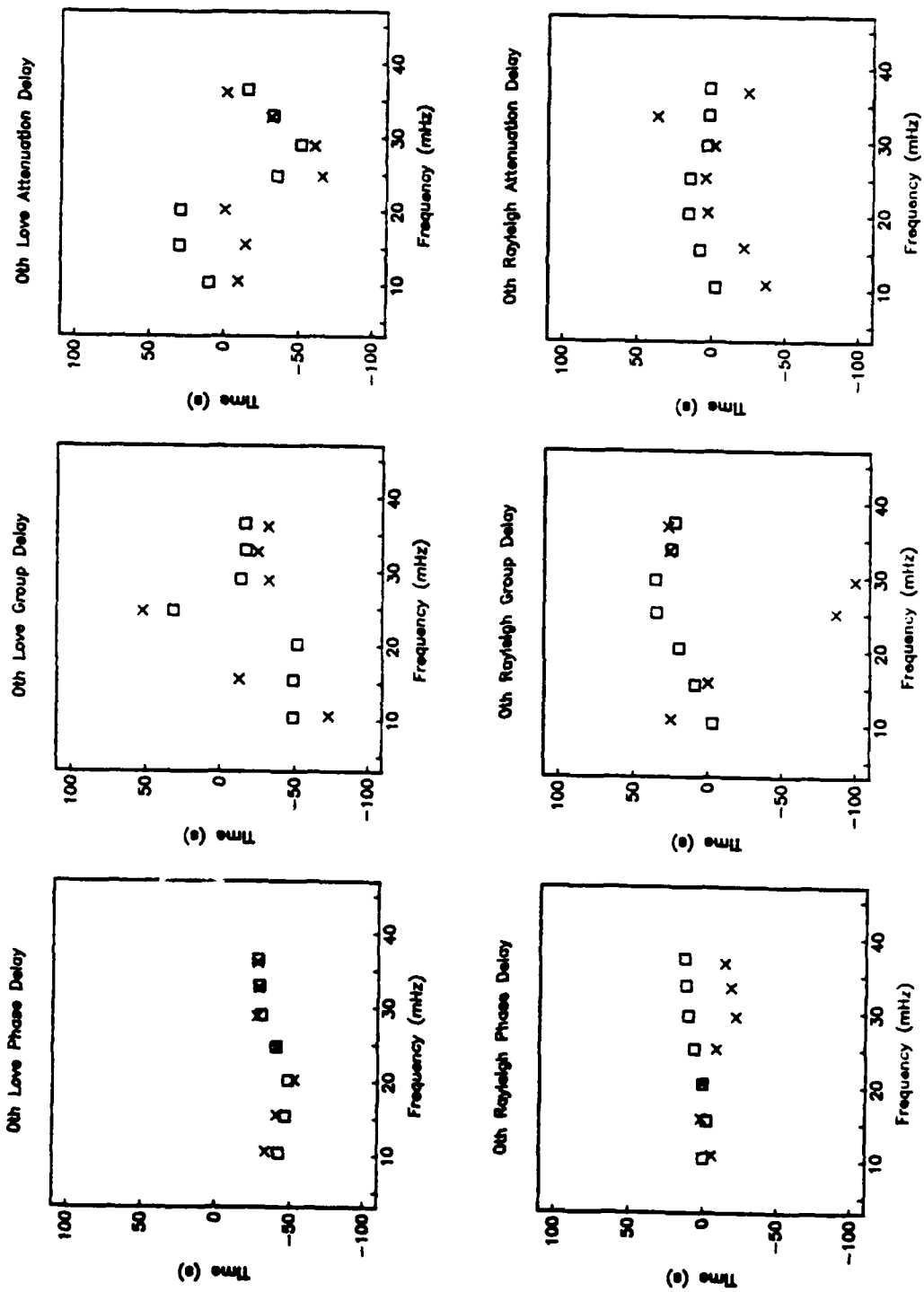

0th Love Attenuation Delay
0th Love Group Delay
0th Love Phase Delay
0th Rayleigh Attenuation Delay
0th Rayleigh Group Delay
0th Rayleigh Phase Delay
Time (s)
Frequency (mHz)
100
50
0
-50
-100
10
20
30
40

Figure 3.14

AD-A-243935

PAGE

#166

MISSING

FROM ORIGINAL

DOCUMENT

AS SENT TO

DTIC

FROM THE

ORIGINATOR

analysis of broad-band seismograms. We illustrate the implementation with the *S* and *SSS* examples and discuss Fréchet kernels obtained from our analysis. These kernels, which are very different from the partial derivatives associated with the infinite-frequency approximation of ray theory, provide new insight into the way wavegroups average the Earth at finite frequencies.

ISOLATION FILTER DESIGN

We motivate the discussion of isolation filters by considering the transverse and vertical-component synthetic seismograms at KONO for the Kamchatka event of 83/08/17 (Figure 4.1). The transverse-component seismogram is composed of four main packets of shear-energy: *S*, *SS*, *SSS*, and the fundamental-mode Love wave. At 63°, the *S*-wave bottoms around 1600 km depth and is followed by *sS* and *ScS*. *SS* also turns in the lower mantle, at approximately 950 km depth. *SSS* is quintuplicated at this distance, giving rise to five arrivals which sum to a large-amplitude pulse, and is sensitive to shear-velocity structure throughout the upper mantle and transition zone. It is closely followed by the fundamental-mode Love wave, which averages the structure of the crust and uppermost mantle. The vertical-component waveforms are more complex, due to *P* and *SV* coupling. For example, the *S*-wave on the vertical component is characterized by a low-amplitude wavetrain, rather than a single, well-defined arrival. Part of this complexity is explained by the presence of shear-coupled *PL* [Caloi, 1948; Oliver and Major, 1960; Oliver, 1961; Chandler *et al.*, 1968; Poupinet and Wright, 1972; Jordan and Frazer, 1975, Frazer, 1977] and part by converted arrivals [Kanasewich *et al.*, 1973; Ward, 1978]. The difficulty of modeling shear-coupled *PL* has led seismologists to abandon the *PSV* system [*e.g.*, Helmberger and Engen, 1974] in favor of *SH*-polarized waveforms, although some recent studies have attempted to utilize *PSV*-polarized waveforms [Baag and Langston, 1985a, 1985b; Langston and Baag, 1986]. The large-amplitude phase on the vertical component is *SSS*. This pulse-like arrival, with a group velocity between 4.4 and 4.6 km/s, is

AD-A-243935

PAGE
#168

MISSING
FROM ORIGINAL
DOCUMENT
AS SENT TO
DTIC
FROM THE
ORIGINATOR

$$\tilde{\alpha}_m(\omega) = \mathrm{Ga}\,[\sigma_o(t_o - \tilde{\tau}_g^m(\omega))] \tag{4.2}$$

The total group delay of the mth branch, $\tilde{\tau}_g^m$, is the sum of contributions from the excitation, propagation, and instrument response:

$$\tilde{\tau}_g^m(\omega) = \tilde{\varphi}_g^m(\omega) + \tilde{t}_g^m(\omega) + \tilde{\varphi}_g^I(\omega) \tag{4.3}$$

where $\tilde{\varphi}_g^m \equiv x\, d\,\tilde{\varphi}_m / d\omega$ is the excitation group delay, $\tilde{t}_g^m \equiv \mathrm{Re}\{dk_m / d\omega\}$ is the propagation group delay, and $\tilde{\varphi}_g^I \equiv d\,\tilde{\varphi}_I / d\omega$ is the instrument group delay. In general, the propagation delay is the dominant term in (4.3), but occasionally the contribution from the excitation term may be significant. The instrument delay does not depend on the branch index and is a slowly varying function of frequency.

It is important to note that $\tilde{\alpha}_m(t)$ is not a group-velocity window in the classical sense: *i.e.*, $\tilde{f}(t)$ is not constructed by windowing the complete synthetic seismogram about a particular group arrival time. Instead, we compute the sum over traveling-wave branches with amplitude coefficients determined by (4.2). Returning to the seismograms in Figure 4.1, we shall use this formulation to construct isolation filters for two very different wavegroups: *S* and *SSS*.

S isolation filter. At 63°, *S* is bottoming in the relatively homogeneous lower mantle and we expect the pulse shape to be simple. This is true on the transverse component, where *S* has a group arrival time of t_o = 1149 s (corresponding to a group velocity of 6.11 km/s). However, the vertical component does not have a single, strong arrival due to the interference of shear-coupled *PL* and converted arrivals. In order to construct an isolation filter for *S*, we chose the group arrival time on the transverse component as t_o. The choice of σ_o^{-1} was made after some experimentation; we found 100 s to be a reasonable value. In

AD-A-243935

PAGE
#170

MISSING
FROM ORIGINAL
DOCUMENT
AS SENT TO
DTIC
FROM THE
ORIGINATOR

the group-velocity diagrams in Figure 1.4 indicates that many branches have Airy-phase behavior at this group velocity, contributing to the large-amplitude arrival. This observation is consistent with the numerical studies of *Sa* by Kovach and Anderson [1964], Schwab *et al.* [1974], and Nakanishi *et al.* [1976].

Figure 4.7 illustrates the isolation filters generated from these amplitude coefficients. On both the transverse and vertical components, the group arrival time algorithm has reproduced the *SSS*-wave, matching the complete synthetic in both amplitude and phase. The periodic blips in the isolation filters are edge effects caused by the phase-velocity cutoff at 7 km/s and are analogous to truncation arrivals in WKBJ seismograms.

FRÉCHET KERNELS

Having demonstrated our ability to construct seismologically useful and interesting isolation filters by weighted sums of traveling-wave branches, we now develop a formalism for the interpretation of the generalized data functionals recovered with their use. In particular, we derive expressions for their Fréchet kernels. Based on the mathematical treatment developed in Chapters 2 and 3, we present formulae for the analysis of broad-band seismograms. In our analysis, we have assumed that the parameters of the applied window are judiciously chosen such that we may neglect any correction terms due to the window in the estimation of the differential phase, group, and amplitude time parameters. In the development below, we assume that the $\tilde{\alpha}_m(\omega)$ are smoothly-varying, real functions of frequency, although the expressions do not depend on a specific form of the weight coefficients.

Autocorrelation function

In Chapters 2 and 3, we considered an isolation filter composed of a single waveform and expanded $C_{\bar{m}\bar{m}}(t)$ in a Gram-Charlier series characterized by a center frequency ω_f and spectral half-width σ_f. We modeled the application of a windowing

AD-A-243935

PAGE
#172

MISSING
FROM ORIGINAL
DOCUMENT
AS SENT TO
DTIC
FROM THE
ORIGINATOR

$$C_{\tilde{f}\tilde{s}}(t) = \sum_m \tilde{\alpha}_m(t) * \tilde{u}_m(t) \otimes \sum_n \tilde{u}_n(t)$$

$$= \frac{1}{2\pi} \sum_m \sum_n \int_{-\infty}^{\infty} \tilde{\alpha}_m(\omega)\, \tilde{u}_m^*(\omega)\, \tilde{u}_n(\omega)\, e^{-i\omega t}\, d\omega \tag{4.5}$$

While the autocorrelation is a symmetric function about zero lag, $C_{\tilde{f}\tilde{s}}(t)$ contains additional cross-terms and the peak may be shifted away from zero lag. We assume that the amplitude spectrum of $C_{\tilde{f}\tilde{s}}(\omega)$, including $\tilde{\alpha}_m(\omega)$, is a slowly varying function of frequency and branch number. Near $\omega = \omega_i'$, we approximate the amplitude spectrum of the individual branch cross-correlations in terms of a scaled version of the autocorrelation spectrum:

$$\tilde{\alpha}_m(\omega)\, |\tilde{A}_n(\omega)|\, |\tilde{A}_m(\omega)|\, |\tilde{I}(\omega)|^2 \approx \tilde{\alpha}_m(\omega_i')\, |\tilde{A}_n(\omega_i')|\, |\tilde{A}_m(\omega_i')|\, C_{\tilde{f}\tilde{f}}(\omega) \tag{4.6}$$

The approximation assumes that most of the cross-term contributions come from nearest-neighbor branches, whose spectral characteristics are similar. As discussed in Chapter 2, this approximation will break down in the case of the cross-correlation between a high-phase-velocity branch with a low-phase velocity branch, because their amplitude spectrum will be quite different. However, the differential group delay in this case will be large, and the windowing operator will substantially reduce this contribution to the sum.

With this parameterization of the amplitude spectrum and the expansion of the phase spectrum in a first-order Taylor series (2.49), we may derive an expression for the filtered cross-correlation between the isolation filter and the complete synthetic:

$$F_i W C_{\tilde{f}\tilde{s}}(t) = \frac{1}{(2\pi)^{5/2}} \frac{\sigma_i'}{\sigma_i \sigma_g} \sum_m \sum_n \tilde{\alpha}_m(\omega_i')\, \tilde{C}_{mn}\, \mathrm{Ga}\left(\sigma_i'(t - \Delta\tilde{\tau}_g^{mn})\right)$$

$$\times \cos\left[\omega_i'(t - \Delta\tilde{\tau}_p^{mn}) - \sigma_i'^2\, \Delta\tilde{\tau}_a^{mn}(t - \Delta\tilde{\tau}_g^{mn})\right] \tag{4.7}$$

where $\tilde{C}_{nm}$, defined in Chapter 3, depends on the amplitude coefficients of the nth and mth branches at ω_i'.

This expression describes the filtered, windowed cross-correlation between the isolation filter and the complete synthetic seismogram as a sum over Gaussian wavelets; each one describing the interaction between two traveling-wave branches in terms of a differential phase, group, and amplitude delay. We expect the sum to be dominated by the autocorrelation terms ($n = m$) and assume that it may be approximated by a single wavelet (3.24), parameterized by an average perturbation in amplitude and phase ($\tilde{C}$, $\Delta\tilde{\tau}_p$, $\Delta\tilde{\tau}_g$, $\Delta\tilde{\tau}_a$) due to interference. As we did in Chapters 2 and 3, we can develop expressions for these average perturbations in terms of the individual contributions of the traveling-wave branches. Illustrating the approach in the simple case when there is no differential attenuation, we find:

$$\tilde{C}^2 = \left[\sum_m \sum_n \tilde{\alpha}_m(\omega_i') \tilde{D}_{nm} \cos(\omega_i' \Delta\tilde{\tau}_p^{nm})\right]^2 + \left[\sum_m \sum_n \tilde{\alpha}_m(\omega_i') \tilde{D}_{nm} \sin(\omega_i' \Delta\tilde{\tau}_p^{nm})\right]^2 \tag{4.8}$$

$$\Delta\tilde{\tau}_p(\omega_i') = \frac{1}{\omega_i'} \tan^{-1}\left(\frac{\sum_m \sum_n \tilde{\alpha}_m(\omega_i') \tilde{D}_{nm} \sin(\omega_i' \Delta\tilde{\tau}_p^{nm})}{\sum_m \sum_n \tilde{\alpha}_m(\omega_i') \tilde{D}_{nm} \cos(\omega_i' \Delta\tilde{\tau}_p^{nm})}\right) \tag{4.9}$$

$$\Delta\tilde{\tau}_g(\omega_i') = \frac{c_0 c_1 + s_0 s_1}{c_0^2 + s_0^2} \tag{4.10}$$

where the coefficients c_0, c_1, s_0, and s_1 are defined in terms of sums over the differential branch phase and group delays:

$$
\begin{aligned}
c_0 &= \sum_m \sum_n \tilde{\alpha}_m(\omega_i') \; \tilde{D}_{nm} \cos(\omega_i' \Delta\tilde{\tau}_p^{nm}) \\
c_1 &= \sum_m \sum_n \tilde{\alpha}_m(\omega_i') \; \tilde{D}_{nm} \cos(\omega_i' \Delta\tilde{\tau}_p^{nm}) \, \Delta\tilde{\tau}_g^{nm} \\
s_0 &= \sum_m \sum_n \tilde{\alpha}_m(\omega_i') \; \tilde{D}_{nm} \sin(\omega_i' \Delta\tilde{\tau}_p^{nm}) \\
s_1 &= \sum_m \sum_n \tilde{\alpha}_m(\omega_i') \; \tilde{D}_{nm} \sin(\omega_i' \Delta\tilde{\tau}_p^{nm}) \, \Delta\tilde{\tau}_g^{nm}
\end{aligned}
\tag{4.11}
$$

and we recall the definition of the $\tilde{D}_{nm}$ from Chapter 3:

$$
\tilde{D}_{nm} = \tilde{A}_{nm} \; \mathrm{Ga}\left(\frac{\sigma_f \sigma_w}{\sigma_g}(t_c - \Delta\tilde{\tau}_g^{nm})\right) \; \mathrm{Ga}\left(\sigma_i' \; (\Delta\tilde{\tau}_g - \Delta\tilde{\tau}_g^{nm})\right) \; / \; \mathrm{Ga}\left(\frac{\sigma_f \sigma_w}{\sigma_g}(t_c - \Delta\tilde{\tau}_g)\right) \tag{4.12}
$$

In summary, we have formulated the cross-correlation function between the isolation filter and the complete synthetic seismogram in terms of a double sum over Gaussian wavelets which describe the interaction between the traveling-wave branches. We have approximated this representation for $F_i C_{\tilde{f}s}(t)$ as a single Gaussian wavelet and derived expressions for the scale constant $\tilde{C}$ and the average phase and group delays $\Delta\tilde{\tau}_p$ and $\Delta\tilde{\tau}_g$ in terms of the individual branch contributions. These formulae are similar in structure to those derived in Chapter 3, as they depend on weighted sums over sines and cosines of $\Delta\tilde{\tau}_p^{nm}$. They differ from earlier expressions, however, since the sum over the branches of the synthetic seismogram is calculated for each branch in the isolation filter and weighted by the $\tilde{\alpha}_m(\omega_i')$.

The observed cross-correlation function

We now present an expression for the filtered, windowed cross-correlation between $\tilde{f}(t)$ and $s(t)$. $F_i W C_{\tilde{f}s}(t)$ may be written as a double sum over the cross-correlations between $\tilde{u}_m$ of the isolation filter and the u_m which comprise the observed seismogram, where we have applied the windowing operator at the peak of $C_{\tilde{f}s}$ ($t_w \equiv t_c + \delta t_c$):

$$F_i WC_{\tilde{f}s}(t) = \frac{1}{(2\pi)^{5/2}} \frac{\sigma_i'}{\sigma_i \sigma_g} \sum_m \sum_n \tilde{\alpha}_m(\omega_i')\ C_{nm}\ \mathrm{Ga}\,(\sigma_i'(t-\delta\tau_g^n-\Delta\tilde{\tau}_g^{nm}))$$

$$\times \cos\,[\omega_i'(t-\delta\tau_p^n-\Delta\tilde{\tau}_p^{nm}) - \sigma_i'^2\,(\delta\tau_a^n+\Delta\tilde{\tau}_a^{nm})(t-\delta\tau_g^n-\Delta\tilde{\tau}_g^{nm})] \qquad (4.13)$$

where C_{nm}, defined in Chapter 3, depends on the amplitude coefficients of the nth and mth branches at ω_i'.

Although (4.13) represents the filtered, windowed cross-correlation between $\tilde{f}(t)$ and $s(t)$ as a sum over many Gaussian wavelets, only a few will make a significant contribution to the correlation function, due to the presence of the window. We represent the sum as a single wavelet (3.34), characterized by perturbations due to interfering waveforms ($\Delta\tilde{\tau}_p$, $\Delta\tilde{\tau}_g$, $\Delta\tilde{\tau}_a$) and perturbations due to deviations of the reference model from the Earth ($\delta\tau_p$, $\delta\tau_g$, $\delta\tau_a$). We can develop expressions for the generalized data functionals, $\delta\tau_p$, $\delta\tau_g$, and $\delta\tau_a$, in terms of the individual contributions of the traveling-wave branches, which will allow us to formulate expressions for the Fréchet kernels as linear sums of the branch parameters. As an example of the approach, we present the formula for the Fréchet kernel for

(ω_i') in the absence of differential attenuation:

$$\delta\tau_p(\omega_i') = \frac{\sum_n \tilde{P}_n\ \delta\tau_p^n(\omega_i')}{\sum_n \tilde{P}_n} \qquad (4.14)$$

where the $\tilde{P}_n$ are defined by

$$\tilde{P}_n = \sum_m \tilde{\alpha}_m(\omega_i')\ \tilde{D}_{nm}\ \cos[\omega_i'(\Delta\tilde{\tau}_p - \Delta\tilde{\tau}_p^{nm})] \qquad (4.15)$$

This equation expresses the average phase perturbation as a linear sum of appropriately scaled branch phase perturbations, that is, as a linear sum of the individual Fréchet kernels at the reference frequency ω_i'. In Chapters 2 and 3, each $\bar{P}_n$ represented the interaction between the single branch of the isolation filter and the nth branch of the synthetic seismogram. In the general formulation of (4.1), $\bar{P}_n$ is a sum over the branches of the isolation filter.

IMPLEMENTATION

We have developed a software package for the design of isolation filters. This package allows us to construct $\bar{f}(t)$ for particular values of t_0 and σ_0^{-1}, accounting for the source, propagation, and instrument contributions. In addition, we may apply phase and group-velocity limits as well as identify those modes which satisfy additional criteria, such as certain ratios of shear-to-compressional energy. Once a set of coefficients are determined, the filter is constructed by normal-mode summation. For the types of filters discussed thus far, the number of modes involved in the sum is substantially less than the total mode set, and isolation filters may be calculated quite efficiently. Once the isolation filter is determined, the processing proceeds as described in Chapter 3. Returning to the seismograms in Figure 4.1, we illustrate this formalism with the isolation filters for *S* and *SSS*.

S isolation filter

Correlation functions. In the first step, we calculate the autocorrelation of the isolation filter and apply the windowing operator and the narrow-band filter. Figure 4.9 displays $C_{\bar{f}\bar{f}}(t)$, $WC_{\bar{f}\bar{f}}(t)$, and $F_iWC_{\bar{f}\bar{f}}(t)$ for several values of ω_i. The autocorrelation of the isolation filter, as defined in (4.4), is a symmetric function which may contain cross-term contributions away from zero lag. In this example, the autocorrelation function of the transverse component *S*-wave is concentrated near the peak, while the vertical-component

function is not. However, the windowing operator eliminates most of the cross-term contributions away from zero lag, allowing us to apply the Gaussian-wavelet approximation to $F_i WC_{\tilde{f}\tilde{f}}(t)$. We use (3.6) to estimate ω_i' and σ_i' by waveform fitting and the resulting models are plotted with dashed lines.

Having determined ω_i' and σ_i' from $F_i WC_{\tilde{f}\tilde{f}}(t)$, we compute the cross-correlation between $\tilde{f}(t)$ and $\tilde{s}(t)$. Figure 4.10 compares $C_{\tilde{f}\tilde{s}}(t)$, $WC_{\tilde{f}\tilde{s}}(t)$, and $F_i WC_{\tilde{f}\tilde{s}}(t)$ for several values of ω_i. Considerable cross-term energy is apparent near zero-lag on the correlation functions, particularly on the vertical component where $C_{\tilde{f}\tilde{s}}(t)$ does not have the same sidelobe structure as $C_{\tilde{f}\tilde{f}}(t)$. Most of the interference is eliminated by the windowing operator and we use the single-wavelet approximation of $F_i WC_{\tilde{f}\tilde{s}}(t)$ (3.24) to estimate the average differential phase, group, and amplitude time delays due to interference. The results obtained from waveform fitting are displayed in Figure 4.11. The values of these parameters are near zero, indicating that the window has successfully removed most of the interference. The filled circles indicate the values of $\Delta\tilde{\tau}_p$ and $\Delta\tilde{\tau}_g$ predicted from the analytic expressions (4.9) and (4.10). The comparison between the measured and predicted values of $\Delta\tilde{\tau}_p$ is within a second or so on both the transverse and vertical components, except at the very lowest frequency. While the predicted values of $\Delta\tilde{\tau}_g$ on the transverse component are within a few seconds of the measured values, with the greatest error occurring at the lowest frequency, the vertical-component results are puzzling. Although the measured values of $\Delta\tilde{\tau}_g$ mimic the dispersion of the predicted values, they are offset by 12 to 15 s in the range from 10 to 20 mHz.

This behavior is due to the breakdown of the wide-window approximation. The vertical-component correlation function is not as localized as the transverse component correlation function and consequently is more sensitive to the width of the window. It is apparent from comparison of Figures 4.9 and 4.10 that $C_{\tilde{f}\tilde{s}} \neq C_{\tilde{f}\tilde{f}}$, even for a single cycle about the maximum. Elimination of the sidelobe structure by the windowing operator

causes the measurement procedure to underestimate the effect of interference. Experimentation has shown that better agreement may be obtained with a wider window.

Finally, we calculate the cross-correlation between $\tilde{f}(t)$ and $s(t)$. Figure 4.12 illustrates $C_{\tilde{f}s}(t)$, $WC_{\tilde{f}s}(t)$, and $F_iWC_{\tilde{f}s}(t)$ for several values of ω_i. While the transverse-component correlation functions has the simple structure of a Gaussian wavelet, the vertical-component correlation function does not. As in Figure 4.10, considerable cross-term energy is apparent in the broad-band correlation function and is removed by the windowing operator. The dashed line indicates the best model derived from waveform fitting of (3.34) and Figure 4.13 displays the values of $\delta\tau_p$, $\delta\tau_g$, and $\delta\tau_a$ recovered from this operation. The solid line indicates the actual values for the EU2-SNA comparison. The transverse-component *S*-wave behaves like a classical body wave, *i.e.*, it is not dispersed, and the recovered parameters match the values predicted from a simple travel-time calculation. The vertical-component arrival is dispersed, however, with values of $\delta\tau_p$ which are a smoothly varying function of frequency.

Fréchet kernels. We may use the expression for the first-order phase perturbation (4.14) to calculate the Fréchet kernels for the *S*-wave isolation filters. Figure 4.14 presents the transverse (*a*) and vertical (*b*) component kernels for velocity and density perturbations at four sample frequencies. The horizontal dashed lines indicate the Moho, the 400 km discontinuity, and the 670 km discontinuity. The solid line is the shear-velocity kernel; the shorter dashed line is the compressional-velocity kernel. The longer dashed line marks the density kernel.

The *SH*-polarized shear-velocity kernels demonstrate that the *S*-wave travel time represents a smoothed average of lower-mantle structure. This averaging becomes localized near the bottoming depth of *S* with increasing frequency, corresponding to the high-frequency turning-point singularity. At this distance, *S* travels nearly vertically through the upper mantle and is relatively insensitive to structure there. The density kernel has low amplitude in the lower mantle, but oscillates about zero in the transition zone and

upper mantle. This oscillation is matched in the shear-velocity kernel and is a function of the reference frequency. While the velocity kernels for an individual branch are non-positive everywhere (as they are for classical travel time), the shear kernels in this figure become positive in several places. This positivity is due to the interaction of interfering wavegroups.

The *PSV*-polarized kernels are an interesting comparison to the *SH* figure. The shear-velocity kernels at 30 and 40 mHz kernels are quite similar for toroidal and spheroidal modes, while the 10 and 20 mHz kernels are very different. Specifically, the 10 and 20 mHz *PSV*-polarized shear-velocity kernels display considerably more sensitivity to upper-mantle structure than the *SH* kernels. This additional sensitivity is due to the presence of shear-coupled *PL* and converted energy in the *S* wavetrain. The difference is less pronounced at higher frequencies because the *S*-wave energy (which has a center frequency of 33 mHz) dominates the wavetrain.

The compressional-velocity kernels indicate that the measurements of phase delay made with an *S*-wave isolation filter contain information about the *P*-wave velocity in the upper mantle. This sensitivity is exhibited at all frequencies, becoming progressively more pronounced with increasing frequency. In fact, the comparison of the density and compressional-velocity kernels indicate that the measurements at 30 and 40 mHz are sensitive to the reflection coefficient at 400 km depth. Woodhouse and Dziewonski [1986] inverted *PSV*-polarized waveforms for *SV* structure and observed a correlation between the lower-mantle shear velocities and independent models of upper-mantle compressional-velocity. They concluded that the *SV* waveforms contained significant amounts of *P*-wave energy. Our methodology provides us with the means of interpreting the compressional-energy contribution to the *SV* waveform.

We may compare the measurements of differential phase delay (squares) with the values predicted from the integration of the Fréchet kernels with the velocity perturbations between EU2 and SNA (filled circles) (Figure 4.15). The comparison for the *SH*

measurement is very good, with phase estimates recovered to a second or so, except for the lowest frequency. The *PSV* comparison is good for frequencies above 25 mHz, but shows some "tears" between the measured and predicted values at lower frequencies. We attribute these errors to the breakdown in the wide-window approximation described in the comparison of the measured and predicted values of $\Delta\tilde{\tau}_g$.

Example. We have applied this methodology to recover estimates of the differential phase delay for *S* between EU2 and the Earth from the Kamchatka seismograms (Figure 4.16*a*). Figure 4.16*b* displays the values of the generalized data functionals estimated by our procedure. The estimates of $\delta\tau_p$ are quite consistent for the *SH* and *PSV* observations, with values around 9 s. This agreement between the transverse and vertical component measurements is surprising in the light of the EU2-SNA comparison (Figure 4.15). The vertical-component measurements in the EU2-SNA example display dispersion, achieving the transverse-component values only at the highest frequencies. We infer from this comparison that the dispersion of shear-coupled *PL* must be sensitive to upper-mantle shear structure and that EU2 is a good model for propagation across northern Eurasia.

The value of 9 s is an unexpectedly large differential travel time for direct *S*. Examination of other seismograms for this event reveals similarly large shifts, suggesting that this variation may be due in part to differences between the actual and assumed source centroid time, rather than differences between EU2 (which uses the lower mantle from PREM [Dziewonski and Anderson, 1981]) and the Earth (the centroid time shift for this event is 13.1 s from the origin time reported by the National Earthquake Information Center in their Preliminary Determination of Epicenters (PDE) [Dziewonski *et al.*, 1984]).

SSS isolation filter

Correlation functions. We begin by calculating the autocorrelation of the isolation filter and applying the windowing and filtering operators (Figure 4.17). Since $\tilde{f}(t)$ is localized in the time domain on both components, $C_{\tilde{f}\tilde{f}}(t)$ is concentrated near zero lag. The waveform

fitting of the Gaussian-wavelet approximation produces estimates of ω_i' and σ_i' which match the filtered, windowed autocorrelation function extremely well.

In the next step, we cross-correlate $\tilde{f}(t)$ and $\tilde{s}(t)$. Figure 4.18 illustrates $C_{\tilde{f}\tilde{s}}(t)$, $WC_{\tilde{f}\tilde{s}}(t)$, and $F_i WC_{\tilde{f}\tilde{s}}(t)$; the values of the average phase, group, and amplitude parameters recovered by this methodology (squares) are displayed in Figure 4.19. The values of $\Delta\tilde{\tau}_p$ and $\Delta\tilde{\tau}_g$ predicted from (4.9) and (4.10) are displayed with filled circles. The comparison of the measured and predicted values of $\Delta\tilde{\tau}_p$ and $\Delta\tilde{\tau}_g$ is excellent on the vertical component, agreeing to within 1 s. On the transverse component, the predicted values of $\Delta\tilde{\tau}_p$ have similar dispersion as the measured values; for example, the theory predicts the large excursion at 10 mHz. However, the 10 and 15 mHz measurements differ by several seconds from the predicted values. The comparison of the measured and predicted values of $\Delta\tilde{\tau}_g$ is less satisfactory, with the variation between the observed and predicted values reaching 10 s.

The cross-correlation betwee $\tilde{f}(t)$ and $s(t)$ is illustrated in Figure 4.20, along with $WC_{fs}(t)$ and $F_i WC_{fs}(t)$, and the estimates of $\delta\tau_p$, $\delta\tau_g$, and $\delta\tau_a$ recovered from the waveform-fitting procedure are displayed in Figure 4.21. On the vertical component, the estimation of $\delta\tau_p$ shows little or no dispersion, while the transverse component measurements display some frequency dependence.

Fréchet kernels. As we did in the example for *S*, we may use the expression for the first-order phase perturbation (4.14) to calculate the Fréchet kernels for the *SSS*-wave isolation filters. Figure 4.22 presents the transverse (*a*) and vertical (*b*) component kernels for velocity and density perturbations at four sample frequencies, with the same conventions as Figure 4.14.

The *SH*-polarized kernels indicate that *SSS* is extremely sensitive to shear velocity in the upper mantle and transition zone. In addition, this sensitivity is strongly frequency dependent, due to the interaction of the triplication branches. For example, the emergence of the reflection from the 400 km discontinuity may be observed in the 30 and 40 mHz

kernels. Because of the strong frequency dependence, these four measurements yield four independent constraints on shear-velocity structure.

The *SH*-polarized shear-velocity kernel at 38 mHz becomes positive at the 670 km discontinuity. This positivity reflects the interaction of two traveling-wave branches which are sensitive to the change in velocity across the discontinuity.

Comparison between the *SH* and *PSV* shear-velocity kernels is quite favorable. Although there are some differences between the mode types, the kernels appear to be averaging structure in the same way. In particular, *SSS* does not display any sensitivity to compressional velocity in this frequency range.

We compare the measured estimates of $\delta\tau_s$ to those predicted by integration of the Fréchet kernels in Figure 4.23. The comparison is very favorable, except for the lowest frequency measurement on the transverse component.

Example. We have used the *SSS* isolation filter to recover estimates of the generalized data functionals on the observed seismogram (Figure 4.24*a* and *b*). The estimates of $\delta\tau_p$, $\delta\tau_g$, and $\delta\tau_a$ on the vertical component are very stable. In particular, the measurement of differential phase delay does not display any dispersion, with values ranging between 4 to 6 s. The estimates of $\delta\tau_p$ from the vertical component are very stable as well, hovering around 30 s. However, the estimates of $\delta\tau_g$ and $\delta\tau_a$ display considerably more scatter. The large discrepancy between the *SH* and *PSV* measured times may be understood by examining Figure 4.24*a*, which compares the isolation filter with the observed and synthetic seismograms. While *SSS* is a clear arrival on the vertical-component record, it does not appear as a large amplitude pulse on the transverse-component record. This variation in waveforms may be due to transition-zone heterogeneity.

DISCUSSION

In this Chapter, we have outlined an approach for constructing isolation filters based on the weighting of traveling-wave branches. However, our technique for the

recovery of generalized data functionals does not depend on method of computation. For example, we could have considered an isolation filter calculated from some other convenient algorithm. Our method is hybrid in the sense that the isolation filter may be calculated by any procedure and need not be exact. The complete synthetic seismogram allows us to correct for errors in the isolation filter as well as the presence of interference.

While many ray-theoretical techniques have been developed specifically for the calculation of body-wave seismograms, we have chosen to construct our isolation filters by weighted sums of traveling-wave branches for several reasons. For example, the WKBJ method [Chapman, 1978; Dey-Sakar and Chapman, 1978; Chapman and Orcutt, 1985] is very popular for its efficient and rapid computation of body waves. We investigated the WKBJ algorithm in some detail and determined that it was not adequate for our needs. In particular, we found that the problems in modeling triplication structure, such as the failure to model low-frequency reflections and wave interactions with interfaces, were exacerbated by the number of surface reflections. More importantly, it is difficult to use WKBJ for modeling *PSV* propagation because of the large number of rays which contribute to arrivals such as shear-coupled *PL*. Although some progress has been made in combining reflectivity codes with WKBJ in order to adequately model *PSV* propagation [Baag and Langston, 1985*a*, 1985*b*; Langston and Baag, 1986], considerable work remains to be done.

On the other hand, our approach allows us to account for the complications of *PSV* propagation explicitly. In the example of the *S*-wave isolation filter, we summed all traveling-wave branches with the appropriate group-velocity weighting. However, we could have easily constructed an isolation filter which did not contain any *P*-wave energy by suppressing modes with a low ratio of *S/P* energy. In either case, the construction of the Fréchet kernels includes the effects of *PSV* propagation. Thus, our approach allows us to handle the problems typically associated with polarization.

We have also shown that this parameter-retrieval methodology is applicable to a wide-range of seismic phases. We considered purely ray-like arrivals such as *S* to purely mode-like arrivals such as the fundamental-mode Love and Rayleigh waves, as well as phases which exhibit both ray-like and mode-like behavior such as *SSS*.

MODE-RAY DUALITY

The relationship between modes and rays, usually referred to as mode-ray duality, has received considerable attention over the years. It is generally recognized that normal modes and rays represent two separate, but complementary, representations of the seismogram. Many investigators have tackled the problem of establishing a relationship between modes and rays [Ben Menahem, 1964; Brune, 1964; Anderssen and Cleary, 1974; Anderssen *et al.*, 1975; Gilbert, 1975; Lapwood, 1975; Anderssen, 1977; Woodhouse, 1978; Nolet and Kennett, 1978; Kennett and Woodhouse, 1978; Kennett and Nolet, 1979]. One approach has focused on the properties of high-frequency modes at low-angular order [Anderssen and Cleary, 1974; Anderssen *et al.*, 1975; Gilbert, 1975; Lapwood, 1975; Anderssen, 1977] and has produced relations between the asymptotic eigenfrequency spacing and vertical rays *ScS* and *PKIKP*. Another approach has focused on the properties of high-frequency modes at high-angular order [Perkeris, 1965; Woodhouse, 1978; Nolet and Kennett, 1978; Kennett and Woodhouse, 1978; Kennett and Nolet, 1979] and has led to the development of constructive-interference conditions for the equivalence between modes and rays. For example, Nolet and Kennett [1978] used a stationary-phase argument to show that normal-mode summation will produce pulse-like arrivals when the group velocity is stationary with respect to angular order.

Despite the extensive theoretical interest in the relationship between modes and rays, there has not been much effort to use normal-modes to calculate body-wave arrivals. As far as we know, the isolation filters displayed in Figures 4.5 and 4.7 represent the first efforts to create body-wave arrivals from standing-wave sums.

SUMMARY

In this chapter, we explored the general form of the isolation filter as a sum over traveling-wave branches convolved with the weight coefficients $\tilde{\alpha}_m(t)$. We considered a method for constructing isolation filters for wavegroups based on the summation of traveling-wave branches with weight coefficients designed to "window" about a particular group arrival time and illustrated this approach with S and SSS. We developed expressions for $F_i WC_{\gamma\gamma}(t)$, $F_i WC_{\gamma\bar{\gamma}}(t)$, and $F_i WC_{\gamma s}(t)$ in terms of sums over traveling-wave branches. We have approximated these sums as a single Gaussian wavelet and derived expressions for the average perturbations due to interfering arrivals ($\Delta\tilde{\tau}_p$, $\Delta\tilde{\tau}_g$, $\Delta\tilde{\tau}_a$). We have also derived a simple expression for the first-order phase perturbation due to differences between the reference model and the Earth.

TABLES

TABLE 4.1 PARAMETERS OF THE THE GROUP ARRIVAL WINDOWS - 83/08/17

Phase	τ_o (s)	σ_o^{-1}(s)	Phase vel.	Comments
S	1149.	100.	3. < c < 13.	
SSS	1589.	111.	3. < c < 7.	fundamental mode excluded

FIGURE CAPTIONS

FIGURE 4.1

Transverse and vertical component synthetic seismograms, calculated from the reference model EU2, for the 83/08/17 Kamchakta event (h = 77 km) at the GDSN station KONO (Δ = 63°) Vertical lines mark the group arrival times of the phases discussed in the text and were picked by eye on the transverse component.

FIGURE 4.2

Frequency *vs.* angular order (ω-l) for toroidal (a) and spheroidal (b) modes, calculated from the model EU2 with the $\tilde{\alpha}_m(\omega)$ for the fundamental-mode surface waves. The branch running from the lower right-hand corner to the upper right-hand corner of each diagram corresponds to fundamental mode (n = 0, where n is the radial order number). The size of the circles indicates the amplitude coefficients of the isolation filter, from zero to one. In the case of the fundamental-mode surface waves, only the coefficients of the n = 0 branch are non-zero. Figure 4.3 displays the isolation filters calculated from the convolution of these amplitude coefficients with the traveling-wave branches.

FIGURE 4.3

Transverse and vertical component seismograms, corresponding to the isolation filter $\tilde{f}(t)$ and the complete synthetic $\tilde{s}(t)$ calculated from the model EU2. An isolation filter for the fundamental-mode surface wave is calculated by summing all the normal modes for the traveling-wave branch n = 0 with $\tilde{\alpha}_0(\omega)$ = 1.

FIGURE 4.4

Frequency *vs.* angular order (ω-l) for toroidal (a) and spheroidal (b) modes, calculated from the model EU2 with the $\tilde{\alpha}_m(\omega)$ for S. The amplitude coefficients were calculated from (4.2) with parameters t_o = 1149 s and σ_o^{-1} = 100 s and a phase-velocity cutoff of 13 km/s. Both the toroidal and spheroidal coefficients form a wedge between phase velocities of 7 and 11 km/s. The highest amplitudes on the toroidal-mode diagram occur in a well-defined region, with a smoothly-scalloped transition to lower amplitudes. The scalloping is due to the presence of discontinuities in EU2. The structure of the spheroidal-mode diagram appears more chaotic, which high and low amplitudes mingled within the wedge. However, the low-amplitude modes align along cross-cutting branches, which correspond to P-waves in the upper mantle. Figure 4.5 displays the isolation filters calculated from these amplitude coefficients.

FIGURE 4.5

Transverse and vertical component seismograms, corresponding to the isolation filter $\tilde{f}(t)$ and the complete synthetic $\tilde{s}(t)$ calculated from the model EU2. The isolation filter for *S* on the transverse component is quite good, reproducing the amplitude and phase of the arrival on the complete synthetic. The vertical component filter does not form a single, well-defined arrival; instead, it reproduces the low-amplitude wavetrain on the complete synthetic.

FIGURE 4.6

Frequency *vs.* angular order (ω-l) for toroidal (a) and spheroidal (b) modes, calculated from the model EU2 with the the $\tilde{\alpha}_m(\omega)$ for *SSS*. The 1st through 8th higher-mode branches dominate the isolation filter for *SSS*, with an excellent correspondence between toroidal and spheroidal modes. The group arrival time was chosen to be 1589 s and the window half-width was 111 s. We removed any fundamental-mode energy which satisfied this criteria and applied a phase-velocity cutoff of 7 km/s. Figure 4.7 illustrates the isolation filters constructed from these amplitude coefficients.

FIGURE 4.7

Transverse and vertical component seismograms, corresponding to the isolation filter $\tilde{f}(t)$ and the complete synthetic $\tilde{s}(t)$ calculated from the model EU2. The isolation filters for *SSS* are quite good, reproducing the amplitude and phase of the arrival on the complete synthetic. Both the transverse and vertical component filters contain periodic arrivals which correspond to edge effects introduced by the phase-velocity cutoff.

FIGURE 4.8

The broad-band autocorrelation of the *S*-wave isolation filter for the transverse (left) and vertical (right) components. $C_{ff}(t)$ (solid line) is a symmetric function, centered at zero lag. The dashed line indicates the best-fitting Gaussian wavelet. Values of ω_f and σ_f are 33 and 9 mHz for the transverse component and 28 and 14 mHz for the vertical component respectively. Because the *SH*-polarized isolation filter is localized in the time domain, its autocorrelation function is concentrated near zero lag and the Gaussian-wavelet approximation provides a good fit near the peak. However, the *PSV*-polarized isolation filter is not localized in the time domain and consequently its autocorrelation is not concentrated near zero lag. $C_{ff}(t)$ contains large-amplitude peaks several cycles away from its maximum and the Gaussian-wavelet approximation cannot model this behavior.

FIGURE 4.9

$C_{ff}(t)$, $W\,C_{ff}(t)$, and $F_i W\,C_{ff}(t)$ for the transverse (left) and vertical (right) components at several center frequencies (1 = 35; 2 = 25; 3 = 15 mHz). The autocorrelation functions are symmetric about zero lag, but contain cross-term

contributions due to branch-branch interactions. The windowing operator reduces the cross-term contribution to the autocorrelation, improving the Gaussian-wavelet approximation. The dashed line (which overlays the solid line) indicates the model obtained by fitting equation (3.6) to the filtered, windowed autocorrelation functions.

FIGURE 4.10

$C_{\overline{fs}}(t)$, $WC_{\overline{fs}}(t)$, and $F_iWC_{\overline{fs}}(t)$ for the transverse (left) and vertical (right) components at several center frequencies (1 = 35; 2 = 25; 3 = 15 mHz). The windowing operator reduces the cross-term contribution to the complete synthetic cross-correlation, improving the Gaussian-wavelet approximation. The dashed line indicates the model obtained by fitting equation (3.24) to the filtered, windowed cross-correlation functions. The estimates of the parameters $\Delta\tilde{\tau}_p$, $\Delta\tilde{\tau}_g$, and $\Delta\tilde{\tau}_a$ recovered from the waveform-fitting procedure are displayed in Figure 4.11.

FIGURE 4.11

Estimates of the parameters $\Delta\tilde{\tau}_p$, $\Delta\tilde{\tau}_g$, and $\Delta\tilde{\tau}_a$ recovered from the waveform fitting of $F_iWC_{\overline{fs}}(t)$ (squares) for the transverse (top) and vertical (bottom) component seismograms and the S isolation filter. The values of these parameters are near zero, indicating that most of the influence of interference has been removed by the windowing operator. The filled circles indicate the values of $\Delta\tilde{\tau}_p$ and $\Delta\tilde{\tau}_g$ predicted from the analytic expressions (4.9) and (4.10). The comparison between the measured and predicted values of $\Delta\tilde{\tau}_p$ is within a second or so on both the transverse and vertical components, except at the very lowest frequency. While the predicted values of $\Delta\tilde{\tau}_g$ on the transverse component are within a few seconds of the measured values, with the greatest error occurring at the lowest frequency, the vertical-component results are puzzling. Although the measured values of $\Delta\tilde{\tau}_g$ mimic the dispersion of the predicted values, they are offset by 12 to 15 s in the range from 10 to 20 mHz. This offset is due to the breakdown of the wind-window approximation.

FIGURE 4.12

$C_{fs}(t)$, $WC_{fs}(t)$, and $F_iWC_{fs}(t)$ for several center frequencies. The windowing operator reduces the cross-term contribution to the observed cross-correlation, improving the Gaussian-wavelet approximation. The dashed line indicates the model obtained by fitting equation (3.34) to the filtered, windowed cross-correlation functions. The estimates of $\delta\tau_p$, $\delta\tau_g$, and $\delta\tau_a$ recovered from the waveform fitting are displayed in Figure 4.13.

FIGURE 4.13

Estimates of the parameters $\delta\tau_p$, $\delta\tau_g$, and $\delta\tau_a$ recovered from the waveform fitting of $F_iWC_{fs}(t)$ (squares) for the transverse (top) and vertical (bottom) component seismograms and the S isolation filter. The solid line indicates the actual values calculated for the EU2/SNA comparison. On the transverse component, the S-wave behaves as a classical body wave, *i.e.*, it is not dispersed, and the recovered parameters match the true values. The transverse-component arrival is dispersed,

however, and the recovered phase delays are a smoothly-varying function of frequency.

FIGURE 4.14

Transverse (*a*) and vertical (*b*) component Fréchet kernels calculated for the first-order perturbation in phase delay (4.14) at four sample frequencies. The horizontal dashed lines indicate the Moho, the 400 km discontinuity, and the 670 km discontinuity. The solid line is the shear-velocity kernel; the shorter dashed line is the compressional-velocity kernel. The longer dashed line marks the density kernel. (*a*) The *SH*-polarized shear-velocity kernels demonstrate that the *S*-wave travel time represents a smoothed average of lower-mantle structure. This averaging becomes localized near the bottoming depth of *S* with increasing frequency, corresponding to the high-frequency turning-point singularity. The shear kernels in this figure become positive in several places, due to the interaction of interfering wavegroups. (*b*) The *PSV*-polarized kernels are an interesting comparison to the *SH* figure. The shear-velocity kernels at 30 and 40 mHz kernels are quite similar for toroidal and spheroidal modes, while the 10 and 20 mHz kernels are very different. The compressional-velocity kernels indicate that the measurements of phase delay made with a *S*-wave isolation filter contain information about the *P*-wave velocity in the upper mantle.

FIGURE 4.15

Comparison between the measurements of differential phase delay (squares) and the values predicted from the integration of the Fréchet kernels with the velocity perturbations between EU2 and SNA (filled circles). The comparison for the *SH* measurement is very good, with phase estimates recovered to a second or so, except for the lowest frequency. The *PSV* comparison is good for frequencies above 25 mHz, but shows some "tears" between the measured and predicted values at lower frequencies. We attribute these errors to the breakdown in the wide-window approximation described in the comparison of the measured and predicted values of $\Delta\tilde{\tau}_g$.

FIGURE 4.16

Application of this methodology to recover estimates of the differential phase delay for *S* between EU2 and the Earth from the Kamchatka seismograms. (*a*) Comparison between the $\tilde{f}(t)$ and $\tilde{s}(t)$ (upper pair) and $\tilde{f}(t)$ and $s(t)$ (lower pair) for the transverse (left) and vertical (right) component records. (*b*) Estimated values of the generalized data functionals recovered by our procedure. The estimates of $\delta\tau_p$ are quite consistent for the *SH* and *PSV* observations, with values around 9 s.

FIGURE 4.17

$C_{ff}(t)$, $WC_{ff}(t)$, and $F_iWC_{ff}(t)$ of *SSS* for the transverse (left) and vertical (right) components at several center frequencies (1 = 35; 2 = 25; 3 = 15). The autocorrelation functions are symmetric about zero lag, but contain cross-term contributions due to branch-branch interactions. The windowing operator reduces the cross-term contribution to the autocorrelation, improving the Gaussian-wavelet

approximation. The dashed line (which overlays the solid line) indicates the model obtained by fitting equation (3.6) to the filtered, windowed autocorrelation functions.

FIGURE 4.18

$C_{\overline{TS}}(t)$, $WC_{\overline{TS}}(t)$, and $F_iWC_{\overline{TS}}(t)$ for the transverse (left) and vertical (right) components at several center frequencies (1 = 35; 2 = 25; 3 = 15). The windowing operator reduces the cross-term contribution to the complete synthetic cross-correlation, improving the Gaussian-wavelet approximation. The dashed line indicates the model obtained by fitting equation (3.24) to the filtered, windowed cross-correlation functions. The estimates of the parameters $\Delta\tilde{\tau}_p$, $\Delta\tilde{\tau}_g$, and $\Delta\tilde{\tau}_a$ recovered from the waveform fitting are displayed in Figure 4.19.

FIGURE 4.19

Estimates of the parameters $\Delta\tilde{\tau}_p$, $\Delta\tilde{\tau}_g$, and $\Delta\tilde{\tau}_a$ recovered from the waveform fitting of $F_iWC_{\overline{TS}}(t)$ (squares) for the transverse (top) and vertical (bottom) component seismograms and the *SSS* isolation filter. The values of these parameters are near zero, indicating that most of the influence of interference has been removed by the windowing operator. The filled circles indicate the values of $\Delta\tilde{\tau}_p$ and $\Delta\tilde{\tau}_g$ predicted from the analytic expressions (4.9) and (4.10). The comparison of the measured and predicted values of $\Delta\tilde{\tau}_p$ and $\Delta\tilde{\tau}_g$ is excellent on the vertical component, agreeing to within 1 s. On the transverse component, the predicted values of $\Delta\tilde{\tau}_p$ have similar dispersion as the measured values; for example, the theory predicts the large excursion at 10 mHz. However, the 10 and 15 mHz measurements differ by several seconds from the predicted values. The comparison of the measured and predicted values of $\Delta\tilde{\tau}_g$ is less satisfactory, with the variation between the observed and predicted values reaching 10 s.

FIGURE 4.20

$C_{TS}(t)$, $WC_{TS}(t)$, and $F_iWC_{TS}(t)$ for several center frequencies. The windowing operator reduces the cross-term contribution to the observed cross-correlation, improving the Gaussian-wavelet approximation. The dashed line indicates the model obtained by fitting equation (3.34) to the filtered, windowed cross-correlation functions. The estimates of $\delta\tau_p$, $\delta\tau_g$, and $\delta\tau_a$ recovered from the waveform fitting are displayed in Figure 4.21.

FIGURE 4.21

Estimates of the parameters $\delta\tau_p$, $\delta\tau_g$, and $\delta\tau_a$ recovered from the waveform fitting of $F_iWC_{TS}(t)$ (squares) for the transverse (top) and vertical (bottom) component seismograms and the *SSS* isolation filter.

FIGURE 4.22

Transverse (*a*) and vertical (*b*) component Fréchet kernels calculated for the first-order perturbation in phase delay (4.14) at four sample frequencies. The horizontal

dashed lines indicate the Moho, the 400 km discontinuity, and the 670 km discontinuity. The solid line is the shear-velocity kernel; the shorter dashed line is the compressional-velocity kernel. The longer dashed line marks the density kernel. The *SH*-polarized kernels indicate that *SSS* is extremely sensitive to shear velocity in the upper mantle and transition zone. In addition, this sensitivity is strongly frequency dependent, due to the interaction of the triplication branches. Comparison between the *SH* and *PSV* shear-velocity kernels is quite favorable. Although there are some differences between the mode types, the kernels appear to average structure in the same way. In particular, *SSS* does not display any sensitivity to compressional velocity in this frequency range.

FIGURE 4.23

Comparison between the measurements of differential phase delay (squares) and the values predicted from the integration of the Fréchet kernels with the velocity perturbations between EU2 and SNA (filled circles). The comparison is very favorable, except for the lowest frequency measurement on the transverse component.

FIGURE 4.24

Application of this methodology to recover estimates of the differential phase delay for *SSS* between EU2 and the Earth from the Kamchatka seismograms. (*a*) Comparison between the $\tilde{f}(t)$ and $\tilde{s}(t)$ (upper pair) and $\tilde{f}(t)$ and $s(t)$ (lower pair) for the transverse (left) and vertical (right) component records. (*b*) Estimated values of the generalized data functionals recovered by our procedure. The estimates of $\delta\tau_p$, $\delta\tau_g$, and $\delta\tau_a$ on the vertical component are very stable. In particular, the measurement of differential phase delay does not display any dispersion, with values ranging between 4 to 6 s. The estimates of $\delta\tau_p$ from the vertical component are very stable as well, hovering around 30 s. However, the estimates of $\delta\tau_g$ and $\delta\tau_a$ display considerably more scatter. The large discrepancy between the *SH* and *PSV* measured times may be understood by examining Figure 4.24*a*, which compares the isolation filter with the observed and synthetic seismograms. While *SSS* is a clear arrival on the vertical-component record, it does not appear as a large amplitude pulse on the transverse-component record. This variation in waveforms may be due to transition-zone heterogeneity.

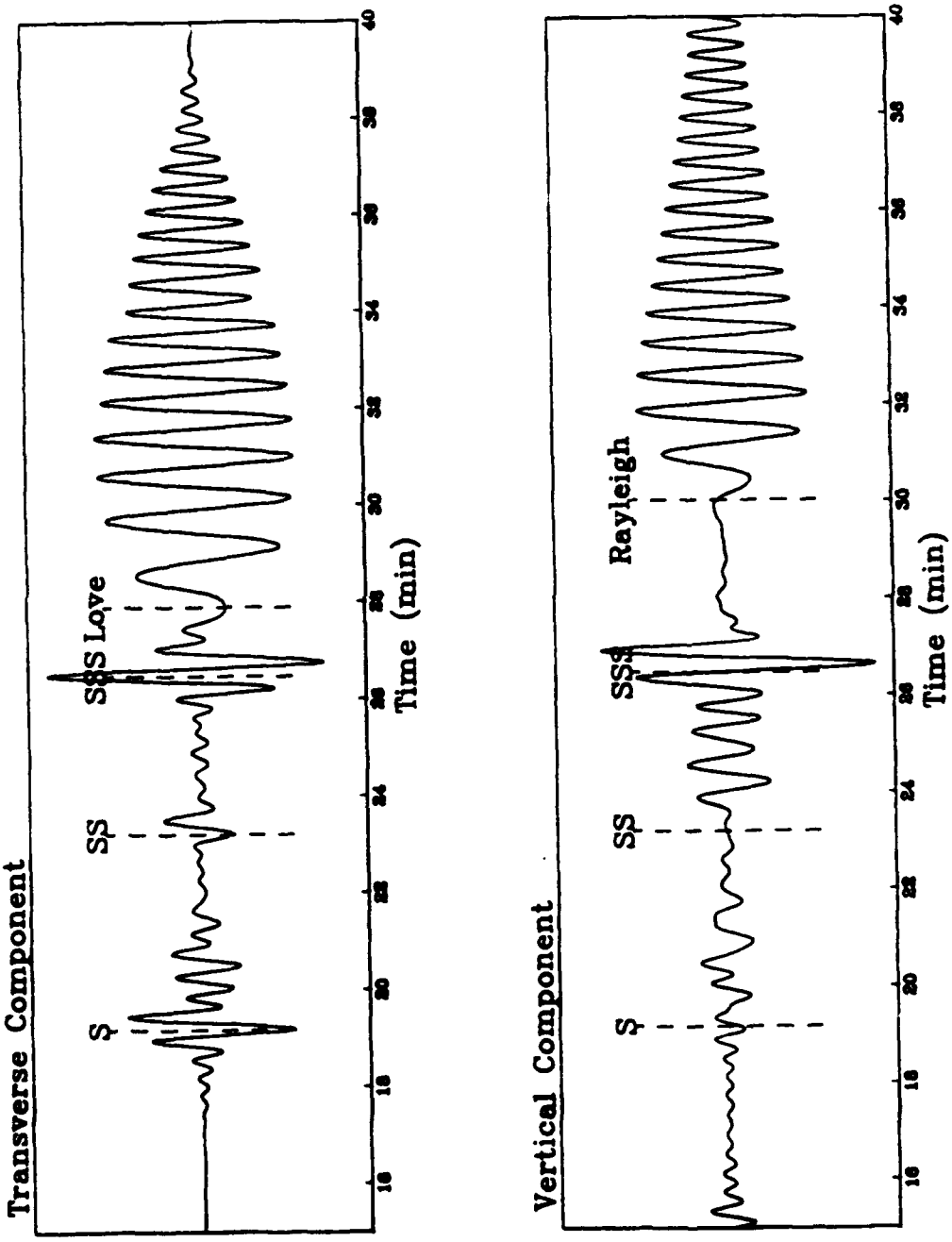

Transverse Component
S
SS
SSS Love
Time (min)
Vertical Component
S
SS
SSS
Rayleigh
Time (min)

Figure 4.1

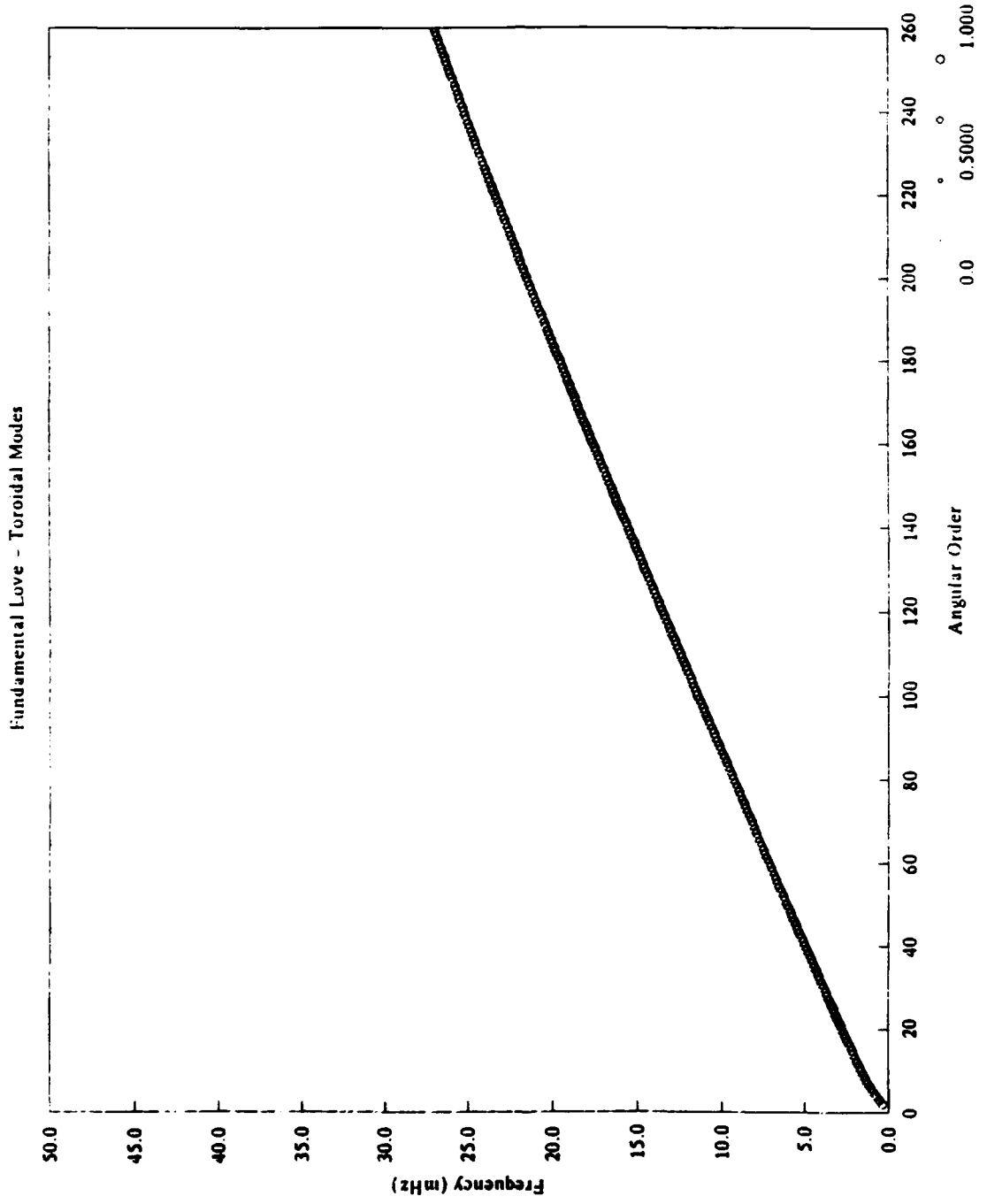

Fundamental Love - Toroidal Modes
Frequency (mHz)
50.0
45.0
40.0
35.0
30.0
25.0
20.0
15.0
10.0
5.0
0.0
0
20
40
60
80
100
120
140
160
180
200
220
240
260
0.0
0.5000
1.000
Angular Order

Figure 4.2a

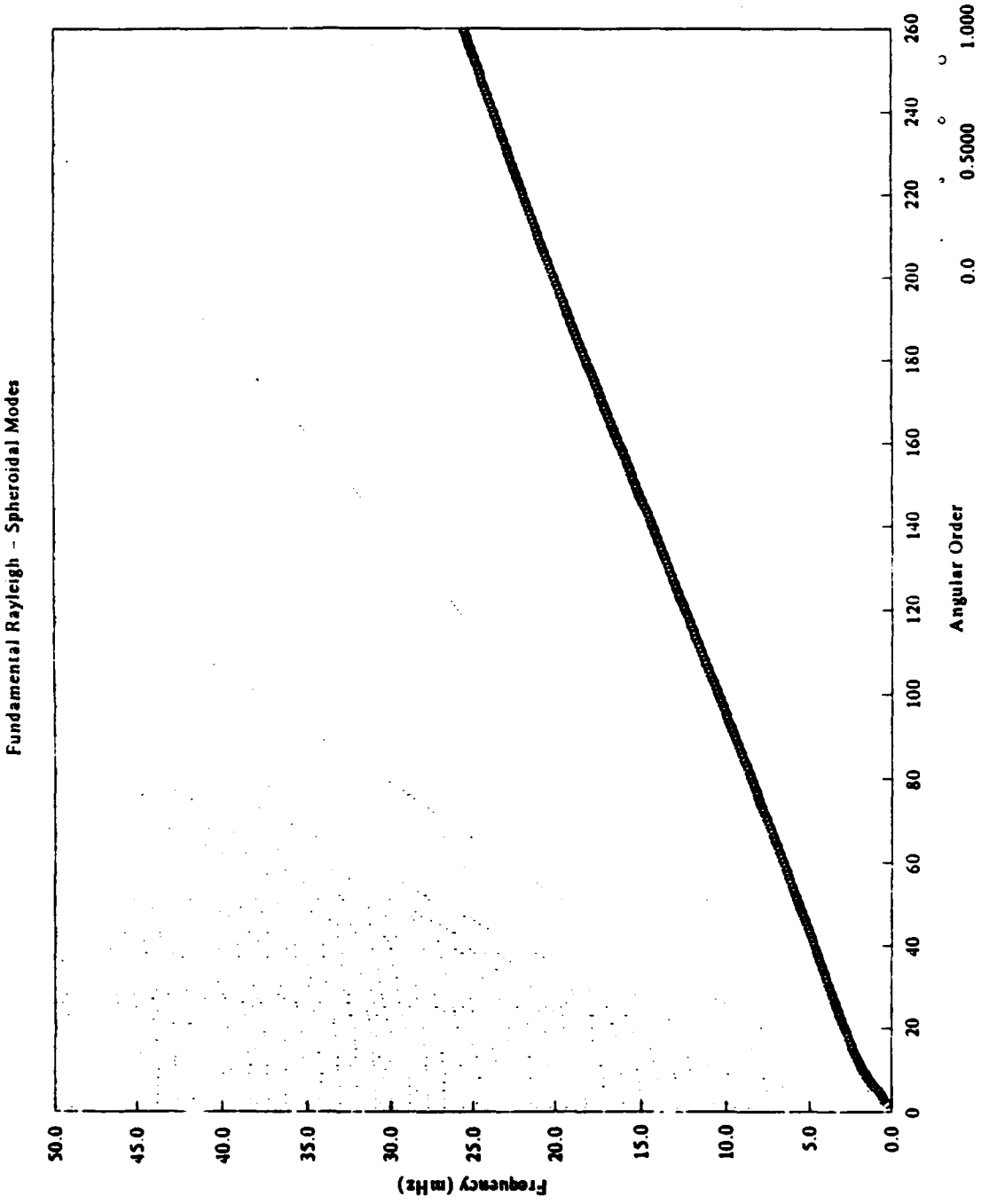

Fundamental Rayleigh - Spheroidal Modes
Frequency (mHz)
50.0
45.0
40.0
35.0
30.0
25.0
20.0
15.0
10.0
5.0
0.0
Angular Order
0
20
40
60
80
100
120
140
160
180
200
220
240
260
0.0
0.5000
1.000

Figure 4.2b

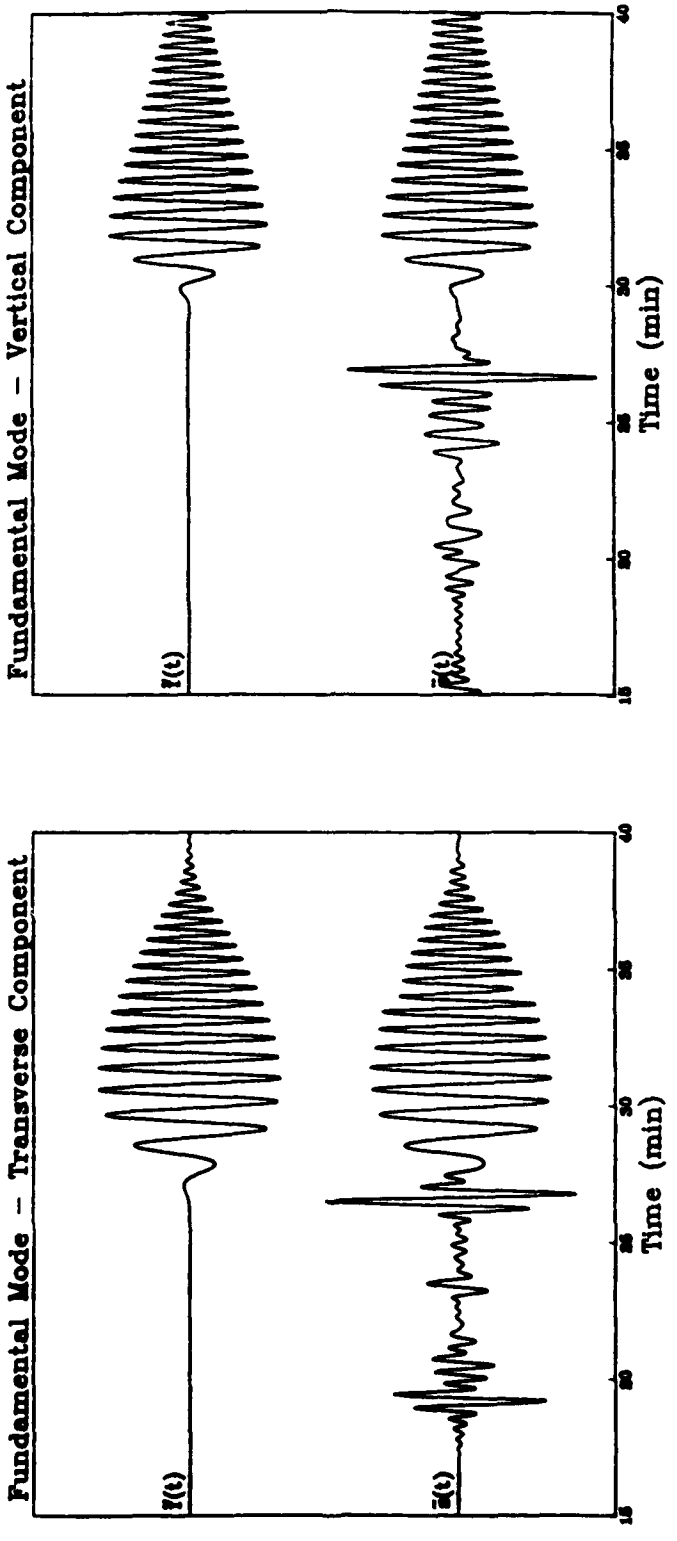

Fundamental Mode – Vertical Component
Z(t)
Time (min)
Fundamental Mode – Transverse Component
R(t)
Time (min)

Figure 4.3

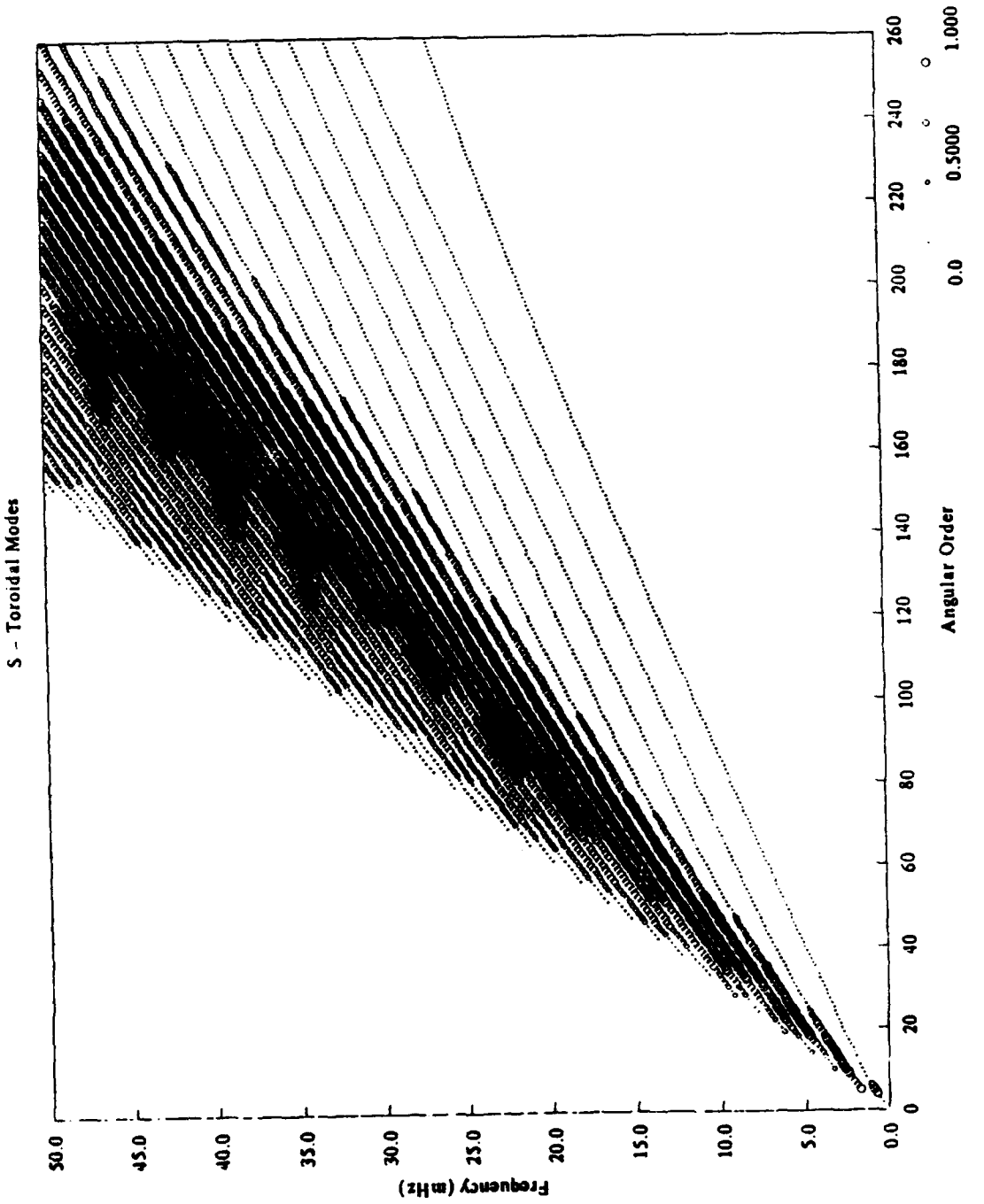

S - Toroidal Modes
Frequency (mHz)
50.0
45.0
40.0
35.0
30.0
25.0
20.0
15.0
10.0
5.0
0.0
0
20
40
60
80
100
120
140
160
180
200
220
240
260
Angular Order
0.0
0.5000
1.000

Figure 4.4a

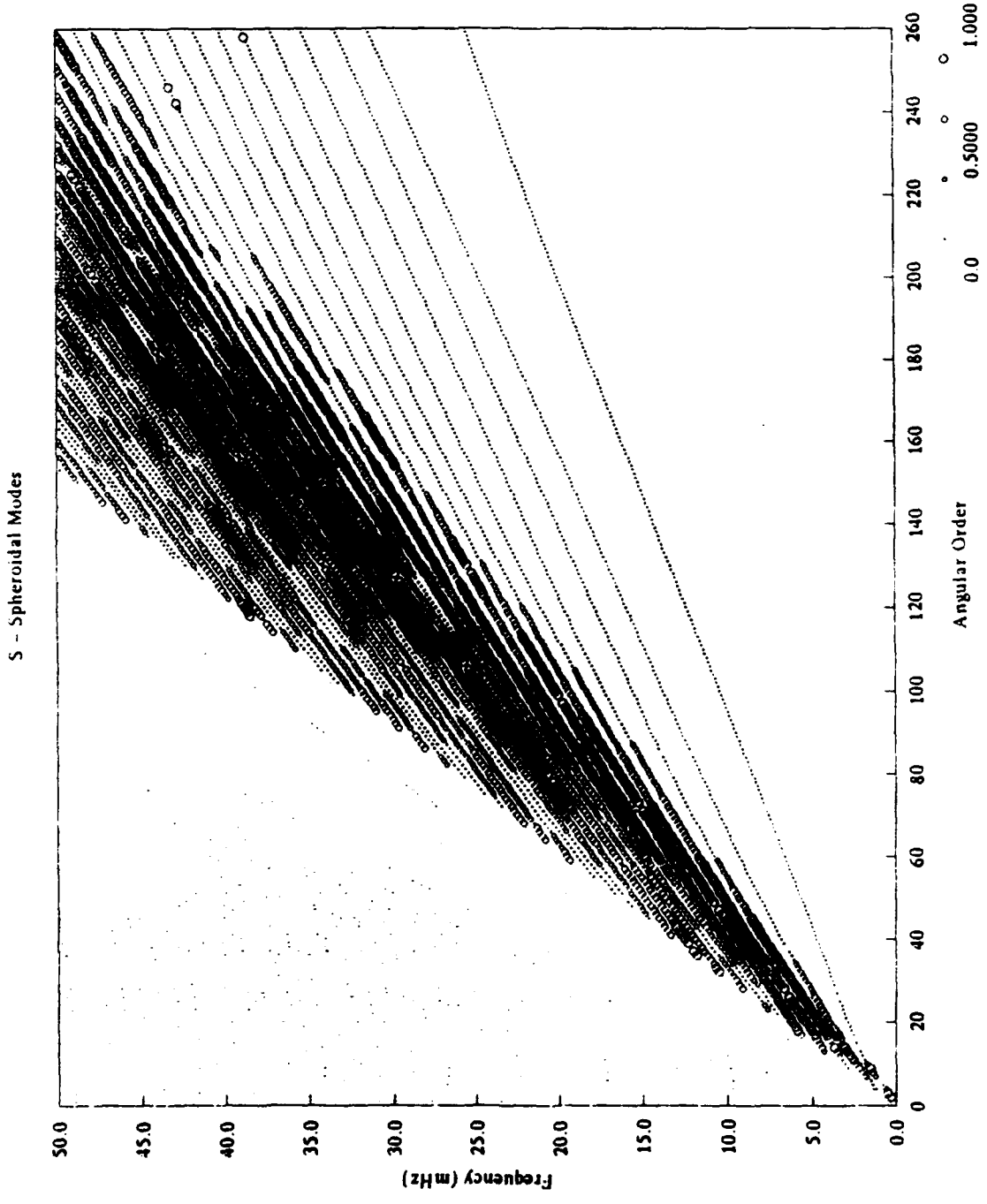

S - Spheroidal Modes
Frequency (mHz)
50.0
45.0
40.0
35.0
30.0
25.0
20.0
15.0
10.0
5.0
0.0
0
20
40
60
80
100
120
140
160
180
200
220
240
260
Angular Order
0.0
0.5000
1.000

Figure 4.4b

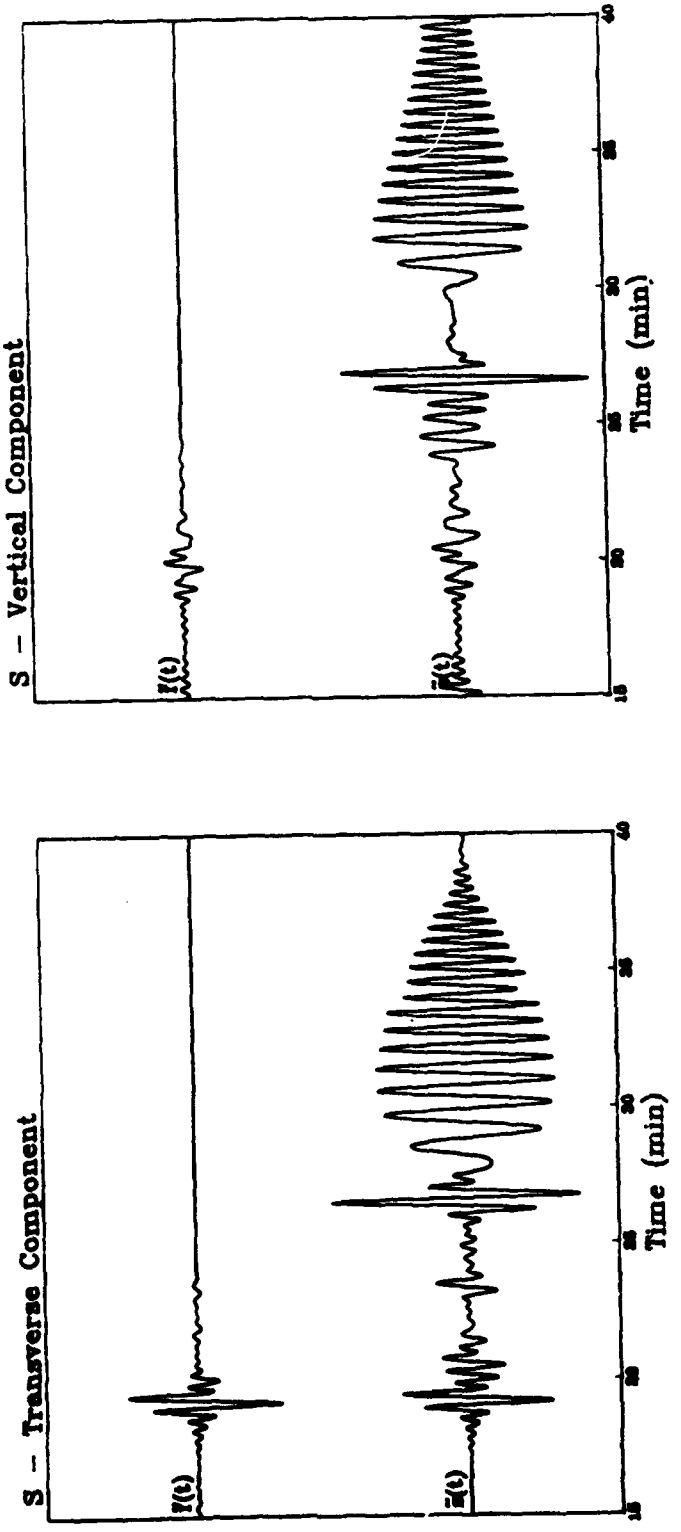

S – Vertical Component
S – Transverse Component
Time (min)
Time (min)

Figure 4.5

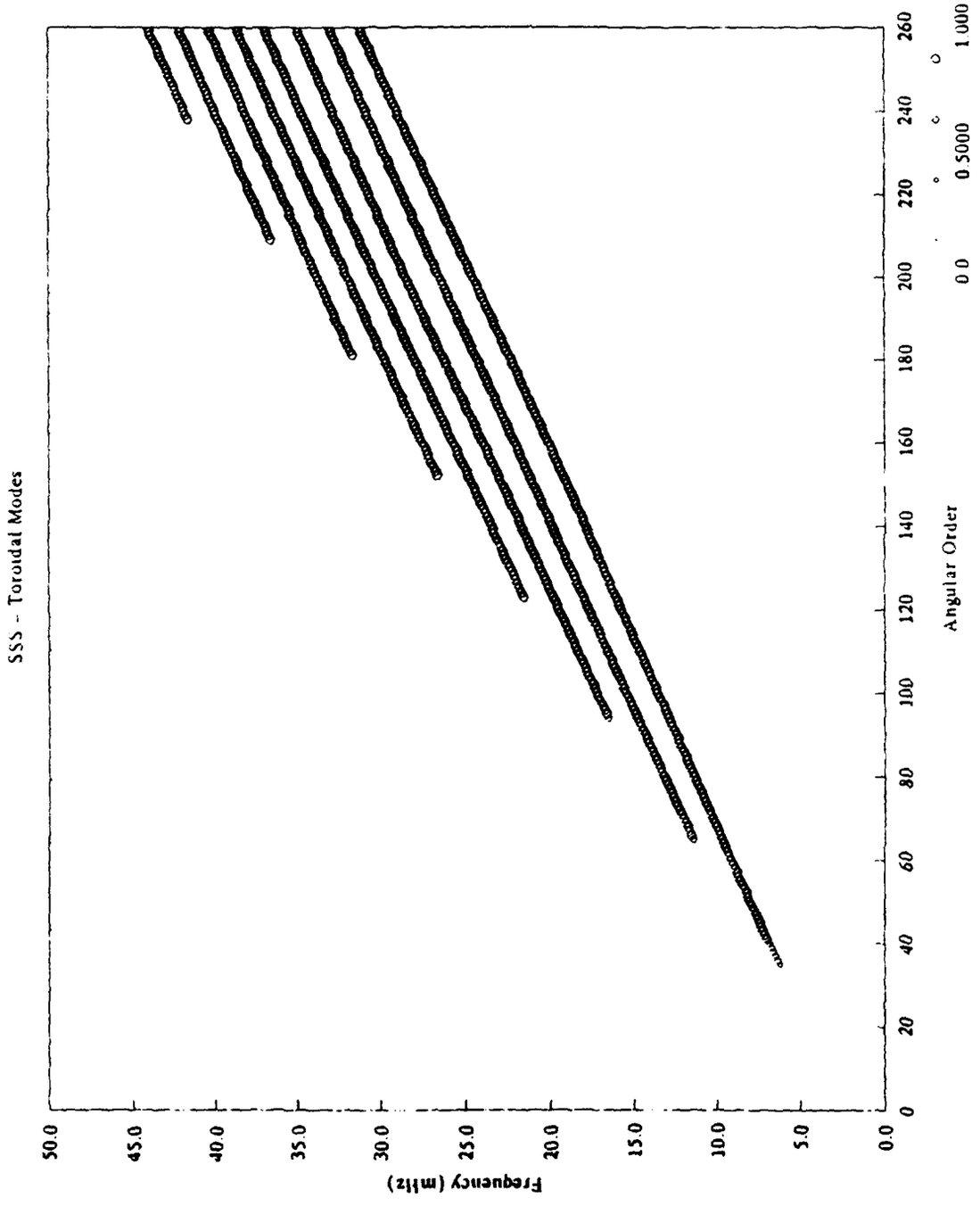

SSS - Toroidal Modes
Frequency (mHz)
50.0
45.0
40.0
35.0
30.0
25.0
20.0
15.0
10.0
5.0
0.0
0
20
40
60
80
100
120
140
160
180
200
220
240
260
Angular Order
0.0
0.5000
1.000

Figure 4.6a

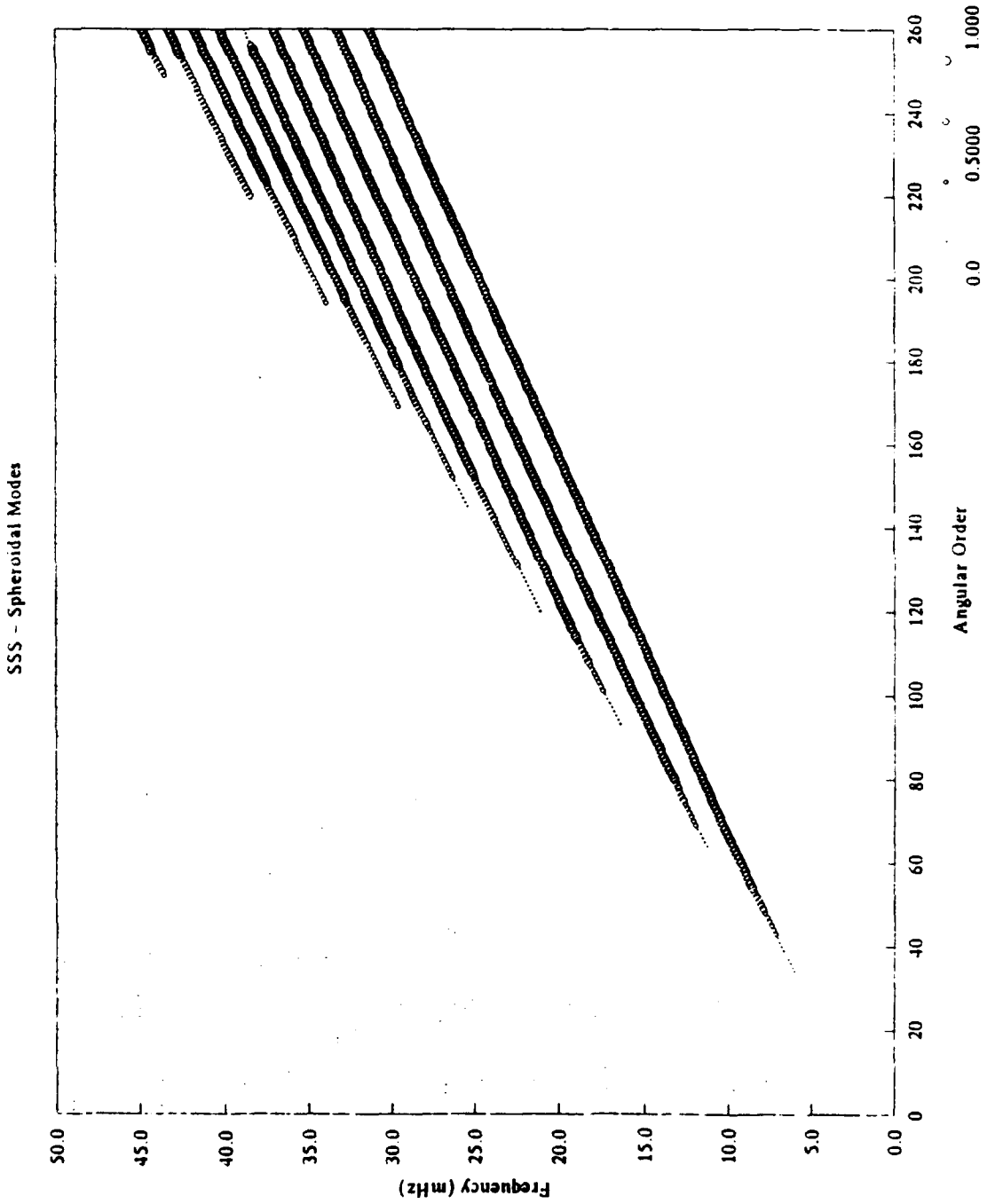

SSS - Spheroidal Modes
Frequency (mHz)
50.0
45.0
40.0
35.0
30.0
25.0
20.0
15.0
10.0
5.0
0.0
0
20
40
60
80
100
120
140
160
180
200
220
240
260
Angular Order
0.0
0.5000
1.000

Figure 4.6b

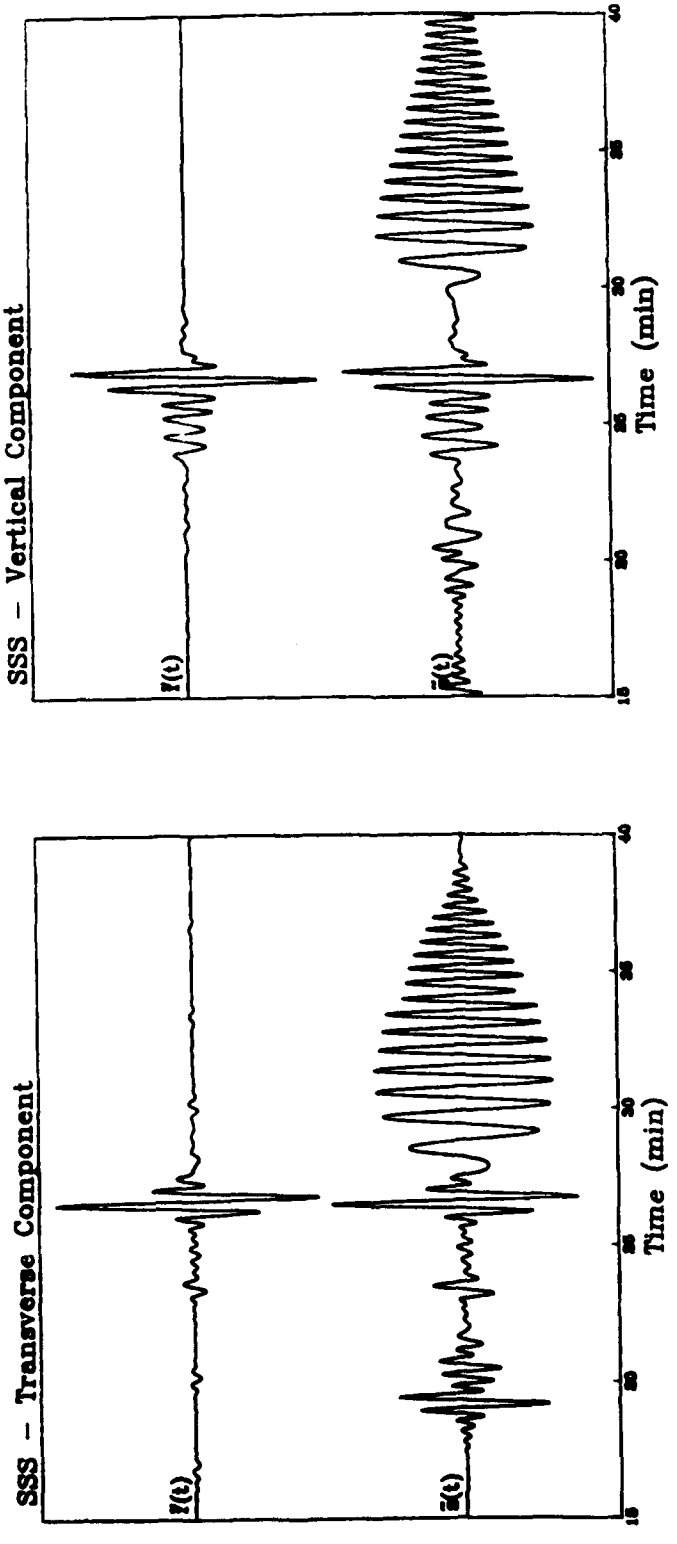

SSS – Vertical Component
SSS – Transverse Component
Time (min)
Time (min)

Figure 4.7

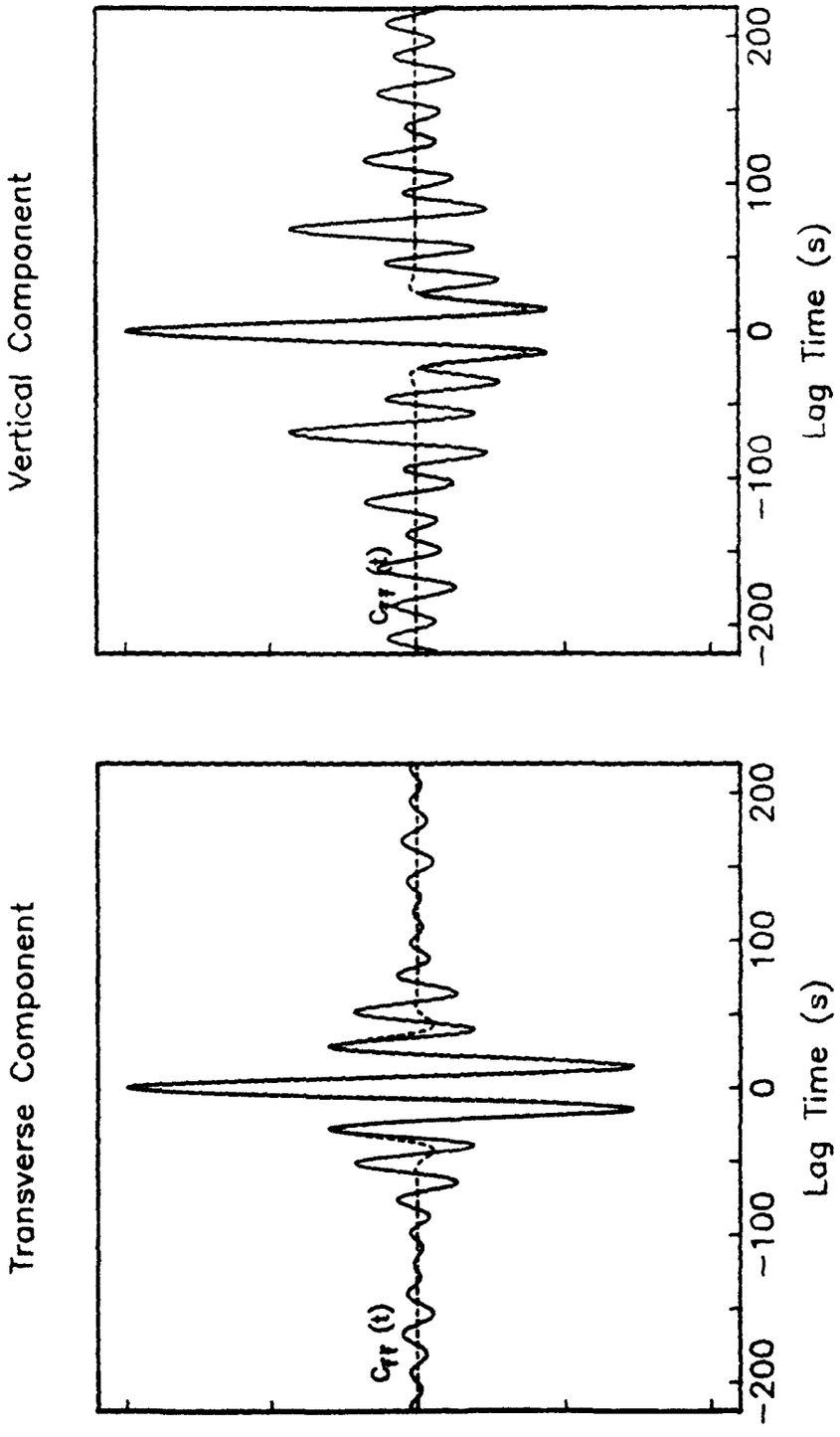

Vertical Component
Transverse Component
Lag Time (s)
Lag Time (s)
-200
-100
0
100
200
-200
-100
0
100
200

Figure 4.8

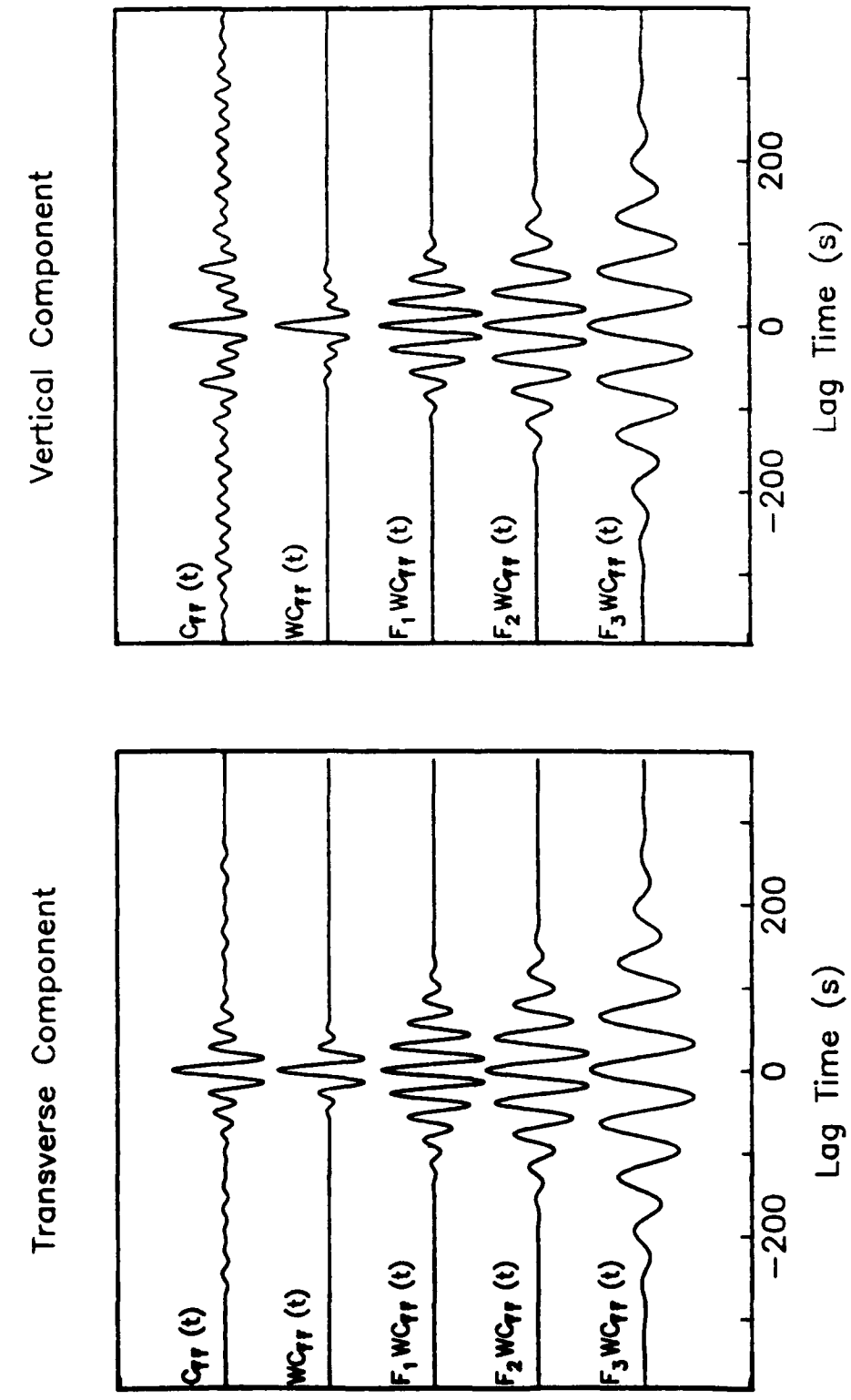

Vertical Component
Transverse Component
C_{TT} (t)
WC_{TT} (t)
F_1WC_{TT} (t)
F_2WC_{TT} (t)
F_3WC_{TT} (t)
-200
0
200
Lag Time (s)

Figure 4.9

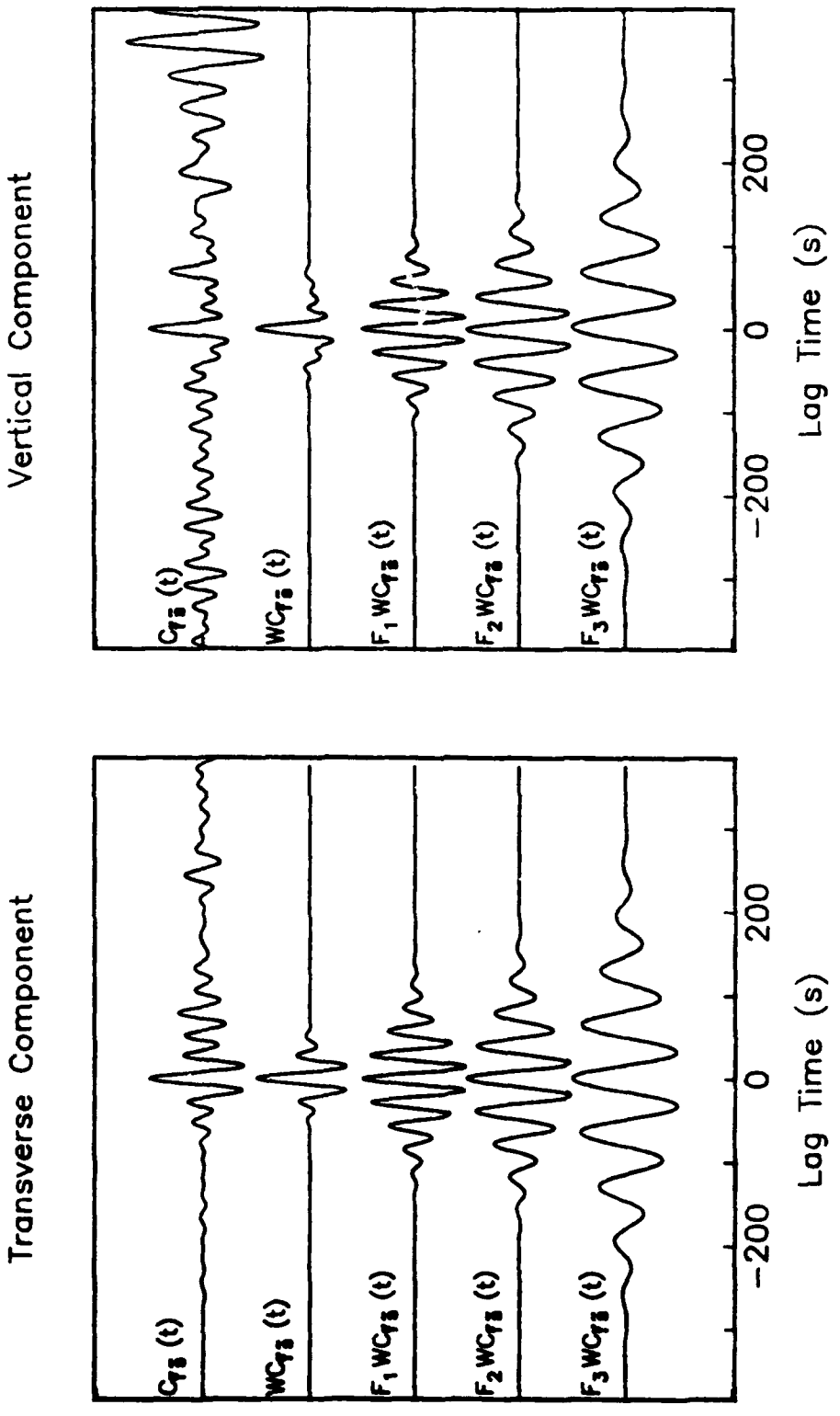

Vertical Component
Transverse Component
C_{TS} (t)
WC_{TS} (t)
F_1WC_{TS} (t)
F_2WC_{TS} (t)
F_3WC_{TS} (t)
-200
0
200
Lag Time (s)

Figure 4.10

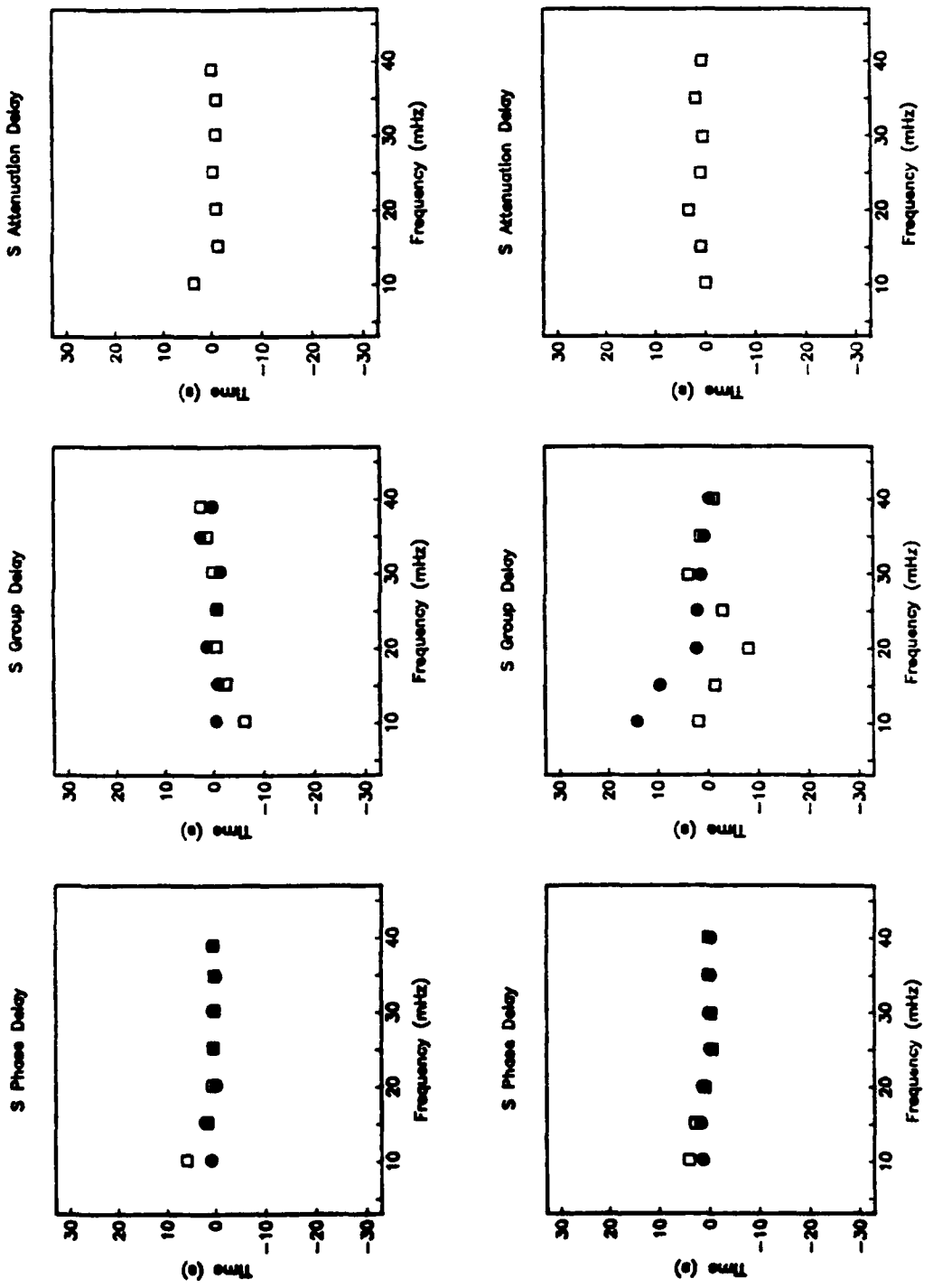

S Attenuation Delay
S Group Delay
S Phase Delay
Time (s)
Frequency (mHz)
30
20
10
0
-10
-20
-30
10
20
30
40

Figure 4.11

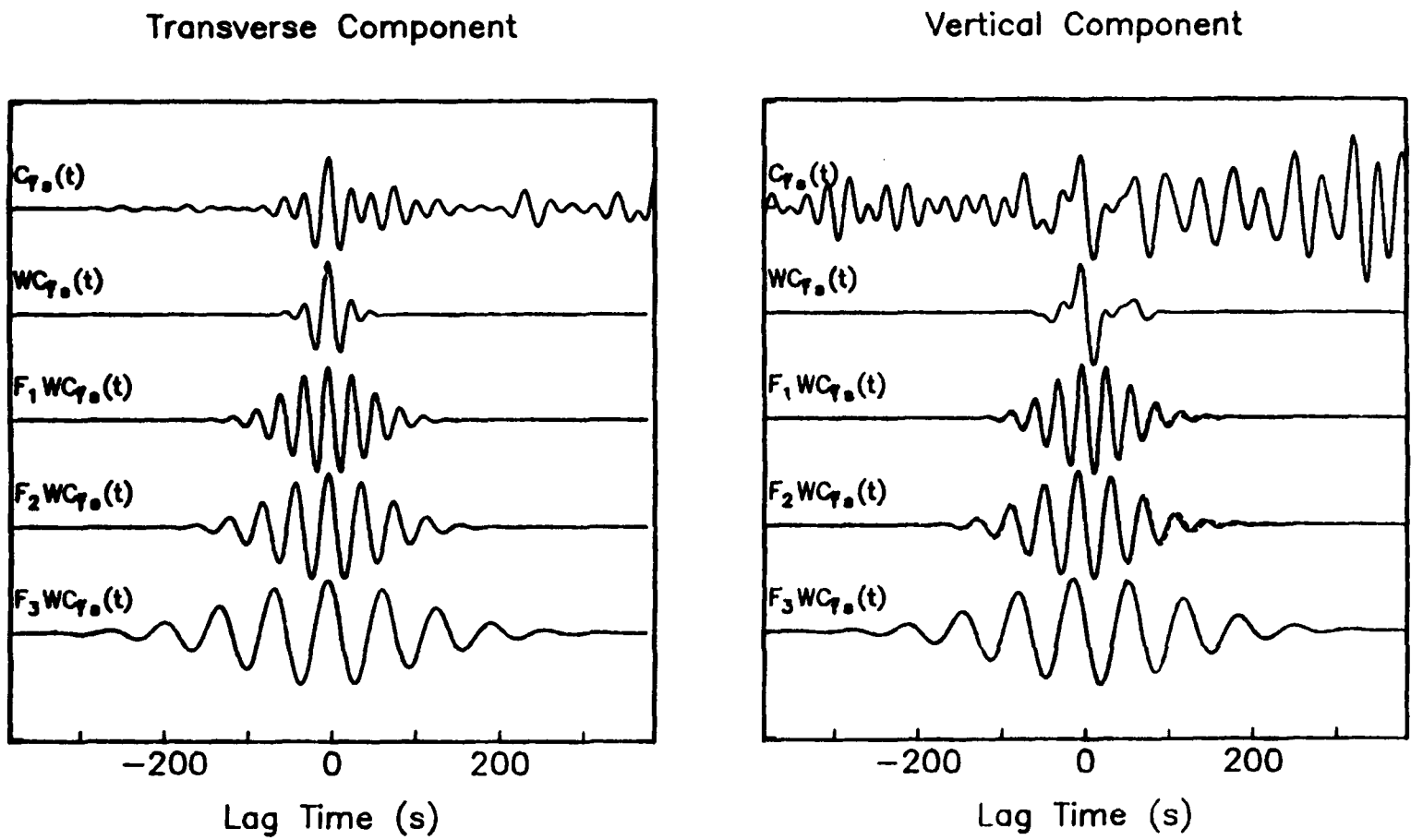

Transverse Component
Vertical Component
$C_{T_0}(t)$
$WC_{T_0}(t)$
$F_1WC_{T_0}(t)$
$F_2WC_{T_0}(t)$
$F_3WC_{T_0}(t)$
−200
0
200
Lag Time (s)

Figure 4.12

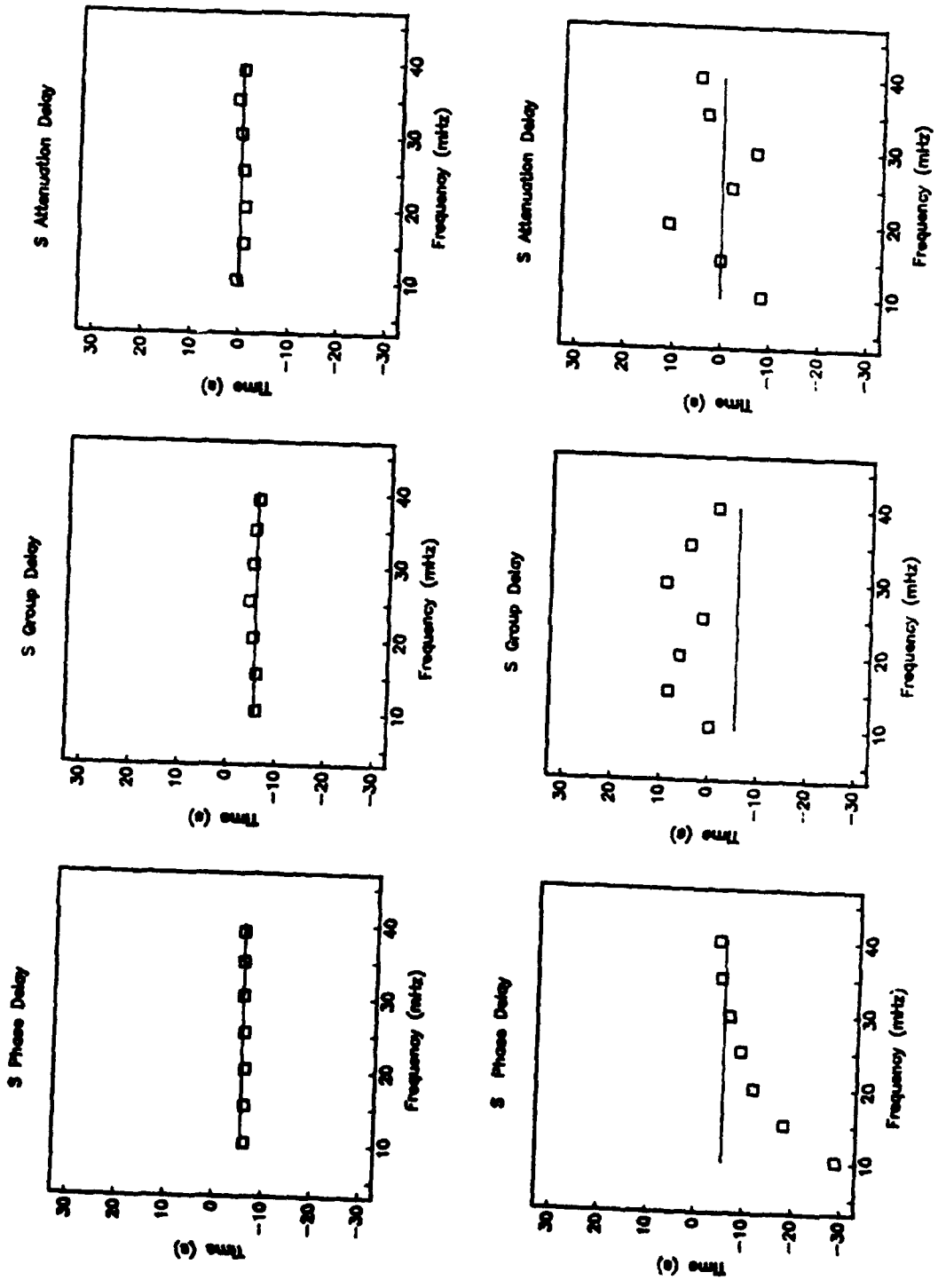

S Attenuation Delay
S Group Delay
S Phase Delay
Time (s)
Frequency (mHz)
30
20
10
0
-10
-20
-30
10
20
30
40

Figure 4.13

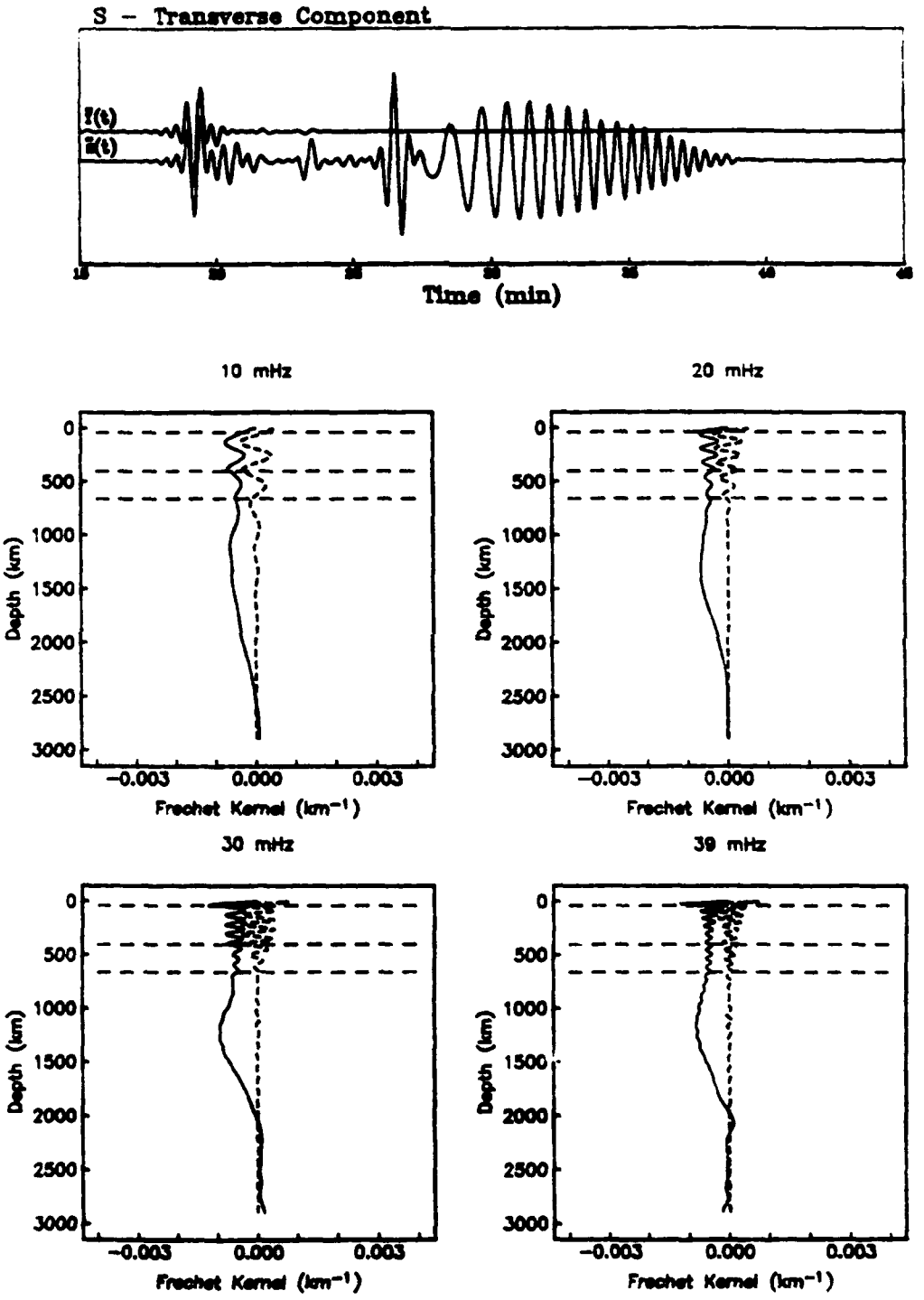

S – Transverse Component
Time (min)
10 mHz
20 mHz
30 mHz
39 mHz
Depth (km)
Frechet Kernel (km⁻¹)
0
500
1000
1500
2000
2500
3000
-0.003
0.000
0.003

Figure 4.14a

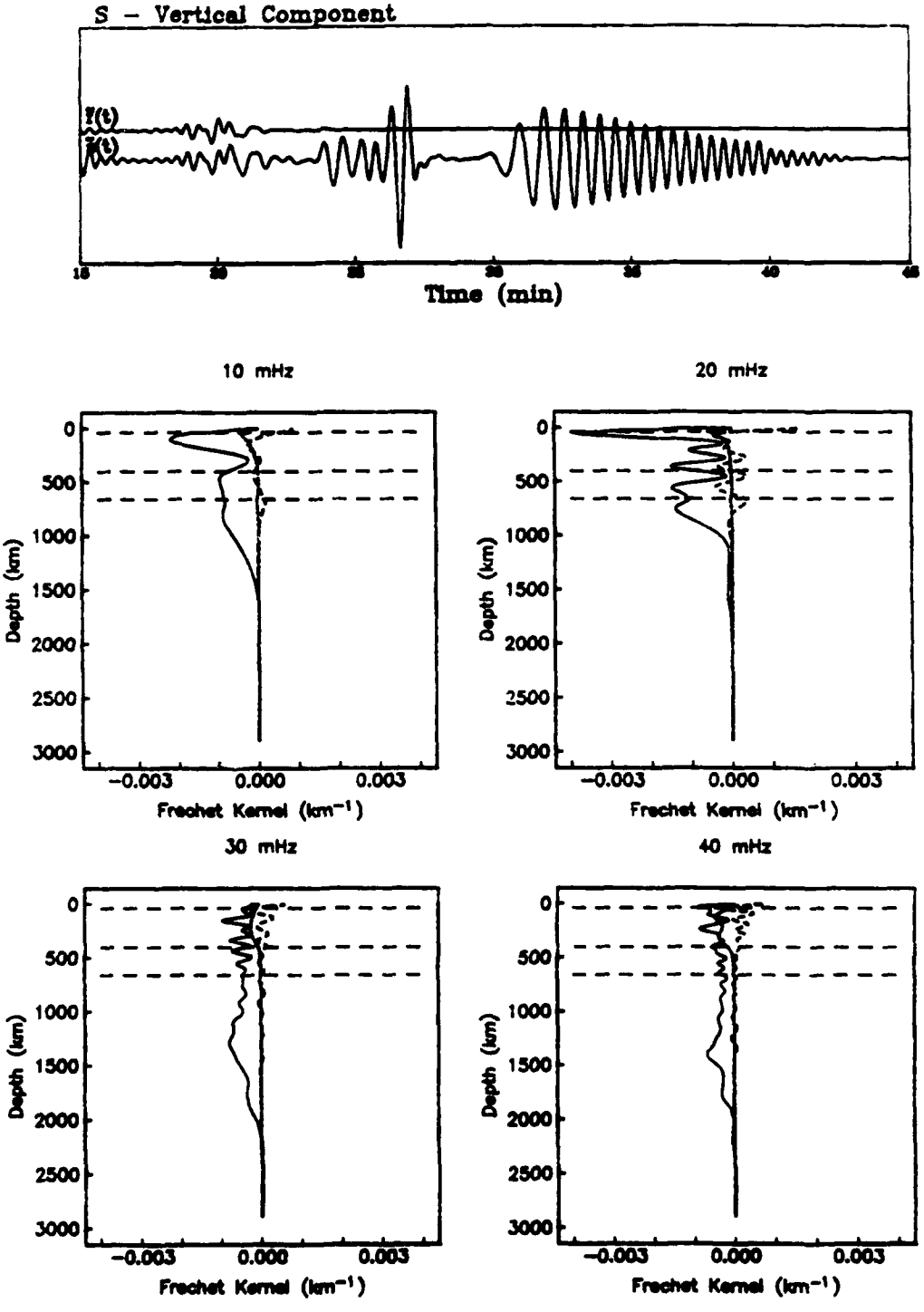

S – Vertical Component
Time (min)
10 mHz
20 mHz
30 mHz
40 mHz
Depth (km)
Frechet Kernel (km⁻¹)
0
500
1000
1500
2000
2500
3000
-0.003
0.000
0.003

Figure 4.14b

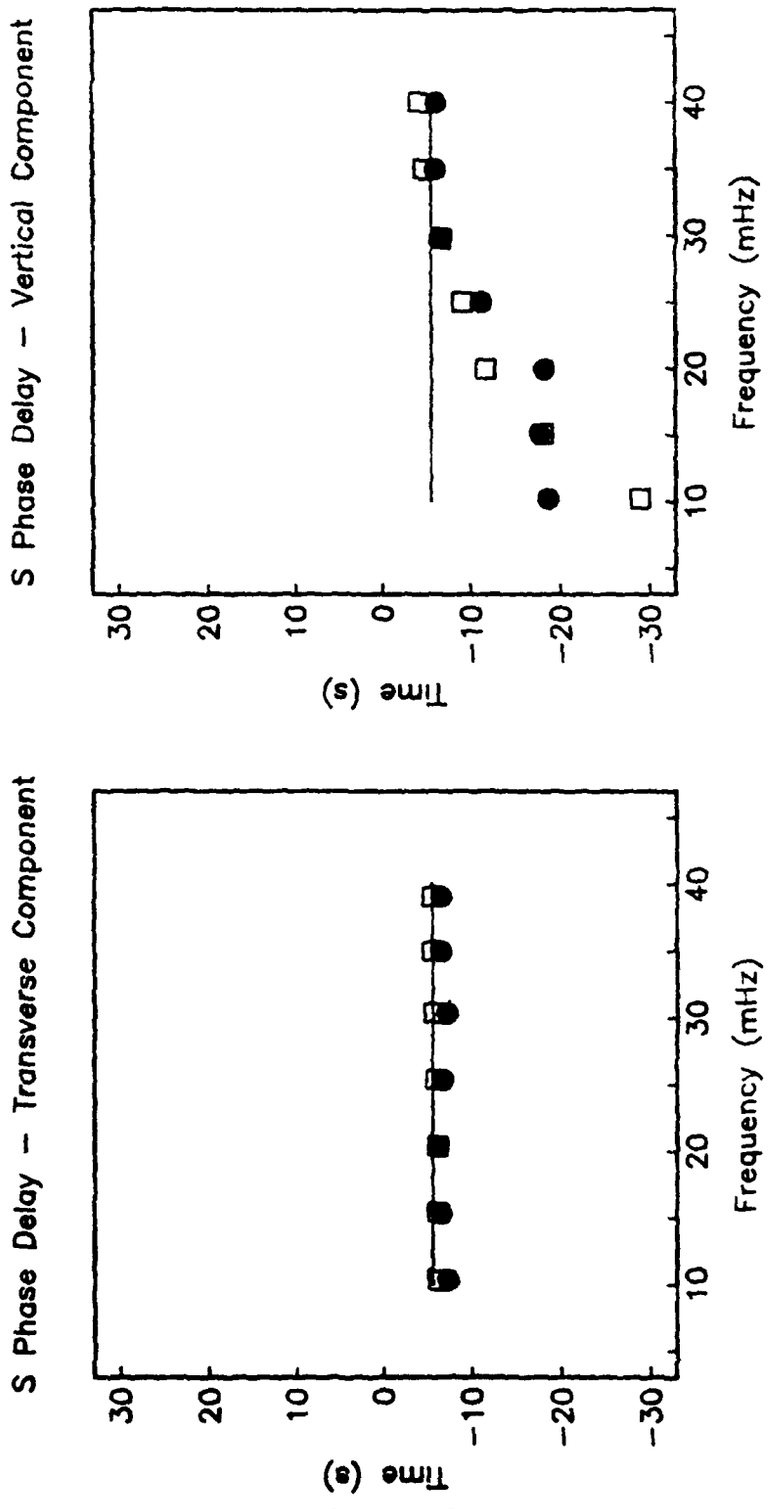

S Phase Delay – Vertical Component
Time (s)
30
20
10
0
-10
-20
-30
Frequency (mHz)
10
20
30
40
S Phase Delay – Transverse Component
Time (s)
30
20
10
0
-10
-20
-30
Frequency (mHz)
10
20
30
40

Figure 4.15

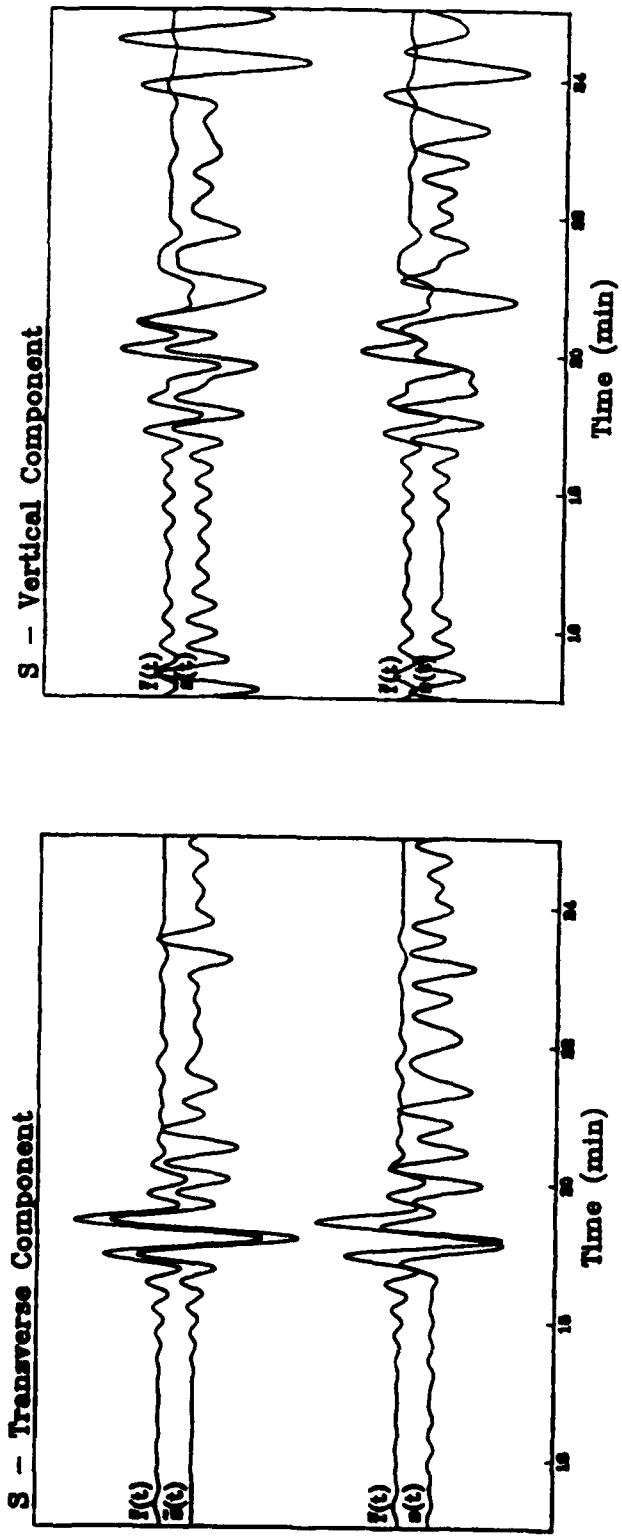

S – Vertical Component
S – Transverse Component
Time (min)
Time (min)

Figure 4.16a

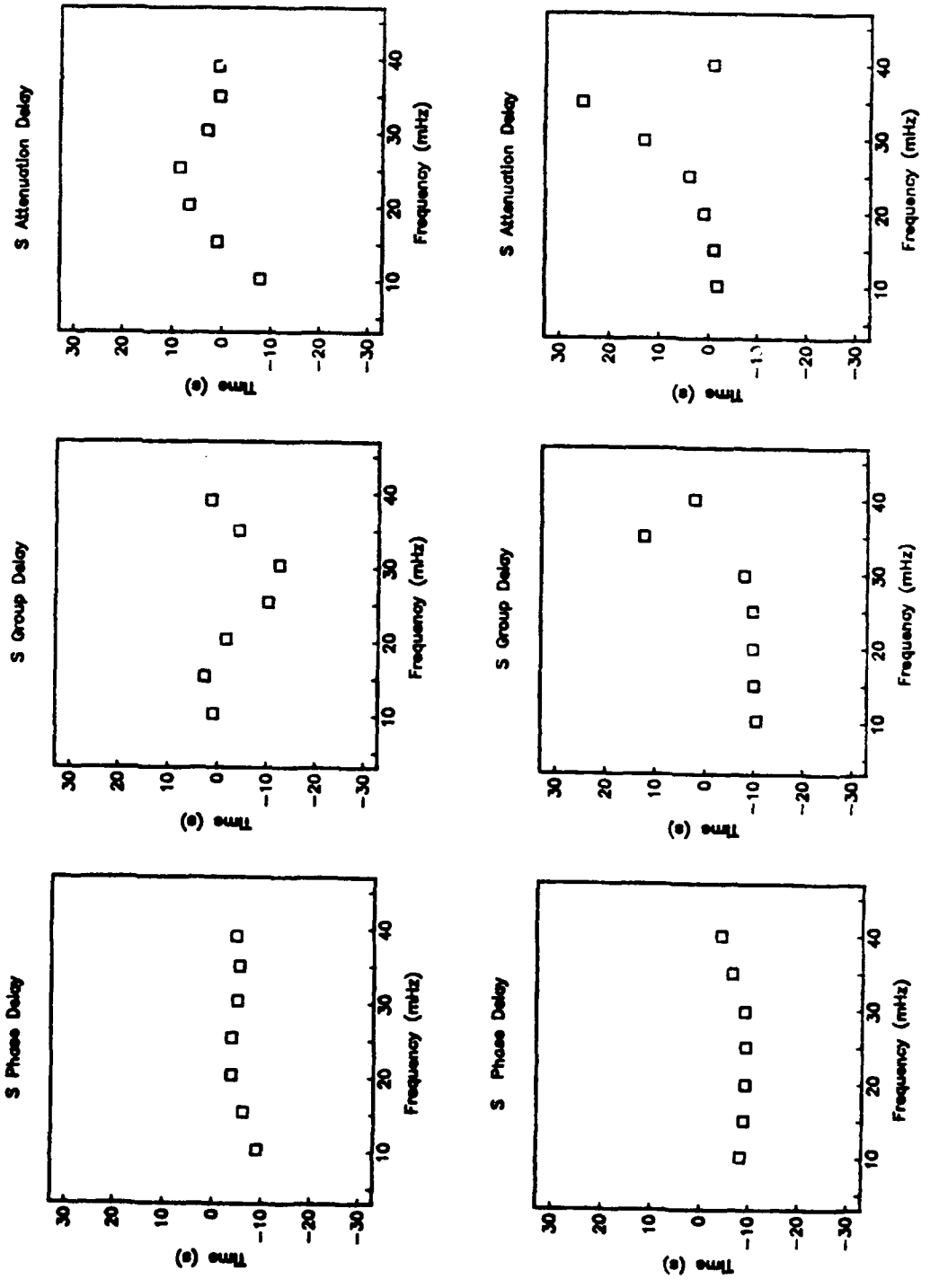

S Attenuation Delay
S Group Delay
S Phase Delay
Time (s)
Frequency (mHz)
30
20
10
0
-10
-20
-30
10
20
30
40

Figure 4.16b

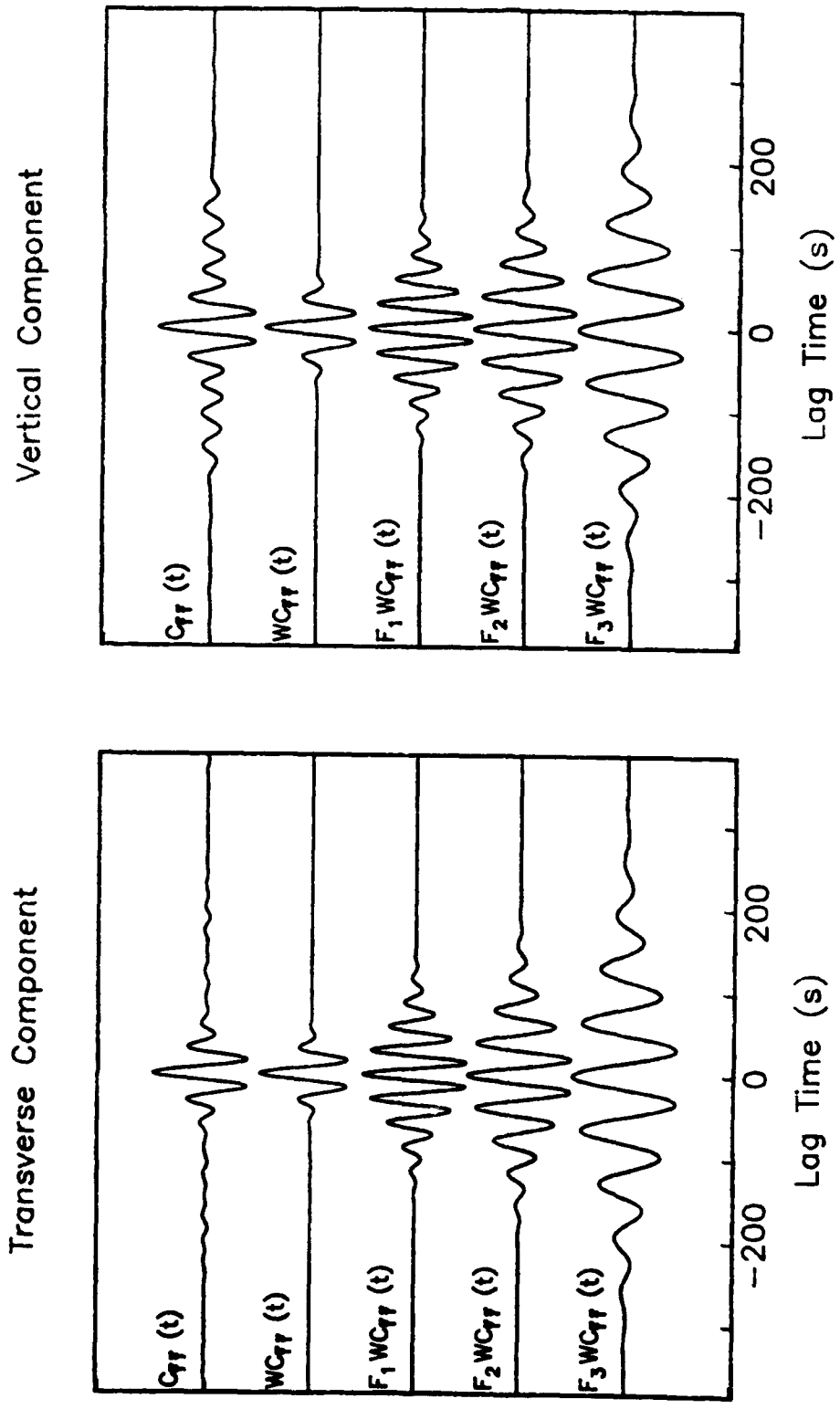

Vertical Component
C_{TT} (t)
WC_{TT} (t)
F_1WC_{TT} (t)
F_2WC_{TT} (t)
F_3WC_{TT} (t)
−200
0
200
Lag Time (s)
Transverse Component
C_{TT} (t)
WC_{TT} (t)
F_1WC_{TT} (t)
F_2WC_{TT} (t)
F_3WC_{TT} (t)
−200
0
200
Lag Time (s)

Figure 4.17

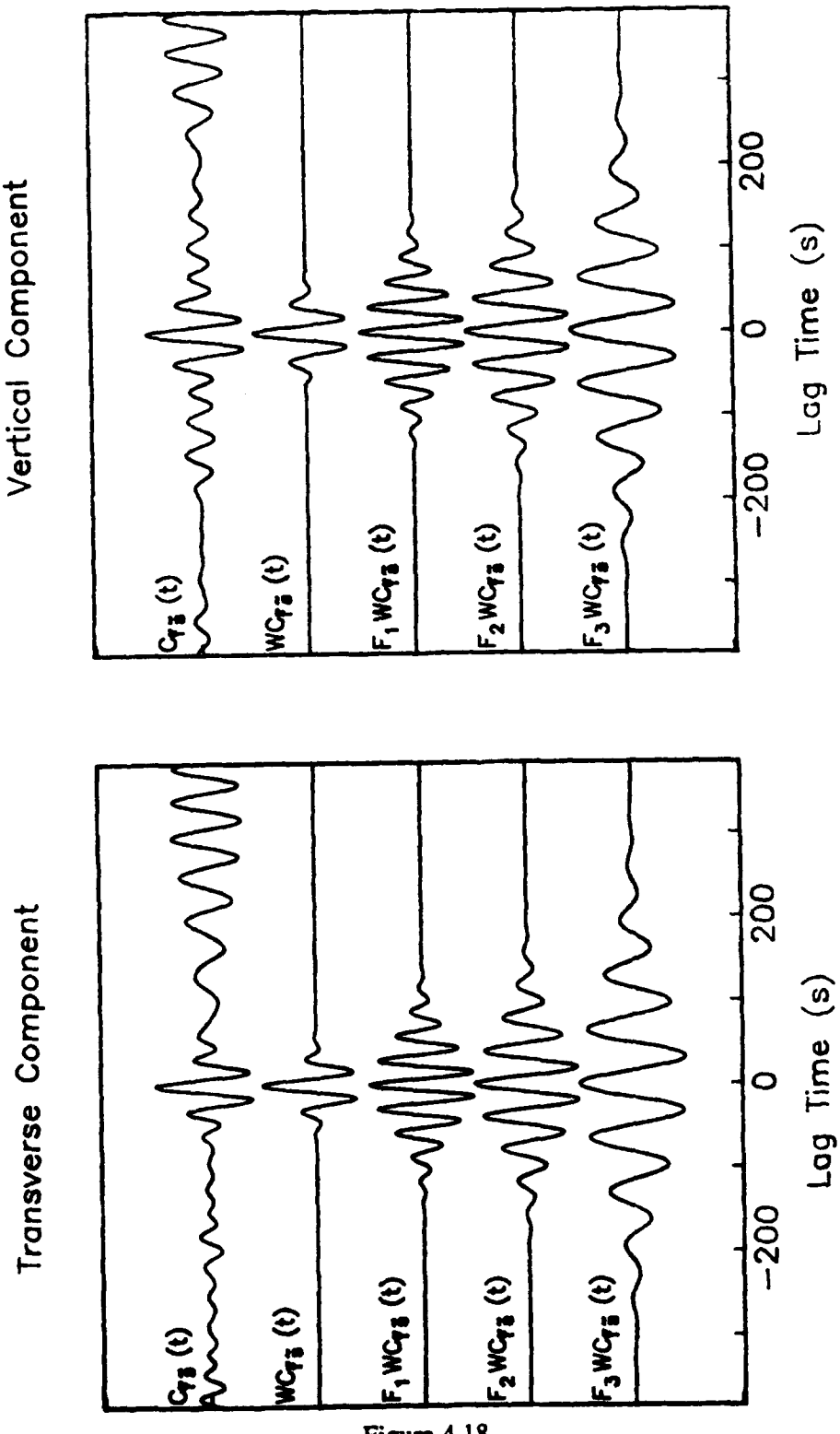

Vertical Component
Transverse Component
C_{TS} (t)
WC_{TS} (t)
F_1WC_{TS} (t)
F_2WC_{TS} (t)
F_3WC_{TS} (t)
-200
0
200
Lag Time (s)

Figure 4.18

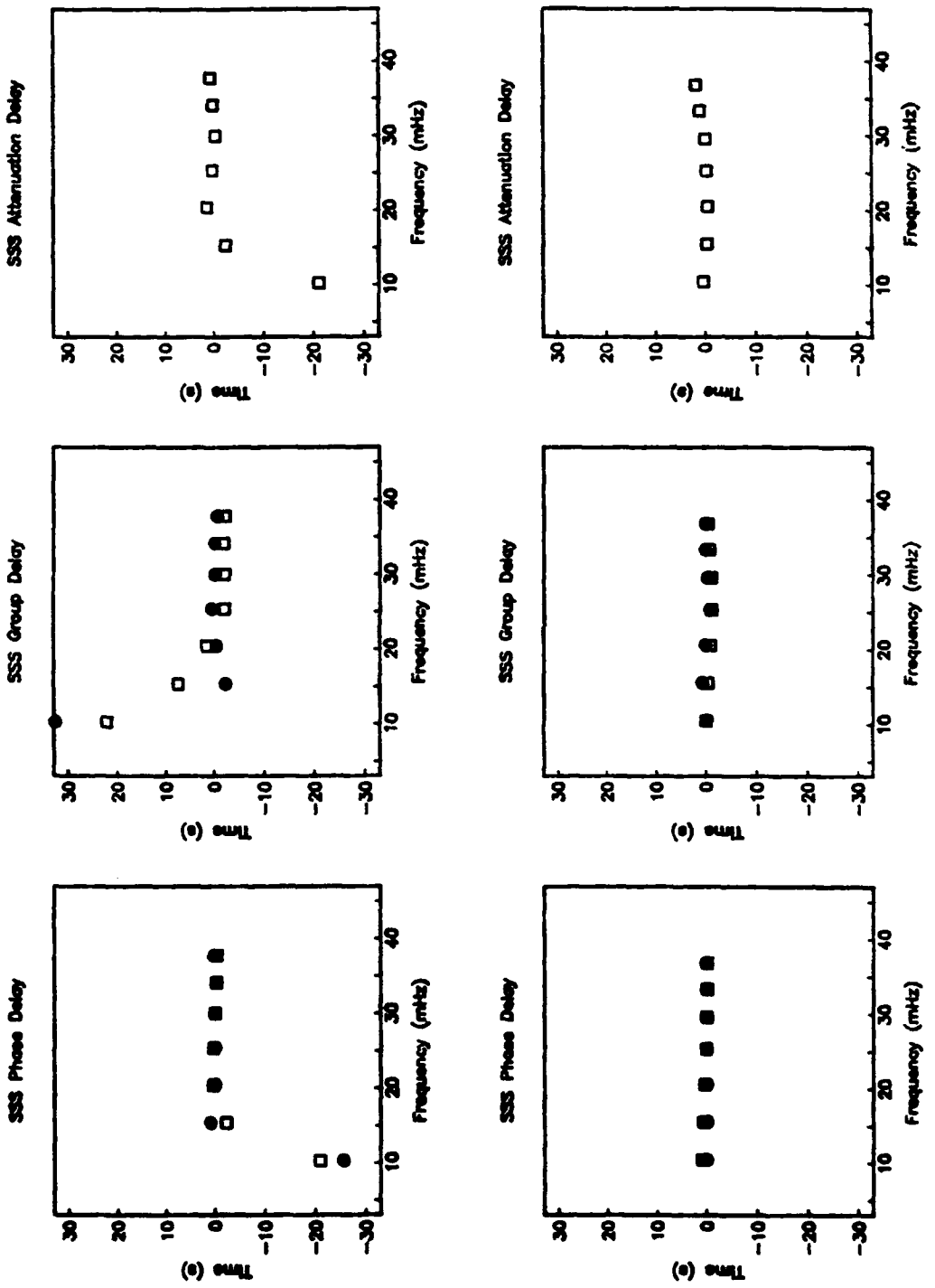

SSS Attenuation Delay
SSS Group Delay
SSS Phase Delay
Frequency (mHz)
Time (s)

Figure 4.19

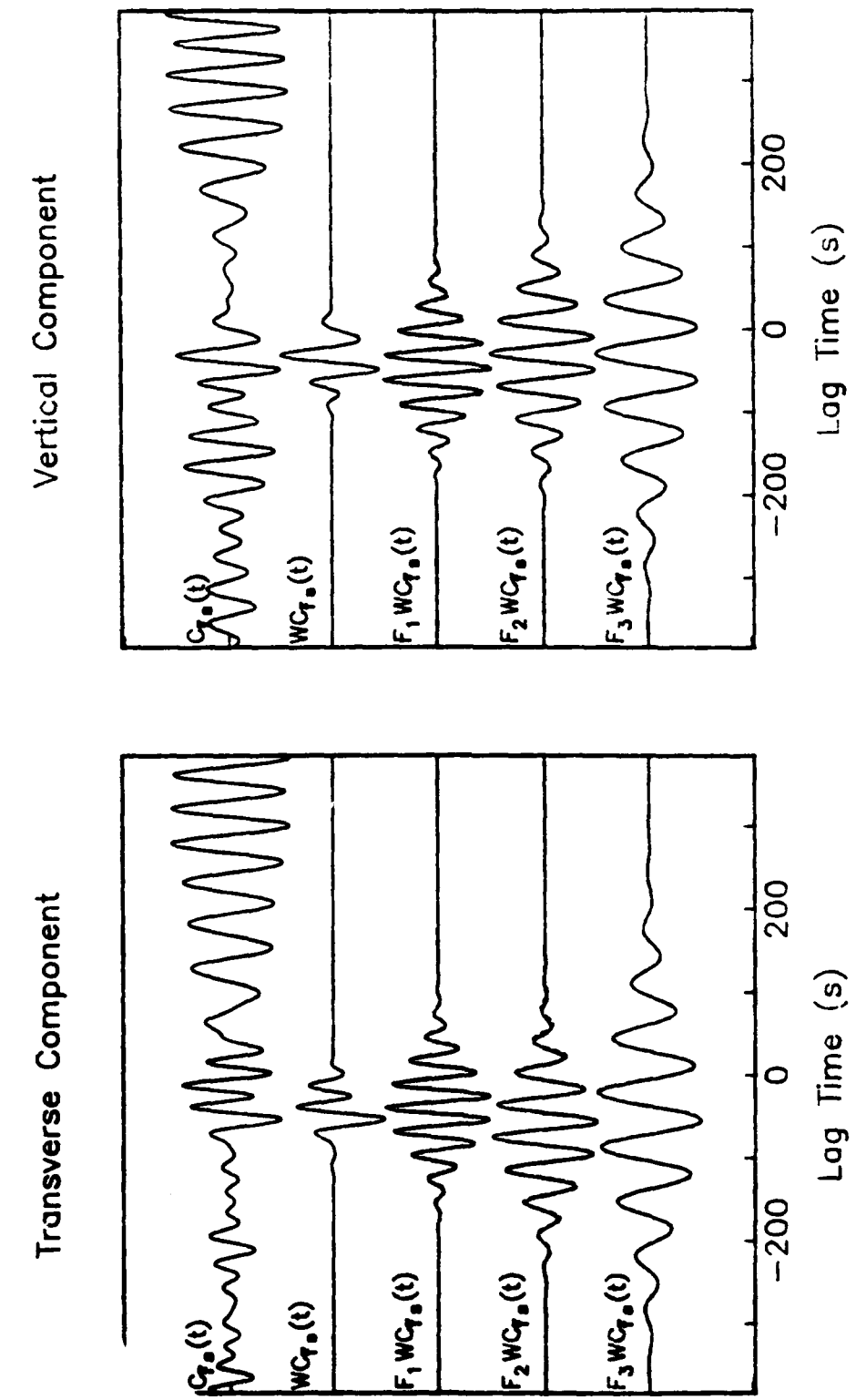

Vertical Component
Transverse Component
$G_{T_0}(t)$
$WG_{T_0}(t)$
$F_1WG_{T_0}(t)$
$F_2WG_{T_0}(t)$
$F_3WG_{T_0}(t)$
−200
0
200
Lag Time (s)

Figure 4.20

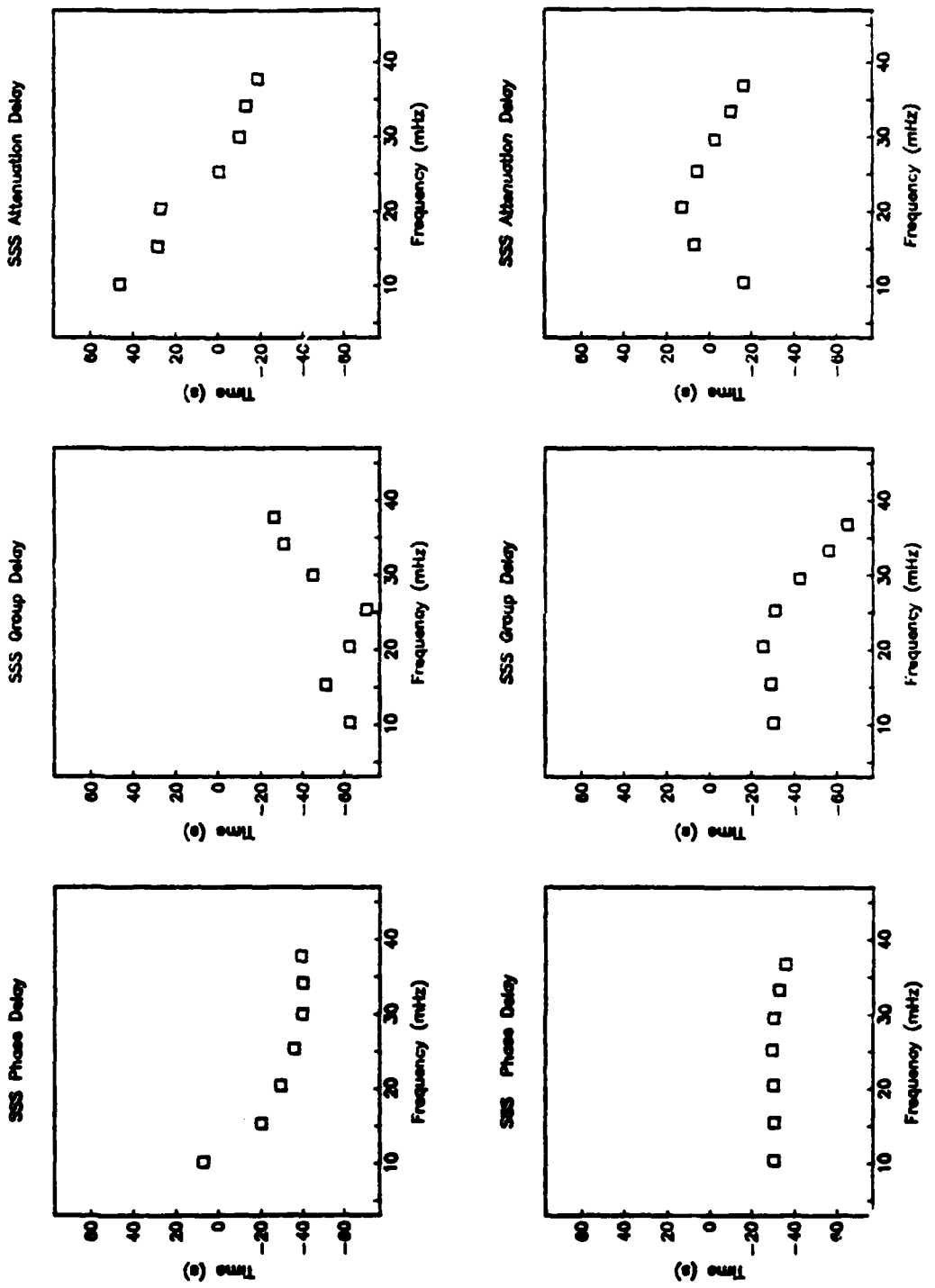

SSS Attenuation Delay
SSS Group Delay
SSS Phase Delay
Time (s)
Frequency (mHz)

Figure 4.21

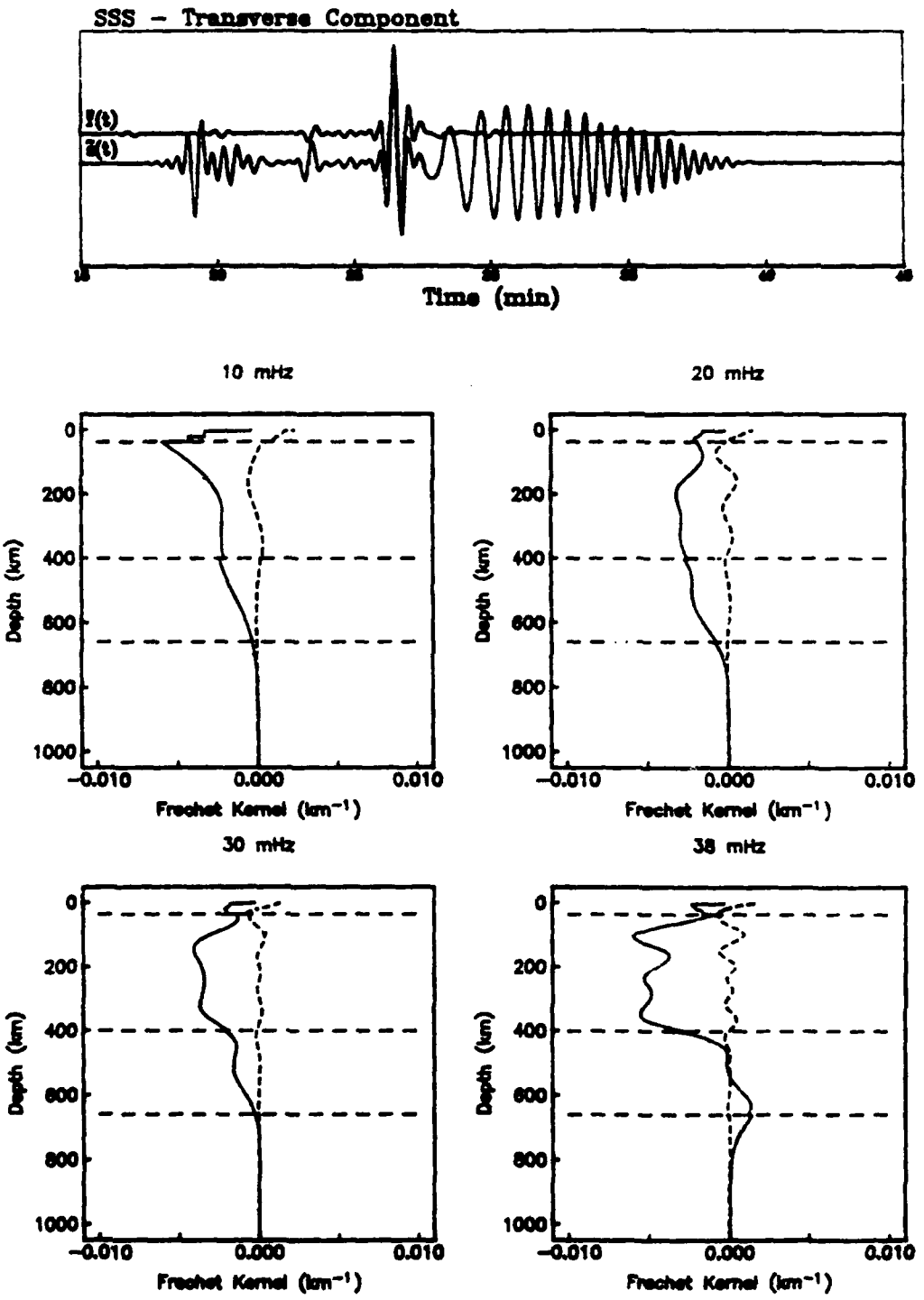

SSS – Transverse Component
Time (min)
10 mHz
20 mHz
30 mHz
38 mHz
Depth (km)
Frechet Kernel (km⁻¹)
0
200
400
600
800
1000
-0.010
0.000
0.010

Figure 4.22a

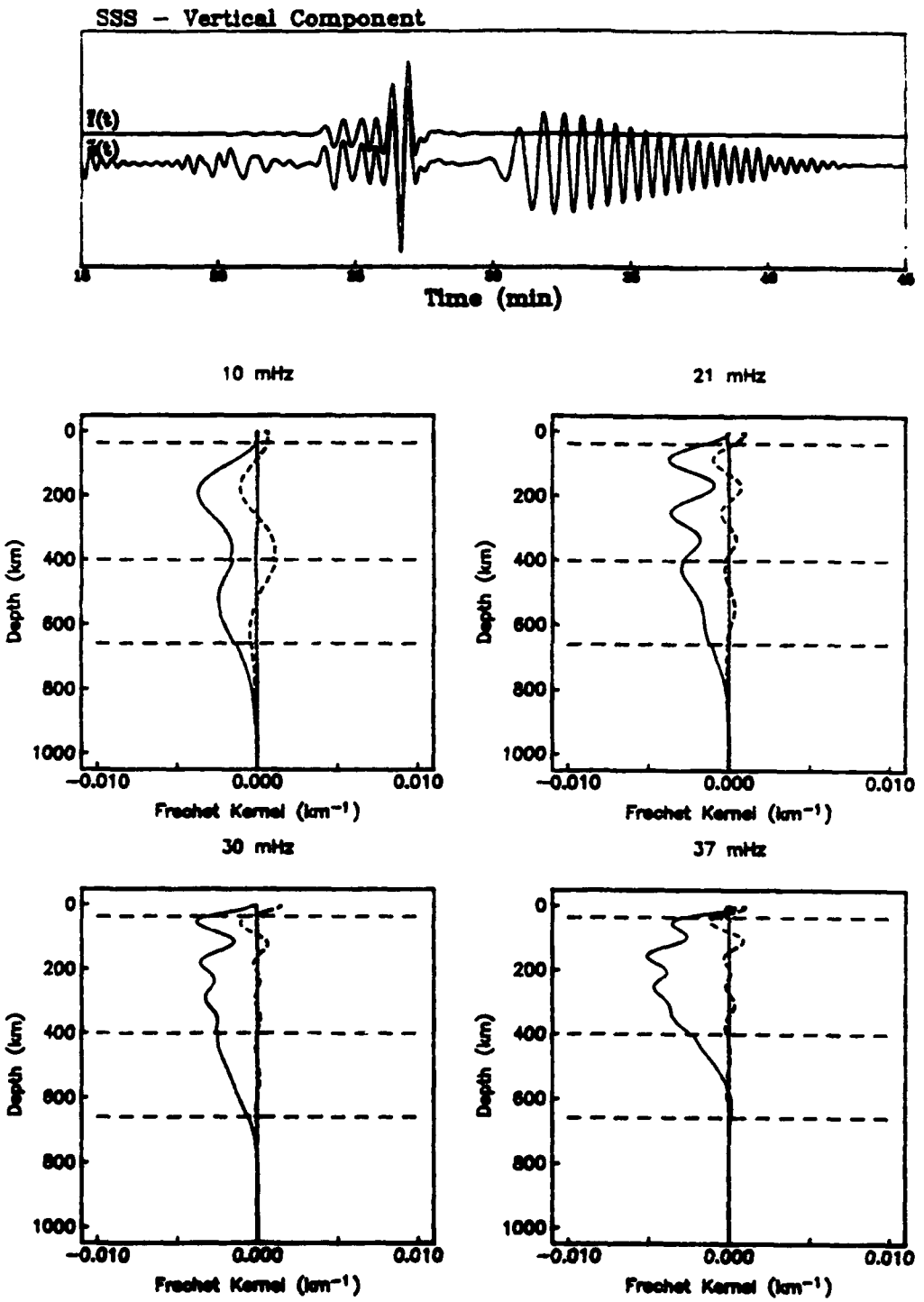

SSS – Vertical Component
Time (min)
10 mHz
21 mHz
30 mHz
37 mHz
Depth (km)
Frechet Kernel (km⁻¹)
0
200
400
600
800
1000
-0.010
0.000
0.010

Figure 4.22b

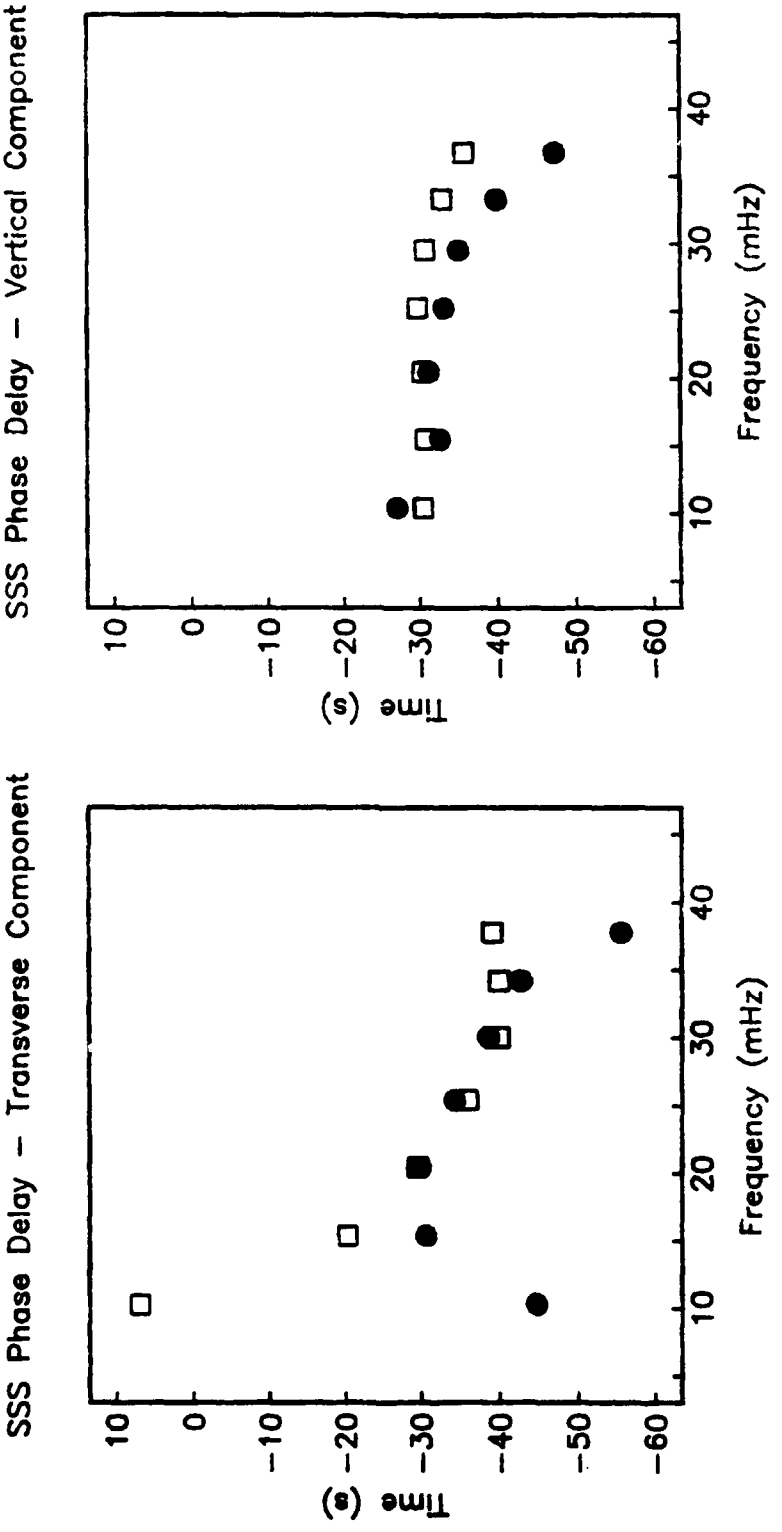

SSS Phase Delay – Vertical Component
Time (s)
10
0
−10
−20
−30
−40
−50
−60
Frequency (mHz)
10
20
30
40
SSS Phase Delay – Transverse Component
Time (s)
10
0
−10
−20
−30
−40
−50
−60
Frequency (mHz)
10
20
30
40

Figure 4.23

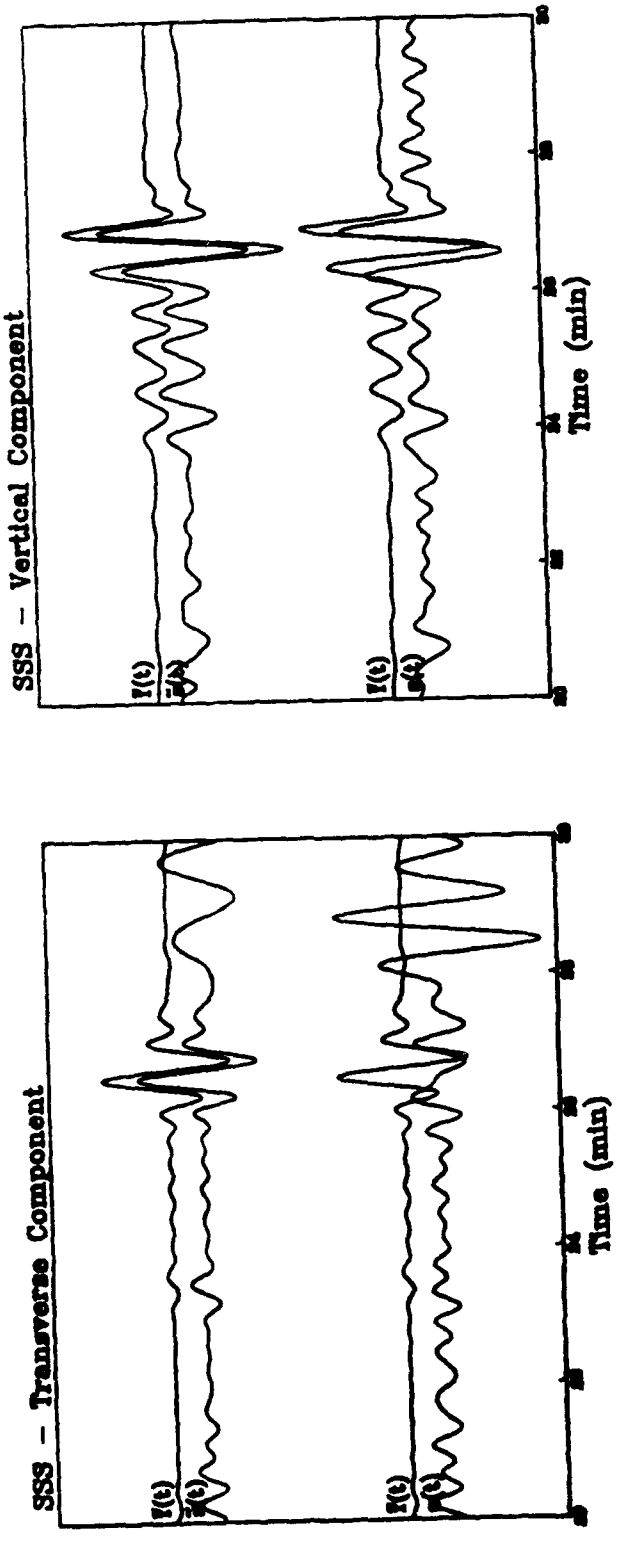

SSS – Vertical Component
I(t)
R(t)
I(t)
R(t)
Time (min)
SSS – Transverse Component
I(t)
R(t)
I(t)
R(t)
Time (min)

Figure 4.24a

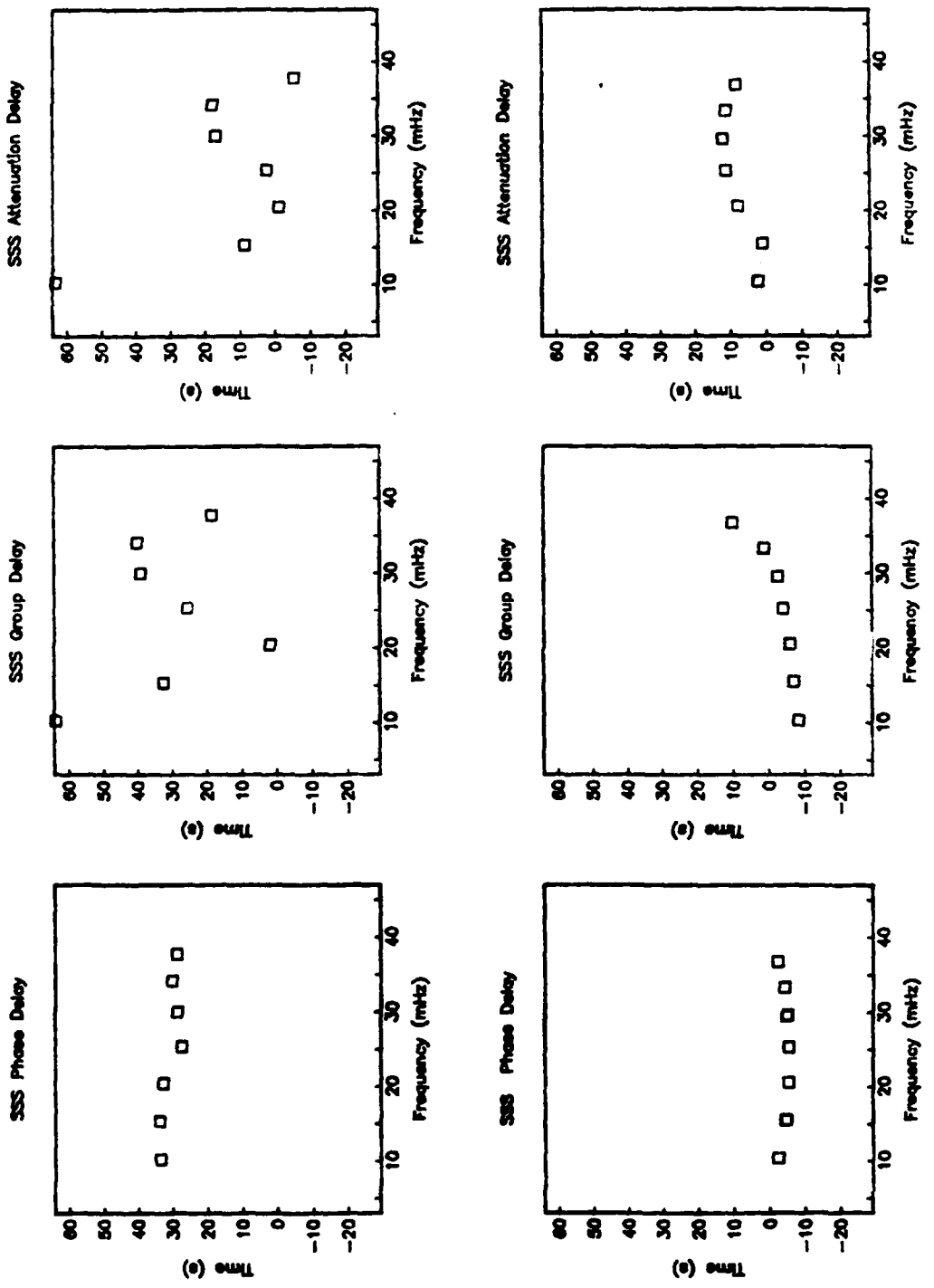

SSS Attenuation Delay
SSS Group Delay
SSS Phase Delay
Time (s)
Frequency (mHz)

Figure 4.24b

CHAPTER 5

FUTURE WORK AND CONCLUSIONS

INTRODUCTION

Cross-correlation has a long history in seismological-data processing, from the "phase-equalization" technique of Aki [1960] for recovery of the source function to the "phase-matched filtering" approach of Herrin and Goforth [1977, 1986] for the reconstruction of group velocity curves. Synthetic seismograms have been used in conjunction with cross-correlation to measure differential travel times [Hart, 1975; Hart and Butler, 1978; Butler, 1979], differential dispersion [Dziewonski *et al.*, 1972; Herrin and Goforth, 1977, 1986] and to isolate wavegroups for waveform inversion [Lerner-Lam and Jordan, 1983, 1987].

Our methodology is most similar to the approach of Lerner-Lam and Jordan [1983, 1987]. They applied isolation filtering techniques to isolate individual wave groups prior to inversion and applied a windowing operator to minimize the contamination of interfering energy. They used the cross-correlation with the complete synthetic seismogram to model wave effects such as caustic phase shifts, diffraction, physical dispersion, and interfering arrivals. However, they inverted the difference between the observed and synthetic correlation functions for model perturbations, while we have developed an analytic representation of the correlation function in terms of a few simple parameters.

Our characterization of the cross-correlation functions, based on an expansion of Hermite polynomials, is a powerful approach. This formulation allows us to express an arbitrary spectrum in the form of a Gaussian with higher-order terms which depend on the

normalized moments of the spectrum. We may carry these terms through our processing procedure and evaluate their contribution. Based on a first-order Taylor series expansion of differential wavenumber, we have shown that the filtered windowed cross-correlation function may be approximated as a harmonic carrier modulated by a Gaussian envelope, parameterized by a center frequency ω_i' and spectral half-width σ_i', perturbations due to interfering arrivals, $\Delta\tilde{\tau}_p$, $\Delta\tilde{\tau}_g$, and $\Delta\tilde{\tau}_a$, and time shifts owing to differences between the reference model and the actual Earth, $\delta\tau_p$, $\delta\tau_g$, and $\delta\tau_a$. In those situations when the differential quadratic term is large, we have developed a processing procedure which allows us to estimate the quadratic contribution.

We have developed a general formalism for isolation filters in terms of weighted sums of traveling-wave branches. This representation is completely general and allows us to compute isolation filters for body-wave phases by designing the $\tilde{\alpha}_m(\omega)$ to select modes on the basis of group arrival time. We demonstrated that this methodology is successful in the computation of isolation filters for *S* and *SSS*. While we can use a ray-theoretical technique such as WKBJ for the calculation of the isolation filters (since the isolation filters need not be exact), our formalism is complete and allows us to model phenomena such as shear-coupled *PL* and converted phases.

This approach is completely general. It allows us to analyze arbitrary waveforms on three-component data over a broad range of frequencies with a self-consistent, uniform methodology.

Before concluding, we wish to discuss several problems in Earth structure to which our waveform-analysis procedure may be readily applied.

FUTURE WORK

Upper-mantle structure

In Chapter 1, we motivated the development of this technique by introducing the EU2-SNA discrepancy (Figure 1.1). In the past, the methods applied to *SH*-polarized waveforms (*e.g.*, the ray-theoretical methods of Grand and Helmberger [1984]) have been very different from the methods applied to *PSV*-polarized waveforms (*e.g.*, the mode-theoretical methods of Lerner-Lam and Jordan [1983]). This has made the comparison of the results, especially discrepancies among models, difficult to evaluate. Our methodology for waveform analysis permits a unified treatment of body waves and surface waves from all three components of ground motion. The preliminary results reported by Gee and Jordan [1988 (Appendix F)] suggest that at least some of the EU2-SNA discrepancy is associated with large-scale, large-amplitude polarization anisotropy in the uppermost mantle. Although this conclusion is hardly surprising given the reported evidence from surface waves for polarization anisotropy in the continental upper mantle [*e.g.*, Crampin and King, 1977; Cara *et al.*, 1980; Dziewonski and Anderson, 1981; Lévèque and Cara, 1983, 1985; Nataf *et al.*, 1986], the large shear-wave splitting in Northern Eurasia was observed using techniques similar to those discussed here. An obvious application of this methodology is to obtain further constraints on the nature of this apparent anisotropy (Figure 5.1). In particular, we shall focus on the question of how much of the splitting is due to true, intrinsic anisotropy caused by crystal alignment and how much is due to fine-scale layering and lateral heterogeneity. The Fréchet kernels examined in Chapter 4 suggest that *SH* and *PSV*-polarized waveforms average Earth structure in fundamentally different ways, at least at low frequencies. Thus, a discrepancy in *SH* and *PSV* travel times is not diagnostic of polarization anisotropy. The solution to these problems lies in the careful

formulation of the inverse problem. In particular, we plan to engage in hypothesis testing in order to determine whether our observations require polarization anisotropy.

Transition-zone heterogeneity

We will also continue the development of the technique to handle the more subtle observations associated with fine-scale upper-mantle structure. For example, because the triplicated body-wave arrivals interfere strongly, their individual travel times cannot be easily measured using single isolation filters. However, we may construct an isolation filter for the complete triplication and measure the resulting frequency-dependent travel times. As demonstrated by the *SSS* example in Chapter 4, we can recover several, independent constraints on velocity structure from a single waveform. We plan to use this technique to study lateral variations in the mid-mantle transition zone. We are particularly interested in the relative amplitude of the topography on the 400-km and 650-km discontinuities and whether it is anticorrelated, as would be expected if these features were phase transitions with Clapeyron slopes of opposite signs. This research will yield information very different from, and complementary to, the data obtained on discontinuity structure from *SH*-polarized mantle reverberations by Revenaugh and Jordan [1987, 1989].

Core-mantle boundary and outer-core structure

The structure of the core and the core-mantle boundary has been the subject of much recent interest [Creager and Jordan, 1986; Ritzwoller *et al.*, 1986, 1988; Morelli and Dziewonski, 1987; Giardini *et al.*, 1987; Young and Lay, 1987a]. A wide-range of resources have been brought to bear on the problem, including high-frequency *P*-wave travel times [Creager and Jordan, 1986; Morelli and Dziewonski, 1987], anomalously split eigenfrequencies [Ritzwoller *et al.*, 1986, 1988; Giardini *et al.*, 1987], and *SH*-polarized waveforms [Lay and Helmberger, 1983; Young and Lay, 1987a; Revenaugh, 1989].

These recent results add to the large literature on the fine structure of the outer core and core-mantle boundary (see Cleary [1974] and Young and Lay [1987b] for extensive reviews).

There is still considerable uncertainty surrounding the velocity gradients in D". The question of whether shear-wave velocity increases or decreases at the base of the mantle is still debated [Cleary *et al.*, 1967; Cleary, 1969; Bolt *et al.*, 1970; Hales and Roberts, 1970; Mitchell and Helmberger, 1972; Mondt, 1977; Okal and Geller, 1979; Doornbos and Mondt, 1979a, 1979b; Mula, 1981; Doornbos, 1983; Lay and Helmberger, 1983; Bolt and Niazi, 1984; Young and Lay, 1987a]. Most of the work supporting a decrease in shear velocity has been carried out on the frequency-dependent amplitude decay and apparent phase velocity of S_{diff}. The difficulty with these approaches is that ray-theoretical methods breakdown in the deep shadow of the core-mantle boundary. Our methodology provides us with the tools to analyze S_{diff} and recover estimates of its differential phase, group, and amplitude time parameters as a function of frequency. Figures 5.2–5.4 illustrate some preliminary results on the analysis of S_{diff}. We have developed a simple relationship between the generalized data functionals and the model parameters, and the interpretation of results from S_{diff} and *SKS* in terms of core-mantle boundary and outer core structure will be straightforward.

Inner-core structure

As a final example, we consider a very specific problem: how do we determine the average shear velocity of the inner core? This is a geophysically interesting problem that has proven to be difficult to solve by standard seismological techniques. Body waves with inner-core shear-wave legs have not been reliably observed (although see Julian *et al.*'s [1973] discussion of array-based searches for *PKJKP*). The best information on inner-core shear velocities currently comes from the eigenfrequencies of a few low-degree spheroidal modes [Dziewonski and Gilbert, 1973], but these observations may be

compromised by anomalous splitting [Masters and Gilbert, 1981; Ritzwoller *et al.*, 1986: Giardini *et al.*, 1987].

We consider two earth models, CORE 11, which we take to be the reference model, and a version of this model, CORE 11', that has the shear velocities of the inner core reduced by 10% (Figure 5.5). (CORE 11 is an unpublished, transversely isotropic structure that fits the spherically averaged eigenfrequencies compiled by the UCSD group [Widmer *et al.*, 1988; G. Masters, personal communication, 1988].) Figure 5.6 shows three time series calculated at the IDA station PFO for the 1982 Banda Sea earthquake between 0 and 1000 minutes after the origin times (*a*) and a high-pass filtered version of the interval from 200 to 800 minutes (*b*). The first is the reference synthetic $\tilde{s}(t)$ computed by complete normal-mode summation from CORE 11; the third models an "observed" seismogram for CORE 11'.

The middle traces in Figure 5.6 correspond to an isolation filter $\tilde{f}(t)$ computed from CORE 11 by summing only those modes that have 5% or more of their shear energy concentrated in the inner core. These modes, of which there are about 1000 whose frequencies are in the band-pass of the IDA instrument, are illustrated on an ω–l diagram in Figure 5.7. The amplitude of any individual mode in $\tilde{f}(t)$ is very small at the earth's surface; for many, such as the inner-core Stoneley modes, it is infinitesimal. However, their summed response has a significant amplitude, especially at higher frequencies; in fact, by high-pass filtering, we can obtain an isolation filter with an RMS amplitude of about 30% of the total response (Figure 5.6*b*). No particularly significant wave groups are obvious in this time interval because the response is distributed over many mode branches.

Figure 5.8 displays the correlation functions $C_{\tilde{f}\tilde{s}}(t)$, $C_{\tilde{f}\tilde{f}}(t)$, and $C_{\tilde{f}s}(t)$ (*a*) and their windowed versions together with their analytical Gaussian approximations (*b*). $C_{\tilde{f}\tilde{s}}(t)$ closely resembles $C_{\tilde{f}\tilde{f}}(t)$ and has its peak very near zero lag, indicating that the distortion due to interference by other modes is small. The envelope of $C_{\tilde{f}s}(t)$, on the other hand, is shifted towards positive lag by about 120 s by the reduced inner-core shear velocity.

Although at the time of this writing we have not yet computed the Fréchet kernel for the phase-delay functional $\delta\tau_p$ associated with this observation, the large group delay evident in the lowermost trace of Figure 5.8 indicates that measurements based on this isolation filter are reasonably sensitive to the inner-core shear velocity.

CONCLUSIONS

In this thesis, we have presented a waveform-analysis procedure for recovering amplitude and phase information from complex seismograms. This methodology is similar to waveform inversion in that it makes use of the ability to calculate synthetic seismograms from realistic Earth models. Unlike waveform-inversion techniques, which invert the difference between an observed and synthetic seismogram for structural perturbations, we use synthetic seismograms to recover information corresponding to well-defined scalar-valued functions of Earth structure: phase delays, group delays, attenuation times, and their generalizations. While the more classical approach to measuring such kinematic properties is often confounded by wave effects such as caustic phase shifts, diffraction, physical dispersion, and interfering arrivals, the use of complete synthetic seismograms allows use to model these phenomena explicitly.

Our approach is based on the analytic representation of the cross-correlations between an isolation filter and the complete synthetic and observed seismograms. We process these correlation functions; first by applying a windowing operator to reduce the effects of interfering arrivals and second by applying a narrow-band filter. We have developed expressions for the processed autocorrelation function $F_i WC_{ff}(t)$, the processed cross-correlation between the isolation filter and the complete synthetic $F_i WC_{fs}(t)$, and the processed cross-correlation between the isolation filter and the observed seismogram $F_i WC_{fs}(t)$. By using a waveform-fitting algorithm, we may estimate the effective center frequency and spectral half-width (ω_i', σ_i'), the perturbations due to interference

$(\Delta\bar{\tau}_p(\omega_i), \Delta\bar{\tau}_g(\omega_i), \Delta\bar{\tau}_a(\omega_i))$, and the perturbations due to differences between the reference model and the Earth $(\delta\tau_p(\omega_i), \delta\tau_g(\omega_i), \delta\tau_a(\omega_i))$.

We illustrated our methodology with several types of isolation filters, spanning the range from fundamental-mode surface waves to body-wave arrivals. In particular, we outlined an approach for constructing isolation filters based on the weighting of traveling-wave branches. However, our technique for the recovery of generalized data functionals does not depend on method of computation. For example, we could have considered an isolation filter calculated from some other convenient algorithm. Our method is hybrid in the sense that the isolation filter may be calculated by any procedure and need not be exact. The complete synthetic seismogram allows us to correct for errors in the isolation filter as well as the presence of interference.

The observables $\delta\tau_p$, $\delta\tau_g$, and $\delta\tau_a$ are functionals of earth structure. In the case of an isolated waveform, the Fréchet kernels are known [Backus and Gilbert, 1967; Woodhouse, 1976; Woodhouse and Dahlen, 1978]. We have developed formulas for the Fréchet kernels for the general form of $\bar{f}(t)$, which express the observed perturbation as a linear sum of the individual perturbations of each traveling-wave branch. Hence, once the generalized data functionals have been measured and their uncertainties determined as a function of frequency (and source-receiver position) by fitting the Gaussian narrow-band expressions to $F_i WC\bar{\gamma}_s(t)$, they may be easily inverted for structural models using standard linearized methods.

Over the last decade, seismologists have made spectacular progress in elucidating the three-dimensional structural variations of the Earth. However, only a small fraction of the information available from existing seismograms has been used for this purpose and many important geophysical questions related to the details of earth structure remain unanswered. While the growth of three-component, broad-band global digital arrays such as Geoscope, Network of Autonomously Registering Stations (NARS), and Global Seismic Network (GSN) will allow seismologists to make further progress on these

problems in the 1990's, the process may be greatly accelerated by the development of techniques which make better use of the existing data. We think the procedures developed in this thesis will contribute significantly in this area.

FIGURE CAPTIONS

FIGURE 5.1

Example of apparent shear-wave splitting in *SS* recorded at the GDSN station KONO from an event in the Hindu Kush [84/07/01, h = 199 km, Δ = 45°]. The upper boxes present the observed seismogram, the complete synthetic seismogram, and the isolation filter for *SS* (left - transverse component; right - vertical component). In this example, the synthetic seismogram and the isolation filter are calculated from the model SNA. The bottom boxes illustrate the resulting correlation functions (solid line), $C_{\tilde{T}\tilde{T}}(t)$, $C_{\tilde{T}\tilde{S}}(t)$, and $C_{\tilde{T}_S}(t)$, with the model derived from waveform fitting (symbols). The broad-band correlation functions indicate a differential shift between the *SH* and *PSV*-polarized waveforms.

FIGURE 5.2

Record section from an event in Tonga [82/12/20, h = 10 km], recorded at the Regional Seismic Test Network (RSTN) stations in North America. The solid line indicates the observed seismogram; the dashed line is the complete synthetic seismogram calculated from the model EU2. In order of increasing distance, the stations are RSSD (Δ = 95°), RSNT (Δ = 99°), RSNT (Δ = 99°), RSON and RSCP (Δ = 104°), and RSNY (Δ = 115°).

FIGURE 5.3

We constructed isolation filters for S_{diff} using the technique described in Chapter 4 and estimated the differential time parameters as a function of frequency. This figure illustrates the differential phase delay at six center frequencies, from 15 to 40 mHz, as a function of distance. These parameters represent raw values $\delta\tau_p$ and have not been corrected for station statics.

FIGURE 5.4

Shear-velocity (*a*) and density (*b*) kernels at four center frequencies. Horizontal dashed lines indicate the Moho, the 400 km discontinuity, the 650 km discontinuity, the top of D", and the core-mantle boundary. The length of the dashes indicate the distance: from 95° (shortest dashes) to 115° (solid line). Notice how the shear-velocity kernels become increasingly localized in D" with increasing distance. We think that the application of our waveform-analysis procedure will allow us to obtain fundamentally new information about the structure at the core-mantle boundary.

FIGURE 5.5

Shear velocity as a function of depth for models CORE 11 (solid line) and CORE 11' (dashed line). CORE 11 is a transversely isotropic structure which represents

the smallest perturbation to PREM [Dziewonski and Anderson, 1981] that satisfies the the spherically-averaged eigenfrequencies compiled by the UCSD group [G. Masters, personal communication, 1988]. CORE 11' is a version of CORE 11 with shear velocities in the inner core reduced by 10%. In our inner core experiment, CORE 11 is used as the reference model. The upper curve in each model is the *SH* velocity, whereas the lower curve is the *SV* velocity.

FIGURE 5.6

Timeseries calculated from CORE 11 and CORE 11' for an event in the Banda Sea (82/06/22, $h = 473$ km) recorded at the IDA station PFO ($\Delta = 117°$) between 0 and 1000 (*a*) and 200 and 800 minutes (*b*) after the origin time. The first is the reference synthetic $\tilde{s}(t)$ computed by complete normal-mode summation from CORE 11; the third models an "observed" seismogram for CORE 11'. The second is an isolation filter $\tilde{f}(t)$ computed from CORE 11 by summing only those modes that have 5% or more of their shear energy concentrated in the inner core.

FIGURE 5.7

Frequency (ω) *vs.* angular order (l) diagram for the transversely isotropic model CORE 11. This figure represents a subset of the spheroidal-mode catalog to 50 mHz. Modes with greater than 5% of their shear energy concentrated in the inner core are indicated with large circles.

FIGURE 5.8

The correlation functions $C_{\tilde{f}\tilde{s}}(t)$, $C_{\tilde{f}\tilde{f}}(t)$, and $C_{\tilde{f}\tilde{s}}(t)$ (*a*) and their filtered, windowed versions (*b*) together with their analytical Gaussian approximations (symbols) are displayed. $C_{\tilde{f}\tilde{s}}(t)$ closely resembles $C_{\tilde{f}\tilde{f}}(t)$ and has its peak very near zero lag, indicating that the distortion due to interference by other modes is small. The envelope of $C_{\tilde{f}\tilde{s}}(t)$, on the other hand, is shifted towards positive lag by about 120 s by the reduced inner-core shear velocity. Although at the time of this writing we have not yet computed the Fréchet kernel for the phase-delay functional $\delta\tau_p$ associated with this observation, it appears that measurements based on this isolation filter should yield data that are reasonably sensitive to the inner-core shear velocity.

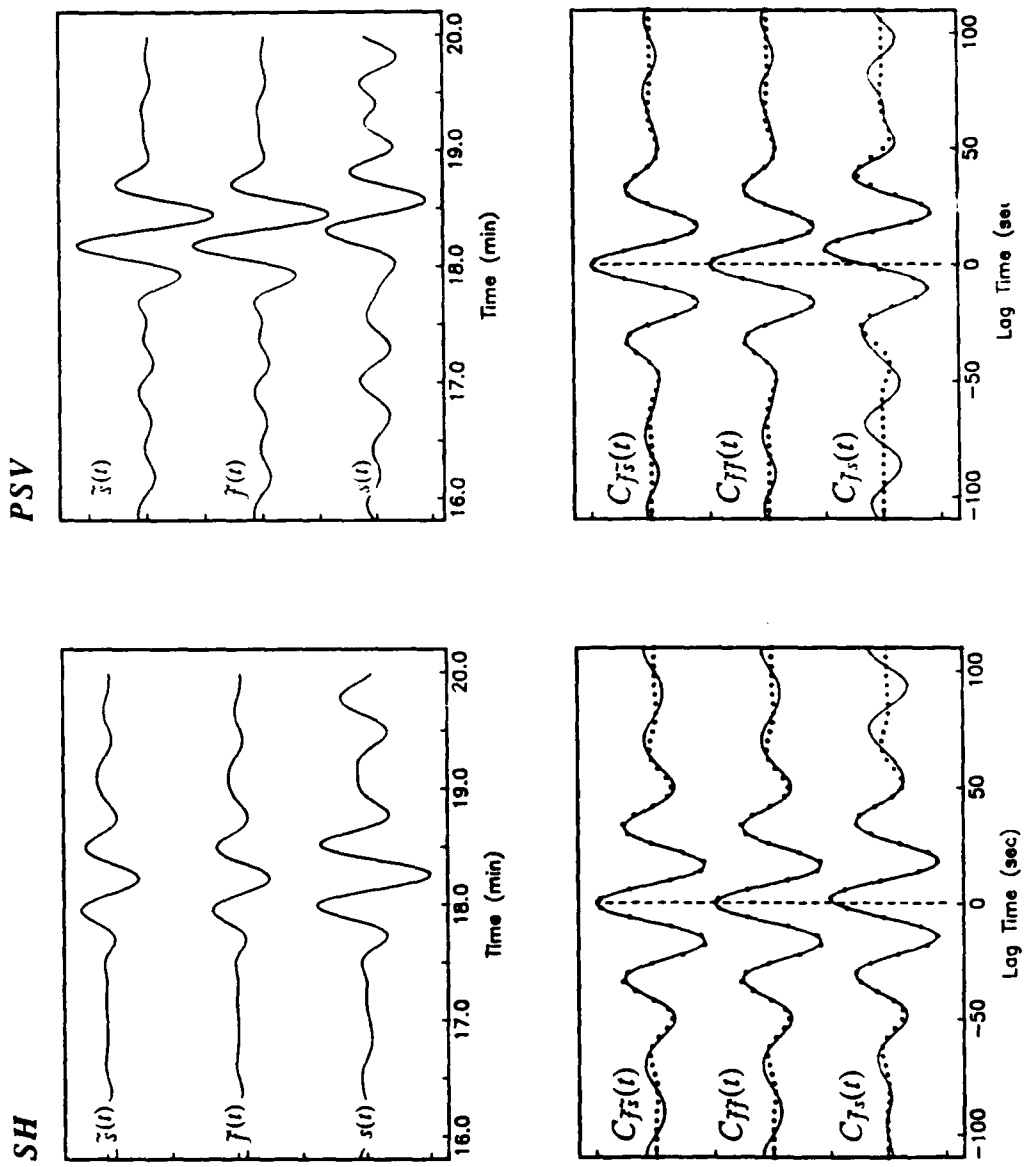


Figure 5.1

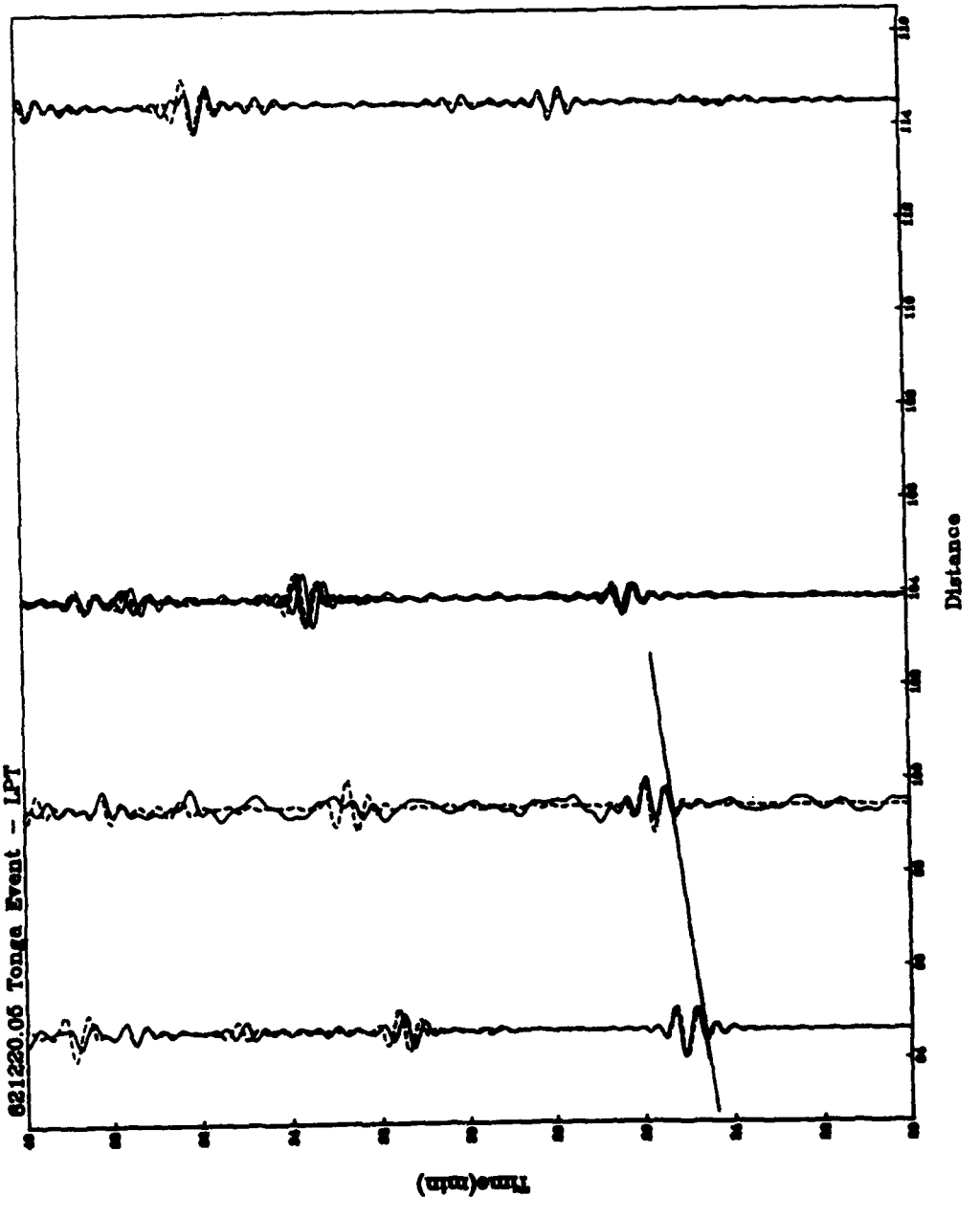


Figure 5.2

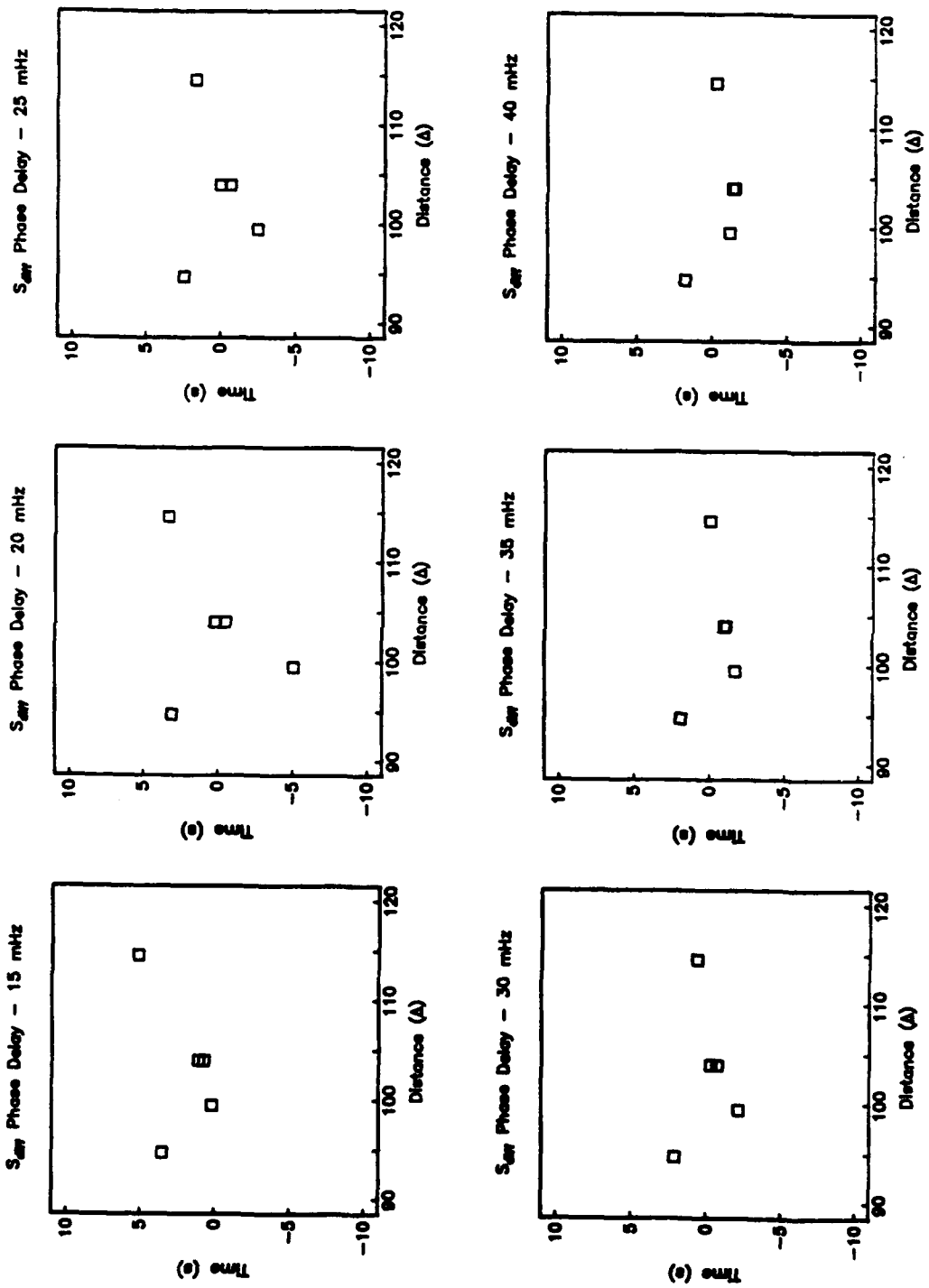

S_{diff} Phase Delay – 25 mHz
S_{diff} Phase Delay – 40 mHz
S_{diff} Phase Delay – 20 mHz
S_{diff} Phase Delay – 35 mHz
S_{diff} Phase Delay – 15 mHz
S_{diff} Phase Delay – 30 mHz
Time (s)
Distance (Å)
10
5
0
−5
−10
90
100
110
120

Figure 5.3

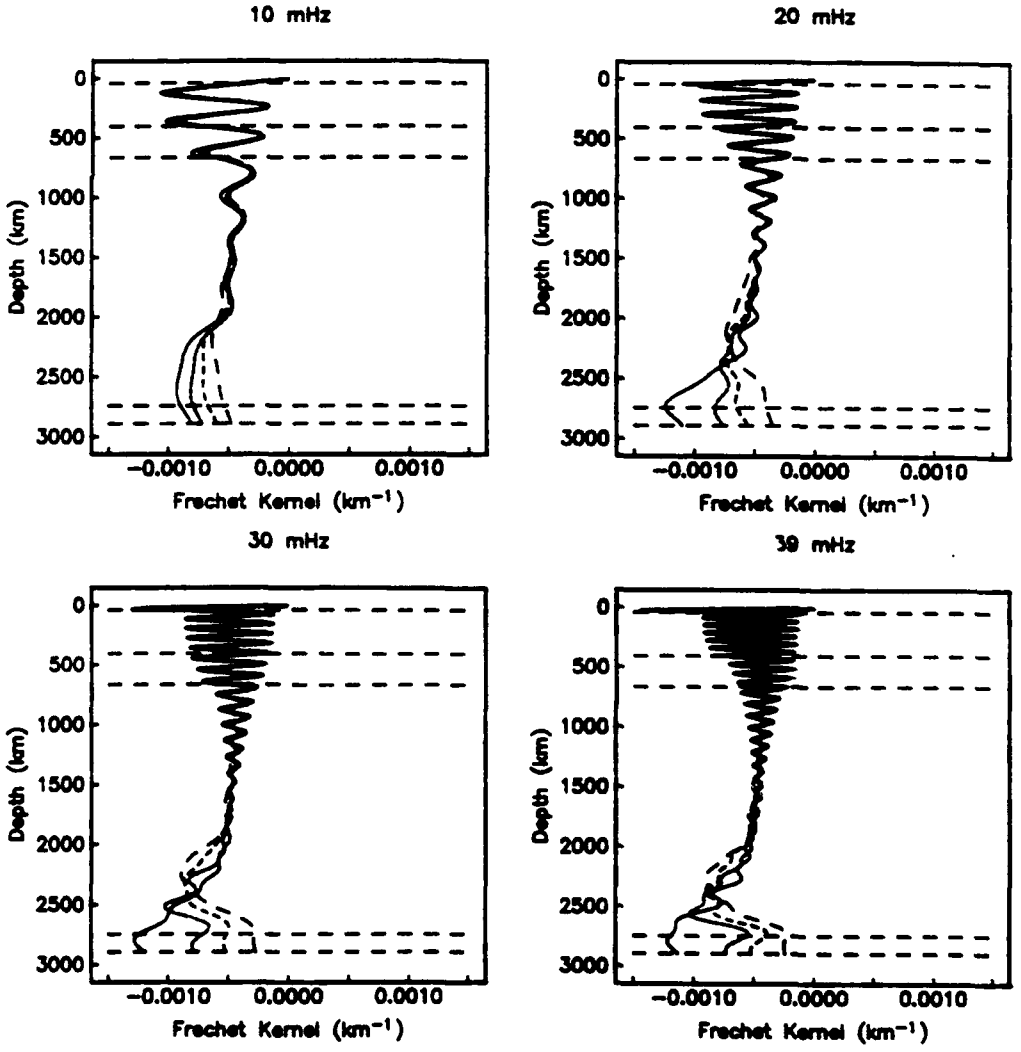

10 mHz
20 mHz
30 mHz
39 mHz
Depth (km)
0
500
1000
1500
2000
2500
3000
−0.0010
0.0000
0.0010
Frechet Kernel (km⁻¹)

Figure 5.4a

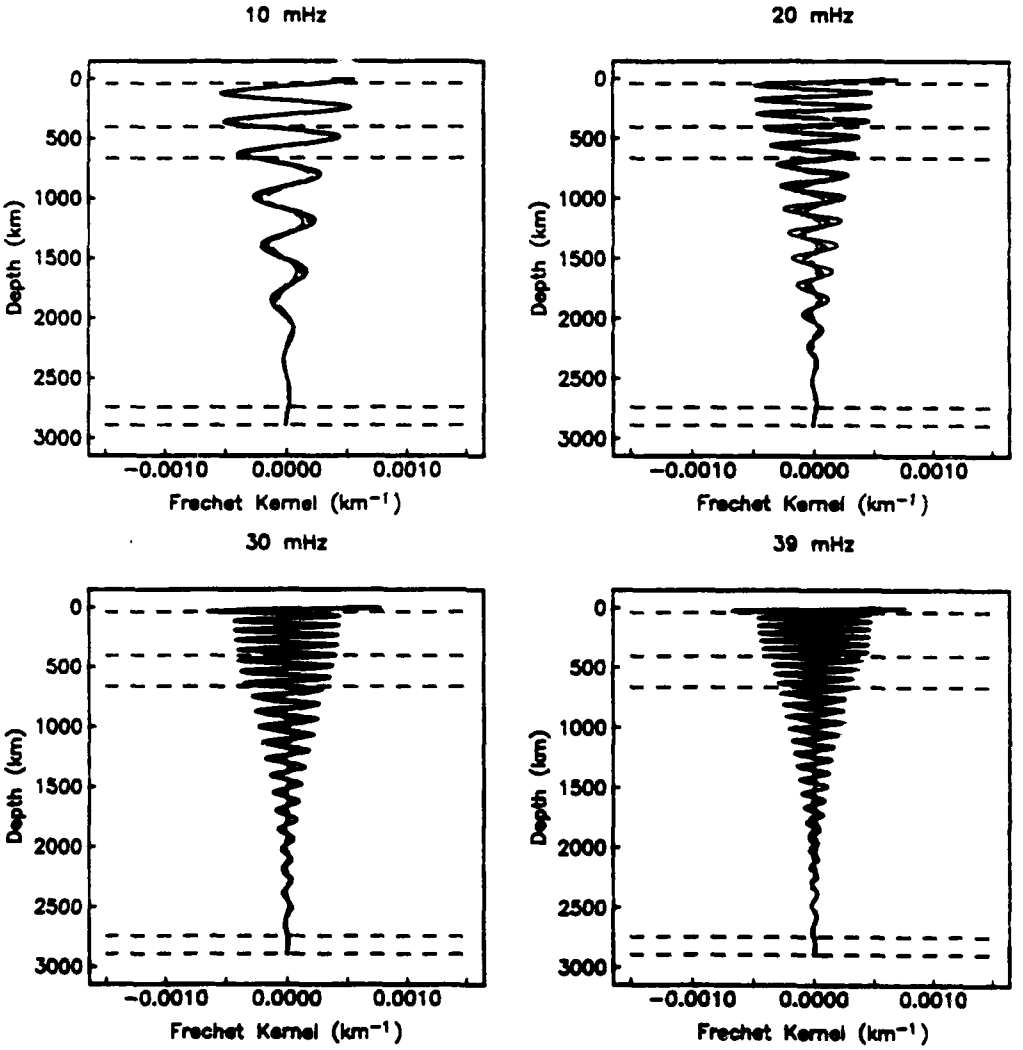

10 mHz
20 mHz
30 mHz
39 mHz
Depth (km)
0
500
1000
1500
2000
2500
3000
−0.0010
0.0000
0.0010
Frechet Kernel (km⁻¹)

Figure 5.4b

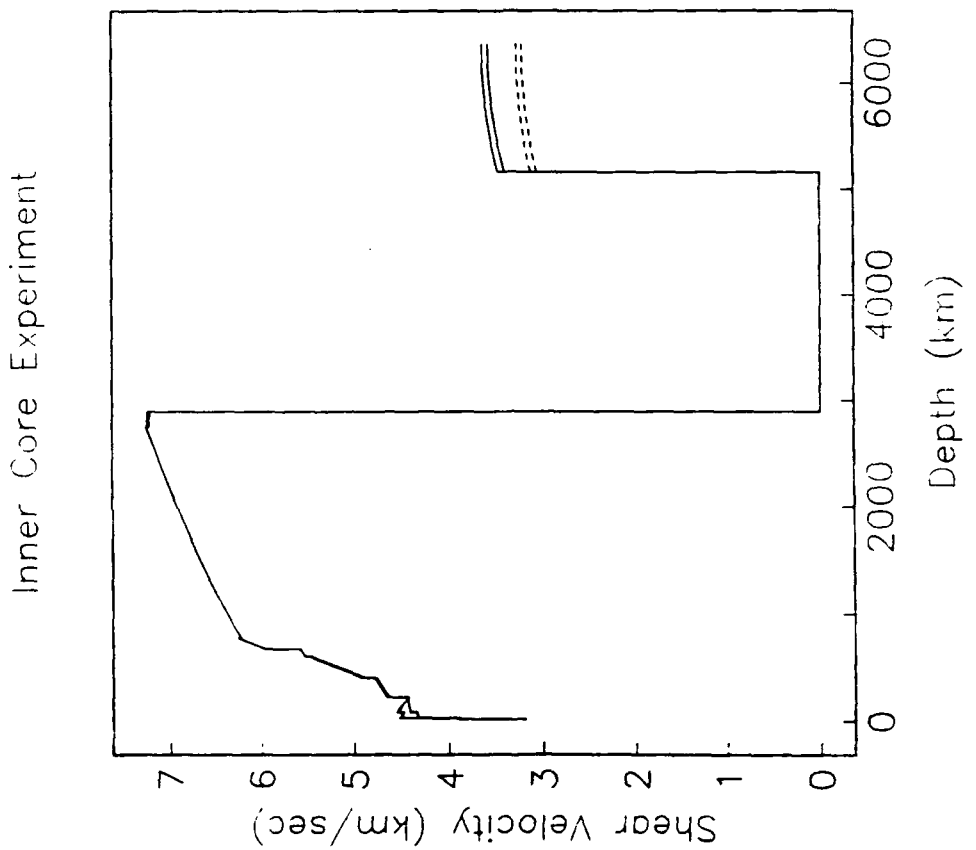

Inner Core Experiment
Shear Velocity (km/sec)
0
1
2
3
4
5
6
7
Depth (km)
0
2000
4000
6000

Figure 5.5

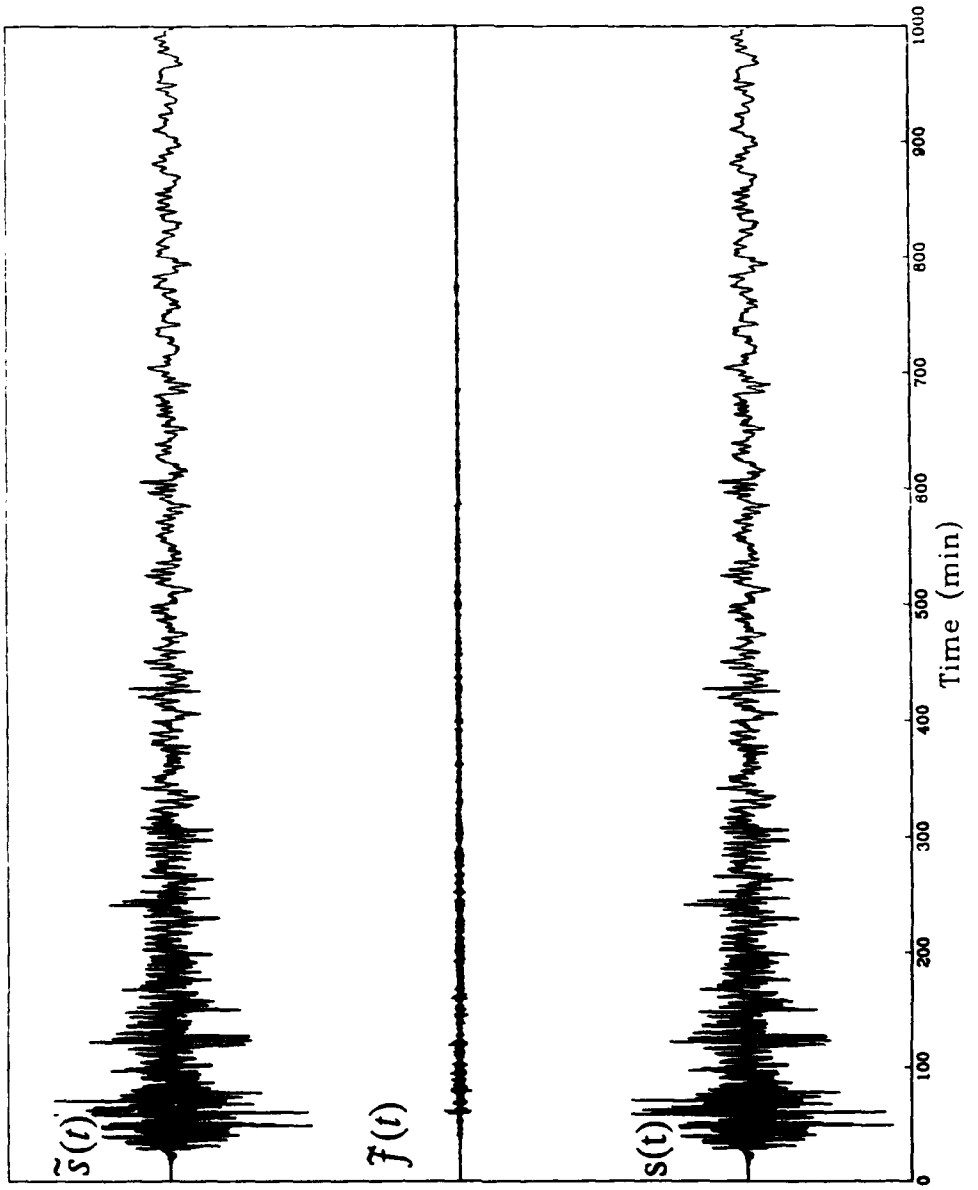

$\tilde{s}(t)$
$\mathcal{F}(t)$
s(t)
0
100
200
300
400
500
600
700
800
900
1000
Time (min)

Figure 5.6a

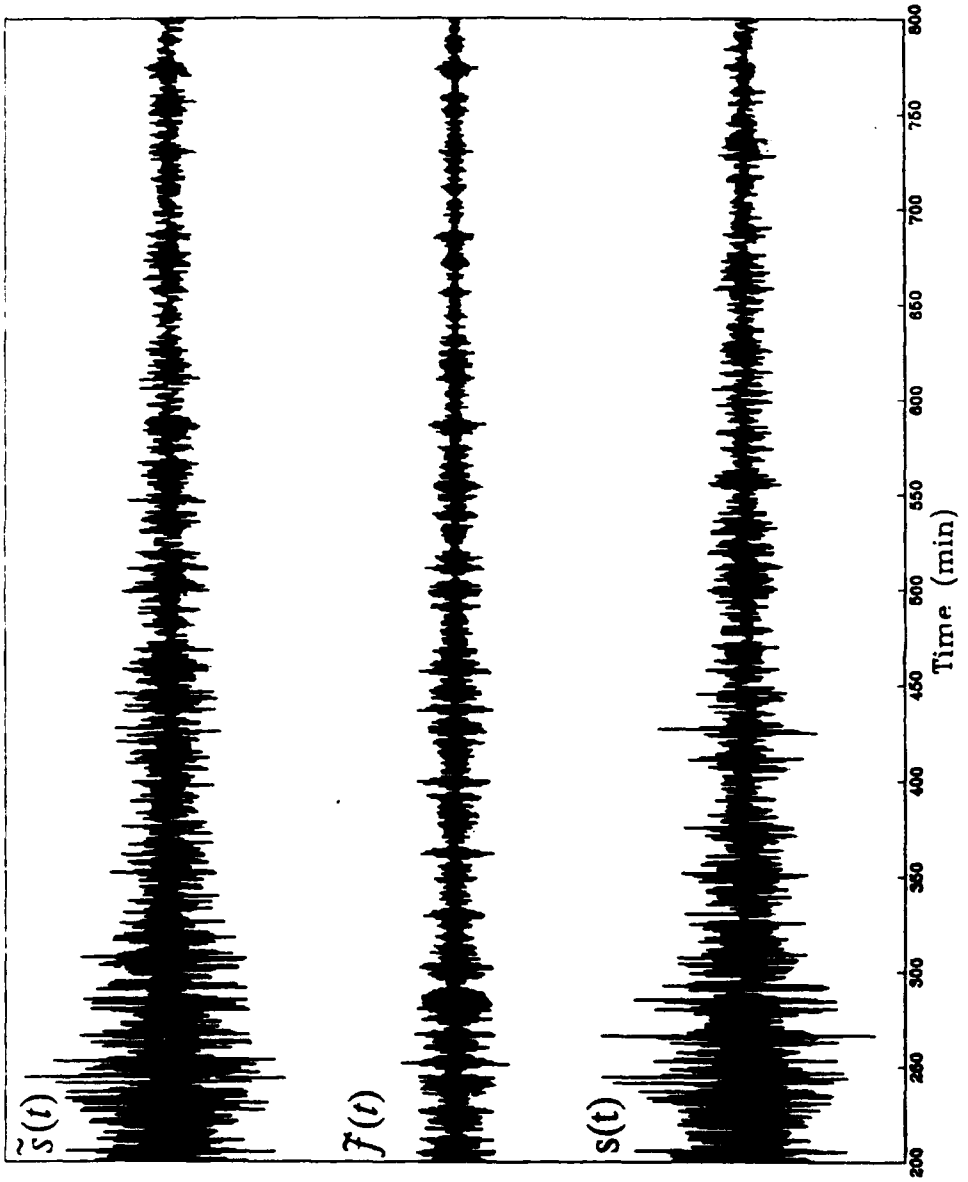

$\tilde{s}(t)$
$\mathcal{F}(t)$
s(t)
200
250
300
350
400
450
500
550
600
650
700
750
800
Time (min)

Figure 5.6b

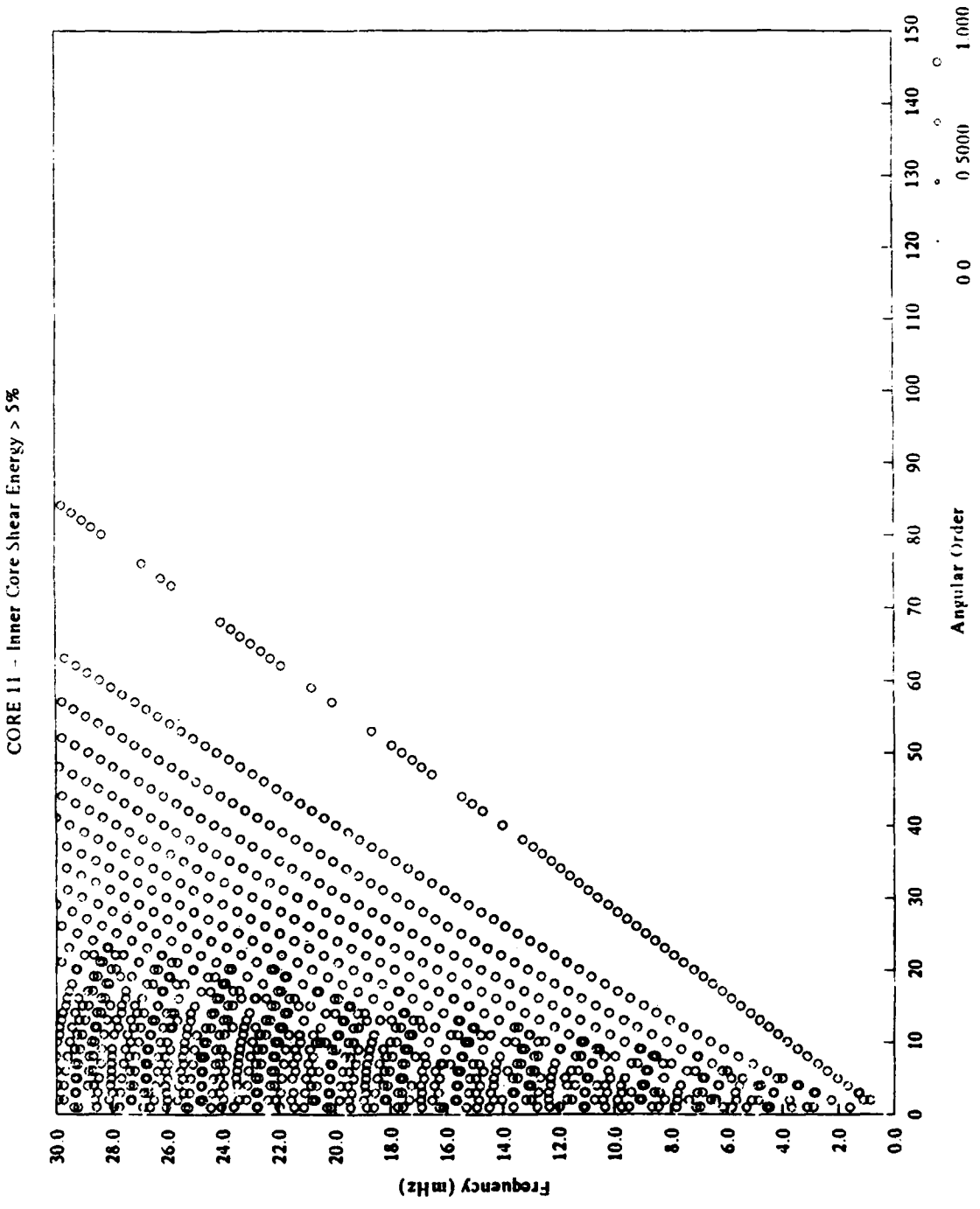

CORE 11 - Inner Core Shear Energy > 5%
Frequency (mHz)
30.0
28.0
26.0
24.0
22.0
20.0
18.0
16.0
14.0
12.0
10.0
8.0
6.0
4.0
2.0
0.0
Angular Order
0
10
20
30
40
50
60
70
80
90
100
110
120
130
140
150
0.0
0.5000
1.000

Figure 5.7

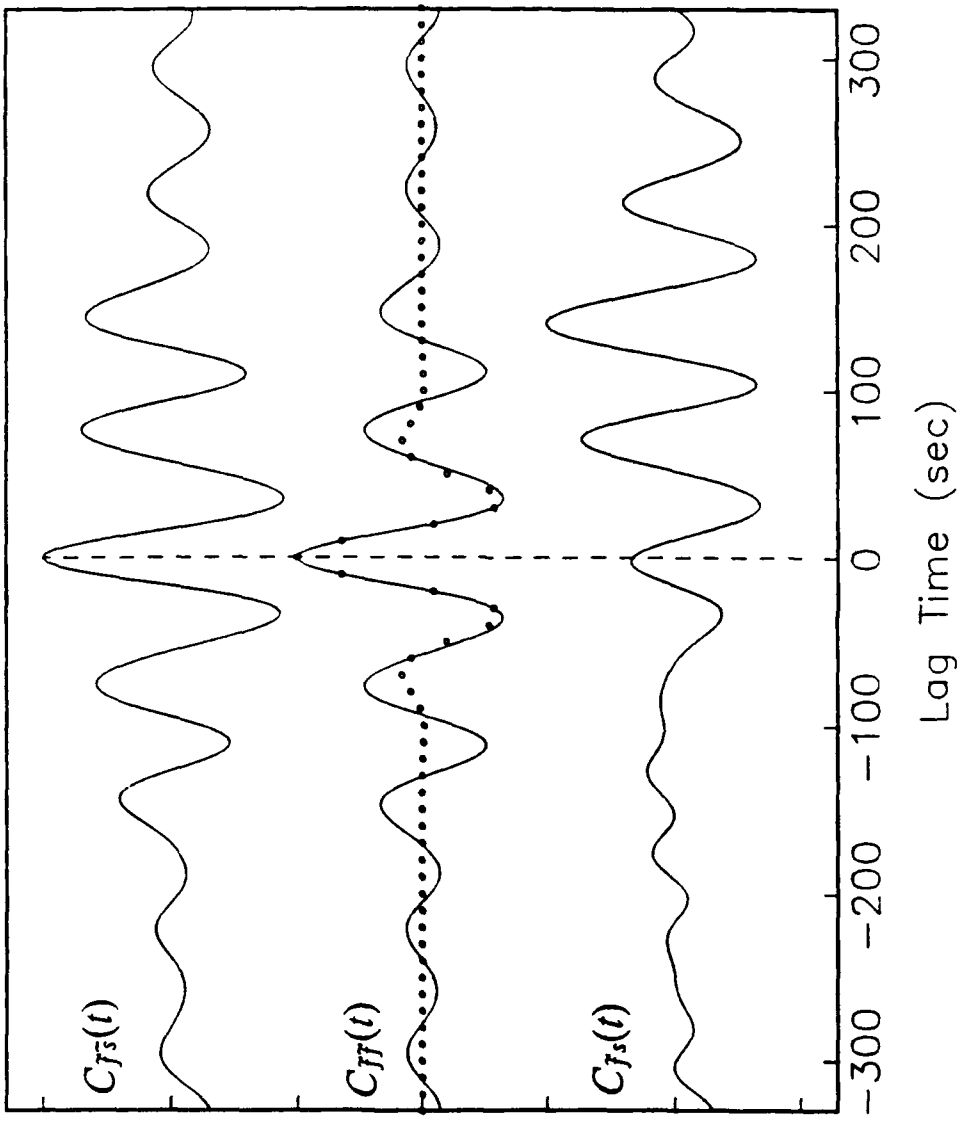


Figure 5.8a

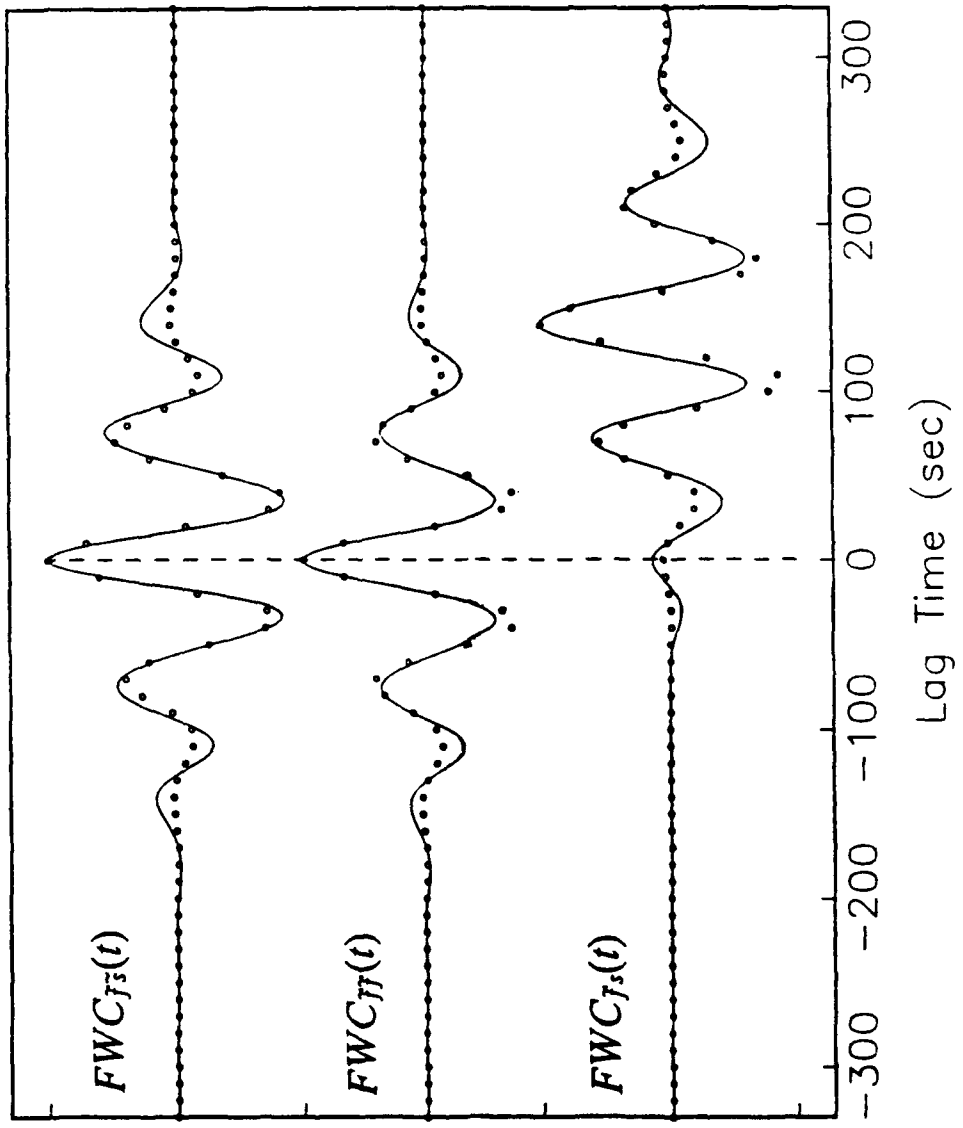

$FWC_{\bar{f}s}(t)$
$FWC_{\bar{f}\bar{f}}(t)$
$FWC_{fs}(t)$
-300
-200
-100
0
100
200
300
Lag Time (sec)

Figure5.8b

APPENDIX A

PROPERTIES OF HERMITE-POLYNOMIAL EXPANSIONS

> Infinite series techniques are a good psychological restorative because they can give a scientist who is stalled on a problem a sense of forward progress as he or she computes the terms.
>
> C.M. Bender and S.A. Orszag, 1978

INTRODUCTION

Hermite polynomials are a member of the family of classical orthogonal polynomials which includes Chebyshev, Gegenbauer, Jacobi, Laguerre, and Legendre. Although they bear the name of Charles Hermite [1864], they were first used by Laplace in 1805 in his *Treatise on Celestial Mechanics*. General discussions of their definition and properties may be found in Hille [1926], Szegö [1926], Palamà [1937], Sansone [1959], and Lebedev [1965]. Hermite polynomials occur in mathematical physics as solutions to time-independent Schrödinger equation for a simple harmonic oscillator and in statistics for the representation of distributions over the interval $[-\infty, \infty]$. In the following sections, we shall present some useful properties of Hermite polynomials and develop the representation of real functions using the Gram-Charlier expansion.

PROPERTIES OF HERMITE POLYNOMIALS

Definition

Hermite polynomials may be derived from the coefficients of the power series expansion of $e^{-t^2/2 + xt}$ in t:

$$e^{-t^2/2 + xt} = e^{x^2/2} \sum_{k=0}^{\infty} \frac{t^k}{k!} \left[\frac{d^k}{dt^k} e^{-(t-x)^2/2} \right]_{t=0}$$

$$= e^{x^2/2} \sum_{k=0}^{\infty} \frac{t^k}{k!} (-1)^k \left[\frac{d^k}{dx^k} e^{-x^2/2} \right]$$

$$= \sum_{k=0}^{\infty} \frac{t^k}{k!} \mathrm{He}_k(x) \qquad \text{(A.1)}$$

$\mathrm{He}_k(x)$ is the Hermite polynomial of degree k and $e^{-t^2/2 + xt}$ is known as the generating function. The $\mathrm{He}_k(x)$ form a complete set over the interval $[-\infty, \infty]$:

$$\mathrm{He}_k(x) \equiv e^{x^2/2} (-1)^k \left[\frac{d^k}{dx^k} e^{-x^2/2} \right] = k! \sum_{j=0}^{[k/2]} \frac{(-1/2)^j \, x^{k-2j}}{j! \, (k-2j)!} \qquad \text{(A.2)}$$

where $[k/2]$ is the largest integer $\leq k/2$, and are orthogonal with respect to a Gaussian weight function:

$$\int_{-\infty}^{\infty} \mathrm{Ga}(x) \, \mathrm{He}_k(x) \, \mathrm{He}_m(x) \, dx = \sqrt{2\pi} \, k! \, \delta_{km} \qquad \text{(A.3)}$$

where $\mathrm{Ga}(x)$ is defined:

$$\mathrm{Ga}(x) \equiv e^{-x^2/2} \qquad \text{(A.4)}$$

The first few polynomials are summarized in Table A.1 and plotted in Figure A.1. The $He_k(x)$ are related to the $H_k(x)$, which are also known are Hermite polynomials, through a simple scaling [Abramowitz and Stegun, 1972]:

$$He_k(x) = 2^{-k/2} H_k(x/\sqrt{2}) \tag{A.5}$$

The $H_k(x)$ nomenclature is more common in mathematical physics; the $He_k(x)$ designation is more prevalent in probability and statistics.

Symmetry theorem

A Hermite polynomial is even or odd depending on degree:

$$He_k(-x) = (-1)^k He_k(x) \tag{A.6}$$

Proof: let $x = -y$ in (A.2)

$$\begin{aligned} He_k(-y) &= k! \sum_{j=0}^{[k/2]} \frac{(-1/2)^j (-y)^{k-2j}}{j!\,(k-2j)!} \\ &= (-1)^k\, k! \sum_{j=0}^{[k/2]} \frac{(-1/2)^j\, y^{k-2j}}{j!\,(k-2j)!} \\ &= (-1)^k He_k(y) \end{aligned}$$

The parity of the Hermite polynomials is demonstrated in Figure A.1.

Argument shift theorem

A Hermite polynomial may be shifted with respect to a location parameter, x_0:

$$\mathrm{He}_k(x-x_0) = \sum_{m=0}^{k} \binom{k}{m} (-x_0)^m \, \mathrm{He}_{k-m}(x) \tag{A.7}$$

Proof: let $x = y - y_0$ in (A.2)

$$\begin{aligned}
\mathrm{He}_k(y-y_0) &= k! \sum_{j=0}^{[k/2]} \frac{(-1/2)^j (y-y_0)^{k-2j}}{j!\,(k-2j)!} \\
&= k! \sum_{j=0}^{[k/2]} \frac{(-1/2)^j}{j!\,(k-2j)!} \sum_{m=0}^{k-2j} \binom{k-2j}{m} (-y_0)^m \, y^{k-m-2j} \\
&= \sum_{m=0}^{k} \binom{k}{m} (-y_0)^m \sum_{j=0}^{[(k-m)/2]} \frac{(-1/2)^j \, y^{k-m-2j}}{j!\,(k-m-2j)!} \\
&= \sum_{m=0}^{k} \binom{k}{m} (-y_0)^m \, \mathrm{He}_{k-m}(y)
\end{aligned}$$

In other words, the shifted Hermite kth polynomial depends on the sum over all the unshifted polynomials up to degree k. More generally, it may be shown:

$$\mathrm{He}_k(y-y_0) = \sum_{m=0}^{k} \binom{k}{m} (y_1-y_0)^m \, \mathrm{He}_{k-m}(y-y_1) \tag{A.8}$$

A corollary of the shift theorem states:

$$\sum_{l=0}^{\infty} c_l \mathrm{He}_l(y-y_1) = \sum_{k=0}^{\infty} d_k \, \mathrm{He}_k(y-y_0) \tag{A.9}$$

where

$$c_l = \sum_{m=0}^{\infty} \binom{m+l}{m} d_{m+l} (y_1 - y_0)^m \tag{A.10}$$

According to this corollary, a sum over Hermite polynomials may be rewritten in terms of a sum over shifted Hermite polynomials, with coefficients which have a power series dependence on the shifted location parameter.

Argument scale theorem

A Hermite polynomial may be adjusted with respect to a scale parameter, a:

$$\mathrm{He}_k\left(\frac{x}{a}\right) = k! \frac{1}{a^k} \sum_{m=0}^{[k/2]} \frac{(-1/2)^m (a^2-1)^m}{m!\,(k-2m)!} \mathrm{He}_{k-2m}(x) \tag{A.11}$$

Proof: let $x = y/a$ in (A.2)

$$\begin{aligned}
\mathrm{He}_k\left(\frac{y}{a}\right) &= k! \sum_{j=0}^{[k/2]} \frac{(-1/2)^j}{j!\,(k-2j)!} \left(\frac{y}{a}\right)^{k-2j} \\
&= k! \frac{1}{a^k} \sum_{j=0}^{[k/2]} \frac{(-1/2)^j y^{k-2j}}{j!\,(k-2j)!} \sum_{m=0}^{j} \binom{j}{m} (a^2-1)^m \\
&= k! \frac{1}{a^k} \sum_{m=0}^{[k/2]} \frac{(-1/2)^m (a^2-1)^m}{m!} \sum_{j=0}^{[(k-2m)/2]} \frac{(-1/2)^j y^{k-2m-2j}}{j!\,(k-2m-2j)!} \\
&= k! \frac{1}{a^k} \sum_{m=0}^{[k/2]} \frac{(-1/2)^m (a^2-1)^m}{m!\,(k-2m)!} \mathrm{He}_{k-2m}(y)
\end{aligned}$$

In other words, the rescaled Hermite kth polynomial depends on the sum of the unscaled polynomials up to degree k. More generally, it may be shown:

$$\mathrm{He}_k\left(\frac{y}{a}\right) = k!\left(\frac{b}{a}\right)^k \sum_{m=0}^{[k/2]} \frac{(-1/2)^m}{m!\,(k-2m)!}\left(\frac{a^2-b^2}{b^2}\right)^m \mathrm{He}_{k-2m}\left(\frac{y}{b}\right) \qquad \text{(A.12)}$$

A corollary of the scale theorem states:

$$\sum_{l=0}^{\infty} c_l \mathrm{He}_l\left(\frac{y}{b}\right) = \sum_{k=0}^{\infty} d_k \mathrm{He}_k\left(\frac{y}{a}\right) \qquad \text{(A.13)}$$

where

$$c_l = \left(\frac{b}{a}\right)^l \sum_{m=0}^{\infty} d_{2m+l} \frac{(2m+l)!\,(-1/2)^m}{k!\,m!}\left(\frac{a^2-b^2}{a^2}\right)^m \qquad \text{(A.14)}$$

Similar to the shift theorem, the scale theorem states that a sum over Hermite polynomials may be rewritten as a sum over scaled Hermite polynomials, with coefficients which have a power-series dependence on the scale parameters.

Fourier transform theorem

The Hermite polynomials have the following Fourier transform pair [Hille, 1926]:

$$\int_{-\infty}^{\infty} \mathrm{Ga}\,(t)\,(-it)^k\,\mathrm{e}^{i\omega t}\,dt = \sqrt{2\pi}\,\mathrm{Ga}\,(\omega)\,\mathrm{He}_k(\omega) \qquad \text{(A.15)}$$

$$\frac{1}{2\pi}\int_{-\infty}^{\infty} \mathrm{Ga}\,(\omega)\,\mathrm{He}_k(\omega)\,\mathrm{e}^{-i\omega t}\,d\omega = \frac{1}{\sqrt{2\pi}}\mathrm{Ga}\,(t)\,(-it)^k \qquad \text{(A.16)}$$

where Ga(ω) is not the Fourier transform of Ga(t), but is defined according to (A.4).

Summary

We have introduced Hermite polynomials and presented some of their properties. In particular, we have enumerated the symmetry, shift, and scale theorems, as well as the Fourier transform pair. The shift and scale theorems will be used extensively in Appendix B; the Fourier transform theorem will be applied throughout the thesis. We shall now demonstrate the usefulness of Hermite polynomials in the representation of real functions.

REPRESENTATION OF REAL FUNCTIONS: GRAM-CHARLIER SERIES

Consider the inverse Fourier transform of a real function $g(t)$

$$g(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} g(\omega)\, e^{-i\omega t}\, d\omega \tag{A.17}$$

Since $g(t)$ is real, $g(\omega)$ will have Hermitian symmetry, $g(-\omega) = g^*(\omega)$ (where $*$ denotes complex conjugate), and we may write

$$g(t) = \frac{1}{2\pi} \mathrm{Re}\, \{ \int_{-\infty}^{\infty} g(\omega)\, \mathrm{H}(\omega)\, e^{-i\omega t}\, d\omega \} \tag{A.18}$$

where $\mathrm{H}(\omega)$ is the Heaviside step function. We expand $g(\omega)$ along the positive axis in terms of an unnormalized Gaussian multiplied by a series of Hermite polynomials:

$$g(\omega)\, \mathrm{H}(\omega) = \frac{1}{\sqrt{2\pi}\, \sigma_g} \mathrm{Ga}\left(\frac{\omega-\omega_g}{\sigma_g}\right) \sum_{k=0}^{\infty} g_k\, \mathrm{He}_k\left(\frac{\omega-\omega_g}{\sigma_g}\right) \tag{A.19}$$

where ω_g is a location parameter and σ_g is a scale parameter. In statistics, an expansion of this form is known as a Gram-Charlier series [Jackson, 1961; Rietz, 1971]. The coefficients may be determined by multiplying each side of (A.19) by $\mathrm{He}_m((\omega-\omega_g)/\sigma_g)$, integrating over the interval $[-\infty, \infty]$, and applying the orthogonality condition:

$$g_k = \frac{1}{k!}\int_0^{\infty} g(\omega)\,\mathrm{He}_k\left(\frac{\omega-\omega_g}{\sigma_g}\right)d\omega \tag{A.20}$$

Substituting the definition of the Hermite polynomials (A.2) into this expression, we see that the coefficients depend on the spectral moments of $g(\omega)$

$$g_k = \sum_{m=0}^{[k/2]} \frac{(-1/2)^m\,\hat{\mu}_{k-2m}(\omega_g)}{m!\,(k-2m)!} \tag{A.21}$$

where $\hat{\mu}_p(\omega_g)$ is the pth normalized one-sided moment of $g(\omega)$ about $\omega = \omega_g$

$$\hat{\mu}_p(\omega_g) = \frac{1}{\sigma_g^p}\int_0^{\infty} g(\omega)\,(\omega-\omega_g)^p\,d\omega \tag{A.22}$$

Table A.2 contains expressions for the first seven coefficients. If the spectrum is normalized such that $\hat{\mu}_0(0) = 1$, then $g_0 = 1$. If the location parameter ω_g is chosen to be the center frequency of $g(\omega)$ (*i.e.*, $\omega_g = \sigma_g\,\hat{\mu}_1(0)$), then $g_1 = 0$. Finally, if the scale parameter σ_g is the half-width of $g(\omega)$ (*i.e.*, $1 = \hat{\mu}_2(\omega_g)$), then $g_2 = 0$. With this parameterization, the zeroth-order term of the expansion is the Gaussian approximation of $g(\omega)$ and the first correction term is third order. We assume this representation in the thesis.

An alternative derivation for the expansion coefficients is based on the χ^2 minimization criterion [Rietz, 1971]:

$$\chi^2 = \int_{-\infty}^{\infty} \mathrm{Ga}\left(\frac{\omega-\omega_g}{\sigma_g}\right)^{-1} \left[\, g(\omega) - \tilde{g}_K\left(\frac{\omega-\omega_g}{\sigma_g}\right)\right]^2 d\omega \tag{A.23}$$

where $\tilde{g}_K((\omega-\omega_g)/\sigma_g)$ represents the Kth partial sum of the Gram-Charlier series:

$$\tilde{g}_K\left(\frac{\omega-\omega_g}{\sigma_g}\right) = \frac{1}{\sqrt{2\pi}\,\sigma_g}\,\mathrm{Ga}\left(\frac{\omega-\omega_g}{\sigma_g}\right) \sum_{k=0}^{K} g_k\,\mathrm{He}_k\left(\frac{\omega-\omega_g}{\sigma_g}\right) \tag{A.24}$$

By taking the derivative of (A.23) with respect to g_k and setting $\partial\chi^2/\partial g_k = 0$, we recover (A.20).

In order to illustrate the uniqueness of the coefficients of a Hermite-polynomial expansion, consider an expansion of $g(\omega)$ with the same center frequency and half-width, but with different coefficients. Such an expansion might be derived by filtering or windowing another function:

$$g(\omega)\,\mathrm{H}(\omega) = \frac{1}{\sqrt{2\pi}\,\sigma_g}\,\mathrm{Ga}\left(\frac{\omega-\omega_g}{\sigma_g}\right) \sum_{l=0}^{\infty} f_l \mathrm{He}_l\left(\frac{\omega-\omega_g}{\sigma_g}\right) \tag{A.25}$$

and the f_k may depend on many other parameters. From (A.20), we know

$$g_k = \frac{1}{k!}\int_{0}^{\infty} g(\omega)\,\mathrm{He}_k\left(\frac{\omega-\omega_g}{\sigma_g}\right) d\omega \tag{A.26}$$

If we substitute our new expansion for $g(\omega)$ into this expression and apply the orthogonality condition:

$$g_k = \frac{1}{k!}\frac{1}{\sqrt{2\pi}\,\sigma_g}\int_{-\infty}^{\infty} \mathrm{Ga}\left(\frac{\omega-\omega_g}{\sigma_g}\right)\mathrm{He}_k\left(\frac{\omega-\omega_g}{\sigma_g}\right)\sum_{l=0}^{\infty} f_l \mathrm{He}_l\left(\frac{\omega-\omega_g}{\sigma_g}\right)d\omega$$

$$= f_k \qquad \text{(A.27)}$$

Thus, the coefficients of a Hermite-polynomial expansion are uniquely determined.

Figure A.2 compares $g(\omega)$ (solid line) with the partial sums of the Gram-Charlier series (dashed line) for K from 2 to 5 in the case of the autocorrelation of the ASRO impulse response. Although the spectrum is strongly peaked, it is asymmetrical. The second partial sum produces the best-fitting Gaussian to $g(\omega)$ by matching the first three moments. The addition of the third-order term improves the fit by shifting the peak and tailoring the flanks. The fourth-order term does not change the partial sum noticeably as most of the misfit is in the odd-order terms. With the addition of the fifth term, the fit on the flanks of the autocorrelation function is better than the fit to the peak. Figure A.3 displays the residual $(g(\omega) - \tilde{g}_K((\omega-\omega_g)/\sigma_g))$ for the first five terms in detail. While the addition of the third-order term reduces the residual everywhere, the fourth and fifth residuals illustrate that the Gram-Charlier series is fitting the flanks of the spectrum as well as the peak. This may be seen from the expression for the χ^2 minimization criterion (A.23), where the inverse Gaussian factor results in an equal weighting of the spectrum. Figures A.4 and A.5 present similar plots for the autocorrelation of the fundamental-mode Love wave displayed in Figure 2.2. This function is more narrow-band and more symmetrical than the autocorrelation of the ASRO impulse response. It is clear from Figure A.4 that the autocorrelation of the Love wave is well-fit by the third partial sum. The

addition of the fourth term improves the fit to the peak, while the addition of the fifth term minimizes the misfit at the flanks at the cost of degrading the fit near the peak. Figure A.5 emphasizes the point that the χ^2 minimization criterion requires that the fit improve uniformly across the band.

Using our Gram-Charlier expansion, the integral for $g(t)$ may be written:

$$g(t) = \frac{1}{2\pi}\frac{1}{\sqrt{2\pi}\,\sigma_g}\,\mathrm{Re}\left\{\int_{-\infty}^{\infty} \mathrm{Ga}\left(\frac{\omega-\omega_g}{\sigma_g}\right)\sum_{k=0}^{\infty} g_k\,\mathrm{He}_k\left(\frac{\omega-\omega_g}{\sigma_g}\right)e^{-i\omega t}\,d\omega\right\} \quad (A.28)$$

One of the advantages of representing a function in terms of Hermite polynomials is ease with which one may move between the time and frequency domains. We may solve this integral using the Fourier transform theorem (A.15-A.16):

$$g(t) = \frac{1}{2\pi}\,\mathrm{Ga}\,(\sigma_g t)\;\mathrm{Re}\left\{\,e^{-i\omega_g t}\sum_{k=0}^{\infty} g_k\;\gamma_g^k\;(-i\omega_g t)^k\,\right\} \quad (A.29)$$

where $\gamma_g = \sigma_g / \omega_g$ is a measure of the relative bandwidth of $g(\omega)$. (A.29) is the canonical form of a Gram-Charlier representation of a real time series and recurs frequently in this thesis.

If $g(t)$ is an even function, then the coefficients g_k are real:

$$\begin{aligned} g(t) &= \frac{1}{2\pi}\,\mathrm{Ga}\,(\sigma_g t)\sum_{k=0}^{\infty} g_k\,\gamma_g^k\,(\omega_g t)^k\cos\left(\omega_g t + k\frac{\pi}{2}\right) \\ &= \frac{1}{2\pi}\,\mathrm{Ga}\,(\sigma_g t)\;\left\{\cos(\omega_g t)\,[1 + g_4\,\gamma_g^4\,(\omega_g t)^4 - g_6\,\gamma_g^6\,(\omega_g t)^6 + \cdots]\right. \\ &\qquad \left. + \sin(\omega_g t)\,[g_3\,\gamma_g^3\,(\omega_g t)^3 - g_5\,\gamma_g^5\,(\omega_g t)^5 + \cdots]\,\right\} \end{aligned} \quad (A.30)$$

where we have normalized the spectrum and chosen ω_g and σ_g such that $g_0 = 1$ and $g_1 = g_2 = 0$. The terms multiplying the cosine represent corrections to the envelope, while the sine terms are corrections to the phase.

Figure A.6 compares $g(t)$ (solid line) with the Gram-Charlier partial sums (dashed line) for K of 2 through 5 for the example of the autocorrelation of the ASRO impulse response. Unlike the expansions in the spectral domain, the addition of terms in the time-domain improves the fit in a localized region. This may be seen more clearly in Figure A.7, which displays the residual for the second through fifth partial sums. The residual decreases rapidly near zero lag with the addition of the higher order terms. Figures A.8 and A.9 are parallel plots for the autocorrelation of the Love wave. The improvement of the fit to $g(t)$ near zero lag with the addition of higher order terms may be understood in terms of the Taylor series expansion of $g(t)$ about zero lag. The relationship between a Taylor series expansion in the time domain and the spectral moments is well-known in time series analysis: an expansion of $g(t)$ in terms of its low order moments in the frequency domain is equivalent to its Taylor series expansion about zero in the time domain.

$$
\begin{aligned}
g(t) &= g(0) + \dot{g}(0)\, t + \ddot{g}(0)\, \frac{t^2}{2} + \cdots \\
&= \frac{1}{2\pi} \mathrm{Re}\, \{ \hat{\mu}_0(0) + \hat{\mu}_1(0)\, \gamma_g\, (-i\omega_g t) + \hat{\mu}_2(0)\, \gamma_g^2\, \frac{(-i\omega_g t)^2}{2} + \cdots \} \\
&= \frac{1}{2\pi} \mathrm{Re}\, \{ e^{-i\omega_g t} [\hat{\mu}_0(\omega_g) + \hat{\mu}_1(\omega_g)\, \gamma_g\, (-i\omega_g t) + \hat{\mu}_2(\omega_g)\, \gamma_g^2\, \frac{(-i\omega_g t)^2}{2} + \cdots] \}
\end{aligned}
\qquad \text{(A.31)}
$$

While the Gram-Charlier series is not equivalent to the Taylor series expansion, it is a series whose coefficients are determined by the one-sided spectral moments of the function. The addition of higher order terms consequently improves the fit near zero lag.

CONCLUSIONS

In this Appendix we have introduced Hermite polynomials and discussed some of their useful properties. Furthermore, we have examined in detail the Gram-Charlier expansion of a real function. We have seen that the coefficients of the expansion are uniquely determined by the one-sided moments of $g(\omega)$ and that the time-domain image may be written as a power series in γ_g, where γ_g is a measure of the relative bandwidth of $g(\omega)$. Finally, we have illustrated that the addition of higher order terms in the frequency domain improves the fit uniformly across the band, which proceeds from the χ^2 minimization criterion, while enhancing the fit near zero in the time domain, which may be understood in terms of the Taylor series expansion of $g(t)$ about zero. This property makes Gram-Charlier expansions ideal for representing correlation functions.

TABLES

TABLE A.1: HERMITE POLYNOMIALS OF DEGREE 0 – 6

k	$\mathrm{He}_k(x)$
0	1
1	x
2	$x^2 - 1$
3	$x^3 - 3x$
4	$x^4 - 6x^2 + 3$
5	$x^5 - 10x^3 + 15x$
6	$x^6 - 15x^4 + 45x^2 - 15$

TABLE A.2 - COEFFICIENTS OF THE GRAM-CHARLIER SERIES FOR $k = 0 - 6$

k	g_k
0	$\hat{\mu}_0(\omega_g)$
1	$\hat{\mu}_1(\omega_g)$
2	$\frac{1}{2!}[\hat{\mu}_2(\omega_g) - \hat{\mu}_0(\omega_g)]$
3	$\frac{1}{3!}[\hat{\mu}_3(\omega_g) - 3\hat{\mu}_1(\omega_g)]$
4	$\frac{1}{4!}[\hat{\mu}_4(\omega_g) - 6\hat{\mu}_2(\omega_g) + 3\hat{\mu}_0(\omega_g)]$
5	$\frac{1}{5!}[\hat{\mu}_5(\omega_g) - 10\hat{\mu}_3(\omega_g) + 15\hat{\mu}_1(\omega_g)]$
6	$\frac{1}{6!}[\hat{\mu}_6(\omega_g) - 15\hat{\mu}_4(\omega_g) + 45\hat{\mu}_2(\omega_g) - 15\hat{\mu}_0(\omega_g)]$

The one-sided normalized moments $\hat{\mu}_p(\omega_g)$ are defined by equation (A.22)

FIGURE CAPTIONS

FIGURE A.1

$He_k(x) / k!$ is plotted for degrees 0 to 6.

FIGURE A.2

This figure illustrates the comparison between $g(\omega)$ and $\tilde{g}_K((\omega-\omega_g)/\sigma_g)$ for the autocorrelation of ASRO impulse response. A Hanning filter has been applied between 0 and 5 and between 40 and 50 mHz. Although the spectrum is strongly peaked ($\gamma_g = 0.5$), it is asymmetrical. The second partial sum produces the best-fitting Gaussian to $g(\omega)$ by matching the first three moments. The addition of the third order term improves the fit by shifting the peak and tailoring the flanks. The fourth partial sum does not change the partial sum markedly as most of the misfit is in the odd-order terms. With the addition of the fifth term, the fit on the flanks of the autocorrelation function is better than the fit to the peak.

FIGURE A.3

The residual between $g(\omega)$ and the Gram-Charlier series is plotted for the second through the fifth partial sums The addition of successive terms of the partial sum improves the fit uniformly across the band, a result of the inverse Gaussian weighting in the χ^2 minimization criterion.

FIGURE A.4

This figure illustrates the comparison between $g(\omega)$ and $\tilde{g}_K((\omega-\omega_g)/\sigma_g)$ for the autocorrelation of the fundamental-mode Love wave displayed in Figure 2.1. This function is more narrow-band ($\gamma_g = 0.3$) and more symmetrical than the autocorrelation of the ASRO impulse response. It is clear from this figure that the autocorrelation of the Love wave is well-fit by the third partial sum. The addition of the fourth term improves the fit to the peak, while the addition of the fifth term minimizes the misfit at the flanks at the cost of degrading the fit near the peak.

FIGURE A.5

The residuals between $g(\omega)$ and the Gram-Charlier series are plotted for the second through the fifth partial sums in the case of the autocorrelation of the fundamental-mode Love wave The addition of successive terms of the partial sum improves the fit uniformly across the band, a result of the inverse Gaussian weighting in the χ^2 minimization criterion.

FIGURE A.6

This figure illustrates the comparison between $g(t)$ and Gram-Charlier partial sums for the autocorrelation of the ASRO impulse response. The addition of higher order terms improves the fit near zero lag.

FIGURE A.7

This figure demonstrates the residual between $g(t)$ and Gram-Charlier partial sums for the autocorrelation of the ASRO impulse response. The addition of higher order terms improves the fit near zero lag.

FIGURE A.8

This figure illustrates the comparison between $g(t)$ and Gram-Charlier partial sums for the autocorrelation of the fundamental-mode Love wave. The addition of higher order terms improves the fit near zero lag.

FIGURE A.9

This figure demonstrates the residual between $g(t)$ and Gram-Charlier partial sums for the autocorrelation of the fundamental-mode Love wave. The addition of higher order terms improves the fit near zero lag.

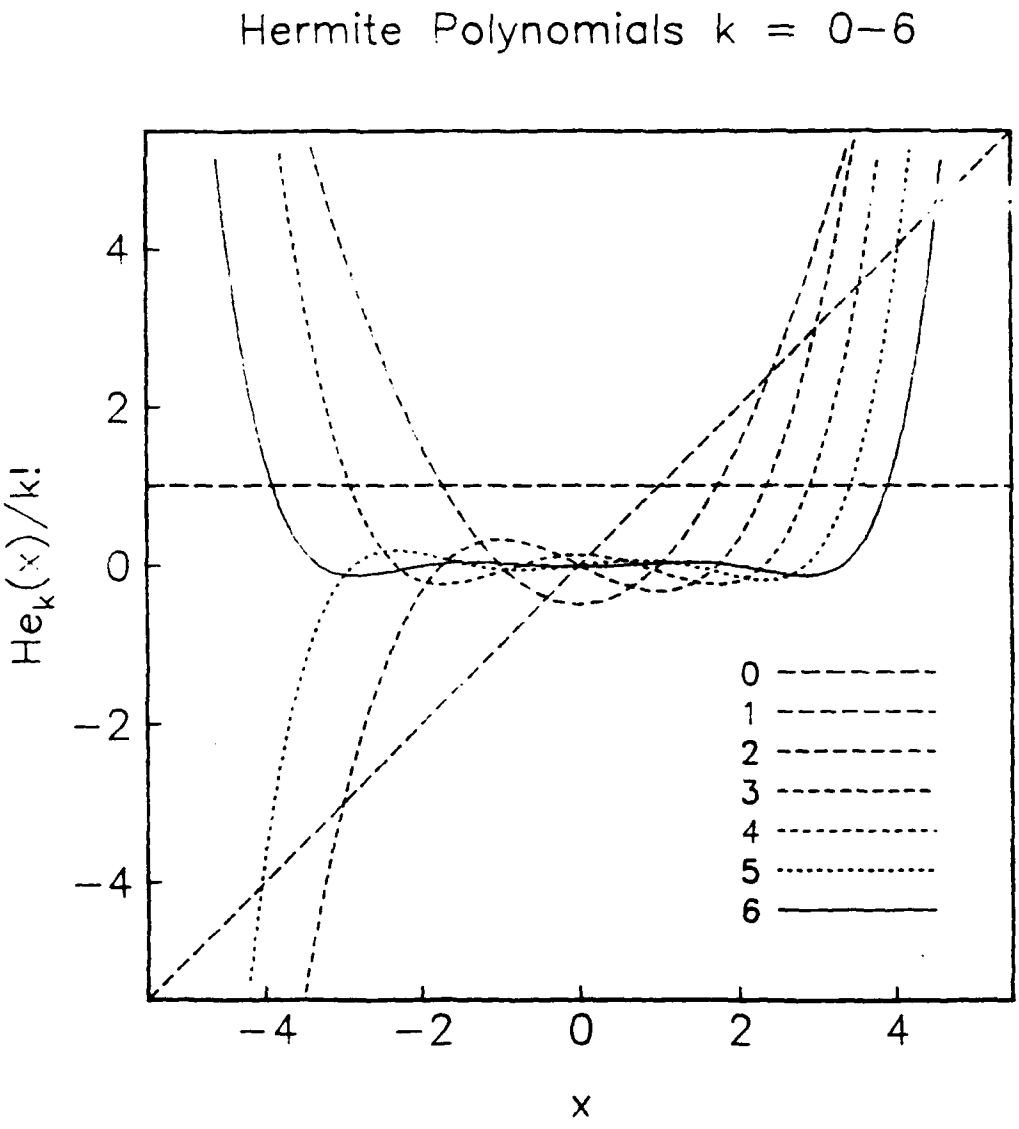


Figure A.1

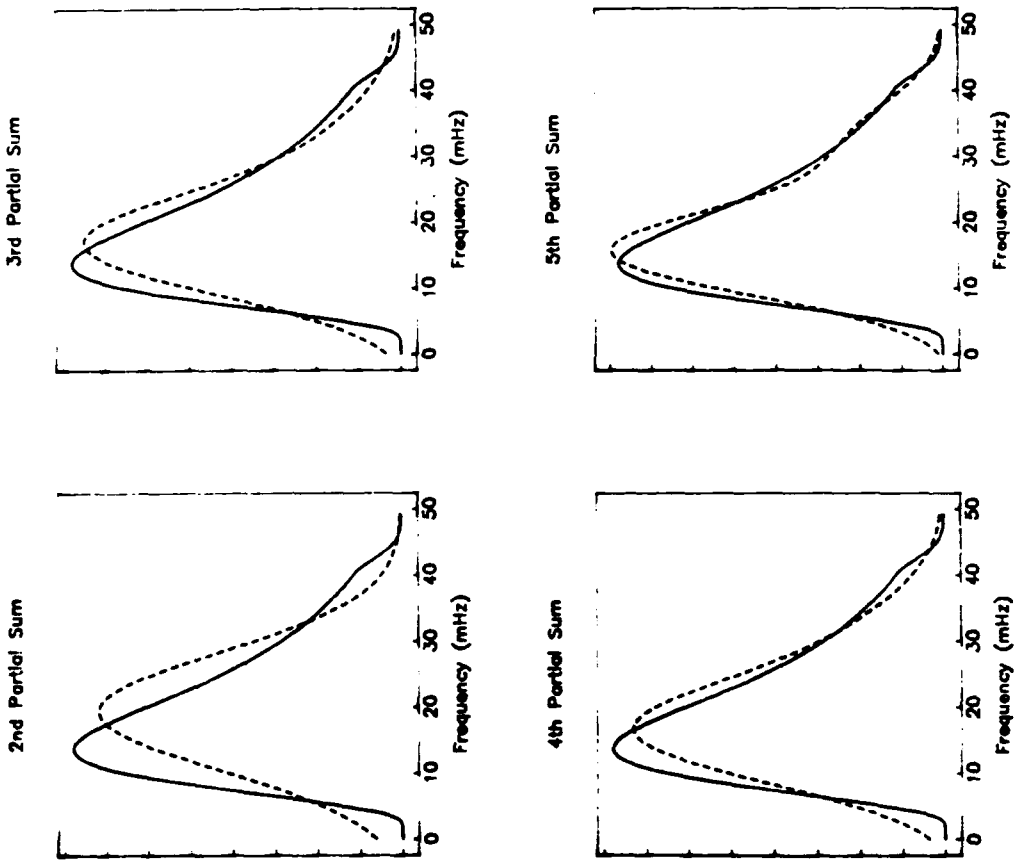

3rd Partial Sum
2nd Partial Sum
5th Partial Sum
4th Partial Sum
Frequency (mHz)
0
10
20
30
40
50

Figure A.2

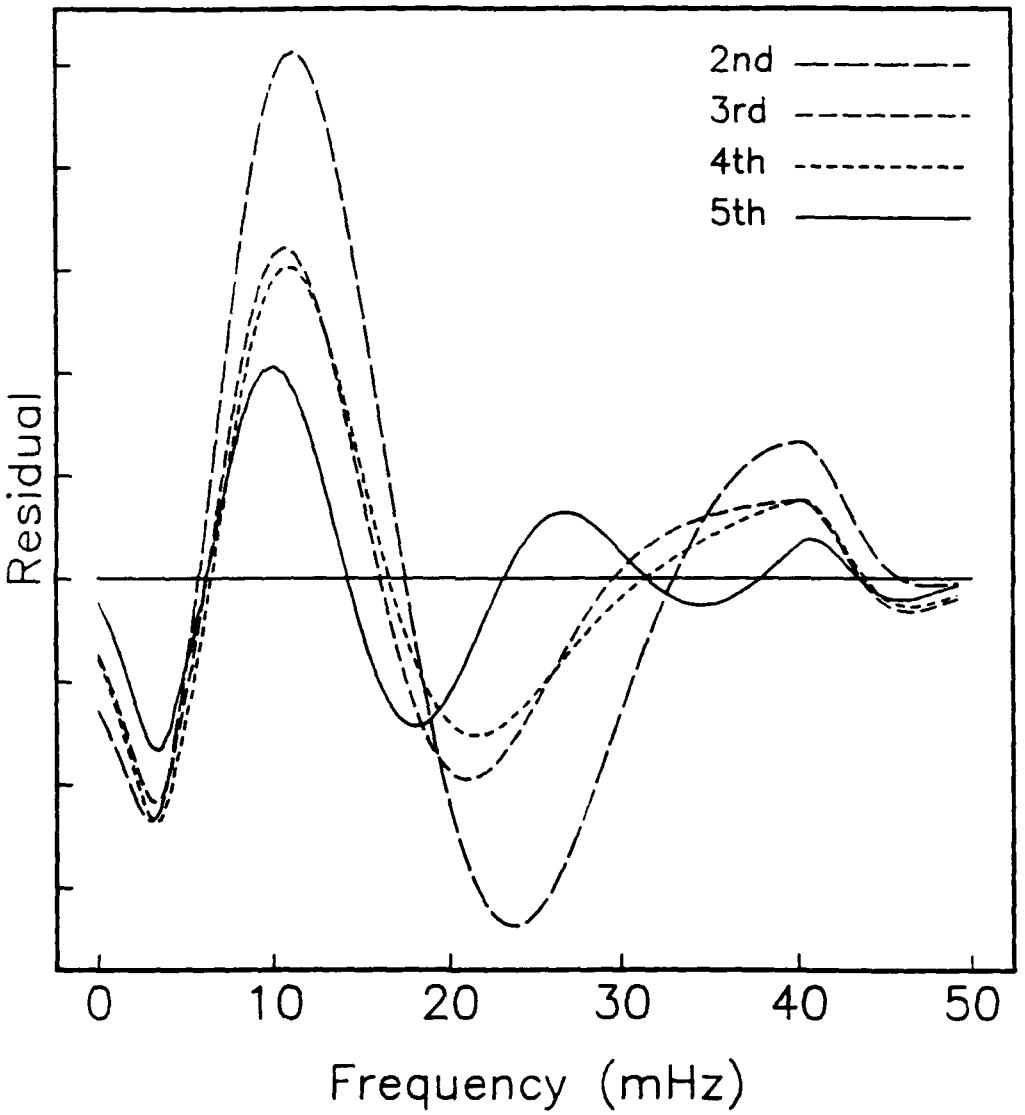


Figure A.3

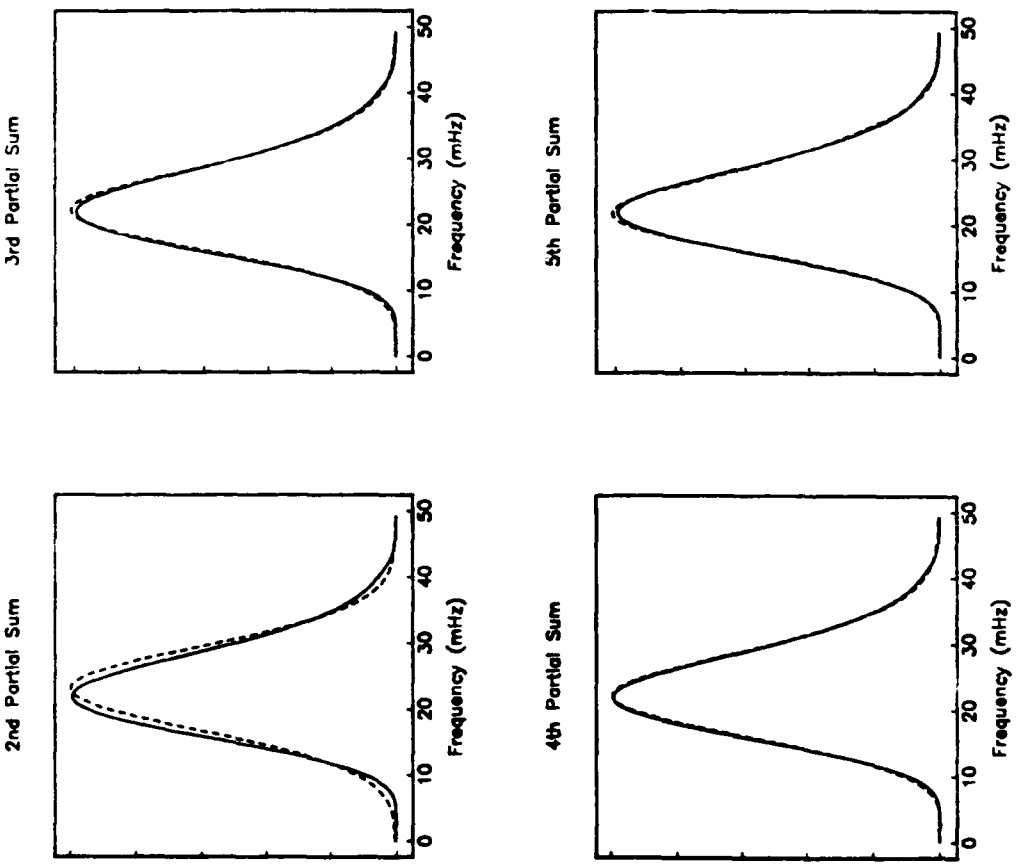

3rd Partial Sum
2nd Partial Sum
5th Partial Sum
4th Partial Sum
Frequency (mHz)
0
10
20
30
40
50

Figure A.4

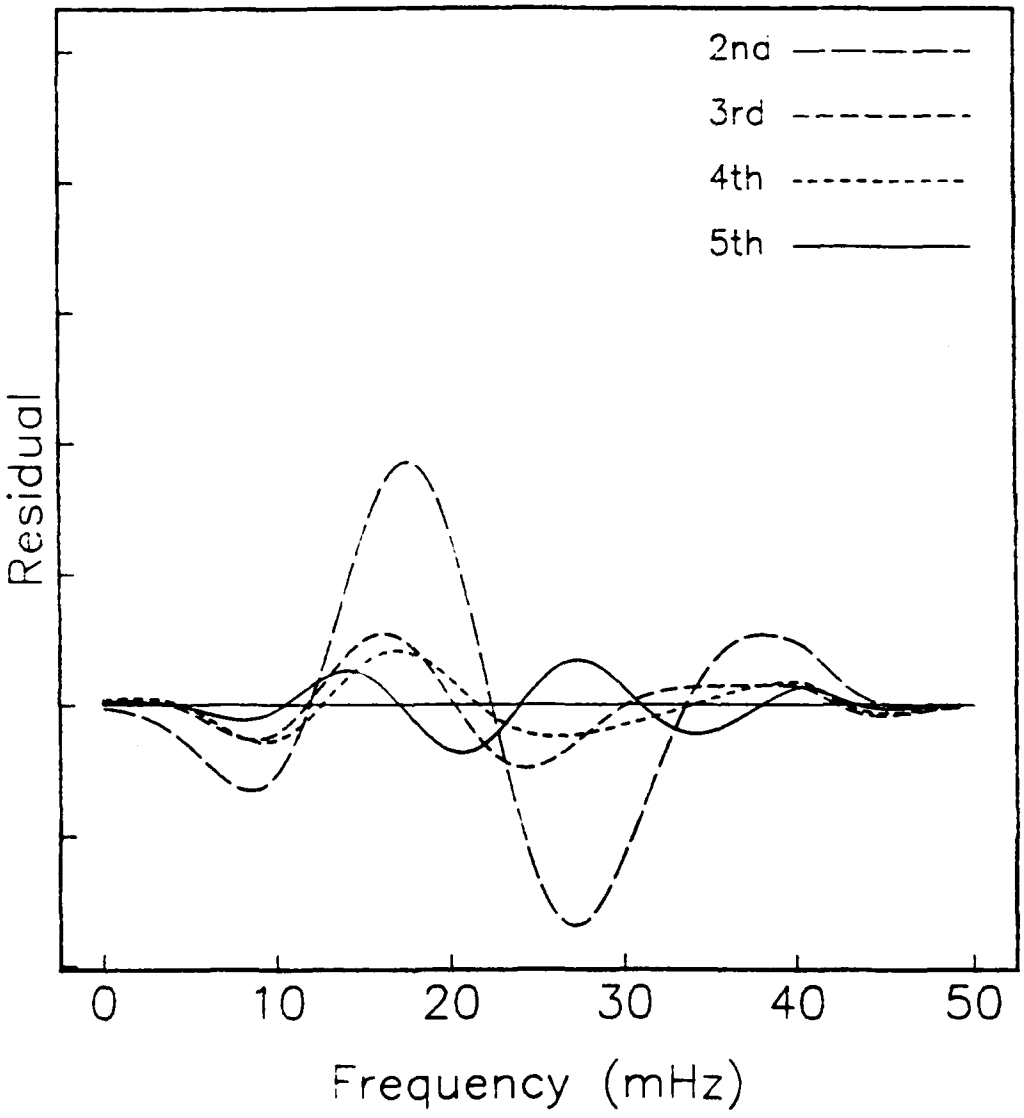


Figure A.5

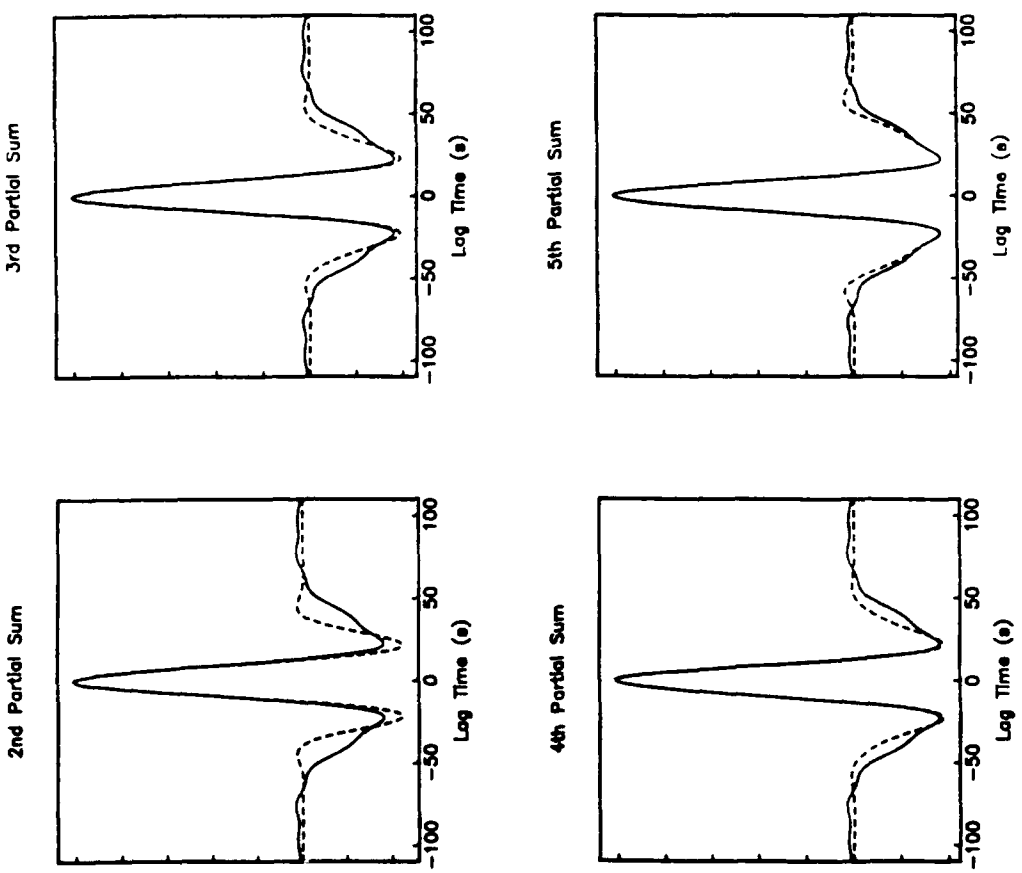

3rd Partial Sum
5th Partial Sum
2nd Partial Sum
4th Partial Sum
Lag Time (s)
100
50
0
-50
-100

Figure A.6

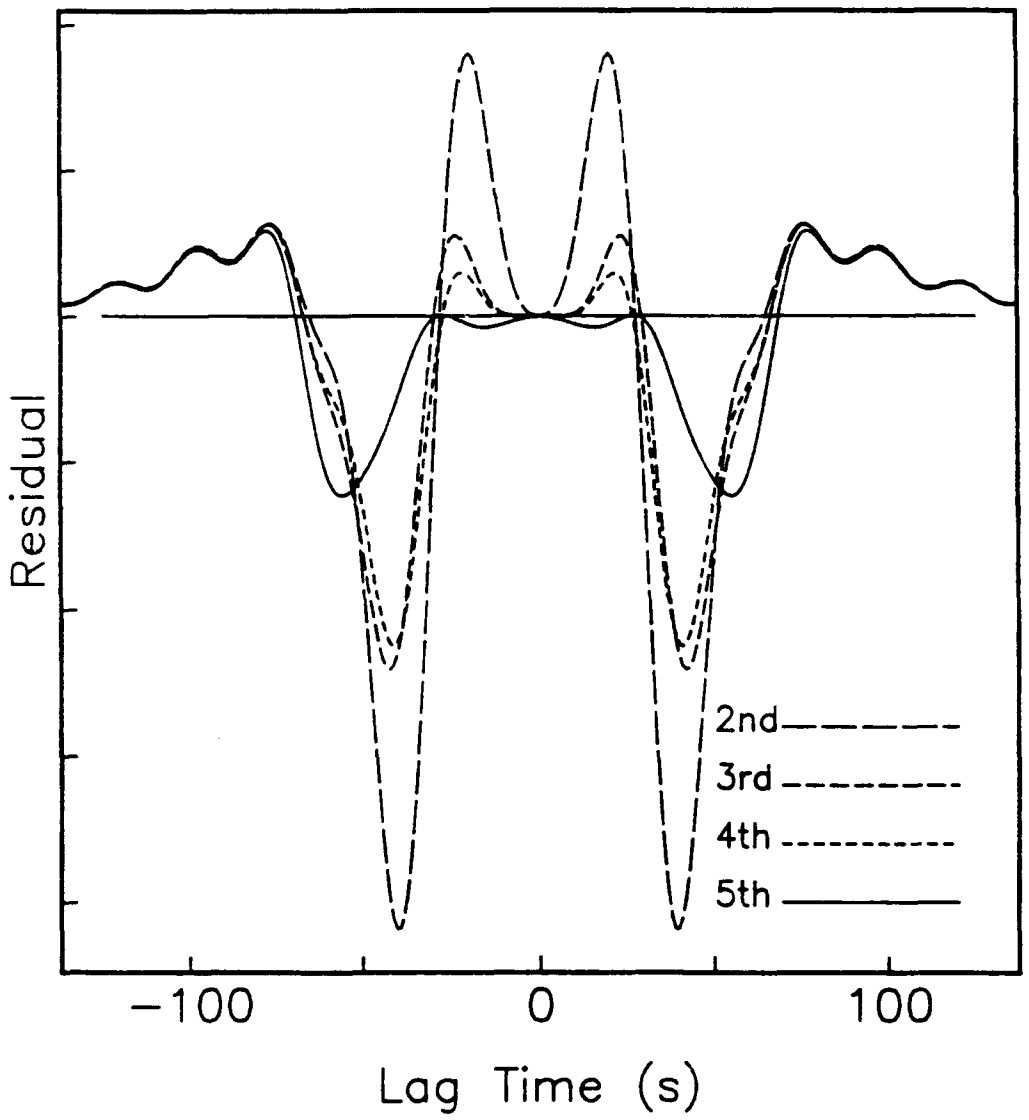


Figure A.7

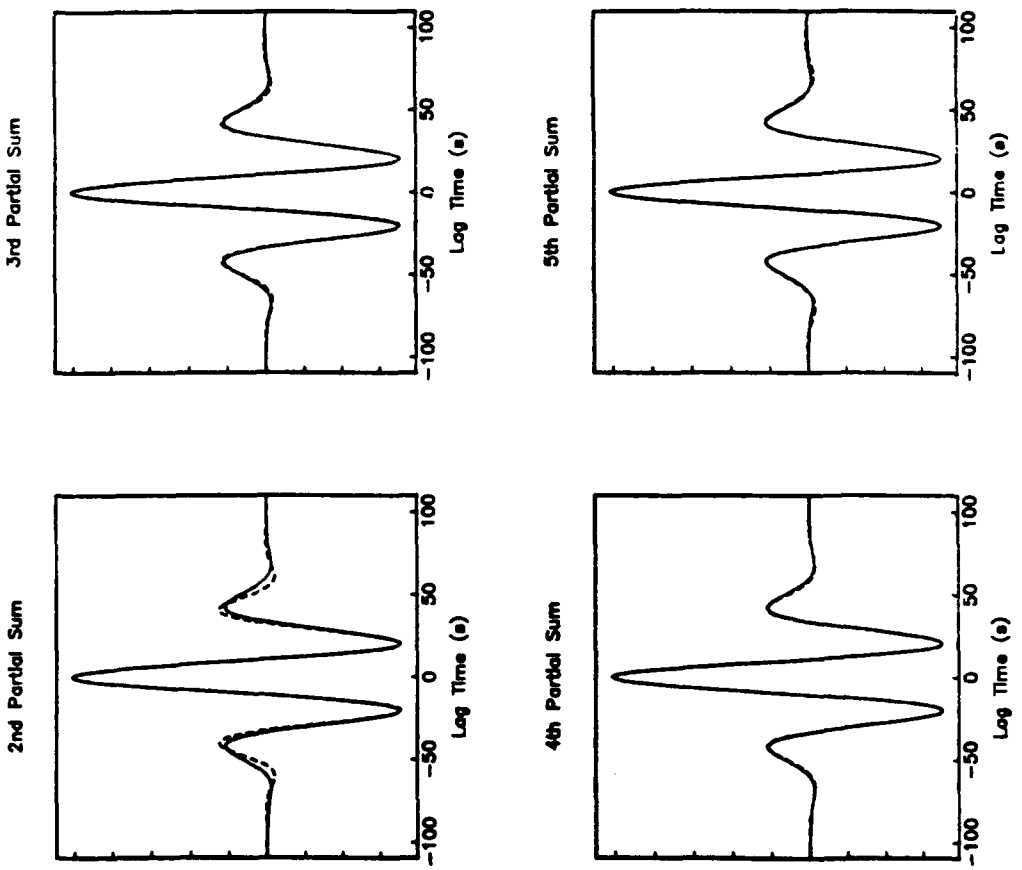

3rd Partial Sum
2nd Partial Sum
5th Partial Sum
4th Partial Sum
Lag Time (s)
100
50
0
-50
-100

Figure A.8

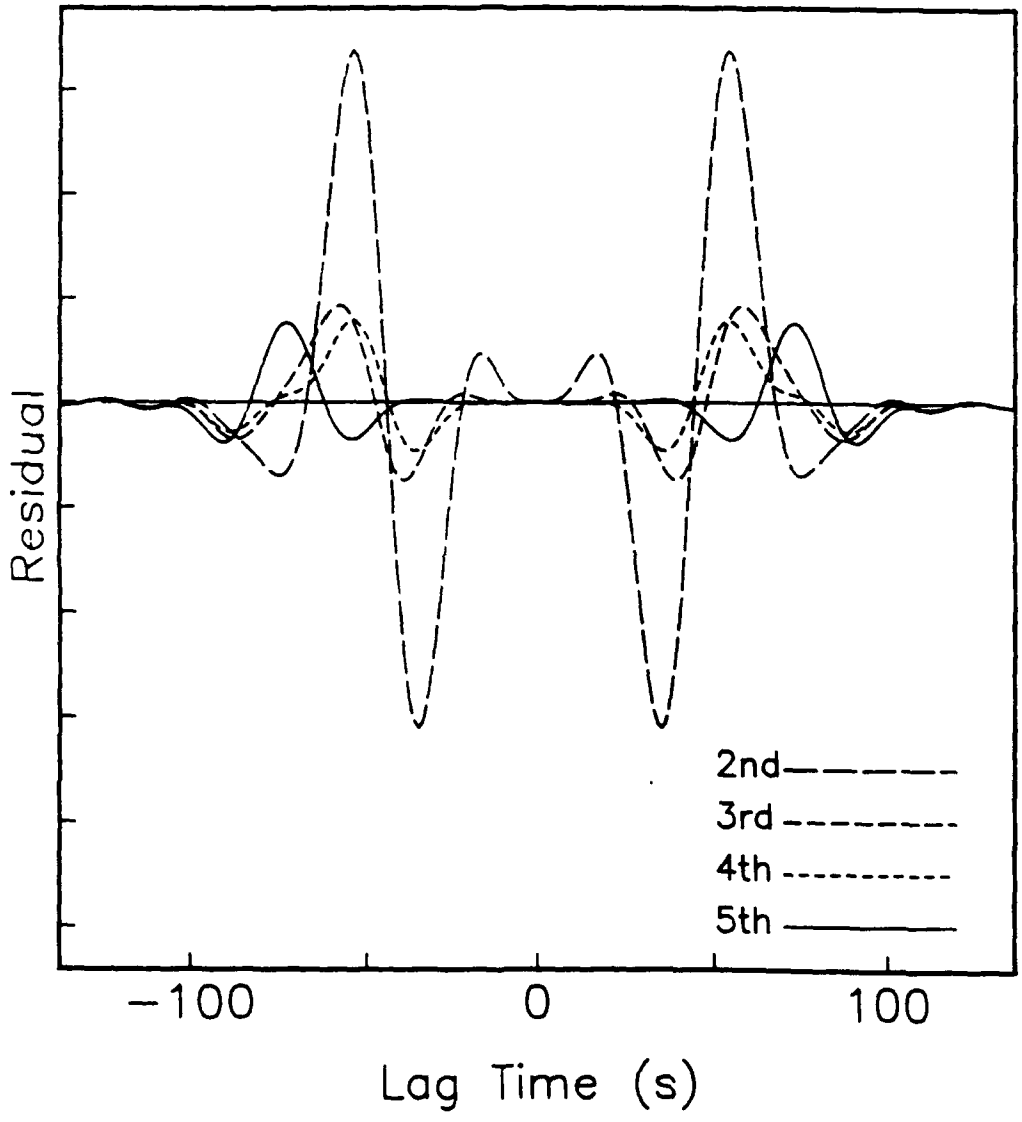


Figure A.9

APPENDIX B

SIGNAL PROCESSING OPERATIONS AND GRAM-CHARLIER SERIES

INTRODUCTION

The multiplication of a windowing operator and the convolution of a filter are standard practices in time-series analysis. In this appendix, we demonstrate that the effects of windowing and filtering on a function may be simply expressed in terms of its Gram-Charlier series. For generality, we shall consider a real function $g(t)$ with an arbitrary time shift $t - t_g$:

$$g(t) = \frac{1}{2\pi} \mathrm{Ga}\,(\sigma_g(t-t_g))\ \mathrm{Re}\ \{\ e^{-i\omega_g(t-t_g)} \sum_{n=0}^{\infty} g_n\ \gamma_g^n\ (-i\omega_g(t-t_g))^n\ \} \qquad \text{(B.1)}$$

where the expansion coefficients are defined in the usual way:

$$g_n = \frac{1}{n!} \int_0^{\infty} g(\omega)\ \mathrm{He}_n\left(\frac{\omega-\omega_g}{\sigma_g}\right) d\omega \qquad \text{(B.2)}$$

As before, we assume that ω_g is the center frequency of $g(\omega)$ and σ_g is the half-width. The coefficients g_n may be complex.

SIGNAL PROCESSING OPERATIONS

Filtering

In these days of high-speed workstations, the application of a filtering operator is usually calculated as a multiplication in the frequency domain. A number of different filters are used in seismological applications, which we may model using a Hermite-polynomial expansion. In order to fix ideas, we begin with the special case of a Gaussian filter before developing the general result.

Gaussian Filter. We assume a Gaussian filter $F_i(\omega)$, characterized by a center frequency ω_i and half-width σ_i:

$$F_i(\omega)\ \mathrm{H}(\omega) = \frac{1}{\sqrt{2\pi}\ \sigma_i}\ \mathrm{Ga}\left(\frac{\omega-\omega_i}{\sigma_i}\right) \tag{B.3}$$

We define $p(\omega)$ to be the product of $F_i(\omega)$ and $g(\omega)$, where we assume no interactions between positive and negative frequencies. In other words, we assume that the products $\mathrm{Ga}((\omega-\omega_i)/\sigma_i)\ \mathrm{Ga}((\omega+\omega_g)/\sigma_g)$ and $\mathrm{Ga}((\omega+\omega_i)/\sigma_i)\ \mathrm{Ga}((\omega-\omega_g)/\sigma_g)$ are small and may be neglected. Thus the inverse Fourier transform for $p(t)$ depends on the product of the Gaussians multiplied by the sum over Hermite polynomials:

$$p(t) = \frac{1}{(2\pi)^2}\frac{1}{\sigma_g\sigma_i}\,\mathrm{Re}\left\{\int_{-\infty}^{\infty} \mathrm{Ga}\left(\frac{\omega-\omega_i}{\sigma_i}\right)\mathrm{Ga}\left(\frac{\omega-\omega_g}{\sigma_g}\right)\sum_{n=0}^{\infty} g_n \mathrm{He}_n\left(\frac{\omega-\omega_g}{\sigma_g}\right)e^{-i\omega(t-t_g)}\,d\omega\right\}$$

$$= \frac{1}{(2\pi)^2}\frac{1}{\sigma_g\sigma_i}\,\mathrm{Ga}\left(\frac{\omega_i-\omega_g}{\sqrt{\sigma_i^2+\sigma_g^2}}\right)\mathrm{Re}\left\{\int_{-\infty}^{\infty} \mathrm{Ga}\left(\frac{\omega-\omega_p}{\sigma_p}\right)\sum_{n=0}^{\infty} g_n \mathrm{He}_n\left(\frac{\omega-\omega_g}{\sigma_g}\right)e^{-i\omega(t-t_g)}\,d\omega\right\} \tag{B.4}$$

where the filtered function $p(\omega)$ depends on a new center frequency ω_p and half-width σ_p:

$$\omega_p = \frac{\sigma_i^2\,\omega_g + \sigma_g^2\,\omega_i}{\sigma_i^2+\sigma_g^2} \tag{B.5}$$

$$\sigma_p^2 = \frac{\sigma_i^2\,\sigma_g^2}{\sigma_i^2+\sigma_g^2} \tag{B.6}$$

and we assume that $p(\omega)$ has been normalized such that $\hat{\mu}_0(0) = 1$.

To solve this integral using the Fourier transform properties of Hermite polynomials, we must rewrite the Hermite-polynomial sum in (B.4) in terms of ω_p and σ_p. We first apply the scale theorem from Appendix A to express the Hermite polynomial in terms of the new scale parameter σ_p:

$$\sum_{l=0}^{\infty} o_l \mathrm{He}_l\left(\frac{\omega-\omega_g}{\sigma_p}\right) = \sum_{n=0}^{\infty} g_n \mathrm{He}_n\left(\frac{\omega-\omega_g}{\sigma_g}\right) \tag{B.7}$$

where

$$o_l = \left(\frac{\sigma_p}{\sigma_g}\right)^l \sum_{m=0}^{\infty} g_{2m+l} \frac{(2m+l)!\,(-1/2)^m}{l!\,m!} \left(\frac{\sigma_g^2-\sigma_p^2}{\sigma_g^2}\right)^m$$

$$= \left(\frac{\sigma_i}{\sqrt{\sigma_i^2+\sigma_g^2}}\right)^l \sum_{m=0}^{\infty} g_{2m+l} \frac{(2m+l)!\,(-1/2)^m}{l!\,m!} \left(\frac{\sigma_g^2}{\sigma_i^2+\sigma_g^2}\right)^m \qquad \text{(B.8)}$$

The coefficients o_l are well-behaved for all values of σ_i and σ_g. We now apply the shift theorem from Appendix A in order to change the location parameter to the Hermite polynomials to the center frequency of the filtered function, ω_p:

$$\sum_{k=0}^{\infty} p_k \mathrm{He}_k\left(\frac{\omega-\omega_p}{\sigma_p}\right) = \sum_{l=0}^{\infty} o_l \mathrm{He}_l\left(\frac{\omega-\omega_g}{\sigma_p}\right) \qquad \text{(B.9)}$$

where

$$p_k = \sum_{j=0}^{\infty} \binom{j+k}{j} \left(\frac{\omega_p-\omega_g}{\sigma_p}\right)^j o_{j+k}$$

$$= \sum_{j=0}^{\infty} \binom{j+k}{j} \left(\frac{\sigma_g(\omega_i-\omega_g)}{\sigma_i\sqrt{\sigma_g^2+\sigma_i^2}}\right)^j o_{j+k} \qquad \text{(B.10)}$$

Now our integral is in the canonical form and we may apply the transform theorem:

$$p(t) = \frac{1}{(2\pi)^2}\frac{1}{\sigma_g\sigma_i}\,\mathrm{Ga}\left(\frac{\omega_i-\omega_g}{\sqrt{\sigma_i^2+\sigma_g^2}}\right)\mathrm{Re}\left\{\int_{-\infty}^{\infty}\mathrm{Ga}\left(\frac{\omega-\omega_p}{\sigma_p}\right)\sum_{k=0}^{\infty}p_k\,\mathrm{He}_k\left(\frac{\omega-\omega_p}{\sigma_p}\right)e^{-i\omega(t-t_g)}\,d\omega\right\}$$

$$= \frac{1}{(2\pi)^{3/2}}\frac{\sigma_p}{\sigma_g\sigma_i}\,\mathrm{Ga}\left(\frac{\omega_i-\omega_g}{\sqrt{\sigma_i^2+\sigma_g^2}}\right)\mathrm{Ga}\left(\sigma_p(t-t_g)\right)\,\mathrm{Re}\left\{e^{-i\omega_p(t-t_g)}\sum_{k=0}^{\infty}p_k\,\gamma_p^k\,(-i\omega_p(t-t_g))^k\right\}$$

(B.11)

where

$$p_k = \left(\frac{\sigma_p}{\sigma_g}\right)^k\sum_{j=0}^{\infty}\binom{j+k}{j}\left(\frac{\omega_p-\omega_g}{\sigma_g}\right)^j\sum_{m=0}^{\infty}g_{2m+j+k}\,\frac{(2m+j+k)!\,(-1/2)^m}{(j+k)!\,m!}\left(\frac{\sigma_g^2-\sigma_p^2}{\sigma_g^2}\right)^m$$

$$= \left(\frac{\sigma_i}{\sqrt{\sigma_i^2+\sigma_g^2}}\right)^k\sum_{j=0}^{\infty}\binom{j+k}{j}\left(\frac{\sigma_g(\omega_i-\omega_g)}{\sigma_i^2+\sigma_g^2}\right)^j\sum_{m=0}^{\infty}g_{2m+j+k}\,\frac{(2m+j+k)!\,(-1/2)^m}{(j+k)!\,m!}\left(\frac{\sigma_g^2}{\sigma_i^2+\sigma_g^2}\right)^m$$

(B.12)

With the application of the shift and scale theorems from Appendix A, we have derived an expression for the coefficients of the filtered function $p(t)$ which depends on the filter parameters. The sum over m arises from the scale theorem and converges for all values of σ_g and σ_i. The sum over j proceeds from the shift theorem, and is well-behaved as long as $\sigma_g\,(\omega_i-\omega_g) < (\sigma_g{}^2+\sigma_i^2)$. This is simply a requirement that the filter not be applied outside of the bandwidth of the signal. Finally, the p_k depend on the ratio $(\sigma_i{}^2 / (\sigma_g{}^2+\sigma_i{}^2))^k$, which is always less than 1. Therefore, the coefficients of the Gram-Charlier series for $p(t)$ are well-behaved.

Full Gram-Charlier Expansion. The Gaussian filter is a special case of the general expansion of $F_i(\omega)$ on the positive axis in terms of Hermite polynomials:

$$F_i(\omega)\,\mathrm{H}(\omega) = \frac{1}{\sqrt{2\pi}\,\sigma_i}\,\mathrm{Ga}\left(\frac{\omega-\omega_i}{\sigma_i}\right)\sum_{m=0}^{\infty} f_m\,\mathrm{He}_m\left(\frac{\omega-\omega_i}{\sigma_i}\right) \tag{B.13}$$

where the coefficients are defined in the usual way:

$$f_m = \frac{1}{m!}\int_0^{\infty} F_i(\omega)\,\mathrm{He}_m\left(\frac{\omega-\omega_i}{\sigma_i}\right)d\omega \tag{B.14}$$

In this case, the integral for $p(t)$ depends on the product of the two expansions:

$$p(t) = \frac{1}{(2\pi)^2}\frac{1}{\sigma_g\sigma_i}\,\mathrm{Ga}\left(\frac{\omega_i-\omega_g}{\sqrt{\sigma_i^2+\sigma_g^2}}\right)\mathrm{Re}\left\{\int_{-\infty}^{\infty}\mathrm{Ga}\left(\frac{\omega-\omega_p}{\sigma_p}\right)\sum_{n=0}^{\infty} g_n\,\mathrm{He}_n\left(\frac{\omega-\omega_g}{\sigma_g}\right)\right.$$
$$\left.\times\sum_{m=0}^{\infty} f_m\,\mathrm{He}_m\left(\frac{\omega-\omega_i}{\sigma_i}\right)e^{-i\omega(t-t_g)}\,d\omega\right\} \tag{B.15}$$

As before we must scale and shift the arguments of the Hermite polynomials in order to solve the integral. We know from the section on the Gaussian filter:

$$\sum_{k=0}^{\infty} p_k\,\mathrm{He}_k\left(\frac{\omega-\omega_p}{\sigma_p}\right) = \sum_{n=0}^{\infty} g_n\,\mathrm{He}_n\left(\frac{\omega-\omega_g}{\sigma_g}\right) \tag{B.16}$$

where the p_k are defined in (B.12). Performing the same operations on the second Hermite-polynomial sum, we find:

$$\sum_{l=0}^{\infty} e_l\,\mathrm{He}_l\left(\frac{\omega-\omega_p}{\sigma_p}\right) = \sum_{m=0}^{\infty} f_m\,\mathrm{He}_m\left(\frac{\omega-\omega_i}{\sigma_i}\right) \tag{B.17}$$

$$e_l = \left(\frac{\sigma_p}{\sigma_i}\right)^l \sum_{j=0}^{\infty} \binom{j+l}{j} \left(\frac{\omega_p-\omega_i}{\sigma_i}\right)^j \sum_{m=0}^{\infty} g_{2m+j+l} \frac{(2m+j+l)!\,(-1/2)^m}{(j+l)!\,m!} \left(\frac{\sigma_i^2-\sigma_p^2}{\sigma_i^2}\right)^m$$

$$= \left(\frac{\sigma_g}{\sqrt{\sigma_i^2+\sigma_g^2}}\right)^l \sum_{j=0}^{\infty} \binom{j+l}{j} \left(\frac{\sigma_i(\omega_g-\omega_i)}{\sigma_i^2+\sigma_g^2}\right)^j \sum_{m=0}^{\infty} g_{2m+j+l} \frac{(2m+j+l)!\,(-1/2)^m}{(j+l)!\,m!} \left(\frac{\sigma_i^2}{\sigma_i^2+\sigma_g^2}\right)^m \quad \text{(B.18)}$$

and the integral now has the form:

$$p(t) = \frac{1}{(2\pi)^2} \frac{1}{\sigma_g\sigma_i} \mathrm{Ga}\left(\frac{\omega_i-\omega_g}{\sqrt{\sigma_i^2+\sigma_g^2}}\right) \mathrm{Re}\,\{ \int_{-\infty}^{\infty} \mathrm{Ga}\left(\frac{\omega-\omega_p}{\sigma_p}\right) \sum_{k}^{\infty} p_k \,\mathrm{He}_k\left(\frac{\omega-\omega_p}{\sigma_p}\right)$$

$$\times \sum_{l=0}^{\infty} e_l \,\mathrm{He}_l\left(\frac{\omega-\omega_p}{\sigma_p}\right) e^{-i\omega(t-t_g)}\, d\omega \,\} \quad \text{(B.19)}$$

In the previous section, we olved the integral using the Fourier transform properties of Hermite polynomials. Here, however, we rearrange the series in order to employ the orthogonality relation:

$$p(t) = \frac{1}{(2\pi)^2} \frac{1}{\sigma_g\sigma_i} \mathrm{Ga}\left(\frac{\omega_i-\omega_g}{\sqrt{\sigma_i^2+\sigma_g^2}}\right) \mathrm{Ga}\,(\sigma_p(t-t_g))\ \mathrm{Re}\,\{e^{-i\omega_p(t-t_g)}$$

$$\times \int_{-\infty}^{\infty} \mathrm{Ga}\left(\frac{\omega-\omega_q}{\sigma_p}\right) \sum_{k=0}^{\infty} q_k \,\mathrm{He}_k\left(\frac{\omega-\omega_q}{\sigma_p}\right) \sum_{m=0}^{\infty} d_m \,\mathrm{He}_m\left(\frac{\omega-\omega_q}{\sigma_p}\right) d\omega \,\}$$

$$= \frac{1}{(2\pi)^{3/2}} \frac{\sigma_p}{\sigma_g\sigma_i} \mathrm{Ga}\left(\frac{\omega_i-\omega_g}{\sqrt{\sigma_i^2+\sigma_g^2}}\right) \mathrm{Ga}\,(\sigma_p(t-t_g))\ \mathrm{Re}\,\{e^{-i\omega_p(t-t_g)} \sum_{k=0}^{\infty} k!\, q_k\, d_k \,\} \quad \text{(B.20)}$$

where $\omega_q = \omega_p - i\sigma_p^2(t - t_g)$. The coefficients q_k and d_k may be written in powers of $(-i\sigma_p(t - t_g))$:

$$q_k = \sum_{n=0}^{\infty} \binom{n+k}{n} p_{n+k} (-i\sigma_p(t-t_g))^n \tag{B.21}$$

$$d_k = \sum_{m=0}^{\infty} \binom{m+k}{m} e_{m+k} (-i\sigma_p(t-t_g))^m \tag{B.22}$$

In order to write the expression for $p(t)$ in the standard form of a time-domain expansion of a Gram-Charlier series, we make use of the product theorem for infinite series:

$$\sum_{k=0}^{\infty} a_k t^k \sum_{m=0}^{\infty} b_m t^m = \sum_{k=0}^{\infty} c_k t^k \tag{B.23}$$

$$c_k = \sum_{m=0}^{k} a_m b_{k-m} \tag{B.24}$$

with the result:

$$p(t) = \frac{1}{(2\pi)^{3/2}} \frac{\sigma_p}{\sigma_g \sigma_i} \mathrm{Ga}\left(\frac{\omega_i - \omega_g}{\sqrt{\sigma_i^2 + \sigma_g^2}}\right) \mathrm{Ga}\left(\sigma_p(t-t_g)\right) \mathrm{Re}\{e^{-i\omega_p(t-t_g)} \sum_{m=0}^{\infty} s_m \gamma_p^m (-i\omega_p (t-t_g))^m\} \tag{B.25}$$

where the coefficients have the following form:

$$s_m = \sum_{k=0}^{\infty} k! \sum_{l=0}^{m} \binom{l+k}{l} p_{l+k} \binom{m-l+k}{m-l} e_{m-l+k} \tag{B.26}$$

Equations (B.25) and (B.26) describe the application of a general filter to a function expanded with Hermite polynomials. It differs from the expression for the Gaussian filter only in its coefficients. In particular, both expressions have a power series dependence on γ_p. Since ω_p is the center frequency and σ_p is the half-width of $p(\omega)$, the first correction term is third order. Thus, if $\gamma_p << 1$, we may neglect terms of third and higher order and $p(t)$ may be described by a cosinusoidal carrier with frequency ω_p modulated by a Gaussian envelope with half-width σ_p^{-1}:

$$p(t) = E(t) \cos\Phi(t) \tag{B.27}$$

where the envelope and carrier function have the form:

$$E(t) = \frac{P_0}{(2\pi)^{3/2}} \frac{\sigma_p}{\sigma_g \sigma_i} \mathrm{Ga}\left(\frac{\omega_i - \omega_g}{\sqrt{\sigma_i^2 + \sigma_g^2}}\right) \mathrm{Ga}\left(\sigma_p (t - t_g)\right) \tag{B.27a}$$

$$\Phi(t) = \omega_g (t - t_g) \tag{B.27b}$$

where P_0 is p_0 in the case of a Gaussian filter and s_0 in the general case of an arbitrary filter.

Windowing

Seismologists are typically concerned with the application of windows in the time domain. We consider an arbitrary, symmetric window $W(t)$, centered at $t = t_w$ with half-width σ_w^{-1}. We expand $W(t)$ in a Gram-Charlier series:

$$W(t) = \frac{1}{2\pi} \mathrm{Ga}\left(\sigma_w (t - t_w)\right) \left\{w_0 - w_2 (\sigma_w (t - t_w))^2 + w_4 (\sigma_w (t - t_w))^4 - \cdots\right\} \tag{B.28}$$

with expansion coefficients that are real and even

$$w_m = \frac{1}{m!} \int_0^{\infty} W(\omega)\, \mathrm{He}_m\left(\frac{\omega}{\sigma_w}\right) d\omega \tag{B.29}$$

Defining $h(t)$ to be the product of $W(t)$ and $g(t)$, we seek an expansion for $h(t)$ in the form of the time-domain representation of a Gram-Charlier series:

$$h(t) = \frac{1}{(2\pi)^2} \mathrm{Ga}\left(\sigma_w(t-t_w)\right)\ \mathrm{Ga}\left(\sigma_g(t-t_g)\right) \sum_{\substack{m=0 \\ m\,\mathrm{even}}}^{\infty} w_m \left(-i\sigma_w\,(t-t_w)\right)^m$$
$$\times\ \mathrm{Re}\left\{\, e^{-i\omega_g(t-t_g)} \sum_{k=0}^{\infty} g_k \left(-i\sigma_g(t-t_g)\right)^k \right\} \tag{B.30}$$

$$= \frac{1}{(2\pi)^2} \mathrm{Ga}\left(\frac{\sigma_g\sigma_w}{\sigma_h}\,(t_w-t_g)\right) \mathrm{Ga}\left(\sigma_h(t-t_h)\right) \sum_{\substack{m=0 \\ m\,\mathrm{even}}}^{\infty} w_m \left(-i\sigma_w\,(t-t_w)\right)^m$$
$$\times\ \mathrm{Re}\left\{\, e^{-i\omega_g(t-t_g)} \sum_{k=0}^{\infty} g_k \left(-i\sigma_g(t-t_g)\right)^k \right\} \tag{B.31}$$

where the envelope is characterized by a new half-width σ_h^{-1} and a new center time t_h:

$$\sigma_h^2 = \sigma_w^2 + \sigma_g^2 \tag{B.32}$$

$$t_h = \frac{\sigma_g^2\, t_g + \sigma_w^2\, t_w}{\sigma_g^2 + \sigma_w^2} \tag{B.33}$$

In order to write (B.31) in the canonical form, we must express the power series expansions in powers of $(t - t_h)$. To do so, we employ a translation theorem for infinite series:

$$\sum_{k=0}^{\infty} a_k (t - t_a)^k = \sum_{k=0}^{\infty} a_k \sum_{m=0}^{k} \binom{k}{m} (t - t_b)^m (t_b - t_a)^{k-m}$$

$$= \sum_{m=0}^{\infty} (t - t_b)^m \sum_{k=m}^{\infty} \binom{k}{m} a_k (t_b - t_a)^{k-m}$$

$$= \sum_{m=0}^{\infty} b_m (t - t_b)^m \tag{B.34}$$

$$b_m = \sum_{k=0}^{\infty} \binom{k+m}{k} a_{k+m} (t_b - t_a)^k \tag{B.35}$$

Applying this theorem to each of the series expansions in (B.31):

$$h(t) = \frac{1}{(2\pi)^2} \mathrm{Ga}\left(\frac{\sigma_g \sigma_w}{\sigma_h} (t_w - t_g)\right) \mathrm{Ga}\left(\sigma_h (t - t_h)\right) \times \mathrm{Re}\left\{ e^{-i\omega_h (t - t_g)} \sum_{l=0}^{\infty} r_l (\sigma_h (t - t_h))^l \sum_{j=0}^{\infty} x_j (\sigma_h (t - t_h))^j \right\} \tag{B.36}$$

where:

$$r_l = \sum_{k=0}^{\infty} \binom{k+l}{k} g_{k+l} (-i)^{k+l} \frac{\sigma_g^{k+l}}{\sigma_h^l} (t_h - t_g)^k \tag{B.37}$$

$$x_j = \sum_{\substack{m=j \\ m\,\text{even}}}^{\infty} \binom{m+j}{m} w_{m+j} (-i)^{m+j} \frac{\sigma_w^{m+j}}{\sigma_h^j} (t_h - t_w)^m \tag{B.38}$$

Making use of the product theorem for infinite series (B.23 and B.24), we rewrite the windowed function:

$$h(t) = \frac{1}{(2\pi)^2} \mathrm{Ga}\left(\frac{\sigma_g \sigma_w}{\sigma_h} (t_w - t_g)\right) \mathrm{Ga}\left(\sigma_h (t - t_h)\right)$$
$$\times \mathrm{Re}\left\{ e^{-i\omega_g (t_h - t_g)} e^{-i\omega_g (t - t_h)} \sum_{k=0}^{\infty} h_k \gamma_h^k (-i\omega_h (t - t_h))^k \right\} \tag{B.39}$$

where $\gamma_h = \sigma_h / \omega_g$ and

$$h_k = i^k \sum_{m=0}^{k} x_m r_{k-m} \tag{B.40}$$

which is in the canonical form for a time-domain representation of a Gram-Charlier series.

We have considered the application of a windowing operator centered at time t_w to a function with a Gaussian envelope centered at time t_g. If t_w is not equal to t_g, the windowed function will be centered at the weighted averaged time t_h.

CONCLUSIONS

In this appendix, we have developed expressions for the standard time-series operations of filtering and windowing. Both applications may be represented as simple Gram-Charlier series.

APPENDIX C

NORMAL-MODE POTPOURRI

INTRODUCTION

In this appendix we discuss various aspects of normal-mode seismology, from ω-l diagrams to Rayleigh's principle to synthetic seismograms, and some details about our calculations. This appendix is not intended to be a comprehensive survey of normal-mode theory; it serves to clarify some references made to normal-mode results in the text. The results presented here are well-known and have been culled from a variety of sources, including Backus and Gilbert [1968], Gilbert [1971, 1976a], Gilbert and Dziewonski [1975], Woodhouse [1976], Woodhouse and Dahlen [1978], and Gilbert [1980].

COMPUTATION OF A NORMAL MODE CATALOG

The differential equations governing the displacements of a self-gravitating Earth form a sixth-order system of equations in the case of spheroidal modes in a solid; a second-order system in the case of toroidal modes. Woodhouse [1988] recently described the solution of these equations in some detail. We use the program MINEOS, due to G. Masters and based on a program originally written by F. Gilbert. This program solves the system of equations using a Runge-Kutta technique.

In our implementation of this program, we typically specify an earth model in terms of the compressional velocity, the shear velocity, and the density (transversely isotropic models are a simple generalization) at a number of depths. In order to avoid spatial aliasing, we generally have chosen a 10 km spacing of the model knots. The work in this

thesis is based on normal-mode catalogs which were calculated to 50 mHz (although we have calculated toroidal-mode catalogs to 100 mHz). For toroidal modes, this encompasses more than 8000 modes; for spheroidal modes, more than 16000 modes. The resulting eigenfunctions require some 300 Mbytes of disk space. After calculation of the Fréchet kernels (see below), we form a stripped set of eigenfunctions. This stripped set is ordered by depth, with all eigenfunctions truncated at 800 km depth (since earthquakes rarely occur at depths greater than 800 km). The final mode tables require approximately 60 Mbytes disk space.

MODELS

In order to facilitate comparisons between the models EU2 and SNA, we considered only variations in shear velocity between the base of the crust and 400 km depth (EU2 and SNA are virtually indistinguishable within the transition zone (Figure 1.1)). Thus the eigenfunctions and eigenfrequencies for each model were calculated with identical compressional velocity and density structures. At depths greater than 750 km, the elastic parameters were taken from PREM [Dziewonski and Anderson, 1981]. The Q model of Masters and Gilbert [1983] was used to calculate the effect of attenuation on eigenfrequency, with an assumed center frequency of 35 mHz. Tables C.1 and C.2 list the elastic parameters of the models used in our normal-mode computations; Table C.3 lists the anelastic parameters.

ω-l DIAGRAMS

Each normal mode is completely specified by a triplet of numbers: the radial order (n), the angular order (l), and the azimuthal order (m). In a spherically-symmetric earth, the $2l + 1$ azimuthal-order numbers are degenerate. For toroidal modes, each radial-order number corresponds to a particular traveling-wave branch and properties such as energy density vary smoothly along a branch. For spheroidal modes, the connection is more

complicated because a given radial-order number may contain more than one traveling-wave branch and properties such as energy density change discontinuously across a branch. The lowest radial-order number ($n = 0$) is the fundamental mode and runs from the lower left-hand corner to the upper right-hand corner of the diagram. For the calculations in this thesis, we used the mantle-wave branches ("R" modes in Okal's [1978] classification) of the spheroidal-mode table. These branches were identified as mantle modes by selecting all modes with phase velocities less than 13.25 km/s (which corresponds to the core-grazing ray) and then removing the Stoneley branches (see Figure 1.3).

RAYLEIGH'S PRINCIPLE

Rayleigh's principle [Backus and Gilbert, 1967; Woodhouse, 1976; Woodhouse and Dahlen, 1978] states that on average the total kinetic energy of the νth normal mode is equal to its total potential energy:

$$\omega_\nu^2 \int_0^a [U_\nu^2(r) + V_\nu^2(r) + W_\nu^2(r)]\, \rho(r)\, r^2 dr = \int_0^a [K_\nu(r)\kappa(r) + M_\nu(r)\mu(r) + \Gamma_\nu(r)]\, dr \qquad \text{(C.1)}$$

where ω_ν is the eigenfrequency; $U_\nu(r)$, $V_\nu(r)$, and $W_\nu(r)$ are the radial scalars; $K_\nu(r)\kappa(r)$ is the compressional-energy density; $M_\nu(r)\mu(r)$ is the shear-energy density; $\int_0^a \Gamma_\nu(r)\, r^2\, dr$ is the gravitational potential energy; and a is the radius of the Earth. The mode index ν is a function of the radial-order number, the angular-order number, and the azimuthal-order number ($U_\nu(r) \equiv {}_nU_l^m(r)$). This notation (which follows that of Gilbert [1980]) assumes the displacement u may be written as a sum of vector spherical harmonics:

$$u(r,\Delta,\xi) = \sum_{n=1}^{\infty} \sum_{l=1}^{\infty} \sum_{m=-l}^{l} {}_nU_l^m(r)\, P_l^m(\Delta,\xi) + {}_nV_l^m(r)\, B_l^m(\Delta,\xi) + {}_nW_l^m(r)\, C_l^m(\Delta,\xi) \qquad \text{(C.2)}$$

where the P_l^m, B_l^m, and C_l^m are defined in terms of Legendre polynomials:

$$\begin{aligned} P_l^m(\Delta,\xi) &= \hat{r}\, Y_l^m(\Delta,\xi) \\ L\, B_l^m(\Delta,\xi) &= r\, \nabla_1\, Y_l^m(\Delta,\xi) \\ L\, C_l^m(\Delta,\xi) &= \nabla_1 \times (r\, Y_l^m(\Delta,\xi)) \end{aligned} \tag{C.3}$$

and $L^2 \equiv l\,(l+1)$. This definition of B_l^m and C_l^m is adopted from Gilbert [1980] and differs from that chosen by Gilbert and Dziewonski [1975], Woodhouse [1976], and Woodhouse and Dahlen [1978] by the scaling factor L. The disagreement about the normalization of the vector spherical harmonics has created considerable confusion among students of normal-mode seismology; I have chosen the notation of Gilbert [1980] in order to simplify the expressions for the energy densities. With this choice, the radial eigenfunctions satisfy the following normalization:

$$\omega_\nu^2 \int_0^a [U_\nu^2(r) + V_\nu^2(r) + W_\nu^2(r)]\, \rho(r)\, r^2 dr = 1 \tag{C.4}$$

$Y_l^m(\Delta,\xi)$ is the normalized surface harmonic:

$$Y_l^m(\Delta,\xi) = (-1)^m \left(\frac{2l+1}{4\pi} \frac{(l-m)!}{(l+m)!} \right)^{1/2} P_l^m(\cos\Delta)\, e^{im\xi} \tag{C.5a}$$

where Δ is epicentral distance and ξ is epicentral longitude, and

$$\int_0^\pi \sin\Delta\, d\Delta \int_{-\pi}^{\pi} Y_l^{*m}(\Delta,\xi)\; Y_{l'}^{m'}(\Delta,\xi)\, d\xi = \delta_{mm'}\, \delta_{ll'} \tag{C.5b}$$

is the appropriate normalization.

Rayleigh's principle is a statement that the eigenfrequencies are stationary with respect to a variation in the eigenfunctions. This stationarity may be exploited to calculate variations in the eigenfrequencies due to small changes in the model parameters:

$$\frac{\delta\omega_\nu}{\omega_\nu} = \mathbf{G}\,\delta\mathbf{m} + \mathrm{O}\left(\|\delta\mathbf{m}\|^2\right)$$
$$= \frac{1}{2}\int_0^a \left[K_\nu(r)\,\kappa(r)\,\frac{\delta\kappa(r)}{\kappa(r)} + M_\nu(r)\,\mu(r)\,\frac{\delta\mu(r)}{\mu(r)} + R_\nu(r)\,\rho(r)\,\frac{\delta\rho(r)}{\rho(r)}\right] dr \quad \text{(C.6)}$$

where G is the Fréchet derivative of eigenfrequency with respect to the model parameters, $\delta\mathbf{m}$ is the perturbation to the model parameters and $\delta\omega_\nu$ is the perturbation to eigenfrequency of the vth normal mode at constant wavenumber. $K_\nu(r)\kappa(r)$ is the Fréchet kernel for the normalized perturbation to bulk modulus $\delta\kappa/\kappa$, $M_\nu(r)\mu(r)$ is the Fréchet kernel for the normalized perturbation to shear modulus $\delta\mu/\mu$, and $R_\nu(r)\rho(r)$ is the Fréchet kernel for the normalized perturbation to density $\delta\rho/\rho$. In this formulation, we have considered only perturbations to the radial distribution of isotropic elastic parameters. These kernels have the following form:

$$K_\nu = (r\partial_r U_\nu + F)^2 \quad \text{(C.7a)}$$

$$M_\nu = \frac{1}{3}(2r\partial_r U_\nu - F)^2 + (r\partial_r(V_\nu + W_\nu) - (V_\nu + W_\nu) + LU_\nu)^2 + (l-1)(l+2)(V_\nu^2 + W_\nu^2) \quad \text{(C.7b)}$$

$$R_\nu = -\omega^2(U_\nu^2 + V_\nu^2 + W_\nu^2) + 2r^2 U_\nu(\partial_r\varphi_\nu + 4\pi\rho G U_\nu - gr^{-1}F) + 2r\,L\,V_\nu\,\varphi_\nu - 8\pi G\int_r^a F\,U_\nu\,\rho\,r^{-1}dr \quad \text{(C.7c)}$$

$$F = 2U_\nu - LV_\nu \tag{C.7d}$$

where $\varphi_\nu(r)$ is the perturbation to gravitational potential, G is the gravitational constant, and g is the acceleration due to gravity.

In this formulation, we have neglected the perturbation in eigenfrequency due to changes in the position of discontinuities. The complete (correct) expressions may be found in Woodhouse and Dahlen [1978]. The formulas for the Fréchet kernels in the case of transverse isotropy may be found in Anderson and Dziewonski [1981].

CALCULATION OF SYNTHETIC SEISMOGRAMS

We may write a synthetic seismogram as a sum over all normal modes:

$$\tilde{s}(t) = \sum_\nu \tilde{u}_\nu(t) \tag{C.8}$$

where each standing wave may be written:

$$\begin{aligned} \tilde{u}_\nu(t) &= \tilde{u}_\nu(r,t) \\ &= a_\nu(r,t) * c_\nu(t) \end{aligned} \tag{C.9}$$

where each wavegroup depends on the receiver position $r = (a, \Delta, \xi)$, the source position $r_0 = (r_0, 0, 0)$, and time t, and we have adopted the epicentral coordinate system. a_ν contains both the source and receiver terms and we refer to it as the excitation coefficient, c_ν represents the propagation operator, and * denotes convolution. Following Gilbert [1971] and Gilbert and Dziewonski [1975], we shall discuss the contributions of these operators to the calculation of synthetic seismograms.

Excitation operator: a_ν

The excitation coefficient may be expressed as the product of both the source term and the receiver term:

$$a_\nu(r,t) = \psi_\nu(r_o,t)\ s_\nu(r) \tag{C.10}$$

Ψ_ν is the source excitation coefficient and contains the spatial and temporal dependence of the source and s_ν is the eigenfunction at the receiver. Ψ_ν may be written as a volume integral over the source region:

$$\psi_\nu = -\int \frac{\mathrm{d}\, f(r_o,t)}{\mathrm{dt}} \cdot s_\nu^*(r_o)\, \mathrm{dV_o} \tag{C.11}$$

where $f(r_o,t)$ is the external force causing the excitation of the normal modes and $s_\nu(r_o)$ is the eigenfunction at the source depth. In this study, we have assumed a step function time dependence of the external force, which allows us to write the excitation coefficient as the double dot product between the moment tensor $\boldsymbol{M}$ and the strain tensor $\boldsymbol{E}_\nu$

$$\psi_\nu = \boldsymbol{M} : \boldsymbol{E}_\nu(r_o) \tag{C.12}$$

Exact expressions for the a_ν may be found in Gilbert and Dziewonski [1975]. We use the Harvard CMT [Dziewonski *et al.*, 1981] solutions to calculate our synthetic seismograms.

Propagation Operator: c_ν

The harmonic time dependence of the synthetic seismogram depends on a transient term which behaves like a sum over decaying cosinusoids plus a static offset term which satisfies the initial conditions:

$$c_\nu(t) = H(t)\,[\cos(\omega_\nu t)\exp(\frac{-\omega_\nu}{2Q_\nu}t) - 1] \qquad \text{(C.13)}$$

where $H(t)$ is the Heaviside step function. In general, we neglect the static offset term and consider only the transient term.

Summation

We have calculated normal-mode catalogs for the models EU2 and SNA which are complete to 50 mHz. For SH, this encompasses 8,000 modes; for PSV, more than 16,000. Although time-intensive, these computations are only made once for each model. Once the eigenfunctions and eigenfrequencies are computed, any number of synthetic seismograms may be calculated. After summation, the synthetic seismograms are convolved with the appropriate instrument response and filtered with a Hanning taper.

TRAVELING-WAVE REPRESENTATION

The representation above is a standing-wave representation. However, it is desirable to express a seismogram as a sum over traveling-wave groups. In Chapter 2, we wrote down an expression for synthetic seismograms as a sum over traveling-wave branches:

$$\tilde{s}(t) = \sum_n \tilde{u}_n(t) \qquad \text{(C.14)}$$

where n refers to the branch index, rather than the radial-order number (in the case of toroidal modes, these are equivalent). Each branch may be written as a sum over angular order l

$$\tilde{u}_n(t) = \sum_{l \in n} {}_n a_l(r)\ {}_n c_l(t)$$

$$= \sum_{l \in n} A_n(l,r)\ P_l^m(\cos\Delta)\ c_n(l,t) \qquad \text{(C.15)}$$

where ${}_n a_l$ is the excitation coefficient (${}_n a_l = A_n\ P_l^m(\cos\Delta)$), c_n represents the propagation operator, and we have assumed a step-function time-dependence of the source. The sum over azimuthal-order number is implicit within the following expressions. Following Gilbert [1976a], we may obtain a traveling-wave representation by employing the Poisson-sum formula to convert the sum over discrete angular order to an integral over wavenumber:

$$\tilde{u}_n(t) = \sum_s \int_0^\infty (-1)^s\ A_n(\lambda-1/2,r)\ P_{\lambda-1/2}^m(\cos\Delta)\ c_n(\lambda-1/2,t)\ e^{2i\pi s\lambda}\ d\lambda \qquad \text{(C.16)}$$

where s is the orbital index of the traveling-wave group and $\lambda = l + 1/2$ is the surface spherical wavenumber. Following Aki and Richards [1980] and Lerner-Lam and Jordan [1983], we shall employ the following asymptotic form of P_l^m:

$$P_{\lambda-1/2}^m(\cos\Delta) \equiv (-\lambda)^m \left(\frac{2}{\pi\,\lambda\,\sin\Delta}\right)^{1/2} \cos[\lambda\Delta - \frac{\pi}{4} + \frac{m\pi}{2}] \qquad \text{(C.17)}$$

which is valid for large angular order (analysis has shown that (C.17) is good approximation to P_l^m for angular order as low as 10 [Lerner-Lam and Jordan, 1983]). By manipulating the asymptotic expressions, we may write:

$$\tilde{u}_n(t) = \int_0^\infty A_n(\lambda-1/2,r)\ c_n(\lambda-1/2,t)\ \sum_s R_s(\lambda,\Delta)\ d\lambda \qquad \text{(C.18)}$$

where $R_s(\lambda,\Delta)$ is the horizontal wavefunction and is defined:

$$R_s = (-\lambda)^m\ (-1)^{(s-1)/2} \left(\frac{2}{\pi\ \lambda\ \sin\Delta}\right)^{1/2} \cos[\lambda((s-1)\pi+\Delta) - \frac{\pi}{4} + \frac{m\pi}{2}] \qquad \text{(C.19)}$$

for s-odd and

$$R_s = (-\lambda)^m\ (-1)^{s/2} \left(\frac{2}{\pi\ \lambda\ \sin\Delta}\right)^{1/2} \cos[\lambda(-s\pi+\Delta) - \frac{\pi}{4} + \frac{m\pi}{2}] \qquad \text{(C.20)}$$

for s-even. In this thesis, we have considered the special case of the minor-arc orbit, that is, we have assumed that $s = 1$ throughout the development of our methodology.

$$\tilde{u}_n(t) = \int_0^\infty A_n(\lambda-1/2,r)\ c_n(\lambda-1/2,t)\ (-\lambda)^m \left(\frac{2}{\pi\ \lambda\ \sin\Delta}\right)^{1/2} \cos[\lambda\Delta - \frac{\pi}{4} + \frac{m\pi}{2}]\ d\lambda \qquad \text{(C.21)}$$

For a quadrapolar source, $m \leq 2$, and we write:

$$\tilde{u}_n(t) = \int_0^\infty |\tilde{A}_n| \exp[\frac{-\omega_n t}{2Q_n}] \cos[\lambda\Delta - \frac{\pi}{4} - \omega_n t + \tilde{\Phi}_n]\ d\lambda \qquad \text{(C.22)}$$

where ω_n and Q_n are functions of wavenumber. To conclude this appendix, we present formulas for the excitation coefficient and phase term in (C.22). We draw upon the notation of Dahlen [1980]:

$$\sum_1 = \left(\frac{\lambda}{2\pi}\right)^{1/2} [M_{rr}\partial_r U(r_o) + 1/2 r_o^{-1}(M_{\theta\theta}+M_{\varphi\varphi})(2U(r_o) - \lambda V(r_o))] \quad \text{(C.23a)}$$

$$\sum_2 = \left(\frac{\lambda}{2\pi}\right)^{1/2} [\partial_r V(r_o) + r_o^{-1}(\lambda U(r_o) - V(r_o))][M_{r\varphi}\sin\xi + M_{r\theta}\cos\xi] \quad \text{(C.23b)}$$

$$\sum_3 = \left(\frac{\lambda}{2\pi}\right)^{1/2} \lambda\, r_o^{-1} V(r_o)[M_{\theta\varphi}\sin 2\xi + 1/2(M_{\theta\theta}-M_{\varphi\varphi})\cos 2\xi] \quad \text{(C.23c)}$$

$$\sum_4 = \left(\frac{\lambda}{2\pi}\right)^{1/2} [\partial_r W(r_o) - r_o^{-1} W(r_o)][M_{r\varphi}\cos\xi - M_{r\theta}\sin\xi] \quad \text{(C.23d)}$$

$$\sum_5 = \left(\frac{\lambda}{2\pi}\right)^{1/2} \lambda\, r_o^{-1} W(r_o)[M_{\theta\varphi}\cos 2\xi - 1/2(M_{\theta\theta}-M_{\varphi\varphi})\sin 2\xi] \quad \text{(C.23e)}$$

where we have dropped the subscripts on U, V, and W. M_{rr}, $M_{\theta\theta}$, $M_{\varphi\varphi}$, $M_{r\theta}$, $M_{r\varphi}$, and $M_{\theta\varphi}$ are the six components of the moment tensor M. We may now write down expressions for the excitation coefficient and phase term for the transverse, vertical, and radial-component seismograms:

Transverse component (toroidal modes)

$$\begin{aligned} |\tilde{A}_n| &= \frac{1}{2\pi} W(a) \left(\frac{1}{\sin\Delta}\right)^{1/2} [\textstyle\sum_4^2 + \sum_5^2]^{1/2} \\ \tilde{\Phi}_n &= \arctan(\textstyle\sum_5 / \sum_4) \end{aligned} \quad \text{(C.24)}$$

Vertical component (spheroidal modes)

$$
\begin{aligned}
|\tilde{A}_n| &= \frac{1}{2\pi}\, U(a) \left(\frac{1}{\sin\Delta}\right)^{1/2} [(\textstyle\sum_1 - \sum_3)^2 + \sum_2^2]^{1/2} \\
\tilde{\Phi}_n &= \arctan(\textstyle\sum_2 / (\sum_1 - \sum_3))
\end{aligned}
\tag{C.25}
$$

Radial component (spheroidal modes)

$$
\begin{aligned}
|\tilde{A}_n| &= \frac{1}{2\pi}\, V(a) \left(\frac{1}{\sin\Delta}\right)^{1/2} [(\textstyle\sum_1 - \sum_3)^2 + \sum_2^2]^{1/2} \\
\tilde{\Phi}_n &= -\arctan((\textstyle\sum_1 - \sum_3) / \sum_2)
\end{aligned}
\tag{C.26}
$$

We make use of these expressions for the excitation amplitude and phase in Chapter 2.

CONCLUSIONS

In this Appendix, I have presented some results from normal-mode seismology in order to clarify various aspects of my thesis and discussed my implementation of specific procedures. While the theoretical results documented here are available in the literature, a full treatment with a self-consistent notation is more difficult to obtain.

Table C.1 Model EU2

R (m)	ρ (kg/m**3)	Vp (m/s)	Vs (m/s)
0.	13088.50	11237.30	3626.46
50000.	13087.96	11236.91	3626.19
100000.	13086.32	11235.74	3625.37
150000.	13083.60	11233.79	3624.02
200000.	13079.79	11231.07	3622.12
250000.	13074.89	11227.56	3619.69
300000.	13068.90	11223.27	3616.71
350000.	13061.83	11218.20	3613.19
400000.	13053.66	11212.36	3609.12
450000.	13044.41	11205.73	3604.52
500000.	13034.06	11198.33	3599.37
550000.	13022.63	11190.14	3593.69
600000.	13010.11	11181.18	3587.46
650000.	12996.50	11171.43	3580.69
700000.	12981.81	11160.91	3573.37
750000.	12966.02	11149.60	3565.52
800000.	12949.14	11137.52	3557.12
850000.	12931.19	11124.66	3548.18
900000.	12912.13	11111.02	3538.71
950000.	12891.99	11096.59	3528.68
1000000.	12870.76	11081.39	3518.12
1050000.	12848.44	11065.41	3507.02
1100000.	12825.03	11048.65	3495.37
1150000.	12800.54	11031.11	3483.18
1200000.	12774.95	11012.79	3470.45
1221500.	12763.61	11004.67	3464.81
1221500.	12166.33	10355.54	0.00
1250000.	12151.57	10338.97	0.00
1300000.	12124.99	10309.58	0.00
1350000.	12097.55	10279.76	0.00
1400000.	12069.23	10249.46	0.00
1450000.	12040.01	10218.66	0.00
1500000.	12009.88	10187.30	0.00
1550000.	11978.82	10155.36	0.00
1600000.	11946.81	10122.78	0.00
1650000.	11913.84	10089.54	0.00
1700000.	11879.89	10055.59	0.00
1750000.	11844.95	10020.90	0.00
1800000.	11809.00	9985.41	0.00
1850000.	11772.02	9949.11	0.00
1900000.	11734.01	9911.93	0.00
1950000.	11694.93	9873.85	0.00
2000000.	11654.78	9834.83	0.00
2050000.	11613.54	9794.83	0.00
2100000.	11571.19	9753.80	0.00
2150000.	11527.72	9711.72	0.00
2200000.	11483.11	9668.53	0.00
2250000.	11437.35	9624.20	0.00
2300000.	11390.42	9578.69	0.00
2350000.	11342.30	9531.96	0.00
2400000.	11292.98	9483.97	0.00
2450000.	11242.45	9434.69	0.00
2500000.	11190.67	9384.06	0.00
2550000.	11137.65	9332.06	0.00
2600000.	11083.36	9278.65	0.00
2650000.	11027.79	9223.78	0.00
2700000.	10970.92	9167.41	0.00
2750000.	10912.73	9109.51	0.00
2800000.	10853.22	9050.04	0.00

2850000.	10792.36	8988.95	0.00
2900000.	10730.13	8926.21	0.00
2950000.	10666.53	8861.78	0.00
3000000.	10601.53	8795.62	0.00
3050000.	10535.12	8727.69	0.00
3100000.	10467.29	8657.95	0.00
3150000.	10398.01	8586.36	0.00
3200000.	10327.27	8512.88	0.00
3250000.	10255.06	8437.47	0.00
3300000.	10181.36	8360.10	0.00
3350000.	10106.15	8280.72	0.00
3400000.	10029.42	8199.29	0.00
3450000.	9951.15	8115.78	0.00
3480000.	9903.44	8064.65	0.00
3480000.	5566.46	13700.80	7242.45
3500000.	5556.43	13695.89	7242.65
3550000.	5531.41	13683.72	7243.12
3600000.	5506.44	13671.71	7243.54
3628000.	5491.48	13664.57	7243.76
3630000.	5491.48	13664.56	7243.72
3650000.	5481.51	13640.31	7234.58
3700000.	5456.60	13580.16	7211.87
3750000.	5431.71	13520.64	7189.32
3800000.	5406.84	13461.67	7166.90
3850000.	5381.97	13403.16	7144.58
3900000.	5357.09	13345.05	7122.34
3950000.	5332.19	13287.25	7100.15
4000000.	5307.27	13229.69	7077.99
4050000.	5282.31	13172.30	7055.83
4100000.	5257.32	13114.98	7033.63
4150000.	5232.27	13057.67	7011.38
4200000.	5207.16	13000.29	6989.05
4250000.	5181.98	12942.76	6966.61
4300000.	5156.72	12885.01	6944.04
4350000.	5131.38	12826.95	6921.29
4400000.	5105.93	12768.51	6898.37
4450000.	5080.39	12709.62	6875.22
4500000.	5054.72	12650.19	6851.83
4550000.	5028.94	12590.15	6828.17
4600000.	5003.02	12529.42	6804.21
4650000.	4976.97	12467.92	6779.93
4700000.	4950.76	12405.58	6755.29
4750000.	4924.39	12342.32	6730.28
4800000.	4897.86	12278.06	6704.87
4850000.	4871.15	12212.73	6679.02
4900000.	4844.25	12146.24	6652.72
4950000.	4817.16	12078.52	6625.93
5000000.	4789.87	12009.50	6598.63
5050000.	4762.36	11939.09	6570.79
5100000.	4734.63	11867.22	6542.38
5150000.	4706.67	11793.81	6513.38
5200000.	4678.48	11718.79	6483.77
5250000.	4650.03	11642.07	6453.50
5300000.	4621.33	11563.58	6422.57
5350000.	4592.36	11483.25	6390.93
5400000.	4563.11	11400.99	6358.57
5450000.	4533.58	11316.72	6325.45
5500000.	4503.76	11230.38	6291.56
5550000.	4473.64	11141.88	6256.86
5596000.	4445.64	11058.41	6224.16
5621000.	4410.00	10962.00	6200.00
5646000.	4401.25	10884.25	6126.67
5671000.	4392.50	10806.50	6053.33
5696000.	4383.75	10728.75	5980.00
5711000.	4380.00	10701.00	5910.00
5713000.	4048.90	10303.96	5531.71
5736000.	4019.91	10206.82	5481.48
5771000.	3975.80	10059.00	5405.03
5821000.	3912.78	9847.83	5295.83
5871000.	3849.77	9636.65	5186.62

5921000.	3786.75	9425.48	5077.41
5970670.	3724.15	9215.70	4968.93
5970670.	3541.09	8646.91	4735.57
6016000.	3513.63	8519.50	4683.86
6061000.	3486.37	8393.03	4632.52
6106000.	3459.11	8266.56	4581.19
6151000.	3431.85	8140.08	4529.85
6151000.	3431.85	8185.55	4527.38
6186000.	3412.42	8185.55	4527.38
6221000.	3392.99	8185.55	4527.38
6256000.	3373.55	8185.55	4527.38
6271000.	3365.23	8185.55	4527.38
6271000.	3365.23	8185.60	4505.56
6311000.	3336.95	8185.60	4505.56
6335157.	3319.87	8185.60	4505.56
6335157.	3000.00	6777.60	3842.30
6351000.	3000.00	6777.60	3842.30
6351000.	2740.00	6077.60	3438.16
6368000.	2740.00	6077.60	3438.16
6368000.	2000.00	2977.60	1416.32
6371000.	2000.00	2977.60	1416.32

Table C.2 Model SNA

R (m)	ρ (kg/m**3)	Vp (m/s)	Vs (m/s)
0.	13088.50	11237.30	3626.46
50000.	13087.96	11236.91	3626.19
100000.	13086.32	11235.74	3625.37
150000.	13083.60	11233.79	3624.02
200000.	13079.79	11231.07	3622.12
250000.	13074.89	11227.56	3619.69
300000.	13068.90	11223.27	3616.71
350000.	13061.83	11218.20	3613.19
400000.	13053.66	11212.36	3609.12
450000.	13044.41	11205.73	3604.52
500000.	13034.06	11198.33	3599.37
550000.	13022.63	11190.14	3593.69
600000.	13010.11	11181.18	3587.46
650000.	12996.50	11171.43	3580.69
700000.	12981.81	11160.91	3573.37
750000.	12966.02	11149.60	3565.52
800000.	12949.14	11137.52	3557.12
850000.	12931.18	11124.66	3548.18
900000.	12912.13	11111.02	3538.71
950000.	12891.99	11096.59	3528.68
1000000.	12870.76	11081.39	3518.12
1050000.	12848.44	11065.41	3507.02
1100000.	12825.03	11048.65	3495.37
1150000.	12800.54	11031.11	3483.18
1200000.	12774.95	11012.79	3470.45
1221500.	12763.61	11004.67	3464.81
1221500.	12166.33	10355.54	0.00
1250000.	12151.57	10338.97	0.00
1300000.	12124.99	10309.58	0.00
1350000.	12097.55	10279.76	0.00
1400000.	12069.23	10249.46	0.00
1450000.	12040.01	10218.66	0.00
1500000.	12009.88	10187.30	0.00
1550000.	11978.82	10155.36	0.00
1600000.	11946.81	10122.78	0.00
1650000.	11913.84	10089.54	0.00
1700000.	11879.89	10055.59	0.00
1750000.	11844.95	10020.90	0.00
1800000.	11809.00	9985.41	0.00
1850000.	11772.02	9949.11	0.00
1900000.	11734.01	9911.93	0.00
1950000.	11694.93	9873.85	0.00
2000000.	11654.78	9834.83	0.00
2050000.	11613.54	9794.83	0.00
2100000.	11571.19	9753.80	0.00
2150000.	11527.72	9711.72	0.00
2200000.	11483.11	9668.53	0.00
2250000.	11437.35	9624.20	0.00
2300000.	11390.42	9578.69	0.00
2350000.	11342.30	9531.96	0.00
2400000.	11292.98	9483.97	0.00
2450000.	11242.45	9434.69	0.00
2500000.	11190.67	9384.06	0.00
2550000.	11137.65	9332.06	0.00
2600000.	11083.36	9278.65	0.00
2650000.	11027.79	9223.78	0.00
2700000.	10970.92	9167.41	0.00
2750000.	10912.73	9109.51	0.00
2800000.	10853.22	9050.04	0.00

2850000.	10792.36	9988.95	0.00
2900000.	10730.13	8926.21	0.00
2950000.	10666.53	8861.78	0.00
3000000.	10601.53	8795.62	0.00
3050000.	10535.12	8727.69	0.00
3100000.	10467.29	8657.95	0.00
3150000.	10398.01	8586.36	0.00
3200000.	10327.27	8512.88	0.00
3250000.	10255.06	8437.47	0.00
3300000.	10181.36	8360.10	0.00
3350000.	10106.15	8280.72	0.00
3400000.	10029.42	8199.29	0.00
3450000.	9951.15	8115.78	0.00
3480000.	9903.44	8064.65	0.00
3480000.	5566.46	13700.80	7242.45
3500000.	5556.43	13695.89	7242.65
3550000.	5531.41	13683.72	7243.12
3600000.	5506.44	13671.71	7243.54
3628000.	5491.48	13664.57	7243.76
3630000.	5491.48	13664.56	7243.72
3650000.	5481.51	13640.31	7234.58
3700000.	5456.60	13580.16	7211.87
3750000.	5431.71	13520.64	7189.32
3800000.	5406.84	13461.67	7166.90
3850000.	5381.97	13403.16	7144.58
3900000.	5357.09	13345.05	7122.34
3950000.	5332.19	13287.25	7100.15
4000000.	5307.27	13229.69	7077.99
4050000.	5282.31	13172.30	7055.83
4100000.	5257.32	13114.98	7033.63
4150000.	5232.27	13057.67	7011.38
4200000.	5207.16	13000.29	6989.05
4250000.	5181.98	12942.76	6966.61
4300000.	5156.72	12885.01	6944.04
4350000.	5131.38	12826.95	6921.29
4400000.	5105.93	12768.51	6898.37
4450000.	5080.39	12709.62	6875.22
4500000.	5054.72	12650.19	6851.83
4550000.	5028.94	12590.15	6828.17
4600000.	5003.02	12529.42	6804.21
4650000.	4976.97	12467.92	6779.93
4700000.	4950.76	12405.58	6755.29
4750000.	4924.39	12342.32	6730.28
4800000.	4897.86	12278.06	6704.87
4850000.	4871.15	12212.73	6679.02
4900000.	4844.25	12146.24	6652.72
4950000.	4817.16	12078.52	6625.93
5000000.	4789.87	12009.50	6598.63
5050000.	4762.36	11939.09	6570.79
5100000.	4734.63	11867.22	6542.38
5150000.	4706.67	11793.81	6513.38
5200000.	4678.48	11718.79	6483.77
5250000.	4650.03	11642.07	6453.50
5300000.	4621.33	11563.58	6422.57
5350000.	4592.36	11483.25	6390.93
5400000.	4563.11	11400.99	6358.57
5450000.	4533.58	11316.72	6325.45
5500000.	4503.76	11230.38	6291.56
5550000.	4473.64	11141.88	6256.86
5596000.	4445.64	11058.41	6224.16
5621000.	4410.00	10962.00	6200.00
5646000.	4401.25	10884.25	6126.67
5671000.	4392.50	10806.50	6053.33
5696000.	4383.75	10728.75	5980.00
5711000.	4380.00	10701.00	5910.00
5713000.	4048.90	10303.96	5531.71
5736000.	4019.91	10206.82	5481.48
5771000.	3975.80	10059.00	5405.03
5921000.	3912.78	9847.83	5295.83
5871000.	3849.77	9636.65	5186.62

5921000.	3786.75	9425.48	5077.41
5970670.	3724.15	9215.70	4968.93
5970670.	3541.09	8646.91	4780.00
5996000.	3525.75	8575.71	4765.00
6016000.	3513.63	8519.50	4757.00
6046000.	3495.46	8435.19	4740.00
6061000.	3486.37	8393.03	4728.00
6106000.	3459.11	8266.56	4685.00
6121000.	3450.02	8224.40	4670.00
6151000.	3431.85	8140.08	4638.00
6151000.	3431.85	8185.55	4638.00
6171000.	3420.75	8185.55	4630.00
6186000.	3412.42	8185.55	4678.00
6196000.	3406.87	8185.55	4710.00
6221000.	3392.96	8185.55	4775.00
6256000.	3373.55	8185.55	4775.00
6271000.	3365.23	8185.55	4775.00
6271000.	3365.23	8185.60	4775.00
6296000.	3347.56	8185.60	4775.00
6311000.	3336.95	8185.60	4784.00
6321000.	3329.88	8185.60	4790.00
6335157.	3319.87	8185.60	4800.00
6335157.	3000.00	6777.60	3842.30
6351000.	3000.00	6777.60	3842.30
6351000.	2740.00	6077.60	3438.16
6368000.	2740.00	6077.60	3438.16
6368000.	2000.00	2977.60	1416.32
6371000.	2000.00	2977.60	1416.32

Table C.3 Q model

R (km)	Qa	Qb
0.00	3666	183300
1221.50	3666	183300
1221.50	999999999	999999999
3480.00	999999999	999999999
3480.00	304	15210
5711.00	304	15210
5713.00	264	13180
5970.67	264	13180
5970.67	118	5898
6371.00	118	5898

APPENDIX D

QUADRATIC DISPERSION

INTRODUCTION

In Chapter 2 we derived expressions for the cross-correlation between an isolation filter and an observed seismogram based on a Taylor series expansion of differential wavenumber. In the case of the filtered cross-correlation F_iC_{ff}, we showed that the quadratic terms could be controlled by the narrow-band filter and thus safely neglected. In the case of the filtered, windowed cross-correlation, however, the application of the windowing operator changes the expansion of differential wavenumber from a local approximation at ω_i' to a global approximation over the band $(\omega_i' - \omega_f)$. We shall investigate this effect in greater detail in this appendix.

QUADRATIC DISPERSION

In the case of quadratic dispersion, the Taylor-series expansion for differential wavenumber has the following form:

$$\delta k_m(\omega)\, x \approx [\delta k_m(\omega_i') + (\omega-\omega_i')\,\delta \dot{k}_m(\omega_i') + \frac{1}{2}(\omega-\omega_i')^2\,\delta \ddot{k}_m(\omega_i') + \cdots]\, x$$

$$\approx \omega_i' \delta\tau_p^m(\omega_i') + (\omega-\omega_i')[\delta\tau_g^m(\omega_i') + i\,\delta\tau_a^m(\omega_i')] + \frac{1}{2}(\omega-\omega_i')^2\,\delta\tau_s^{m\,2}(\omega_i') + \cdots \quad \text{(D.1)}$$

where the "superdot" denotes the derivative with respect to ω. $\delta\tau_p^m(\omega_i') \equiv x\ \mathrm{Re}\{\delta k_m(\omega_i')\} / \omega_i'$ is the differential phase delay at the center frequency ω_i'; $\delta\tau_g^m(\omega_i') \equiv x\ \mathrm{Re}\{\delta \dot{k}_m(\omega_i')\}$

is the differential group delay; $\delta\tau_a^m(\omega_i') \equiv x\ \mathrm{Im}\{\dot{\delta k}_m(\omega_i')\}$ is a differential amplitude factor; and $\delta\tau_s^{m\,2}(\omega_i') \equiv x\ \ddot{\delta k}_m(\omega_i')$ measures the differential curvature of the dispersion function and is, in general, a complex constant.

ANALYSIS OF NARROW-BAND SEISMOGRAMS

In Chapter 2, we developed an analysis technique for narrow-band seismograms. In the case of quadratic dispersion, our expression for $F_i C_{\tilde{u}u}(t)$ has the form:

$$F_i C_{\tilde{u}u}(t) = \mathrm{Re}\Big\{ \frac{1}{(2\pi)^{3/2}} \frac{\sigma_z}{\sigma_i \sigma_f} D_0^m\ \mathrm{Ga}\left(\frac{\omega_i-\omega_f}{\sqrt{\sigma_i^2+\sigma_f^2}}\right) \exp[(\sigma_z \delta\tau_a^m)^2/2]\ \mathrm{Ga}\,(\sigma_z(t-\delta\tau_g^m)) \times \sum_{j=0}^{\infty} c_j\,(\sigma_z(t-\delta\tau_g^m))^j\ \exp[-i(\omega_i'(t-\delta\tau_p^m) - \sigma_z^2\delta\tau_a^m(t-\delta\tau_g^m) + j\tfrac{\pi}{2})] \Big\} \quad \text{(D.2)}$$

where the c_j coefficients are complex and depend on the coefficients of the filtered autocorrelation, the differential attenuation, and differential curvature:

$$c_j = \sum_{l=0}^{\infty} \binom{l+j}{l} b_{l+j} (-\sigma_z \delta\tau_a^m)^l \quad \text{(D.3)}$$

In this expression for the filtered cross-correlation function, the quadratic dispersion term enters only through the effective half-width σ_z:

$$\sigma_z^2 = \frac{\sigma_i'^2\ \sigma_s^{m\,2}}{\sigma_i'^2 + \sigma_s^{m\,2}} \quad \text{(D.4)}$$

where $\sigma_s^{m2} \equiv i/\delta\tau_s^{m2}$. If $\sigma_i'^2 << \sigma_s^{m2}$, then $\sigma_z = \sigma_i'$, allowing us to control the effect of quadratic dispersion through the parameters of the narrow-band filter. Of course, $\delta\tau_s^{m2}$ is

proportional to the epicentral distance, which means effect of quadratic dispersion will increase with distance. The maximum value of $\mathrm{Re}\{\delta\tau_s^{m^2}\}$ from EU2-SNA at 70° for the fundamental Love wave is 1489 s^2 at 22 mHz; the maximum value for the fundamental Rayleigh wave is 1640 s^2 at 38 mHz. These values imply that a γ_i' of 0.1 or so is sufficient to control the contribution of quadratic dispersion in this extreme case.

ANALYSIS OF BROAD-BAND SEISMOGRAMS

As we discussed in Chapter 3, the addition of a windowing operator in our processing scheme limits the approximation of differential dispersion. In particular, our approximation in (D.1) is no longer an expansion in the neighborhood of ω_i', instead, the approximation spans the band $(\omega_i' - \omega_f)$. Using our Hermite-polynomial expansion, we derive an expression for $F_iWC_{\tilde{u}\tilde{u}}(t)$ in the case of quadratic dispersion. This requires carrying the quadratic term through the operations of windowing and filtering. Because the quadratic term is complex-valued, the resulting expression contains frequency and bandwidth parameters which are complex-valued. We write this formula as a Gaussian envelope times a harmonic function, where we take the real value of their product:

$$F_iWC_{\tilde{u}\tilde{u}}(t) = \mathrm{Re}\,\{E(t)\exp[-i\Phi(t)]\} \tag{D.5}$$

where $E(t)$ and $\Phi(t)$ are defined by:

$$E(t) = E\ \mathrm{Ga}\,(\sigma_z(t-\delta\tau_g^m)) \tag{D.5a}$$

$$\Phi(t) = \omega_i'(t-\delta\tau_p^m) - (\omega_i' - (\omega_z - \sigma_z^2\delta\tau_a^m))(t-\delta\tau_g^m) \tag{D.5b}$$

and we have made the wide-window approximation. The envelope coefficient has the form:

$$E = \frac{1}{(2\pi)^{5/2}} \mathrm{Re}\left\{\frac{\sigma_x\sigma_z}{\sigma_i\sigma_f\sigma_y} D_0^m \, \mathrm{Ga}\left(\frac{\omega_i'-\omega_f}{\sqrt{\sigma_s^2+\sigma_f^2}}\right) \mathrm{Ga}\left(\frac{\sigma_x\sigma_w}{\sigma_y}(\delta\tau_c-\delta\tau_g^m)\right) \mathrm{Ga}\left(\frac{\omega_i-\omega_x}{\sqrt{\sigma_i^2+\sigma_y^2}}\right)\right. \quad \text{(D.5c)}$$

The constants ω_z,σ_z and σ_y are complex:

$$\omega_z = \frac{\sigma_i^2\omega_x + (\sigma_w^2+\sigma_x^2)\omega_i}{\sigma_i^2+\sigma_w^2+\sigma_x^2} \quad \text{(D.6)}$$

$$\sigma_z^2 = \frac{\sigma_i^2(\sigma_w^2+\sigma_x^2)}{\sigma_i^2+\sigma_w^2+\sigma_x^2} \quad \text{(D.7)}$$

$$\sigma_y^2 = \sigma_x^2 + \sigma_w^2 \quad \text{(D.8)}$$

The influence of quadratic dispersion appears through the parameters ω_x and σ_x, which depend on the differential quadratic contribution:

$$\omega_x = \frac{\sigma_s^{m\,2}\omega_f + \sigma_f^2\omega_i'}{\sigma_s^{m\,2}+\sigma_f^2} = \frac{\omega_f - i\gamma_d^2\omega_i'}{1-i\gamma_d^2} \quad \text{(D.9)}$$

$$\sigma_x^2 = \frac{\sigma_s^{m\,2}\sigma_f^2}{\sigma_s^{m\,2}+\sigma_f^2} = \frac{\sigma_f^2}{1-i\gamma_d^2} \quad \text{(D.10)}$$

where, as we recall from Chapter 2, $\sigma_s^{m\,2} \equiv i/\delta\tau_s^{m\,2}$ and $\delta\tau_s^{m\,2}(\omega_i') \equiv x\,\delta\ddot{k}_m(\omega_i')$. $\gamma_d^{\,2} \equiv \sigma_f^{\,2}/\delta\tau_s^{m\,2}$ is a measure of the effect of quadratic dispersion. If $\gamma_d^{\,2} << 1$, the contribution of

quadratic dispersion will be negligible. This was the case in Chapter 2, where the application of the narrow-band filter ensured that the spectral half-width of the correlation function was small.

Since we may parameterize the effect of quadratic dispersion on $F_i WC_{\overline{uu}}(t)$ exactly, we could use the complete expression for quadratic dispersion in our waveform-fitting algorithm and invert for the quadratic term at each ω_i', in addition to $\delta\tau_p^n$, $\delta\tau_g^n$, and $\delta\tau_a^n$. In Chapter 3, we described a procedure through which we could estimate the contribution of quadratic dispersion. Using this approach, we obtained unbiased estimates of the generalized data functionals. In the future, we anticipate developing a different correction scheme for quadratic dispersion. In particular, we wish to develop an approach which will not affect the measurement procedure. We anticipate using the quadratic expression (D.5) in conjunction with our estimate of the peak of the cross-correlation function to derive correction terms for the effect of quadratic dispersion. Such corrections could be applied after the measurements were made.

APPENDIX E

GRAM-CHARLIER EXPANSIONS

INTRODUCTION

In Chapters 2 and 3, we developed a number of expansions using Gram-Charlier series. In general, we did not include explicit expressions for the coefficients because of the their complexity. In this appendix, we summarize the various Gram-Charlier expansions discussed in this thesis and include expressions for their coefficients. The expressions presented here are those for an isolated waveform; those for the complete seismograms are algebraically more complex, but hardly more enlightening. In this appendix, we have dropped the dependence on the branch index m.

COEFFICIENTS OF THE GRAM-CHARLIER SERIES

Autocorrelation of the Isolation Filter

The expansion of the autocorrelation function, $C_{\bar{u}\bar{u}}(t)$, of the isolation filter is given in Chapter 2 (2.11):

$$C_{\bar{u}\bar{u}}(t) = \frac{1}{2\pi}\,\mathrm{Ga}\,(\sigma_f t)\sum_{k=0}^{\infty} a_k\,\gamma_f^k\,(\omega_f t)^k \cos(\omega_f t + k\tfrac{\pi}{2}) \qquad \text{(E.1)}$$

where the a_k are real:

$$a_k = \frac{1}{k!}\int_0^{\infty} C_{\tilde{u}\tilde{u}}(\omega)\,\mathrm{He}_k\left(\frac{\omega-\omega_f}{\sigma_f}\right)d\omega \qquad \text{(E.2)}$$

If $C_{\tilde{u}\tilde{u}}(\omega)$ is normalized such that $\hat{\mu}_0(0) = 1$, then $a_0 = 1$. If the location parameter ω_f is chosen to be the center frequency of $C_{\tilde{u}\tilde{u}}(\omega)$ (*i.e.*, $\omega_f = \sigma_f \hat{\mu}_1(0)$), then $a_1 = 0$. Finally, if the scale parameter σ_f is the half-width of $C_{\tilde{u}\tilde{u}}(\omega)$ (*i.e.*, $1 = \hat{\mu}_2(\omega_f)$), then $a_2 = 0$. Thus, with these choices, the zeroth-order term of the Hermite-polynomial expansion is the best-fitting Gaussian, and the first correction term is third-order. $\gamma_f \equiv \sigma_f / \omega_f$ measures the relative bandwidth of the autocorrelation function.

Cross-Correlation of an Isolated Waveform - Quadratic Dispersion

The cross-correlation function, $C_{\tilde{u}u}(t)$, of an isolated waveform in the case of quadratic dispersion may be written in the following form:

$$C_{\tilde{u}u}(t) = \mathrm{Re}\{\frac{1}{2\pi} D_0 \frac{\sigma_x}{\sigma_f} \exp(\sigma_x^2\delta\tau_a^2/2 + (\omega_i'-\omega_x)\delta\tau_a)\ \mathrm{Ga}\,(\sigma_x(t-\delta\tau_g))$$
$$\sum_{k=0}^{\infty} b_k\, \gamma_x^k\, (\omega_x(t-\delta\tau_g))^k\ \exp[-i(\omega_i'(t-\delta\tau_p) - (\omega_i'-\omega_x+\sigma_x^2\delta\tau_a)(t-\delta\tau_g) + k\,\frac{\pi}{2})] \qquad \text{(E.3)}$$

where the b_k are complex and defined:

$$b_k = \sum_{l=k}^{\infty} \binom{l}{k} (-\sigma_x\delta\tau_a)^{l-k} \sum_{j=l}^{\infty} \binom{j}{l} \left(\frac{\omega_x-\omega_f}{\sigma_x}\right)^{j-l} \sum_{m=j}^{\infty} a_m \left(\frac{\sigma_x}{\sigma_f}\right)^m \frac{m!\ (-1/2)^{(m-j)/2}}{j!\ ((m-j)/2)!} \left(\frac{\sigma_f^2-\sigma_x^2}{\sigma_x^2}\right)^{(m-j)/2} \qquad \text{(E.4)}$$

and $(m-j)/2$ assumes only integer values. $\gamma_x \equiv \sigma_x/\omega_x$, where σ_x and ω_x are defined in (D.9) and (D.10).

Cross-Correlation of an Isolated Waveform - Linear Dispersion

The cross-correlation function, $C_{\bar{u}u}(t)$, of an isolated waveform in the case of linear dispersion was discussed in Chapters 2 and 3 and may be written in the following form:

$$C_{\bar{u}u}(t) = \frac{1}{2\pi} D_0 \exp(\sigma_f^2 \delta\tau_a^2/2 + (\omega_i' - \omega_f)\delta\tau_a) \; \mathrm{Ga}\,(\sigma_f(t-\delta\tau_g)) \sum_{k=0}^{\infty} c_k \gamma_f^k (\omega_f(t-\delta\tau_g))^k \cos(\omega_i'(t-\delta\tau_p) - (\omega_i' - \omega_f + \sigma_f^2\delta\tau_a)(t-\delta\tau_g) + k\frac{\pi}{2}) \qquad \text{(E.5)}$$

where the c_k are real and defined:

$$c_k = \sum_{l=k}^{\infty} a_l \binom{l}{k} (-\sigma_f \delta\tau_a)^{l-k} \qquad \text{(E.6)}$$

As in the case of the autcorrelation function, the expansion parameter is γ_f.

Filtering Operator

The general filtering operator, $F_i(\omega)$, is defined in Chapter 2 (2.14):

$$F_i(\omega)\,\mathrm{H}(\omega) = \frac{1}{\sqrt{2\pi}\,\sigma_i} \mathrm{Ga}\left(\frac{\omega-\omega_i}{\sigma_i}\right) \sum_{m=0}^{\infty} f_m \mathrm{He}_m\left(\frac{\omega-\omega_i}{\sigma_i}\right) \qquad \text{(E.7)}$$

where the f_m may be complex:

$$f_m = \frac{1}{m!} \int_0^{\infty} F_i(\omega)\,\mathrm{He}_m\left(\frac{\omega-\omega_i}{\sigma_i}\right) d\omega \qquad \text{(E.8)}$$

We assume that $f_0 = 1$ and $f_1 = f_2 = 0$.

Filtered Autocorrelation Function

The expression for the autocorrelation function convolved with a Gaussian filter, $F_i C_{\tilde{u}\tilde{u}}(t)$, is defined in Chapter 2 (2.16):

$$F_i C_{\tilde{u}\tilde{u}}(t) = \frac{1}{(2\pi)^{3/2}} \frac{\sigma_i'}{\sigma_f \sigma_i} \mathrm{Ga}\left(\frac{\omega_i - \omega_f}{\sqrt{\sigma_i^2 + \sigma_f^2}}\right) \mathrm{Ga}(\sigma_i' t) \sum_{l=0}^{\infty} a_l' \gamma_i'^{\,l} (\omega_i' t)^l \cos(\omega_i' t + l\frac{\pi}{2}) \qquad \text{(E.9)}$$

where the a_l' are real:

$$a_l' = \sum_{m=l}^{\infty} \binom{m}{l} \left(\frac{\omega_i' - \omega_f}{\sigma_i'}\right)^{m-l} \sum_{k=m}^{\infty} a_k \left(\frac{\sigma_i'}{\sigma_f}\right)^k \frac{k!\,(-1/2)^{(k-m)/2}}{m!\,((k-m)/2)!} \left(\frac{\sigma_f^2 - \sigma_i'^2}{\sigma_i'^2}\right)^{(k-m)/2} \qquad \text{(E.10)}$$

and $(k-m)/2$ assumes only integer values. γ_i' measures the relative bandwidth of the filtered autcorrelation; ω_i' and σ_i' are defined in (2.17) and (2.18).

Filtered Cross-Correlation Function - Quadratic Dispersion

The expression for the filtered cross-correlation $F_i C_{\tilde{u}u}(t)$ is given in Chapter 2 for the case of quadratic dispersion (2.25):

$$F_i C_{\tilde{u}u}(t) = \mathrm{Re}\{ \frac{1}{(2\pi)^{3/2}} \frac{\sigma_z}{\sigma_f \sigma_i} D_0 \ \mathrm{Ga}\left(\frac{\omega_i - \omega_f}{\sqrt{\sigma_i^2 + \sigma_f^2}}\right) \exp(\sigma_z^2 \delta\tau_a^2/2) \ \mathrm{Ga}(\sigma_z(t - \delta\tau_g))$$
$$\sum_{j=0}^{\infty} b_j' (\sigma_z(t - \delta\tau_g))^j \exp[-i(\omega_i'(t - \delta\tau_p) - \sigma_z^2 \delta\tau_a (t - \delta\tau_g) + j\frac{\pi}{2})]\} \qquad \text{(E.11)}$$

where the b_j' are complex and defined:

$$b_j' = \sum_{l=j}^{\infty} a_l' \binom{l}{j} (-\sigma_z \delta\tau_a)^{l-j} \qquad \text{(E.12)}$$

The differentaial quadratic term enters these expressions through σ_z, which is defined in (2.27).

Filtered Cross-Correlation Function - Linear Dispersion

The expression for the filtered cross-correlation is given in Chapter 2 for the cae of linear dispersion (2.28):

$$F_i C_{au}(t) = \frac{1}{(2\pi)^{3/2}} \frac{\sigma_i'}{\sigma_f \sigma_i} D_0 \exp(\sigma_i'^2 \delta\tau_a^2/2)\ \mathrm{Ga}\left(\frac{\omega_i - \omega_f}{\sqrt{\sigma_i^2 + \sigma_f^2}}\right) \mathrm{Ga}\,(\sigma_i'(t-\delta\tau_g))$$
$$\sum_{j=0}^{\infty} c_j' \gamma_i'^j (\omega_i'(t-\delta\tau_g))^j \cos(\omega_i'(t-\delta\tau_p) - \sigma_i'^2 \delta\tau_a(t-\delta\tau_g) + j\frac{\pi}{2}) \qquad \text{(E.13)}$$

where the c_j are real:

$$c_j' = \sum_{l=j}^{\infty} a_l' \binom{l}{j} (-\sigma_i' \delta\tau_a)^{l-j} \qquad \text{(E.14)}$$

Windowing Operator

The expansion of the windowing operator $W(t)$ is given in Chapter 3

$$W(t) = \frac{1}{2\pi} \mathrm{Ga}\left(\sigma_w(t-t_w)\right) \sum_{\substack{n=0 \\ n\,\mathrm{even}}}^{\infty} w_n(-i\sigma_w(t-t_w))^n \qquad \text{(E.15)}$$

and the coefficients w_n are real:

$$w_n = \frac{1}{n!} \int_0^{\infty} W(\omega)\, \mathrm{He}_n\left(\frac{\omega}{\sigma_w}\right) d\omega \qquad \text{(E.16)}$$

We assume that $w_0 = 1$ and $w_2 = 0$.

Windowed Autocorrelation Function

The expansion of the windowed autocorrelation function of the isolated waveform, $WC_{\tilde{u}\tilde{u}}(t)$, is of the form:

$$WC_{\tilde{u}\tilde{u}}(t) = \frac{1}{(2\pi)^2} \mathrm{Ga}\left(\sigma_g t\right) \sum_{l=0}^{\infty} p_l\, \gamma_g^l\, (\omega_f t)^l \cos(\omega_f t + l\frac{\pi}{2}) \qquad \text{(E.17)}$$

where σ_g^{-1} is effective half-width of the windowed autocorrelation ($\sigma_g^2 = \sigma_f^2 + \sigma_w^2$) and we have assumed that t_w is zero on the case of the autocorrelation. $\gamma_g \equiv \sigma_g / \omega_f$ is a measure of the relative bandwidth of the windowed correlation function. The p_l are defined:

$$p_l = \sum_{\substack{n=0 \\ n\,\mathrm{even}}}^{l} w_n\, a_{l-n}\left(\frac{\sigma_w^n\, \sigma_f^{l-n}}{\sigma_g^l}\right) \qquad \text{(E.18)}$$

The Windowed Cross-Correlation Function

The expansion of the windowed cross-correlation function, $WC_{\tilde{u}u}(t)$, has the following form in the case of linear dispersion:

$$WC_{\tilde{u}u}(t) = \frac{1}{(2\pi)^2} D_0 \exp\{\sigma_f^2 \delta\tau_a^{\hat{}}/2 + (\omega_i' - \omega_f)\delta\tau_a\} \, \mathrm{Ga}\left(\frac{\sigma_f \sigma_w}{\sigma_g}(\delta t_c - \delta\tau_g')\right) \, \mathrm{Ga}\,(\sigma_g(t-\delta\tau_g')) \sum_{l=0}^{\infty} q_l \, \gamma_g^l \, (\omega_f(t-\delta\tau_g'))^l \cos(\omega_i'(t-\delta\tau_p) - (\omega_i' - \omega_f + \sigma_f^2 \delta\tau_a)(t-\delta\tau_g) + l\frac{\pi}{2})\} \quad \text{(E.19)}$$

where the q_l are defined:

$$q_l = i^l \sum_{j=0}^{l} \left[\sum_{\substack{n=j \\ n \text{ even}}}^{\infty} \binom{n}{j} w_n (-i)^n \frac{\sigma_w^n}{\sigma_g^j} (\delta\tau_g' - \delta t_c)^{n-j} \right] \left[\sum_{m=l-j}^{\infty} \binom{m}{l-j} b_m (-i)^m \frac{\sigma_f^m}{\sigma_g^{l-j}} (\delta\tau_g' - \delta\tau_g)^{m-(l-j)} \right] \text{(E.20)}$$

$\delta\tau_g'$ is defined in (3.12); we have assumed that the windowing operator is centered at the peak of the cross-correlation functions, $t_w \equiv \delta t_c$.

The Filtered Windowed Autocorrelation Function

The expansion of the filtered, windowed autocorrelation function of the isolation filter is given in Chapter 3 for the case of the Gaussian filter (3.6):

$$F_i WC_{\tilde{u}\tilde{u}}(t) = \frac{1}{(2\pi)^{5/2}} \frac{\sigma_i'}{\sigma_i \sigma_g} \mathrm{Ga}\left(\frac{\omega_i - \omega_f}{\sqrt{\sigma_i^2 + \sigma_g^2}}\right) \mathrm{Ga}\,(\sigma_i' t) \sum_{k=0}^{\infty} d_k \, \gamma_i'^k \, (\omega_i' t)^k \cos(\omega_i' t + k\frac{\pi}{2}) \text{(E.21)}$$

where the d_k are defined:

$$d_k = \sum_{j=k}^{\infty} \binom{j}{k} \left(\frac{\omega_i' - \omega_f}{\sigma_i'} \right)^{j-k} \sum_{l=j}^{\infty} p_l \frac{l!\,(-1/2)^{(l-j)/2}}{j!\,((l-j)/2)!} \left(\frac{\sigma_i'}{\sigma_g} \right)^{l} \left(\frac{\sigma_g^2 - \sigma_i'^2}{\sigma_i'^2} \right)^{(l-j)/2} \tag{E.22}$$

and $(l-j)/2$ assumes only integer values. ω_i' and σ_i' are defined in (3.7) and (3.8).

The Filtered Windowed Cross-Correlation Function - Linear Dispersion

The expansion of the filtered, windowed cross-correlation function is given in (3.13) for the case of linear dispersion and a Gaussian filter

$$\begin{aligned} F_i WC_{aa}(t) = {} & \frac{1}{(2\pi)^{5/2}} \frac{\sigma_i'}{\sigma_i \sigma_g} D_0 \exp\{\sigma_f^2 \delta\tau_a^2/2 + (\omega_i' - \omega_f)\delta\tau_a\} \, \mathrm{Ga}\left(\frac{\sigma_f \sigma_w}{\sigma_g} (\delta\tau_c - \delta\tau_g) \right) \\ & \mathrm{Ga}\left(\frac{\omega_i - (\omega_f - \sigma_f^2 \delta\tau_a))}{\sqrt{\sigma_i^2 + \sigma_g^2}} \right) \mathrm{Ga}\,(\sigma_i'(t - \delta\tau_g')) \; \mathrm{Re}\,\{ \sum_{k=0}^{\infty} e_k \, \gamma_i'^{k} \, (-i\omega_i'(t - \delta\tau_g'))^k \\ & \exp[-i((\omega_i'(t - \delta\tau_p') - \sigma_i'^2 \delta\tau_a'(t - \delta\tau_g')))]\} \end{aligned} \tag{E.23}$$

where the e_k are defined:

$$e_k = \sum_{j=k}^{\infty} \binom{j}{k} \left(\frac{\omega_i' - (\omega_f - \sigma_f^2 \delta\tau_a)}{\sigma_i'} \right)^{j-k} \sum_{l=j}^{\infty} q_l \frac{l!\,(-1/2)^{(l-j)/2}}{j!\,((l-j)/2)!} \left(\frac{\sigma_i'}{\sigma_g} \right)^{l} \left(\frac{\sigma_g^2 - \sigma_i'^2}{\sigma_i'^2} \right)^{(l-j)/2} \tag{E.24}$$

and $(l-j)/2$ assumes only integer values. The effective differential time parameters are defined in (3.14).

CONCLUSIONS

In this Appendix, we have presented expressions for the complete Gram-Charlier expansions of $C_{\bar{u}\bar{u}}(t)$, $C_{\tilde{u}u}(t)$, $W(t)$, $WC_{\bar{u}\bar{u}}(t)$, $WC_{\tilde{u}u}(t)$, $F_i(\omega)$, $F_i C_{\bar{u}\bar{u}}(t)$, and $F_i C_{\tilde{u}u}(t)$. We obtained these expressions by representing the autocorrelation of the isolation filter, the

windowing operator, and the filter in a Gram-Charlier series. Manipulation of the Hermite-polynomial expansions is algebraically complex, but conceptually simple. We have made extensive use of the theorems presented in Appendices A and B in order to derive these formulae.

APPENDIX F

POLARIZATION ANISOTROPY AND FINE-SCALE STRUCTURE OF THE EURASIAN UPPER MANTLE

Saying there are two kinds of continents is like saying that there are two kinds of women.

T.H. Jordan, DARPA Seismic Symposium, May, 1985

Note

This appendix was published as a short note in *Geophysical Research Letters* in August 1988. The notation, equation numbers and figure numbers have not been revised for consistency in this thesis. This paper presents results from a study of multiply reflected shear waves and fundamental-mode surface waves for two stable continental paths in Eurasia. The differential travel times used in this study were not measured using the waveform-fitting algorithm described in the last section, but were determined from the peak of the cross-correlation function and corrected for differential dispersion and attenuation.

ABSTRACT

We have observed shear-wave splitting, with $(t_{SH}-t_{SV})/t_{SV}$ up to 1.5%, on long-period records of multiply reflected *S* waves bottoming in the upper mantle beneath the Russian and Siberian platforms. The dispersion of Love and Rayleigh waves over these paths shows discrepancies of comparable or larger magnitude with respect to smooth, isotropic (SI) structures, consistent with a model of the uppermost mantle having significant apparent vertical anisotropy. When combined with evidence from *Sn* observations for the frequency dependence of shear velocity, these data suggest a fine-scale ("rough") structure beneath stable Eurasia which can be represented by a random field

whose two-point correlation function has a characteristic vertical wavenumber much larger than its characteristic horizontal wavenumber. We fit the data with a rough, isotropic (RI) model having an rms shear velocity fluctuation that varies from 14% in the uppermost mantle to zero at 400-km depth. These fluctuations are larger than the variation in isotropically averaged parameters expected for even a diverse assemblage of upper-mantle ultrabasic rocks, which we take to be evidence for some sort of intrinsic (local) anisotropy. Our observation that the shear-wave splitting is significantly smaller for paths along the Alpine-Himalayan front suggests that the large-scale variations in the stochastic parameters of a rough, anisotropic (RA) model may be related to continental deep structure.

INTRODUCTION

Although seismological studies have begun to illuminate the fine structure of the upper mantle, very little can yet be said to quantify the spatial distribution of wave-speed heterogeneity and anisotropy. A particularly interesting region for the study of small-scale structure is the continental upper mantle, where comparisons between data and the predictions of smooth, isotropic (SI) earth models reveal several discrepancies:

Love-Rayleigh (LR) discrepancy. The problems of satisfying the dispersion of Love and Rayleigh waves by SI structures is well documented [McEvilly, 1964; Cara *et al.*, 1980; Lévêque and Cara, 1983]. The discrepancy appears to be global in nature, and some form of radial anisotropy (our preferred term for the type of anisotropy whose contrapositive is transverse isotropy) is usually invoked to explain it [*e.g.*, Anderson and Dziewonski, 1982]. The significance of the LR discrepancy for the continents has been debated [James, 1971; Mitchell, 1984].

SNA-EU2 discrepancy. Structures of the upper mantle beneath stable cratons derived from waveform matches to *SH*-polarized waves, specifically the SNA model of Grand and Helmberger [1984], have consistently higher v_S values in the upper 400 km than structures

derived from *PSV*-polarized waves, specifically the EU2 model of Lerner-Lam and Jordan [1987]. Lerner-Lam and Jordan argue that the EU2-SNA discrepancy cannot be entirely explained by path differences and must involve polarization anisotropy in the continental upper mantle.

Sn discrepancy. The apparent velocities of high-frequency (~1 Hz) *Sn* waves on both horizontal and vertical components are typically 100–300 m/s higher than the average shear velocity v_S of the uppermost mantle derived from low-frequency (~.01 Hz) Rayleigh waves. Although not systematically treated in the literature, this problem is evident in the comparison of the *Sn* velocities of Heustis *et al.* [1973] with Rayleigh-wave models such as EU2 (Figure 1).

In this paper, we report some new observations of shear-wave splitting for multiple-*S* waves bottoming beneath Eurasia that are relevant to these discrepancies, and we attempt to explain them in terms of a stochastic model of fine-scale upper-mantle structure.

OBSERVATIONS

We have measured the travel times (both phase and group delays) of direct and multiply reflected *S* waves and fundamental- and higher-mode surface waves on 40 three-component records from the Global Digital Seismic Network (GDSN) stations KONO and GRFO for two corridors across Eurasia (Figure 2). The northern Eurasia corridor is nearly identical to that used in deriving EU2; it includes the marginal basins and active foldbelts east of the Verkhoyansk suture, as well as the stable cratons of the Siberian and Russian platforms. The southwestern Eurasia corridor includes two paths, one crossing the central part of the Russian platform from Hindu Kush events to KONO, and one traversing the southwestern margin of the platform along the Alpine-Himalayan front to GRFO. Large variations in *SS* travel times are observed across the transition from the Russian platform to the Alpine-Himalayan orogenic belt, the latter yielding times delayed by as much as 20-30 s

[Rial *et al.*, 1984]. Grand and Helmberger [1985] find that SNA satisfies the data from *SH* body waves propagating across the central Russian platform.

Our technique for measuring travel times is based on the ability to synthesize complete seismograms by normal-mode summation [Gee and Jordan, manuscript in preparation]. A narrow-band isolation filter for a particular wave group, such as a body-wave pulse or a fundamental-mode surface wave, is computed by a convenient method (we use both ray and mode theoretic algorithms) and cross-correlated with both the observed seismogram and the complete normal-mode synthetic for an appropriate source location and mechanism. The difference between the peak times of the observed and synthetic cross-correlation functions is measured and corrected for differential dispersion and attenuation to obtain the differential phase-delay time Δt at the center frequency f_0 of the isolation filter. Δt thus measures the difference between the true arrival time and the model-predicted time at f_0. This phase-isolation technique provides a self-consistent methodology for measuring the travel times of complex wave groups on three-component seismograms. The complications handled by the technique include caustic phase shifts, interference among multiple arrivals, dispersion, attenuation, and differences in polarization.

For the body-wave data presented in this paper, the center frequencies of the isolation filters range from 25-30 mHz, and their bandwidths from 8-10 mHz; the standard errors of measurement are typically 1-2 s, excluding the bias due to unmodeled interference. The surface waves were measured in the band $10 \le f_0 \le 25$ mHz using isolation filters with bandwidths of $0.15 f_0$, yielding an experimental precision of $\sim 0.1 f_0^{-1}$, or about 5 s for a 20-mHz observation.

Examples of seismograms on which these measurements have been made are presented in Figures 3 and 4. The best match between data and synthetics is obtained by using EU2 as the reference model for computing the isolation filters and complete synthetics for the northern Eurasia corridor and SNA for the southwestern Eurasia corridor (although the results of the data analysis are essentially independent of this choice). EU2

provides a good fit to the Rayleigh waves and *PSV*-polarized waveforms of multiply reflected *S* phases for the northern Eurasia corridor, but is too slow for the Love waves and *SH*-polarized phases with turning points in the upper mantle (Figure 3). SNA is generally consistent with the waveforms and travel times of *SH*-polarized body phases for paths to KONO across the Russian platform, but is too fast for the *PSV*-polarized *SS* phases with turning points in the upper mantle (Figure 4).

On the other hand, the path to GRFO along the southwestern margin of the Russian platform does not show this polarization difference; the observed travel times exceed those predicted by SNA by about 15 s on both components, consistent with the *SH* observations of Rial *et al.* [1984]. Shifting the synthetics by this amount aligns the *SH* and *PSV* waveforms equally well. Therefore, the data-synthetic comparisons indicate strong shear-wave splitting for the northern Eurasia and central Russian platform paths, but not for paths along the Alpine-Himalayan front.

To quantify the polarization difference in the data-model residual for a specific source-receiver pair, we define an "apparent splitting time" by

$$\Delta\tau = \Delta t_{SH} - \Delta t_{PSV} \tag{1}$$

where Δt_{SH} is the phase-delay time measured for a particular wave group on the transverse component and Δt_{PSV} is the phase-delay time for the corresponding wave group on the vertical component. In the case of body waves, we employ the standard phase notation as a subscript; e.g., $\Delta\tau_{SS}$ is the difference between the travel-time residual for an *SS* arrival on the transverse component and its residual on the vertical component. For surface waves, however, we choose Δt_{SH} to be the Love-wave phase delay and Δt_{PSV} to be the Rayleigh-wave phase delay at the same f_0 and let $\Delta\tau_{LR}$ denote the difference. Hence, the values of $\Delta\tau$ for body waves are essentially independent of the reference model (to the extent that it

correctly predicts the interference effects), while the values of $\Delta\tau_{LR}$ depend on the differential dispersion between Love and Rayleigh waves of the reference model.

Figure 5 summarizes the data for northern Eurasia (circles) and the Russian platform (crosses). The apparent splitting time, expressed as a percentage of the total *PSV* travel time, is plotted against the horizontal slowness of the wavegroup computed from the EU2 reference model. The northern Eurasia body wave data include observations of *S*, *SSS*, *SSSS*, and *SSSSS* between 60° and 90°. The cluster of points with ray parameters between .09 and .12 s/km are *S* waves that bottom in the lower mantle and do not exhibit any significant splitting. However, the multiply reflected *S* waves which turn in the upper mantle and transition zone, such as *SSS* from 65° to 75° and *SSSS* from 75° to 85°, are split with values of $\Delta\tau$ up to 12 s. For the body waves, the largest values of the ratio $\Delta\tau/t_{PSV}$ are the Russian platform observations of *SS* in the range 43° to 46°. The LR discrepancy for northern Eurasia, which has a mean value of $-1.8\% \pm .08\%$, is somewhat larger in magnitude than that for the Russian platform ($-1.1\% \pm .09\%$). There is an indication of systematic variations in $\Delta\tau_{LR}$ with frequency, but the present data set is inadequate to quantify the effect.

As shown in Figure 5, the apparent splitting times for the northern Eurasia and Russian platform paths vary systematically with horizontal slowness: phases most sensitive to velocity perturbations in the uppermost mantle display the greatest apparent splitting. For comparison, we also plot the apparent splitting times computed from two models: (1) Dziewonski and Anderson's [1981] radially anisotropic earth model PREM, modified to have an EU2 crust, and (2) an isotropic, finely layered stochastic model which we now describe.

INTERPRETATION

The body-wave measurements are very consistent with the shear-wave splitting computed from PREM. However, the surface-wave observations show a smaller LR

discrepancy than predicted by this radially anisotropic structure. It is certainly possible to derive a smooth, anisotropic (SA) structure that fits the data by relatively small perturbations to the five elastic moduli profiles that characterize PREM. We defer this exercise to a later, more detailed report. However, we find it interesting to note that most aspects of the data can be fit by a rough, isotropic (RI) model whose deviations from an SI structure are given by a single function of depth.

We consider a stochastic model of the upper mantle $\mathbf{m} = [\lambda, \mu, \rho]$ in which the variation of isotropic elastic constants and density with vector position $\mathbf{x} = [x_1, x_2, x_3]$ is a small perturbation to a horizontally layered structure $\mathbf{m}(x_3)$, where x_3 is depth. In any particular layer, we write this small perturbation as the product of a constant triplet $[\delta\lambda, \delta\mu, \delta\rho]$ and a homogeneous, scalar-valued, Gaussian random field $f(\mathbf{x})$ with zero mean and autocovariance function

$$C_{ff}(\mathbf{x}) = \langle f(\mathbf{x}+\mathbf{x}')\, f(\mathbf{x}') \rangle = e^{-\sqrt{k_1^2 x_1^2 + k_2^2 x_2^2 + k_3^2 x_3^2}} \qquad (2)$$

Such a random field is fractal with Hausdorff dimension 7/2; the two-dimensional version of (2) has proven useful in the study of seafloor morphology, where the parameters $\{k_i\}$ are the characteristic wavenumbers of the abyssal hills [Goff and Jordan, 1988].

We assume the characteristic vertical wavenumber is large compared to the two horizontal wavenumbers: $k_1 \sim k_2 << k_3$; i.e., the depth variations have much shorter scale lengths than the geographic variations. Then, for waves of length $l >> k_3^{-1}$, a theoretical argument due to Backus [1962] may be extended to show that the medium responds like a homogeneous, radially anisotropic solid. In the special case where the root-mean-square (rms) perturbations are described by a single, small fluctuation parameter $\varepsilon = \delta\lambda/\lambda = \delta\mu/\mu << 1$, Backus's averaging procedure yields to $O(\varepsilon^2)$,

$$v_{PV}^2 = \bar{v}_P^2(1-\varepsilon^2), \quad v_{PH}^2 = \bar{v}_P^2(1-\bar{\lambda}^2/\bar{\sigma}^2\varepsilon^2)$$
$$v_{SV}^2 = \bar{v}_S^2(1-\varepsilon^2), \quad v_{SH}^2 = \bar{v}_S^2 \tag{3}$$
$$\eta = 1 - (\bar{\sigma}/\bar{\lambda} - \bar{\lambda}/\bar{\sigma})\varepsilon^2$$

where $\sigma \equiv \lambda + 2\mu$.

These equations allow us to calculate an SA model [v_{PH}, v_{PV}, v_{SH}, v_{SV}, η] that is equivalent to the RI model [v_P, v_S, ε] in the long-wavelength limit. An example is given in Table 1. In deriving this structure, we took v_{PV} and v_{SV} from EU2 and computed v_{PH}, v_{SH} and η from the values of ε listed in the table. We found that the LR discrepancy could be satisfied by an uppermost mantle with $\varepsilon = .28$. This yields a long-wavelength *SH-SV* velocity difference of 4.2% for a ray angle of 90°(horizontal path), which accounts for most of the SNA-EU2 discrepancy (Figure 1). The *SH-SV* difference described by (3) is a strong function of ray angle, becoming negative for angles steeper than about 52°. Consequently, to explain the magnitude of the shear-wave splitting observed for rays with bottoming depths in the transition zone, it was necessary to maintain a relatively large value of ε to depths greater than 200 km. In our example structure, ε decreases linearly from .22 at 220 km to zero at 400 km. Given the scatter, the model is consistent with most of the observations, although it predicts a somewhat smaller amount of splitting than is observed for rays bottoming near the 670-km discontinuity for the northern Eurasian paths.

This modeling exercise establishes the magnitude of the velocity heterogeneity needed to explain the splitting data by an RI mechanism. It is large: a value of $\varepsilon = .28$ corresponds to a 14% rms fluctuation in the isotropic wave speeds. This exceeds, probably by a substantial amount, the variation in isotropically averaged parameters expected for even a diverse assemblage of upper-mantle ultrabasic rocks. We take this to be evidence for some sort of intrinsic (i.e., local) anisotropy, for example, the alignment of olivine crystals. In future modeling work, we intend to generalize the Gaussian

stochastic description represented by (2) to include the tensor properties of local anisotropy; i.e., we will consider rough, anisotropic (RA) structures.

In such a stochastic description, the qualitative difference between SA and RA structures is reduced to a quantitative difference in the characteristic wavenumbers $\{k_i\}$. The long-wavelength splitting data presented in this paper do not constrain these wavenumbers. However, the fact that the first arrival times from short-period *Sn* waves yield high apparent velocities is evidence that the characteristic vertical scale of the heterogeneity, k_3^{-1}, lies between the lengths of the long-period and short-period waves, say between 100 km and 10 km. In addition to giving apparent anisotropy at long wavelengths, characteristic scale lengths in this intermediate range provide high-velocity "micropaths" at short wavelengths [*cf.* Flatte, 1979], which could account for the *Sn* discrepancy. This hypothesis is consistent with the thinking of Fuchs and Schultz [1976], who have suggested that an RA structure is needed to account for the "shingling" of *P* arrivals observed on long refraction profiles in Europe.

In conclusion, it appears that stochastic descriptions of fine-scale heterogeneity can potentially explain the LR, *Sn*, and SNA-EU2 discrepancies, as well as other aspects of wave propagation in the continental upper mantle. Our observation that the shear-wave splitting is significantly smaller for paths along the Alpine-Himalayan front than for cratonic paths suggests that the large scale variations in the stochastic parameters may be related to continental deep structure.

ACKNOWLEDGEMENTS

V. Cormier provided the code to compute travel times in a radially anisotropic structure. We thank him and J. Goff for useful discussions. This reasearch was supported by DARPA and AFGL under contract F19628-87-K-0040. L. Gee was supported by an Air Force Graduate Fellowship.

TABLES

Table 1. RI and equivalent SA models for Eurasian paths.

Depth (km)	ε	v_{PH} (km/s)	v_{PV} (km/s)	v_{SH} (km/s)	v_{SV} (km/s)	η
40-100	.28	8.48	8.19	4.69	4.51	0.85
100-200	.22	8.36	8.19	4.64	4.53	0.90
220	.22	8.32	8.14	4.64	4.53	0.90
400	0	8.65	8.65	4.74	4.74	1.00

All values linearly interpolated between 220 and 400 km.

FIGURE CAPTIONS

FIGURE 1

Shear velocity as a function of depth for continental models EU2 and SNA. Stippling indicates the range of regional *Sn* velocities typical of continental cratons [Heustis et al., 1973]. The dashed line labeled EU2' is the *SH* velocity structure corresponding to the RI model in Table 1.

FIGURE.2

Azimuthal equidistant projection centered on KONO, illustrating the source-receiver geometry. Triangles are earthquake locations; octagons are receiver locations. Shields and stable platforms (shaded) are from Jordan [1981].

FIGURE 3.

Comparison of transverse and vertical component seismograms from northern Eurasian paths with EU2 synthetics. (*a*) Shear-wave splitting in *SSSS* at GRFO for the 01 Feb 84 Sea of Okhotsk event. Observed seismograms shifted by +2.5 s to align the *SSSS* pulse on the vertical component; $\Delta\tau_{SSSS}$ is 6 s. (*b*) *LR* discrepancy at KONO for the 25 Aug 83 Kyushu event. Observed seismograms shifted by +3.2 s to align the Rayleigh wave; $\Delta\tau_{LR}$ is 39 s for f_0 = 20 mhz.

FIGURE 4.

Example of shear-wave splitting in *SS* for a path crossing the Russian platform from the 01 Jul 84 Hindu Kush event. The traces for GRFO have been shifted by +2.5 s and the traces for KONO have been shifted by +1.5 s to align the *SS* pulses with the SNA synthetic on the transverse component. $\Delta\tau_{SS}$ is 5 s for KONO, but less than 1 s for GRFO.

FIGURE 5.

The ratio $\Delta\tau/t_{PSV}$ as a function of horizontal slowness for the data and models discussed in this study. Circles are measurements from the northern Eurasia corridor, and crosses from the Russian platform path. The zero line corresponds to the isotropic reference model EU2. Calculated values from PREM, modified to have an EU2 crust, are indicated by solid lines. Short dashed lines labeled EU2' are calculated from the RI model in Table 1.

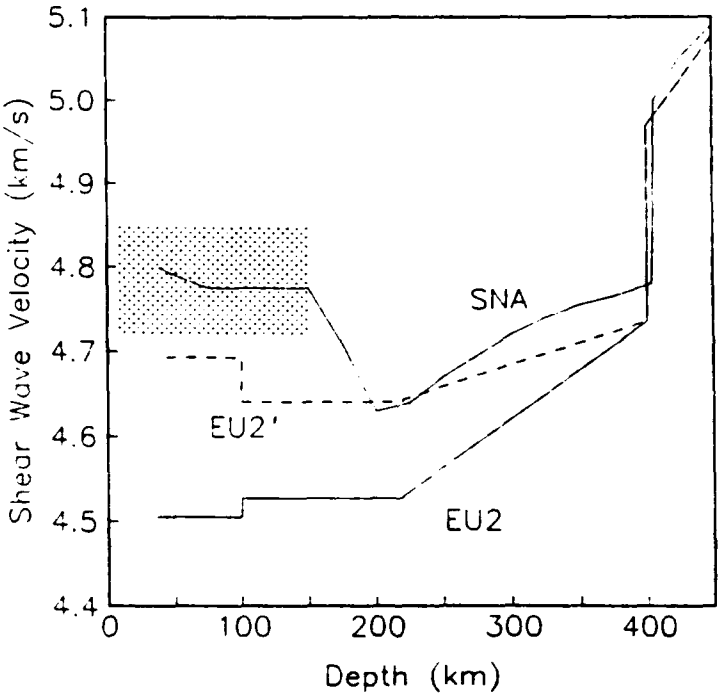

5.1
5.0
4.9
4.8
4.7
4.6
4.5
4.4
Shear Wave Velocity (km/s)
SNA
EU2'
EU2
0
100
200
300
400
Depth (km)

Figure 1

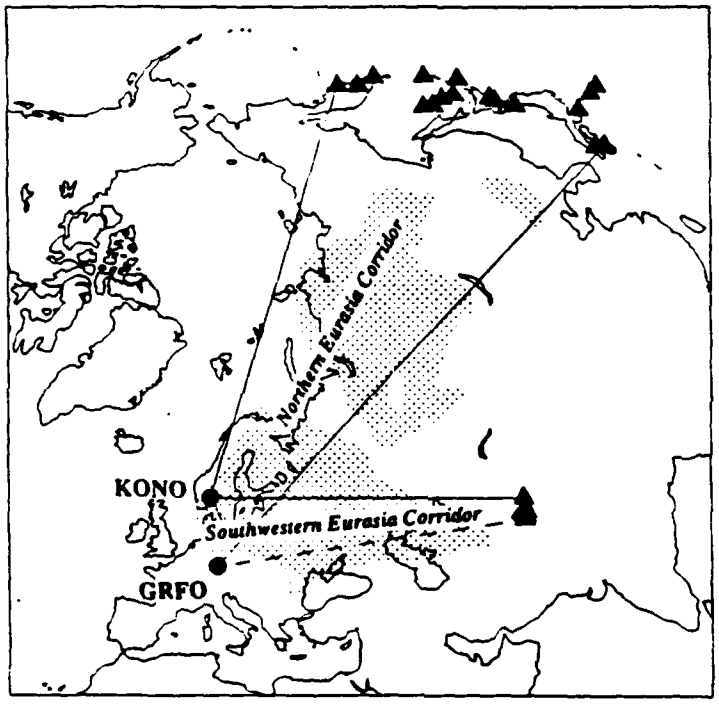

Northern Eurasia Corridor
KONO
Southwestern Eurasia Corridor
GRFO

Figure 2

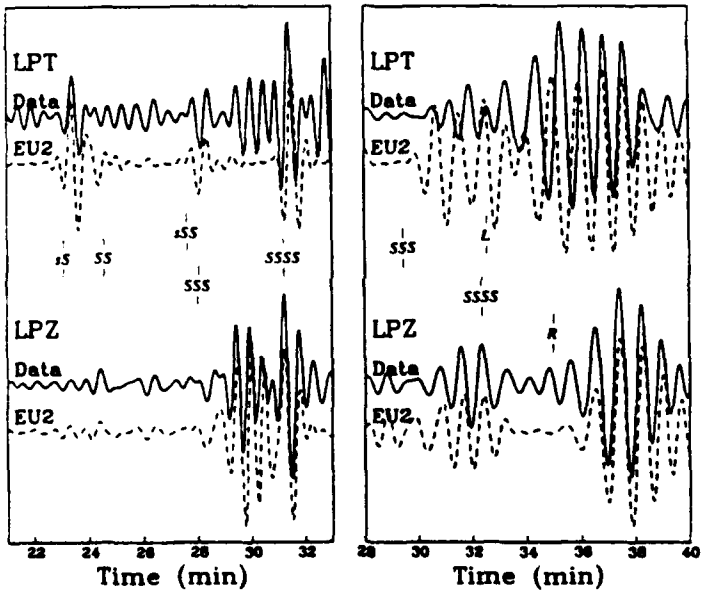

LPT
Data
EU2
sS
SS
sSS
SSS
SSSS
LPZ
Data
EU2
22
24
26
28
30
32
Time (min)
LPT
Data
EU2
SSS
L
SSSS
LPZ
R
Data
EU2
28
30
32
34
36
38
40
Time (min)

Figure 3

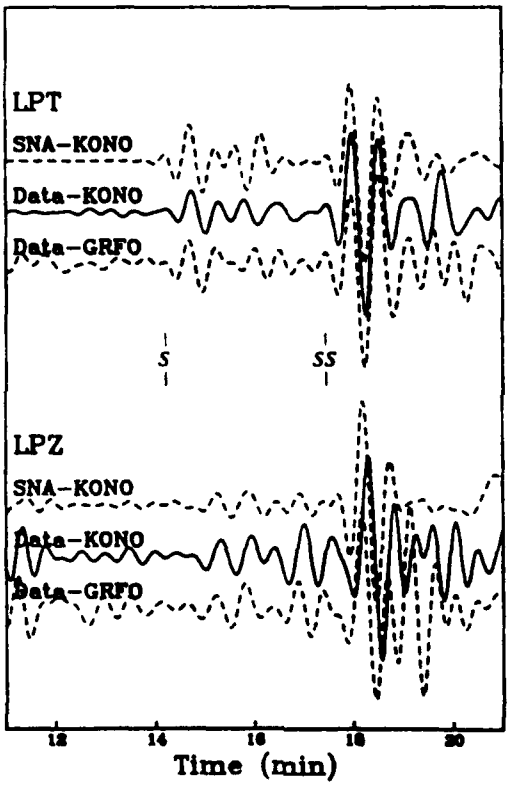

LPT
SNA-KONO
Data-KONO
Data-GRFO
S
SS
LPZ
SNA-KONO
Data-KONO
Data-GRFO
12
14
16
18
20
Time (min)

Figure 4

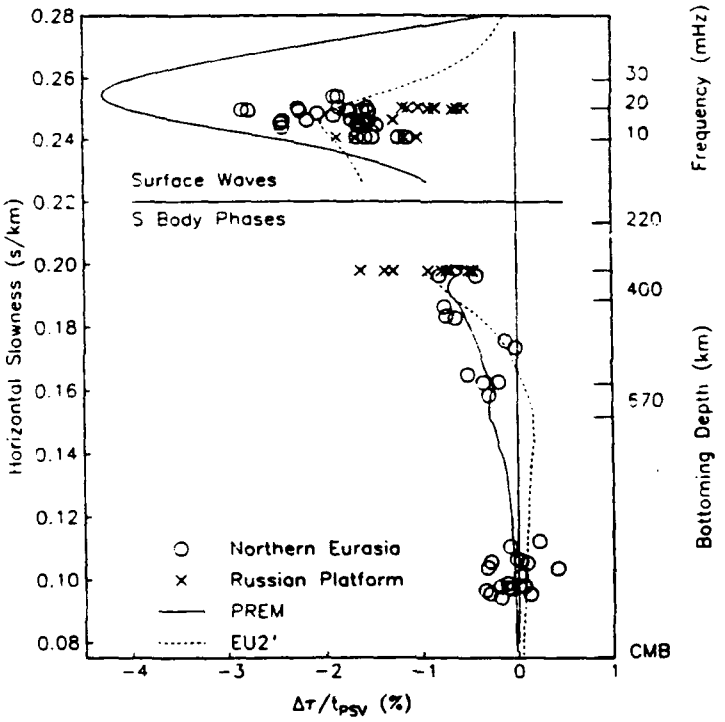

0.28
0.26
0.24
0.22
0.20
0.18
0.16
0.14
0.12
0.10
0.08
Horizontal Slowness (s/km)
Surface Waves
S Body Phases
Northern Eurasia
Russian Platform
PREM
EU2'
−4
−3
−2
−1
0
1
Δτ/t_PSV (%)
30
20
10
220
400
670
CMB
Frequency (mHz)
Bottoming Depth (km)

Figure 5

REFERENCES

Abramowitz, M. and I. Stegun, *Handbook of mathematical functions with formulas, graphs, and mathematical tables*, 10th ed., NBS Applied Mathematics Series 55, 1046p., 1972.

Aki, K., Study of earthquake mechanism by a method of phase equalization applied to Rayleigh and Love waves, *J. Geophys. Res.*, *65*, 729-750, 1960.

Aki, K., Study of Love and Rayleigh waves from earthquakes with fault plane solutions or with known faulting. Part 1. A phase difference method based on a new model of earthquake source, *Bull. Seis Soc. Am.*, *54*, 511-527, 1964.

Aki, K., Seismological evidences for the existence of soft thin layers in the upper mantle under Japan, *J. Geophys. Res.*, *73*, 585-594, 1968.

Aki, K., and K. Kaminuma, Phase velocity of Love waves in Japan, 1., Love waves from the Aleutian shock of March 9, 1957, *Bull. Earthquake Res. Inst.*, *41*, 243-259, 1963.

Aki, K., and P. Richards, *Quantitative Seismology*, vol 1., Freeman and Co., San Francisco, 557p., 1980.

Anderson, D. L., and A. M. Dziewonski, Upper mantle anisotropy: evidence from free oscillations, *Geophys. J. R. astr. Soc.*, *69*, 383-404, 1982.

Anderssen, R. S., The effect of discontinuities in density and shear velocity on the asymptotic overtone structure of torsional eigenfrequencies of the Earth, *Geophys. J. R. astr. Soc.*, *50*, 303-309, 1977.

Anderssen, R. S., and J. R. Cleary, Asymptotic structure in torsional free oscillations of the Earth - I. Overtone structure, *Geophys. J. R. astr. Soc.*, *39*, 241-268, 1974.

Anderssen, R. S., J. R. Cleary, and A. M. Dziewonski, Asymptotic structure in eigenfrequencies of spheroidal normal modes of the Earth, *Geophys. J. R. astr. Soc.*, *43*, 1001-1006, 1975.

Baag, C.-E. and C. A. Langston, Shear-coupled *PL*, *Geophys. J. R. astr. Soc.*, *80*, 363-385, 1985a.

Baag, C.-E. and C. A. Langston, A WKBJ spectral method for computation of *SV* synthetic seismograms in a cylindrically symmetric medium, *Geophys. J. R. astr. Soc.*, *80*, 387-417, 1985b.

Backus, G. E., Long-wave elastic anisotropy produced by horizontal layering, *J. Geophys. Res.*, *67*, 4427-4440, 1962.

Backus, G. E., and J. F. Gilbert, Numerical applications of a formalism for geophysical inverse problems, *Geophys. J. R. astr. Soc.*, *13*, 247-276, 1967.

Backus, G. E., and J. F. Gilbert, The resolving power of gross Earth data, *Geophys. J. R. astr. Soc.*, *16*, 169-205, 1968.

Båth, M., Propagation of *Sn* and *Pn* to teleseismic distances, *Pageoph*, *64*, 19-30, 1966.

Båth, M., and A. Lopez Arroyo, *Pa* and *Sa* waves and the upper mantle, *Geofis pura e appl.*, *56*, 67-92, 1963.

Ben Menahem, A., Mode-ray duality, *Bull. Seis Soc. Am.*, *54*, 1315-1321, 1964.

Bolt, B. A., M. Niazi, Somerville, M. R., Diffracted *ScS* and the shear velocity at the core boundary, *Geophys. J. R. astr. Soc.*, *19*, 299-305, 1970.

Bolt, B. A., and M. Niazi, *S* velocities in D" from diffracted *SH*-waves at the core boundary, *Geophys. J. R. astr. Soc.*, *79*, 825-834, 1984.

Boore, D. M., Effect of higher mode contamination on measured Love wave phase velocities, *J. Geophys. Res.*, *74*, 6612-6616, 1969.

Bracewell, R., *The Fourier transform and its applications*, 2nd ed., McGraw-Hill Company, New York, 444p., 1978.

Brune, J. N., Travel times, body waves, and normal modes of the Earth, *Bull. Seis Soc. Am.*, *54*, 2099-2128, 1964.

Brune, J. N., and J. Dorman, Seismic waves and Earth structure in the Canadian shield, *Bull. Seis Soc. Am*, *53*, 167-210, 1963.

Butler, R., A source of bias in multiple *ScS* differential travel times determined by waveform correlation, *Geophys. Res. Lett.*, *4*, 593-595, 1977.

Butler, R., Shear wave travel times from *SS*, *Bull. Seis Soc. Am.*, *69*, 1715-1732, 1979.

Caloi, P., Sur l'origine des ondes de type superficiel associees aux ondes *S*, *SS*, *SSS*, *Bur. Cent. Seism. Int., Serie A*, *17*, 235-241, 1948.

Caloi, P., Onde longitudinali e trasversali guidate dall'astenosfera, *Rend. Sc. fis. mat. e nat.*, *XV*, 352-357, 1953.

Capon, J., High-resolution frequency-wavenumber spectrum analysis, *Proc. IEEE*, *57*, 1408-1418, 1969.

Cara, M., Observations d'ondes *Sa* de type *SH*, *Pageoph*, *114*, 141-157, 1976.

Cara, M., Lateral variations of *S* velocity in the upper mantle from higher Rayleigh modes, *Geophys. J. R. astr. Soc.*, *57*, 649-670, 1979.

Cara, M., A. Nercessian, and G. Nolet, New inferences from higher mode data in western Europe and northern Eurasia, *Geophys. J. R. astr. Soc.*, *61*, 459-478, 1980.

Cara, M., and J. J. Lévêque, Waveform inversion using secondary observables, *Geophys. Res. Lett.*, *14*, 1046-1049, 1987.

Chandler, R., L. E. Alsop, and J. Oliver, On the synthesis of shear-coupled *PL*, *Bull. Seis Soc. Am.*, *58*, 1849-1877, 1968.

Chapman, C. H., A new method for computing seismograms, *Geophys. J. R. astr. Soc.*, *54*, 481-518, 1978.

Chapman, C. H., and J. A. Orcutt, The computation of body wave synthetic seismograms in laterally homogeneous media, *Rev. of Geophys.*, *23*, 105-163, 1985.

Cleary, J. R., The D" region, *Phys. Earth Planet. Inter.*, *19*, 13-27, 1974.

Cleary, J. R., The *S* velocity at the core-mantle boundary, from observations of diffracted *S*, *Bull. Seis Soc. Am.*, *59*, 1399-1405, 1969.

Crampin, S., and D. W. King, Evidence for anisotropy in the upper mantle beneath Eurasia from the polarization of higher mode seismic surface waves, *Geophys. J. R. astr. Soc.*, *49*, 59-85, 1977.

Creager, K. C., and T. H. Jordan, Aspherical structure of the core-mantle boundary from *PKP* travel times, *Geophys. Res. Lett.*, *13*, 1497-1500, 1986.

Dahlen, F. A., Surface waves and the ellipticity of the Earth, *J. Geophys. Res.*, *80*, 4895-4903, 1975.

Dahlen, F. A., A uniformly valid asymptotic representation of normal mode multiplet spectra on a laterally heterogeneous Earth, *Geophys. J. R. astr. Soc.*, *62*, 225-247, 1980.

Dey-Sakar, S. K., and C. H. Chapman, A simple method for the computation of body wave seismograms, *Bull. Seis Soc. Am.*, *68*, 1577-1593, 1978.

Doornbos, D. J., Present seismic evidence for a boundary layer at the base of the mantle, *J. Geophys. Res.*, *88*, 3498-3505, 1983.

Doornbos, D. J., and J. C. Mondt, Attenuation of *P* and *S* waves diffracted around the core, *Geophys. J. R. astr. Soc.*, *57*, 353-379, 1979a.

Doornbos, D. J., and J. C. Mondt, *P* and *S* waves diffracted around the core and the velocity structure and the base of the mantle, *Geophys. J. R. astr. Soc.*, *57*, 381-395, 1979b.

Dorman, J., M. Ewing, and J. Oliver, Study of shear-velocity distribution in the upper mantle by mantle Rayleigh waves, *Bull. Seis Soc. Am.*, *50*, 87-115, 1960.

Dziewonski, A. M., Mapping the lower mantle: Determination of lateral heterogeneity in *P*-velocity up to degree and order 6, *J. Geophys. Res.*, *89*, 5929-5952, 1984.

Dziewonski, A. M., S. Bloch, and M. Landisman, A technique for the analysis of transient seismic signals, *Bull. Seis Soc. Am.*, *59*, 427-444, 1969.

Dziewonski, A. M., and M. Landisman, Great circle Rayleigh and Love wave dispersion from 100 to 900 seconds, *Geophys. J. R. astr. Soc.*, *19*, 37-91, 1970.

Dziewonski, A. M., J. Mills, and S. Bloch, Residual dispersion measurement - a new method of surface-wave analysis, *Bull. Seis Soc. Am.*, *62*, 129-139, 1972.

Dziewonski, A. M., and J. F. Gilbert, Observations of normal modes for 84 recordings of the Alaskan earthquake of 1964 March 28. Part II: Further remarks based on new spheroidal overtone data, *Geophys. J. R. astr. Soc.*, *35*, 401-437, 1973.

Dziewonski, A. M., B. H. Hager, and R. J. O'Connell, Large-scale heterogeneities in the lower mantle, *J. Geophys. Res.*, *82*, 239-255, 1977.

Dziewonski, A. M., and D. L. Anderson, Preliminary reference Earth model, *Phys. Earth Planet. Inter.*, *25*, 297-356, 1981.

Dziewonski, A. M., T.-A. Chou, and J. H. Woodhouse, Determination of earthquake source parameters from waveform data for studies of global and regional seismicity, *J. Geophys. Res.*, *86*, 2825-2952, 1981.

Dziewonski, A. M., and J. M. Steim, Dispersion and attenuation of mantle waves through waveform inversion, *Geophys. J. R. astr. Soc.*, *70*, 503-527, 1982.

Dziewonski, A. M., J. E. Franzen, and J. H. Woodhouse, Centroid-moment tensor solutions for July-September, 1983, *Phys. Earth Planet. Inter.*, *34*, 1-8, 1984.

Flatté, S. M. (ed.), *Sound Transmission through a Fluctuating Ocean*, Cambridge, 299 pp., 1979.

Forsyth, D. W., The early structural evolution and anisotropy of the oceanic upper mantle, *Geophys. J. R. astr. Soc.*, *43*, 103-162, 1975.

Frazer, L. N., Synthesis of shear-coupled *PL*, *Ph.D. Thesis*, Princeton University, 54p., 1977.

Fuchs, K., and K. Schulz, Tunneling of low-frequency waves through the subcrustal lithosphere, *J. Geophys.*, *42*, 175-190, 1976.

Gee, L. S., and T. H. Jordan, Polarization anisotropy and fine-scale structure of the Eurasian upper mantle, *Geophys. Res. Lett.*, *15*, 824-827, 1988.

Giardini, D., X.-D. Li, and J. H. Woodhouse, Three-dimensional structure of the Earth from splitting in free-oscillation spectra, *Nature*, *325*, 405-411, 1987.

Gilbert, J. F., Excitation of normal modes of the Earth by earthquake sources, *Geophys. J. R. astr. Soc.*, *22*, 223-226, 1971.

Gilbert, J. F., Some asymptotic properties of the normal modes of the Earth, *Geophys. J. R. astr. Soc.*, *43*, 1007-1011, 1975.

Gilbert, J. F., The representation of seismic displacements in terms of travelling waves, *Geophys. J. R. astr. Soc.*, *44*, 275-280, 1976a.

Gilbert, J. F., Differential kernels for group velocity,*Geophys. J. R. astr. Soc.*, *44*, 649-660, 1976b.

Gilbert, J. F., An introduction to low-frequency seismology, in *Physics of the Earth's Interior: Proceedings of the International School of Physics "Enrico Fermi," Course LXXVII*, eds. A.M. Dziewonski and E. Boschi, Soc. Italiana de Fisisca, Bologna, 41-126, 1980.

Gilbert, J. F., and A. M. Dziewonski, An application of normal mode theory to the retrieval of structural parameters and source mechanisms from seismic spectra, *Phil. Trans. R. Soc. A.*, *278*, 187-269, 1975.

Goff, J. A., and T. H. Jordan, Stochastic modeling of seafloor morphology: Inversion of Seabeam data for second order statistics, *J. Geophys. Res.*, *93*, 13589-13609, 1988.

Grand, S. P., and D. V. Helmberger, Upper mantle shear structure of North America, *Geophys. J. R. astr. Soc.*, *76*, 399-438, 1984.

Grand, S. P., and D. V. Helmberger, Upper mantle shear structure beneath Asia from multi-bounce *S* waves, *Phys. Earth Planet. Int.*, *41*, 154-169, 1985.

Hales, A. L., and S. Bloch, Upper mantle structure: Are low velocity zones thin?, *Nature*, *221*, 930-933, 1969.

Hales, A. L., and J. L. Roberts, The travel times of *S* and *SKS*, *Bull. Seis Soc. Am.*, *60*, 461-489, 1970.

Hart, R. S., Shear velocity in the lower mantle from explosion data, *J. Geophys. Res.*, *80*, 4889-4894, 1975.

Hart, R. S., and R. Butler, Shear-wave travel times and amplitudes for two well-constrained earthquakes, *Bull. Seis Soc. Am.*, *68*, 973-985, 1978.

Helmberger, D. V., and G. R. Engen, Upper mantle shear structure, *J. Geophys. Res.*, *79*, 4017-4028, 1974.

Hermite, C., Sur un nouveau développement ensérie de fonctions, *Comp. Rend. Acad. Sci.*, *55*, 93-100 and 266-273, 1864.

Herrin, E., and T. Goforth, Phase-matched filters: application to the study of Rayleigh waves, *Bull. Seis Soc. Am.*, *67*, 1259-1275, 1977.

Herrin, E., and T. Goforth, Phase analysis of Rayleigh waves from the Shagan River test site in the USSR, *Bull. Seis Soc. Am.*, *76*, 1739-1754, 1986.

Heustis, S., P. Molnar, and J. Oliver, Regional *Sn* velocities and shear velocity in the upper mantle, *Bull. Seis. Soc. Am.*, *63*, 469-475, 1973.

Hille, E., A class of reciprocal functions, *Annals of Math.*, *27*, 427-464, 1926.

Jackson, D., *Fourier series and orthogonal polynomials*, Carus Mathematical Monographs, no. 6, Mathematical Association of America, Monash, WI, 5th ed., 234p, 1961.

James, D. E., Anomalous Love wave phase velocities, *J Geophys. Res.*, *76*, 2077-2083, 1971.

Jeffreys, H., The times of *P*, *S*, and *SKS* and the velocities of *P* and *S*, *Mon. Nat. R. astr. Soc., Geophys. Suppl.*, *4*, 498-533, 1939.

Jordan, T. H., Earth structure from seismological observations, in *Physics of the Earth's Interior: Proceedings of the International School of Physics "Enrico Fermi," Course LXXVII*, eds. A.M. Dziewonski and E. Boschi, Soc. Italiana de Fisisca, Bologna, 1-40, 1980.

Jordan, T. H., Global tectonic regionalization for seismological data analysis, *Bull. Seis. Soc. Am.*, *71*, 1131-1141, 1981.

Jordan, T. H., and L. N. Frazer, Crustal and upper mantle structure from *Sp* phases, *J. Geophys. Res.*, *80*, 1504-1518, 1975.

Jordan, T. H., and S. A. Sipkin, Estimation of the attenuation operator for multiple *ScS* waves, *Geophys. Res. Lett.*, *4*, 167-170, 1977.

Julian, B. R., D. Davies, and R. M. Sheppard, *PKJKP*, *Nature*, *235*, 317-318, 1972.

Kaminuma, K., The crust and upper mantle structure in Japan, 3, An anisotropic model of the structure in Japan, *Bull. Earthquake Res. Inst.*, *44*, 511-518, 1966.

Kanasewich, E. R., T. Alpaslan, and F. Hron, The importance of *S*-wave precursors in shear-wave studies, *Bull. Seis Soc. Am.*, *63*, 2167-2176, 1973.

Kennett, B. L. N., and J. H. Woodhouse, On high-frequency spheroidal modes and the structure of the upper mantle, *Geophys. J. R. astr. Soc.*, *55*, 333-350, 1978.

Kennett, B. L. N., and G. Nolet, The influence of upper mantle discontinuities on the toroidal free oscillations of the Earth, *Geophys. J. R. astr. Soc.*, *56*, 283-308, 1979.

Kirkwood, S. C., and S. Crampin, Surface wave propagation in an ocean basin with an anisotropic upper mantle: observations of polarization anomalies, *Geophys. J. R. astr. Soc.*, *64*, 487-497, 1981.

Kovach, R. L., and D. L. Anderson, Higher modes surface waves and their bearing on the structure of the Earth's mantle, *Bull. Seis. Soc. Am.*, *54*, 161-182, 1964.

Kuo, B.-Y., D. Forsyth, and M. Wyse[illegible]ion, Lateral heterogeneity and azimuthal anisotropy in the North Atlantic determined from *SS-S* differential travel times, *J. Geophys. Res.*, *92*, 6421-6436, 1987.

LaCoss, R. T., E. J. Kelly, and M.N. Toksöz, Estimation of seismic noise structure using arrays, *Geophysics*, *34*, 21-38, 1969.

Landisman, M., A. M. Dziewonski, and Y. Satô, Recent improvements in the analysis of surface wave observations, *Geophys. J. R. astr. Soc.*, *17*, 369-403, 1969.

Langston, C.A. and C.-E. Baag, The validity of ray theory approximations for the computation of teleseismic *SV* waves, *Bull. Seis Soc. Am.*, *75*, 1719-1727, 1986.

Lapwood, E. R., The effect of discontinuities in density and rigidity on torsional eigenfrequencies of the Earth, *Geophys. J. R. astr. Soc.*, *40*, 453-464, 1975.

Lay, T., and D. V. Helmberger, A lower mantle *S* wave triplication and the shear-velocity structure of D", *Geophys. J. R. astr. Soc.*, *75*, 799-837, 1983.

Lebedev, N. N., *Special functions and their applications*, ed. R. Silverman, Prentice-Hall, Englewood Cliffs, NJ, 308p., 1965

Lerner-Lam, A. L., and T. H. Jordan, Earth structure from fundamental and higher-mode waveform analysis, *Geophys. J. R. astr. Soc.*, *75*, 759-797, 1983.

Lerner-Lam, A. L., and T. H. Jordan, How thick are the continents?, *J. Geophys. Res.*, *92*, 14007-14026, 1987.

Lévêque, J. J., and M. Cara, Long-period Love wave overtone data in North America and the Pacific Ocean: new evidence for upper mantle anisotropy, *Phys. Earth Planet. Inter.*, *33*, 164-179, 1983.

Masters, G., and J. F. Gilbert, Structure of the inner core inferred from observations of its spheroidal shear modes, *Geophys. Res. Lett.*, *8*, 569-571, 1981.

Masters, G., and J. F. Gilbert, Attenuation in the Earth at low frequencies, *Phil. Trans. R. Soc. A.*, *308*, 479-522, 1983.

Masters, G., T. H. Jordan, P. G. Silver, and J. F. Gilbert, Aspherical earth structure from fundamental spheroidal-mode data, *Nature*, *298*, 609-613, 1982.

McEvilly, T. V., Central US crust-upper mantle structure from Love and Rayleigh wave phase velocity inversion, *Bull. Seis. Soc. Am.*, *54*, 1997-2015, 1964.

Mendiguren, J., Identification of free oscillation spectral peaks for 1970 July 31, Colombian deep shock using the excitation criterion, *Geophys. J. R. astr. Soc.*, *33*, 281-321, 1973.

Mellman, G. R., A method of body-wave waveform inversion for the determination of earth struct re, *Geophys. J. R. astr. Soc.*, *62*, 481-504, 1980.

Mitchell, B. J., On the inversion of Love and Rayleigh-wave dispersion and implications for earth structure and anisotropy, *Geophys. J. R. astr. Soc.*, *76*, 233-241, 1984.

Mitchell, B. J., and D. V. Helmberger, Shear velocities at the base of the mantle from observations of *S* and *ScS*, *J. Geophys. Res.*, *78*, 6009-6020, 1973.

Mondt, J. C., *SH* waves: theory and observations for epicentral distances greater than 90 degrees, *Phys. Earth Planet. Inter.*, *15*, 46-59, 1977.

Morelli, A., and A. M. Dziewonski, Topography of the core-mantle boundary and lateral heterogeneity of the inner core, *Nature*, *325*, 678-683, 1987.

Mula, A. H. G., Amplitudes of diffracted long-period *P* and *S* waves and the velocities and *Q* structure at the base of the mantle, *J. Geophys. Res.*, *86*, 4999-5011, 1981.

Nakanishi, I., and D. L. Anderson, Worldwide distribution of group velocity of mantle Rayleigh waves as determined by spherical harmonic inversion, *Bull. Seis Soc. Am.*, *72*, 1185-1194, 1982.

Nakanishi, I., and D. L. Anderson, Measurement of mantle wave velocities and inversion for lateral heterogeneity and anisotropy, I. Analysis of great circle phase velocities, *J. Geophys. Res.*, *88*, 10267-10283, 1983.

Nakanishi, I., and D. L. Anderson, Measurement of mantle wave velocities and inversion for lateral heterogeneity and anisotropy, II. Analysis by the single station method, *Geophys. J. R. astr. Soc..*, *78*, 573-618, 1984.

Nakanishi, K. K., F. Schwab, and E. G. Kausel, Interpretation of *Sa* on a continental structure, *Geophys. J. R. astr. Soc.*, *47*, 211-223, 1976.

Nataf, H.-C., I. Nakanishi, and D. L. Anderson, Anisotropy and shear-velocity heterogeneities in the upper mantle, *Geophys. Res. Lett.*, *11*, 109-112, 1984.

Nataf, H.-C., I. Nakanishi, and D. L. Anderson, Measurements of mantle wave velocities and inversion for lateral heterogeneities and anisotropy, 3, Inversion, *J. Geophys. Res.*, *91*, 7261-7307, 1986.

Nolet, G., Higher Rayleigh modes in western Europe, *Geophys. Res. Lett.*, *2*, 60-62, 1975.

Nolet, G., The upper mantle structure under western Europe inferred from the dispersion of Rayleigh modes, *J. Geophys.*, *43*, 265-285, 1977.

Nolet, G., and B. L. N. Kennett, Normal-mode representations of multiple-ray reflections in a spherical earth, *Geophys. J. R. astr. Soc.*, *53*, 219-226, 1978.

Nolet, G., J. van Trier, and R. Huisman, A formalism for nonlinear inversion of seismic surface waves, *Geophys. Res. Lett.*, *13*, 26-29, 1986.

Okal, E. A., A physical classification of the Earth's spheroidal modes, *J. Phys. Earth*, *26*, 75-103, 1978.

Okal, E. A., and D. L. Anderson, A study of inhomogeneities in the upper mantle by multiple *ScS* travel-time residuals, *Geophys. Res. Lett.*, *2*, 1975.

Okal, E. A., and R. J. Geller, Shear-wave velocity at the base of the mantle from profiles of diffracted *SH* waves, *Bull. Seis Soc. Am.*, *69*, 1039-1053, 1979.

Oldham, R. D., The constitution of the interior of the Earth, as revealed by earthquakes, *Quarterly Journal, Geological Society*, *62*, 465-475, 1906.

Oliver, J., On the long period character of shear waves, *Bull. Seis Soc. Am.*, *51*, 1-12, 1961.

Oliver, J., and M. Ewing, Normal modes of continental surface waves, *Bull. Seis Soc. Am.*, *48*, 33-49, 1958.

Oliver, J., and M. Major, Leaking modes and the *PL* phase, *Bull. Seis Soc. Am.*, *50*, 165-180, 1960.

Palamà, G., Sui polinomi di Legendre di Laguerre e di Hermite, *Reale Istituto Lombardo di Scienze e Lettere, Rendiconti, LXX*, 147-191, 1937.

Pekeris, C. L., Asymptotic theory of the free oscillations of the Earth, *Proceedings of the National Academy of Sciences*, *53*, 1254-1257, 1965.

Peterson, J., Preliminary observations of noise spectra at the SRO and ASRO stations, *U.S. Geological Survey Open-File Report 80-992*, 25p., 1980.

Peterson, J., H. M. Butler, L. G. Holcomb, and C. R. Hutt, The seismic research observatory, *Bull. Seis Soc. Am.*, *66*, 2049-2068, 1976.

Peterson, J., C. R. Hutt, and L. G. Holcomb, Test and calibration of the seismic research observatory, *U.S. Geological Survey Open-File Report 80-187*, 86p., 1980.

Peterson, J., and C. R. Hutt, Test and calibration of the Digital World-Wide Standardized Seismograph, *U.S. Geological Survey Open-File Report 82-1087*, 170p., 1982.

Pilant, W. L., and L. Knopoff, Observations of multiple seismic events, *Bull. Seis Soc. Am.*, *54*, 19-39, 1964.

Poupinet, G., and C. Wright, The generation and properties of shear-coupled *PL* waves, *Bull. Seis Soc. Am.*, *62*, 1699-1710, 1972.

Press, F., and M. Ewing, Waves with *Pn* and *Sn* at great distances, *Proceedings of the National Academy of Sciences*, *41*, 24-27, 1955.

Revenaugh, J. S., The nature of mantle layering from first-order reverberations, *Ph.D. thesis*, Massachusetts Institute of Technology, 268p., 1989.

Revenaugh, J. S., and T. H. Jordan, Observations of first-order mantle reverberations, *Bull. Seis Soc. Am.*, *77*, 1704-1717, 1987.

Revenaugh, J. S., and T. H. Jordan, A study of mantle layering beneath the western Pacific, *J. Geophys. Res.*, *94*, 5787-5813, 1989.

Rial, J. A., S. P. Grand, and D. V. Helmberger, A note on lateral variation in upper mantle shear-wave velocity across the Alpine front, *Geophys. J. R. Astr. Soc.*, *77*, 639-654, 1984.

Rietz, H. L., *Mathematical statistics*, Carus Mathematical Monographs, no. 3, Mathematical Association of America, La Salle, IL, 6th ed., 181p, 1971.

Ritzwoller, M., G. Masters, and J. F. Gilbert, Observations of anomalous splitting and their interpretation in terms of aspherical structure, *J. Geophys. Res.*, *91*, 10203-10228, 1986.

Ritzwoller, M., G. Masters, and J. F. Gilbert, Constraining aspherical structure with low-degree interaction coefficients: application to uncoupled multiplets, *J. Geophys. Res.*, *93*, 6369-6396, 1988.

Sailor, R. V., and A. M. Dziewonski, Measurements and interpretation of normal mode attenuation, *Geophys. J. R. astr. Soc.*, *53*, 559-581, 1978.

Sansone, G., *Orthogonal functions*, Pure and Applied Mathematics, vol. IX, Intersciences Publishers, New York, 411p, 1959.

Schlue, J. W., and L. Knopoff, Shear-wave polarization anisotropy in the Pacific basin, *Geophys. J. R. astr. Soc.*, *49*, 145-165, 1977.

Schwab, F., E. G. Kausel, and L. Knopoff, Interpretation of *Sa* for a shield structure, *Geophys. J. R. astr. Soc.*, *36*, 737-742, 1974.

Shaw, P. R., Waveform inversion of explosion seismology data, *Ph.D. thesis*, University of California at San Diego, 189p., 1983.

Sheehan, A. F., and S. C. Solomon, Evidence from *SS-S* travel times for long wavelength variations in mantle structure beneath the north Atlantic (abstract), *EOS Trans. AGU*, *70*, 1227, 1989.

Silver, P. G.., and W. W. Chan, Implications for continental structure and evolution from seismic anisotropy, *Nature*, *335*, 1988.

Sipkin, S. A., and T. H. Jordan, Lateral heterogeneity of the upper mantle determined from the travel times of multiple *ScS*, *J. Geophys. Res.*, *81*, 6307-6320, 1976.

Sipkin, S. A., and T. H. Jordan, Multiple *ScS* travel times in the western Pacific: Implications for mantle heterogeneity, *J. Geophys. Res.*, *85*, 853-861, 1980.

Stark, M. and D. Forsyth, The geoid, small-scale convection, and differential travel time anomalies of shear waves in the Central Indian Ocean, *J. Geophys. Res.*, *88*, 2273-2288, 1983.

Szegö, G., Beiträge zur theorie der Laguerreschen polynome. I.: Entwicklungssätze, *Math. Zeitschrift*, *25*, 87-115, 1926.

Szegö, G., *Orthogonal Polynomials*, American Mathematical Society Colloquim Publications, vol. 23, American Mathematical Society, Providence, 4th ed., 421p., 1959.

Tanimoto, T., Waveform inversion of mantle Love waves: the Born seismogram approximation, *Geophys. J. R. astr. Soc.*, *78*, 641-660, 1984.

Tanimoto, T., The 3-D shear wave structure in the mantle by overtone waveform inversions - II. Inversion of X-waves, R-waves, and G-waves, *Geophys. J. R. astr. Soc.*, *93*, 321-334, 1987.

Tarantola, A., A strategy for nonlinear elastic inversion of seismic reflection data, *Geophysics*, *51*, 1893-1903, 1986.

Thatcher, W., and J. N. Brune, Higher mode interference and observed anomalous apparent Love wave phase velocities, *J. Geophys. Res.*, *74*, 6603-6611, 1969.

Ward, S. N., Long-period reflected and converted upper-mantle phases, *Bull. Seis Soc. Am.*, *68*, 133-153, 1978.

Widmer, R., G. Masters, and F. Gilbert, The spherically symmetric earth: observational aspects and constraints on new models (abstract), *EOS Trans. AGU* , *69*, 1310, 1988.

Woodhouse, J. H., On Rayleigh's principle, *Geophys. J. R. astr. Soc.*, *46*, 11-22, 1976.

Woodhouse, J. H., Asymptotic results for elastodynamic propagator matrices in plane stratified and spherically stratified earth models, *Geophys. J. R. astr. Soc.*, *54*, 263-280, 1978.

Woodhouse, J. H., The calculation of eigenfrequencies and eigenfunctions of the free oscillations of the Earth and Sun, in *Seismological Algorithms*, ed. D. J. Doornbos, Academic Press, London, 321-370, 1988.

Woodhouse, J. H., and F. A. Dahlen, The effect of a general aspherical perturbation on the free oscillations of the Earth, *Geophys. J. R. astr. Soc.*, *53*, 335-354, 1978.

Woodhouse, J. H., and A. M. Dziewonski, Mapping the upper mantle: three-dimensional modeling of earth structure by inversion of seismic waveforms, *J. Geophys. Res.*, *89*, 5953-5986, 1984.

Woodhouse, J. H., and A. M. Dziewonski, Three dimensional mantle models based on mantle wave and long period body wave data (abstract), *EOS Trans. AGU*, *67*, 307, 1986.

Woodward, R. L., and G. Masters, Global upper mantle structure from long-period differential travel times, *J. Geophys. Res.*, submitted, 1989.

Young, C. J., and T. Lay, Evidence for a shear velocity discontinuity in the lower mantle beneath India and the Indian ocean, *Phys. Earth Planet. Inter.*, *49*, 37-53, 1987a.

Young, C. J., and T. Lay, The core-mantle boundary, *Ann. Rev. Earth Planet. Sci.*, *15*, 25-46, 1987b.

Yu, G. K., and B. J. Mitchell, Regionalized shear velocity models of the Pacific upper mantle from observed Love and Rayleigh wave dispersion, *Geophys. J. R. astr. Soc.*, *57*, 311-341, 1979.

Attachment 3

AFOSR-TR- 91 0982

THESIS BY:

ELENA PLANTE

THE UNIVERSITY OF ARIZONA

Subcontract No.# S-789-000-006

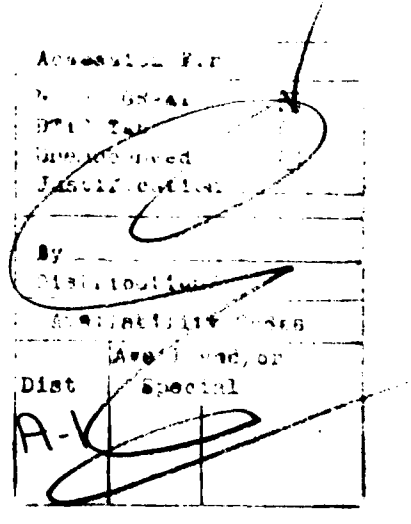
CEREBRAL CONFIGURATIONS AMONG THE PARENTS AND SIBLINGS OF LANGUAGE-DISORDERED BOYS

by

Elena Margaret Plante

A Dissertation Submitted to the Faculty of the
DEPARTMENT OF SPEECH AND HEARING SCIENCES
In Partial Fulfillment of the Requirements
For the Degree of
DOCTOR OF PHILOSOPHY
In the Graduate College
THE UNIVERSITY OF ARIZONA

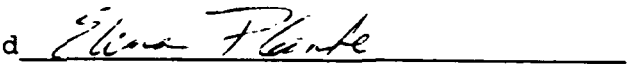
1 9 9 0



STATEMENT BY AUTHOR

This dissertation has been submitted in partial fulfillment of requirements for an advanced degree at the University of Arizona and it deposited in the University Library to be made available to borrowers under the rules of the Library.

Brief quotations from this dissertation are allowable without special permission, provided that accurate acknowledgment of source is made. Requests for permission for extended quotation from or reproduction of this manuscript in whole or in part may be granted by the copyright holder.

Signed 

Acknowledgements

I thank each of my doctoral committee members, Drs. Anna Binkewicz, Judith Lauter, Theodore Glattke, and Steven Rapscak for providing advise and professional opportunities. Dr. Joachim Seeger and Mrs. Rebecca Vance were instrumental in providing the support and resources necessary to establish the line of research that led to this dissertation. I thank Dr. Linda Swisher who created a setting in which I was challenged to learn and who provided the resources, guidance, and opportunities to do so.

Table of Contents

List of Illustrations

List of Tables

Abstract

Four families that include a specifically language-impaired (SLI) boy were studied to test the hypothesis that developmental language disorders are biologically transmittable and to further describe the neuroanatomical correlates of the disorder. A majority of the parents and siblings of the SLI boys also experienced communication difficulty (i.e., difficulty with speech, language, or academic skills). Evidence of communication difficulty was paired on an individual basis with neuroanatomical data obtained through quantitative analysis of magnetic resonance imaging (MRI) scans. Atypical perisylvian asymmetries were documented in a majority of the parents ($p < .05$) and were closely associated ($p = .84$) with a history of communication difficulty. These findings provide evidence that the disorder is biologically transmittable. In addition, language-disordered siblings of SLI boys also had atypical perisylvian asymmetries. This finding suggests that atypical perisylvian asymmetries reflect biological factors that place some families at risk for language impairment.

Measures of seven additional cerebral regions established that areas outside the perisylvian are often atypical in size. These measures demonstrate that neuroanatomical effects were bilateral and widespread. Thus, the neuroanatomical profile for developmental language disorder

differs from the profile typically associated with cases of acquired language disorder, which typically result from damage to the left perisylvian area in a premorbidly normal brain. In contrast, neuroanatomical correlates of developmental language disorder reflect a probable disturbance of prenatal brain development.

CHAPTER I

Introduction

Neuroanatomical background

Neuroanatomical studies provide information into the biological foundations of developmental language disorders. Two types of neuroanatomical studies are available: autopsy and imaging. Autopsy studies offer the best opportunity for detailed study of both gross anatomy and the underlying cellular architectonics. However, with autopsy studies, the behavioral features of a disorder are necessarily described retrospectively. Thus, adequate documentation of behavior is often unavailable. In contrast, noninvasive in-vivo imaging studies afford the opportunity to study the cerebral anatomy of subjects whose behavioral strengths and deficits can be documented. This advantage is offset by the relatively poor resolution of imaging studies. Current imaging capabilities are only capable of detecting lesions or developmental abnormalities that are sufficiently widespread that they alter gross anatomy. Therefore, while imaging studies can be used to document an apparent departure from the typical size or configuration of an anatomical region, autopsy examination is necessary to determine whether changes in the cellular architecture has produced the effect.

Neuroanatomical studies have established a strong relation between certain acquired language disorders in adults and lesions of the left perisylvian cortex (see Rubens, 1984 for a review). Preliminary evidence from neuroanatomical studies of subjects with a developmental disorder involving impaired language suggests a brain-language relation. Alterations in both the cellular and gross anatomy have been noted. Landau, Goldstein, and Kleffner (1960) documented a bilateral gross degeneration of the insulae and operculae along with polymicrogyri in the posterior sylvian region in a multiply handicapped boy whose major behavioral deficit was a severe language disorder. Subcortical degeneration of the cerebral peduncles and geniculate nuclei were also noted in this child.

Alterations of both cellular and gross anatomy were also noted in a study of four males with developmental dyslexia (Galaburda, Sherman, Rosen, Aboiwtz, & Geschwind, 1985). In at least three of the four subjects, the dyslexia was associated with speech or language difficulties. Autopsy revealed cortical ectopias, which were most numerous in the left perisylvian area in all four subjects. One subject also had polymicrogyri in the posterior sylvian region. The plana temporale, which typically is larger in the left than in the right hemisphere (Chi, Dooling, & Gilles, 1977a; Geschwind & Levitsky, 1968; Wada, Clarke, & Hamm, 1975; Witelson & Pallie, 1973), was symmetrical in all four

subjects.

Perisylvian symmetry has also been documented in an autopsy study of a girl who had a severe expressive language disorder along with attention deficit disorder (Cohen, Campbell, & Yaghmai, 1988). The atypical symmetry was accompanied by a single dysplastic region in the left frontal opercular region. A comparable finding occurred in an magnetic resonance imaging (MRI) study of four boys with specific language impairment (SLI), each of whom had an expressive impairment involving grammatical morphemes (Plante, Swisher, Vance, & Rapscak, 1990). The perisylvian areas of two of the boys were symmetrical, whereas a reversed (right > left) asymmetry was noted for a third.

In summary, both autopsy and imaging studies draw attention to a variety of perisylvian findings. Frank abnormalities at the cellular level occurred bilaterally in some subjects (Galaburda et al, 1985; Landau et al, 1960) and unilaterally in one (Cohen et al, 1988). When atypical asymmetries were documented, these resulted from a variety of left-right perisylvian configurations. In the subjects described as having dyslexia, symmetry resulted because the left planum temporale was of the expected size, while the right planum was atypically large (Galaburda et al, 1985). A similar finding was also noted for one SLI boy (Plante, Swisher, & Vance, 1989). In this child, the right perisylvian area was larger than average while the left was

of the expected size. Additional SLI boys with atypical perisylvian asymmetries had different perisylvian configurations. One boy had a right perisylvian area that was average size whereas the left was smaller than expected. Another had left and right perisylvian areas that were both smaller than average (Plante et al, 1990).

The nature of these neuroanatomical findings indicates the abnormalities occurred during the course of prenatal brain development (Galaburda et al, 1985; Plante et al, 1989). For example, polymicrogyri can be experimentally induced by lesioning the mammalian brain during the period of cell migration. The extent of gyral abnormality is related to the time of lesion, with the more extensive abnormalities resulting from an early time of lesion (Dvorak & Feit, 1977; Dvorak, Feit, & Jurankova, 1978). Cortical ectopias are also thought to occur during the period of cell migration as the result of misplaced migratory neurons (Caviness, Evrard, & Lyon, 1978; Sherman, Galaburda, & Geschwind, 1983). The presence of atypical cerebral asymmetries also suggests a prenatal effect. The normal left > right pattern of perisylvian asymmetries first appears during the third trimester (Chi et al, 1977a). After this time, the left > right pattern of plana asymmetry predominates in infants and children (Chi et al, 1977a; Wada, et al 1975; Witelson & Pallie, 1973) and adults (Geschwind & Levitsky, 1968; Wada et al, 1975; Witelson &

Pallie, 1973).

Several autopsy studies have addressed the issue of possible gender differences in the pattern of asymmetry. Most studies have not found gender differences for size of the right and left plana temporale or the degree of asymmetry of cerebral structures (Chi et al, 1977a, Koff, Naeser, Pieniadz, Foundas, & Levine, 1986; McShane, Risse, & Rubens, 1984; Wada et al, 1975). Two studies (Bear, Schiff, Saver, Greenberg, & Freeman, 1986; Wada et al, 1975) report that subgroups of subjects who have an atypical, right > left asymmetry tend to have more female than male members.

Atypical asymmetries have been documented for regions that extend beyond the perisylvian areas. Rosenberger and Hier (1980) used computerized tomography (CT) in a study of learning-disabled subjects. These subjects all had verbal intelligence test scores that were lower than their performance scores. Some subjects also had a history of "delayed speech." Approximately one half of the subjects with a history of "delayed speech" also had a reversed (right > left) asymmetry of the occipital poles. A third of the remaining learning-disabled subjects also had a reversed asymmetry pattern.

Using magnetic resonance imaging (MRI), Jernigan et al (1987) studied specifically language-impaired (SLI) children. In approximately half of the children, an atypical pattern of asymmetry (right > left) was noted for a

cerebral region bounded by the sylvian fissure anteriorly and the occipital poles posteriorly. Plante et al (1990) documented atypical asymmetries (left > right) of the cerebral hemispheres in three of four SLI boys. These studies suggest that the developmental effects producing atypical neuroanatomical patterns in language-disordered children may be relatively widespread. The anatomical areas that contribute to this atypical hemispheric asymmetry have not yet been identified.

A description of the relative effects on neuroanatomy would provide further insight into the mechanism(s) that produced the atypical neuroanatomical patterns. Studies of language-disordered children reveal neuroanatomical variability across subjects. Such inter-subject variability may reflect etiological differences. Alternatively, anatomical variation may be related to interactions between the causal agent and the stage of brain maturation in any given subject (Plante et al, 1989). The gyral configurations of the perisylvian area mature relatively late (Chi, Dooling, & Gilles, 1977b). A finding that late-developing structures are most often atypical in size would suggest the time of greatest neuroanatomical effect occurs during late gestation. If the probability that a structure will be atypical increases in parallel with the general maturational gradient for gyri, intersubject heterogeneity might actually reflect differences in the onset and severity

of the causal agent(s). Thus, differences in neuroanatomical effect would represent a range of one biological effect that occurs at different times during development.

Evidence for transmission of language disorders

The atypical neuroanatomical findings in language-disordered individuals may be the result of an agent that is transmitted through families. If this is true, the parents and siblings of language-disordered children should show the neuroanatomical effect as well. Although such a transmittable neuroanatomical effect has yet to be demonstrated in the families of language-disordered children, there is behavioral evidence that the disorder runs through families. Early evidence of familial tendencies toward developmental language disorders came from reports of selected families that include multiple cases of a speech or language disorder (Arnold, 1961; Samples & Lane, 1985).

This familial tendency has been substantiated by studies examining groups of language-disordered children and their families. Several studies have relied on questionnaires completed by a parent of language-disordered and control children, detailing the language and academic histories of all family members (Bishop & Edmundson, 1986; Neils & Aram,

1986; Tallal, Ross, & Curtiss, 1989a & b; Tomblin, 1989). These studies report that if a child is language disordered, more language or learning problems are reported for first-degree relatives than if a child is not language-disordered. Tomblin (1989), reported that 51% of the impaired probands had at least one language-impaired relative. A relative was considered language disordered only if he or she had received speech or language therapy. Tallal et al (1989a & b), using a criterion of self-report of language disorder, reading difficulties, or academic failure to indicate signs of impairment found that 77% of SLI probands had at least one impaired relative. Fathers reported some form of impairment more often than mothers (Neils & Aram, 1986; Tallal, et al, 1989a; Tomblin, 1989). However, there is some evidence to suggest that impaired mothers have more language-disordered children than do impaired fathers or control parents (Tallal et al, 1989a). Across studies, language-disordered probands have more impaired brothers than sisters (Neils & Aram, 1986; Tallal, et al, 1989a; Tomblin, 1989). This may be confounded by a skewed sex ratio favoring boys in families of language-impaired children (Tallal et al, 1989b).

The questionnaire method may include a reporting bias favoring families that already include a language-disordered child. These families are likely to be more sensitive to the signs of disordered language than other families. It is

possible that the results of the familial studies reflect inflated estimates in the families of language-disordered children, and underestimates the prevalence of language disorder in the families of controls. Standardized testing provides a more objective method of assessing the prevalence of disordered language than does self report. Standardized testing was used to determine differences in language skills for normal and reading-impaired children and their families (DeFries, Singer, Foch, & Lewitter, 1978). Reading-impaired children, as a group, performed poorly relative to controls on measures of language. Parents and siblings of reading-impaired children obtained lower test scores for reading, spelling, and language-based measures than did controls. Language skills showing significant differences between groups included grammatical closure, auditory memory, and verbal analogies. Thus, a familial tendency towards poor language skills has been documented using objective as well as subjective methods.

In any group of language-disordered children, not all will have a family history for the disorder. Byrne, Willerman, and Ashmore (1974) adapted a model originally used to describe distributions of mentally retarded children and suggested that there may be two groups of language-disordered children which can be identified by factors such as severity and family history. They found that children with moderately impaired language skills generally had a

lower socioeconomic level and more language-disordered relatives than children with severely impaired language skills. They suggest that the moderately impaired children were more likely to represent the extremes of the normal distribution for language skills. Furthermore, they suggested the children with severe language impairment may represent a biologically different population, in which the disorder is more likely to be secondary to acquired lesions or trauma.

The findings of this study should be interpreted with caution. The measure used to document language impairment in these children was the Illinois Test of Psycholinguistic Abilities (Kirk, McCarthy & Kirk, 1968). Of this battery's eight subtests, only two appear to reflect skills that are weak in language impaired children (Fundudis, Kolvin, & Garside, 1979). The remaining six subtests probably reflect general cognition. Therefore, the authors may have inadvertently been testing distributions of children with moderate and severe cognitive skill deficits rather than language skill deficits. A later study that hypothesized an association between the severity of a language disorder and family history failed to confirm such a relationship (Neils & Aram, 1986).

Implications of a transmittable effect

Factors such as trauma or stroke are unlikely

explanations for multiple cases of impairment within a family. A family tendency towards language or learning disorders suggests a genetic contribution to the disorder although the available data does not clearly conform to classic models of genetic transmission (Neils & Aram, 1986; Tallal et al, 1989a; Tomblin, 1989). Tomblin (1989) points out that rearing practices that restrict language development cannot be ruled out when multiple family members have poor language skills. However, rearing practices would not explain the atypical neuroanatomical findings in SLI children. An environmental factor, such as a toxin, can affect multiple family members. To date, there is no evidence to implicate any toxin as a common cause of developmental language disorders. Both parents and children would have to be exposed to the same toxic effects in utero to explain similar neuroanatomical findings across generations.

A fourth explanation proposes hormonal effects on development (Geschwind & Behan, 1982; Plante et al, 1989; Tallal et al, 1989b). A hormonal explanation is consistent with the constellation of behavioral and biological characteristics associated with developmental language disorders. A hormonal effect could explain transmission through families (cf. Perakis & Stylianopoulou, 1986) as well as the atypical sex ratio in families of language-impaired children (Tallal et al, 1989b). Subjects with

developmental disorders involving gonadal hormones tend to have poor language skills and learning difficulties (McCardle & Wilson, 1987; Perlman, 1973). Gonadal hormones are known to influence brain development in animals, and may explain the anatomical effects seen in language-impaired subjects (Galaburda et al, 1985; Plante et al, 1990).

If a transmittable factor produces language impairment and also alters brain development, the relatives of language-impaired probands should also have a high rate of atypical cerebral configurations. Preliminary evidence supports this hypothesis. Atypical brain configurations have been noted in the normally developing twin sister of a language-impaired boy (Plante et al, 1989). Atypical perisylvian asymmetries were also found in a male who reported having a brother who had received language therapy. This individual was one of a group of males who had agreed to have an MRI scan as part of a research program (Plante, unpublished data). A more systematic study of the siblings of language-disordered children is needed to verify these preliminary findings. Atypical neuroanatomical findings in parents of language-disordered children would strengthen the argument that a transmittable effect contributes to developmental language disorders.

The consistency of neuroanatomical configurations across subjects is also potentially important. A consistent pattern of findings across subjects who share the same

behavioral diagnosis would suggest a single effect is operating on the developing brain. In a previous study (Plante et al, 1990), atypical cerebral asymmetries were found across subjects, with some degree of individual variability noted. This variability may represent a range of one biological effect, or separate effects operating across subjects. When subjects consist of first-degree relatives (parents and their children), their biological backgrounds can be considered highly similar. Should such subjects show atypical cerebral configurations, inter-subject heterogeneity is more likely to reflect a range produced by one biological effect rather than separate effects across family members.

Statement of Purpose and Significance

This study will describe the neuroanatomical profiles in families containing a specifically language-impaired boy, in order to assess the type and frequency of atypical anatomical profiles. The presence of atypical cerebral configurations in parents of specifically language-impaired boys would support the theory that this type of language impairment is the result of a biological factor that is transmitted through families. The neuroanatomical profiles among first-degree relatives will help to establish a range that is likely to reflect a single biological agent. The

identification of neuroanatomical regions that are frequently atypical will provide insight into the nature of the presumed biological agent.

Research Questions

The study is designed to answer the following research questions:

1. Do atypical perisylvian asymmetries occur in parents of SLI boys?
2. To what extent are atypical perisylvian asymmetries associated with evidence suggesting a language disorder in the parents of SLI boys.
3. What is the range of cerebral effect as seen among first-degree relatives?
4. What neuroanatomical regions, in addition to the perisylvian areas, are atypical in individuals with a personal or family history of language impairment.

CHAPTER 2

Method

Subjects

Members of four families that include at least one boy with SLI (the proband) participated in this study. A brief description of each family follows.

Family 1:

A 46-year-old father and 39-year-old mother, along with three male children (ages 12;10, 7;0, and 5;4) are the members of Family 1. All family members consider themselves to be right handed, although the parents report left handedness among their blood relatives. All three sons were diagnosed as language impaired and were enrolled in therapy programs at the time of study. The mother reported the oldest son was born two months prematurely and spent his first month in an intensive care unit. He was identified as language impaired when he entered first grade. He has a history of poor articulation skills and difficulty with expressive language. At the time of study, he was experiencing difficulty with reading and writing as well as with expressive language skills. He was taking Ritalin at the time of study to manage hyperactivity. The second son's birth was described as unremarkable but he required surgery to correct an imperforate anus. He has undergone additional surgery on the scrotum and testicles and has had tests to

evaluate bladder control. He was enrolled in therapy for impaired language and poor articulation at age three years. These skills were still impaired at the time of study. The youngest son, the proband, was delivered three weeks prematurely and required oxygen after birth. He was enrolled in language therapy at age three years. His articulation skills were age appropriate. His expressive language skills were significantly below the level expected of a child his age.

Both parents report that they had language/learning problems as children similar to what they have observed in their sons. The father repeated first grade and entered the military service after grade nine. He later received a high school diploma and was employed as a materials handler at the time of study. The mother, a housewife, has a high school diploma and completed two years of post-secondary education. The father has four additional sons from two previous marriages. Two of these sons have developmental disabilities. Both parents have nephews who have been diagnosed as learning disabled. Both parents report twins (fraternal and identical) among their blood relatives.

Family 2:

Family 2 includes a 37-year-old father and a 31-year-old mother along with their two sons (ages 6;10 and 5;0) and a daughter (age 2;0). The father is left handed; all other family members are right handed. The mother reported each

of her pregnancies were unremarkable. The first-born son appeared to be developing normally at the time of study. The daughter was too young to participate in this study, but also appeared to be developing language normally. The second-born son, the proband, was diagnosed as SLI and was enrolled in a therapy program at the time of this study. He was being treated for a severe articulation disorder and impaired language. His family first became concerned about his speech and language development when he was approximately two years of age.

Neither parent reported childhood difficulties with language or academic skills. The father completed a college degree and was employed as an engineer. He reported that he had "stuttered" until age five years. He also reported his sister had stuttered as well. The mother completed a college degree and was a housewife at the time of study. No developmental disorders were reported among the mother's relatives. The mother reported fraternal twins among her relatives.

Family 3:

The third family includes a father (age 57), a mother (age 39), two daughters (ages 11;5 and 8;6), and one son (age 9;6) who served as the proband for this family. The mother reported her pregnancies were unremarkable. General health of all children is reported to be unremarkable. The oldest daughter appeared to be developing normally. The

proband has a history of speech and language impairment. His mother first noticed a problem with language development when he was approximately age two years. He started receiving therapy for poor articulation and impaired language at age three and was enrolled in a school serving "dyslexic" students at the time of study. The youngest daughter was receiving tutoring for reading in school and was being evaluated for language impairment at the time of study. She declined participation in this study because of apprehension about being inside the MRI scanner.

The father reported he had difficulties with speech and language development similar to those he observes in his son. The father reported his own speech was hard to understand until age six and he subsequently had difficulty learning to read in school. He reported that he does not enjoy reading. He completed nine years of schooling and was employed as a bus driver. He has a brother who also had difficulty with speech and language development and has twin sons from an earlier marriage. The mother completed nine years of school and managed a small business at the time of study. She was born prematurely by approximately 6 weeks. She reported that she had a "lisp" as a young child. She reported no problems with language or academic skills.

Family 4:

The fourth family includes a father (age 46), mother (age 42), a son (the proband, age 8;2) and daughter (age

6;2). All family members consider themselves to be right handed. The mother reported that pregnancy and birth for each child was unremarkable. Both children had a history of middle-ear infections, but general health has otherwise been unremarkable. The proband has a history of delayed speech onset (first words appeared at approximately 19 months) and was diagnosed as SLI. He was enrolled in a school serving "dyslexic" children at the time of study. According to standardized testing, the daughter had significantly delayed articulation skills for her age. Although she performed well on most language tests, she was verbally reticent. An analysis of spontaneous language revealed mildly impaired expressive language skills and word-finding problems. She was not receiving speech or language services at the time of study. The parents had experienced one miscarriage prior to the birth of their first child.

The mother considered herself dyslexic and reported she had used "baby talk" as late as kindergarten. She recalled having some difficulty learning in grade school and reported that reading takes her a "long time" and is "tiring." She also reported that she had a unilateral hearing loss as a child and attended a school for deaf and blind children for first grade. She reported that she "out grew" the hearing loss. She passed a pure-tone screening test at the time of study. The mother has a high school diploma and completed a year of post-secondary education. She was a housewife at

the time of study. The father reported he had difficulty learning to read and reports he does not enjoy reading. He holds a masters degree and was teaching at the college level at the time of study. The father reported that his brother had difficulty with speech and language and considers him to be dyslexic. The mother has a nephew who is receiving educational services for dyslexia.

Subject recruitment and selection

Subjects were recruited by soliciting referrals of boys with SLI from agencies and professionals serving such children in the Tucson area. The term SLI was defined to the referral sources as a language impairment in the absence of deficits in cognitive, sensory, motor, and social-emotional functioning.

After referral to the study, a diagnosis of SLI was confirmed through standardized testing and clinical judgement. A battery of standardized, norm-referenced tests was administered to each proband. The results of standardized testing are reported in Table 1. The test battery is described in Appendix A. The standardized tests chosen were those that reflected the skill areas of typical strengths (nonverbal, semantics) and weakness (morpho-syntactic, speech articulation) that characterize SLI in children. The manuals provided sufficient information so that a z-score could be calculated. They were relatively strong psychometrically compared with other available

instruments (McCauley & Swisher, 1984), or had proven utility in identifying language-impaired children All tests were administered by a certified speech-language pathologist or a graduate student under the supervision of a speech-language pathologist. Test score reliability was calculated on a point-to-point basis for all scorable test items. Median point-to-point reliability was 97 percent with a range between 75 and 100 percent agreement.

For children 7;11 years and younger, a spontaneous language sample was obtained for analysis. A twenty-minute play session with the child and an examiner was used to gather the language sample. Children were given a choice of activities and toys to play with during this time. The play session was video taped and a corpus of the child's spontaneous language was transcribed from video tape. Four minutes of each child's language sample was selected at random to assess transcription reliability. Point-to-point reliability for transcribed words ranged between 98 and 99 percent for the samples obtained. A corpus of sentences were analyzed using Developmental Sentence Scoring (DSS)(Lee, 1974). For all but one child (sibling 1 in Family 4), the sentences included in the analysis were 50 consecutively occurring sentences that were spoken after the first five minutes of the play session had elapsed. The remaining child had so few sentences that utterances from the first five minutes had to be included to obtain a sample

of 47 sentences. Point-to-point reliability for scorable items on the DSS ranged between 71 and 96 percent. Disagreements in scoring were resolved through joint review of the items in question.

To be selected for study, a proband had to demonstrate a significant impairment (lower 5 percent of the normative sample) in the comprehension or use of the morpho-syntactic components of language in the presence of normal nonverbal skills (upper 93 percent of the normative sample). In addition, to be selected for study, each proband had to pass a hearing screening at the time of testing. A passing performance consisted of reliable responses to 20 dB HL pure tones for at least three of the following frequencies: 500, 1000, 2000, and 4000 Hz. No selection criteria were set for performance on tests of vocabulary or articulation, or for the language sample analysis.

Two speech-language pathologists not otherwise connected with this study were asked to review compiled records for each potential proband. Records included past evaluations and therapy reports; tests of morpho-syntactic skills administered as part of this study were excluded. Each proband was judged language impaired by both speech-language pathologists.

When a potential proband was identified, the family was then invited to participate in this study. Families selected met three additional characteristics: children were

Table 1.
Test profiles for probands and siblings.

	Family 1			Family 2		Family 3		Family 4	
	sib 1	sib 2	proband	sib 1	proband	sib 1	proband	proband	sib
gender:	male	male	male	male	male	female	male	male	female
age:	12;10	7;0	5;4	6;10	5;0	11;5	9;6	8;2	6;2
Tests:									
Nonverbal:									
K-ABC	-0.93	-0.15	-0.47	0.00	0.67	-1.00	-0.07	-0.13	-0.33
Vineland									
Daily living	-1.93	0.66	0.07	1.07	0.33	1.87	0.66	0.27	0.27
Socialization	-1.33	-0.80	0.33	0.04	0.13	0.67	-0.53	-0.33	-0.27
Motor	*	*	-0.80	*	0.87	*	*	*	*
Morpho-syntactic:									
TACL-R									
Word classes	*	-0.25	-1.28	0.39	-0.92	*	0.47	-1.48	-0.41
Grammatical Morphemes	*	-0.71	0.74	-0.50	-2.33	*	-1.28	-1.18	0.92
Elaborated Sentences	*	-0.20	0.02	1.75	-1.23	*	0.05	-0.39	0.47
TOKEN	0.00	-0.60	NC	-0.26	-1.05	-0.99	-1.94	-1.63	0.07
ITPA									
Grammatical Closure	*	-0.83	-0.67	-0.38	-2.71	*	-0.79	-0.79	2.73

Table continues

	Family 1			Family 2		Family 3		Family 4	
gender:	sib 1 male	sib 2 male	proband male	sib 1 male	proband male	sib 1 female	proband male	proband male	sib female
age:	12;10	7;0	5;4	6;10	5;0	11;5	9;6	8;2	6;2
NSST-E	*	-6.46	-2.78	-0.38	-3.33	*	*	*	-0.24
CELF-R									
Oral Direct.	-0.66	*	*	*	*	-2.05	-2.33	-1.64	*
Form. Sent.	*	*	*	*	*	-0.67	-0.68	-1.67	*
Recall. Sent.	-2.33	*	*	*	*	0.33	-0.68	0.00	*
Sent. Assembly	-1.33	*	*	*	*	-0.00	-1.00	-1.34	*
Semantics:									
PPVT-R	-0.33	0.22	0.00	-0.27	-2.07	0.27	0.00	0.54	0.67
OWEPVT	*	0.27	-0.47	-0.60	-0.46	0.67	0.33	0.40	0.40
CELF-R									
Word Classes	*	*	*	*	*	-0.33	-2.05	-0.67	*
Semantic Rel.	-1.00	*	*	*	*	-0.33	-0.68	-1.34	*
Articulation:									
Templin-Darley	*	0.48	-1.07	1.81	-1.98	*	*	0.52	-1.64
Spontaneous language:									
DSS	*	*	-1.00	-1.57	-1.82	*	*	*	-1.33

* indicates that test was not given because it is not appropriate for the child's age.
NC indicates the test was discontinued because of unreliable responses.

monolingual English-speakers, both biological parents were willing to participate, and family members had no known neurological or developmental conditions that are known to alter brain morphology from normal (e.g. stroke, trauma, seizures). Nineteen potential probands were referred to the study. Four were rejected because their language skills were not significantly impaired. Three had other handicapping conditions (e.g. attention deficit, hearing loss) in addition to impaired language. Two had a history of seizures and one was excluded because his mother had a history of seizures. Two qualified as SLI but were excluded from the study because the proband or a parent was unable to complete the MRI scan due to claustrophobic reactions. Three did not have both biological parents available for study. The subjects of the present study are the first four consecutive families that were not excluded on the bases described above.

When a family was selected, a case history was taken for each family member. Case histories covered medical, birth, developmental, educational, and family background in addition to speech and language history (see Appendix B). Information concerning the personal and family history for the mothers and the fathers in each family are reported in Table 2.

Behavioral documentation for siblings

The battery of standardized tests described in Appendix

Table 2.
Family history characteristics as reported by the parents of language-disordered boys.

	Family 1		Family 2		Family 3		Family 4	
	maternal	paternal	maternal	paternal	maternal	paternal	maternal	paternal
Parental history:								
Academic failure	no	grade 1	no	no	no	no	no	no
Reading difficulty	yes	yes	no	no	no	yes	yes	yes
Speech or language difficulties	no	no	no	"stuttered"	"lisp"	speech & language	"baby talk"	no
Schooling completed	highschool + 2yrs	grade 9 +G.E.D.	college (bachelor's)	college (bachelor's)	grade 9	grade 9	highschool + 1 yr	college (master's)
Handedness	right	right	right	left	left	right	right	right
Place of birth	New Ulm MN	Toledo OH	Salt Lake City UT	Columbus OH	Brisbane Austrailia	Sedalia NC	Pittsburg PA	Canton IL
Birth order	3rd	3rd	3rd	2nd	1st	4th	2nd	1st
Mother's age at Subject's birth	26	31	31	16	27	30	31	22
Maternal miscarriages/ stillbirths	2	none	none	none	none	none	none	none

Table continues

	Family 1		Family 2		Family 3		Family 4	
	maternal	paternal	maternal	paternal	maternal	paternal	maternal	paternal
Risk factors during pregnancy	"drug for miscarriage" father smoked	father smoked	father smoked	maternal stress	father smoked	father smoked	father smoked	father smoked
Risk factors at birth	caesarian	none	none	none	premature (6 weeks)	none	breach	forceps
Risk factors postnatally	none	none	none	none	otitis	none	none	none
Family history:								
Left handedness	yes	no	no	no	yes	yes	no	no
Developmental disability	speech/ language hyperactivity stuttering	speech/ language learning disability hyperactivity	none	stuttering	none	speech/ language	dyslexia	dyslexia
Twinning	yes	yes	yes	no	yes	yes	no	no

A was used to describe verbal and nonverbal skills in the siblings of each proband. The results of standardized testing for these subjects are provided in Table 1.

Procedures

Subjects were scanned in the axial plane through the full volume of the cerebral cortex. The slice angle was standardized for each subject by aligning the slice angle parallel to the frontal and occipital poles from a sagittal prescan view. Scans were gathered on a Toshiba 0.5 Tesla magnet. A spin-echo scan sequence (TR 2800, TE 90) was used that produces good distinctions between grey and white matter.

Scans from the subjects were compared with a group of comparison scans from volunteers who were without a history of language impairment. These scans (11 from males, 6 from females) were selected from a data bank (Plante, unpublished data) of volunteers for whom case history information was available. Selected scans were all those from males and females who, according to self-report, lack a personal or family history of developmental disorders, twinning, or neurological conditions that are known to alter brain morphology from normal. The characteristics of the individuals from whom these scans were obtained are described in Table 3. Scans included an axial view set at a similar angle as the scans from subjects in this study.

Table 3.
Characteristics of comparison group.

Personal background:

Gender 11 males, 6 females

Ages: 20 to 47; median = 29

Handedness: right handed in all cases

Birth history:

Mother's age at birth: 20 - 38 years; median = 24

Birth order: 1 - 6; median = 2

Maternal miscarriages: 3 for male subjects
1 for female subjects

Pregnancy risk factors:
father smoked in 3 case for females
mother smoked in 1 case for females
males were not asked this question

Birth risk factors: 1 male was premature by 8 weeks

Family history:

Left handedness: 2 males
1 female

Developmental disorders: none reported
(this was a selection criterion)

Twinning: none reported
(this was a selection criterion)

Education:

years of education: 14-21 years (college)
median = 16 years (bachelor's degree)

special school services:
1 male (gifted & talented program)
1 female (articulation therapy for "r" only)

grade school failure: none reported

reading difficulty: none reported

Scans were clinically evaluated by a neuroradiologist prior to quantitative analysis. No parenchymal lesions were reported for any subject. One comparison scan from a female control was noted to have two punctate regions of hyperintensity in the centrum semiovale on the T2-weighted, axial image. These were not visible on a T1-weighted coronal scan. The scans were dummy coded by a research assistant so that those measuring the scans were unaware of the subject identity or language status. Scans were measured by five assistants who were not familiar with the subjects or hypotheses under study and me.

MRI films from these scans were computer digitized and measured using JAVA software (Jandel Scientific, 1988). The instrumentation used for quantitative analysis is provided in Appendix C. A calibration figure of 300 mm^2 was used to establish the degree of measurement error attributable to the resolution of the computer system and human error under optimal conditions. The mean value obtained for this figure was 300.62 mm^2 with a standard error of 0.16 mm^2. Measurement error under optimal conditions was approximately 0.6 percent.

A series of regional brain volumes was obtained for each subject and control scan. Measures were developed based upon both neuroanatomical divisions visible on the MRI scans and consideration of the time of development for cortical regions (cf. Chi et al, 1977b) The procedures for

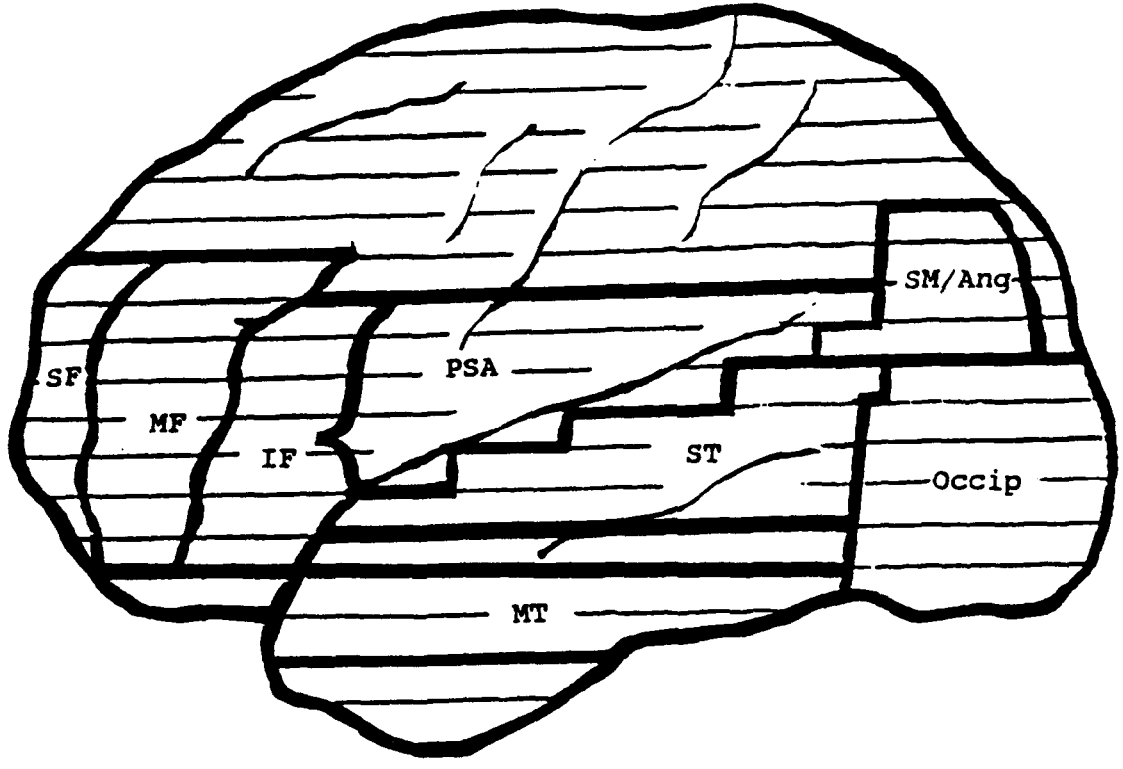
measuring each region are given in Appendix D. These regions are illustrated in Figures 1 and 2. All regional measures are expressed as a proportion of total brain size to stabilize the values for differences in head sizes across subjects.

A quotient reflecting the degree of asymmetry or symmetry was calculated for each pair of homologous cerebral areas measured. This quotient corresponds to the right regional volume divided by the left. Asymmetries (right > left or left > right) and symmetry (left = right) are classified with reference to the intertester measurement error for each measure. Perfect symmetry results in a quotient of 1.00. When measurement error is taken into account, asymmetry can be defined as values that exceed values of 1.00 +/-1.64 SE for quotient values.

Measurement reliability was assessed using a Pearson product-moment correlation for each anatomical measure completed. Reliability was assessed for every scan completed for a family member and for at least 20 percent of all comparison scans (selected at random). Acceptable reliability was defined as an r-squared value of .70 (r = .84) and above for a series of measures. The r values for each neuroanatomical region are provided in Table 4.

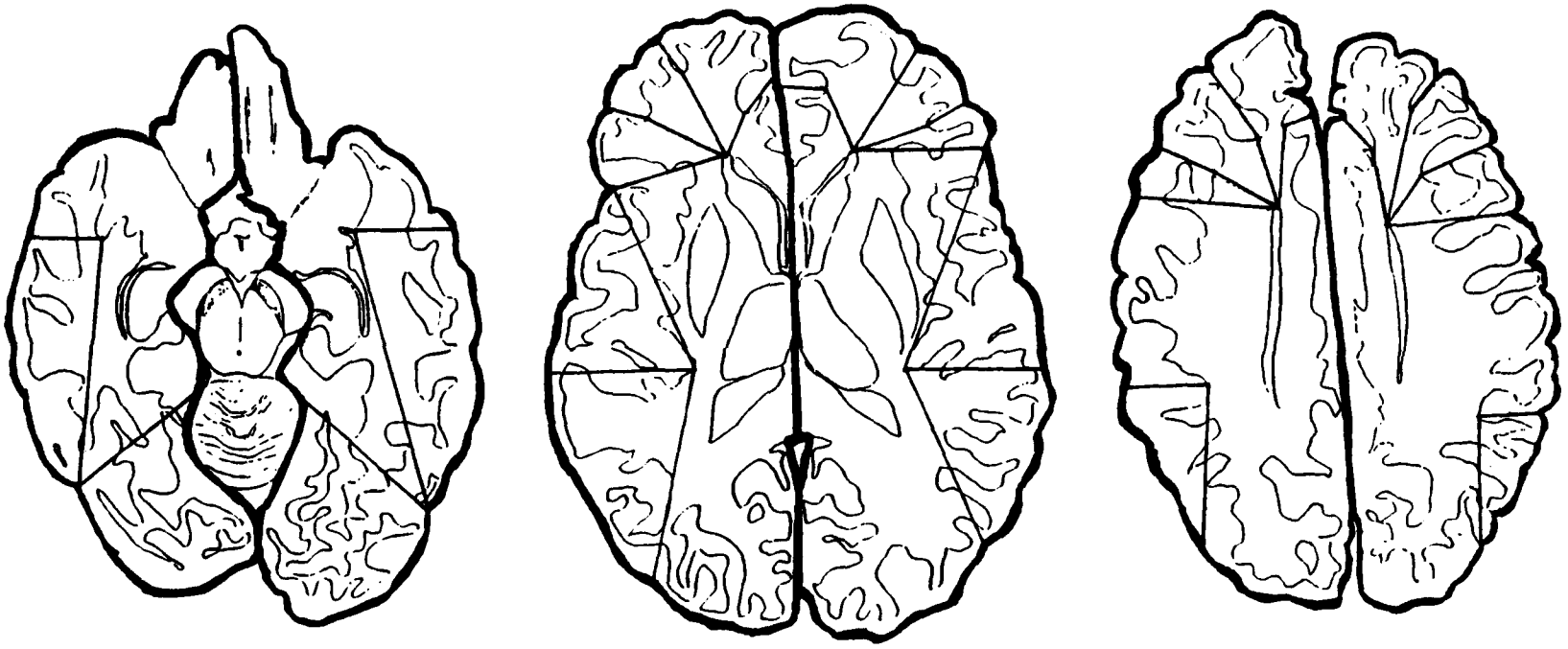
A pilot study of three SLI boys who served as subjects in a previous study (Plante et al, 1990) indicated that certain of these brain regions were most likely to be

Figure 1. Sagittal reconstruction of slices comprising area measures



SF = Superior frontal area; MF = Middle frontal area; IF = Inferior frontal region; PSA = Perisylvian area; ST = Superior temporal area; MT = Middle temporal area; SM/Ang = Supramarginal/angular area; Occip = Occipital area

Figure 2. Axial views of areas measured at three levels in the brain



atypical (see Appendix E). These include an asymmetry of the perisylvian area as well as proportional volumes of the superior, middle, and inferior frontal areas, superior and middle temporal areas, occipital areas, supramarginal/angular areas, and perisylvian areas.

Chapter III

Results

Comparison scans

Prior to comparisons with probands and their family members, the anatomical measures for males and females were tested for gender differences. No significant ($p < .10$, two-tailed test) differences for scans from male and female members of the comparison group were found for either proportional volumes or for degree of asymmetry. Consequently, all comparison scans were grouped together for comparison with probands and their family members. The proportional volumes for each regional measure in the comparison group are reported in Table 4. These values will serve as the reference for evaluating proportional volumes from the scans of family members. The asymmetries for each region are reported in Table 5. In eight of the nine areas measured, no consistent asymmetry was detected for a majority of the comparison scans. The perisylvian area was the one exception ($M = 0.92$, $SD = 0.06$). In a majority ($n = 10$) of comparison scans, a left > right asymmetry was documented, which is the predicted asymmetry based on previous studies (Chi, Dooling, & Gilles, 1977b; Geschwind & Levitsky, 1968; Wada, Clarke, & Hamm, 1975; Witleson & Pallie, 1973). Because a left equals right (L = R) configuration was the less frequent configuration, it will

Table 4.
Regional volumes from comparison scans.

	Regional volumes (area/cerebrum)						Reliability
	Right			Left			
	Mean	SD	SEM	Mean	SD	SEM	(r)
Inf. Frontal	0.86	0.20	0.08	0.84	0.21	0.08	.85
Mid. Frontal	1.41	0.39	0.14	1.43	0.40	0.14	.88
Sup. Frontal	1.82	0.43	0.13	1.77	0.45	0.14	.91
Mid. Temporal	1.16	0.33	0.07	1.15	0.29	0.06	.96
Sup. Temporal	1.61	0.37	0.08	1.51	0.31	0.07	.95
Perisylvian	1.82	0.21	0.06	1.97	0.19	0.05	.93
Supramar/Ang.	1.25	0.26	0.09	1.22	0.30	0.11	.87
Occipital	3.75	0.66	0.16	3.86	0.68	0.17	.94
Hemispheres	49.51	0.01	0.001	50.49	0.01	0.001	.99

Table 5.
Asymmetries and symmetries* from comparison scans.

	Asymmetry Mean	Number L > R	L = R	R > L	Symmetry criteria
Inf. Frontal	1.06	0	14	3	1.00 +/-0.24
Mid. Frontal	0.99	1	14	2	1.00 +/-0.18
Sup. Frontal	1.04	5	5	7	1.00 +/-0.05
Mid. Temporal	1.01	3	11	2	1.00 +/-0.14
Sup. Temporal	1.07	0	13	4	1.00 +/-0.18
Perisylvian	0.92	10	7	0	1.00 +/-0.05
Supramar/Ang.	1.05	0	14	3	1.00 +/-0.22
Occipital	0.98	3	13	1	1.00 +/-0.07
Hemispheres	1.02	1	11	5	1.00 +/-0.03

*Symmetry is set at 1.64(SEM) for each measure of asymmetry.

be referred to as "atypical." Likewise, because right > left configuration was not seen in the comparison group, this configuration will also be referred to as "atypical."

Subject scans

Perisylvian asymmetries

Table 6 displays the perisylvian quotients obtained for individual members of each family. As for comparison scans, quotients of 1.00 +/-0.05 correspond to a judgement of symmetry.

Parents

Atypical perisylvian asymmetries were documented in seven of the eight parents (M = 1.04, SD = 0.10). The probability of these asymmetries occurring by chance, based on the rate of perisylvian symmetry among comparison, scans is .016. The difference between the distributions of perisylvian asymmetries from parent and comparison scans was statistically significant as well ($t = 3.73$, $df = 23$; $p < .01$). This finding reflects the fact that the perisylvian asymmetry distribution from comparison scans ranged from L > R (n = 11) to L = R (n = 7) configurations, whereas the parents had a range of configurations that extended from L > R (n = 1) through L = R (n = 4) to R > L (n = 3).

Probability levels calculated separately for mothers and for fathers are useful to evaluate whether either mothers or fathers alone produce a statistically significant

Table 6.
Perisylvian asymmetries from members of families that include SLI boys.

	Quotient	Asymmetry Configuration*	Classification
Family 1			
mother	1.06	R > L	atypical
father	1.19	R > L	atypical
sibling-1	0.95	R = L	atypical
sibling-2	1.02	R = L	atypical
proband	0.83	L > R	typical
Family 2			
mother	0.96	L = R	atypical
father	1.16	R > L	atypical
sibling	0.82	L > R	typical
proband	0.91	L > R	typical
Family 3			
mother	1.06	R > L	atypical
father	1.04	R = L	atypical
sibling	0.95	R = L	atypical
proband	1.01	R = L	atypical
Family 4			
mother	0.99	R = L	atypical
father	0.88	L > R	typical
proband	1.10	R > L	atypical
sibling	0.81	L > R	typical

* asymmetries of within 1.00 +/-0.05 are classified as L = R

finding. All mothers had atypical perisylvian asymmetries ($p = .029$; two having a L = R configuration and two having a R > L configuration. Fathers in Families 1, 2, and 3 also had atypical perisylvian asymmetries ($p = .279$).

Conditional probabilities (Fleiss, 1981) describe the degree to which atypical perisylvian asymmetries are likely to be found in mothers and in fathers whose spouses have the trait. Although the sample studied here is small, this probability provides an estimate of the degree of concordance for perisylvian asymmetries among parents that may be useful in future studies. Because all mothers had atypical perisylvian asymmetries, the conditional probability for atypical perisylvian asymmetries in mothers when fathers have the trait is 1.00. The probability of a father having atypical perisylvian asymmetry when the spouse has the trait is .75.

Probands and Siblings

Two probands also had atypical perisylvian configurations (one each L = R, R > L). These probands belonged to families 3 and 4. Because proband-1 was the subject of another study (Plante et al, 1990), his perisylvian configuration (L > R) was known prior to his inclusion in this study. Therefore, he was excluded in all probability calculations involving the perisylvian areas. The remaining three probands are too few in number to produce statistically significant probability levels.

Four of the five siblings also had atypical perisylvian configurations (L = R in all cases). No probability level is given for this rate because the sibling data are dependent.

An over-all conditional probability for offspring of parents with atypical perisylvian asymmetry was calculated. This probability describes the degree to which atypical perisylvian asymmetries in parents are associated with the same trait in children. When at least one parent has an atypical perisylvian asymmetry, the likelihood of at least one offspring having the trait is .75. The likelihood of the proband having the trait is .66. (Proband-1 has been excluded from this calculation.)

Proportional volumes

Proportional volumes (displayed in Tables 7 - 10) were obtained for eight cerebral areas in the left and right hemisphere in order to document the pattern of cortical involvement. Proportional volumes for the perisylvian area are useful in that they illucidate the nature of atypical perisylvian asymmetries. Proportional volumes in additional cerebral regions describe the extent to which atypical effects on the brain can be found outside the perisylvian areas. The proportional volumes for each subject were converted to z-scores so that they may be evaluated relative to the range seen in the comparison group.

Table 7.
Z-scores for proportional volumes for eight cerebral areas in members of Family-1.

	Mother		Father		Sib-1		Sib-2		Proband	
	rt	lf	rt	lf	rt	lf	rt	lf	rt	lf
Superior Frontal	**-2.10**	-1.56	-0.98	-1.11	-0.55	0.17	0.82	0.35	0.08	0.35
Middle Frontal	-1.38	-1.75	-0.21	-0.78	-0.15	-0.11	0.28	-0.47	0.74	0.08
Inferior Frontal	-0.61	-0.13	-1.31	-0.98	0.53	0.64	-0.45	-0.77	-1.00	0.74
Perisylvian	**2.10**	0.81	**2.19**	-0.33	**2.04**	**2.05**	**2.29**	1.49	0.05	1.30
Superior Temporal	0.72	0.38	-1.80	1.86	0.96	1.32	-0.91	-1.21	0.23	-0.12
Middle Temporal	**2.78**	**2.43**	0.58	0.04	**2.53**	1.34	-0.08	-0.47	**-1.99**	**-2.43**
Supramarginal/Angular	0.52	1.45	0.86	0.39	1.03	1.04	0.77	0.36	1.78	0.87
Occipital	0.39	0.54	-1.17	-1.25	-0.23	-1.03	-1.56	-0.52	1.22	**2.64**

Underlined values are those for which z equals or exceeds +/-1.64 ($p < .10$, two-tailed test)
Bolded values are those for which z equals or exceeds +/-1.96 ($p < .05$, two-tailed test)

Table 8.
Z-scores for proportional volumes for eight cerebral areas in members of Family-2.

	Mother		Father		Sib		Proband	
	rt	lf	rt	lf	rt	lf	rt	lf
Superior Frontal	0.53	1.17	-1.24	-1.24	**2.39**	**3.03**	-0.24	0.39
Middle Frontal	0.77	0.40	0.04	-0.67	**2.48**	1.91	0.96	0.35
Inferior Frontal	**2.24**	1.04	-0.33	-0.14	1.26	0.72	-0.49	0.28
Perisylvian	-0.86	-1.38	0.80	-1.34	-0.59	0.52	0.03	0.57
Superior Temporal	-1.39	-1.59	-1.69	-1.63	-1.27	-1.32	-1.81	**-2.10**
Middle Temporal	-1.65	-1.30	0.18	0.04	**-2.41**	**-2.33**	-0.40	-0.35
Supramarginal/Angular	0.93	0.30	-0.13	-1.05	1.67	0.81	1.57	0.09
Occipital	1.26	1.62	0.10	0.11	-1.55	-1.18	1.07	0.92

Underlined values are those for which z equals or exceeds +/-1.64 ($p < .10$, two-tailed test)
Bolded values are those for which z equals or exceeds +/-1.96 ($p < .05$, two-tailed test)

Table 9.
Z-scores for proportional volumes for eight cerebral areas in members of Family-3.

	Mother		Father		Sib		Proband	
	rt	lf	rt	lf	rt	lf	rt	lf
Superior Frontal	-0.64	-0.31	-0.36	0.17	1.62	1.58	0.18	0.45
Middle Frontal	-0.08	-0.52	0.23	-0.01	1.55	1.69	0.32	-0.33
Inferior Frontal	0.19	0.21	1.67	0.62	**1.99**	0.16	-0.93	-0.90
Perisylvian	0.19	-1.11	-0.03	-1.17	0.81	0.62	**2.11**	1.47
Superior Temporal	-1.00	-1.21	**-3.05**	**-3.83**	-0.16	-0.80	1.04	0.27
Middle Temporal	1.35	0.84	0.84	1.86	**-2.09**	**-2.80**	1.88	1.01
Supramarginal/Angular	0.47	0.44	1.81	0.88	1.24	1.74	1.64	1.52
Occipital	1.62	1.86	0.45	0.62	-1.22	-1.42	1.56	1.87

Underlined values are those for which z equals or exceeds +/-1.64 ($p < .10$, two-tailed test)
Bolded values are those for which z equals or exceeds +/-1.96 ($p < .05$, two-tailed test)

Table 10.
Z-scores for proportional volumes for eight cerebral areas in members of Family-4.

	Mother		Father		Proband		Sib	
	rt	lf	rt	lf	rt	lf	rt	lf
Superior Frontal	-0.95	-0.73	1.64	0.89	-0.61	-0.79	1.63	1.91
Middle Frontal	-0.65	-0.09	1.00	0.32	-0.56	-0.91	1.85	1.69
Inferior Frontal	0.43	0.12	-1.22	-0.38	0.30	0.29	0.79	-0.41
Perisylvian	-0.89	-1.68	1.74	**2.72**	**2.66**	1.03	-0.03	1.39
Superior Temporal	-1.24	-1.68	1.00	1.52	-1.65	-1.55	0.20	0.06
Middle Temporal	-1.55	-1.64	-0.46	-0.13	-0.17	0.07	**-2.20**	**-2.46**
Supramarginal/Angular	0.54	0.47	-0.70	-1.03	1.23	0.10	-0.76	-0.22
Occipital	-0.43	-0.27	0.91	1.90	0.82	0.85	-1.17	-0.13

Underlined values are those for which z equals or exceeds +/-1.64 ($p < .10$, two-tailed test)
Bolded values are those for which z equals or exceeds +/-1.96 ($p < .05$, two-tailed test)

With the use of a stringent criteria for deviance ($p < .05$, two-tailed test), no more than one deviant z-score was obtained for any area in the left or right hemisphere in comparison scans. For family members, at least one deviant z-score was documented for every region measured, with the exception of the supramarginal/angular gyrus area. The number of family members with deviant z-scores exceeded, by at least one, the number of deviant z-scores in the comparison group for the perisylvian, superior temporal, and middle temporal areas.

The right perisylvian areas, for members of three families (1, 3, & 4), were significantly larger ($p < .05$, two tailed) than those from comparison scans. The left perisylvian was significantly larger than in the comparison scans for members of two families (1 & 4). No statistical comparison can be made for these data due to data dependency among family members.

For the middle temporal area, z-scores for at least one member of each family were significantly different compared with the distribution of the comparison group. In all cases, deviant z-scores were obtained bilaterally. In most cases, the middle temporal areas were significantly smaller than those obtained from comparison scans. Family 1 included two members for whom this area was significantly larger than expected while one member showed the opposite effect (significantly smaller).

For the superior temporal area and the superior frontal area, members of two families had significantly deviant z-scores unilaterally or bilaterally. This number of subjects exceeds the number in the comparison group by one for both areas.

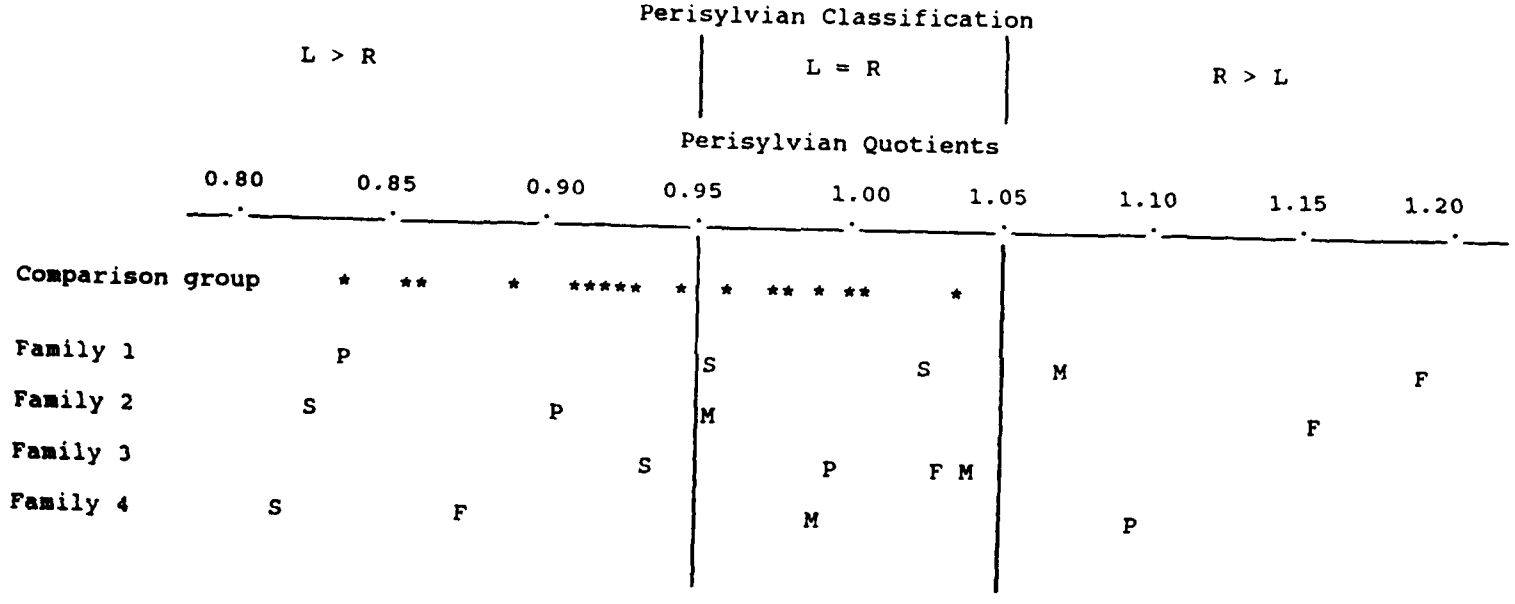
The use of a less stringent requirement for deviancy ($p < .10$, two-tailed test), increased the discrepancy between number of deviant z-scores found for subjects and in the comparison group for each of the areas identified above. This discrepency in the z-score distributions between subjects and comparisons provides converging evidence for a stable neuroanatomical effect in the areas identified.

Neuroanatomical variability within families

Marked variability was noted for the neuroanatomical profiles within each of the families. Figure 3 illustrates that the range of perisylvian quotients in each family overlapped the range documented in the comparison group. In addition, at least one quotient that was outside the range in the comparison group was documented in every family. In three of the families, the range seen in biologically related individuals exceeded the range documented for the unrelated individuals who served as comparisons. The ranges of perisylvian asymmetries obtained for the four families were not markedly different.

Similarly, there was notable variability for the rates

Figure 3.
Range of perisylvian asymmetries in the comparison group and in members of families of language-disordered boys.



F = father; M = mother; P = proband; S = sibling

of deviant proportional volumes among family members, as seen in Tables 7 - 10. In areas were deviant z-scores from subjects exceeded those from the comparison group, deviant scores tended to be distributed across the different families.

Association of atypical neuroanatomy and impaired language

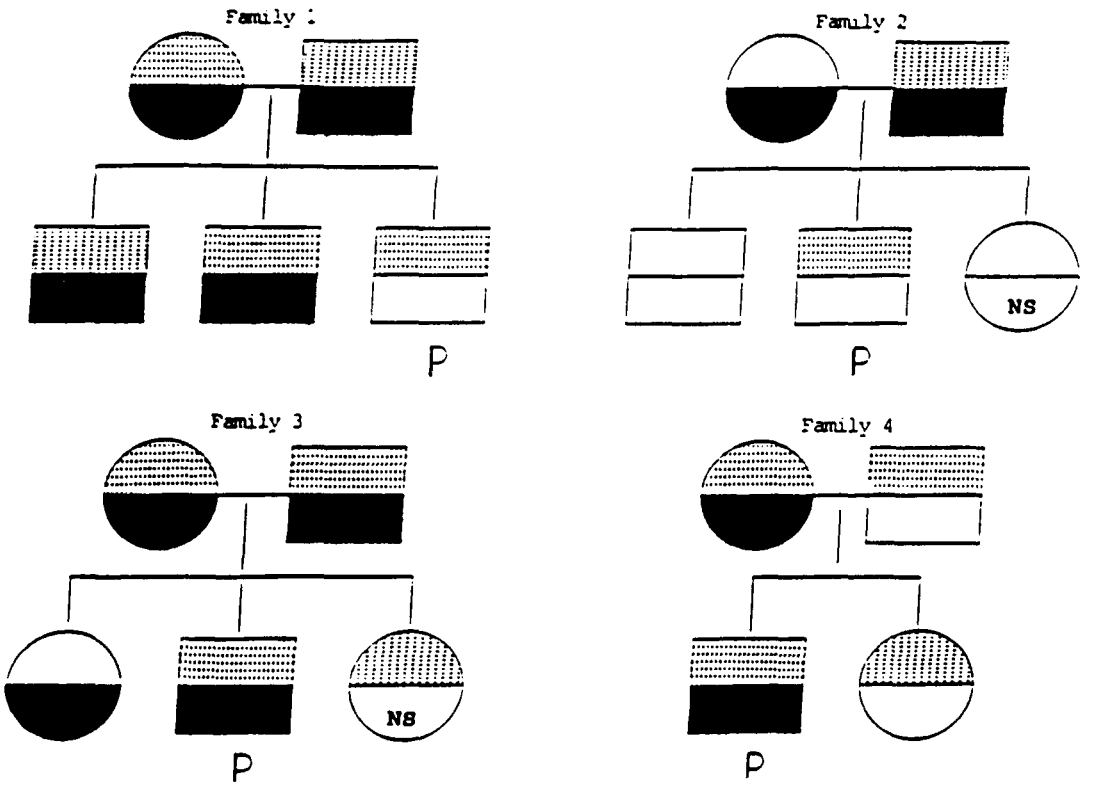
Perisylvian asymmetries

Figure 4 illustrates the family constellations of perisylvian findings (typical or atypical) in relation to indications of language impairment in each family member. Several measures are available to describe the association between atypical perisylvian asymmetries and evidence of impaired language. This association will be considered separately for parents and for children because different methods were used to determine history of childhood language impairment in these to groups of subjects.

Parents

Parents were considered to have a history of early communication difficulty if they self-reported any of the following characteristics: early speech or language difficulty, difficulty with reading, or academic failure (cf. Tallal, Ross, & Curtis, 1989a & b). One method that describes the relation between these factors and perisylvian asymmetry is a measure of the relative "sensitivity" (Fleiss, 1981, p. 6) of a history of communication

Figure 4. Constellations of atypical perisylvian asymmetry and history of communication difficulty in four families.



difficulty as an indicator of the presence of atypical perisylvian asymmetries in parents. The conditional probability that describes sensitivity is .86.

The converse of the sensitivity of a measure is it's "specificity" (Fleiss, 1981, p. 6). A proportional probability for specificity describes the degree to which atypical perisylvian asymmetries are specific to individuals with a history of communication difficulties. When parents who reported communication difficulties are compared with comparisons who lack such a history, the specificity level is .59.

A third method to describe the relation between atypical perisylvian asymmetries and a history of communication difficulty is a "proportion of attributable risk" (Fleiss, 1981, p.75). This statistic estimates the proportion of the probability of having a given condition that can be attributed to the presence of a precursor condition. It is used to estimate the degree to which early communication difficulty is attributable to the presence of atypical perisylvian asymmetries in the combined set of parents and the comparison group. The proportion of attributable risk estimated from both the parents and the comparison group is .74. This value probably underestimates attributable risk in parents of language-impaired children due to the fact that the rate of the risk factor (atypical perisylvian asymmetries) is significantly higher in this

population than in the normal population (Fleiss, 1981, p. 76).

Siblings

Most of the siblings studied had impaired language as documented by standardized testing. Three of the siblings with impaired language also had atypical perisylvian areas (families 1 & 4). A fourth sibling, in family 3, had atypical perisylvian asymmetry without a documented language impairment. The statistical procedures used with parents to describe the relation between language impairment and atypical perisylvian asymmetries require independent selection of cases. Therefore, they cannot be applied to data from siblings.

Proportional volumes

Parents

The mother in Family 1 had deviant z-scores in the superior frontal gyrus region and in the middle temporal gyrus region. This parent also had a history of communication difficulty. The conditional probability of deviant z-scores in either area given a history of communication difficulty is .14. No other region was tested using conditional probabilites because deviant scores occurred with equal frequency in parent and comparison scans.

Probands and Siblings

For the middle temporal gyrus region, one proband (in Family 1) and four siblings (one in each family) had significantly deviant z-scores. Three of the siblings also had a documented language impairment. For the superior temporal gyrus region, one proband (in Family 2) had a deviant z-score. No language impairment was documented for this child. One sibling (in Family 2) without documented language impairment also had a deviant z-score in the superior frontal gyrus region.

CHAPTER 4

Discussion

Perisylvian findings

Atypical perisylvian asymmetries were documented in seven of eight parents of SLI boys. This statistically significant finding establishes that a neuroanatomical effect previously associated with SLI in boys (Plante et al, 1990) is also found among the parents of such children. The majority of these parents also had children with atypical perisylvian asymmetries. Thus atypical perisylvian asymmetries appeared in two generations within families that have a history of language disorder.

In five cases, the atypical perisylvian asymmetries occurred because the right perisylvian areas were larger than expected, in relation to total brain size while the left perisylvian area was of the expected proportional size. For these subjects, this perisylvian configuration did not vary with the type of atypical perisylvian asymmetry; a large right perisylvian area occurred in the presence of both right > left and right = left perisylvian asymmetries. A sixth subject had perisylvian areas that were larger than expected bilaterally, which produced perisylvian symmetry. One subject had a left perisylvian area that was disproportionately large while the right was of the expected size. In this subject, this pattern produced the typical

left > right asymmetry. Larger than typical volumes provide strong evidence that an effect altered cerebral development, rather than damaged a brain that had developed normally. This pattern of results has been previously reported for an SLI boy (Plante et al, 1989) and for four males who had developmental dyslexia (Galaburda et al, 1985). In the present study, this pattern occurred for both parents and children who had atypical perisylvian asymmetries. Thus atypical perisylvian findings, which differed in degree of asymmetry across subjects, were qualitatively similar in terms of proportional volume.

Perisylvian-language relations

The assumption that perisylvian abnormalities are linked to developmental language disorders is supported from both data-based and theoretical positions. Models based on anatomical correlates of acquired language disorders highlight the importance of left hemisphere perisylvian structures (e.g. Geschwind, 1979). As discussed above, previous reports have linked developmentally impaired language to disturbances to the cellular architectonics (Cohen et al, 1988; Galaburda et al, 1985) and in alterations of gross anatomy (Cohen et al, 1988; Galaburda et al, 1985; Plante et al, 1989; Plante et al, 1990) in the perisylvian areas.

The relation between developmental language disorders and

atypical perisylvian configurations is further supported by the self-reported histories of the parents who were studied. Most parents reported problems indicative of early communication difficulty and possible language disorder. Communication difficulty was defined as a reported childhood difficulty with speech or language, academic failure, or difficulty with reading (cf. Tallal et al, 1989a & b). When parents reported such difficulty, the probability of their having atypical perisylvian configurations was relatively high (.86), indicating that such a history is a sensitive indicator of the presence of an atypical brain.

Because some individuals who did not report a history of communication difficulty also had atypical perisylvian asymmetries, this neuroanatomical marker, as measured by MRI, cannot be considered specific to language disorder alone. However, the relation is sufficiently strong as to suggest that the atypical perisylvian asymmetries represent a risk factor that increases the likelihood ($p = .74$) that an individual will experience difficulty with communication skills. These data suggest that perisylvian asymmetries reflect biological factors that place some families at risk for language disorder. The notion of familial risk is supported by the multiple cases of language disorder within families as well as the indications of similar problems in parents and their children.

Language disorder as a transmitted effect

The results of this study are consistent with the hypothesis that a transmittable, biological effect contributes to the expression of a developmental language disorder. Support for this hypothesis requires a demonstration that the behavioral and biological characteristics both occur across generations within families, and both are associated within individual members of families. Finding atypical perisylvian asymmetries in parents and their children establishes that this biological marker occurs across generations in families affected by language disorder. Likewise, indications of impaired language, based on testing in children and self report in adults, also occurred across generations in these families. Finally, the rate of both the biological and behavioral factors suggests that a proportion of the risk for communication difficulty can be attributed to the presence of atypical perisylvian asymmetries. Thus, atypical perisylvian asymmetries have been linked with the behavioral disorder in the present study and both appear to be transmitted from parents to children in the families of SLI boys.

The results of this study confirm and extend previous behavioral studies that examine the issue of transmission of language disorder through families. The high prevalence of early communication difficulties among the parents is

consistent with previous studies that used self report of communication difficulty to establish the prevalence of possible language disorder among parents of language-impaired children (Byrne et al, 1974; Neils & Aram, 1986; Tallal et al, 1989a & b; Tomblin, 1989). The findings of the present study extend previous reports by pairing self report with a neuroanatomical correlate of language disorder which is unknown to the parents. The relatively high agreement between neuroanatomical and behavioral features provide converging evidence that the high rate of communication difficulties reported for parents is not the sole result of over-reporting by parents.

Previous studies have suggested that fathers of language-impaired children are more likely than mothers to indicate signs of disorder in themselves (Neils & Aram, 1986; Tallal et al, 1989a & b). In this study, one more father than mother reported signs of communication difficulties. In contrast, one more mother than father had atypical perisylvian asymmetries. Because the majority of mothers and fathers had both atypical perisylvian asymmetries and behavioral indicators of language disorder, this study does not provide clear evidence concerning the mode of transmission of these features. The relatively high concordance between spouses for atypical perisylvian asymmetries in this small sample indicates future studies must have a relatively large sample size in order to asses

the relative contribution to the transmission of language disorder by either mothers or fathers.

Like the high rate of communication difficulty among the parents studied, the high rate of language disorder among the siblings of SLI boys is not unusual, according to reports in the literature (Bishop & Edmundson, 1986; Neils & Aram, 1986; Tallal et al, 1989a & b; Tomblin, 1989). Unlike previous studies of familial constellation of language disorder, in the present study language impairment in siblings was determined by standardized testing rather than by parental report. Siblings who evidenced significantly impaired language also had atypical perisylvian asymmetries. Thus, the data for siblings, who were not selected for study on the basis of language scores, suggest that the presence of language disorder is predictive of the presence of atypical perisylvian asymmetries.

In one sibling's case, atypical perisylvian asymmetries were documented in the absence of evidence for impaired language. This dissociation was previously documented in a normally developing twin of an SLI boy (Plante et al, 1989). A percentage of individuals without a personal history of impaired language also have atypical perisylvian asymmetries. The presence of atypical perisylvian asymmetries without a documented language disorder suggest that atypical perisylvian asymmetries are not in themselves sufficient for the expression of language disorder, but are

compatible with the notion of familial risk.

A neuroanatomical effect that occurs across generations within the same family is unlikely to be the result of accidental events such as trauma or toxins. The kind of evidence that would advance trauma or toxins as potential explanations of the perisylvian effects was lacking in this group of subjects. For example, brain damage would be expected either to leave visibly detectable evidence (e.g. parenchymal lesions, ventricular enlargement), or measurable decreases in the perisylvian volumes due to tissue loss. No evidence of damage was reported during routine clinical examination of MRI scans. More importantly, perisylvian volumes were not smaller but larger than usual in many cases. As mentioned previously, this is the opposite effect that damage would be expected to produce.

Although not tested exhaustively, toxins appear to be an unlikely explanation as well. A toxin would have to act during prenatal development to explain the perisylvian findings. Responses to case histories identified only one potential toxin, parental exposure to cigarette smoke, that occurred across families. Since this factor occurred in the comparison group as well, it cannot be considered unique to language-disordered individuals. The likelihood of an unreported, environmental toxin accounting for the perisylvian findings is reduced by the fact that parents and offspring in each family were born in different parts of the

country.

Tomblin (1989) has hypothesized that multiple cases of language disorder within families might be explained by poor rearing practices. The present findings are incompatible with this hypothesis as atypical perisylvian asymmetries reflect prenatal effects. Although environmental input will influence cellular connections (see Cowan, 1979 for examples), input does not change gyral configurations at the level of gross anatomy. Thus, less than optimal language experiences could not account for the atypical neuroanatomy observed. The presence of a prenatal neuroanatomical effect rules out parental rearing practices as a potential causal factor, although they may be maintaining factors in the expression of developmental language disorder in some cases.

Range of neuroanatomical effect

Variability of neuroanatomical findings across language-disordered children may signal a range of effect produced by a single cause, or different causes. Because of the similarity among genetically related family members for biological background, the range of neuroanatomical findings among family members is likely to reflect the effects of similar biological factors. Considered in this context, the range of neuroanatomical findings within the families studied can serve as a benchmark for evaluating neuroanatomical heterogeneity across unrelated language-

impaired individuals. It is therefore notable that the ranges for perisylvian quotients among related family members were broader than the range seen in the comparison group for three of the four families studied. This suggests that relatively similar biological backgrounds are associated with substantial neuroanatomical variability.

The perisylvian quotients did not appear to be sensitive to the degree of language disorder across members of the four families. For example, the sibling in Family 3, who lacks a documented language disorder, has a higher quotient than either of the probands in Families 1 and 2. Part of the reason for the apparent insensitivity of the degree of perisylvian asymmetry to the degree of behavioral disorder is related to the limits of MRI as a tool for documenting neuroanatomical effects. As mentioned in the introduction, MRI detects only deviations in gross anatomy. This characteristic of MRI can result in effects that seem counter-intuitive. For example, in many of the subjects studied, the proportional volumes for the left perisylvian areas were of the expected size while the right tended to be significantly larger than expected. This pattern of findings has been previously documented for the perisylvian area in an SLI boy (Plante et al, 1989) and for the plana temporale in dyslexic subjects (Galaburda et al, 1985).

This pattern appears to be contradictory to the brain-language relations documented in cases of acquired language

disorders which are associated with unilateral lesions of the left perisylvian area. However, microscopic examination of the brains of dyslexics revealed cellular disturbances predominantly in the left perisylvian area (Galaburda et al, 1985). Thus, examination of gross anatomy, the level detectable with MRI, implicated the opposite side of the brain compared with examination of cellular anatomy in the same subjects. The implications of this limitation of MRI as a tool is that many forms and combinations of cellular-level effects will not produce a detectable effect at the level of gross anatomy. This limitation probably also contributes to the apparent range of neuroanatomical effects among family members.

Pattern of cerebral involvement

The idea of a range of neuroanatomical effect is again reflected by the data indicating neuroanatomical involvement outside of the perisylvian areas. Although in every family, atypical proportional volumes were obtained for every member, the areas so identified varied across family members. Some cerebral areas were more frequently deviant than others across both individual subjects and across families. These included, along with the perisylvian areas, the superior frontal, superior temporal, and middle temporal areas.

The most frequently obtained deviant proportional volumes

were for the left and right middle temporal area. This area tended to be smaller than expected, compared with control scans. In all cases, when deviant scores were obtained, they were obtained bilaterally. This is clear evidence of bilateral involvement in the subjects studied. This is not unexpected, as an effect on the developing brain is likely to affect both the left and right hemispheres. Indeed, autopsy studies of individual with disorders that included impaired language have revealed bilateral cellular-level abnormalities (Galaburda et al, 1985; Landau et al, 1960).

Other deviant proportional volumes were obtained in the superior temporal area and superior frontal area. In both cases, both disproportionately large and disproportionately small volumes were obtained across subjects. Such scores were obtained unilaterally and bilaterally. Findings in spatially disparate regions such as the superior frontal and perisylvian areas provide evidence of widespread cerebral involvement associated with developmental language disorder. This neuroanatomical feature, like bilateral involvement is not unexpected since an in-utero effect is likely to affect all brain areas, to varying extents.

Because gyral patterns are determined during the prenatal period of cell migration (see Cowan, 1979 for an overview of cerebral development), disturbances in the size or configuration of gyri are likely to reflect an alteration in cerebral development. The fact that neuroanatomical areas

were both significantly larger and smaller than expected across family members may reflect an interaction between the stage of brain development and the agent(s) producing the neuroanatomical effect (Galaburda et al, 1985; Plante et al, 1989). It is also the case, as discussed previously, that certain combinations of adverse developmental effects will not result in a cerebral effect that is detectable at the level of gross anatomy.

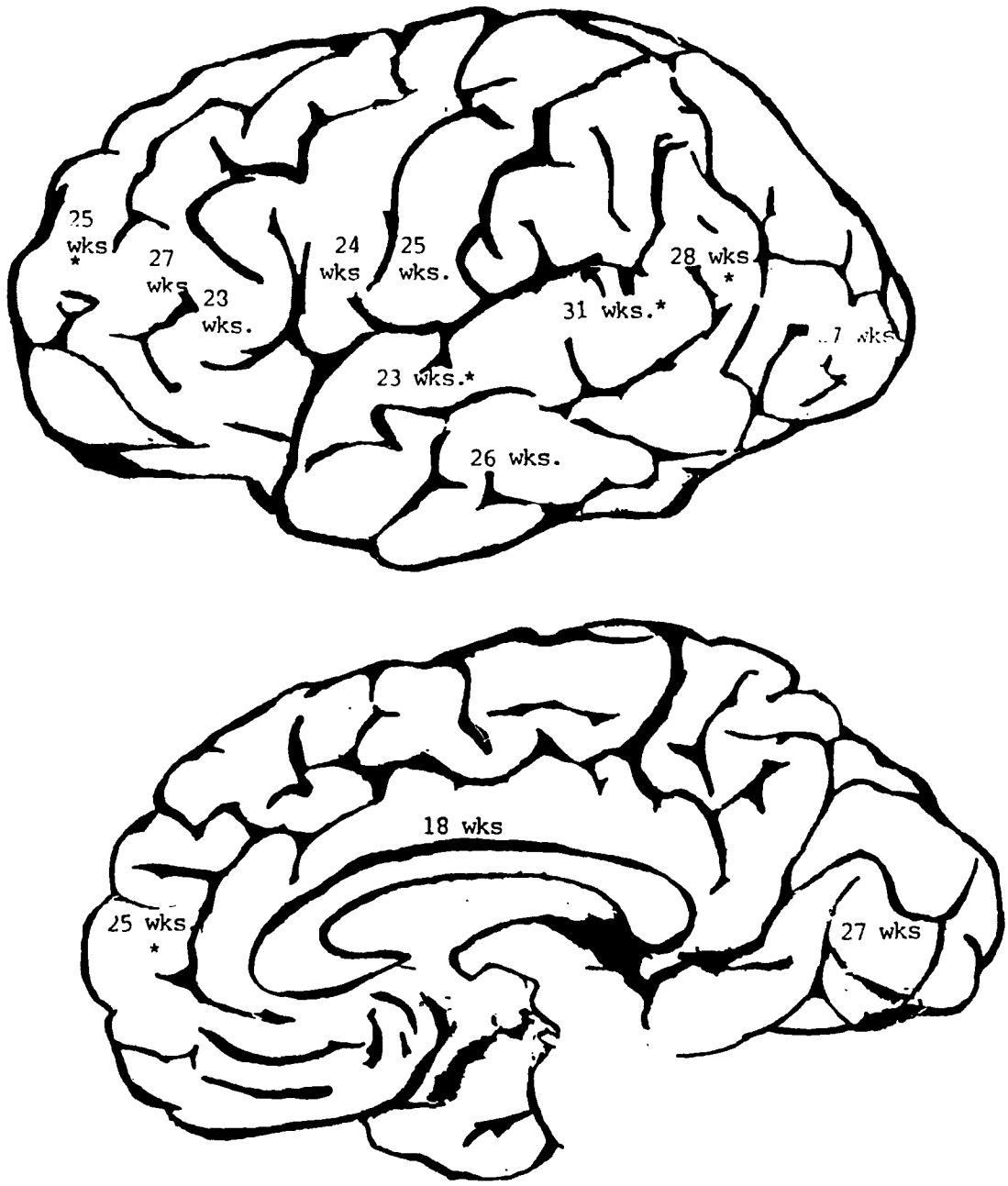
An effect that occurs early, relative to gyral development, would be expected to interfere with the period of cell genesis and migration for that gyrus. If cell genesis or migration is limited, the predicted effect would be a smaller than usual gyral volume. An effect that interferes primarily with the next stage of cerebral development, programed cell death, would result in gyral regions that are abnormally large. In the normally developing brain many more cells are generated and migrate than are needed. Excess cells, including misplaced and nonfunctional cells, are eliminated during a period of cell death (see Cowan, Fawcett, O'Leary, & Stanfield, 1984 for a review of regressive events in neurogenesis). As stated previously, the strongest evidence from this study for an effect that interacts with brain development comes from the finding of disproportionately large cerebral areas as in the perisylvian areas. Such a finding could not be explained by damage to a brain that had otherwise developed normally, but

is consistent with a failure of regressive events. The presence of neuronal ectopias documented in dyslexic subjects (Galaburda et al, 1985) also suggests a failure of the regressive events that normally eliminate heterotopic cells. In addition, the presence and location of such cells suggest an effect operated on cell migration as well as cell survival.

Given evidence for a developmental effect, the areas that were most often atypical in the subjects studied can be used to hypothesize about the course of altered cerebral development. One approach is to consider the regions where a neuroanatomical effect was documented relative to the time course of gyral development. According to data presented by Chi and colleagues, (1977b)(see Figure 5), the gyral regions most frequently implicated in this study are neither the earliest or latest to appear. The superior temporal gyrus was first identified in 25-50% of the brains studied during the 23 week of gestation. The superior frontal gyrus and middle temporal gyrus followed at 25 and 26 weeks respectively. The perisylvian area covers several regions that are first identified between 23 to 31 weeks.

An inter-hemispheric difference exists for development of three of these four areas. In each case, the gyri appear later in the left hemisphere than in the right. Only one other cerebral region has such a inter-hemispheric difference in the course of development. This is the area

Figure 5. Time course of prenatal gyral development[1]



* indicates areas for which gyri in the left hemisphere appear after the homologous gyri in the right hemisphere.

[1] based on Chi, Dooling, & Gilles, 1977b.

of the supramarginal and angular gyri, which first appears at 28 weeks gestation. It has been suggested that such left-right differences leave areas of the left hemisphere more vulnerable to adverse effects on development, compared with the homologous areas in the right hemisphere (Geschwind & Behan, 1982). It is also possible that the interhemispheric difference in the time of development increased the probability that alterations in cerebral development would result in an effect detectable by MRI in one or the other hemisphere.

Unlike the majority of areas that were most often atypical in the family members studied, no left-right difference in development was reported by Chi et al (1977b) for the middle temporal gyrus. The pattern of deviant proportional volumes for this area presents clear evidence of a bilateral effect on brain development. Findings for this area were always similar across hemispheres within any given subject. This similarity across hemispheres is consistent with the fact that these areas develop at the same rate and should be equally affected by any adverse effects.

Hormones as a contributing agent

One potentially transmittable factor, which may or may not be genetically mediated, is gonadal hormones. As reviewed in the introduction, a variety of studies of language- and learning-disabled subjects suggest a hormonal role in the

origins of these disorders (Galaburda et al, 1985; Geschwind & Behan, 1982; Plante et al, 1989; Tallal et al, 1989b). Likewise, the action of gonadal hormones may explain the cerebral effects and certain aspects of background history in the subjects of this study.

One indicator of a possible hormonal role in the transmission of language disorders is the high prevalence of dizygotic twinning (cf. Milham, 1964) among the families studied. This familial characteristic has also been documented in previous studies of individuals with impaired language skills (Galaburda et al, 1985; Plante et al, 1989; Plante et al, 1990). The presence of dizygotic twinning in the family history suggests ambient conditions existed that increased the likelihood that family members were exposed to high levels of gonadal hormones in utero. The same ambient conditions would also place other siblings at risk for developmental disorders commonly associated with impaired language. Such conditions would explain the prevalence of language and learning disorders among the siblings of the probands.

In the absence of evidence for lesion-induced alterations of asymmetries, a hormonally-mediated change in brain development is an attractive explanation for the atypical anatomical findings. Experimental manipulations of either endogenous or exogenous hormone levels alter neuronal volumes in the developing brain (Diamond, Dowling, &

Johnson, 1981; Dodson, Shryne, & Gorski, 1988, Gorski, Gordon, Shryne, & Southam, 1978; Jacobson, Csernus, Shryne, & Gorski, 1981; Pappas, Diamond, & Johnson, 1978; Pfaff, 1966; Sandhu, Cook, & Diamond, 1986). Cortical hormone receptors are present during the prenatal period in monkeys (Handa, Connolly, & Resko, 1988; Pomerantz, Fox, Sholl, Vito, & Goy, 1985; Sholl & Kim, 1989;), and in rats (MacLusky, Lieberburg, & McEwen, 1979; McEwen, Plapinger, Chaptal, Gerlach, & Wallach, 1975). These conditions provide an opportunity for hormones to influence cortical development.

One way in which gonadal hormones are known to influence brain development is by increasing survival of cells (Bloch & Gorski, 1988a & b; Dodson et al, 1988; Gorski et al, 1978; Jacobson et al, 1981; Matsumoto & Arai, 1976) and cellular connections (Yu, 1989). Because the presence of 5 alpha-reductase (which is critical to the conversion of one type of testosterone to an estrogen prior to cellular uptake) increases over the course of gestation in monkeys (Resko, Connolly, & Roselli, 1988), the greatest hormone influence in these areas would be expected during late gestation. A late hormone effect would be predicted to influence cell survival more than cell proliferation or migration. The data from these animal studies demonstrate the potential for gonadal hormones to increase cortical volumes during the prenatal period. Thus, animal models are available that

might be applicable to the effects demonstrated in the human subjects of the present study.

Implications for brain-behavior relations in SLI

Current models of brain-language relations reflect information available from the results of lesions to the premorbidly normal, mature brain (e.g. Geschwind, 1979). These models typically emphasize the importance of left perisylvian areas structures and the connecting pathways between these structures and others in cases of acquired language disorder. Studies of acquired language disorders in children (see Aram & Whitaker, 1988 for a review) appear to support the applicability of such models, again emphasizing the importance of perisylvian structures in cases of acquired language disorder. However, a model of developmentally disordered brain-language relations has not yet been put forward. For purposes of this discussion, a "developmental" disorder will refer to disordered behavior that occurs subsequent to an alteration of brain development. The data from this study indicate a model of developmentally-disordered language would have both similarities and differences with models based on cases of acquired language disorder.

Like models based on acquired language disorders, a model for developmental language disorders would emphasize the importance of the perisylvian areas. In this and previous

studies, individuals with developmental disorders that included impaired language had either cellular level pathology (Cohen et al, 1988; Galaburda et al, 1985; Landau et al, 1960) or alterations in the normal pattern of asymmetry for this area (Cohen et al, 1988; Galaburda et al, 1985; Plante et al, 1989; Plante et al, 1990). Because asymmetries of the perisylvian area reflect an asymmetry in the underlying cytoarchetectonic organization (Galaburda & Sanides, & Geschwind, 1978), a disturbance of the asymmetry is likely to reflect altered cytoachetectonics (e.g. Galaburda et al, 1985). This disturbance may be responsible for the increased difficulty language-disordered children have in developing the skills that other children seem to master with relative ease. Therefore, although the language-disordered subjects studied to date lack evidence of acquired brain damage to the perisylvian area, a developmental alteration of this cerebral region appears capable of placing these individuals at increased risk for the disorder.

The importance of the perisylvian area in contributing to the expression of a language disorder can be evaluated relative to other areas of abnormality that were documented in language-disordered individuals. Only one autopsy study documented extra-perisylvian effects but in a child with multiple handicaps (Landau et al, 1960). This calls into question the degree to which such findings can be considered

correlates of a language disorder. The present study is the first imaging study to examine specific cerebral areas in addition to perisylvian structures. Although other cerebral regions were identified as atypical, no one area was as consistently associated with evidence of a language disorder as was the perisylvian.

The extra-perisylvian findings provide a point of departure between models of acquired and developmental language disorders and their basis in the brain. In acquired language disorders, unilateral damage to the left perisylvian area is sufficient for the expression of a language disorder. The evidence from this study, and an autopsy report (Galaburda et al, 1985), suggest both widespread and bilateral involvement is typical in cases of developmental language disorder. It is unknown whether the widespread and bilateral damage is a prerequisite for the expression of the language disorder or merely an artifact of the interaction of the causal agent(s) with the course of brain development. It is possible that the widespread and bilateral involvement is one factor limiting the brain's ability to compensate and overcome the structural limitations of the left perisylvian area, thus explaining the persistent nature of the disorder in many of the subjects studied here and elsewhere (e.g. Aram, Ekleman, & Nation, 1984; Aram & Nation, 1980; Griffiths, 1969; Hall & Tomblin, 1978).

The widespread and bilateral involvement provides a potential neuroanatomical explanation for the nonlinguistic deficits documented in groups of specifically language-impaired children (Johnston & Smith, 1989; Johnston & Weismer, 1983; Kamhi, 1981; Kamhi, Catts, Koenig, & Lewis, 1984; Kamhi, Catts, Mauer, Apel, & Gentry, 1988; Nelson, Kamhi, & Apel, 1987; Savich, 1984). Although nonverbal skills have been addressed from a variety of theoretical frameworks, certain similarities recur across studies in the types of tasks that specifically language-impaired children perform poorly. These tend to be tasks that involve spatial imagery and rotation. Others have interpreted the association of both language and these types of nonverbal deficits as reflecting a generalized deficit in symbolic manipulation (Kamhi, 1981). In light of the current neuroanatomical findings, the behavioral constellation of verbal and select nonverbal deficits co-occur in SLI children because these children have bilateral brain involvement. From this perspective, it is not surprising that language deficits are associated with the type of nonverbal deficits frequently associated with right hemisphere damage.

Appendix A

Standardized Test Battery

Nonverbal measures:

Performance criterion:

All nonverbal scales appropriate to the child's age were administered. To be considered for this study, probands had to score above -1.5 SD according to the test's normative data. No exclusion criterion was used for siblings on this measure.

Kaufman Assessment Battery for Children (K-ABC) (Kaufman & Kaufman, 1983)

Test description:

The K-ABC consists of a series of subtests that require nonverbal responses. The subtests were designed to tap skills that require "simultaneous" and "sequential" processing skills. The subtests include Triangles, which requires the child to construct a pictured figure using blue and yellow triangles. Matrix Analogies requires the child to select a geometric figure that corresponds to a given figure in the same way as a previously presented pair are related. Spatial Memory requires a child to remember the locations of pictured items. Hand Movements requires the child to repeat a series of hand movements demonstrated by the examiner. Photo Series requires the child to order a series of photographs to show a sequence of events.

Vineland Social Maturity Scale (Vineland) (Sparrow, Balla, & Cicchetti, 1984)

Test description:

The Vineland is a parent-interview scale that samples daily living and social skills to assess the child's typical behaviors. Parents of children ages 5;11 and below are also surveyed concerning their child's motor skills.

Language measures

Morpho-syntactic:

Performance criterion:

To qualify as a proband, a boy had to score at or below

-1.64 SD below the normative sample mean on one or more of the morpho-syntactic tests. No criterion was set for siblings.

Clinical Evaluation of Language Functioning-Revised (CELF-R) (Semmel, Wiig, & Secord, 1987)

Six subtests contribute to computation of an overall language level for children 8;0 and above. Oral Directions taps the child's ability to follow directions of varying length and complexity. Word Classes requires the child to identify which two of four words go together semantically. Semantic Relations requires the child to identify spatial, temporal, and attributional relations. Formulated Sentences requires the child to make up a sentence using a given word. Recalling Sentences requires the child to repeat verbatim sentences spoken by the examiner. Sentence Assembly requires the child to make two sentences from a group of written words.

Subtests of the CELF-R that are known to discriminate between language-impaired and language-normal children (Semmel, Wiig, & Secord, 1987) were administered to children ages 8;0 and above. These include Oral Directions, Semantic Relations, Recalling Sentences, and Sentence Assembly. Additional subtests that contribute to the computation of an overall language level were also administered to most children.

Illinois Test of Psycholinguistic Abilities--Grammatical Closure subtest (ITPA-GC) (Kirk, McCarthy, & Kirk, 1968)

In giving the ITPA-GC, the examiner prompts the child to say words that contain grammatical morphemes using a sentence closure task.

Northwestern Syntax Screening Test--Expressive subtest (NSST-E) (Lee, 1969)

The NSST-E requires the child first to listen to two sentences that correspond to two pictures. Then, the child is asked to repeat the sentences verbatim.

Test of Auditory Comprehension of Language-Revised (TACL-R) (Carrow-Woodfolk, 1985)

The TACL-R requires the child to point to a picture that corresponds to an orally presented word or sentence. The TACL-R has three subtests: Word Classes and Relations, Grammatical Morphemes, and Elaborated Sentences.

Token Test for Children (Token) (DiSimoni, 1978)

The Token has five subtests that require the child to touch or manipulate colored shapes in response to oral directions. The first four subtests increase the length of directions while the fifth increases the linguistic complexity of the directions.

Vocabulary:

Performance criterion: None used.

Peabody Picture Vocabulary Test-Revised (PPVT-R) (Dunn & Dunn, 1981)

The PPVT-R requires the child to point to one of four pictures that corresponds to a word said by the examiner.

Expressive One Word Picture Vocabulary Test (EOWPVT) (Gardner, 1979)

The EOWPVT requires the child to name a word when shown its picture.

Articulation:

Performance criterion: None used.

Templin-Darley Test of Articulation (Templin-Darley) (Templin & Darley, 1968)

The Templin-Darley samples the child's single word articulation skills through picture naming. Normative data is available for children ages 8 and below.

Spontaneous Language Analysis

Performance criterion: None used.

Developmental Sentence Scoring (DSS) (Lee, 1974)

The DSS analyses 50 consecutive sentences (utterances that have a subject and predicate) from a spontaneous speech sample elicited by a clinician. A DSS was used on language samples from children ages 7;11 and younger.

Appendix B

Case Histories

Child Language Laboratory
MRI and Language Impairment in Children
Case History

CHILD'S FORM

Date:_____

Person completing this form_____

Relation to child:_____

IDENTIFYING INFORMATION:

Child's name:_____

Address:_____

_____

Phone:_____

Family: Father_____
years of education:_____
Occupation:_____

Mother_____
years of education:_____
Occupation:_____

Children: Names	Sex	Age	Grade

Are there any other persons living in your household

Names:	Relationship to family:

Child's doctor:_____
phone:_____

Is your child currently receiving speech-language therapy:_____
For how long:_____
Child's Speech-Language Therapist:_____
Clinic name:_____
phone:_____

STATEMENT OF THE PROBLEM:

Describe in your own words what problem you child is having with speech,

_____

_____

_____

When was the problem first noticed:_____

_____

Who first noticed the problem:_____

What changes in your child's language or speech have you noticed since then:

_____

_____

Do you have any thoughts on the cause of the problem? Please describe:

_____

_____

What have you done to try to help your child's speech:_____

_____

Has it helped:_____

Is your child aware of having difficulty with speech or language:_____

_____

If so, how does he react:_____

_____

Do any relatives have problems with speech or language:_____

If so, relation to child:_____ age_____

_____ age_____

Do any relatives stutter or stammer:_____

If so, relation to child:_____ age_____

_____ age_____

Do any relatives receive special services in school:_____

If so, relation to child:_____

_____

_____

Do any relatives have have a developmental disorder (e.g. Dyslexia, Autism Down syndrome, attention deficit disorder, hyperactivity, learning disabilities or others)

If so, relation to child	type of developmental disorder
_____	_____
_____	_____
_____	_____

What hand does your child prefer to use: right____ left____ either hand____
Are any of the child's family members left handed_____
If so, how many family members_____
Are any of the child's relatives left handed_____
If so, how many relatives_____

SPEECH, LANGUAGE AND HEARING DEVELOPMENT

Did the child make babbling or cooing sounds during the first 6 months of life

_____

At what age did the child say his first word:_____

What was his first words:_____

Did the child keep adding words once he started to talk:_____

At what age did he start using 2 and 3 word sentences:_____

Examples:_____

_____

Did speech learning ever seem to stop for a period of time:_____

If so, when_____

Do you have any thoughts on the possible cause:_____

_____

_____

Does your child talk frequently_____, occasionally_____, never_____.
Does he prefer to talk_____, gesture_____, talk and gesture_____.
Does he frequently use sounds only_____, single words_____,
2 word sentences_____, 3 word sentences_____,
more than 3 word sentences_____. Examples:_____

_____

Can he tell a simple story:_____

Does your child make sounds incorrectly:_____ If so, which ones_____

_____

How well is he understood by his parents:_____

by sisters and brothers_____

by relatives and strangers_____

Will he get common objects when he's asked to:_____

Does he ever have trouble remembering what you have told him:_____

If so, when does this happen:_____

Does your child use any books or records:_____

Does he enjoy being read to:_____

Describe any recent changes in your child's speech_____

_____

_____

BIRTH HISTORY:

This is our biological_____, foster_____, adopted_____ child.
How many pregnancies has the mother had_____
Has the mother had any miscarriages_____, stillbirths_____, abortions____

If so, which pregnancies_____
Which pregnancy was this child_____.
Mother's age at the time of the child's birth:_____.

Where was the child born (town)_____.
Where was the mother born (town) _____.
Where was the father born (town) _____.

Were there any medical problems before this pregnancy_____,
during the pregnancy _____.

If so, what_____ _____

Did the mother have any of the following during the pregnancy:
German measles_____ Toxemia_____ Anemia _____
Accidents/injuries_____ Kidney infections_____

Did the mother take any prescription or nonprescription drugs (including alcohol) during the pregnancy:_____ If so what_____

Did the mother smoke during pregnancy_____
If so, how many cigarettes per day_____
Did the father smoke during the pregnancy_____
If so, how many cigarettes per day _____
Does either parent or any member of the household currently smoke_____

If so who_____ how much_____

_____ _____

Was the child delivered full term_____, premature_____ months_____
Was delivery normal_____, cesarian_____, breach_____
Length of hard labor_____. Were forceps used:_____
Was the mother given drugs during delivery_____

What was the child's APGAR score (if known)_____
Any birth injuries_____
Was the child a blue baby_____
Did the child require oxygen_____
Was the child an Rh baby_____
Did the child receive any special medications or treatment at birth_____

If so, what_____

How long were the mother and child in the hospital_____

Any comments about the pregnancy or birth_____

_____

DEVELOPMENTAL HISTORY:

Did your child have feeding problems_____ describe_____ _____

_____

Did your child have sleeping problems_____ describe_____

_____

When did your child first:

sit unsupported_____	crawl_____
reach for an object_____	stand_____
walk unaided_____	run_____
bladder trained_____	bowel trained_____
night trained_____	

MEDICAL HISTORY:

Please check if your child, any members of his immediate family (parents brothers, or sisters) or any other relatives (grandparents, aunts, uncles cousins) have had any of the following. Only report on blood relatives.

	child	family	relatives
Allergies			
Asthma			
Blood disease			
Bowel disorders			
Chronic colds			
Celiac disease			
Colitis			
Convulsions			
Dermatomyositis			
Diabetes Mellitus			
Diphtheria			
Earaches			
Ear infections			
Encephalitis			
Epilepsy			
Chronic Headaches			
Hay fever			
Head injuries			
Heart problems			
High fevers			
Influenza			
Kartagener's syndrome			
Mastoiditis			
Mastoidectomy			
Measles			
Meningitis			
Migraine headaches			
Multiple sclerosis			
Muscle disorder			
Myasthenia gravis			
Myxedema			
Pneumonia			
Polio			
Regional ileitis (Chron's)			
Rheumatic fever			
Rheumatoid arthritis			
Scarlet fever			
Thyroid disease			
Uveitis			
Young's syndrome			
Whooping cough			

Has your child had middle ear infections_____

At what age was the first one_____

Did they occur frequently_____ describe the frequency_____

_____

Was he seen by a doctor concerning these infections_____

Did he ever have tubes inserted into his ear drums_____

for how long_____

Are the ear infections still occurring_____ how frequently_____

_____

If not, at what age did they stop_____

Did the child ever have seizures or convulsions_____

If so, when_____

how frequently_____

was medical attention sought_____

what was the outcome_____

Describe any other illnesses, accidents, injuries, operations and hospitalizations the child has had (include the age of the child at the time)._____

_____

_____

_____

Is the child now under any medical treatment or taking any medication_____

If so, describe:_____

_____

_____

Is the child's health good_____

Are there any twins in the family or among relatives _____

relation to child _____

_____

Are the twins on the _____ mother's or _____ father's side of the family
Are the twins _____ identical or _____ fraternal.

EDUCATIONAL HISTORY:

Did or does your child attend day care_____, Nursery school_____, Kindergarten_____, grade school_____

For what subjects is he an average student_____

above average_____

below average_____

Does he receive special help in school_____ describe_____

_____

What is your impression of your child's learning ability_____

_____

Has your child received any special education, psychological or hearing services at school_____ If so, describe (include child's age at the time

_____

_____

What things does the child do particularly well _____

_____

Does he have any special interests or talents_____

_____

Child Language Laboratory
MRI and Language Impairment in Children
Case History

ADULT FORM

Date:_____

IDENTIFYING INFORMATION:

Your name:_____

Address:_____

_____

Phone:_____

Your occupation:_____

Your Family:
Father: years of education:_____
Occupation:_____

Mother: years of education:_____
Occupation:_____

Number of brothers_____ Number of sisters_____
Number of half-brothers_____ Number of half-sisters_____

Your children: Names	Sex	Age	Grade
_____	_____	_____	_____
_____	_____	_____	_____
_____	_____	_____	_____
_____	_____	_____	_____
_____	_____	_____	_____

Are there any other persons living in your household

Names:	Relationship to family:
_____	_____
_____	_____

Your doctor:_____
phone:_____

Do you have any metal in your body (e.g. surgical clips, pins) _____

Are you currently pregnant _____

Do you have a serious problem with claustrophobia _____

SPEECH, LANGUAGE AND HEARING DEVELOPMENT

Were you a late talker? _____ yes _____ no _____ don't know
if yes, when did you begin to use words _____

Did you have any difficulty with speech or language as a child?
____ yes _____ no _____ don't know

if yes, describe_____

_____

Did you have any difficulty pronouncing certain sounds?
____ yes ____ no ____ don't know

if yes, describe _____

_____

Did you stutter? ____ yes ____ no ____don't know

if yes, until what age _____

Did you have any trouble learning to read? ____ yes ____ no

if yes, describe_____

_____

Do you enjoy reading today? ____ yes ____ no ____ I don't read much

Did you receive any special services in school? ____ yes ___no

if yes, describe_____

Did you as a child have any difficulties similar to those you see in your child? _____ yes _____ no

if yes, describe_____

_____

_____

_____

Do you have any sisters, brothers, nieces or nephews with similar speech or langauge difficulties? ____ yes ____ no

if yes, relation to you _____

_____

Do any relatives have problems with speech or language:_____

If so, relation to you: _____ age_____

_____ age_____

_____ age_____

_____ age_____

Do any relatives stutter or stammer:_____

If so, relation to you: _____ age_____

_____ age_____

Do any relatives receive special services in school:_____

If so, relation to you: _____

_____

_____

Do any relatives have have a developmental disorder (e.g. Dyslexia, Autism Down syndrome, attention deficit disorder, hyperactivity, learning disabilities or others)

If so, relation to child — type of developmental disorder

_____ _____

_____ _____

_____ _____

_____ _____

What hand do you prefer to use: right____ left____ either hand____

Are any of your family members left handed_____
If so, how many family members_____

Are any of your relatives left handed_____
If so, how many relatives_____

BIRTH HISTORY:

Were you adopted? ____ yes ____ no
Were any of your relatives adopted? ____yes _____ no

if yes, relation to you _____

How many pregnancies did your mother have_____
Did your mother had any miscarriages_____, stillbirths_____
If so, which pregnancies_____
Which pregnancy were you_____
Your mother's age at the time of your birth:_____

Where were you born (town)_____
Where was your mother born (town) _____
Where was your father born (town) _____.

Were there any medical problems when your mother was pregnant with you?
_____ yes _____ no _____ don't know

If so, what_____

Was your mother under unusual stress while pregnant with you _____

_____

_____

Did your mother take any prescription or nonprescription drugs (including alcohol) during the pregnancy:_____ yes ____ no ____ don't know

If so what_____

Did your mother smoke during pregnancy_____
If so, how many cigarettes per day_____
Did your father smoke during the pregnancy_____
If so, how many cigarettes per day _____

Were you delivered ____ full term _____ premature(_____ months)
Was delivery _____ normal _____ cesarian _____ breach _____ don't know
Any birth injuries_____
Were there any medical concerns at birth ____ yes ____ no ____ don't know
If so, what_____

Any comments about the pregnancy or birth_____

_____

DEVELOPMENTAL HISTORY:

Did you have any difficulty with

____ sleeping ____ standing ____ toilet training
____ eating ____ walking

MEDICAL HISTORY:

Please check if you, any members of your immediate family (parents, brothers sisters, or children) or any other relatives (grandparents, aunts, uncles cousins) have had any of the following. Only report on blood relatives.

	you	family	relatives
Allergies			
Asthma			
Blood disease			
Bowel disorders			
Chronic colds			
Celiac disease			
Colitis			
Convulsions			
Dermatomyositis			
Diabetes Mellitus			
Diphtheria			
Earaches			
Ear infections			
Encephalitis			
Epilepsy			
Chronic Headaches			
Hay fever			
Head injuries			
Heart problems			
High fevers			
Influenza			
Kartagener's syndrome			
Mastoiditis			
Mastoidectomy			
Measles			
Meningitis			
Migraine headaches			
Multiple sclerosis			
Muscle disorder			
Myasthenia gravis			
Myxedema			
Pneumonia			
Polio			
Regional ileitis (Chron's)			
Rheumatic fever			
Rheumatoid arthritis			
Scarlet fever			
Thyroid disease			
Uveitis			
Young's syndrome			
Whooping cough			

Did you have middle ear infections as a child ___ yes ___ no ____don't know
Did any of your family or relatives have middle ear infections?
____ yes ____ no ____ don't know
if yes, relation to you _____

_____

Have you ever have seizures or convulsions ____ yes ____ no

If so, when_____

how frequently_____

was medical attention sought_____

what was the outcome_____

Describe any other illnesses, accidents, injuries, operations and hospitalizations you have had (include the age at the time)._____

_____

Are you now under any medical treatment or taking any medication_____.

If so, describe:_____

Have you ever had any of the following:

____ stroke ____ head trauma ____ epilepsy

Are there any twins in the family or among relatives _____

relation to you _____

_____

Are the twins in ____your family _____mother's side or _____father's side of the family
Are the twins _____ identical or _____ fraternal.

EDUCATIONAL HISTORY:

I completed ____ grade school, ____ highschool, ____ technical school, ____ college

My last degree obtained was _____.

For what subjects were you an average student?

_____

above average _____

below average _____

My favorite subject was _____

My least favorite subject was_____

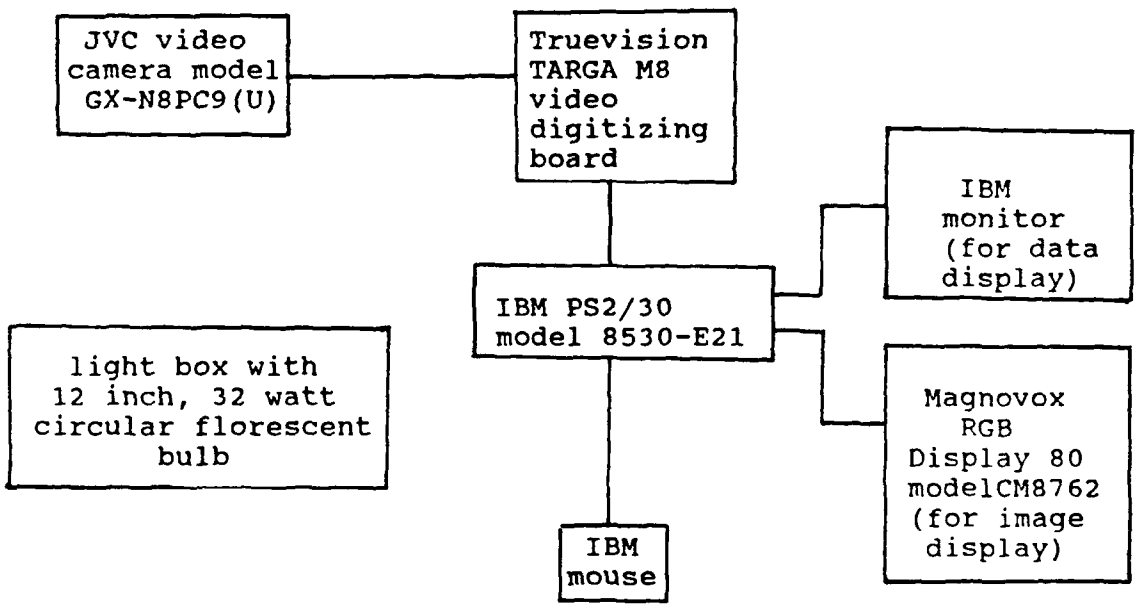
Did you enjoy academics? _____ if no, explain _____

_____

Do you have any further comments on your school experiences? _____

_____

Instrumentation



Appendix D

MRI Measurement Protocols

I. Calibration.

A. Height and level of two lightboxes were measured and verified as identical within the limits of measurement. Lightboxes were used interchangeably.

B. Two carpenters levels, placed on the camera and lightbox, were used to verify that the camera tilt was level with the lightbox (and counter-top) tilt.

C. A calibration check was performed daily.

1. The plane of the camera and light-box were judged to be parallel by reading levels placed on each.

2. A calibration figure of 300 millimeters squared was placed at a random angle on the lightbox and measured. This was repeated twice so that three calibration measures were taken a day.

II. General measurement procedures.

A. All images were place on the center of the screen for measurement.

B. Measures were done from the inferior to the superior slices.

C. Areas in the right hemisphere (left side of MRI) were measured before areas in the left.

D. Homologous areas in the left and right hemisphere were measured in the same sitting.

E. When magnetic inhomogeneities are present, the person measuring adjusted the brightness/contrast to compensate for the differences in pixel intensity across inhomogeneous regions.

F. When measuring perpendicular to the midline, only the medial third of the midline was used as the reference for perpendicular.

G. Partial volumes were excluded from measures.

H. Blood vessels and ventricular volumes were included in measures.

I. When sulci are used as landmarks, the grey-white margin was used to define the end of a sulcus in T2 weighted images. When this margin does not form a narrow, distinct point, the midpoint between two adjacent fiber tracts was used. In cases where a grey-white margin was not visible, the visible end of the sulcus was used.

J. Persons measuring scans were trained on MRI scans of subjects not used in this study. The measurement rules and anatomy books (primarily DeArmond, Fusco, & Dewey, 1976 and Takayoshi & Asao, 1978) were available to assist in identifying anatomical landmarks.

III. Hemispheric measures.

A. Measures started from the anterior pole of the hemisphere and follow the cortical edge.

B. On slices where the hemispheres are connected, a straight line connected the anterior and posterior junctures of the hemispheres.

C. The brainstem was excluded from hemisphere measures.

D. In a small number of scans where full cerebral volumes were not included on axial scans, the coronal view was used to obtain hemispheric volumes. Data, presented below, from individuals who had both axial and coronal scans that included the full cerebral volume indicated no difference in the hemisphere volumes obtained using either view, and that volumes obtained from either view were highly correlated (r=.98 right hemisphere; r=.94 left hemisphere).

IV. The Frontal Lobe

A. Superior Frontal Gyrus area (SFG).

1. The SFG began on the first slice in which the orbital gyrus is no longer visible.
2. It started at the sulcus posterior to the medial branch of the SFG.
3. The starting point was connected, with a straight line, to the anterior-lateral edge of the head of the caudate nucleus. If this point was not visible, the anterior-lateral edge of the lateral ventricles was used.
4. This point was connected, with a straight line, to the superior frontal sulcus.
5. The measure then followed the cortical edge to the starting point.
6. The measure ended on the last slice in which the body of the caudate nucleus was visible, below the area of the centrum semiovale.

B. Middle Frontal Gyrus area (MFG).

1. The MFG began on the first slice in which the orbital gyrus was no longer visible.
2. It started at the edge of the cortex at the juncture of the SFG and MFG.
3. From there it followed the superior frontal sulcus to the grey-white margin.
4. This point was connected, with a straight line, to the anterior-lateral edge of the caudate nucleus (see also B-3 above).
5. This point was connected, with a straight

line, to the grey-white margin of the inferior frontal sulcus.

6. The measure followed the sulcus to the cortical edge.

7. The cortical edge was traced to the starting point.

8. The measure ended on the last slice in which the body of the caudate nucleus is visible, below the area of the centrum semiovale.

C. Inferior Frontal Gyrus area (IFG).

1. The IFG began on the first slice in which the orbital gyrus was no longer visible.

2. It started at the cortical edge at the juncture of the MFG and IFG.

3. From this point, it followed the inferior frontal sulcus to the grey-white margin.

4. This point was connected, with a straight line, to the anterior-lateral edge of the head of the caudate nucleus (see also B-3 above).

5. This point was connected, with a straight line, to the juncture of the IFG and the pars triangularis.

6. The cortical edge was traced to the starting point.

8. The measures ended when the IFG was no longer visible.

D. Anterior Cingulate area.

1. The anterior cingulate area began on the first slice in which the genu of the corpus callosum was visible.

2. It started at exterior edge of the cingulate sulcus, defined as the first sulcus anterior to the corpus callosum.

3. From this point, it followed the sulcus interiorly to the grey-white margin.

4. This point was connected, with a straight line, to the most lateral point of the anterior horn of the caudate nucleus.

5. This point was connected, with a straight line, to the juncture of the corpus callosum and cingulate gyrus.

6. The edge of the cortex was traced to the starting point.

7. The measures ended on the most superior slice in which the body of the corpus callosum is visible.

V. Perisylvian area (PSA).

A. The PSA began on the first slice on which a sylvian

fissure was seen, typically just superior to the middle cerebral arteries.

B. The most medial-posterior pixel of grey mater at the end of the sylvian fissure was located. When multiple Hescl's gyri were visible, the sylvian was defined as posterior to the first Hescl's gyrus (c.f. Witleson & Pallie, 1971).

C. This point was connected, with a line drawn perpendicular to the midline, to the cortical edge.

D. The cortical edge was followed anteriorly to the medial-most edge of the temporal poles or to the posterior margin of the IFG.

E. This point was connected, with a straight line, to the starting point.

F. The measures ended on the last slice in which the insula was visible.

VI. Temporal Lobe.

A. Middle Temporal Gyrus area (MTG).

1. The MTG began on the most inferior slice on which some portion of the occipital pole and gyrus rectus was visible.

2. The most medial-posterior pixel of grey matter corresponding to the most anterior and lateral lobule of the temporal lobe was located.

3. This point was connected, with a line drawn perpendicular to the midline, to the cortical edge.

4. The cortical edge was followed posteriorly to the temporal-occipital sulcus, defined as anterior to the second girl convolution from the occipital pole (this gyrus may have two visible branches).

5. This point was connected, with a straight line, to the starting point.

6. The MTG ended at the level of the perisylvian area.

C. Superior Temporal Gyrus area (STG).

1. The STG began on the second slice of the PSA.

2. The most medial-posterior pixel of grey mater at the end of the sylvian fissure was located. When multiple Hescl's gyri are visible, the sylvian was defined as posterior to the first Hescl's gyrus.

3. This point was connected, with a line drawn perpendicular to the midline, to the cortical edge.

4. The measure followed the cortical edge posteriorly to the temporal-occipital sulcus, defined as anterior to the second girl convolution from the occipital pole (this gyrus may have two visible branches).

5. This point was connected, with a straight line, to the starting point.

6. The STG ended on the slice below the level on which the splenium of the corpus callosum is visible.

VI. Occipital area (Occip.).

A. The Occip. began on first slice on which occipital poles are visible.

B. Measurement started at the juncture of the temporal and occipital lobes (see V. B-4 above).

C. This point was connected with the anterior juncture of the cerebellum and cerebrum.

D. The measure traced the edge of the cortex, around the occipital poles, to the starting point.

E. The Occip ended on the slice below the level on which the splenium of the corpus callosum was visible.

VII. Parietal Lobe

A. Supramarginal/angular gyrus area (SMG/ANG).

1. The SMG/ANG began on the first slice in which the splenium of the corpus callosum is seen completely bridging the hemispheres.

2. Measurement started at the most medial-posterior margin of the sylvian fissure (see also V. B-4 above).

3. This point was connected, with a line drawn perpendicular to the midline, to the cortical edge.

4. The measure followed the cortical edge posteriorly to the interparietal sulcus, defined as anterior to the second girl convolution from the occipital pole (these gyri may have two visible branches).

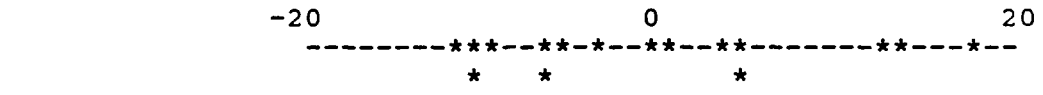
5. This point was connected, with a straight line, to the starting point.

6. The SMG/ANG ended on the last slice below the centrum semiovale.

Differences between hemispheric measures taken from brains cut in coronal and axial sections.

Subject	Coronals		Axials		Differences	
	Right	Left	Right	Left	Right	Left
1	370.71	369.47	365.66	351.40	5.05	18.07
2	372.15	351.81	378.56	361.95	-6.41	-10.14
3	357.72	347.57	368.55	352.93	-10.83	-5.36
4	333.83	329.17	334.27	332.30	-0.44	-3.13
5	387.72	366.92	386.37	376.86	1.35	-9.94
6	392.58	385.41	398.44	394.17	-5.86	-8.76
7	309.26	308.97	294.80	295.73	14.46	13.24
8	394.08	392.64	388.97	386.03	5.11	6.61

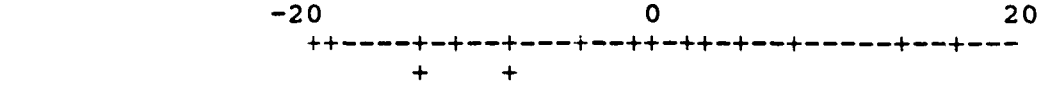
Distribution of Differences between Coronal and Axial Measures



Inter-operator differences in axial hemispheric measures.

Subject	Axial-1		Axials-2		Differences	
	Right	Left	Right	Left	Right	Left
1	370.72	369.47	378.56	373.15	-7.84	-3.68
2	378.56	361.95	397.97	381.94	-19.41	-19.99
3	368.55	352.93	381.13	365.64	-12.58	-12.71
4	334.27	332.30	344.82	340.26	-10.55	-7.96
5	386.37	376.86	387.77	376.61	-1.40	-0.25
6	398.44	394.17	391.91	389.57	6.53	4.60
7	294.80	295.73	292.49	292.48	2.31	3.25
8	388.97	386.03	374.48	369.01	14.49	17.02

Distribution of Inter-operator Differences



Correlated t test for difference between inter-operator differences in axial measures and coronal-axial differences.

right hemisphere differences $t = 0.79$, $df = 7$, $p > .10$
left hemisphere differences $t = 0.45$, $df = 7$, $p > .10$

Pearson product moment correlation for axial and coronal hemisphere measures.

right hemisphere $r = 0.98$, $df = 6$, $p < .01$
left hemisphere $r = 0.94$, $df = 6$, $p < .01$

Appendix E

A Pilot Study of Regional Brain Measures

Method

Subjects

Three boys who were identified as specifically language impaired with deficits in the comprehension or use of grammatical morphemes served as subjects for this pilot study. These subjects were previously described in a separate paper (Plante, Swisher, Vance, and Rapcsak, 1990). None served as subjects in this dissertation study.

Scans from these three boys were compared with comparison scans from 10 normal male volunteers. Comparison scans met the same selection criteria detailed in the methods section of this dissertation.

Procedures

Measurement protocols for ten regions of the brain were developed on the normal volunteers. Regional measures were developed with consideration for differences in their time of development (cf. Chi, Dooling, & Gilles, 1977). These protocols are detailed in Appendix D. Measurement protocols used anatomical landmarks that could be clearly identified on all brains. This accomplished several goals. First, it increased the likelihood that the protocol could be followed reliably across subjects. It also insured measurements would be more sensitive to individual variations in brain configuration than more arbitrary divisions of the brain (e.g. grids). Regional brain volumes were converted into proportions of the total brain volume to eliminate the effects of brain size on the measurements.

A protocol for any given area was considered reliable when intra-operator reliable measures were obtained on all control scans used in this pilot effort. Measures were considered reliable when volumes for all control scans obtained on two sessions correlated at or above $r = .83$ (70% shared variance) using a Pearson's product-moment correlation. Reliability figures for measurements on both subject and control scans ranged between .87 and .99.

Regions above the level of the centrum semiovale were not measured. This was partially due to the limitations of using axial scans. Regions towards the top of the brain were affected by partial volume effects due to the marked curvature of the brain on the more superior slices. This effect limited the extent to which anatomical landmarks could be reliably identified. Similar problems interfered with measurement of gyrus rectus, therefore it was excluded from the group of measures.

Results

The results of the pilot study are given in the table below. A z-score was calculated for each regional brain volume. This facilitated a comparison of the subject scans and comparison scans. Z-scores beyond +/-1.30 ($p < .20$) in any subject were considered to indicate a measure of potential interest for additional study. The following measures met this criterion: inferior frontal area, middle frontal area, superior frontal area, superior temporal area, middle temporal area, perisylvian area, supramarginal area, occipital lobe. Several measures, the inferior frontal area, and the superior and middle temporal areas were significantly ($z > +/-1.96$; $p < .05$) different from normal in at least one subject. The earliest developing area, the anterior cingulate area, was not markedly different in the SLI boys than in the comparison group. Measures for this area will not be reported for dissertation subjects.

Table
Proportional volumes of brain regions from three SLI-GI boys.

Brain region	Adam		Brian		Cory		controls	
	volume	z	volume	z	volume	z	mean	SD
Rt. Cingulate	0.51	0.19	0.47	-0.06	0.48	0.00	0.48	0.16
Lf. Cingulate	0.57	0.41	0.47	-0.18	0.51	0.06	0.50	0.17
Rt. I. Frontal	0.81	0.14	1.10	**2.21**	0.55	**-1.71**	0.79	0.14
Lf. I. Frontal	0.73	-0.42	1.05	1.26	0.81	0.00	0.81	0.19
Rt. M. Frontal	1.34	0.68	1.63	**1.66**	1.27	0.41	1.15	0.29
Lf. M. Frontal	1.29	0.19	1.32	0.31	1.03	-0.81	1.24	0.26
Rt. S. Frontal	1.64	0.18	1.88	0.87	1.45	-0.38	1.58	0.34
Lf. S. Frontal	1.52	0.08	1.92	**1.31**	1.44	-0.15	1.49	0.33
Rt. S. Marginal	1.24	0.08	1.49	1.08	1.17	-0.20	1.22	0.25
Lf. S. Marginal	1.14	-0.25	1.53	**1.70**	1.18	-0.05	1.19	0.20
Rt. Perisylvian	2.17	0.59	1.86	-0.38	2.43	**1.41**	1.98	0.32
Lf. Perisylvian	2.21	0.19	1.77	-1.19	2.12	-0.09	2.15	0.32
Rt. Occipital	4.69	0.38	3.55	**-1.78**	4.04	-0.85	4.49	0.53
Lf. Occipital	5.26	**1.38**	4.02	-0.87	3.95	-1.00	4.50	0.55
Rt. M. Temporal	1.09	-0.67	1.12	-0.42	1.15	-0.17	1.17	0.12
Lf. M. Temporal	0.99	**-1.90**	0.87	**-3.10**	1.15	-0.30	1.18	0.10
Rt. S. Temporal	0.72	**-2.17**	0.88	**-1.72**	0.74	**-2.11**	1.50	0.36
Lf. S. Temporal	0.62	**-2.53**	0.89	**-1.72**	1.18	-0.85	1.46	0.33

Reference

American National Standards Institution (1969). Specification for Audiometers. ANSI S3.6-1969. New York: American National Standards Institute, Inc.

Aram, D.M., Ekelman, B.L., & Nation, J.E. (1984). Preschoolers with language disorders: Ten years later, Journal of Speech and Hearing Research, 27, 232-244.

Aram, D.M., & Nation, J.E. (1980). Preschool language disorders and subsequent language and academic difficulties. Journal of Communication Disorders, 13, 159-170.

Aram, D.M., & Whittaker, H.A. (1988). Cognitive sequelae of unilateral lesions acquired in early childhood. In D.L. Molfese & S.J. Segalowitz (Eds.), Brain Lateralization in Children: Developmental Implications. New York: Guilford Press.

Arnold, G.E. (1961). The genetic background of developmental language disorders. Folia Phoniatrica, 13, 246-254.

Bear, D., Schiff, D., Saver, J., Greenberg, M., & Freeman, R. (1986). Quantitative analysis of cerebral asymmetries Archives of Neurology, 43, 598-603.

Bishop, D.V.M. & Edmundson, A. (1986). Is otitis media a major cause of specific developmental language disorders? British Journal of Disorders of Communication, 21, 321-338.

Bloch, G.J. & Gorski, R.A. (1988a). Cytoarchitectonic analysis of the SDN-POA of the intact and gonadectomized rat. Journal of Comparative Neurology, 275, 604-612.

Bloch, G.J. & Gorski, R.A. (1988b). Estrogen/progesterone treatment in adulthood affects the size of several components of the medial preoptic area in the male rat. Journal of Comparative Neurology, 275, 613-622.

Byrne, B.M., Willerman, L., & Ashmore, L.L. (1974). Severe and moderate language impairment: Evidence for distinctive etiologies. Behavioral Genetics, 4, 331-345

Carrow-Woodfolk, E. (1985). Test of Auditory Comprehension of Language-Revised Allen TX: DLM Teaching Resources.

Caviness, V.S., Evrard, P., & Lyon, G. (1978). Radial neuronal assemblies, ectopia and necrosis of developing

cortex: A case analysis. Acta Neuropathologica, 41, 67-72.

Chapman, R.S. (1981). Computing mean length of utterance in morphemes. In Miller JF, Assessing Langauge Production in Children. Baltimore: University Park Press.

Chi, J.G., Dooling, E.C., & Gilles, F.H. (1977a). Left-right asymmetries of the temporal speech areas of the human fetus. Archives of Neurology, 34, 346-348.

Chi, J.G., Dooling, E.C., & Gilles, F.H. (1977b). Gyral development of the human brain. Annals of Neurology, 1, 86-93.

Cohen, M., Campbell, R., & Yaghmai, (1988, January). Neuropathological abnormalities in developmental dysphasia. Paper presented at the International Neuropsychological Society Meeting, New Orleans, LA.

Cowan, W.M. (1979). The development of the brain. Scientific American, 241, 112-133.

Cowan, W.M., Fawcett, J.W., O'Leary, D.M., Stanfield, B.B. (1984). Regressive events in neurogenesis. Science, 225, 1258-1265.

DeArmond, S.J., Fusco, M.M., & Dewey, M.M. (1976). Structure of the Human Brain. New York: Oxford Press.

DeFries, J.C., Singer, S.M., Foch, T.T., & Lewitter, F.I. (1978). Familial nature of reading disability. British Journal of Pychiatry, 132, 361-367.

Diamond, M.C., Dowling, G.A., & Johnson, R.E. (1981). Morphologic cerebral cortical asymmetry in male and female rats. Experimental Neurology, 71, 261-268.

DiSimoni, F. (1978). Token Test for Children Hingham, MA: Teaching Resources.

Dodson, R.E., Shryne, J.E., & Gorski, R.A. (1988). Hormonal modification of the number of total and late-arising neurons in the central part of the medial preoptic nucleus of the rat. Journal of Comparative Neurology, 275, 623-629.

Doughty, C., Booth, J.E., McDonald, P.G., & Parrott, R.F. (1975). Effects of oestradiol-17 , oestradiol benzoate and the synthetic oestrogen RU 2858 on sexual differentiation in the neonatal female rat. Journal of Endocrinology, 67, 419-424.

Dunn, L.M. & Dunn, L.M. (1981). Peabody Picture Vocabulary Test-Revised Circle Pines, MN: American Guidance Service.

Dvorak, K. & Feit, J. (1977). Migration of Neuroblasts through partial necrosis of the cerebral cortex in newborn rats--Contribution to the problems of morphological development and developmental period of cerebral microgyria. Acta Neuropathologica, 38, 203-212.

Dvorak, K., Feit, J., & Jurankova, Z. (1978). Experimentally induced focal microgyria and status verrucosus deformis in rats--Pathogenesis and interrelation histological and autoradiographical study. Acta Neuropathologica, 44, 121-129.

Fleiss, J.L. (1981). Statistical Methods for Rates and Proportions. New York: John Wiley & Sons.

Fundudis, T., Kolvin, I., & Garside, R.F. (1979). A follow-up study: predictive importance--cognitive language and educational development. In Speech Retarded and Deaf Children: Their psychological development. New York: Academic Press, pp 51-66.

Galaburda, A.M., Sanides, F., & Geschwind, N. (1978). The human brain: Cytoarchitectonic left-right asymmetries in the temporal speech region. Archives of Neurology, 35, 812-817.

Galaburda, A.M., Sherman, G.F., Rosen, G.D., Aboitiz, F., & Geschwind, N. (1985). Developmental dyslexia: four consecutive patients with cortical anomalies. Annals of Neurology, 18, 222-233.

Gardner, M. (1979). Expressive One Word Picture Vocabulary Test Navato, CA: Academic Therapy Publications.

Geschwind, N. (1979). Specialization of the human brain. Scientific American, 241, 180-201.

Geschwind, N. & Behan, P. (1982). Left handedness: association with immune disease, migraine, and developmental learning disorder. Proceedings of the New York Academy of Sciences, 7, 5097-5100.

Geschwind, N. & Levitsky, W. (1968). Human brain: Asymmetries in temporal speech region. Science, 161, 186-187.

Gorski, R.A., Gordon, J.H., Shryne, J.E., & Southam, A.M. (1978). Evidence for a morphological sex difference

within the medial preoptic area of the rat brain. Brain Research, 148, 333-346.

Griffiths, C.P.S., (1969). A follow-up study of children with disorders in speech. British Journal of Disorders of Communication, 4, 46-56.

Hall, P.K. & Tomblin, J.B. (1978). A follow-up study of children with articulation and language disorders. Journal of Speech and Hearing Disorders, 43, 227-241.

Handa, R.J., Connelly, P.B., & Resko, J.A. (1988). Ontogeny of cytosolic androgen receptors in the brain of the fetal rhesus monkey. Endocrinology, 122, 1890-1896.

Jandel Scientific (1988). JAVA: Jandel Video Analysis Software. Corte Madera, CA: Author.

Jacobson, C.D., Csernus, V.J., Shryne, J.E., & Gorski, R.A. (1981). The influence of gonadectomy, androgen exposure, or a gonadal graft in the neonatal rat on the volume of the sexually dimorphic nucleus of the preoptic area. The Journal of Neuroscience, 1, 1142-1147.

Jernigan, T.L., Tallal, P.A., & Hesselink, J. (1987, November). Cerebral morphology on magnetic resonance imaging in developmental dysphasia. Paper presented at the annual meeting of the Society for Neuroscience, New Orleans, LA.

Johnston, J.R. & Smith, L.B. (1989). Dimensional thinking in language impaired children. Journal of Speech and Hearing Research, 32, 33-38.

Johnston, J.R. & Weismer, S.E. (1983). Mental rotation abilities in language-disordered children. Journal of Speech and Hearing Research, 26, 397-403.

Kamhi, A.G. (1981). Nonlinguistic symbolic and conceptual abilities of language impaired and normally developing children. Journal of Speech and Hearing Research, 24, 446-453.

Kamhi, A.G., Catts, H.W., Koenig, L.A., & Lewis, B.A. (1984). Hypothesis-testing and nonlinguistic symbolic abilities in language-impaired children. Journal of Speech and Hearing Disorders, 49, 169-176.

Kamhi, A.G., Catts, H.W., Mauer, D., Apel, K., & Gentry, B.F. (1988). Phonological and spatial processing abilities in langauge- and reading-impaired children. Journal of Speech and Hearing Disorders, 53, 316-327.

Kaufman, A.S. & Kaufman, N.L. (1983). Kaufman Assessment Battery for Children Circle Pines, MN: American Guidance Service.

Kirk, S.A., McCarthy, J.J., & Kirk, W.D. (1968). Illinois Test of Psycholinguistic Abilities Chicago: University of Illinois Press.

Koff, E., Naeser, M.A., Piendiaz, J.M., Foundas, A.L., Levine, H.L. (1986). Computed tomographic scan hemispheric asymmetries in right- and left-handed male and female subjects. Archives of Neurology, 43, 487-491.

Landau, W.M. Goldstein, R., & Kleffner, F.R. (1960). Congenital aphasia: A clinicopathologic study. Neurology, 10, 915-921.

Lee, L.L. (1974). Developmental Sentence Scoring. Evanston: Northwestern University Press.

Lee, L.L. (1969). Northwestern Syntax Screening Test Evanston, IL: Northwestern University Press.

Lieberburg, I., MacLusky, N., & McEwan, B. (1980). Cytoplasmic and nuclear estradiol-17 binding in male and female rat brain: regional distribution, temporal aspects and metabolism. Brain Research, 193, 487-503.

Matsumoto, A.M. & Arai, Y. (1976). Effect of estrogen on early postnatal development of synaptic formation in the hypothalamic arcuate nucleus of female rats. Neuroscience Letters, 2, 79-82.

MacLusky, N., Lieberburg, I., & McEwan, B. (1979). The development of estrogen receptor systems in the rat brain and pituitary: perinatal development. Brain Research, 178, 143-160.

McCauley, R.J. & Swisher, L. (1984b). Psychometric review of language and articulation tests for preschool children. Journal of Speech and Hearing Disorders, 49, 34-42.

McCardle, P. & Wilson, B.E. (1987, November). Hormonal influence on language development in physically advanced children. Paper presented at the American Speech and Hearing convention, New Orleans.

McEwan, B.S., Plapinger, L., Chaptal, C., Gerlach, J., & Wallach, G. (1975). Role of fetoneonatal estrogen binding proteins in the association of estrogen with neonatal brain

cell nuclear receptors. Brain Research, 96, 400-406.

McShane, D., Risse, G.L., & Rubens, A.B. (1984). Cerebral asymmetries on CT scan in three ethnic groups. International Journal of Neurosciences, 23, 69-74.

Milham S. (1964). Pituitary gonadotrophin and dizygotic twinning. The Lancet, 1, 556.

Miller, J.F. (1981). Assessing Language Production in Children. Baltimore: University Park Press.

Neils, J. & Aram, D.M. (1986). Family history of children with developmental language disorders. Perceptual and Motor Skills, 63, 655-658.

Nelson, L.K., Kamhi, A.G., & Apel, K. (1987). Cognitive strengths and weaknesses in language-impaired children: One more look. Journal of Speech and Hearing Disorders, 52, 36-43.

Pappas, C.T.E., Diamond, M.C., & Johnson, R.E. (1978). Effects of ovariectomy and differential experience on rat cerebral cortical morphology. Brain Research, 154, 53-60.

Perakis, A. & Stylianopoulou, F. (1986). Effects of a prenatal androgen peak on rat brain sexual differentiation. Journal of Endocrinology, 108, 281-285.

Perlman, S.M. (1973). Cognitive abilities of children with hormone abnormalities: Screening by psychoeducational tests. Journal of Learning Disabilities, 6, 26-33.

Pfaff, D.W. (1966). Morphological changes in the brains of adult male rats after neonatal castration. Journal of Endocrinology, 36, 415-416.

Plante, E. (1989) [Quantitative measurements of brain structures from magnetic resonance images of normal volunteers] Unpublished data.

Plante, E., Swisher, L., & Vance, R. (1989). Anatomical correlates of normal and impaired language in a set of dizygotic twins. Brain and Language, 37, 643-655.

Plante, E., Swisher, L., Vance, R., & Rapcsak, S. (1990).
MRI Findings in Four Consecutive Cases of Specific Language Impairment. Manuscript submitted for publication.

Pomerantz, S.M., Fox, T.O., Sholl, S.A., Vito, C.C., & Goy, R.W. (1985). Androgen and estrogen receptors in fetal

rhesus monkey brain and anterior pituitary. Endocrinology, 116, 83-89.

Resko, J.A., Connolly, P.B., & Roselli, C.E. (1988). Testosterone 5 reductase activity in neural tissue of fetal rhesus macaques. Journal of Steroid Biochemistry, 29, 429-434.

Rosenberger, P.B. & Hier, D.B. (1980). Cerebral asymmetries and verbal intellectual deficits. Annals if Neurology, 8, 300-304.

Rubens, A.B. (1984). The neuroanatomy of disorders of auditory comprehension. Seminars in Neurology, 4, 174-178.

Samples, J.M. & Lane, V.W. (1985). Genetic possibilities in six siblings with specific language learning disorders. ASHA, , 27-32.

Sandhu, S., Cook, P., & Diamond, M.C. (1986). Rat cerebral cortical estrogen receptors: Male-female; right-left. Experimental Neurology, 92, 186-196.

Scholl, S.A., & Kim, K.L. (1989). Estrogen receptors in the rhesus monkey brain during fetal development. Developmental Brain Research, 50, 189-196.

Sparrow, S.S., Balla, D.A., & Cicchetti, D.V. (1984). Vineland Adaptive Behavior Scales Circle Pines, MN: American Guidance Service.

Savich, P.A. (1984). Anticipatory imagery ability in normal and language-disabled children. Journal of Speech and Hearing Research, 27, 494-501.

Sherman, G.F., Galaburda, A.M., & Geschwind, N. (1983). Ectopic neurons in the brain of the autoimmune mouse: a neuropathological model of dyslexia? Society of Neuroscience Abstracts, 9, 939.

Takayoshi, M. & Asao, H. (1978). An Atlas of the Human Brain for Computerized Tomography. New York: Igaku Shoin.

Tallal, P., Ross, R., & Curtiss, S. (1989a). Familial aggregation in specific language impairment. Journal of Speech and Hearing Disorders, 54, 167-173.

Tallal, P., Ross, R., & Curtiss, S. (1989b). Unexpected sex-ratios in families of language/learning-impaired children. Neuropsychologia, 27, 987-998.

Templin, M.C. & Darley, F. (1968). Templin-Darley Test

of Articulation Iowa City: The University of Iowa.

Tomblin, J.B. (1989). Familial concentration of developmental language impairment. Journal of Speech and Hearing Disorders, 54, 287-295.

Wada, J.A., Clarke, R., & Hamm, A. (1975). Cerebral hemispheric asymmetry in humans. Archives of Neurology, 32, 239-246.

Witelson, S.F. & Pallie, W. (1973). Left hemisphere specialization for language in the newborn. Brain, 96, 641-646.

Yu, W.A. (1989). Administration of testosterone attenuates neuronal loss following axotomy in the brain-stem motor nuclei of female rats. The Journal of Neuroscience, 9, 3908-3914.

Attachment 4

AFOSR-TR- 91 0982

THESIS BY:

DAVID A. WAGNER

STANFORD UNIVERSITY

Subcontract No.# S-789-000-009

FRACTURE CHARACTERIZATION FOR THERMOINELASTICITY

A DISSERTATION

SUBMITTED TO THE DEPARTMENT OF MECHANICAL ENGINEERING

AND THE COMMITTEE ON GRADUATE STUDIES

OF STANFORD UNIVERSITY

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

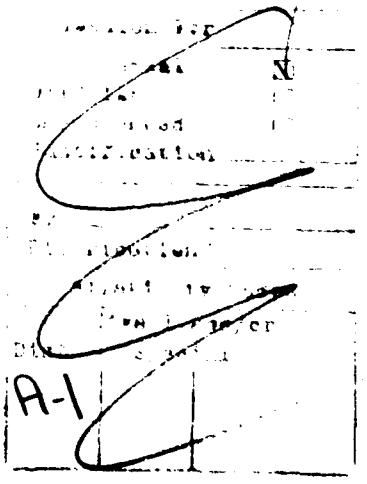
FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

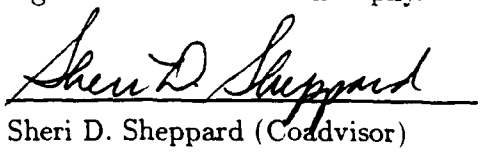
By

David Anthony Wagner

June 1990

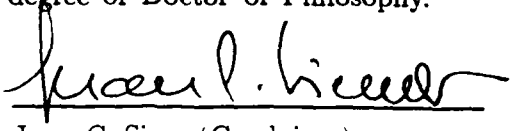


I certify that I have read this dissertation and that in my opinion it is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.



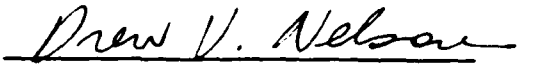
Sheri D. Sheppard (Coadvisor)

I certify that I have read this dissertation and that in my opinion it is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.



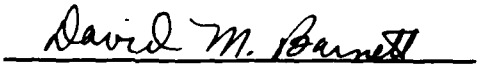
Juan C. Simo (Coadvisor)

I certify that I have read this dissertation and that in my opinion it is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.



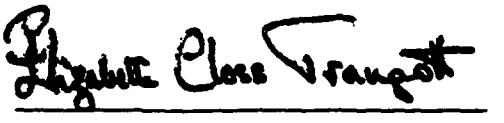
Drew V. Nelson

I certify that I have read this dissertation and that in my opinion it is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.



David M. Barnett (Material Science)

Approved for the University Committee on Graduate Studies:



Dean of Graduate Studies

Abstract

This research develops and verifies the path domain independent integral **S** that is exactly the energy release rate for an extending crack within a thermoinelastic material response region. Emanating from thermoinelastic continuum mechanics and Noether's theorem from classical field theory, the **S** integral defines the force acting on an extending crack and represents a conservation law for a crack free body. Limited physical experiments and computational investigations verify the **S** integral for uncoupled thermoinelasticity. **S** offers a parameter to improve the understanding of the strength and reliability of materials subjected to thermomechanical loadings.

The theoretical development produces the **S** path domain independent integral for two quasi static cases: uncoupled thermoinelasticity and fully coupled thermoinelasticity. Proofs demonstrate the path domain independence and the total energy release rate aspects of the **S** integral.

A limited experimental program and an associated computational investigation verifies that the **S** integral characterizes fracture for uncoupled thermoinelasticity. Experiments on a crack free aluminum sheet and finite element results from a simple dogbone specimen verify the **S** conservation law. Fracture resistance tests on aluminum 2024 demonstrate that **S** equals the crack driving force under thermomechanical loadings. The fracture resistance experiments consider two specimen geometries and two specimen thicknesses.

Acknowledgements

The author gratefully acknowledges all those who contributed to this research effort in both substance and spirit. The U. S. Air Force Office of Scientific Research funded this effort through the Laboratory Graduate Fellowship Program. The insights and guidance of Drs. Sheppard and Simo assisted in the timely completion of this research. The helpful comments of Drs. Nelson and Barnett improved the quality of the final dissertation. The interest, support and guidance of Dr. George P. Sendeckyj (Wright Research and Development Center) facilitated all aspects of this fascinating research.

The author also gratefully acknowledges the Fatigue, Fracture and Reliability Group of the Structural Integrity Branch within the Structures Division of the Flight Dynamics Laboratory at Wright Research and Development Center, Wright Patterson Air Force Base for supporting the experimental program. Specifically, the individual contributions of Larry Bates, Don Cook, Harold Stalnaker and the numerous technicians at the lab enabled the experiments to be conducted.

This dissertation is dedicated to Anna Marie Jarocki Wagner for her untiring love and encouragement throughout this research.

Table of Contents

List of Tables

List of Illustrations

1

Introduction to Thermomechanical Fracture

The National Aerospace Plane *might* carry passengers from Washington D. C. to Tokyo in less than two hours. Electrical power *might* be generated with improved efficiency at higher operating temperatures. Nuclear power plants *might* be designed and built more economically with improved reliability and safety. Aircraft jet engines *might* produce greater thrust by operating at higher temperatures with improved turbine design. These future engineering triumphs and many others depend at least in part on an improved understanding of the strength and reliability of materials subjected to complex thermomechanical loadings. The research herein develops and verifies a parameter that improves this understanding.

These engineering advances demand fault tolerant designs. The critical structures and components must withstand operating loads without failure between maintenance inspections. Understanding the thermomechanical conditions likely to cause crack extension is essential in determining critical crack length criteria and recommending inspection intervals.

Fracture and crack growth characterizations under general thermoinelastic conditions present difficult challenges to the analyst and designer. Neither the stress

intensity factor, **K**, from linear elastic fracture mechanics nor the **J** integral from elastic plastic fracture mechanics adequately describes the crack driving force for general thermomechanical loadings.

Certain parameters characterize crack tip force and crack advance criterion for the creep aspects of inelastic material response. The C^* energy parameter developed by Landes and Begley [1976] * following a suggestion by Goldman and Hutchinson [1975], and the ΔT_c integral proposed by Atluri [1982] provide insight concerning creep fracture. The $\mathbf{J}^{*SH}$ integral formulated by Simo and Honein [1990] addresses crack advance in elasto-viscoplastic material. The strain range partitioning method introduced by Manson et al. [1971] and popularized by Halford et al. [1973] suggests a procedure to estimate crack growth and fatigue life during cyclic loading with combined creep and plasticity. However, no single parameter completely characterizes the crack driving force for general thermoinelasticity.

The **S** path domain independent integral parameter developed in this dissertation describes the force on a singularity, such as a crack tip, under thermoinelastic material response. Severe thermal gradients coupled with mechanical loads are considered. Classical field theory provides the foundation for the S parameter. The **S** integral gives the theoretical crack driving force which, at the instant of incipient fracture, should also equal the material fracture resistance. Since **S** is the total energy released per unit crack advance, the parameter represents a necessary but not sufficient condition for fracture.

Nonetheless, the limited fracture resistance experiments performed on aluminum 2024-T351 plate and 2024-T3 sheet suggest that the **S** integral adequately describes the crack driving force. Critical **S** values, calculated at initial crack extension, from tests with thermal gradients match isothermal fracture toughness values.

---

* Complete reference given in Appendix **REFERENCES**

The experiments also demonstrate the detrimental effect a thermal gradient can have on fracture resistance.

The **S** parameter can be readily computed from elastic plastic finite element analysis results for use in analysis and design of critical components. The computation of the path domain independent integral **S** has been successfully implemented as a postprocessor to the existing FEAP finite element code (Taylor [1977] and Zienkiewicz and Taylor[1989]) for the uncoupled thermoinelastic case.

1.1 Motivation for Research

Improved reliability assessments for the design of critical structures and components require a parameter characterizing the crack driving force for fracture under thermomechanical loading. This need results from the increasing demands placed on both established and new materials in industrial applications such as airframe structures, turbine blades and power generation components, among others. The lack of a parameter solidly based on thermoinelastic continuum mechanics to characterize incipient fracture and crack growth appears most dramatically in those loading regimes where transient thermal stresses and severe thermal gradients determine crack behavior.

Accurate reliability assessments are crucial for ensuring safety. The potential for loss of human life, severe environmental impacts and costly property damage due to failure of a critical structure or component demands that engineers adequately understand fracture and crack growth due to general thermoinelastic material response. The energy based **S** integral provides a parameter for evaluating crack behavior under thermomechanical loading.

From finite element analysis, a designer would calculate the maximum **S** parameter for a specified loading condition and assumed material. The designer

would check that the chosen material has adequate fracture resistance, measured as J_{1c} that depends on material thickness. Similarly, knowing the ΔS range for an assumed thermomechanical loading history, the designer would estimate the number of load cycles between when a crack can be reliably detected during inspection and when that crack will grow to a critical length. This number of cycles would translate into a recommended inspection interval accounting for reasonable safety factors.

In this discussion, the characterizing parameter **S** plays the role for thermoinelastic fracture that **K**, the stress intensity factor, plays for linear elastic fracture mechanics and that **J** plays for elastic plastic fracture mechanics. Naturally, this dissertation does not provide fracture resistance **S** values for many materials, nor does it present any exponential crack growth law empirical factors. Through the theoretical development of **S** based on thermoinelastic mechanics and the few tests on 2024 aluminum this dissertation strongly suggests that **S** provides a parameter characterizing the crack driving force. The S parameter potentially can provide designers and analysts information concerning the likelihood of fracture and/or crack advance during thermoinelastic material response.

1.2 Examples of Thermomechanical Loadings

The path domain independent **S** integral developed herein addresses fracture under severe thermal gradients coupled with mechanical loads. Advances, for example, in the aerospace, power utility and automobile engine industries have created these demanding thermomechanical loading regimes.

Within the aerospace industry, many present and future airframe structures and jet engine components must operate reliably under severe thermomechanical loadings. For hypervelocity vehicles such as the National Aerospace Plane (Bylinsky [1986]) designs predict high thermal gradients and transients through the wing

thickness. With the wing leading edge and foremost top surface reaching temperatures above 3500°F and the liquid hydrogen fuel stored in the wing at cryogenic temperatures the structural components in the wing potentially experience thermal gradients in the range of 1000°F per inch (Gabor [1986]).

The wing structure must safely carry the aerodynamic loads in addition to the severe thermal gradients. The induced thermal stresses added to the fluctuating mechanical stresses cause local inelastic regions at the crack tips that may govern component life. Naturally, the design process must investigate fracture potential under these combined thermomechanical loadings. Currently, no theoretically based fracture mechanics parameter addresses thermoinelastic material response for the fracture potential associated with these loading conditions.

For jet engine turbine components, the thermal cycling associated with each flight determines safe life. Marchand et al. [1988] report that none of the three common fracture parameters, the stress intensity factor, **K**, the strain intensity factor, or the **J** integral, successfully correlate crack growth data for tests on turbine engine hot section materials. A modified stress intensity factor, ΔK_σ, which includes the hardening/softening material behavior as well as load shedding considerations yields the best data correlation for the two nickel based alloys studied. Since the **S** integral proposed herein captures the thermoinelastic material response, it might provide a characterizing parameter for crack growth data reduction.

Within the power utilities, turbines and other components face severe thermomechanical loadings. Power generating turbines operate at higher temperatures to increase power generating efficiency. However, this leads to higher thermal stresses and the potential for accelerated crack growth. Gamble and Paris [1976] successfully predict allowable thermal fatigue crack growth limits for root cracks in gas turbine discs. For cracks initiated by thermal shock then grown by internal

pressure and thermal gradient fluctuations at re-entrant corners in cooling conduits of boiling water reactors. Smith [1986] reports stable growth regimes for the curved crack fronts commonly found in service inspections. These efforts combine experimental results and linear elastic fracture mechanics analyses. No single parameter captures the crack behavior under the thermoinelastic material response.

The thermal shock problem associated with a loss of coolant accident in a nuclear reactor is an area of extensive study. The potential catastrophic consequences of such an accident motivates the investigations. The Nuclear Regulatory Commission, reactor vendors and power utilities aggressively investigate potential pressurized thermal shock failures in pressurized water reactors. These investigations date from 1974. A number of specific programs, for example Bryan et al. [1988], and summary studies, for example Cheverton et al. [1988], conclude that designs for thermal shock can be based on the stress intensity factor with proper allowance for the observed scatter in crack initiation and arrest values. A true thermomechanical energy parameter such as **S** might provide for a more effective design of critical components.

Within the automobile industry since the 1970's, the fatigue failures of exhaust valves and valve stems represent one of the major failure mechanisms for the internal combustion engine (Vitcha [1973]). Today, as fuel economy dictates lighter vehicles, thermal fatigue of many engine components must be considered in design.

The cracking of hot glasses or dishware due to incidental minor impacts, the shattering of hot plates from splashing with cold water and the fracture of stemmed glassware after repeated dishwasher cycles all represent common household examples of fracture and crack growth under thermomechanical loadings.

Thus, thermomechanical loadings are prevalent throughout various industries. The reliability of a critical structure or component often depends on crack

behavior under general thermoinelastic material response. The availability of a single parameter characterizing the crack driving force for thermoinelastic fracture and crack growth will enable engineers to more accurately account for general thermomechanical loadings thus permitting more economical designs. The S integral developed in this dissertation provides such a parameter.

1.3 Study Scope and Limitations

The goal of this investigation is to formulate, verify and calculate the path domain independent integral, S, that characterizes the crack driving force for thermoinelasticity. Though by no means exhaustive, this study suggests that S may provide the needed energy based insight into thermoinelastic crack behavior that designers require.

The second chapter discusses the state of the art in characterizing parameters for thermomechanical fracture mechanics. Beginning with the J integral, Chapter 2 reviews a few of the most common energy terms referenced in the literature that address fracture criteria, especially for inelastic material response regimes. After a brief recap of the various forms of the J integral, the chapter discusses those parameters that characterize creep fracture and finally presents the few energy integrals proposed for thermomechanical fracture.

The third chapter formulates the path domain independent integral S for the quasi-static case. Based on thermoinelastic continuum mechanics and a material conservation law developed from classical field theory, S equals the force acting on a singularity. Thus S defines the total energy released per unit crack advance. Though the method for generating S applies to dynamic and quasi-static regimes, this effort only addresses the quasi-static case. This development draws heavily on the work of Simo and Honein [1990] on discrete conservation laws for inelasticity.

Chapter 3 concludes by defining S for the quasi-static case including body forces for both coupled and uncoupled thermoinelasticity. The coupled case assumes that the temperature field induces a strain field and that the quasi-static strain rates induce a temperature field. The uncoupled case employs the typical assumptions for metals, i.e., the temperatures induce strain but slowly varying strains do not induce temperature change. Later chapters show this uncoupled assumption to produce satisfactory results in evaluating test data for aluminum specimens. The inelasticity theory uses internal variables to model elastic-viscoplastic response with combined isotropic and kinematic hardening.

Chapter 4 describes the experimental investigation, confirms that S represents a thermoinelastic material conservation law and discusses the results from the fracture toughness tests. Limited tests on aircraft 2024 aluminum demonstrate the detrimental effect a thermal gradient has on the load carrying capacity of a member. For the quasi static case, approximately a 65°/F per inch thermal gradient reduces the applied tension to cause initial crack extension by approximately fifty percent. This experimental investigation does not address fatigue crack growth from cyclic thermomechanical loading. This important phenomenon, crucial for successful engineering design, provides an interesting topic for further research.

The fifth chapter presents the numerical calculation of S via the finite element method. A postprocessor to the general finite element analysis program, FEAP, computes the S path domain independent integral. Investigating the finite element based S integral on a crack free body subjected to thermomechanical loading demonstrates the conservation law nature of S. Computations model eight selected fracture toughness experiments showing the S integral characterizes fracture under thermomechanical loading. The S integrals computed from the finite element analyses match those estimated from the experimental force displacement

trace for the isothermal cases verifying the **S** calculation. Furthermore, the **S** integrals calculated for the thermomechanical loading cases agree with the values from the isothermal cases suggesting that **S** characterizes incipient fracture.

The sixth chapter summarizes the investigation recapitulating the key points of **S** integral formulation plus experimental and computational verification. Next, the final chapter discusses the limitations of the **S** path domain independent integral. The chapter concludes by suggesting future research in the crucial area of thermoinelastic fracture mechanics.

2

Background on Fracture Criteria

Energy based fracture criteria hold a highly regarded place in fracture mechanics. These criteria address the macroscopic thermodynamic energy balance associated with extending a crack. Path or path domain integrals typically define the critical parameters for fracture. Over the years the criteria have developed from the simplest case, including only the material surface energy, to the present effort that includes the full thermoinelastic energy fields.

2.1 Early Energy Concepts

Griffith in his classic works [1921,1925] pioneered energy based fracture criteria. Under elastic conditions, Griffith suggests that crack propagation occurs if the elastic energy released upon crack extension provides sufficient energy for crack growth. Hence, the elastic energy release rate, **G**, (also called the crack driving force per unit crack extension) required for crack growth can be expressed as,

$$\mathbf{G} = \frac{dU}{da} = \frac{dW_g}{da} = R \tag{2.1}$$

where a is the crack length. U is the elastic energy, W_g is the energy required for crack growth, and R is the resistance per unit crack advance.

Since Griffith studied fracture of glass, his criterion theorizes the fracture resistance simply equals the surface energy required to form two new surfaces as the crack extends a unit distance. Hence, Griffith set the resistance to twice the surface energy, γ.

$$R = 2\gamma. \tag{2.2}$$

This criterion requires knowing the surface energy, γ, of the material. While providing reasonable accuracy for brittle fracture, the Griffith criterion proves inadequate for ductile fractures that commonly occur in metals.

Most critical engineering structures in the first half of this century failed by ductile fracture of metal components. Irwin [1948] and Orowan [1955] independently noted that the energy required for unit crack advance in metal far exceeds the surface energy of the two new free surfaces. They included an empirical factor within the Griffith criterion to account for the plastic energy at the crack tip associated with ductile fracture. Denoting this plasticity factor as R^{plas}, the Irwin criterion states

$$R = 2\gamma + R^{\mathrm{plas}}. \tag{2.3}$$

While conceptually correct, the Irwin ductile fracture criterion proves ineffective for design due to difficulties in obtaining the plastic energy term. This empirical correction term depends on the material, component and crack geometry as well as the loading history.

Following the efforts of Griffith and Irwin, researchers have developed energy based fracture parameters that apply to arbitrary geometries and loadings. These energy criteria often appear as path and path domain independent integrals. Herein, path independent integral refers to an integral expression that has the same value regardless of the integration path (a line integral for two dimensions or a surface integral for three dimensional cases). Likewise, path domain independent integral

refers to the sum of a path integral plus an integral over the enclosed domain (line and area integral for two dimensions or surface and volume for three dimensions) that has the same value regardless of the choice of path.

The following sections briefly review some of the integral parameters discussed in the literature for fracture mechanics. The review begins by defining and describing the **J** integral then addresses some of the parameters proposed to overcome various restrictions on **J**. Some of the integrals carry an energy release rate interpretation, while others only represent a convenient parameter that characterizes crack extension.

The discussions draw on the original cited references and the insightful comparison articles of Kim and Orange [1988], Kanninen and Popelar [1985], Hellen and Blackburn [1986] and Atluri [1986]. While less than exhaustive, this review captures the flavor of the integral parameters, briefly discusses their backgrounds and highlights certain limitations.

One final note on energy based fracture criteria concludes this introduction. The integral parameters discussed herein are typically based on a continuum mechanics description of the crack tip fields. Since at the crack tip such noncontinuum effects as void initiation, microcracking and microvoid coalescence occur as described by Broberg [1971], the continuum based energy criteria can only be a necessary but not sufficient condition for fracture. Given the necessary energy condition, fracture will occur only if the material *exactly* at the crack tip is ready to fail and become two separate pieces under the acting stresses and strains. Fortunately, for most materials discussed in the literature the noncontinuum fracture process zone adds only a negligible energy to the total system. Thus, the continuum based parameters accurately predict fracture.

2.2 The J Integral

One of the most widely researched topics over the past twenty years in analytical fracture mechanics has been the J integral. This path independent integral fracture parameter has allowed impressive advances in elastic plastic fracture mechanics through J_R fracture resistance curves.

Independently developed and popularized for fracture mechanics by Rice [1968], though formulated earlier in the works of Eshelby [1956], Sanders [1960] and Cherapanov [1961], the J integral equals the energy release per unit crack length for a unit extension in an elastic material. The following expression defines J for elastostatics with no body forces and homogeneous material,

$$\mathbf{J} = \int_{\Gamma} \{W\boldsymbol{n} - (\nabla \boldsymbol{u})^T(\boldsymbol{\sigma}\boldsymbol{n})\}\, d\Gamma. \tag{2.4}$$

where W is the strain energy density, $\boldsymbol{n}$ is the outward unit normal to Γ, Γ is a simple curve enclosing the crack tip, see Figure 2.1, $\boldsymbol{\sigma}$ is the stress tensor and $\boldsymbol{u}$ is the displacement vector.

This vector representation of J follows Eshelby's [1956] original expression for the force on a singularity inside Γ. Knowles and Sternberg [1972] show that this vectorized J represents a conservation law for self similar crack extension.

The J integral exactly equals the negative of the change in potential energy per unit crack advance (see Rice [1968], Eshelby [1975] and Budiansky and Rice [1973]). As such, J characterizes the driving force for crack extension.

For a body loaded only into the linear elastic regime, the J integral component in the direction of crack extension, $\mathbf{J}_x$, identically equals the Griffith elastic energy release rate, G. The energy release rate, in turn, can be expressed as a function of the mode I (opening) stress intensity factor, $\mathbf{K}_I$, (see Broek[1987] for example),

$$\text{linear elastic plane stress} \qquad \mathbf{J}_x = \mathbf{G} = \frac{\mathbf{K}_I^2}{E}\ , \tag{2.5a}$$

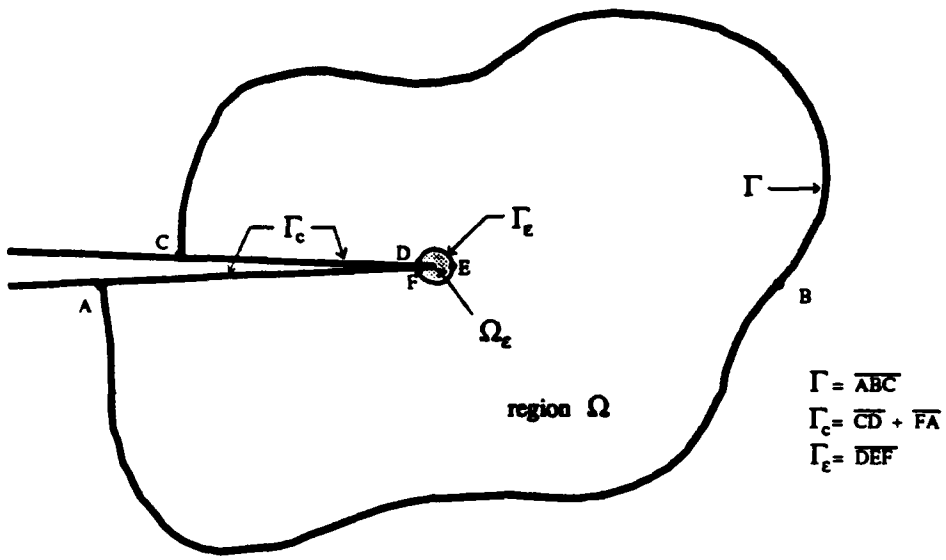


Figure 2.1 *Crack region for fracture integrals.*

$$\text{linear elastic plane strain} \qquad J_x = G = \frac{K_I^2}{E}(1-\nu^2) \ , \qquad (2.5b)$$

where E is the Young's modulus and ν is Poisson's ratio.

The theoretical foundations of **J** remain valid for nonlinear elasticity. Thus, the **J** integral addresses elastic plastic fracture for proportional loading since for this loading case the material response can be accurately described by a nonlinear elastic stress strain relation.

The potential energy change definition of **J** permits **J** to be evaluated experimentally from the load vs deflection curves for cracked specimens. One calculation method (Begley and Landes [1972]) examines the change in area under the load vs deflection curve up to some specified deflection for numerous samples with slightly different initial crack lengths. Other calculation methods determine **J** from a single sample (Rice et al. [1973], Landes et al. [1989] and ASTM E813 [1988]) in which the crack is carefully advanced. These methods separate the calculation into elastic

and plastic (nonlinear elastic to be precise) portions,

$$J_I = J_I^{elas} + J_I^{plas}, \tag{2.6}$$

where J_I^{elas} is the elastic energy release rate given by equation (2.5) and J_I^{plas} is determined from the load vs deflection curve. Chapter 4 presents and illustrates two procedures for calculating J_I.

The J integral can also be calculated from finite element analyses. The path independence of the J integral permits its evaluation along contours removed from the intricate stress and strain fields in the immediate crack tip vicinity. For example, the early study by Kobayashi et al. [1973] shows excellent agreement between the critical value of J at incipient fracture calculated from experiments and generated by finite element analysis.

Interestingly, although elasticity provides the foundation for the J integral, J addresses elastic plastic fracture mechanics under certain limited conditions. Proportional loading is the crucial condition. Under proportional loading the stain energy density of an elastic plastic material remains single valued in strain and identically equals the strain energy density of a nonlinear elastic material. Hence, the J integral characterizes ductile fracture under proportional loading of an elastic plastic material.

The theoretical foundation for J becomes invalid for incremental plasticity since the strain energy density ceases to be single valued during any unloading. Certainly, some unloading occurs in the material directly behind an extending crack tip. The J integral does not address plastic unloading, incremental plasticity, viscous material behavior, body forces, arbitrary material inhomogeneity or thermal strains.

Many critical structures and components under severe loadings experience accumulating inelastic strains due to cyclic loadings and/or increasing temperatures.

Since these structures and components require reliability evaluation for overall safety and economy and J integral analysis cannot be applied, many fracture mechanics parameters address one or more of the aspects limiting the J integral.

Lamba [1975] proposes a cyclic definition of J to apply to fatigue crack initiation from a notch. The fatigue cyclic parameter ΔJ is given as,

$$\Delta J = \int_{\Gamma} \{\Psi(\Delta\epsilon)\boldsymbol{n} - (\nabla\Delta\boldsymbol{u})^T(\Delta\boldsymbol{\sigma}\boldsymbol{n})\}\, d\Gamma\,, \tag{2.7}$$

where $\Delta\boldsymbol{u}, \Delta\epsilon,$ and$\Delta\boldsymbol{\sigma}$ are increments of displacement, strain and stress and $\Psi(\Delta\epsilon)$ is the strain potential such that $\Delta\boldsymbol{\sigma} = \partial\Psi/\partial\Delta\epsilon$.

The ΔJ integral retains the path independence property and the theoretical limitations of the J integral discussed in this section. Compared to Neuber [1961] notch analysis the ΔJ parameter yields improved results as the notch becomes sharper for predominantly elastic loads. Recent studies, Huang and Pelloux [1980], Hatanaka et al. [1989] and Jablonski [1989], report good correlation of low cycle fatigue crack growth rates, da/dN, where a is the crack length and N is the number of cycles with ΔJ on Hastelloy-X, medium carbon steel, 304 stainless steel and HY 100 alloy. These studies demonstrate ΔJ's usefulness for fatigue analyses regardless of the lack of theoretical foundation.

2.3 Complementary Energy Version of J

Another path independent integral energy parameter for elastostatics is the I integral developed by Bui [1974]. The I integral represents the dual of the J integral since I considers the complementary energy of a cracked body. Bui [1974] defines I for two dimensional elastostatics with homogeneous material in the absence of body forces as,

$$\boldsymbol{I} = \int_{\Gamma} \{-U^B(\boldsymbol{\sigma})\boldsymbol{n} + \boldsymbol{u}^T((\nabla\boldsymbol{\sigma})\boldsymbol{n})\}\, d\Gamma\,, \tag{2.8}$$

where $U^B(\sigma)$ defines the stress function energy such that the strain is given by $\epsilon = \partial U^B/\partial\sigma$.

The I integral identically equals the negative of the stress potential energy change with increasing crack length.

$$I = -\frac{dQ}{da}, \tag{2.9}$$

$$Q = -\int_\Omega U^B(\sigma)d\Omega + \int_{\partial\Omega_u} \boldsymbol{u}^d(\sigma\boldsymbol{n})\,d\Gamma, \tag{2.10}$$

where $\boldsymbol{u}^d$ are the specified displacements along the boundary $\partial\Omega_u$. For the case where convex functions define both the strain energy and the stress energy then I identically equals J since the stress potential equals the strain potential.

Similar to the experimental determination of J, the load displacement curve from a fracture test determines I. The accuracy of finite element based numerical evaluation of I increases for hybrid or stress based finite element formulations. The traditional displacement based finite elements yield less accurate stress gradients than displacement gradients and therefore is not computes as accurately as J.

2.4 Two Definitions of the J* Integral

Attempting to overcome the proportional loading restriction imposed on J, Blackburn [1972] proposed the J* integral defined as (see Figure 2.1),

$$\mathrm{J}^* = \int_{\Gamma_\epsilon} \{\frac{1}{2}(\sigma : \nabla\boldsymbol{u})\boldsymbol{n} - (\nabla\boldsymbol{u})^T(\sigma\boldsymbol{n})\}\,d\Gamma \ , \tag{2.11}$$

or equivalently via Green's theorem and elastostatic equilibrium in the absence of body forces,

$$\begin{aligned}\mathrm{J}^* = & \int_{\Gamma+\Gamma_\epsilon} \{\frac{1}{2}(\sigma : \nabla\boldsymbol{u})\boldsymbol{n} - (\nabla\boldsymbol{u})^T(\sigma\boldsymbol{n})\}\,d\Gamma \\ & + \lim_{\epsilon\to 0}\int_{\Omega-\Omega_\epsilon} \{\frac{1}{2}(\sigma : \nabla^2\boldsymbol{u}) - \frac{1}{2}(\nabla\sigma : \nabla\boldsymbol{u})\}\,d\Omega \ .\end{aligned} \tag{2.12}$$

For an elastic material the area integral vanishes and $\mathbf{J}^*$ identically equals $\mathbf{J}$. However, for the antiplane shear problem and for the Dugdale (Tresca yield based) plane stress problem using infinitesimal strains, $\mathbf{J}^*$ equals $\mathbf{J}/2$. Furthermore, for power law hardening materials $\mathbf{J}^*$ does not support an energy release rate interpretation. Using the explicit term $\boldsymbol{\sigma} : \nabla \boldsymbol{u}$ expands the scope of $\mathbf{J}^*$ beyond the limitations of $\mathbf{J}$ but obscures the physical meaning.

Blackburn [1972] suggests using $\mathbf{J}^*$ as a fracture parameter for those nonelastic cases for which $\mathbf{J}$ is inappropriate, such as cases including inelastic load reversals. Comparing a critical value of $\mathbf{J}^*$ calculated from another known case to the situation in question determines the likelihood of fracture.

Finite element results provide the necessary information for calculating $\mathbf{J}^*$. Since the area integral in equation (2.12) vanishes in the elastic region, only the inelastic elements (gauss point areas) enter the area integral term. Numerically investigating $\mathbf{J}^*$ for cases containing creep, unloading from an inelastic state, thermal strains and material inhomogeneity, Blackburn et al. [1977] find the expected result, that $\mathbf{J}^*$ remains path independent over various contours.

Supporting $\mathbf{J}^*$ as a fracture parameter, Batte et al. [1983] calculate a critical stress intensity factor using the $\mathbf{J}^*$ integral from finite element models of fracture toughness tests on a 1% CrMoV IP steam turbine rotor forging. The computed stress intensity factors based on $\mathbf{J}^*$ show little dependence on crack length or contour choice and fall within +0% to +20% of the experimentally determined values.

Blackburn [1985] specifically investigates using $\mathbf{J}^*$ and $\Delta\mathbf{J}^*$ for cyclic loading. The change of $\mathbf{J}^*$ with incremental crack length increase defines $\Delta\mathbf{J}^*$. The parameters accurately determine the cumulative and current crack tip state and are recommended by Blackburn for the unloading and reloading situation. In this respect $\mathbf{J}^*$ appears robust as a fracture parameter.

Thermal strains enter $\mathbf{J}^*$ as defined in equation (2.12) directly through the displacement gradient. This causes $\mathbf{J}^*$ to depend unreasonably on an isothermal temperature change. Apparently to alleviate this potential problem, Hellen and Blackburn [1986] propose redefining $\mathbf{J}^*$ in the presence of a temperature field as,

$$\mathbf{J}^* = \int_{\Gamma_\epsilon} \{\frac{1}{2}(\boldsymbol{\sigma} : (\nabla \boldsymbol{u} - \boldsymbol{\epsilon}^{th}))\boldsymbol{n} - (\nabla \boldsymbol{u})^T(\boldsymbol{\sigma}\boldsymbol{n})\}\, d\Gamma \ , \tag{2.13}$$

where $\boldsymbol{\epsilon}^{th}$ equals the thermal strain tensor. Unfortunately, little discussion and no examples accompany this recommendation.

Schmitt and Kienzler [1989] modify the **J** integral in a manner quite similar to $\mathbf{J}^*$ as defined by Blackburn [1972] to apply to materials described by incremental plasticity. They define the new path domain independent integral as,

$$\mathbf{J}^s = \int_\Gamma \{W^s \boldsymbol{n} - (\nabla \boldsymbol{u})^T(\boldsymbol{\sigma}\boldsymbol{n})\}\, d\Gamma + \int_\Omega \nabla \boldsymbol{\sigma} : \nabla \boldsymbol{u}\, d\Omega \ . \tag{2.14}$$

where W^s is the stress work density defined as, $\boldsymbol{\sigma} = \partial W^s / \partial \nabla \boldsymbol{u}$. Hence, W^s depends on the deformation history. For an elastic material $\mathbf{J}^s$ identically equals **J**. For incremental plasticity the principle of virtual work leads to the interpretation of $\mathbf{J}^s$ as a work dissipation rate akin to the energy release rate interpretation of **J** for elasticity.

Defining $\mathbf{J}^s$ in terms of the history dependent stress work density provides an unambiguous evaluation of the parameter. With this definition, $\mathbf{J}^s$ applies to incremental plasticity, nonproportional loadings and elastic unloadings. However, thermal strains again enter through the displacement gradient and stress thus potentially leading to erroneous conclusions.

Recently, in their work on discrete conservation laws and path domain independent integrals, Simo and Honein [1990] generalize the classical **J** integral to discrete incremental plasticity and viscoplasticity. Constructing the Noether

[1918,1971] quantity associated with the translation symmetry group for the discrete (algorithmic) Lagrangian. Simo and Honein [1990] develop the path domain independent integral,

$$\mathbf{J}^{*(SH)} = \int_{\Gamma} \Big\{ \big(W(\boldsymbol{\epsilon}_{n+1} - \boldsymbol{\epsilon}^{vp}_{n+1}) + \boldsymbol{q}_{n+1} : \boldsymbol{D}^{-1} : (\frac{1}{2}\boldsymbol{q}_{n+1} - \boldsymbol{q}_n) + \boldsymbol{\sigma}_{n+1} : (\boldsymbol{\epsilon}^{vp}_{n+1} - \boldsymbol{\epsilon}^{vp}_n) - \frac{\Delta t}{\eta}\gamma^{+}(f(\boldsymbol{\sigma}_{n+1}, \boldsymbol{q}_{n+1}))\big)\mathbf{1} - \nabla \boldsymbol{u}^{T}_{n+1}\boldsymbol{\sigma}_{n+1} \Big\} \boldsymbol{n}\, d\Gamma + \int_{\Omega} \big\{ \boldsymbol{\sigma}_{n+1} : \nabla \boldsymbol{\epsilon}^{vp}_n - \boldsymbol{q}_{n+1} : \boldsymbol{D}^{-1} : \nabla \boldsymbol{q}_n \big\}\, d\Omega \,, \tag{2.15}$$

for viscoplasticity where W is the stored energy function such that the stress $\boldsymbol{\sigma} = \partial W/\partial \boldsymbol{\epsilon}$. $\boldsymbol{\epsilon}$ is the total strain tensor, $\boldsymbol{\epsilon}^{vp}$ is the viscoplastic strain tensor, $\boldsymbol{q}$ is the tensor of hardening variables (back stress and isotropic hardening scalar), $\boldsymbol{D}$ is the positive definite hardening moduli tensor, Δt is the time step for the discrete interval $[n, n+1]$, η is the fluidity parameter, γ^{+} is the penalty function for viscoplastic regularization and $f(\boldsymbol{\sigma}, \boldsymbol{q})$ is the yield function. The subscripts n and $n+1$ denote the variable values at the beginning (n) or end ($n+1$) of the interval. For plasticity the plastic strain tensor $\boldsymbol{\epsilon}^{p}$ replaces $\boldsymbol{\epsilon}^{vp}$ and the yield function obeys $f(\boldsymbol{\sigma}, \boldsymbol{q}) \leq 0$.

This version of the $\mathbf{J}^{*}$ integral, denoted as $\mathbf{J}^{*(SH)}$, equals the total energy released per unit crack extension. Simo and Honein [1990] prove this using an argument following the Budiansky and Rice [1973] proof that the classic J integral equals the potential energy release rate.

This $\mathbf{J}^{*(SH)}$ path domain independent integral extends the classical J integral to nonproportional loading cases treated by incremental plasticity and viscoplasticity. The discrete Lagrangian and discrete conservation law work of Simo and Honein [1990] forms the basis for the S integral development in Chapter 3.

the work of Simo and Honein [1990] on discrete conservation laws for inelasticity.

2.5 Creep Crack Growth with Integral Parameters

High temperature creep behavior of components such as power generating turbine blades and rotors violate the limitations on the **J** integral. A parameter that uniquely characterizes the crack tip stress and strain rate fields during creep response provides an improved correlation of creep crack growth rates compared to the linear elastic stress intensity factor. This section discusses two such parameters, $\mathbf{C}^*$ and $\mathbf{F}$, while other more general parameters proposed by Stonesifer and Atluri [1982a,b] and Brust et al. [1985] are discussed in the next section.

For nonlinear steady state (secondary) creep, the $\mathbf{C}^*$ parameter popularized by Landes and Begley [1976] (following a suggestion by Goldman and Hutchinson [1975]) characterizes the crack tip stress and strain rate fields. The rate form of the **J** integral defines $\mathbf{C}^*$ as,

$$C^* = \int_\Gamma \{W^* \boldsymbol{n} - (\nabla \dot{\boldsymbol{u}})^T (\boldsymbol{\sigma} \boldsymbol{n})\}\, d\Gamma, \tag{2.16}$$

where W^* is the strain energy rate density defined as,

$$W^* = \int_0^{\dot{\boldsymbol{\epsilon}}} \boldsymbol{\sigma}\, d\dot{\boldsymbol{\epsilon}}. \tag{2.17}$$

The $\mathbf{C}^*$ integral relation retains the path independence and all the limitations of the **J** integral. Also, $\mathbf{C}^*$ equals the power difference between two identically loaded bodies having incrementally different crack lengths. Similar to **J**, $\mathbf{C}^*$ can be evaluated from the load vs displacement rate trace of steady state creep fracture tests. Naturally, $\mathbf{C}^*$ can also be evaluated from finite element results.

In the study by Landes and Begley [1976], the $\mathbf{C}^*$ parameter successfully correlates steady state creep crack growth rates on center cracked panels of Fe-Ni-Cr superalloy (discaloy). Using $\mathbf{C}^*$ reduces the scatter in crack growth rate by a factor of six compared to using the stress intensity factor. Nikbin et al. [1976]

report similar findings for creep crack growth tests conducted on aluminum alloy RR58 and a chromium-molybdenum-vanadium steel.

In a creative application of the technique Eshelby [1956] uses to generate the force on a singularity, Atkinson and Smelser [1982] develop an invariant integral **F** for coupled time dependent thermoviscoelastic solids. For special cracked strip problems this integral yields the crack tip stress and displacement fields.

The integral results from the tensor akin to the energy momentum tensor (see Eshelby [1956]) but with the Lagrangian constructed from the Laplace transform of the coupled equations of motion. As such, the integral generates quantities such as the Laplace transformed stress, displacement, temperature and stress intensity factor. Naturally, except for certain problems the procedure becomes algebraically complicated. Furthermore, results in terms of the Laplace transforms require some effort to convert to useful real time results.

2.6 The ΔT Integral Family

Atluri [1982] develops incremental integrals for plasticity, $\Delta\mathbf{T}$, and rate sensitive inelasticity, $\Delta\mathbf{T}_c$. These integrals generalize the conservation law for self similar crack extension of Knowles and Sternberg [1972], and hence the **J** integral, to include finite deformations, inelasticity, body forces, material inertia and arbitrary crack face conditions. Kanninen and Popelar [1985] observe that thermal effects can be easily accommodated by modifying the body force to include the term $-\alpha E\nabla\theta/(1-2\nu)$ where α is the thermal expansion coefficient and θ is the temperature change. Later modifications to $\Delta\mathbf{T}$ by Atluri et al. [1984] produce the incremental integrals $\Delta\mathbf{T}_\mathbf{p}^*$ and $\Delta\mathbf{T}_\mathbf{p}$. Summing $\Delta\mathbf{T}$ over the history yields the $\mathbf{T}^*$ parameter investigated by Brust et al. [1985].

Beginning from the conservation law, Atluri [1982] develops $\Delta \mathbf{T}$ for the general case of classical elastoplasticity. The $\Delta \mathbf{T}$ parameter equals the difference in combined incremental potential energy per unit crack length in the time interval t to $t + \Delta t$ of two cracked bodies identical in shape and load history differing only slightly in crack length. Though developed for finite deformations, for consistency within this review attention is restricted to the infinitesimal strain theory. The infinitesimal strain version of $\Delta \mathbf{T}$ becomes,

$$\Delta \boldsymbol{T} = \int_{\Gamma} \{(\Delta W)\boldsymbol{n} - (\Delta(\nabla \boldsymbol{u}))^{T}((\boldsymbol{\sigma} + \Delta \boldsymbol{\sigma})\boldsymbol{n})\}\, d\Gamma$$

$$- \int_{\Omega} \{\nabla \boldsymbol{\sigma} : \Delta \boldsymbol{\epsilon} + \rho \boldsymbol{b} \bullet \Delta \boldsymbol{\epsilon}\}\, d\Omega\,, \tag{2.18}$$

where ρ is the material density, $\boldsymbol{b}$ is the body force per unit mass, ΔW is the increment in total stress working density in the time interval from t to $t + \Delta t$ defined by,

$$\Delta W = (\boldsymbol{\sigma} + \frac{1}{2}\Delta \boldsymbol{\sigma}) : \Delta \boldsymbol{\epsilon}\,. \tag{2.19}$$

The $\Delta \mathbf{T}$ integral measures the severity of the crack tip stress and strain fields including the effects of near-tip plasticity and loading/unloading zones. For paths entirely in either the loading or unloading zone, Atluri [1982] proves the path domain independence of $\Delta \mathbf{T}$.

Similar to $\Delta \mathbf{T}$, the parameter $\Delta \mathbf{T}_c$ addresses rate sensitive inelastic materials. For elasto-viscoplasticity the $\Delta \mathbf{T}_c$ parameter has the same energy interpretation and the same path domain independence as $\Delta \mathbf{T}$. Actually, the definition of $\Delta \mathbf{T}_c$ matches $\Delta \mathbf{T}$ with the stress terms and the incremental stress working density correctly defined for an elasto-viscoplastic material.

The parameter $\Delta \mathbf{T}_c$ addresses nonsteady creep crack growth. Stonesifer and Atluri [1982a,b] investigate $\Delta \mathbf{T}_c$ based on finite element results and compare it to

the steady state creep parameter $\mathbf{C}^*$. The comparison with the $\mathbf{C}^*$ integral requires a steady state version of the incremental parameter $\Delta\mathbf{T}_c$ for mode I crack growth defined as $\dot{\mathbf{T}}_c$,

$$\dot{T}_c = \lim_{\Delta t \to 0} \frac{\Delta T_c}{\Delta t} \approx \frac{\Delta T_c}{\Delta t} \quad . \tag{2.20}$$

The comparisons conclude that $\dot{T}_c$ accurately characterizes mode I (opening) crack growth for nonsteady as well as steady state creep. The $\dot{T}_c$ parameter has an energy interpretation akin to $\Delta\mathbf{T}$ and can be easily calculated from numerical models. For steady state creep in plane stain $\dot{T}_c$ remains within two percent of $\mathbf{C}^*$ and falls eleven to fourteen percent above $\mathbf{C}^*$ for plane stress.

Atluri et al. [1984] present two modified versions of the $\Delta\mathbf{T}$ integral for incremental plasticity. The new integrals, denoted as $\Delta\mathbf{T}_p^*$ and $\Delta\mathbf{T}_p$ express the elastic plastic boundary contribution in terms of field variables. The following expressions define these new parameters.

$$\begin{aligned}\Delta T_p^* = &\int_{\Gamma+\Gamma_\epsilon} \{\Delta W \boldsymbol{n} - \Delta(\nabla \boldsymbol{u})^T(\boldsymbol{\sigma} + \Delta\boldsymbol{\sigma})\boldsymbol{n} - (\nabla \boldsymbol{u})^T \Delta\boldsymbol{\sigma}\boldsymbol{n}\}\, d\Gamma \\ &+ \int_{\Omega-\Omega_\epsilon} \{\Delta\boldsymbol{\sigma} : (\nabla\boldsymbol{\epsilon} + \frac{1}{2}\Delta(\nabla\boldsymbol{\epsilon})) - \Delta\boldsymbol{\epsilon} : (\nabla\boldsymbol{\sigma} + \frac{1}{2}\Delta(\nabla\boldsymbol{\sigma}))\}\, d\Omega, \end{aligned} \tag{2.21}$$

$$\begin{aligned}\Delta T_p = &\int_{\Gamma+\Gamma_\epsilon} \{\Delta W \boldsymbol{n} - \nabla(\Delta \boldsymbol{u})^T(\boldsymbol{\sigma} + \Delta\boldsymbol{\sigma})\boldsymbol{n} - (\nabla \boldsymbol{u})^T \Delta\boldsymbol{\sigma}\boldsymbol{n}\}\, d\Gamma \\ &+ \int_{V_b-\Omega} \{\Delta\boldsymbol{\epsilon} : (\nabla\boldsymbol{\sigma} + \frac{1}{2}\Delta(\nabla\boldsymbol{\sigma})) - \Delta\boldsymbol{\sigma} : (\nabla\boldsymbol{\epsilon} + \frac{1}{2}\Delta(\nabla\boldsymbol{\epsilon}))\}\, d\Omega \\ = &\int_S \{\Delta W \boldsymbol{n} - \nabla(\Delta \boldsymbol{u})^T(\boldsymbol{\sigma} + \Delta\boldsymbol{\sigma})\boldsymbol{n} - (\nabla \boldsymbol{u})^T(\Delta\boldsymbol{\sigma})\boldsymbol{n}\}\, d\Gamma\,, \end{aligned} \tag{2.22}$$

where V_b is the total volume of the body and S is the external boundary including the crack faces.

The integral $\Delta\mathbf{T}_p^*$ includes the volume between the contours Γ and Γ_ϵ whereas the $\Delta\mathbf{T}_p$ integral includes the entire volume less the volume inside Γ. The parameter

$\Delta \mathbf{T}_p$ can be experimentally evaluated for proportional loading by measuring the incremental area between load vs displacement curves for two identical specimens with only slightly different crack lengths. Directly measuring the displacement and traction data along the entire specimen boundary also generates $\Delta \mathbf{T}_p$ via equation (2.21). Atluri et al. [1984] suggest using a mixed experimental (to determine $\Delta \mathbf{T}_p$) and computational (to obtain $\Delta \mathbf{T}_p^*$ from $\Delta \mathbf{T}_p$ plus a volume integral) program to determine fracture potential.

The $\Delta \mathbf{T}_p^*$ integral measures the severity of the deformation at the crack tip. During proportional loading $\Delta \mathbf{T}_p$ measures the immediate crack tip fields and equals the rate of change of incremental energy per unit crack advance. For nonproportional loading neither $\Delta \mathbf{T}_p^*$ nor $\Delta \mathbf{T}_p$ have clear physical meanings.

For the limiting case of nonlinear quasi static in a homogeneous material elasticity the parameters reduce to the incremental $\Delta \mathbf{J}$ integral in the interval t to $t + \Delta t$.

$$\Delta T_p^* \quad = \quad \Delta T_p \quad = \quad \Delta \mathbf{J} \,. \tag{2.23}$$

$$\Delta \mathbf{J} = \int_\Gamma \{\Delta W \boldsymbol{n} - \nabla(\Delta \boldsymbol{u})^T(\boldsymbol{\sigma} + \Delta \boldsymbol{\sigma})\boldsymbol{n} - (\nabla \boldsymbol{u})^T(\Delta \boldsymbol{\sigma} \boldsymbol{n})\}\, d\Gamma. \tag{2.24}$$

Also, for proportional loading in elastostatics the sum of the integrals over the monotonic loading history identically equals the $\mathbf{J}$ integral.

$$\sum \Delta T_p^* \quad = \quad \sum \Delta \Gamma_p \quad = \quad \mathbf{J} \,. \tag{2.25}$$

A further addition to the $\Delta \mathbf{T}$ integral family sums the incremental integral $\Delta \mathbf{T}_p^*$ over the loading history. The summation, defined as $\mathbf{T}^*$, see Brust et al. [1985], serves as a fracture initiation condition similar to $\mathbf{J}_{Ic}$.

The summation parameter remains well defined for unloading and subsequent reloading. In fact, the $\mathbf{T}^*$ parameter computed from finite element fracture model of

A533B steel compact tension specimen by Brust et al. [1986] successfully predicts continued crack extension upon reloading following elastic unloading after crack extension. This example highlights the history dependence inherent in $\mathbf{T}^*$.

To address nonsteady creep crack growth, Brust and Atluri [1986] suggest a refinement to the $\dot{T}_c$ parameter defined by equation (2.21). The new rate parameter, $\dot{T}^*$, characterizes creep fracture and has the near field form,

$$\begin{aligned} \dot{T}^* &= \lim_{\Delta t \to 0} \frac{\Delta T^*}{\Delta t} , \\ &= \lim_{\epsilon \to 0} \int_{\Gamma_\epsilon} \{\dot{W}\boldsymbol{n} - (\nabla \dot{\boldsymbol{u}})^T(\boldsymbol{\sigma} + \Delta\boldsymbol{\sigma})\boldsymbol{n} - (\nabla \boldsymbol{u})^T \dot{\boldsymbol{\sigma}}\boldsymbol{n}\} \, d\Gamma , \end{aligned} \tag{2.26}$$

where the $(\dot{\bullet})$ signifies the time derivative and $\dot{W}$ is the rate of stress working density. Numerical studies suggest that $\dot{T}^*$ accurately characterizes creep growth under nonsteady creep but not for pure power law creep where $\dot{T}^*$ becomes constant.

2.7 The $\hat{\mathbf{J}}$ Integral Includes Fracture Process Zone

Kishimoto et al. [1980] extend the **J** integral notion to address the fracture process zone, see Broberg [1971], which plays a crucial role in fracture. Within the small fracture process zone at the crack tip voids and microcracks initiate, grow and coalesce permitting crack extension. Classical continuum mechanics does not apply to these effects. However, the total system energy including the fracture process zone energetics must be balanced.

The proposed $\hat{\mathbf{J}}$ integral accounts for the fracture process zone and the effects of plastic deformations, body forces, thermal strains and material inertia. Denoting the rate of energy change in the fracture process zone by the path domain independent integral $\hat{\mathbf{J}}$. Kishimoto et al. [1980] define this parameter for elastostatics as,

$$\hat{\mathbf{J}} = -\int_{\Gamma_\epsilon} (\nabla \boldsymbol{u})^T(\boldsymbol{\sigma}\boldsymbol{n}) \, d\Gamma , \tag{2.27a}$$

$$\hat{\mathbf{J}} = \int_{\Gamma+\Gamma_\epsilon} \{W^e \boldsymbol{n} - (\nabla \boldsymbol{u})^T (\boldsymbol{\sigma} \boldsymbol{n})\}\, d\Gamma + \int_\Omega \{\boldsymbol{\sigma} : \nabla \boldsymbol{\epsilon}^* - \rho \boldsymbol{b} \nabla \boldsymbol{u}\}\, d\Omega \,. \tag{2.27b}$$

where W^e is the elastic strain energy density such that $\boldsymbol{\sigma} = \partial W^e / \partial \boldsymbol{\epsilon}^e$ and the strain permits the decomposition $\boldsymbol{\epsilon} = \boldsymbol{\epsilon}^e + \boldsymbol{\epsilon}^*$ where $\boldsymbol{\epsilon}^e$ is the elastic strain and $\boldsymbol{\epsilon}^*$ is the inelastic strain, typically thermal, plastic and/or viscoplastic components.

For the elastic case $\hat{\mathbf{J}}$ reduces to the **J** integral. The $\hat{\mathbf{J}}$ integral measures the work done on the crack tip by the surrounding material.

The $\hat{\mathbf{J}}$ integral development assumes that the fracture process zone remains autonomous during infinitesimal crack advance. Autonomous here means that the fracture process zone does not depend on crack geometry, body shape, loading condition; it remains constant in dimension and moves with the same speed as the crack tip. Hence, the relation $d\boldsymbol{u}/da = 0$, where $\boldsymbol{u}$ is the displacement and a is the crack length, holds on the contour Γ_ϵ.

Beginning from total system energy balance, Aoki et al. [1981b] show that $\hat{\mathbf{J}}$ equals the energy release rate associated with the translation of the fracture process zone. Here again, the restrictive assumptions on the fracture process zone outlined above apply to the development.

Using finite element calculations, Aoki et al. [1982] demonstrates the path domain independence of $\hat{\mathbf{J}}$ and claim that $\hat{\mathbf{J}}$ addresses incipient fracture in thermal stress fields. They suggest developing fracture resistance curves based on $\hat{\mathbf{J}}$ for thermally loaded structures.

2.8 Other Integral Parameters for Thermal Effects

A number of other path and path domain independent integrals address including thermal strains in the **J** integral. These integrals consider specific cases and hence do not include the general effects included in $\mathbf{J}^*$, $\hat{\mathbf{J}}$, and the ΔT integrals.

Ainsworth et al. [1978] first explicitly considered thermal strains in the path domain independent integral $\mathbf{J}^{\theta}$ for elastostatics defined as,

$$\mathbf{J}^{\theta} = \int_{\Gamma} \{W^{*\theta}\boldsymbol{n} - (\nabla \boldsymbol{u})^{T}(\boldsymbol{\sigma}\boldsymbol{n})\}\, d\Gamma + \int_{\Omega} \boldsymbol{\sigma} : \nabla \epsilon^{th}\, d\Omega\,, \tag{2.28}$$

where ϵ^{th} is the thermal strain tensor component of the total strain decomposition $\epsilon = \epsilon^{e} + \epsilon^{th}$, and $W^{*\theta}$ is the strain energy density defined as,

$$W^{*\theta} = \int_{0}^{\epsilon^{e}} \boldsymbol{\sigma}\, d\epsilon \quad \text{i.e.,} \quad \boldsymbol{\sigma} = \frac{\partial W^{\theta}}{\partial \epsilon^{e}}\,. \tag{2.29}$$

The $\mathbf{J}^{\theta}$ integral has the same energy release rate interpretation as $\mathbf{J}$. Similarly, $\mathbf{J}^{\theta}$ shares the shortcomings of $\mathbf{J}$ such as its foundation in nonlinear elasticity. Naturally, in the absence of thermal gradient $\mathbf{J}^{\theta}$ identically equals the $\mathbf{J}$ integral.

Numerical examples studied by Ainsworth et al. [1978] demonstrate the path domain independence of $\mathbf{J}^{\theta}$ and its relation to the stress intensity factor. Hence, $\mathbf{J}^{\theta}$ establishes a fracture load consistent with the critical potential energy release rate criterion.

Bass and Bryson [1983] include body forces in the $\mathbf{J}^{\theta}$ integral by redefining the volume integral. Their path domain independent integral $\mathbf{J}^{\delta}$ also equals the energy release rate per unit crack advance,

$$\mathbf{J}^{\delta} = \int_{\Gamma} \{W^{e}\boldsymbol{n} - (\nabla \boldsymbol{u})^{T}(\boldsymbol{\sigma}\boldsymbol{n})\}\, d\Gamma + \int_{\Omega} \{\boldsymbol{\sigma} : \nabla \epsilon^{th} - \rho \boldsymbol{b} \bullet \nabla \boldsymbol{u}\}\, d\Omega\,. \tag{2.30}$$

Wilson and Yu [1979] independently develop a path domain independent integral including thermal strains valid for isotropic, homogeneous linear elastic material free from body forces and inertia effects. They define their modification to the $\mathbf{J}$ integral as,

$$\mathbf{J}^{w} = \int_{\Gamma} \{W^{w}\boldsymbol{n} - (\nabla \boldsymbol{u})^{T}(\boldsymbol{\sigma}\boldsymbol{n})\}\, d\Gamma - \frac{E\alpha}{(1-2\nu)} \int_{\Omega} \{\frac{1}{2}\nabla(\theta \mathrm{tr}\epsilon) - \mathrm{tr}\epsilon\, \nabla\theta\}\, d\Omega\,. \tag{2.31}$$

where $W^w = \frac{1}{2}\boldsymbol{\sigma} : \boldsymbol{\epsilon}$ for the linear elastic material. The $\mathbf{J}^w$ integral exactly equals the $\mathbf{J}^\theta$ integral suitably restricted as discussed above.

Wilson and Yu [1979] check the path domain independence of $\mathbf{J}^w$ on numerical examples and generate stress intensity factors. Chen and Chen [1981] confirm that $\mathbf{J}^w$ addresses mixed mode fracture problems showing the relations between J_1^w, J_2^w and the mode I and mode II stress intensity factors.

Interestingly, McCartney [1979] discusses $\mathbf{J}^w$ in terms of thermodynamics showing that,

$$\boldsymbol{J}^w = \int_\Gamma \{\rho\Psi\boldsymbol{n} - (\nabla\boldsymbol{u})^T(\boldsymbol{\sigma}\boldsymbol{n})\}\, d\Gamma + \int_\Omega \rho S \nabla T_a \, d\Omega \,, \tag{2.32}$$

where Ψ is the Helmholtz free energy density, S equals the entropy and T_a is the absolute temperature. Nguyen [1981] and Germain et al. [1983] show that,

$$\mathbf{J}^w = -\frac{\partial \phi}{\partial a} \,, \tag{2.33}$$

where ϕ equals the global thermodynamic potential. Hence, as previously discussed, the path domain independent integral $\mathbf{J}^w$ carries a well founded physical energy interpretation.

For a very particular case of steady state temperature field, linear thermoelasticity with homogeneous isotropic material in the absence of body forces and material inertia, Gurtin [1979] generates a true path independent integral. Stated as a conservation law for a singularity free solid,

$$\begin{aligned}\boldsymbol{J}^G = \int_\Gamma \{ & W^G\boldsymbol{n} - (\nabla\boldsymbol{u})^T(\boldsymbol{\sigma}\boldsymbol{n}) - \frac{\alpha^2(3\lambda+2\mu)^2}{2(\lambda+\mu)}\,\theta^2\boldsymbol{n} \\ & + \frac{\alpha\mu(3\lambda+2\mu)}{(\lambda+\mu)}(\theta\nabla\boldsymbol{u} - \boldsymbol{u}\nabla\theta)\boldsymbol{n}\,\}\, d\Gamma = 0 \,,\end{aligned} \tag{2.34}$$

where $W^G = \mu\boldsymbol{\epsilon} : \boldsymbol{\epsilon} + \frac{1}{2}\lambda(\mathrm{tr}\boldsymbol{\epsilon})^2$, λ is the Lame parameter and μ is the shear modulus. In the absence of a thermal gradient ($\nabla\theta = 0$) the integral $\mathbf{J}^G$ reduces to the $\mathbf{J}$ integral since W^G includes temperature terms in θ.

The path independence of $\mathbf{J}^G$ depends on the quantities $\theta u_{1,2}$ and $\theta_{,2} u_1$ maintaining continuity across the crack face. Kim and Orange [1988] suggest changing the integration path from Γ to $\Gamma + \Gamma_c$ to avoid this sever restriction.

2.9 Summary of Integral Fracture Parameters

All of the parameters discussed in this chapter ranging from the simple Griffith fracture resistance to Atluri's $\Delta \boldsymbol{T}$ integrals seek to describe a single critical condition for crack advance. Table 2.1 summarizes the integral fracture parameters following the criteria discussed by Kim and Orange [1988]. The Table identifies the various conditions included in each parameter and highlights the physical meaning associated with each integral.

Many of the integrals afford the desirable interpretation of a potential energy release rate. These parameters, in principal, can be determined experimentally via a load vs displacement record of a fracture toughness test. For all the integrals listed on Table 2.1 finite element results supply the field information necessary to evaluate the path and path domain independent integrals.

For evaluations of critical structures and components, a true fracture parameter must correlate various types of crack propagation behavior. Crack advance must depend on the critical parameter value regardless of crack size and component geometry. An integral with a sound physical meaning, such as the energy release rate, that characterizes the crack tip severity appears most likely to satisfy these conditions. The **S** integral developed in Chapter 3 provides such a parameter.

All of the integrals discussed in this chapter characterize linear elastic fracture. Most also successfully address monotonic loading into the inelastic regime with negligible nonproportionality. However, for the incremental inelastic case including load cycling and accumulating inelastic strain, only the $\mathbf{J}^*$, $\hat{\mathbf{J}}$ and the $\Delta \boldsymbol{T}$ integral

Table 2.1
Fracture Integral Comparisons

Integral	Reference	Conditions Included: Thermal Strain	Body Forces	Material Inhomogeneity	Inelastic Regime: Prop. Load	Inelastic Regime: Nonprop. Load/Unload	Integral Computations (2-D Case)	Physical Meaning
J	Rice [1968]	no	no	no	yes	no	line	potential energy change
J	Lamba [1975]	no	no	no	yes	no	line	potential energy change
ΔI	Bui [1974]	no	no	no	yes	no	line	potential energy change
J^*	Blackburn [1972]	yes	no	yes	yes	yes	line + area	none [1]
J^*	Hellen and Blackburn [1986]	yes	no	yes	yes	yes	line + area	none
J^s	Schmitt and Kienzler [1989]	yes	no	yes	yes	yes	line + area	work dissipation rate
$J^{*(SIF)}$	Simo and Honein [1990]	no	yes	yes	yes	yes	line + area	total energy release
C^*	Landes and Begley [1976]	yes	no	yes	yes	no	line	power difference
ΔT	Atluri [1982]	no	yes	yes	yes	yes	line + area	incremental pot energy difference
ΔT_p^*	Atluri et al. [1984]	yes	no	yes	yes	yes	line + area	none
ΔT_p	Atluri et al. [1984]	yes	no	yes	yes	yes	line + area	none
T^*	Brust et al. [1985]	yes	no	yes	yes	yes	line + area	none
$\dagger^*$	Brust and Atluri [1986]	yes	no	yes	yes	yes	line + area	none
$\hat{J}$	Kishimoto et al. [1980]	yes	yes	yes	yes	yes	line + area	rate of work performed on tip
J^θ	Ainsworth et al. [1978]	yes	no	no	yes	no	line + area	thermodynamic potential change
J^δ	Bass and Bryson [1983]	yes	yes	no	yes	no	line + area	thermodynamic potential change
J^w	Wilson and Yu [1979]	yes	no	no	no	no	line + area	thermodynamic potential change
J^G	Gurtin [1979]	yes	no	no	no	no	line	thermodynamic potential change

Note: 1. none implies no clear physical meaning for nonproportional loading

family maintain their theoretical foundations. Unfortunately these integrals, except $\hat{\mathbf{J}}$, lose their physical meaning for this situation.

The $\hat{\mathbf{J}}$ integral appears to satisfy all requirements for a valid, useful thermoinelastic fracture parameter. However, since it does not relate to a change in some potential, it can not readily be evaluated experimentally. Since **S** emanates from the total Lagrangian, it equals the change in total energy per unit crack advance and characterizes the force acting on the crack tip.

The literature does not contain a single fracture integral parameter that maintains a sound physical interpretation while addressing the general case of thermoinelasticity. The S path domain independent integral developed and investigated in the subsequent chapters supplies a parameter that addresses the general thermoinelastic case. The S integral includes all the conditions listed on Table 2.1 and has the physical meaning of the change in total energy per unit crack advance.

3

Path Domain Independent Integral Formulation for Thermoinelasticity

The path domain independent **S** integral provides an energy based fracture parameter for thermoinelasticity. The parameter characterizes fracture under general thermoinelastic material response. The **S** integral equals the crack driving force at incipient fracture. Derived from the Noether quantity associated with the translation symmetry group for the discrete (algorithmic) Lagrangian density, the **S** integral carries the physical interpretation of an energy release rate. As such, the parameter also represents a conservation law for thermoinelastic fields.

The first section describes general conservation laws beginning with an illustrative case. Rigorous development of material conservation laws in classical field theory stems from applying Noether's theorem to the total Lagrangian (or Hamiltonian) defining the system. A brief discussion including a few relevant references and a simple example of Noether's theorem provides sufficient background for developing the conservation law most interesting in fracture mechanics. Eshelby's law [1956] calculates the force on an elastic singularity from the Noether quantity associated with the symmetry group of coordinate translations. This discussion and a simple elastic example establish the foundation for the **S** integral.

The subsequent development formulates the path domain independent integral S for the infinitesimal strain, quasi static case. The development follows the work of Simo and Honein [1990] and utilizes the time discretized field quantities ideal for later computational efforts. Both the uncoupled and coupled strain–temperature cases are presented. The uncoupled case assumes that a temperature field induces a strain field but the strain field does not influence the temperature. The absolute static case meets this assumption. Naturally, the coupled case includes the full strain–temperature interaction.

After formulating the S integral for linear thermoinelasticity, a proof demonstrates the integral's path independence. The physical interpretation of S as a total energy release per unit crack advance also follows directly from the formulation. A second proof succinctly provides the justification for this attractive attribute.

Finally, with suitable restrictions S degrades to the $\mathbf{J}^*$ integral for isothermal inelasticity, the J integral for elasticity and the $\mathbf{J}^\theta$ and the $\mathbf{J}^\delta$ integrals for uncoupled thermoelasticity.

3.1 Integral Conservation Law Formulation

Innumerable integral conservation laws exist within mechanics. Specifically for fracture mechanics many integral conservation laws relate the energy change per unit crack advance to a material dependent critical value for fracture. The previous chapter discusses some of these integral laws.

A quick digression removes much of the notational complexity masking the basic nature of integral conservation laws. Simply integrating any pointwise equality over any regular region produces an integral conservation law. Moreover, if the relation involves the divergence of a quantity then employing Green's theorem yields a path domain integral law.

Beginning with pointwise conservation of linear momentum for statics and integrating over some volume Ω produces a trivial path domain independent integral conservation law relating the tractions around any closed surface and the total body force within the surface,

$$\mathrm{div}\boldsymbol{\sigma} + \rho\boldsymbol{b} = 0\,, \tag{3.1a}$$

$$\int_{\Omega} (\mathrm{div}\boldsymbol{\sigma} + \rho\boldsymbol{b})\, d\Omega = 0\,, \tag{3.1b}$$

$$\int_{\partial\Omega} \boldsymbol{\sigma}\boldsymbol{n}\, d\Gamma + \int_{\Omega} \rho\boldsymbol{b}\, d\Omega = 0\,, \tag{3.1c}$$

where $\boldsymbol{\sigma}$ is the (Cauchy) stress tensor, ρ is the material's mass density, $\boldsymbol{b}$ is the body force per unit mass and $\boldsymbol{n}$ is the outward unit normal to the curve $\partial\Omega$.

Many of the fracture mechanics path and path domain independent integrals discussed in Chapter 2 as well as the S integral developed herein emanate from the Eshelby law associated with the energy momentum tensor. This formulation directly shows the integral to be the force on all singularities within the integration region.

For further clarity, the subsequent discussion develops the energy momentum tensor basis of the J integral. This development closely follows the efforts of Eshelby [1951,1956,1975] and Knowles and Sternberg [1972].

Within classical field theory Noether's two theorems (Noether [1918, 1971], Gelfand and Fomin [1963], Goldstein [1980] and Rosen [1974] present discussions of these important theorems) relate one parameter symmetry groups and associated conservation laws to variational problems. A simplified version of Noether's general theorem states that if the scalar valued function describing the fields remains invariant under some infinitesimal superimposed variable transformation then there exists a conservation law associated with the symmetry group.

For mechanics, this simplified version of Noether's theorem states that if the system's Lagrangian (or Hamiltonian) functional remains invariant with respect to

a one parameter transformation, i.e., the functional does not explicitly depend on a specific variable, then an associated conservation law exists.

In elasticity, Gunther [1962], Knowles and Sternberg [1972], and Fletcher [1976] among others employ versions of Noether's theorem to obtain conservation laws. Olver [1984a,b] presents the complete set of conservation laws for three dimensional linear, homogeneous, isotropic elastostatics in the absence of body forces. Naturally, one of these laws is the Eshelby law establishing the divergence free nature of the energy momentum tensor and hence the **J** integral.

The Eshelby law for an elastic region containing a singularity (the interesting case for fracture mechanics) produces the force on the singularity. This force also equals the decrease in potential energy per unit motion of the singularity.

The Noether quantity associated with the infinitesimal translation symmetry group generated by $\partial/\partial \boldsymbol{x}$ is the energy momentum tensor. For homogeneous, isotropic elastostatics this simply means that the Lagrangian does not explicitly depend on the coordinate position, $\boldsymbol{x}$.

In the absence of body forces, the Lagrangian density for infinitesimal strain, homogeneous, isotropic elastostatics is identically the stored energy function.

$$\Pi = \int_{\Omega} W \, d\Omega - \int_{\partial_{\sigma}\Omega} \bar{\boldsymbol{t}} \cdot \boldsymbol{u} \, d\Gamma \,, \tag{3.2a}$$

$$\mathsf{L} = \int_{\Omega} W \, d\Omega \,, \tag{3.2b}$$

$$L = W(\boldsymbol{u}, \nabla \boldsymbol{u}) \,, \tag{3.2c}$$

where Π gives the total potential energy, L defines the total Lagrangian, L is the Lagrangian density, W is the stored energy function, $\hat{\boldsymbol{t}} = \boldsymbol{\sigma}\boldsymbol{n}$ is the traction vector and $\partial_{\sigma}\Omega$ is the prescribed traction boundary.

The Noether theorem for the translation symmetry group $\boldsymbol{x}^* \rightarrow \boldsymbol{x} + \boldsymbol{b}$ with

$\|\boldsymbol{b}\|/\|\boldsymbol{x}\| \ll 1$ yields the energy momentum tensor, $\boldsymbol{\Sigma}$, defined as,

$$\boldsymbol{\Sigma} = L\mathbf{1} - (\nabla \boldsymbol{u})^T \frac{\partial L}{\partial \nabla \boldsymbol{u}} . \tag{3.3}$$

The divergence free nature of the energy momentum tensor, $\mathrm{div}\boldsymbol{\Sigma} = 0$, constitutes Eshelby's conservation law for a region without singularities;

$$0 = \int_{\Omega} \mathrm{div}\boldsymbol{\Sigma}\, d\Omega = \int_{\partial\Omega} \boldsymbol{\Sigma}\boldsymbol{n}\, d\Gamma . \tag{3.4}$$

For a region including a singularity, such as a crack tip, the Eshelby law yields the force on the singularity, identically the **J** integral,

$$\mathbf{J} = \int_{\partial\Omega} \boldsymbol{\Sigma}\boldsymbol{n}\, d\Gamma = \int_{\partial\Omega} \{W\boldsymbol{n} - (\nabla \boldsymbol{u})^T (\frac{\partial L}{\partial \nabla \boldsymbol{u}})\boldsymbol{n}\}\, d\Gamma , \tag{3.5}$$

since $\partial L/\partial \nabla \boldsymbol{u}$ equals the stress tensor, $\boldsymbol{\sigma}$.

For the case in which the Noether quantity is not divergence free, such as for inhomogeneous material or in the presence of thermal strains, then the conservation law becomes a path domain independent integral as described in equation (3.1).

The procedure outlined in this section provides the framework for developing the **S** integral. Beginning from a valid Lagrangian density, invoking Noether's theorem for the translation symmetry group produces a conserved quantity akin to the energy momentum tensor for elasticity. Applying Green's theorem to the integral conservation law associated with the divergence of the Noether quantity yields the **S** integral. Thus, **S** is identically the total force on the singularities within the integration region per unit translation of these singularities. Section 3.5 proves this claim for a traction free cavity following the work of Budiansky and Rice [1973] for the **J** integral.

3.2 S Integral for Uncoupled Thermoinelasticity

For slow loading rates in most metals the uncoupled thermoinelasticity assumption yields acceptable results. The uncoupled case assumes that a slowly varying strain field does not induce a temperature change. The slow loading rate assumption forms the foundation for stress intensity factor and **J** integral analyses. The fracture resistance experiments detailed in Chapter 4 examine slow displacement controlled crack extension and hence permit the uncoupled assumption.

This uncoupled case treats the temperature field equations as ancillary to the mechanics problem. The temperature throughout the body is assumed to be known and given at all times. Thus, the Lagrangian density does not need to produce the differential equations governing the temperature field. Although temperature is a parameter in the Lagrangian, it is not an independent variable. In this regard the temperature treatment parallels that of the material properties, the body force or traction boundary conditions.

This **S** integral development considers quasi static conditions and infinitesimal small strains. Neglecting inertia effects typically has less than a ten percent influence on the total energy release rate (Kanninen and Popelar [1985]). Addressing the infinitesimal strain case minimizes the notational difficulties in the development without loss of the concept. Furthermore, and perhaps most convincing, these assumptions generate the case of interest for most structures and components in potential fracture situations. The **S** integral development draws heavily on the work of Simo and Honein [1990] on discrete Lagrangian formulation and discrete conservation laws for inelasticity.

For convenience, this development assumes the typical isotropic relation that temperature change only produces volumetric strain;

$$\boldsymbol{\epsilon}^{th} = \alpha\theta\mathbf{1} \tag{3.6}$$

where ϵ^{th} is the thermal strain tensor, α is the coefficient of thermal expansion, θ is the temperature difference from some reference temperature and $\mathbf{1}$ is the rank two identity tensor.

The total strain tensor permits the decomposition;

$$\epsilon = \frac{1}{2}\{\nabla \boldsymbol{u} + \nabla \boldsymbol{u}^T\} = \epsilon^e + \epsilon^{th} + \epsilon^p \qquad \text{for plasticity} \tag{3.7a}$$

$$\epsilon = \frac{1}{2}\{\nabla \boldsymbol{u} + \nabla \boldsymbol{u}^T\} = \epsilon^e + \epsilon^{th} + \epsilon^{vp} \qquad \text{for viscoplasticity} \tag{3.7b}$$

where $\boldsymbol{u}$ is the displacement vector, ϵ^e is the elastic strain tensor, ϵ^p is the plastic strain tensor and ϵ^{vp} is the viscoplastic strain tensor.

A stored energy function $W(\epsilon^e)$ governs the stress response. Irreversible plastic flow depends on the total strain history, ϵ, and two internal variables, the plastic strain, ϵ^p, or the viscoplastic strain, ϵ^{vp}, and hardening parameters q. Table 3.1 summarizes the governing equations for classical infinitesimal rate independent plasticity and penalty regularized viscoplasticity. Simo and Hughes [1989], Simo et al. [1989], Strang [1986], Luenberger [1984], Perzyna [1971] and Hill [1960] among others provide the full background for this inelasticity development.

Table 3.1
Equations Governing Plasticity and Viscoplasticity

	Plasticity	Viscoplasticity
Stress Strain Relation	$\boldsymbol{\sigma} = \frac{\partial W(\epsilon^e)}{\partial \epsilon}$	$\boldsymbol{\sigma} = \frac{\partial W(\epsilon^e)}{\partial \epsilon}$
Associative Flow Rule	$\dot{\epsilon}^p = \dot{\gamma}\frac{\partial f(\boldsymbol{\sigma},q)}{\partial \boldsymbol{\sigma}}$	$\dot{\epsilon}^{vp} = \dot{\gamma}\frac{<f(\boldsymbol{\sigma},q)>}{\eta}\frac{\partial f}{\partial \boldsymbol{\sigma}}$
Hardening Law	$\dot{q} = -\dot{\gamma}D\frac{\partial f(\boldsymbol{\sigma},q)}{\partial q}$	$\dot{q} = -\dot{\gamma}\frac{<f(\boldsymbol{\sigma},q)>}{\eta}D\frac{\partial f}{\partial q}$
Yield Condition	$f(\boldsymbol{\sigma},q) \leq 0$	
Loading/Unloading Cond	$\dot{\gamma} \geq 0, \dot{\gamma}f = 0, \dot{\gamma}\dot{f} = 0$	

In Table 3.1 γ is the inelastic consistency parameter, $\boldsymbol{D}$ is the hardening moduli coefficient tensor, f is the yield function, η is the fluidity parameter and a superposed dot, as in $\dot{\gamma}$, denotes time differentiation.

The Lagrangian density must produce the relations in Table 3.1 in addition to the equilibrium equations and natural boundary conditions,

$$\text{div } \boldsymbol{\sigma} + \rho \boldsymbol{b} = 0 \qquad \text{in} \quad \Omega\,, \tag{3.8a}$$

$$\boldsymbol{\sigma n} = \bar{\boldsymbol{t}} \qquad \text{on} \quad \partial_{\boldsymbol{\sigma}} \Omega\,, \tag{3.8b}$$

where $\bar{\boldsymbol{t}}$ is the prescribed traction boundary conditions.

The total free energy Π at any current time t can be defined in terms of the Helmholtz free energy Ψ as,

$$\begin{aligned} \Pi_t(\boldsymbol{u}, \epsilon^p, \boldsymbol{q}) = \int_\Omega \Psi(\boldsymbol{\epsilon} - \epsilon^{th} - \epsilon^p, \boldsymbol{q})\, d\Omega \\ - \int_\Omega \boldsymbol{b} \cdot \boldsymbol{u}\, d\Omega - \int_{\partial_{\boldsymbol{\sigma}} \Omega} \bar{\boldsymbol{t}} \cdot \boldsymbol{u}\, d\Gamma \end{aligned} \tag{3.9}$$

where two quadratic functions define the Helmholtz free energy,

$$\Psi(\epsilon, \epsilon^p, \boldsymbol{q}) = W(\boldsymbol{\epsilon} - \epsilon^{th} - \epsilon^p) + \frac{1}{2} \boldsymbol{q} : \boldsymbol{D}^{-1} : \boldsymbol{q}. \tag{3.10}$$

The total free energy for the viscoplastic case is identical to equation (3.10) with ϵ^{vp} replacing ϵ^p in the Helmholtz free energy.

The total free energy change in the time interval $[t_n, t_{n+1}]$ involves not only the change in the Helmholtz free energy and the potential energy of the loadings but also the energy dissipation associated with irreversible inelastic work. For a purely mechanical process, the second law of thermodynamics requires nonnegative entropy production so. (Simo and Hughes [1989] and Simo and Honein [1990])

$$\mathcal{D} = -\dot{\Psi} + \boldsymbol{\sigma} : \dot{\boldsymbol{\epsilon}} \geq 0 \tag{3.11}$$

where $\mathcal{D}$ is the entropy production per unit volume which identically equals the instantaneous energy dissipation.

Differentiating the Helmholtz free energy expression in (3.10) and invoking Coleman's method (Coleman and Noll [1963]) to obtain $\boldsymbol{\sigma} = \partial\Psi/\partial\boldsymbol{\epsilon}$ leaves the familiar expression for inelastic dissipation,

$$\mathcal{D} = \boldsymbol{\sigma} : \dot{\boldsymbol{\epsilon}}^p - \dot{\boldsymbol{q}} : \boldsymbol{D}^{-1} : \boldsymbol{q} - \dot{\gamma} f \geq 0 \qquad \text{for plasticity,} \qquad (3.12a)$$

$$\mathcal{D} = \boldsymbol{\sigma} : \dot{\boldsymbol{\epsilon}}^{vp} - \dot{\boldsymbol{q}} : \boldsymbol{D}^{-1} : \boldsymbol{q} - \frac{1}{\eta}\gamma^+(f) \geq 0 \qquad \text{for viscoplasticity,} \qquad (3.12b)$$

where the term $\dot{\gamma} f = 0$ for plasticity has no effect on the dissipation whereas the term $\frac{1}{\eta}\gamma^+(f)$ for viscoplasticity includes the viscous dissipation. Here $\gamma^+(f)$ is a penalty function defined as;

$$\gamma^+(x) = \begin{cases} 0, & \text{for } x \leq 0, \\ \frac{1}{2}x^2 & \text{for } x > 0, \end{cases} \qquad (3.13a)$$

$$\text{and} \quad \frac{d}{dx}\gamma^+(x) = < x >, \qquad (3.13b)$$

where $< x >$ is the McCauly bracket function of the argument x.

With these dissipation definitions a Lagrangian function associated with energy dissipation in the body up to the current time can be defined following the work of Simo and Honein [1990] and Simo et al. [1989] as;

$$L_t^d = \int_0^t \int_\Omega \mathcal{D}\, d\Omega\, d\tau \,. \qquad (3.14)$$

At this point the Lagrangians defining the total free energy and the energy dissipation combine to yield a single Lagrangian function expressing the total free energy in the body at any time. This total Lagrangian density is what will yield a material conservation law by using Noether's theorem.

Discretizing the dissipation Lagrangian for the time interval $t \in [t_n, t_{n+1}]$ produces.

$$L_{t_{n+1}}^d = L_{t_n}^d + \int_{t_n}^{t_{n+1}} \int_\Omega \mathcal{D}\, d\Omega\, d\tau \,. \qquad (3.15)$$

Evaluating this time integral by the backwards Euler scheme yields the discrete expression for the plastic dissipation Lagrangian in terms of variable values at the beginning (n) and end $(n+1)$ of the time interval,

$$L^d_{n+1} = L^d_n + \int_\Omega \{(\epsilon^p_{n+1} - \epsilon^p_n) : \sigma_{n+1} - \gamma_{n+1} f_{n+1} - (q_{n+1} - q_n) : D^{-1} : q_{n+1}\}\, d\Omega \,. \tag{3.16}$$

where $\gamma_{n+1} = \dot{\gamma}\Delta t$. For viscoplasticity the strain ϵ^{vp}_m replaces ϵ^p_m and the term $\frac{\Delta t}{\eta}\gamma^+_{n+1}(f_{n+1})$ replaces $\gamma_{n+1} f_{n+1}$.

As discussed in Simo et al. [1989] and Simo and Honein [1990] a functional describing the total available free energy at time t_n in terms of state variables at time t_{n+1} produces the discrete expression,

$$\begin{aligned}\hat{\Pi}_n(\chi_{n+1}) &= \Pi_{n+1}(\chi_{n+1}) + (L^d_{n+1}(\chi_{n+1}) - L^d_n) \\ &= \int_\Omega \hat{L}_n(\chi_{n+1})\, d\Omega + \int_{\partial\Omega} \hat{L}^{BC}_n(\chi_{n+1})\, d\Gamma \,.\end{aligned} \tag{3.17}$$

where χ_t is the set of state variables at time t, i.e., $\chi_{n+1} = \{u_{n+1}, \epsilon^p_{n+1}, q_{n+1}, \gamma_{n+1}\}$ for plasticity. For viscoplasticity, the strain ϵ^{vp} replace ϵ^p. The values χ_τ, for $\tau \in [0, t_n]$, up to time t_n are assumed given and fixed. Also, this uncoupled case assumes that θ_t is given and fixed *for all time* since the temperature development is ancillary to the mechanics problem.

The discrete Lagrangian densities for plasticity are defined as:

$$\begin{aligned}\hat{L}_n(\chi_{n+1}) = W(\epsilon_{n+1} - \epsilon^{th}_{n+1} - \epsilon^p_{n+1}) &+ \frac{1}{2} q_{n+1} : D^{-1} : q_{n+1} - \rho b_{n+1} \cdot u_{n+1} \\ &- \gamma_{n+1} f(\sigma_{n+1}, q_{n+1}) + (\epsilon^p_{n+1} - \epsilon^p_n) : \sigma_{n+1} \\ &- (q_{n+1} - q_n) : D^{-1} : q_{n+1}\end{aligned} \tag{3.18a}$$

$$\hat{L}^{BC}_n(\chi_{n+1}) = -\bar{t}_{n+1} \cdot u_{n+1}. \tag{3.18b}$$

For viscoplasticity, the strain ϵ^{vp} replace ϵ^p and the term $\frac{\Delta t}{\eta}\gamma^+(f_{n+1})$ replaces $\gamma_{n+1} f_{n+1}$.

These discrete Lagrangian densities associated with the total available energy functional must yield the correct equilibrium conditions and inelasticity evolution expressions as Euler Lagrange equations. Employing standard calculus of variations methods (Gelfand and Fomin [1963]), the stationary conditions of the functional (3.17) result in the following weak forms,

$$\delta\hat{\Pi}(\boldsymbol{\chi}_{n+1};\delta\boldsymbol{u}) = \int_{\Omega}\{\nabla(\delta\boldsymbol{u}) : \frac{\partial W}{\partial\boldsymbol{\epsilon}_{n+1}} - \rho\boldsymbol{b}_{n+1}\cdot\delta\boldsymbol{u}\}\,d\Omega - \int_{\partial_{\boldsymbol{\sigma}}\Omega}\bar{\boldsymbol{t}}_{n+1}\cdot\delta\boldsymbol{u}\,d\Gamma = 0\,, \tag{3.19a}$$

$$\delta\hat{\Pi}(\boldsymbol{\chi}_{n+1};\delta\boldsymbol{\epsilon}^{p}) = \int_{\Omega}\{\boldsymbol{\sigma}_{n+1} - \frac{\partial W}{\partial\boldsymbol{\epsilon}_{n+1}} + (\boldsymbol{\epsilon}^{p}_{n+1} - \boldsymbol{\epsilon}^{p}_{n}) : \frac{\partial\boldsymbol{\sigma}_{n+1}}{\partial\boldsymbol{\epsilon}_{n+1}} - \gamma_{n+1}\frac{\partial f_{n+1}}{\partial\boldsymbol{\sigma}_{n+1}} : \frac{\partial\boldsymbol{\sigma}_{n+1}}{\partial\boldsymbol{\epsilon}_{n+1}}\} : \delta\boldsymbol{\epsilon}^{p}\,d\Omega = 0\,, \tag{3.19b}$$

$$\delta\hat{\Pi}(\boldsymbol{\chi}_{n+1};\delta\boldsymbol{q}) = \int_{\Omega}\{\boldsymbol{D}^{-1} : (\boldsymbol{q}_{n+1} - \boldsymbol{q}_{n}) + \gamma_{n+1}\frac{\partial f_{n+1}}{\partial\boldsymbol{q}_{n+1}}\}\cdot\delta\boldsymbol{q}\;d\Omega = 0\,, \tag{3.19c}$$

$$\delta\hat{\Pi}(\boldsymbol{\chi}_{n+1};\delta\gamma) = \int_{\Omega} f_{n+1}\delta\gamma\,d\Omega = 0\,. \tag{3.19d}$$

Recalling that from Coleman's method $\boldsymbol{\sigma} = \partial\Psi/\partial\boldsymbol{\epsilon} = \partial W/\partial\boldsymbol{\epsilon}$, and using integration by parts on equation (3.19a) to obtain $\delta\boldsymbol{u}$ from $\nabla(\delta\boldsymbol{u})$ yields the discrete version of the Euler Lagrange equations within the domain Ω,

$$\mathrm{div}\boldsymbol{\sigma}_{n+1} + \rho\boldsymbol{b}_{n+1} = 0\,, \tag{3.20a}$$

$$\boldsymbol{\epsilon}^{p}_{n+1} = \boldsymbol{\epsilon}^{p}_{n} + \gamma_{n+1}\frac{\partial f_{n+1}}{\partial\boldsymbol{\sigma}_{n+1}}\,, \tag{3.20b}$$

$$\boldsymbol{q}_{n+1} = \boldsymbol{q}_{n} - \gamma_{n+1}\boldsymbol{D}^{-1} : \frac{\partial f_{n+1}}{\partial\boldsymbol{q}_{n+1}}\,, \tag{3.20c}$$

$$f_{n+1} \le 0.\quad \gamma_{n+1} \ge 0.\quad f_{n+1}\gamma_{n+1} = 0\,. \tag{3.20d}$$

and the natural boundary condition on the traction surface $\partial_{\boldsymbol{\sigma}}\Omega$;

$$\boldsymbol{\sigma}_{n+1}\boldsymbol{n} = \bar{\boldsymbol{t}}_{n+1}. \tag{3.20e}$$

The valid Lagrangian density equation (3.18) now serves as the basis functional for Noether's theorem. Invoking Noether's theorem for the infinitesimal translation symmetry group produces the following quantity that reduces to the energy momentum tensor for elasticity;

$$\boldsymbol{\Sigma}(\boldsymbol{\chi}_{n+1}) = \hat{L}(\boldsymbol{\chi}_{n+1})\mathbf{1} - (\nabla\boldsymbol{\chi}_{n+1})^T \frac{\partial \hat{L}}{\partial(\nabla\boldsymbol{\chi}_{n+1})}. \tag{3.21}$$

For the uncoupled thermoinelastic case $\partial\hat{L}/\partial(\nabla\boldsymbol{\chi}_{n+1}) = \partial\hat{L}/\partial(\nabla\boldsymbol{u}_{n+1}) = \boldsymbol{\sigma}_{n+1}$.

Thus, for uncoupled thermoplasticity the Noether quantity associated with the translation symmetry group is of interest for fracture mechanics is defined by,

$$\begin{aligned}\boldsymbol{\Sigma}^p_{n+1} = \{&W(\boldsymbol{\epsilon}_{n+1} - \boldsymbol{\epsilon}^{th}_{n+1} - \boldsymbol{\epsilon}^p_{n+1}) - \rho\boldsymbol{b}_{n+1}\cdot\boldsymbol{u}_{n+1} + \boldsymbol{q}_{n+1} : \boldsymbol{D}^{-1} : (\boldsymbol{q}_n - \frac{1}{2}\boldsymbol{q}_{n+1}) \\ &+ (\boldsymbol{\epsilon}^p_{n+1} - \boldsymbol{\epsilon}^p_n) : \boldsymbol{\sigma}_{n+1} - \gamma_{n+1}f_{n+1}\}\mathbf{1} - (\nabla\boldsymbol{u}_{n+1})^T\boldsymbol{\sigma}_{n+1}.\end{aligned} \tag{3.22}$$

For viscoplasticity the strain $\boldsymbol{\epsilon}^{vp}_m$ replaces $\boldsymbol{\epsilon}^p_m$ and the term $\frac{\Delta t}{\eta}\gamma^+(f_{n+1})$ replaces $\gamma_{n+1}f_{n+1}$. This Noether quantity is akin to the energy momentum tensor for elasticity. This expression, $\boldsymbol{\Sigma}^p_{n+1}$, is the simple extension of the Noether quantity developed by Simo and Honein [1990] to include uncoupled thermal strains.

As demonstrated in Section 3.1. a material conservation follows from this Noether quantity. Taking the divergence of the Noether quantity $\boldsymbol{\Sigma}^p_{n+1}$ for plasticity and simplifying by using the Euler Lagrange equations (3.20 a-e) produces the following expression,

$$\begin{aligned}\operatorname{div}\boldsymbol{\Sigma}^p_{n+1} = \; &-\boldsymbol{\sigma}_{n+1} : \mathbf{1}(\alpha\nabla\theta_{n+1}) - \rho\nabla\boldsymbol{b}_{n+1}\cdot\boldsymbol{u}_{n+1} + \boldsymbol{q}_{n+1} : \boldsymbol{D}^{-1} : (\nabla\boldsymbol{q}_n) \\ &- \boldsymbol{\sigma}_{n+1} : (\nabla\boldsymbol{\epsilon}^p_n) + \{\frac{1}{2}\boldsymbol{\epsilon}^e_{n+1} : (\nabla\frac{\partial\boldsymbol{\sigma}_{n+1}}{\partial\boldsymbol{\epsilon}_{n+1}}) : \boldsymbol{\epsilon}^e_{n+1} - (\nabla\rho)\boldsymbol{b}_{n+1}\cdot\boldsymbol{u}_{n+1} \\ &+ \boldsymbol{q}_{n+1} : (\nabla\boldsymbol{D}^{-1}) : (\boldsymbol{q}_n - \frac{1}{2}\boldsymbol{q}_{n+1}) - \boldsymbol{\sigma}_{n+1} : \mathbf{1}\theta_{n+1}(\nabla\alpha)\}\end{aligned} \tag{3.23}$$

where the last four terms in brackets emanate from material inhomogeneity. For viscoplasticity the strain $\boldsymbol{\epsilon}^{vp}_m$ replaces $\boldsymbol{\epsilon}^p_m$.

An integral conservation law develops from using Green's theorem on one term of $\int_\Omega \mathrm{div}\boldsymbol{\Sigma}\, d\Omega - \int_\Omega \mathrm{div}\boldsymbol{\Sigma}\, d\Omega = 0$ but not the other. In this manner the integral conservation law $\mathrm{S} = 0$ is formed for a region free of singularities. For completeness the full definition of the S integral for uncoupled thermoinelastic material response follows,

for uncoupled thermo-plasticity,

$$\begin{aligned}\mathrm{S}_{n+1} = \int_{\partial\Omega} \Big(&\{W(\epsilon_{n+1} - \epsilon^{th}_{n+1} - \epsilon^{p}_{n+1}) - \rho \boldsymbol{b}_{n+1}\cdot\boldsymbol{u}_{n+1} + \boldsymbol{q}_{n+1} : \boldsymbol{D}^{-1} : (\boldsymbol{q}_n - \frac{1}{2}\boldsymbol{q}_{n+1}) \\ &+ (\epsilon^{p}_{n+1} - \epsilon^{p}_{n}) : \boldsymbol{\sigma}_{n+1}\}\mathbf{1} - (\nabla \boldsymbol{u}_{n+1})^T \boldsymbol{\sigma}_{n+1} \Big)\, \boldsymbol{n}\, d\Gamma \\ + \int_\Omega \Big\{ &\boldsymbol{\sigma}_{n+1} : \mathbf{1}(\alpha \nabla\theta_{n+1}) + \rho \nabla \boldsymbol{b}_{n+1}\cdot\boldsymbol{u}_{n+1} - \boldsymbol{q}_{n+1} : \boldsymbol{D}^{-1} : (\nabla \boldsymbol{q}_n) \\ &+ \boldsymbol{\sigma}_{n+1} : (\nabla \epsilon^{p}_{n}) - \{\frac{1}{2}\epsilon^{e}_{n+1} : (\nabla \frac{\partial \boldsymbol{\sigma}_{n+1}}{\partial \epsilon_{n+1}}) : \epsilon^{e}_{n+1} - (\nabla\rho)\boldsymbol{b}_{n+1}\cdot\boldsymbol{u}_{n+1} \\ &+ \boldsymbol{q}_{n+1} : (\nabla \boldsymbol{D}^{-1}) : (\boldsymbol{q}_n - \frac{1}{2}\boldsymbol{q}_{n+1}) - \boldsymbol{\sigma}_{n+1} : \mathbf{1}\theta_{n+1}(\nabla\alpha)\} \Big\}\, d\Omega\,, \end{aligned} \tag{3.24a}$$

for uncoupled thermo-viscoplasticity,

$$\begin{aligned}\mathrm{S}_{n+1} = \int_{\partial\Omega} \Big(&\{W(\epsilon_{n+1} - \epsilon^{th}_{n+1} - \epsilon^{vp}_{n+1}) - \rho \boldsymbol{b}_{n+1}\cdot\boldsymbol{u}_{n+1} + \boldsymbol{q}_{n+1} : \boldsymbol{D}^{-1} : (\boldsymbol{q}_n - \frac{1}{2}\boldsymbol{q}_{n+1}) \\ &+ (\epsilon^{vp}_{n+1} - \epsilon^{vp}_{n}) : \boldsymbol{\sigma}_{n+1} - \frac{\Delta t}{\eta}\gamma^{+}(f_{n+1})\}\mathbf{1} - (\nabla \mathbf{u} \quad ;)^T \boldsymbol{\sigma}_{n+1} \Big)\, \boldsymbol{n}\, d\Gamma \\ + \int_\Omega \Big\{ &\boldsymbol{\sigma}_{n+1} : \mathbf{1}(\alpha \nabla\theta_{n+1}) + \rho \nabla \boldsymbol{b}_{n+1}\cdot\boldsymbol{u}_{n+1} - \boldsymbol{q}_{n+1} : \boldsymbol{D}^{-1} : (\nabla \boldsymbol{q}_n) \\ &+ \boldsymbol{\sigma}_{n+1} : (\nabla \epsilon^{vp}_{n}) - \{\frac{1}{2}\epsilon^{e}_{n+1} : (\nabla \frac{\partial \boldsymbol{\sigma}_{n+1}}{\partial \epsilon_{n+1}}) : \epsilon^{e}_{n+1} - (\nabla\rho)\boldsymbol{b}_{n+1}\cdot\boldsymbol{u}_{n+1} \\ &+ \boldsymbol{q}_{n+1} : (\nabla \boldsymbol{D}^{-1}) : (\boldsymbol{q}_n - \frac{1}{2}\boldsymbol{q}_{n+1}) - \boldsymbol{\sigma}_{n+1} : \mathbf{1}\theta_{n+1}(\nabla\alpha)\} \Big\}\, d\Omega\,. \end{aligned} \tag{3.24b}$$

3.3 S Integral for Coupled Thermoinelasticity

General fracture situations require the fully coupled thermoinelastic case. A changing strain field induces a temperature change in all materials. For rapid fracture in metals this temperature increase at the crack tip can be significant, see Atluri et al. [1986], Kuang and Atluri [1985], and Sih and Tzou [1986]. Also, some of the advanced nonmetallic materials such as graphite epoxy composites appear to have coupled temperature strain behavior.

The coupled treatment includes the relations describing the temperature and heat flux as governing Euler Lagrange equations. These equations emanate from the balance of linear momentum and the thermodynamic balance of energy for the system. They involve the time derivative of the strain, internal variables and temperature fields. The uncoupled case allows the temperature to be determined *a priori*. So, for the uncoupled case the associated Lagrangian does not have temperature and heat equations as Euler Lagrange equations.

Following the work on coupled thermoelasticity by Carlson [1972], the equations governing coupled thermoplasticity, without an internal heat source, linearized for the small strain and small temperature change θ from some reference temperature T_r are:

$$\operatorname{div}\boldsymbol{\sigma} + \rho\boldsymbol{b} = 0 \qquad \text{in} \quad \Omega\,, \tag{3.25a}$$

$$T_r\boldsymbol{\beta} : (\dot{\boldsymbol{\epsilon}} - \dot{\boldsymbol{\epsilon}}^p) + c_\epsilon\dot{\theta} - \boldsymbol{\sigma} : \dot{\boldsymbol{\epsilon}}^p + \boldsymbol{q} : \boldsymbol{D}^{-1} : \dot{\boldsymbol{q}} + \boldsymbol{m} : \dot{\boldsymbol{q}} + \operatorname{div}\boldsymbol{h} = 0 \quad \text{in} \quad \Omega\,, \tag{3.25b}$$

$$\bar{\theta} + T_r = \bar{T} \qquad \text{on} \quad \partial_T\Omega\,, \tag{3.25c}$$

$$\boldsymbol{h} = \bar{h} \qquad \text{on} \quad \partial_h\Omega\,. \tag{3.25d}$$

$$\boldsymbol{\sigma}\,\boldsymbol{n} = \bar{t} \qquad \text{on} \quad \partial_\sigma\Omega\,, \tag{3.25e}$$

$$\boldsymbol{u} = \bar{\boldsymbol{u}} \qquad \text{on} \quad \partial_u\Omega\,, \tag{3.25f}$$

where T_1 is the present temperature, T_r is the reference temperature, θ is the temperature change, $T_1 + \theta = T_r$, r is the heat source, $\boldsymbol{h}$ is the heat flux, $\bar{T}$, $\bar{\boldsymbol{h}}$ and $\bar{\boldsymbol{u}}$ are the prescribed boundary temperature, displacement and heat flux respectively. A superimposed dot, as in $\dot{\epsilon}^p$, denotes time differentiation. $\boldsymbol{\beta}$ is the material property tensor coupling strain and temperature, c_ϵ is the specific heat at constant strain and $\boldsymbol{m}$ is the material property tensor relating temperature, entropy and the hardening variables. Specifically, $\boldsymbol{m}$ is related to the reference temperature and the Helmholtz free energy by $\boldsymbol{m} = -T_r \partial^2 \Psi / (\partial T \partial \boldsymbol{q})$. Nowacki [1969] and Carlson [1972] present concise linear thermoelastic developments that define c_ϵ and $\boldsymbol{\beta}$ in terms of the Helmholtz free energy as $c_\epsilon = -T_r \partial^2 \Psi / (\partial T \partial T)$ and $\boldsymbol{\beta} = -\partial^2 \Psi / (\partial T \partial \boldsymbol{\epsilon})$. For linearized thermo-viscoplasticity $\boldsymbol{\epsilon}^{vp}$ replaces $\boldsymbol{\epsilon}^p$.

Equations (3.25 a-f) together with the internal variable evolution expressions from Table 3.1 form the full set of field equations for the linearized version of coupled thermoinelasticity. Equation (3.25 b) is the linearized balance of energy for a thermoinelastic material where the local entropy production (Classius–Plank inequality) is nonzero. For a thermoelastic material, without an internal heat source, where the local entropy production is zero, this balance of energy expression reduces to $T_r \boldsymbol{\beta} : \dot{\boldsymbol{\epsilon}} + c_\epsilon \dot{\theta} + \mathrm{div} \boldsymbol{h} = 0$, see e.g., Nowacki [1969] or Carlson [1972].

This development considers a region without an explicit heat source, see equation (3.25b). The thermal coupling with the strain rate provides an implicit mechanism for heat generation within the region. This assumption reduces the complexity of the subsequent development and applies to most fracture situations.

As in the uncoupled case, this development assumes the typical isotropic relation that temperature change only produces volumetric strains, equation (3.6), and the strain decomposition, equation (3.7a,b).

Following the work of Biot [1956,1959] (see e.g., Fung [1965] page 404 ff. for

a review) a variable transformation simplifies the problem. Introducing the vector quantity $\boldsymbol{H}$ permits the development of a valid Lagrangian density. The vector $\boldsymbol{H}$, proportional to the entropy flow or entropy displacement, is defined as,

$$-\text{div}\,\boldsymbol{H} = T_r\boldsymbol{\beta} : (\boldsymbol{\epsilon} - \boldsymbol{\epsilon}^p) + c_\epsilon\theta - \boldsymbol{\sigma} : \boldsymbol{\epsilon}^p + \boldsymbol{q} : \boldsymbol{D}^{-1} : \boldsymbol{q} + \boldsymbol{m} : \boldsymbol{q}\,, \tag{3.26a}$$

$$\text{such that} \qquad \dot{\boldsymbol{H}} = \boldsymbol{h} = -\boldsymbol{\kappa}\nabla\theta\,, \tag{3.26b}$$

$$\text{div}\,\dot{\boldsymbol{H}} = \text{div}\boldsymbol{h}\,, \tag{3.26c}$$

where $\boldsymbol{\kappa}$ is the thermal conductivity tensor assuming the Fourier law for heat conduction. For the viscoplastic case the viscoplastic strain tensor $\boldsymbol{\epsilon}^{vp}$ replaces the plastic strain tensor $\boldsymbol{\epsilon}^p$.

Three quantities combine to form the total available energy functional; the thermoinelastic potential, the loading energy term and the dissipation.

The thermoinelastic potential, Φ^{th}, for a system at the reference state equals the total Helmholtz free energy. Imposing a temperature change $d\theta$ from some reference temperature $T_1 = T_r + \theta$ requires the addition of heat $c_\epsilon\, d\theta$ and changes the thermoinelastic potential by P_c^{th} where,

$$P_c^{th} = \int_\Omega \int_0^\theta \frac{c_\epsilon\theta}{T_r + \theta}\, d\theta\, d\Omega\,, \tag{3.27}$$

which for the small temperature change assumption, $\theta \ll T_r$, reduces to

$$P_c^{th} \approx \frac{1}{2}\int_\Omega \frac{c_\epsilon\theta^2}{T_r}\, d\Omega\,. \tag{3.28}$$

This change P_c^{th} represents the heat that may be transformed into useful work. This term equals the total of the heat $c_\epsilon\, d\theta$ multiplied by the Carnot efficiency $\theta/(T_r + \theta)$ integrated over the temperature change from 0 to θ (see Fung [1965] page 406 for a full discussion).

The total thermoinelastic potential is the sum of the total Helmholtz free energy at the reference temperature and the term P_c^{th}, the change in potential due to a temperature change holding all other state variables constant. Thus

$$P^{th} = \int_{\Omega} \{\Psi|_{\theta=0} + \frac{1}{2}\frac{c_\epsilon}{T_r+\theta}\theta^2\}\, d\Omega\,. \tag{3.29}$$

Assuming that equation (3.10) describes the Helmholtz free energy, denoting $V = W|_{\theta=0}$, invoking the small temperature change assumption and expressing the result in terms of $\boldsymbol{H}$ yields,

$$P^{th} = \int_{\Omega} \Big(V(\boldsymbol{\epsilon}-\boldsymbol{\epsilon}^p) + \frac{1}{2}\boldsymbol{q} : \boldsymbol{D}^{-1} : \boldsymbol{q} \tag{3.30}$$

$$+ \frac{1}{2T_r c_\epsilon}[T_r\boldsymbol{\beta} : (\boldsymbol{\epsilon}-\boldsymbol{\epsilon}^p) + \mathrm{div}\boldsymbol{H} - \boldsymbol{\sigma} : \boldsymbol{\epsilon}^p + \boldsymbol{q} : \boldsymbol{D}^{-1} : \boldsymbol{q} + \boldsymbol{m} : \boldsymbol{q}]^2 \Big)\, d\Omega\,,$$

for plasticity and for viscoplasticity $\boldsymbol{\epsilon}^{vp}$ replaces $\boldsymbol{\epsilon}^p$.

The loading potential is simply,

$$P^L = -\int_{\Omega} (\rho \boldsymbol{b}\cdot\boldsymbol{u})\, d\Omega - \int_{\partial\Omega} (\bar{\boldsymbol{t}}\cdot\boldsymbol{u} - \frac{\bar{\theta}}{T_r}\boldsymbol{H}\cdot\boldsymbol{n})\, d\Gamma\,. \tag{3.31}$$

The dissipation functional develops from the rate of entropy production. Including the Helmholtz free energy and the specific entropy, S, the expression for the Classius–Duhem inequality (entropy production) can be rewritten to produce the dissipation inequality, see, for example, Sarti and Medri [1985],

$$\mathcal{D} = -(\dot{\Psi} + S\dot{T}_r) + \boldsymbol{\sigma} : \dot{\boldsymbol{\epsilon}} - \frac{1}{T_r}\boldsymbol{h}\cdot(\nabla T_1) \geq 0\,. \tag{3.32}$$

Differentiating the Helmholtz free energy and invoking Coleman's method to obtain $\boldsymbol{\sigma} = \partial\Psi/\partial\boldsymbol{\epsilon} = \partial V/\partial\boldsymbol{\epsilon} - \boldsymbol{\beta}\,\theta$ where $\theta = \frac{1}{c_\epsilon}[T_r\boldsymbol{\beta} : (\boldsymbol{\epsilon}-\boldsymbol{\epsilon}^p) + \mathrm{div}\boldsymbol{H} - \boldsymbol{\sigma} : \boldsymbol{\epsilon}^p + \boldsymbol{q} : \boldsymbol{D}^{-1} : \boldsymbol{q} + \boldsymbol{m} : \boldsymbol{q}]$ and $S = \partial\Psi/\partial T$ yields the dissipation functional for linearized coupled thermoplasticity,

$$\mathcal{D} = \boldsymbol{\sigma} : \dot{\boldsymbol{\epsilon}}^p - \dot{\boldsymbol{q}} : \boldsymbol{D}^{-1} : \boldsymbol{q} - \dot{\gamma} f + \frac{1}{2T_r}\dot{\boldsymbol{H}}^T\boldsymbol{\kappa}^{-1}\dot{\boldsymbol{H}} \geq 0\,, \tag{3.33a}$$

or for linearized coupled thermo-viscoplasticity,

$$\mathcal{D} = \boldsymbol{\sigma} : \dot{\epsilon}^{vp} - \dot{\boldsymbol{q}} : \boldsymbol{D}^{-1} : \boldsymbol{q} - \frac{1}{\eta}\gamma^{+}(f) + \frac{1}{2T_r}\dot{\boldsymbol{H}}^{T}\boldsymbol{\kappa}^{-1}\dot{\boldsymbol{H}} \geq 0 \,. \tag{3.33b}$$

As in the uncoupled case the total dissipation up to the current time expressed as a Lagrangian functional is,

$$L^{D} = \int_0^t \int_{\Omega} \mathcal{D} \, d\Omega \, d\tau \,. \tag{3.34}$$

Discretizing the dissipation Lagrangian for the time interval $t \in [t_n, t_{n+1}]$ and evaluating the resulting integral via the backwards Euler scheme produces the discrete dissipation Lagrangian functional,

for plasticity

$$\begin{aligned} L^{d}_{n+1} = L^{d}_{n} + \int_{\Omega} \Big\{ & (\epsilon^{p}_{n+1} - \epsilon^{p}_{n}) : \boldsymbol{\sigma}_{n+1} - \gamma_{n+1} f_{n+1} \\ & - (\boldsymbol{q}_{n+1} - \boldsymbol{q}_{n}) : \boldsymbol{D}^{-1} : \boldsymbol{q}_{n+1} \\ & + \frac{1}{2T_r \Delta t}(\boldsymbol{H}_{n+1} - \boldsymbol{H}_{n})^{T}\boldsymbol{\kappa}^{-1}(\boldsymbol{H}_{n+1} - \boldsymbol{H}_{n}) \Big\} \, d\Omega. \end{aligned} \tag{3.35a}$$

for viscoplasticity

$$\begin{aligned} L^{d}_{n+1} = L^{d}_{n} + \int_{\Omega} \Big\{ & (\epsilon^{vp}_{n+1} - \epsilon^{vp}_{n}) : \boldsymbol{\sigma}_{n+1} - \frac{\Delta t}{\eta}\gamma^{+}_{n+1}(f_{n+1}) \\ & - (\boldsymbol{q}_{n+1} - \boldsymbol{q}_{n}) : \boldsymbol{D}^{-1} : \boldsymbol{q}_{n+1} \\ & + \frac{1}{2T_r \Delta t}(\boldsymbol{H}_{n+1} - \boldsymbol{H}_{n})^{T}\boldsymbol{\kappa}^{-1}(\boldsymbol{H}_{n+1} - \boldsymbol{H}_{n}) \Big\} \, d\Omega. \end{aligned} \tag{3.35b}$$

Expressing the total available energy at time t_n in terms of the state variables at time t_{n+1} produces the discrete expression given by equation (3.17), see Simo et al. [1989] and Simo and Honein [1990]. For the linearized coupled thermoplasticity case the discrete Lagrangian density is,

$$\hat{L}_n(\boldsymbol{\chi}_{n+1}) = V(\boldsymbol{\epsilon}_{n+1} - \boldsymbol{\epsilon}^{p}_{n+1}) + \frac{1}{2}\boldsymbol{q}_{n+1} : \boldsymbol{D}^{-1} : \boldsymbol{q}_{n+1} - \rho \boldsymbol{b}_{n+1} \cdot \boldsymbol{u}_{n+1}$$

$$+\frac{1}{2T_r c_\epsilon}[T_r\boldsymbol{\beta}:(\boldsymbol{\epsilon}_{n+1}-\boldsymbol{\epsilon}^p_{n+1})+\mathrm{div}\boldsymbol{H}_{n+1}-\boldsymbol{\sigma}_{n+1}:\boldsymbol{\epsilon}^p_{n+1}$$
$$+\boldsymbol{q}_{n+1}:\boldsymbol{D}^{-1}:\boldsymbol{q}_{n+1}+\boldsymbol{m}:\boldsymbol{q}_{n+1}]^2$$
$$-\gamma_{n+1}f_{n+1}+(\boldsymbol{\epsilon}^p_{n+1}-\boldsymbol{\epsilon}^p_n):\boldsymbol{\sigma}_{n+1}-(\boldsymbol{q}_{n+1}-\boldsymbol{q}_n):\boldsymbol{D}^{-1}:\boldsymbol{q}_{n+1}$$
$$+\frac{1}{2T_r\Delta t}(\boldsymbol{H}_{n+1}-\boldsymbol{H}_n)^T\boldsymbol{\kappa}^{-1}(\boldsymbol{H}_{n+1}-\boldsymbol{H}_n)\,, \qquad (3.36a)$$

and

$$L^{BC}_n(\boldsymbol{\chi}_{n+1})=-\bar{\boldsymbol{t}}_{n+1}\cdot\boldsymbol{u}_{n+1}+\frac{1}{T_r}\bar{\theta}\boldsymbol{H}_{n+1}\cdot\boldsymbol{n}\,. \qquad (3.36b)$$

In this case, $\boldsymbol{\chi}_{n+1}=\{\boldsymbol{u}_{n+1},\boldsymbol{H}_{n+1},\boldsymbol{\epsilon}^p_{n+1},\boldsymbol{q}_{n+1},\gamma_{n+1}\}^T$. For the coupled thermo-viscoplasticity case the strain $\boldsymbol{\epsilon}^{vp}_m$ replaces $\boldsymbol{\epsilon}^p_m$ and the term $(\Delta t/\eta)\gamma^+_{n+1}$ replaces $\gamma_{n+1}f_{n+1}$.

These discrete Lagrangian densities yield the correct equilibrium conditions and evolution expressions as Euler Lagrange equations. The stationary conditions for the functional (3.17) result in the following weak forms for the linearized coupled thermoplasticity case where the variable substitution $\theta_{n+1}=\frac{-1}{c_\epsilon}[T_r\boldsymbol{\beta}:(\boldsymbol{\epsilon}_{n+1}-\boldsymbol{\epsilon}^p_{n+1})+\mathrm{div}\boldsymbol{H}_{n+1}-\boldsymbol{\sigma}_{n+1}:\boldsymbol{\epsilon}^p_{n+1}+\boldsymbol{q}_{n+1}:\boldsymbol{D}^{-1}:\boldsymbol{q}_{n+1}+\boldsymbol{m}:\boldsymbol{q}_{n+1}]$ has been employed for notational convenience,

$$\delta\Pi(\boldsymbol{\chi}_{n+1};\delta\boldsymbol{u})=\int_\Omega\Big\{(\partial_\epsilon V-\theta_{n+1}\boldsymbol{\beta}):\nabla(\delta\boldsymbol{u})-\rho\boldsymbol{b}_{n+1}\cdot\delta\boldsymbol{u}\Big\}\,d\Omega$$
$$-\int_{\partial_{\boldsymbol{\sigma}}\Omega}\bar{\boldsymbol{t}}_{n+1}\cdot\delta\boldsymbol{u}\,d\Gamma=0\,, \qquad (3.37a)$$

$$\delta\Pi(\boldsymbol{\chi}_{n+1};\delta\boldsymbol{H})=\int_\Omega\Big\{\frac{-\theta_{n+1}}{T_r}\mathrm{div}(\delta\boldsymbol{H})+\frac{1}{T_r\Delta t}(\boldsymbol{H}_{n+1}-\boldsymbol{H}_n)^T\boldsymbol{\kappa}^{-1}\delta\boldsymbol{H}\Big\}\,d\Omega$$
$$+\int_{\partial_\theta\Omega}\frac{\bar{\theta}}{T_r}\delta\boldsymbol{H}\cdot\boldsymbol{n}\,d\Gamma=0\,, \qquad (3.37b)$$

$$\delta\Pi(\boldsymbol{\chi}_{n+1};\delta\boldsymbol{\epsilon}^p)=\int_\Omega\Big\{\boldsymbol{\sigma}_{n+1}(1+\frac{\theta_{n+1}}{T_r})-\partial_\epsilon V+\boldsymbol{\beta}\theta_{n+1}+(\boldsymbol{\epsilon}^p_{n+1}-\boldsymbol{\epsilon}^p_n):\partial_\epsilon\boldsymbol{\sigma}$$
$$-\gamma_{n+1}\partial_{\boldsymbol{\sigma}}f_{n+1}:\partial_\epsilon\boldsymbol{\sigma}\Big\}:\delta\boldsymbol{\epsilon}^p\,d\Omega=0\,, \qquad (3.37c)$$

$$\delta\Pi(\boldsymbol{\chi}_{n+1};\delta\boldsymbol{q})=\int_\Omega\Big\{\boldsymbol{D}^{-1}:(\boldsymbol{q}_{n+1}(1+\frac{2\theta_{n+1}}{T_r})-\boldsymbol{q}_n)$$

$$+ \gamma_{n+1}\partial_q f_{n+1} \Big\} \cdot \delta \boldsymbol{q}\, d\Omega = 0\,, \tag{3.37d}$$

$$\delta \Pi(\boldsymbol{\chi}_{n+1}; \delta\gamma) = \int_\Omega f_{n+1}\delta\gamma \; d\Omega = 0\,. \tag{3.37e}$$

where, for example, $\partial_\epsilon V$ denotes $\partial V/\partial\epsilon$ in the compact notation.

Recalling that from Coleman's method, $\boldsymbol{\sigma} = \partial\Psi/\partial\boldsymbol{\epsilon} = \partial_\epsilon V - \boldsymbol{\beta}\theta$, relying on the small temperature change assumption such that $(1 + \theta/T_r) \approx 1$ and invoking integration by parts on equations (3.37 a,b) produces the following Euler Lagrange equations within the region Ω,

$$\mathrm{div}\boldsymbol{\sigma}_{n+1} + \rho\boldsymbol{b}_{n+1} = 0\,, \tag{3.38a}$$

$$\frac{\boldsymbol{\kappa}}{T_r}\nabla\theta_{n+1} + \frac{1}{T_r\Delta t}(\boldsymbol{H}_{n+1} - \boldsymbol{H}_n) = 0\,, \tag{3.38b}$$

$$\boldsymbol{\epsilon}^p_{n+1} = \boldsymbol{\epsilon}^p_n + \gamma_{n+1}\frac{\partial f_{n+1}}{\partial\boldsymbol{\sigma}_{n+1}}\,, \tag{3.38c}$$

$$\boldsymbol{q}_{n+1} = \boldsymbol{q}_n - \gamma_{n+1}\boldsymbol{D}^{-1}\frac{\partial f_{n+1}}{\partial\boldsymbol{q}_{n+1}}\,, \tag{3.38d}$$

$$f_{n+1} \le 0, \quad \gamma_{n+1} \ge 0, \quad f_{n+1}\gamma_{n+1} = 0\,. \tag{3.38e}$$

and the natural boundary condition on the surface, $\partial\Omega$,

$$\boldsymbol{\sigma}_{n+1}\boldsymbol{n} = \bar{\boldsymbol{t}}_{n+1}\,, \quad \text{on} \quad \partial_{\boldsymbol{\sigma}}\Omega \tag{3.38f}$$

$$\frac{1}{T_r}\theta_{n+1} - \frac{\bar{\theta}}{T_r} = 0\,, \quad \text{on} \quad \partial_T\Omega \tag{3.38g}$$

for $\delta\boldsymbol{u}$ =0 on $\partial_u\Omega$ and $\partial\boldsymbol{H}$ =0 on $\partial_h\Omega$.

Similarly, for the thermo–viscoplastic case the weak forms produce Euler Lagrange equations identical to (3.38 a,b,f,g) with equation (3.38 c.e) replaced by,

$$\boldsymbol{\epsilon}^{vp}_{n+1} = \boldsymbol{\epsilon}^{vp}_n + \frac{\Delta t}{\eta} < f_{n+1} > \frac{\partial f_{n+1}}{\partial\boldsymbol{\sigma}_{n+1}}\,, \tag{3.38c}$$

$$\boldsymbol{q}_{n+1} = \boldsymbol{q}_n - \frac{\Delta t}{\eta} < f_{n+1} > \boldsymbol{D}^{-1}\frac{\partial f_{n+1}}{\partial\boldsymbol{q}_{n+1}}\,, \tag{3.38d}$$

Multiplying equation (3.38b) by $T_r\boldsymbol{\kappa}$ and taking the divergence produces,

$$\begin{aligned}\operatorname{div}(\boldsymbol{\kappa}\nabla\theta_{n+1}) - \frac{1}{\Delta t}[T_r\boldsymbol{\beta} : ((\boldsymbol{\epsilon}_{n+1} - \boldsymbol{\epsilon}_n) - (\boldsymbol{\epsilon}^p_{n+1} - \boldsymbol{\epsilon}^p_n)) + c_\epsilon(\theta_{n+1} - \theta_n) \\ - \boldsymbol{\sigma}_{n+1} : (\boldsymbol{\epsilon}^p_{n+1} - \boldsymbol{\epsilon}^p_n) + \boldsymbol{q}_{n+1} : \boldsymbol{D}^{-1} : (\boldsymbol{q}_{n+1} - \boldsymbol{q}_n) \\ + \boldsymbol{m} : (\boldsymbol{q}_{n+1} - \boldsymbol{q}_n)] \;=\; 0, \end{aligned} \tag{3.39}$$

which is the discrete form of the energy balance equation (3.25 b). Thus, the discrete Lagrangian (3.37) is valid for thermoinelasticity.

As presented in Section 3.1, the Noether quantity associated with the infinitesimal translation symmetry group for a particular Lagrangian density functional is,

$$\boldsymbol{\Sigma}(\chi) \;=\; \hat{L}(\chi)\mathbf{1} - (\nabla\chi)^T\frac{\partial\hat{L}}{\partial(\nabla\chi)}\,. \tag{3.40}$$

For linear coupled thermoplasticity with $\chi_{n+1} = \{\boldsymbol{u}_{n+1}, \boldsymbol{H}_{n+1}, \boldsymbol{\epsilon}^p_{n+1}, \boldsymbol{q}_{n+1}, \gamma_{n+1}\}$, and using the variable transformation $\theta_{n+1} = \frac{-1}{c_\epsilon}[T_r\boldsymbol{\beta} : (\boldsymbol{\epsilon}_{n+1} - \boldsymbol{\epsilon}^p_{n+1}) + \operatorname{div}\boldsymbol{H}_{n+1} - \boldsymbol{\sigma}_{n+1} : \boldsymbol{\epsilon}^p_{n+1} + \boldsymbol{q}_{n+1} : \boldsymbol{D}^{-1} : \boldsymbol{q}_{n+1} + \boldsymbol{m} : \boldsymbol{q}_{n+1}]$ for notational convenience,

$$\frac{\partial\hat{L}}{\partial(\nabla\chi_{n+1})} = \{\boldsymbol{\sigma}_{n+1}, \frac{-1}{T_r}\theta_{n+1}, 0, 0, 0\}^T, \tag{3.41}$$

and for the thermo–viscoplasticity case $\boldsymbol{\epsilon}^{vp}_m$ replaces $\boldsymbol{\epsilon}^p_m$.

Thus, the Noether quantity for thermoplasticity is,

$$\begin{aligned}\boldsymbol{\Sigma}^p_{n+1} = \Big[& V(\boldsymbol{\epsilon}_{n+1} - \boldsymbol{\epsilon}^p_{n+1}) - \rho\boldsymbol{b}_{n+1}\cdot\boldsymbol{u}_{n+1} + (\boldsymbol{q}_n - \frac{1}{2}\boldsymbol{q}_{n+1}) : \boldsymbol{D}^{-1} : \boldsymbol{q}_{n+1} \\ & + \frac{c_\epsilon}{2T_r}[\theta_{n+1}]^2 + (\boldsymbol{\epsilon}^p_{n+1} - \boldsymbol{\epsilon}^p_n) : \boldsymbol{\sigma}_{n+1} - \gamma_{n+1}f_{n+1} \\ & + \frac{1}{2T_r\Delta t}(\boldsymbol{H}_{n+1} - \boldsymbol{H}_n)^T\boldsymbol{\kappa}^{-1}(\boldsymbol{H}_{n+1} - \boldsymbol{H}_n)\Big]\mathbf{1} \\ & - (\nabla\boldsymbol{u}_{n+1})^T\boldsymbol{\sigma}_{n+1} + (\nabla\boldsymbol{H}_{n+1})^T\frac{1}{T_r}[\theta_{n+1}]\,. \end{aligned} \tag{3.42}$$

Naturally, for viscoplasticity the strain $\boldsymbol{\epsilon}^{vp}_m$ replaces the plastic strain $\boldsymbol{\epsilon}^p_m$ and the term $(\Delta t/\eta)\gamma^+_{n+1}(f_{n+1})$ replaces the term $\gamma_{n+1}f_{n+1}$.

As discussed in Section 3.1, a material conservation law follows from this Noether quantity. Taking the divergence of $\boldsymbol{\Sigma}^{p}_{n+1}$ for plasticity and simplifying using Euler Lagrange equations (3.39 a-g) produces the following divergence for a material with homogeneous properties.

$$\begin{aligned}
\mathrm{div}\boldsymbol{\Sigma}^{p}_{n+1} = & -\boldsymbol{\sigma}_{n+1} : \nabla\boldsymbol{\epsilon}^{p}_{n} - \rho\nabla\boldsymbol{b}_{n+1}\cdot\boldsymbol{u}_{n+1} + \nabla\boldsymbol{q}_{n} : \boldsymbol{D}^{-1} : \boldsymbol{q}_{n+1} \\
& + \frac{1}{T_r\Delta t}(\nabla\boldsymbol{H}_{n+1} - \nabla\boldsymbol{H}_{n})^{T}\boldsymbol{\kappa}^{-1}(\boldsymbol{H}_{n+1} - \boldsymbol{H}_{n}) \\
& - (\nabla\boldsymbol{H}_{n+1})^{T}\frac{1}{T_r c_\epsilon}[T_r\boldsymbol{\beta} : (\nabla\boldsymbol{\epsilon}_{n+1} - \nabla\boldsymbol{\epsilon}^{p}_{n+1}) + \nabla(\mathrm{div}\boldsymbol{H}_{n+1}) \\
& - \boldsymbol{\sigma}_{n+1} : \nabla\boldsymbol{\epsilon}^{p}_{n+1} + \boldsymbol{q}_{n+1} : \boldsymbol{D}^{-1} : \nabla\boldsymbol{q}_{n+1} + \boldsymbol{m} : \nabla\boldsymbol{q}_{n+1}\,] . \qquad (3.43)
\end{aligned}$$

For an inhomogeneous material terms involving the material property gradients must also be included. Naturally, for viscoplasticity $\boldsymbol{\epsilon}^{vp}_{m}$ replaces $\boldsymbol{\epsilon}^{p}_{m}$.

As in the uncoupled case, using Green's theorem produces the path domain integral conservation law S=0. For a homogeneous coupled thermoplastic material,

$$\begin{aligned}
\mathbf{S}_{n+1} = & \int_{\partial\Omega}\Big(\{V(\boldsymbol{\epsilon}_{n+1} - \boldsymbol{\epsilon}^{p}_{n+1}) - \rho\boldsymbol{b}_{n+1}\cdot\boldsymbol{u}_{n+1} + (\boldsymbol{q}_{n} - \frac{1}{2}\boldsymbol{q}_{n+1}) : \boldsymbol{D}^{-1} : \boldsymbol{q}_{n+1} \\
& + \frac{c_\epsilon}{2T_r}[\theta_{n+1}]^{2} + (\boldsymbol{\epsilon}^{p}_{n+1} - \boldsymbol{\epsilon}^{p}_{n}) : \boldsymbol{\sigma}_{n+1} \\
& + \frac{1}{2T_r\Delta t}(\boldsymbol{H}_{n+1} - \boldsymbol{H}_{n})^{T}\boldsymbol{\kappa}^{-1}(\boldsymbol{H}_{n+1} - \boldsymbol{H}_{n})\,\}\mathbf{1} \\
& - (\nabla\boldsymbol{u}_{n+1})^{T}\boldsymbol{\sigma}_{n+1} + (\nabla\boldsymbol{H}_{n+1})^{T}\frac{1}{T_r}[\theta_{n+1}]\Big)\,\boldsymbol{n}\,d\Gamma \\
& + \int_{\Omega}(\rho\nabla\boldsymbol{b}_{n+1}\cdot\boldsymbol{u}_{n+1} - \nabla\boldsymbol{q}_{n}\cdot\boldsymbol{D}^{-1}\boldsymbol{q}_{n+1} + \boldsymbol{\sigma}_{n+1} : \nabla\boldsymbol{\epsilon}^{p}_{n} \\
& - \frac{1}{T_r\Delta t}(\nabla\boldsymbol{H}_{n+1} - \nabla\boldsymbol{H}_{n})^{T}\boldsymbol{\kappa}^{-1}(\boldsymbol{H}_{n+1} - \boldsymbol{H}_{n}) \qquad (3.44) \\
& + (\nabla\boldsymbol{H}_{n+1})^{T}\frac{1}{T_r c_\epsilon}[T_r\boldsymbol{\beta} : (\nabla\boldsymbol{\epsilon}_{n+1} - \nabla\boldsymbol{\epsilon}^{p}_{n+1}) + \nabla(\mathrm{div}\boldsymbol{H}_{n+1}) \\
& - \boldsymbol{\sigma}_{n+1} : \nabla\boldsymbol{\epsilon}^{p}_{n+1} + \boldsymbol{q}_{n+1} : \boldsymbol{D}^{-1} : \nabla\boldsymbol{q}_{n+1} + \boldsymbol{m} : \nabla\boldsymbol{q}_{n+1}\,])\,d\Omega .
\end{aligned}$$

For a coupled thermo-viscoplastic homogeneous material $\boldsymbol{\epsilon}^{vp}_{m}$ replaces $\boldsymbol{\epsilon}^{p}_{m}$ and the term $\int_{\partial\Omega} -(\Delta t/\eta)\gamma^{+}_{n+1}(f_{n+1})\,\boldsymbol{n}\,d\Gamma$ is included in this expression for $\mathbf{S}_{n+1}$.

3.4 Path Domain Independence of S

The S fracture parameter developed in the previous sections emanates from a material conservation law. Thus, S is a path domain independent integral. The following brief argument demonstrates the attractive path domain independence of the fracture parameter S.

The path domain independence demonstration hinges on the geometrical construction sketched for two dimensions on Figure 3.1. The argument holds for three dimensions as well. The region Ω_1 has the region Ω_2 as a subset, $\Omega_2 \subset \Omega_1$. A singularity in the form of a crack tip exists in region Ω_2 and therefore also in region Ω_1. Region Ω_3 is the region within Ω_1 excluding Ω_2, i.e., $\Omega_3 = \Omega_1 - \Omega_2$. Region Ω_3 is the broken "doughnut" area shown on Figure 3.1, region ABCDEFA. Although region Ω_3 encloses the space containing the crack tip, region Ω_3 *does not contain the singularity.*

For convenience, the S integral as defined in equations (3.24) and (3.44) is written in shorthand notation as,

$$S_{\Omega_m} = \int_{\partial\Omega_m} \boldsymbol{\Sigma n}\, d\Gamma + \int_{\Omega_m} \mathbf{f}\, d\Omega \tag{3.45}$$

where S_{Ω_m} is the S integral associated with region Ω_m and not the component of S in the direction Ω_m. Also, $\boldsymbol{\Sigma}$ is the Noether quantity given by (3.22) or (3.42) and $\mathbf{f}$ is the negative of the divergence of the Noether quantity, $\mathbf{f} = -\text{div}\,\boldsymbol{\Sigma}$, given by equation (3.23) or (3.43).

With this shorthand notation, and the regions depicted on Figure 3.1 the path domain independence of S follows directly. Since the region Ω_3 does not contain the singularity, the conservation law states that $S_3 = 0$. For regions Ω_1 and Ω_2 containing the singularity the S integral does not equal zero. Hence,

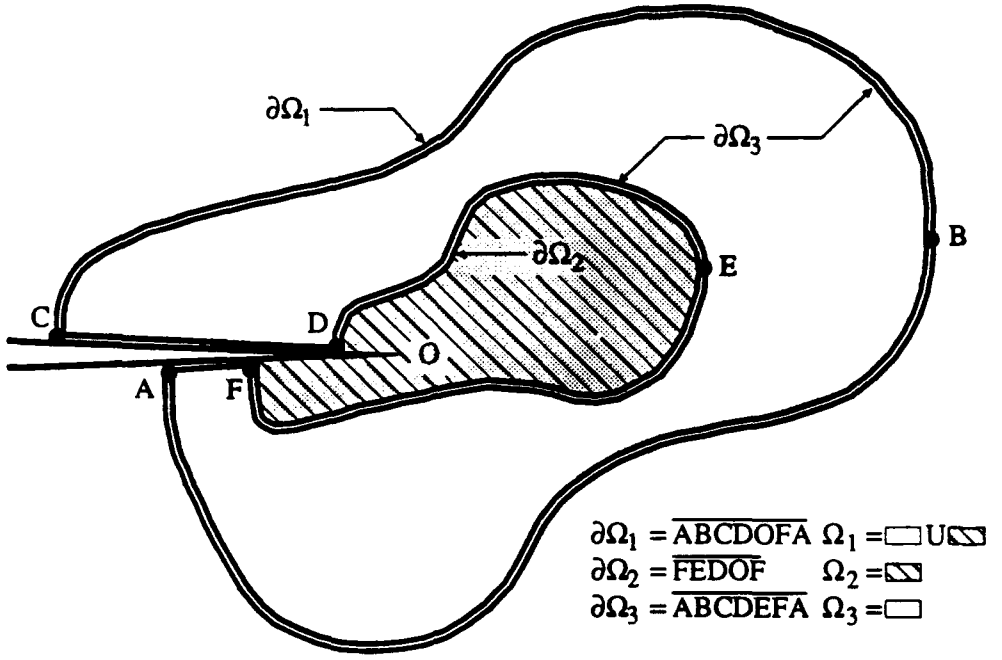


Figure 3.1 *Arbitrary region Ω_1 containing another region Ω_2 that contains a singularity (crack tip).*

$$0 \neq \boldsymbol{S}_{\Omega_1} = \int_{\partial\Omega_1} \boldsymbol{\Sigma n}\, d\Gamma + \int_{\Omega_1} \mathbf{f}\, d\Omega$$

$$= \int_A^B \boldsymbol{\Sigma n}\, d\Gamma + \int_B^C \boldsymbol{\Sigma n}\, d\Gamma + \int_C^D \boldsymbol{\Sigma n}\, d\Gamma + \int_D^O \boldsymbol{\Sigma n}\, d\Gamma + \int_O^F \boldsymbol{\Sigma n}\, d\Gamma + \int_F^A \boldsymbol{\Sigma n}\, d\Gamma$$

$$+ \int_{ABCDEFA} \mathbf{f}\, d\Omega + \int_{FEDOF} \mathbf{f}\, d\Omega, \tag{3.46a}$$

$$0 \neq \boldsymbol{S}_{\Omega_2} = \int_{\partial\Omega_2} \boldsymbol{\Sigma n}\, d\Gamma + \int_{\Omega_2} \mathbf{f}\, d\Omega$$

$$= \int_F^E \boldsymbol{\Sigma n}\, d\Gamma + \int_E^D \boldsymbol{\Sigma n}\, d\Gamma + \int_D^O \boldsymbol{\Sigma n}\, d\Gamma + \int_O^F \boldsymbol{\Sigma n}\, d\Gamma$$

$$+ \int_{FEDOF} \mathbf{f}\, d\Omega, \tag{3.46b}$$

$$0 = S_{\Omega_3} = \int_{\partial\Omega_3} \boldsymbol{\Sigma n}\, d\Gamma + \int_{\Omega_3} \mathbf{f}\, d\Omega$$

$$= \int_A^B \boldsymbol{\Sigma n}\, d\Gamma + \int_B^C \boldsymbol{\Sigma n}\, d\Gamma + \int_C^D \boldsymbol{\Sigma n}\, d\Gamma + \int_D^E \boldsymbol{\Sigma n}\, d\Gamma + \int_E^F \boldsymbol{\Sigma n}\, d\Gamma + \int_F^A \boldsymbol{\Sigma n}\, d\Gamma$$

$$+ \int_{ABCDEFA} \mathbf{f}\, d\Omega\,. \tag{3.46c}$$

Noting the sense of the integration paths, it follows that;

$$S_{\Omega_1} - S_{\Omega_2} = S_{\Omega_3} = 0\,. \tag{3.47}$$

So, the path domain integral S_{Ω_1} must equal the path domain integral S_{Ω_2} for this arbitrary construction. This argument proves the path domain independence property of the S integral.

3.5 S as an Energy Release Rate

The S path domain independent integral carries an energy release rate interpretation in the same manner as the J integral. Inspired by the work of Budiansky and Rice [1973] and following the arguments of Simo and Honein [1990] the S integral equals the change in total energy available due to a unit crack extension. This intuitively follows from the development of S as a force on the crack tip. The development below substantiates this intuition.

This development addresses an arbitrary region, Ω_A containing a traction free cavity C, modeling the crack, as depicted in Figure 3.2. The boundary conditions on $\partial\Omega_A$ remain fixed in the time interval $t \in [t_n, t_{n+1}]$. As presented in equation (3.9). $\Pi_n(\chi_n)$ and $\Pi_{n+1}(\chi_{n+1})$ define the total energy available at times t_n and t_{n+1} respectively. Furthermore, $D(\chi_n, \chi_{n+1})$ defines the total energy dissipation, apart from the fracture energy, due to entropy production during the time interval $t \in [t_n, t_{n+1}]$.

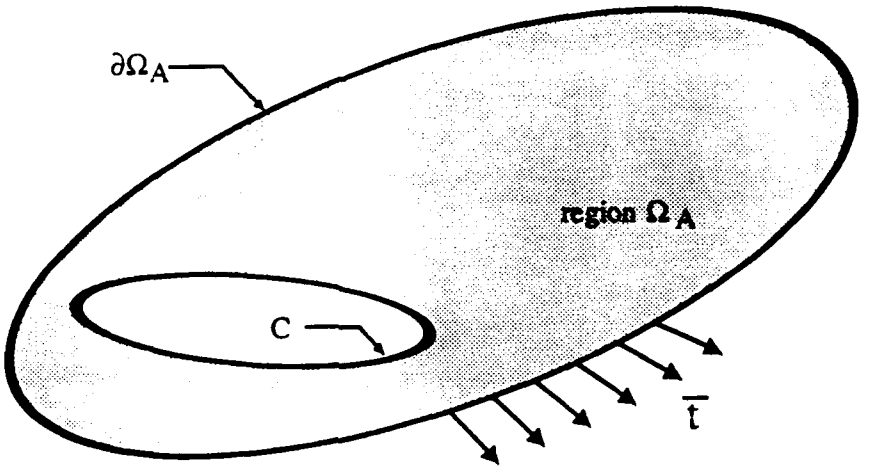


Figure 3.2 *Arbitrary region containing a traction free cavity.*

The homogeneous translation of the cavity C by an amount,

$$\boldsymbol{x} \in C \mapsto \boldsymbol{x} + \nu \boldsymbol{y}\,, \qquad \nu > 0\,, \tag{3.48}$$

during the time interval $t \in [t_n, t_{n+1}]$ mathematically extends the crack in the direction $\boldsymbol{y}$. The state variables $\boldsymbol{\chi}_n$ remain unchanged as they are initial conditions at the beginning of the time interval. The fracture process adds an additional term to the energy balance,

$$\begin{aligned} \Pi_n(\boldsymbol{\chi}_n) = \Pi_{n+1}(\boldsymbol{\chi}_{n+1}(\boldsymbol{x} + \nu\boldsymbol{y})) + D(\boldsymbol{\chi}_n, \boldsymbol{\chi}_{n+1}(\boldsymbol{x} + \nu\boldsymbol{y})) \\ + F(\boldsymbol{\chi}_n, \boldsymbol{\chi}_{n+1}(\boldsymbol{x} + \nu\boldsymbol{y})) \end{aligned} \tag{3.49}$$

where $F(\boldsymbol{\chi}_n, \boldsymbol{\chi}_{n+1})$ is the fracture energy released due to the crack extension. Recalling equation (3.17), $\hat{\Pi}_n$ defines the total energy available at time t_n in terms of the state variables at time t_{n+1}, the fracture energy can be expressed as;

$$F(\boldsymbol{\chi}_n, \boldsymbol{\chi}_{n+1}(\boldsymbol{x} + \nu\boldsymbol{y})) = -(\hat{\Pi}_n(\boldsymbol{\chi}_n, \boldsymbol{\chi}_{n+1}(\boldsymbol{x} + \nu\boldsymbol{y})) - \Pi_n(\boldsymbol{\chi}_n)). \tag{3.50}$$

Invoking the notion of the directional derivative of the fracture energy functional F in the direction $\boldsymbol{y}$ leads to the total energy release rate;

$$< \frac{\delta F}{\delta \boldsymbol{x}} ; \boldsymbol{y} >= -\frac{d}{d\nu}|_{\nu=0} \hat{\Pi}_n(\boldsymbol{\chi}_n, \boldsymbol{\chi}_{n+1}(\boldsymbol{x} + \nu \boldsymbol{y})). \tag{3.51}$$

Simo and Honein [1990] summarize this procedure stating that the directional derivative, $< \frac{\delta F}{\delta \boldsymbol{x}} ; \boldsymbol{y} >$, with respect to variations in $\boldsymbol{x}$, of the total energy available at time t_n expressed in terms of the state variables $\boldsymbol{\chi}_{n+1}$ at time t_{n+1} equals the total energy released due to a homogeneous translation of a defect.

The explicit calculation of equation (3.51) involves the Lagrangian density $\hat{L}$ associated with $\hat{\Pi}$, as defined in equations (3.18) and (3.37) and the transport theorem from continuum mechanics, see Gurtin [1982] for an insightful presentation. Following Budiansky and Rice [1973] and Simo and Honein [1990] this calculation considers only dead loading. Taking the directional derivative of the Lagrangian density in the direction of $\boldsymbol{y}$ produces the following expressions for the uncoupled and coupled cases where the variable substitution $\theta_{n+1} = \frac{-1}{c_\epsilon}[T_r \boldsymbol{\beta} : (\boldsymbol{\epsilon}_{n+1} - \boldsymbol{\epsilon}^p_{n+1}) + \mathrm{div}\boldsymbol{H}_{n+1} - \boldsymbol{\sigma}_{n+1} : \boldsymbol{\epsilon}^p_{n+1} + \boldsymbol{q}_{n+1} : \boldsymbol{D}^{-1} : \boldsymbol{q}_{n+1} + \boldsymbol{m} : \boldsymbol{q}_{n+1}]$ has been employed for notational convenience.

for the uncoupled case,

$$\partial_x \hat{L} \cdot \boldsymbol{y}|_{\boldsymbol{\chi}=\mathrm{const}} = \boldsymbol{\sigma}_{n+} \; : (\partial_x(\boldsymbol{\epsilon}_{n+1}) \cdot \boldsymbol{y}) - \rho \boldsymbol{b}_{n+1} \cdot (\partial_x(\boldsymbol{u}_{n+1}) \cdot \boldsymbol{y}) \,, \tag{3.52a}$$

for the coupled case,

$$\begin{aligned} \partial_x \hat{L} \cdot \boldsymbol{y}|_{\boldsymbol{\chi}=\mathrm{const}} =& \partial_{\boldsymbol{\epsilon}} V_{n+1} : \partial_x \boldsymbol{\epsilon}_{n+1} \cdot \boldsymbol{y} - \rho \boldsymbol{b}_{n+1} \cdot (\partial_x \boldsymbol{u}_{n+1} \cdot \boldsymbol{y}) \\ & - \frac{1}{T_r}(\theta_{n+1}) \cdot (T_r \boldsymbol{\beta} : \partial_x \boldsymbol{\epsilon}_{n+1} \cdot \boldsymbol{y} + \partial_x(\mathrm{div}\boldsymbol{H}_{n+1}) \cdot \boldsymbol{y} \\ & \qquad - \boldsymbol{\sigma}_{n+1} : \partial_x(\boldsymbol{\epsilon}^p_{n+1}) \cdot \boldsymbol{y} + \boldsymbol{q}_{n+1} : \boldsymbol{D}^{-1} : \partial_x(\boldsymbol{q}_{n+1}) \cdot \boldsymbol{y} \\ & \qquad + \boldsymbol{m} : \partial_x(\boldsymbol{q}_{n+1}) \cdot \boldsymbol{y}) \\ & + \frac{1}{T_r \Delta t}(\partial_x \boldsymbol{H}_{n+1} \cdot \boldsymbol{y})^T \boldsymbol{\kappa}^{-1}(\boldsymbol{H}_{n+1} - \boldsymbol{H}_n) \,. \end{aligned} \tag{3.52b}$$

Invoking the transport theorem yields the directional derivatives.

for the uncoupled case

$$-<\frac{\partial F}{\partial \boldsymbol{x}}:\boldsymbol{y}>=\int_{\Omega_A}\Big\{\boldsymbol{\sigma}_{n+1}:(\partial_x\boldsymbol{\epsilon}_{n+1}\cdot\boldsymbol{y})-\rho\boldsymbol{b}_{n+1}\cdot(\partial_x\boldsymbol{u}_{n+1}\cdot\boldsymbol{y})\Big\}\,d\Omega$$
$$-\boldsymbol{y}\cdot\int_C\hat{L}_n\boldsymbol{n}\,d\Gamma-\int_{\partial\Omega_A}\bar{\boldsymbol{t}}\cdot(\partial_x\boldsymbol{u}_{n+1}\cdot\boldsymbol{y})\,d\Gamma\,, \qquad (3.53a)$$

for the coupled case noting $\theta_{n+1}=-\frac{1}{c_\epsilon}[T_r\boldsymbol{\beta}:(\boldsymbol{\epsilon}_{n+1}-\boldsymbol{\epsilon}^p_{n+1})+\mathrm{div}\boldsymbol{H}_{n+1}]$,

$$-<\frac{\partial F}{\partial \boldsymbol{x}}:\boldsymbol{y}>=\int_{\Omega_A}\Big\{(\partial_\epsilon V_{n+1}-\theta_{n+1}\boldsymbol{\beta}):\partial_x\boldsymbol{\epsilon}_{n+1}\cdot\boldsymbol{y}-\frac{\theta_{n+1}}{T_r}(\partial_x(\mathrm{div}\boldsymbol{H}_{n+1}\cdot\boldsymbol{y})$$
$$-\boldsymbol{\sigma}_{n+1}:\partial_x(\boldsymbol{\epsilon}^p_{n+1})\cdot\boldsymbol{y}+\boldsymbol{q}_{n+1}:\boldsymbol{D}^{-1}:\partial_x(\boldsymbol{q}_{n+1})\cdot\boldsymbol{y}$$
$$+\boldsymbol{m}:\partial_x(\boldsymbol{q}_{n+1})\cdot\boldsymbol{y})-\rho\boldsymbol{b}_{n+1}\cdot(\partial_x\boldsymbol{u}_{n+1}\cdot\boldsymbol{y})$$
$$+\frac{1}{T_r\Delta t}(\partial_x\boldsymbol{H}_{n+1}\cdot\boldsymbol{y})^T\boldsymbol{\kappa}^{-1}(\boldsymbol{H}_{n+1}-\boldsymbol{H}_n)\Big\}\,d\Omega$$
$$-\boldsymbol{y}\cdot\int_C\hat{L}_n\boldsymbol{n}\,d\Gamma-\int_{\partial\Omega_A}(\bar{\boldsymbol{t}}\cdot(\partial_x\boldsymbol{u}_{n+1}\cdot\boldsymbol{y})+\frac{\theta_{n+1}}{T_r}\boldsymbol{n})\,d\Gamma\,. \qquad (3.53b)$$

Now, using the divergence theorem to convert the $\int_{\partial\Omega}[\bullet]\boldsymbol{n}\,d\Gamma$ surface integrals to $\int_\Omega \mathrm{div}[\bullet]\,d\Omega$ volume integrals produces the directional derivatives in terms of surface integrals around the cavity,

for the uncoupled case,

$$-<\frac{\partial F}{\partial \boldsymbol{x}}:\boldsymbol{y}>=-\boldsymbol{y}\cdot\int_C(\hat{L}_n\mathbf{1}-(\partial_x\boldsymbol{u}_{n+1})^T\boldsymbol{\sigma}_{n+1})\boldsymbol{n}\,d\Gamma\,. \qquad (3.54a)$$

for the coupled case ,

$$-<\frac{\partial F}{\partial \boldsymbol{x}}:\boldsymbol{y}>=-\boldsymbol{y}\cdot\int_C(\hat{L}_n\mathbf{1}-(\partial_x\boldsymbol{u}_{n+1})^T\boldsymbol{\sigma}_{n+1}-(\nabla\boldsymbol{H}_{n+1})^T\frac{\theta_{n+1}}{T_r})\boldsymbol{n}\,d\Gamma\,. \qquad (3.54b)$$

However, applying the divergence theorem to the previously derived material conservation law (3.21) given by equations (3.22, 3.43) for the region Ω_A leads to;

for the uncoupled thermoplasticity case,

$$\int_C \boldsymbol{\Sigma}_{n+1}(\boldsymbol{\chi}_{n+1})\boldsymbol{n}\,d\Gamma = \int_{\partial\Omega_A} \boldsymbol{\Sigma}^p_{n+1}(\boldsymbol{\chi}_{n+1})\boldsymbol{n}\,d\Gamma + \int_{\Omega_A} \big(\alpha\boldsymbol{\sigma}_{n+1} : \mathbf{1}(\nabla\theta_{n+1}) + \rho\nabla\boldsymbol{b}_{n+1}\cdot\boldsymbol{u}_{n+1} + \boldsymbol{\sigma}_{n+1} : (\nabla\boldsymbol{\epsilon}^p_n) - \boldsymbol{q}_{n+1}\cdot\boldsymbol{D}^{-1}(\nabla\boldsymbol{q}_n)\big)\,d\Omega = \mathbf{S}\,. \tag{3.55a}$$

for the coupled case.

$$\int_C \boldsymbol{\Sigma}_{n+1}(\boldsymbol{\chi}_{n+1})\boldsymbol{n}\,d\Gamma = \int_{\partial\Omega_A} \boldsymbol{\Sigma}^p_{n+1}(\boldsymbol{\chi}_{n+1})\boldsymbol{n}\,d\Gamma + \int_{\Omega_A} \Big(\rho\nabla\boldsymbol{b}_{n+1}\cdot\boldsymbol{u}_{n+1} - \boldsymbol{q}_{n+1}\cdot\boldsymbol{D}^{-1}(\nabla\boldsymbol{q}_n) + \boldsymbol{\sigma}_{n+1} : (\nabla\boldsymbol{\epsilon}^p_n) - \frac{1}{T_r\Delta t}(\nabla\boldsymbol{H}_{n+1} - \nabla\boldsymbol{H}_n)^T\boldsymbol{\kappa}^{-1}(\boldsymbol{H}_{n+1} - \boldsymbol{H}_n) + (\nabla\boldsymbol{H}_{n+1})^T\frac{1}{T_t c_\epsilon}[T_r\boldsymbol{\beta} : (\nabla\boldsymbol{\epsilon}_{n+1} - \nabla\boldsymbol{\epsilon}^p_{n+1}) + \nabla(\mathrm{div}\boldsymbol{H}_{n+1}) - \boldsymbol{\sigma}_{n+1} : \nabla\boldsymbol{\epsilon}^p_{n+1} + \boldsymbol{q}_{n+1} : \boldsymbol{D}^{-1} : \nabla\boldsymbol{q}_{n+1} + \boldsymbol{m} : \nabla\boldsymbol{q}_{n+1}]\Big)\,d\Omega = \mathbf{S}\,. \tag{3.55b}$$

Recalling the definitions of $\boldsymbol{\Sigma}^p_{n+1}$ for both the uncoupled and coupled cases, this development shows that **S** carries an energy release rate interpretation. This quality provides an illuminating physical reality to the **S** path domain independent integral.

3.6 Reduction of S for Simple Cases

Realizing the complexity of equations (3.24 a,b) for the path domain independent integral **S** for uncoupled thermoinelasticity, this section demonstrates that for suitably restricted cases **S** reduces to recognizable expressions from Chapter 2.

For isothermal inelasticity the S integral reduces to the $\mathbf{J}^*$ integral from Simo and Honein [1990], equation (2.15). Considering a homogeneous material,

$$\begin{aligned} \boldsymbol{S} = \int_{\partial\Omega} \Big\{ & \big(W(\epsilon_{n+1} - \epsilon^{vp}_{n+1}) - \boldsymbol{q}_{n+1} : \boldsymbol{D}^{-1} : (\frac{1}{2}\boldsymbol{q}_{n+1} - \boldsymbol{q}_n) \\ & \boldsymbol{\sigma}_{n+1} : (\epsilon^{vp}_{n+1} - \epsilon^{vp}_n) - \frac{\Delta t}{\eta}\gamma^+(f_{n+1})\big)\mathbf{1} - \nabla \boldsymbol{u}^T_{n+1}\boldsymbol{\sigma}_{n+1} \Big\} \boldsymbol{n}\, d\Gamma \\ & + \int_\Omega \left\{ \boldsymbol{\sigma}_{n+1} : \nabla \epsilon^{vp}_n - \boldsymbol{q}_{n+1} : \boldsymbol{D}^{-1} : \nabla \boldsymbol{q}_n \right\} d\Omega \;=\; \mathbf{J}^{*(SH)} . \end{aligned} \tag{3.56}$$

For basic quasi static elasticity and monotonic loading elastoplasticity, S reduces to the J integral, equation (2.4). Considering an elastostatic case with homogeneous, isotropic material in the absence of body forces,

$$\boldsymbol{S} = \int_{\partial\Omega} \left\{ W_{n+1}\boldsymbol{n} - (\nabla \boldsymbol{u}_{n+1})^T \boldsymbol{\sigma}_{n+1}\boldsymbol{n} \right\} d\Gamma \;=\; \boldsymbol{J} . \tag{3.57}$$

Including the thermal strain $\epsilon^{th} = \alpha\theta\mathbf{1}$ in the restricted case considered for equation (3.57), shows that S reduces to $\hat{\mathbf{J}}$, equation (2.27 b), for $\epsilon^* = \epsilon^{th}$ and $\mathbf{J}^\theta$, equation (2.28).

$$\begin{aligned} \boldsymbol{S} = & \int_{\partial\Omega} \left\{ W_{n+1}\boldsymbol{n} - (\nabla \boldsymbol{u}_{n+1})^T \boldsymbol{\sigma}_{n+1}\boldsymbol{n} \right\} d\Gamma \\ & + \int_\Omega \boldsymbol{\sigma}_{n+1} : (\alpha\mathbf{1}\nabla\theta_{n+1})\, d\Omega \;=\; \hat{\mathbf{J}} \;=\; \mathbf{J}^\theta . \end{aligned} \tag{3.58}$$

Finally, expanding this case to include body forces, S reduces to $\hat{\mathbf{J}}$ and $\mathbf{J}^b$, equation (2.30).

$$\begin{aligned} \boldsymbol{S} = & \int_{\partial\Omega} \left(W_{n+1}\boldsymbol{n} - (\nabla \boldsymbol{u}_{n+1})^T \boldsymbol{\sigma}_{n+1}\boldsymbol{n} \right) d\Gamma \\ & + \int_\Omega \left(\boldsymbol{\sigma}_{n+1} : (\alpha\mathbf{1}\nabla\theta_{n+1}) - \rho \boldsymbol{b}_{n+1} \cdot \nabla \boldsymbol{u}_{n+1} \right\} d\Omega \;=\; \hat{\mathbf{J}} \;=\; \mathbf{J}^b . \end{aligned} \tag{3.59}$$

4

Thermomechanical Fracture Experiments

Experiments performed at the Flight Dynamics Laboratory of Wright Research and Development Center demonstrated the effect a thermal gradient has on fracture resistance. Under the influence of a thermal gradient, the mechanical tension required to extend a crack decreased as compared to the isothermal case.

The test program was not intended to exhaustively investigate the thermal gradient effect on fracture resistance. Rather, the experiments attempted to supply a limited proof of concept for the S integral.

4.1 Overview of Experiments

A limited test program investigated the fracture behavior of aluminum 2024 subjected to uniaxial tension and approximately a 65°F/in. thermal gradient. The program contained 21 fracture resistance tests and three tests verifying the S integral conservation law presented in Chapter 3. The fracture resistance tests investigated two specimen thicknesses (0.123 in. and 0.491 in.) and two specimen geometries (single edge notch and center cracked plate). The conservation law verifications tested an uncracked 0.123 in. thick sheet.

The fracture resistance results demonstrate that approximately a 65°F/in. thermal gradient on aluminum 2024 reduces the mechanical tension required for initial crack extension by more than 50% as compared to the isothermal case. Furthermore, these results show that the **J** integral characterizing fracture resistance, as calculated from the load displacement traces, does not adequately address cases including a thermal gradient. While the computed **J** from the isothermal tests matches the expected fracture toughness values, those from the thermal gradient tests average less than 40% of the expected value. (Section 4.5 and specifically Table 4.2 discuss the test results.)

Following an analytical procedure suggested by Kumar et al. [1984] for incorporating a thermal gradient field, the **J** integral can be calculated for proportional loading and a steady state thermal gradient. This procedure, endorsed by the Electric Power Research Institute (EPRI) in EPRI NP-3607, includes the thermal gradient effect by first superimposing a stress field generated by the thermal gradient on the uncracked body onto the mechanical stress field, then calculating the **J** integral. This method yields accurate results for simple temperature distributions coupled with monotonic mechanical loading.

Unfortunately, insufficient displacement gradient field resolution prohibits calculating the **S** value directly from the test data. However, finite element models approximate the fracture resistance tests and supply information for computing the **S** integral that properly includes the thermal gradient effect.

The three conservation law tests' results experimentally verify that **S** represents a conserved quantity as claimed in Chapter 3. This important verification results solely from the test data collected during the three conservation tests.

This chapter begins by discussing the testing program's objectives. Next, Chapter 4 describes the experimental specimens, equipment and instrumentation.

Following the conservation law experimental verification, the fracture resistance tests' summary highlights the effect a thermal gradient has on crack extension and J integral computation. The final section discusses the test program's limitations and suggests an alternative instrumentation package that potentially could capture all of the S integral information.

4.2 Testing Objectives

The experimental program's objective was to collect data to validate the path domain independent integral, S, as a fracture characterization parameter for thermoinelastic material response.

A two part experimental program addressed the testing objective. The first part investigated the integral's use as a conservation law for a body without a crack. The second part examined S as a parameter characterizing the crack driving force at initial crack extension.

The experimental program results verify that the S integral characterizes thermoinelastic fracture. For the homogeneous material properties, uncoupled thermoplastic case treated in the experiments, the following special case of equation (3.24 a) defines the S integral,

$$\begin{aligned} \mathbf{S} = \int_{\partial\Omega} \big[W(\epsilon_{n+1} - \epsilon^{th}_{n+1} - \epsilon^{p}_{n+1}) + q_{n+1} : D^{-1} : (q_n - \frac{1}{2} q_{n+1}) \\ + (\epsilon^{p}_{n+1} - \epsilon^{p}_{n}) : \sigma_{n+1} \big] n - (\nabla u_{n+1})^T \sigma_{n+1} n \, d\Gamma \\ + \int_{\Omega} \big[\sigma_{n+1} : (\alpha \mathbf{1} \nabla \theta_{n+1} + \nabla \epsilon^{p}_{n}) - (q_{n+1} : D^{-1}) : \nabla q_n \big] \, d\Omega \, , \end{aligned} \tag{4.1}$$

where W is the stored energy function such that $\sigma = \partial W / \partial \epsilon$ defines the stress tensor and ϵ is the total strain tensor, q is the tensor of isotropic and kinematic hardening variables, D is the hardening moduli coefficient tensor, u is the displacement vector, ϵ^p is the plastic strain tensor and ϵ^{th} is the thermal strain tensor

defined as $\epsilon^{th} = \alpha\theta\mathbf{1}$ with α as the thermal expansion coefficient and θ as the temperature change from the zero strain reference temperature. The subscripts indicate the time increment associated with each parameter.

The conservation test results demonstrate that the experimentally determined S integral in a crack free specimen remains approximately zero under three loading conditions: isothermal elasticity, thermoelasticity and thermoinelasticity (Section 4.4 provides complete details). The conservation law developed from theory in Chapter 3 claims that S equals zero for any thermomechanical loading in a singularity free body.

The fracture resistance results demonstrate the effect of a thermal gradient on crack extension. The S integral calculated from the load vs displacement trace for the isothermal tests and from the finite element models for the thermal gradient tests attains approximately the same critical value at initial crack extension (Section 4.5 presents the details). This verifies that, for the test material, S characterizes the thermoinelastic energy release rate for a unit crack extension as proven from the theoretical development in Chapter 3.

The testing program provides data to meet the objective of validating S as a fracture characterization parameter for uncoupled thermoinelastic material response. The limited test results suggest that the S path domain independent integral is a conservation law for a crack free body and equals the crack driving force for a cracked body under thermomechanical loads.

This Chapter continues with a discussion of the experimental set up and procedures. Subsequent sections address the conservation law confirmation and the fracture toughness characterization.

4.3 Description of Experiments

The following experimental description provides information on the specimen material, specimen geometry, loading scheme, instrumentation and data acquisition system. Subsequent sections detail the test loading control sequences while discussing the conservation and fracture resistance test results.

4.3.1 Specimen Material and Geometry

The test program examined one half inch thick aluminum 2024-T351 plate and one eighth inch thick 2024-T3 sheet. The specimens measured 24.0 in. long and 3.875 in. wide. The width permitted secure gripping in the 4.0 in. wide Instron hydraulic grips. The length allowed 6.0 in. total grip length, an additional three inches for the distance between the edge of the hydraulic grip body and the gripping dog wedges, space for the thermal gradient generating assemblies exceeding three times the width and some clearance. The fracture specimens included a machined notch 1.850 in. long for the single edge notch (SEN) geometry and 0.600 in. long for the center cracked plate (CCP) geometry.

The aluminum 2024 had constant material properties over the test temperature range of 70°F to 320°F as detailed in Appendix A1. The heated specimen edge reached approximately 320°F during the fracture tests with a thermal gradient of 65°F/in. Following recommendations in the Military Standardization Handbook [1983], the aluminum was assumed homogeneous, isotropic and immune to precipitation age hardening at the test temperatures. The 10,200 ksi Young's modulus (E) and 0.32 Poisson's ratio (ν) obtained from material property tests meeting ASTM [1988] E111 and E132 procedures matched the values in the Military Standardization Handbook and the Metals Handbook [1985]. The 0.2% offset yield strength of 44.4 ksi, the 63.2 ksi ultimate strength, the 17.3% elongation and the 19.3% reduction of area from ASTM E8 tests also agreed with published values. Material test

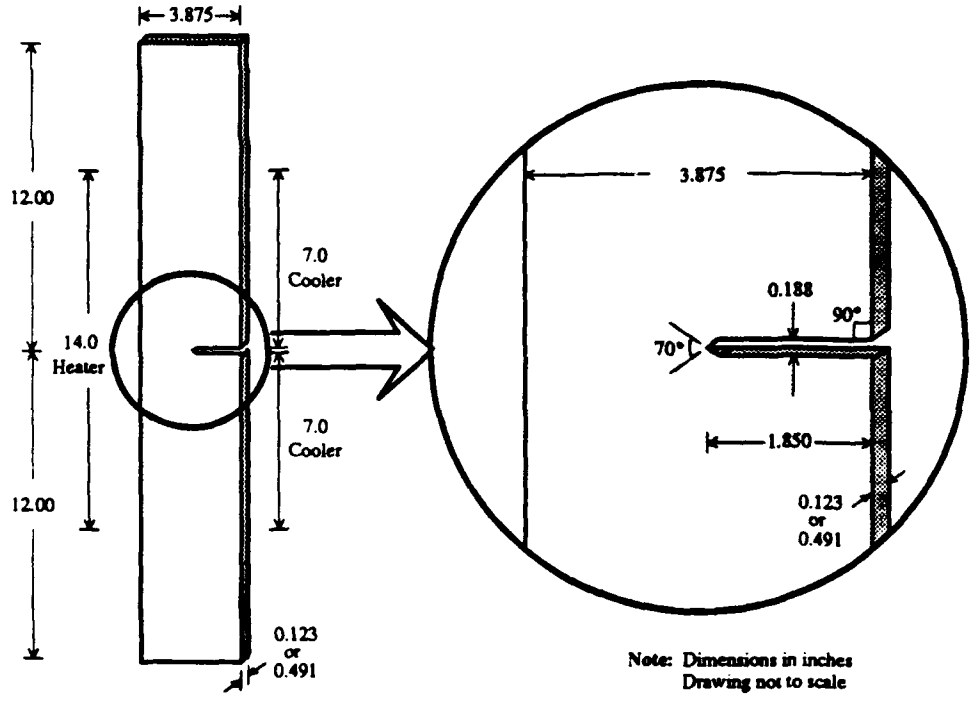


Figure 4.1 *Single edge notch (SEN) specimen geometry.*

data provided parameters for a linear isotropic and kinematic hardening plasticity model plus a Ramberg-Osgood constitutive model as described in Appendix A1.

For the SEN specimen geometry, the 1.850 in. machined notch cuts through approximately half the specimen width as depicted on Figure 4.1. Cutting the notch 0.188 in. wide allowed for installing a crack opening displacement gage. However, this gage was not used in the actual instrumentation. For the CCP specimen geometry, the 0.600 in. long machined central notch allowed a 1.0 in. wide strain gage rosette array to be attached in the crack tip vicinity while permitting a clear 0.50 in. along the specimen edge for installing the heaters and coolers. (Section 4.3.3 provides more instrumentation details.) The conservation test specimen was simply a 24.0 in. x 3.875 in. x 0.123 in. blank without any machined notch.

Starter prefatigue cracks generated in all the fracture specimens developed a

truly sharp crack and moved the tip away from any cold worked area surrounding the machined notch. The starter prefatigue crack grew from the notch tip under force control. For the 0.491 in. thick 2024-T351 specimens, cycling ranged from zero to five thousand pounds to produce the 0.125 in. target starter crack after approximately 40,000 cycles. For the 0.123 in. thick 2024-T3 specimens cycling between zero and one thousand pounds produced the desired final length of approximately 0.125 in. within 30,000 cycles. The low cycling loads created a negligible plastic region ahead of the crack tip as recommended in ASTM E813.

4.3.2 Mechanical and Thermal Loading

A one hundred kip capacity Instron screw type testing machine supplied the mechanical axial loading for all the tests. For the conservation law and fracture resistance tests the machine operated in displacement control with a crosshead speed of 0.05 in./min. Hydraulically operated grips at 3000 psi clamped the top and bottom 3.0 in. of each specimen for effective load transfer.

The test program required both isothermal fracture tests at elevated temperatures and tests with a lateral thermal gradient. One or two simple 1200 watt heat blowers impinging on the specimen around the notch area raised the aluminum temperature to the 170° to 210°F test temperatures for the isothermal elevated temperature tests. The blowers had rudimentary controls on air flow sufficient to maintain specimen temperature within ±3°F for the test duration. Following ASTM E111 recommendations, the specimens soaked at the elevated temperature for approximately one hour per inch thickness prior to testing.

Heaters and coolers attached to either edge along the central 14.0 in. of the specimens generated the desired thermal gradient. Using tap water for the cooling fluid and having approximately a 65°F/in. thermal gradient acting across the 3.875 in. specimen width limited the specimen hot side to 320°F. This arrangement

permitted modeling the aluminum with temperature independent material properties as previously discussed. Four 400 W electric resistance heaters provided the heat source along one edge while water cooled aluminum blocks removed the heat from the other edge. Coating the specimen edges with silicon heat sink compound improved heat flow without increasing the clamping force of the two #10 bolts holding each heater and cooler assembly onto the specimen. Appendix A2 provides complete details of the heater and cooler design and operation.

4.3.3 Specimen Instrumentation

Instrumentation on the specimens provided the data for the computing the fracture resistance and for finite element model comparisons. The load cell internal to the Instron testing machine supplied the applied force information. All of the fracture specimens ca[illegible] a Fractomat 10mm range crack gage to monitor crack length during the experiments. A linear variable displacement transducer (LVDT) attached to the specimen within one half inch of the grips provided the load line displacement necessary for the potential energy calculation required for determining J from the isothermal tests' load displacement curves. Furthermore, strain gages supplied strain information and thermocouples monitored temperature on some of the specimens.

The monitoring gages and transducers, with the exception of the thermocouples, fed their information to a 143 channel analog to digital converter. The converter had a full range resolution of +2048 and -2047 counts at 50 mV input. This gave the overall data system a precision slightly better than 0.05% of the full range value. A DEC Vax 11-780 computer interrogated and recorded the information from each channel approximately once per second.

Difficulties with the thermocouple cold junction reference box prohibited accurate temperature recording through the data system. An analog Doric box

converted the thermocouple voltage to Fahrenheit temperature registered on a light emitting display. The thermocouple temperatures were manually recorded at the beginning and end of each steady state temperature test.

The Instron internal load cell operated in the 10, 20, 50 or 100 kip range depending on the specimen. For the 0.491 in. thick specimens, the SEN geometry required the 50 kip range and the CCP geometry needed the 100 kip capacity. For the 0.123 in. thick specimens, the SEN geometry used the 10 kip range while the CCP geometry ran with the 20 kip capacity. The Instron users manual reported the load cell's accuracy at $\pm 0.25\%$ of the full range, however considering the complete data acquisition system an overall accuracy of one percent appears defendable.

The Fractomat 10 mm constant current crack gage changed output voltage as the advancing crack tore the thin foil gage. Operating at 50 mA gave the required output voltage range for the 10 mm (0.394 in.) full gage range. The Fractomat crack gage offered a 0.01 mm (0.0004 in.) precision through the data system. Calibration of four of the gages with a micrometer on the prefatigue and final resistance crack lengths revealed an accuracy of ± 0.002 in.

The LVDT had ± 0.125 in. full stroke range with a 50 mV output at maximum displacement. A precision of 61μin. due to the analog to digital conversion resolution of one part in 2047, provided ample sensitivity. Checking the LVDT with a reference micrometer every 0.025 in. within its range demonstrated an accuracy of better than ± 0.001 in. The LVDT was bolted to the specimen within a half inch of the grips with #6 bolts (0.1360 in. hole diameter). This positioning captured virtually all of the specimen's axial displacement for the potential energy calculation.

Strain gages bonded to the specimen provided strain information with an overall accuracy of approximately one percent. High temperature M-Bond 610

adhesive (325°F one hour cure) bonded the thermally compensated, 350 ohm, 0.062 in. gage length rosettes, CEA-13-062-WR-350, to the aluminum. The three legs at 0°, 45° and 90° orientations provided the information to calculate the strain tensor components at each gage location. These CEA-13-062-WR-350 gages were chosen because they were the smallest gages the lab had in stock and had experience using. The Flight Dynamics Laboratory used these gages extensively in their tests of thin plates (Sendeckyj [1989]).

A Wheatstone bridge arrangement conditioned and balanced each gage leg signal for the analog to digital processing. The bridge was tuned for small strain accuracy such that approximately one percent strain yielded full scale output on the analog to digital converter.

Type J thermocouples, iron vs constantan -328°F to 1712°F operating range, provided the temperature information with approximately one degree Fahrenheit accuracy and precision. Welding the thermocouples to small gold foil pads welded to the specimens averted difficulties typically encountered in welding the iron thermocouples directly to the aluminum specimens.

Only certain specimens carried strain gages and thermocouples. All specimens subjected to the thermal gradient loading required the full strain gage (ten or twelve rosettes) and thermocouple (five) instrumentation to provide the information for finite element S integral calculation. Additionally, one of the isothermal fracture specimens carried the full instrumentation to compare fracture toughness calculated from the S integral via finite element results to that calculated from the load displacement trace.

Post test strain gage data reduction applied the minor temperature correction to the strain gage data. Using the thermally compensated strain gage rosettes limited this correction to less than 0.0001 in./in. (100 micro strain). These small

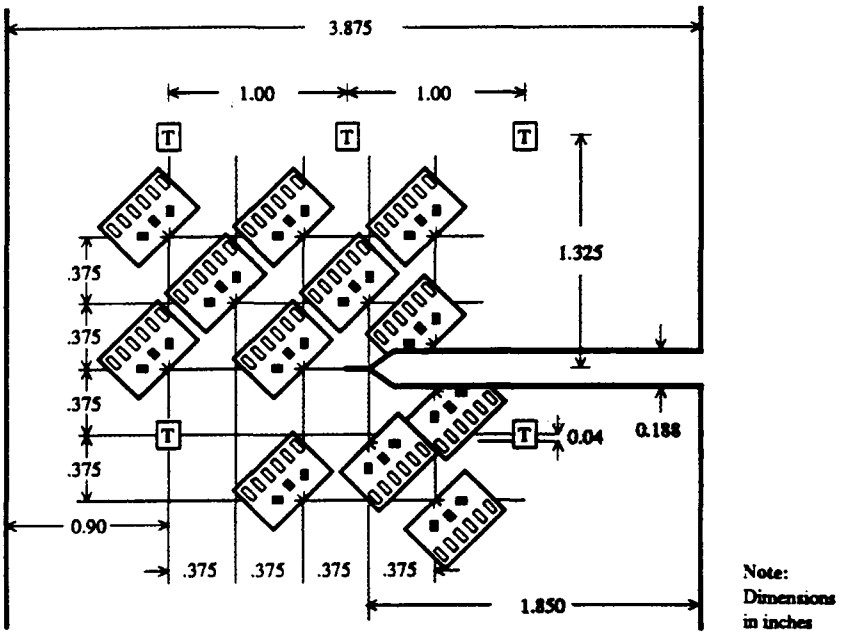


Figure 4.2 *Instrumentation arrangement showing thermocouples,* **T**, *and CEA-13-062-WR-350 strain gage rosettes for the SEN specimen.*

corrections represented less than approximately 3% on a typical gage. The results from the material property tests (see Appendix A1) supported ignoring any zero drift correction.

The strain gage pattern on the SEN specimen fitted the twelve gages over a small area to gather data for comparing test conditions to finite element model results before calculating the S intergal. The five point thermocouple pattern confirmed the linear thermal gradient and checked the gradient along the specimen axis. The strain gage and thermocouple instrumentation pattern depicted on Figure 4.2 for the SEN specimen provided the necessary information. The side opposite these gages carried the crack gage and LVDT required by all fracture specimens. For the CCP geometry, the gage pattern shifted slightly to accommodate the two crack tips as shown on Figure 4.3. Naturally, the CCP specimens carried two Fractomat crack gages to measure total crack length.

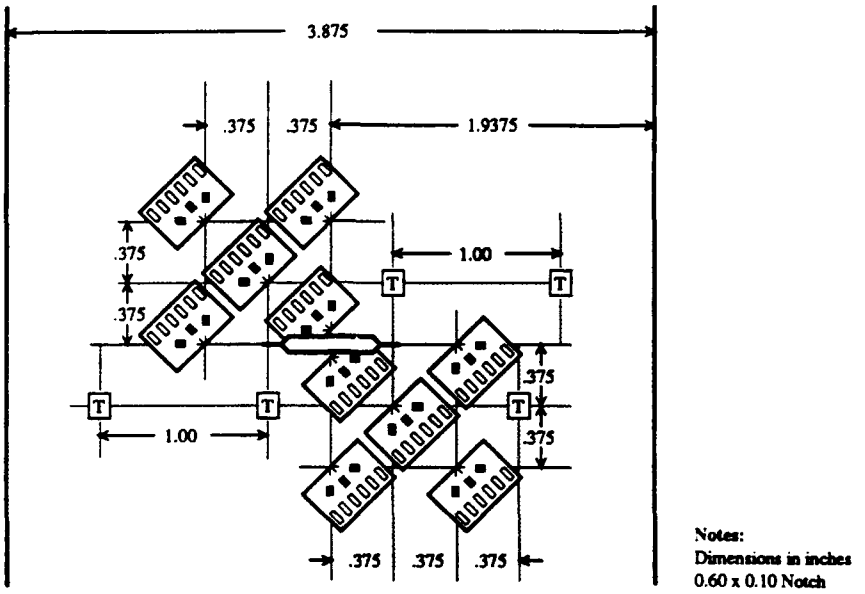


Figure 4.3 *Instrumentation arrangement showing thermocouples,* **T**, *and CEA-13-062-WR-350 strain gage rosettes for the CCP specimen.*

4.4 Experimental Verification of Conservation Law

Three tests confirmed that the S integral represents a thermoinelastic material conservation law. These tests supplied proof of the important concept from Chapter 3. Since S emanated from Noether's theorem from classical field theory and equaled the force on all singularities within a region, for a crack free body S must equal zero. Three separate loadings on a single crack free specimen verified that S equals zero for elastic, thermoelastic and thermoinelastic material response.

The 24.0in. x 3.875in. x 0.123in.aluminum 2024-T3 specimen without a machined notch was instrumented with eleven strain gage rosettes and five thermocouples in a pattern similar to that for the SEN specimen. The two rosettes along the notch in Figure 4.2 were replaced by a single rosette located at the intersection of the grid lines. This gage pattern supplied sufficient information to calculate the S integral solely from experimental data.

The conservation law tests investigated three separate load regimes. First, the fully instrumented notch free specimen underwent a simple elastic uniaxial tension load reaching approximately 60% of the 0.2% offset yield strength. Second, a 72°F/in. thermal gradient acted across the 3.875 in. wide specimen while uniaxial tension brought the average total axial strain to approximately 0.003in./in. Third, the specimen subjected to the thermal gradient was loaded into the plastic range, unloaded and reloaded twice with increasing total displacement on each reloading. These loadings enabled the conservation nature of the S integral to be verified for elastic, thermoelastic and thermoinelastic material response regimes.

The strain gage rosettes and the thermocouples provided the data for calculating the S integral. Reducing the rosette leg data at each gage location produced the strain tensor at eleven spatial points for discrete time points during the loading. The strain tensor was assumed to be valid at the intersection of the individual leg axes comprising the rosette.

The thermocouple data permitted assigning a temperature to each gage location. Since the temperature varied negligibly along the specimen axis and linearly across the specimen width simple linear interpolation estimated the temperature at each gage.

Using the strain tensor and temperature, the plasticity model discussed in Appendix A1 provided the stress history at each gage location. Since the crack free specimen was loaded under displacement control, the displacement gradient component $u_{2,1}$ remained zero throughout the test. This permitted uniquely separating the shear strain into its two displacement gradient parts.

The x component of the S integral computed from the strain gage and thermocouple data from the crack free specimen remains at approximately zero for each of the three loading scenarios. The ratio of the x component of S, denoted S_x, to the

maximum contribution of the x component of the path integral from any segment defined by two successive strain gage locations determines the experimental error in S_x. Normalizing by the exact value, zero, would give an undefined error. The maximum x component path integral contribution is defined as the maximum over all segments defining the closed contour (two successive strain gage locations) of the x component of the path integral, $\mathbf{P}_{\alpha\beta}$, with,

$$\mathbf{P}_{\alpha\beta} = \int_{G_\alpha}^{G_\beta} \left[W(\epsilon_{n+1} - \epsilon^{th}_{n+1} - \epsilon^{p}_{n+1}) + \boldsymbol{q}_{n+1} : \boldsymbol{D}^{-1} : (\boldsymbol{q}_n - \frac{1}{2}\boldsymbol{q}_{n+1}) + (\epsilon^{p}_{n+1} - \epsilon^{p}_{n}) : \boldsymbol{\sigma}_{n+1} \right] \boldsymbol{n} - (\nabla \boldsymbol{u}_{n+1})^T \boldsymbol{\sigma}_{n+1} \boldsymbol{n} \, d\Gamma \tag{4.2}$$

where G_α and G_β are successive strain gage locations along some closed contour.

Based on the isothermal elastic loading test data, the S_x integral varies from 0.7% to 2.8% (absolute value) of the maximum x component of $\mathbf{P}_{\alpha\beta}$ depending on the chosen integration contour. Computing S_x at one instant during the increasing load regime, at the peak tension and during each unloading and reloading portion adds robustness to this investigation.

Using the thermoelastic loading data, the S_x integral for a particular contour surrounding eight gage locations falls within 12.5% (absolute value) of the maximum x component of $\mathbf{P}_{\alpha\beta}$. The thermocouple data show a 72°F/in. thermal gradient across the specimen and the load cell data indicate 14.8 kips tension for this S_x calculation.

At each loading peak (18.7, 20.8 and 22.1 kips) and at each unloading trough (12.8, 15.0 and 13.9 kips) the S_x integral calculated from the thermoplastic (72°F/in. thermal gradient) data remains less than 15% (absolute value) of the maximum x component of $\mathbf{P}_{\alpha\beta}$. This experimentally demonstrates that the **S** integral retains its conservation law nature during plastic loading, unloading and reloading sequences.

The 15% errors in the S_x calculations most likely emanate from the coarse strain gage grid which necessitates a simplistic gradient approximation and permits

only a gross area integration of the stress and strain terms. Assuming the strain reduced from the individual legs to be valid at the intersection of the leg axes also adds to the error. Data from a more refined mesh using smaller stacked rosette strain gages would improve the S_x integral accuracy.

These results experimentally verify the conservation law nature of the S integral fracture parameter. As expected, based on the theoretical development of S in Chapter 3, the S integral remains zero under nonmonotonic thermomechanical loading that produces thermoinelastic material response with thermal strain and accumulating inelastic strain.

4.5 Fracture Tests Show Thermal Gradient Effect

Twenty-one tests investigated the effect a thermal gradient had on fracture resistance. The results from this limited program concluded that the thermal gradient tested exacerbated crack extension. The tensile force required for initial crack extension dropped by more than 50% for the tests with the approximately 65°F/in. thermal gradient as compared to the isothermal tests.

As expected, the simple apparent fracture resistance (J_Q) calculation based on the experimental load vs displacement trace fails to incorporate the thermal gradient effect. However, the procedure suggested by Kumar et al. [1984] (EPRI NP-3607) adequately accounts for the linear thermal gradient in the J_Q calculation by considering a superimposed elastic stress field equivalent to the thermally generated stress field.

The strain gage data analysis can not resolve the displacement gradients with sufficient accuracy for calculating the S integral directly from the experimental data. The critical x component of S calculated solely from the experimental data at initial crack extension falls 50% or more below the value calculated from the

isothermal load vs displacement trace. Insufficient strain and displacement gradient resolution probably contribute the most to this error. The collected data provide the comparison and verification for the finite element model computations used to estimate the S_r integral for the tests.

The test procedure followed the ASTM E813 [1988] specification for determining the apparent fracture toughness. Each specimen was loaded in tension until crack extension began. Then the load was reduced by approximately 25% to determine the elastic compliance. Next, the tension increased until between 0.004 in. and 0.010 in. crack extension occurred followed by a 25% unloading. Approximately seven of these crack extension and compliance determination cycles comprised one test as depicted on Figure 4.4.

The experimental program tested one half inch thick aluminum 2024-T351 plate and one eighth inch thick 2024-T3 sheet in both the center cracked plate (CCP) and single edge notch (SEN) geometries. Each specimen listed on Table 4.1 underwent the fracture toughness loading sequence described in ASTM E813.

4.5.1 Thermal Gradient Promotes Crack Extension

Examining the tension to cause initial crack extension highlights the thermal gradient effect. With approximately a 65°F/in. thermal gradient acting across the 3.875 in. wide specimens, the tension to cause initial crack extension of 0.004 in. to 0.010 in. falls to less than 50% of the isothermal value, (see Table 4.1). The results from the single test with the 26°F/in. gradient (SEN-2.7) shows a tension for initial 0.010 in. crack extension between the isothermal and the 61°F/in. average value as anticipated. The 0.004 in. crack extension represents a small growth yet one large enough (ten times the crack gage precision) to be distinct from data system noise. The 0.010 in. extension is approximately the extension for calculating the fracture toughness value as explained in Section 4.5.2.

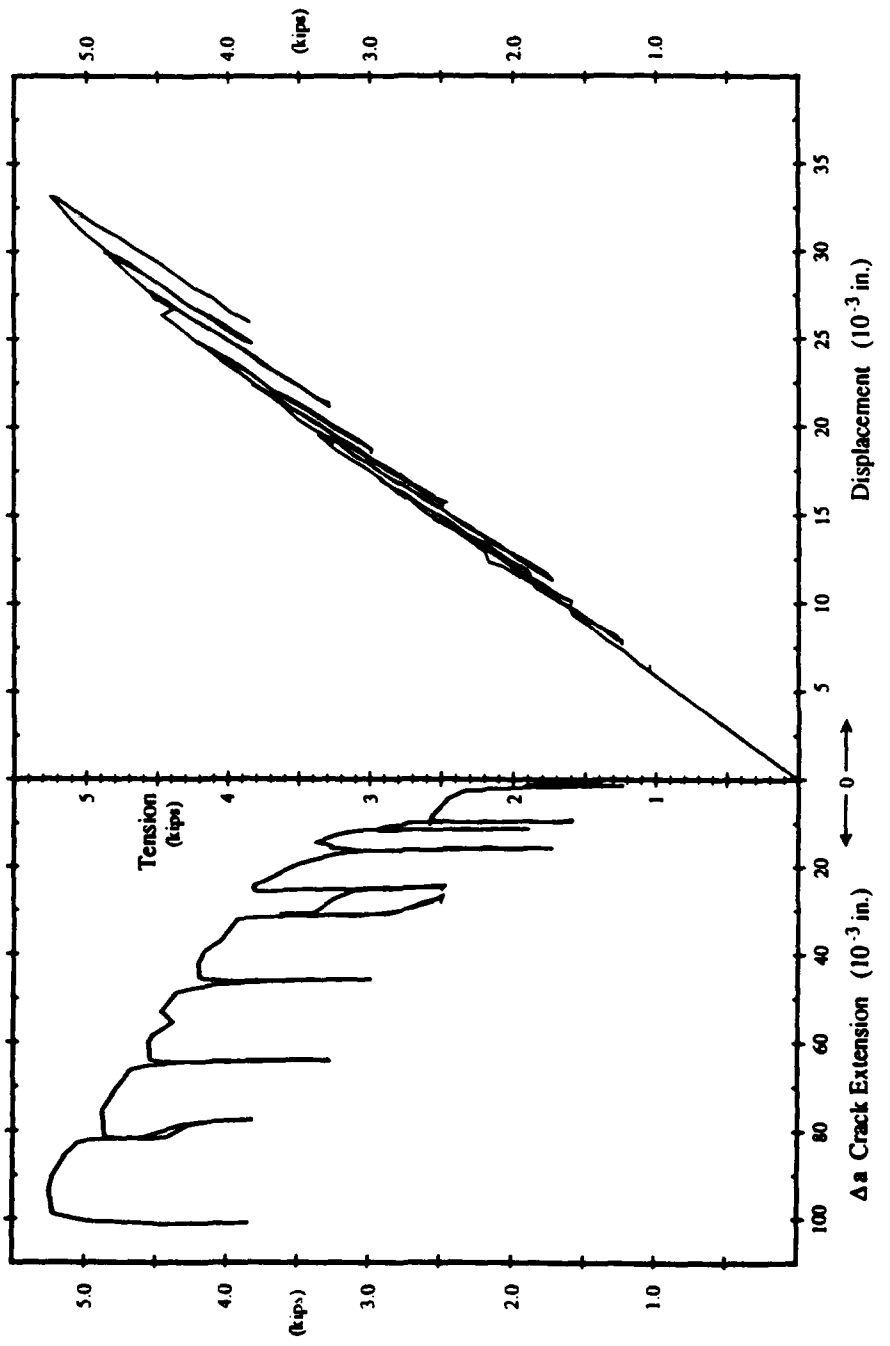


Figure 4.4 *Load - Displacement - Crack Growth for SEN-8.2*

Table 4.1
Fracture Resistance Specimens

Spec.	Temp/Grad	Dimensions (in. x in.)	Machined Notch (in.)	Initial Crack (in.)	Tension at $\Delta a_{to\ Hot}$ 0.004 in. (kips)	0.010 in. (kips)
1/2 inch thick 2024-T351 Aluminum						
CCP-2.1	72 °F	3.877 x 0.495	0.589	0.868	42.4	51.8
CCP-2.2	191 °F	3.879 x 0.490	0.596	0.845	40.5	46.3
CCP-2.3	193 °F	3.875 x 0.495	0.595	0.868	38.3	46.1
SEN-2.1	72 °F	3.875 x 0.493	1.831	1.967	11.0	14.5
SEN-2.2	200 °F	3.875 x 0.490	1.841	1.961	7.8	8.8
SEN-2.3	200 °F	3.876 x 0.489	1.846	1.985	9.1	10.6
CCP-2.4	60 °F/in.	3.878 x 0.494	0.575	0.886	23.4	26.9
CCP-2.5	58 °F/in.	3.875 x 0.492	0.575	0.883	21.8	24.7
SEN-2.4	59 °F/in.	3.879 x 0.489	1.846	1.970	1.9	4.5
SEN-2.5	62 °F/in.	3.875 x 0.490	1.840	1.974	3.5	5.1
SEN-2.6	61 °F/in.	3.875 x 0.490	1.820	1.932	4.0	4.4
SEN-2.7	26 °F/in.	3.874 x 0.494	1.820	1.955	3.8	6.9
1/8 inch thick 2024-T3 Aluminum						
CCP-8.1	200 °F	3.876 x 0.123	0.595	0.855	8.2	10.5
CCP-8.2	72 °F	3.876 x 0.125	0.592	0.909	9.1	13.7
SEN-8.1	172 °F	3.875 x 0.123	1.850	1.969	2.5	2.5
SEN-8.2	183 °F	3.875 x 0.123	1.847	1.963	2.7	3.0
SEN-8.3	207 °F	3.875 x 0.123	1.850	1.966	3.8	5.0
CCP-8.4	64 °F/in.	3.875 x 0.124	0.575	0.859	4.1	5.2
SEN-8.4	72 °F/in.	3.875 x 0.123	1.850	1.979	1.5	1.5
SEN-8.5	69 °F/in.	3.875 x 0.123	1.850	2.100	1.2	1.3
SEN-8.6	69 °F/in.	3.875 x 0.123	1.849	1.967	1.2	1.3

The SEN specimen results demonstrate the thermal gradient effect with the average tension for 0.010 in. initial crack extension falling from 11.3k to 4.7k for the thick specimens and from 3.5k to 1.4k for the thin specimens, see Figures 4.5 and 4.6. The large variation in the initial crack extension vs applied tension traces stems from a number of sources. First, only one of the three isothermal tests on each specimen thickness is from a room temperature test. The other tests are elevated temperature isothermal tests that used blowers to heat the specimens. Uneven heating around the crack tip and through the specimen thickness likely adds to the variation apparent on the traces. Second, having a crack gage on only one side of each specimen adds error due to neglecting any nonuniformity in the crack front through the specimen thickness.

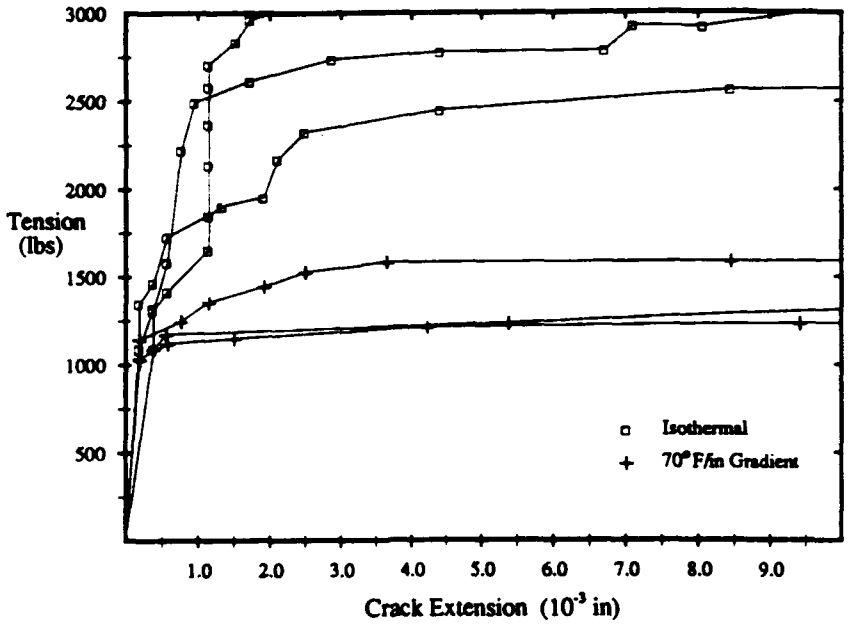


Figure 4.5 *Crack extension vs applied tension for the 0.123 in. thick 2024-T3 SEN specimens.*

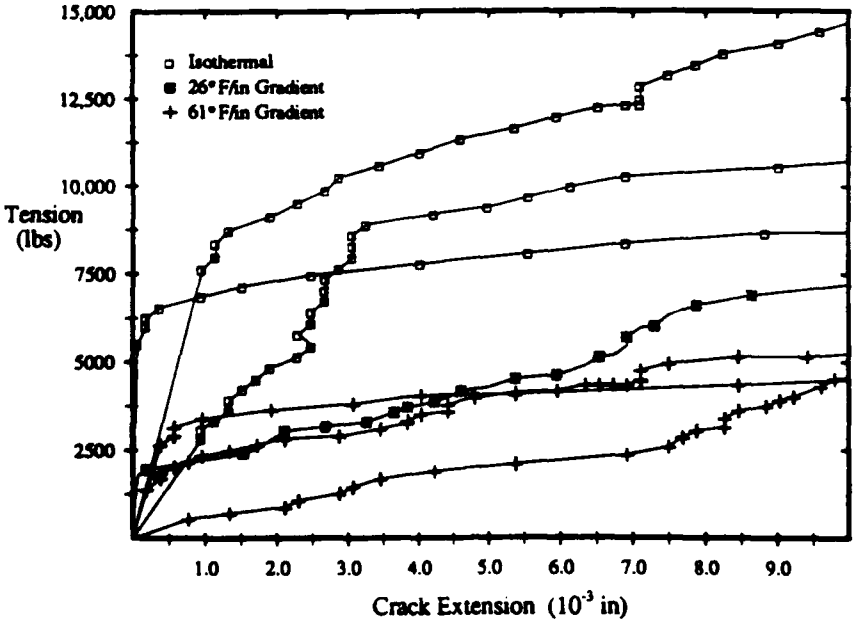


Figure 4.6 *Crack extension vs applied tension for the 0.491 in. thick 2024-T351 SEN specimens.*

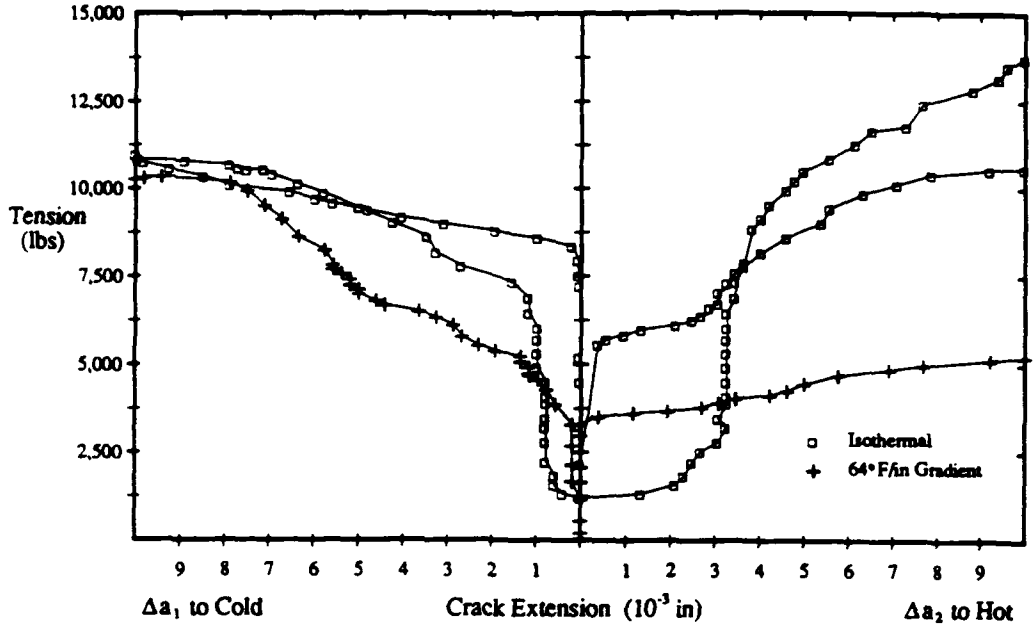


Figure 4.7 *Crack extension vs applied tension for the 0.123 in. thick 2024-T3 CCP specimens.*

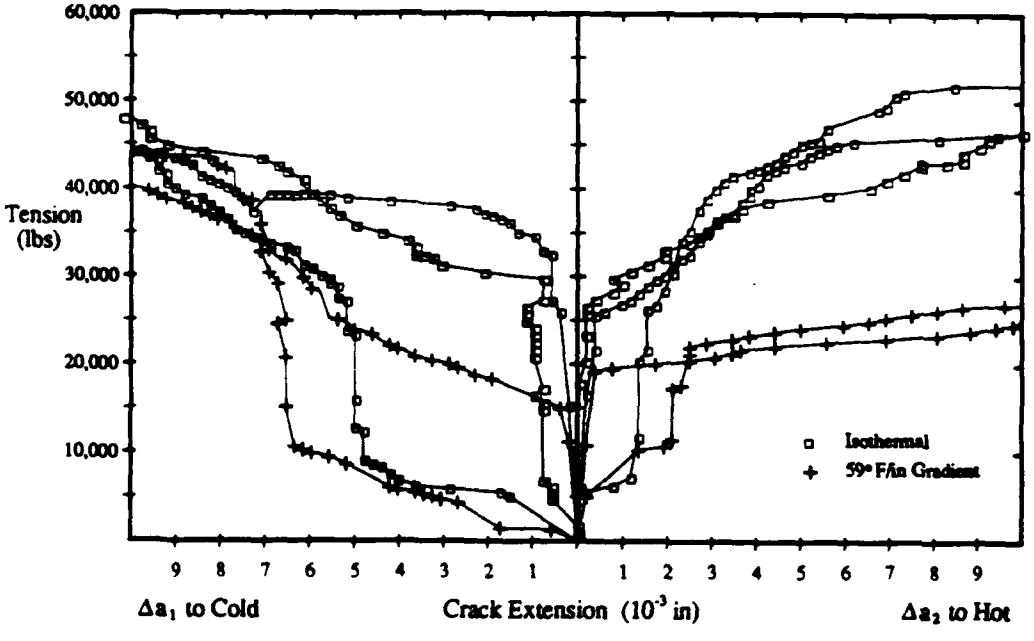


Figure 4.8 *Crack extension vs applied tension for the 0.491 in. thick 2024-T351 CCP specimens.*

As shown on Figures 4.7 and 4.8, the crack extension plots for the CCP specimens demonstrate different behavior depending on the direction of crack advance for the gradient tests. In fact, the variations in the crack extension vs applied tension traces for these few tests undermines any conclusions drawn from theses traces. Generally, at crack extensions approaching 0.010 in. the extension towards the cooled edge (Δa_1 to Cold) shows only a minor difference from the isothermal tests. However, the extension towards the heated edge (Δa_2 to Hot) shows approximately a 50% tension decrease from the isothermal tests. These results disagree with intuition and the finite element based calculations of crack driving force on the respective crack tips. Details of these calculations are presented in Chapter 5. Intuition suggests that the crack would extend preferentially toward the cooled edge since the axial tension increases in that direction.

The CCP crack extension vs applied tension traces at the initial 0.005 in. extension shows opposite trends for the thin and thick specimens under combined thermal gradient and tension loading. For the thin specimen the crack extends preferentially toward the heated edge. However, for the thick specimen the initial crack extension occurs preferentially towards the cooled edge (as expected).

Considering the variations in the isothermal test traces and the likely error from reporting crack length from only one specimen side, this inconsistency from these few tests requires no more investigation here. Naturally, the extension of each end of a CCP specimen crack in a thermal gradient field requires further investigation.

4.5.2 Thermal Gradient Effect on Fracture Toughness

Calculations based on the experimental load vs displacement trace (e.g., Figure 4.4) provide the x component of the J integral, denoted by J_x, as a function of crack advance for each isothermal specimen. This J_x vs crack extension curve's

intersection with the 0.008 in. offset line paralleling the crack tip blunting line determines the apparent fracture toughness, J_Q for the aluminum 2024-T351 plate and the 2024-T3 sheet via the procedure outlined in Broek [1987].

The J_z integral calculation methods suggested by Rice et al. [1973] and Landes et al. [1989] are used to calculate J_z (which identically equals S_z for the isothermal case) from the load vs displacement trace. Both methods split the J_z integral into elastic and plastic portions. Thus,

$$J_z = J_z^{elas} + J_z^{plas}, \tag{4.3}$$

where J_z^{elas} is the elastic energy release rate and J_z^{plas} depends on the area under the load vs displacement curve.

The Griffith energy release rate based on the stress intensity factor defines the elastic portion. Thus,

$$J_z^{elas} = \frac{K_I^2}{E'} \tag{4.4}$$

where J_z^{elas} is the energy release rate, K_I is the mode I stress intensity factor (from Sih [1973] or Murakami [1987] for example) and E' is the effective Young's modulus with $E'=E$ for plane stress and $E' = E/(1-\nu^2)$ for plane strain.

The area for the plastic portion of J_z is the integral of the load over the plastic displacement change. Rice et al. [1973] defines the plastic portion as,

$$J_z^{plas} = \frac{1}{b}\left[2\int_0^{v_{pl}} P\, d\tilde{v}_{pl} - P\, v_{pl}\right] \tag{4.5}$$

where b is the uncracked ligament, $W-a$, with W the specimen width (3.875 in.) and a the crack length, v_{pl} is the total plastic displacement, and P is the force per unit length along the crack front ($P = T/B$ for SEN and $P = T/(2B)$ for CCP with T being the total applied tension load and B the specimen width, 0.123in. or 0.491in.).

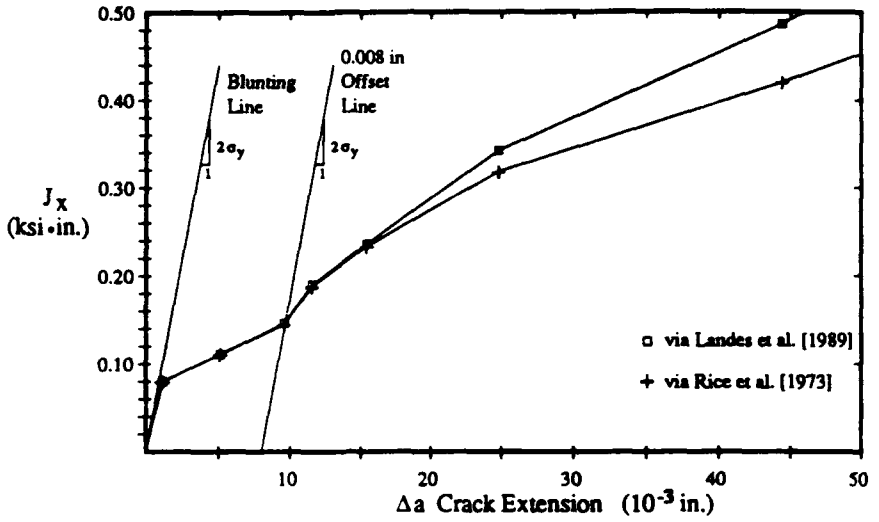


Figure 4.9 *Crack resistance curve for SEN-8.2.*

The method suggested by Landes et al. [1989] includes geometry calibration factors based on Kumar et al. [1981] (EPRI NP-1931). The plastic J_x^{plas} integral calculation from Landes et al. is,

$$J_x^{plas} = \frac{\eta_{pl}}{Bb} \int_0^{v_{pl}} T \, d\tilde{v}_{pl} \tag{4.6}$$

where,

$$\text{for SEN} \quad \eta_{pl} = \frac{0.933}{\gamma} \frac{b}{W} \left(\frac{n+1}{n} \right) \frac{h_1}{h_3} \tag{4.7a}$$

$$\text{for CCP} \quad \eta_{pl} = \left(1 - \frac{2a}{W} \right) \left(\frac{n+1}{n} \right) \frac{h_1}{h_3} \tag{4.7b}$$

where the geometry factor γ and the functions h_1 and h_3 are given by Kumar et al. [1981] and n is the Ramberg-Osgood exponent with n=11.88 for 2024 aluminum as discussed in Appendix A1. Using these two different expressions for J_x^{plas} and the elastic energy release rate J_x^{elas} determines the total J_x from the load displacement trace. Figure 4.9 shows an example of the J_x vs crack extension resistance curve from one of the SEN tests.

The intersection of the resistance curve and the 0.008 in. offset line provides the apparent fracture toughness value J_Q for each test. A linear curve fit to the first few J_r points beyond the 0.008 in. offset line is used to estimate the apparent fracture toughness J_Q (see Broek [1987]).

As presented on Table 4.2, the isothermal test results estimate an apparent fracture toughness J_Q of 0.13 ksi·in. (average of block A) for the half inch thick aluminum 2024-T351 plate and 0.17 ksi·in. (average of block C) for the eighth inch thick aluminum 2024-T3 sheet. The two calculation methods for J_r^{plas} used in determining J_Q agree well within the overall scatter of the results. The crack length measurement probably contributes most to the scatter.

Since the 0.491 in. thickness (B) of the 2024-T351 plate exceeds the critical thickness for plane strain testing ($B \geq 25\,J_r/\sigma_y$ where σ_y is the 0.2% offset yield strength of 44.4 ksi), the 0.13 ksi·in. toughness value meets the requirements for plane strain fracture toughness J_{Ic}. This value agrees with the published value of 0.13±0.01 ksi·in. from Damage Tolerance Design Handbook [1983]. Furthermore, the 0.17 ksi·in. value for the 0.123 in. thick 2024-T3 sheet agrees with the 0.20 ksi·in. value obtained from the plane stress correction to J_{Ic} suggested by Broek [1987] and confirmed in the examples of Pettit and Van Orden [1979] and Sullivan et al. [1973].

As expected, naively computing J_r from the load displacement trace for the tests with a thermal gradient produces inaccurate estimates for the fracture toughness. Simply using the Griffith elastic energy release rate for J_r^{elas} and either the Rice et al. [1973] or Landes et al. [1989] procedure for calculating J_r^{plas} ignores the thermal gradient's contribution to the crack diving force. Neglecting the fracture energy produced by the thermal gradient yields a fracture toughness estimate less than one half the actual value as determined from the isothermal tests

Table 4.2
Fracture Toughness Estimates

1/2 inch thick 2024-T351 Aluminum

	Spec.	Temp/Grad	Exp Load vs Disp J_Q^{Rice} (1) (ksi-in.)	Exp Load vs Disp J_Q^{Landes} (2) (ksi-in.)	EPRI (3,4) J_x (ksi-in.)	FEM S_x (ksi-in.)
(A)	CCP-2.1	72 °F	0.118	0.120	0.119	0.129
	CCP-2.2	191 °F	0.094	0.098	0.099	
	CCP-2.3	193 °F	0.142	0.135	0.157	
	SEN-2.1	72 °F	0.153	0.173	0.149	0.146
	SEN-2.2	200 °F	0.134	0.153	0.135	
	SEN-2.3	200 °F	0.140	0.132	0.149	
		Average:	0.130	0.135	0.135	
(B)	CCP-2.4	60 °F/in.	0.047	0.046	0.087	
	CCP-2.5	58 °F/in.	0.035	0.035	0.088	0.122
	SEN-2.4	59 °F/in.	0.045	0.049	0.215	0.145
	SEN-2.5	62 °F/in.	0.058	0.060	0.250	
	SEN 2.6	61 °F/in.	0.025	0.025	0.178	
		Average:	0.025	0.025	0.163	
	SEN-2.7	26 °F/in.	0.025	0.025	0.140	

1/8 inch thick 2024-T3 Aluminum

	Spec.	Temp/Grad	Exp Load vs Disp J_Q^{Rice} (1) (ksi-in.)	Exp Load vs Disp J_Q^{Landes} (2) (ksi-in.)	EPRI (3,4) J_x (ksi-in.)	FEM S_x (ksi-in.)
(C)	CCP-8.1	200 °F	0.148	0.152	0.146	0.148
	CCP-8.2	72 °F	0.184	0.183	0.192	
	SEN-8.1	172 °F	0.216	0.216	0.219	
	SEN-8.2	183 °F	0.145	0.145	0.148	0.149
	SEN-8.3	207 °F	0.164	0.166	0.172	
		Average:	0.171	0.172	0.175	
(D)	CCP-8.4	64 °F/in.	0.032	0.032	0.155	0.155
	SEN-8.4	72 °F/in.	0.063	0.064	0.175	
	SEN-8.5	69 °F/in.	0.041	0.045	0.199	
	SEN-8.6	69 °F/in.	0.030	0.024	0.172	0.164
		Average:	0.030	0.024	0.175	

Notes: 1. Rice et al. [1973] 2. Landes et al. [1989]
3. Kumar et al. [1981] 4. Kumar et al. [1984]

and confirmed in the literature (see Table 4.2 blocks B and D).

For comparison, the J_r integral value associated with the crack extension at the apparent fracture toughness value is also determined via the EPRI procedures (see Kumar et al. [1981, 1984]) and presented on Table 4.2. Once the intersection of the J_r vs crack extension curve (calculated by either Rice et al. [1973] or Landes et al. [1989] J_r^{plas} procedures) and the crack tip blunting line determine the crack extension (and hence the crack length) for J_Q, the associated tension from the load vs crack length record completes the information required for the EPRI J_r estimation procedure. Invoking the EPRI procedure produces a J_r value equivalent to the apparent fracture toughness J_Q (see Table 4.2).

The isothermal EPRI based J_r integral values agree closely with the values determined from the experimental load vs displacement traces as shown on Table 4.2 blocks A and C. This confirms the adequacy of EPRI NP-1931 (Kumar et al. [1981]) for determining J_r for the two specimen geometries for isothermal loading.

The EPRI NP-3607 procedure suggested by Kumar et al. [1984] includes the thermal gradient effects in the J_r calculation. This procedure incorporates the stress field due to the thermal gradient in the elastic portion of J_r, i.e., J_r^{elas}, the elastic energy release rate, by the superposition method of linear elastic fracture mechanics. The stress intensity factor for the cracked specimen subjected to the stress field generated by the thermal gradient on an uncracked body adds to the stress intensity factor due to the mechanical tension on the cracked specimen in determining K_I and hence J_r^{elas}. Then J_r^{plas} is calculated as prescribed in EPRI NP-1931 (Kumar et al. [1981]).

While the EPRI NP-3607 procedure adequately incorporates the thermal gradient effects for these monotonic loading tests, with a simple linear temperature distribution and uniaxial tension, the procedure can not address cyclic plastic

loading or the other generalities included in the **S** thermoinelastic integral. The EPRI procedure neglects the thermal gradient stress field interaction captured in the **S** integral. Additionally, the EPRI method does not address the accumulation of inelastic strain associated with cyclic loading beyond the elastic yield limit as does the **S** integral.

Calculating the S_r integral from eight of the fracture resistance tests completes the apparent fracture toughness investigation. Finite element analyses supply the information to compute the S_r integral from equation (4.1). The finite element models use the tension (and associated displacement) and crack length at initial crack extension for consistency in determining the equivalent apparent fracture toughness J_Q via S_r. Chapter 5 discusses the modeling and integral computation details. The eight test cases from Table 4.2 address one test from each specimen geometry, material thickness and loading set.

The finite element based apparent fracture toughness estimates (S_r on Table 4.2) suggest that the **S** integral characterizes the crack driving force for inelastic and thermoinelastic material response. The agreement between S_r and J_Q derived from the load vs displacement trace for the isothermal tests (Table 4.2 blocks A and C) adds credibility to the contention that S_r for the thermal gradient tests defines the apparent fracture toughness for initial crack extension. The agreement between S_r for the isothermal cases and for the thermal gradient cases (Table 4.2 blocks A and B or blocks C and D) demonstrates that apparent fracture toughness J_Q can be considered a material parameter.

The general agreement between the finite element based S_r and J_r computed by the EPRI procedures on Table 4.2 confirms the EPRI estimates for these simple loadings. (This agreement is expected since the EPRI parameters are based on finite element analyses, see Kumar et al. [1981].) The differences between the linearly

superimposed elastic stress field in the EPRI procedures and the complete thermal gradient stress interaction in the S_r calculation may cause the discrepancies for the thermal gradient cases.

4.6 Limitations and Experimental Improvements

This test program addressed the effect of a thermal gradient on fracture for a few limited cases. The program demonstrated the conservation law nature of the S integral, a critical property from the theoretical development. The bulk of the experiments showed the thermal gradient effect on crack extension during fracture resistance tests, highlighted the requirement of accurately including the thermal gradient in calculating the fracture resistance and provided data for comparison with finite element models determining the S_r integral for eight selected fracture resistance tests.

These few tests did not fully investigate the thermal gradient effect on fracture but rather provided limited experimental evidence supporting the S integral. Many more tests on various materials, geometries and thermomechanical loading histories would have been required to complete the experimental confirmation of S.

An improved experimental instrumentation package substituting Moire interferometry for the strain gage rosette grid would have provided the data necessary for an experimental calculation of the S integral. Working directly from the planar displacement field, the displacement gradients could have been estimated throughout the crack tip vicinity. These displacement gradients, in turn, would provide the strain field which with the temperatures and the plasticity model in Appendix A1 would have produced the stress field. With these fields known, equation (4.1) would have computed the S integral directly from the experimental data.

5

Computational Implementation and Verification

Many analysis and design studies rely on the finite element method (FEM) for information concerning complex loading on intricate structures and components. Certainly, all but the most simple thermomechanical fracture problems require FEM to estimate the displacements, stresses and strains in cracked bodies. Hence, for the **S** integral to be applicable to engineering design and analysis situations, it needs to be a derivative of finite element results.

This Chapter demonstrates that the **S** integral can be calculated from finite element output information. The first section identifies and defines the necessary quantities for the calculation in terms of the finite element equations. Then the second section presents the algorithm for calculating the **S** path domain independent integral as a FEM postprocessor using nodal and gauss point information.

Two sets of computational experiments validate the **S** integral as a thermoinelastic fracture characterization parameter. The first exercise verifies the conservation law nature of **S**. This important property developed theoretically in Chapter 3 and proved experimentally in Chapter 4 is demonstrated computationally in Section 5.3. The second exercise computes the x component of **S**, S_x, for eight of

the fracture toughness experiments as discussed in Chapter 4. The four isothermal cases (where S_x identically equals J_x) verify that the computationally generated S_x matches the value obtained from the load vs displacement trace. The agreement between S_x, calculated at initial crack extension, for the isothermal cases and for the thermal gradient cases supports **S** as the crack driving force for thermoinelastic material response. These computational exercises validate the path domain independent integral **S** as a characterizing parameter for thermoinelastic fracture.

5.1 FEM Implementation

The computational investigation of S revolves around finite element results generated by a single element type, a nine node two dimensional quadrilateral. Using this element in the FEAP program (Taylor [1977], Zienkiewicz and Taylor [1989]), supplies results from the conservation law and fracture toughness computational investigations. The element includes routines to provide the nodal and gauss point output information required in the S_x path and domain integral calculations.

The nine node quadrilateral element analyzes two dimensional heat conduction and plane stress or plane strain problems including uncoupled thermoinelasticity. The element models isotropic thermal strain,

$$\begin{pmatrix} \epsilon_{11}^{th} \\ \epsilon_{22}^{th} \\ \epsilon_{12}^{th} \\ \epsilon_{33}^{th} \end{pmatrix} = \alpha\,\theta \begin{pmatrix} 1 \\ 1 \\ 0 \\ 1 \end{pmatrix} \tag{5.1}$$

where ϵ_{ij}^{th} are the components of the thermal strain in the vector form suitable for FEM implementation. α is the isotropic thermal expansion coefficient and θ is the temperature change from some strain free temperature. This formulation neglects any temperature change due to a time varying strain field and hence addresses the uncoupled strain temperature case.

The constitutive model uses the total strain decomposition from equation (3.7a), (i.e., $\epsilon = \epsilon^e + \epsilon^{th} + \epsilon^p$), to separate the strain into elastic, thermal and plastic portions. The model invokes elastoplastic evolution equations (see Chapter 3, Table 3.1 and Appendix A1) with linear isotropic and kinematic hardening to compute the plastic strains. For simplicity, the model assumes that all constitutive parameters are independent of temperature. Simo and Hughes [1988] presents the algorithmic details included in the constitutive model.

The element employs the standard nine node Lagrangian isoparametric shape functions (see, for example, Hughes [1987]) to interpolate the displacement field within each element. This formulation permits modeling the $1/\sqrt{r}$ strain singularity near the crack tip for elastic fracture mechanics and the $1/r$ singularity for perfect plasticity as discussed by Barsoum [1977].

The model strain singularity results from collapsing one side of the quadrilateral onto the singular point, thus forming a triangle. The midside node on each side emanating from the singular point is located at one quarter the distance from the singular point to the nonsingular corner node. Enforcing displacement compatibility of all the nodes collapsed onto the singular point produces the $1/\sqrt{r}$ strain singularity for elastic fracture mechanics. However, permitting the displacements of all the individual nodes collapsed onto the singularity point to vary with respect to one another yields the $1/r$ strain singularity for perfectly plastic fracture. (Hutchinson [1968] and Rice and Rosengren [1968] present the asymptotic strain singularity at the crack top for a general hardening material.)

Within the plane strain formulation the element uses a three field method to capture the incompressible nature of plastic flow. As suggested by various FEM researchers (for example, Simo et al. [1985], Hughes [1987], Simo and Hughes [1988] and Zienkiewicz and Taylor [1989]), solving for the dilatational strain and

the pressure fields within the element in addition to the nodal displacements removes the difficulty of solving the displacement equations associated with incompressible behavior. The plane strain option employs dilatational strain and pressure fields within the element that are linear in the isoparametric space variables.

The element supplies the variable values to compute the S_x integral. The path integral calculation requires information at the FEM nodes and the domain integral calculation demands information at the gauss integration points within the elements. The element outputs the variable values listed on Table 5.1 to calculate the x component of the thermoplastic **S** integral defined by equation (4.1) and rewritten here in index notation (where i, j, k, l range over 1,2,3).

$$S_x = S_x^{\text{path}} + S_x^{\text{domain}}, \tag{5.2a}$$

$$S_x^{\text{path}} = \int_{\partial\Omega} \left\{ \frac{1}{2}\epsilon^e_{ij}\big|_{n+1}\sigma_{ij}\big|_{n+1} + q_{ij}\big|_{n+1} D^{-1}_{ijkl}\ (q_{kl}\big|_n - \frac{1}{2}q_{kl}\big|_{n+1}) + (\epsilon^p_{ij}\big|_{n+1} - \epsilon^p_{ij}\big|_n)\,\sigma_{ij}\big|_{n+1} \right\} n_x - u_{i,x}\big|_{n+1}\sigma_{ij}\big|_{n+1}\, n_j \, d\Gamma, \tag{5.2b}$$

$$S_x^{\text{domain}} = \int_{\Omega} \alpha\,\sigma_{ij}\big|_{n+1}\,\delta_{ij}\,\theta_{,x}\big|_{n+1} + \sigma_{ij}\big|_{n+1}\,\epsilon^p_{ij,x}\big|_n - q_{ij}\big|_{n+1} D^{-1}_{ijkl}\ q_{kl,x}\big|_n \, d\Omega, \tag{5.2c}$$

where δ_{ij} is the Kronecker delta. For the path integral; the term $\frac{1}{2}\epsilon^e_{ij}\big|_{n+1}\sigma_{ij}\big|_{n+1}\, n_x$ is the stored energy contribution, $q_{ij}\big|_{n+1} D^{-1}_{ijkl}\ (q_{kl}\big|_n - \frac{1}{2}q_{kl}\big|_{n+1})\, n_x$ is the hardening contribution including both the potential and dissipation, $(\epsilon^p_{ij}\big|_{n+1} - \epsilon^p_{ij}\big|_n)\,\sigma_{ij}\big|_{n+1}\, n_x$ is the plastic dissipation contribution and $u_{i,x}\big|_{n+1}\sigma_{ij}\, n_x$ is the traction energy contribution. Similarly, for the domain integral: $\alpha\ \sigma_{ij}\big|_{n+1}\,\delta_{ij}\,\theta_{,x}\big|_{n+1}$ is the thermal gradient energy contribution, $\sigma_{ij}\big|_{n+1}\,\epsilon^p_{ij,x}\big|_n$ is the plastic strain gradient dissipation contribution and $q_{ij}\big|_{n+1} D^{-1}_{ijkl}\ \ q_{kl,x}\big|_n$ is the hardening gradient dissipation contribution.

The nodal quantities listed on Table 5.1 are smoothed from the variables computed at the element's 3 x 3 gauss integration points by the least squares projection method as suggested by Simo [1988] and Zienkiewicz and Taylor [1989]. The method minimizes the functional,

$$F = \int_{\Omega} ||\hat{\sigma} - \sigma||^2 \, d\Omega \,, \tag{5.3}$$

where $\hat{\sigma}$ is the stress field obtained from the nodal stress projections and σ is the stress field generated by the stresses calculated at the 3 x 3 gauss points in each element. The implementation uses the row sum lumping technique to diagonalize the projection matrix.

Table 5.1
FEM Output Information for S_x Integral

Nodal Output		Gauss Point Output	
Spatial Coordinates	$x_i \big\|_{n+1}$	Weighted Area	A
Displacements	$u_i \big\|_{n+1}$	Thermal Gradient	$\theta_{,x} \big\|_{n+1}$
Temperature Diff	$\theta \big\|_{n+1}$	Stresses	$\sigma_{ij} \big\|_{n+1}$
Displacement Grad	$u_{i,x} \big\|_{n+1}$	Hardening Variables	$q_{ij} \big\|_{n+1}$
Stresses	$\sigma_{ij} \big\|_{n+1}$	Plastic Strain Grad	$\epsilon^p_{ij,x} \big\|_n$
Plastic Strains	$\epsilon^p_{ij} \big\|_{n+1}$, $\epsilon^p_{ij} \big\|_n$	Hardening Var Grad	$q_{ij,x} \big\|_n$
Hardening Variables	$q_{ij} \big\|_{n+1}$, $q_{ij} \big\|_n$		

The gauss integration point information falls into two groups: variables that are computed within the standard element formulation (area, stresses, thermal gradient and hardening variables), and variables that must be computed from gauss point information (the gradients of plastic strain and hardening variables). The element formulates the variables in the first set from readily available information.

The gradient variables in the second set are calculated using the derivatives of the gauss point shape functions. The gauss point shape function are unity at a specific gauss point and zero at the other gauss points. The derivatives with respect to global space directions are calculated by the chain rule (see, for example, Hughes [1987]), from the isoparametric shape functions.

With these enhanced output routines, the nine node element provides the information necessary to calculate the x component of the S integral.

5.2 S_x Postprocessor Calculation

All the finite element analysis output information listed on Table 5.1 combines with the material properties to calculate the S_x integral. The postprocessor algorithm uses equations (5.2 b,c) to form the path and domain integrals.

The postprocessor developed herein assumes that the integration contour runs among element boundaries and includes a finite number of complete elements. The path integration algorithm assumes that the quadrilateral's midside nodes, along the integration path, fall exactly midway between the corner nodes. This linear side with central midside node restriction only applies to element edges included in the path contour and not generally throughout the mesh. The domain integral calculation only restricts the domain to complete elements for computational ease.

The total path integral, S_x^{path}, is comprised of contributions from each element edge that defines the contour,

$$S_x^{path} = \sum_{a=1}^{n_{edges}} S_x^{path}\Big|_a \quad , \tag{5.4}$$

where $S_x^{path}\Big|_a$ is the S_x contribution for a single element edge along the contour defined by three nodes.

Calculating the single element edge contribution begins by computing the thermoplastic Noether quantity akin to the energy momentum tensor at each node.

This quantity at a given node is exactly the path integrand from equation (5.2b),

$$\mathrm{E}_n = \left\{ \frac{1}{2}\epsilon^e_{ij}\big|_{n+1}\sigma_{ij}\big|_{n+1} + q_{ij}\big|_{n+1} D^{-1}_{ijkl} \left(q_{kl}\big|_{n+1} - \frac{1}{2}q_{kl}\big|_n\right) + \left(\epsilon^p_{ij}\big|_{n+1} - \epsilon^p_{ij}\big|_n\right)\sigma_{ij}\big|_{n+1} \right\} n_x - u_{i,x}\big|_{n+1}\sigma_{ij}\big|_{n+1} n_j \quad , \qquad (5.5)$$

where n_i are the components of the right hand outward unit normal vector to the path contour along the element edge. The node numbers defining each element side must conform to the counter-clockwise path integration convention where the integration domain lies to the left of an observer moving along the integration path in the direction of integration.

The E_n values from each node along the element edge and the edge length combine to form the element edge contribution to S_x. With three nodes defining the element edge, the algorithm uses the trapezoidal rule to calculate $\mathrm{S}_x^{\mathbf{path}}\big|_a$,

$$\mathrm{S}_x^{\mathbf{path}}\big|_a = \frac{1}{6}\left\{ \mathrm{E}_{\mathbf{corner\ 1}} + 4\,\mathrm{E}_{\mathbf{midside}} + \mathrm{E}_{\mathbf{corner\ 2}} \right\} ds \quad , \qquad (5.6)$$

Where *ds* is the distance along the element edge.

The domain integral calculations sums the contribution to $\mathrm{S}_x^{\mathbf{domain}}$ from each element contained within the contour integration path (in the left hand sense as previously discussed). Hence,

$$\mathrm{S}_x^{\mathbf{domain}} = \sum_{e=1}^{n_{\mathbf{elements}}} \mathrm{S}_x^{\mathbf{domain}}\big|_e \quad . \qquad (5.7)$$

where $\mathrm{S}_x^{\mathbf{domain}}\big|_e$ is the S_x contribution for a single element.

The domain integration uses 3 x 3 gauss quadrature to compute the area integral $\mathrm{S}_x^{\mathbf{domain}}\big|_e$. Hence,

$$\mathrm{S}_x^{\mathbf{domain}}\big|_e = \sum_{l=1}^{9} \mathrm{S}_x^{\mathbf{domain}}\big|_l \, j_l \, w_l \quad , \qquad (5.8)$$

where j_l is the Jacobian value at gauss point l from the area mapping, w_l is the gauss point weight and $S_x^{\text{domain}}|_l$ is the domain integrand evaluated at the gauss point l. The weighted area from Table 5.1 provides the product $j_l\, w_l$ at each gauss point within the element.

The domain integrand is calculated directly from equation (5.2c),

$$S_x^{\text{domain}}|_l = \alpha\, \sigma_{ij}|_{n+1}\, \delta_{ij}\, \theta_{,x}|_{n+1} + \sigma_{ij}|_{n+1}\, \epsilon^p_{ij,x}|_n - q_{ij}|_{n+1} D^{-1}_{ijkl}\, q_{kl,x}|_n \quad , \tag{5.9}$$

using the values from gauss point l and the material properties α and D_{ijkl}.

The postprocessor algorithm discussed in this section calculates the x component of the **S** integral from finite element results. The calculations follow directly from the integral development in Chapter 3. Subsequent sections demonstrate the conservation law nature of **S** and the thermoinelastic fracture characterization nature of the integral.

5.3 Conservation Law Computational Verification

A computational investigation verifies the path domain independent **S** integral as a conservation law for thermoinelastic material response. Using finite element results and the postprocessor algorithm, the calculated x component of **S** approximately equals zero for any closed contour containing no singularities. The computational approximation ($S_x = 0$) improves as modeling refinements improve the FEM solution accuracy.

This computational experiment investigates S_x for a dogbone shaped region loaded under displacement and temperature control. The prescribed loading results in a well developed plastic zone through the thin section of the dogbone while the

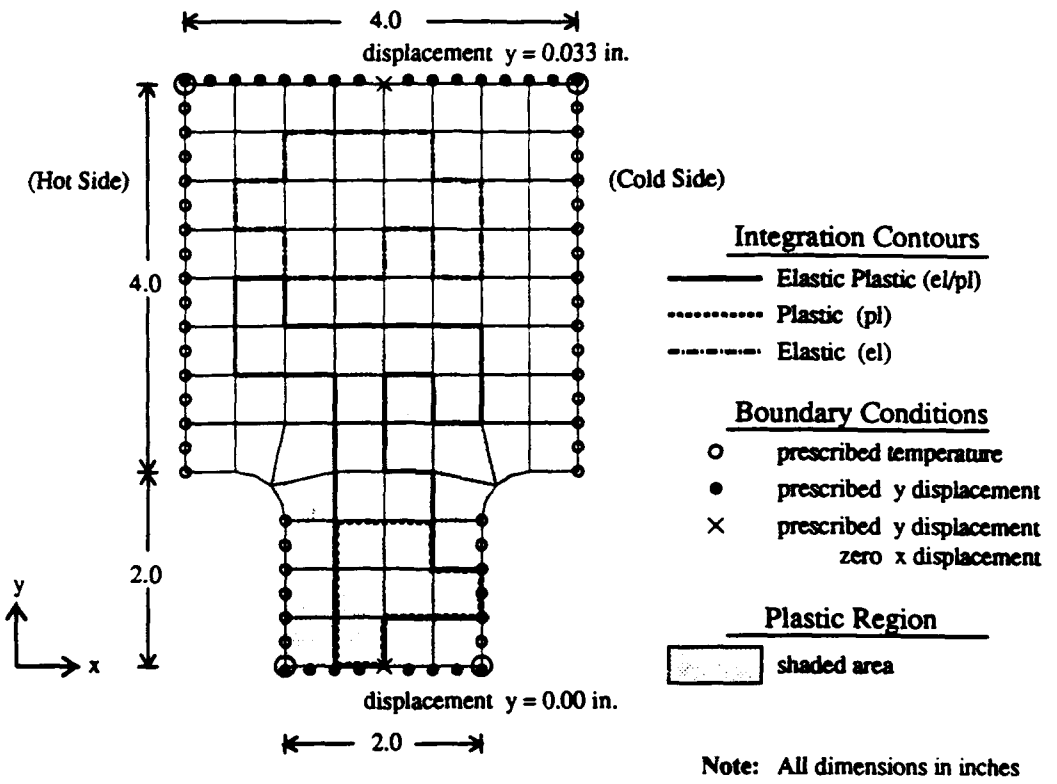


Figure 5.1 *Base case dogbone finite element specimen with three* S *integration contours.*

majority of the thick section remains elastic. Figure 5.1 presents the base case model and shows the elastic and plastic zones.

The dogbone model carries a 50°F/in. thermal gradient across the width. The displacement at the top edge is set to 0.0330 in. to develop the plastic zone in Figure 5.1. The model uses 320 elements. All but the six elements at the transition region are square with 0.50 in. sides. The material model parameters are presented on Table 5.2. The base case model uses the plane strain material relations. The temperature and displacement loading is applied over 20 equal load steps for the base case model.

This investigation examines the S_z integral for three contours; one entirely contained in the plastic zone, one in the elastic region and one including both

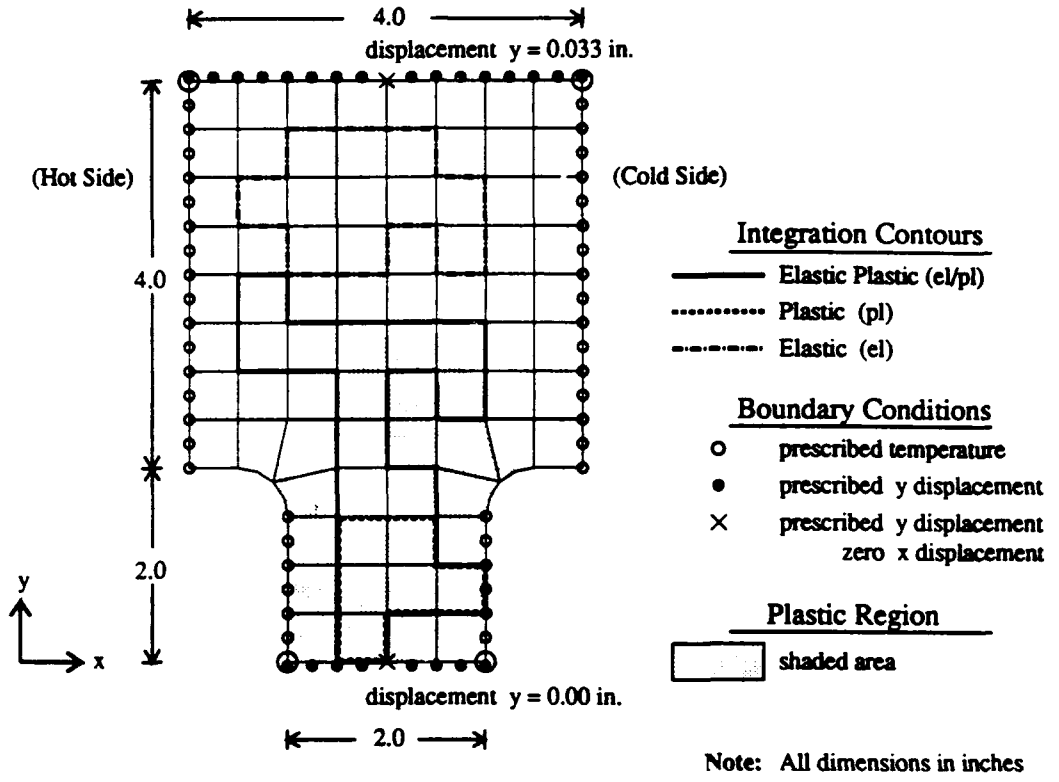


Figure 5.1 *Base case dogbone finite element specimen with three* **S** *integration contours.*

majority of the thick section remains elastic. Figure 5.1 presents the base case model and shows the elastic and plastic zones.

The dogbone model carries a 50°F/in. thermal gradient across the width. The displacement at the top edge is set to 0.0330 in. to develop the plastic zone in Figure 5.1. The model uses 320 elements. All but the six elements at the transition region are square with 0.50 in. sides. The material model parameters are presented on Table 5.2. The base case model uses the plane strain material relations. The temperature and displacement loading is applied over 20 equal load steps for the base case model.

This investigation examines the S_z integral for three contours; one entirely contained in the plastic zone, one in the elastic region and one including both

elastic and plastic areas. The S_x integral approximately equals zero for each of these contours. Comparing the calculated S_x integral to the largest individual contribution from any path or domain integral measures the error in the computed S_x value. Figure 5.1 also shows the three integration contours in relation to the plastic zone in the model.

For the base case model, the S_x integral remains less than approximately 2.0% of the largest contribution for any of the three arbitrary contours. Table 5.3 presents the contributions to each integral.

Table 5.2
Constitutive Parameters for Representative Material

Young's Modulus	$E = 10{,}000\ ksi$	Yield Point	$\bar{\sigma}_y = 40.0\ ksi$
Poisson's Ratio	$\nu = 0.32$	Isotropic Hard Coef	$K = 600\ ksi$
Thermal Exp Coef	$\alpha = 13 \times 10^{-6}/°F$	Kinematic Hard Coef	$H = 600\ ksi$
Strain Free Temp	$T_0 = 75\ °F$		

Three variations on the base case comprise a parameter study on load step size, mesh refinement and material relation. The first variation considers 50 load steps instead of 20, the second variation uses a model with 1280 elements instead of 320 and the third variation invokes the plane stress material relations instead of plane strain. These cases demonstrate the robustness of the S integral computation for the conservation law.

Using the base case model but applying the same temperature and displacement load in 50 equal steps instead of 20 examines the effect a finer loading procedure has on the S_x integral. The finer loading produces minor changes in the computed history dependent plastic strains and hardening variables. The modifications alter the S_x value for the combined elastic and plastic zone contour only slightly as shown on Table 5.3.

Table 5.3
Conservation Law Integral Contributions

Case[1]	Contour	Path Integral Values[2] W_p (ksi • in)	PS_p (ksi • in)	H_p (ksi • in)	T_p (ksi • in)	S_x^{path} (ksi • in)	Domain Integral Values[3] θ_d (ksi • in)	PS_d (ksi • in)	H_d (ksi • in)	S_x^{domain} (ksi • in)	Total S_x (ksi • in)	Error[4]
A	el	0.02124	0.0	0.0	-.04025	**0.06149**	-.06146	0.0	0.0	**-.06146**	**0.00003**	0.05 %
A	pl	0.01688	0.00427	0.02124	0.09988	**-.05749**	-.08524	0.16643	0.02693	**0.05427**	**-.00323**	1.94 %
A	el/pl	0.06796	0.00297	0.02295	0.06951	**0.02825**	-.16864	0.17364	0.02718	**-.02619**	**0.00206**	1.26 %
B	el/pl	0.06800	0.00847	0.02129	0.06950	**0.02437**	-.16851	0.16932	0.02700	**-.02218**	**0.00219**	1.22 %
C	el/pl	0.05251	0.00217	0.02621	0.05301	**0.02789**	-.17400	0.17514	0.02957	**-.02843**	**0.00054**	0.31 %
D	el	0.01764	0.0	0.0	-.02482	**0.04247**	-.04118	0.0	0.0	**-.04118**	**0.00129**	3.13 %
D	pl	0.00400	0.00062	0.00862	0.00865	**0.00459**	-.04934	0.05892	0.01025	**-.00067**	**0.00392**	6.64 %
D	el/pl	0.03192	0.00807	0.01101	0.03478	**0.01622**	-.10540	0.10320	0.01394	**-.01614**	**0.00008**	0.08 %

Notes; 1. Cases A Base Case; 320 elements, 20 load steps, plane strain
B Load Case; 320 elements, 50 load steps, plane strain
C Mesh Case; 1280 elements, 20 load steps, plane strain
D Stress Case; 320 elements, 20 load steps, plane stress

2. *Path Integral Terms, eqn (5.2 b)*
   - W_p Stored Energy
   - PS_p Plastic Dissipation
   - H_p Hardening Contribution
   - T_p Traction Energy

3. *Domain Integral Terms, eqn (5.2 c)*
   - θ_d Thermal Gradient Energy
   - PS_d Plastic Strain Gradient Dissipation
   - H_d Hardening Gradient Dissipation

4. S_x / *Maximum Contribution* x 100 %

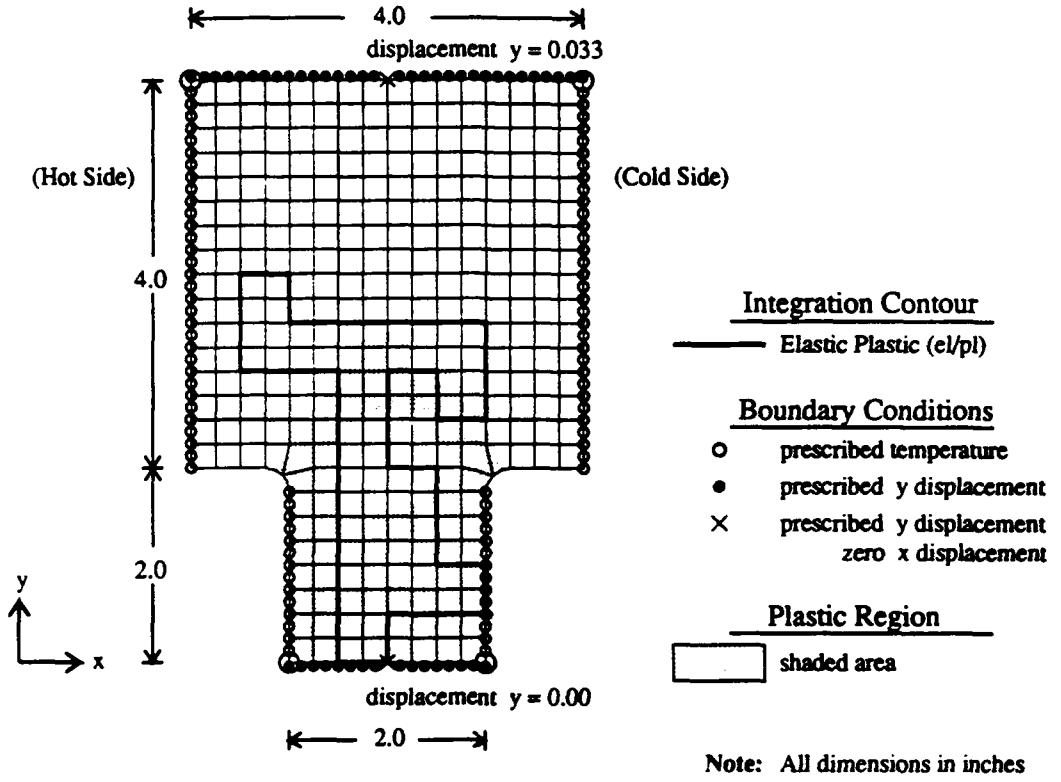


Figure 5.2 *Refined dogbone finite element mesh with* S *integration contour including elastic and plastic elements.*

Modeling the dogbone with 1280 elements, all but six 0.25 in. square, improves the computational accuracy of S_x. This model uses 20 equal steps to apply the temperature and displacement load and assumes plane strain. The accuracy of the S_x integral computed over the combined elastic and plastic contour equivalent to the base case (see Figure 5.2) improves from 1.26% to 0.31%, see Table 5.3. This improvement follows from the increased accuracy associated with the finer mesh.

Changing the base case from plane strain to plane stress material behavior alters the computed S_x value but doesn't significantly change the accuracy. The results on Table 5.3 demonstrate that both the plane strain and plane stress assumptions produce an S_x value approximately zero for an uncracked body.

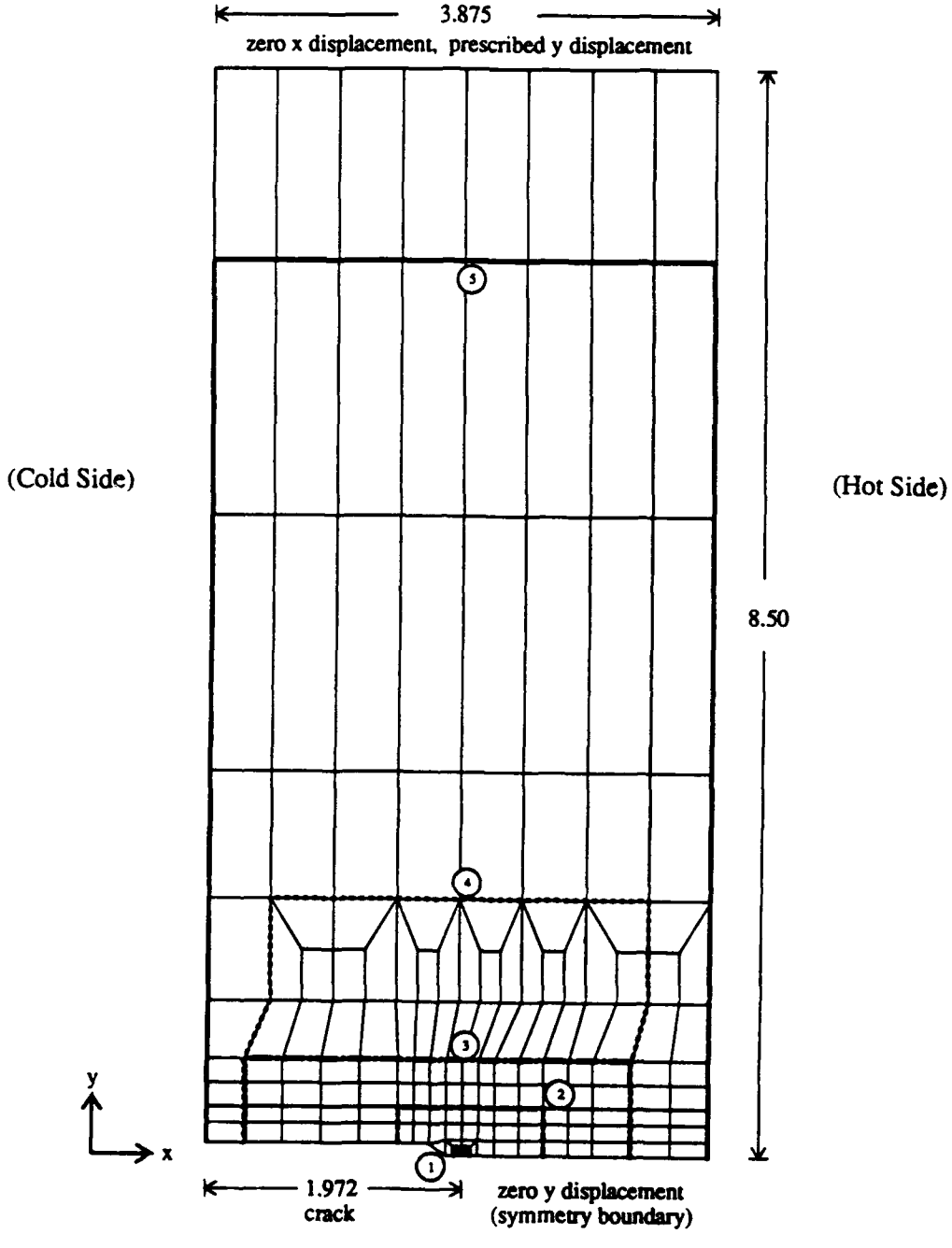
These computational experiments verify that the path domain independent S integral is a conserved quantity for uncoupled thermoinelastic material response.

5.4 Fracture Toughness Comparisons

Examining the S_x integral calculated from finite element models of eight fracture toughness experiments as discussed in Section 4.5 verifies that S characterizes the crack driving force for thermomechanical fracture. In four isothermal comparisons (where S_x equals J_x), the FEM based S_x fracture toughness estimate agrees with the value obtained from the load vs displacement trace. This confirms that S characterizes the crack driving force for isothermal inelasticity. The agreement between FEM based S_x for the four thermal gradient tests and the isothermal values confirms that S characterizes fracture under thermoinelastic material response. Table 4.2 summarizes the fracture toughness estimates.

Nonuniform sized elements comprise the models. Refining the mesh at the crack tips produces models capturing the material behavior in the crack tip region. The single edge notch (SEN) specimen model uses 709 nodes and 164 elements, see Figure 5.3. The SEN model has a crack length of 1.972 in. The center cracked plate (CCP) model with two crack tips has 841 nodes and 194 elements, see Figure 5.4. For the CCP model, the centered crack measures 0.880 in. long.

The models examine a stationary crack geometry using four nine node quadrilateral elements degenerated to triangles at the crack tips. The triangular elements have their midside nodes located at the quarter points of the sides emanating from the crack tip as discussed in Section 5.1. The crack tip nodes' boundary conditions produce the $1/r$ strain singularity which approximates the aluminum behavior. In both the SEN and CCP models the longest edge of the isosceles triangle crack tip elements is 0.08485 in.



Note: All dimensions in inches

Figure 5.3 *SEN specimen FEM mesh with five* S *integration contours.*

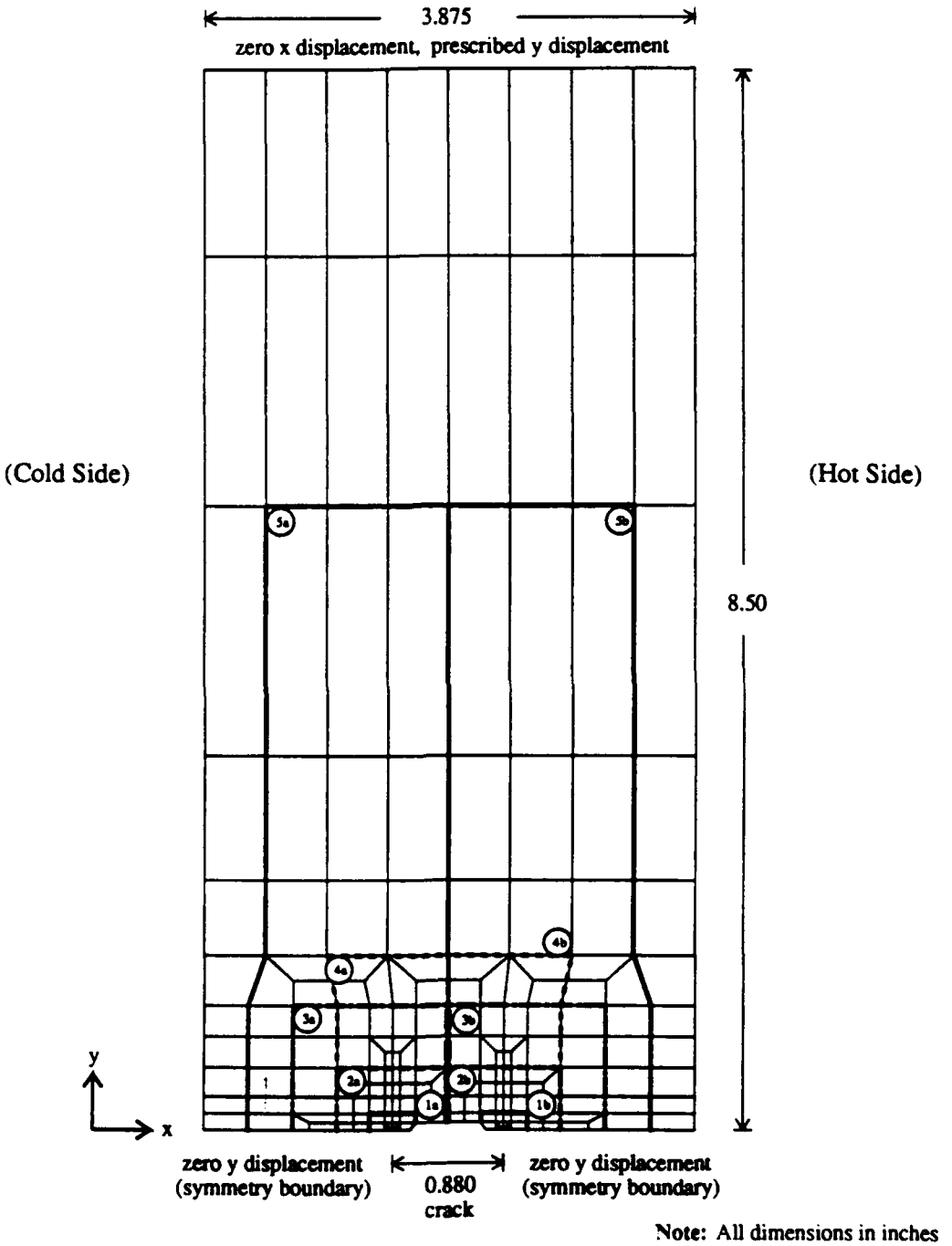


Figure 5.4 *CCP specimen FEM mesh with five* **S** *integration contours.*

The calculated S_r value differs only slightly for the five integration contours shown on Figures 5.3 and 5.4. This investigation reports the S_r integral results from contour 3 on both specimens for consistency. A later parameter study demonstrates that S_r is independent of the chosen contour. The CCP specimen model requires two separate contours, one for each crack tip, to calculate the total S_r integral since S_r gives the force for unit crack extension in the positive $+x$ direction.

The finite element models attempt to accurately approximate the material behavior. The models use the constitutive parameters determined in Appendix A1 and summarized in Table 5.4 to describe the 2024 aluminum. The 0.123 in. thick specimen tests are investigated with the plane stress material assumptions. Plane strain equations model the 0.491 in. thick 2024-T351 plate specimens.

Table 5.4
2024-T3 or T351 Aluminum Constitutive Parameters

Young's Modulus	$E = 10{,}200\ ksi$	Yield Point	$\bar{\sigma}_y = 42.3\ ksi$
Poisson's Ratio	$\nu = 0.32$	Isotropic Hard Coef	$K = 615\ ksi$
Thermal Exp Coef	$\alpha = 13 \times 10^{-6}/°F$	Kinematic Hard Coef	$H = 615\ ksi$
Strain Free Temp	$T_0 = 80\ °F$		

The four isothermal fracture toughness cases chosen for FEM evaluation have crack lengths associated with the J_Q apparent toughness approximately equal to the model values. Since only half the specimen is modeled, the FEM analyses use half the measured displacement as input.

Table 5.5 summarizes the agreement in load and crack driving force S_r (for isothermal cases S_r equals J_r) between the experimental (using equation 4.5) and stationary crack finite element results. The Table reports total values accounting for the model symmetry. Additionally, since specimen SEN-8.1 carried strain gages, a separate comparison examines the FEM strains. The excellent agreement in the

Table 5.5

Test and FEM Comparisons for Isothermal Cases

		Experiment				FEM			
Specimen	Thick (in.)	Crack Length (in.)	Disp (in.)	Tension (kips)	J_x (ksi • in.)	Crack Length (in.)	Disp (in.)	Tension (kips)	S_x (ksi • in.)
CCP-2.1	.495	0.879	0.0336	42.73	0.118	0.880	0.0336	42.78	0.129
SEN-2.1	.493	1.972	0.0134	11.99	0.153	1.972	0.0134	12.02	0.146
CCP-8.1	.123	0.860	0.0348	9.98	0.148	0.880	0.0348	9.96	0.148
SEN-8.2	.123	1.972	0.0132	2.74	0.145	1.972	0.0132	2.69	0.149

strains ϵ_{11} and ϵ_{22} between the experiment and FEM results is shown on Table A3.1 in Appendix A3.

The agreement between the experimentally determined J_x and S_x calculated from finite element results confirms that S characterizes the crack driving force for uncoupled inelasticity. The agreement also demonstrates the appropriateness of computing the S integral from finite element results.

Modeling the thermal gradient fracture toughness tests requires judicious choice of thermal boundary conditions. Though the heater and cooler assemblies add and subtract heat from the specimen, the FEM analyses use prescribed temperature boundary conditions to generate the thermal gradient. Specifying a constant temperature for all the nodes in locations covered by the heaters and coolers produces the most accurate results. Matching the recorded thermal gradient and the total thermal expansion at zero load but clamped ends establishes accurate thermal boundary conditions.

Tests in which the total crack length at initial crack extension approximately

matches the FEM crack lengths are chosen for FEM investigation. The data from these tests serve as the comparison bases for the FEM analyses. As explained in Section 4.5, choosing a load and crack length at initial crack extension (≈ 0.010 in.) implied by the J_Q analysis provides a point for comparing S_r that gives an equivalent apparent fracture toughness for the thermal gradient tests.

Table 5.6 highlights the agreement in load between the experimental and finite element results. The Table also presents the FEM input information and the S_r values. The hot and cold temperatures listed under the FEM heading on Table 5.6 each act over 0.5 in. at the specimen edge so the thermal gradient is the difference divided by 2.875 in. As with Table 5.5, Table 5.6 reports total values considering the model symmetry. Tables A3.2 though A3.5 in Appendix A3 show the agreement in the strains ϵ_{11} and ϵ_{22} between the test data and FEM results.

Comparing the crack driving force values at initial crack extension for similar specimen thicknesses from both the isothermal and thermal gradient tests demonstrates that the **S** integral characterizes fracture for thermoinelastic material response. The agreement between the fracture resistance at initial crack extension supports the fracture toughness as a thickness dependent material parameter. This combined experimental and computational investigation verifies the theoretically derived **S** integral for uncoupled thermoinelasticity as a fracture characterization parameter.

For completeness, a parameter study investigates S_r's sensitivity to integration contour choice. Table 5.7 shows that while the path and domain integral terms change dramatically, the total S_r integral differs only slightly depending on the integration contour. The large discrepancy in S_r for the smallest contour, encompassing only four elements (eight considering symmetry) at the crack tip, suggests that an accurate S_r calculation requires a more refined mesh within the contour.

Table 5.6

Test and FEM Comparisons for Thermal Gradient Cases

		Experiment				FEM							
Specimen	Thick (in.)	Crack Length (in.)	Thermal Gradient (°F/in.)	Disp (in.)	Tension (kips)	Crack Length (in.)	Hot Temp (°F)	Cold Temp (°F)	Disp (in.)	Tension (kips)	to Hot S_x (ksi • in.)	to Cold S_x (ksi • in.)	Total S_x (ksi • in.)
CCP-2.5	.492	0.901	58	0.0600	38.71	0.880	280	115	0.0600	37.83	0.044	0.078	0.122
SEN-2.4	.489	1.973	59	0.0380	1.27	1.972	280	110	0.0380	1.17	0.145		0.145
CCP-8.4	.124	0.876	64	0.0564	9.98	0.880	285	96	0.0564	9.95	0.063	0.092	0.155
SEN-8.6	.123	1.973	69	0.0284	1.23	1.972	270	72	0.0284	0.83	0.164		0.164

Table 5.7
Parameter Study on Integration Contours

	Contour	S_x^{path} (ksi • in.)	S_x^{domain} (ksi • in.)	Total S_x (ksi • in.)	Difference w.r.t. Contour 3
SEN-8.2	1	0.02952	0.05538	0.08490	-41.49 %
	2	0.13396	0.01446	0.14842	2.29 %
	3	0.14982	-0.00472	0.14510	0.00 %
	4	0.21924	-0.07542	0.14384	-0.87 %
	5	0.57851	-0.43192	0.14660	1.03 %
CCP-2.5	1a	0.03455	0.04255	0.07709	-1.05 %
	2a	0.03051	0.04706	0.07757	-0.45 %
(to Cold)	3a	0.04193	0.03598	0.07791	0.00 %
	4a	0.03456	0.04130	0.07586	-2.63 %
	5a	0.09363	-0.01452	0.07910	1.53 %
	1b	0.02101	0.02173	0.04274	-0.77 %
	2b	0.01984	0.02310	0.04294	-0.32 %
(to Hot)	3b	0.02443	0.01864	0.04308	0.00 %
	4b	0.02265	0.02020	0.04285	-0.53 %
	5b	0.03204	0.01102	0.04306	-0.04 %

The agreement of S_x calculated for the four larger contours computationally verifies the path domain independence of S proven theoretically in Section 3.4.

The finite element based S_x integral for the CCP thermal gradient specimens presents an interesting disparity with the experimental observations. The experimental data report that under a thermal gradient and axial tension load the crack tip facing the heated edge extended preferentially to the tip facing the cooled edge. However, the FEM based S_x calculations presented on Table 5.6 show that the crack driving force towards the cooled edge exceeds the crack driving force towards the heated edge. This FEM result agrees with physical intuition that suspects the tip advancing into the greater axial tension field (towards the cooled edge) to extend preferentially to the tip advancing into a lesser axial tension field.

A mislabeled crack gage channel appears to be the most likely explanation for

the disparity. Other more subtle explanations include; unequal grip pressure across the specimen width due to nonuniform thermal expansion could have produced a nonuniform axial mechanical tension, or perhaps a consistent difference in the prefatigue starter crack could have enhanced one crack tip growth.

Though this matter does not modify any of the **S integral** conclusions presented in this Chapter. it calls for further experimental investigation.

6

Summary and Conclusions

This dissertation develops the first parameter to characterize the crack driving force for general thermomechanical loadings. The energy based **S** path domain independent integral characterizes fracture for thermoinelastic material response. The **S** integral addresses the thermomechanical conditions likely to cause crack extension. As such, **S** offers a parameter to improve engineers' understanding of the strength and reliability of materials subjected to complex thermomechanical loadings.

This chapter summarizes the development and verification of the **S** integral plus it discusses the limitations on this work and suggests future research. The first section highlights the key points from the theoretical formulation presented in Chapter 3. The second section discusses the experimental and computational verification of **S** detailed in Chapters 4 and 5. The third section presents the limitations of the **S** integral and this dissertation. Finally, the fourth section suggests potential future research in the crucial area of thermoinelastic fracture mechanics exercising the **S** integral.

6.1 Formulation Highlights

The **S** path domain independent integral characterizes the crack driving force for general thermoinelastic fracture. The integral results from the Noether quantity associated with infinitesimal translation symmetry group generated by $\partial/\partial x$ acting on the discrete (algorithmic) Lagrangian density for thermoinelastic fields. As such, **S** equals the change in total energy per unit translation of a given singularity or, the energy release per unit crack advance. **S** defines a conserved quantity that must equal zero for any integration contour encompassing a crack free region.

Beginning from a valid Lagrangian density, applying Noether's theorem produces a quantity akin to the energy momentum tensor for elasticity. Using Green's theorem on the integral conservation law associated with the Noether quantity produces the **S** integral. For a region *without any* singularities, the **S** integral must equal zero. For a region *with* singularities, **S** is the total force on the singularities within the integration region per unit translation of these singularities.

With this procedure to generate the path domain independent **S** integral, Chapter 3 develops the valid algorithmic Lagrangian densities and **S** integrals for two quasi static cases: uncoupled thermoinelasticity and fully coupled linearized thermoinelasticity. The quasi static restriction assumes that material inertia energy can be neglected in the total energy description. The uncoupled case assumes the temperature field is given and known *a priori* and therefore is not a variable in the Lagrangian density. This is the typical case for most metals and for quasi static loadings. The fully coupled case includes the temperature/heat equations as Euler Lagrange equations associated with the Lagrangian density. This case, where a time varying strain field induces a temperature change, addresses certain composite materials and impact loadings. Equations (3.24) and (3.44) present the **S** integrals for uncoupled and coupled thermoinelasticity.

6.2 Experimental and Computational Verification

A limited experimental program and an associated computational investigation verifies that the **S** integral characterizes thermoinelastic fracture. The verification program addresses the two critical aspects of the **S** integral: that it is a conservation law for thermoinelasticity and that it characterizes the crack driving force. Experiments on a crack free aluminum sheet and finite element results from a dogbone specimen verify **S**'s conservation law nature. Fracture resistance tests suggest that **S** equals the crack driving force. The fracture resistance experiments consider two specimen geometries, single edge notch and center cracked plate, and two specimen thicknesses, nominally one half and one eighth inch, for a single material, aluminum 2024. Though far from an extensive program, these physical and computational investigations verify the **S** integral.

The theoretical development in Chapter 3 proves that **S** must equal zero for a crack free region. Three experiments and a single computation study verify this crucial aspect. The experiments consider a 24 in. x 3.875 in. x 0.123 in. aluminum 2024 sheet under three loading conditions: elasticity, thermoelasticity and thermoinelasticity. The finite element investigation considers a dogbone specimen loaded into the inelastic range by axial tension and an imposed thermal gradient.

The physical experiments verify that **S** remains approximately zero for thermoinelastic material response as shown in Section 4.4. The final thermoinelastic loading places the specimen in tensile yield then unloads and reloads twice with increasing total displacement on each reloading. For all of the load regimes, elastic, thermoelastic and thermoinelastic with unloading and reloading, the x component of **S** remains less than 15% of the maximum path integral component comprising the total integral. As discussed in Section 4.4, the error probably results from the coarse strain gage grid instrumenting the specimen.

The finite element investigation shows that **S** equals zero for an extensive inelastic zone. Modeling a dogbone specimen, the finite element analysis imposes a thermal gradient across the specimen then loads the dogbone until the entire neck section yields. Computing the x component of **S** to equal zero in the elastic, inelastic or an arbitrary region demonstrates the robustness of the computational evaluation of the **S** integral conservation law. Section 5.3 details the parameter study that expands the plane strain base case investigation to include load and mesh refinements and plane stress.

The fracture resistance experiments confirm that **S** gives the crack driving force for thermoinelastic fracture. Under isothermal conditions **S** equals the classic **J** integral. For thermal gradient conditions, when the **J** integral loses validity, the **S** integral still yields the crack driving force. At initial crack extension the **S** integral equals the material's apparent fracture toughness expressed as J_Q.

Isothermal test results verify the material's fracture toughness. At initial crack extension, evaluating the **J** integral directly from the load displacement trace or computing the **S** integral from finite element results yields approximately the same critical fracture toughness value for the aluminum 2024 specimens as the Damage Tolerance Design Handbook [1983] reports. Sections 4.5.2 and 5.4 present complete details. This confirms both the experimental **J** integral evaluation and the computational **S** integral calculation.

Under combined thermal gradient and tension loading, the **J** integral evaluated from the load vs displacement trace implies a fracture toughness value far below the published toughness. However, the **S** integral computed from finite element results continues to yield a toughness value approximately equal to the published value. This confirms that for thermoinelastic fracture the **S** integral gives the crack driving force. For thermomechanical loads the **S** integral characterizes fracture.

6.3 Limitations on S

The limitations on the scope of this dissertation far exceed the limitations on the S integral. Only a few restrictions limit the S integral, all emanating from the theoretical foundations. This dissertation represents only a limited investigation into the applicability of S.

Naturally, S is limited to those regimes addressed in the theoretical development. Since S is based on the Lagrangian describing thermoinelastic systems, S is limited to quasi static systems. Beginning from the Hamiltonian including material inertia would generate a path domain independent integral including true dynamic effects. Altering the body force in S to include the material inertia would address the dynamic fracture question but not robustly since the Lagrangian would not produce the correct Euler Lagrange equations for dynamics.

Since S for fracture gives the energy release due to extending the singularities within the integration region, S's applicability to crack initiation from notches is not well founded. The strain concentration at a notch is not a singularity so evaluating S around a closed contour containing the notch yields the force per unit extension of the notch in a self similar manner. A crack emanating from a notch has a geometry drasticly different from the notch geometry.

Finally, since the S integral is based on continuum mechanics, S does not address noncontinuum effects present at the crack tip. The exact micromechanics of the crack tip are not considered in S. Therefore, S cannot evaluate microcracking, microvoid coalescence or void initiation immediately adjacent to the crack tip. This same restriction acts on all continuum based fracture parameters.

The scope of this dissertation represents only a limited investigation into the applicability of the S integral. Unfortunately, this work does not investigate the use of ΔS as a parameter characterizing fatigue crack growth. Since the S

integral remains valid for thermoinelastic loading, unloading and reloading, the range $\Delta \mathbf{S}$ likely characterizes fatigue crack growth. Certainly, the success of $\Delta \mathbf{J}$ in correlating low cycle fatigue crack growth rates as discussed in Section 2.2 implies that $\Delta \mathbf{S}$ will successfully address fatigue crack growth. (This is especially true since $\Delta \mathbf{J}$ is not well founded theoretically whereas $\mathbf{S}$ and $\Delta \mathbf{S}$ remain theoretically valid for unloading from an inelastic state.)

The testing program results succeed in providing limited proof for the $\mathbf{S}$ integral. A more extensive program, addressing other materials, geometries and loadings would provide further evidence supporting the $\mathbf{S}$ integral. The testing program loading only include steady state thermal gradients. Testing with cyclic or random thermal fields would add robustness to the $\mathbf{S}$ integral verification.

The finite element investigation only considers a steady state crack. Using higher level fracture elements that model crack extension would improve the fracture resistance modeling with $\mathbf{S}$. Including temperature dependent inelasticity would also enhance the computational $\mathbf{S}$ verification.

Finally, this dissertation only directly verifies the uncoupled case of $\mathbf{S}$. The fully coupled thermoinelastic $\mathbf{S}$ integral is verified simply by analogy with the uncoupled case.

The areas for future research address all these limitations.

6.4 Future Research

The fault tolerant design of critical structures and components subjected to thermomechanical loadings demands further research. The $\mathbf{S}$ integral provides a well founded parameter characterizing thermoinelastic fracture. Researching conditions leading to fracture under thermomechanical loads with the $\mathbf{S}$ integral will provide insight and understanding into the strength and reliability of materials. Though

many aspects of thermoinelastic fracture require further research, only a few general areas are mentioned here.

Fatigue crack growth under thermomechanical loadings appears as a crucial area for further S integral based research. Components including jet engine turbine blades and power plant piping experience severe nonsteady thermomechanical loads. The range ΔS will most probably characterize fatigue crack growth. Investigating combinations of steady state and cyclic (or random) thermal fields and mechanical loadings deserves attention with the **S** integral.

Naturally, continued testing will add robustness to the **S** integral confirmation. Using Moire interferometry to obtain displacements over a very fine area including the crack tip would permit direct experimental confirmation of **S**. Or, somehow measuring the total energy in a body under thermomechanical loadings would also potentially permit direct experimental **S** integral calculation. Initial testing might investigate other materials and fracture specimen geometries. Further testing might address a steady mechanical load with an increasing thermal load to achieve crack extension.

On the computational front, including crack extension elements in the finite element analysis would enhance the fracture resistance modeling. These elements would model the inelastic unloaded region behind an extending crack tip. Also, with these elements the full crack extension vs **S** resistance curve could be estimated for a given material. Further finite element investigations might include more complex temperature dependent inelasticity models to accurately capture a broader range of material behavior.

Finally, on the theoretical front, developing conservation laws and fracture parameters for the coupled thermoinelastic case including large temperature deviations from some reference temperature would be a major contribution. The fully

nonlinear thermoinelastic equations represent a formidable challenge for future continuum and fracture mechanics research.

As more structures and components use composite materials the area of coupled thermoinelasticity will gain importance. Fracture of materials that generate nonnegligible temperatures at reasonable strain rates will require investigation with the fully coupled version of the S integral. Future research might experimentally and/or computationally verify the S integral for fully coupled thermoinelastic material behavior.

This dissertation develops the S path domain independent integral that characterizes thermoinelastic fracture. This parameter is the first fracture integral to accurately provide the crack driving force for fully coupled and uncoupled thermoinelasticity. A limited experimental and computational investigation verifies the S integral as a characterizing parameter for uncoupled thermoinelasticity. Hence, this dissertation provides the first theoretically founded and verified parameter for investigation the important area of thermomechanical fracture.

Appendix A1

Aluminum Material Characterization

The experimental program discussed in Chapter 4 investigates thermomechanical fracture of aluminum 2024. This alloy, which is commonly used in airframes, (Metals Handbook [1985]), has stable material properties in the T3 and T351 heat treated states and has been the subject of prior fracture and fatigue studies (e.g., Ibrahim [1989], Hahn and Simon [1973] and Walker [1970]).

The stable constitutive properties over the test temperature range from 72° to 320°F reported in the Metals Handbook [1985] and the general availability of the alloy made it an appropriate choice for the program.

This Appendix details the material characterization of the aluminum 2024-T3 one eighth inch thick sheet and the 2024-T351 one half inch thick plate. The fracture toughness test results, the conservation law example and the finite element modeling use the material properties and the constitutive model to describe the aluminum. After enumerating the goals of the material characterization program, this Appendix describes the tests and results then discusses the constitutive models for the alloy.

A1.1 Characterization Program Goals

The material characterization program provides parameters and constitutive models that describe the stress vs strain relation for aluminum 2024 in the T3 and T351 heat treated states.

Specifically, tests determined the Young's modulus and Poisson's ratio along with the parameters for a linear isotropic plus kinematic hardening elastic plastic model and a Ramberg-Osgood model. Following recommendation from the Military Standardization Handbook [1983] and implications from previous investigations (Ibrahim [1989] and Hahn and Simon [1973]) this program assumed the aluminum 2024 behaved isotropically and was homogeneous. Furthermore, the program verified that the parameters show less than ten percent variation over the fracture test temperature range of 72° to 320°F. Hence, a single set of temperature independent parameters adequately described the aluminum constitutive behavior for the fracture and conservation law tests.

The Metals Handbook [1985] states that for aluminum 2024 the coefficient of linear thermal expansion over the temperature range for the fracture tests averages 13×10^{-6} in./in./°F. No tests in the material characterization program address this parameter. So, the thermal expansion coefficient is assumed to match the value from the Metals Handbook.

A1.2 Test Description

Fourteen tests comprised the material characterization program. The tests conformed to ASTM [1988] specifications for Young's modulus, E111, Poisson's ratio, E132, and tension properties, E8. Additional compression loading on two tests separated the isotropic from the kinematic hardening coefficients for the plasticity constitutive model.

The loading direction for all the material test specimens matched the longitudinal tension loading direction of the fracture and conservation law test specimens. Cutting all specimens from the single sheets of each the 2024-T3 and the 2024-T351 at the same time ensured that the longitudinal axis of all the specimens remained

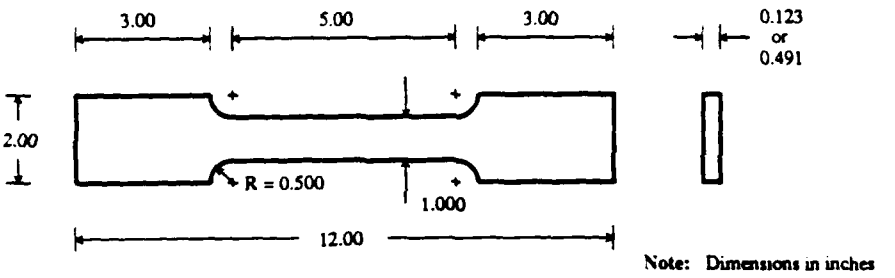


Figure A1.1 *Dogbone material test specimen geometry .*

constant along the rolling direction of the sheets. Characterizing the material by the properties in the same direction as the fracture tests' tension axis minimized error due to any real anisotropy ignored by the isotropic models.

The dogbone specimens for the material tests followed the ASTM guidelines from specification E8 with the grip ends slightly enlarged to fit in the Instron testing machine hydraulic grips. The geometry detailed on Figure A1.1 provided nominally a 4.5 in. gage length for the room temperature tests and a 2.5 in. uniform temperature gage length for the elevated temperature tests. The heating arrangement determined the smaller gage length for the elevated temperature tests.

Instrumentation within the Instron testing machine and attached to the specimens provided the data for determining the material parameters. The load cell internal to the Instron machine provided the applied force information. Strain gages supplied strain data for all tests and thermocouples monitored temperatures on the elevated temperature tests.

The material tests used the same data acquisition system as the fracture tests described in Chapter 4. The load cell and strain gages fed their data to an analog to digital converter and in turn to the DEC Vax 11-780 computer approximately once per second. The analog to digital converter provided a resolution of approximately 0.05% of full range. An analog Doric box converted the thermocouple voltage to

Fahrenheit temperature, with a one degree precision, on a light emitting display for manual recording.

The Instron internal load cell operated at the 10 kip range for the one eighth inch thick 2024-T3 specimen tests. The 50 kip range provided the necessary data for the one half inch thick 2024-T351 specimen tests. The Instron manual reported the load cell accuracy at 0.25% of the full range, however, considering the the complete data acquisition system an overall accuracy of one percent (100 lbs on 10 kips and 500 lbs on 50 kips) appeared defensible.

The strain gage rosettes and thermocouples described in Chapter 4 provided the strain and temperature data. Two CEA-13-062-WR-350 rosettes bonded in mirror images of each other on either side of the specimen supplied the axial and transverse strain information to compute the Young's modulus and Poisson's ratio. Three type J thermocouples monitored the middle 2.5 in. specimen section during the elevated temperature tests.

Visually reading the gage length between two die punch marks with a 0.01 in. resolution ruler and a 7X magnifying glass before and after tensile separation determined the total specimen separation. Before and after readings with a 0.001 in. precision micrometer at the specimen separation site provided the information for the reduction of area calculation.

As with the fracture and conservation law tests, the Instron mechanical screw axial load testing machine provided the mechanical loading. The displacement controlled tests ran with a crosshead speed of 0.05 in./min until the axial strain exceeded the one percent data system maximum strain. For the ultimate tensile strength portion of the material tests the crosshead speed was increased to 0.2 in./min since only the maximum load needed to be recorded.

For the elevated temperature tests, a simple 1200 watt air blower heated the

specimen. A 6 in. diameter acrylic cylinder 4 in. tall with an insulated top and bottom that fit around the thin dogbone section contained the hot air around the specimen. A metal T section placed in front of the blower nozzle diverted the hot air from impinging on one of the strain gages. A manually operated cool air vent on the blower grossly controlled the specimen temperature by modifying the hot air temperature.

The acrylic container limited the elevated temperature tests to a maximum of 280°F. Thus, the target temperature for the elevated temperature material tests was 260°F, somewhat below the 320°F maximum temperature for the fracture and conservation law tests, but sufficient to verify the temperature independence of the material properties over the test temperature range.

Following ASTM specifications in E111, once the three thermocouples centered over the middle 2.5 in. of the dogbone specimen had reached the desired temperature, the specimen soaked at temperature for one hour per inch thickness. This improved the likelihood of uniform temperature throughout the specimen thickness within the 2.5 in. middle section. Throughout the soaking period all three thermocouples remained within 4°F of an average temperature. However, slight variations in the cool air vent controlling the test temperature affected the average temperature by as much as 17°F.

A1.3 Test Procedure and Matrix

The material characterization program consisted of eight tests on the one half inch thick aluminum 2024-T351 plate, four at room temperature and four at an elevated temperature of approximately 260°F. Four room temperature tests and two elevated temperature tests comprised the six tests conducted on the one eighth inch thick aluminum 2024-T3 sheet. All but two of the 2024-T351 tests (one test

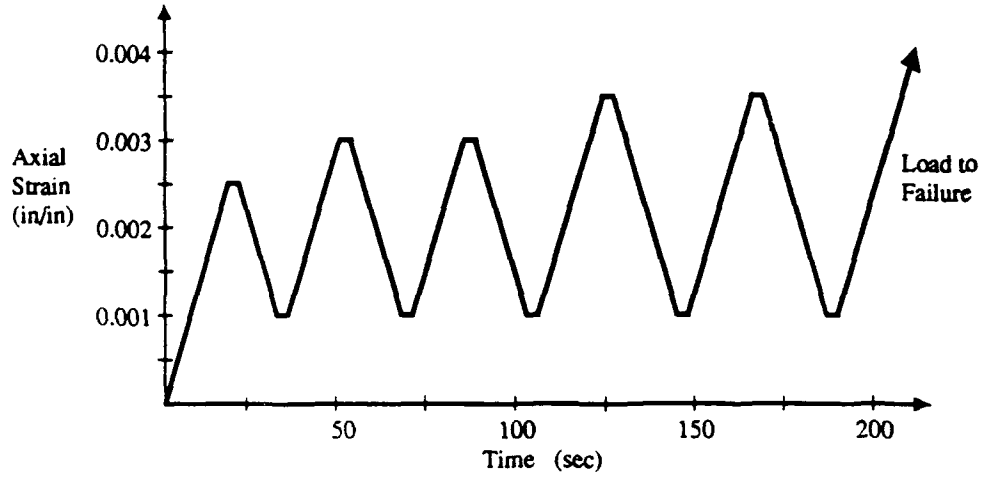


Figure A1.2 *Target Loading for determining Young's Modulus and Poisson's Ratio .*

at room temperature and one at an elevated temperature) combined the ASTM specifications E111, E132 and E8 for determining Young's modulus, Poisson's ratio as well as the tensile properties of yield and ultimate tensile strength, percent elongation and percent reduction of area.

The material property test data provided the basis for determining the Ramberg-Osgood and the elastic plastic constitutive model parameters. The final two 2024-T351 tests began with the E111 and E132 test procedures then, instead of loading to ultimate tensile strength, underwent fully reversed tension and compression strain controlled cycling with increasing strain. These tests separated the isotropic from the kinematic hardening coefficient by determining the growth of the hysteresis loop and the position of its center relative to the zero stress axis.

The elastic property determination portion of the material tests contained five elastic tension reversals at increasing strain as suggested by ASTM E111. Manually pausing then reversing the Instron crosshead direction at each strain extremum, shown on Figure A1.2, provided the necessary stress and strain data. Averaging the

Table A1.1
Material Test Descriptions

Spec.	Dimensions (in x in)	Temp. (° F)	Loading History	Comments
1/2 inch thick 2024-T351 Aluminum				
P2-1	1.001 x 0.491	70	Figure A1.2* then to failure	
P2-2	1.004 x 0.491	70	repeat	Strain gages debond at low plastic strain
P2-3	1.002 x 0.491	70	repeat	
P2-4	0.998 x 0.490	70	Figure A1.2 then fully reversed cycling	Strain gages debond after 1.5 cycles
P2-6	0.997 x 0.492	260	Figure A1.2 then fully reversed cycling	Initial imperfections produce bending
P2-7	0.999 x 0.491	245	Figure A1.2 then to failure	
P2-8	1.002 x 0.491	253	repeat	
P2-10	1.006 x 0.493	250	repeat	Test stopped at 1% strain to check gages
1/8 inch thick 2024-T3 Aluminum				
P8-1	1.001 x 0.123	70	Figure A1.2 then to failure	
P8-2	1.001 x 0.122	70	repeat	
P8-3	1.000 x 0.123	70	repeat	
P8-4	1.000 x 0.124	70	repeat	
P8-5	0.999 x 0.124	260	Figure A1.2 then to failure	
P8-6	1.001 x 0.124	277	repeat	

* Figure A1.2 presents five elastic tension cycles

axial and transverse strains from the two mirrored rosettes at each strain extremum removed any slight bending strain potentially induced by specimen misalignment. The ASTM E111 and E132 data reduction procedures were then used to calculate the Young's modulus and the Poisson's ratio respectively.

For all but the two fully reversed tests mentioned above, after the fifth elastic cycle, the crossheads continued tensioning the specimen to failure. The axial engineering stress vs strain trace up to approximately one percent strain provided the data for the constitutive modeling. For completeness, the ASTM E8 evaluation procedures determined the 0.2% offset yield strength, the nominal ultimate strength, the percent elongation and the percent reduction of area.

Table A1.1 details the test specimen original dimensions, mentions the loading history and notes any testing difficulties. The repetitions of the tests improves the statistical confidence in the results.

Table A1.2
Material Property Summary

Spec.	Temp. (°F)	E-mod. (ksi)	Pois.	σ_y(0.2%) (ksi)	$\bar{\sigma}_y$ (ksi)	H+K (ksi)	σ_{ult} (ksi)	% Elong.	Gage (in)	% Reduction of area
1/2 inch thick 2024-T351 Aluminum										
P2-1	70	10459	0.329	45.6	42.3	1031	65.8	16.7	4.51	16.9
P2-2	70	10473	0.321				65.9	18.6	4.51	16.7
P2-3	70	10410	0.323	45.1	43.3	863	65.8	16.4	4.50	16.7
P2-4	70	10577	0.322	45.0	42.3	1400 (40% H and 60% K)				
P2-6	260	10200	0.321	43.5	42.2	1412 (70% H and 30% K)				
P2-7	245	10157	0.309	44.1	41.6	1153	60.0	19.5	2.00	18.1
P2-8	253	10246	0.321	45.1	43.2	1103	60.5	19.0	2.00	18.0
P2-10	250	9854	0.312	44.4	41.7	1341				
1/8 inch thick 2024-T3 Aluminum										
P8-1	70	10394	0.332	45.7	43.3	1040	65.2	16.2	4.51	20.1
P8-2	70	10508	0.327	46.0	43.9	1331	65.9	15.6	4.50	19.5
P8-3	70	10399	0.326	45.5	42.9	1326	65.0	16.7	4.49	20.5
P8-4	70	10276	0.325	45.6	43.6	1473	64.5	16.2	4.50	20.8
P8-5	260	9645	0.318	43.5	41.1	1214	60.1	17.9	2.01	22.8
P8-6	277	9664	0.316	41.5	38.8	1271	56.6	17.2	1.98	22.2
Average		10200	0.322	44.4	42.3	1230	63.2	17.3		19.3
std. dev.		300	0.006	1.2	1.4	180	3.3	1.3		2.2

A1.4 Test Results and Constitutive Models

The material property tests indicate that a single set of constitutive parameters adequately describes the aluminum over the temperature range, 70° to 260°F. This finding agrees with information in the Metals Handbook [1985].

The ASTM E111, E132 and E8 procedures calculate the material properties. From the fourteen tests, the Young's modulus (E) averages 10,200 ksi, Poisson's ratio (ν) averages 0.32 and the 0.2% offset yield strength is 44.4 ksi. The ultimate tensile strength of the aluminum 2024 is 63.2 ksi, the elongation at failure is 17.3% and the reduction of area is 19.3%. Though not relevant to the constitutive modeling, these last three values verify the testing procedure and material behavior by their agreement with the Metals Handbook [1985]. Table A1.2 summarizes the material property test results.

In addition to the properties suggested by the ASTM specifications, three

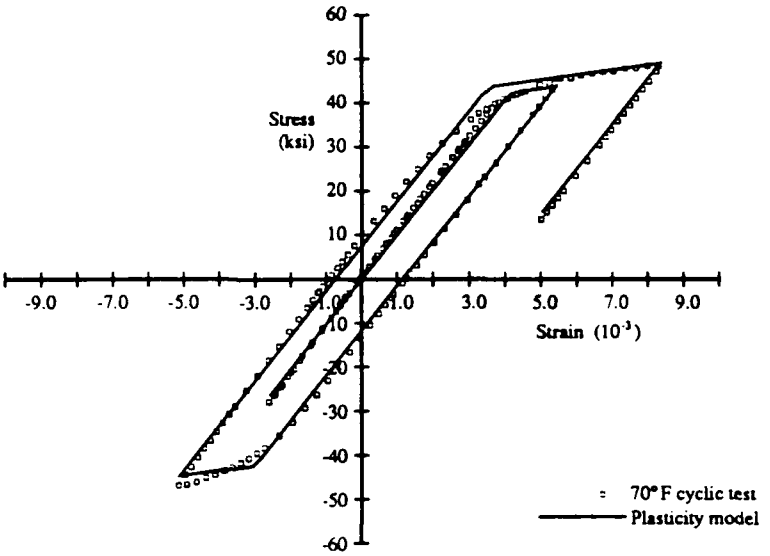


Figure A1.3 *Cyclic stress strain data and plasticity model for aluminum 2024.*

parameters obtained from the material tests define the rate independent elastic plastic constitutive model with combined linear isotropic and kinematic hardening. This model assumes the following forms for the yield function and hardening flow rules based on classical rate independent plasticity (Simo and Hughes [1989], Simo et al. [1988], Luenberger [1984] and Hill [1960]).

$$f = \|dev(\boldsymbol{\sigma}) - \boldsymbol{q}\| - \sqrt{\frac{2}{3}}(\bar{\sigma}_y + K\,\hat{q}) \qquad (A1.1a)$$

$$\dot{\hat{q}} = \dot{\gamma}\frac{(dev(\boldsymbol{\sigma}) - \boldsymbol{q})}{\|dev(\boldsymbol{\sigma}) - \boldsymbol{q}\|} \qquad (A1.1b)$$

$$\dot{\boldsymbol{q}} = -\dot{\gamma}\boldsymbol{H}\frac{(dev(\boldsymbol{\sigma}) - \boldsymbol{q})}{\|dev(\boldsymbol{\sigma}) - \boldsymbol{q}\|} \qquad (A1.1c)$$

where f is the yield function, γ is the consistency parameter such that $f \leq 0$, $\gamma \geq 0$ and $f\gamma = 0$, the superposed dot as in $\dot{\hat{q}}$ denotes time differentiation, $\boldsymbol{\sigma}$ is the stress, $\boldsymbol{q}$ is the back stress, $\hat{q}$ is the isotropic hardening scalar variable, K is the linear isotropic hardening coefficient, $\bar{\sigma}_y$ is the yield point defining the elastic to plastic transition, and $\boldsymbol{H} = H\mathbf{1}$ where H is the linear kinematic hardening coefficient.

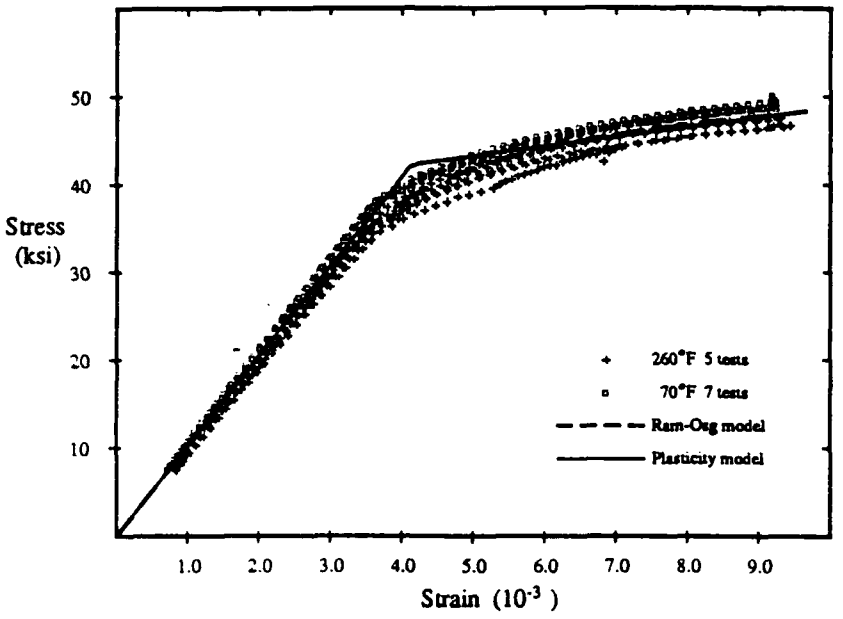


Figure A1.4 *Uniaxial stress strain data and models for aluminum 2024.*

The parameters $\bar{\sigma}_y$, K and H were obtained from visual graphical interpretation of the stress vs strain plots. The yield point $\bar{\sigma}_y$ defines the intersection of the elastic and linear strain hardening portions of a bilinear representation of the stress vs strain trace. The sum $H + K$ relates to the slope of the stress vs strain trace in the strain hardening regime. The parameters H and K can only be separated from cyclic tension and compression fully reversed tests since K governs the growing size of the hysteresis loop between the peak stresses and H shifts of the center of the hysteresis loop along the stress axis. The separation between H and K only matters for fully reversed yielding. None of the fracture or conservation law tests consider this fully reversed loading so separating H and K receives low priority in the material tests. The tests determine that the values $\bar{\sigma}_y = 42.3$ ksi. $H = 615$ ksi and $K = 615$ ksi adequately match the data. Figure A1.3 depicts the bilinear model fit to one cyclic test.

In addition to the plasticity model described above, the uniaxial stress strain

data up to approximately one percent strain generates a Ramberg-Osgood power law constitutive model. The Ramberg-Osgood model parameters enter into the EPRI J integral estimation methods detailed in Chapter 4. The Ramberg-Osgood stress strain relation is,

$$\frac{\epsilon}{\epsilon_y} = \frac{\sigma}{\sigma_y} + \alpha^{RO}(\frac{\sigma}{\sigma_y})^n \qquad (A1.2)$$

where σ and ϵ are the uniaxial stress and strain respectively, $\sigma_y = 44.4$ ksi is the uniaxial 0.2% offset yield strength and $\epsilon_y = 0.00435$ is the strain normalizing parameter computed via Young's modulus. The dimensionless Ramberg-Osgood parameters are $\alpha^{RO} = 0.42$ and $n = 11.88$. Figure A1.4 compares all of the uniaxial tension test data to this Ramberg-Osgood model and the plasticity model discussed above. Naturally the Ramberg-Osgood model cannot match unloading and reloading as it is only a nonlinear elastic constitutive model.

Appendix A2

Heater and Cooler Design

Heaters and coolers attached to the specimen edges generate the thermal gradient for the nonuniform temperature conservation law and fracture resistance tests. The thermal gradient runs along the direction of crack extension, i.e., along the 3.875 in. specimen width.

This Appendix describes the heater and cooler design and operation. The first section discusses the design constraints for the thermal gradient and capacities for the heater and cooler. The second section describes the heater assembly while the third section presents the cooler device design. The final section details the performance of the heater and cooler package on the two specimen thicknesses.

A2.1 Thermal Gradient Constraints

Naturally, the testing program demonstrates the effect a thermal gradient has on crack extension more dramatically as the thermal gradient increases. Thus, to demonstrate the greatest effect, the program seeks the maximum the thermal gradient subject to two constraints; limiting the heated edge to 320°F to preserve the temperature independent material property assumption and using tap water or compressed air to remove heat from the cooled edge.

Preserving the temperature independent material property assumption in the area surrounding the crack tip simplifies the later fracture analyses. Furthermore, limiting the maximum temperature to 320°F limits the precipitation age hardening

to negligible values for the test durations according to the Metals Handbook [1985]. The unavailability of cooling apparatus other than those based on tap water limited the cool side temperature to a minimum of about 70°F. Compressed air's small heat capacity compared to water makes it the less attractive coolant choice. (The lab would have required major modifications to provide liquid nitrogen, gaseous nitrogen, chilled water or another coolant to the testing area.)

Coolant availability limits the cooled side temperature while preserving temperature independent material properties limits the heated side temperature. Thus, the maximum thermal gradient across the 3.875 in. specimen width is approximately 65°F/in. Fortunately, preliminary finite element results indicate that approximately half the fracture resistance **S** value results from thermal gradient terms with a 65°F/in. thermal gradient.

Assuming linear one dimensional heat conduction for the aluminum specimen suffices to size the heater and cooler capacities. As noted in Chapter 4 specimen dimensions and testing machine grip design suggest generating the thermal gradient within the central 14 in. section. The 14 in. length ensures a uniform gradient, free of end effects in the region surrounding the crack. Using 190 W/m/°C for the aluminum 2024 thermal conductivity (Metals Handbook [1985]) and the half inch specimen thickness implies the following heat capacity for the heaters and coolers,

$$q = -k\ A\ \frac{d\theta}{dx} \qquad (A2.1a)$$

$$q = (-190\frac{\mathrm{W}}{\mathrm{m}\cdot^{\circ}\mathrm{C}})(0.0127\mathrm{m} \bullet 0.3556\mathrm{m})(-1412\frac{^{\circ}\mathrm{C}}{\mathrm{m}})$$

$$= 1212\ \mathrm{W} \qquad (A2.1b)$$

where q is the required heating or cooling capacity for the heaters and coolers, k is the thermal conductivity, A is the edge area for this simple one dimensional model and $d\theta/dx$ is the thermal gradient.

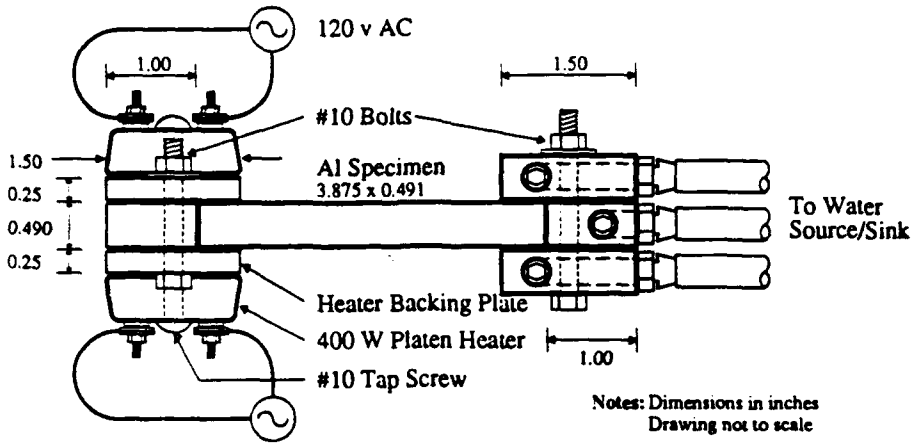


Figure A2.1 *Plan view of heater and cooler assemblies clamped onto aluminum specimen.*

Thus, to generate the 65°F/in. thermal gradient from 70°F to 320°F across the 3.875 in. specimen width, along a 14 in. length of the half inch thick specimen, approximately 1200 W of heat must be added to the hot side and subtracted from the cool side.

The overall design includes electrical resistance heaters to supply the heat input and water cooled blocks to remove the heat. The heater and cooler assemblies clamp onto the test specimen's outer 0.5 in. edges for effective heat transfer. Figure A2.1 shows a plan view of the heater and cooler assemblies clamped to a thick fracture specimen. For the thin specimen the heater spacer plate would be 0.122 in. wide instead of the 0.490 in. width shown for the thick specimen and the cooler spacer would not have a water channel.

A2.2 Heater Assembly Description

The heater assembly consists of three major components: two plates with attached heaters and a spacer plate. The backing and spacer plates are machined from aluminum for effective heat transfer. Figure A2.2 sketches the heater assembly.

Four cast aluminum 400 W electrical resistance platen heaters (6.0 in x

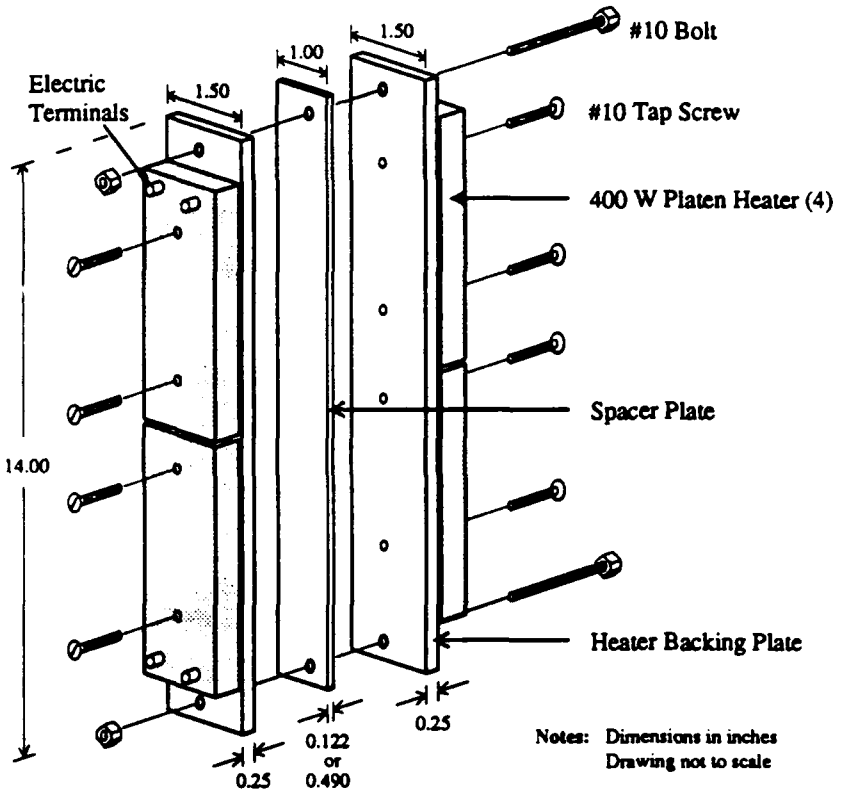


Figure A2.2 *Electric resistance heater assembly.*

1.50 in. x 0.50 in.) supply the heat. The total 1600 W capacity exceeds the 1200 W design value and permits for losses ignored in the sizing calculation. Sets of two heaters are mounted on 14.0in x 1.50in. x 0.25in. aluminum backing plates with #10 tap screws as shown on Figure A2.2. A thin layer of silicon heat sink compound between the heaters and the backing plates improves the heat transfer and reduces local "hot spots" caused by point contacts.

The spacer plate thickness is 0.001 in. thinner than the specimen to permit the clamping force to hold the heater assembly to the specimen. The spacer plate also provides a heat transfer surface against the specimen edge. Silicon heat sink compound on the front and back of the spacer plate improves the heat flow from the heater backing plates to the spacer plate and hence to the specimen edge.

Two 0.1875 in. diameter #10 bolts torqued only snug tight clamp the heater assembly onto the outer 0.50 in. specimen edge. The specimen thickness plus 0.50 in. on both the front and the back faces of the specimen provide the heat transfer surfaces along the 14.0 in. heated length. Coating the specimen edges with silicon heat sink compound improves heat flow without increasing the clamping force. Clamping the heater assembly to the specimen in this manner had less than 0.1% effect on the load displacement trace of the blank conservation test specimen.

Each set of two heaters is wired in series and connected to a 220 volt power outlet through a rheostat. The 30 ampere rheostat controls the voltage supplied to the heaters. This simple arrangement permits constant heat generation at a controlled level. The heaters effectively and reliably produce the constant heat input for the thermal gradient.

A2.3 Cooler Assembly Description

Water cooled aluminum blocks remove heat from the specimen edge opposite the heaters to generate the thermal gradient. The full cooler assembly consists of two sets of three components each. Having two cooler sets permits cooling the single edge notch specimen without interfering with the notch opening displacement by attaching separate cooler sets above and below the notch. Each cooler set consists of two cooler blocks, depicted on Figure A2.3, and a spacer plate slightly thinner than the specimen width. Each set attaches to the outer 0.50 in. of the specimen edge.

The four coolers are each 7.0 in. x 1.50 in. x 0.50 in. aluminum blocks drilled with a 0.297 in. diameter water path as shown on Figure A2.3. A $\frac{1}{8}$in. thread plug seals the water path. The path connects to the water supply and sink via $\frac{1}{8}$in. thread x $\frac{1}{4}$in. hose barb.

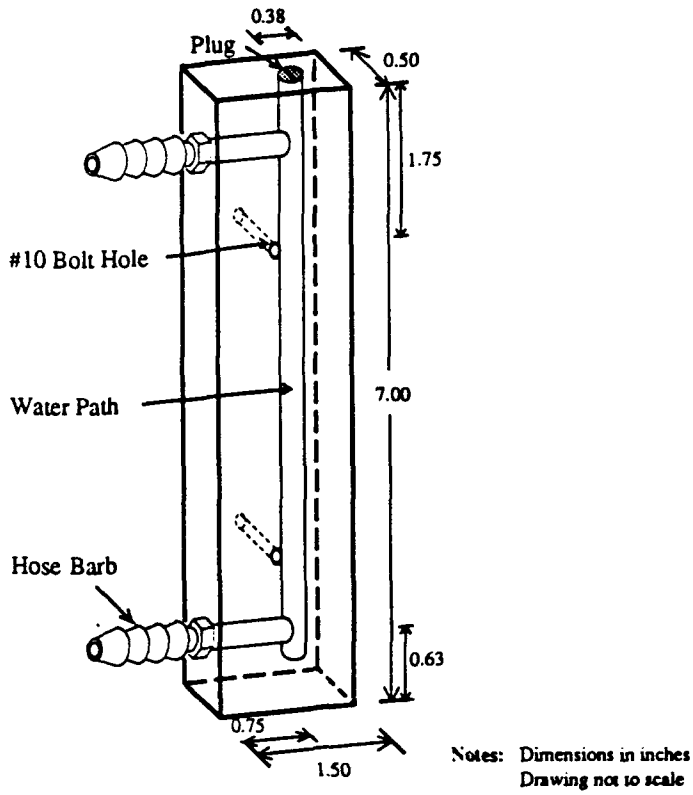


Figure A2.3 *Single aluminum cooler block that connects to tap water supply.*

The spacer blocks measure 7.0 in. long by 1.00 in wide to permit clamping to the specimen edge in the same manner as the heater assembly. The spacer blocks are 0.001 in. thinner than the specimens to allow a clamping force to hold the cooler assemblies onto the outer 0.50 in. of the specimen edge. A thin layer of silicon heat sink compound coating the spacer blocks improves the heat transfer from the specimen edge. The thick spacer block (0.490 in.) has an internal water path similar to the cooling blocks to enhance specimen edge cooling. The thin spacer block (0.122 in.) has no cooling water path.

A simple heat exchanger analysis estimates the required water flow rate to remove 1200 W from along a 14.0 in. edge. Using 4.18 kJ/kg·°C for the heat capacity of water (e.g., Chapman [1987]) and assuming a 2°F change in water temperature

along the 14.0 in. specimen edge (less than a 0.2°F/in. thermal gradient along the specimen axis) implies the following water flow rate,

$$\dot{m} = \frac{q}{c_w \, \Delta\theta} \qquad (A2.2a)$$

$$\dot{m} = \frac{1.2\,\text{kJ/sec}}{4.18 \frac{\text{kJ}}{\text{kg}\cdot^\circ\text{C}} \text{ x } 1.1\,^\circ\text{C}}$$

$$= 0.26 \text{ kg/sec} = 4.3 \text{ gal/min} \qquad (A2.2b)$$

where $\dot{m}$ is the mass flow rate, q is the heat transferred, c_w is the heat capacity of the fluid and $\Delta\theta$ is the fluid temperature change from heat exchanger inlet to outlet. This simple analysis estimates that 4.3 gal/min water flow removes 1200 W from the specimen while raising the cooling water temperature by 2°F.

Tap water at common supply pressure (approximately 40 psi) provides adequate volume, approximately five gallons per minute through the entire cooling assembly, to cool the specimen edge to the water temperature, typically 70°F. Each cooling block is individually connected to a supply and return manifold. While this creates an octopus-like collection of 1/4 inch general purpose hoses, it provides for effective cooling. The water flow runs in opposite directions in the two outer coolers of each set above and below the notch. In the central edge coolers for the thick specimens, the water flows from the specimen middle towards the grips. This flow arrangement minimizes the thermal gradient along the specimen length.

A2.4 Heater and Cooler Operations

Though simple in design the electric resistance heaters and tap water coolers produced approximately a 65°F/in. thermal gradient across the fracture specimens. The 14.0 in. long heater and cooler assemblies generated this gradient in the central section of the specimens.

The temperature field reached steady state in less than ten minutes from activating the heaters and coolers. Thermocouples on the test specimens supplied the temperature information as described in Chapter 4.

Controlling the rheostat effectively maintained the heated edge close to the 320°F temperature limit. For the 0.123 in. thick specimens running the heater rheostat at approximately 45% power generated the desired thermal gradient without exceeding 320°F on the hot edge. For the 0.491 in. thick specimens, running the rheostat at 85% power provided the testing condition.

For all specimens the cooling water operated at full flow. The cooler operations limited the generated thermal gradient. A more efficient cooling assembly would have permitted a greater thermal gradient to be generated without exceeding the 320°F hot side upper temperature limit.

The heater and cooler assemblies effectively generated the thermal gradient for the experimental program. Minor difficulties in local heat transfer created a negligible thermal gradient along the specimen axis of less than 1°F/in. The gradient remained stable for even the long duration (1 hour) fracture resistance tests.

Appendix A3

Experiment / FEM Strain Comparison

This Appendix compares the transverse ($\epsilon_{11} = \epsilon_{xx}$) and longitudinal ($\epsilon_{22} = \epsilon_{yy}$) strains measured in five experiments to those computed in the finite element analysis. All four of the thermal gradient tests and one of the isothermal tests from the eight experiments modeled with finite elements have strain data in the crack tip vicinity from the strain gage rosettes. The agreement of the experimentally measured transverse and longitudinal strain components with those computed from the finite element analysis confirms that the finite element models adequately depict the strain field near the crack tip for each case.

As described in Chapter 4, all the thermal gradient test specimens carried strain gage rosettes for strain field comparison with the finite element model used to calculate the S_r integral at initial crack extension. Also, one isothermal test specimen carried strain gage rosettes to verify the finite element model. As shown on Figures A3.1 and A3.2, the strain gage rosettes covered an area surrounding the crack tip. This pattern offered ten (for the CCP specimens) or twelve (for the SEN specimens) comparison points for the strain field.

The comparison verifies the accuracy of the finite element model for the tests examined. Since the strain fields calculated by the finite element analyses agree with the strain fields measured in the experiments, the S_r integrals calculated from the finite element results accurately represent the actual crack driving forces in the experiments.

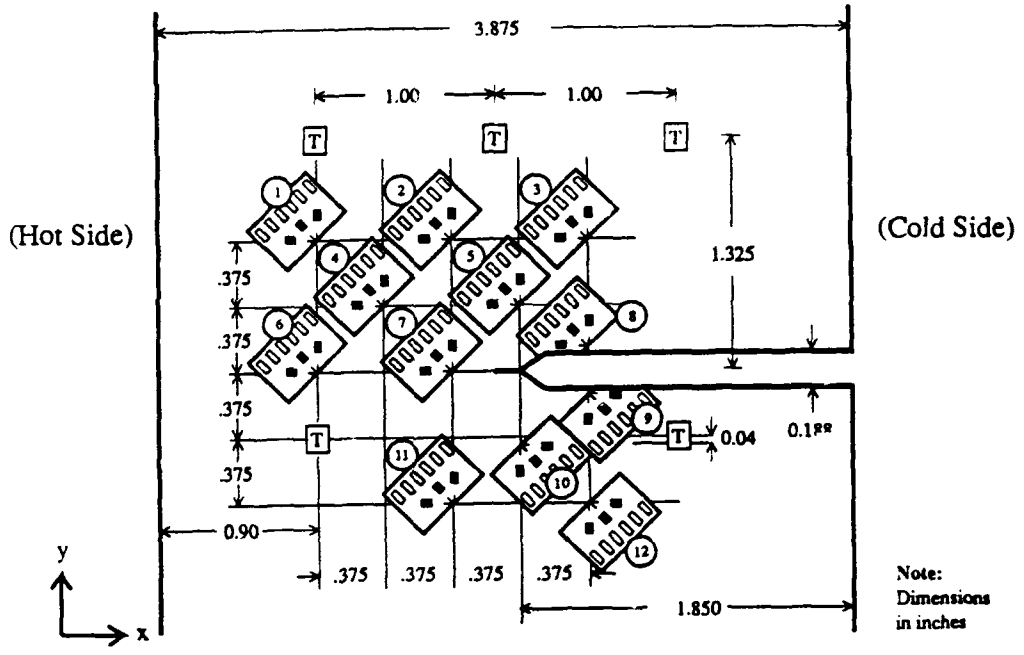


Figure A3.1 *SEN specimen strain gage numbering*

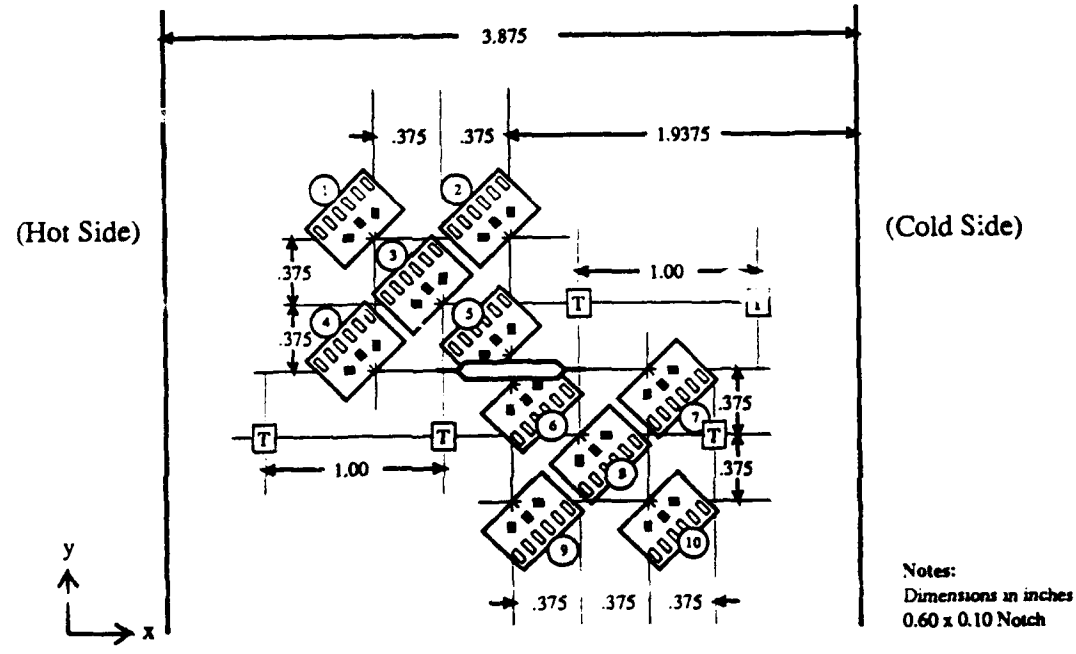


Figure A3.2 *CCP specimen strain gage numbering*

Table A3.1
SEN-8.1 / FEM Strain Comparison

Gage	ε_{xx}^{Exp}	ε_{xx}^{FEM}	Error [1]	ε_{yy}^{Exp}	ε_{yy}^{FEM}	Error
1	-.0002055	-.0002036	-0.9 %	.0009033	.0009726	7.7 %
2	-.0004566	-.0003896	-14.7 %	.0013085	.0012176	-6.9 %
3	-.0000313	-.0000303	-3.5 %	.0003565	.0003256	-8.7 %
4	-.0000046	-.0000044	-4.3 %	.0012805	.0012019	-6.1 %
5	-.0003313	-.0003197	-3.5 %	.0010370	.0009959	-4.0 %
6	.0003012	.0003143	4.4 %	.0006834	.0006819	-0.2 %
7	.0007936	.0007875	-0.8 %	.0021520	.0022267	3.5 %
8	.0000722	.0000720	-0.3 %	.0000277	.0000267	-3.6 %
9	-.0000914	-.0000865	-5.3 %	.0000292	.0000267	-8.6 %
10	.0001096	.0000937	-14.5 %	.0010026	.0009634	-3.9 %
11	-.0005469	-.0005268	-3.7 %	.0016961	.0015602	-8.0 %
12	.0000642	.0000673	4.8 %	.0004514	.0004558	1.0 %
		cor. coef.	= 0.999		cor. coef.	= 0.996
		slope	= 0.97		slope	= 0.99

Note 1. Error = (FEM - Exp)/Exp

The point by point comparisons report the error computed by,

$$\text{Error} = \frac{\epsilon^{FEM} - \epsilon^{Experiment}}{\epsilon^{Experiment}} \qquad (A3.1)$$

where ϵ is the strain component from either the finite element analysis or the experiment as indicated by the superscript. The finite element strains are linearly inperpolated from strains at node points closest to the actual gage locations. Many of the gage locations exactly match nodal locations. The correlation coefficient and the slope indicate the overall agreement between the measured and calculated strain fields. Specifically, the least squares fit to the equation,

$$(\text{FEM Strain}) = A\,(\text{Exp. Strain}) \qquad (A3.2)$$

where A is the slope relating the finite element strain to the strain measured in the experiment provides the slope parameter listed on the Tables A3.1 through A3.5.

Table A3.2
CCP-2.5 / FEM Strain Comparison

Gage	ε_{xx}^{Exp} [1]	ε_{xx}^{FEM} [2]	Error [3]	ε_{yy}^{Exp}	ε_{yy}^{FEM}	Error
1	-.0000897	-.0001018	13.6 %	.0018993	.0018391	-3.2 %
2	.0000252	.0000169	-32.9 %	.0012285	.0014878	21.1 %
3	-.0012125	-.0006106	-49.6 %	.0014702	.0017023	15.8 %
4	.0002843	.0003182	11.9 %	.0024582	.0022402	-8.8 %
5	-.0006270	-.0005772	-7.9 %	.0006940	.0006858	-1.2 %
6	-.0007102	-.0007717	8.7 %	.0007292	.0068589	-6.0 %
7	-.0008394	-.0008211	-2.2 %	.0028643	.0030694	7.2 %
8	-.0012340	-.0012462	-1.0 %	.0026787	.0024649	-8.0 %
9	-.0001694	-.0001513	-10.7 %	.0012741	.0014618	14.7 %
10	-.0009892	-.0012185	23.2 %	.0023518	.0027209	15.7 %
		cor. coef.	= 0.920		cor. coef.	= 0.967
		slope	= 0.92		slope	= 1.03

Notes 1. Gage strain (thermally compensated)
2. Mechanical strain (total-thermal)
3. Error = (FEM - Exp)/Exp

Table A3.3
SEN-2.4 / FEM Strain Comparison

Gage	ε_{xx}^{Exp} [1]	ε_{xx}^{FEM} [2]	Error [3]	ε_{yy}^{Exp}	ε_{yy}^{FEM}	Error
1	.0000455	.0000509	10.2 %	.0001848	.0001990	7.7 %
2	-.0005772	-.0004409	-23.6 %	.0011396	.0013925	22.2 %
3	.0001285	.0001247	-3.0 %	.0007565	.0007473	-1.2 %
4	.0007162	.0009811	37.0 %	.0008837	.0009405	6.4 %
5	.0000414	.0000586	41.5 %	.0010775	.0010895	1.1 %
6	.0010051	.0013321	33.0 %	-.0001789	-.0001939	8.4 %
7	.0014046	.0016325	16.2 %	.0021777	.0021777	0.0 %
8	.0009534	.0010808	13.3 %	.0000207	.0000240	15.9 %
9	.0004278	.0005046	17.9 %	.0000368	.0000240	-34.8 %
10	.0004282	.0005001	16.8 %	.0009211	.0012875	39.0 %
11	-.0005487	-.0004409	-19.6 %	.0019199	.0019703	2.6 %
12	.0001644	.0002500	52.1 %	.0006970	.0007473	7.2 %
		cor. coef.	= 0.992		cor. coef.	= 0.909
		slope	= 1.12		slope	= 1.13

Notes 1. Gage strain (thermally compensated)
2. Mechanical strain (total-thermal)
3. Error = (FEM - Exp)/Exp

Table A3.4
CCP-8.4 / FEM Strain Comparison

Gage	ε_{xx}^{Exp} [1]	ε_{xx}^{FEM} [2]	Error [3]	ε_{yy}^{Exp}	ε_{yy}^{FEM}	Error
1	-.0008799	-.0008448	-4.0 %	.0018976	.0016291	-14.1 %
2	-.0002466	-.0002720	10.3 %	.0011892	.0011986	0.8 %
3	-.0010038	-.0010006	-3.2 %	.0012350	.0013136	6.4 %
4	-.0002391	-.0002338	-2.2 %	.0027302	.0031697	16.1 %
5	-.0015214	-.0013566	-10.8 %	.0002538	.0002784	9.7 %
6	-.0014892	-.0014724	-1.1 %	.0002389	.0002784	16.5 %
7	-.0003894	-.0004290	-1.1 %	.0035687	.0031271	-12.4 %
8	-.0010772	-.0011443	6.2 %	.0019048	.0020757	9.0 %
9	-.0002874	-.0003007	4.6 %	.0012050	.0011856	-1.6 %
10	-.0010643	-.0011005	3.4 %	.0024732	.0025622	3.6 %
		cor. coef.	= 0.993		cor. coef.	= 0.975
		slope	= 0.98		slope	= 0.99

Notes 1. Gage strain (thermally compensated)
2. Mechanical strain (total-thermal)
3. Error = (FEM - Exp)/Exp

Table A3.5
SEN-8.6 / FEM Strain Comparison

Gage	ε_{xx}^{Exp} [1]	ε_{xx}^{FEM} [2]	Error [3]	ε_{yy}^{Exp}	ε_{yy}^{FEM}	Error
1	.0000913	.0000896	-1.9 %	.0001477	.0001416	-4.1 %
2	-.0005915	-.0005774	-2.4 %	.0012656	.0012076	-4.6 %
3	-.0000046	-.0000069	50.0 %	.0004204	.0004057	-3.5 %
4	.0002760	.0003267	10.8 %	.0010749	.0009732	-9.5 %
5	-.0001195	-.0001145	4.2 %	.0009713	.0009401	-3.2 %
6	.0006021	.0006097	1.3 %	-.0002924	-.0002919	-0.2 %
7	.0016850	.0016372	-2.8 %	.0032865	.0029480	-10.3 %
8	.0004329	.0004374	1.0 %	-.0000726	-.0000718	-1.1 %
9	.0001189	.0001145	-3.7 %	-.0000645	-.0000718	11.3 %
10	.0002693	.0002460	-8.7 %	.0008723	.0009011	3.3 %
11	-.0005521	-.0005774	4.6 %	.0016804	.0014637	-12.9 %
12	.0001518	.0001272	-16.2 %	.0003684	.0004057	10.1 %
		cor. coef.	= 0.999		cor. coef.	= 0.998
		slope	= 0.99		slope	= 0.91

Notes 1. Gage strain (thermally compensated)
2. Mechanica strain (total-thermal)
3. Error = (FEM - Exp)/Exp

References

Ainsworth, R.A., Neale, B.K., Price, R.H., [1978], "Fracture Behaviour in the Presence of Thermal Strains," *Tolerance of Flaws in Pressurized Components; Conference, London, Mechanical Engineering Publications for the Institution of Mechanical Engineers*, I Mech E, pp. 171-178.

Aoki, S., Kishimoto, K. and Sakata, M., [1981a], "Crack-Tip Singularity and Strain Singularity in Thermally Loaded Elastic-Plastic Material," *Journal of Applied Mechanics*, **48**, June, pp. 428-429.

Aoki, S., Kishimoto, K. and Sakata, M., [1981b], "Energy-Release Rate in Elastic-Plastic Fracture Problems," *Journal of Applied Mechanics*, **48**, December, pp. 825-829.

Aoki, S., Kishimoto, K. and Sakata, M., [1982], "Elastic-Plastic Analysis of Crack in Thermally-Loaded Structures," *Engineering Fracture Mechanics*, **16**, No. 3, pp. 405-413.

ASTM, [1988], *Annual Book of ASTM Standards*, Volume 3.01 Metals-Mechanical Testing; Elevated and Low Temperature Tests; Metallography, American Society for Testing and Materials, Philadelphia, PA.

Atkinson, C. and Smelser, R.E., [1982], "Invariant Integrals for Thermo- Viscoelasticity and Applications," *International Journal of Solids, Structures*, **18**, No. 6, pp. 533-549.

Atluri, S. N., [1982], "Path-Independent Integrals in Finite Elasticity and Inelasticity, with Body Forces, Inertia, and Arbitrary Crack-Face Conditions," *Engineering Fracture Mechanics*, **16**, No. 3, 341-364.

Atluri, S.N., [1986], "Computational Methods in the Mechanics of Fracture," in *Computational Methods in the Mechanics of Fracture, Volume 2 in Computational Methods in Mechanics*, S.N. Atluri editor, North-Holland, Elsevier Science Publishers, pp. 121-165.

Atluri, S.N., Nakagaki, M., Nishioka, T. and Kuang, Z.B., "Crack-Tip Parameters and Temperature Rise in Dynamic Crack Propagation," *Engineering Fracture Mechanics*, **23**, No. 1, pp. 167-182

Atluri, S.N., Nishioka, T. and Nakagaki, M., [1984], "Incremental Path-Independent Integrals in Inelastic and Dynamic Fracture Mechanics," *Engineering Fracture Mechanics*, **20**, No. 2, pp. 209-244.

Barsoum, R.S., [1977], "Triangular Quarter-Point Elements as Elastic and Perfectly-Plastic Crack Tip Elements," *International Journal for Numerical Methods in Engineering*, **11**, pp. 85-98.

Bass, B.R. and Bryson, J.W., [1983], "Energy Release Rate Techniques for Combined Thermo-Mechanical Loading," *International Journal of Fracture*, **22**, pp. R3-R7.

Batte, A.D., Blackburn, W.S., Elsender, A., Hellen, T. and Jackson, A., [1983], "A Comparison of the J^* Integral With Other Methods of Post Yield Fracture Mechanics," *International Journal of Fracture*, **21**, pp. 49-66.

Begley, J.A. and Landes, J.D., [1972], "The J-Integral as a Fracture Criterion," *Fracture Toughness, Part II, ASTM STP 514*, American Society for Testing and Materials, Philadelphia, pp. 1-20.

Biot, M.A., [1956], "Thermoelasticity and Irreversible Thermodynamics," *Journal of Applied Physics*, **27**, No. 3, March, pp. 240-253.

Biot, M.A., [1959], "New Thermomechanical Reciprocity Relations with Applications to Thermal Stresses," *Journal of Aero-Space Science*, **26**, No. 7, pp. 401-408.

Blackburn, W.S., [1972], "Path Independent Integrals to Predict Onset of Crack Instability in an Elastic Material," *International Journal of Fracture Mechanics*, **8**, pp. 343-346.

Blackburn, W.S., [1985], "Contour Integrals Around Crack Tips for Reversed Loading," *International Journal of Fracture*, **28**, pp. R73-R78.

Blackburn, W.S., Jackson, A.D. and Hellen, T.K., [1977], "An Integral Associated With the State of a Crack Tip in a Non-Elastic Material," *International Journal of Fracture*, **13**, No. 2, April, pp. 183-200.

Broberg, K.B., [1971], "Crack-Growth Criteria and Non-Linear Fracture Mechanics," *Journal of Mechanics and Physics of Solids*, **19**, pp. 407-418.

Broek, D., [1987], *Elementary Engineering Fracture Mechanics*, Martinus Nijhoff Publishers, Boston, MA.

Brust, F.W. and Atluri, S.N., [1986], "Studies on Creep Crack Growth Using the $\dot{T}^*$ Integral," *Engineering Fracture Mechanics*, **23**, No. 3, pp. 551-574.

Brust, F.W., McGowan, J.J. and Atluri, S.N., [1986], "A Combined Numerical / Experimental Study of Ductile Crack Growth after a Large Unloading, Using T^*, J and CTOA Criteria," *Engineering Fracture Mechanics*, **23**, No. 3, pp. 537-550.

Brust, F.W., Nishioka, T., Atluri, S.N. and Nakagaki, M., [1985], "Further Studies on Elastic-Plastic Stable Fracture Utilizing the T^* Integral," *Engineering Fracture Mechanics*, **22**, No. 6, pp. 1079-1103.

Bryan, R. H., Merkle, J. G., Nanstad, R. K. and Robinson, G. C., [1988], "Pressurized Thermal Shock Experiments with Thick Vessels," *Fracture Mechanics: Nineteenth Symposium, ASTM STP 969*, T. A. Cruse, Ed., American Society for Testing and Materials, Philadelphia, pp. 767-783.

Budiansky, B. and Rice, J.R., [1973], "Conservation Laws and Energy-Release Rates," *Journal of Applied Mechanics*, **40**, March, pp. 201-203.

Bui, H.D., [1974], "Dual Path Independent Integrals in the Boundary-Value Problems of Cracks," *Engineering Fracture Mechanics*, **6**, pp. 287-296.

Bylinsky, G., [1986], "The 10,000-mph Airliner," *Fortune*, Dec. 8. 1986, pp. 50-60.

Carlson, D.E., [1972] "Thermoelasticity," in *Handbuch der Phsyik* edited by Truesdell, C.A., Springer Verlag, Berlin.

Chapman, A.J., [1987], *Fundamentals of Heat Transfer*, Macmillan Publishing Company, New York.

Chen, W.H. and Chen, K.T., [1981], "On the Study of Mixed Mode Thermal Fracture Using Modified J_k Integrals," *International Journal of Fracture*, **17**, pp. R99-R103.

Chen, F.H.K. and Shield, R.T., [1977], "Conservation Laws in Elasticity of the J-Integral Type," *Zeitschrift fur angewandte Mathematik und Physik, (Journal of Applied Mathematics and Physics)*, **28**, No. 1, pp. 1-22.

Cherapanov, G.P., [1967], "Crack Propagation in Continuous Media," *Journal of Applied Mathematics and Mechanics*, **31**, No. 3, pp. 503-512, translation of Russian journal *PMM*, **31**, No. 3, pp. 476-488.

Cheverton, R. D., Iskandr, S. K. and Ball, D. G., [1988], "Review of Pressurized-Water-Reactor-Related Thermal Shock Studies," *Fracture Mechanics: Nineteenth Symposium, ASTM STP 969*, T. A. Cruse, Ed., American Society for Testing and Materials, Philadelphia, pp. 752-766.

..... [1983], *Damage Tolerance Design Handbook*, edited by, J. Gallagher, Battelle, Columbus, OH

Eshelby, J.D., [1951], "The Force on an Elastic Singularity," *Philosophical Transactions of the Royal Society*, A 244, November, pp. 87-111.

Eshelby, J.D., [1956], "The Continuum Theory of Lattice Defects," *Solid State Physics*, **3**, pp. 79-144.

Eshelby, J.D. [1975], "The Elastic Energy-Momentum Tensor," *Journal of Elasticity*, **5**, Nos. 3-4, November, pp. 321-335.

Fletcher, D.C., [1976], "Conservation Laws in Linear Elastodynamics," *Archive for Rational Mechanics and Analysis*, **60**, pp. 329-353.

Fung, Y.C., [1965], *Foundations of Solid Mechanics*, Prentice Hall, Chapters 12-14.

Gabor, A., [1986], "Space Plane Gets a 'GO' From Reagan," *U.S.News & World Report*, February 17, 1986, 65-66.

Gamble, R. M. and Paris, P. C., [1976], "Cyclic Crack Growth Analysis for Notched Structures at Elevated Temperatures," *Mechanics of Crack Growth, ASTM STP 590*, American Society for Testing and Materials, pp. 345-367.

Gelfand, I.M. and Fomin, S.V., [1963], *Calculus of Variations*, translated by R.A. Silverman, Prentice Hall, Inc., Englewood Cliffs, New Jersey.

Germain, P. Nguyen, Q.S. and Suquet, P., [1983], "Continuum Thermodynamics," *Journal of Applied Mechanics*, **50**, December, pp. 1010-1020.

Goldman, N.L. and Hutchinson, J.W., [1975], "Fully-Plastic Crack Problems: The Center Cracked Strip Under Plane Strain," *International Journal of Solids and Structures*, **11**, No. 5, pp. 575-592.

Goldstein, H., [1980], *Classical Mechanics*, Second Edition, Addison-Wesley.

Griffith, A.A., [1921], "The Phenomena of Rupture and Flow in Solids," *Philosophical Transactions of the Royal Society of London*, A 221, pp. 163-197.

Griffith, A.A. [1925], "The Theory of Rupture," *Proceedings 1st International Congress on Applied Mechanics*, Biezeno and Burgers editors, Waltman, pp. 55-63.

Günther, W., [1962], "Über einige Randintegrale der Elastomechanik," *Abhandlungen, Braunschweiger Wissenschaftliche Gesellschaft*, **14**, pg. 53 ff.

Gurtin, M.E., [1979], "On a Path-Independent Integral for Thermoelasticity," *International Journal of Fracture*, **15**, pp. R169-R170.

Gurtin, M.E., [1981], *Introduction to Continuum Mechanics*, Academic Press, Inc., Harcourt Brace Tovanovich, Publishers, New York.

Halford, G. R., Hirschberg, M. H. and Manson, S. S., [1973], "Temperature Effects on the Strainrange Partitioning Approach for Creep Fatigue Analysis," *Fatigue at Elevated Temperatures, ASTM STP 520*, American Society for Testing and Materials, pp. 658-669.

Hahn, C.T. and Simon, R., [1973], "A Review of Fatigue Crack Growth in High Strength Aluminum Alloys and the Relevant Metallurgical Factors," *Engineering Fracture Mechanics*, **5**, pp. 523-540.

Hatanaka, K., Fujimitsu, T. and Shiraishi, S., [1989], "An Analysis of Surface Crack Growth in Circumferentially Grooved Components Under Low-Cycle Fatigue," *JSME International Journal*, Series I, **32**, No. 2, pp. 245-255.

Hellen, T.K. and Blackburn, W.S., [1986], *Non-Linear Fracture Mechanics and Finite Elements*, Central Electricty Generating Board, Berkeley Nuclear Laboratories, Gloucestershire, July.

Hill, R., [1960], *The Mathematical Theory of Plasticity*, Oxford University Press, Oxford, U.K. (Latest Edition, 1983).

Hughes, T.J.R., [1987], *The Finite Element Method*, Prentice Hall, Inc., Englewood Cliffs, New Jersey.

Hutchinson, J.W., [1968], "Singular Behavior at the End of a Tensile Crack in a Hardening Material," *Journal of Mechanics and Physics of Solids*, **16**, pp. 13-31.

Ibrahim, F.K., [1989], "The Effects of Stress Ratio, Compressive Peak Stress and Maximum Stress Level on Fatigue Behavior of 2024-T3 Aluminum Alloy," *Fatigue and Fracture of Engineering Materials and Structures*, **12**, No. 1, pp. 1-8.

Irwin, G.R., [1948], "Fracture Dynamics," *Fracturing of Metals*, American Society of Metals, pp. 147-166.

Jablonski, D.A., [1989], "An Experimental Study of the Validity of a Delta J Criterion for Fatigue Crack Growth," *Nonlinear Fracture Mechanics: Volume I - Time-Dependent Fracture, ASTM STP 995*, A. Saxena, J.D. Landes and J.L. Bassani editors, American Society for Testing and Materials, Philadelphia, pp. 361-387.

Kanninen, M.F. and Popelar, C.H., [1985], *Advanced Fracture Mechanics*, Oxford University Press, New York.

Kim K.S. and Orange T.W., [1988], "A Review of Path-Independent Integrals in Elastic-Plastic Fracture Mechanics," *Fracture Mechanics: Eighteenth Symposium, STP 945*, D.T. Read and R.P. Reed editors, American Society for Testing and Materials, Philadelphia, pp. 713-729.

Kishimoto, K., Aoki, S. and Sakata, M., [1980], "On the Path Independent Integral $\hat{J}$," *Engineering Fracture Mechanics*, **13**, pp. 841-850.

Knowles, J.K. and Sternberg, E., [1972], "On a Class of Conservation Laws in Linearized and Finite Elastostatics," *Archive for Rational Mechanics and Analysis*, **44**, pp. 187-211.

Kobayashi, A., Chiu, S. and Beeuwkes, R., [1973], "A Numerical and Experimental Investigation of the Use of the J-Integral," *Engineering Fracture Mechanics*, **5**, pp. 293-305.

Kubo, S., Yafuso, T., Nohara, M., Ishimaru, T. and Ohji, K., [1989], "Investigation on Path-Integral Expression of the J-Integral Range Using Numerical Simulations of Fatigue Crack Growth," *JSME International Journal*, Series I, **32**, No. 2, pp. 237-244.

Kumar, V., German, M.D. and Shih, G.C., [1981], *An Engineering Approach for Elastic-Plastic Fracture Analysis*, NP-1931, Electric Power Research Institute, Palo Alto, California.

Kumar, V., German, M.D., Wilening, W.W., Andrews, W.R., deLorenzi, H.D. and Mowbray, D.F., [1984], *Advances in Elastic-Plastic Fracture Analysis*, NP-3607, Electric Power Research Institute, Palo Alto, California.

Lamba, H.S., [1975], "The J-Integral Applied to Cyclic Loading," *Engineering Fracture Mechanics*, 7, pp. 693-703.

Landes, J. D. and Begley, J. A., [1976], "A Fracture Mechanics Approach to Creep Crack Growth," *Mechanics of Crack Growth, ASTM STP 590*, American Society for Testing and Materials, Philadelphia, pp. 128-148.

Landes, J.D., McCabe, D.E. and Ernst, H.A., [1989], "Geometry Effects on the R-Curve," *Nonlinear Fracture Mechanics: Volume I - Time-Dependent Fracture, ASTM STP 995*, A. Saxena, J.D. Landes and J.L. Bassani editors, American Society for Testing and Materials, Philadelphia, pp. 123-143.

Luenberger, D.G., [1984], *Linear and Nonlinear Programming*, Addison-Wesley Publishing Company, Menlo Park, California.

Marsden, J.E. and Hughes, T.J.R., [1985], *Mathematical Foundations of Elasticity*, Prentice Hall, Englewod Cliffs, New Jersey.

Manson, S. S., Halford, G. R. and Hirschberg, M. H., [1971], *Symposium on Design for Elevated Temperature Environment*, American Society for Mechanical Engineers, New York, pp. 12-24.

Marchand, N. J., Pelloux, R. M. and Ilschners, B., [1988], "A Fracture Mechanics Criterion for Thermal-Mechanical Fatigue Crack Growth of Gas Turbine Materials," *Engineering Fracture Mechanics*, **31**, No. 3, pp. 535-551.

McCartney, L.N., [1979], "Discussion: 'The Use of the J-Integral in Thermal Stress Crack Problems,' by W.K. Wilson and I.W. Yu," *International Journal of Fracture*, **15**, pp. R217-R221.

..... [1985], *Metals Handbook*, Ninth Edition, Volume 2, Properties and Selection: Nonferrous Alloys and Pure Metals, American Society for Metals, Metals Park, OH.

Murakami, Y., [1987], *Stress Intensity Factors Handbook*, Volume I, Pergamon Press, New York.

Neuber, H., [1961], "Theory of Stress Concentration for Shear-Strained Prismatical Bodies with Arbitrary Non-Linear Stress-Strain Law," *Journal of Applied Mechanics*, **28**, pp. 544-550.

Nguyen, Q.S., [1981], "A Thermodynamic Description of the Running Crack Problem," *Proceedings, IUTAM Symposium on Three Dimensional Constitutive Relations and Ductile Fracture*, edited by A. Nemat–Nasser, North-Holland Publishing, Amsterdam, The Netherlands, pp. 315-330.

Nikbin, K.M., Webster, G.A. and Turner, C.E., [1976], "Relevance of Nonlinear Fracture Mechanics to Creep Cracking," *Cracks and Fracture, ASTM STP 601*, American Society for Testing and Materials, Philadelphia, pp. 47-62.

..... [1983], *Military Standardization Handbook*-5D, National Standards Association, Inc., Bethesda, MD

Noether, E., [1918], "Invariante Variationsprobleme," *Gottinger Nachrichten Mathematisch - Physicalische Klasse*, **2**, pp. 235 ff.

Noether, E., [1971], "Invariant Variational Problems," *Transport Theory and Statistical Physics*, **1**, pp. 183-207, translation of Noether [1918] by M. A. Travel

Nowacki, W., [1986], *Thermoelasticity*, Second Edition, translated by H. Zorski, Pergamon Press, Oxford, and PWN Polish Scientific Publishers, Warszawa.

Olver, P.J., [1984], "Conservation Laws in Elasticity; I. General Results," *Archive for Rational Mechanics and Analysis*, **85**, No. 2, pp. 111-129.

Olver, P.J., [1984], "Conservation Laws in Elasticity; II. Linear Homogeneous Isotropic Elastostatics," *Archive for Rational Mechanics and Analysis*, **85**, No. 2, pp. 131-160.

Orowan, E., [1955], "Energy Criteria of Fracture," *Welding Journal*, **34**, pp. 157s-160s.

Paris, P.C., Gomez, M.P. and Anderson, W.E., [1961], "A Rational Analytic Theory of Fatigue," *The Trends in Engineering*, **13**, pp. 9-14.

Perzyna, P., [1971], "Thermodynamic Theory of Viscoplasticity," in *Advances in Applied Mechanics*, **11**, Academic Press, New York.

Pettit, D.E. and Van Orden, J.M., [1979], "Evaluation of Temperature Effects on Crack Growth in Aluminum Sheet Materials," *Fracture Mechanics, ASTM STP 667*, C.W. Smith editor, American Society for Testing and Materials, pp. 106-124.

Rice, J.R., [1968], "A Path Independent Integral and the Approximate Analysis of Strain Concentration by Notches and Cracks," *Journal of Applied Mechanics*, **35**, No. 2, pp. 379-386.

Rice, J.R. and Rosengren, G.F., [1968], "Plane Strain Deformation Near a Crack Tip in a Power-Law Hardening Material," *Journal of Mechanics and Physics of Solids*, **16**, pp. 1-12.

Rice, J.R., Paris, P.C. and Merkle, J.G., [1973], "Some Further Results of J-Integral Analysis and Estimates," *Progress in Flaw Growth and Fracture Toughness Testing, ASTM STP 536*, American Society for Testing and Materials, Philadelphia, pp. 231-245.

Rosen, J., [1974], "Generalized Noether's Theorem. II. Application," *Annals of Physics* , **82**, pp. 70-88.

Sanders, J.L., [1960], "On the Griffith-Irwin Fracture Theory," *Journal of Applied Mechanics*, **27**, June, pp. 352-353.

Sarti, G.C. and Medri, G., [1985], "Thermodynamic Basis for Viscoelastic and Non-Isothermal Fracture Mechanics," *Theoretical and Applied Fracture Mechanics*, **4**, pp. 175-179.

Schmitt, W. and Kienzler, R., [1989], "The J-Integral Concept for Elastic-Plastic Material Behavior," *Engineering Fracture Mechanics*, **32**, No. 3, pp. 409-418.

Sendeckyj, G.P., [1989] Private Communications, February-July.

Sih, G.C., [1973], *Handbook of Stress - Intensity Factors*, Institute of Fracture and Solid Mechanics, Lehigh University, Bethlehem, Pennsylvania.

Sih, G.C. and Tzou, D.Y., [1986], "Heating Preceded by Cooling Ahead of Crack: Macrodamage Free Zone," *Theoretical and Applied Mechanics*, **6**, pp. 103-111.

Smith, C. W., [1986], "Cracking at Nozzle Corners in the Nuclear Pressure Vessel Industry," *Case Histories Involving Fatigue and Fracture Mechanics, ASTM STP 918*, C. M. Hudson and T. P. Rich, Eds., American Society for Testing and Materials, Philadelphia, pp. 31-45.

Simo, J.C., and Honein, T., [1990], "Variational Formulation, Discrete Conservation Laws and Path-Domain Independent Integrals for Elasto-Viscoplasticity," to appear in *Journal of Applied Mechanics*

Simo, J.C., Kennedy, J.G. and Taylor, R.L., [1989], "Complementary and Mixed Finite Element Formulations for Elastoplasticity," *Computer Methods in Applied Mechanics and Engineering*, **74**, pp. 177-206.

Simo, J.C. and Hughes, T.J.R., [1989], *Elastoplasticity and Viscoplasticity; Computational Aspects*, in publication.

Simo, J.C., Kennedy, J.G. and Govindjee, S. [1988], "Non-Smooth Multisurface Plasticity and Viscoplasticity. Loading/Unloading Conditions and Numerical Algorithms ", *International Journal for Numerical Methods in Engineering*, **26**, No. 10, pp. 2161-2185.

Simo, J.C., [1988], "A Framework for Finite Strain Elastoplasticity Based on Maximum Plastic Dissipation and the Multiplicative Decomposition, Part II: Computational Aspects," *Computer Methods in Applied Mechanics and Engineering*, **68**, pp. 1-31.

Simo, J.C., Taylor, R.L. and Pister, K.S., [1985], "Variational and Projection Methods for the Volume Constraint in Finite Deformation Elasto-Plasticity," *Computer Methods in Applied Mechanics and Engineering*, **51**, pp. 177-208.

Stonesifer, R.B. and Atluri, S.N., [1982a], "On a Study of the $(\Delta T)_c$ and C^* Integrals for Fracture Analysis Under Non-Steady Creep," *Engineering Fracture Mechanics*, **16**, No. 5, pp. 625-643.

Stonesifer, R.B. and Atluri, S.N., [1982b], "Moving Singularity Creep Crack Growth Analysis with the $(\Delta T)_c$ and C^* Integrals," *Engineering Fracture Mechanics*, **16**, No. 6, pp. 769-782.

Strang, G., [1986], *Introduction to Applied Mathematics*, Wellesley-Cambridge Press, Wellesley, Massachusettes.

Sullivan, A.M., Freed, C.N. and Stoop, J., [1973], "Comparison of R-Curve Determined from Different Specimen Types," *Fracture Toughness Evaluation by R-Curve Methods*, ASTM STP 527, American Society for Testing and Materials, pp.85-104.

Taylor, R. L., [1977], "Computer Procedures for Finite Element Analysis," Chapter 24 in *The Finite Element Method*, Third Edition, by O. C. Zienkiewicz, McGraw Hill, London, pp. 677-757.

Vitcha, E. T., [1973], "High-Temperature Fatigue Testing of Automotive Valve Steels," *Fatigue at High Temperatures, ASTM STP 520*, American Society for Testing and Materials, Philadelphia, pp. 231-241.

Walker, K., [1970], "The Effect of Stress Ratio During Crack Propagation and Fatigue for 2024-T3 and 7075-T6 Aluminum," *Effects of Environment and Complex Load History on Fatigue Life, ASTM STP 462*, American Society for Testing and Materials, Philadelphia, pp. 1-14.

Wilson, W.K. and Yu, I.W., [1979], "The Use of the J-Integral in Thermal Stress Crack Problems," *International Journal of Fracture*, **15**, No. 4, August, pp. 377-387.

Zienkiewicz, O.C. and Taylor, R.L., [1989] *The Finite Element Method*, Fourth Edition, Mcgraw Hill, London.

Attachment 5

AFOSR-TR- 91 0982

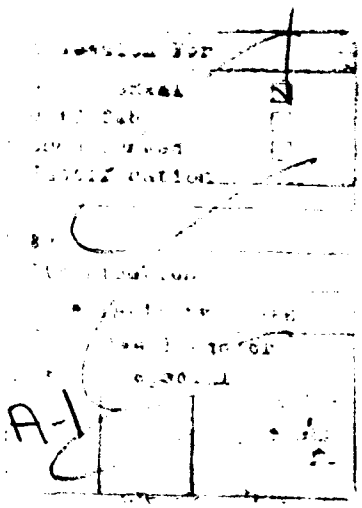
THESIS BY:

ANDREW C. BARTLETT

UNIVERSITY OF MASSACHUSETTS

Subcontract No.# S-789-000-013

VERTEX AND EDGE THEOREMS WHICH SIMPLIFY CLASSICAL ANALYSES OF LINEAR SYSTEMS WITH UNCERTAIN PARAMETERS



A Dissertation Presented

by

ANDREW C. BARTLETT

Submitted to the Graduate School of the

University of Massachusetts in partial fulfillment

of the requirements for the degree of

DOCTOR OF PHILOSOPHY

SEPTEMBER 1990

Department of Electrical and Computer Engineering

VERTEX AND EDGE THEOREMS WHICH SIMPLIFY CLASSICAL ANALYSES OF LINEAR SYSTEMS WITH UNCERTAIN PARAMETERS

A Dissertation Presented

by

ANDREW C. BARTLETT

Approved as to style and content:

Christopher V. Hollot, Chair

Theodore E. Djaferis, Member

Douglas P. Looze, Member

Yossi Chait, Member

Keith R. Carver, Department Head
Electrical and Computer Engineering

ACKNOWLEDGEMENT

Since the Fall of 1987, my doctoral studies have been supported by a United States Air Force Office of Scientific Research Laboratory Graduate Fellowship. My studies were also supported by a 1988 Air Force Summer Fellowship with the Control Analysis Group in the Flight Dynamics Laboratory at Wright-Paterson AFB. I would like to thank the Air Force Office of Scientific Research for giving me the opportunity to be part of both of these programs. I would also like to thank the Flight Dynamics Laboratory for selecting me to participate in these programs and for hosting me during my summer visit. I would especially like to thank my fellowship mentor, Dr. Siva S. Banda, for the advice and encouragement he has given me. The helpfulness of the fellowship administrators at Universal Energy Systems is also gratefully acknowledged.

The staff of the Department of Electrical and Computer Engineering at the University of Massachusetts has also been very helpful. I appreciate all the friendly assistance they have given me. In particular, I am indebted to Pat Zuzgo for the excellent job she has done in coordinating the funding for my fellowship.

I would like to take this opportunity to belatedly acknowledge Prof. Clas A. Jacobson formerly with Rensselaer Polytechnic Institute and now with Northeastern University. During my early work on the results contained in Chapter 3 of this dissertation, I had many helpful discussions with Clas, and he gave me a considerable amount of encouragement. Clas should have been acknowledged by me in [Bartlett, Hollot, Huang, 1988]. This oversight was due to my ignorance of proper etiquette and was in no way due to a lack of appreciation.

The people I wish to thank above all are the faculty members of my dissertation committee, Professors Yossi Chait, Douglas P. Looze, Theodore E. Djaferis, and

Christopher V. Hollot. They have all helped to educate, guide, and encourage me. Most of all, I wish to express my gratitude to my advisor, Prof. Hollot. I can say without a doubt that if it had not been for Kris Hollot, I would not have made any of the contributions to the analysis of uncertain control systems that are contained in this dissertation.

ABSTRACT

VERTEX AND EDGE THEOREMS
WHICH SIMPLIFY CLASSICAL ANALYSES OF
LINEAR SYSTEMS WITH UNCERTAIN PARAMETERS

SEPTEMBER 1990

ANDREW C. BARTLETT

B.S.E.E., UNIVERSITY OF MASSACHUSETTS

M.E., RENSSELAER POLYTECHNIC INSTITUTE

PH.D., UNIVERSITY OF MASSACHUSETTS

Directed by: Professor Christopher V. Hollot

This dissertation addresses the problem of analyzing both continuous and discrete-time, linear time-invariant systems with uncertain parameters. The investigation considers four classifications of uncertain systems that are loosely called interval families, affine uncertainties, polytopes, and multiaffine uncertainties. The focus is on classical analyses: stability, pole locations, frequency response, and time response. The goal of each analysis is to determine the worst-case behavior of the system over all possible values of the parameter vector. To simplify these analyses, the existence of relationships between the extreme behavior of the system and the "prominent" values of the parameters is investigated. Because the "prominent" parameter values are analogous to the vertices or the edges of a box, results which show that the worst-case behavior can be determined using only "prominent" parameters are referred to as vertex and edge theorems. For each class of systems and each analysis problem, the existence of vertex and edge theorems is reviewed.

For stability and pole location analyses, edge theorems are presented for interval families, affine uncertainties, and polytopes. These edge theorems are a contribution of

this dissertation. For multiaffine uncertainties, no stability or pole location edge theorems exist. In general, stability and pole location vertex theorems do not exist for any of the four system classes. The main exception is Kharitonov's stability vertex theorem for continuous-time interval families.

For frequency response determination, the contribution of this dissertation is an edge theorem for interval families, affine uncertainties, and polytopes. For multiaffine uncertainties, no frequency response edge theorem exists. For all four classes of systems, frequency response vertex theorems also do not exist.

A steady state time response vertex theorem is presented for all four classes of uncertain systems. For affine uncertainties, polytopes, and multiaffine uncertainties, it is shown that a transient response vertex theorem does not exist. These two results are contributions of this dissertation. The existence or absence of a transient response vertex theorem for interval families is still an open question. The availability of transient response edge theorems remains an open problem for all four classes of uncertain systems.

TABLE OF CONTENTS

LIST OF FIGURES

NOTATION

j	$\sqrt{-1}$
$\mathbb{C}$	all complex numbers, the complex plane
$\mathbb{R}$	all real numbers, the real line
$\mathbb{R}^m$	all m-tuples of reals
$\mathbb{S}$	the Riemann Sphere, $\mathbb{C} \cup \{ \infty \}$
$\mathbb{C}[s]$	all complex polynomials
$\mathbb{C}_n[s]$	all complex polynomials of degree n
$\mathbb{C}[s]^2$	all ordered pairs (2-tuples) of complex polynomials
$\mathbb{R}[s]$	all real polynomials
$\mathbb{R}_n[s]$	all real polynomials of degree n
$\mathbb{R}_n^{monic}[s]$	all real monic polynomials of degree n
$\mathbb{R}[s]^2$	all ordered pairs (2-tuples) of real polynomials
$u(s)$	transform of the system input, Laplace transform for continuous-time, Z-transform for discrete-time
$y(s)$	transform of the system output, Laplace transform for continuous-time, Z-transform for discrete-time
$d(s)$	the system's characteristic polynomial, also called the denominator polynomial
$n(s)$	the system's numerator polynomial
$(n(s),d(s))$	numerator-denominator polynomial-pair description for the system
δ	the real m-tuple of uncertain system parameters, $\delta = (\delta_1, \delta_2, \ldots, \delta_m)$
Δ	the set of possible values for the uncertain parameter δ
D	mapping from $\delta \in \mathbb{R}^m$ to $d(s) \in \mathbb{C}[s]$

N	mapping from $\delta \in \mathbb{R}^m$ to $n(s) \in \mathbb{C}[s]$
T	mapping from $\delta \in \mathbb{R}^m$ to $(n(s),d(s)) \in \mathbb{C}[s]^2$, $T(\delta) = (N(\delta),D(\delta))$
$\mathcal{D}$, $D(\Delta)$	the set of possible characteristic polynomials for the system
$\mathcal{N}$, $N(\Delta)$	the set of possible numerator polynomials for the system
$\mathcal{T}$, $T(\Delta)$	the set of possible polynomial-pair descriptions for the system
$\mathcal{P}$	a set of polynomials, equals either $\mathcal{D}$ or $\mathcal{N}$
$\text{Num}(\mathcal{T})$	$\{ n(s) \in \mathbb{C}[s] : (n(s),d(s)) \in \mathcal{T} \}$
$\text{Den}(\mathcal{T})$	$\{ d(s) \in \mathbb{C}[s] : (n(s),d(s)) \in \mathcal{T} \}$
$\text{Box}\{\delta^L,\delta^H\}$	interval family of real m-tuples given by $\{ (\delta_1, \delta_2, \ldots, \delta_m) \in \mathbb{R}^m : \delta_i^L \le \delta_i \le \delta_i^H, \ i = 1, 2, \ldots, m \}$
$\text{Vert}\{\delta^L,\delta^H\}$	vertices of $\text{Box}\{\delta^L,\delta^H\}$ $\{ (\delta_1, \delta_2, \ldots, \delta_m) \in \mathbb{R}^m : \delta_i \in \{\delta_i^L,\delta_i^H\}, \ i = 1, 2, \ldots, m \}$
$\text{Edge}\{\delta^L,\delta^H\}$	edges of $\text{Box}\{\delta^L,\delta^H\}$ $\{ (\delta_1, \delta_2, \ldots, \delta_m) \in \mathbb{R}^m : \delta_i \in \{\delta_i^L,\delta_i^H\}, \ i = 1, 2, \ldots, j-1, j+1, \ldots, m, \ \delta_j^L \le \delta_j \le \delta_j^H, \ j = 1, 2, \ldots, m \}$
$\text{Affine}\{\mathbb{U},\mathbb{V}\}$	the set of all affine mappings from $\mathbb{U}$ to $\mathbb{V}$
$\text{Multi}\{\mathbb{R}^m,\mathbb{V}\}$	the set of all multiaffine mappings from $\mathbb{R}^m$ to $\mathbb{V}$
$\text{Vert}(\Delta)$	all vertices of Δ when Δ is an interval family of real m-tuples
$\text{Edge}(\Delta)$	the set of all parameters contained in edges of Δ when Δ is an interval family of real m-tuples
$\text{Vert}(\mathcal{D})$	all vertex polynomials of $\mathcal{D}$ when $\mathcal{D}$ is a polytope of polynomials
$\text{Vert}(\mathcal{N})$	all vertex polynomials of $\mathcal{N}$ when $\mathcal{N}$ is a polytope of polynomials
$\text{Vert}(\mathcal{T})$	all vertex polynomial-pairs of $\mathcal{T}$ when $\mathcal{T}$ is a polytope of polynomial-pairs

Edge($\mathbb{D}$)	the set of all polynomials contained in edges of $\mathbb{D}$ when $\mathbb{D}$ is a polytope of polynomials
Edge($\mathbb{N}$)	the set of all polynomials contained in edges of $\mathbb{N}$ when $\mathbb{N}$ is a polytope of polynomials
Edge($\mathbb{T}$)	the set of all polynomial-pairs contained in edges of $\mathbb{T}$ when $\mathbb{T}$ is a polytope of polynomial-pairs
Root($\mathbb{P}$)	root space of the set of polynomials $\mathbb{P}$, $\{ s \in \mathbb{C} : p(s) = 0, p(s) \in \mathbb{P} \}$
Root($\mathbb{D}$)	the set of all possible poles of the system
Root($\mathbb{N}$)	the set of all possible zeros of the system
∂Root($\mathbb{D}$)	the boundary of Root($\mathbb{D}$) $\subset \mathbb{C}$
G	a stability region
∂G	the boundary of G $\subset \mathbb{C}$
G_H	the open left half plane
G_S	the open unit disk
G-stable	$\mathbb{P} \subset \mathbb{C}[s]$ is G-stable if Root($\mathbb{P}$) $\subset$ G
Val($\mathbb{D}$,z)	the value set of $\mathbb{D} \subset \mathbb{C}[s]$, $\{ p(z) \in \mathbb{C} : p(s) \in \mathbb{D} \}$.
Val($\mathbb{T}$,z)	the value set of $\mathbb{T} \subset \mathbb{C}[s]^2$, $\{ (n(z),d(z)) \in \mathbb{C} : (n(s),d(s)) \in \mathbb{T} \}$
conv(A)	convex hull of the set A
A \ B	$\{ x : x \in A, x \notin B \}$
arg[c]	the argument of $c \in \mathbb{C}$
abs[c]	the absolute value of $c \in \mathbb{C}$
∂A	the boundary of the set A in $\mathbb{C}$ (∞ excluded)
A / B	$\{ x/y : x \in A, y \in B \}$
Nyq($\mathbb{T}$,z)	$\{ n(z)/d(z) \in \mathbb{C} : abs[n(z)/d(z)] < \infty, (n(s),d(s)) \in \mathbb{T} \}$
Mag($\mathbb{T}$,z)	$\{ abs[n(z)/d(z)] : 0 < abs[n(z)/d(z)] < \infty, (n(s),d(s)) \in \mathbb{T} \}$

$[\delta]_i$	the ith component of the p-tuple δ
$A \times \mathbb{B}$	Cartesian product of A and $\mathbb{B}$
max A	maximum element of the set $A \subset \mathbb{R}$
min A	minimum element of the set $A \subset \mathbb{R}$
$\mathcal{L}^{-1}\{F(s)\}$	inverse Laplace transform over s of F(s), the region of convergence is assumed to be an open right half plane
$\mathcal{Z}^{-1}\{F(z)\}$	inverse Z-transform over z of F(z), the region of convergence is assumed to be an unbounded open annulus
$Khar_+(\mathbb{P})$	the positive frequency Kharitonov polynomials of $\mathbb{P}$ when $\mathbb{P}$ is an interval family of polynomials, see Definition 3.2
$Khar_-(\mathbb{P})$	the negative frequency Kharitonov polynomials of $\mathbb{P}$ when $\mathbb{P}$ is an interval family of polynomials, see Definition 3.2
$Khar(\mathbb{P})$	$Khar_+(\mathbb{P}) \cup Khar_-(\mathbb{P})$
$Aux_+(\mathbb{P},jw)$	the set of points where $\partial Val(\mathbb{P},jw)$ intersects the real or imaginary axis plus the intersection of $Val(\mathbb{P},jw)$ with the origin when $\mathbb{P}$ is an interval family of polynomials and $w \geq 0$
$Aux_-(\mathbb{P},jw)$	the set of points where $\partial Val(\mathbb{P},jw)$ intersects the real or imaginary axis plus the intersection of $Val(\mathbb{P},jw)$ with the origin when $\mathbb{P}$ is an interval family of polynomials and $w \geq 0$

CHAPTER 1

INTRODUCTION

1.1 Overview

This dissertation addresses the problem of analyzing both continuous-time and discrete-time, finite-dimensional, linear, time-invariant systems[1] with uncertain parameters. This work will focus on classical analyses: stability, pole-zero locations, frequency response, and time response. A worst-case philosophy will be employed, so the purpose of each analysis will be to determine the extreme behavior of the system over all possible values of the parameters. For example, a stability or pole location analysis will typically seek to ascertain if there are any parameter values for which the poles of the system have unacceptable locations in the complex plane. A frequency response analysis will attempt to determine quantities like the maximum possible bandwidth of the system or the minimum possible distance of the Nyquist plot to the origin. A time response analysis, for a given input, will seek information such as the maximum possible value of a particular output during the transient period or in the steady state. In an effort to simplify these analyses, this dissertation will investigate the existence of relationships between the extreme behavior of the system and "prominent" values of the parameters. For the cases that are considered, these "prominent" parameter values are analogous to either the vertices or the edges of a box. For this reason, results which show that the worst-case behavior of a system can be determined using only "prominent" parameters are referred to as vertex theorems and edge

[1] Throughout the remainder of this dissertation unless noted to the contrary, systems will refer to only finite-dimensional, linear, time-invariant systems.

theorems. The existence of several theorems of this type will be demonstrated, and counterexamples to several similar conjectures will also be given.

1.2 Descriptions of Systems with Parameter Uncertainties

The analyses that are investigated in this dissertation will use polynomials to describe each uncertain system. The polynomials come from the system description

$$d(s)\, y(s) = n(s)\, u(s)$$

which uses the polynomials d(s) and n(s) to relate the transform of a given output y to the transform of a given input u. The appropriate transform is the Laplace transform for continuous-time systems and the Z-transform for discrete-time systems. Regardless of which transform is used, d(s) will be called both the **characteristic polynomial** and the **denominator polynomial**, and n(s) will be called the **numerator polynomial**. When a system depends on an m-tuple of parameters $\delta = (\delta_1, \delta_2, \ldots, \delta_m)$, the polynomials which describe the system will also depend on δ. The two mappings D and N from the set of all real m-tuples $\mathbb{R}^m$ to the set of all complex polynomials $\mathbb{C}[s]$ will be used to define d(s) and n(s), respectively, for each value of δ. When uncertainties such as measurement error or manufacturing tolerance are present, the value of each parameter may not be known exactly. Typically, δ will only be known to lie in some **parameter set** $\Delta \subset \mathbb{R}^m$. As a result, the polynomials d(s) and n(s) which describe the behavior of the system from the input u to the output y will only be known to lie in sets of polynomials. This means that to determine worst-case behavior an entire **set of system descriptions** must be used.

The composition of the set of descriptions will vary depending on the type of analysis. To check for possible instabilities or to determine all possible pole locations, the set of polynomials

$$\mathbb{D} = D(\Delta) = \{ D(\delta) : \delta \in \Delta \}$$

is the easiest to use. To determine all possible zero locations, the set of polynomials

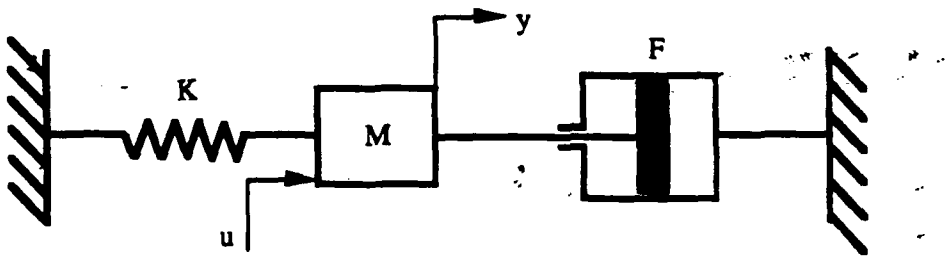
$$\mathbb{N} = N(\Delta) = \{ N(\delta) : \delta \in \Delta \}$$

is the best choice. Some analyses will need to know both n(s) and d(s). For these cases, the mapping T from $\mathbb{R}^m$ to the set of all ordered pairs (2-tuples) of complex polynomials $\mathbb{C}[s]^2$ defined by $T(\delta) = (N(\delta), D(\delta))$ will be helpful. Using T, the set

$$\mathbb{T} = T(\Delta) = \{ T(\delta) : \delta \in \Delta \}$$

can be easily specified. The set $\mathbb{T}$ provides all the information needed to determine the worst-case frequency or time response. Two examples are given below which illustrate how the mappings N, D, and T and the sets Δ, $\mathbb{D}$, $\mathbb{N}$, and $\mathbb{T}$ are generated for physical systems.

Example 1.1 Consider the spring-mass-dashpot system shown in Figure 1.1 having mass M, spring constant K, and damping coefficient F.



Spring-Mass-Dashpot System.

Figure 1.1

The following description

$$(Ms^2+Fs+K)\, y(s) = u(s)$$

relates the Laplace transform of the input force u and the output position y. The three mappings $D: \mathbb{R}^3 \to \mathbb{C}[s]$, $N: \mathbb{R}^3 \to \mathbb{C}[s]$, and $T: \mathbb{R}^3 \to \mathbb{C}[s]^2$ are defined by

$$D((K,M,F)) = Ms^2+Fs+K,$$

$$N((K,M,F)) = 1,$$

and

$$T((K,M,F)) = (1, Ms^2+Fs+K),$$

respectively. If the the parameters are uncertain and satisfy the bounds

$$1 \le K \le 2; \qquad 1.5 \le M \le 1.6; \qquad 3 \le F \le 5$$

(in appropriate units), then the set of parameters is given by

$$\Delta = \{ \ \delta \in \mathbb{R}^3 \ : \ \delta = (K,M,F), \ 1 \le K \le 2, \ 1.5 \le M \le 1.6, \ 3 \le F \le 5 \ \}.$$

The mappings D, N, and T along with the set Δ provide the information needed to define the following sets

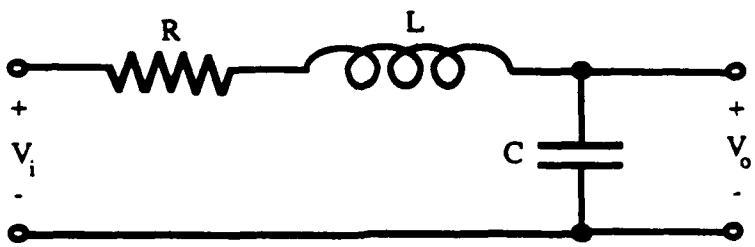
$$\mathbf{D} = \{ \ Ms^2+Fs+K \ : \ 1 \le K \le 2, \ 1.5 \le M \le 1.6, \ 3 \le F \le 5 \ \},$$

$$\mathbf{N} = \{ \ 1 \ \},$$

and

$$\mathbf{T} = \{ \ (\ 1, Ms^2+Fs+K \ ) \ : \ 1 \le K \le 2, \ 1.5 \le M \le 1.6, \ 3 \le F \le 5 \ \}.$$

Example 1.2 Consider the RLC circuit displayed in Figure 1.2.



RLC Circuit.

Figure 1.2

For this RLC circuit, the following description

$$(LCs^2+RCs+1)\, v_o(s) = v_i(s)$$

relates the Laplace transform of the input voltage v_i and the output voltage v_o. The three mappings $D: \mathbb{R}^3 \to \mathbb{C}[s]$, $N: \mathbb{R}^3 \to \mathbb{C}[s]$, and $T: \mathbb{R}^3 \to \mathbb{C}[s]^2$ are defined by

$$D((R,L,C)) = LCs^2+RCs+1,$$

$$N((R,L,C)) = 1,$$

and

$$T((R,L,C)) = (1, LCs^2+RCs+1),$$

respectively. If the the parameters are uncertain and satisfy the bounds

$$1 \le R \le 2; \qquad 3 \le L \le 4; \qquad 5 \le C \le 6$$

(in appropriate units), then the set of all possible parameters is given by

$$\Delta = \{ \delta \in \mathbb{R}^3 : \delta = (R,L,C),\ 1 \le R \le 2,\ 3 \le L \le 4,\ 5 \le C \le 6 \}.$$

For this system, the sets of descriptions $\mathbf{D}$, $\mathbf{N}$, and $\mathbf{T}$ are defined as follows

$$\mathbf{D} = \{ LCs^2+RCs+1 : 1 \le R \le 2,\ 3 \le L \le 4,\ 5 \le C \le 6 \},$$

$$\mathbf{N} = \{ 1 \},$$

and

$$\mathbf{T} = \{ (1, LCs^2+RCs+1) : 1 \le R \le 2,\ 3 \le L \le 4,\ 5 \le C \le 6 \}.$$

For the uncertain systems considered in this dissertation, the sets $\mathbf{D}$, $\mathbf{N}$, and $\mathbf{T}$ provide essentially all the information about the system that is required to carry out the desired analyses. The mappings D, N, and T in combination with the set Δ also provide the same information in a different form. The only other information that must be specified is whether the system is discrete or continuous-time. For these reasons, systems can be treated throughout the remainder of this dissertation using only the sets of polynomials $\mathbf{D}$ and $\mathbf{N}$, the set of polynomial-pairs $\mathbf{T}$, the set of real m-tuples Δ, the

mappings D and N from real m-tuples to polynomials, and the mapping T from real m-tuples to polynomial-pairs.

1.3 Classes of Polynomial Sets and Polynomial-Pair Sets

Examples 1.1 and 1.2 show that it is quite natural for the sets D, N, and/or T to contain an infinite number of possible system descriptions. When these sets are infinite, it will be impossible to carry out a worst-case analysis using the **brute force** method of individually testing each and every description. Because a feasible alternative to the brute force method is required, this dissertation focuses on a few general but highly structured classes of description sets. This structure makes it possible to analyze these sets without using the brute force approach. The classifications for D, N, and T that are of interest will be formally presented in Chapter 2. These classifications are designed to include the types of polynomial sets and polynomial-pair sets needed to describe physical systems that have a reasonably simple dependence on uncertain parameters.

Some of the classifications given in Chapter 2 will be defined in terms of how the set Δ and the mappings D, N, and T generate the sets D, N, and T. For these classifications, the set of parameters Δ will be required to be an **interval family of real m-tuples**. In both Example 1.1 and 1.2, the set of parameters Δ is an interval family of real 3-tuples. The classifications for D, N, and T are further refined by requiring the mapping D, N, and T to be one of two types of mappings. These two choices called **multiaffine mappings** and **affine mappings** are defined in Chapter 2. In both Example 1.1 and 1.2, all three mappings D, N, and T are multiaffine. The three mappings D, N, and T in Example 1.1 are affine, but the two mappings D and T in Example 1.2 are not affine. These examples correctly suggest that affine mappings

are a special type of multiaffine mappings. The two choices of mappings define two classifications for sets of polynomials and two classifications for sets of polynomial-pairs. A set of polynomials $\mathbb{D} = D(\Delta)$ or $\mathbb{N} = N(\Delta)$ is classified as being **generated by multiaffine uncertainties** if Δ is an interval family and if D or N, respectively, is a multiaffine mapping. If D or N is also an affine mapping, then $\mathbb{D} = D(\Delta)$ or $\mathbb{N} = N(\Delta)$, respectively, is also said to be **generated by affine uncertainties.** Similarly, a set of polynomial-pairs $\mathbb{T} = T(\Delta)$ is classified as being **generated by multiaffine uncertainties** if Δ is an interval family and if T is a multiaffine map. If T is also an affine map, then $\mathbb{T} = T(\Delta)$ is also said to be **generated by affine uncertainties.** The sets D, N, and T in Example 1.1 are examples of sets generated by affine uncertainties. The sets D and T in Example 1.2 are not generated by affine uncertainties, but they are generated by multiaffine uncertainties.

The remaining classifications that are defined in Chapter 2 are stated directly in terms of the sets D, N, and T and do not dependent on the set Δ or the mappings D, N, and T. Two classifications of sets of polynomials and two classifications of sets of polynomial-pairs are defined in this way. The first two of these classes are called **polytopes of polynomials** and **polytopes of polynomial-pairs.** The second two classes are special types of polytopes called **interval families of polynomials** and **interval families of polynomial-pairs.** These classes will be formally defined in Chapter 2. At this point, it will be adequate to note some examples of these classes. The set D in Example 1.1 is an interval family of polynomials, so it must also be a polytope of polynomials. The set D in Example 1.2 is not an interval family of polynomials, but it is a polytope of polynomials. The set T in Example 1.1 is an interval family of polynomial-pairs. The set T in Example 1.2 is not an interval family of polynomial-pairs, but it is a polytope of polynomial-pairs.

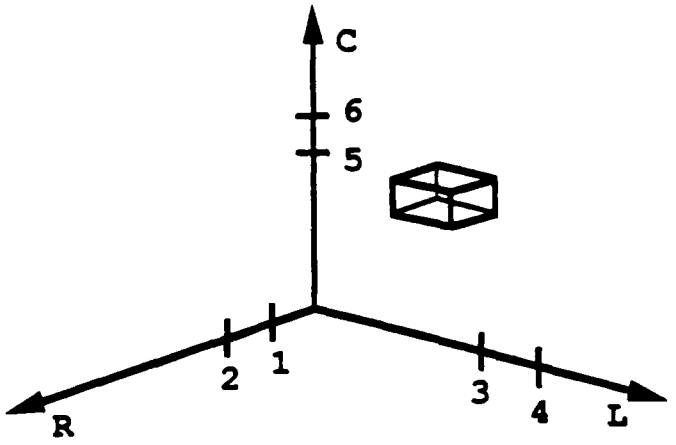
Examples 1.1 and 1.2 show that some physical systems do indeed have polynomial and polynomial-pair description sets that fall into the classifications discussed above. The number of physical systems whose description sets fall into one or more of the classifications above is quite large. Dasgupta and Anderson [1987] have shown that a significant class of physical systems are describe by sets of polynomials and polynomial-pairs that are generated by multiaffine uncertainties. This provides a strong motivation to investigate multiaffine uncertainties.

The reason description sets generated by affine uncertainties are treated separately from the multiaffine case is that significant simplification of the analysis process is possible when attention is restricted to this subclass. The difference in the analysis effort required by the affine case verses the multiaffine case is the reason for considering polytopes. All description sets generated by affine uncertainties are polytopes, so polytopes are a more general class. Despite being more general, polytopes require essentially the same analysis effort as sets generated by affine uncertainties. This fact is useful because some families generated by multiaffine uncertainties are polytopes. These special types of multiaffine uncertainties can be analyzed with the same amount of effort that is required to analyze affine uncertainties. Furthermore, the smallest convex set that contains a family of polynomials or polynomial-pairs generated by multiaffine uncertainties is always a polytope. This overbounding polytope is easily identified, and it is useful for analyzing the original family. An analysis carried out using the overbounding polytope will be conservative, but its use is often justified by the decrease in analysis effort. The subclass of polytopes called interval families are treated as special cases because dramatic simplifications of some important analyses are possible when attention is restricted to these families.

1.4 Vertex and Edge Theorems

It was stated previously that the sets of descriptions (polynomials or polynomial-pairs) considered in this dissertation would be highly structured and that this structured would be used for a non-brute force analysis. Four classes of these structured sets were specified for both polynomials and polynomial-pairs. This section will discuss some of the prominent features of these sets and indicate how these features will be used in a feasible analysis method. The prominent features of sets of polynomials and sets of polynomial-pairs generated by affine and multiaffine uncertainties will be discussed first. Afterwards, the prominent features of polytopes and interval families of polynomials and polynomial-pairs will be pointed out. Finally, the intended strategy for using the prominent features in a non-brute force analysis will be outlined.

For the sets $D(\Delta)$, $N(\Delta)$, or $T(\Delta)$ to be classified as being generated by affine or multiaffine uncertainties, the set Δ is required to be an interval family of real m-tuples. These types of sets are highly structured and have some prominent features. For example, interval families in $\mathbf{R}^3$ are shaped exactly like rectangular boxes, so the most prominent features of a real 3-dimensional interval family are its vertices and its edges. To illustrate this point, the cube shaped interval family Δ from Example 1.2 is shown in Figure 1.3. This figure should provide the reader with an intuitive feel for the vertices and edges of an interval family in $\mathbf{R}^3$. For $m \neq 3$, the vertices and edges of an interval family of real m-tuples can be easily described algebraically. The exact definitions will be provided in Chapter 2. If Δ is an interval family, then for convenient reference, the set of all its vertices will be denoted by $\mathbf{Vert}(\Delta)$ and the set of all real m-tuples contained in all edges of Δ will be represented by the notation $\mathbf{Edge}(\Delta)$. The sets of real m-tuples $\mathrm{Vert}(\Delta)$ and $\mathrm{Edge}(\Delta)$ will play an important analysis role.

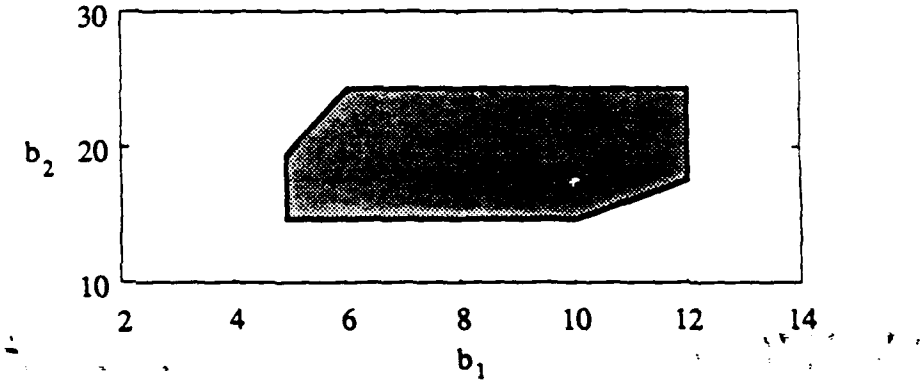


Interval family of real 3-tuples.

Figure 1.3

As with interval families of real m-tuples, vertices and edges can also be defined for interval families and polytopes of polynomials and polynomial-pairs. The rectangular box analogy used above can also be employed for interval families of polynomials and polynomial-pairs. A geometric analogy for polytopes can use not only rectangular boxes but also pyramids and other flat-sided, convex figures that have obvious definitions for vertices and edges. This can be illustrated with the polytope of polynomials D from Example 1.2. Each polynomial d(s) in D can be represented with a real 2-tuple by letting the polynomial $d(s) = b_2s^2+b_1s+1$ correspond to the pair of coefficients (b_2, b_1). Using this coefficient space representation, the set D will correspond to the set of coefficients illustrated in Figure 1.4. The set of coefficients in Figure 1.4 can also represent the set of polynomial-pair descriptions T from Example 1.2 by letting the pair of coefficients (b_2, b_1) correspond to the polynomial-pair $(1, b_2s^2+b_1s+1)$. The vertices and edges of the set shown in Figure 1.4 correspond

to the vertices and edges of ⊐ and ⊤. This geometric idea is the basis for the algebraic definitions of polytope edges and vertices that will be given in Chapter 2. When ⊐ or ℕ is a polytope of polynomials, the set of all its vertices will be denoted as **Vert**(⊐) or **Vert**(ℕ), respectively, and the set of all polynomials that are contained in edges of the polytope will be represented by **Edge**(⊐) or **Edge**(ℕ), respectively. If ⊤ is a polytope of polynomial-pairs, its vertices and edges will be denoted by **Vert**(⊤) and **Edge**(⊤). For the general case of polytopes, the edges will play an important role in analysis. For the special case of interval families of polynomials and polynomial-pairs, the vertices will be even more important.



The polytope of polynomials D from Example 1.2 represented in coefficient space. The pair of coefficients (b_2, b_1) corresponds to the polynomial $b_2s^2+b_1s+1$.

Figure 1.4

All the different types of vertices and edges discussed above will utilized to find feasible alternatives to the brute force method of analyzing sets of polynomial and polynomial-pair descriptions. Recall that in analyzing a set of descriptions the objective is to determine the extreme or worst-case behavior of the system over all possible

descriptions. The primary goal of the research leading up to this dissertation was to find out if the extreme behavior could be determined using only a few "prominent" descriptions. Examples of "prominent" description are the "vertices" Vert($\supset$), Vert($\mathbb{N}$), Vert($\top$), D(Vert(Δ)), N(Vert(Δ)), and T(Vert(Δ)); and the "edges" Edge($\supset$), Edge($\mathbb{N}$), Edge($\top$), D(Edge(Δ)), N(Edge(Δ)), and T(Edge(Δ)). For any of the analyses, one of the most desirable products of this line of investigation would be a result which showed that the extreme behavior of any description set in a given class could be determined using only the "vertex" descriptions. A result of this type will be referred to as a **vertex theorem.** If the "vertex" descriptions do not determine the worst-case behavior, then it would be desirable to have a result which showed that the "edge" descriptions provide sufficient information about a system's extreme behavior. A result of this type will be called an **edge theorem.** For each of the classical analyses mentioned above, this dissertation will present results in the form of vertex theorems, edge theorems, counterexamples to vertex conjectures, or counterexamples to edge conjectures. A significant portion of these theorems and counterexamples are a contribution of the author.

1.5 Pole-Zero Locations

Using the sets of polynomial descriptions $\mathbb{D}$ and $\mathbb{N}$, the pole and zero location analysis problems mentioned previously can be defined more clearly. This clarification is efficiently achieved using the root space mapping Root: $\mathbb{C}[s] \rightarrow \mathbb{C}$ that is defined for any $\mathbb{P} \subset \mathbb{C}[s]$ by

$$\text{Root}(\mathbb{P}) = \{\ s \in \mathbb{C}\ :\ p(s) = 0,\ p(s) \in \mathbb{P}\ \}.$$

In terms of root spaces, the goal of a zero location analysis is to determine Root($\mathbb{N}$) while a pole location analysis must determine Root($\mathbb{D}$). The two root spaces have very

different meanings in the analysis of a system, but in terms of computation, they are equivalent problems. For this reason, it will be sufficient to investigate non-brute force ways of computing Root(𝔻).

The problem of computing Root(D(Δ)) for the case when D(Δ) contains only real polynomials, Δ is an interval family of real scalars, and D is an affine mapping[2] was considered in [Evans, 1948]. Evans developed the well known Root Locus method for easily constructing an accurate approximation of Root(D(Δ)). This technique showed that D(Vert(Δ)) provides sufficient information to determine the portion of Root(D(Δ)) on the real axis but not the non-real portion of the root space (for all four classes of polynomial sets considered in this dissertation, the vertices do not generally provide sufficient information to determine the root space). A direct extension of root locus techniques to families of polynomials generated by several parameters is generally not feasible.

For the case of multiple parameters, the work of [Zeheb, Walach, 1981] can be used to compute Root(D(Δ)). The work of Zeheb and Walach is very powerful because they place extremely mild restrictions on D and Δ. Describing their result is difficult, because it is so general. Generally speaking, the method begins by looking for parameters over the entire set Δ that satisfy certain algebraic conditions. For the parameters Δ_1 that satisfy the conditions, Root(D(Δ_1)) is computed. Next, attention is restricted to smaller regions of Δ. Parameters Δ_2 that satisfy a new set of algebraic conditions on the smaller regions are found and then Root(D(Δ_2)) is computed. This procedure is continued on smaller and smaller regions of Δ. For the case when Δ is an interval family, the smallest region is Vert(Δ) and the second smallest region is Edge(Δ). The eventual result of this procedure is the set

[2]When the domain is R, affine mappings and multiaffine mappings are equivalent.

$$\text{Root}(D(\Delta_1)) \cup \text{Root}(D(\Delta_2)) \cup \ldots = \text{Root}(D(\Delta_1 \cup \Delta_2 \cup \ldots)) \subset \text{Root}(D(\Delta)).$$

This set covers the boundary of Root(D(Δ)) and divides the complex plane into a finite number of regions. Each of these regions is either contained in or disjoint from Root(D(Δ)). Zeheb and Walach provide a procedure for testing each region to determine which of the two cases holds. Once each region is checked, a complete description of Root(D(Δ)) is available. Because of its generality, this procedure is a very powerful tool, but in some special cases, it is more cumbersome than necessary.

For the special case when $\mathbb{D}$ is a polytope of polynomials, this dissertation will present a theorem which shows that the boundary of Root($\mathbb{D}$), denoted ∂Root($\mathbb{D}$), is a subset of Root(Edge($\mathbb{D}$)). This result is frequently referred to as the (root version of the) **Edge Theorem**. The original version of the Edge Theorem [Bartlett, Hollot, Huang, 1988] is a joint contribution of the author. Using the Edge Theorem to find the boundary of Root($\mathbb{D}$) is simpler than using Zeheb and Walach's method because it avoids solving their algebraic conditions. For the task of finding all the interior points of Root($\mathbb{D}$), Zeheb and Walach's method of testing regions is applicable. It was shown by the author [Bartlett, 1990a] that, for polytopes of polynomials, it is very easy to determine which of the "nonboundary" regions are included in Root(D) and which are excluded. In Chapter 3, the Edge Theorem will be presented, and its use will be demonstrated.

The root version of the Edge Theorem does not extend to sets of polynomials generated by multiaffine uncertainties. An example given by [Barmish, Fu, Saleh, 1988] will be used to show this in Chapter 3. A similar example given by [Ackermann, Hu, Kaesbauer, 1990] shows exactly the same thing. For the multiaffine case, [Zeheb, Walach, 1981] is the best alternative to the brute force approach of computing Root(D).

1.6 Stability

Because this dissertation restricts attention to linear time-invariant systems, stability can be defined using the root space of the characteristic polynomial set $\mathcal{D}$. To allow a variety of problem formulations, the definition of stability will be a general one that is stated in terms of a stability region G that is some subset of the complex plane. The definition states that a set of polynomials $\mathcal{D}$ is **G-stable** if and only if $\text{Root}(\mathcal{D}) \subset G$. For discrete-time systems, the most important stability region is the **open unit disk** G_S. When $\mathcal{D}$ is G_S-stable, it is said that all the polynomials in $\mathcal{D}$ are **strictly Schur**. For continuous-time systems, the most important stability region is the **open left half plane** G_H. When $\mathcal{D}$ is G_H-stable, it is said that all the polynomials in $\mathcal{D}$ are **strictly Hurwitz**. From these definitions, the goal of a worst-case stability analysis is clearly to determine if $\text{Root}(\mathcal{D}) \subset G$. The following subsections will review several methods for carrying out worst-case stability analyses on the four classes of polynomial sets discussed in Section 1.3.

1.6.1 The Root Space Method

From the definition above, one obvious way to check if the set of polynomials $\mathcal{D}$ is G-stable is to first compute $\text{Root}(\mathcal{D})$ and then compare it with G. Stability can therefore be handled using the root location analysis methods discussed above in Section 1.5, but this is not necessarily the best approach. Consider the problem of determining the stability of a single polynomial p(s). For many stability regions G, there are tests that can determine the G-stability of p(s) with much less effort than computation of $\text{Root}(\{p(s)\})$ would require. For G_H-stability, the tests of Hermite, Routh, and Hurwitz are well known examples. For G_S-stability, tests due to Schur,

Cohn, and Jury are also well known. Tests for G_H-stability and G_S-stability can also be used for other stability regions by employing various transformations [Sondergeld, 1983]. Not all stability regions admit efficient alternatives to root computation, but many of the important ones do. Based on the single polynomial case, it is reasonable to expect that there may be methods of determining the stability of a set of polynomial $\mathcal{D}$ that are more computationally efficient than computing Root($\mathcal{D}$).

1.6.2 Classical Algorithms

One alternative to computing Root($\mathcal{D}$) to test stability of $\mathcal{D}$ is to directly extend the methods of Hermite, Routh, Hurwitz, Schur, Cohn, and Jury to sets of polynomials. This extension is similar for all the tests. One may interpret each of these tests as defining several functions $F_0, F_1, \ldots, F_r$ from the set of all real polynomials[3] of degree n, $\mathbb{R}_n[s]$, to $\mathbb{R}$ such that any polynomial p(s) in $\mathbb{R}_n[s]$ is stable[4] if and only if the real numbers $F_0(p(s)), F_1(p(s)), \ldots, F_r(p(s))$ are all positive. In Hurwitz's method for example, the numbers $F_0(p(s)), F_1(p(s)), \ldots, F_r(p(s))$ would be the leading principle minors of the Hurwitz matrix constructed using the coefficients of p(s). When extended to sets of polynomials of fixed degree[5], the classical tests state that $\mathcal{D} \subset \mathbb{R}_n[s]$ is stable if and only if the real sets $F_0(\mathbf{D}), F_1(\mathbf{D}), \ldots, F_r(\mathbf{D})$ contain only positive numbers. When $\mathbf{D} = D(\Delta)$, showing that $F_0(\mathbf{D}), F_1(\mathbf{D}), \ldots, F_r(\mathbf{D})$ are sets of positive numbers is equivalent to showing that the composite mappings $F_0 \bullet D, F_1 \bullet D, \ldots, F_r \bullet D$ are all positive over the set Δ.

---

[3] More complicated versions of each test are available to handle complex polynomials.
[4] G_H-stable or G_S-stable as appropriate.
[5] Sets containing polynomials of more than one degree can be analyzed, but more effort is generally required.

Showing that $F_0 \circ D, F_1 \circ D, \ldots, F_r \circ D > 0$ over the set Δ can be an extremely complicated task even when D is a fairly simple mapping. If D is a multiaffine or even an affine mapping, the functions $F_0 \circ D, F_1 \circ D, \ldots, F_m \circ D$ will be multivariate polynomials of the individual parameters $\delta_1, \delta_2, \ldots, \delta_m$. The positivity over the set Δ of these multivariate polynomials can be tested using various finite step algorithms [see Anderson, Scott, 1977; Bickart, Jury, 1978; Walach, Zeheb, 1980], but these algorithms are in general more complex than is feasible to implement [see the discussion of complexity by Walach and Zeheb, 1980]. By first testing any one polynomial in $\mathbb{D}$ for stability, the stability conditions $F_0(\mathbb{D}), F_1(\mathbb{D}), \ldots, F_r(\mathbb{D}) > 0$ can be reduced to a few "critical" stability conditions $F_i(\mathbb{D}), F_j(\mathbb{D}), \ldots > 0$ [Frazer, Duncan, 1929; Jury, 1974], but even this simplification will not make this approach feasible in general.

The extension of the classical algorithms to test the stability of a family generated by affine or multiaffine uncertainties may not be feasible in general, but it is feasible for the special case when only a one uncertain parameter is present. In this case, the critical conditions $F_i \circ D, F_j \circ D, \ldots$ will be univariate polynomials of the individual parameter δ_1. It is easy to test the positivity of a univariate polynomial over a real interval Δ, so the direct extensions of Hermite, Routh, Hurwitz, Schur, Cohn, and Jury can be used in this case. The eigenvalue tests of [Bialas, 1985; Fu, Barmish, 1988; Bartlett, Hollot, 1988; Ackermann, Barmish, 1988, Saydy, Tits, Abed, 1988], the resultant test of [Bose, 1989], and part of the test conditions in [Kraus, Anderson, Jury, Mansour, 1988] are all essentially results of this type. In summary, the classical stability tests are in general excessively difficult to use on sets of polynomials, but they are relatively easy to use in the special cases when the set of polynomials is generated by a single multiaffine uncertainty.

1.6.3 Vertex and Edge Stability Theorems

The fact that the classical stability algorithms are difficult to use with many uncertain parameters but easy to use with less than two uncertain parameters provides motivation for considering vertex and edge theorems. A vertex theorem would equate stability of the whole set of descriptions to the stability of a finite number of "vertices." Each of these vertices could be tested using the classical algorithms without modifications. An edge theorem would equate stability of the whole family to stability of a finite number of edges. Each edge is essentially a set of descriptions generated by a single multiaffine uncertainty, so the stability of each edge could be tested using a one parameter test [Bialas, 1985; Fu, Barmish, 1988; Bartlett, Hollot, 1988; Ackermann, Barmish; 1988, Saydy, Tits, Abed, 1988; Bose, 1989]. There are several vertex and edge theorems available that can be used in this manner.

Vertex stability theorems have a long history dating back to the previous century. Chebyshev presented a vertex G_H-stability theorem in 1892, and Markov presented a similar but more general vertex G_H-stability theorem in 1894 [see Gantmacher, 1959, vol. 2, pp. 240-248]. A recent exposition on Markov's theorem is given by [Hollot, 1989]. Markov's theorem is not directly applicable to the classes of description sets considered in this dissertation because this theorem is stated in terms of polynomials represented in Markov parameter space rather than in $\mathbb{R}[s]$ or $\mathbb{C}[s]$. The possibility of mapping a set $\mathbb{D} \subset \mathbb{C}[s]$ into Markov parameter space and then applying Markov's theorem on the image has not been fully investigated, so it is not clear how difficult the process would be or how much conservatism it would introduce.

Vertex results are also available for G_S-stability. At the beginning of this century, Perron, 1907, and Frobenius, 1909, developed powerful theorems concerning the largest real eigenvalues of positive matrices. These results have lead directly to a

simple G_S-stability[6] vertex theorem due to Wielandt, 1950 [see Gantmacher, 1959, vol. 2, p. 57]. Wiedlant's result has been extended and strengthened recently by [Mayer, 1989]. Like the original theorems, these results are stated in terms of the eigenvalues of square matrices, but this not a significant detraction because it is a simple matter to transform a polynomial root problem into a matrix eigenvalue problem. The main drawback of the Wielandt-Mayer theorem is that it provides necessary and sufficient conditions only for certain types of interval families. In general, the results can only be used in a conservative fashion. Like Markov's theorem, the Wielandt-Mayer theorem is useful, but it doesn't provide a general solution to any of the problems consider in this work.

A theorem due to Kharitonov [1978a&b] provides a complete solution to a problem considered in this dissertation. The weak version of Kharitonov's theorem shows that an interval family of polynomials is G_H-stable if and only if all its vertices are G_H-stable. The strong version of Kharitonov's theorem shows that an interval family of polynomials is G_H-stable if and only if eight specific vertices are G_H-stable. Kharitonov's amazing result and several other vertex theorems which extend or partially extend his work will be reviewed in Chapter 3. Unfortunately, many desirable extensions of Kharitonov's theorem do not exist. Even the weak version does not extend to polynomial sets that are generated by affine uncertainties. The example in [Kochenburger, 1953] shows a family of polynomials generated by a single affine uncertainty that is not G_H-stable even though it has G_H-stable vertices. Kharitonov's weak theorem for interval families does extend to some other stability regions [Soh, Berger, 1988; Petersen, 1989; Fu, 1989a], but it does not extend to many important stability regions including the open unit disk. The example in [Hollot, Bartlett, 1986]

[6]This theorem is valid for disks centered at the origin of any radius.

shows an interval family of polynomials that is not G_S-stable even though it has G_S-stable vertices. For those problems in which a vertex theorem like Kharitonov's does not hold, the best that can be hoped for is an edge theorem.

Based on the Edge Theorem for root locations, it was shown by [Bartlett, Hollot, Huang, 1988] that an edge theorem also holds for several stability problems. Their work showed that if $\mathbb{D}$ is a polytope of monic real polynomials of degree n and if G is a simply connected set, then $\mathbb{D}$ is G-stable if and only if Edge($\mathbb{D}$) is G-stable. This theorem and subsequent extensions are referred to as **stability versions of the Edge Theorem**. Chapter 3 will discuss the extensions of this theorem [Fu, Barmish, 1989; Hollot, Looze, Bartlett, 1990; Barmish, Sideris, 1989] which have sought to relax the assumptions on $\mathbb{D}$ and G in various ways. Chapter 3 will also present examples from [Bartlett, 1990a] which show that the assumptions cannot simply be removed. For stability problems that violate the previously used assumptions, a modified stability version of the Edge Theorem that includes a **precondition** [Bartlett, 1990a] can be employed. Chapter 3 will show how to use the precondition and Edge Theorem to determine the G-stability of a polytope of polynomials.

As might be expected from the root location problem, the stability version of the Edge Theorem is generally not valid for sets generated by multiaffine uncertainties. A counterexample given by [Barmish, Fu, Saleh, 1988] clearly shows this fact. Since edges are not sufficient, one might want to know if some larger "geometrically inspired" subset such as faces would be sufficient to determine stability. A counterexample in [Ackermann, 1989] shows that no proper subset of this type is sufficient. For the multiaffine case, the only way to identify the crucial parameters to test would be to use problem dependent algebraic conditions like those given in [Zeheb, 1990]. This algebraic approach is more or less equivalent to the classical approach

discussed in Section 1.6.2, so it has the potential to be very complex when many uncertain parameters are present.

1.6.4 The Zero Exclusion Principle

Another alternative to computing Root(𝔻) is to test stability using the "zero exclusion principle" [Frazer, Duncan, 1929]. To test the G-stability of a connected set of polynomials of fixed degree 𝔻 for any open set G, this method requires two steps[7]. The first step is to check the G-stability of any polynomial p(s) in 𝔻 (p(s) is sometimes called the nominal polynomial). If p(s) is not G-stable, then 𝔻 is not G-stable and there is no need to go to the second step. If p(s) is G-stable, then the next step must be carried out. The second step is to determine if there is any complex number z on the boundary of G, denoted ∂G, such that zero is in the value set

$$Val(\mathbb{D},z) = \{ p(z) \in \mathbb{C} : p(s) \in \mathbb{D} \}.$$

If there does exist a $z \in \partial G$ such that $0 \in Val(\mathbb{D},z)$, then 𝔻 is not stable. If $0 \notin Val(\mathbb{D},z)$ for all $z \in \partial G$, then 𝔻 is G-stable. The zero exclusion method has two commonly used interpretations. The Nyquist interpretation explains the condition $0 \notin Val(\mathbb{D},z)$ for all $z \in \partial G$ as preventing a destabilizing change in the number of encirclements of the Nyquist plot. The root continuity interpretation explains the condition $0 \notin Val(\mathbb{D},z)$ for all $z \in \partial G$ as preventing stable poles from migrating into unstable regions. Both interpretations have motivated results in the literature.

One obvious way to use the zero exclusion principle is to show Val(𝔻,z) in the complex plane for each frequency $z \in \partial G$. If 𝔻 is an interval family of polynomials

[7]If 𝔻 is not connected or contains polynomials of different degrees, then this method can be used by separating 𝔻 into several connected subsets each of which contains only polynomials of the same degree. The first step of this method will have to be repeated separately for each subset. The second step does not need to be altered.

and if $z \in \partial G_H$, then $Val(\mathbb{D},z)$ is simple to construct because it is a rectangle with corners that are easily identified using Kharitonov's polynomials [see Dasgupta, 1988; Minnichelli, Anagnost, Desoer, 1989]. If $\mathbb{D}$ is a polytope of polynomials, then for any frequency z, $Val(\mathbb{D},z)$ is simply the convex polygon whose vertices correspond to the vertices of $\mathbb{D}$, i.e. $Val(\mathbb{D},z)$ equals the convex hull of $Val(Vert(\mathbb{D}),z)$, denoted $conv(Val(Vert(\mathbb{D}),z))$, [Barmish, 1989]. For polynomial sets generated by multiaffine uncertainties, $Val(\mathbb{D},z)$ is a subset of $conv(Val(Vert(\mathbb{D}),z))$, but it does not necessarily equal this convex polygon [see the Mapping Theorem, Zadeh, Desoer, 1963, page 476]. Due to the lack of convexity, $Val(\mathbb{D},z)$ can be fairly difficult to determine for the multiaffine case. Ackermann, Hu, and Kaesbauer [1990] illustrate ways to reduce the amount of difficulty encountered in using this graphical stability test on polynomial sets generated by multiaffine uncertainties.

An alternative to visually checking if $0 \in Val(\mathbb{D},z)$ is to use an algebraic algorithm. For polytopes of polynomials, a non-iterative algorithm of this type is given by [Barmish, 1989]. An iterative algorithm that can handle sets generated by multiaffine uncertainties is given by [de Gaston, Safonov, 1988][8]. The main problem suffered by these two algorithms and the visual approach discussed above is the need to sweep over all $z \in \partial G$.

Several methods based on the zero exclusion principle but avoiding frequency sweeps have been proposed. For polytopes of polynomials, [Zeheb, 1989] provides a method for identifying a finite number of crucial boundary frequencies. The condition $0 \notin Val(\mathbb{D},z)$ needs to be checked only on these frequencies. In [Djaferis, Hollot, 1989], a similar method (for $G = G_H$) is given that takes advantage of the special structure of sets of polynomials generated by affine uncertainties. Motivated by the

[8] A paper by [Sideris, de Gaston, 1986] extend this algorithm to handle polynomic uncertainty structures as well as multiaffine uncertainties.

algorithms of [de Gaston, Safonov, 1988; Sideris, de Gaston, 1986], Sideris and Sánchez Peña [1989] provide an algorithm that can test the G_H-stability of sets generated by multiaffine uncertainties without the need for a frequency sweep[9].

1.6.5 Parameter Space Methods

Another way to test stability of a family of polynomials is to use parameter space methods [see Naimark, 1947; Siljak, 1969; Fam, Meditich, 1978; Ackermann 1980; Siljak, 1989]. For a set of polynomials $\mathbb{D} = D(\Delta)$ and an open stability region G, this method maps the stability boundary $\partial G \subset \mathbb{C}$ back into the parameter space $\mathbb{R}^m$. The image is the set

$$\mathbb{B} = \{ \delta \in \mathbb{R}^m : p(z) = 0,\ p(s) = D(\delta),\ z \in \partial G \}.$$

If $\mathbb{B} \cap \Delta = \emptyset$, if $\mathbb{D}$ is a connected set of polynomials of a fixed degree, and if $\mathbb{D}$ contains at least one G-stable polynomial, then $\mathbb{D}$ is guaranteed to be G-stable. The condition $\mathbb{B} \cap \Delta = \emptyset$ is the parameter space counterpart of the zero exclusion principle. For m = 2, it is typically not too difficult to determine $\mathbb{B}$ and to check if $\mathbb{B}$ intersects Δ. This is often carried out graphically. For larger m = 3, displaying $\mathbb{B}$ becomes difficult, and for m > 3, a graphical representation of $\mathbb{B}$ becomes impossible. There are various ways to circumvent this dimensionality problem, but in general, this approach is very difficult to use when several parameters are present. This is true even for simple sets of polynomials such as interval families.

---

[9]This algorithm avoids frequency sweeping by using the Routh-Hurwitz criteria rather than the zero exclusion principle. This algorithm is noted in this section rather than the classical algorithms section because it is generally considered to be an extension of the zero exclusion based algorithms of [de Gaston, Safonov, 1988, Sideris, de Gaston, 1986]. Sideris and Sánchez Peña present this method as a G_H-stability algorithm, but it has broader applications because the method provides an algorithm for testing the positivity of a multivariate polynomial over an interval family of real m-tuples.

1.6.6 Lyapunov Methods

Lyapunov's direct method provides another means of testing the stability of uncertain systems. The reader is referred to the recent survey by [Siljak, 1989] for an introduction to these methods and an extensive bibliography of related papers. Lyapunov's method has been used mainly for G_H-stability, but it has the advantage that it can handle nonlinear and time varying uncertainties. It has also been used to design feedback controllers that will stabilize uncertain systems. The main drawback of this method is that it provides mainly sufficient conditions. For the classes of uncertain systems considered in this dissertation, the Lyapunov methods referenced in [Siljak, 1989] will give conservative stability analysis results.

1.7 Frequency Response

A worst-case frequency response analysis can have many objectives such as determining a system's minimum bandwidth or its maximum level of amplification. Traditionally, for systems with no uncertain parameters, these objectives were achieved using two steps. First, at several appropriately selected frequencies along the stability boundary, the frequency response of the system's model was computed and then displayed visually using the Nyquist plane, the Nichols chart, or the Bode magnitude and phase plots.. The second step was simply to obtain the desired information about the system by inspecting the graphical information. This traditional approach can also be used for systems with parameter uncertainties [Horowitz, 1963]. In the presence of uncertainties, the first step involves the computation of a set of responses at each frequency rather than a single response. The set of responses are also displayed graphically using any of the three common methods. The second step can still be

carried out by inspection but the information obtained may have a different interpretation and form. For example, the visual information may indicate a range of bandwidths rather one bandwidth. The main difficulty with this type of analysis is the computational burden of carrying out the first step. For this reason, this dissertation will focus on methods of simplifying the computation of the frequency response.

Using the set of polynomial-pairs T, the desired frequency response information can be defined more clearly. This information is defined separately for each frequency z. A Nyquist plane analysis requires the set

$$\mathrm{Nyq}(T,z) = \left\{ \frac{n(z)}{d(z)} \in \mathbb{C} \;:\; \mathrm{abs}\left[\frac{n(z)}{d(z)}\right] < \infty,\ (n(s),d(s)) \in T \right\}.$$

For use with a Nichols chart analysis, the set

$$\mathrm{Nic}(T,z) = \left\{ \left(20\mathrm{Log}\left(\mathrm{abs}\left[\frac{n(z)}{d(z)}\right]\right), \arg\left[\frac{n(z)}{d(z)}\right] \right) \in \mathbb{R}^2 \;:\; 0 < \mathrm{abs}\left[\frac{n(z)}{d(z)}\right] < \infty,\ (n(s),d(s)) \in T \right\}$$

is needed. The set Nic(T,z) is often called a **template** [Horowitz, 1982]. Bode magnitude and phase plots utilize the two sets

$$\mathrm{Mag}(T,z) = \left\{ \mathrm{abs}\left[\frac{n(z)}{d(z)}\right] \in \mathbb{R} \;:\; \mathrm{abs}\left[\frac{n(z)}{d(z)}\right] < \infty,\ (n(s);d(s)) \in T \right\}$$

and

$$\mathrm{Arg}(T,z) = \left\{ \arg\left[\frac{n(z)}{d(z)}\right] \in \mathbb{R} \;:\; 0 < \mathrm{abs}\left[\frac{n(z)}{d(z)}\right] < \infty,\ (n(s),d(s)) \in T \right\}.$$

These sets must be computed at each boundary frequency of interest.

Worst-case frequency response analyses have been carried out for many years, but until recently, very few feasible methods of computing Nyq(T,z), Nic(T,z), Mag(T,z), or Arg(T,z) have been presented. For this reason, approximations of the frequency response sets are often used. These approximations are obtained by

gridding the set of systems descriptions and computing only the responses of those models on the grid. The grid might be formed by using several evenly spaced values for each uncertain parameter. It is easy to see that gridding can be computationally intensive. For example, if there are 5 uncertain parameter and if 10 values are used for each, then it will be necessary to compute the frequency response of 10^5 models. The computational demand can be reduced by using fewer grid points, but that will reduce the accuracy of the approximation. In some cases, this tradeoff of accuracy and computational burden has no satisfactory compromise.

Alternatives to gridding can be obtained by exploiting any and all structure in the set of possible descriptions. The works of Horowitz [for example Horowitz, 1963 & 1982] indicate that there are an assortment of ways to intelligently utilize this structure. By using these methods, gridding can be reduced but generally not eliminated. Because these works [Horowitz, 1963 & 1982] focus on using the frequency response rather than computing it, the ways of avoiding gridding appear mainly in examples and have not been developed into a general theory. Recently, for the case of interval families of polynomial-pairs, [Bailey, Panzer, 1988] have given general methods for efficient computation of frequency response sets . For special types of affine uncertainties, [Bailey, Hui, 1989] present algorithms for easily computing the frequency response sets. These algorithms can also be used with more general types of affine uncertainties, but a conservative overbound of the frequency response sets will be obtained. For general parameter uncertainties, [Bailey, Panzer, Gu, 1988] noted that the frequency response sets can also be computed using nonlinear programming, but they point out that this method is very difficult. This dissertation will show that vertex and edge theorems exist which can be used to avoid gridding and nonlinear programming in some general cases.

Results concerning the availability of vertex and edge theorems for frequency response calculation will be presented in Chapter 4. For polytopes of polynomial-pairs, it will be shown that the Nyquist, Bode, and Nichols plots of $\overline{T}$ can be determined using only Edge($\overline{T}$). This edge theorem provides nonconservative results for a larger class of uncertain systems than handled by the algorithm of [Bailey, Hui, 1989]. This edge theorem is an original contribution of the author [Bartlett, 1990b]. Special results concerning the response of interval families of polynomial-pairs to purely imaginary frequencies $z = jw$ will also be presented in Chapter 4. It will be shown that the Nyquist and Nichols plots can be determined using at most 32 specific edge-like subsets instead of all edges and that the Bode Magnitude plots can be determined using only 18 vertex-like descriptions. These interval family results were discussed in [Bartlett, 1990b]. The formulation of the later two results is original, but *their contributions are not* particularly significant because the methods of [Bailey, Panzer, 1988] will compute the frequency response of interval families with a similar amount of effort. The later two results are included in this dissertation because they provide a nice vertex/edge theorem interpretation for the interval family case. Concerning polynomial-pair sets generated by multiaffine uncertainties, an example (based on a stability counterexample due to [Barmish, Fu, Saleh, 1988]) will be given which shows that the edges do not provide sufficient information to determine the frequency response. The results in Chapter 4 answer most questions concerning the existence or nonexistence of frequency response vertex and edge theorems for the four classes of polynomial-pairs considered in this dissertation.

1.8 Time Responses

In addition to sinusoidal responses, this dissertation will also consider the problem of analyzing the response of an uncertain system to an input comprised of steps, ramps, and various other signals. It will be shown in Chapter 5 that, for continuous and discrete-time stable systems represented by sets of polynomial-pairs $\mathcal{T} = T(\Delta)$ generated by multiaffine uncertainties, the extreme values of the steady state response can be determined using only the vertex descriptions $T(Vert(\Delta))$. For the special case when $\mathcal{T}$ is a interval family of polynomial-pairs, the maximum and minimum steady state response can be determined using only small easily identified subsets of $Vert(\mathcal{T})$. These steady state results will be illustrated by examples. Concerning transient response analyses, there are currently no vertex or edge theorems available, but some negative results exist. Chapter 5 will present examples which show that if $\mathcal{T}$ is generated by affine uncertainties, then the maximum overshoot of the step response does not necessarily occur at a vertex. It is not currently known whether or not edges would provide the desired transient response information for sets of polynomial-pairs generated by affine uncertainties (or more general classes of polynomial-pairs). It is also not known whether or not a transient response edge or vertex theorem exists for interval families of polynomial-pairs. The steady state response theorems and the transient response counterexamples in Chapter 5 are all independent original results of this author [Bartlett, 1990c].

The steady state vertex theorems in Chapter 5 are most closely related to the frequency response results in [Bartlett, 1989 & 1990b; Fu, 1989b]. For an uncertain system with affine uncertainties responding to a sinusoidal input (possibly a step), the steady state amplitude and phase of the sinusoidal output can be determined using [Bartlett, 1989 & 1990b; Fu, 1989b]. The results in Chapter 5 cannot be used with

pure sinusoidal inputs, but they can be used with steps and many other inputs that produce a constant steady state output. On the area of overlap, i.e. step responses, the results in Chapter 5 are stronger than those in [Bartlett, 1989 & 1990b; Fu, 1989b] in two respects. First, Chapter 5 contains vertex theorems rather than edge theorems, so the methods of Chapter 5 are much simpler to use. Second, the results in [Bartlett, 1989 & 1990b; Fu, 1989b] only handle affine uncertainties while Chapter 5 allows multiaffine uncertainties. Other approaches, besides vertex and edge theorems, have been used to investigate the responses of systems with parametric uncertainties. For example, [Oppenheimer, Michel, 1988] have applied interval analysis techniques to this problem. The results in [Oppenheimer, Michel, 1988] can handle transient response as well as steady state, but they provide only approximate bounds on the response. The steady state vertex theorems in Chapter 5 provide nonconservative bounds. This discussion shows that the steady state vertex theorems in Chapter 5 are a contribution to the area of steady state analysis of uncertain systems.

1.9 Organization

For several important classes of uncertain systems, this dissertation presents results concerning the existence of vertex and edge theorems related to many fundamental classical analyses. The notion of vertex and edge theorems has been discussed in this introduction. The four classes of uncertain systems that are investigated are presented in Chapter 2. Chapter 3 address two closely related fundamental analysis problems, stability and pole-zero locations. The topic of frequency response analysis is cover in Chapter 4. Time response analysis is covered in Chapter 5. For each class of uncertain systems that is considered, Chapter 6 will review which analysis problems are known to have vertex or edge theorems, which

problems are known not to admit vertex or edge theorems, and which problems are still open concerning the existence of vertex or edge theorems. Chapter 6 will also summarize the contributions the author has made to this broad range of fundamental linear systems analysis problems.

CHAPTER 2

CLASSIFICATION OF UNCERTAINTY STRUCTURE

2.1 Introduction

This dissertation studies the analysis of systems with parameter uncertainties that can be described by characteristic polynomial sets $\mathbb{D}$, numerator polynomial sets $\mathbb{N}$, and numerator-denominator polynomial-pair sets $\mathbb{T}$. This investigation will focus on those systems whose description sets are contained in a few general classes. The main purpose of this chapter is to define these classifications. In all, four classes of polynomial sets and the four classes of polynomial-pair sets will be introduced. This chapter will also point out the highly structured nature of these classes. This structure will be used to describe the vertices and edges of description sets contained in these classes. The vertex and edge descriptions will play a key role in the analysis of systems in the given classes.

2.2 Generation Based Classifications

For each system with uncertain parameters, the set of possible parameter values $\Delta \subset \mathbf{R}^m$ and the mappings $D: \mathbf{R}^m \to \mathbf{C}[s]$, $N: \mathbf{R}^m \to \mathbf{C}[s]$, and $T: \mathbf{R}^m \to \mathbf{C}[s]^2$ define the description sets

$$\mathbb{D} = D(\Delta) \qquad \mathbb{N} = N(\Delta) \qquad \mathbb{T} = T(\Delta).$$

These relationships indicate that a natural way to define classifications for $\mathbb{D}$, $\mathbb{N}$, and $\mathbb{T}$ is to restrict the mappings D, N, and T, respectively, and the set Δ. This section will

define two classes of polynomial sets and two classes of polynomial-pair sets in this manner.

2.2.1 Interval Families of Real m-Tuples

Only one class of parameter sets will be used to restrict Δ. This class is called interval families of real m-tuples and is defined as follows.

Definition 2.1 A set of parameters Δ is said to be an **interval family of real m-tuples** if there exist two real m-tuples

$$\delta^L = (\delta_1^L, \delta_2^L, \ldots , \delta_m^L)$$

$$\delta^H = (\delta_1^H, \delta_2^H, \ldots , \delta_m^H)$$

such that

$$\Delta = \mathrm{Box}\{\delta^L,\delta^H\}$$

where

$$\mathrm{Box}\{\delta^L,\delta^H\} = \{ (\delta_1, \delta_2, \ldots , \delta_m) \in \mathbf{R}^m \ : \ \delta_i^L \le \delta_i \le \delta_i^H, \ i = 1, 2, \ldots , m \}.$$

Interval families of real m-tuples are highly structured, so it is easy to describe their vertices and edges.

Definition 2.2 The **set of all vertices** of the interval family $\Delta = \mathrm{Box}\{\delta^L,\delta^H\}$ is defined as

$$\mathrm{Vert}\{\delta^L,\delta^H\} = \{ (\delta_1, \delta_2, \ldots , \delta_m) \in \mathbf{R}^m \ : \ \delta_i \in \{\delta_i^L,\delta_i^H\}, \ i = 1, 2, \ldots , m \}.$$

The **set of all edge m-tuples** of the interval family $\Delta = \mathrm{Box}\{\delta^L,\delta^H\}$ is defined as

$$\text{Edge}\{\delta^L,\delta^H\} = \{\ (\delta_1, \delta_2, \ldots, \delta_m) \in \mathbb{R}^m\ :\ \delta_i \in \{\delta_i^L,\delta_i^H\},$$

$$i = 1, 2, \ldots, j\text{-}1, j\text{+}1, \ldots, m,\ \delta_j^L \le \delta_j \le \delta_j^H,\quad j = 1, 2, \ldots, m\ \}.$$

It is quite natural for the set Δ to be an interval families of real m-tuples. Anytime the parameters are independent, the set of all possible parameter values will be a member of this class.

2.2.2 Affine and Multiaffine Mappings

Two types of mappings will be used to restrict D, N, and T. The first type of mappings are called affine operators, and they are defined as follows.

Definition 2.3 Let $\mathbf{V}_1$ and $\mathbf{V}_2$ be real vector spaces. The mapping $F: \mathbf{V}_1 \to \mathbf{V}_2$ is said to be an **affine mapping** if for $y, z \in \mathbf{V}_1$ and $\lambda \in \mathbf{R}$

$$F[\lambda y+(1-\lambda)z] = \lambda F[y] + (1-\lambda)F[z].$$

The set of all such mappings from $\mathbf{V}_1$ to $\mathbf{V}_2$ will be denoted by $\text{Affine}\{\mathbf{V}_1,\mathbf{V}_2\}$.

Affine operators are closely related to linear operators. In fact, if $F: \mathbf{V}_1 \to \mathbf{V}_2$ is an affine operator, then the function $G: \mathbf{V}_1 \to \mathbf{V}_2$ given by $G[x] = F[x] - F[0]$ is a linear operator. The second type of mappings are generalizations of affine operators for the special case when the domain is $\mathbf{R}^m$. The mappings are called multiaffine operators and are defined below.

Definition 2.4 Let $\mathbb{V}$ be a real vector space. The mapping $F: \mathbb{R}^m \to \mathbb{V}$ is said to be a **multiaffine mapping** if for $i \in \{1, 2, \ldots, m\}$, $\lambda \in \mathbb{R}$, and $y = (y_1, y_2, \ldots, y_m)$, $z = (z_1, z_2, \ldots, z_m) \in \mathbb{R}^m$

$$F[(y_1, y_2, \ldots, y_{i-1}, \lambda y_i+(1-\lambda)z_i, y_{i+1}, \ldots, y_m)] =$$

$$\lambda F[(y_1, y_2, \ldots, y_{i-1}, y_i, y_{i+1}, \ldots, y_m)]$$

$$+ (1-\lambda)F[(y_1, y_2, \ldots, y_{i-1}, z_i, y_{i+1}, \ldots, y_m)].$$

The set of all such mappings from $\mathbb{R}^m$ to $\mathbb{V}$ will be denoted $\mathrm{Multi}\{\mathbb{R}^m,\mathbb{V}\}$.

Example 2.1 Let $F_1: \mathbb{R}^3 \to \mathbb{R}$, $F_2: \mathbb{R}^3 \to \mathbb{C}[s]$, and $F_3: \mathbb{R}^3 \to \mathbb{C}[s]^2$ be the mappings such that

$$F_1[(x,y,z)] = 3x + 2y - z + 1$$

$$F_2[(x,y,z)] = (3x+y+z+1)s^2 + (-4y-4z+2)s + (-7x+y-7)$$

$$F_3[(x,y,z)] = ((z-11)s^2 + 9s + (6y),\ (22x+13)s^2 + (5x-5y-5z+1)s + (-2)).$$

Let $G_1: \mathbb{R}^3 \to \mathbb{R}$, $G_2: \mathbb{R}^2 \to \mathbb{C}[s]$, and $G_3: \mathbb{R}^2 \to \mathbb{C}[s]^2$ be the mappings such that

$$G_1[(x,y,z)] = 3xyz + 2xy + xz - yz + 6x - 7y - 9z - 11$$

$$G_2[(x,y)] = (3xy-x+y+1)s^2 + (-4xy+2)s + (-7x+y-7)$$

$$G_3[(x,y)] = ((xy+x+4y-11)s^2 + 9s + (6y),\ (xy-22x+13)s^2 + (5x-5y-5z+1)s + (-2)).$$

These functions can be classified as follows

$$F_1 \in \mathrm{Affine}\{\mathbb{R}^3,\mathbb{R}\} \subset \mathrm{Multi}\{\mathbb{R}^3,\mathbb{R}\}$$

$$F_2 \in \mathrm{Affine}\{\mathbb{R}^3,\mathbb{C}[s]\} \subset \mathrm{Multi}\{\mathbb{R}^3,\mathbb{C}[s]\}$$

$$F_3 \in \mathrm{Affine}\{\mathbb{R}^3,\mathbb{C}[s]^2\} \subset \mathrm{Multi}\{\mathbb{R}^3,\mathbb{C}[s]^2\}$$

$$G_1 \notin \mathrm{Affine}\{\mathbb{R}^3,\mathbb{R}\} \qquad G_1 \in \mathrm{Multi}\{\mathbb{R}^3,\mathbb{R}\}$$

$$G_2 \notin \mathrm{Affine}\{\mathbb{R}^2,\mathbb{C}[s]\} \qquad G_2 \in \mathrm{Multi}\{\mathbb{R}^2,\mathbb{C}[s]\}$$

$$G_3 \notin \mathrm{Affine}\{\mathbb{R}^2,\mathbb{C}[s]^2\} \qquad G_3 \in \mathrm{Multi}\{\mathbb{R}^2,\mathbb{C}[s]^2\}.$$

2.2.3 Description Sets Generated by Affine Uncertainties

This section will use the definitions of interval families of real m-tuples and of affine mappings to specify a class of polynomial sets and a class of polynomial-pair sets. The description sets in both of these classes are said to be generated by affine uncertainties. These classes are define as follows.

Definition 2.5 The set of polynomial-pairs $\mathbf{T} = T(\Delta)$ is said to be **generated by affine uncertainties** if Δ is an interval family of real m-tuples and if $\mathbf{T} \in \text{Affine}\{\mathbf{R}^m, \mathbf{C}[s]^2\}$.

Definition 2.6 The set of polynomials $\mathbf{D} = D(\Delta)$ is said to be **generated by affine uncertainties** if Δ is an interval family of real m-tuples and if $\mathbf{D} \in \text{Affine}\{\mathbf{R}^m, \mathbf{C}[s]\}$. Similarly, the set of polynomials $\mathbf{N} = N(\Delta)$ is said to be generated by affine uncertainties if Δ is an interval family of real m-tuples and if $\mathbf{N} \in \text{Affine}\{\mathbf{R}^m, \mathbf{C}[s]\}$.

Examples of description sets generated by affine uncertainties are given by **T**, **D**, and **N** from Example 1.1.

When referring to a set generated by affine uncertainties $N(\Delta)$, $D(\Delta)$, or $T(\Delta)$, the term **vertex** descriptions will refer to the sets $N(\text{Vert}(\Delta))$, $D(\text{Vert}(\Delta))$, or $T(\text{Vert}(\Delta))$, respectively. Similarly, the term **edge** descriptions will refer to the sets $N(\text{Edge}(\Delta))$, $D(\text{Edge}(\Delta))$, or $T(\text{Edge}(\Delta))$, respectively.

2.2.4 Description Sets Generated by Multiaffine Uncertainties

This section will use the definition of interval families of real m-tuples and of multiaffine mappings to specify a class of polynomial sets and a class of polynomial-pair sets. The description sets in both of these classes are said to be generated by multiaffine uncertainties. These classes are define as follows.

Definition 2.7 The set of polynomial-pairs $\mathbf{T} = T(\Delta)$ is said to be **generated by multiaffine uncertainties** if Δ is an interval family of real m-tuples and if $T \in \text{Multi}\{\mathbf{R}^m, C[s]^2\}$.

Definition 2.8 The set of polynomials $\mathbf{D} = D(\Delta)$ is said to be **generated by multiaffine uncertainties** if Δ is an interval family of real m-tuples and if $D \in \text{Multi}\{\mathbf{R}^m, C[s]\}$. Similarly, the set of polynomials $\mathbf{N} = N(\Delta)$ is said to be generated by multiaffine uncertainties if Δ is an interval family of real m-tuples and if $N \in \text{Multi}\{\mathbf{R}^m, C[s]\}$.

Examples of description sets generated by multiaffine uncertainties are given by **T**, **D**, and **N** in both Example 1.1 and 1.2.

When referring to a set generated by multiaffine uncertainties $N(\Delta)$, $D(\Delta)$, or $T(\Delta)$, the term **vertex** descriptions will refer to the sets $N(\text{Vert}(\Delta))$, $D(\text{Vert}(\Delta))$, or $T(\text{Vert}(\Delta))$, respectively. Similarly, the term **edge** descriptions will refer to the sets $N(\text{Edge}(\Delta))$, $D(\text{Edge}(\Delta))$, or $T(\text{Edge}(\Delta))$, respectively.

2.3 Image Based Classifications

This section will define two classifications of polynomial sets and two classifications of polynomial-pair sets. These definitions are stated directly in terms of the sets $\mathbb{T}$, $\mathbb{D}$, and $\mathbb{N}$, so the definitions don't depend on the set Δ or on the mappings D, N, or T. The four classifications presented in this section are called polytopes of polynomials, polytopes of polynomial-pairs, interval families of polynomials, and interval families of polynomial-pairs.

2.3.1 Polytopes

A class of highly structured sets called polytopes will be defined in this section. The conventions of [Grunbaum, 1967; Brønsted, 1983] will primarily be used; the main difference is that the definitions below are stated for any real vector space rather than just for $\mathbf{R}^n$. This extension is easily achieved because the definition of a polytope relies mainly on the notion of convexity which can be defined for any real vector space.

Definition 2.9 Let **V** be a real vector space. A set $\mathbf{A} \subset \mathbf{V}$ is said to be a **convex set** if

$$\{ \lambda_1 x + \lambda_2 y : \lambda_1, \lambda_2 \in \mathbf{R},\ \lambda_1, \lambda_2 \geq 0,\ \lambda_1 + \lambda_2 = 1 \} \subset \mathbf{A}$$

for all $x, y \in \mathbf{A}$.

Definition 2.10 Let **V** be a real vector space. The **convex hull** of a set $\mathbf{A} \subset \mathbf{V}$, denoted conv(A), is defined to be the intersection of all convex sets that contain **A**.

Using the convex hull operation, polytopes as well as their vertices and edges are defined as follows.

Definition 2.11 Let $\mathbb{V}$ be a real vector space. A set $\mathbf{A} \subset \mathbb{V}$ is said to be a **polytope** if there exists a finite collection of vectors $y_1, y_2, \ldots, y_r \in \mathbb{V}$ such that

$$\mathbf{A} = \text{conv}(\{ y_1, y_2, \ldots, y_r \}).$$

Definition 2.12 Let $\mathbf{V}$ be a real vector space, and let $\mathbf{A} \subset \mathbf{V}$ be a polytope. The **set of all vertices** of $\mathbf{A}$, denoted Vert($\mathbf{A}$), is the set such that $\mathbf{A} = \text{conv}(\text{Vert}(\mathbf{A}))$ and $\mathbf{A} \neq \text{conv}(\mathbf{B})$ if $\mathbf{B}$ is a proper subset of Vert($\mathbf{A}$). A vector y is called a vertex of $\mathbf{A}$ if $y \in \text{Vert}(\mathbf{A})$.

Definition 2.13 Let $\mathbf{V}$ be a real vector space, let $\mathbf{A} \subset \mathbf{V}$ be a polytope, and let x, y be two distinct vertices of $\mathbf{A}$. The set conv($\{x, y\}$) is called an **edge** of $\mathbf{A}$ if

$$\text{conv}(\{x, y\}) \cap \text{conv}(\text{Vert}(\mathbf{P}) \backslash \{x, y\}) = \emptyset.$$

The union of all edges of $\mathbf{A}$ and Vert($\mathbf{A}$) will be denoted as Edge($\mathbf{A}$).

Several examples of polytopes have already appeared in this dissertation. The sets Δ in Example 1.1 and 1.2 are not only interval families of real 3-tuples they are also polytopes in $\mathbf{R}^3$. This is not surprizing because interval families of real m-tuples are a subclass of polytopes in $\mathbf{R}^m$. The definitions of vertices and edges for interval families of real m-tuples and for polytopes in $\mathbf{R}^m$ are consistent because $\text{Vert}\{\delta^L, \delta^H\} = \text{Vert}(\text{Box}\{\delta^L, \delta^H\})$ and $\text{Edge}\{\delta^L, \delta^H\} = \text{Edge}(\text{Box}\{\delta^L, \delta^H\})$. In addition to polytopes of real m-tuples, polytopes of polynomials and polynomial-pairs have also appeared.

Example 2.2 Recall that the sets $\mathcal{D}$ and $\mathcal{T}$ in Example 1.2 are given by

$$\mathcal{D} = \{\ LCs^2+RCs+1\ :\ (R,L,C) \in \text{Box}\{(1,3,5),\ (2,4,6)\}\ \}$$

$$\mathcal{T} = \{\ (\ 1,\ LCs^2+RCs+1)\ :\ (R,L,C) \in \text{Box}\{(1,3,5),\ (2,4,6)\}\ \}.$$

Both of these sets are polytopes. Their vertices are given by

$$\text{Vert}(\mathcal{D}) = \{\ 15s^2+5s+1,\ \ 20s^2+5s+1,\ \ 24s^2+6s+1,$$

$$15s^2+10s+1,\ 18s^2+12s+1,\ 24s^2+12s+1\ \}$$

$$\text{Vert}(\mathcal{T}) = \{\ (\ 1,\ 15s^2+5s+1),\ (\ 1,\ 20s^2+5s+1),\ (\ 1,\ 24s^2+6s+1),$$

$$(\ 1,\ 15s^2+10s+1),\ (\ 1,\ 18s^2+12s+1),\ (\ 1,\ 24s^2+12s+1)\ \}.$$

The six edges of $\mathcal{D}$ are given by the following sets

$$\text{conv}(\ \{\ 15s^2+5s+1,\ 20s^2+5s+1\}\ )$$

$$\text{conv}(\ \{\ 20s^2+5s+1,24s^2+6s+1\}\ )$$

$$\text{conv}(\ \{\ 24s^2+6s+1,\ 24s^2+12s+1\}\ )$$

$$\text{conv}(\ \{\ 24s^2+12s+1,\ 18s^2+12s+1\}\ )$$

$$\text{conv}(\ \{\ 18s^2+12s+1,\ 15s^2+10s+1\}\ )$$

$$\text{conv}(\ \{\ 15s^2+10s+1,\ 15s^2+5s+1\}\ ).$$

The set Edge($\mathcal{D}$) equals the union of these six edges. The six edges of $\mathcal{T}$ are

$$\text{conv}(\ \{\ (\ 1,\ 15s^2+5s+1),\ (\ 1,\ 20s^2+5s+1)\ \}\ )$$

$$\text{conv}(\ \{\ (\ 1,\ 20s^2+5s+1),\ (\ 1,\ 24s^2+6s+1)\ \}\ )$$

$$\text{conv}(\ \{\ (\ 1,\ 24s^2+6s+1),\ (\ 1,\ 24s^2+12s+1)\ \}\ )$$

$$\text{conv}(\ \{\ (\ 1,\ 24s^2+12s+1),\ (\ 1,\ 18s^2+12s+1)\ \}\ )$$

$$\text{conv}(\ \{\ (\ 1,\ 18s^2+12s+1),\ (\ 1,\ 15s^2+10s+1)\ \}\ )$$

$$\text{conv}(\ \{\ (\ 1,\ 15s^2+10s+1),\ (\ 1,\ 15s^2+5s+1)\ \}\ ).$$

The set Edge($\mathcal{T}$) equals the union of these six polynomial-pair edges.

The sets $\mathcal{D}$ and $\mathcal{T}$ from Example 1.1 are also polytopes. In fact, they are very special types of polytopes. These special polytopes will be defined in the next section.

2.3.2 Interval Families of Polynomials and Polynomial-Pairs

In addition to polytopes, two more classes of uncertainties will be defined in terms of $\mathbb{T}$, $\mathbb{D}$, and $\mathbb{N}$. These classes are known as interval families of polynomials and interval families of polynomial-pairs. They are similar to interval families of real m-tuples which have already been introduced for describing parameter sets. These two new classes are define as follows.

Definition 2.14 The set of transfer functions $\mathbf{T}$ is called an **interval family of polynomial-pairs** if it can be represented in the form

$$\begin{aligned}\mathbf{T} = \{ \; (\; & (a_r+jb_r)s^r + (a_{r-1}+jb_{r-1})s^{r-1} + \ldots + (a_1+jb_1)s + (a_0+jb_0), \\ & (c_q+jd_q)s^q + (c_{q-1}+jd_{q-1})s^{q-1} + \ldots + (c_1+jd_1)s + (c_0+jd_0) \;) \; : \\ & a_0^L \le a_0 \le a_0^H, \quad a_1^L \le a_1 \le a_1^H, \quad \ldots, \quad a_r^L \le a_r \le a_r^H, \\ & b_0^L \le b_0 \le b_0^H, \quad b_1^L \le b_1 \le b_1^H, \quad \ldots, \quad b_r^L \le b_r \le b_r^H, \\ & c_0^L \le c_0 \le c_0^H, \quad c_1^L \le c_1 \le c_1^H, \quad \ldots, \quad c_q^L \le c_q \le c_q^H, \\ & d_0^L \le d_0 \le d_0^H, \quad d_1^L \le d_1 \le d_1^H, \quad \ldots, \quad d_q^L \le d_q \le d_q^H \; \}.\end{aligned}$$

Definition 2.15 The set of polynomials $\mathbf{D}$ is called an **interval family of polynomials** if it can be represented in the form

$$\begin{aligned}\mathbf{D} = \{ \; & (c_q+jd_q)s^q + (c_{q-1}+jd_{q-1})s^{q-1} + \ldots + (c_1+jd_1)s + (c_0+jd_0) \;) \; : \\ & c_0^L \le c_0 \le c_0^H, \quad c_1^L \le c_1 \le c_1^H, \quad \ldots, \quad c_q^L \le c_q \le c_q^H, \\ & d_0^L \le d_0 \le d_0^H, \quad d_1^L \le d_1 \le d_1^H, \quad \ldots, \quad d_q^L \le d_q \le d_q^H \; \}.\end{aligned}$$

A similar definition of interval families of polynomials hold for $\mathbb{N}$.

Interval families are special types of polytopes, so their vertices and edges are given by Definitions 2.12 and 2.13, respectively.

Example 2.3 Recall that $\mathbb{D}$ and $\mathbb{T}$ from Example 1.1 have the form

$$\mathbb{D} = \{\ Ms^2+Fs+K\ :\ 1 \le K \le 2,\ 1.5 \le M \le 1.6,\ 3 \le F \le 5\ \},$$

$$\mathbb{T} = \{\ (1, Ms^2+Fs+K)\ :\ 1 \le K \le 2,\ 1.5 \le M \le 1.6,\ 3 \le F \le 5\ \}.$$

Clearly, $\mathbb{D}$ is an interval family of polynomials, and $\mathbb{T}$ is an interval family of polynomial-pairs. The vertices of these two sets are given by

$$\text{Vert}(\mathbf{D}) = \{\ 1.5s^2+3+1,\ 1.5s^2+3+2,\ 1.5s^2+5+1,\ 1.5s^2+5+2,$$
$$1.6s^2+3+1,\ 1.6s^2+3+2,\ 1.6s^2+5+1,\ 1.6s^2+5+2\ \}$$

$$\text{Vert}(\mathbf{T}) = \{\ (1, 1.5s^2+3+1),\ (1, 1.5s^2+3+2),\ (1, 1.5s^2+5+1),\ (1, 1.5s^2+5+2),$$
$$(1, 1.6s^2+3+1),\ (1, 1.6s^2+3+2),\ (1, 1.6s^2+5+1),\ (1, 1.6s^2+5+2)\ \}.$$

The twelve edges of $\mathbf{D}$ are

$$\{\ Ms^2+3+1\ :\ 1.5 \le M \le 1.6\ \} \qquad \{\ Ms^2+3+2\ :\ 1.5 \le M \le 1.6\ \}$$
$$\{\ Ms^2+5+1\ :\ 1.5 \le M \le 1.6\ \} \qquad \{\ Ms^2+5+2\ :\ 1.5 \le M \le 1.6\ \}$$
$$\{\ 1.5s^2+F+1\ :\ 3 \le F \le 5\ \} \qquad \{\ 1.5s^2+F+2\ :\ 3 \le F \le 5\ \}$$
$$\{\ 1.6s^2+F+1\ :\ 3 \le F \le 5\ \} \qquad \{\ 1.6s^2+F+2\ :\ 3 \le F \le 5\ \}$$
$$\{\ 1.5s^2+3+K\ :\ 1 \le K \le 2\ \} \qquad \{\ 1.5s^2+5+K\ :\ 1 \le K \le 2\ \}$$
$$\{\ 1.6s^2+3+K\ :\ 1 \le K \le 2\ \} \qquad \{\ 1.6s^2+5+K\ :\ 1 \le K \le 2\ \}.$$

The set $\text{Edge}(\mathbf{D})$ equals the union of these twelve edges. The twelve edges of $\mathbf{T}$ are

$$\{\ (1, Ms^2+3+1)\ :\ 1.5 \le M \le 1.6\ \} \qquad \{\ (1, Ms^2+3+2)\ :\ 1.5 \le M \le 1.6\ \}$$
$$\{\ (1, Ms^2+5+1)\ :\ 1.5 \le M \le 1.6\ \} \qquad \{\ (1, Ms^2+5+2)\ :\ 1.5 \le M \le 1.6\ \}$$
$$\{\ (1, 1.5s^2+F+1)\ :\ 3 \le F \le 5\ \} \qquad \{\ (1, 1.5s^2+F+2)\ :\ 3 \le F \le 5\ \}$$
$$\{\ (1, 1.6s^2+F+1)\ :\ 3 \le F \le 5\ \} \qquad \{\ (1, 1.6s^2+F+2)\ :\ 3 \le F \le 5\ \}$$
$$\{\ (1, 1.5s^2+3+K)\ :\ 1 \le K \le 2\ \} \qquad \{\ (1, 1.5s^2+5+K)\ :\ 1 \le K \le 2\ \}$$

$\{ (1, 1.6s^2+3+K) : 1 \le K \le 2 \}$ $\{ (1, 1.6s^2+5+K) : 1 \le K \le 2 \}$.

The set Edge($\mathcal{T}$) equals the union of these twelve polynomial-pair edges.

2.4 Relation Between Polytopes and Image Based Classifications

This section will review some useful relationships between polytopes and sets generated by affine and multiaffine uncertainties. These relationships are based on the following facts.

Fact 2.1 Let $\mathbf{V}$ be a real vector spaces. If $F: \mathbb{R}^m \to \mathbf{V}$ is an affine mapping and if $\text{Box}\{\delta^L,\delta^H\} \subset \mathbb{R}^m$ then the set $F(\text{Box}\{\delta^L,\delta^H\})$ is a polytope in $\mathbf{V}$, and the following two set relations hold

$$\text{Vert}\big(F(\text{Box}\{\delta^L,\delta^H\})\big) \subset F(\text{Vert}\{\delta^L,\delta^H\})$$
$$\text{Edge}\big(F(\text{Box}\{\delta^L,\delta^H\})\big) \subset F(\text{Edge}\{\delta^L,\delta^H\}).$$

Fact 2.1 shows that sets of polynomials and polynomial-pairs generated by affine uncertainties are a subclass of polytopes of polynomials and polynomial-pairs, respectively. Description sets generated by multiaffine uncertainties are not a subclass of polytopes, but they can be related to polytopes using a result known as the Mapping Theorem [Zadeh,Desoer, 1963, p. 476]. A generalized version of the Mapping Theorem is stated in the following fact.

Fact 2.2 Let $\mathbf{V}$ be a real vector space. If $F: \mathbb{R}^m \to \mathbf{V}$ is a multiaffine mapping and if $\text{Box}\{\delta^L,\delta^H\} \subset \mathbb{R}^m$ then

$$\text{conv}\big(F(\text{Box}\{\delta^L,\delta^H\})\big) = \text{conv}\big(F(\text{Vert}\{\delta^L,\delta^H\})\big).$$

Proof Using $\delta^L = (\delta_1^L, \delta_2^L, \ldots, \delta_m^L)$ and $\delta^H = (\delta_1^H, \delta_2^H, \ldots, \delta_m^H)$ define the sets

$$A_0 = \text{Box}\{(\delta_1^L, \ldots, \delta_m^L), (\delta_1^H, \ldots, \delta_m^H)\}$$

$$A_1 = \text{Vert}\{(\delta_1^L), (\delta_1^H)\} \times \text{Box}\{(\delta_2^L, \ldots, \delta_m^L), (\delta_2^H, \ldots, \delta_m^H)\}$$

$$\vdots$$

$$A_i = \text{Vert}\{(\delta_1^L, \ldots, \delta_i^L), (\delta_1^H, \ldots, \delta_i^H)\} \times \text{Box}\{(\delta_{i+1}^L, \ldots, \delta_m^L), (\delta_{i+1}^H, \ldots, \delta_m^H)\}$$

$$\vdots$$

$$A_m = \text{Vert}\{(\delta_1^L, \ldots, \delta_m^L), (\delta_1^H, \ldots, \delta_m^H)\}.$$

For $i = 0, 1, \ldots, m-1$, it should be clear that

$$A_{i+1} \subset A_i$$

and hence

$$\text{conv}(F[A_{i+1}]) \subset \text{conv}(F[A_i]). \tag{1}$$

Let $x = (x_1, x_2, \ldots, x_m)$ be an arbitrary element in A_i $(i < m)$. For the case when $\delta_{i+1}^L = \delta_{i+1}^H$, it should be apparent that

$$A_{i+1} = A_i$$

and that

$$\text{conv}(F[A_i]) \subset \text{conv}(F[A_{i+1}]). \tag{2}$$

For the case when $\delta_{i+1}^L \neq \delta_{i+1}^H$, it is true that

$$x = \lambda\,(x_1, \ldots, x_i, \delta_{i+1}^H, x_{i+2}, \ldots, x_m) + (1-\lambda)\,(x_1, \ldots, x_i, \delta_{i+1}^L, x_{i+2}, \ldots, x_m)$$

for

$$\lambda = (x_{i+1} - \delta_{i+1}^L) / (\delta_{i+1}^H - \delta_{i+1}^L).$$

Since $F: \mathbb{R}^m \rightarrow V$ is a multiaffine mapping, it follows that

$$F[x] = \lambda\, F[(x_1, \ldots, x_i, \delta_{i+1}^H, x_{i+2}, \ldots, x_m)]$$
$$+ (1-\lambda)\, F[(x_1, \ldots, x_i, \delta_{i+1}^L, x_{i+2}, \ldots, x_m)].$$

This implies that

$$F[x] \in \text{conv}(F[A_{i+1}]).$$

Since x was an arbitrary element in $\mathbb{A}_i$, it follows that

$$F[\mathbf{A}_i] \subset \text{conv}(F[\mathbf{A}_{i+1}]). \tag{3}$$

Definition 2.10 and from equation (3) imply that

$$\text{conv}(F[\mathbf{A}_i]) \subset \text{conv}(F[\mathbf{A}_{i+1}]). \tag{4}$$

Equations (1), (2), and (4) show that

$$\text{conv}(F[\mathbf{A}_i]) = \text{conv}(F[\mathbf{A}_{i+1}]). \tag{5}$$

Since equation (5) is valid for any $i < m$, it is follows that

$$\text{conv}(F[\mathbf{A}_0]) = \text{conv}(F[\mathbf{A}_1]) = \ldots = \text{conv}(F[\mathbf{A}_i]) = \ldots = \text{conv}(F[\mathbf{A}_m]).$$

The equality of the first and last term completes the proof. □

Fact 2.2 shows that a set of polynomials or polynomial-pairs generated by multiaffine uncertainties can easily be overbounded by a polytope. Furthermore, this *overbounding polytope is the smallest convex set that contains the description set.* In some cases, multiaffine uncertainties will actually produce a set of descriptions that is a polytope, so no overbounding would be needed. This was the case for both $\mathbb{D}$ and $\mathbf{T}$ in Example 2.2.

2.5 Conclusion

This chapter has defined four classifications for sets of polynomials and four classifications for sets of polynomial-pairs. These classes were chosen to achieve two opposing goals. The first goal was for each class to include the types of polynomial sets and polynomial-pair sets needed to describe physical systems that have a reasonably simple dependence on uncertain parameters. The second goal was for all the sets contained in each class to be as highly structured as possible. Sets generated by multiaffine uncertainties are the most general class, so they do the best job of

achieving the first goal. Polytopes do the second best job, sets generated by affine uncertainties do the third best, and finally, interval families do the worst job of achieving the first goal. In terms of achieving the second goal, the ordering of best to worst is reversed. Interval families are the most structured sets, and sets generated by multiaffine uncertainties are the least structured. This tradeoff of generality and structure is the reason all four classes will be investigated in later chapters. It will be easiest to simplify the analysis of sets in the most structured classes, but analysis results for the most general classes are in many ways more useful.

CHAPTER 3

STABILITY AND POLE LOCATION ANALYSES

3.1 Introduction

This chapter will discuss the existence or nonexistence of vertex and edge theorems for stability and pole location analyses of the classes of uncertain systems defined in Chapter 2. The two analyses are treated together because their goals are very closely related. Given a uncertain system represented by a set of characteristic polynomials $\mathbb{D}$, the goal of a pole location analysis is to find the set of all possible poles of the system

$$\text{Root}(\mathbb{D}) = \{ s \in \mathbb{C} : p(s) = 0, p(s) \in \mathbb{D} \}.$$

The goal of a stability analysis is to determine if Root($\mathbb{D}$) is contained in a given region G of the complex plane. For these two analyses, a variety of vertex theorems, edge theorems, and counterexamples to vertex and edge conjectures will be given.

3.2 Some Polynomial Notation

To make it easier to present the results in this area, some polynomial nomenclature will be reviewed. As previously noted, the set of all polynomials in the complex variable s with coefficients from the complex field $\mathbb{C}$ will be denoted $\mathbb{C}[s]$, and the set of all polynomials with coefficients restricted to the real field $\mathbb{R}$ will be denoted $\mathbb{R}[s]$. Each polynomial p(s) in $\mathbb{C}[s]$ ($\mathbb{R}[s]$) has the form

$$p(s) = c_n s^n + c_{n-1} s^{n-1} + \ldots + c_1 s + c_0$$

where n is a nonnegative integer and each coefficient c_i is an element of $\mathbb{C}$ ($\mathbb{R}$). If $c_n \neq 0$ and $c_i = 0$ for all $i > n$ then the degree of p(s), denoted deg(p(s)), equals n. If $p(s) = 0$ then $deg(p(s)) \equiv -\infty$. The degree of a set of polynomials $\mathbb{D}$ is the set of integers $deg(\mathbb{D}) = \{ deg(p(s)) : p(s) \in \mathbb{D} \}$. The set of all complex polynomials of degree n is defined as follows

$$\mathbb{C}_n[s] = \{ p(s) \in \mathbb{C}[s] : deg(p(s)) = n \}.$$

If $deg(p(s)) = n$ and $c_n = 1$ then p(s) is called a monic polynomial. The set of all monic complex polynomials of degree n is denoted

$$\mathbb{C}_n^{monic}[s] = \{ p(s) \in \mathbb{C}[s] : deg(p(s)) = n, c_n = 1 \}.$$

The set of nth order real polynomials $\mathbb{R}_n[s]$ and the set of nth order real monic polynomials $\mathbb{R}_n^{monic}[s]$ can be defined similarly. Finally, the set of values taken by a collection of polynomials $\mathbb{D}$ when evaluated a $z \in \mathbb{C}$ is given by

$$Val(\mathbb{D},z) = \{ p(z) \in \mathbb{C} : p(s) \in \mathbb{D} \}$$

The use of this definition is motivated by the value set ideas of [Barmish, 1988].

3.3 An Edge Theorem for Root Locations of Polytopes of Polynomials

This section will discuss edge theorems which reduce the effort needed to compute Root($\mathbb{D}$) in the case when $\mathbb{D}$ is a polytope of polynomials. The first edge theorem concerning Root($\mathbb{D}$) to appear in the literature is due to [Bartlett, Hollot, Huang, 1988]. They impose the restriction that the polytope $\mathbb{P}$ must be a subset of $\mathbb{R}_n^{monic}[s]$. Under this assumption, it was shown that the boundary of Root($\mathbb{D}$) defined on $\mathbb{C}$ (denoted ∂Root($\mathbb{D}$)) is a subset of Root(Edge($\mathbb{D}$)). Later works on Root($\mathbb{D}$) have addressed the problem of removing the restriction on $\mathbb{D}$ imposed by [Bartlett, Hollot, Huang, 1988]. [Hollot, Looze, Bartlett, 1990] have removed the monic n-th order restriction on $\mathbb{D}$. They proved that the edge theorem of [Bartlett,

Hollot, Huang, 1988] remains valid for any polytope $\mathbb{D} \subset \mathbb{R}[s]$. [Bartlett, 1990a] discussed how the extension from real to complex polynomials can be easily achieved based on the work in [Bartlett, Hollot, Huang, 1988; Hollot, Looze, Bartlett, 1987; Fu and Barmish, 1989]. This extension is stated in the following theorem. A new proof motivated by the value set ideas of [Barmish, 1988; Sideris, Barmish, 1989] is given[1].

Theorem 3.1 If $\mathbb{D}$ is a polytope in $\mathbb{C}[s]$ then the boundary of Root($\mathbb{D}$) is a subset of Root(Edge($\mathbb{D}$)).

Proof Because $\text{Val}(\cdot,x) : \mathbb{C}[s] \rightarrow \mathbb{C}$ is an affine mapping for any fixed value of $x \in \mathbb{C}$ and because $\mathbb{D}$ is a polytope of polynomials, it follows that Val($\mathbb{D}$,x) is a polytope in $\mathbb{C}$ and that

$$\partial\text{Val}(\mathbb{D},x) = \text{Edge}(\text{Val}(\mathbb{D},x)) \subset \text{Val}(\text{Edge}(\mathbb{D}),x) \qquad (1)$$

for all $x \in \mathbb{C}$. Now, let z be an arbitrary complex number such that $z \notin \text{Root}(\text{Edge}(\mathbb{D}))$. This implies that $0 \notin \text{Val}(\text{Edge}(\mathbb{D}),z)$. Let the edges of $\mathbb{D}$ be denoted as $E_1, E_2, \ldots, E_m$, and for each i, define r_i to be the minimum distance from the origin to the line segment $\text{Val}(E_i,z)$. Use these quantities to define the positive real number $r = \min\{r_1, r_2, \ldots, r_m\}$. Next, define the set

$$\text{Ball}[x,a] = \{ y \in \mathbb{C} : \text{abs}(x-y) \leq a \}.$$

Let the vertices of $\mathbb{D}$ be denoted as $v_1(s), v_2(s), \ldots, v_n(s)$, and for each i, define ε_i to be the largest positive real number less than or equal to one such that

$$\{ v_i(x) \in \mathbb{C} : x \in \text{Ball}[z,\varepsilon_i] \} \subset \text{Ball}[v_i(z),r/2].$$

Note that ε_i exists because $v_i(s)$ is a continuous function of s. Now define

[1]This new line of proof suggests that the theorem can be extended from polytopes of polynomials to polytopes of any functions that are continuous at all points in the complex plane. This extension is not considered because it is outside the scope of this dissertation.

$\varepsilon = \min\{\varepsilon_1,\varepsilon_2, \ldots , \varepsilon_n\}$. Let E_k be an arbitrary edge of $\supset$, and let v_i, v_j be the vertices of this edge. From the definition of ε, it is clear that

$$\text{abs}(v_i(z)-v_i(c)) \le r/2 \qquad\qquad \text{abs}(v_j(z)-v_j(c)) \le r/2$$

for all $c \in \text{Ball}[z,\varepsilon]$. By the triangle inequality, this implies that, for all $c \in \text{Ball}[z,\varepsilon]$ and for all $\lambda \in [0,1]$,

$$\text{abs}\big(\lambda v_i(z)+(1-\lambda)v_j(z)-\lambda v_i(c)-(1-\lambda)v_j(c) \big) \le \lambda r/2+(1-\lambda)r/2 = r/2.$$

For all $c \in \text{Ball}[z,\varepsilon]$, this implies that every point on the line segment $\text{Val}(E_k,c)$ is a distance of less than or equal to $r/2$ from the line segment $\text{Val}(E_k,z)$ which in turn implies that every point on $\text{Val}(E_k,c)$ is a distance of at least $r_k - r/2 \ge r/2 > 0$ from the origin. This implies that $c \notin \text{Root}(\text{Edge}(\mathbb{D}))$ for all $c \in \text{Ball}[z,\varepsilon]$. From this fact, equation (1) implies that either $\text{Ball}[z,\varepsilon] \subset \text{Root}(\mathbb{D})$ or $\text{Ball}[z,\varepsilon] \cap \text{Root}(\mathbb{D}) = \emptyset$. In either case, this shows that $z \notin \partial\text{Root}(\mathbb{D})$ which completes the proof. □

Using Fact 3.1, Theorem 3.1 immediately implies a slightly different edge theorem for sets of polynomials generated by affine uncertainties. This result is stated in the following corollary.

Corollary 3.1 If the set of polynomials $D(\Delta)$ is generated by affine uncertainties then the boundary of $\text{Root}(D(\Delta))$ is a subset of $\text{Root}(D(\text{Edge}(\Delta)))$.

Theorem 3.1 has come to be known as the root version of the Edge Theorem. It is one of the main contributions of the author, but it should be pointed out that this is not a sole contribution of the author. The original and most important work on this theorem was performed in collaboration with C. V. Hollot and Huang Lin. The extension to non-monic polynomials was carried out with C. V. Hollot and D. P. Looze. The extension to complex polynomials and the latest proof are

independent work, but these are very minor contributions in comparison to the earlier work. In addition, it should be pointed that Fu and Barmish extended a stability version of the Edge Theorem to complex polynomials of fixed degree before any of the other extensions were made.

The Edge Theorem indicates how to find the boundary of the root space, ∂Root($\mathbb{D}$), using only the **edge root locus** Root(Edge($\mathbb{D}$)) [or Root(D(Edge(Δ)))], but it does not indicate how to completely determine the **root space** Root($\mathbb{D}$). For complete knowledge, it is still necessary to determine the "pseudo-interior" Root($\mathbb{D}$) \ Root(Edge($\mathbb{D}$)). The work of [Zeheb, Walach, 1981] provides a finite step procedure for solving a similar root location problem for more general families of polynomials. It was pointed out by the author [Bartlett, 1990a] that Zeheb and Walach's result is not only applicable to the polytope "pseudo-interior" problem, but it is also quite easy to implement in this special case. The next theorem paraphrases Zeheb and Walach's result as applied to polytopes of polynomials. In combination, the results of [Zeheb, Walach, 1981] and the Edge Theorem provide the best available method of obtaining complete knowledge of Root(**D**).

Theorem 3.2 [see Zeheb, Walach, 1981 for proof] Let $V_1, V_2, \ldots, V_q$ be a finite collection of open sets such that $V_1 \cup V_2 \cup \ldots \cup V_q = \mathbb{C} \setminus$ Root(Edge($\mathbb{D}$)), each V_i is a connected set, and $V_i \cap V_k = \emptyset$ for $i \neq k$. If $z_i \in V_i$ and $z_i \in$ Root($\mathbb{D}$), then $V_i \subset$ Root(**D**). Conversely, if $z_i \in V_i$ and $z_i \notin$ Root($\mathbb{D}$) then $V_i \cap$ Root(**D**) = $\emptyset$.

Given Root(Edge($\mathbb{D}$)), the interior of Root($\mathbb{D}$) can be determined by testing if a finite number of points are contained in Root(**D**). For the complicated families of polynomials considered by [Zeheb, Walach, 1981], this can be quite a chore. For a polytope of polynomials, this task is very simple. Testing $z_i \in$ Root($\mathbb{D}$) is equivalent to

testing

$$0 \in \text{conv}\big(\text{Val}(\text{Vert}(\mathbb{D}),z_i) \big).$$

As noted by Barmish, conv(Val(Vert($\mathbb{D}$),z_i)) is simply a convex polygon in $\mathbb{C}$. Determining whether or not this polygon covers the origin is easy. It can be done graphically or by some algebraic method such as in [Barmish, 1988]. The complete determination of Root($\mathbb{D}$) using Theorem 3.1 and 3.2 is illustrated by the following example.

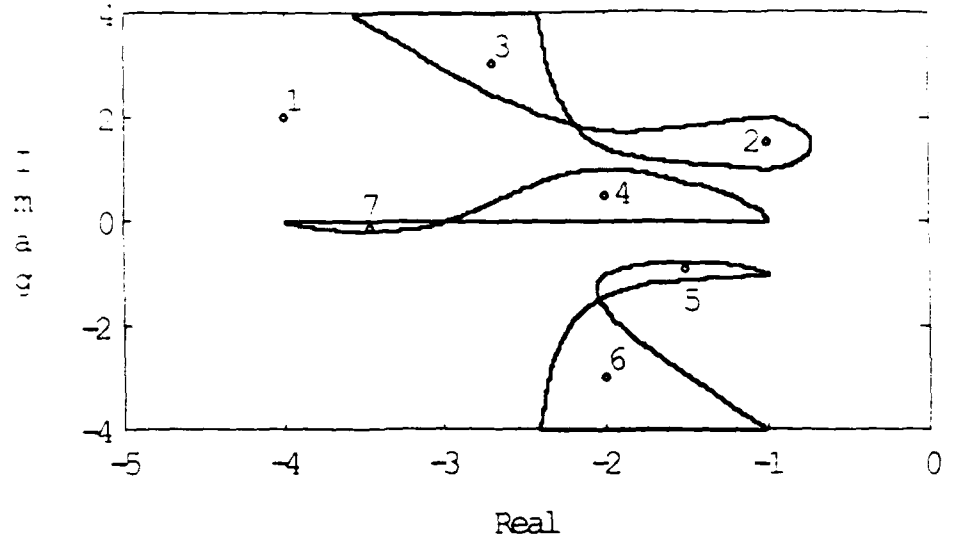
Example 3.1 Let Vert($\mathbb{D}$) be comprised of the following three vertex polynomials

$$p_1(s) = s + 1$$

$$p_2(s) = (1\text{-}j)s^3 + (5\text{-}j7)s^2 + (7\text{-}j19)s + (3\text{-}j2\ )$$

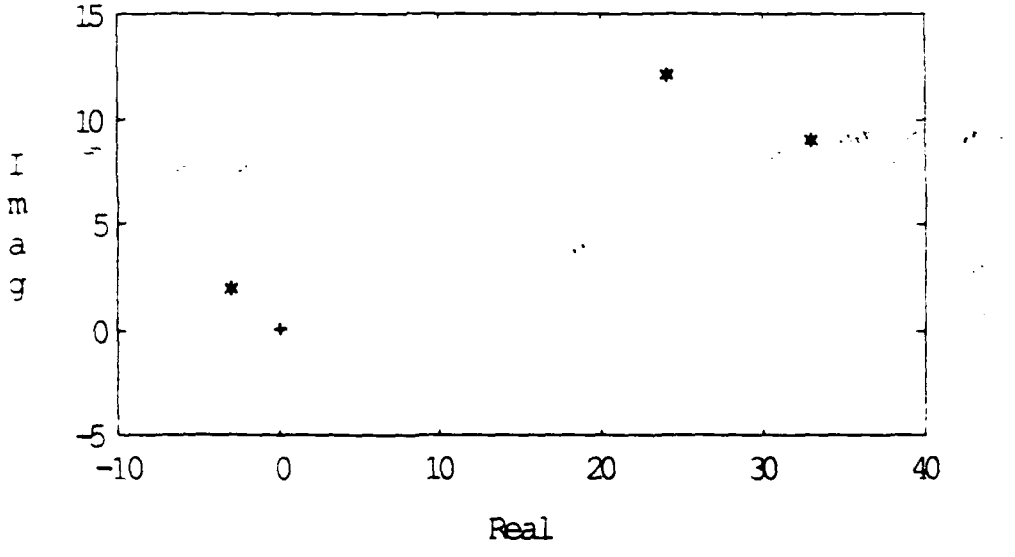
$$p_3(s) = s^3 + 6s^2 + 10s + 8.$$

For the polytope $\mathbb{D}$ = conv(Vert($\mathbb{D}$)), a bounded approximation of the edge root locus Root(Edge($\mathbb{D}$)) is shown in Figure 3.1. It appears that the edge root locus divides the complex plane into seven regions. A test point z_i for each region is marked by an "o" in Figure 3.1. For region 1, the test point is $z_1 = -4 + j\,2$. The plot of conv(Val(Vert($\mathbb{D}$),z_1)) = conv($\{p_1(z_1), p_2(z_1), p_3(z_1)\}$) is shown in Figure 3.2. The polygon does not contain the origin, so none of the points in region 1 are contained in Root($\mathbb{D}$). For region 2, the test point is $z_2 = -1 + j\,1.5$. The plot of conv(Val(Vert($\mathbb{D}$),z_2)) = conv($\{p_1(z_2), p_2(z_2), p_3(z_2)\}$) is shown in Figure 3.3. The polygon contains the origin, so every point in region 2 is contained in Root($\mathbb{D}$). In the same manner as for regions 1 and 2, regions 3-7 were all determined to be subsets of Root($\mathbb{D}$). The complete root locus of the polytope is displayed in Figure 3.4.



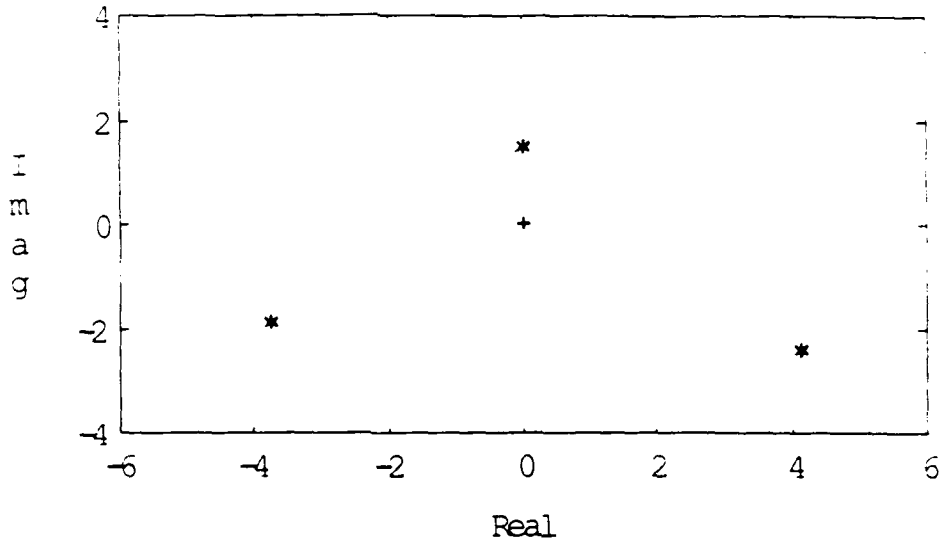
Edge root locus Root(Edge(𝔻)) and Test Points for Example 3.1.

Figure 3.1



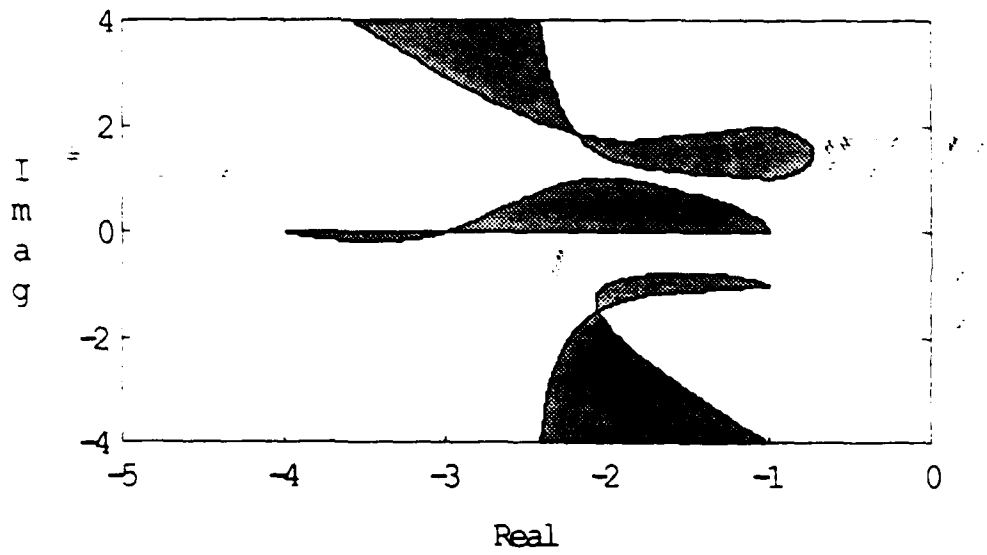
Test showing that region 1 of Figure 3.1 is <u>not</u> included Root(𝔻) for Example 3.1.

Figure 3.2



Test showing that region 2 of Figure 3.1 is included in Root(𝔻) for Example 3.1.

Figure 3.3



Root space of the polytope of polynomials for Example 3.1.

Figure 3.4

As a final comment on root location results for polytopes of polynomials, it is noted that Root(Vert($\mathbb{D}$)) does not in general provide sufficient information to determine Root($\mathbb{D}$). This should be obvious from looking at Root($\mathbb{D}$) for any nontrivial example such as the one above. This implies that no vertex theorem exists for determining Root($\mathbb{D}$) except possibly in special cases. Even if $\mathbb{D}$ is restricted to the class of polynomial sets generated by affine uncertainties or to the class of interval families of polynomials, a root location vertex theorem still does not exist . If the goal is to determine only Root($\mathbb{D}$) $\cap$ $\mathbb{R}$ and if $\mathbb{D} \subset \mathbb{R}[s]$ then a vertex theorem should be obtainable using the rules of root locus construction [Evans, 1948]. Theorems in this vein where discussed in [Hollot, Bartlett, 1987].

3.4 No Root Location Edge Theorem for Multiaffine Uncertainties

This section will give a counterexample which shows that the root space of a set of polynomials D(Δ) generated by multiaffine uncertainties cannot be determined from the root space of D(Edge(Δ)). This counterexample is not an original contribution of the author; it is included here simply for the sake of completeness. This counterexample follows almost immediately from an example by [Barmish, Fu, Saleh, 1988]. A similar example is also given in [Ackermann, Hu, Kaesbauer, 1990].

Example 3.2 Consider the following multiaffine mapping D: $\mathbb{R}^2 \rightarrow \mathbb{C}[s]$

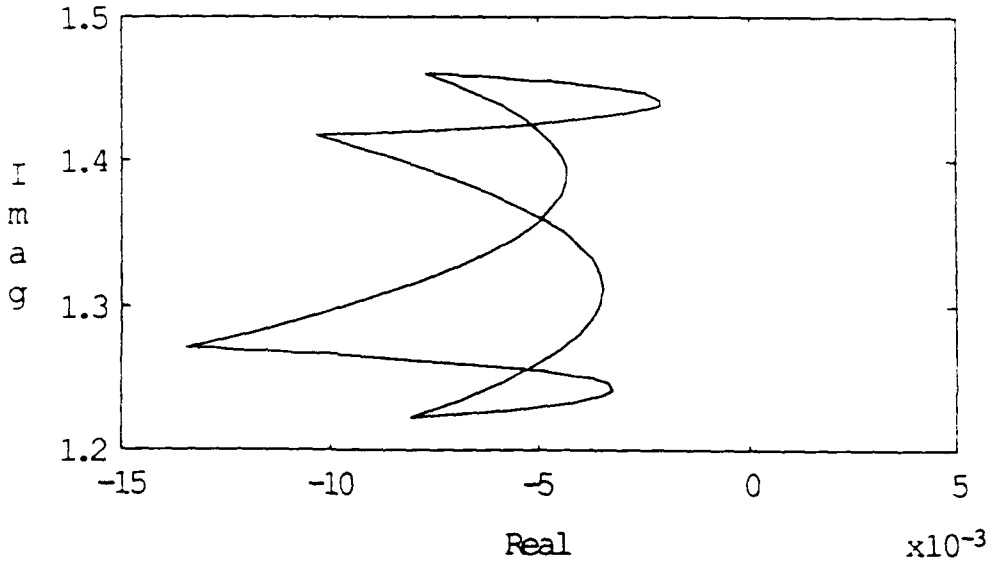
$$D((x,y)) = s^4 + (2.56 + x + y)s^3 + (2.871 + 2.06\,x + 1.561y + xy)s^2$$
$$+ (3.164 + 4.841x + 1.56y + 1.06xy)s + (1.853 + 3.773x + 1.985y + 4.032xy)$$

that was given by [Barmish, Fu, Saleh, 1988]. Let the set of possible parameters be

$$\Delta = \{ (x,y) \in \mathbb{R}^2 \ : \ 0.2 \le x \le 0.7,\ 0 \le y \le 3 \}.$$

For all $(x,y) \in \Delta$, the polynomial D((x,y)) has two real roots and one pair of complex

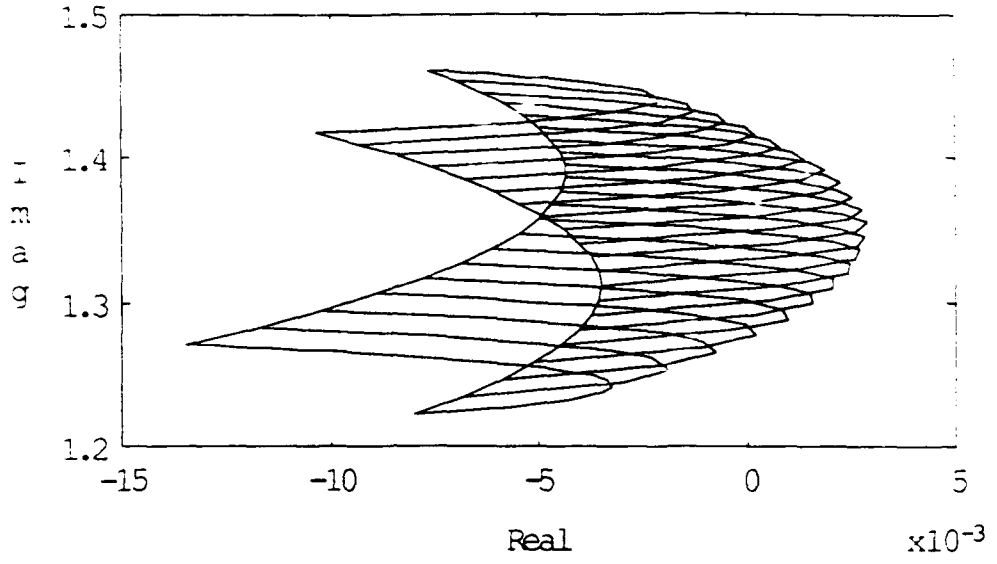
conjugate roots. Focusing only on the complex root with positive imaginary part, Figure 3.5.shows Root(D(Edge(Δ))), and Figure 3.6 shows a grid approximation of Root(D(Δ)). These two figures show that the boundary of the root space of D(Δ) is not given by the root space of D(Edge(Δ)).



Root(D(Edge(Δ))) for Example 3.2.

Figure 3.5

Since it is not generally true that ∂Root(D(Δ)) ⊂ Root(Edge(Δ)) when D(Δ) is generated by multiaffine uncertainties, it is hard to conceive of any simple way to determine the root space of D(Δ) from the root space of D(Edge(Δ)).



Grid approximation of Root(D(Δ)) for Example 3.2.

Figure 3.6

3.5 Edges Alone Don't Imply Stability of a Polytope of Polynomials

Determination of Root(𝔻) is one way of ascertaining if 𝔻 is G-stable, but in many cases, it is not the most efficient way. This is one of the reasons why there is a penchant in the controls community to restate the root version of the Edge Theorem above directly in terms of stability. To some extent, this conversion has already been successfully achieved, but it is only valid for certain classes of G regions. Using a direct restatement, the stability version of the Edge Theorem says that a polytope of polynomials 𝔻 is G-stable if and only if Edge(𝔻) is G-stable. [Bartlett, Hollot, and Huang, 1988] showed that the stability version is valid if $\mathbb{D} \subset \mathbb{R}_n^{\text{monic}}[s]$ and if G is simply connected on ℂ. [Fu, Barmish, 1989] proved that the stability version is valid if $\mathbb{D} \subset \mathbb{C}_n[s]$ and if the complement of G, denoted G^C, is unbounded and path

connected on the Riemann Sphere, $\mathbb{S}$. [Hollot, Looze, Bartlett, 1990] proved that the stability version[2] remains valid if $\mathbb{D} \subset \mathbb{R}[s]$ and if G is conjugate symmetric, simply connected on $\mathbb{S}$. Unfortunately, neither the restrictions on the set of polynomials nor on the stability regions can be removed without invalidating the theorem.

Some conjectures have appeared in the literature which remove these restrictions. Unfortunately, these conjectures are incorrect. In [Hollot, Looze, Bartlett, 1987], it was stated that the stability version[3] of the Edge Theorem is valid if $\mathbb{D} \subset \mathbb{R}[s]$ and if G is simply connected on $\mathbb{C}$. [Barmish, 1988a; Sideris, Barmish, 1989] showed by counterexample that this statement is not true. [Soh, Foo, 1989] have proposed a Generalized Edge Theorem which handles a larger class of functions than polynomials. When restricted to polynomials, the Generalized Edge Theorem states that for any polytope $\mathbb{D} \subset \mathbb{C}[s]$ and any simply connected region G every element in $\mathbb{D}$ has m roots in G if and only if every element in Edge($\mathbb{D}$) has m roots in G. For the case when $m = 0$, this theorem implies that $\mathbb{D}$ is G^C-stable if and only if Edge($\mathbb{D}$) is G^C-stable. In this section, it will be shown by counterexample that this statement is not true. The error in the Generalized Edge Theorem is not limited to the case when $m = 0$. Other counterexamples for $m \neq 0$ will also be given. The counterexamples in this section and in [Barmish, 1988] clearly show that the restrictions on the Edge Theorems in [Bartlett, Hollot, Huang, 1988; Fu, Barmish 1989; Hollot, Looze, Bartlett, 1990] cannot simply be removed. These counterexamples are contributions of the author [Bartlett, 1990a].

The first counterexample will address the removal of conditions on the polytope. As noted above, previous stability versions of the Edge Theorem all contain restrictions on the polytopes. [Fu, Barmish, 1989] require that $\mathbb{D} \subset \mathbb{C}_n[s]$ and

---

[2] In [Hollot et al 1990], Root(**A**) is the closure on **S** of { $z \in \mathbf{C} : p(z)=0, p(s) \in \mathbf{A} \subset \mathbf{R}[s]$ }.
[3] Unlike [Hollot et al 1990], [Hollot et al 1987] uses the standard definition of Root(**A**).

[Hollot, Looze, Bartlett, 1990] require that $\mathcal{D} \subset \Re[s]$. The next example will show that these restrictions can't be removed even for basic stability regions.

Example 3.3 Given the following polynomials

$$p_0(s) = 3\,s + 0.5$$

$$p_1(s) = (-3 + j3)\,s + (-1.5 + j2)$$

$$p_2(s) = (-3 - j3)\,s + (-1.5 - j2),$$

define $D: \mathbb{R}^2 \rightarrow \mathbb{C}[s]$ to be the mapping such that

$$D(\,(\delta_1,\delta_2)\,) = p_0(s) + \delta_1 p_1(s) + \delta_2 p_2(s),$$

and define

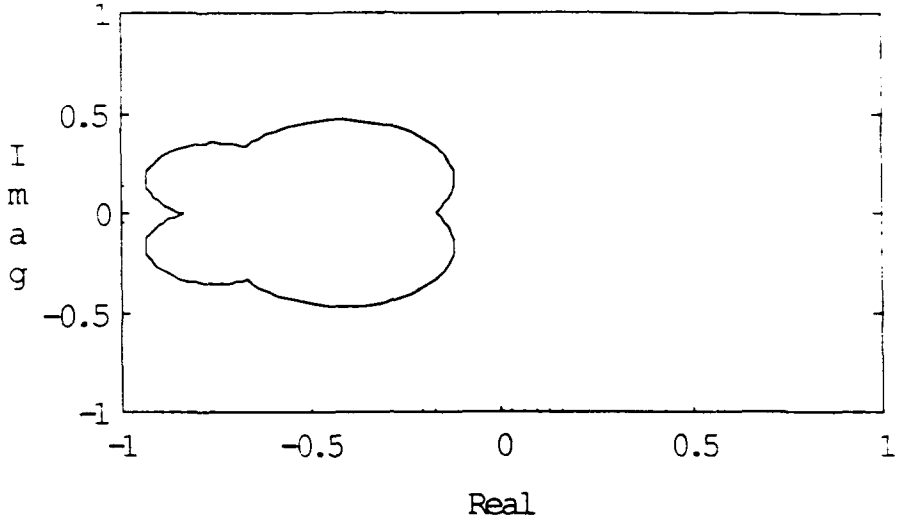
$$\Delta = \text{Box}\{\,(0,0),\,(1,1)\}.$$

The set Δ is an interval family and D is an affine mapping, so the set of polynomials

$$\mathcal{D} = D(\,\text{Box}\{(0,0),\,(1,1)\}\,)$$

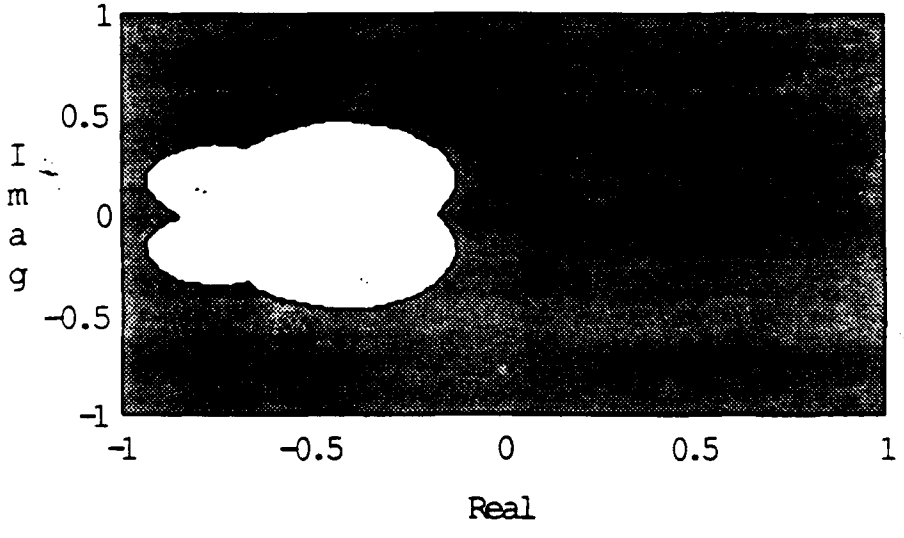
is generated by affine uncertainties. From Fact 2.1, it is known that $\mathcal{D}$ is also a polytope of polynomials. The edge root locus Root(Edge($\mathcal{D}$)) of this polytope is represented by the solid lines in Figure 3.7. The dotted lines in Figure 3.7 represent the jw-axis and the unit circle. The complete root space Root($\mathcal{D}$) is represented by the shaded region in Figure 3.8. From the two figures, it is clear that $\mathcal{D}$ is neither G_H-stable nor G_S-stable even though Edge($\mathcal{D}$) is both G_H-stable and G_S-stable.

The next counterexample will consider the conditions on the stability region. In addition to restrictions on the polytope of polynomials, previous stability versions of the Edge Theorem also contain restrictions on the stability region G. [Fu, Barmish, 1989] require that G^C be path connected and unbounded while [Hollot, Looze, Bartlett, 1990] require that G be complex symmetric, simply connected on S. It has already been shown by [Barmish, 1988] that the restrictions in [Hollot, Looze, Bartlett, 1990]



The edge root locus Root(Edge(𝔻)) for Example 3.3.

Figure 3.7



The root space Root(𝔻) for Example 3.3.

Figure 3.8

can't be removed. The next example will show that the stability region restrictions in [Fu, Barmish, 1989] also cannot be removed.

Example 3.4 Given the following polynomials

$$p_0(s) = s^2 + (-3.7900 - j4.7700)s + (-1.8800 + j9.1000)$$

$$p_1(s) = (-0.7875 + j0.5125)s + (2.5500 + j0.6250)$$

$$p_2(s) = (0.4158 + j0.8118)s + (0.7128 - j2.4948),$$

define $D: \mathbb{R}^2 \rightarrow \mathbb{C}[s]$ to be the mapping such that

$$D((\delta_1,\delta_2)) = p_0(s) + \delta_1 p_1(s) + \delta_2 p_2(s),$$

and define

$$\Delta = \text{Box}\{ (0,0), (1,1)\}.$$

The set Δ is an interval family and D is an affine mapping, so the set of polynomials

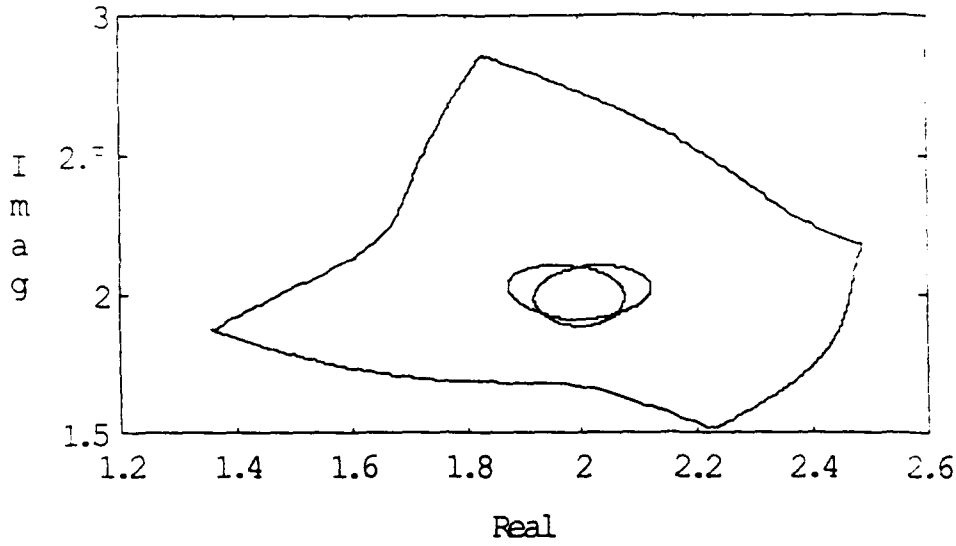
$$\mathbb{D} = D(\text{Box}\{(0,0), (1,1)\})$$

is said to be generated by affine uncertainties. From Fact 2.1, it is known that $\mathbb{D}$ is also a polytope of polynomials. The edge root locus Root(Edge($\mathbb{D}$)) of this polytope is shown in Figure 3.9. The complete root space Root($\mathbb{D}$) is represented by the shaded region in Figure 3.10. Let G be the stability region whose complement is

$$G^C = \{ s \in \mathbf{C} : | s - 2 - j 2 | \leq 0.04 \}.$$

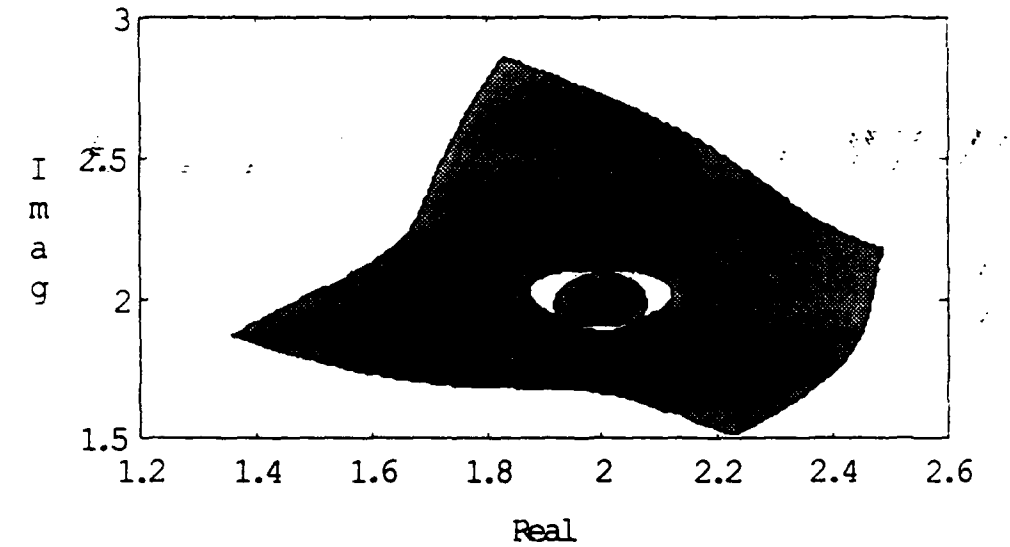
Note that the complement of G is simply connected, but it is not unbounded. The edge root locus Root(Edge($\mathbb{D}$)) and the region of instability G^C (lightly shaded) are shown in Figure 3.11. The edge root locus is completely outside of G^C, so Edge($\mathbb{D}$) is G-stable. From Figure 3.10 and 3.11, it is clear that Root($\mathbb{D}$) contains G^C, so $\mathbb{D}$ is not G-stable. This shows that the restriction on G given in [Fu, Barmish, 1989] can't be removed.

The Generalized Edge Theorem of [Soh, Foo, 1989] restricts the stability region but does not restrict the polytope. The Generalized Edge Theorem states that for



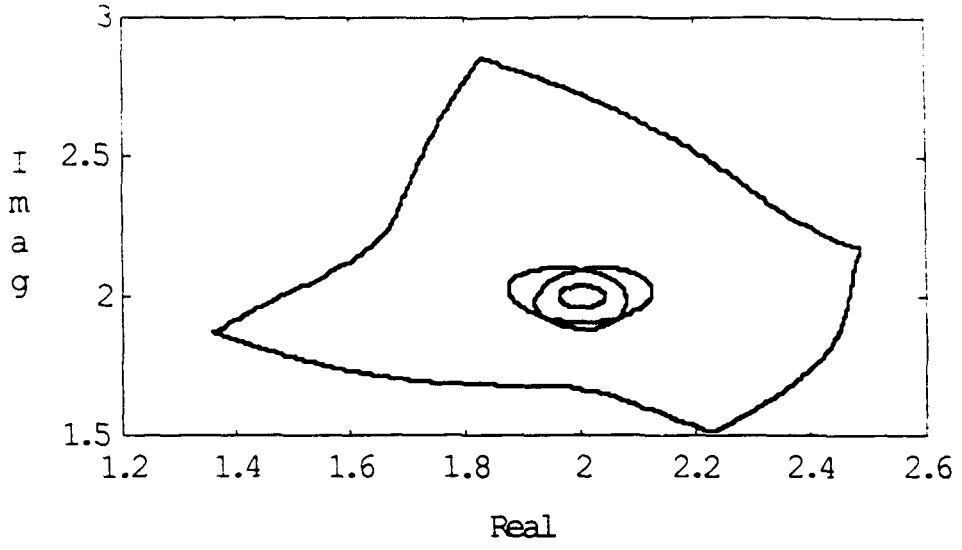
The edge root locus Root(Edge(D)) for Example 3.4.

Figure 3.9



The root space Root(D) for Example 3.4.

Figure 3.10



The edge root locus and the region of instability for Example 3.4.

Figure 3.11

any polytope $\mathbb{D} \subset \mathbb{C}[s]$ and any simply connected region G every element in $\mathbb{D}$ has m roots in G if and only if every element in Edge($\mathbb{D}$) has m roots in G. In Example 3.3, every edge polynomial has one root (its only root) in the open left half plane, but some polynomials in $\mathbb{D}$ have their only root in the right half plane. This shows that for m = max deg($\mathbb{D}$), the Generalized Edge Theorem is incorrect. For the case when m = 0, the Generalized Edge Theorem implies that if G^C is simply connected then $\mathbb{D}$ is G-stable if and only if Edge($\mathbb{D}$) is G-stable. Both Example 3.3 and 3.4 provide counterexamples to this statement. The error in the theorem is not limited to the cases when m = 0 or m = max deg($\mathbb{P}$). The following example disproves the intermediary case $0 < m < \max \deg(\mathbb{P})$.

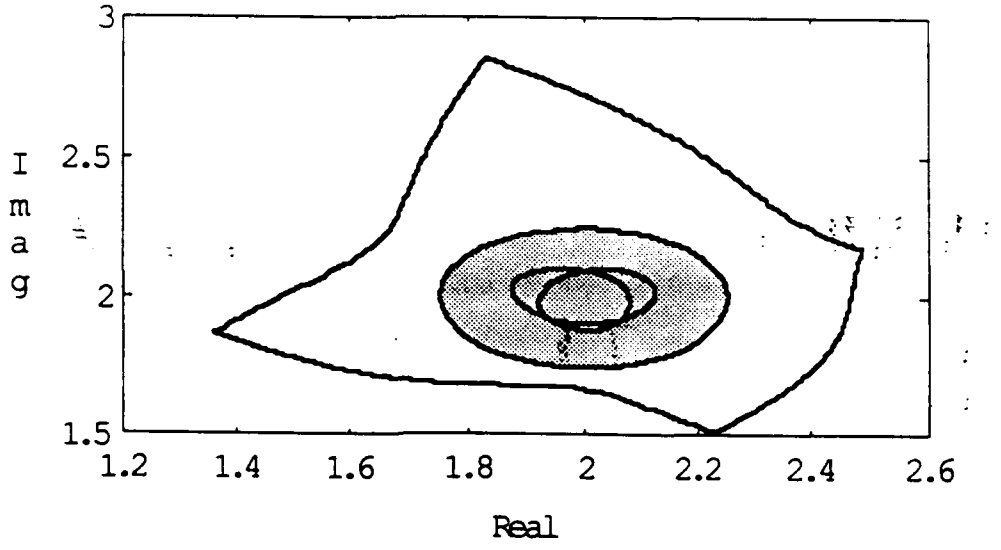
Example 3.5 Let $\mathbb{D}$ be the polytope given in Example 3.4. Let the stability region be

$$G = \{ s \in \mathbb{C} : |s - 2 - j2| \leq 0.25 \}.$$

Note that G is simply connected. The edge root locus Root(Edge(⊃)) and G (lightly shaded) are shown in Figure 3.12 (compare with Figure 3.9 to distinguish Root(Edge(𝔻)) from ∂G). The edge root locus has two connected regions one in G and one outside of G. Since the roots vary continuously [Marden, 1949; Zedek, 1965], each edge polynomial will have at least one root in each of the two connected regions. Since every polynomial in Edge(𝔻) has two roots, one must be in G and one must be outside of G. The Generalized Edge Theorem implies that every polynomial in ⊃ will have exactly one root in G. However, the polynomial

$$D((0.5, 0.5)) = p_0(s) + 0.5p_1(s) + 0.5p_2(s),$$

has 2 roots in G. This contradicts the theorem.



The edge root locus and the region G for Example 3.5.

Figure 3.12

These examples clearly show that the root version of the Edge Theorem, Theorem 3.1, can't be directly restated in terms of stability without restrictions on the polytope of polynomials and on the stability region G. Examples 3.3, 3.4, and 3.5 also disprove another conjecture. The sets $\mathbb{D} = D(\Delta)$ in these examples are generated by special affine uncertainties which have the property that $Edge(D(\Delta)) = D(Edge(\Delta))$. Because of this property, these examples show that, like Theorem 3.1, Corollary 3.1 can't be directly restated in terms of stability without the use of restrictions. The next section will seek to overcome the need to use these restrictions.

3.6 With a Precondition, Edges Imply Stability of a Polytope

It will be necessary to amend some additional conditions to the stability version of the Edge Theorem in order to remove the restrictions on the polytope of polynomials $\mathbb{D}$ and on the stability region G. In this section, a simple precondition for the Edge Theorem [Bartlett, 1990a] will be presented. If the precondition is not satisfied, then $\mathbb{D}$ has roots outside of G. If the precondition is satisfied, then G-stability of $Edge(\mathbb{D})$ implies G-stability of $\mathbb{D}$. So by using this result, a modified stability version of the Edge Theorem would state that $\mathbb{D}$ is G-stable if and only if $\mathbb{D}$ satisfies the precondition and $Edge(\mathbb{D})$ is G-stable. This version is valid for any region G and any polytope in $\mathbb{C}[s]$. The precondition is defined below.

Definition 3.1 For an arbitrary stability region G, let $U_1, U_2, U_3, \ldots$ be a collection of sets such that each U_i is a connected set[4] and $U_1 \cup U_2 \cup U_3 \cup \ldots = G^C$. Let $z_1, z_2, z_3, \ldots \in \mathbb{C}$ be an arbitrary collection of points such that $z_i \in U_i$ for each i.

[4] $U_i \subset \mathbb{C}$ is a connected set if $A \cap U_i = \emptyset$ or $B \cap U_i = \emptyset$ for all pairs of disjoint open sets $A, B \subset \mathbb{C}$.

The set of polynomials $\mathbb{D}$ is said to satisfy the **precondition** for G if $\{ z_1, z_2, z_3, \ldots \} \cap \text{Root}(\mathbb{D}) = \emptyset$.

The sets $U_1, U_2, U_3, \ldots$ should be chosen so that as few sets as possible are required. This will reduce the amount of work required to check the precondition. For cases when there exist a finite number of connected sets $U_1, U_2, \ldots, U_r$ such that $U_1 \cup U_2 \cup \ldots \cup U_r = G^C$, the condition $\{z_1, z_2, \ldots, z_r\} \cap \text{Root}(\mathbb{D}) = \emptyset$ requires only r simple tests. These test are the same as those shown in Figure 3.2 and 3.3 from Example 3.1. If G^C can only be decomposed into a infinite number of connected sets, then the precondition will require an infinite number of simple tests. In this infinite case, direct computation of Root($\mathbb{D}$) via Theorem 3.1 and 3.2 seems more feasible than using the precondition and the stability version of the Edge Theorem presented below.

The following theorem is a modified stability version of the Edge Theorem [Bartlett, 1990a]. It is an original contribution of the author.

Theorem 3.3 Let $\mathbb{D}$ be a polytope of polynomials, and let G be an arbitrary stability region in C. $\mathbb{D}$ is G-stable if and only if $\mathbb{D}$ satisfies the precondition for G and Edge($\mathbb{D}$) is G-stable.

Proof (Necessity) By Definition 3.1, if $\mathbb{D}$ does not satisfy the precondition for G, then there exists a $z \in G^C$ which is also an element of Root($\mathbb{D}$). This would imply that $\mathbb{D}$ was not G-stable, so the precondition is a necessary condition for G-stability of $\mathbb{D}$. Since Edge($\mathbb{D}$) $\subset \mathbb{D}$, it is clear that G-stability of Edge($\mathbb{D}$) is a necessary condition for G-stability of $\mathbb{D}$.

(Sufficiency) Assume that $\mathbb{D}$ satisfies the precondition for G and assume that Edge($\mathbb{D}$) is G-stable. This implies that there exists a collection of sets $U_1, U_2, U_3, \ldots$ and a collection of points $z_1, z_2, z_3, \ldots$ such that each U_i is a connected set, $U_1 \cup U_2 \cup U_3 \cup \ldots = G^C$, $z_i \in U_i$ for each i, and $\{z_1, z_2, z_3, \ldots\} \cap \text{Root}(\mathbb{D}) = \emptyset$. For an arbitrary i, it will be shown that $\mathbb{D}$ is U_i^C-stable. Since $U_i \subset G^C$, Edge($\mathbb{D}$) is U_i^C-stable. As in Theorem 3.2, let $V_1, V_2, \ldots, V_q$ be a finite collection of open sets such that $V_1 \cup V_2 \cup \ldots \cup V_q = \mathbb{C} \setminus \text{Root}(\text{Edge}(\mathbb{D}))$ and $V_v \cap V_k = \emptyset$ for $v \neq k$. Edge($\mathbb{D}$) is U_i^C-stable, so U_i^C contains Root(Edge($\mathbb{D}$)) or equivalently $\text{Root}(\text{Edge}(\mathbb{D})) \cap U_i = \emptyset$. This implies that $U_i \subset V_1 \cup V_2 \cup \ldots \cup V_q$. Since U_i is connected and the V_k are disjoint open sets, there can only be one value of k such that $U_i \cap V_k \neq \emptyset$. This implies that $U_i \subset V_k$ and consequently $z_i \in V_l$. Since $\{z_1, z_2, z_3, \ldots\} \cap \text{Root}(\mathbb{D}) = \emptyset$ and hence $z_i \notin \text{Root}(\mathbb{D})$, Theorem 3.2 implies that $V_k \cap \text{Root}(\mathbb{D}) = \emptyset$. The intersection $U_i \cap \text{Root}(\mathbb{D})$ must also be empty, therefore $\mathbb{D}$ is U_i^C-stable. Since i was arbitrary, $\mathbb{D}$ is U_i^C-stable for all i. This implies that $\mathbb{D}$ is $(U_1^C \cap U_2^C \cap U_3^C \cap \ldots)$-stable. By DeMorgan's Law, $(U_1^C \cap U_2^C \cap U_3^C \cap \ldots) = (U_1 \cup U_2 \cup U_3 \cup \ldots)^C$, which equals $(G^C)^C = G$. This implies that $\mathbb{D}$ is G-stable. □

The following corollary to Theorem 3.3 follows directly from Fact 2.1.

Corollary 3.2 Let the set of polynomials $D(\Delta)$ be generated by affine uncertainties, and let G be an arbitrary stability region in $\mathbb{C}$. $D(\Delta)$ is G-stable if and only if $D(\Delta)$ satisfies the precondition for G and $D(\text{Edge}(\Delta))$ is G-stable.

One of the main reasons the theorem above is a contribution is based on computational considerations. There are methods for determining the G-stability of

Edge(𝔻) that are significantly more computationally efficient than computing Root(Edge(𝔻)) [see Bialas, 1985; Fu, Barmish,1988; Bartlett, Hollot, 1988; Ackermann, Barmish, 1988; Saydy, Tits, Abed, 1988; Bose, 1989]. These edge tests can be combined with the stability version of the Edge Theorem to efficiently determine if 𝔻 is G-stable provided that the restrictions on G and 𝔻 are satisfied. If the restrictions are violated, then the computationally advantageous edge tests can't be utilized to determine the G-stability of 𝔻 unless the precondition is used.

The usefulness of the precondition is illustrated by the next example. A edge stability test due to [Bose, 1989] will be used to determine if Edge(𝔻) is stable. The polytope of polynomials considered in the example cannot be handled by previous stability versions of the Edge Theorem, so stability of Edge(𝔻) alone will not imply stability of 𝔻. By testing the precondition and applying Theorem 3.3, it will be possible to use the results of Bose's test to determine the stability of the entire polytope. By using this edge test and the precondition instead of computing Root(Edge(𝔻)), the amount of computations required is greatly reduced. These saving could not have been realized without Theorem 3.3.

Example 3.6 Let 𝔻 be the polytope of polynomials for which Vert(𝔻) is comprised of the following three vertices

$$p_1(s) = s + 1$$

$$p_2(s) = s^2 + (3 + j) s + (2 + j 2)$$

$$p_3(s) = s^2 + (4 - j) s + (3 - j 3).$$

Let the stability region be the open left half plane G_H. The instability region G_H^C is clearly comprised of a single connected set, so the precondition requires only one test. When the origin is used as the precondition test point, the value set is $Val(\mathbb{D},0) = conv(\{1, 2+j2, 3-j3\})$. This polygon is contained in the open right half

plane, so the origin is not covered by the value set and hence the precondition is satisfied. To show that the edges are G_H-stable, the resultant method of [Bose, 1989] will be used. The first condition is that the vertices must be G_H-stable. By computing Root(Vert($\mathbb{D}$)) = {-1, -2, -3, -1-j, -1+j}, it is clear that the vertices are G_H-stable. The second condition requires the formation of a resultant $\Delta(\lambda)$ for each edge (see [Bose, 1989] for details). The edge will be stable if and only if $\Delta(\lambda)$ has no zeros in the interval (0,1). For the three edges of this polytope, the resultants are

$$\Delta_{1,2}(\lambda) = -4\,\lambda^3 + 22\,\lambda^2 - 37\,\lambda + 20$$

$$\Delta_{1,3}(\lambda) = -18\,\lambda^3 + 78\,\lambda^2 - 110\,\lambda + 51$$

$$\Delta_{2,3}(\lambda) = -11\,\lambda^3 + 37\,\lambda^2 - 57\,\lambda + 51.$$

By computing the root spaces of these polynomials

$$\text{Root}(\ \{\Delta_{1,2}(\lambda)\}\ ) = \{\ 2.9156,\ 1.2922 \pm j\ 0.2124\ \}$$

$$\text{Root}(\ \{\Delta_{1,3}(\lambda)\}\ ) = \{\ 1.9133,\ 1.2100 \pm j\ 0.1293\ \}$$

$$\text{Root}(\ \{\Delta_{2,3}(\lambda)\}\ ) = \{\ 1.9227,\ 0.7205 \pm j\ 1.3756\ \},$$

it is determined that the resultants have no roots in (0,1), so the edges are all G_H-stable. Since Edge($\mathbb{D}$) is G_H-stable and $\mathbb{D}$ satisfies the precondition for G_H, Theorem 3.3 implies that $\mathbb{D}$ is G_H-stable.

In addition to G-stability, the previous section raised the issue of determining if all polynomials in a polytope have m roots in a given region of the complex plane. [Soh, Foo, 1989] conjectured that for any polytope $\mathbb{D} \subset \mathbb{C}[s]$ and any simply connected region G every element in $\mathbb{D}$ has m roots in G if and only if every element in Edge($\mathbb{D}$) has m roots in G. Examples 3.3, 3.4, and 3.5 showed that this conjecture is not true as stated. The results in this section suggest ways of correcting this conjecture, but this may not be the best way of carrying out this type of analysis.

An alternate approach is to recast this problem in terms of stability. The first step in this approach is to show that the polytope $\mathbb{D}$ is (G $\cup$ B)-stable where B is a set that is separable[5] from G (there is often a natural choice for B based on other goals of the overall analysis). Next, the continuity of root loci [Marden, 1949; Zedek, 1965] is used to show that (under certain conditions) if any polynomial in $\mathbb{D}$ has m roots in G then every polynomial in $\mathbb{D}$ has m roots in G. Conditions under which this last statement is guaranteed to be true are when G is bounded and/or when there exists an n such that $\mathbb{D} \subset \mathbb{C}_n[s]$. This alternate approach is not original to this dissertation; it has been used by others [for example, see the discussion of the dominant pole location problem by Fu, Barmish, 1989].

For the special case of polytopes of polynomials, the approach described above is one possible way of carrying out the analysis suggested in [Soh, Foo, 1989] (It also extends [Soh, Foo, 1989] because G is not required to be simply connected). In addition to polynomials, [Soh, Foo, 1989] were also interested in the zeros of other functions. [Fu, Olbrot, Polis, 1989] have already extended a stability version of the Edge Theorem to "quasi" or "delay" polynomials. The extension to polytopes of more general functions appears to be more difficult. The results and counterexamples above will be useful in any attempt to achieve this extension, but that undertaking is beyond the scope of this dissertation.

[5]G and B are called separable if there exist disjoint open sets V and U such that D $\subset$ V and B $\subset$ U.

3.7 No Stability Edge Theorem for Multiaffine Uncertainties

This section will review a counterexample due to [Barmish, Fu, Saleh, 1988]. This counterexample shows that the stability of a polynomial set $D(\Delta)$ generated by multiaffine uncertainties cannot be determined from the stability of its edges.

Example 3.7 Consider the following multiaffine mapping $D: \mathbb{R}^2 \rightarrow \mathbb{C}[s]$

$$D((x,y)) = s^4 + (2.56 + x + y)s^3 + (2.871 + 2.06\,x + 1.561y + xy)s^2$$
$$+ (3.164 + 4.841x + 1.56y + 1.06xy)s + (1.853 + 3.773x + 1.985y + 4.032xy)$$

and the set of possible parameters

$$\Delta = \{ (x,y) \in \mathbb{R}^2 : 0 \le x \le 1, 0 \le y \le 3 \}.$$

It was shown in [Barmish, Fu, Saleh, 1988] that $Root(D(\Delta)) \not\subset G_H$ even though $Root(D(Edge(\Delta))) \subset G_H$.

The counterexample above is very strong because $D(\Delta)$ is a basic set of polynomials containing only 4-th order real polynomials and because the stability region G_H is one of the two most basic stability regions. Based on the strength of this counterexample, it is unlikely that a G_H-stability edge theorem for multiaffine uncertainties exists without excessive restrictions on $D(\Delta)$. One might wonder if a stability edge theorem for multiaffine uncertainties exists for other stability regions. The following counterexample shows that one does not hold for G_S. The absence of a stability edge theorem for the two most important stability regions G_H and G_S does not generate much optimism concerning edge theorems for other stability regions.

Example 3.8 Consider the multiaffine mapping D: $\mathbb{R}^2 \rightarrow \mathbb{C}[s]$ and the interval family of 2-tuples from Example 3.7. Define the multiaffine mapping P: $\mathbb{R}^2 \rightarrow \mathbb{C}[z]$ as follows

$$\begin{aligned} P(\,(x,y)\,) \;=\; & (z+1)^4 + (2.56 + x + y)(z+1)^3(z-1) \\ & + (2.871 + 2.06\,x + 1.561y + xy)(z+1)^2(z-1)^2 \\ & + (3.164 + 4.841x + 1.56y + 1.06xy)(z+1)(z-1)^3 \\ & + (1.853 + 3.773x + 1.985y + 4.032xy)(z-1)^4. \end{aligned}$$

For any $(x,y) \in \Delta$, the two polynomials $d(s) = D(\,(x,y)\,)$ and $p(z) = P(\,(x,y)\,)$ are related by the equation

$$p(z) = (z-1)^4\, d\left(\frac{z+1}{z-1}\right).$$

This relation shows that the roots of d(s) (except those at s = 1) are mapped to the roots of p(z) via the bilinear transformation. Note that all the coefficients of all the polynomials in $D(\Delta)$ are positive, so none of the polynomials in $D(\Delta)$ have a root a s = 1. This implies that, for any $(x,y) \in \Delta$, if the polynomial $D(\,(x,y)\,)$ has m roots in G_H and 4-m roots in G_H^C, then the polynomial $P(\,(x,y)\,)$ will have m roots in G_S and 4-m roots in G_S^C. From the fact that $\text{Root}(D(\Delta)) \not\subset G_H$ and $\text{Root}(D(\text{Edge}(\Delta))) \subset G_H$, [Barmish, Fu, Saleh, 1988], it follows that $\text{Root}(P(\Delta)) \not\subset G_S$ even though $\text{Root}(P(\text{Edge}(\Delta))) \subset G_S$. This straight forward extension of the [Barmish, Fu, Saleh, 1988] example contradicts a G_S-stability edge conjecture for sets of n-th order real polynomials generated by multiaffine uncertainties.

For polynomial sets generated by multiaffine uncertainties, it is doubtful that a modified stability edge theorem can be obtained through the use of a simple "precondition." This pessimism is based on the fact that $\partial\text{Root}(D(\Delta))$ is generally not a subset of $\text{Root}(\text{Edge}(\Delta))$ when $D(\Delta)$ is generated by multiaffine uncertainties. Without

this root space property, it is not possible to follow the precondition approach that was used with polytopes of polynomials.

3.8 Counterexamples to Vertex Theorems for Affine Uncertainties

For a set of polynomials $D(\Delta)$ generated by affine uncertainties, this section will provide two counterexamples which show that G-stability of $D(Vert(\Delta))$ does not imply that $D(\Delta)$ is G-stable. The first counterexample will show that G_H-stability of a set of 3rd order real polynomials generated by affine uncertainties is not implied by G_H-stability of the vertex polynomials. This is equivalent to showing that the set of all G_H-stable real polynomials is not convex. It is not clear at what point this lack of convexity was first discovered, but this fact easily follows from work conducted as far back as the 1950's. In [Truxal, 1955, pp. 595-598], an example due to [Kochenburger, 1953] of a conditionally stable control system is presented which implies that the set of G_H-stable 5-th order real polynomials is not convex. A simpler 3rd order example based on the work of [Bialas, Garloff, 1985] is given next.

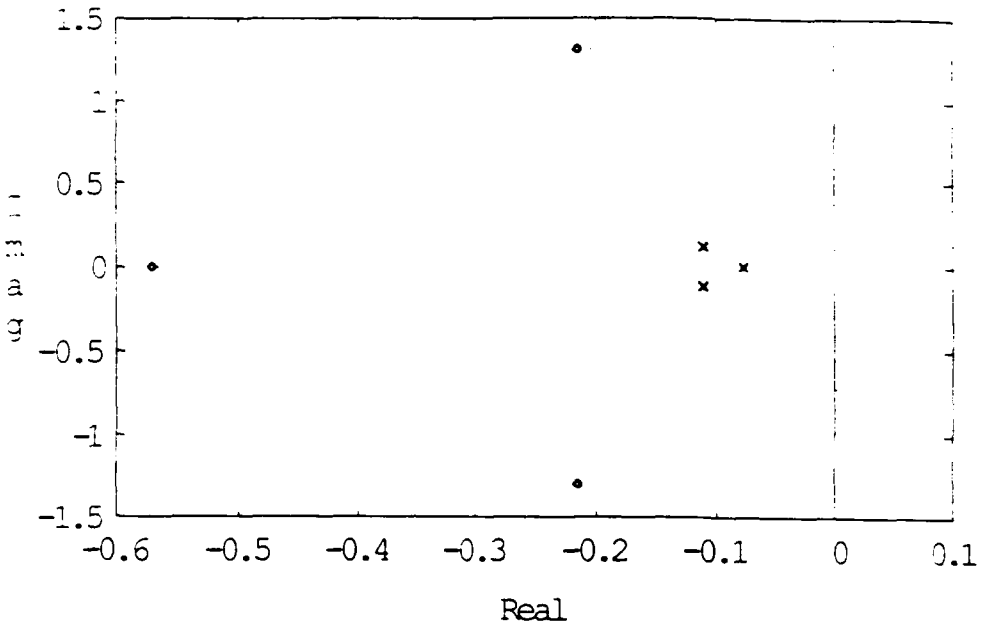
Example 3.8 Consider the following affine mapping $D: \mathbb{R} \rightarrow \mathbb{R}[s]$

$$D(\delta) = s^3 + (1-0.7\delta)s^2 + (2-0.957\delta)\, s + (1-0.998\delta)$$

and the set of possible parameter values

$$\Delta = \{ \delta \in \mathbb{R} : 0 \leq \delta \leq 1 \}.$$

Figure 3.13 shows the roots of the vertices $D(1)$ and $D(0)$ represented by x's and o's, respectively. The dotted line connecting the x's and the o's represents $Root(D(\Delta))$. The figure clearly shows that $D(\Delta)$ is not G_H-stable even though $D(Vert(\Delta))$ is G_H-stable.



Root space for Example 3.8.

Figure 3.13

For sets of polynomials generated by affine uncertainties, a counterexample to a G_H-stability vertex conjecture can easily be transformed in to a counterexample to a G_S-stability vertex conjecture by using the bilinear transformation. For this reason, once it is known that the set of G_H-stable n-th order real polynomials is not convex, it follows almost immediately that the set of G_S-stable n-th order real polynomials is also not convex. It is not clear when this fact was first pointed out. Clearly, [Ackermann, 1980, Figure 4] shows that the set of G_S-stable 3rd order real polynomials is not convex. A slightly modified version of a 4th order counterexample due to [Hollot, Bartlett, 1986] is given next.

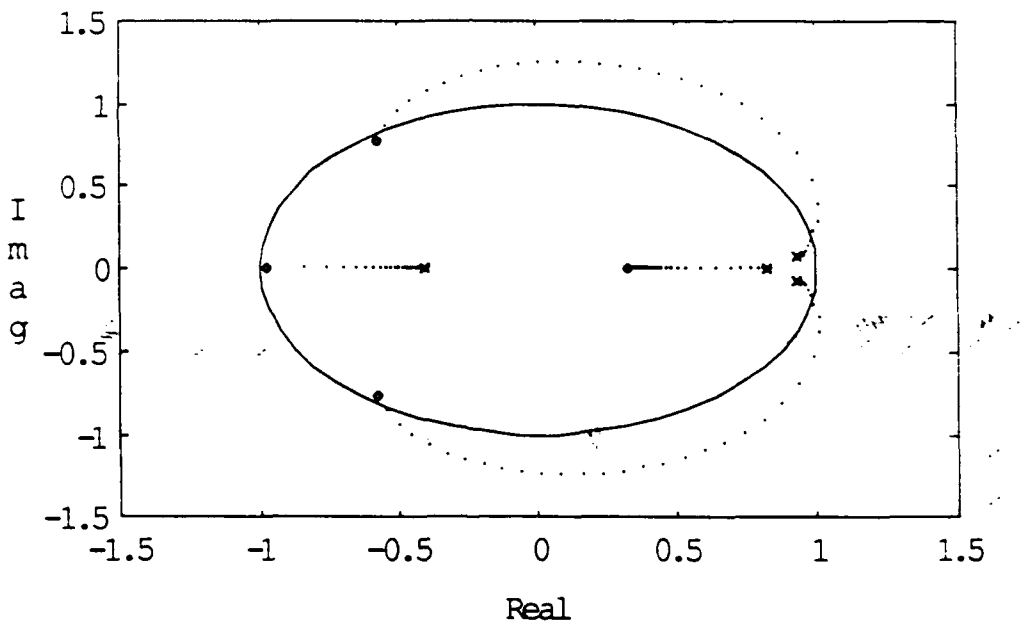
Example 3.9 Consider the following affine mapping $D: \mathbb{R} \rightarrow \mathbb{R}[s]$

$$D(\delta) = s^4 + \delta s^3 + 1.35 s^2 + 0.243 s - 0.2916$$

and the set of possible parameter values

$$\Delta = \{ \delta \in \mathbb{R} : -2.299 \leq \delta \leq 1.79 \}.$$

Figure 3.14 shows the unit circle represented by a solid line and the roots of the vertices D(-2.299) and D(1.79) represented by x's and o's, respectively. The roots of the vertices are inside the unit disk, so the vertices are G_S-stable. The dotted line connecting the x's and the o's represents Root(D(Δ)). The figure clearly shows that D(Δ) is not G_S-stable even though D(Vert(Δ)) is G_S-stable.



Root space for Example 3.9

Figure 3.14

The two examples above show that in general G-stability of the vertices of a polynomial set generated by affine uncertainties does not imply G-stability of the whole set. It may be possible that a vertex conjecture does hold for certain stability regions but not for the two most important regions. Example 3.8 shows that a G_H-stability vertex conjecture is false, and Example 3.9 shows that a G_S-stability vertex conjecture is also false. In addition to the affine case, these examples obviously disprove similar conjectures for polytopes of polynomials and polynomial sets generated by multiaffine uncertainties. The case of interval families of polynomials will be considered in the next section.

3.9 Vertex Theorems for the Stability of Interval Families

This section will review the existence and non-existence of G-stability vertex theorems for interval families of polynomials. For arbitrary stability regions, G-stability vertex theorems do not exist. This is demonstrated by Example 3.9 above which shows that an interval family is not G_S-stable even though its vertices are G_S-stable. Despite this counterexample, G-stability vertex theorems do hold for several important stability regions. The first of these theorems concerns G_H and is due to [Kharitonov, 1978a,b].

Theorem 3.4 [see Kharitonov, 1978a & b for proof] Let $\mathbb{D}$ be an interval family of polynomials. $\mathbb{D}$ is G_H-stable if and only if $\mathrm{Vert}(\mathbb{D})$ is G_H-stable.

This is known as the **weak version of Kharitonov's Theorem.** Kharitonov also proved a much stronger theorem. This stronger version relates stability of the family to the stability of the polynomials defined below.

Definition 3.2 For the interval family

$$\mathcal{D} = \{ \quad (c_q+jd_q)s^q + (c_{q-1}+jd_{q-1})s^{q-1} + \ldots + (c_1+jd_1)s + (c_0+jd_0) \ :$$

$$c_0^L \le c_0 \le c_0^H, \qquad c_1^L \le c_1 \le c_1^H, \qquad \ldots, \qquad c_q^L \le c_q \le c_q^H,$$

$$d_0^L \le d_0 \le d_0^H, \qquad d_1^L \le d_1 \le d_1^H, \qquad \ldots, \qquad d_q^L \le d_q \le d_q^H \ \}.$$

define the following eight polynomials

$$p_{HR+}(s) = c_0^H + jd_1^L s + c_2^L s^2 + jd_3^H s^3 + c_4^H s^4 + jd_5^L s^5 + c_6^L s^6 + jd_7^H s^7 + c_8^H s^8 + \ldots$$

$$p_{LR+}(s) = c_0^L + jd_1^H s + c_2^H s^2 + jd_3^L s^3 + c_4^L s^4 + jd_5^H s^5 + c_6^H s^6 + jd_7^L s^7 + c_8^L s^8 + \ldots$$

$$p_{HI+}(s) = jd_0^H + c_1^H s + jd_2^L s^2 + c_3^L s^3 + jd_4^H s^4 + c_5^H s^5 + jd_6^L s^6 + c_7^L s^7 + jd_8^H s^8 + \ldots$$

$$p_{LI+}(s) = jd_0^L + c_1^L s + jd_2^H s^2 + c_3^H s^3 + jd_4^L s^4 + c_5^L s^5 + jd_6^H s^6 + c_7^H s^7 + jd_8^L s^8 + \ldots$$

$$p_{HR-}(s) = c_0^H + jd_1^H s + c_2^L s^2 + jd_3^L s^3 + c_4^H s^4 + jd_5^H s^5 + c_6^L s^6 + jd_7^L s^7 + c_8^H s^8 + \ldots$$

$$p_{LR-}(s) = c_0^L + jd_1^L s + c_2^H s^2 + jd_3^H s^3 + c_4^L s^4 + jd_5^L s^5 + c_6^H s^6 + jd_7^H s^7 + c_8^L s^8 + \ldots$$

$$p_{HI-}(s) = jd_0^H + c_1^L s + jd_2^L s^2 + c_3^H s^3 + jd_4^H s^4 + c_5^L s^5 + jd_6^L s^6 + c_7^H s^7 + jd_8^H s^8 + \ldots$$

$$p_{LI-}(s) = jd_0^L + c_1^H s + jd_2^H s^2 + c_3^L s^3 + jd_4^L s^4 + c_5^H s^5 + jd_6^H s^6 + c_7^L s^7 + jd_8^L s^8 + \ldots$$

These polynomials are used to define the following sets

$$\text{HiLow}_+(\mathcal{D}) = \{ \ p_{HR+}(s), \ p_{LR+}(s), \ p_{HI+}(s), \ p_{LI+}(s), \ 0 \ \}$$

$$\text{Khar}_+(\mathcal{D}) = \{ \ p_{HR+}(s)+p_{HI+}(s), \ p_{HR+}(s)+p_{LI+}(s), \ p_{LR+}(s)+p_{LI+}(s), \ p_{LR+}(s)+p_{HI+}(s) \ \}$$

$$\text{HiLow}_-(\mathcal{D}) = \{ \ p_{HR-}(s), \ p_{LR-}(s), \ p_{HI-}(s), \ p_{LI-}(s), \ 0 \ \}$$

$$\text{Khar}_-(\mathcal{D}) = \{ \ p_{HR-}(s)+p_{HI-}(s), \ p_{HR-}(s)+p_{LI-}(s), \ p_{LR-}(s)+p_{LI-}(s), \ p_{LR-}(s)+p_{HI-}(s) \ \}$$

$$\text{Khar}(\mathcal{D}) = \text{Khar}_+(\mathcal{D}) \cup \text{Khar}_-(\mathcal{D}).$$

The next theorem is **the strong version of Kharitonov's Theorem.**

Theorem 3.5 [see Kharitonov, 1978a,b, and Bose, Shi, 1987 for proof]
Let $\mathcal{D}$ be an interval family of polynomials. $\mathcal{D}$ is G_H-stable if and only if Khar($\mathcal{D}$) is

G_H-stable. Furthermore, if $\mathcal{D} \subset \mathcal{R}[s]$, then $\mathcal{D}$ is G_H-stable if and only if $Khar_+(\mathcal{D})$ is G_H-stable.

The strong version of Kharitonov's Theorem has not been extended to any other stability region, but the weak version has been extended. For the stability region

$$G_\theta = \{ s \in \mathbb{C} : \pi+\theta \le \arg(s) \le \pi-\theta \}$$

where $0 \le \theta \le \pi/2$, [Soh, Berger, 1988] showed that an interval family of polynomials $\mathcal{D} \subset \mathbb{C}_n[s]$ is G_θ-stable if and only if Vert($\mathcal{D}$) is G_θ-stable. This result is generalized by the following theorem due to [Petersen, 1989] (some further generalizations of this theorem have also been considered by [Fu, 1989a]).

Theorem 3.6 [see Petersen, 1989 for proof] Given the following sets

$$H_1 = \{ s \in \mathbb{C} : \mathrm{Re}[(s+\sigma_1)\exp(-j\theta_1)] < 0 \}$$

$$H_2 = \{ s \in \mathbb{C} : \mathrm{Re}[(s+\sigma_2)\exp(-j\theta_2)] < 0 \}$$

$$\vdots$$

$$H_k = \{ s \in \mathbb{C} : \mathrm{Re}[(s+\sigma_k)\exp(-j\theta_k)] < 0 \}$$

$$B_1 = \{ s \in \mathbb{C} : \mathrm{abs}(s+c_1) < r_1 \}$$

$$B_2 = \{ s \in \mathbb{C} : \mathrm{abs}(s+c_2) < r_2 \}$$

$$\vdots$$

$$B_q = \{ s \in \mathbb{C} : \mathrm{abs}(s+c_q) < r_q \}$$

where $\sigma_1, \sigma_2, \ldots, \sigma_k \ge 0$; $\theta_1, \theta_2, \ldots, \theta_k \in [-\pi/2,\pi/2]$; and $c_i \ge r_i > 0$ for $i = 1,2,\ldots,q$. Let $\mathcal{D} \subset \mathbb{C}_n[s]$ be an interval family of polynomials, and let

$$G = H_1 \cap H_2 \cap \ldots \cap H_k \cap B_1 \cap B_2 \cap \ldots \cap B_q.$$

$\mathcal{D}$ is G-stable if and only if Vert($\mathcal{D}$) is G-stable.

The stability vertex theorems in this section are primarily useful for continuous-time systems. For discrete-time systems, the best available results are the stability versions of the Edge Theorem.

3.10 Conclusion

This chapter has given a thorough review of the existence of vertex and edge theorems pertaining to root location and stability analyses of the four classes of polynomial families considered in this dissertation. It was shown that, for polynomial sets generated by multiaffine uncertainties, no vertex or edge theorems are available to simplify root location or stability analyses. This is true even for the two special cases of G_H and G_S-stability analyses and for the special case of real polynomials. For polytopes of polynomials and polynomial sets generated by affine uncertainties, the availability of edge theorems and the absence of vertex theorems were pointed out for both root location and stability analyses. Care must be taken with the use of this stability edge theorem. Either assumptions on the polytope and on the stability region must be satisfied or a precondition must be used. For interval families of polynomials, it was demonstrated that root location analyses and general cases of stability analyses can be simplified by edge theorems but not by vertex theorems. This is true even for the special case of G_S-stability analysis. However, for some special stability regions including the open left half plane G_H, it was shown that powerful vertex theorems exist for the stability analysis of interval families of polynomials.

CHAPTER 4

FREQUENCY RESPONSE ANALYSES

4.1 Introduction

This chapter will discuss frequency response analyses of the four classes of uncertain systems defined in Chapter 2. The main goal of this chapter is to review the availability of vertex and edge theorems that simplify the computation of the frequency response sets

$$\mathrm{Nyq}(\mathbf{T},z)=\left\{ \left[\frac{n(z)}{d(z)}\right] \in \mathbb{C} \;:\; \mathrm{abs}\left[\frac{n(z)}{d(z)}\right] < \infty,\ (n(s),d(s)) \in \mathbf{T} \right\}$$

and

$$\mathrm{Mag}(\mathbf{T},z)=\left\{ \mathrm{abs}\left[\frac{n(z)}{d(z)}\right] \in \mathbb{R} \;:\; \mathrm{abs}\left[\frac{n(z)}{d(z)}\right] < \infty,\ (n(s);d(s)) \in \mathbf{T} \right\}$$

where $\mathbf{T}$ is a polynomial-pair set and z is any complex frequency. Several vertex theorems, edge theorems, and counterexamples to similar conjectures will be presented.

The results in this chapter are directly applicable to Nyquist plane and Bode magnitude plot analyses. These results are also indirectly applicable to analyses using the Nichols chart and the Bode phase plot. This is due to the fact that theorems which simplify the computation of Nyq(T,z) will automatically provide similar theorems for computation of the **template** [Horowitz, 1982]

$$\mathrm{Nic}(\mathbf{T},z)=\left\{ \left(20\mathrm{Log}\left(\mathrm{abs}\left[\frac{n(z)}{d(z)}\right]\right), \arg\left[\frac{n(z)}{d(z)}\right] \right) \in \mathbb{R}^2 \;:\; 0 < \mathrm{abs}\left[\frac{n(z)}{d(z)}\right] < \infty,\ (n(s),d(s)) \in \mathbf{T} \right\}$$

and the set

$$\mathrm{Arg}(\mathbb{T},z) = \left\{ \arg\left[\frac{n(z)}{d(z)}\right] \in \mathbb{R} \ : \ 0 < \mathrm{abs}\left[\frac{n(z)}{d(z)}\right] < \infty, \ (n(s),d(s)) \in \mathbb{T} \right\}$$

where $\mathbb{T}$ is a polynomial-pair set and z is any complex frequency. These theorems will not be stated explicitly, but computation of templates will be illustrated by example.

4.2 An Edge Theorem for Polytopes of Polynomial-Pairs

This section will present a frequency response edge theorem for polytopes of polynomial-pairs. This edge theorem is obtained by relating the frequency response of a set of polynomial-pairs to the roots of a set of polynomials. This is the purpose of the following lemma.

Lemma 4.1 If $\mathbb{T} \subset \mathbb{C}[s]^2$ and $z \in \mathbb{C}$ such that

$$(0,0) \notin \mathrm{Val}(\mathbb{T},z) = \{ \ (n(z),d(z)) \in \mathbb{C}^2 \ : \ (n(s),d(s)) \in \mathbb{T} \ \}$$

then

$$\mathrm{Nyq}(\mathbb{T},z) = \mathrm{Root}(\ \{ \ d(z)x-n(z) \in \mathbb{C}[x] \ : \ (n(s),d(s)) \in \mathbb{T} \ \} \ ).$$

Proof The assumption $(0,0) \notin \mathrm{Val}(\mathbb{T},z)$ implies that for any $(n(s),d(s)) \in \mathbb{T}$ if $d(z) = 0$, then $n(z) \neq 0$. In turn, it follows that $d(z)x-n(z) \in \mathbb{C}[x]$ has a root if and only if $d(z) \neq 0$ and that $\mathrm{abs}[n(z)/d(z)] < \infty$ if and only if $d(z) \neq 0$. This implies the set equalities

$$\mathrm{Nyq}(\mathbb{T},z) = \{ \ n(z)/d(z) \in \mathbb{C} \ : \ d(z) \neq 0, \ (n(s),d(s)) \in \mathbb{T} \ \}$$

and

$$\mathrm{Root}(\ \{ \ d(z)x-n(z) \in \mathbb{C}[x] \ : \ (n(s),d(s)) \in \mathbb{T} \ \} \ ) =$$
$$\mathrm{Root}(\ d(z)x-n(z) \in \mathbb{C}[x] \ : \ d(z) \neq 0, \ (n(s),d(s)) \in \mathbb{T} \ \} \ ).$$

The lemma now follows from the fact that if $d(z) \neq 0$ then $n(z)/d(z)$ is the only root of the polynomial $d(z)x-n(z) \in \mathbb{C}[x]$. □

Lemma 1 has recast the frequency response problem in terms of polynomial root locations. This will allow use of Theorem 3.1 if $\mathbb{T}$ is appropriately structured.

Theorem 4.1 Given $\mathbb{T} \subset \mathbb{C}[s]^2$ and $z \in \mathbb{C}$ such that $(0,0) \notin \mathrm{Val}(\mathbb{T},z)$. If $\mathbb{T}$ is a polytope, then

$$\partial \mathrm{Nyq}(\mathbb{T},z) \subset \mathrm{Nyq}(\mathrm{Edge}(\mathbb{T}),z).$$

Proof Lemma 4.1 shows that

$$\mathrm{Nyq}(\mathbb{T},z) = \mathrm{Root}(\{ d(z)x-n(z) \in \mathbb{C}[x] : (n(s),d(s)) \in \mathbb{T} \}). \quad (1)$$

Because $\mathbb{T}$ is a polytope in $\mathbb{C}[s]^2$ and because $d(z)x-n(z)$ can be treated as an affine mapping from $\mathbb{C}[s]^2$ to $\mathbb{C}[x]$, it follows that the set

$$\mathbb{P} = \{ d(z)x-n(z) \in \mathbb{C}[x] : (n(s),d(s)) \in \mathbb{T} \} \quad (2)$$

will be a polytope in $\mathbb{C}[x]$ and that

$$\mathrm{Edge}(\mathbb{P}) \subset \{ d(z)x-n(z) \in \mathbb{C}[x] : (n(s),d(s)) \in \mathrm{Edge}(\mathbb{T}) \}. \quad (3)$$

Since $\mathbb{P}$ is a polytope, Theorem 3.1 can be used to show that

$$\partial \mathrm{Root}(\mathbb{P}) \subset \mathrm{Root}(\mathrm{Edge}(\mathbb{P})). \quad (4)$$

Combining equations (3) and (4) gives

$$\partial \mathrm{Root}(\mathbb{P}) \subset \mathrm{Root}(\{ d(z)x-n(z) \in \mathbb{C}[x] : (n(s),d(s)) \in \mathrm{Edge}(\mathbb{T}) \}). \quad (5)$$

From equations (1) and (2), it follows that

$$\partial \mathrm{Nyq}(\mathbb{T},z) = \partial \mathrm{Root}(\mathbb{P}). \quad (6)$$

Applying Lemma 4.1 in the other direction gives

$$\mathrm{Root}(\{ d(z)x-n(z) \in \mathbb{C}[x] : (n(s),d(s)) \in \mathrm{Edge}(\mathbb{T}) \}) = \mathrm{Nyq}(\mathrm{Edge}(\mathbb{T}),z). \quad (7)$$

Substituting equations (6) and (7) into (5) gives

$$\partial Nyq(\mathbb{T},z) \subset Nyq(Edge(\mathbb{T}),z)$$

which completes the proof. □

The following corollary to Theorem 4.1 is easily obtained using Fact 2.1.

Corollary 4.1 Given $T(\Delta) \subset \mathbb{C}[s]^2$ and $z \in \mathbb{C}$ such that $(0,0) \notin Val(T(\Delta),z)$. If $T(\Delta)$ is generated by affine uncertainties, then

$$\partial Nyq(T(\Delta),z) \subset Nyq(T(Edge(\Delta)),z).$$

Theorem 4.1 shows that for polytopes of polynomial-pairs most of the important information about the frequency response set $Nyq(\mathbb{T},z)$ can be determined by using just the edge descriptions. To determine the entire response set, it is necessary to determine the "pseudo-interior" $Nyq(\mathbb{T},z) \setminus Nyq(Edge(\mathbb{T}),z)$. The following theorem indicates an easy procedure for finding these points. The proof follows directly from the definition of boundary, so it is omitted.

Theorem 4.2 Assume that $\mathbb{T} \subset \mathbb{C}[s]^2$ and $z \in \mathbb{C}$ such that $(0,0) \notin Val(\mathbb{T},z)$, and assume that $\mathbb{T}$ is a polytope. Let c be any point in V where V is a connected subset of $\mathbb{C}$ such that $V \cap Nyq(Edge(\mathbb{T}),z) = \emptyset$. If the polytope of complex numbers

$$\{d(z)c-n(z) \in \mathbb{C} : (n(s),d(s)) \in \mathbb{T}\} = conv(\{d(z)c-n(z) \in \mathbb{C} : (n(s),d(s)) \in Vert(\mathbb{T})\})$$

contains the origin, then $V \subset Nyq(\mathbb{T},z)$. Conversely, If

$$0 \notin conv(\{d(z)c-n(z) \in \mathbb{C} : (n(s),d(s)) \in Vert(\mathbb{T})\}),$$

then $V \cap Nyq(\mathbb{T},z) \neq \emptyset$.

The application of Theorem 4.1 and 4.2 to find the response set of a polytope of polynomial-pairs is illustrated by the following example.

Example 4.1 Consider the set of polynomial-pairs

$$\mathbb{T} = \{ (n(s)+a_0n_0(s)+a_1n_1(s)+a_2n_2(s),\ d(s)+a_0d_0(s)+a_1d_1(s)+a_2d_2(s)) : a_0, a_1, a_2 \in [-0.1, 0.1] \}$$

where

$$n(s) = s^4 + 3\,s^3 + 3\,s^2 + 3\,s + 2$$

$$n_0(s) = 0.1\,s^3 + 0.1\,s^2 - 0.2\,s + 0.3$$

$$n_1(s) = -0.2\,s^2 - 0.1$$

$$n_2(s) = -0.3\,s^3 - 0.1\,s^2 + 0.2\,s$$

$$d(s) = s^5 + 15\,s^4 + 99\,s^3 + 495\,s^2 + 1850\,s + 3000$$

$$d_0(s) = 0$$

$$d_1(s) = -30\,s^4 + 20\,s$$

$$d_2(s) = 1000\,s^2 + 1000\,s + 2000.$$

The edge frequency response Nyq(Edge($\mathbb{T}$), j1) is shown in Figure 4.1. The edge response divides the Nyquist plane into seven bounded regions and one unbounded region. To determine which of these regions is part of Nyq($\mathbb{T}$, j1), Theorem 4.2 is used. For the test point $c = j3\times10^{-5}$, Figure 4.2 shows that

$$0 \notin \text{conv}(\{d(z)c-n(z) \in \mathbb{C} : (n(s),d(s)) \in \text{Vert}(\mathbb{T})\}).$$

This implies that Nyq($\mathbb{T}$, j1) and the region containing c (the unbounded region) share no points in common. For the test point $c = 0$, Figure 4.3 shows that

$$0 \in \text{conv}(\{d(z)c-n(z) \in \mathbb{C} : (n(s),d(s)) \in \text{Vert}(\mathbb{T})\}).$$

The bounded region containing c is thereforea subset of Nyq($\mathbb{T}$, j1). By similar tests, it can be shown that the other six bounded regions are contained in Nyq($\mathbb{T}$, j1). A complete representation of Nyq($\mathbb{T}$, j1) is shown in Figure 4.4. The equivalent template Nic($\mathbb{T}$, j1) is shown in Figure 4.5. The template is unbounded because at least one of the possible descriptions of the system has a zero at $z = j1$.

Now consider the frequency z = j3. The edge frequency response Nyq(Edge($\mathbb{T}$), j3) is shown in Figure 4.6. Using the test point c = -0.024-j0.024, Figure 4.7 shows that the unbounded region in Figure 4.6 is not part of Nyq($\mathbb{T}$, j3) because

$$0 \notin \text{conv}(\{ d(z)c-n(z) : (n(s),d(s)) \in \text{Vert}(\mathbb{T}) \}).$$

By similar tests, it can be shown that all the bounded regions are contained in Nyq($\mathbb{T}$, j3). A complete representation of Nyq($\mathbb{T}$, j3) is shown in Figure 4.8. The equivalent template Nic($\mathbb{T}$, j3) is shown in Figure 4.9.

For the frequency z = j5.05, the edge frequency response Nyq(Edge($\mathbb{T}$), j5.05) is shown in Figure 4.10. The edge response divides the Nyquist plane into two large regions and a few thin regions that are hidden by the curve. Figure 4.11 shows that

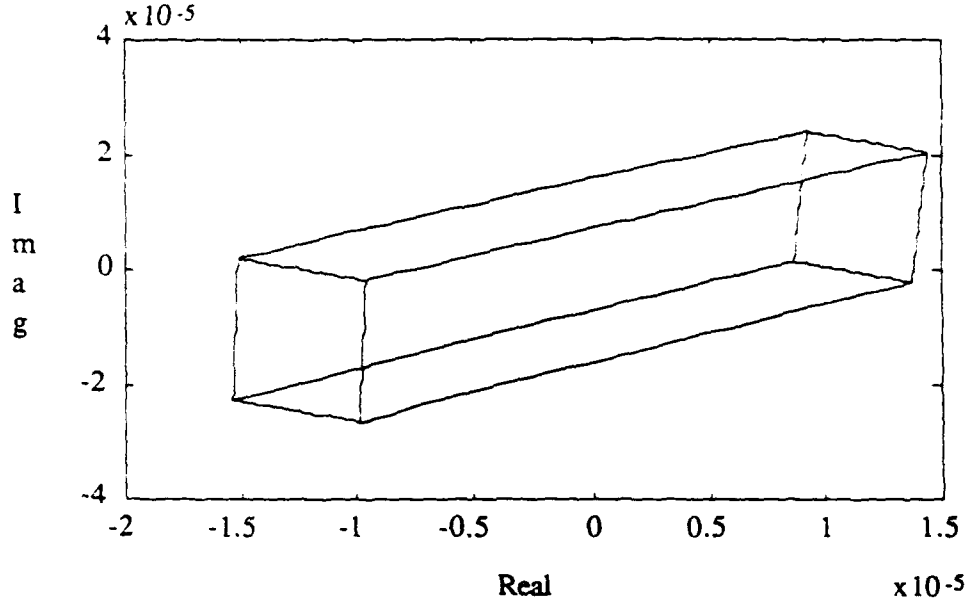
$$0 \notin \text{conv}(\{ d(z)(-1-j0.3)-n(z) : (n(s),d(s)) \in \text{Vert}(\mathbb{T}) \})$$

and hence the large bounded region does not intersect Nyq($\mathbb{T}$,j5.05). Figure 4.12 shows that

$$0 \in \text{conv}(\{ d(z)(-1-j)-n(z) : (n(s),d(s)) \in \text{Vert}(\mathbb{T}) \})$$

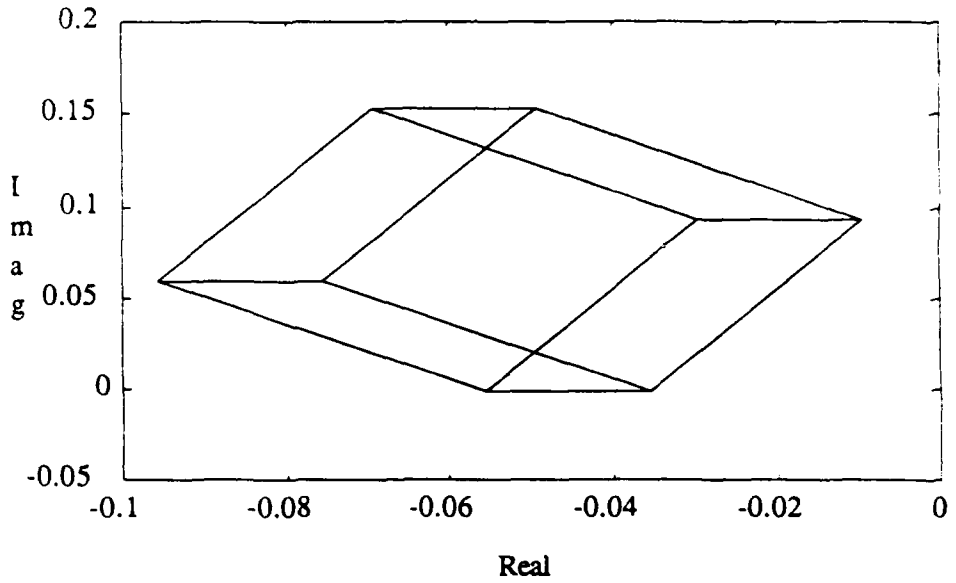
and hence the large unbounded region is contained in Nyq($\mathbb{T}$,j5.05). The thin regions hidden by the curve are of little significance, so Figure 4.13 provides a representation of Nyq($\mathbb{T}$,j5.05) that is accurate enough for almost any analysis. By similar procedures, one can produce Figure 4.14 which shows Nic($\mathbb{T}$,j5.05). Figures 4.13 and 4.14 indicates that at least one model of the system has a pole at j5.05.

The results given in this section can only be applied at frequencies where there is no pole-zero cancelation, i.e. (0,0) $\notin$ Val($\mathbb{T}$,z). This condition is very hard to test algebraically, so it may seem like a major drawback to Theorem 4.1 and 4.2. It is not a drawback because this condition really does not need to be checked in advance. If there is a pole-zero cancellation at z, then its existence will become apparent when the



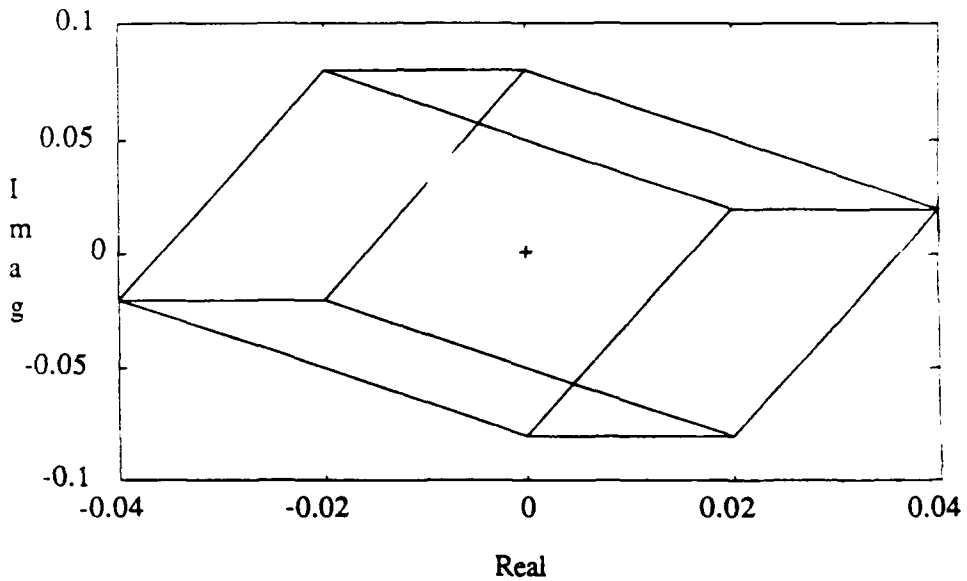
Edge frequency response at z = j1 for Example 4.1

Figure 4.1



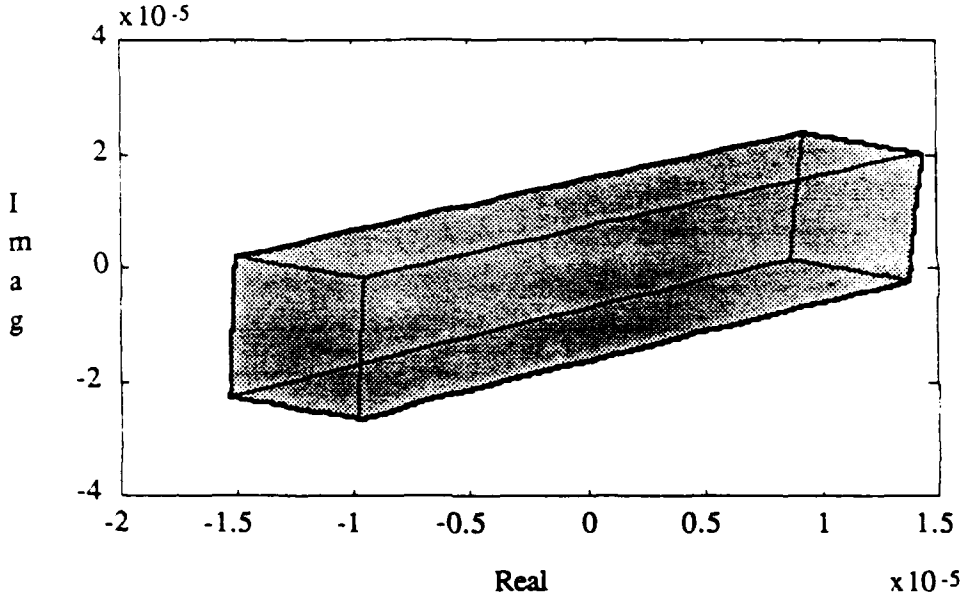
Test showing that the unbounded region in Figure 4.1 is not part of the frequency response at $z = j1$ of Example 4.1.

Figure 4.2



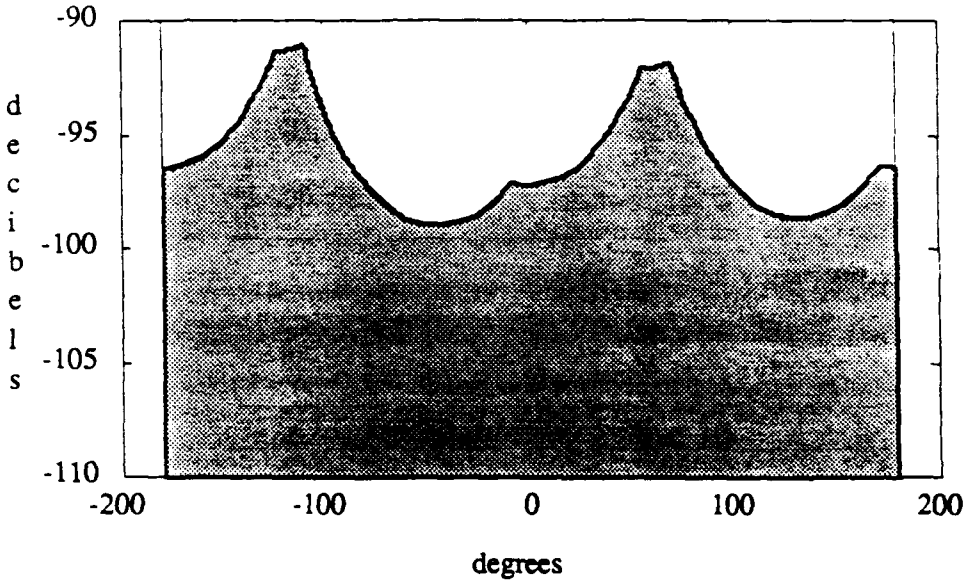
Test showing that the region in Figure 4.1 containing the point c = 0 is part of the frequency response at z = j1 of Example 4.1.

Figure 4.3



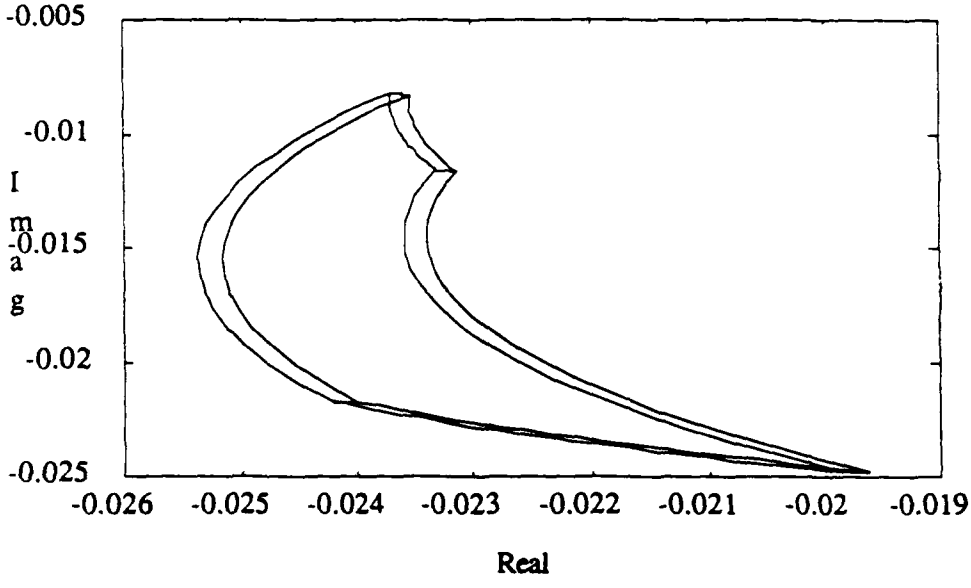
Frequency response at z = j1 of Example 4.1.

Figure 4.4



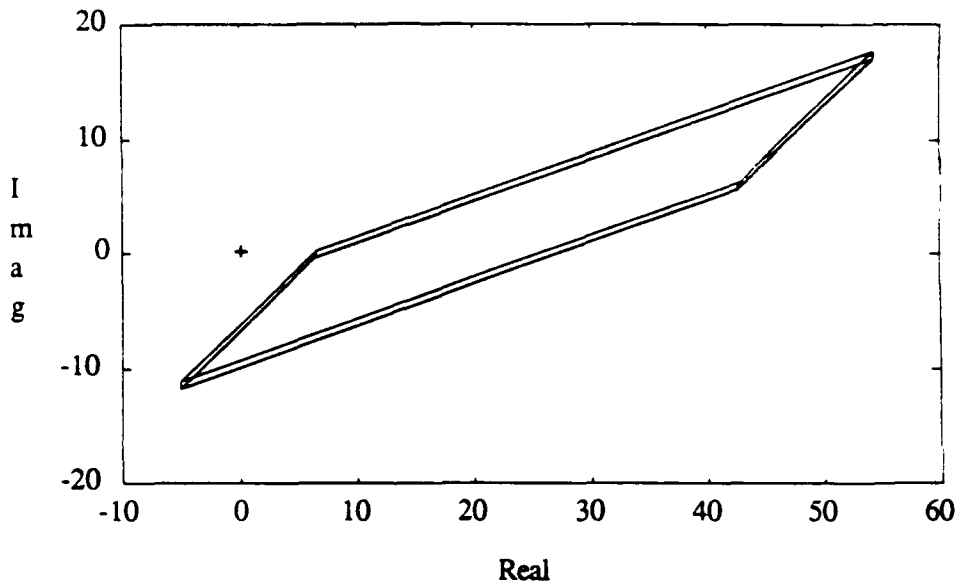
Template at z = j1 of Example 4.1.

Figure 4.5



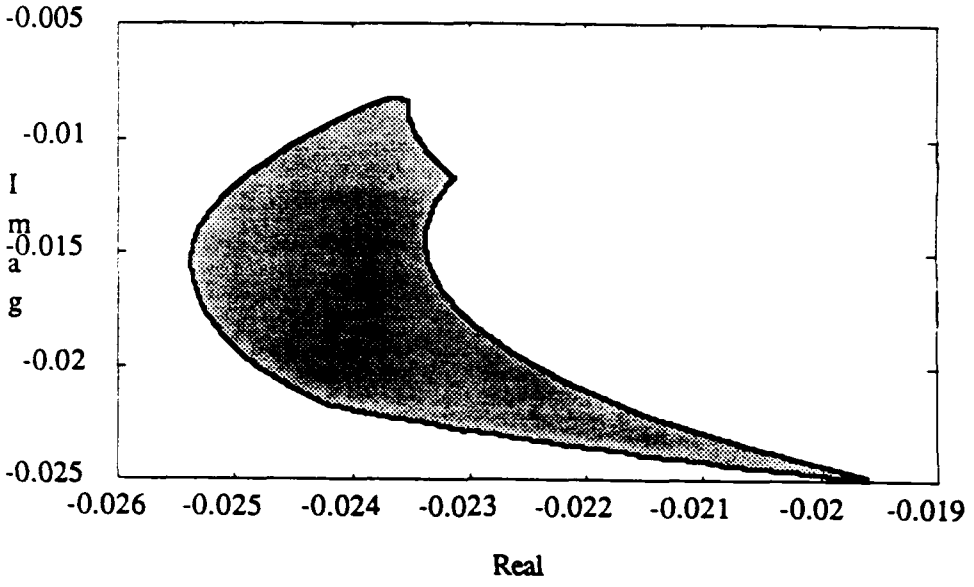
Edge frequency response at z = j3 of Example 4.1.

Figure 4.6



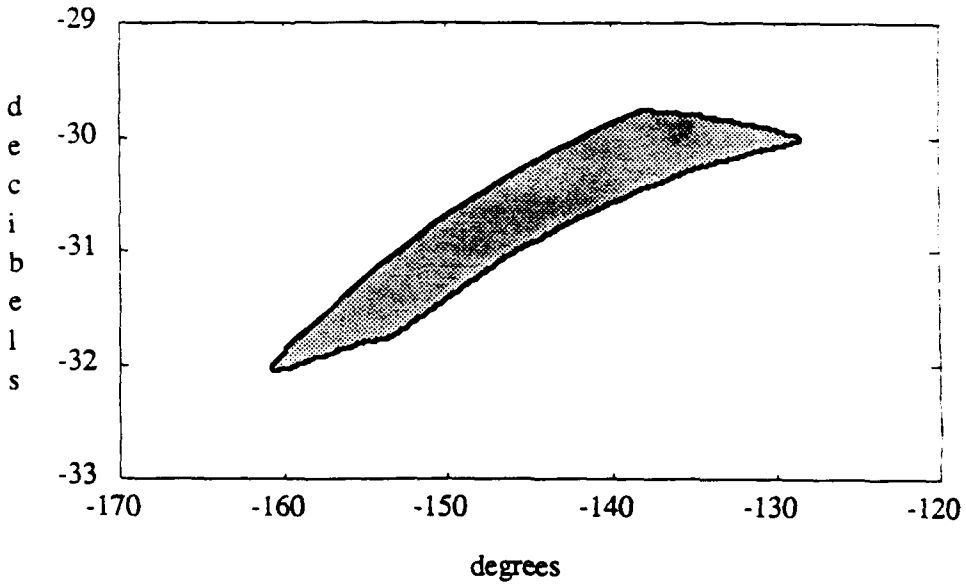
Test showing that the unbounded region in Figure 4.6 is not part of the frequency response at z = j3 of Example 4.1.

Figure 4.7



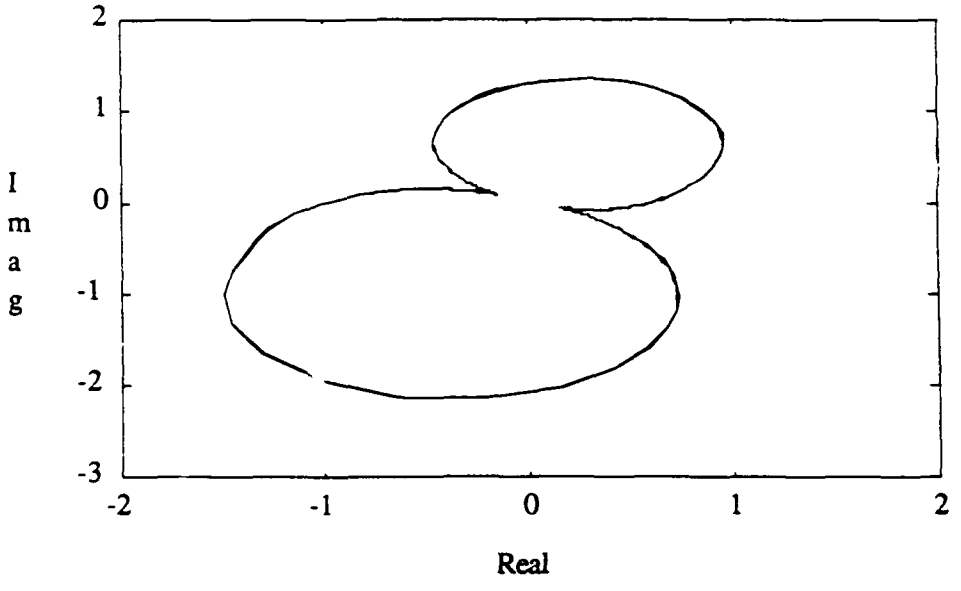
Frequency response at z = j3 of Example 4.1.

Figure 4.8



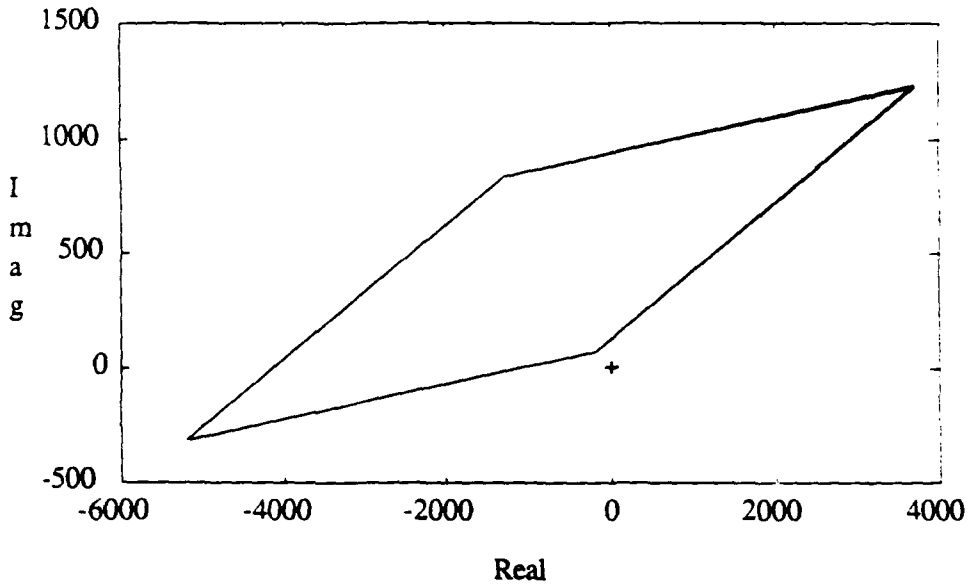
Template at z = j3 of Example 4.1.

Figure 4.9



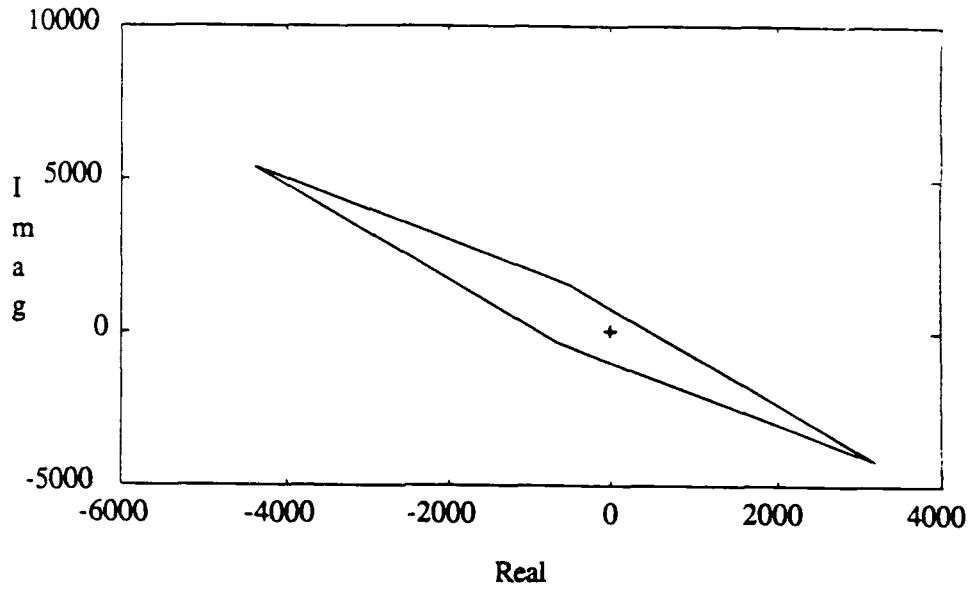
Edge frequency response at z = j5.05 of Example 4.1.

Figure 4.10



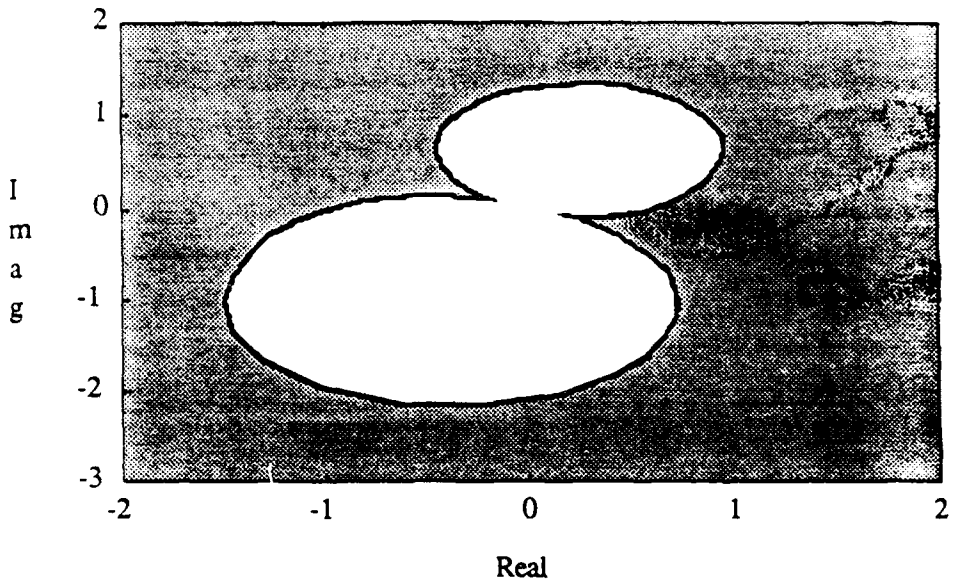
Test showing that the large bounded region in Figure 4.10 is not part of the frequency response at z = j5.05 of Example 4.1.

Figure 4.11



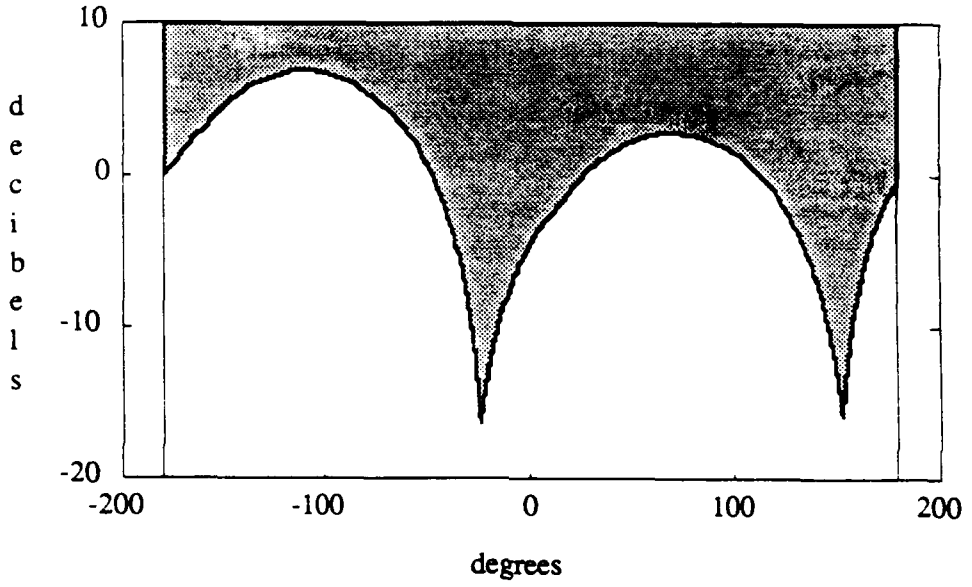
Test showing that the large unbounded region in Figure 4.10 is part of the frequency response at z = j5.05 of Example 4.1.

Figure 4.12



Frequency response at z = j5.05 of Example 4.1.

Figure 4.13



Template at z = j5.05 of Example 4.1.

Figure 4.14

procedure illustrated in Example 4.1 is followed. The analyst will find that according to the theory every region bounded by Nyq(Edge($\mathbb{T}$),z) is contained in Nyq($\mathbb{T}$,z), so the set of all possible responses would seem to be the entire complex plane $\mathbb{C}$. This occurs because $(0,0) \in \text{Val}(\mathbb{T},z)$ and hence $\text{conv}(\{d(z)c-n(z) : (n(s),d(s)) \in \mathbb{T}\})$ will contain the origin for every $c \in \mathbb{C}$. Once the cancellation is detected, it must be accepted that the results in this section are not applicable to this system. This is not of much consequence because the existence of an unstable pole-zero cancellation makes the satisfaction of frequency domain specifications a moot point.

The theorems and example in this section are taken from [Bartlett, 1990b]. These results represent the primary original contribution of the author to frequency response analyses of systems with uncertain parameters.

4.3 No Edge Theorem for Multiaffine Uncertainties

This section will give a counterexample which shows that the frequency response of a set of polynomial-pairs $T(\Delta)$ generated by multiaffine uncertainties cannot be determined from the frequency response of $T(\text{Edge}(\Delta))$. Two classical texts from 1963 indicate that it was known to the authors that no edge theorem exists for this problem [Horowitz, 1963, p. 161, Figure 4.12-3b; Zadeh, Desoer, 1963, p. 475, Figure 9.17.5]. The example given in this section is based on a more recent paper by [Barmish, Fu, Saleh, 1988].

Example 4.2 Recall the multiaffine mapping $D: \mathbb{R}^2 \rightarrow \mathbb{C}[s]$

$$D((x,y)) = s^4 + (2.56 + x + y)s^3 + (2.871 + 2.06x + 1.561y + xy)s^2$$
$$+ (3.164 + 4.841x + 1.56y + 1.06xy)s + (1.853 + 3.773x + 1.985y + 4.032xy)$$

that was given by [Barmish, Fu, Saleh, 1988]. Using this operator, define the following multiaffine mapping T: $\mathbb{R}^2 \rightarrow \mathbb{C}[s]^2$

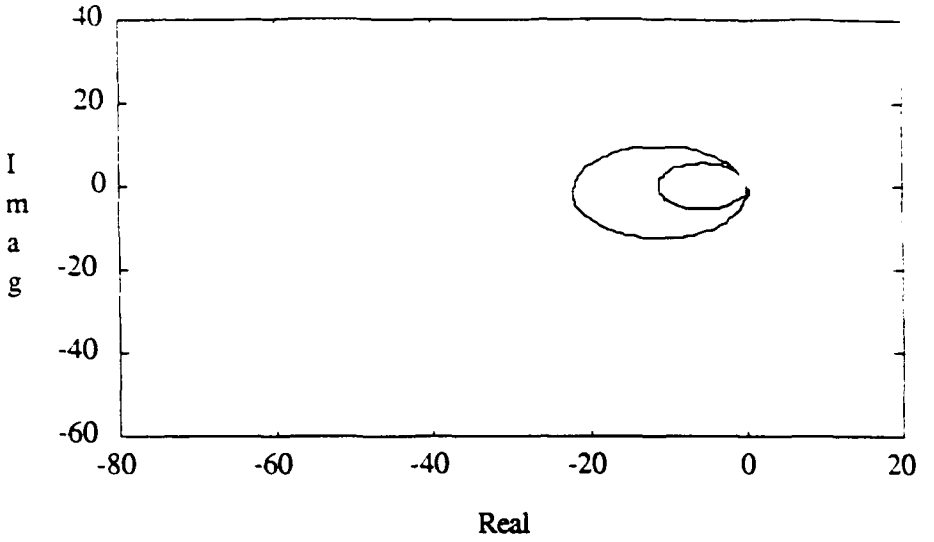
$$T((x,y)) = (1, D((x,y))).$$

Let the set of possible parameters be

$$\Delta = \{ (x,y) \in \mathbb{R}^2 : 0.2 \leq x \leq 0.7, 0 \leq y \leq 3 \}.$$

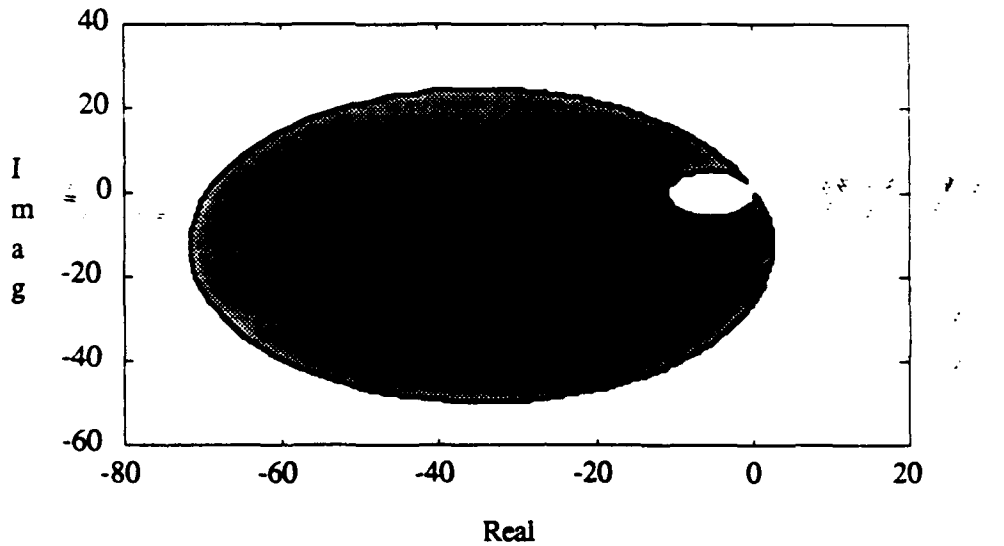
Figure 4.15 shows Nyq(T(Edge(Δ)), j1.26), and Figure 4.16 shows an accurate approximation of Nyq(T(Δ), j1.26). These two figures show that the boundary of the frequency response at z = j1.26 of T(Δ) is not given by the frequency response at z = j1.26 of T(Edge(Δ)). Since it is not generally true that ∂Nyq(T(Δ), z) $\subset$ Nyq(T(Edge(Δ)), z) when T(Δ) is generated by multiaffine uncertainties, it is hard to conceive of any simple way to determine the frequency response at z of T(Δ) from the frequency response at z of T(Edge(Δ)). Figures 4.15 and 4.16 also show that max Mag(T(Δ),j1.26) is finite and larger than max Mag(T(Edge(Δ)),j1.26). This implies that Mag(T(Δ),z) cannot in general be determine from Mag(T(Edge(Δ)),j1.26).

For the case of multiaffine uncertainties, the example above shows that in general Nyq(T(Δ),z) can't be determined from Nyq(T(Edge(Δ)),z). It also shows that an edge theorem does not hold for the special case when z is restricted to the jw-axis. One might wonder it there is an edge theorem for the special case when z is restricted to the unit circle. The answer is no. The example above could be used to show this by replacing s with x/1.26. Figures 4.15 and 4.16 would give the response of this new example at x = j which is on the unit circle.



Nyq(T(Edge(Δ)), j1.26) for Example 4.2.

Figure 4.15



Approximation of Nyq(T(Δ), j1.26) for Example 4.2.

Figure 4.16

4.4 No Vertex Theorem for Affine Uncertainties

For the case when the set of polynomial-pairs $T(\Delta)$ is generated by affine uncertainties, Section 4.2 showed that the frequency response could be determined using just the edges $T(Edge(\Delta))$. One might wonder if the frequency response can be determined using the smaller set $T(Vert(\Delta))$. Example 4.1 showed that the set of complex numbers $Nyq(T(\Delta),z)$ can have a variety of complicated shapes. Given this complexity, it is hard to imagine how $Nyq(T(\Delta),z)$ could be easily determined using the finite set of complex numbers $Nyq(T(Vert(\Delta)),z)$. Because the Bode magnitude plot uses only real numbers, it would seem much more likely that the set $Mag(T(\Delta),z)$ could be determined from the finite set $Mag(T(Vert(\Delta)),z)$. Despite the advantage of using only real numbers, this section will provide an example which shows that in general $Mag(T(Vert(\Delta)),z)$ does not provide adequate information concerning $Mag(T(\Delta),z)$.

Example 4.3 Given the polynomials

$$p_1(s) = s^3+0.3665s^2+0.0940s+0.0235$$

$$p_2(s) = s^3+0.7025s^2+1.0452s+0.5191$$

$$p_3(s) = s^3+3s^2+3s+1,$$

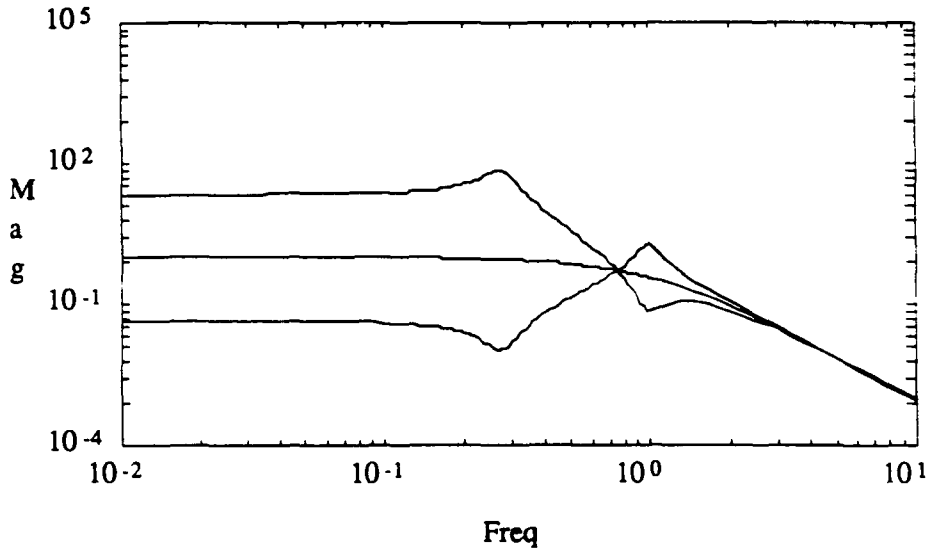
define the affine mapping $T: \mathbb{R}^2 \rightarrow \mathbb{C}[s]$ as follows

$$T((x,y)) = \left(xp_1(s)+(1-x)p_2(s),\ p_3(s)(yp_1(s)+(1-y)p_2(s)) \right).$$

Let the set of parameter values be

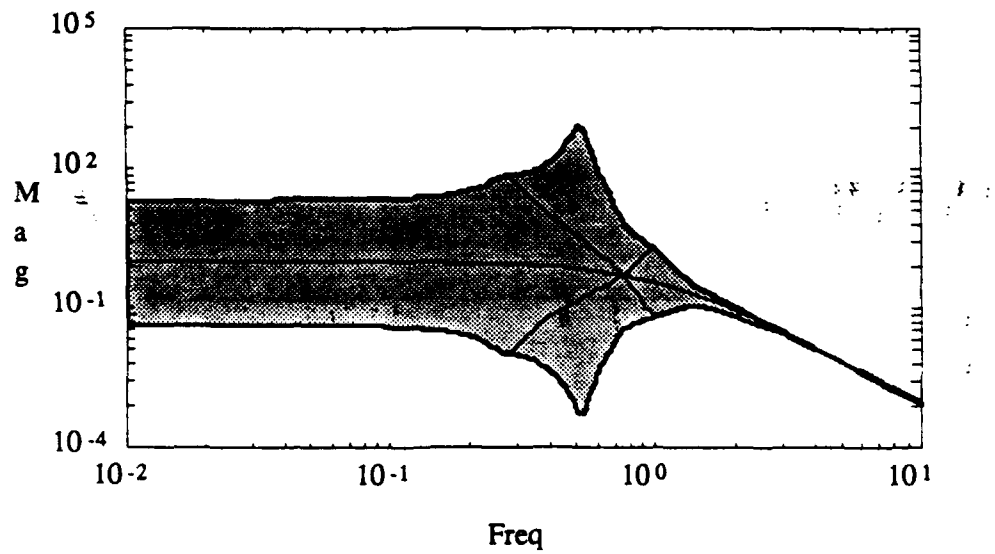
$$\Delta = \{ (x,y) \in \mathbb{R}^2 : 0 \le x \le 1, 0 \le y \le 1 \}.$$

Figure 4.17 shows the Bode magnitude plot of the vertices $Mag(T(Vert(\Delta)),jw)$ for $w \in [0.01, 10]$. Figure 4.18 shows the complete Bode magnitude plot $Mag(T(\Delta),jw)$ for several values of $w \in [0.01, 10]$. Comparison of the two figures clearly shows that



The Bode magnitude plot of the vertices Mag(T(Vert(Δ)),jw) for Example 4.3.

Figure 4.17



The complete Bode magnitude plot Mag(T(Δ),jw) for Example 4.3.

Figure 4.18

Mag(T(Vert(Δ)),jw) does not provide sufficient information to determine Mag(T(Δ),jw).

For the case of affine uncertainties, the example above shows that in general Mag(T(Δ),z) can't be determined from Mag(T(Vert(Δ)),z). It also shows that an vertex theorem does not hold for the special case when z is restricted to the jw-axis. One might wonder it there is an vertex theorem for the special case when z is restricted to the unit circle. The answer is no. The example above could be used to show this by replacing s with (x+1)/(x-1). Along a portion the unit circle, the Bode magnitude response of this new example in x would be very similar to Figures 4.17 and 4.18.

4.5 A Simplified Edge Theorem for Continuous-Time Interval Families

In Section 4.2, it was shown that if $\mathbb{T}$ is a polytope of polynomial-pairs, then the frequency response Nyq($\mathbb{T}$,z) can be determined using only the edges of $\mathbb{T}$. This section will show that, for the special case when $\mathbb{T}$ is an interval family of polynomial-pairs and z = jw is a frequency on the imaginary axis, the set Nyq($\mathbb{T}$,jw) can be determined using just 32 edge-like subsets of $\mathbb{T}$ rather than using all possible edges. This result is essentially a simplification of Theorem 4.1 for frequency responses analyses of continuous-time systems represented by interval families of polynomial-pairs.

The results in this section are based on observations relating the value set of interval families of polynomials to the value sets of the Kharitonov polynomials (Definition 3.2). In particular, it is known that if $\mathbb{P}$ is an interval family of polynomials, then for w ≥ 0,

$$\mathrm{Val}(\mathbb{P},jw) = \mathrm{Val}\big(\mathrm{conv}(\mathrm{Khar}_+(\mathbb{P})),jw\big) = \mathrm{conv}\big(\mathrm{Val}(\mathrm{Khar}_+(\mathbb{P}),jw)\big), \qquad (8)$$

and for $w \leq 0$,

$$\mathrm{Val}(\mathbb{P},jw) = \mathrm{Val}\big(\mathrm{conv}(\mathrm{Khar}_-(\mathbb{P})),jw\big) = \mathrm{conv}\big(\mathrm{Val}(\mathrm{Khar}_-(\mathbb{P}),jw)\big). \qquad (9)$$

This relationship is known from the works of several authors such as [Bose, Shi, 1987; Dasgupta, 1988; Barmish, 1989; Minnichelli, Anagnost, Desoer, 1989; Chapellat, Bhattacharyya, 1989]. In order to exploit this relationship for frequency response analyses, the interval family of polynomial-pairs $\mathbb{T}$ will be split into two sets of polynomials

$$\mathrm{Num}(\mathbb{T}) = \{\ n(s) \in \mathbb{C}[s]\ :\ (n(s),d(s)) \in \mathbb{C}[s]^2\ \}$$

$$\mathrm{Den}(\mathbb{T}) = \{\ d(s) \in \mathbb{C}[s]\ :\ (n(s),d(s)) \in \mathbb{C}[s]^2\ \}.$$

It is easily seen by comparing Definitions 2.14 and 2.15 that $\mathrm{Num}(\mathbb{T})$ and $\mathrm{Den}(\mathbb{T})$ are interval families of polynomials and that $\mathbb{T}$ equals the cartesian product of these two sets $\mathrm{Num}(\mathbb{T}) \times \mathrm{Den}(\mathbb{T})$. These relationships are used by the following theorem.

Theorem 4.3 Given $\mathbb{T} \subset \mathbb{C}[s]^2$ and $w \in \mathbb{R}$ such that $(0,0) \notin \mathrm{Val}(\mathbb{T},jw)$. If $\mathbb{T}$ is an interval family of polynomial-pairs then, for $w \geq 0$,

$$\mathrm{Nyq}(\mathbb{T},jw) = \mathrm{Nyq}\Big(\ \mathrm{conv}\big(\ \mathrm{Khar}_+(\mathrm{Num}(\mathbb{T})) \times \mathrm{Khar}_+(\mathrm{Den}(\mathbb{T}))\ \big),\ jw\Big),$$

and for $w \leq 0$,

$$\mathrm{Nyq}(\mathbb{T},jw) = \mathrm{Nyq}\Big(\ \mathrm{conv}\big(\ \mathrm{Khar}_-(\mathrm{Num}(\mathbb{T})) \times \mathrm{Khar}_-(\mathrm{Den}(\mathbb{T}))\ \big),\ jw\Big).$$

$\mathrm{Khar}_+(\cdot)$ and $\mathrm{Khar}_-(\cdot)$ are given in Definition 3.2.

Proof First, consider the case when $w \geq 0$. Because $\mathbb{T}$ is assumed to be an interval family, it is true that

$$\mathbb{T} = \mathrm{Num}(\mathbb{T}) \times \mathrm{Den}(\mathbb{T}).$$

From the assumption $(0,0) \notin \mathrm{Val}(\mathbb{T},jw)$, it can be seen that

$$\mathrm{Nyq}(\mathbb{T},jw) = \mathrm{Nyq}(\mathrm{Num}(\mathbb{T}) \times \mathrm{Den}(\mathbb{T}),jw) = \mathrm{Val}(\mathrm{Num}(\mathbb{T}),jw) \,/\, \mathrm{Val}(\mathrm{Den}(\mathbb{T}),jw).$$

Since Num(𝕋) and Den(𝕋) are interval families of polynomials, it follows from equation (8) that

$$\mathrm{Nyq}(\mathbb{T},jw) = \mathrm{Val}\big(\mathrm{conv}(\mathrm{Khar}_+(\mathrm{Num}(\mathbb{T}))),jw\big) \,/\, \mathrm{Val}\big(\mathrm{conv}(\mathrm{Khar}_+(\mathrm{Den}(\mathbb{T}))),jw\big).$$

This in turn implies that

$$\mathrm{Nyq}(\mathbb{T},jw) = \mathrm{Nyq}\big(\ \mathrm{conv}(\mathrm{Khar}_+(\mathrm{Num}(\mathbb{T}))) \times \mathrm{conv}(\mathrm{Khar}_+(\mathrm{Den}(\mathbb{T}))),\ jw\big).$$

Since

$$\mathrm{conv}\big(\mathrm{Khar}_+(\mathrm{Num}(\mathbb{T}))\big) \times \mathrm{conv}\big(\mathrm{Khar}_+(\mathrm{Den}(\mathbb{T}))\big)$$
$$= \mathrm{conv}\big(\ \mathrm{Khar}_+(\mathrm{Num}(\mathbb{T})) \times \mathrm{Khar}_+(\mathrm{Den}(\mathbb{T}))\ \big),$$

it follows that

$$\mathrm{Nyq}(\mathbb{T},jw) = \mathrm{Nyq}\big(\ \mathrm{conv}\big(\ \mathrm{Khar}_+(\mathrm{Num}(\mathbb{T})) \times \mathrm{Khar}_+(\mathrm{Den}(\mathbb{T}))\ \big),\ jw\big).$$

This completes the proof for the case when $w \geq 0$. The proof for the second case, $w < 0$, is identical to the proof for $w \geq 0$ except that the set $\mathrm{Khar}_+(\cdot)$ is replaced by the set $\mathrm{Khar}_-(\cdot)$. □

The next theorem follows directly from the combination of Theorem 4.1 and 4.3, so it is given without proof.

Theorem 4.4 Given $\mathbb{T} \subset \mathbb{C}[s]^2$ and $w \in \mathbb{R}$ such that $(0,0) \notin \mathrm{Val}(\mathbb{T},jw)$. If $\mathbb{T}$ is an interval family of polynomial-pairs then

$$\partial\mathrm{Nyq}(\mathbb{T},jw) \subset \mathrm{Nyq}(\mathrm{Edge}(\mathbb{K}),z).$$

where $\mathbb{K}$ is a polytope of polynomial-pairs given by

$$\mathbb{K} = \mathrm{conv}\big(\ \mathrm{Khar}_+(\mathrm{Num}(\mathbb{T})) \times \mathrm{Khar}_+(\mathrm{Den}(\mathbb{T}))\ \big),$$

for $w \geq 0$, and by

$$\mathbb{K} = \mathrm{conv}\big(\ \mathrm{Khar}_-(\mathrm{Num}(\mathbb{T})) \times \mathrm{Khar}_-(\mathrm{Den}(\mathbb{T}))\ \big).$$

for $w < 0$.

The boundary of Nyq(𝕋,jw) can be found using either Theorem 4.1 or 4.4. Theorem 4.1 uses the $(r+q)2^{r+q-1}$ edges of 𝕋, and Theorem 4.4 uses the 32 edges of the polytope 𝕂. The edges of 𝕂 are subsets of 𝕋, and they are convex combinations of vertices of 𝕋. Despite these edge-like properties, the edges of 𝕂 are generally not edges of 𝕋. This fact slightly complicates the comparison of the two methods, but it is still clear that Theorem 4.4 is significantly simpler to use that Theorem 4.1.

Once an overbound of ∂Nyq(𝕋,jw) is found, the interior points of Nyq(𝕋,jw) can be found using Theorem 4.2, but there is a simpler method. The following theorem indicates how to determine the interior points using only 16 vertices of 𝕋 rather than all possible vertices. The proof follows directly from Theorem 4.3 and the definition of boundary, so it is omitted.

Theorem 4.5 Assume that $\mathbb{T} \subset \mathbb{C}[s]^2$ and $w \in \mathbb{R}$ such that $(0,0) \notin \text{Val}(\mathbb{T},jw)$, and assume that 𝕋 is an interval family of polynomial-pairs. Let c be any point in V where V is a connected subset of ℂ such that $V \cap \partial\text{Nyq}(\mathbb{T},jw) = \emptyset$. For $w \geq 0$, if the polytope of complex numbers

$$\text{conv}\Big(\{ d(jw)c-n(jw) : n(s) \in \text{Khar}_+(\text{Num}(\mathbb{T})), d(s) \in \text{Khar}_+(\text{Den}(\mathbb{T})) \} \Big)$$

contains the origin, then $V \subset \text{Nyq}(\mathbb{T},z)$. Conversely, if

$$0 \notin \text{conv}\Big(\{ d(jw)c-n(jw) : n(s) \in \text{Khar}_+(\text{Num}(\mathbb{T})), d(s) \in \text{Khar}_+(\text{Den}(\mathbb{T})) \} \Big),$$

then $V \cap \text{Nyq}(\mathbb{T},z) \neq \emptyset$. For $w < 0$, if the polytope of complex numbers

$$\text{conv}\Big(\{ d(jw)c-n(jw) : n(s) \in \text{Khar}_-(\text{Num}(\mathbb{T})), d(s) \in \text{Khar}_-(\text{Den}(\mathbb{T})) \} \Big)$$

contains the origin, then $V \subset \text{Nyq}(\mathbb{T},jw)$, and if

$$0 \notin \text{conv}\Big(\{ d(jw)c-n(jw) : n(s) \in \text{Khar}_-(\text{Num}(\mathbb{T})), d(s) \in \text{Khar}_-(\text{Den}(\mathbb{T})) \} \Big),$$

then $V \cap \text{Nyq}(\mathbb{T},z) \neq \emptyset$.

As was pointed out in Chapter 1, the edge theorem formulation of the result in this section is a contribution of the author [Bartlett, 1990b]. However, the overall contribution of these results is not particularly significant because [Bailey, Panzer, 1988] have already provided a simple method for computing Nyq(T,jw) when T is an interval family.

4.6 A Vertex-Like Theorem for Continuous-Time Interval Families

For the case when the set of polynomials T(Δ) is generated by affine uncertainties, Section 4.4 showed that the Bode magnitude plot could not be determined using just the vertices. Although it has not been shown, it is also true that, for the special case when $\mathbb{T}$ is an interval family of polynomial-pairs, the set Mag(T,z) can't be determined just from the set Mag(Vert(T),z). Even though a "true" vertex theorem does not exist, this section will show that, when $\mathbb{T}$ is an interval family of polynomial-pairs and z =jw is a frequency on the imaginary axis, the set Mag(T,jw) can be determined using only a few vertex-like polynomials. The results given in this section use the following definition.

Definition 4.1 For an interval family of polynomials $\mathbb{P}$, define the following sets which are comprised of 5 or fewer complex numbers

$$\text{Aux}_+(\mathbb{P}, jw) = \text{conv}\Big(\text{Val}(\text{Khar}_+(\mathbb{P}), jw)\Big) \cap \text{Val}(\text{HiLow}_+(\mathbb{P}), jw)$$

$$\text{Aux}_-(\mathbb{P}, jw) = \text{conv}\Big(\text{Val}(\text{Khar}_-(\mathbb{P}), jw)\Big) \cap \text{Val}(\text{HiLow}_-(\mathbb{P}), jw)$$

where $\text{Khar}_+(\cdot)$, $\text{Khar}_-(\cdot)$, $\text{HiLow}_+(\cdot)$, and $\text{HiLow}_-(\cdot)$ are given in Definition 3.2.

The usefulness of the sets $Aux_+(\mathbb{P},jw)$ and $Aux_-(\mathbb{P},jw)$ will become apparent in the following lemma.

Lemma 4.1 If $\mathbb{P}$ is an interval family of polynomials, then for $w \geq 0$,

$$\max \text{abs}[\text{Val}(\mathbb{P},jw)] = \max \text{abs}\left[\text{Val}(\text{Khar}_+(\mathbb{P}), jw) \right]$$

$$\min \text{abs}[\text{Val}(\mathbb{P},jw)] = \min \text{abs}\left[\text{Val}(\text{Khar}_+(\text{Num}(\mathbb{T})), jw) \cup \text{Aux}_+(\text{Num}(\mathbb{T}), jw) \right],$$

and for $w < 0$,

$$\max \text{abs}[\text{Val}(\mathbb{P},jw)] = \max \text{abs}\left[\text{Val}(\text{Khar}_-(\mathbb{P}), jw) \right]$$

$$\min \text{abs}[\text{Val}(\mathbb{P},jw)] = \min \text{abs}\left[\text{Val}(\text{Khar}_-(\text{Num}(\mathbb{T})), jw) \cup \text{Aux}_-(\text{Num}(\mathbb{T}), jw) \right].$$

Proof First, consider the case when $w \geq 0$. The problem of finding $\max \text{abs}[\text{Val}(\mathbb{P},jw)]$ and $\min \text{abs}[\text{Val}(\mathbb{P},jw)]$ is equivalent to finding the maximum and minimum, respectively, of

$$J(r,i) = \sqrt{r^2 + i^2}$$

subject to the constraint

$$r+ji \in \text{Val}(\mathbb{P},jw)$$

Because $\mathbb{P}$ is an interval family of polynomials, the constraint set $\text{Val}(\mathbb{P},jw)$ is equal to the polytope of complex numbers

$$\text{conv}\big(\text{Val}(\text{Khar}_+(\mathbb{P}),jw)\big).$$

Since the function J is convex, its maximum will occur at a vertex of the polytopic constraint set. The vertices of the constraint set $\text{conv}\big(\text{Val}(\text{Khar}_+(\mathbb{P}),jw)\big)$ are a subset of $\text{Val}(\text{Khar}_+(\mathbb{P}),jw)$, so it follows that

$$\max \text{abs}[\text{Val}(\mathbb{P},jw)] = \max \text{abs}\left[\text{Val}(\text{Khar}_+(\mathbb{P}), jw) \right].$$

This completes the maximum portion of the proof for $w \geq 0$.

The minimum portion of the proof is more complicated. The first step is to use the observations of [Bose, Shi, 1987] to replace the set constraint $r+ji \in Val(\mathbb{P},jw)$ by the four inequality constraints

$$p_{HR+}(jw) \geq r$$

$$p_{HI+}(jw) \geq i$$

$$r \geq p_{LR+}(jw)$$

$$i \geq p_{LI+}(jw).$$

From these constraints, it can be seen that there are nine possible values for the minimum of J. The first possibility is the unconstrained minimum $J(0,0) = 0$, but this value can only be the minimum if $0+j0 \in Val(\mathbb{P},jw)$. The second possibility is the minimum when the constraint $p_{HR+}(jw) = r$ is active. This value $J(p_{HR+}(jw),0) = abs(p_{HR+}(jw))$ can only be the minimum if $p_{HR+}(jw) \in Val(\mathbb{P},jw)$. The third possibility is the minimum when the constraint $p_{LR+}(jw) = r$ is active. This value $J(p_{LR+}(jw),0) = abs(p_{LR+}(jw))$ can only be the minimum if $p_{LR+}(jw) \in Val(\mathbb{P},jw)$. The fourth possibility is the minimum when the constraint $p_{HI+}(jw) = i$ is active. This value $J(0,p_{HI+}(jw)) = abs(p_{HI+}(jw))$ can only be the minimum if $p_{HI+}(jw) \in Val(\mathbb{P},jw)$. The fifth possibility is the minimum when the constraint $p_{LI+}(jw) = i$ is active. This value $J(0,p_{LI+}(jw)) = abs(p_{LI+}(jw))$ can only be the minimum if $p_{LI+}(jw) \in Val(\mathbb{P},jw)$. Of the nine possibile values, the four remaining ones $J(p_{HR+}(jw),p_{HI+}(jw))$, $J(p_{HR+}(jw),p_{LI+}(jw))$, $J(p_{LR+}(jw),p_{HI+}(jw))$, and $J(p_{LR+}(jw),p_{LI+}(jw))\}$ occur when constraints on both r and i are active. These nine possibilities including the conditional requirements are contained in the equation

$$\min abs[Val(\mathbb{P},jw)] = \min abs\left[Val\left(Khar_{+}(Num(\mathbb{T})), jw\right) \cup Aux_{+}\left(Num(\mathbb{T}), jw\right) \right].$$

This completes the proof for $w \geq 0$. The proof for $w < 0$ is identical except that all '+' subscripts are replaced by '-' (minus) subscripts. □

This lemma easily leads to the following theorem. The vertex-like formulation of this result is a contribution of the author [Bartlett, 1990b].

Theorem 4.5 Given $\mathbb{T} \subset \mathbb{C}[s]^2$ and $w \in \mathbb{R}$ such that $(0,0) \notin \mathrm{Val}(\mathbb{T},jw)$. If $\mathbb{T}$ is an interval family of polynomial-pairs, then for $w \geq 0$,

$$\sup \mathrm{Mag}(\mathbb{T},jw) = \frac{\max \mathrm{abs}\left[\mathrm{Val}\left(\mathrm{Khar}_+(\mathrm{Num}(\mathbb{T})), jw\right) \right]}{\min \mathrm{abs}\left[\mathrm{Val}\left(\mathrm{Khar}_+(\mathrm{Den}(\mathbb{T})), jw\right) \cup \mathrm{Aux}_+\left(\mathrm{Den}(\mathbb{T}), jw\right) \right]}$$

$$\inf \mathrm{Mag}(\mathbb{T},jw) = \frac{\min \mathrm{abs}\left[\mathrm{Val}\left(\mathrm{Khar}_+(\mathrm{Num}(\mathbb{T})), jw\right) \cup \mathrm{Aux}_+\left(\mathrm{Num}(\mathbb{T}), jw\right) \right]}{\max \mathrm{abs}\left[\mathrm{Val}\left(\mathrm{Khar}_+(\mathrm{Den}(\mathbb{T})), jw\right) \right]},$$

and for $w < 0$,

$$\sup \mathrm{Mag}(\mathbb{T},jw) = \frac{\max \mathrm{abs}\left[\mathrm{Val}\left(\mathrm{Khar}_-(\mathrm{Num}(\mathbb{T})), jw\right) \right]}{\min \mathrm{abs}\left[\mathrm{Val}\left(\mathrm{Khar}_-(\mathrm{Den}(\mathbb{T})), jw\right) \cup \mathrm{Aux}_-\left(\mathrm{Den}(\mathbb{T}), jw\right) \right]}$$

$$\inf \mathrm{Mag}(\mathbb{T},jw) = \frac{\min \mathrm{abs}\left[\mathrm{Val}\left(\mathrm{Khar}_-(\mathrm{Num}(\mathbb{T})), jw\right) \cup \mathrm{Aux}_-\left(\mathrm{Num}(\mathbb{T}), jw\right) \right]}{\max \mathrm{abs}\left[\mathrm{Val}\left(\mathrm{Khar}_-(\mathrm{Den}(\mathbb{T})), jw\right) \right]}.$$

Proof Because $\mathbb{T} = \mathrm{Num}(\mathbb{T}) \times \mathrm{Den}(\mathbb{T})$ and because $(0,0) \notin \mathrm{Val}(\mathbb{T},jw)$, it is easily seen that

$$\mathrm{Mag}(\mathbb{T},jw) = \left\{ \mathrm{abs}[n(jw)/d(jw)] < \infty \ : \ n(s) \in \mathrm{Num}(\mathbb{T}),\ d(s) \in \mathrm{Den}(\mathbb{T}) \right\}.$$

From this, it follows that

$$\sup \text{Mag}(\mathbb{T},jw) = \frac{\max \text{abs}\left[\text{Val}(\text{Num}(\mathbb{T}), jw) \right]}{\min \text{abs}\left[\text{Val}(\text{Den}(\mathbb{T}), jw) \right]}$$

$$\inf \text{Mag}(\mathbb{T},jw) = \frac{\min \text{abs}\left[\text{Val}(\text{Num}(\mathbb{T}), jw) \right]}{\max \text{abs}\left[\text{Val}(\text{Den}(\mathbb{T}), jw) \right]},$$

Both Num($\mathbb{T}$) and Den($\mathbb{T}$) are interval families of polynomials, so the theorem now follows directly from Lemma 4.1. □

Even though a "true" vertex theorem does not exist, this section has shown a simple way to determine the Bode magnitude plot of an interval family of polynomial-pairs using only a few vertex-like polynomials. Based on this result, one might wonder if there is a similar vertex-like method of determining the Bode magnitude plots of polynomial-pair sets that are polytopes, are generated by affine uncertainties, or are generated by multiaffine uncertainties. The answer is no. The results in this section rely on the fact that an interval family of polynomials is always equal to the cartesian product of two sets of polynomials. For the three other classes of polynomial-pairs, this is not generally true, so the approach used in this section can not be extended to these classes.

4.7 Conclusion

For the four classes of polynomial-pair families considered in this dissertation, this chapter has given a review of the existence of vertex and edge theorems pertaining to determination of the frequency response sets Nyq($\mathbb{T}$,z) and Mag($\mathbb{T}$,z). It was shown that, for sets generated by multiaffine uncertainties, neither of the two frequency response sets could be determined using only the frequency responses of the edge

polynomial-pairs. This was true even if z was restricted to the unit circle or to the jw-axis. For sets of polynomial-pairs that are polytopes or are generated by affine uncertainties, it was shown that an edge theorem but not a vertex theorem could be used to determine the frequency response sets. A vertex theorem does not exist even for the special case when z is restricted to the jw-axis or to the unit circle. This edge theorem for polytopes obviously applies to the special case of interval families. It is also generally true that, for interval families of polynomial-pairs, no vertex theorem can be used to determine the frequency response sets. However, the analysis of interval families can be greatly simplified if attention is restricted to frequencies on the jw-axis. This restriction is too limiting for the analysis of discrete-time systems, but for continuous-time systems, this restriction is of essentially no consequence. One of the simplifying results shows that an interval family of polynomial-pairs and a polytopic subset of it have identical frequency response sets. By combining this fact and the edge theorem for polytopes, it was shown that the frequency response sets of an interval family could be determined using only the edges (32 or less) and the vertices (16 or less) of this special polytopic subset. A further simplify result, showed that, for interval families, the extreme values of the frequency response set Mag($\underline{\underline{T}}$,jw) could be determined using only 18 polynomials that were easily derived from the 16 vertices of the special polytopic subset. For the special case of z restricted to the unit circle, it may be possible to determine the Bode magnitude plot of an interval family using a vertex-like theorem, but it will certainly be much more complicated than for frequencies on the jw-axis. All the vertex and edge results presented in this chapter make a theoretical contribution to the frequency domain analysis of uncertain systems.

CHAPTER 5

TIME RESPONSE ANALYSES

5.1 Introduction

This chapter will discuss time response analyses of uncertain systems. The results in this chapter are primarily intended for analyzing the response to step inputs, but these results are also applicable to many other types of inputs. The main result of this chapter will show that, for continuous and discrete-time stable systems represented by sets of polynomial-pairs generated by multiaffine uncertainties, the extreme values of the steady state response can be determined using only the vertices. For the special case of interval families of polynomial-pairs, a simplified steady state vertex theorem will be presented. The treatment of transient response analyses in this chapter will be limited to counterexamples concerning vertex conjectures for polynomial-pair sets generated by affine uncertainties.

5.2 A Steady State Vertex Theorem for Multiaffine Uncertainties

This section will provide vertex theorems for the steady state analysis of uncertain systems whose polynomial representations are generated by multiaffine uncertainties. For simplicity, these theorems assume that the system is real, i.e. its numerator and denominator polynomials are real. These theorems also require that the input be real valued and that its transform is a real rational function. One final restriction essentially requires the output to be bounded. These three assumption are all very reasonable, and they allow a variety of important analyses to be carried out. For

example, the steady state step response of any real stable system (with multiaffine uncertainties) can be analyzed using these vertex theorems.

Continuous-time and discrete-time systems use two different transforms, so analysis of their time responses will be handled by two different theorems. Despite this separation, the theorems are nearly identical and rely on the same line of proof. The following definition and lemmas are presented to take advantage of this duplicity and to efficiently prove the two theorems which follow.

Definition 5.1 Define the following two functions

$$L_i(\lambda) = \begin{Bmatrix} ([\lambda]_1, \ldots, [\lambda]_{i-1}, [\delta^H]_i, [\lambda]_{i+1}, \ldots, [\lambda]_p) & \text{if } [\lambda]_i = [\delta^H]_i \\ ([\lambda]_1, \ldots, [\lambda]_{i-1}, [\delta^L]_i, [\lambda]_{i+1}, \ldots, [\lambda]_p) & \text{if } [\lambda]_i \neq [\delta^H]_i \end{Bmatrix}$$

$$H_i(\lambda) = \begin{Bmatrix} ([\lambda]_1, \ldots, [\lambda]_{i-1}, [\delta^L]_i, [\lambda]_{i+1}, \ldots, [\lambda]_p) & \text{if } [\lambda]_i = [\delta^L]_i \\ ([\lambda]_1, \ldots, [\lambda]_{i-1}, [\delta^H]_i, [\lambda]_{i+1}, \ldots, [\lambda]_p) & \text{if } [\lambda]_i \neq [\delta^L]_i \end{Bmatrix}.$$

For simplicity of notation, the dependence of $L_i(\lambda)$ and $H_i(\lambda)$ on δ^L and δ^H has not been explicitly indicated.

Lemma 5.1 If the two conditions

i) $a, b \in \text{Multi}\{\mathbb{R}^m, \mathbb{R}\}$

ii) $b(\delta) \neq 0$ for all $\delta \in \text{Box}\{\delta^L, \delta^H\}$

are satisfied, then

1) $\max\{ a(\delta)/b(\delta) : \delta \in \text{Box}\{\delta^L, \delta^H\} \} = \max\{ a(\delta)/b(\delta) : \delta \in \text{Vert}\{\delta^L, \delta^H\} \}$

2) $\min\{ a(\delta)/b(\delta) : \delta \in \text{Box}\{\delta^L, \delta^H\} \} = \min\{ a(\delta)/b(\delta) : \delta \in \text{Vert}\{\delta^L, \delta^H\} \}$.

Proof For simplicity of notation, let the components of δ be denoted as $(\delta_1, \delta_2, \ldots, \delta_m)$. Condition (i) allows $a(\delta)$ and $b(\delta)$ to be written in the form

$$a(\delta) = [\ a((\delta_1,\ldots,\delta_{i-1},1,\delta_{i+1},\ldots,\delta_m)) - a((\delta_1,\ldots,\delta_{i-1},0,\delta_{i+1},\ldots,\delta_m))\ ]\ \delta_i + a((\delta_1,\ldots,\delta_{i-1},0,\delta_{i+1},\ldots,\delta_m))$$

$$b(\delta) = [\ b((\delta_1,\ldots,\delta_{i-1},1,\delta_{i+1},\ldots,\delta_m)) - b((\delta_1,\ldots,\delta_{i-1},0,\delta_{i+1},\ldots,\delta_m))\ ]\ \delta_i + b((\delta_1,\ldots,\delta_{i-1},0,\delta_{i+1},\ldots,\delta_m)).$$

Using these representations, the derivative of $a(\delta)/b(\delta)$ with respect to δ_i is given by

$$\frac{\partial}{\partial\delta_i}\left(\frac{a(\delta)}{b(\delta)}\right) = \left(\frac{a((\delta_1,\ldots,\delta_{i-1},1,\delta_{i+1},\ldots,\delta_m))\ b((\delta_1,\ldots,\delta_{i-1},0,\delta_{i+1},\ldots,\delta_m))}{b((\delta_1,\delta_2,\ldots,\delta_m))^2}\right)$$

$$-\left(\frac{a((\delta_1,\ldots,\delta_{i-1},0,\delta_{i+1},\ldots,\delta_m))\ b((\delta_1,\ldots,\delta_{i-1},1,\delta_{i+1},\ldots,\delta_m))}{b((\delta_1,\delta_2,\ldots,\delta_m))^2}\right).$$

Condition (ii) implies that this derivative exists and that it is zero if and only if its numerator is zero. Since this numerator is independent of δ_i, the derivative is zero at an arbitrary point $\lambda = (\lambda_1, \lambda_2, \ldots, \lambda_m)$ if and only if it is zero at all points in the set $\mathrm{Box}\{L_i(\lambda),H_i(\lambda)\}$. This implies that if the derivative is zero, then $a(\delta)/b(\delta)$ is constant for all $\delta \in \mathrm{Box}\{L_i(\lambda),H_i(\lambda)\}$. If the derivative is not zero, then λ cannot be an extreme point of $a(\delta)/b(\delta)$ over $\mathrm{Box}\{L_i(\lambda),H_i(\lambda)\}$ unless λ equals $L_i(\lambda)$ or $H_i(\lambda)$. Together, these two conditions imply that

$$\min\left\{\frac{a(L_i(\lambda))}{b(L_i(\lambda))}, \frac{a(H_i(\lambda))}{b(H_i(\lambda))}\right\} \le \frac{a(\lambda)}{b(\lambda)} \le \max\left\{\frac{a(L_i(\lambda))}{b(L_i(\lambda))}, \frac{a(H_i(\lambda))}{b(H_i(\lambda))}\right\}. \qquad (1)$$

Since $L_i(\lambda)$ and $H_i(\lambda)$ are elements of $\mathrm{Vert}\{L_i(\varepsilon^L),H_i(\varepsilon^H)\}$ for all $\lambda \in \mathrm{Vert}\{\varepsilon^L,\varepsilon^H\}$, equation (1) implies the following

$$\min\{\, a(\delta)/b(\delta) : \delta \in \mathrm{Vert}\{\, L_m(L_{m-1}(\ldots L_1(\lambda)\ldots)), H_m(H_{m-1}(\ldots H_1(\lambda)\ldots))\,\}\,\}$$
$$\leq a(\lambda)/b(\lambda) \leq$$
$$\max\{\, a(\delta)/b(\delta) : \delta \in \mathrm{Vert}\{\, L_m(L_{m-1}(\ldots L_1(\lambda)\ldots)), H_m(H_{m-1}(\ldots H_1(\lambda)\ldots))\,\}\,\}.$$

At this point, it is noted that

$$\mathrm{Vert}\{\, L_m(L_{m-1}(\ldots L_1(\lambda)\ldots)), H_m(H_{m-1}(\ldots H_1(\lambda)\ldots))\,\} \subset \mathrm{Vert}\{\delta^L; \delta^H\}.$$

This last relationship implies that the following equation

$$\min\left\{ \frac{a(\delta)}{b(\delta)} : \delta \in \mathrm{Vert}\{\delta^L, \delta^H\} \right\} \leq \frac{a(\lambda)}{b(\lambda)} \leq \max\left\{ \frac{a(\delta)}{b(\delta)} : \delta \in \mathrm{Vert}\{\delta^L, \delta^H\} \right\}$$

is true for arbitrary $\lambda \in \mathrm{Box}\{\delta^L; \delta^H\}$ and therefore the lemma is true. □

Lemma 5.2 Let $r \in \mathbb{R}$ and $P \in \mathrm{Multi}\{\mathbb{R}^m, \mathbb{R}[s]\}$. If there exists $\lambda \in \mathrm{Box}\{\delta^L, \delta^H\}$ such that $\mathrm{Val}(P(\lambda), r) \neq 0$, then there exists $\varepsilon \in \mathrm{Vert}\{\delta^L, \delta^H\}$ such that $\mathrm{Val}(P(\varepsilon), r) \neq 0$.

Proof Let $a: \mathbb{R}^m \to \mathbb{R}$ be the function such that $a(\delta) = \mathrm{Val}(P(\delta), r)$, and let $b: \mathbb{R}^m \to \mathbb{R}$ be the function such that $b(\delta) = 1$. It is true that $a, b \in \mathrm{Multi}\{\mathbb{R}^m, \mathbb{R}\}$, so Lemma 5.1 implies that there exist $\varepsilon_1, \varepsilon_2 \in \mathrm{Vert}\{\delta^L, \delta^H\}$ such that $\mathrm{Val}(P(\varepsilon_1), r) \leq \mathrm{Val}(P(\lambda), r) \leq \mathrm{Val}(P(\varepsilon_2), r)$. Since $\mathrm{Val}(P(\lambda), r) \neq 0$, $\mathrm{Val}(P(\varepsilon_1), r)$ or $\mathrm{Val}(P(\varepsilon_2), r)$ must be different from zero. □

Lemma 5.3 Let $r_0 \in \mathbb{R}$, $\delta^L, \delta^H \in \mathbb{R}^m$, and $\lambda \in \mathrm{Box}\{\delta^L, \delta^H\}$. If the following conditions

i) $P \in \mathrm{Multi}\{\mathbb{R}^m, \mathbb{R}[s]\}$

ii) $Val(P(\lambda),r_0) = 0$

iii) $Val(P(\delta),r) \neq 0$ for all $r > r_0$, $\delta \in Box\{\delta^L, \delta^H\}$

are satisfied, then there exist $\varepsilon^L, \varepsilon^H \in Vert\{\delta^L, \delta^H\}$ such that

1) $\lambda \in Box\{\varepsilon^L, \varepsilon^H\}$

2) $Val(P(\delta),r_0) = 0$ for all $\delta \in Box\{\varepsilon^L, \varepsilon^H\}$.

Proof First, it will be demonstrated that if $Val(P(\delta),r_0) = 0$ for all $\delta \in Box\{\rho^L, \rho^H\} \subset Box\{\delta^L, \delta^H\}$ then $Val(P(\delta),r_0) = 0$ for all $\delta \in Box\{L_i(\rho^L), H_i(\rho^H)\}$. To prove this statement, it will be shown that if σ is any element in $Box\{L_i(\rho^L),H_i(\rho^H)\}$, then $Val(P(\sigma),r_0) = 0$. If $\sigma \in Box\{\rho^L, \rho^H\}$, then clearly $Val(P(\sigma),r_0) = 0$. If $\sigma \notin Box\{\rho^L, \rho^H\}$, then define

$$\psi = ([\sigma]_1, \ldots, [\sigma]_{i-1}, x, [\sigma]_{i+1}, \ldots, [\sigma]_p)$$

where x is one of the elements in the interval $[\,[\rho^L]_i, [\rho^H]_i\,] \setminus \{\,[\delta^L]_i, [\delta^H]_i\,\}$. For the selected value of ψ, consider the closed bounded interval

$$V(r) = Val(P(Box\{L_i(\psi),H_i(\psi)\}),r).$$

Since $Box\{L_i(\psi), H_i(\psi)\} \subset Box\{\delta^L, \delta^H\}$, condition (iii) implies that V(r) does not contain the origin for all $r > r_0$. Condition (i) ensures that the end points of V(r) are continuous functions of r. These last two facts combine to show that $V(r_0)$ is contained in $(-\infty,0]$ or $[0,\infty)$. From the definition of ψ, it can be seen that $\psi \in Box\{\rho^L, \rho^H\}$, so $Val(P(\psi),r_0) = 0$ and hence $V(r_0)$ contains the origin. It follows that $\min V(r_0)$ or $\max V(r_0)$ equals zero. Using this fact and Lemma 5.1 (setting $a(\delta) = Val(P(\delta),r_0)$ and $b(\delta) = 1$) implies that

$$0 \in Val(P(Vert\{L_i(\psi),H_i(\psi)\}),r_0) = \{\, Val(P(L_i(\psi)),r_0), Val(P(H_i(\psi)),r_0)\,\}.$$

The function $Val(P(\delta),r_0)$ has been shown to equal zero on at least two of the three distinct points: $\psi, L_i(\psi), H_i(\psi) \in Box\{L_i(\psi), H_i(\psi)\}$. Because $Val(P(\delta),r_0)$ is an

affine function when restricted to this one dimensional set, it can only equal zero at two distinct points if it equals zero for all $\delta \in \text{Box}\{L_i(\psi), H_i(\psi)\}$. Since $\sigma \in \text{Box}\{L_i(\psi),H_i(\psi)\}$, the goal of showing that $\text{Val}(P(\sigma),r_0)$ has been achieved.

To prove the lemma, begin by noting that $\text{Val}(P(\delta),r_0) = 0$ for all $\delta \in \{\lambda\} = \text{Box}\{\lambda,\lambda\}$. Since the first step is true, the fact demonstrated above can be used inductively. This process shows that $\text{Val}(P(\delta),r_0) = 0$ for all

$$\delta \in \text{Box}\{\ L_p(\ \dots\ L_2(L_1(\lambda))\ \dots\ ),\ H_p(\ \dots\ H_2(H_1(\lambda))\ \dots\ )\ \}.$$

This set clearly contains λ, so it only remains to show that $\varepsilon^L = L_p(\ \dots\ L_2(L_1(\lambda))\ \dots\ )$ and $\varepsilon^H = H_p(\ \dots\ H_2(H_1(\lambda))\ \dots\ )$ are contained in $\text{Vert}\{\delta^L, \delta^H\}$. This final requirement follows from the Definition 5.1. □

Lemma 5.4 If the two conditions

i) $\mathfrak{N}, \mathfrak{D} \in \text{Multi}\{\mathbb{R}^m,\mathbb{R}[s]\}$

ii) the rational function $(s-r)/\mathfrak{D}(\delta)$ is analytic on the real interval $[r,\infty)$ for all $\delta \in \text{Box}\{\delta^L, \delta^H\}$

are satisfied, then

$$\min\left\{\lim_{s\to r}\frac{(s-r)\ \text{Val}(\mathfrak{N}(\delta),s)}{\text{Val}(\mathfrak{D}(\delta),s)} : \delta \in \text{Box}\{\delta^L;\delta^H\}\right\}$$

$$= \min\left\{\lim_{s\to r}\frac{(s-r)\ \text{Val}(\mathfrak{N}(\delta),s)}{\text{Val}(\mathfrak{D}(\delta),s)} : \delta \in \text{Vert}\{\delta^L;\delta^H\}\right\}$$

and

$$\max\left\{\lim_{s\to r}\frac{(s-r)\ \text{Val}(\mathfrak{N}(\delta),s)}{\text{Val}(\mathfrak{D}(\delta),s)} : \delta \in \text{Box}\{\delta^L;\delta^H\}\right\}$$

$$= \max\left\{\lim_{s\to r}\frac{(s-r)\ \text{Val}(\mathfrak{N}(\delta),s)}{\text{Val}(\mathfrak{D}(\delta),s)} : \delta \in \text{Vert}\{\delta^L;\delta^H\}\right\}.$$

Proof Let λ be an arbitrary element in $\text{Box}\{\delta^L,\delta^H\}$, and consider two cases for λ. First, consider the case when $\text{Val}(\mathfrak{D}(\lambda),r) = 0$. Lemma 5.3 implies that there exist $\varepsilon^L, \varepsilon^H \in \text{Vert}\{\delta^L, \delta^H\}$ such that $\lambda \in \text{Box}\{\varepsilon^L, \varepsilon^H\}$ and $\text{Val}(\mathfrak{D}(\delta),r) = 0$ for all $\delta \in \text{Box}\{\varepsilon^L, \varepsilon^H\}$. Using l'Hôpital's Rule, this implies that for all $\delta \in \text{Box}\{\varepsilon^L, \varepsilon^H\}$

$$\lim_{s\to r} \frac{(s-r)\ \text{Val}(\mathfrak{N}(\delta),s)}{\text{Val}(\mathfrak{D}(\delta),s)} = \frac{\text{Val}(\mathfrak{N}(\delta),r)}{\text{Val}\left(\frac{\partial \mathfrak{D}(\delta)}{\partial s},r\right)}.$$

Condition (ii) guarantees that

$$\text{Val}\left(\frac{\partial \mathfrak{D}(\delta)}{\partial s},r\right) \neq 0$$

for all $\delta \in \text{Box}\{\varepsilon^L, \varepsilon^H\}$, so Lemma 5.1 implies that the value of this limit for all $\delta \in \text{Box}\{\varepsilon^L, \varepsilon^H\}$ is bounded by its extreme values on the set $\text{Vert}\{\varepsilon^L, \varepsilon^H\}$. Combining this with the fact that $\text{Vert}\{\varepsilon^L, \varepsilon^H\}$ is a subset of $\text{Vert}\{\delta^L,\delta^H\}$ implies that the following equation is true

$$\min\left\{\lim_{s\to r} \frac{(s-r)\ \text{Val}(\mathfrak{N}(\delta),s)}{\text{Val}(\mathfrak{D}(\delta),s)} : \delta \in \text{Vert}\{\delta^L;\delta^H\}\right\} \le \lim_{s\to r} \frac{(s-r)\ \text{Val}(\mathfrak{N}(\lambda),s)}{\text{Val}(\mathfrak{D}(\lambda),s)} \le$$

$$\max\left\{\lim_{s\to r} \frac{(s-r)\ \text{Val}(\mathfrak{N}(\delta),s)}{\text{Val}(\mathfrak{D}(\delta),s)} : \delta \in \text{Vert}\{\delta^L;\delta^H\}\right\}. \qquad (2)$$

Now, consider the second case when $\text{Val}(\mathfrak{D}(\lambda),r) \neq 0$. Lemma 5.2 implies that there exists $\rho \in \text{Vert}\{\delta^L,\delta^H\}$ such that $\text{Val}(\mathfrak{D}(\rho),r) \neq 0$. Since the denominator is nonzero at these points,

$$\lim_{s\to r} \frac{(s-r)\ \text{Val}(\mathfrak{N}(\lambda),s)}{\text{Val}(\mathfrak{D}(\lambda),s)} = 0 = \lim_{s\to r} \frac{(s-r)\ \text{Val}(\mathfrak{N}(\rho),s)}{\text{Val}(\mathfrak{D}(\rho),s)}.$$

This shows that equation (2) is also valid for this second case. Since equation (2) is valid for arbitrary $\lambda \in \text{Box}\{\delta^L,\delta^H\}$, the lemma follows. □

Theorem 5.1 Let u be a right-handed, continuous-time, input signal such that

$$u(t) = \mathfrak{L}^{-1}\left\{ \frac{u_n(s)}{u_d(s)} \right\}$$

where $u_n(s)$, $u_d(s) \in \mathbb{R}[s]$. If the two conditions

i) $N, D \in \text{Multi}\{\mathbb{R}^m, \mathbb{R}[s]\}$

ii) the rational function $s/[D(\delta)u_d(s)]$ is analytic in the closed right half plane for all $\delta \in \text{Box}\{\delta^L, \delta^H\}$

are satisfied, then

1)
$$\min\left\{ \lim_{t\to\infty} \mathfrak{L}^{-1}\left\{ \frac{N(\delta)u_n(s)}{D(\delta)u_d(s)} \right\} : \delta \in \text{Box}\{\delta^L;\delta^H\} \right\}$$
$$= \min\left\{ \lim_{t\to\infty} \mathfrak{L}^{-1}\left\{ \frac{N(\delta)u_n(s)}{D(\delta)u_d(s)} \right\} : \delta \in \text{Vert}\{\delta^L;\delta^H\} \right\}$$

2)
$$\max\left\{ \lim_{t\to\infty} \mathfrak{L}^{-1}\left\{ \frac{N(\delta)u_n(s)}{D(\delta)u_d(s)} \right\} : \delta \in \text{Box}\{\delta^L;\delta^H\} \right\}$$
$$= \max\left\{ \lim_{t\to\infty} \mathfrak{L}^{-1}\left\{ \frac{N(\delta)u_n(s)}{D(\delta)u_d(s)} \right\} : \delta \in \text{Vert}\{\delta^L;\delta^H\} \right\}.$$

Proof Condition (ii) allows use of the Final Value Theorem which shows that

$$\lim_{t\to\infty} \mathfrak{L}^{-1}\left\{ \frac{N(\delta)u_n}{D(\delta)u_d} \right\} = \lim_{s\to 0}\left\{ \frac{s\ \text{Val}(N(\delta)u_n(s), s)}{\text{Val}(D(\delta)u_d(s), s)} \right\}. \tag{3}$$

Using equation (3), the proof can now be completed by invoking Lemma 5.4 with $\mathfrak{N}(\delta) = N(\delta)u_n(s)$, $\mathfrak{D}(\delta) = D(\delta)u_d(s)$, and $r = 0$. □

Theorem 5.2 Let u be a right-handed, discrete-time, input signal such that

$$u(k) = \mathcal{Z}^{-1}\left\{ \frac{u_n(s)}{u_d(s)} \right\}$$

where $u_n(s), u_d(s) \in \mathbb{R}[s]$. If the two conditions

i) $N, D \in \text{Multi}\{\mathbb{R}^m, \mathbb{R}[s]\}$

ii) the rational function $(s-1)/[D(\delta)u_d(s)]$ is analytic outside the open unit disk for all $\delta \in \text{Box}\{\delta^L, \delta^H\}$

are satisfied, then

1)
$$\min\left\{ \lim_{k\to\infty} \mathcal{Z}^{-1}\left\{ \frac{N(\delta)u_n(s)}{D(\delta)u_d(s)} \right\} : \delta \in \text{Box}\{\delta^L;\delta^H\} \right\}$$
$$= \min\left\{ \lim_{k\to\infty} \mathcal{Z}^{-1}\left\{ \frac{N(\delta)u_n(s)}{D(\delta)u_d(s)} \right\} : \delta \in \text{Vert}\{\delta^L;\delta^H\} \right\}$$

2)
$$\max\left\{ \lim_{k\to\infty} \mathcal{Z}^{-1}\left\{ \frac{N(\delta)u_n(s)}{D(\delta)u_d(s)} \right\} : \delta \in \text{Box}\{\delta^L;\delta^H\} \right\}$$
$$= \max\left\{ \lim_{k\to\infty} \mathcal{Z}^{-1}\left\{ \frac{N(\delta)u_n(s)}{D(\delta)u_d(s)} \right\} : \delta \in \text{Vert}\{\delta^L;\delta^H\} \right\}.$$

Proof Condition (ii) allows use of the Final Value Theorem which shows that

$$\lim_{k\to\infty} \mathcal{Z}^{-1}\left\{ \frac{N(\delta)u_n(s)}{D(\delta)u_d(s)} \right\} = \lim_{s\to 1}\left\{ \frac{(s-1)\ \text{Val}(\ N(\delta)u_n(s),\ s)}{\text{Val}(\ D(\delta)u_d(s),\ s)} \right\}. \qquad (4)$$

Using equation (4), the proof can now be completed by invoking Lemma 5.4 with $\mathcal{N}(\delta) = N(\delta)u_n(s)$, $\mathcal{D}(\delta) = D(\delta)u_d(s)$, and $r = 1$. □

Theorems 5.1 and 5.2 are the main results of this chapter. Both theorems are original contributions of the author [Bartlett, 1990c]. Application of the continuous-time version of these vertex results is illustrated by the following example.

Example 5.1 The following is the transfer function of a single axis of a computerized numerical control (CNC) positioning system [Uliana, 1985]. The system contains a DC motor which produces translational motion through gear reduction and a lead screw.

$$H(s) = \begin{bmatrix} \dfrac{[K_t K_g P] / 2\pi}{s^3(JL_a)+s^2(JR_a+BL_a)+s(BR_a+K_tK_e)} & \dfrac{[s(L_aK_gP)+(R_aK_gP)] / 2\pi}{s^3(JL_a)+s^2(JR_a+BL_a)+s(BR_a+K_tK_e)} \\ \dfrac{sJ+B}{s^2(JL_a)+s(JR_a+BL_a)+(BR_a+K_tK_e)} & \dfrac{K_e}{s^2(JL_a)+s(JR_a+BL_a)+(BR_a+K_tK_e)} \end{bmatrix}$$

Let the vector of parameters equal

$$\delta = (R_a, K_t, J, L_a, B, K_e, K_g, P).$$

The parameters are only known to lie in the interval family Box$\{\delta^L,\delta^H\}$ where

$$\delta^L = (0.7, 10, 0.05, 0.005, 0.01, 1.1, 0.16, 9.5)$$

$$\delta^H = (0.8, 11, 2, 0.015, 0.02, 1.2, 0.17, 10.5).$$

The inputs to the system $U = [V \;\; G]^T$ are the armature voltage and the torque loading. The outputs of the system $Y = [X \;\; I]^T$ are the tool position and the armature current.

Now, consider the response of the tool position X to the torque disturbance $G(t) = [e^{-t}\cos(t)]x(t)$ where x(t) is the unit step. In the transform relation from G to X, the numerator and denominator polynomials are given by the multiaffine mappings

$$N(\delta) = s(L_aK_gP)+(R_aK_gP)$$

$$D(\delta) = 2\pi\,[s^3(JL_a)+s^2(JR_a+DL_a)+s(DR_a+K_tK_e)].$$

The Laplace transform of the input G is given by the ratio of the two polynomials

$$u_n = s + 1$$

$$u_d = s^2 + s + 1$$

All the coefficients of the 2nd order polynomials u_d and $D(\delta)/s$ are positive, so both polynomials are G_H stable. This implies that $s/(D(\delta)u_d)$ is analytic in the closed right

half plane. All the conditions are satisfied, so Theorem 5.1 is applicable. By evaluating the steady state response at the vertices $Ver\{\delta^L,\delta^H\}$, the maximum and minimum possible steady state values of the tool position are found to be 0.0103 and 0.0064, respectively.

With a little algebraic manipulation, Theorem 5.1 can also be used to determine to response of a single output to multiple inputs. Furthermore, uncertainty in the inputs can also be handles. To see this, consider the response of the armature current to a unit step armature voltage and a step torque disturbance of amplitude $e \in [0.5, 2]$. The transform of the output equals

$$I(s) = \frac{K_t+sL_ab+R_ae}{s^3(JL_a)+s^2(JR_a+DL_a)+s(DR_a+K_tK_e)}.$$

To use Theorem 5.1, this problem will be converted to the standard form. To this end, the parameter vector will be expanded to include e. Let

$$\delta_* = (R_a, K_t, J, L_a, B, K_e, K_g, P, e).$$

$$\delta_*^L = (0.7, 10, 0.05, 0.005, 0.01, 1.1, 0.16, 9.5, 0.5)$$

$$\delta_*^H = (0.8, 11, 2, 0.015, 0.02, 1.2, 0.17, 10.5, 2).$$

The system will be treated as if it is single-input single-output with numerator and denominator polynomials given by the multiaffine mappings

$$N_*(\delta_*) = K_t+sL_ab+R_ae$$

$$D_*(\delta_*) = s^2(JL_a)+s(JR_a+DL_a)+(DR_a+K_tK_e).$$

The input will be treated as if it is a scalar unit step with $u_{*n} = 1$ and $u_{*d} = s$. The polynomial $(D_*(\delta_*)s)/s$ is G_H-stable for all $\delta_* \in Box\{\delta_*^L,\delta_*^H\}$, so Theorem 5.1 can be used to find the range of steady state values of

$$I(s) = \frac{N_*(\delta_*)}{D_*(\delta_*)s}$$

for $\delta_* \in \text{Box}\{\delta_*^L, \delta_*^H\}$. Using $\text{Vert}\{\delta_*^L, \delta_*^H\}$, it is found that the maximum possible steady state value of I is 1.0538 and the minimum is 0.8589.

5.3 A Simplified Steady State Vertex Theorem for Interval Families

The steady state analysis vertex theorem for systems with multiaffine uncertainties can also be used to analyze systems represented by interval families of polynomial-pairs. This section will show that it is actually sufficient to use only a few special vertices of an interval family in order to determine the extreme values of the steady state response. This simplification is given by the following theorems.

Theorem 5.3 Let

$$(a_r^x s^r + \ldots + a_3^x s^3 + a_2^x s^2 + a_1^x s + a_0^x,\ c_q^x s^q + \ldots + c_3^x s^3 + c_2^x s^2 + c_1^x s + c_0^x)$$

be an element in the interval family of real polynomial-pairs

$$\mathcal{T} = \{ (a_r s^r + \ldots + a_3 s^3 + a_2 s^2 + a_1 s + a_0,\ c_q s^q + \ldots + c_3 s^3 + c_2 s^2 + c_1 s + c_0) :$$
$$a_r^L \le a_r \le a_r^H,\ \ldots,\ a_2^L \le a_2 \le a_2^H,\ a_1^L \le a_1 \le a_1^H,\ a_0^L \le a_0 \le a_0^H,$$
$$c_r^L \le c_r \le c_r^H,\ \ldots,\ c_2^L \le c_2 \le c_2^H,\ c_1^L \le c_1 \le c_1^H,\ c_0^L \le c_0 \le c_0^H \},$$

and let u be a right-handed, continuous-time, input signal such that

$$u(t) = \mathcal{L}^{-1}\left\{ \frac{u_n(s)}{u_d(s)} \right\}$$

where $u_n(s), u_d(s) \in \mathbb{R}[s]$. If the rational function $s/[d(s)u_d(s)]$ is analytic in the closed right half plane for all $d(s) \in \text{Den}(\mathcal{T})$ then

1)
$$\min\left\{ \lim_{t\to\infty} \mathfrak{L}^{-1}\left\{ \frac{n(s)u_n(s)}{d(s)u_d(s)} \right\} : (n(s);d(s)) \in \mathbb{T} \right\}$$

$$= \min\left\{ \lim_{t\to\infty} \mathfrak{L}^{-1}\left\{ \frac{n(s)u_n(s)}{d(s)u_d(s)} \right\} : (n(s);d(s)) \in \mathbb{V} \right\}$$

2)
$$\max\left\{ \lim_{t\to\infty} \mathfrak{L}^{-1}\left\{ \frac{n(s)u_n(s)}{d(s)u_d(s)} \right\} : (n(s);d(s)) \in \mathbb{T} \right\}$$

$$= \max\left\{ \lim_{t\to\infty} \mathfrak{L}^{-1}\left\{ \frac{n(s)u_n(s)}{d(s)u_d(s)} \right\} : (n(s);d(s)) \in \mathbb{V} \right\}$$

where

$$\mathbb{V} = \{ (a_r{}^x s^r + \ldots + a_2{}^x s^2 + a_1{}^x s + a_0,\ c_q{}^x s^q + \ldots + c_3{}^x s^3 + c_2{}^x s^2 + c_1 s + c_0) :$$
$$a_0 \in \{a_0{}^L, a_0{}^H\}, \qquad c_1 \in \{c_1{}^L, c_1{}^H\}, \qquad c_0 \in \{c_0{}^L, c_0{}^H\} \}.$$

Proof The condition that $s/[d(s)u_d(s)]$ is analytic in the closed right half plane for all $d(s) \in \mathrm{Den}(\mathbb{T})$ allows use of the Final Value Theorem which shows that

$$\lim_{t\to\infty} \mathfrak{L}^{-1}\left\{ \frac{(a_r s^r + \ldots + a_3 s^3 + a_2 s^2 + a_1 s + a_0)\, u_n(s)}{(c_q s^q + \ldots + c_3 s^3 + c_2 s^2 + c_1 s + c_0)\, u_d(s)} \right\}$$

$$= \lim_{s\to 0}\left[\frac{s\,(a_r s^r + \ldots + a_3 s^3 + a_2 s^2 + a_1 s + a_0)\, u_n(s)}{(c_q s^q + \ldots + c_3 s^3 + c_2 s^2 + c_1 s + c_0)\, u_d(s)} \right]$$

$$= \lim_{s\to 0}\left[\frac{s\,(a_0)\, u_n(s)}{(c_1 s + c_0)\, u_d(s)} \right] \qquad (5)$$

for all

$$(a_r s^r + \ldots + a_3 s^3 + a_2 s^2 + a_1 s + a_0,\ c_q s^q + \ldots + c_3 s^3 + c_2 s^2 + c_1 s + c_0) \in \mathbb{T}.$$

Equation (5) implies that the steady state value does not directly depend on a_r, ... a_3, a_2, a_1, c_q, ... c_3, and c_2. In other words,

$$\left\{ \lim_{t\to\infty} \mathcal{Z}^{-1}\left\{ \frac{n(s)u_n(s)}{d(s)u_d(s)} \right\} : (n(s);d(s)) \in \mathbb{T} \right\}$$

$$= \left\{ \lim_{t\to\infty} \mathcal{Z}^{-1}\left\{ \frac{n(s)u_n(s)}{d(s)u_d(s)} \right\} : (n(s);d(s)) \in \mathbb{X} \right\} \qquad (6)$$

where

$$\mathbb{X} = \{ \; (a_r{}^x s^r + \ldots + a_2{}^x s^2 + a_1{}^x s + a_0,\ c_q{}^x s^q + \ldots + c_3{}^x s^3 + c_2{}^x s^2 + c_1 s + c_0) \; :$$
$$a_0 \in [a_0{}^L, a_0{}^H], \qquad c_1 \in [c_1{}^L, c_1{}^H], \qquad c_0 \in [c_0{}^L, c_0{}^H] \; \}.$$

Theorem 5.1 can be used to show that

$$\min\left\{ \lim_{t\to\infty} \mathcal{Z}^{-1}\left\{ \frac{n(s)u_n(s)}{d(s)u_d(s)} \right\} : (n(s);d(s)) \in \mathbb{X} \right\}$$
$$= \min\left\{ \lim_{t\to\infty} \mathcal{Z}^{-1}\left\{ \frac{n(s)u_n(s)}{d(s)u_d(s)} \right\} : (n(s);d(s)) \in \mathbb{V} \right\} \qquad (7)$$

$$\max\left\{ \lim_{t\to\infty} \mathcal{Z}^{-1}\left\{ \frac{n(s)u_n(s)}{d(s)u_d(s)} \right\} : (n(s);d(s)) \in \mathbb{X} \right\}$$
$$= \max\left\{ \lim_{t\to\infty} \mathcal{Z}^{-1}\left\{ \frac{n(s)u_n(s)}{d(s)u_d(s)} \right\} : (n(s);d(s)) \in \mathbb{V} \right\}. \qquad (8)$$

The proof now follows from the combining of equation (6-8). □

Theorem 5.4 Consider the interval family of real polynomial-pairs

$$\mathbb{T} = \{ \; (a_r s^r + \ldots + a_3 s^3 + a_2 s^2 + a_1 s + a_0,\ c_q s^q + \ldots + c_3 s^3 + c_2 s^2 + c_1 s + c_0) \; :$$
$$a_r{}^L \le a_r \le a_r{}^H, \ \ldots, \ a_2{}^L \le a_2 \le a_2{}^H, \ a_1{}^L \le a_1 \le a_1{}^H, \ a_0{}^L \le a_0 \le a_0{}^H,$$
$$c_r{}^L \le c_r \le c_r{}^H, \ \ldots, \ c_2{}^L \le c_2 \le c_2{}^H, \ c_1{}^L \le c_1 \le c_1{}^H, \ c_0{}^L \le c_0 \le c_0{}^H \; \}.$$

Let u be a right-handed, discrete-time, input signal such that

$$u(k) = \mathcal{Z}^{-1}\left\{ \frac{u_n(s)}{u_d(s)} \right\}$$

where $u_n(s)$, $u_d(s) \in \mathbb{R}[s]$. If the rational function $(s-1)/[d(s)u_d(s)]$ is analytic outside the open unit disk for all $d(s) \in \mathrm{Den}(\mathbb{T})$ then

1)
$$\min\left\{ \lim_{k\to\infty} \mathcal{Z}^{-1}\left\{ \frac{n(s)u_n(s)}{d(s)u_d(s)} \right\} : (n(s);d(s)) \in \mathbb{T} \right\}$$
$$= \min\left\{ \lim_{k\to\infty} \mathcal{Z}^{-1}\left\{ \frac{n(s)u_n(s)}{d(s)u_d(s)} \right\} : (n(s);d(s)) \in \mathbb{V} \right\}$$

2)
$$\max\left\{ \lim_{k\to\infty} \mathcal{Z}^{-1}\left\{ \frac{n(s)u_n(s)}{d(s)u_d(s)} \right\} : (n(s);d(s)) \in \mathbb{T} \right\}$$
$$= \max\left\{ \lim_{k\to\infty} \mathcal{Z}^{-1}\left\{ \frac{n(s)u_n(s)}{d(s)u_d(s)} \right\} : (n(s);d(s)) \in \mathbb{V} \right\}$$

where

$$\mathbb{V} = \{ (a_r{}^L s^r + \ldots + a_2{}^L s^2 + a_1{}^L s + a_0{}^L,\ c_q{}^L s^q + \ldots + c_2{}^L s^2 + c_1{}^L s + c_0{}^L),$$
$$(a_r{}^L s^r + \ldots + a_2{}^L s^2 + a_1{}^L s + a_0{}^L,\ c_q{}^H s^q + \ldots + c_2{}^H s^2 + c_1{}^H s + c_0{}^H),$$
$$(a_r{}^H s^r + \ldots + a_2{}^H s^2 + a_1{}^H s + a_0{}^H,\ c_q{}^L s^q + \ldots + c_2{}^L s^2 + c_1{}^L s + c_0{}^L),$$
$$(a_r{}^H s^r + \ldots + a_2{}^H s^2 + a_1{}^H s + a_0{}^H,\ c_q{}^H s^q + \ldots + c_2{}^H s^2 + c_1{}^H s + c_0{}^H) \}.$$

Proof The condition that $(s-1)/[d(s)u_d(s)]$ is analytic outside the open unit circle for all $d(s) \in \mathrm{Den}(\mathbb{T})$ allows use of the Final Value Theorem which shows that

$$\lim_{k\to\infty} \mathcal{Z}^{-1}\left\{ \frac{(a_r s^r + \ldots + a_3 s^3 + a_2 s^2 + a_1 s + a_0)\, u_n(s)}{(c_q s^q + \ldots + c_3 s^3 + c_2 s^2 + c_1 s + c_0)\, u_d(s)} \right\}$$
$$= \lim_{s\to 1}\left[\frac{(s-1)\,(a_r s^r + \ldots + a_3 s^3 + a_2 s^2 + a_1 s + a_0)\, u_n(s)}{(c_q s^q + \ldots + c_3 s^3 + c_2 s^2 + c_1 s + c_0)\, u_d(s)} \right]$$
$$= \left(\sum_{i=0}^{r} a_i \right) u_n(1) \lim_{s\to 1}\left[\frac{(s-1)}{(c_q s^q + \ldots + c_3 s^3 + c_2 s^2 + c_1 s + c_0)\, u_d(s)} \right] \quad (9)$$

for all

$$(a_r s^r + \ldots + a_3 s^3 + a_2 s^2 + a_1 s + a_0,\ c_q s^q + \ldots + c_3 s^3 + c_2 s^2 + c_1 s + c_0) \in \mathbb{T}.$$

From the inequality

$$\left(\sum_{i=0}^{r} a_i^L\right) \le \left(\sum_{i=0}^{r} a_i\right) \le \left(\sum_{i=0}^{r} a_i^H\right),$$

it becomes clear that equation (9) can only take on an extreme value when

$$a_r s^r + \ldots + a_3 s^3 + a_2 s^2 + a_1 s + a_0 \in \mathrm{Num}(\mathbb{V})$$

$$= \left\{ \; a_r^L s^r + \ldots + a_2^L s^2 + a_1^L s + a_0^L, \;\; a_r^H s^r + \ldots + a_2^H s^2 + a_1^H s + a_0^H \; \right\}.$$

This implies that

$$\min\left\{ \lim_{k\to\infty} \mathcal{Z}^{-1}\left\{ \frac{n(s)u_n(s)}{d(s)u_d(s)} \right\} : (n(s);d(s)) \in \mathbb{T} \right\}$$

$$= \min\left\{ \lim_{k\to\infty} \mathcal{Z}^{-1}\left\{ \frac{n(s)u_n(s)}{d(s)u_d(s)} \right\} : (n(s);d(s)) \in \mathrm{Num}(\mathbb{V}) \times \mathrm{Den}(\mathbb{T}) \right\} \qquad (10)$$

$$\max\left\{ \lim_{k\to\infty} \mathcal{Z}^{-1}\left\{ \frac{n(s)u_n(s)}{d(s)u_d(s)} \right\} : (n(s);d(s)) \in \mathbb{T} \right\}$$

$$= \max\left\{ \lim_{k\to\infty} \mathcal{Z}^{-1}\left\{ \frac{n(s)u_n(s)}{d(s)u_d(s)} \right\} : (n(s);d(s)) \in \mathrm{Num}(\mathbb{V}) \times \mathrm{Den}(\mathbb{T}) \right\} \qquad (11)$$

To complete the proof, it must be shown that Den($\mathbb{T}$) can be replaced by Den($\mathbb{V}$) in equations (10) and (11). This will be achieved by considering two separate cases.

First, consider the case when $u_d(1) = 0$. For this case, the condition that $(s-1)/[d(s)u_d(s)]$ is analytic outside the open unit circle for all $d(s) \in \mathrm{Den}(\mathbb{T})$ implies that

$$\left[\frac{\partial u_d(s)}{\partial s}\right]_{s=1} = u_d'(1) \neq 0$$

$$d(1) = \sum_{i=0}^{q} c_i \neq 0 \qquad (12)$$

for all $d(s) = (c_q s^q + \ldots + c_3 s^3 + c_2 s^2 + c_1 s + c_0) \in \mathrm{Den}(\mathbb{T})$. From these facts, equation (9) can be simplified to

$$\lim_{k\to\infty} \mathcal{Z}^{-1}\left\{ \frac{(a_r s^r + \ldots + a_3 s^3 + a_2 s^2 + a_1 s + a_0)\, u_n(s)}{(c_q s^q + \ldots + c_3 s^3 + c_2 s^2 + c_1 s + c_0)\, u_d(s)} \right\}$$

$$= \left(\sum_{i=0}^{r} a_i\right)\left(\sum_{i=0}^{r} c_i\right)^{-1}\left(\frac{u_n(1)}{u_d'(1)}\right) \qquad (13)$$

for all

$$(a_r s^r + \ldots + a_3 s^3 + a_2 s^2 + a_1 s + a_0,\ c_q s^q + \ldots + c_3 s^3 + c_2 s^2 + c_1 s + c_0) \in \mathbb{T}.$$

From the inequality

$$\left(\sum_{i=0}^{r} c_i^L\right) \le \left(\sum_{i=0}^{r} c_i\right) \le \left(\sum_{i=0}^{r} c_i^H\right)$$

combined with equation (12), it follows that

$$\left(\sum_{i=0}^{r} c_i^L\right)^{-1} \ge \left(\sum_{i=0}^{r} c_i\right)^{-1} \ge \left(\sum_{i=0}^{r} c_i^H\right)^{-1}.$$

From this latter inequality, it becomes clear that equation (13) can only take on an extreme value when

$$c_r s^r + \ldots + c_3 s^3 + c_2 s^2 + c_1 s + c_0 \in \mathrm{Den}(\mathbb{V})$$

$$= \left\{ \ c_r^L s^r + \ldots + c_2^L s^2 + c_1^L s + c_0^L,\ c_r^H s^r + \ldots + c_2^H s^2 + c_1^H s + c_0^H \ \right\}.$$

This achieves the goal of showing that Den($\mathbb{T}$) can be replaced by Den($\mathbb{V}$) in equations (10) and (11) for the case when $u_d(1) = 0$.

Now, consider the second case when $u_d(1) \neq 0$. For this case, Den($\mathbb{T}$) and Den($\mathbb{V}$) will both be divided into two sets

$$\mathbb{D}_{\neq T} = \{\ d(s) \in \mathrm{Den}(\mathbb{T}) \ :\ d(1) \neq 0\ \}$$

$$\mathbb{D}_{=T} = \{\ d(s) \in \mathrm{Den}(\mathbb{T}) \ :\ d(1) = 0\ \}$$

$$\mathbb{D}_{\neq V} = \{\ d(s) \in \mathrm{Den}(\mathbb{V}) \ :\ d(1) \neq 0\ \}$$

$$\mathbb{D}_{=V} = \{\ d(s) \in \mathrm{Den}(\mathbb{V}) \ :\ d(1) = 0\ \}.$$

For all real values of x greater than or equal to zero, the following inequality holds

$$\left[c_r{}^L s^r + \ldots + c_2{}^L s^2 + c_1{}^L s + c_0{}^L \right]_{s=x}$$
$$\leq \left[c_r s^r + \ldots + c_2 s^2 + c_1 s + c_0 \right]_{s=x}$$
$$\leq \left[c_r{}^H s^r + \ldots + c_2{}^H s^2 + c_1{}^H s + c_0{}^H \right]_{s=x}. \qquad (14)$$

For x=1, this inequality implies that $\mathbb{D}_{\neq T}$ is nonempty if and only if $\mathbb{D}_{\neq V}$ is nonempty. This combined with the fact that equation (9) equals zero for all polynomial-pairs in $\text{Num}(\mathbb{V}) \times \mathbb{D}_{\neq T}$ and $\text{Num}(\mathbb{V}) \times \mathbb{D}_{\neq V}$ implies that

$$\left\{ \lim_{k\to\infty} \mathcal{Z}^{-1}\left\{ \frac{n(s)u_n(s)}{d(s)u_d(s)} \right\} : (n(s);d(s)) \in \text{Num}(\mathbb{V}) \times \mathbb{D}_{\neq T} \right\}$$

$$= \left\{ \lim_{k\to\infty} \mathcal{Z}^{-1}\left\{ \frac{n(s)u_n(s)}{d(s)u_d(s)} \right\} : (n(s);d(s)) \in \text{Num}(\mathbb{V}) \times \mathbb{D}_{\neq V} \right\}. \qquad (15)$$

The condition that $(s-1)/[d(s)u_d(s)]$ is analytic outside the open unit circle for all $d(s) \in \text{Den}(\mathbb{T})$ implies that $c_r{}^L$ and $c_r{}^H$ are both positive or both negative and it implies that

$$\left[c_r s^r + \ldots + c_2 s^2 + c_1 s + c_0 \right]_{s=x} \neq 0$$

for all $x > 1$, $d(s) \in \text{Den}(\mathbb{T})$. *These together indicate that*

$$\left[c_r{}^L s^r + \ldots + c_2{}^L s^2 + c_1{}^L s + c_0{}^L \right]_{s=x}$$
$$\left[c_r{}^H s^r + \ldots + c_2{}^H s^2 + c_1{}^H s + c_0{}^H \right]_{s=x}$$

are both positive or both negative for all $x > 1$. Continuity and equation (14) imply that $d(s) \in \text{Den}(\mathbb{T})$ has a root at 1 only if $d(s)$ equals

$$c_r{}^L s^r + \ldots + c_2{}^L s^2 + c_1{}^L s + c_0{}^L$$

or

$$c_r{}^H s^r + \ldots + c_2{}^H s^2 + c_1{}^H s + c_0{}^H.$$

This implies that $\mathbb{D}_{=T}$ equals $\mathbb{D}_{=V}$, so

$$\left\{ \lim_{k\to\infty} \mathcal{Z}^{-1}\left\{ \frac{n(s)u_n(s)}{d(s)u_d(s)} \right\} : (n(s);d(s)) \in \text{Num}(\mathbb{V}) \times \mathbb{D}_{=T} \right\}$$

$$= \left\{ \lim_{k\to\infty} \mathcal{Z}^{-1}\left\{ \frac{n(s)u_n(s)}{d(s)u_d(s)} \right\} : (n(s);d(s)) \in \text{Num}(\mathbb{V}) \times \mathbb{D}_{=V} \right\}. \qquad (16)$$

$\text{Den}(\mathbb{T}) = \mathbb{D}_{\neq T} \cup \mathbb{D}_{=T}$ and $\text{Den}(\mathbb{V}) = \mathbb{D}_{\neq V} \cup \mathbb{D}_{=V}$, so equations (15) and (16) show that $\text{Den}(\mathbb{T})$ can be replaced by $\text{Den}(\mathbb{V})$ in equations (10) and (11) for the case when $u_d(1) \neq 0$. This completes the second case and the proof. □

This section has shown that the steady state response of stable systems represented by interval families of polynomial-pairs can be determined using only a small number of vertices. For the continuous-time case, eight vertices are needed, and for the discrete-time case only four vertices are required.

5.4 Counterexamples to a Transient Response Vertex Conjecture

The steady state vertex results presented in the Section 5.2 are cause to speculate that a transient response vertex result might exist for stable systems with multiaffine uncertainties. This section will show that the crucial features of the transient response cannot be determined from the vertices alone even when only affine uncertainties are allowed. This negative answer to the conjecture is demonstrated by examples in both continuous-time and discrete-time.

Example 5.2 Consider a continuous-time system whose polynomial-pair description is given by the affine mapping $T: \mathbb{R} \rightarrow \mathbb{R}[s]^2$ defined by

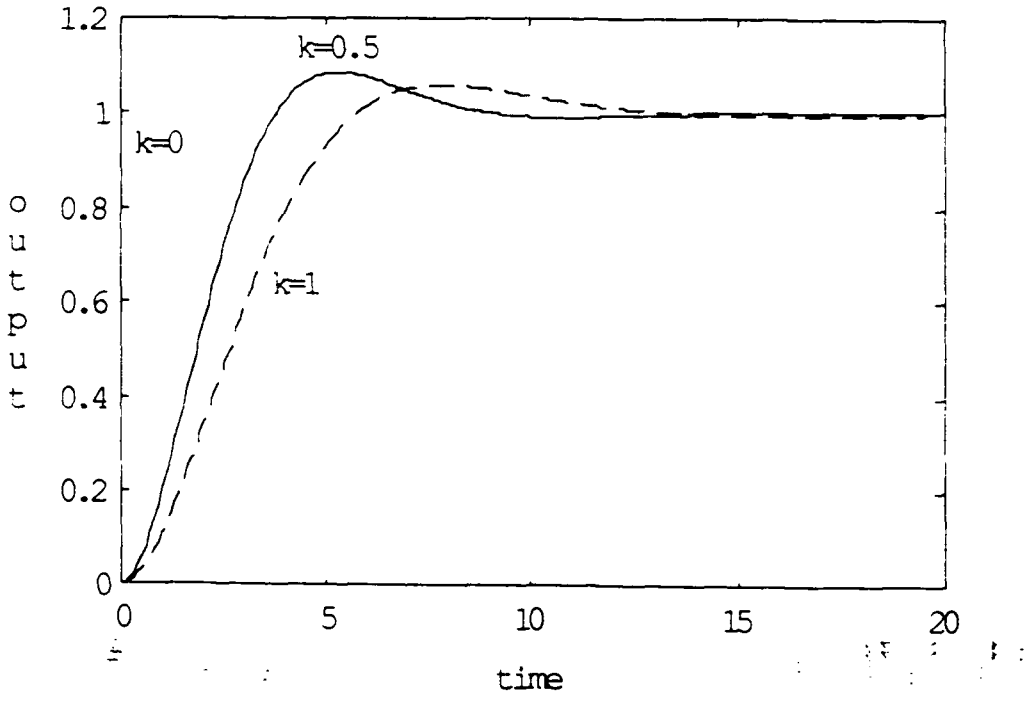
$$T(k) = (1,(3.4\ k+0.1)s^2+(1.7\ k+0.8)s+1).$$

The value of the parameter k is only known to lie in the set

$$\Delta = [0,1].$$

This system is G_H-stable for all $k \in \Delta$. The continuous-time step response of this system when k takes on the vertex values 0 and 1 is shown in Figure 5.1. The step response for the non-vertex value k = 0.5 is also shown. Figure 5.1 shows that the

maximum possible value of the peak overshoot does not occur at a vertex. This implies that, in general, the vital statistics of the transient response of a continuous-time stable system with affine uncertainties cannot be determined from the response of the vertex descriptions.



Continuous-time example showing that maximum possible value of the step response does not necessarily occur at a vertex for stable systems with affine uncertainties.

Figure 5.1

A similar statement can be made for discrete-time systems.

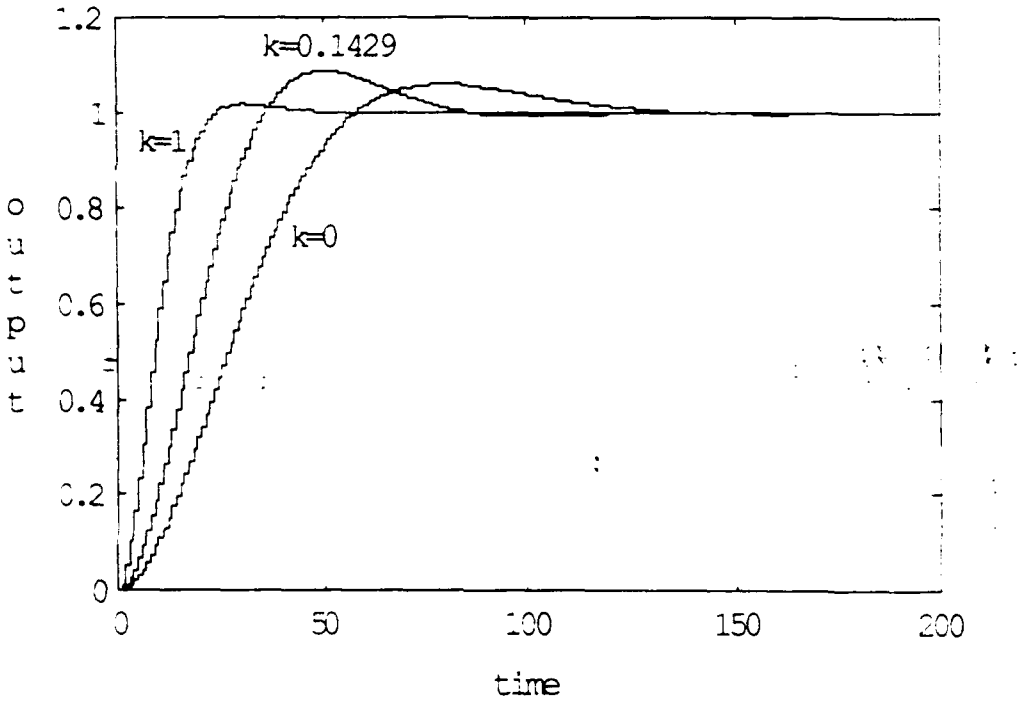
Example 5.3 Consider a discrete-time system whose polynomial-pair description is given by the affine mapping $T: \mathbb{R} \rightarrow \mathbb{R}[s]^2$ defined by

$$T(s,\delta) = \Big(\; (0.0130\,k + 0.0014)s + (0.0099\,k + 0.0014),$$
$$s^2 + (0.1915\,k - 1.9283)s + (-0.1686\,k + 0.9311) \; \Big)$$

The value of the parameter k is only known to lie in the set

$$\Delta = [0,1].$$

This system is G_S-stable for all $k \in \Delta$. The discrete-time step responses of this system for k equal to 0, 0.1429, and 1 are shown in Figure 5.2. Clearly, the maximum peak overshoot does not occur at a vertex. This implies that the vital statistics of the transient response of a discrete-time uncertain system cannot be completely determined from the response of the vertex descriptions.



Discrete-time example showing that maximum possible value of the step response does not necessarily occur at a vertex for stable systems with affine uncertainties.

Figure 5.2

This section has shown that vertices are not sufficient for a complete transient response analysis of a stable system whose set of polynomial-pair descriptions is generated by affine uncertainties. Whether or not some larger subset such as edges or faces would provide all the desired information is still an open question. Another open question is whether a vertex result exists for more restrictive uncertainty classes such as interval families.

5.5 Conclusion

This chapter presented new vertex theorems applicable to the steady state analysis of both continuous-time and discrete-time stable uncertain systems. For sets of polynomial-pairs generated by multiaffine uncertainties, it has been shown that an exact worst-case analysis can be carried out using only the vertices. This automatically implies a similar vertex theorem for sets of polynomial-pairs generated by affine uncertainties. With a little extra work, the multiaffine vertex result could be used to prove a steady state vertex result for polytopes of polynomial-pairs. In addition, it was shown that a steady state analysis could be carried out for discrete-time and continuous-time interval families of polynomial-pairs using only four vertices and eight vertices, respectively.

This chapter also showed that the vertex results for polynomial-pair sets generated by affine uncertainties were only valid for steady state analysis and not for a complete transient response analysis. Both continuous-time and discrete-time counter-examples were presented. These examples showed that the maximum peak overshoot of a stable system with affine uncertainties does not necessarily occur at a vertex. This suggests two possible directions for future research. One is to look for transient

response vertex results for more restrictive uncertainty classes such as interval families. The other direction is to investigate the possibility of transient response edge results.

CHAPTER 6

CONCLUSION

This dissertation has reviewed the existence of vertex and edge theorems for the analysis of both continuous and discrete-time, finite-dimensional, linear, time-invariant systems with uncertain parameters. The focus has been on four classes of uncertain systems and several types of classical analyses. The four classifications of uncertain systems are loosely called interval families, affine uncertainties, polytopes, and multiaffine uncertainties. The analyses are pole locations, stability, frequency response, and time response. Several vertex and edge theorems as well as counterexamples to similar conjectures have been presented. A few open questions have also been pointed out. Several of the results concerning these topics are joint or independent original contributions of the author.

Determination of pole locations was the first analysis topic that was considered. It was shown that all possible poles of an uncertain system represented by a polytope of characteristic polynomials could be determined using only the edge characteristic polynomials. This result is one of the main joint contributions of the author [Bartlett, Hollot, Huang, 1988; Hollot, Looze, Bartlett, 1990; Bartlett, 1990a]. This result automatically provides a pole location edge theorem for systems represented by sets of polynomials generated by affine uncertainties and by interval families of polynomials. For systems represented by sets of characteristic polynomials generated by multiaffine uncertainties, it was shown that the edges polynomials do not provide sufficient information to easily determine all possible pole locations. This fact follows almost immediately from [Barmish, Fu, Saleh, 1988] and is explicitly shown by [Ackermann, Hu, Kaesbauer, 1990]. For all four classes of polynomial sets, it is easily seen from

traditional root locus examples that the vertices do not provide sufficient information to easily determine all possible system poles. These results on pole locations naturally suggest results concerning stability analyses.

Stability analysis was the second topic that was covered. It was pointed out that under certain conditions the G-stability of all polynomials in a polytope would be implied by the G-stable of all edge polynomials. This edge theorem is valid if the polytope and the stability region both satisfy certain assumptions [Bartlett, Hollot, Huang, 1988; Fu, Barmish, 1988; Sideris, Barmish, 1989; Hollot, Looze, Bartlett, 1990] or alternately if a precondition is used [Bartlett, 1990a]. The author is a joint contributor to the original version of this edge theorem and to some of the subsequent revisions. Unfortunately, this edge theorem does not extend to multiaffine uncertainties [Barmish, Fu, Saleh, 1988]. This is true even for the important special cases when G is to equal G_H or G_S and when all the possible characteristic polynomials are real. Besides edge results, vertex theorems were also discussed. A counterexample similar to that of [Kochenburger, 1953] was given which showed that G_H-stability of a set of real polynomial generated by affine uncertainties is not implied by G_H-stability of its vertex polynomials. From this example, it is straight forward to generate a counterexample to a similar vertex conjecture concerning G_S-stability. It follows that vertex stability theorems also do not exist for the more general cases of polytopes of polynomials or for polynomial sets generated by affine uncertainties. For the more restrictive case of interval families of polynomials, it is also true that G_S-stability of the vertices does not imply G_S-stability of the whole family [Hollot, Bartlett, 1986]. However, for some special stability regions including G_H, it was pointed that powerful vertex theorems exist for the stability analysis of interval families of polynomials [Kharitonov, 1978a&b; Petersen, 1989; Fu, 1989]. The result for G_H-stability is

particularly amazing because it only requires that the stability of eight special vertices be tested [Kharitonov, 1978a&b].

Frequency response analysis was the third topic that was considered. Given a set of polynomial-pair descriptions $\mathbb{T}$ for an uncertain system and a frequency z, the goal was to determine all possible responses in Nyquist plane form Nyq($\mathbb{T}$,z) or in Bode magnitude form Mag($\mathbb{T}$,z). It was shown that, for sets generated by multiaffine uncertainties, neither of the two frequency response sets could be determined using only the frequency responses of the edge polynomial-pairs. This was true even if z was restricted to the unit circle or to the jw-axis. For sets of polynomial-pairs that are polytopes or are generated by affine uncertainties, it was shown that an edge theorem could be used to determine the two frequency response sets. This edge theorem is an independent original contribution of the author [Bartlett, 1990b]. In addition, it was shown that a vertex theorem for determining Nyq($\mathbb{T}$,z) or Mag($\mathbb{T}$,z) does not exist even for the special case when z is restricted to the jw-axis or to the unit circle. The frequency response edge theorem for polytopes obviously applies to the special case of interval families. It is also generally true that, for interval families of polynomial-pairs, no vertex theorem can be used to determine the frequency response sets. However, the analysis of interval families can be greatly simplified if attention is restricted to frequencies on the jw-axis. This restriction is too limiting for the analysis of discrete-time systems, but for continuous-time systems, this restriction is of essentially no consequence. One of the simplifying results shows that an interval family of polynomial-pairs and a polytopic subset of it have identical frequency response sets. By combining this fact and the edge theorem for polytopes, it was shown that the frequency response sets of an interval family could be determined using only the edges (32 or less) and the vertices (16 or less) of this special polytopic subset. A further simplify result, showed that, for interval families, the extreme values of the frequency

AD-A-243935

PAGE
140
MISSING
FROM ORIGINAL
DOCUMENT
AS SENT TO
DTIC
FROM THE
ORIGINATOR

This dissertation has reviewed the availability of vertex and edge theorems which simplify pole location analyses, stability analyses, frequency response analyses, and time response analyses for four classes of uncertain systems. Among the contributions of the author that were included in this review, three are the most significant. The first of these main contributions is the collection of edge theorems for determining pole locations and stability of polytopes of polynomials [Bartlett, Hollot, Huang, 1988; Fu, Barmish, 1988; Sideris, Barmish, 1989; Hollot, Looze, Bartlett, 1990; Bartlett, 1990a]. The second contribution is the edge theorem for determining the frequency response of polytopes of polynomial-pairs [Bartlett, 1990b]. The third of these main results is the vertex theorem for determining the steady state time response of polynomial-pair sets generated by multiaffine uncertainties [Bartlett, 1990c]. As was intended, all three of these contribution can be used to significantly reduce the effort required to carry out worst case analyses on systems with certain general types of parameter uncertainties.

BIBLIOGRAPHY

Ackermann, J. E., 1980, "Parameter Space Design of Robust Control Systems," IEEE Transactions on Automatic Control, vol. 25, no. 6, pp. 1058-1072.

Ackermann, J. E., 1989, "Does It Suffice to Check a Subset of Multilinear Parameters in Robustness Analysis?," submitted to the Proceedings of the 1990 Conference on Decision and Control.

Ackermann, J. E. and Barmish, B. R., 1988, "Robust Schur Stability of a Polytope of Polynomials," IEEE Transactions on Automatic Control, vol. 33, no. 10, pp. 984-986.

Ackermann, J. E., Hu, H. Z., and Kaesbauer, D., 1990, "Robustness Analysis: A Case Study," IEEE Transactions on Automatic Control, vol. 35, no. 3, pp. 352-356.

Anderson, B. D. O. and Scott, R. W., 1977, "Output Feedback Stabilization–Solution by Algebraic Geometry Methods," Proceedings of the IEEE, vol. 65, no. 6, pp. 849-861.

Astrom, K. J. and Wittenmark, B., 1984, Computer Controlled Systems, Prentice-Hall, Inc., Englewood Cliffs, NJ.

Bailey, F. N., and Hui, C. H., 1989, "A Fast Algorithm for Computing Parametric Rational Functions," IEEE Transaction on Automatic Control, vol. 34, no. 11, pp. 1209-1212.

Bailey, F. N., and Panzer, D., 1988, "A Fast Algorithm for Computing Interval Rational Functions," Proceedings of the 1988 American Control Conference.

Bailey, F. N., Panzer, D., and Gu, G., 1988, "Two Algorithms for Frequency Domain Design of Robust Control Systems," International Journal of Control, vol. 48, no. 5, pp. 1787-1806.

Barmish, B. R., 1988a, private communication.

Barmish, B. R., 1988b, "New Tools for Robustness Analysis," Proceedings of the 1988 Conference on Decision and Control.

Barmish, B. R., 1989, "A Generalization of Kharitonov's Four-Polynomial Concept for Robust Stability Problems with Linearly Dependent Coefficient Perturbations," IEEE Transactions on Automatic Control, vol. 34, no. 2, pp. 157-165.

Barmish, B. R., Fu, M., and Saleh, S., 1988, "Stability of a Polytope of Matrices: Counterexamples," IEEE Transactions on Automatic Control, vol. 33, no. 6, pp. 569-572.

Barnett, S., 1984, Matrices in Control Theory, Robert E. Krieger Publishing Co., Malabar, Florida.

Bartlett, A. C., 1989, "Analysis and Design Methods for the Control of Uncertain Linear Systems," A Dissertation Proposal, University of Massachusetts, Amherst, November.

Bartlett, A. C., 1990a, "A Precondition for the Edge Theorem," Proceedings of the 1990 American Control Conference.

Bartlett, A. C., 1990b, "Nyquist, Bode, and Nichols Plots of Uncertain Systems," Proceedings of the 1990 American Control Conference.

Bartlett, A. C., 1990c, "Vertex Results for the Steady State Analysis of Uncertain Systems," to appear in the Proceedings of the 1990 Conference on Decision and Control.

Bartlett, A. C. and Hollot, C. V., 1988, "A Necessary and Sufficient Condition for Schur Invariance and Generalized Stability of Polytopes of Polynomials," IEEE Transactions on Automatic Control, vol. 33, no. 6, pp. 575-578.

Bartlett, A. C., Hollot, C. V., and Huang Lin, 1988, "Root Locations of an Entire Polytope of Polynomials: It Suffices to Check the Edges," Mathematics of Control, Signals, and Systems, vol. 1, no. 1, pp. 61-71.

Bialas, S., 1985, "A Necessary and Sufficient Condition for the Stability of Convex Combinations of Stable Polynomials or Matrices," Bulletin of the Polish Academy of Sciences: Technical Sciences, vol. 33, no. 9-10, pp. 472-480.

Bialas, S. and Garloff, J., 1985, "Convex Combinations of Stable Polynomials," Journal of the Franklin Institute, vol. 319, pp. 373-377.

Bickart, T. A. and Jury, E. I., 1978, "Real Polynomials: Nonnegativity and Positivity," IEEE Transactions on Circuits and Systems, vol. 25, no. 9, pp. 676-684.

Bose, N. K., 1985, "A System-Theoretic Approach to Stability of Sets of Polynomials," in *Linear Algebra and Its Role in Systems Theory* (Contemporary Mathematics, vol. 47), Providence, RI: American Math. Soc., pp 25-34.

Bose, N. K., 1989, "Tests for Hurwitz and Schur Properties of Convex Combinations of Complex Polynomials," IEEE Transactions on Circuits and Systems, vol. 36, no. 9, pp.1245-1247.

Bose, N. K. and Kim, K. D., 1989, "Boundary Implications for Frequency Response of Interval FIR Filters," Proceedings of the 1989 Conference on Decision and Control.

Bose, N. K. and Shi, Y. Q., 1987, "A Simple General Proof of Kharitonov's Generalized Stability Criterion," IEEE Transactions on Circuits and Systems, vol. 34, no. 8, pp. 1233-1237.

Brønsted, A., 1983, An Introduction to Convex Polytopes, Springer-Verlag, New York.

Chapellat, H. and Bhattacharyya, S. P., 1989, "A Generalization of Kharitonov's Theorem: Robust Stability of Interval Plants," IEEE Transactions on Automatic Control, vol. 34, no. 3, pp. 306-312.

Churchill, R. V. and Brown, J. W., 1984, Complex Variables and Applications, McGraw-Hill, New York.

Cieslik, J., 1987, "On Possibilities of the Extension of Kharitonov's Stability Test for Interval Polynomials to the Discrete-Time Case," IEEE Transactions on Automatic Control, vol. 32, pp. 237-238, March.

Dasgupta, S., 1988, "Kharitonov's Theorem Revisited," Systems & Control Letters, vol. 11, pp. 381-384.

Dasgupta, S. and Anderson, B. D. O., 1987, "Physically Based Parameterizations of Designing Adaptive Algorithms," Automatica, vol. 23, no. 4, pp. 469-477.

de Gaston, R. R. E. and Safonov, M. G., 1988, "Exact Calculation of the Multiloop Stability Margin," IEEE Transactions on Automatic Control, vol. 33, no. 2, pp 156-171.

Djaferis, T. E., and Hollot, C. V., 1989, "The Stabiltiy of a Family of Polynomials Can Be Deduced from a Finite Number $O(k^3)$ of Frequency Checks," IEEE Transactions on Automatic Control, vol. 34, no. 9, pp. 982-986.

Evans, W. R., 1948, "Graphical Analysis of Control Systems," AIEE Transactions, vol. 67, pp. 547-551.

Fam, A. T. and Meditch, J. S., 1978, "A Canonical Parameter Space for Linear System Design," IEEE Transactions on Automatic Control, vol. 23, no. 3, pp. 454-458.

Frazer, R. A. and Duncan, W. J., 1929, "On the Criteria for the Stability of Small Motions," Proceedings of the Royal Society, A, vol. 124, pp. 642-654.

Fu, M., 1989a, "A Class of Kharitonov Regions for Robust Stability of Dynamic Systems, Proceedings of the 1989 American Control Conference.

Fu, M., 1989b, "Computing the Frequency Response of Linear Systems with Parametric Perturbations," Technical Report EE8972, Dept. of Elect. Eng. and Comp. Science, University of Newcastle, N. S. W. 2308, Australia, December.

Fu, M. and Barmish, B. R., 1988, "Maximal Unidirectional Perturbation Bounds for Stability of Polynomials and Matrices," Systems & Control Letters, vol. 11, pp. 173-179.

Fu, M. and Barmish, B. R., 1989, "Polytopes of Polynomials with Zeros in a Prescribed Set," IEEE Transactions on Automatic Control, vol. 34, no. 5, pp. 544-546.

Fu, M., Olbrot, A. W., and Polis, M. P., 1989, "Robust Stability for Time-Delay Systems: The Edge Theorem and Graphical Tests," IEEE Transactions on Automatic Control, vol. 34, no. 8, pp. 813-820.

Gantmacher, F. R., 1959, The Theory of Matrices, Volume I and II, Chelsea Publishing Company, New York.

Gosh, B. K., 1985, "Some New Results on the Simultaneous Stabilizability of a Family of Single Input, Single Output Systems," Systems & Control Letters, vol. 6, pp. 39-45.

Grunbaum, B., 1967, Convex Polytopes, Interscience Publishers, London.

Hollot, C. V., 1989, "Kharitonov-Like Results in the Space of Markov Parameters," IEEE Transactions on Automatic Control, vol. 34, no. 5, pp. 536-538.

Hollot, C. V. and Bartlett, A. C., 1986, "Some Discrete-time Counterparts to Kharitonov's Stability Criterion for Uncertain Systems," IEEE Transactions on Automatic Control, vol. 31, no. 4, pp. 355-357.

Hollot, C. V. and Bartlett, A. C., 1987, "On the Eigenvalues of Interval Matrices," Proceedings of the 1987 Conference on Decision and Control.

Hollot, C. V., Looze, D. P., and Bartlett, A. C., 1987, "Unmodeled Dynamics: Performance and Stability via Parameter Space Methods," Proceedings of the 1987 Conference on Decision and Control.

Hollot, C. V., Looze, D. P., and Bartlett, A. C., 1990, "Parametric Uncertainty and Unmodeled Dynamics: Analysis via Parameter Space Methods," to appear in Automatica.

Hollot, C. V., and Yang, F., 1989, "Robust Stabilization of Interval Plants Using Load or Lag Compensators," Proceedings of the 1989 Conference on Decision and Control.

Holohan, A. M. and Safonov, M. G., 1989, "On Computing The M.I.M.O. Real Structured Stability Margin," Proceedings of the 1989 Conference on Decision and Control.

Horowitz, I. M., 1963, Synthesis of Feedback Systems, Academic Press, New York.

Horowitz, I. M., 1982, "Quantitative Feedback Theory," IEE Proceedings, vol. 129, pt. D, no. 6, pp. 215-226.

Jury, E. I., 1974, Inners and Stability of Dynamic Systems, Wiley, New York, 1974.

Kharitonov, V. L., 1978a, "Asymptotic Stability of an Equilibrium Position of a Family of Systems of Linear Differential Equations," *Differential'nye Uravneniya*, vol. 14, no. 11, pp.1483-1485.

Kharitonov, V. L., 1978b, "On a generalization of a stability criterion," *Izv. Akad. Nauk. Kazakh. SSR Ser. Fiz. Mat.*, vol. 1, pp 53-57, (in Russian).

Kochenburger, R. J., 1953, "Limiting in Feedback Control Systems," Transactions of the AIEE, vol. 72, part II, Applications and Industry, pp. 180-194.

Kraus, F. J., Anderson, B. D. O., Jury, E. I., and Mansour, M., 1988, "On the Robustness of Low-Order Schur Polynomials," IEEE Transactions on Circuits and Systems, vol. 35, no. 5, pp. 570-577.

Ljung, L., 1988, A Progress Report from IFAC's Technical Committee on Theory, Automatica, vol. 24, no. 4, pp. 573-583.

Luenberger, D. G., 1984, Linear and Nonlinear Programming, Addison-Wesley Pub. Co., Reading, Massachusetts.

Marden, M., 1949, "The Geometry of the Zeros of a Polynomial in a Complex Variable," Math Surveys, no. 3, Amer. Math. Soc., Providence, R. I.

Mayer, G., 1984, "On the Convergence of Powers of Interval Matrices," Linear Algebra and Its Applications, vol. 58, pp. 201-216.

Minnichelli, R. J., Anagnost, J. J., and Desoer, C. A., 1989, "An Elementary Proof of Kharitonov's Stability Theorem with Extensions," IEEE Transactions on Automatic Control, vol. 34, no. 9, pp. 995-998, September.

Munkres, J. R., 1975, Topology: A First Course, Prentice-Hall, Englewood Cliffs, New Jersey.

Naimark, Y. I., 1949, Stability of Linearized Systems, Leningrad Aeronautical Engineering Academy, Leningrad.

Oppenheimer, E. P., and Michel, A. N., 1988, "Application of Interval Analysis Techniques to Linear Systems: Part III–Initial Value Problems," IEEE Transactions on Circuits and Systems, vol. 35, no. 10, pp. 1275-1290.

Petersen, I. R., 1989, "A Class of Stability Regions for which a Kharitonov-Like Theorem Holds," IEEE Transactions on Automatic Control, vol. 34, no. 10, pp. 1111-1115.

Petersen, I. R., 1987, "A New Extension to Kharitonov's Theorem," Proceedings of the 1987 Conference on Decision and Control.

Saydy, L., Tits, A. L., and Abed, E. H., 1988, "On the Generalized Stability of Convex Hull of Two Matrices," Proceedings of the 1988 Conference on Decision and Control.

Sideris, A. and Barmish, B. R., 1989, "An Edge Theorem for Polytopes of Polynomials which Can Drop in Degree," Systems & Control Letters, vol. 13, pp. 233-238.

Sideris, A. and de Gaston, R. R. E., 1986, "Multivariable Stability Margin Calculation with Uncertain Correlated Parameters," Proceedings of the 1986 Conference on Decision and Control.

Sideris, A. and Sánchez Peña, R. S., 1989, "Fast Computation of the Multivariable Stability Margin for Real Interrelated Uncertain Parameters," .IEEE Transactions on Automatic Control, vol. 34, no. 12, pp. 1272-1276.

Siljak, D. D., 1969, Nonlinear Systems: The Parameter Analysis and Design, Wiley, New York.

Siljak, D. D., 1989, "Parameter Space Methods for Robust Control Design: A Guided Tour," IEEE Transactions on Automatic Control, vol. 34, no. 7, pp 674-688.

Soh, C. B. and Berger, C. S., 1988, "Damping Margins of Polynomials with Perturbed Coefficients," IEEE Transactions on Automatic Control, vol. 33, no. 5, pp. 509-511.

Soh, Y. C. and Foo, Y. K., 1989, "Generalized Edge Theorem," Systems & Controls Letters, vol. 12, pp. 219-224.

Sondergeld, K., 1983, "A Generalization of the Routh-Hurwitz Stability Criteria and an Application to a Problem in Robust Controller Design," IEEE Transactions on Automatic Control, vol. 28, no. 10, pp. 965-970.

Truxal, J.G., 1955, Automatic Feedback Control System Synthesis, McGraw-Hill, New York.

Uliana, C., 1985, "Model Reference and Optimal Control Applied to Computerized Numerical Control Systems," ME Project Report, RPI, December.

Walach, E. and Zeheb, E., 1980, "Sign Test of Multivariable Real Polynomials," IEEE Transactions on Circuits and Systems, vol. 27, no. 7, pp. 619-625.

Zadeh, L. A. and Desoer, C. A., 1963, Linear Systems Theory, McGraw-Hill, New York.

Zedek, M., 1965, "Continuity and Location of Zeros of Linear Combinations of Polynomials," Proceedings of the American Math. Society, vol. 16, pp. 78-84.

Zeheb, E., 1989, "Necessary and Sufficient Conditions for the Root Clustering of a Polytope of Polynomials in a Simply Connected Domain," IEEE Transactions on Automatic Control, vol. 34, no. 9, pp. 986-990.

Zeheb, E., 1990, "Necessary and Sufficient Conditions for Robust Stability of a Continuous System - The Continuous Dependency Case Illustrated Via Multilinear Dependency," IEEE Transactions on Circuits and Systems, vol. 37, no. 1, pp. 47-53.

Zeheb, E. and Walach, E., 1981, "Zero Sets of Multiparameter Functions and Stability of Multidimensional Systems," IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 29, no. 2, pp. 197-206.

Attachment 7

AFOSR-TR- 91 0988

THESIS BY:

D. J. KNUDSEN

CORNELL UNIVERSITY

Subcontract No.# S-789-000-038

GEOPHYSICAL RESEARCH LETTERS, VOL. 17, NO. 7, PAGES 921-924, JUNE 1990

DISTINGUISHING ALFVÉN WAVES FROM QUASI-STATIC FIELD STRUCTURES ASSOCIATED WITH THE DISCRETE AURORA: SOUNDING ROCKET AND HILAT SATELLITE MEASUREMENTS

D. J. Knudsen, M. C. Kelley, and G. D. Earle[1]

School of Electrical Engineering, Cornell University

J. F. Vickrey

SRI International

M. Boehm[2]

Space Sciences Laboratory, University of California, Berkeley

Abstract. We present and analyze sounding rocket and HILAT satellite measurements of the low frequency (< 1 Hz) electric and magnetic fields $\delta\mathbf{E}$ and $\delta\mathbf{B}$ perpendicular to the Earth's magnetic field $\mathbf{B}_0$ in the auroral oval. By examining the time-domain field data it is often difficult to distinguish temporal fluctuations from static structures which are Doppler shifted to a non-zero frequency in the spacecraft frame. However, we show that such a distinction can be made by constructing the impedance function $Z(f) = \mu_0|\delta E(f)/\delta B(f)|$. Using $Z(f)$ we find agreement with the static field interpretation below about 0.1 Hz in the spacecraft frame, i.e. $Z(f) = \Sigma_p{}^{-1}$ where Σ_p is the height-integrated Pedersen conductivity of the ionosphere. Above 0.1 Hz we find $Z(f) > \Sigma_p{}^{-1}$, which we argue to be due to the presence of Alfvén waves incident from the magnetosphere and reflecting from the lower ionosphere, forming a standing wave pattern. These waves may represent an electromagnetic coupling mechanism between the auroral acceleration region and the ionosphere.

Introduction

Satellite measurements of high-latitude fluctuating electric and magnetic fields measured perpendicular to $\mathbf{B}_0$ often show a high degree of correlation. The nature of the correlation has been explained by treating the fields as static, with the fluctuations resulting from the motion of the satellite through spatial structures with scale sizes from a few hundred meters to hundreds of kilometers. Assuming that there are no variations in the zonal direction, the ratio of the zonal magnetic to meridional electric field amplitudes will be proportional to the height-integrated Pedersen conductivity in the ionosphere Σ_p [*Sugiura et al.*, 1982; *Sugiura*, 1984; *Smiddy et al.*, 1984]. Since the spectrum of fluctuating fields measured in such an experiment is generally a monotonically decreasing function of frequency, the correlation method emphasizes the largest scales in the system.

The purpose of this study is to investigate spatial and/or temporal scales smaller than those studied by the authors listed above and which are closer to the regime associated with discrete auroral arcs. In so doing we address the possibility that the fluctuation fields may be due to Alfvén waves. In the wave model, the fluctuation amplitudes are related to the characteristic impedance of Alfvén waves, $Z_A = \mu_0 V_A$, where μ_0 is the permeability of free space and V_A is the Alfvén wave phase velocity. This effect was measured in two events lasting a few seconds each by *Chmyrev et al.* [1985] using the Intercosmos-Bulgaria-1300 satellite. Since the wave model can include reflections from the ionosphere, the relation between $\delta\mathbf{E}$ and $\delta\mathbf{B}$ depends upon the conductivity of the ionosphere and upon the electrical length (i.e. the number of Alfvén wavelengths) between the measurement point and the ionosphere as well.

The connection between Alfvén waves and auroral arcs has been discussed by many authors, including *Hasegawa* [1976], *Goertz and Boswell* [1979], *Haerendel* [1983], and *Seyler* [1988]. Alfvén waves have been measured in and above the ionosphere, most recently by *Boehm et al.* [1990] who identified step-like waves and near-coherent oscillations in time-domain rocket data, and by *Erlandson et al.* [1989] who measured electric and magnetic field spectral enhancements at micropulsation frequencies using the Viking satellite. Due in part to the Doppler shift of spatial structures into the Alfvén wave frequency band, the presence of Alfvén waves in either the time-domain or spectral data is often not so obvious. In this paper we show that the impedance function $Z(f) = \mu_0|\delta E(f)/\delta B(f)|$ can be used to distinguish the presence of propagating temporal phenomena in cases for which the temporal nature would not otherwise be obvious.

Method of Analysis

In the absence of reflections an Alfvén wave is characterized by an impedance function $Z(f)$ which is independent of frequency and equal to the characteristic impedance of the medium, $Z_A = \mu_0 V_A$. Z_A is typically much greater than $\Sigma_p{}^{-1}$, and the detection of measured values of $Z(f)$

[1]Currently at SAIC, McLean, Virginia

[2]Currently at the Max-Planck-Institute for Extraterrestrial Physics, Garching

Paper number 90GL00672
0094-8276/90/90GL-00672$03.00

AD-A-243935

PAGE
#922

MISSING
FROM ORIGINAL
DOCUMENT
AS SENT TO
DTIC
FROM THE
ORIGINATOR

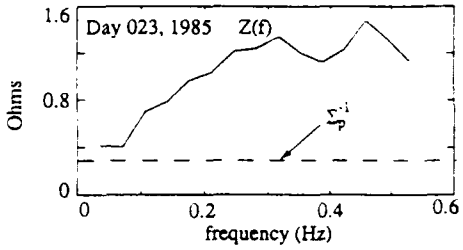


Fig. 2. Measured impedance spectrum $Z(f)$ computed from the data shown in Figure 1.

have plotted Σ_p^{-1} as deduced from Sondrestrom radar data, shown by the dotted line. Note that $Z(f)$ tends towards Σ_p^{-1} in the low frequency limit. A trend towards increasing impedance at higher frequencies is clearly visible. Quasi-static fields and a uniform Σ_p would produce the flat line at $Z(f) = \Sigma_p^{-1}$.

We now examine HILAT data for which the satellite was passing southward through the polar cap and auroral oval between 16:30 and 17:00 MLT. The solar zenith angle was about 60°, and the ionosphere was therefore sunlit. HILAT orbits at about 800 km, with a velocity of 7.4 km/s. The northward electric field is derived from an ion drift meter measuring cross-track ion **E×B** velocities, which periodically switches between 16 and 32 s^{-1} sample rates. The eastward magnetic field perturbations were measured with a fluxgate magnetometer which was sampled at 20 s^{-1}. These instruments are described in detail by *Potemra et al.* [1984] and *Rich et al.* [1984]. Both field quantities were averaged to 2 Hz, or one measurement every 3.7 km. The meridional electric and zonal magnetic fields are shown in Figure 3. The slow variation in the magnetic field throughout the pass is due to thermal stress on and attitude changes of the satellite, but is well below the range of frequencies we are considering. The magnetometer resolution is about 15 nT, but averaging gives an effective resolution somewhat lower than this. The resolution is still marginal for Alfvén wave measurements, so we must choose times in which sufficient magnetic field fluctuations are available to assure that we are measuring signal rather than noise. This is the case in the shaded interval in Figure 3.

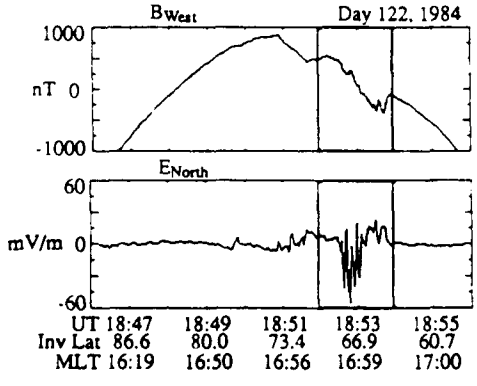


Fig. 3. Zonal magnetic and meridional electric fields measured with the HILAT satellite.

Proceeding as described above we obtained 14 separate spectra which were averaged to find the impedance function $Z(f)$ shown in Figure 4. Σ_p^{-1} (shown by the dotted line) was estimated based on the solar and particle input to the E region [*Robinson and Vondrak*, 1984]. For this sunlit case Σ_p was quite uniform. As with the rocket measurement, $Z(f) = \Sigma_p^{-1}$ near $f = 0$, and again an increase in $Z(f)$ above Σ_p^{-1} is evident above about 0.1 Hz.

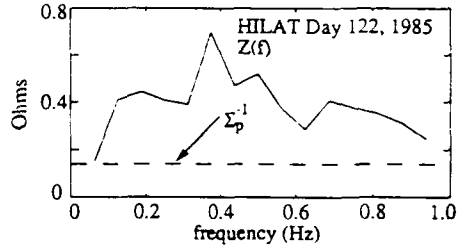


Fig. 4. $Z(f)$ computed from the HILAT satellite data.

Discussion

Above 0.1 Hz the impedance functions for the two cases we have presented show a marked increase over what would be expected from a simple static electric field/Birkeland current model. This result alone argues for the importance of Alfvén waves in the electromagnetic structure of the disturbed auroral oval.

If the field fluctuations vary in time, then they will propagate as Alfvén waves and reflect off of the lower ionosphere to form a standing wave pattern. As in a transmission line, the wave impedance has a maximum one quarter wavelength from a conducting load. Of course in the ionosphere the Alfvén wavelength changes with varying plasma density, but we can estimate the frequency of maximum impedance f_{max} by neglecting partial reflections off of the F-region density gradient and assuming a single reflection from the top of the E region (at z_{min}). The result is

$$\frac{1}{f_{max}} = 4\int_{z_{min}}^{z_{max}} \frac{dz}{V_A(z)} \qquad (1)$$

z_{max} is the height at which the measurement is taken. If we assume an O^+ plasma and use the density profile taken with the Sondrestrom radar at the time of the Black Brant rocket launch, (1) gives f_{max} = 0.37 Hz. For the satellite data we estimated the density profile by using Sondrestrom radar data for days close to the pass and at the same local time. Unfortunately, no radar data were available at the time of the satellite pass. In this case (1) gives f_{max} = 0.25 Hz. Comparing these values to Figures 2 and 4 shows that deviations of $Z(f)$ from Σ_p^{-1} are indeed maximum at roughly a few tenths of Hz.

Variations with frequency in the impedance spectrum $Z(f)$ can arise for reasons other than the presence of Alfvén waves. A detailed study of these possibilities using a realistic numerical model for a distributed, reflecting ionosphere is in progress and will be published elsewhere.

If we perform a statistical analysis on the time-domain data without filtering as done by *Sugiura et al.* [1982], we find for

AD-A-243935

PAGE
924
MISSING
FROM ORIGINAL
DOCUMENT
AS SENT TO
DTIC
FROM THE
ORIGINATOR

ALFVÉN WAVES AND STATIC FIELDS IN MAGNETOSPHERE/IONOSPHERE COUPLING: *IN-SITU* MEASUREMENTS AND A NUMERICAL MODEL

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

David J. Knudsen

August, 1990

ALFVÉN WAVES AND STATIC FIELDS IN MAGNETOSPHERE/IONOSPHERE COUPLING: *IN-SITU* MEASUREMENTS AND A NUMERICAL MODEL

David J. Knudsen, Ph.D.

Cornell University 1990

Perturbation electric and magnetic fields carry in excess of 10^{10} to 10^{12} W of electrical power between the magnetosphere and high-latitude ionosphere. Most of this power is generated by the solar wind. The ionosphere at large spatial and temporal scales acts as a dissipative slab which can be characterized by its height-integrated Pedersen conductivity Σ_P, so that the power flux into the ionosphere due to a quasi-static electric field E is given by $\Sigma_P E^2$.

The energy transferred to the ionosphere by time-varying electromagnetic fields in the form of Alfvén waves is more difficult to calculate because density and conductivity gradients can reflect energy. Thus, field resonances and standing wave patterns affect the magnitude and altitude distribution of electrical energy dissipation. We use a numerical model to calculate the frequency-dependent electric field reflection coefficient of the ionosphere and show that the ionosphere does not behave as a simple resistive slab for electric field time scales less than a few seconds.

Time variation of spacecraft-measured high-latitude electric and perturbation magnetic fields is difficult to distinguish from spatial

structuring that has been Doppler-shifted to a non-zero frequency in the spacecraft frame. However, by calculating the frequency-dependent amplitude and phase relations between fluctuating electric and magnetic fields we are able to show that low frequency fields (< 1 Hz) measured by an auroral sounding rocket traveling parallel to the auroral oval are due to standing Alfvén waves rather than quasi-static structures. Comparing the field fluctuations with electron energy measurements indicates that the waves occur near auroral arcs.

We include satellite data in our study as well. The amplitude relations between electric and magnetic field measurements taken by the HILAT satellite (traveling perpendicular to the auroral oval at an altitude of 800 km) show that the field fluctuations are due largely to Doppler-shifted quasi-static structures, but in some cases standing Alfvén waves also contribute.

Biographical Sketch

David J. Knudsen was born in Rapid City, South Dakota on 24 April, 1963. He attended elementary school in Wichita, Kansas, and junior high and high schools in Fort Dodge, Iowa. He graduated from Fort Dodge Senior High School in 1981. He then entered the electrical engineering program at Iowa State University. While he was an undergraduate, David held a summer internship at the Ford Aerospace and Communications Corp. in Newport Beach, California, and he worked as a student engineer at WOI Television in Ames, Iowa. He graduated with the B. S. degree in May, 1985. Since June, 1985 he has worked with the Space Plasma Physics Group at Cornell University.

The author married Ellen S. Miller in June, 1986, and they have a son, Aaron, who was born on 21 September, 1989.

To Ellen, Aaron,

and our families,

for making this project seem

like an adventure

rather than an ordeal.

Acknowledgments

It is my pleasure to extend my sincere thanks to Prof. M. C. Kelley for serving as chairman of my special committee. Despite an amazing number of other projects, Mike is always willing to spend an hour or two in lively discussions about Poynting flux or auroral physics. I also appreciate the effort he put forth in introducing me to the space physics community, and in helping me find post-doctoral employment.

I would also like to thank the other members of my committee: Prof. D. Hammer and Prof. D. T. Farley, whose excellent classroom instruction and careful attention to this project are much appreciated, and Dr. J. F. Vickrey. I have very much enjoyed working with Dr. Vickrey, who has been extremely helpful during the entire duration of my graduate program. Jim has supplied a huge amount of experimental data and many good ideas for interpreting it.

Thanks are due to Prof. C. E. Seyler for the interest and time he has devoted to this project. I have also enjoyed working with Prof. P. M. Kintner, Prof. W. E. Swartz, Dr. Jason Providakes, and Prof. N. Otani.

Dr. Greg Earle, Dr. Wayne Scales, Prof. James LaBelle, Dr. Robert Pfaff, Dr. Carl Siefring, and Dr. Charles Cornish deserve thanks for helping me to define my research program in the early stages, and for many good insights into graduate student life. I am indebted to Dr. John Sahr and Dr. James Providakes for allowing me to remain mostly computer-illiterate by offering lots of computer help, and also for being good office mates and friends from my first days at Cornell. Many thanks to Phil Erickson and Dr. Joe Pingree, who did most of the

calculations in Chapter 6, and who were, along with Richard Brittain, Chuck Swenson, and Glenn Berg, very generous with computer advice over the years. I've also very much enjoyed working with Tim Hall, Brett Isham, Robert Green, Steve Powell, Tim Wheeler, Cho-Hoi Hui, Dave Hysell, Jorge Vago, John Cho, Marian Silberstein, and Jian Ding. I gratefully acknowledge the friendly competence of Laurie Shelton, Sally Bird, and Susan Swartz in keeping our research group running smoothly.

Terri Dabbs, Nancy Walker, Mary McCready, Craig Heinselman, and Denise Rust of SRI International have been very helpful in supplying and processing HILAT satellite and Sondrestrom radar data. Thanks are also due to Dr. M. Boehm, Dr. B. McFadden, and Prof. C. Carlson of the University of California at Berkeley for providing the sounding rocket data presented in Chapters 3 and 5.

I am grateful to Dr. H. Carlson and Dr. E. Weber of the Air Force Geophysics Laboratory for helping me apply for and obtain an Air Force Office of Scientific Research Graduate Laboratory Fellowship, which I have held for the past two years.

Finally, I would like to acknowledge the support of close friends and family. Grant Heffelfinger is responsible in large part for my coming to Cornell. Paul Bay, Gretta Anderson, Sam Otter, and Caverlee Cary deserve much of the credit for my staying here. And most importantly, I would like to express my deep love and gratitude to Ellen, Aaron, Mom, Dad, and the rest of our families, to whom this dissertation is dedicated.

Table of Contents

List of Tables

List of Figures

CHAPTER 1
INTRODUCTION

1.1 Background

Eighteenth century scientists might have been satisfied with the idea of an infinite void, save for a relative few heavenly bodies, extending outward from somewhere above the cloud tops had it not been for the tantalizing geophysical clues provided by nature. In the early 1700s the Earth's magnetic field was well understood, static and predictable enough to use as a navigational tool, but why did G. Graham's finer measurements in 1722 reveal fluctuations? Balfour Stewart was able to show in 1882 that the magnetic fluctuations were due to electrical currents above the Earth. Of course this finding was probably stranger than the B-field fluctuations themselves, but Marconi's trans-Atlantic radio broadcast in 1901 corroborated the idea of a conducting layer high in the atmosphere.

The origin of the auroras borealis and australis had to be and still is one of the more interesting puzzles posed to those fortunate enough to view them, especially after the turn of the century when triangulation measurements placed the displays between 100 and 1000 km above the Earth's surface. The electrical nature of the aurora was known from the magnetic perturbations measured at ground level and associated with individual auroral arcs, but not until spacecraft flights through auroras in the 1960s was it known with certainty that energetic electrons (many

keV) crashing into the neutral atmosphere from above were responsible for the optical fluorescence.

Two discoveries in the 1950s pushed the envelope containing the known part of the Earth's environment well beyond the 1000 km upper boundary of the visible aurora. The first was that whistlers, the descending tone, audio-frequency electromagnetic waves known since World War I, were due to lightning discharges. More importantly, their dispersion was found to be the result of propagation through charged particles permeating space tens of thousands of km above the Earth's surface. Secondly, satellite measurements in this same region led to the discovery of the very high energy particles comprising the Van Allen radiation belts.

The present view is that the Earth's near space environment must be viewed as a complete system which includes the Sun's outermost atmosphere. One of the main goals of present day research is to construct a self-consistent model of that system's energy source (which is to say the Sun itself), the means of energy transmission (photons and the particles and magnetic fields of the solar wind), and the energy deposition in and associated dynamics of the Earth's magnetosphere, ionosphere, and neutral atmosphere. We might separate the effort to construct this model into two parts: 1) a detailed study of all the separate constituents of the system, and 2) an examination of how these constituents are interconnected. By and large this dissertation falls into the "connecting" category, where the two pieces we are trying to fit together are the magnetosphere and ionosphere.

The magnetosphere extends from about ten to hundreds of Earth radii away (depending on whether you are "upwind" or "downwind"

from the Earth) down to the ionosphere starting at roughly 1000 km in altitude, where collisions with the neutral atmosphere begin to take control of the charged particle dynamics. The study of magnetosphere-ionosphere coupling pertains mostly to high latitudes because that is where geomagnetic field lines extend from the ionosphere deep into the magnetosphere. At lower latitudes magnetic field lines are shorter and do not extend as far from the Earth.

Energy is exchanged between the magnetosphere and ionosphere in two main ways, via kinetic energy of charged particles which we will consider briefly in Chapter 3, and by quasi-static and wave related electric and magnetic fields, which we will emphasize throughout. Each of these energy sources can at certain times and places dominate the other, but the total power carried by each of them is roughly 10^{10} to 10^{12} W. Although this is 5 to 7 orders of magnitude smaller than the energy flux into the sunlit polar cap from solar photons, most of the solar flux goes directly into the neutral atmosphere at relatively low altitudes, leaving magnetosphere-ionosphere energy exchange to determine a significant part of the ionospheric plasma (and high-altitude neutral atmosphere) dynamics at high latitudes. Of course much of the ionospheric plasma is produced by photoionization in the first place, but its bulk motions, structuring, and instabilities are due in large part to particle precipitation and electric fields. In the polar winter, these sources dominate the energetics and dynamics of the upper atmosphere. Interactions with the neutral atmosphere (e.g. from gravity waves) and chemical processes also play a role.

On average the ionosphere acts as a dissipative load attached to the magnetospheric energy source, which in turn is driven by the interaction

between the magnetosphere and the solar wind. The situation can be quite complex because magnetospheric electric fields driving currents in the ionosphere are often accompanied by beams of energetic particles, which modify the conductivity of the ionosphere. This change in the load can affect the ionospheric electric fields and conceivably the particle precipitation itself, creating a feedback effect which is not yet fully understood and which we will not attempt to address. Other complications arise when neutral winds drive the ionospheric plasma, creating dynamo electric fields and switching the role of the magnetosphere from source to load.

1.2 Purpose and Organization of This Work

The central topic we address in this dissertation is the interpretation of spacecraft measurements of low frequency (less than 1 Hz) electric and perturbation magnetic fluctuations above the high-latitude ionosphere. As mentioned in the previous section, magnetic perturbations have been measured on the ground for centuries. *Birkeland* [1908] noted that these perturbations were especially strong under auroral arcs, and suggested they were caused by electric currents flowing along geomagnetic field lines associated with the aurora. The discovery of the Alfvén wave [*Alfvén*, 1950] allowed for the interpretation that B-field fluctuations measured on the ground were due to Alfvén wave resonances in the Earth's dipole field. Early satellite measurements [see *Zmuda et al.*, 1966] showed magnetic fluctuations in the polar region, and these also were interpreted predominantly in terms of Alfvén waves. *Cummings and Dessler* [1967] called into question the Alfvén wave interpretation by arguing that it was not possible for Alfvén waves to be localized as

satellite measurements had shown, and they again proposed, as Birkeland had, that quasi-static field-aligned currents were primarily responsible for the magnetic fluctuations.

Despite their arguments that quasi-static currents could explain spacecraft magnetic field measurements, Cummings and Dessler acknowledged the possibility that localized Alfvén waves are possible when accompanied by field-aligned currents. Nonetheless, the static model was generally accepted following the publication of *Cummings and Dessler* [1967]. In the last decade, the validity of the static field model has been re-examined. Obliquely propagating Alfvén waves known as "shear" or "slow" mode waves [*Stix*, 1962] carry field-aligned currents, and it is now known that these waves are indeed very important in magnetospheric electrodynamics. We argue in this dissertation that the static field model alone does not adequately describe high-latitude electromagnetic fields, and that it is appropriate to include the effect of shear Alfvén waves in studies of high-latitude and auroral dynamics.

In some sense the static Birkeland current model is a limiting case of the shear Alfvén wave model. However, there are important differences between the static and wave models because time-varying fields allow for reflections, resonances, and interference. It is difficult to discern from measurements on board a spacecraft if electric and magnetic field fluctuations are due to waves or to localized static disturbances which are Doppler-shifted to a finite frequency in the spacecraft frame. *Sugiura et al.*, [1982] used the Dynamics Explorer satellite to show that the electric and magnetic field fluctuations above the auroral zone are often highly correlated, and they used a novel approach to distinguish between the wave and static field cases by

calculating the ratio of the r.m.s. field amplitudes $\mu_0 E_{rms}/\delta B_{rms}$. They found the value of this ratio to be equal to the inverse of the height-integrated Pedersen conductivity of the ionosphere Σ_P^{-1}, and we will show in Chapter 3 that this is expected from static electric fields and associated Birkeland currents.

A problem with the analysis of *Sugiura et al.* [1982] is that E and δB power spectra are usually monotonically decreasing with frequency, so the lowest frequencies are emphasized in a correlation analysis. In Chapter 5 we remedy this problem by calculating the ratio of electric and magnetic field amplitudes *as a function of frequency*. We will also study the frequency-dependent phase relation between E and δB. In so doing we find that Alfvén wave dynamics play an important role in electrodynamical coupling between the magnetosphere and ionosphere for time scales less than about 10 s. We also show that, in at least one sounding rocket experiment, most of the Alfvén wave energy lies near regions of auroral precipitation.

The Alfvén wave model can be sub-divided into two categories: traveling waves and standing waves. This distinction is important because standing waves are indicative of reflections, and an understanding of these reflections is crucial in determining the fraction of electrical energy incident from the magnetosphere that is dissipated in the ionosphere. Our analysis in Chapter 5 shows that standing waves are present in both the satellite and rocket data (taken at 800 km and near 600 km, respectively).

The reflection characteristics of the ionosphere are complicated and change with wave frequency as a result of the strong altitude dependence of the plasma density and collision frequencies, and of the fact that the

thickness of the ionospheric "load" is on the order of an Alfvén wavelength. In order to accurately compute the amplitude and phase relations between E and δB fields due to Alfvén waves reflecting from the ionosphere, one must in general use a numerical model. We devote Chapter 4 to the development and general results of such a model. The model we choose was originally used by *Hughes* [1974] to predict properties of ground-based magnetometer measurements.

The electric field reflection coefficient of the ionosphere is often estimated by treating the ionosphere as a slab reflector with height-integrated Pedersen conductivity Σ_P, and the region above the ionosphere as a homogeneous "transmission line" with characteristic impedance $Z_A = \mu_0 V_A$ where V_A is the Alfvén velocity. The E-field reflection coefficient of the ionosphere is then given by $(\Sigma_P^{-1} - Z_A)/(\Sigma_P^{-1} + Z_A)$ [see for example *Paul and Nassar*, 1987]. As one of the main results of Chapter 4 we show that this approximation is not valid for time scales less than a few seconds, and we present plots of the reflection coefficient for small time scales and for different ionospheric density profiles.

While Chapters 4 and 5 treat time-varying electric and magnetic fields, Chapter 3 is devoted to interpretation of satellite and sounding rocket estimates of ionospheric Joule heating and Poynting flux into and out of the ionosphere in the DC limit. The Poynting flux method has not been used extensively, but it has several advantages over other electromagnetic energy measurements and we discuss those in detail. There are several factors which complicate all energy measurements; neutral winds and temporal variations are particularly important. We treat the topic of neutral winds in Chapter 3.

Having established in Chapter 5 the importance of Alfvén waves in the high-latitude ionosphere, we turn in Chapter 6 to a brief study of the effects that large amplitude waves can have on the interpretation of incoherent scatter radar data. But first we will review the literature and some of the physical concepts pertaining to Alfvén waves, in Chapter 2.

CHAPTER 2

BACKGROUND AND REVIEW

2.1 The Solar Wind and Magnetosphere

Much of the previous work in magnetosphere-ionosphere coupling has emphasized the region extending from the top of the ionosphere up to several thousand km. The ionosphere itself is often modeled simply as a conducting slab characterized by its height-integrated Pedersen conductivity (see references in Section 2.6). We will take the opposite approach by thinking of the magnetosphere only as an "upper boundary condition" which supplies electric fields, currents, and particles to the ionosphere and treating in detail the interaction between those energy sources and the ionosphere. In Chapter 4 we will use a detailed model of the ionosphere as input to a numerical model, so we will save a review of the ionosphere's morphology until then and concentrate now on the magnetosphere and some of the ways it can produce Alfvén waves -- the magnetosphere-ionosphere coupling mode which will receive most of the attention in this thesis.

Figure 2.1 shows a cross section of the magnetosphere in the noon-midnight meridian plane when the $\hat{z}$ component of the interplanetary magnetic field (IMF) is southward, which allows the IMF to penetrate to the magnetopause and connect with the Earth's magnetic field. The solar wind impinging from the left is comprised mostly of protons and electrons (≈ 5 cm^{-3}, $T_i \approx T_e \approx 10$ eV) traveling at about 500 km/s. The kinetic energy flux from this bulk flow is thus $\approx 5\times10^{-4}$ W/m^2. Assuming

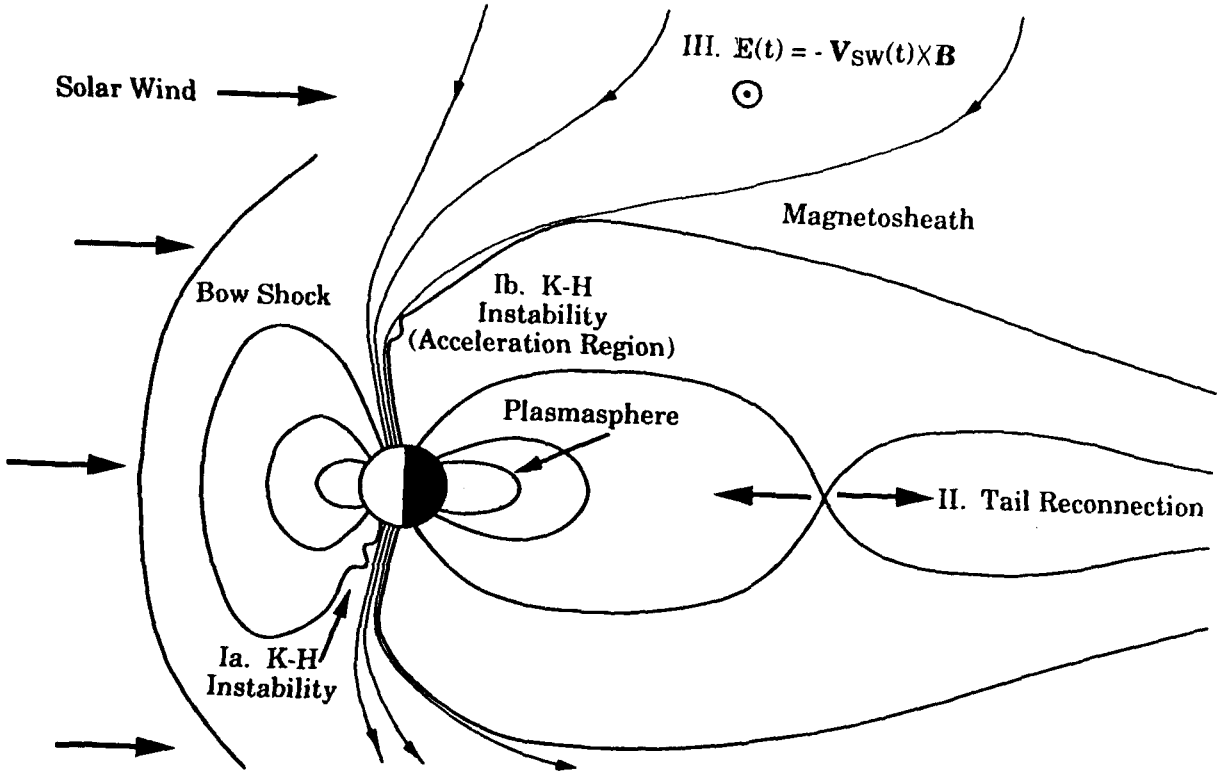


Figure 2.1 Possible sources of magnetospheric Alfvén waves which propagate to the ionosphere: I. Kelvin-Helmholtz instability. II. Tail reconnection and associated substorms. III. A "gusty" solar wind can generate a time-varying electric field.

that the magnetosphere has a cross section of $\pi\times$ (10 R_{Earth})^2 leaves on the order of 10^{13} W of available solar wind power. Solar photons carry more than 10^6 times more energy flux (1400 W/m^2), but the "capture area" is 100 times less since it is the Earth rather than the magnetosphere which intercepts the energy. The solar wind has a magnetic field of about 5 nT which means it can carry energy in the form of Alfvén waves in addition to the kinetic energy.

Although energy input from the solar wind to the Earth is usually 4-5 orders of magnitude smaller than from solar photon flux, solar wind energy is also much more variable and can have important effects on industry, electrical power systems, communications, and space operations and astronaut safety. This point is illustrated nicely by the huge solar flare and subsequent geomagnetic storm during the weeks following March 6, 1989 [*Allen et al.*, 1989]. The aurora, normally visible over the northern U. S. and Canada, was seen clearly in Georgia, Texas, and New Mexico. HF communications and navigation systems (< 50 MHz) which rely on reflection from the ionosphere were useless. Most of Quebec Province experienced a power blackout for up to 9 hours, large voltage swings appeared on undersea cables, and huge currents induced in pipelines caused concerns about corrosion. In a particularly bizarre event a Navy ship had to go to a backup shore-based radio system, causing automatic garage door openers in a California coastal suburb to start opening and closing.

When the solar wind has a velocity component perpendicular to magnetospheric field lines, electric fields are created and the process is called an "MHD dynamo" (see Figure 2.1). Magnetic field lines in collisionless plasmas are equipotentials, thus the dynamo electric fields

are mapped throughout the magnetosphere where they create a large scale plasma convection pattern (see *Stern* [1977] and references therein). If the solar wind is uniform and constant the convection pattern can remain in a steady state. But changes in the solar wind affect the entire convection pattern, and the new equilibrium is found after different parts of the system exchange energy in the form of Alfvén waves. Magnetic field lines which have fixed ends (e.g. at the Earth) can set up standing Alfvén waves which oscillate at resonant frequencies. The resulting magnetic perturbations can be measured with ground-based magnetometers, and the phenomenon has traditionally been studied under the name "micropulsations."

The solar wind is certainly not the only source of changes in the magnetosphere, and is not necessarily the most important. A leading explanation for the origin of micropulsations is the Kelvin-Helmholtz instability, which causes waves on the boundary between two fluids in relative motion. Examples are waves on lakes and flags flapping in the breeze. See *Melrose* [1986] for a more detailed discussion of the K-H instability.

Magnetospheric substorms are another source of Alfvén waves. Over a period of hours or days the magnetosphere can store energy in the form of magnetic fields in its tail. This is thought to take place through the process of dayside reconnection of geomagnetic field lines with the IMF and subsequent deposition in the magnetotail. At some point the tail thins and field lines there reconnect, sweeping part of the tail downstream in the solar wind and snapping the remaining part back towards the Earth. The result is a huge flux of particle and Alfvén wave energy towards the polar caps, accompanied by increases in auroral

displays, electrojet currents, and more. (See the review paper on substorms by *McPherron* [1979].)

There are other proposed sources of magnetospheric Alfvén waves for which references can be found in the review paper by *Hughes* [1982]. There one can also find a good review of micropulsation observations from spacecraft.

The observations presented in this thesis are taken within the auroral oval because that is where much of the magnetospheric energy is deposited. It is this energy, not the aurora itself, with which we are primarily concerned, but the aurora is an important part of the environment we study and we will devote the next section to a brief overview.

2.2 The Aurora

A large part of the study of both space and laboratory plasmas is devoted to the instruments used to make diagnostic measurements. Only after a huge amount of work can one piece together the measured fields, potentials, drifts, etc., into a coherent picture of some physical phenomenon. The aurora is one of the few examples of a plasma physics experiment that can be observed simply by looking upward at the right latitude, and the results are displayed in 3-D and brilliant color across the entire sky. Unfortunately, even this dynamic display did not provide enough information for early researchers to understand the cause of the aurora, and today after thousands of satellite passes above auroral displays, dozens of rocket flights through them, and countless hours of ground-based radar observations, there are many remaining questions.

The auroral oval is actually more of a torus centered near the magnetic pole with representative boundaries extending from 65°-75° in magnetic latitude. The oval thickens and extends equatorward during magnetic activity. Within the oval are smaller "curtains" or "arcs" extended in the zonal (E-W) direction for hundreds or thousands of km and ranging from 1-100 km in latitudinal thickness. The arcs in turn have twists, folds, and intensity enhancements which race along the edge, and which have been found to be due to the Kelvin-Helmholtz instability [*Hallinan and Davis*, 1970]. The morphology of the aurora is quite a large subject in itself, but a nice overview with color photographs is given by *Eather* [1980].

The term "aurora" encompasses many different phenomena which in general are characterized by light emitted in the upper atmosphere (100-1000 km) due to electrons incident from above and colliding with neutral atoms. The resulting fluorescence extends from infrared to ultraviolet and beyond, but the visible aurora is due to electrons with a kinetic energy of 1-10 keV directed along the geomagnetic field. In the early 1970s these electron beams were measured by satellites and rockets, and their energy spectra were found to be roughly monoenergetic. This and other measurements led to the discovery that the electrons were accelerated by a quasi-static potential drop maintained from 1-2 Earth radii above the surface (see *Akasofu*, [1981] and references therein).

Since the plasma in the acceleration zone is collisionless, it is hard to explain the existence of electric fields parallel to **B** that last for tens of minutes. The origin of this energization is still under debate, and is possibly the strongest motivator behind current auroral research. The leading theories for the potential drop associated with small scale intense

arcs involve anomalous resistivity, double layers, and kinetic Alfvén waves.

Plasma wave turbulence can arise from the intense field-aligned currents which are known to exist over the aurora. The turbulence can in some cases mimic the effects of collisions, allowing a finite conductivity along field lines. This effect is termed "anomalous resistivity" and is thought to play a role in sustaining the kV potentials through which auroral electrons are accelerated.

Double layers are small structures (several Debye lengths long) which are driven by field-aligned currents and, most importantly, support a net potential drop across themselves. The potential difference is on the order of the electron temperature, which is only about 1 eV. As small as these structures are, they have been observed by the S3-3 and Viking satellites [*Temerin et al.*, 1982; *Boström et al.*, 1988; *Koskinen et al.*, 1989], and if thousands of them occur on a single field line they might possibly explain the electron acceleration.

A third contender in explaining auroral electron energies is the kinetic Alfvén wave theory. Since we will deal with Alfvén waves throughout much of this thesis, we will spend the next few sections reviewing their linear theory, and in the process we will talk a little about Alfvén waves as a possible auroral acceleration mechanism.

2.3 Linear Theory of Alfvén Waves

Without a static magnetic field, electromagnetic waves cannot propagate below the electron plasma frequency. Alfvén waves are electromagnetic waves which exist in a magnetized plasma at frequencies below all of the cyclotron frequencies. They propagate as

perturbations along the static **B**-field in the plasma. A useful intuitive picture of Alfvén waves comes from thinking of **B**-field lines as taut strings which propagate perturbations when plucked. The linear dispersion relation for Alfvén waves is derived in most plasma theory textbooks from the equations of magnetohydrodynamics. We will instead stress the physical picture of the mechanism underlying Alfvén waves, and by examining the motions of single particles in the presence of low frequency electric fields we will find the dielectric response of a plasma to low frequency perturbations.

In Figure 2.2 an ion and an electron are shown (schematically) in the presence of a $\hat{z}$ - directed static magnetic field $\mathbf{B}_0$. Both particles are initially on the line $y = 0$, but when a static electric field $E_y\hat{y}$ is applied they begin to **E**×**B** drift in the $\hat{x}$ direction. Since this drift is the same for electrons and ions, there is no current in the $\hat{x}$ direction. Notice however that the *average* positions of the particles have separated in $\hat{y}$, i.e. in the direction of the electric field.

We can draw an analogy between this situation and the polarization of a dielectric solid. Applying an electric field to a slab of dielectric material causes the individual atoms in the dielectric to polarize, which gives them a dipole moment $e\mathbf{d}$ where e is the fundamental unit of charge and **d** is an effective separation of positive and negative charges. The dipole moment per unit volume is known as the polarization $\mathbf{P} = ne\mathbf{d}$ where n is the density of atoms. The electric flux density vector is the sum of the "free space" flux density plus the polarization of the material, i.e. $\mathbf{D} = \varepsilon_0\mathbf{E} + \mathbf{P}$. Finally, the dielectric constant of the material ε is defined by $\mathbf{D} = \varepsilon\mathbf{E}$, thus

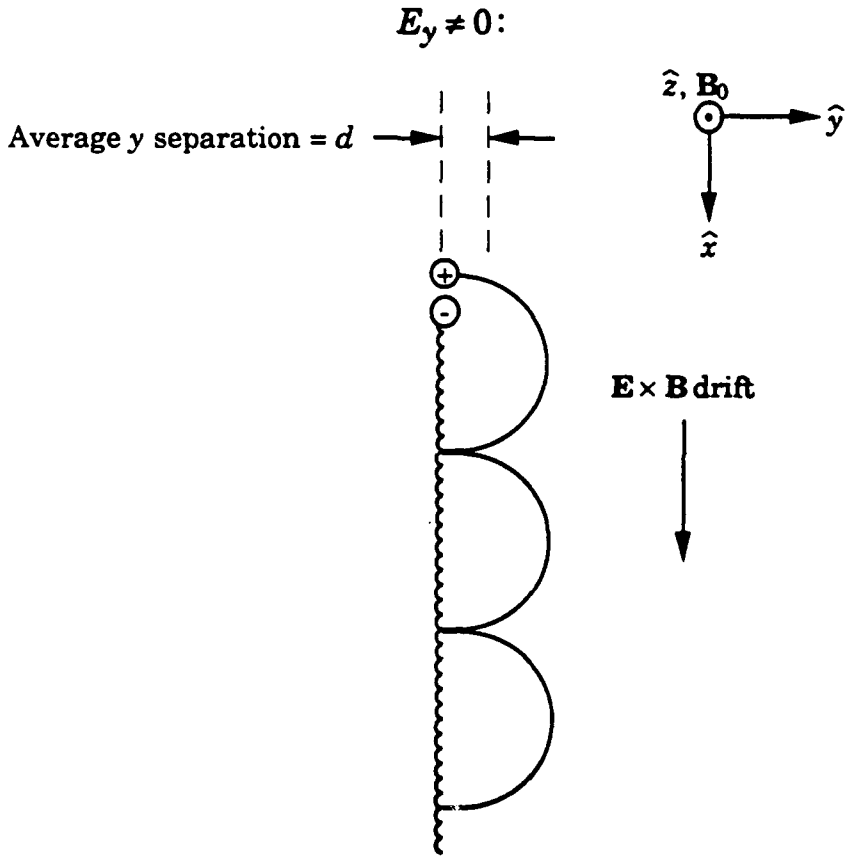


Figure 2.2 Applying a static electric field **E** perpendicular to $\mathbf{B}_0$ causes a dielectric response in the plasma by displacing charges relative to each other in the direction of **E**.

$$\varepsilon = \varepsilon_0 + ne|\mathbf{d}|/|\mathbf{E}| \tag{2.1}$$

This type of analysis is commonly used in introductory electromagnetics texts [e.g. *Paul and Nasar*, 1987] to explain the dielectric behavior of solids, and the picture is useful over a wide band of frequencies. The model breaks down at high frequencies, where the dielectric "constant" becomes frequency dependent.

In plasmas, the dielectric response is very dependent on frequency due to the various fundamental frequencies in the plasma such as the plasma and cyclotron frequencies. It follows then that plasma textbooks do not use the polarization model reviewed above to find the dielectric function for plasmas. As we will show, it turns out that the analogy with dielectric solids *does* predict the dielectric response of a magnetized plasma at frequencies well below the plasma and cyclotron frequencies, in the regime of Alfvén waves. We choose this approach since it is not commonly used in the literature and it does provide some useful insights.

Our main job now in finding the low frequency polarizability of a magnetized plasma is to find the average total displacement (labeled d in Figure 2.2) between ions and electrons after an electric field is imposed. To do this we can solve the equations of motion for a single ion in the presence of static magnetic and electric fields. Using the field directions shown in Figure 2.2,

$$\frac{dv_x}{dt} = \frac{q_i}{m_i} B_0 v_y \tag{2.2a}$$

and

$$\frac{dv_y}{dt} = \frac{q_i}{m_i}(E_y - B_0 v_x) \tag{2.2b}$$

Eliminating v_y gives a second-order equation for v_x:

$$\frac{d^2 v_x}{dt^2} = \Omega_i^2 \left(\frac{E_y}{B_0} - v_x \right) \tag{2.3}$$

Equation 2.3 has the general solution

$$v_x(t) = A\sin(\Omega_i t) + B\cos(\Omega_i t) + (E_y/B_0) \tag{2.4}$$

This may now be substituted into (2.2a) to find $v_y(t)$:

$$v_y(t) = A\cos(\Omega_i t) - B\sin(\Omega_i t) \tag{2.5}$$

If we require $v_x(0) = v_y(0) = 0$ then $A = 0$, $B = -E_y / B_0$. Finally, we can integrate the velocities and impose $x(0) = y(0) = 0$ to obtain:

$$x(t) = \frac{E_y}{B_0} \left(t - \frac{1}{\Omega_i} \sin(\Omega_i t) \right) \tag{2.6a}$$

$$y(t) = \frac{E_y}{B_0} \cdot \frac{1}{\Omega_i} \left(1 - \cos(\Omega_i t) \right) \tag{2.6b}$$

Equation 2.6a shows that the ion gyrates at the cyclotron frequency Ω_i with a radius $\rho = E_y/(B_0\Omega_i)$. Notice that this is the "usual" gyroradius with the thermal velocity v_{th} replaced by the **E×B** drift velocity. Superimposed on the gyro motion is the **E×B** drift, represented by the first term in Equation 2.6a. Our main interest is in the first term in Equation 2.6b. That is the term which remains after averaging over the fast cyclotron motion, leaving an offset in the $\hat{y}$ direction equal to one "**E×B** gyro radius." This offset is proportional to E_y, and is the polarization effect we are looking for. We can now substitute $|\mathbf{d}| = E_y/(B_0\Omega_i)$ into Equation 2.1 to obtain:

$$\varepsilon = \varepsilon_0 + nm_i/B_0^2 \tag{2.7}$$

It is important to remember that (2.7) was derived assuming that the applied electric field is perpendicular to $\mathbf{B}_0$. Our simple analogy with dielectric solids breaks down for electric fields parallel to $\mathbf{B}_0$, although we can correct this problem, as we will show later. We have neglected the $\hat{y}$ displacement of the electrons because it is smaller than the ion displacement by m_e/m_i, as is evident from Equation 2.6b.

Now consider an electromagnetic plane wave propagating along $B_{0\hat{z}}$ with a $\hat{y}$-directed electric field and a perturbation magnetic field associated with the wave in the $-\hat{x}$ direction. What is the phase velocity of such a wave? In this case we can use the simple relation

$$v_{ph} = \frac{c}{\sqrt{\varepsilon/\varepsilon_0}} \tag{2.8}$$

Combining (2.7) and (2.8) shows that electromagnetic waves propagating along $\mathbf{B}_0$ travel at the Alfvén velocity, i.e. $v_{ph} = V_A$, where

$$V_A = \frac{B_0}{\sqrt{\mu_0 n m_i}} \tag{2.9}$$

We have assumed that $V_A^2 << c^2$, which is true in the ionosphere.

To summarize, a low frequency electromagnetic wave traveling along the background magnetic field in a plasma polarizes the plasma by displacing the ion gyro orbits in the direction of the wave electric field. This interaction leads to a high refractive index $n = c/V_A$. In the ionosphere, n is typically several hundred. Some plasma textbooks treat Alfvén waves by solving for the *current* caused by the changing centers of gyration of the ions. This current is known as the polarization current J_{pol} and can be found using $J_{pol} = ne\dot{y}$. However, we cannot use $\dot{y} = v_y$ from (2.5) because we assumed that E_y was constant in time, leaving only motions at the cyclotron frequency in the $\hat{y}$ direction. That is, (2.5) is

accurate only to zeroth order in ω. It turns out that the polarization drift can be found from (2.6b) when one takes the time derivative of the "constant" of (the spatial) integration, $E_y/(B_0\Omega_i)$, yielding

$$J_{pol} = (\partial E_y/\partial t)/(\mu_0 V_A^2) \tag{2.10}$$

In effect this is a perturbation solution good to first order in ω/Ω_i. A somewhat more rigorous derivation of the polarization drift can be found in various plasma textbooks, e.g. *Nicholson* [1983].

2.4 Oblique Propagation of Alfvén Waves

Now that we have established the basic mechanism underlying Alfvén waves, we are ready to add another detail: propagation directions with a component perpendicular to $\mathbf{B}_0$. This is important for use in subsequent chapters because the Alfvén waves we will study are confined in at least one spatial direction perpendicular to $\mathbf{B}_0$, either by the auroral oval or by individual auroral arcs.

To begin we assume a homogeneous plasma with a background magnetic field $\mathbf{B}_0 = B_0\hat{z}$. We allow the propagation vector to have along-$\mathbf{B}_0$ and off-$\mathbf{B}_0$ components: $\mathbf{k} = k_x\hat{x} + k_z\hat{z}$. Thus we may assume that all field quantities vary as $\exp(i\omega t - ik_x x - ik_z z)$. In this case, Maxwell's wave equation

$$\nabla\times\nabla\times\mathbf{E} + \mu_0\frac{\partial\mathbf{J}}{\partial t} + \frac{1}{c^2}\frac{\partial^2\mathbf{E}}{\partial t^2} = 0 \tag{2.11}$$

can be expressed in component form as follows:

$$x: \quad k_z^2 E_x - k_x k_z E_z + i\omega\mu_0 J_x - \frac{\omega^2}{c^2} E_x = 0 \tag{2.12a}$$

$$y: \quad \left(k_x^2 + k_z^2\right) E_y + i\omega\mu_0 J_y - \frac{\omega^2}{c^2} E_y = 0 \tag{2.12b}$$

$$z: \quad k_x^2 E_z - k_x k_z E_x + i\omega\mu_0 J_z - \frac{\omega^2}{c^2} E_z = 0 \tag{2.12c}$$

To solve for the Alfvén wave dispersion relations, we need to express the currents J_x, J_y, and J_z in terms of the electric field components. For the two currents perpendicular to $\mathbf{B}_0$, J_x and J_y, we can use the expression for the polarization current from Equation (2.10). The polarization current cannot operate parallel to $\mathbf{B}_0$, so we need to refer to the particle equations of motion to find the current:

$$\frac{dv_{j,z}}{dt} = \frac{q_j}{m_j} E_z \tag{2.13}$$

The subscript j is a species index. Using $J_z = nev_z$, we obtain (neglecting $m_e << m_i$)

$$J_z = \frac{\omega_{pe}^2}{i\omega} \varepsilon_0 E_z \tag{2.14}$$

Notice that since E_x and E_y are decoupled in Equations 2.12, the two resulting wave modes are linearly polarized. We are now ready to eliminate E_x and E_z from (2.12a) and (2.12c) to find the dispersion relation for Alfvén waves with the perpendicular electric field E_x in the same direction as the across-$\mathbf{B}_0$ component of the propagation vector, k_x. The geometry of this case is illustrated in the top part of Figure 2.3, and the corresponding dispersion relation is

$$k_z^2 = \frac{\omega^2}{V_A^2}\left(1 + k_x^2 c^2/\omega_{pe}^2\right) \text{ (slow mode)} \tag{2.15}$$

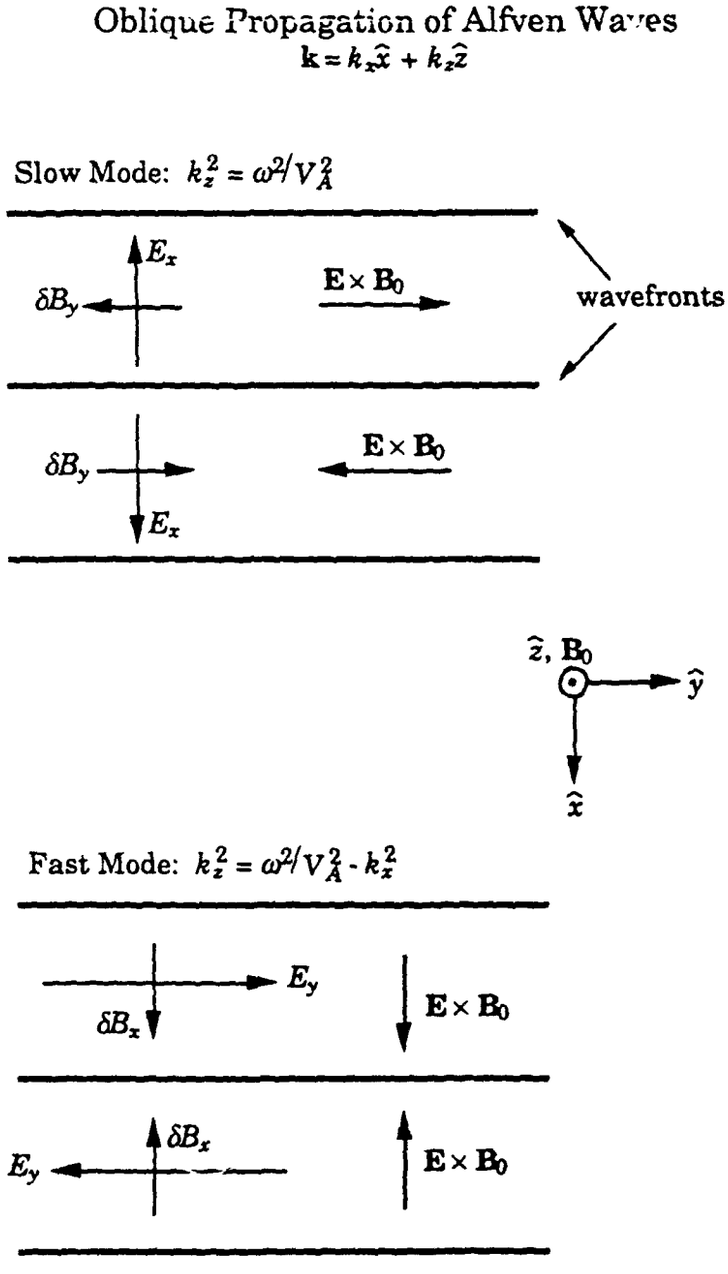


Figure 2.3 The two Alfvén wave modes as defined by *Stix* [1962]. For typical auroral parameters the fast mode is evanescent, causing the meridional electric field E_x to dominate.

This mode is referred to in the literature either as the "slow mode" or the "shear mode" Alfvén wave [*Stix*, 1962]. The term "slow mode" comes from the fact that (neglecting for convenience $k_x^2c^2/\omega_{pe}^2 << 1$) (2.15) can be re-written $\omega/k = V_A\cos(\theta)$, where θ is the angle between **k** and $\hat{z}$. The phase velocity ω/k becomes arbitrarily small as the propagation direction becomes perpendicular to $\mathbf{B}_0$. Note however that the projection of the phase velocity *along* $\mathbf{B}_0$ is always V_A. The reason for the term "shear mode" can be seen in Figure 2.3. Since the perpendicular electric field is along $\hat{x}$, the plasma will **E**×**B** drift along $\hat{y}$. But the direction of the drift reverses in alternating phases of the wave as one moves along $\hat{x}$, causing the **E**×**B** drift to be sheared.

When the electric field is in the $\hat{y}$ direction, the **E**×**B** drift is along $\hat{x}$, and from the bottom part of Figure 2.3 one can see that there is a non-zero divergence or *compression* in the **E**×**B** drift. For this case the Alfvén wave is referred to as "compressional" or "fast." We will use the latter term. The dispersion relation follows directly from (2.12b):

$$k_z^2 = \frac{\omega^2}{V_A^2} - k_x^2 \text{ (fast mode)} \tag{2.16}$$

The interesting thing about (2.16) is that if k_x is larger than ω/V_A, k_z becomes imaginary. Thus the fast mode is evanescent at low frequencies, much as an electromagnetic wave in a waveguide cannot propagate below the cutoff frequency. This phenomenon lends insight to the auroral ionosphere. Structuring in the aurora is usually in the N-S (or meridional) direction, with the arcs and associated electric potentials elongated in the zonal (E-W) direction. To a first approximation we can ignore zonal variations and choose $\lambda_x = 2\pi/k_x$ to be between 1 km (roughly the minimum meridional thickness of an auroral arc) and 1000 km (the

width of the auroral oval). Using typical ion densities in the ionosphere, we find that Alfvén waves contained in the auroral oval are below the fast mode cutoff for $f = \omega/2\pi$ less than 1 Hz. In other words, if we assume that the wave properties do not vary in the zonal direction, the result is that the normal mode with a zonal electric field cannot propagate, leaving only the meridional electric field in the slow mode. Spacecraft measurements confirm that the meridional electric field is dominant in the aurora, although to be sure most studies assume that this field is quasi-static. One of our goals is to find the relative importance of static versus Alfvénic electric fields. We will find in Chapter 4 that if we include the effect of collisions, the slow and fast modes are coupled, and slow mode Alfvén waves propagating into the auroral ionosphere from above will drive the fast mode in the E-region, giving a non-zero zonal electric field (but which is still much smaller than the meridional electric field).

2.5 Kinetic Alfvén Waves and Parallel Electric Fields

There has been quite a bit of interest in the slow mode Alfvén wave (2.15) because it has associated with it an electric field parallel to $\mathbf{B}_0$. As mentioned in Section 2.2, parallel electric fields are known to play an important role in one of the still unsettled problems concerning the aurora: the acceleration of auroral electrons through a large potential drop, often as much as 10 kV or more. Since ideally the conductivity parallel to $\mathbf{B}_0$ is infinite in a collisionless plasma, it is difficult to explain the presence of such a large potential drop for *tens of minutes*, which is roughly the lifetime of auroral arcs. Alfvén waves have been considered as a source of parallel electric fields because a finite parallel conductivity

is not required to maintain a parallel electric field. E_z is created as a result of the inertia of the electrons which carry the current J_z. This can be shown from the $\hat{z}$ component of Ampere's Law, using (2.14) to eliminate J_z in favor of E_z:

$$k_x B_y = \frac{\omega_{pe}^2}{c^2} \cdot \frac{E_z}{\omega} \tag{2.17}$$

The electron mass and therefore the inertial effect is contained in the electron plasma frequency ω_{pe}. The term "kinetic Alfvén wave" is used to describe waves for which k_x is large enough so that electron dynamics along $\mathbf{B}_0$ affect the phase velocity of the slow mode, and allow for finite E_z as indicated in (2.17). Unfortunately, if we assume $\omega = 2\pi/(10$ minutes) as a representative auroral arc frequency, then inserting even "optimistic" values into (2.17) will not yield nearly enough of a parallel electric field to explain electron acceleration in auroral arcs. However, *C. Seyler* [personal communication, 1990] has suggested that the electrons associated with an arc drifting in the meridional direction with velocity V_d will experience an effective frequency $\omega = k_x V_d$. The electromagnetic skin depth c/ω_{pe} is representative of the smallest horizontal scales associated with auroral arcs (~ 1 km) so setting $\lambda_x = c/\omega_{pe}$ allows us to estimate an upper limit for E_z using (2.17),

$$E_z = V_d B_y \tag{2.18}$$

A typical arc drift in the ionosphere is a few hundred m/s, and when mapped up to the electron acceleration region it is near 1 km/s. Spacecraft routinely measure zonal magnetic fields of a few hundred nT. Thus a hundred μV/m parallel electric field is possible in a drifting arc. This electric field must be maintained over at least 10,000 km to obtain kV

potential drops, which is possible. At present the debate over the auroral acceleration mechanism is continuing, and the importance of Alfvén waves versus other mechanisms has yet to be established.

So far in this review of Alfvén waves we have assumed that the electrons are cold. However, this assumption breaks down when the thermal velocity of electrons v_{th} is fast enough so the electrons can shield parallel potential variations. This requires that $v_{th} >> V_A$, or equivalently $\beta >> m_e/m_i$. (β is the ratio of the electron pressure nkT_e to the magnetic field pressure $B_0^2/2\mu_0$.) When this is satisfied the electrons will adjust to the wave potential ϕ according to a Boltzmann distribution:

$$n_e = n_0 exp(e\phi/kT_e) \tag{2.19}$$

For small ϕ we can expand this and differentiate with respect to z to obtain

$$\frac{\partial n_e}{\partial z} = -n_0 eE_z/kT_e \tag{2.20}$$

We're interested in small horizontal scales (as we will see in the result), so we can neglect $k_z << k_x$ in the equation of current continuity

$$\nabla \cdot \mathbf{J} + \frac{\partial \rho_c}{\partial t} = 0 \tag{2.21}$$

which results in

$$-ik_x J_x = i\omega(\delta n_e)e \tag{2.22}$$

We can use the polarization current from (2.10) to eliminate J_x, and (2.20) to substitute E_z for δn_e, which gives

$$E_z = - k_x k_z \rho_i^2 E_x \tag{2.23}$$

where $\rho_i = c_s / \Omega_i$ is the ion gyro radius. Finally, this combined with (2.12a) gives the dispersion relation for kinetic Alfvén waves with $\beta >> m_e/m_i$:

$$\frac{\omega^2}{k_z^2} = V_A^2 \left(1 + k_x^2 \rho_i^2\right) \tag{2.24}$$

This dispersion relation was derived by *Hasegawa* [1977] and was applied to the aurora by *Goertz and Boswell* [1979]. Notice the similarity to the cold plasma dispersion relation in (2.15), with the ion gyro radius ρ_i replacing the electromagnetic skin depth c/ω_{pe} as the horizontal length scale at which the phase velocity is significantly different from the plane Alfvén wave case, and also the scale at which the parallel electric field becomes important. However, a large k_x causes an increase in the phase velocity ω/k_z in (2.24), but a decrease in (2.15).

To estimate the electric field let $k_x \rho_i = 1$ in (2.23), which corresponds to Alfvén waves with horizontal scales on the order of 1 km. In this case $E_z/E_x = k_z/k_x$. If we assume as before a drifting arc which generates in the arc frame a frequency $\omega = k_x V_d$ then $E_z/E_x = V_d/\sqrt{2} V_A$. Perpendicular electric fields of hundreds of mV/m, arc drifts of hundreds of m/s, and Alfvén velocities appropriate for a low density ($n_e \sim$ 100) hydrogen plasma can produce parallel electric fields on the order of 100 μV/m, just as in the cold plasma case. If Alfvén waves associated with drifting arcs are responsible for auroral electron acceleration, then the mechanism should be able to operate over a range of temperatures. In particular, we have shown that the cases $\beta = 0$ and $\beta >> m_e/m_i$ can generate similar parallel electric fields. However, in obtaining this

result we have assumed in the cold plasma case that $\lambda_x = c/\omega_{pe}$, and in the warm case that $\lambda_x = \rho_i$. (The cold approximation is probably more relevant for auroral acceleration.) For nominal parameters the proton gyroradius in the acceleration region is roughly a factor or 10 smaller than c/ω_{pe}, and one can compare (2.23) and (2.17) with the aid of the appropriate dispersion relations (2.15 and 2.24) to find $E_{z,cold}/E_{z,warm} \approx c^2/\omega_{pe}{}^2\rho_i{}^2$. Thus for a fixed horizontal scale the electric field in the cold plasma case is larger. We have not treated the intermediate case $0 < \beta < m_e/m_i$, but it is reasonable to assume that the magnitude of parallel electric fields from waves in this regime would fall somewhere between the cases we have studied.

2.6 Alfvén Waves and Magnetosphere-Ionosphere Coupling

The simplest (and often entirely sufficient) approach to M-I coupling comes from thinking of a voltage generator in the magnetosphere (e.g. the MHD generator from the solar wind-magnetosphere interaction) applying an electric field $\mathbf{E}_\perp$ which maps to the ionosphere and drives a Pedersen current in the ionosphere $\mathbf{J}_\perp = \sigma_P\mathbf{E}_\perp$ where σ_P is the Pedersen conductivity. (These relationships will be discussed in more detail in the next chapter). Any horizontal divergence of this current in the ionosphere must be closed by field-aligned currents to and from the magnetosphere as required by current continuity (see Section 3.6).

This simple model breaks down for many reasons. *Vickrey et al.* [1986] found that at horizontal scales between 3 and 80 km the magnetosphere acts more like a current source than a voltage source. This has important consequences concerning the boundary conditions appropriate for simulations of M-I coupling, as discussed by *Lysak* [1986].

Weimer et al. [1985] simultaneously measured electric fields at two altitudes with the DE 1 and 2 spacecraft and found that even for static electric fields, field line mapping in the collisionless region above the ionosphere is ineffective for horizontal scales below 100 km. This result stems from the presence of the parallel electric fields and the requirement that static electric fields must be curl-free. In this thesis we will discuss other complications in the simple coupling model above which result from neutral winds in the ionosphere and time varying electric fields.

Mallinckrodt and Carlson [1978] were among the first to realize that any changes in either the magnetospheric generator or ionospheric load must be communicated along magnetic field lines by Alfvén waves. They modeled the wave as it propagated towards the ionosphere through the changing refractive index which results from the Earth's dipole magnetic field and the increasing plasma density. The ionosphere is treated as a slab of conducting material characterized by its integrated conductivity Σ_P that causes a reflected wave resulting from the mismatch between Σ_P and the intrinsic impedance of Alfvén waves above the ionosphere. They also pointed out that incident and reflected Alfvén waves with small horizontal scales will only interfere close to the ionosphere, because plasma convection will carry the reflected wave away from the incident part.

Goertz and Boswell [1979] looked in detail at the front edge of an electric field pulse applied suddenly in the magnetosphere. They showed that such a pulse will propagate at the Alfvén velocity and that a parallel fringing field will exist at the leading edge of the pulse. If the pulse reflects enough times between the ionosphere and magnetosphere it can

repeatedly accelerate electrons. *Lysak and Carlson* [1981] showed that for Alfvén wave pulses with typical parameters, electrostatic ion cyclotron wave turbulence can cause large effective collision frequencies along **B** in the magnetosphere. *Lysak and Dum* [1983] included the effects of this turbulence in a time-dependent MHD simulation of Alfvén wave propagation and found that a magnetospheric region of wave turbulence can support parallel electric fields and decouple the magnetospheric generator from an ionospheric load. This decoupling may be a useful idea in explaining the fact that arcs are often found to drift at a velocity not equal to the **E**×**B** drift velocity in the ionosphere. *Lysak* [1986] further improved on these MHD simulations by dynamically changing the conductivity of the ionosphere as a result of energetic electrons incident from above, creating a feedback effect between the magnetosphere and ionosphere.

Tests of these competing and complimentary theories are suffering from an absence of experimental data pertaining to the relative importance of quasi-static and electromagnetic coupling. One of the primary goals of this thesis is to fill this experimental gap.

CHAPTER 3

MEASURING ENERGY COUPLING BETWEEN THE MAGNETOSPHERE AND HIGH-LATITUDE IONOSPHERE

3.1 Introduction

Energy is efficiently transferred between the solar wind-magnetosphere system and high-latitude ionosphere because of the geomagnetic field $\mathbf{B}_0 = B_0\hat{z}$. The magnetic field facilitates energy exchange in two ways. First, in the absence of electric fields perpendicular to $\mathbf{B}_0$ the field lines constrain charged particle motion to the $\hat{z}$ direction, and in this chapter we will show, in agreement with previous work (e.g. *Foster et al.* [1983]), that the resulting field-aligned flux of kinetic energy integrated over the high-latitude region can exceed 10^{10} W. The second mode of energy transfer is electrical. Magnetic field lines act as "wires" that carry electrical current to the ionosphere, where Joule heating in the dense, partially ionized medium dissipates the incident energy. *Vickrey et al.* [1982] showed that although the daily averages of the energy flux from particles and Joule heating are comparable in the auroral oval, there is a tendency for the two to be anticorrelated. Based on Chatanika incoherent scatter radar measurements, those authors found the morning sector (westward electrojet) particle energy deposition rate to be generally larger than that in the pre-midnight sector eastward electrojet. The Joule heating rate has the opposite asymmetry about midnight.

This tendency for anticorrelation has one straightforward explanation. Where the particle flux is relatively energetic, as it is in the morning sector as compared to the evening sector, ionization is produced at lower altitudes. At these altitudes the conductivity tensor is such that electric fields and associated currents are mostly perpendicular to each other, which limits the Joule dissipation $\mathbf{J}\cdot\mathbf{E}$. (We will discuss the conductivity tensor in detail in the next section.) When precipitating electrons are less energetic, they increase the plasma density at somewhat higher altitudes, where $\mathbf{J}$ and $\mathbf{E}$ are more nearly parallel. Thus, less kinetic energy deposition leads to more Joule dissipation [*Kelley et al.*, 1990].

This interrelationship is likely more than coincidental. It is reasonable to expect that the ionosphere/magnetosphere system operates in a feedback mode. For example, Joule heating requires field-aligned currents, and when these currents exceed a certain threshold various plasma waves can become destabilized. The waves may then convert electrical to kinetic energy through wave/particle interactions, which might explain the anticorrelation between kinetic and electrical energy deposition rates in the ionosphere.

The possibility for such an important interrelationship between electrical and kinetic energy input to the ionosphere is one of the reasons for the study we have conducted. In addition, on a global scale Joule heating is thought to be larger than particle energy deposition because it is spread over a wider range of latitudes. In the summer polar cap, for example, when B_z is southward, considerable Joule heating occurs while any particles that precipitate are generally soft. It is crucial therefore to

develop remote sensing techniques to determine the Joule heat input to the upper atmosphere.

Sensing of particle precipitation is straightforward and is regularly performed by polar orbiting spacecraft. The electromagnetic input, on the other hand, is not routinely monitored and its estimation usually requires severe approximations such as neglect (or very simplistic modeling) of the atmospheric wind; e.g. *Vickrey et al.* [1982]. In this chapter we show that the electromagnetic energy flux into the atmosphere can be reliably measured remotely by polar orbiting spacecraft at altitudes in the range 400-1000 km using the vertical component of the Poynting flux, and we present examples of its measurement using both the HILAT satellite and a sounding rocket. Since this measurement is of a local quantity, no assumptions are required concerning the relative orientation of the spacecraft velocity and current sheets such as are needed in determination of Birkeland currents. Moreover, knowledge of the ionospheric conductivity and conductivity gradients are not necessary for the measurement of the energy input.

The concept of the Poynting flux as a diagnostic tool in the study of time-varying electromagnetic waves is well established. As discussed by *Feynman et al.* [1964], under certain circumstances the Poynting flux provides a valid conceptual measure of energy flow even for steady or DC electric and magnetic fields. In Section 3.2 we give a brief derivation and a discussion of the concept as applied to geophysical systems. We show that a local measurement of $(\delta \mathbf{E}_\perp \times \delta \mathbf{B}_\perp)/\mu_0$ at typical ionospheric satellite altitudes yields the local electromagnetic power input to the Earth's atmosphere. It is important to note that no geometric assumptions are

necessary to find this quantity, unlike those required to determine, say, J_z from magnetic field measurements along a trajectory. That is, although $\mu_0 J_z$ can in principle be found from $\partial B_y/\partial x$ - $\partial B_x/\partial y$, one must in practice neglect the y derivative for a satellite moving in the x direction. The Poynting flux technique can also be used to detect energy flow *out* of the Earth's ionosphere, which can occur where a neutral wind dynamo is present.

While measurements of kinetic and electrical energy flow between the magnetosphere and ionosphere are not new, the various methods used have their own advantages and problems which until recently have not been carefully compared. In this chapter, Sections 3.3 and 3.4 are devoted to a discussion of two methods for measuring the electrical energy flux at high latitudes. Later in the chapter we will present energy flow calculations using data from the HILAT satellite and from a sounding rocket. Our purpose will not be to present an exhaustive survey of energy flow measurements from spacecraft, but rather we will develop and compare techniques for doing so. We begin with a discussion of the theory behind these methods.

3.2 Techniques for Measuring Electromagnetic Energy Flow Between the Ionosphere and Magnetosphere

A satellite such as HILAT, with the capability to measure electric and magnetic fields simultaneously, can be a very useful tool to monitor the rate of electromagnetic energy flow into or out of the ionosphere at the magnetosphere-ionosphere interface. However, as with any measurement of a physical parameter, the process is imperfect and we must anticipate the various complications and errors which will arise.

We will begin with a simple model of the magnetosphere/ionosphere energy exchange. We limit our discussion to high latitudes, so we can assume that the geomagnetic field $\mathbf{B}_0$ is vertical. In this chapter we will assume that the energy sources are constant in time, but we will relax this requirement in Chapter 4. And we will first treat the case in which there is no neutral wind.

In the static case, the amount of electromagnetic energy generated or dissipated in some volume V is the Joule dissipation:

$$W = \iiint_V \mathbf{J}\cdot\mathbf{E}\, dV \tag{3.1}$$

The plasma fluid equations allow us to relate the current density $\mathbf{J}$ and electric field $\mathbf{E}$ through the conductivity tensor $\boldsymbol{\sigma}$:

$$\mathbf{J} = \boldsymbol{\sigma}\cdot\mathbf{E} \tag{3.2}$$

where

$$\boldsymbol{\sigma} = \begin{pmatrix} \sigma_P & \sigma_H & 0 \\ -\sigma_H & \sigma_P & 0 \\ 0 & 0 & \sigma_0 \end{pmatrix} \tag{3.3}$$

and

$$\sigma_0 = \varepsilon_0 \sum_j \frac{\omega_{pj}^2}{\nu_j} \tag{3.4a}$$

$$\sigma_P = \varepsilon_0 \sum_j \frac{\nu_j \omega_{pj}^2}{\nu_j^2 + \Omega_j^2} \tag{3.4b}$$

$$\sigma_H = \varepsilon_0 \sum_j \frac{\Omega_j \omega_{pj}^2}{\nu_j^2 + \Omega_j^2} \tag{3.4c}$$

j is a species index and σ_0, σ_P, and σ_H are known as the direct (or specific), Pedersen, and Hall conductivities respectively. We will derive a generalized version of σ in Chapter 4 which allows for time variation of field quantities.

Although σ_0 is generally much larger than σ_P or σ_H , the field-aligned electric field E_z is much smaller than E_x and E_y for the range of physical parameters that are of interest to us, and we can safely neglect the dissipation term $J_z E_z$. One consequence of the small electric field parallel to $\mathbf{B}_0$ is that the perpendicular fields are approximately constant along $\mathbf{B}_0$. This allows us to carry out the z integration in (3.1) to obtain

$$W = \iint_A \Sigma_p \left(E_x^2 + E_y^2\right) dx\, dy \tag{3.5}$$

where

$$\Sigma_p = \int_{ionosphere} \sigma_p\, dz \tag{3.6}$$

Equation (3.5) can be used to estimate the energy dissipated in the ionosphere. Since the electric field maps along $\mathbf{B}_0$, it can be measured either by satellites or sounding rockets in or above the ionosphere, or by high altitude balloons since the electric field maps into the lower atmosphere [*Mozer and Serlin*, 1969]. Σ_P is more difficult to determine since it requires altitude profiles of the plasma and neutral atmosphere densities. Ionospheric plasma at high latitudes can be produced by electron precipitation from the magnetosphere in an unpredictable and time dependent manner, and once produced the plasma can quickly convect away. Thus without a direct measurement from an incoherent

scatter radar, for example, estimates of Σ_P are limited by poor knowledge of the plasma density.

Poynting's theorem can be used to eliminate the need for a Σ_P estimate in ionospheric energy deposition measurements as long as the magnetic field can be measured at the same time as the electric field. We give a formal derivation of Poynting's theorem and an example of its application to energy dissipation in a simple resistor in Appendix A. Poynting's theorem states that

$$W = \iint_A \mathbf{S} \cdot d\hat{s} = \iiint_V \left(\mathbf{J} \cdot \mathbf{E} + \mathbf{E} \cdot \frac{\partial \mathbf{D}}{\partial t} + \mathbf{H} \cdot \frac{\partial \mathbf{B}}{\partial t} \right) dV \tag{3.7a}$$

where $\mathbf{S} = (\mathbf{E} \times \mathbf{B})/\mu_0$, the vector $d\hat{s}$ is normal to an element of surface area and points into the volume V everywhere, $\mathbf{D}$ is the electric flux density, and $\mathbf{H}$ is the magnetic field intensity. In the case of DC energy flow, $\partial/\partial t = 0$ and thus

$$W = \iint_A \mathbf{S} \cdot d\hat{s} = \iiint_V \mathbf{J} \cdot \mathbf{E} \, dV \tag{3.7b}$$

For our application to the problem of magnetosphere-ionosphere coupling, we first consider the volume enclosed by the surface of the Earth and a "cap" covering all latitudes above say 50 degrees. The cap is located at an altitude that is not crucial but which is between 400 and 1000 km. (Below,we assume that it is the HILAT satellite's orbital altitude of 800 km.) This height is chosen to be high enough that particle collisions are rare but below any region of significant field-aligned electric fields associated with the aurora.

We assume that the zonal component of the perpendicular electric field goes smoothly to zero at the boundary of this region, that the magnetic field lines are everywhere vertical, and we ignore curvature of the magnetic field lines over this height range. The volume of interest, shown in Figure 3.1, is then bounded by the high altitude cap, S_1, the surface of the Earth, S_2, and the surface S_3 linking the cap and the Earth. Since the Earth is a good conductor the electric field vanishes on S_2 and the Poynting flux is zero across it. If no thunderstorms are located near the boundary then we can assume that the fair weather electric field is vertical and the Poynting flux across S_3 is also zero. This implies that the entire electromagnetic power dissipated in the volume may be found by integrating the Poynting flux across S_1. Since S_1 is perpendicular to $\mathbf{B}_0$, the power input to the Earth's atmosphere in the high latitude zone is given by

$$W = \frac{1}{\mu_0} \iint_{S_1} \left(\mathbf{E}_\perp \times \delta\mathbf{B}_\perp\right) \cdot d\hat{s} \tag{3.8}$$

where $\mathbf{E}_\perp$ is the perpendicular electric field on S_1 and $\delta\mathbf{B}_\perp$ is the deviation of the total magnetic field from the undisturbed value in the plane perpendicular to $\mathbf{B}_0$.

We now argue that the cross product of these two quantities gives the local value of the energy flow rate into the atmosphere. Consider an infinitesimal element of the surface S_1 and the volume it subtends between S_1 and the Earth. The contribution at the Earth vanishes as before. Since we know that the flow in the high altitude ionosphere is incompressible, it follows that the integral of $(\mathbf{E}_\perp \times \delta\mathbf{B}_\perp) \cdot d\hat{s}$ over the sides of the volume vanishes there. Furthermore, since large scale DC electric fields map without distortion along field lines into and through the E

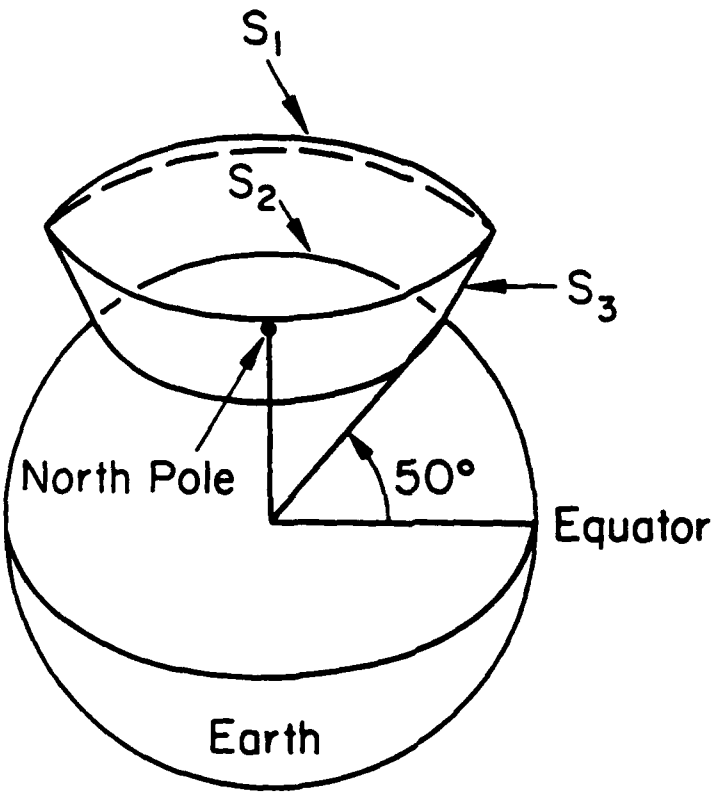


Figure 3.1 S_1 is the surface through which magnetospheric Poynting flux **S** enters the upper atmosphere. S_2 is the Earth's surface, which is a good conductor and therefore requires $E_\perp = 0$ and hence $\mathbf{S} = 0$. Assuming that the magnetic field lines are straight and vertical, and the fair weather electric field is vertical implies that $\mathbf{S} \cdot \hat{n}_3 \equiv 0$ where $\hat{n}_3$ is a unit vector normal to the surface S_3.

region and vertically deep into the atmosphere [*Mozer and Serlin*, 1969], where they go smoothly to zero near the Earth's surface, the integral of $(\mathbf{E}_\perp \times \delta\mathbf{B}_\perp)\cdot d\hat{s}$ vanishes on these edges as well.

This discussion of the amount of Poynting flow across various surfaces of the volume in Figure 3.1 is necessary because we need to establish that by integrating the Poynting flux across the high altitude surface S_1 we are actually measuring a *divergence* in the Poynting flux -- i.e. Poynting flux is entering the ionosphere from above, but it can't "escape" from the sides (S_2) or through the Earth's surface (S_3). A single measurement of Poynting flux alone can be meaningless. For example, imagine placing a net electrical charge on the surface of the Earth, which would cause a radial electric field. At the magnetic equator the Earth's dipole field $\mathbf{B}_0$ crossed into the radial electric field gives a non-zero Poynting vector $\mathbf{S}$. However, if we integrate $\mathbf{S}$ over any closed surface we find no divergence and therefore no net dissipation or generation of electrical power, as we can see from the definition of $\mathbf{S}$ with the aid of a vector identity:

$$\mu_0 \nabla\cdot\mathbf{S} \equiv \nabla\cdot(\mathbf{E}\times\mathbf{B}_0) = \mathbf{B}_0\cdot(\nabla\times\mathbf{E}) - \mathbf{E}\cdot(\nabla\times\mathbf{B}_0) \tag{3.9}$$

In the case at hand $\nabla\times\mathbf{E} = 0$, $\nabla\times\mathbf{B}_0 = \mu_0\mathbf{J} = 0$, and $\nabla\cdot\mathbf{S} = 0$. Only when there is a non-zero current $\mathbf{J}$ can there be a net divergence of Poynting flux for a steady state system. We will show with a specific example in Section 3.6 that Poynting flux into the ionosphere is associated with Birkeland (field-aligned) currents that close as Pedersen currents.

3.3 The Role of the Neutral Winds in Energy Flow

Thus far we have assumed that all of our measurements are made in the Earth-fixed reference frame and that the neutral wind is zero. We

have presupposed an externally applied electric field and studied the energy dissipated in a conducting ionosphere. If the neutral atmosphere is in motion however, it can also be a source of electric current and Poynting flux. In fact if we examine the case with an external medium acting as a passive load with an effective height-integrated conductivity Σ_E, a wind generated current source yields an upward Poynting flux above the ionosphere, and the integral of $(\mathbf{E}\times\delta\mathbf{B})\cdot d\hat{s}$ over the surface of the external load is equal to the volume integral of $\mathbf{J}\cdot\mathbf{E}$ within the load. In this case $\mathbf{J}\cdot\mathbf{E} < 0$ in the ionosphere, indicating that the neutral wind in the ionosphere is acting as a generator and supplying electrical energy to the external load. Even when the ionosphere is acting as a load, the neutral wind affects its interaction with any external energy sources. We will look at two examples of an ionospheric load with a neutral wind.

For the first example consider an external generator which applies an electric field E_G across geomagnetic field lines. The generator is not ideal and has associated with it an internal conductance which we model as a thin strip extended in y characterized by σ_G (mho/m) such that the potential across it and current density through it are related by $V\sigma_G = I$ (A/m), as shown in Figure 3.2. Magnetic field lines connect the generator to an ionospheric load region with height-integrated conductivity Σ_P and constant neutral wind U_y. We will treat the ionosphere as a simple slab, with $\Sigma_P = 0$ outside of the gray region. The current density in the $\hat{x}$ direction is given by $J_x = \sigma_P(E_x + U_yB_0)$ (A/m^2) in the ionosphere, and in the generator by $(E_x - E_G)d\sigma_G$ (A/m) where d is the width of the system in the $\hat{x}$ direction. Current continuity requires

$$I = |\mathbf{I}| = \Sigma_P(E_x + U_yB_0) = (E_G - E_x)d\sigma_G \text{ (A/m)} \qquad (3.10)$$

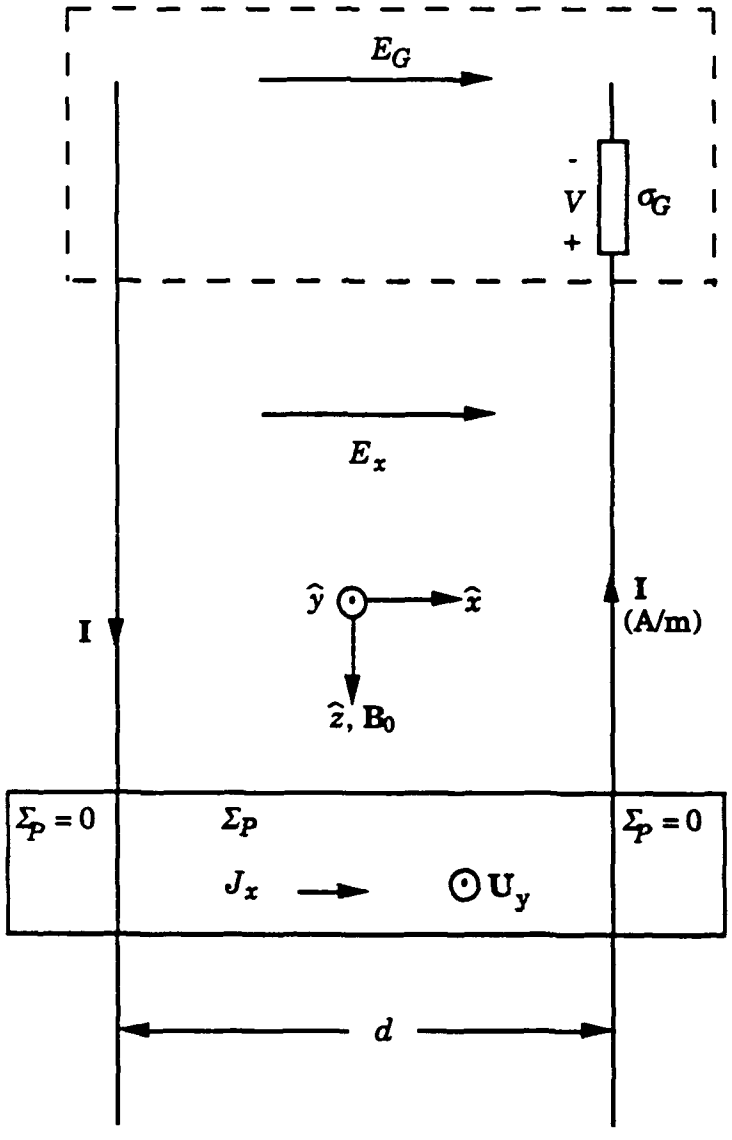


Figure 3.2 Cross-section in the meridional plane of a simplified model illustrating the effect of neutral winds on ionospheric dissipation of magnetospherically applied electric field energy.

with the direction of **I** shown in the figure. The perturbation magnetic field at the center of the diagram, which is assumed to be very far from the cross field currents at either end, is $\delta\mathbf{B} = \mu_0 I\hat{y}$. The Poynting flux measured in the Earth-fixed frame is thus $\mathbf{S} = \mathbf{E}\times\delta\mathbf{B}/\mu_0 = \Sigma_P E_x(E_x + U_y B_0)\hat{z}$.

With (3.10) we can solve for $E_x = (dE_G\sigma_G - U_y B_0\Sigma_P)/(\Sigma_P + d\sigma_G)$ which allows us to write the Poynting vector as

$$\mathbf{S} = \hat{z}\left(\frac{d\sigma_G\Sigma_P}{d\sigma_G + \Sigma_P}\right)^2\left(\frac{E_G}{\Sigma_P} - \frac{U_y B_0}{d\sigma_G}\right)(E_G + U_y B_0) \tag{3.11}$$

Clearly for large σ_G the external source will dominate and the Poynting flux will be downward. Likewise for $E_G = 0$, electrical energy generated by the neutral wind will flow upward.

The $\hat{z}$ component of the Poynting vector is equal to the height-integrated value of $\mathbf{J}\cdot\mathbf{E}$ in the ionosphere

$$\int_{ionosphere} \mathbf{J}\cdot\mathbf{E}\, dz = \Sigma_P E_x(E_x + U_y B_0) \tag{3.12}$$

$\mathbf{J}\cdot\mathbf{E}$ is the Joule dissipation when measured in the frame of the neutral wind, but in the present case we have written the electrodynamic quantities in a frame with a non-zero wind. We will now discuss the meaning of the quantity $\mathbf{J}\cdot\mathbf{E}$ in the presence of a neutral wind.

As long as the wind velocity is non-relativistic, the frame in which the magnetic field is measured does not matter. This can be seen from the equations which transform electric and magnetic fields from the neutral wind frame (primed coordinates) to the Earth-fixed frame (not primed), where the neutral wind velocity is **U**:

$$\mathbf{E}' = \mathbf{E} + \mathbf{U}\times\mathbf{B} \tag{3.13a}$$

$$\mathbf{B}' = \mathbf{B} \tag{3.13b}$$

We have neglected $\mathbf{U}\times\mathbf{E}/c^2 << \mathbf{B}$. A discussion of these transformations can be found in Chapter 2 of *Kelley* [1989]. Equation (3.13b) implies that $\mathbf{J}' = \mathbf{J}$. With these transformations we can write the Joule dissipation $\mathbf{J}'\cdot\mathbf{E}'$ in the Earth-fixed frame as $\mathbf{J}\cdot(\mathbf{E} + \mathbf{U}\times\mathbf{B}_0)$. A vector identity allows us to write this as

$$\mathbf{J}\cdot\mathbf{E} = \mathbf{J}'\cdot\mathbf{E}' + \mathbf{U}\cdot(\mathbf{J}\times\mathbf{B}_0) \tag{3.14}$$

hence in a frame where the neutral wind is non-zero $\mathbf{J}\cdot\mathbf{E}$ can be interpreted as the Joule heat plus the work done on the neutral wind by the $\mathbf{J}\times\mathbf{B}$ force.

To summarize, in the presence of a neutral wind the Poynting flux measured by a satellite can determine the amount of energy flowing into or out of the ionosphere. If the neutral wind is acting as a generator, some of the energy will be dissipated by the ionosphere and the Poynting flux will measure only the net outward energy flow. Nonetheless, it is this very exchange of energy which is crucial in the study of magnetosphere-ionosphere coupling.

A potentially confusing aspect of Poynting flux measurements is the fact that their value depends on the reference frame in which they are measured. From Equations (3.13) we can see that the magnetic field does not change with reference frame but the electric field depends on the velocity of the measuring platform. We will consider one additional example which will help to clarify the reason for this.

Consider two ionospheres in conjugate hemispheres connected by the geomagnetic field $\mathbf{B}_0$. Figure 3.3 shows the configuration with the

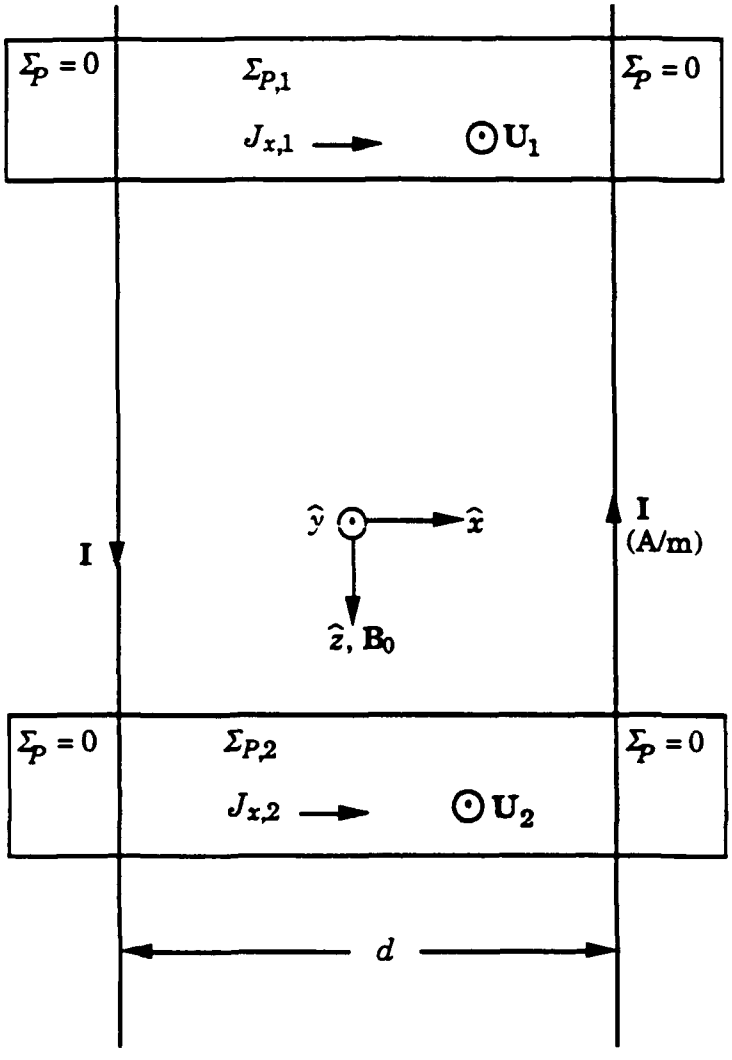


Figure 3.3 Cross-section in the meridional plane of a simplified model illustrating the interaction of neutral wind dynamos in conjugate hemispheres.

field lines straightened into a rectangular geometry. The two ionospheres are labeled "1" and "2" and each has its own height-integrated Pedersen conductivity $\Sigma_{P,j}$ and neutral wind $U_{y,j}$ where j indexes the ionospheres. The conducting region is limited in the $\hat{x}$ direction to a width d, and outside of this region we assume zero conductivity. All quantities extend infinitely in the $\hat{y}$ direction as before.

The neutral winds will carry plasma across geomagnetic field lines and in general will create some electric field E_x which will in turn drive currents in both ionospheres given respectively by

$$J_{x,j} = \sigma_{P,j}(E_x + U_{y,j}B_0) \tag{3.15}$$

Current continuity at the edges of the conducting troughs requires two infinite sheet currents

$$\mathbf{I} = -\Sigma_{P,1}(E_x + U_{y,1}B_0)\hat{z} = \Sigma_{P,2}(E_x + U_{y,2}B_0)\hat{z} \tag{3.16}$$

where again $\mathbf{I}$ has units of A/m. This allows us to calculate the electric field as a function of the U_y:

$$E_x = -\frac{\left(\Sigma_{P,1}U_{y,1} + \Sigma_{P,2}U_{y,2}\right)B_0}{\Sigma_{P,1} + \Sigma_{P,2}} \tag{3.17}$$

Using this with (3.16) allows us to solve for magnitude of the sheet currents:

$$|\mathbf{I}| = \frac{\Sigma_{P,1}\Sigma_{P,2}}{\Sigma_{P,1} + \Sigma_{P,2}}(U_{y,2} - U_{y,1})B_0 \tag{3.18}$$

Electric fields and winds cause the ionospheric plasma to collide with neutral atmosphere particles which results both in increased temperature and changes in the bulk velocity of the neutral wind. The electromagnetic coupling between the two ionospheres will act to cause

the neutral winds to be equal eventually. If enough time passes so that they become equal, no current will flow between the two hemispheres, as we can see from (3.17).

In the case for which $U_{y,1} = 0$, the E_x generated by $U_{y,2}$ will act to accelerate the neutral wind in ionosphere "1" and the resulting drag on $U_{y,2}$ will slow it down. In the Earth-fixed frame it is natural to think of "2" as a generator and "1" as a load. Of course this is entirely dependent upon reference frame. If we happen to be traveling along with the neutral wind in "2" then the now non-zero wind in "1" will cause $U_{y,2}$ to accelerate, and the role of generator and load are reversed.

We will now calculate the Poynting flux that would be measured by a spacecraft flying between the two ionospheres. The magnetic perturbation caused by the current sheets is $\delta \mathbf{B} = \mu_0 I \hat{y}$ and thus (assuming $\Sigma_{P,1} = \Sigma_{P,2}$ for simplicity) the resulting Poynting vector $E_x \delta H_y \hat{z}$ can be written

$$\mathbf{S} = \frac{\Sigma_P \left(U_{y,1}^2 - U_{y,2}^2 \right) B_0^2}{4} \hat{z} \tag{3.19}$$

Equation (3.19) tells us that if one of the ionospheres has zero neutral wind then the Poynting vector is directed towards it. This is consistent with the fact that a zero velocity neutral wind also has zero bulk kinetic energy. Thus any plasma motion driven by electric fields must increase the kinetic energy of the neutrals, and the neutral atmosphere acts as a load. The neutral wind in the conjugate hemisphere acts as a generator and the bulk kinetic energy of the neutral wind there decreases.

Of course in the case of an ionospheric load not all of the Poynting flux energy increases the bulk velocity of the neutrals, and likewise in a neutral wind generator not all of the kinetic energy supplied by the wind

appears as Poynting flux. In both cases the neutral gas and the plasma are heated. This leads us to interpret the Poynting vector measured above the ionosphere in the Earth-fixed frame as the rate at which electromagnetic energy from the magnetosphere is causing the neutral wind in the ionosphere (also measured in the Earth-fixed frame) to change its kinetic energy per unit area plus the rate at which that energy is heating the atmosphere:

$$\frac{1}{\mu_0}\left(\mathbf{E}_\perp \times \delta\mathbf{B}_\perp\right) = \frac{\partial}{\partial t}\left[\int_{ionosphere} \frac{\rho U^2}{2}\, dz + \Delta Q\right] \tag{3.20}$$

where ΔQ is the heat per unit volume supplied *to* the atmosphere and ρ is the mass density of the neutrals. We are neglecting the kinetic energy of the plasma since its density is many orders of magnitude less than the neutral density in the ionosphere. The Poynting vector is positive when directed towards the ionosphere.

The conclusion we can draw from the above discussion is that the Poynting vector $\mathbf{E}_\perp \times \delta\mathbf{B}_\perp/\mu_0$ measured in the Earth-fixed frame can be used to detect whether the neutral atmosphere below the measurement platform is gaining energy (i.e. is load-like) or losing energy (generator-like) *in the Earth-fixed frame*. Since the kinetic energy of the neutral wind depends on the velocity squared, its time rate of change depends on the neutral wind velocity, thus the time derivative of the kinetic energy and the Poynting vector are frame-dependent.

There are two situations which can cause $\mathbf{S} \approx 0$, namely $\mathbf{E}_\perp = 0$ and $\delta\mathbf{B}_\perp = 0$. In the first case there is still energy exchange between the ionosphere and a conjugate ionosphere or the magnetosphere if we move to a different frame. If $\delta\mathbf{B}_\perp = 0$ there is no energy exchange in *any* frame.

In this sense magnetic perturbations associated with field-aligned currents are more indicative of energy coupling to the ionosphere than are electric fields.

A disadvantage that Poynting flux measurements have is that the perturbation magnetic field $\delta\mathbf{B}_\perp$ is usually found by subtracting a model of the geomagnetic field $\mathbf{B}_0$ from the total field $\mathbf{B}$ measured by the satellite magnetometer. Since $\mathbf{B}_0$ is roughly 40,000 nT and perturbations from field-aligned currents are typically several hundred nT, small errors in the model can greatly affect $\delta\mathbf{B}_\perp$ measurements. However, model errors manifest themselves as DC (or low frequency) offsets in $\delta\mathbf{B}_\perp$, thus in the following section we will ignore the Poynting flux from low-frequency fields by high-pass filtering the data.

3.4 Satellite Observations of Kinetic Energy and Poynting Flux

In this section we present examples of Poynting flux measurements in the high latitude ionosphere. The instruments used were not optimized for measurement of this parameter and yet the results are quite reasonable. We believe that when interpreted as discussed in the previous section that these examples support the idea that Poynting flux is a useful diagnostic quantity, and we hope that other researchers pursue this concept with more sensitive instruments and better behaved measurement platforms. Since the electric fields detected are considerably larger than UB_0 in the high latitude E region we will ignore neutral wind effects.

The first two examples come from the HILAT satellite. HILAT is a polar-orbiting, real-time satellite which means it cannot store data on board. Thus measurements are available only when the satellite is in

view of a ground receiving station. The data we will present were recorded at the station in Sondrestrom, Greenland (67° N), so each pass is centered about this geographic latitude. HILAT orbits at about 800 km with a velocity of 7.4 km/s, and it is in view of a ground station for about 10 minutes per pass, depending on its elevation angle at closest approach.

HILAT measures the vector magnetic field using a fluxgate magnetometer sampled at 20 s^{-1} and deduces the electric field from a cross-track ion drift meter switching between 16 and 32 s^{-1}. We have averaged the data to 1.5 s^{-1}. These and other instruments on HILAT are described in detail elsewhere [*Potemra et al.*, 1984; *Rich et al.*, 1984]. Since the satellite is in a high-inclination orbit (81°), its orbital velocity (defined to be in the $\hat{x}$ direction) is mostly meridional. $\hat{z}$ is taken to be downward, and $\hat{y}$ completes the right-hand system in HILAT coordinates. The cross-track ion drift allows one to calculate the meridional electric field. The in-track drift component yields the cross-track (zonal) electric field but is only available once per second since it requires a sweep of the retarding potential analyzer. In the auroral zone, the zonal electric field is generally much smaller than E_x and adds only a small correction to the Poynting flux found from the meridional electric field alone.

The spacecraft is gravity gradient stabilized but suffers attitude perturbations from thermal stress. Examples of the magnetic field data which we have used in the two satellite orbits presented here are given in Figure 3.4a and show the attitude problem very clearly. The upper panel shows magnetic field data obtained on Day 164, 1984. The sinusoidal modulation of the signal is due to one of the unfortunate attitude

HILAT Magnetic Field Data

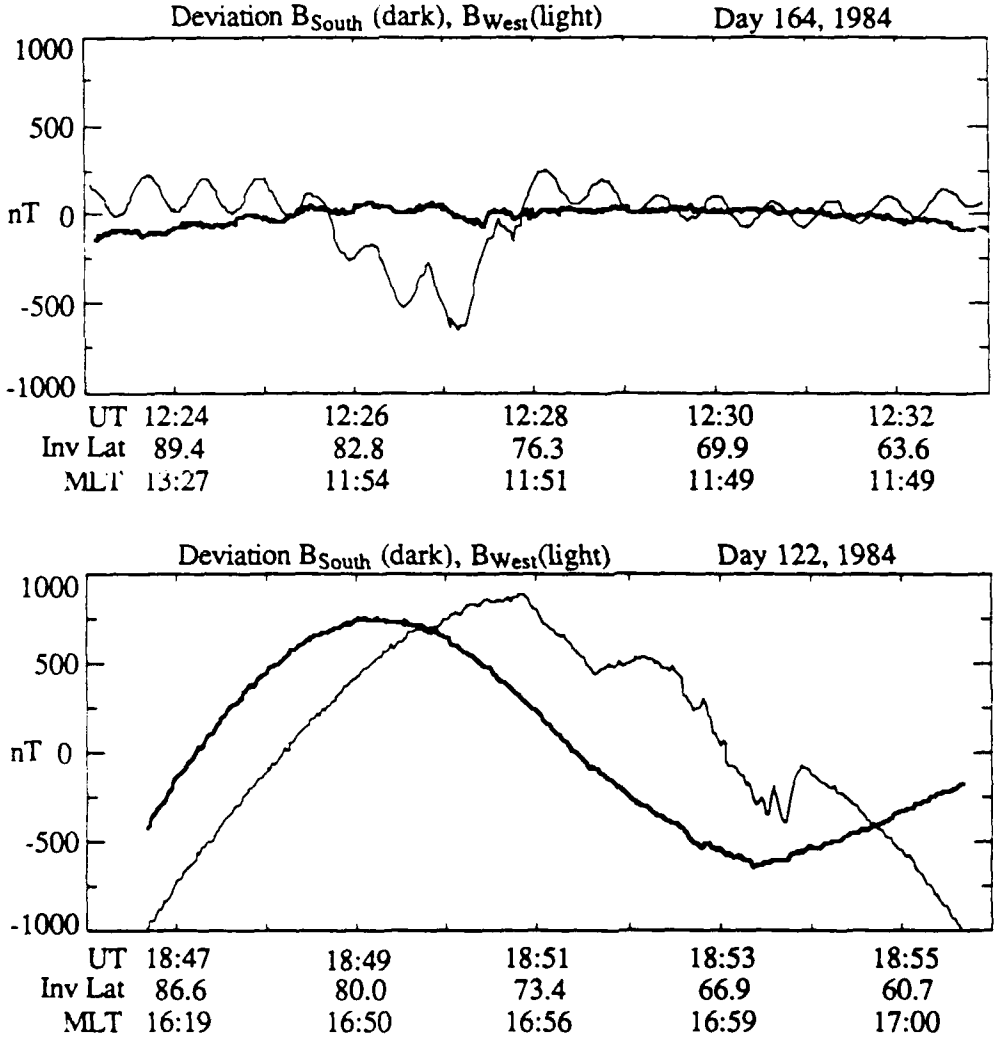


Figure 3.4a Meridional and zonal magnetic field perturbations for two HILAT passes. The 40 second variation in the Day 164 data and the longer period variation in the Day 122 data are due to attitude oscillations of the satellite. Superimposed on these are clear signatures of large-scale field-aligned currents.

oscillation modes of the spacecraft. Another mode is clearly seen in the second panel using data obtained on Day 122 of the same year. Here a very long period attitude oscillation is seen in the signal.

It is clear from both satellite passes that signals of geophysical significance are present. On both days the spacecraft passes through large scale regions of field-aligned current as ascertained from the derivative of the magnetometer signal, ignoring the sinusoidal oscillations of the satellite. In the analysis below we have filtered the signals to remove these perturbing influences. A notch filter (order 20 digital Butterworth) with a center frequency of .0286 Hz and a bandwidth of 0.01 Hz was used for Day 164; a high pass filter allowing only frequencies above 0.0029 Hz to contribute was used for both days. This necessary filtering precludes measurement of the largest scale size Poynting flux input to the high latitude system.

We will see that even after filtering out the lowest frequencies in the HILAT data, the Poynting flux is still mostly downward, which is what we would expect for an ionospheric load. One reason that filtering does not destroy the Poynting flux measurement is that the satellite is above the auroral oval for only a fraction of the entire pass, therefore perturbations associated with the auroral oval (and which represent most of the Poynting flux) are above the filter cutoff. The fractional orbit acquisitions from a real time satellite system such as HILAT are not suited for fully global measurements, and we are therefore restricted to studies such as auroral oval crossings. The electric field after filtering for these two passes is presented in Figure 3.4b.

The Poynting flux measured on Day 164 is presented in Figure 3.5a along with the field-aligned current and precipitating electrons in two

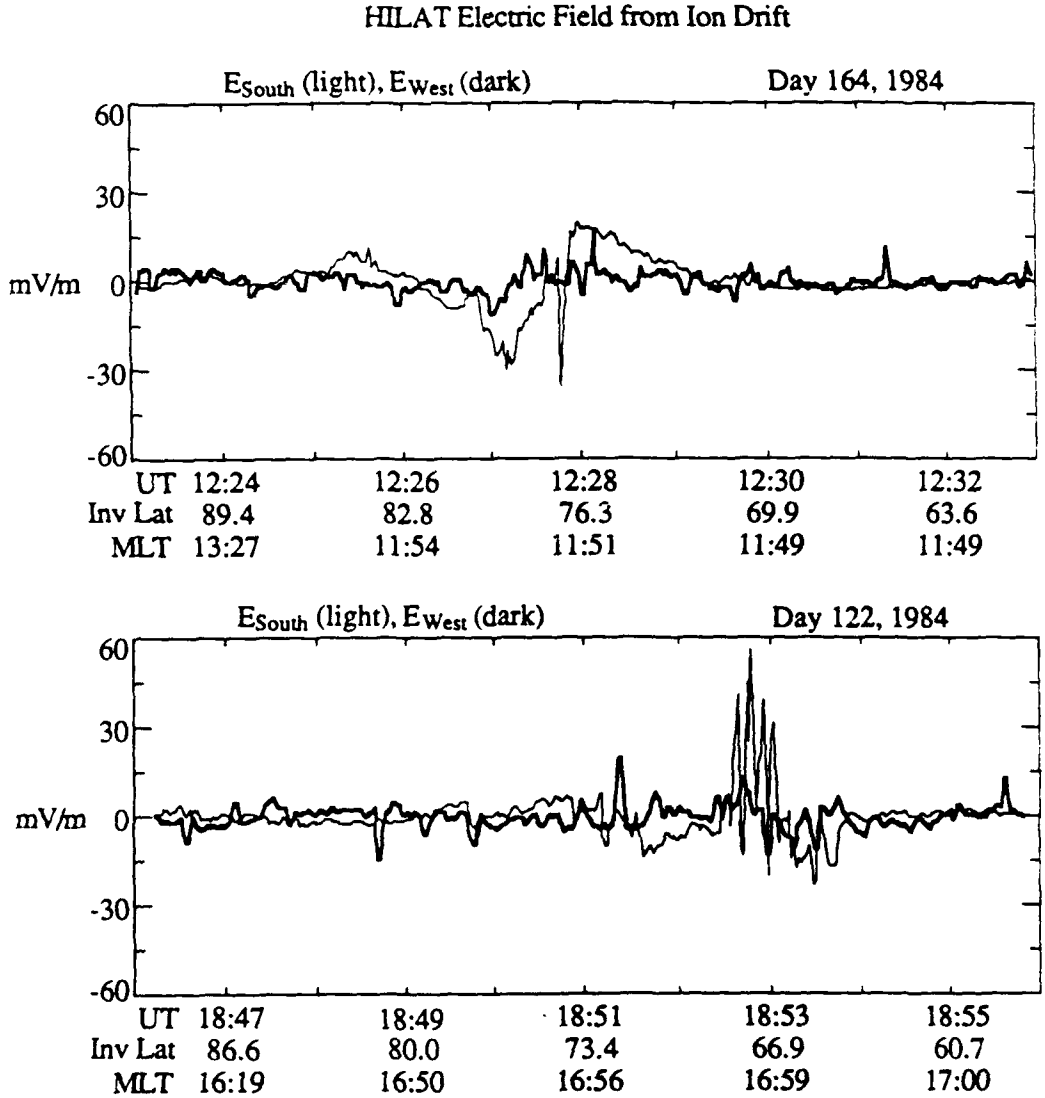


Figure 3.4b Meridional and zonal electric fields for the two HILAT passes shown in Figure 3.4a.

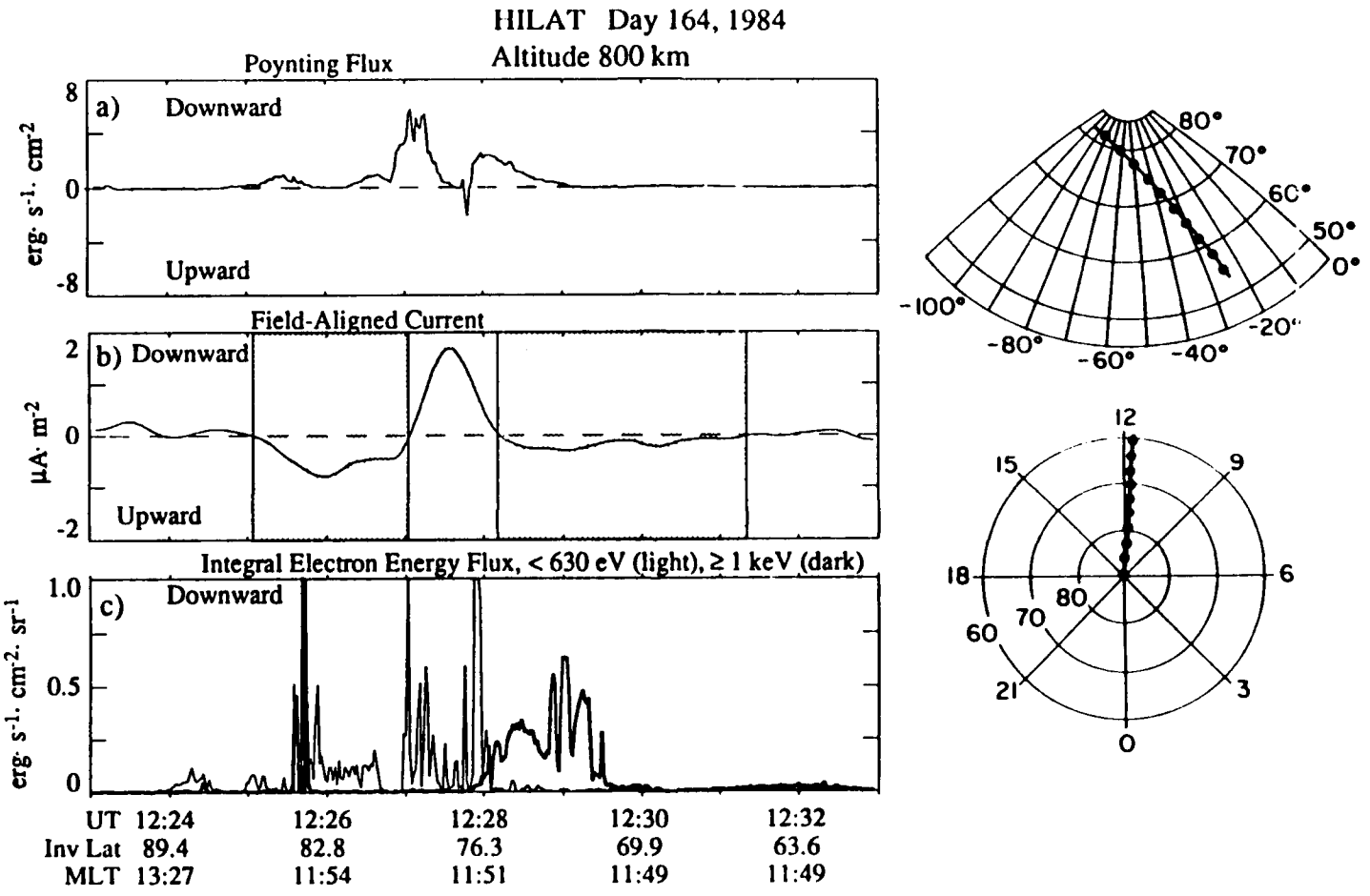


Figure 3.5 Comparison of electromagnetic and particle energy inputs into the ionosphere and the large-scale field-aligned current structure for a summer noon descending pass.

energy bands. The latter are plotted positive for downward energy flow since the detector looks upward. The inset shows the pass in a magnetic local time invariant-latitude format. The satellite was acquired in the polar cap and passed over the dayside auroral oval just before local noon.

We have plotted the Poynting flux in ergs/(cm^{2}s) to conform to the usual notation in presenting particle fluxes in the aurora. The power flux is almost entirely downward throughout the pass, with the single exception of a brief burst of upward flux near 12:28 UT. The average vertical electrical power flux during the pass is equal to 0.45 ergs/(cm^{2}s) (or 0.45 mW/m^2) downward, while an upper limit for the average kinetic power flux due to the electrons was 0.70 ergs/(cm^{2}s). To obtain this value a distribution of downgoing electrons isotropic over 2π steradians was assumed. Comparison of the 45° and vertical electron sensors (not shown here) indicate this is a reasonable assumption. Integrating along the trajectory yields 2100 W/m (electromagnetic) and 3300 W/m (kinetic energy). To give some perspective we can estimate the total power from both sources into the entire auroral oval region by assuming that the energy input is independent of local time. This yields 7.9×10^{10} W. This value is a lower limit since we cannot determine the Poynting flux at the largest scales.

Figure 3.5b is the field-aligned current derived from $\partial B_y/\partial x = \mu_0 J_z$ assuming that variation in the $\hat{y}$ direction is unimportant. Because a derivative is required some smoothing has been necessary. We restrict attention to the three shaded current sheets in the center of the figure and not the small variations outside this region as they may be due to the filter. The existence of three sheets is quite common in the noon sector [*Iijima and Potemra*, 1976]. The upward current sheet at invariant

latitudes below about 72° is co-located with fairly hard electron precipitation, as shown in Figure 3.5c, as well as convection toward the noon meridian. The Poynting flux near the central (downward) current sheet was greater than the precipitating electron energy flux, even assuming that the down-going particle energies are distributed over 2π steradians.

It is interesting to note that the flux of soft electrons in the region of downward current is anticorrelated with the measured current density. Furthermore the precipitating electrons carry current of the opposite sign to that measured. A lower limit to the current carried by the soft electrons can be estimated by assuming that the perpendicular energy of the electrons is small so that they all fall within the aperture of the detector (6° by 4°, or 7.3×10^{-3} sr). The current nev_e due to a 1 erg/(cm^{2}s·sr) flux can be found by multiplying by the angular area of the aperture and dividing by the estimated average energy of the electrons. The current caused by, say, 10 eV electrons is 0.7 μA/m^2. The implication is that the upward thermal electron flux must have been fairly large or a considerable ion precipitation was occurring to counter the upward current from the soft electron precipitation.

The Day 164, 1984 orbit was such that the ionosphere was sunlit over the entire trajectory. By taking into account the solar depression angle, the electron density and the conductivity of the E region can be determined [*Robinson and Vondrak*, 1984]. Although it is not particularly important in this case, we have also estimated the contribution of particle precipitation to the conductivity by assuming that the observed electron flux has been present long enough for a steady state electron density profile to be reached. With this estimate for Σ_P and the

observed electric field from the ion drift meter we can estimate the Joule heating in the ionosphere and compare it to the Poynting flux as shown in Figure 3.6. The ratio of the Joule heating rate to Poynting flux magnitude in Figure 3.6c shows that the two quantities coincide roughly within a factor of two. The Joule dissipation estimate relies on a model for the neutral atmosphere to compute collision frequencies. This model in turn uses an estimate of the thermospheric temperature, but in general this parameter is difficult to determine, and this uncertainty is one possible explanation for the deviation from unity of the Poynting Flux to Joule heat ratio.

The ratio is "spikey" in places, which is to be expected when dividing two noisy quantities, but we will show in Section 3.6 that the spike near 12:27:45 UT coincides with a burst of temporally varying fields, which invalidates one of the assumptions allowing us to use the Joule dissipation in Equation 3.5. Poynting flux, on the other hand, is a valid way to measure energy fluxes in electromagnetic waves. It is important to note that in this case the Poynting flux is upwards, which means that the ionosphere is either reflecting or generating instead of dissipating energy, and this is not discernable from the Joule heating calculation alone. (Since we cannot measure Poynting flux on the largest scales it is possible that the measured upward Poynting flux is actually an upward perturbation on a large scale downward flux.) For these reasons and those discussed in Section 3.3 we argue that Poynting flux as a tool is superior to Joule heating estimates. However, we must qualify this statement for HILAT satellite measurements since the magnetometer resolution is about 15 nT. For small signal levels the quantization noise

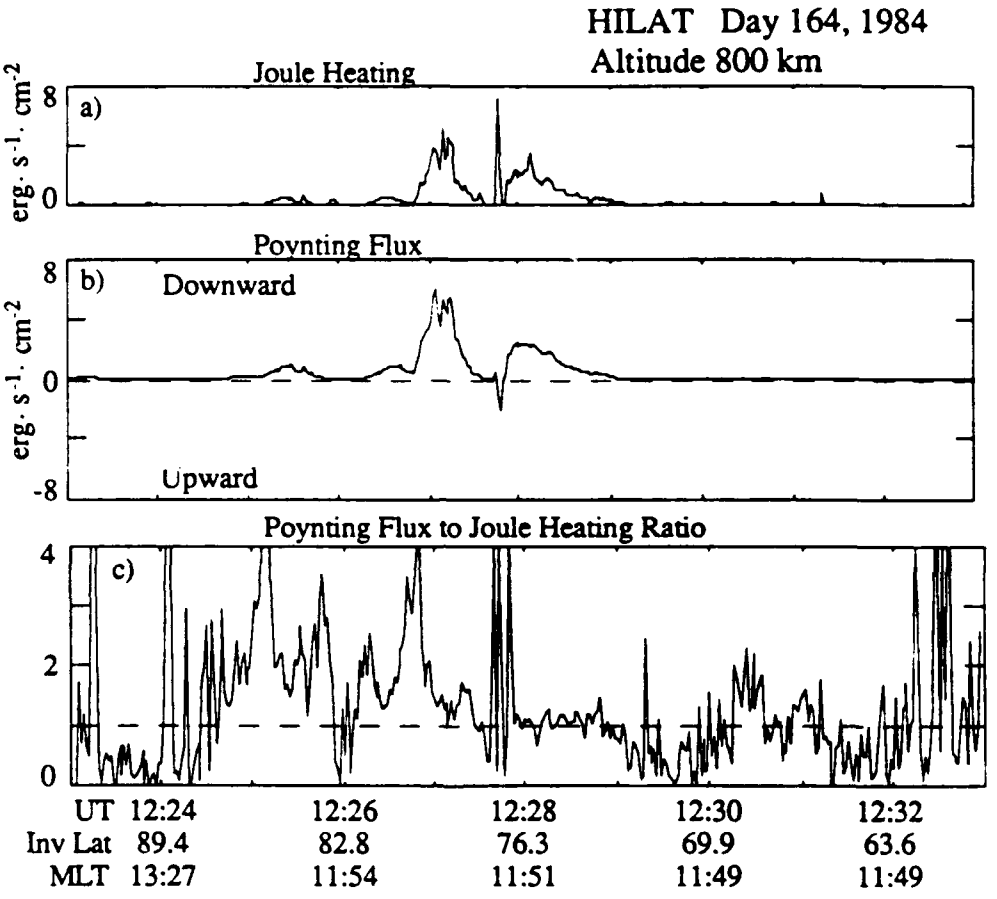


Figure 3.6 Comparison of the Joule heating rate and Poynting flux for the HILAT pass shown in Figure 3.5.

can cause an anomalously large magnetic field (and therefore Poynting flux) estimate.

Data from the second event are presented in Figure 3.7 in a format nearly identical to that used in Figure 3.5. As can be seen in the raw data in Figure 3.4a and in the smoothed Birkeland currents in the second panel, several current sheets were detected during this dusk pass through the auroral oval. This is unusual, at least as far as the literature indicates. All significant upward currents are co-located with a burst of electron precipitation and anti-sunward convection while downward current regions were associated with sunward flow and no precipitation. There was still significant Poynting flux in regions where the particle input was low and the Birkeland currents downward. In the central downward current sheet, regions of both upward and downward Poynting flux were found. Figure 3.8 shows this effect in an expanded plot of the meridional electric field, the Poynting flux, and the electron energy flux for the period 18:52-18:54 UT. The magnitude of the average electromagnetic power density over the entire pass was 0.22 ergs/(cm^2s); the rate of kinetic energy input was 2.0 ergs/(cm^2s). The integrated values over the pass are 980 W/m and 8900 W/m. Again we assumed an angular spread of 2π steradians in the kinetic energy based on vertical and 45° electron energy measurements. Assuming no variation of energy input with local time in this case gives 1.5×10^{11} W for the total electrical and mechanical power input into the auroral oval.

In Figure 3.9 we plot the Joule heating rate, the Poynting flux, and the ratio of the two quantities for Day 122, 1984. There appears to be good agreement except near 18:53 UT, where the Joule heating estimate is much larger than the Poynting flux. For the periods in which magnetic

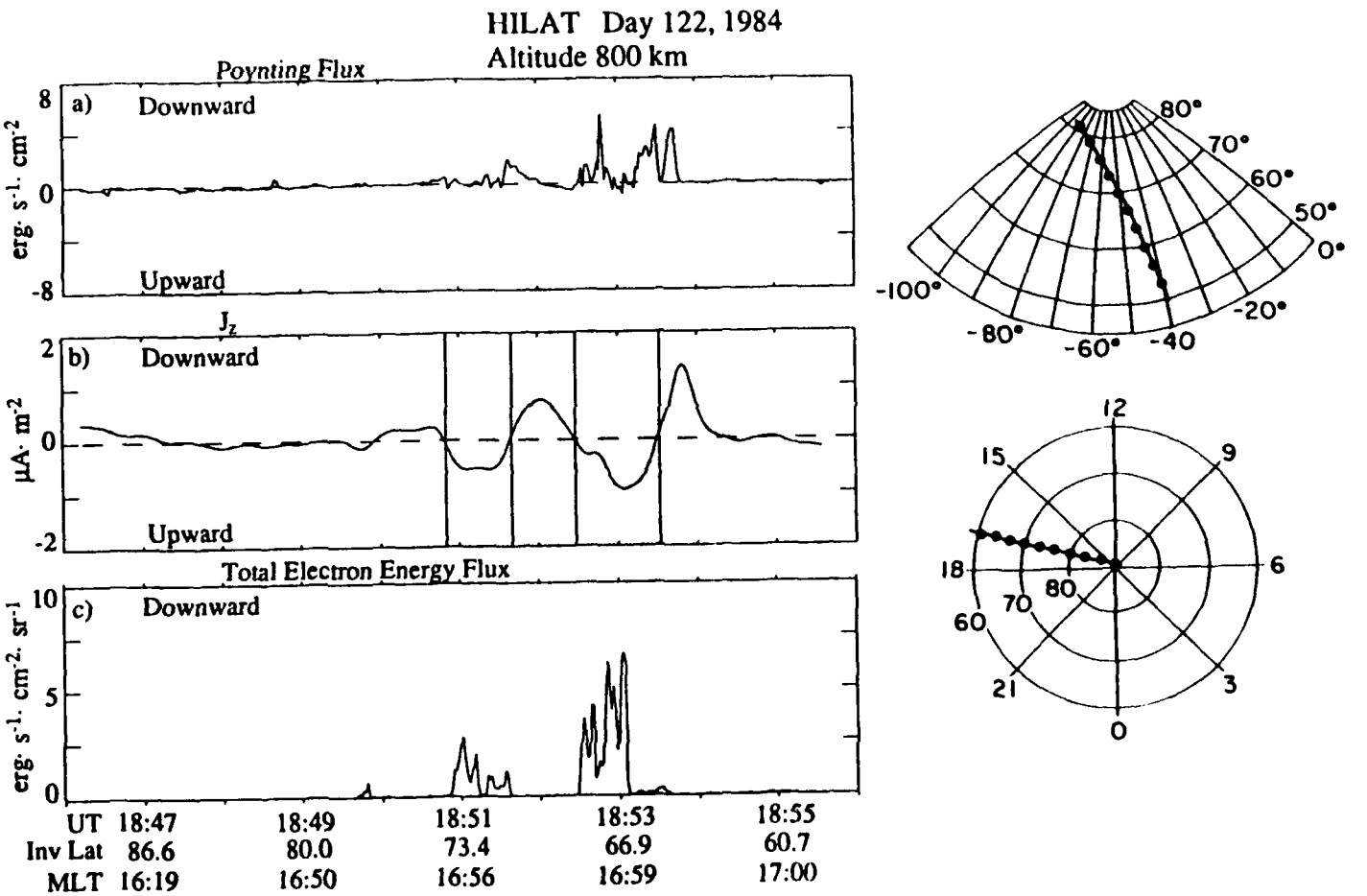


Figure 3.7 Comparison of electromagnetic and particle energy inputs into the ionosphere and the large-scale field-aligned current structure for a HILAT pass through the afternoon sector.

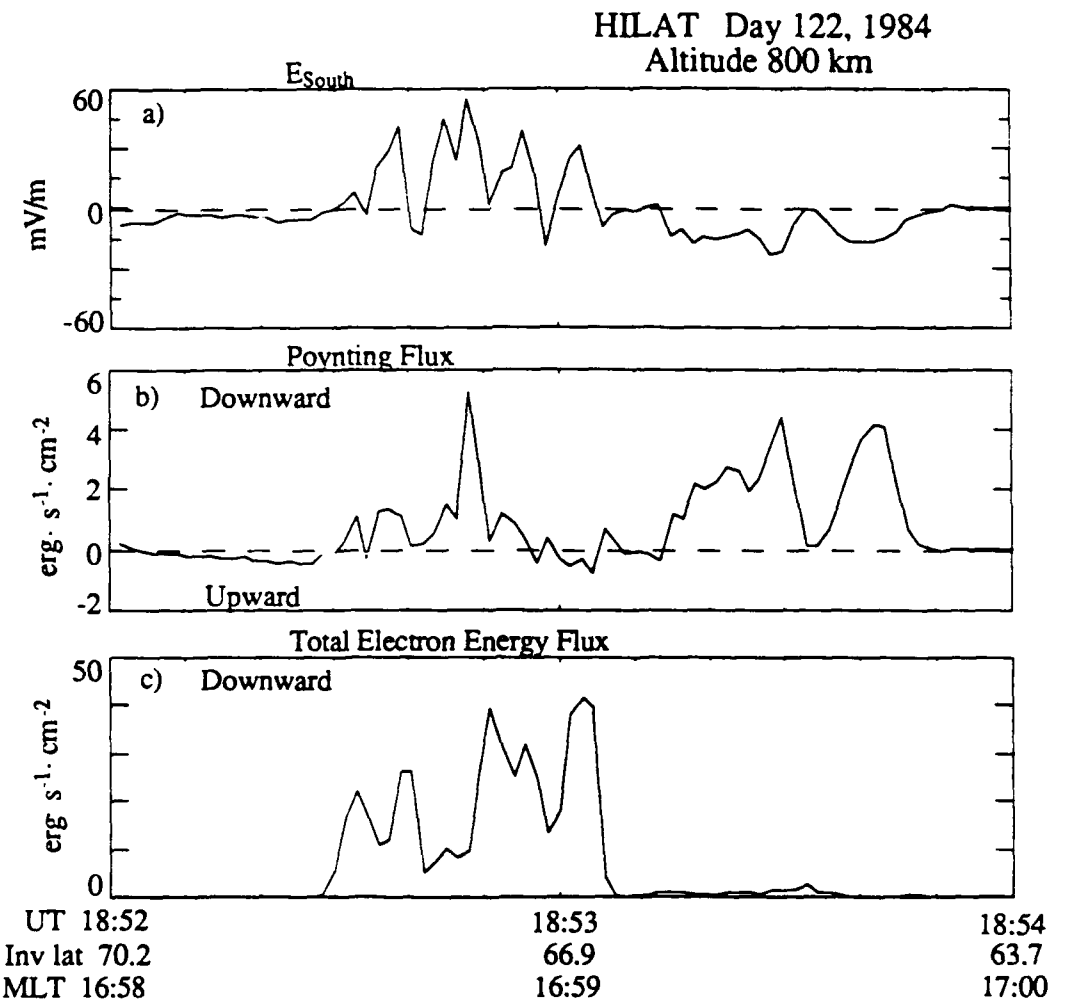


Figure 3.8 Expanded view of a two minute interval during the HILAT pass on Day 122, 1984 during which upward Poynting flux was observed.

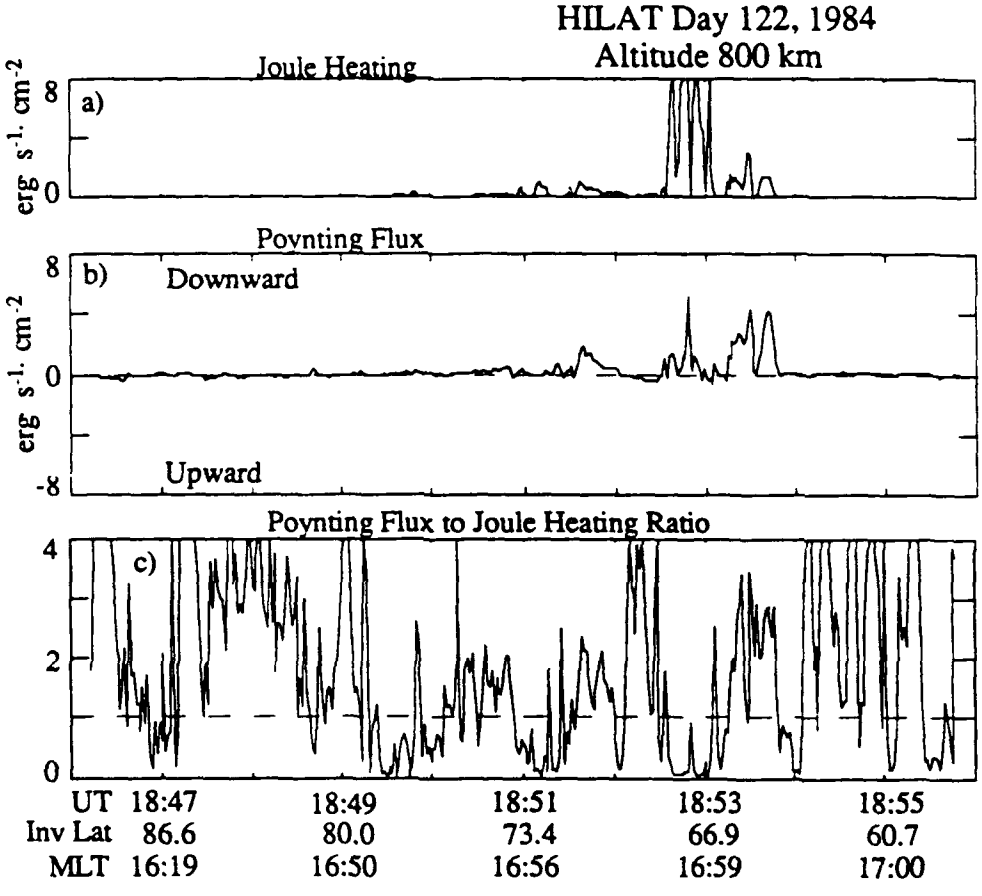


Figure 3.9 Comparison of the Joule heating rate and Poynting flux for the HILAT pass shown in Figure 3.7.

field fluctuations exceed the minimum resolution of the magnetometer, we attribute the disagreement between the two measurements mainly to errors in the conductivity estimate from the particle flux and hence in the Joule heating estimate.

3.5 Sounding Rocket Observations

As part of the 1985 NASA Greenland I campaign a Black Brant X sounding rocket was launched from Sondrestrom eastward into the dayside auroral oval and remained inside the oval for the entire upleg. Other results from that flight have been published by *Boehm et al.* [1990]. The measurements we present were taken during the upleg of the rocket flight at altitudes between 400 and 770 km. Electric fields perpendicular to $\mathbf{B}_0$ were measured with perpendicular 3 m electric field booms, and magnetic measurements were taken with a fluxgate magnetometer. To obtain electric fields below the rocket spin frequency, electric field measurements were fit to a sine wave at the spin frequency, then averaged to obtain two measurements per rocket spin period. The resulting sample period is 0.887 s. The measured $\delta\mathbf{E}$ and $\delta\mathbf{B}$ fields perpendicular to $\mathbf{B}_0$ for the ascending half of the flight are shown in the top two panels of Figure 3.10.

To illustrate the amount of electromagnetic power flowing between the magnetosphere and ionosphere during the flight we plot the field-aligned component of the Poynting vector in Figure 3.10c. We must be especially wary of the Poynting flux estimate in this case because 1) small errors in the geomagnetic field model can cause large errors in estimates of the perturbation magnetic fields used in the Poynting flux calculations, and 2) we cannot distinguish magnetic field perturbations

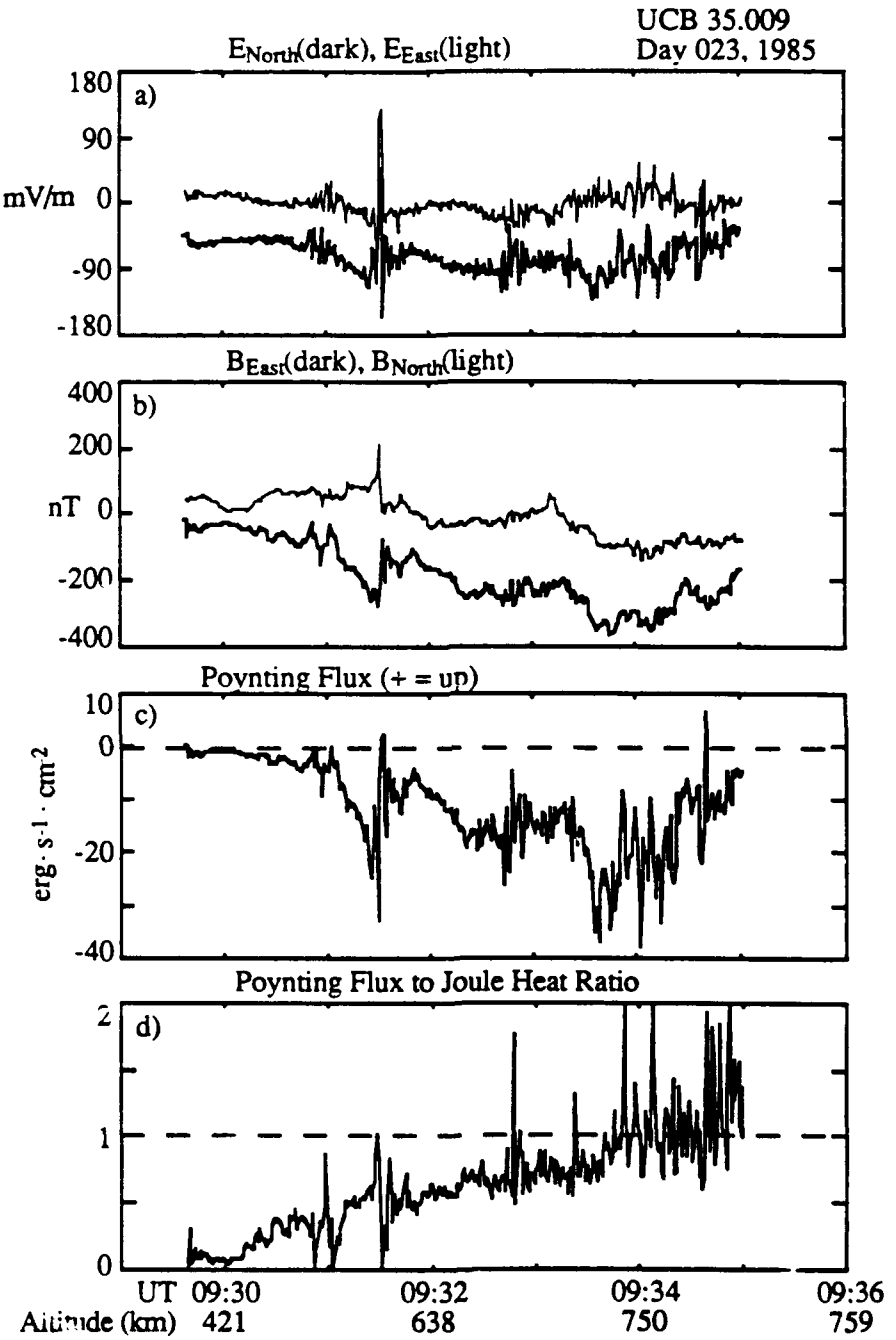


Figure 3.10 Data taken from a Black Brant X sounding rocket launched from Sondrestrom, Greenland on 23 January, 1985. The rocket traveled eastward along the auroral oval. (Data are courtesy of C. Carlson, B. McFadden, and M. Boehm at the University of California, Berkeley.)

due to large scale current systems from an error in the zero order magnetic field model. We can detect the associated electric field however since there is no zero order electric field. This effect almost certainly explains the large difference between the Poynting flux and $\Sigma_P E_\perp^2$ early in the flight where $\mathbf{E} \neq 0$ but the model subtraction yields $\delta\mathbf{B} \approx 0$. The high-pass filter technique we used for the satellite data does not work in this case, because the data are taken completely within the auroral oval. Recall that in the satellite case the auroral oval crossing was only a fraction of the total duration of the pass, causing it to lie in the filter passband. We still have some confidence in the validity of the rocket-measured Poynting flux, however, because it agrees reasonably well with $\Sigma_P E_\perp^2$ as we can see from the Poynting flux to Joule heat ratio shown in Figure 3.10d. The slow trend causing this ratio to increase over the course of the flight could be due to an error in the perturbation magnetic field estimate, or to the fact that we used a constant value of $\Sigma_P = 3$ mhos throughout the flight. The rocket flew eastward into regions of increasing sunlight, which would cause Σ_P to increase throughout the flight. But due to the general agreement between the two power flow estimates we will assume that the perturbation magnetic fields are not too contaminated by low frequency field model errors. Notice that the magnitude of the Poynting flux is consistently tens of ergs/(cm^{2}s), which is several times larger than the peak power fluxes from the two satellite passes discussed in the previous section.

The Poynting flux is predominantly downward except for 2 short intervals, near 09:30:30 and 09:34:40 U.T. In Chapter 5 we will show that much of the electric field energy during the flight is dominated by standing Alfvén waves. Standing waves can produce both upward and

downward Poynting flux during different phases of their cycle, and this could possibly account for the observed upward Poynting flux.

3.6 Time-Domain Measurements of Auroral Field Impedances

We have shown that Joule dissipation in the ionosphere implies electric and perturbation magnetic fields perpendicular to $\mathbf{B}_0$ above the ionosphere. In order to understand the origin of the perturbation fields it is helpful to consider a simple model of current closure through the ionosphere. In this model we assume that fields and ionospheric parameters such as density and collision frequencies vary in the meridional ($\hat{x}$) direction only. As we will show in the next chapter, no variation in the zonal ($\hat{y}$) direction implies that the zonal electric field E_y is much smaller in magnitude than the meridional field E_x, and we will therefore neglect it. If we apply a meridional electric field $E_x(x)\hat{x}$ above the ionosphere, that field will map into the ionosphere and drive a current $\mathbf{J} = J_x\hat{x} + J_y\hat{y}$. Since $\partial/\partial y = 0$ the current continuity equation $\nabla\cdot\mathbf{J} = 0$ can be written

$$J_z = \frac{\partial}{\partial x}\int_{ionosphere} J_x\, dz \tag{3.21}$$

where J_z is the field-aligned current above the ionosphere. Ampere's law in the region above the ionosphere is $\mu_0 J_z = \partial B_y/\partial x$, which we can apply to (3.21) to obtain

$$S_z = E_x B_y/\mu_0 = \int_{ionosphere} J_x E_x\, dz \tag{3.22}$$

We have assumed E_x is constant in altitude, and the constant of integration over x is taken to be zero to ensure that $S_z = 0$ in the absence of

Joule dissipation. As we might expect, the $\hat{z}$ component of the Poynting vector gives the height-integrated Joule dissipation per unit area in the ionosphere. The point we wish to make with this example is that a Poynting vector measured *above* the ionosphere correctly measures the Joule dissipation *in* the ionosphere because the zonal perturbation magnetic field B_y is caused by field-aligned currents which close through the ionosphere.

In addition to Joule dissipation in the ionosphere, the electromagnetic fields above the ionosphere can be used to predict another useful quantity. Replacing J_x in (3.22) with $\Sigma_P E_x$ gives the following result:

$$\frac{E_x}{B_y} = \frac{1}{\mu_0 \Sigma_P} \tag{3.23}$$

This allows us to remotely estimate the height-integrated Pedersen conductivity of the ionosphere directly below the spacecraft. Again, we have assumed no variation in the zonal ($\hat{y}$) direction, negligible E_y, quasi-static fields, and no neutral winds.

Notice that we were able to eliminate the x derivative in (3.21), which means that no assumptions are necessary concerning the scale length of variations in x. Whether we measure fields associated with an auroral arc or with the entire auroral oval, (3.23) holds. However, we will show in the next chapter with the aid of a numerical model that (3.23) is violated at scales less than a few km, due to the fact the small scale electric fields do not map without attenuation along the geomagnetic field.

In Figures 3.11 and 3.12 we plot $\mu_0 |E_x/B_y|$ along with those same values multiplied by Σ_P as a function of time for the two satellite passes

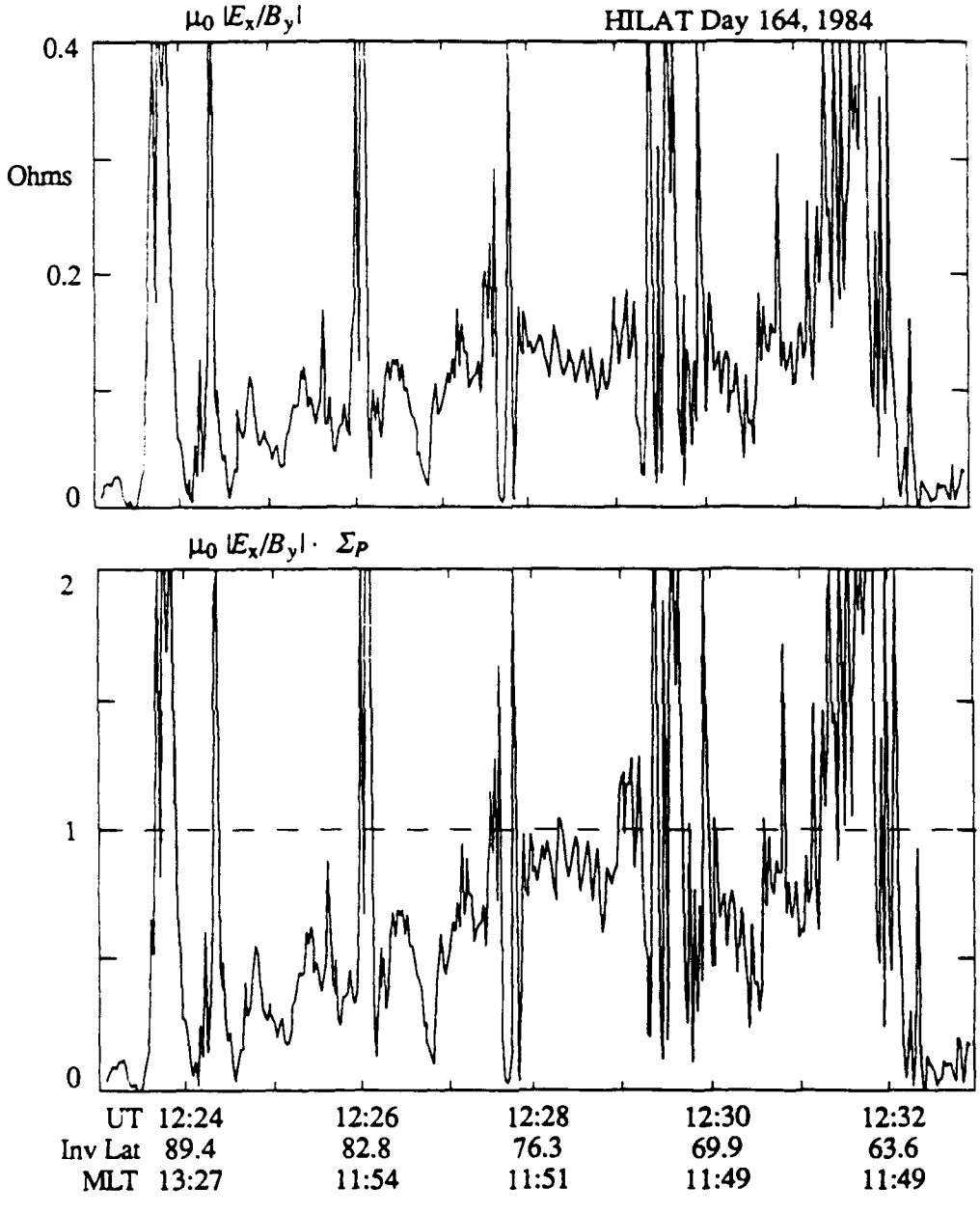


Figure 3.11 Electromagnetic field impedance as a function of time calculated from the Day 164, 1984 data. The impedances in the lower panel are normalized to Σ_P^{-1}.

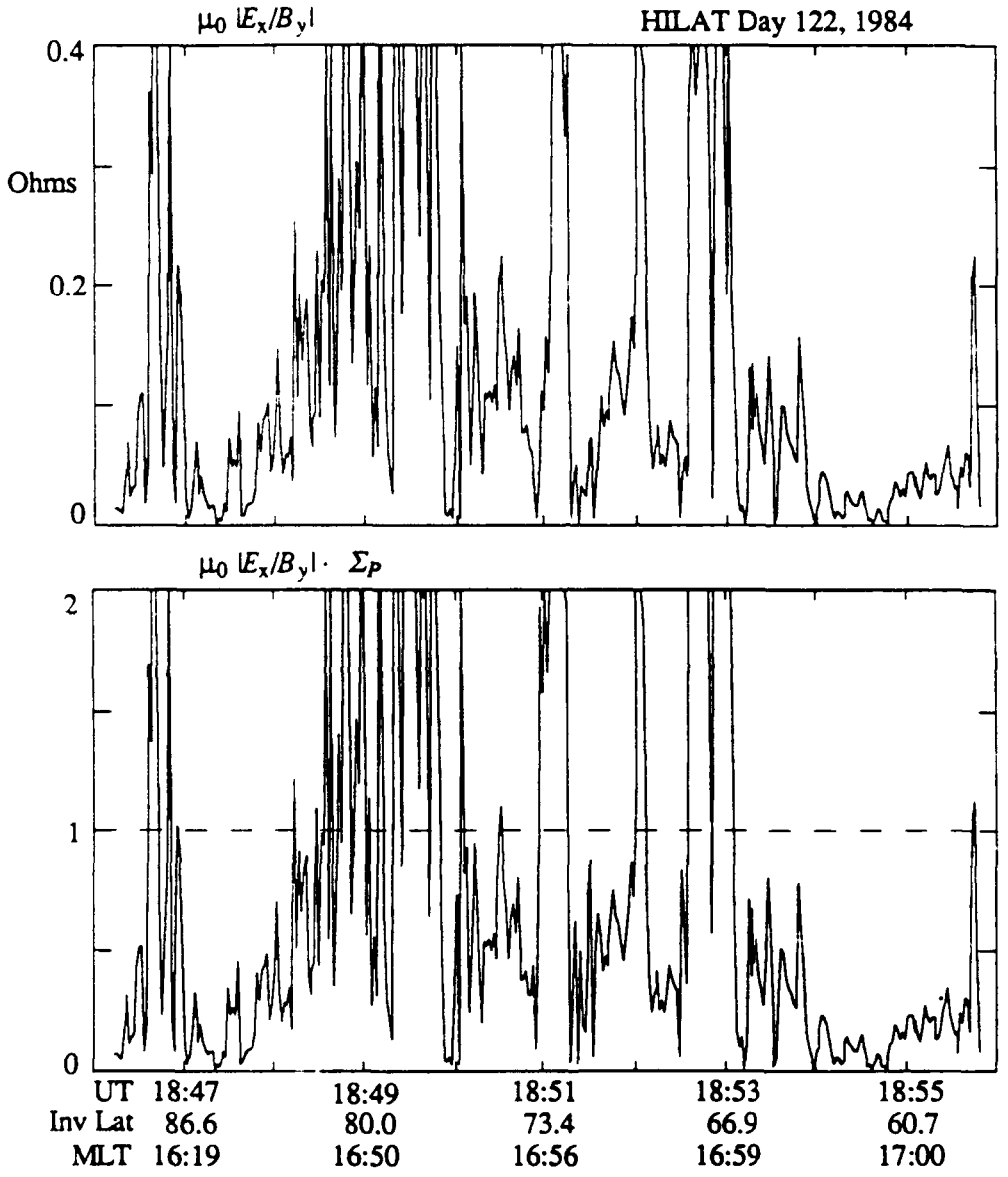


Figure 3.12 Electromagnetic field impedance as a function of time calculated from the Day 122, 1984 data. The impedances in the lower panel are normalized to Σ_P^{-1}.

discussed in the previous section. Deviations from unity in Figures 3.11b and 3.12b indicate either that our Σ_P estimates are in error, that one or more of the assumptions leading to Equation (3.23) are violated, or that the magnetic field fluctuations are below the resolution of the instrument. In one of the cases we can identify the exact cause of the deviation. Figure 3.13 shows the meridional electric field plotted at the full time resolution of the instrument for the time period near 12:27:45 UT on Day 164, 1984. The electric field shows a wave-like burst with a peak amplitude of over 100 mV/m. The coherent nature of the burst is indicative of temporal variation. Thus the spike in Figure 3.11b just before 12:28 UT can be attributed to a breakdown in our assumption of static fields.

A similar increase in $\mu_0 |E_x/B_y|$ over Σ_P^{-1} occurs in the sounding rocket data, shown in Figure 3.14. As in Figure 3.10d, the slow trend is probably due to our inadequate estimate of Σ_P or to errors in the geomagnetic field model. The electromagnetic fields associated with the huge deviation near 09:31:30 UT have been identified as an Alfvén wave by *Boehm et al.* [1990]. Thus in a least two cases, one measured from a satellite and one from a sounding rocket, a substantial increase in $\mu_0 |E_x/B_y|$ over Σ_P coincides with temporally varying fields. In Chapter 5 we will develop a technique which can help to determine the importance of time varying fields (i.e. Alfvén waves) in time series data for which no coherent wave structures are evident. But first we will investigate the details of the interaction of Alfvén waves with the ionosphere, which is the topic of Chapter 4.

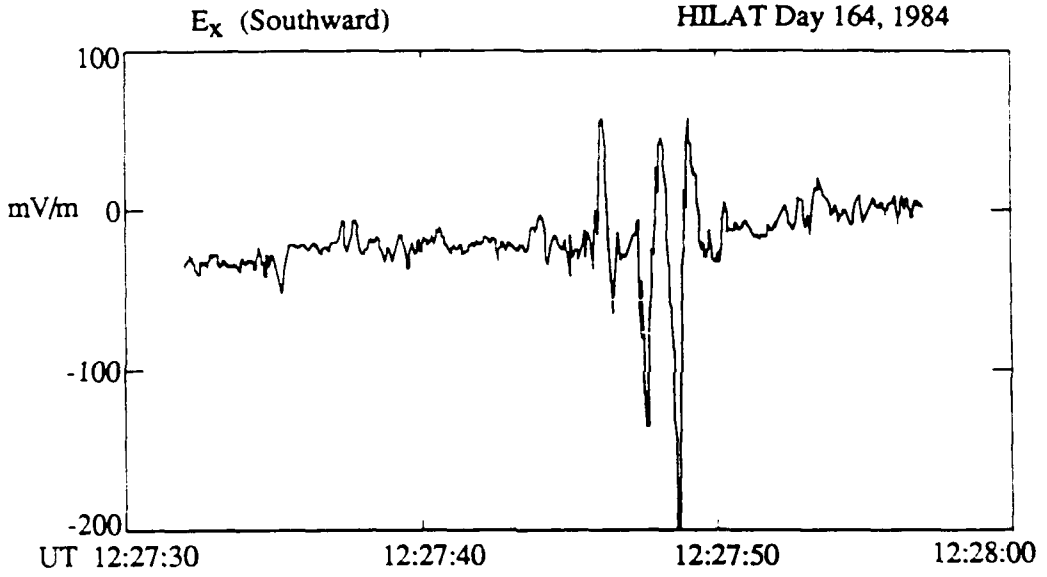


Figure 3.13 Electric field during a short event measured by HILAT on Day 164, 1984, plotted at the full time resolution of the instrument (16 s^{-1} or 32 s^{-1}). The coherent nature of the burst is indicative of time variation (i.e. an Alfvén wave) rather that spatial structuring.

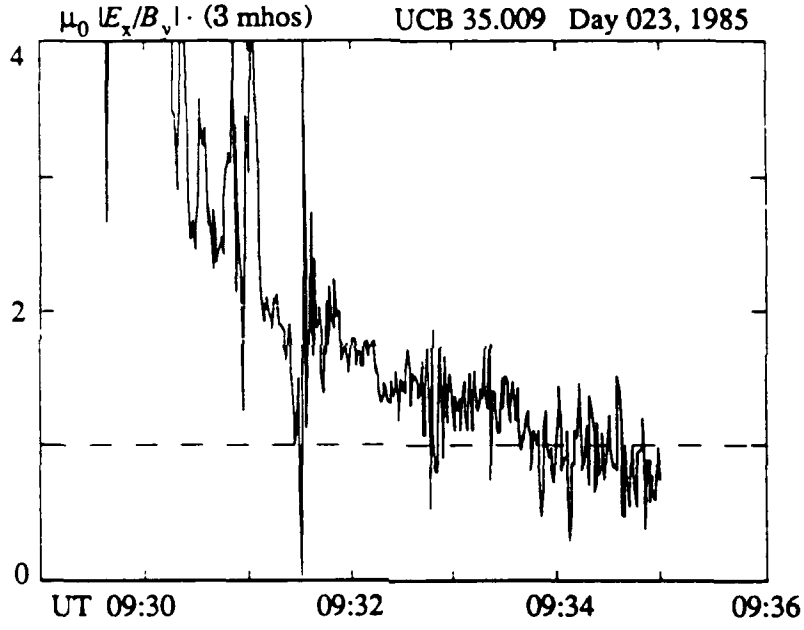


Figure 3.14 Electromagnetic field impedance as a function of time calculated from the sounding rocket data shown in Figure 3.10. The impedances are normalized to a constant value of Σp^{-1} = (3 mhos)$^{-1}$.

CHAPTER 4

A NUMERICAL MODEL OF ALFVÉN WAVES INTERACTING WITH THE HIGH-LATITUDE IONOSPHERE

4.1 Introduction

If the electric and magnetic fields carrying energy from the solar wind and magnetosphere to the high-latitude ionosphere do not vary in time, we can assume that for the most part magnetic field lines are equipotentials, and the energy dissipated in the ionosphere is $\Sigma_P E^2$ (neglecting neutral winds), as we discussed in the last chapter. Now consider time varying fields, but with frequencies less than, say, 10 Hz. In the magnetosphere and ionosphere these waves fall in the Alfvén wave regime, and we have to consider wave-related behaviors like reflections, interference, and particle inertial effects.

As an Alfvén wave propagates towards the ionosphere, it encounters a steeply changing refractive index due to variations in density, composition, and collision frequency. The plasma density varies over several orders of magnitude between 1000 km and the Earth's surface, and at low frequencies this distance can be much less than an Alfvén wavelength. Ion and electron collisions begin to play an important role below a few hundred kilometers, and below a certain altitude (which we will calculate later in the chapter) they control the charged particles to such an extent that what once was an Alfvén wave above the ionosphere cannot now interact with the charged particles, and the wave travels towards the Earth's surface as a "light wave", i.e. $\omega/k = c$. Finally, since

the surface of the Earth is a good conductor, it reflects most of the wave energy. The point to be taken from this description is that the amplitude and phase of an Alfvén wave at one point depend strongly on the plasma characteristics at other points, and those characteristics change rapidly with altitude. Thus in general we need a numerical model to accurately describe Alfvén wave propagation through the ionosphere.

As discussed in Section 2.6, there have been many numerical simulations of Alfvén waves traveling through the density gradient within a few Earth radii of the auroral ionosphere. All of these models treat various phenomena in the collisionless region between the source of Alfvén waves and the ionosphere, but they treat the ionosphere as a single slab characterized by its height-integrated conductivity. In contrast, we will ignore the region above the ionosphere except to treat it as a source of Alfvén waves. We will then model the details of the interaction between these incident Alfvén waves and the ionosphere, with realistic density and collision frequency profiles. We will also find the conditions under which the ionosphere can be modeled as a conducting slab, and our results can be used to provide more realistic boundary conditions for simulations like those mentioned in Chapter 2.

Turning from simulations of Alfvén waves above the ionosphere to waves *in* the ionosphere, we find an enormous amount of literature on the subject. *Budden* [1985] is a good general reference. One of the first computer solutions of Alfvén waves propagating through the ionosphere was carried out by *Francis and Karplus* [1960], who calculated the amount of ionospheric heating by Alfvén waves at 45° latitude for frequencies less than 4 Hz. *Prince and Bostick* [1964] found the amount of attenuation for waves below 10 Hz as they propagate between the

magnetopause and the Earth's surface. These works used relatively coarse models of the ionospheric profile, and treated only vertically incident plane waves. *Greifinger* [1972] , *Hughes* [1974], and *Hughes and Southwood* [1976] allowed for obliquely propagating waves and found field amplitude profiles and reflection coefficients for waves incident from the magnetosphere onto the ionosphere for a variety of ionospheric conditions.

The numerical model that we develop in this chapter is nearly identical to that presented by *Hughes* [1974] and *Hughes and Southwood* [1976], although those authors dealt with wave periods of many minutes, while we emphasize periods of several seconds. Our main purpose for developing the model is to carefully compare its predictions with satellite measurements, which we present in Chapter 5 and which are treated by none of the above references. While developing the model we hope to emphasize an understanding of the physical reasons behind different features in the modeled Alfvén waves.

4.2 Derivation

The propagation of electric and magnetic fields **E** and **B** with frequency ω is described by Maxwell's curl equations

$$\nabla \times \mathbf{E} = -i\omega \mathbf{B} \tag{4.1a}$$

$$\nabla \times \mathbf{B} = \mu_0 \sigma \mathbf{E} + \frac{i\omega}{c^2} \mathbf{E} \tag{4.1b}$$

$\sigma = \sigma(\omega)$ is the frequency-dependent conductivity, which for our needs can be derived from the linearized fluid equations of motion

$$i\omega \mathbf{v}_j = \frac{q_j}{m_j}\left(\mathbf{E} + \mathbf{v}_j \times \mathbf{B}_0\right) - \nu_j \mathbf{v}_j \tag{4.2}$$

where j is a species index. The usual theory of cold plasma waves (see for example *Stix* [1962]) is also derived from the fluid equations of motion, but without the collision term. Notice that collisions can be thought of as creating an effective wave frequency $\omega' = \omega - i\nu$. We take $\mathbf{B}_0$ to be in the $\hat{z}$ direction, solve for the $\mathbf{v}_j$ terms and substitute into the definition of conductivity:

$$\mathbf{J} = \sum_j n_j q_j \mathbf{v}_j = \sigma \cdot \mathbf{E} \tag{4.3}$$

The resulting σ has the following form:

$$\sigma = \begin{pmatrix} \sigma_1 & \sigma_2 & 0 \\ -\sigma_2 & \sigma_1 & 0 \\ 0 & 0 & \sigma_0 \end{pmatrix} \tag{4.4}$$

where

$$\sigma_0 = \varepsilon_0 \sum_j \frac{\omega_{pj}^2}{(i\omega + \nu_j)} \tag{4.5a}$$

$$\sigma_1 = \varepsilon_0 \sum_j \frac{(i\omega + \nu_j)\omega_{pj}^2}{\left[(i\omega + \nu_j)^2 + \Omega_j^2\right]} \tag{4.5b}$$

$$\sigma_2 = \varepsilon_0 \sum_j \frac{\Omega_j \omega_{pj}^2}{\left[(i\omega + \nu_j)^2 + \Omega_j^2\right]} \tag{4.5c}$$

σ_0 describes the relation between **J** and **E** when they are both parallel to each other, and to $\mathbf{B}_0$. For $\omega = 0$ it is known as the "direct" conductivity. In the small ω limit σ_1 is the Pedersen conductivity (Equation 3.4b), and describes dissipative currents ($\mathbf{J}\cdot\mathbf{E} > 0$) which are perpendicular to $\mathbf{B}_0$. If we neglect $\omega << \Omega_i$ in the denominator and let $\nu = 0$ we find $\sigma_1 = \varepsilon_0 i\omega\omega_{pe}^2 / \Omega_i^2 = i\omega/\mu_0 V_A^2$, where V_A is the Alfvén velocity. Thus σ_1 carries

the information allowing us to model Alfvén waves while σ_2 is the generalization of the Hall conductivity ($\mathbf{J} \perp \mathbf{E} \perp \mathbf{B}_0$) for $\omega \neq 0$. We have written (4.5c) in such a way that Ω_j carries the sign of the charge, so that it is negative for electrons.

To solve Equations 4.1 we assume a flat Earth surface, upward $\hat{z}$, periodic variation of the fields in the $\hat{x}$ direction, and no variation in the $\hat{y}$ direction. The ambient magnetic field $\mathbf{B}_0$ is vertical, and σ is assumed to be homogeneous in x and y. Under these conditions we can eliminate E_z and B_z by substituting the $\hat{z}$ component of Equation (4.1b) ($-ik_xB_y = (\mu_0\sigma_0 + i\omega/c^2)E_z$) into the $\hat{y}$ component of (4.1a), and the $\hat{z}$ component of (4.1a) ($\omega B_z = k_xE_y$) into the $\hat{y}$ component of (4.1b). The result is 4 equations in 4 unknowns:

$$\frac{\partial E_x}{\partial z} = -\left(\frac{k_x^2}{\mu_0\sigma_0 + i\omega/c^2} + i\omega\right)B_y \tag{4.6a}$$

$$\frac{\partial E_y}{\partial z} = i\omega B_x \tag{4.6b}$$

$$\frac{\partial B_x}{\partial z} = -\mu_0\sigma_2 E_x + \left(\mu_0\sigma_1 + \frac{i\omega}{c^2} - \frac{ik_x^2}{\omega}\right)E_y \tag{4.6c}$$

$$\frac{\partial B_y}{\partial z} = -\left(\mu_0\sigma_1 + \frac{i\omega}{c^2}\right)E_x - \mu_0\sigma_2 E_y \tag{4.6d}$$

The integration in z is necessary because all of the conductivities vary with altitude. *Hughes* [1974] solved for E_z and B_z explicitly by integrating $\nabla\cdot\mathbf{B} = 0$ and the current continuity equation $\nabla\cdot\mathbf{J} + \partial\rho_c/\partial t = 0$ along with Equations (4.1), which makes a total of six equations to integrate. The price we pay for integrating two fewer equations is that we must apply the additional constraint $\omega \neq 0$, as is obvious from the term proportional to ω^{-1} in (4.6c). The physical reason for this constraint is

that Equations (4.6) satisfy $\nabla \cdot \mathbf{B} = 0$ identically as long as $\omega \neq 0$, which can be seen by taking the divergence of (4.1a). Thus there is no need to solve $\nabla \cdot \mathbf{B} = 0$ explicitly. In a similar way it can be shown that (4.1b) satisfies the current continuity equation for $\omega \neq 0$. We can still treat "DC" fields by making ω very small. We solve Equations (4.6) using an adaptive step size, 4th order Runge-Kutta ODE solver [*Press et al*, 1986].

4.3 Boundary Conditions

At the upper boundary of the modeled region (1000 km) we can neglect σ_2 as long as $\omega >> \nu_i$ and $\omega << \Omega_i$ (recall that Ω_j is a signed quantity). Equations (4.6a) and (4.6d) then decouple from (4.6b) and (4.6c), and if we assume no variation in σ_0 and σ_1 with z above the upper boundary we find solutions for E_x and B_y varying as $\exp(ik_{z,s}z)$ where

$$k_{z,s} = \sqrt{-\left(\frac{k_x^2}{\mu_0\sigma_0 + i\omega/c^2} + i\omega\right)\left(\mu_0\sigma_1 + i\omega/c^2\right)} \tag{4.7}$$

In the MHD limit ($\nu << \omega << \Omega_i$) we can take $\sigma_0 \rightarrow \infty$ and $\sigma_1 = i\omega/(\mu_0 V_A^2)$ where V_A is the Alfvén velocity. Neglecting $1/c^2 << 1/V_A^2$ results in $k_{z,s}^2 = \omega^2/V_A^2$, which is the dispersion relation for the slow Alfvén mode [*Stix*, 1962]. By replacing $\partial/\partial z$ with $ik_{z,s}$ in Equation (4.6d) we obtain a relation between E_x and B_y which serves as the upper boundary for the slow mode.

Turning to E_y and B_x we find from (4.6b) and 4.(6c) that they vary as $\exp(ik_{z,f})$ where

$$k_{z,f} = \sqrt{-\left(i\omega\mu_0\sigma_1 - \omega^2/c^2 + k_x^2\right)} \tag{4.8}$$

Taking the MHD limit this time gives $k_{z,f}^2 = \omega^2/V_A^2 - k_x^2$, which is the Alfvén fast mode. The fast mode is evanescent for $\omega^2/V_A^2 < k_x^2$, which is

satisfied for the range of parameters that we will use at the top of the model region. The source of fast mode energy is in the ionosphere where $\sigma_2 \neq 0$, so we choose boundary conditions at the top of the model region such that the fast mode attenuates with increasing altitude. Substituting $k_{z,f}$ for $\partial/\partial z$ in (4.6d) determines B_x for arbitrary starting values of E_y.

At the bottom of the model region we find two independent solutions to Equations (4.6) by first setting $B_x = 0$, then $B_y = 0$. The electric field is also zero at the perfectly conducting surface. The two solutions are integrated up to about 200 km, where they are matched to the upward and downward propagating slow mode waves and the upwardly evanescent fast mode wave, which have been integrated downward. The reason for matching the solutions at 200 km altitude is that the field amplitudes maximize there, and numerical integration of second order differential equations is most stable in the direction that the solution increases. After the 4 independent solutions have been found, a linear combination is found that allows for a downward-propagating slow wave which has a unit amplitude at 1000 km. The coefficients a_j used to take the linear combination of solutions are found from

$$\begin{pmatrix} E_{x,g1} & E_{x,g2} & -E_{x,s,up} & -E_{x,f} \\ E_{y,g1} & E_{y,g2} & -E_{y,s,up} & -E_{y,f} \\ B_{x,g1} & B_{x,g2} & -B_{x,s,up} & -B_{x,f} \\ B_{y,g1} & B_{y,g2} & -B_{y,s,up} & -B_{y,f} \end{pmatrix} \cdot \begin{pmatrix} a_1 \\ a_2 \\ a_3 \\ a_4 \end{pmatrix} = \begin{pmatrix} E_{x,s,down} \\ E_{y,s,down} \\ B_{x,s,down} \\ B_{y,s,down} \end{pmatrix} \tag{4.9}$$

where subscripts "g1" and "g2" denote the two solutions starting from the ground, "S" and "F" are the slow and fast modes, and "down" and "up" indicate propagation direction. The field values are all taken at the solution matching altitude.

Figure 4.1 summarizes the boundary conditions. At the top of the simulation region we impose a unit amplitude slow wave ($\omega^2 = k_z^2 V_A^2$) propagating downward. This wave is also the reference for zero phase angle in the system. While this mode has $E_y = B_x = 0$ at 1000 km, these field components become non-zero as the wave propagates into regions where $\sigma_2 \neq 0$, which explains why they cannot be neglected at the solution matching altitude and thus appear in the column vector on the right side of (4.9). There is also an upward-propagating slow wave which has an amplitude at 1000 km a_3, given by Equation 4.9. Since the downward slow wave has unit amplitude, a_3 is the "voltage reflection coefficient" ($E_{reflected}/E_{incident}$) at 1000 km for slow mode Alfvén waves incident on the ionosphere.

In addition to the incident and reflected slow waves, there is the evanescent fast wave which is driven due to coupling of energy from the slow wave in the E region where the conductivity σ_2 is non-negligible. Since the energy source for this mode is below 1000 km in altitude, we choose the solution which decays with increasing altitude. After finding the two modes below 200 km shown in Figure 4.1, they are matched to the 3 upper solutions using Equation 4.9.

4.4 Model Input

The physical description of the atmosphere and ionosphere is contained in the conductivities in Equation 4.5. To find the conductivities we need altitude profiles from 0 to 1000 km of electron density, ion composition, collision frequencies, and the geomagnetic field.

We simplify the geomagnetic field model by restricting our attention to high latitudes, where we can assume that $\mathbf{B}_0$ is vertical. In the next

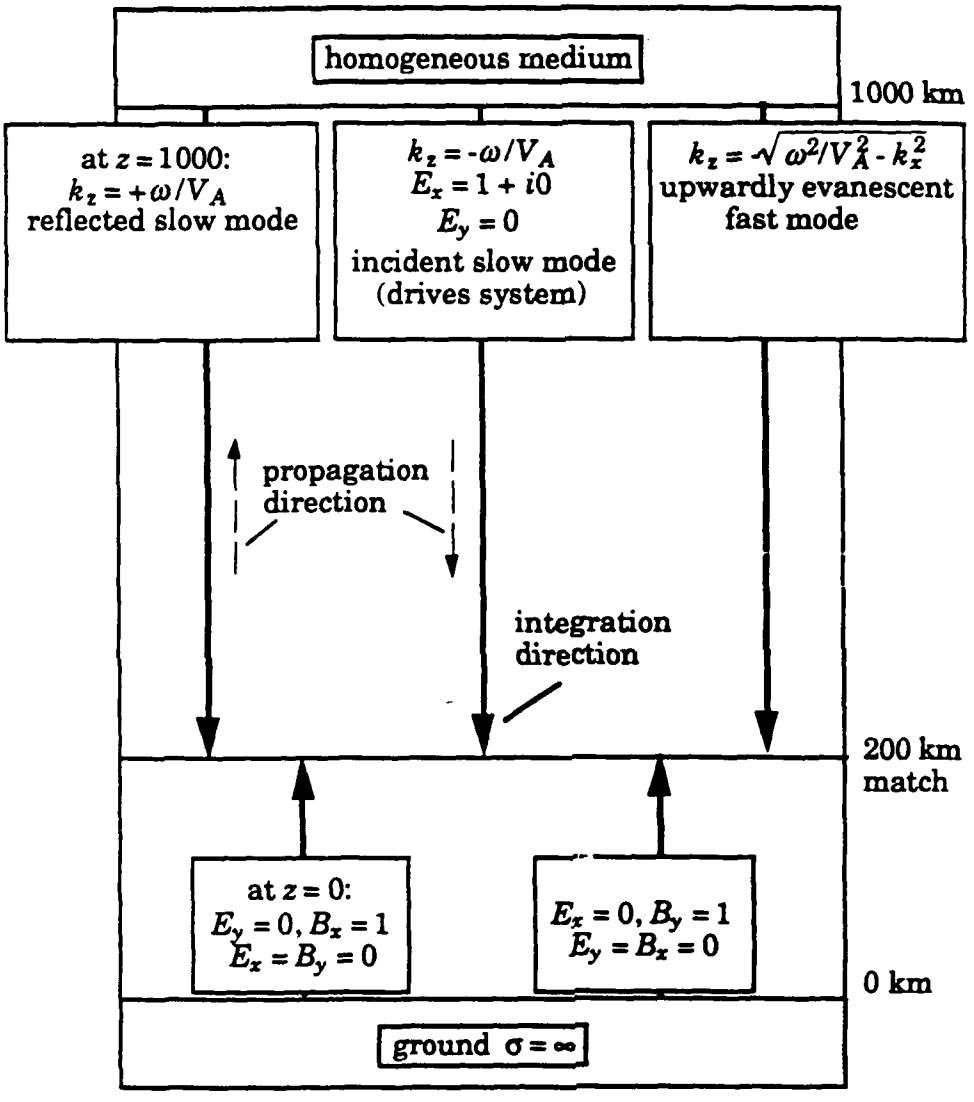


Figure 4.1 Schematic representation of the five independent solutions which are combined into a single solution for the electromagnetic fields between 0 and 1000 km.

chapter we will compare the result of the model with experiments carried out above Sondrestrom, Greenland, so in our model we use Sondrestrom's surface magnetic field of 0.56×10^{-4} T (downward). In a dipole field , $\mathbf{B}_0 \propto r^{-3}$, and we let our model magnetic field fall off accordingly with altitude.

The most important input into our model is the plasma density profile above 95 km. At high latitudes this is very difficult to model. The plasma density profile created by photoionization can be predicted by a Chapman production function, but this plasma can have a very long lifetime in the F region, and it can convect far away from its point of production. In addition to photoionization, precipitation of energetic electrons produces a significant amount of the high latitude plasma, but in an unpredictable manner. Later, when we compare the model results with experimental measurements, we will use measurements of the electron density profile from the Sondrestrom Incoherent Scatter Radar. Our purpose in this chapter is to see how the interaction of Alfvén waves with the ionosphere changes with various inputs, so we will use each of the 3 different model electron density profiles shown in Figures 4.2 - 4.4. With each of the profiles is plotted the conductivities σ_0, σ_1, and σ_2 with $\omega = 0$. The profile in Figure 4.2a, labeled "EF", is a typical daytime profile with both an E and F region. The "F" profile in Figure 4.3a lacks an E region, and therefore has associated with it a low Pedersen conductivity. This profile could occur, for example, after sunset when the E region quickly recombines leaving an F region only. The "E" profile in Figure 4.4a could occur in darkness with electron precipitation leading to the ionization bump below 200 km. The E-region ionization in the "E" and "EF" profiles is modeled with a Gaussian:

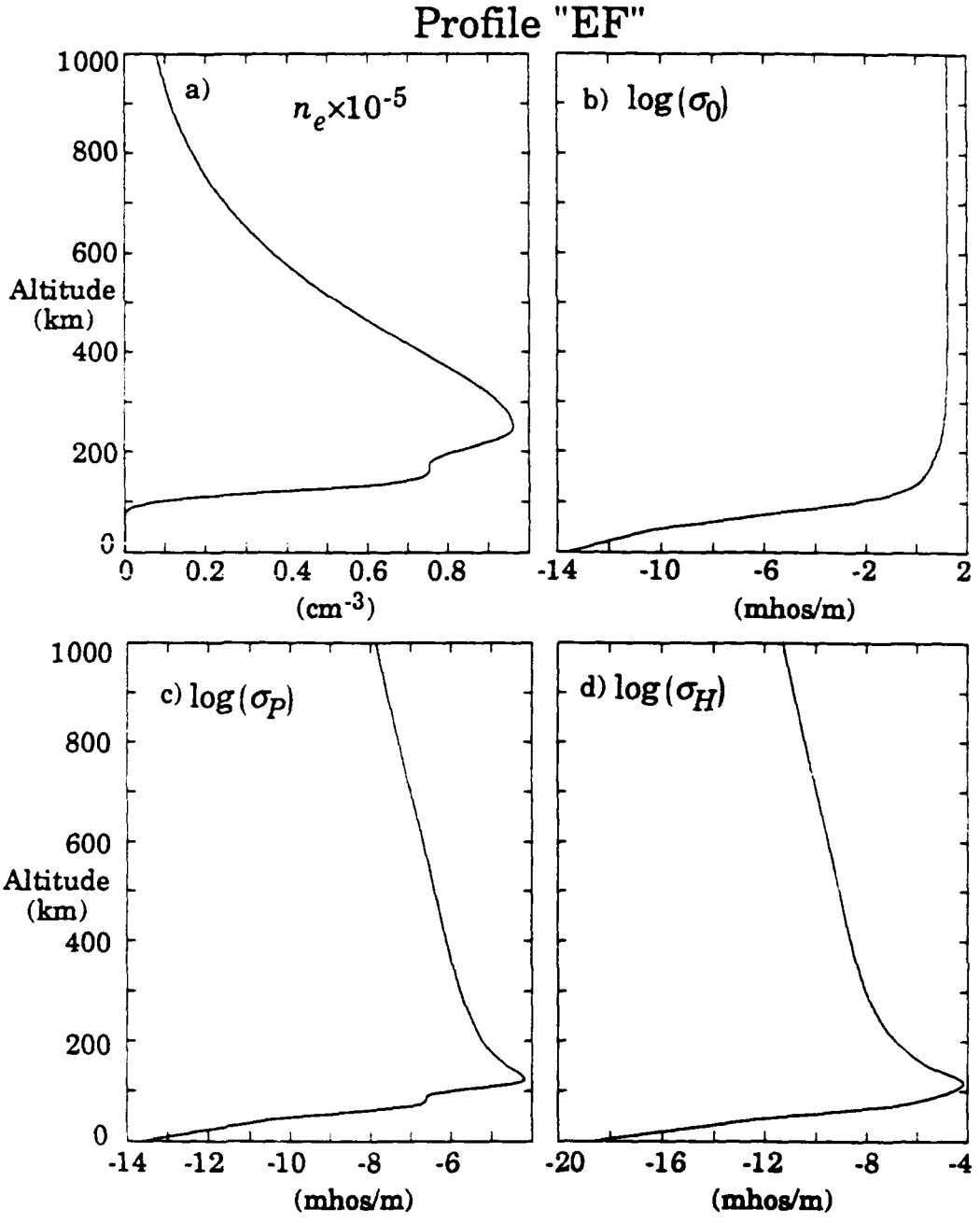


Figure 4.2 a) A typical electron density profile and the associated b) direct, c) Pedersen and d) Hall conductivity profiles for a sunlit, daytime ionosphere.

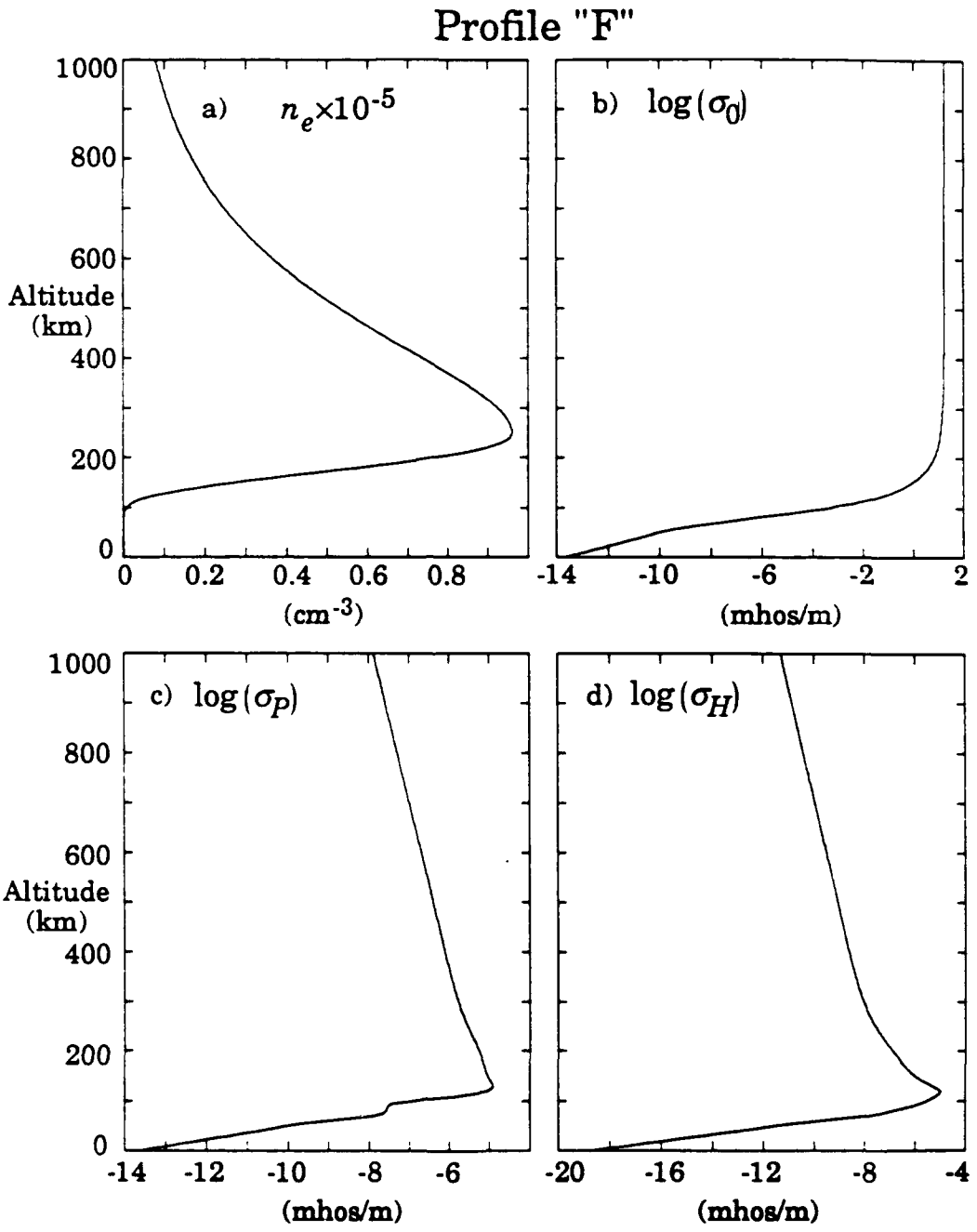


Figure 4.3 a) A typical electron density profile and the associated b) direct, c) Pedersen and d) Hall conductivity profiles for a post-sunset ionosphere with an F region only.

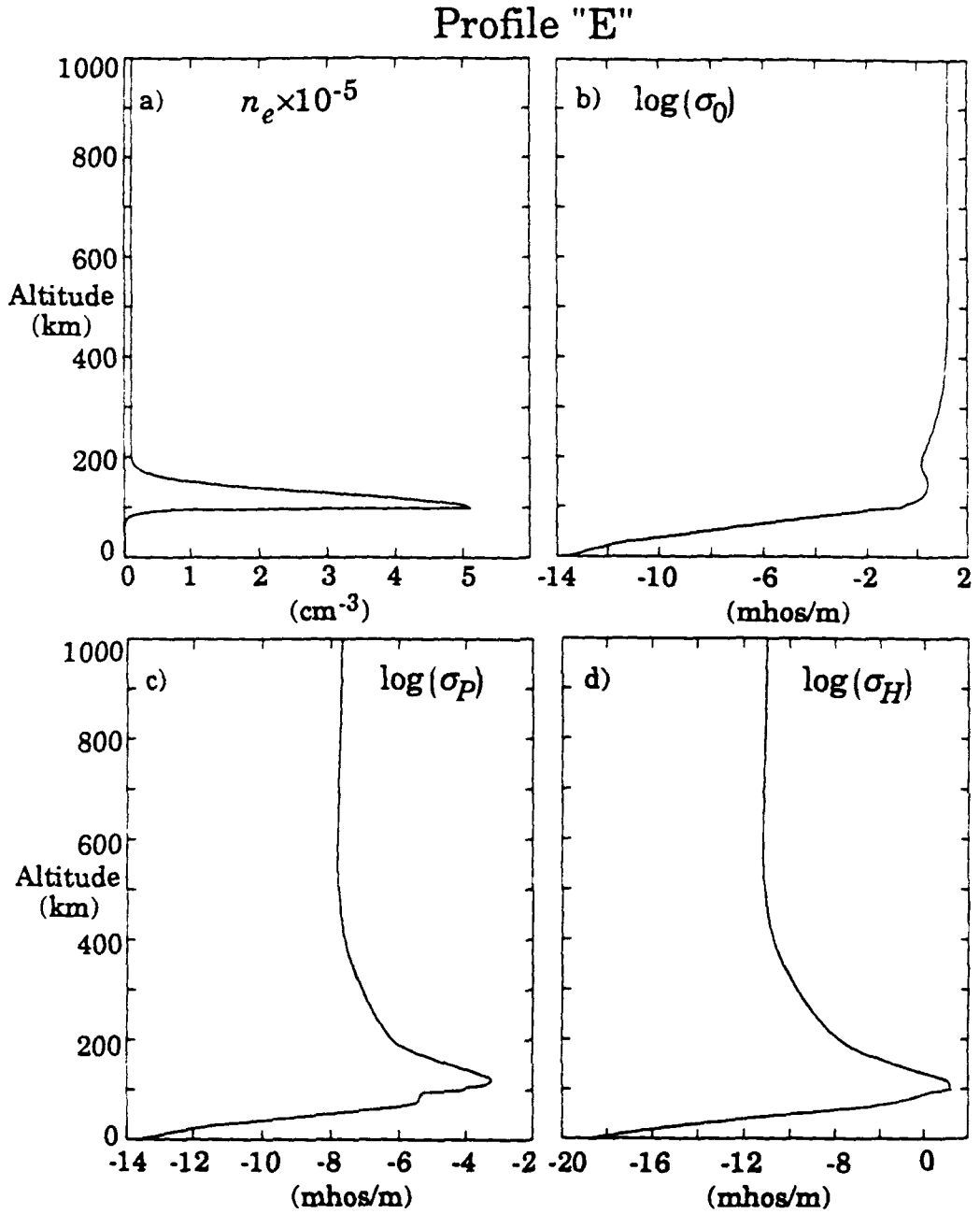


Figure 4.4 a) A typical electron density profile and the associated b) direct, c) Pedersen and d) Hall conductivity profiles for a nighttime ionosphere caused by relatively energetic electron precipitation.

$$n_{e,E}(z) = N_E \exp(-z^2/a^2) \tag{4.10}$$

The F-region ionization is from a Chapman profile [*Banks and Kockarts*, 1973],

$$n_{e,F}(z) = N_F \exp(0.5(1 - z' - \exp(-z')\sec\chi) \tag{4.11}$$

where $z' = (z - z_{max})/H$, z_{max} is the altitude of maximum ionization density, H is the scale height of the neutral atmosphere, and χ is the angle between the local zenith and the sun. For the model density profiles we have used $\sec(\chi) = 1$, and we adjust the value of H to give a "reasonable" profile based on the measured profiles we will present in the next chapter. We found that to mimic the experimental data, different scale heights above and below the F-region peak altitude were sometimes necessary. The various parameters used to create the model profiles are given in Table 4.1.

We calculate ion-neutral collision frequencies using the formulas in the appendix of *Schunk and Walker* [1973] in conjunction with the *Jacchia* [1971] neutral atmosphere model. The neutral atmosphere model uses an assumed thermospheric temperature as input, and although this parameter can vary widely we use 1000 K throughout this chapter and the next. *Banks and Kockarts* [1973] supply expressions for electron-neutral and electron-ion collision rates. We add these two quantities to obtain an effective electron collision frequency, i.e. $\nu_e = \nu_{en} + \nu_{ei}$. Strictly speaking, this is incorrect because the collisional drag term in the fluid equations is proportional to the velocity difference of the colliding species, and we have written (4.3) in a way that assumes that charged particles are colliding with particles at rest. The term ν_{ei} represents electron collisions with ions that are not necessarily at rest, although at altitudes where ν_{en} is not the dominant source of collisions it

Table 4.1. Density model input paramters as defined in Equations 4.10 and 4.11.

	Profile E	Profile F	Profile EF
N_E (cm^{-3})	5×10^5	0	5×10^4
E-region topside, scale, a_{top} (km)	40		40
E-region bottomside scale, a_{bottom} (km)	5		40
E-region z_{max} (km)	100		140
N_F (cm^{-3})		9.5×10^4	9.5×10^4
F-region topside scale, H_{top} (km)		120	120
F-region bottomside scale, H_{bottom} (km)		60	60
F-region z_{max} (km)		250	250
Background density (cm^{-3})	10^4	10^3	10^3
Σ_P (mhos)	14	0.93	2.6

is reasonable to neglect the ion velocity. We also neglect collisions between ions of different species.

Below 95 km we increase collision frequencies and decrease charged particle densities exponentially with a scale height of 6 km, corresponding to the scale height of the neutral atmosphere. However, we do not allow the electron density to decrease below 1 cm^{-3} in order to make conductivities within 10 km of the ground consistent with Figure 20.3 of *Sagalyn and Burke* [1985]. Varying the density and collision frequencies in the lower atmosphere has little effect on the Alfvén wave electromagnetic fields above the ionosphere. Figures 4.5a-c show the electron and ion collision frequencies between 0 and 1000 km altitude for the 3 model profiles in Figures 4.2-4.4. The gross features of the profiles are quite similar, although ν_e in the "E" profile does not decrease with decreasing altitude above 500 km as in the "EF" and "F" profiles. This is because electron-ion collisions dominate at the highest altitudes, and the high altitude ion density is constant in altitude for profile "E".

We use 2 ion species in our model. Far above 180 km we assume that O^+ is the only constituent, and it changes smoothly to NO^+ over about 40 km (centered at 180 km) as shown in Figure 4.5d. Of course there can be many other positively and negatively charged ions in the D and E regions of the ionosphere. *Hughes* [1974] chose to incorporate negative ions into his numerical model, but his purpose was to find Alfvén wave fields on the Earth's surface, so he needed an accurate measure of the attenuation of waves in the D region. In our case we will be comparing the model to data taken at several hundred km in altitude, and from working with the model we found that for our purposes the exact composition of the D region was unimportant.

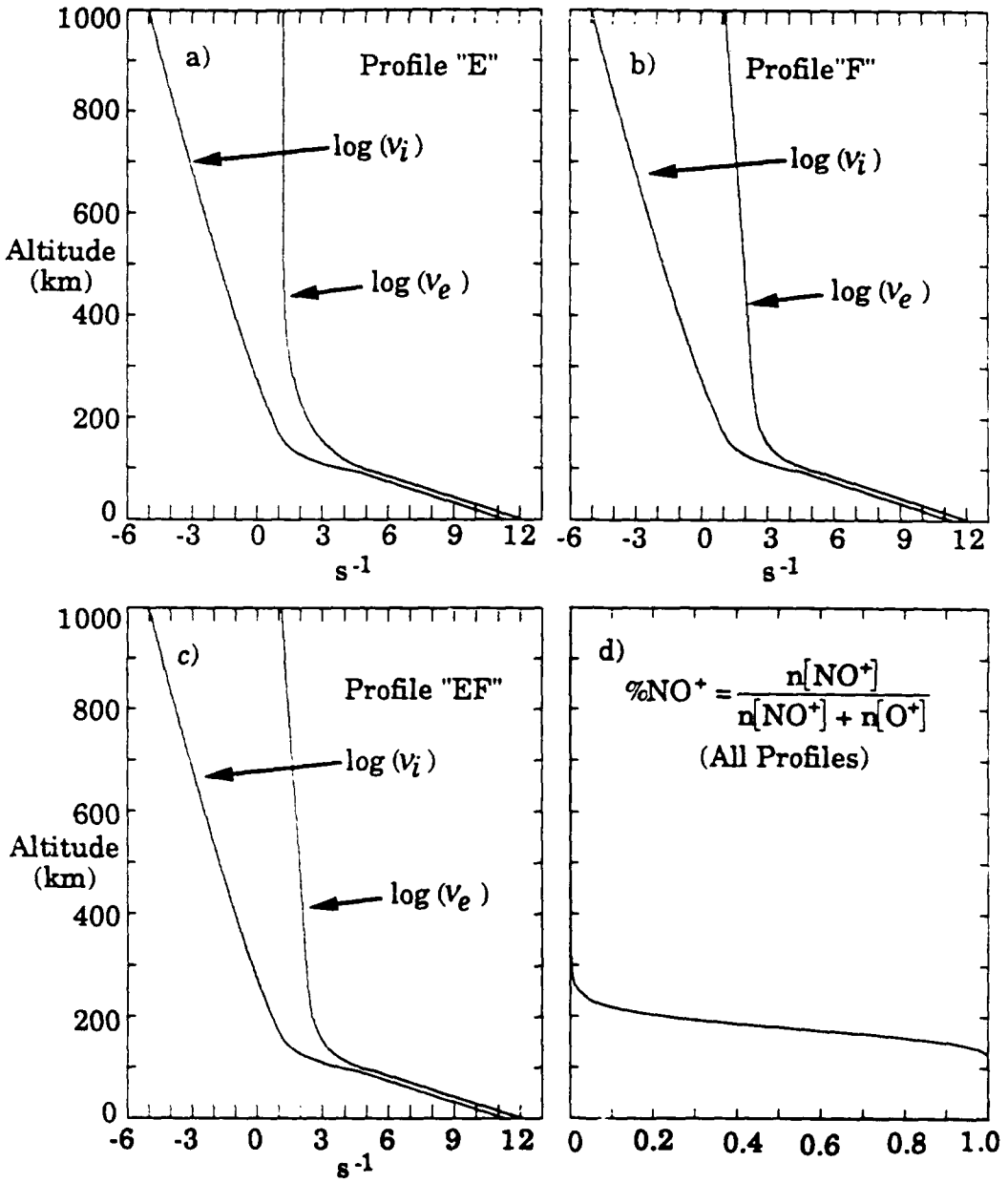


Figure 4.5 a-c) Electron and ion collision frequency profiles for the density profiles shown in Figures 4.2-4.4. d) Relative concentration of O^+ and NO^+ as a function of altitude used for input to the numerical model.

Charged molecular oxygen, O_2^+, can be important in the E region, but its molecular mass (32 a.m.u.) is very close to that of NO^+ (30 a.m.u.), causing only a small change in the gyrofrequency and Alfvén velocity. The collision frequency for O_2^+ is different than that of NO^+, but not enough to significantly change the results of the model. Later we will use the model to illustrate that small changes in collision frequencies have only a minor effect on the fields above the ionosphere.

Our purpose in this chapter is to illustrate with the numerical model the effects of various ionospheric features on Alfvén waves incident from the magnetosphere. One of the main subjects we will investigate will be the Alfvén wave reflection coefficient (a_3 in Equation 4.9) as a function of frequency for various ionospheric models. But first we will try in the next 2 sections to provide a general idea of some the phenomena associated with the Alfvén wave/ionosphere interaction by examining altitude profiles of the electric and magnetic field amplitudes at two different frequencies.

4.5 Quasi-Static Fields in the Ionosphere

Before we present output from the numerical model, we must make a few comments concerning our plotting conventions. We plot altitude profiles of the fields with 3 different curves. The first is the real part of the field, to give an idea of the number of wavelengths contained in each plot. We also plot both the amplitude of the field and the negative of the amplitude, which creates a wave "envelope". We do this because if we plot only the real part of the wave and the (positive) amplitude, it is often difficult to distinguish between the two.

A second convention we use is to plot all magnetic flux densities after multiplying them by the speed of light c. This means that magnetic and electric fields are plotted in the same units (V/m). Furthermore, since the incident slow mode electric field has a unit amplitude, cB_y is related to the refractive index of the slow mode. It is equal to the refractive index of the slow mode in the special case of no reflections.

Our model is constrained to non-zero frequencies, but we can make the frequency small enough to study the behavior of fields which are effectively "DC" in the ionosphere and lower atmosphere. It turns out that $f = 10^{-3}$ Hz is sufficiently low because at that frequency the spatial extent of the ionosphere is a very small fraction of an Alfvén wavelength. We have calculated the electric and magnetic field profiles with the "EF" density profile (Figure 4.2) for $f = 10^{-3}$ Hz and a horizontal spatial scale $\lambda_x = 1000$ km.

Turning to Figure 4.6a, we see that the meridional electric field E_x is constant at a value of about $E_x = 0.25$ until very close to $z = 0$, where it suddenly falls to zero as it must since we assume a perfectly conducting ground. At high altitudes the constant electric field E_x can be explained by referring to Equation 4.6a. When the neutral density becomes small, so do the collision frequencies, thus σ_0 becomes large. Since we have used a very small ω, Equation 4.6a reduces to

$$\partial E_x / \partial z \approx 0 \tag{4.12}$$

Equation 4.12 is equivalent to the statement that "large scale static electric fields map along field lines". We can see from the plot of the zonal electric field E_y in Figure 4.6b that σ_0 is not infinite, because it is evident that E_y does *not* map along field lines. Note that E_y is two orders

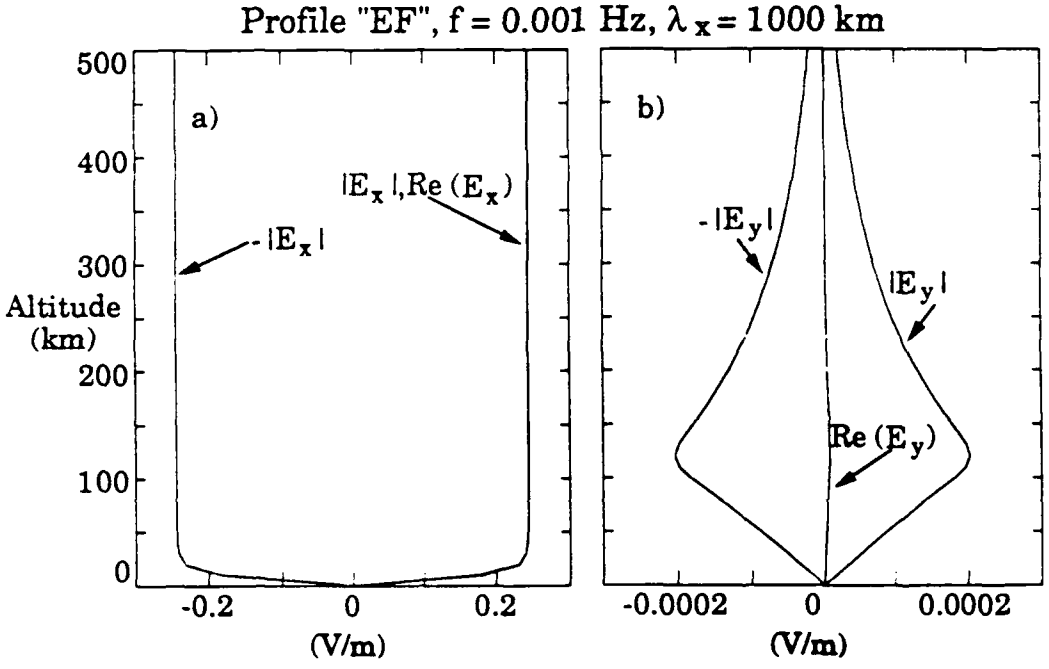


Figure 4.6 a) Meridional and b) zonal electric field profiles in the quasi-static limit.

of magnitude smaller than E_x. The reason for the different behaviors of the meridional and zonal electric fields is that we have assumed no variation in the $\hat{y}$ direction, thus there can be no divergence in E_y. This means that the E_y is not sustained by electric charge concentrations building up along magnetic field lines, which is the case for E_x. Instead, we can think of E_y as a sort of "fringing field" which decreases with distance from its source in the E region, where σ_2 is large. At such a low frequency the electric fields must be essentially curl-free, so the change in E_y with altitude implies a non-zero parallel electric field E_z, which is only possible in our DC approximation for non-zero σ_0. Both the zonal and parallel electric fields have very small amplitudes compared to the meridional electric field.

It is instructive to derive an expression for E_x near the Earth's surface in the DC approximation. At low altitudes we may safely neglect $\Omega << \nu$ (i.e. the medium is completely collision dominated), which from Equations 4.5 means that $\sigma_0 \approx \sigma_1$. We may also neglect σ_2 because it decreases as ν^{-2}, while σ_0 and σ_1 decrease as ν^{-1}. If we assume that the charged particle density near the ground is constant and $\nu_e \propto \exp(-z/H)$ where H is the neutral atmosphere scale height, we may write

$$\sigma_0, \sigma_1 \approx S_0 \exp(z/H) \tag{4.13}$$

If we neglect ω as unimportant and use (4.13) then (4.6a) and (4.6d) can be written

$$E_x' = -\frac{k_x^2}{\mu_0 S_0} e^{-z/H} B_y \tag{4.14a}$$

$$B_y' = -\mu_0 S_0 e^{z/H} E_x \tag{4.14b}$$

A prime denotes differentiation by z. If we multiply Equation (4.14a) by $e^{(z/2H)}$, (4.14b) by $e^{-(z/2H)}$, and recognize that $(e^{(z/2H)}E_x)' = (e^{(z/2H)}E_x)' - (2H)^{-1}e^{(z/2H)}E_x$, we can solve for $E_x(z)$:

$$E_x(z) = A\, e^{-z/2H} \sinh\left(z\sqrt{(2H)^{-2} + k_x^2}\right) \tag{4.15}$$

We have eliminated a "cosh" term from (4.15) to satisfy $E_x(0) = 0$. In the limit $H << \lambda_x$, (4.15) reduces to $E_x(z) = A(1 - e^{-z/H})$. Thus, in agreement with Figure 4.7a, we see that the DC meridional electric field maps from the magnetosphere through the lower atmosphere, and to within H of the Earth's surface.

One final detail concerning the DC meridional electric field is the magnitude of the field, or "A" in (4.15). The value in Figure 4.6a of 0.25 comes from the fact that we are applying an incident field of unit magnitude far from the ionosphere, but our model includes wave reflections. We can infer then that a reflected wave of amplitude -0.75 is interfering with the incident wave, and therefore the low frequency electric field reflection coefficient for the "EF" profile we have used is also -0.75. A simple way to check this is to treat the ionosphere as a conducting slab with Σ_P = 2.6 mhos for profile "EF" and the region above the ionosphere as a transmission line with characteristic impedance $Z_A = \mu_0 V_A$ = 2.8 Ω at 1000 km for profile "EF". The resulting electric field reflection coefficient is then $(\Sigma_P^{-1} - Z_A)/(\Sigma_P^{-1} + Z_A)$ = -0.76.

We turn now to the quasi-DC zonal magnetic field cB_y shown in Figure 4.6d. As with the meridional electric field E_x, cB_y "maps" along $\mathbf{B}_0$ at high altitudes, but it does not penetrate into the lower atmosphere as does E_x. The reason for this is that at low frequencies and high altitudes, cB_y is created by field-aligned currents. Around a couple of

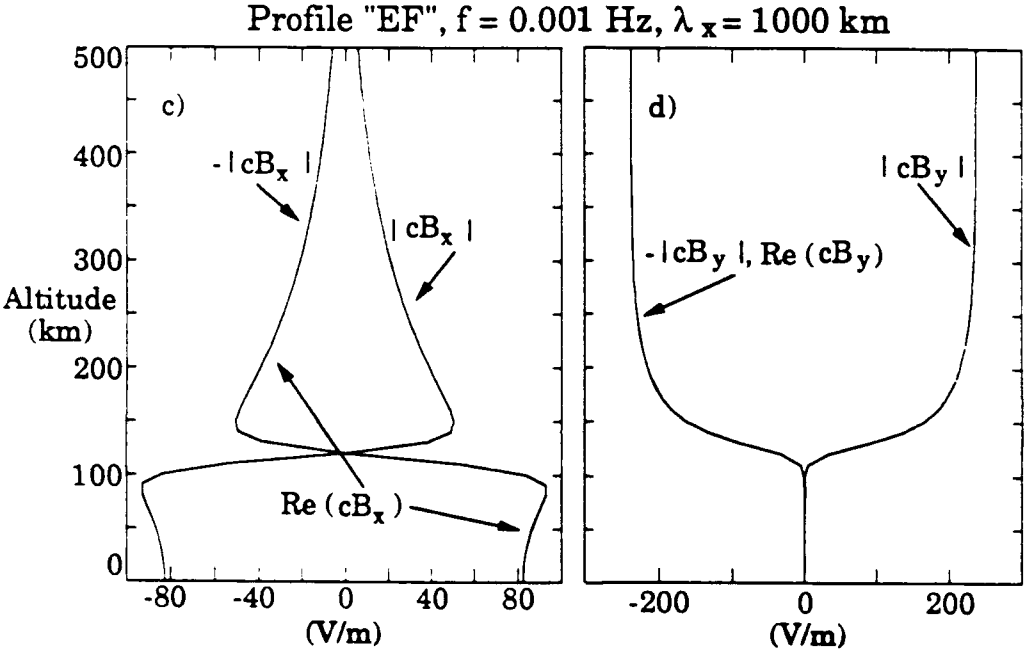


Figure 4.6 c) Meridional and d) zonal perturbation magnetic field profiles in the quasi-static limit.

hundred km altitude the field-aligned currents begin to close horizontally as the Pedersen conductivity becomes important. By the bottom of the ionosphere (around 100 km) the field-aligned currents have been completely closed by the Pedersen currents.

The absence of field-aligned currents alone does not explain the lack of any magnetic field cB_y below 90 km. That is, even though the field-aligned current sheets end at about 100 km, one might expect the magnetic field due to them to extend below 100 km. It turns out that the magnetic perturbation from the Pedersen current layer exactly cancels the field-aligned current contribution below the current system, so that no zonal magnetic field can "leak out" into the lower atmosphere. This is similar to the case of the magnetic field outside an infinitely long solenoid, which is identically zero everywhere. In Appendix B we prove that the magnetic field is zero outside of an idealized auroral arc consisting of two semi-infinite field-aligned current sheets which are connected by a thin Pedersen current layer. The Hall current in an auroral arc creates a magnetic field B_x that does have a magnetic signature on the ground.

The fact that the zonal magnetic field cB_y disappears below the E region is consistent with the fact that almost no Alfvén wave energy is dissipated in the lower atmosphere. This can be seen from the Poynting vector $\mathbf{E}\times\mathbf{H}$. Since $cBy = 0$ in the lower atmosphere, the component $E_x \times H_y$ is also zero. The other part of the vertical component of the Poynting vector is also zero because even though the fast mode fields E_y and cB_x are both non-zero in the lower atmosphere, they are 90° out-of-phase, thus the time average of their product is zero. For example, at an altitude of 50 km, $E_y = 8.79\times10^{-5}\angle 87.9°$ V/m and $cB_x = 86.7\angle -2.1°$ V/m. Electric fields

measured with balloons were compared to ground-based magnetic field measurements by *Mozer and Manka* [1971], who showed $\mathbf{E}_\perp$ and $\delta\mathbf{B}_\perp$ were parallel, thus it follows that the vertical component of the Poynting vector was zero in the lower atmosphere.

The fast mode fields E_y and cB_x (Figures 4.6b and c) maximize near their source in the E region, and they fall off with a scale length roughly equal to the horizontal scale λ_x. This can be seen from the zero frequency limit of the fast mode dispersion relation in Equation (2.15), i.e. $k_z^2 = -k_x^2$. The narrow null in cBx (Figure 4.6c) is due to the fact that there is a zonal Hall current driven by the meridional electric field E_x. One would expect cB_x to have opposite signs above and below the Hall current layer, thus the field must go through zero within the current sheet. Notice that unlike the zonal field cB_y, cB_x is non-zero at the ground.

4.6 A 1 Hz Alfvén Wave in the Ionosphere

At 1 Hz the wave nature of electromagnetic fields in the ionosphere is very evident. Another way of saying this is that the thickness of the lower atmosphere/ionosphere system is on the order of an Alfvén wavelength for frequencies on the order of 1 Hz. Thus static electric field mapping ideas are not applicable. This is evident in Figures 4.7a-d, where we plot the four horizontal field components at 1 Hz, using the same "EF" density profile (Figure 4.2a) as in the previous section.

Referring first to the meridional electric field E_x we see that at higher altitudes the magnitude of the field is much greater than in the quasi-static case, and it is not constant with altitude. This is because we are seeing the standing wave pattern caused by the interfering incident and reflected slow mode Alfvén waves. It appears from Figure 4.7a that

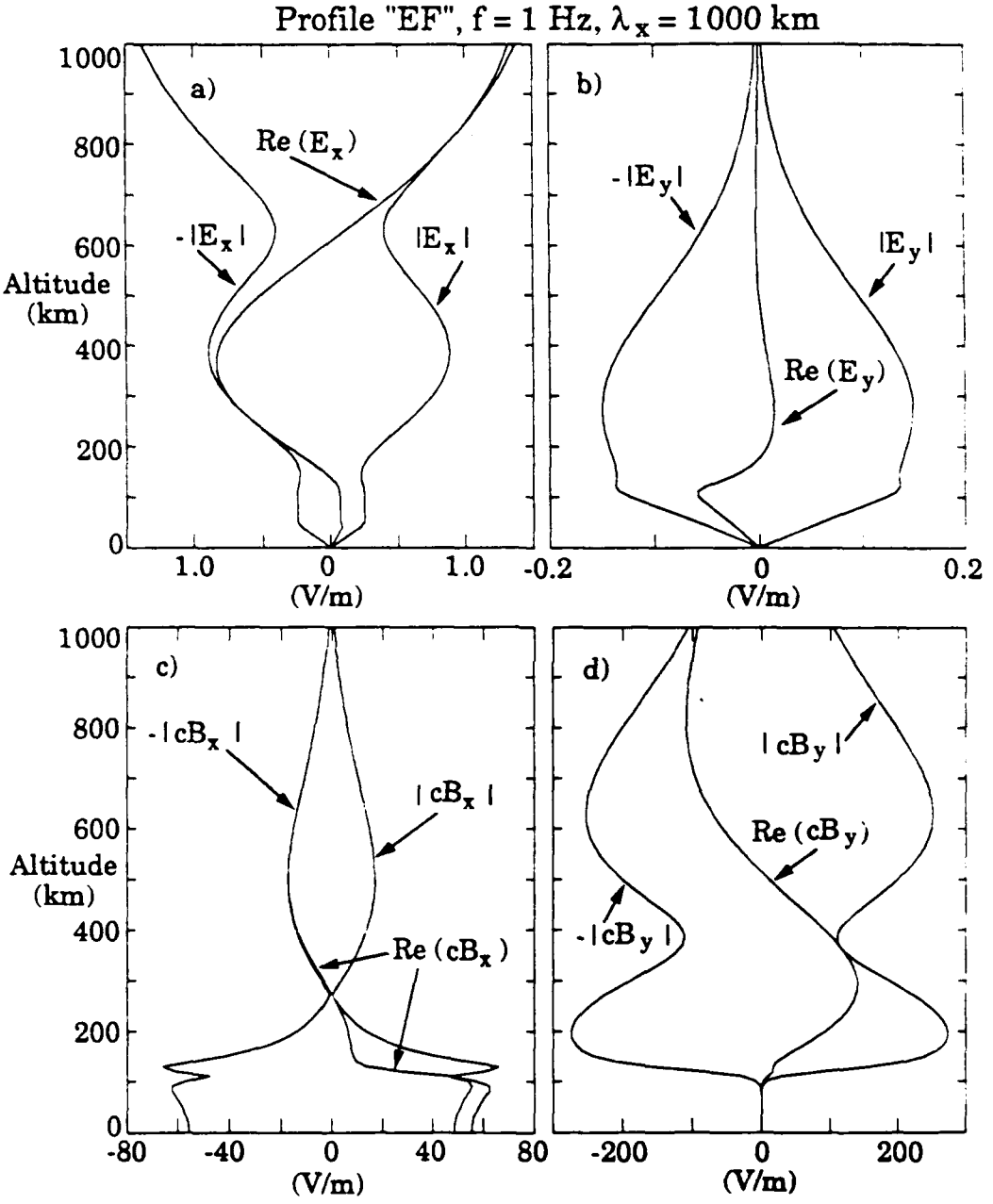


Figure 4.7 a) Meridional electric, b) zonal electric, c) meridional magnetic, and d) zonal magnetic field profiles due to a 1 Hz Alfvén wave reflecting from the ionosphere.

most of the reflection takes place at or above 100 km, and at 1000 km constructive interference leads to an amplitude greater than the incident wave amplitude (unity). The reflection coefficient at 1 Hz for this density profile is -0.4, so we expect a peak amplitude of 1.4 at half-wavelength intervals above 1000 km.

Below 100 km E_x is constant until about 40 km, then it begins to drop off, reaching zero at the ground. The zonal magnetic field cB_y in Figure 4.7b is similar to the quasi-static case except for the standing wave pattern above 100 km. As before, there is no zonal magnetic perturbation at the ground.

There are two notable differences between the 1 Hz and quasi-static fast mode fields E_y and cB_x (Figures 4.7c and d). First, E_y at 1 Hz is 2-3 orders of magnitude greater than in the quasi-static case, and secondly there is a "bulge" in the 1 Hz field amplitudes above 200 km. This is due to the fact that the 1 Hz Alfvén wavelength in the F region peak is about 690 km, which makes it is *less* than the horizontal spatial scale λ_x of 1000 km. From Equation 2.16 this makes k_z real. Thus the fast mode Alfvén wave is not evanescent in the F region. As the density decreases above the F region, k_z again becomes imaginary. If we decrease λ_x to be less than 690 km the bulge in the field amplitudes disappears, but the fields in the E region and below do not change much.

As in the quasi-static case, the vertical component of the Poynting vector for a 1 Hz wave in the lower atmosphere is zero because $cBy = 0$ and E_y and cB_x are almost 90° out-of-phase. For example, at an altitude of 50 km, $E_y = 5.9\times10^{-2}\angle119.8°$ V/m and $cB_x = 58.3\angle28.9°$ V/m. The phase difference is thus 90.9°. Again, this is consistent with the fact that almost no Alfvén wave energy is dissipated below the E region.

To provide a qualitative picture of the variation in electric field profiles versus frequency and altitude we have plotted surface contours of the meridional and zonal electric fields in Figures 4.8a and 4.8b. We have used λ_x = 1000 km. As the frequency exceeds 1 Hz the zonal field suddenly "turns on" and resonates in the F-region cavity. This resonance is driven at the expense of meridional electric field energy, as is evident from the slight decrease in the meridional field amplitude near 1.2 Hz.

At this point we are ready to study the effects of varying ionospheric parameters on Alfvén waves. We want to investigate the Alfvén wave/ionosphere interaction at many different frequencies; hence, the field amplitude profiles shown in the last two sections are not the most useful format for comparing numerical results. Instead we will use the frequency-dependent complex reflection coefficient for the slow mode Alfvén wave.

4.7 Reflection of Alfvén Waves from Different Ionospheric Density Profiles

The complex reflection coefficient Γ for the slow mode Alfvén wave is defined by

$$\Gamma = E_{x,up}(z = 1000)/E_{x,down}(z = 1000) \tag{4.16}$$

where the subscripts "up" and "down" refer to propagation direction, and the electric fields are evaluated at 1000 km. In our model we impose $E_{x,down}(z = 1000) = 1 + i0$, thus $\Gamma = E_{x,up}(z = 1000) = a_3$, where a_3 is found from Equation 4.9. Figures 4.9 a and b show the magnitude and phase of

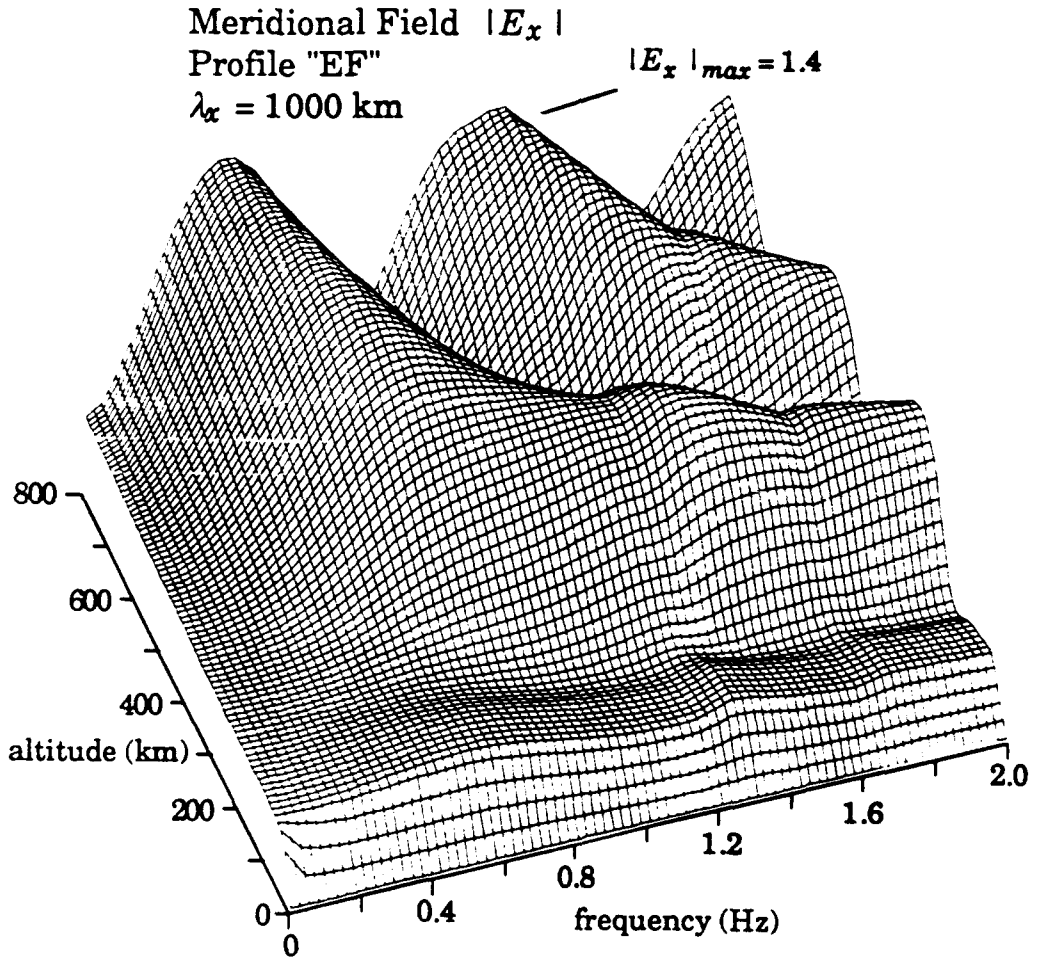


Figure 4.8a Surface plot showing the variation of the magnitude of the meridional electric field $|E_x|$ as a function of frequency and altitude.

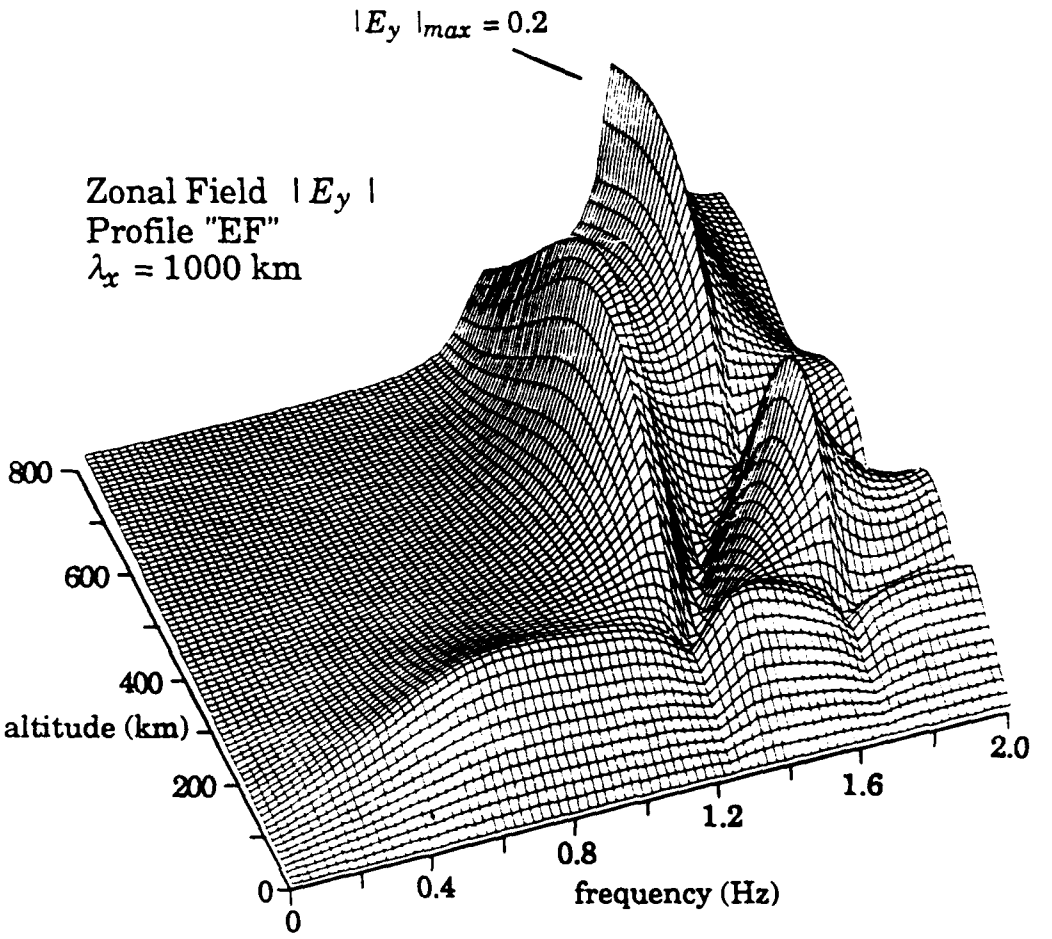


Figure 4.8b Surface plot showing the variation of the magnitude of the zonal electric field $|E_y|$ as a function of frequency and altitude.

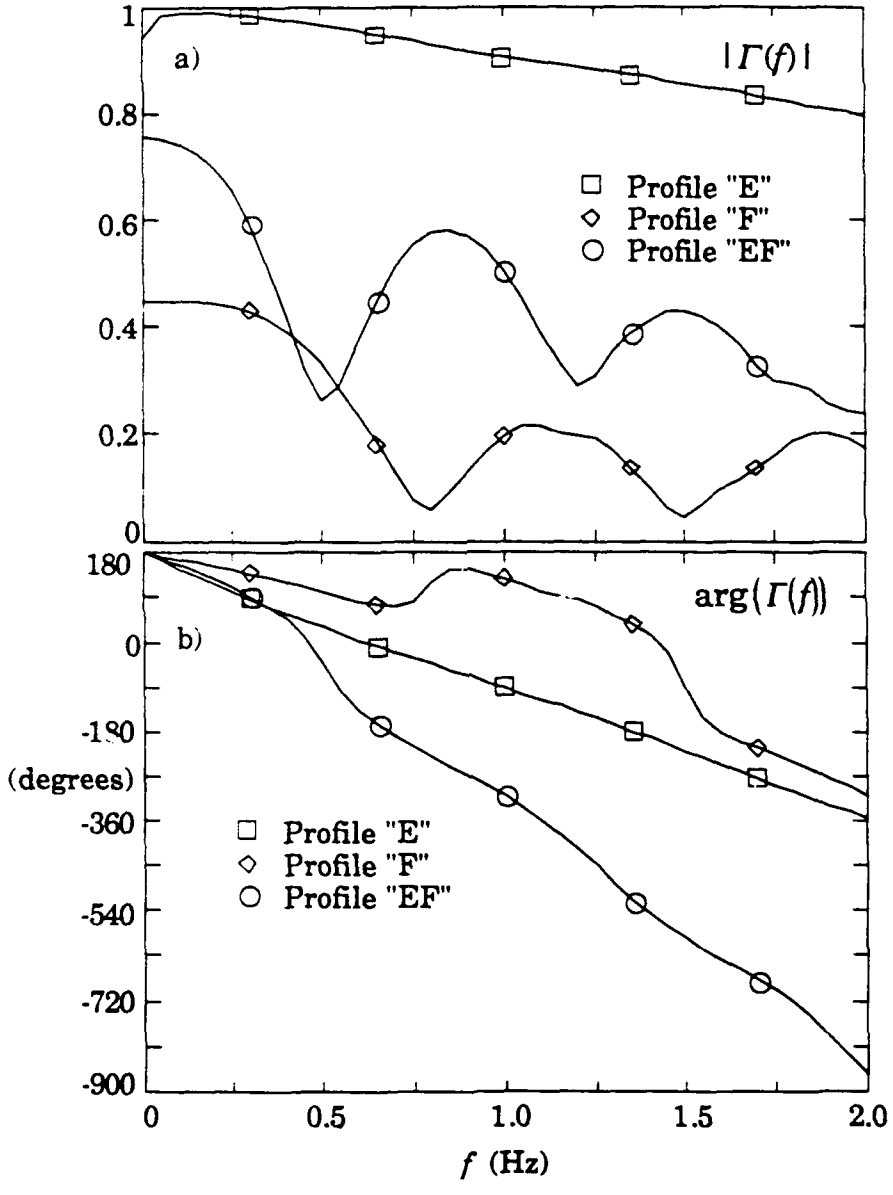


Figure 4.9 a) Magnitude and b) phase of the complex electric field reflection coefficient Γ at 1000 km altitude for the three ionospheric density profiles shown in Figures 4.2-4.4.

Γ as a function of frequency for each of the three density profiles in Figures 4.2 - 4.4. We use λ_x = 1000 km for these calculations.

In the low frequency limit for all three profiles the reflection coefficient Γ can be predicted with the transmission line analogy discussed in Section 4.5 by $(\Sigma_P^{-1} - Z_A)/(\Sigma_P^{-1} + Z_A)$ where again $Z_A = \mu_0 V_A$. The height-integrated Pedersen conductivities in each case are given in Table 4.1. As the frequency increases the "E" and "EF" profile reflection coefficients go through pronounced dips. This is a result of the fact that the ionosphere is a distributed load, thus the slab reflection model is not appropriate above roughly 0.1 Hz. A lower reflection coefficient Γ is a result of an increase in Joule dissipation given by $(E_x^2 + E_y^2)\Re(\sigma_1) + E_z^2\Re(\sigma_0)$. The real parts of the conductivities (denoted by the symbol $\Re$) are mostly unaffected by changes in frequency. Thus the increase in Joule dissipation is due to an increase in electric field magnitudes within the conducting layer. At the minimum Γ frequencies in Figure 4.9a the electric field standing wave pattern is such that E_x is increased in the E region. This increase in E_x can be seen in the surface plot in Figure 4.8a.

Superimposed on the dips in Γ in the "EF" and "F" curves is a general trend towards smaller reflection coefficients at higher frequencies. This is the propagation loss which increases as the electrical length of propagation increases. In a homogeneous medium this loss increases exponentially with electrical length. A second energy sink at higher frequencies is the zonal electric field, which increases because it is no longer evanescent, at least in regions of high density.

Figure 4.9b shows the phase angle in degrees of $\Gamma(f)$. If we again use a transmission line/load analogy, the overall negative slope in $\Gamma(f)$ gives an effective reflection altitude for the Alfvén waves. If we let d denote the

distance from 1000 km to the reflection point, then the phase angle ϕ between the upward and downward propagating slow Alfvén waves at 1000 km is given by $\phi = 2(360°)d/\lambda$, thus $d\phi/df = 720°d/V_A$. Applying this to the curve for the "E" profile in Figure 4.9b gives d = 820 km, or a reflection altitude of 180 km, which corresponds to the steep conductivity gradient on the top side of the E region.

In all 3 density profiles we have examined, the lower atmosphere plasma density profile was the same, namely an exponential decrease in charge density with decreasing altitude below 95 km. We have chosen a 6 km scale height for this decrease. We found that doubling the scale height to 12 km had no discernable effect on the 0-2 Hz reflection coefficient for the "EF" density profile, hence we conclude that accurate modeling of charge density in the lower atmosphere is unimportant for our purposes.

4.8 The Effect of Collisions on $\Gamma(f)$

Small reflection coefficients are due to efficient dissipation of electrical energy in the form of Joule heat, which is the result of particle collisions. If both electron and ion collisions are absent, then $\Gamma(f) \equiv 1$. To determine the relative importance of the two types of collisions, we again calculated $\Gamma(f)$ for the "EF" density profile in Figure 4.2a, but we multiplied and divided the ion collision frequency ν_i by 3 at all altitudes, as shown in Figure 4.10. We have re-plotted the original $\Gamma(f)$ from Figure 4.9a for reference. At low frequencies, more ion collisions increase Σ_P and consequently the reflection coefficient. The behavior is not as intuitively predictable above 1 Hz.

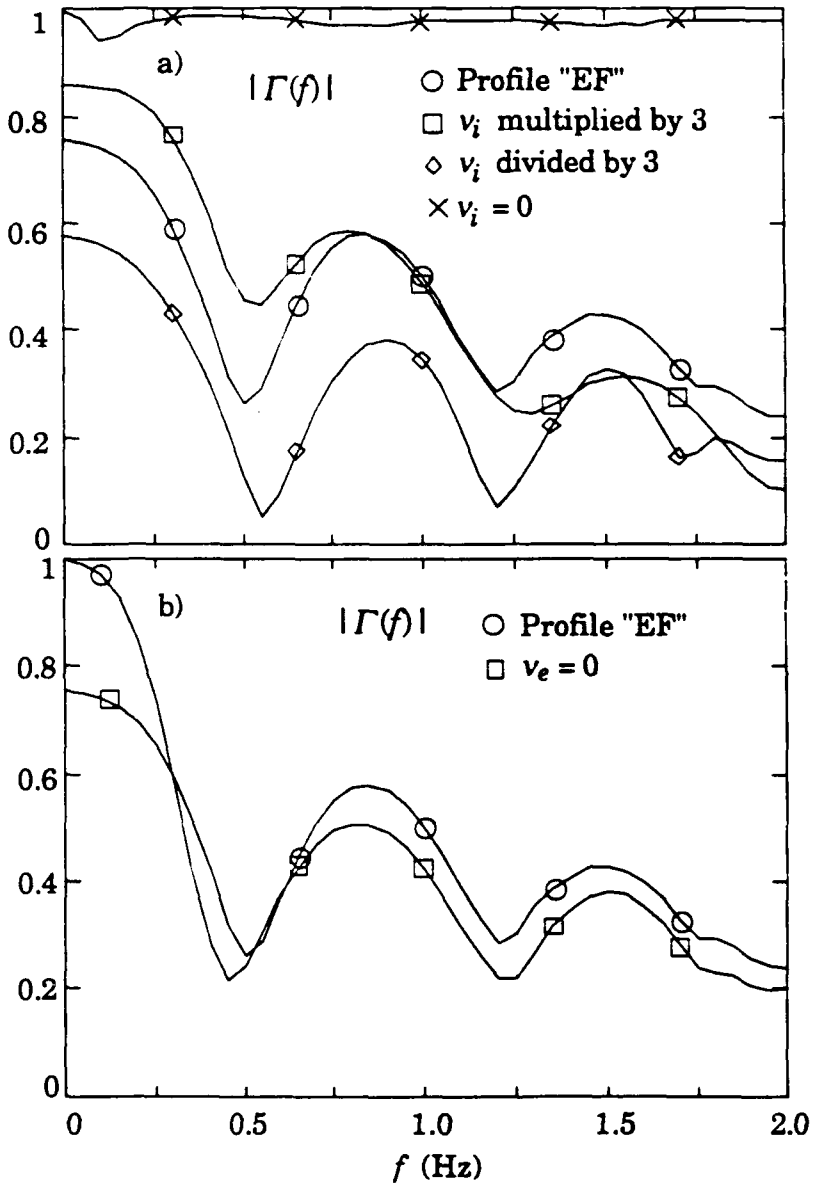


Figure 4.10 Illustration of the effects that changes in ion collision frequencies (upper panel) and electron collision frequencies (lower panel) can have on the magnitude of $|\Gamma|$.

The Γ curve resulting from $\nu_i = 0$ in Figure 4.10a shows that electron collisions account for only a small part of the energy absorption, allowing most of the incident wave to reflect. Figure 4.10b also illustrates the slight effect of electron collisions by comparing the original $\Gamma(f)$ along with the curve for $\nu_e = 0$. Above 0.2 Hz the two curves differ by only a few percent. We can conclude that ion collisions are mainly responsible for the dissipation of Alfvén wave energy.

Notice in Figure 4.10a that decreasing the ion collision frequency by a factor of three reduces $|\Gamma|$, but eliminating ν_i completely increases $|\Gamma|$ to near unity. This seemingly contradictory behavior is a result of the fact that ion collisions play two distinct roles -- they both reflect and absorb energy. Dividing ν_i by three decreases Σ_P and allows for a better impedance match to the region above the ionosphere. As a result, electric fields penetrate deeper into the E region, and more energy is dissipated. But at some point, decreasing ν_i will limit the amount of E-region attenuation and the wave will simply reflect off of the Earth's surface and back into the magnetosphere.

4.9 $\Gamma(\lambda_x)$

In previous sections we have varied ionospheric parameters but assumed a constant horizontal spatial scale λ_x of 1000 km. The reflection coefficient for the slow mode Alfvén wave is mostly unaffected as long as $\lambda_x > 10$ km. To investigate smaller scales with the numerical model we have to neglect the fast mode completely because the fast mode at x scales near 10 km is very evanescent, and over the 1000 km simulation region the code must integrate through hundreds of e-folds in amplitude, which is computationally prohibitive. Neglecting the fast mode probably doesn't

cause significant errors in our results because its amplitude is quite small for small λ_x. Figure 4.11 shows $\Gamma(\lambda_x)$ between 10 and 1 km for quasi-DC fields, and although the dependence is not strong, a clear trend towards smaller Γ below 2.0 km scales is evident. A possible reason for this decrease is that electric fields with horizontal spatial scales less than a few km map poorly through the E region. This scale-size dependence of electric field mapping is well known [see for example *Farley*, 1960]. A consequence of this poor mapping is that the electric field "sees" less of the conductivity in the E region, so the effective height-integrated conductivity is less. This in turn means the ionospheric impedance is more nearly matched to the Alfvén impedance above the ionosphere, and an impedance match implies a smaller reflection coefficient.

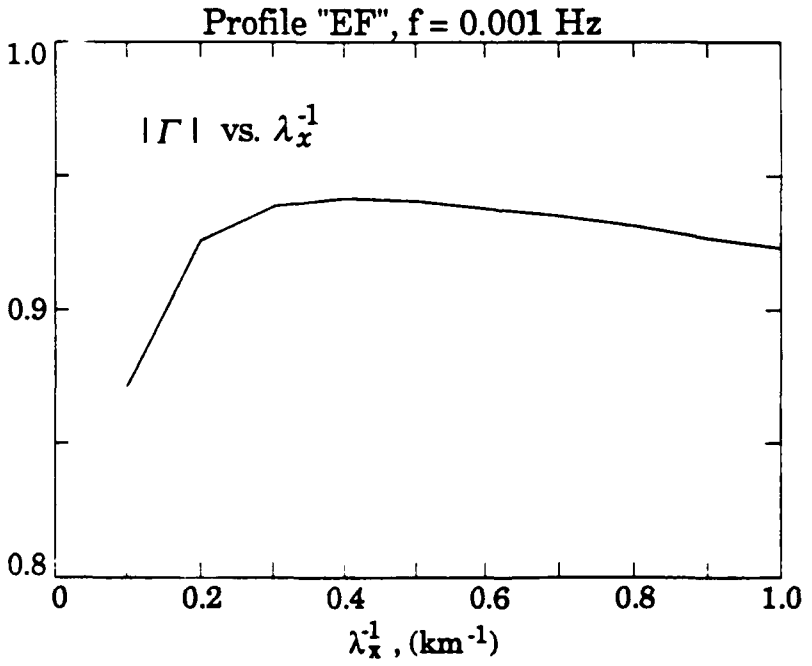


Figure 4.11 Reflection coefficient magnitude, $|\Gamma|$, as a function of inverse horizontal spatial scale λ_x^{-1}.

CHAPTER 5

ROCKET AND SATELLITE MEASUREMENTS OF ALFVÉN WAVES ASSOCIATED WITH THE DISCRETE AURORA

5.1 Introduction

The connection between Alfvén waves and auroral arcs has been discussed by many authors [*Hasegawa*, 1976; *Goertz and Boswell* 1979; *Haerendel* 1983; *Seyler* 1988], mainly because Alfvén waves are a possible mechanism for auroral electron acceleration. One of the purposes of this chapter is to establish experimentally (at least in a few examples) that Alfvén waves occur near auroral arcs. Our main goal is more general, however. We will develop and apply techniques to analyze auroral electric and perturbation magnetic field data and extract information concerning the source, propagation, dissipation, and relative amounts of spatial and temporal structuring of those fields.

Zmuda et al. [1966] were among the first to measure magnetic fluctuations with a satellite. The measurements were taken at 1100 km with the 1963 38C satellite and the authors interpreted the fluctuations as Alfvén waves. Many authors have identified Alfvén waves in satellite measurements taken at altitudes of several Earth radii, most recently *Erlandson et al.* [1990], who measured very coherent burst of elliptically polarized waves below 1 Hz with the Viking satellite. *Iyemori and Hayashi* [1989] also found coherent bursts of Alfvén waves with the Magsat satellite orbiting at 400 km, but purposefully neglected latitudes

above 65° "because of the difficulty in distinguishing the waves from the small-scale field-aligned current structures."

Sugiura et al. [1982], *Sugiura* [1984], and *Smiddy et al.* [1984] measured electric and magnetic fields perpendicular to $\mathbf{B}_0$ with the DE-2 satellite near the ionosphere (< 1000 km) at auroral latitudes and found them to be highly correlated but not coherent as in the above studies. Furthermore, they found that the ratio $Z = \mu_0 E_{rms}/\delta B_{rms}$ was not equal to $\mu_0 V_A$ as one would expect from Alfvén waves but instead $Z = \Sigma_P^{-1}$. As we discussed at the end of Chapter 3, this can be explained by a static electric field/Birkeland current model with fluctuations arising from the motion of the spacecraft through spatial structures with scale sizes from hundreds of meters to hundreds of km. Since the spectrum of fluctuating fields measured by a spacecraft traversing the auroral oval is generally a monotonically decreasing function of frequency, the bulk of the spectral energy lies at low frequencies. A correlation analysis of electromagnetic field fluctuations measured in the oval will therefore emphasize the largest scales in the system.

Rather than forming a single r.m.s. measure of field fluctuations, we will relate the amplitudes and phases of electric and magnetic fields as a function of frequency, which allows us to investigate spatial and/or temporal scales that are smaller than those considered in previous studies of auroral fields and closer to the regime associated with discrete auroral arcs and Alfvén waves.

There are several ways to detect Alfvén waves in the ionosphere. The most direct is to look for coherent electromagnetic field oscillations in the time-domain data. We have already seen an example of this in HILAT satellite data in Chapter 3 (Figure 3.13). Recently *Boehm et al.*

[1990] found Alfvén waves in the time-domain data from a Black Brant X sounding rocket launched into the morning auroral oval (see Section 3.5). One can also detect Alfvén waves by forming frequency-time sonograms and looking for narrow-band enhancements in the electric and magnetic field spectra. For example *Erlandson et al.* [1990] used this technique with Viking satellite data.

The frequencies at which spectral enhancements occur provide information pertaining to the source and propagation of Alfvén waves which has to date not been fully exploited. The peaks may represent the frequency of the source supplying the Alfvén waves, or properties of the medium through which they have traveled. For example *Lysak* [1988] showed that the exponential density decrease thousands of km above the ionosphere can give rise to resonant excitations in the ionosphere. Another cause for structuring in frequency is interference between waves incident from the magnetosphere and reflected from the ionosphere, i.e. standing waves. One way to distinguish between these possibilities is to form the quotient of the electric and magnetic field spectra. Spectral enhancements due to the source of Alfvén waves will appear in both the electric and magnetic field spectra. Dividing the spectra gives what we call the "impedance function" $Z(f)$, and we will show in the next section that any structuring it has in frequency must be due to standing Alfvén waves. In this chapter we will also examine the frequency-dependent phase relation between the meridional electric and zonal magnetic fields E_x and δB_y, and show that this can also be used to identify standing wave patterns. These techniques have the advantage that in some cases they can detect Alfvén waves when their presence in time-domain or spectral data alone is not obvious.

5.2 The Impedance Function

Let $\tilde{E}_x(f)$ and $\delta\tilde{B}_y(f)$ be the complex Fourier transforms of the meridional electric and zonal magnetic fields measured in the frame of a moving satellite or sounding rocket. The complex impedance $\tilde{Z}(f)$ is defined by

$$\tilde{Z}(f) = \mu_0 \tilde{E}_x(f) / \delta\tilde{B}_y(f) \qquad (5.1)$$

We will use $Z(f)$ to denote the magnitude of $\tilde{Z}(f)$, and this is what we will call the "impedance function". In practice we find the impedance function from $\mu_0(P_E/P_B)^{1/2}$ where P_E and P_B are the electric and magnetic field power spectra of $E_x(t)$ and $\delta B_y(t)$.

We will compare $Z(f)$ from spacecraft-measured electromagnetic fields above the aurora with the predictions of three ideal models: 1) Structured static fields and Birkeland currents, 2) traveling Alfvén waves with no horizontal structure, and 3) standing Alfvén waves with no horizontal structure.

In the static field model the frequency f measured in the spacecraft frame is due entirely to the motion of the spacecraft through spatial structures with scale size λ_x. If the spacecraft velocity in the x direction is V_s, then $f\lambda_x = V_s$. In this case $Z(f) = \Sigma_P^{-1}$ for all spatial scales greater than a few km. This is violated at small scales because the electric fields do not map completely through the E region, thus the height-integrated Pedersen current decreases.

A traveling Alfvén plane wave will have an impedance function which is constant in frequency and which is equal to the characteristic impedance of Alfvén waves, $Z_A = \mu_0 V_A$. As we mentioned in the introduction, $Z(f)$ is constant in this case even when the Alfvén wave

source supplies waves at preferred frequencies because spectral enhancements occur in both P_B and P_E and therefore divide out. Obliquely propagating traveling waves have a modified impedance when the horizontal wavelength λ_x approaches the electromagnetic skin depth c/ω_{pe}, and this can cause $Z(f)$ to vary with frequency. We will consider this possibility in more detail in the discussion section at the end of this chapter.

Previous authors have measured the ratio of electric and magnetic field spectra and have found values between Z_A and Σ_P^{-1}. One of the first to do this was *Gurnett et al.* [1984] who used the Dynamics Explorer 1 satellite to measure the refractive index cB/E (proportional to $Z^{-1}(f)$) between 1.78 and 31.1 Hz. These authors pointed out that a static field-aligned current model was not sufficient at all frequencies. *Berthelier et al.* [1989] also calculated the refractive index versus f using the AUREOL-3 satellite and found a refractive index that was structured in frequency but again was larger than could be expected from Alfvén waves. They concluded that the fields must have been due to Doppler-shifted static fields. As we will see, both the structuring in frequency and the seemingly high refractive index are consistent with the standing Alfvén wave model.

Near a boundary such as the ionosphere, an Alfvén wave will be partially reflected due to the high conductivity, and the incident and reflected waves will form a standing wave pattern. In this case the impedance function will vary with distance from the reflection point.

The vertically changing field impedances in a standing wave pattern are essentially impossible to detect directly in the Earth's auroral zone since satellite trajectories are horizontal, and rockets cannot make a

vertical cut through a distance of several times $\lambda/4$ quickly enough to unambiguously separate temporal and spatial variations. (Future satellite or space shuttle experiments with tethered probes may make a direct measurement of a standing wave pattern possible.) This is not true in the case of whistler mode waves which have much shorter wavelengths. For example *Siefring et al.* [1990] have reported standing VLF waves at two frequencies below a sporadic E layer. In this case the transmitters were on the ground and at a fixed frequency so the analysis was straightforward. Here we must cope with a geophysical source of unknown character which co-exists with a spatially turbulent convection electric field, and in which are imbedded field-aligned current sources. Furthermore, in our case the reflecting surface is almost certainly structured at scales which in the satellite frame generate frequencies in the Alfvén wave regime. Despite these difficulties, evidence for standing Alfvén waves can be found from the impedance function because different frequencies correspond to different electrical lengths above the ionosphere; that is, higher frequencies in $Z(f)$ correspond to larger electrical distances above the ionosphere.

For a uniform reflecting sheet of integrated conductivity Σ_P in contact with a uniform medium characterized by a constant impedance Z_A the electric and magnetic fields due to an Alfvén wave can be written

$$\tilde{E}(z, t) = \tilde{E}_{incident}\left(e^{i(\omega t + k_z z)} + \Gamma e^{i(\omega t - k_z z)}\right) \qquad (5.2a)$$

$$\tilde{B}(z, t) = \tilde{B}_{incident}\left(e^{i(\omega t + k_z z)} - \Gamma e^{i(\omega t - k_z z)}\right) \qquad (5.2b)$$

where a tilde denotes a complex quantity, $\omega/k_z = V_A$ and Γ is the electric field reflection coefficient $(\Sigma_P^{-1} - Z_A)/(\Sigma_P^{-1} + Z_A)$. (See for example *Ramo, Whinnery, and Van Duzer* [1965] or other texts on plane wave

propagation or transmission lines.) The ratio of (5.2a) to (5.2b) gives the complex impedance function

$$\tilde{Z}(\omega) = \mu_0 V_A \left[\frac{1 + \Gamma e^{-2i\omega z / V_A}}{1 - \Gamma e^{-2i\omega z / V_A}}\right] \tag{5.3}$$

The magnitude of the impedance function $Z(f)$ varies between Z_A/S and $Z_A S$ where S is the standing wave ratio given by $S = (1 + |\Gamma|)/(1 - |\Gamma|)$. For ionospheric reflections with $Z_A > \Sigma_P^{-1}$ the minimum and maximum impedances above the ionosphere are Σ_P^{-1} and $Z_A^2 \Sigma_P$ and are separated by a distance $\lambda_z/4$ where $\lambda_z = V_A/f$.

Of course the ionosphere above the reflecting E region is not homogeneous as we have assumed above, and later in the chapter we will attempt to predict the behavior of measured impedance spectra using the numerical model of Chapter 4. However, a simple and reasonable estimate of the frequency f_{max} of the first peak in $Z(f)$ can be found by neglecting partial reflections off of F-region density gradients and assuming a single reflection from the top of the E region (at z_{min}). (This assumption is valid under a WKB approximation.) The result is

$$\frac{1}{f_{max}} = 4 \int_{z_{min}}^{z_{max}} \frac{dz}{V_A(z)} \tag{5.4}$$

which reduces to $f_{max} = V_A/[4(z_{max} - z_{min})]$ for constant V_A. Here z_{max} is the height at which the fields are measured.

Knudsen et al. [1990] showed in two examples, one from a sounding rocket and one from the HILAT satellite, in which measured peaks in the impedance functions were predicted by (5.4). We will present some of those results along with new data later in the chapter. When calculating impedance spectra from sounding rocket and satellite data, large fluctuations due to noise arise when dividing the electric and magnetic

field spectra. To reduce the amount of statistical variation in the measured impedance spectra, we split the time series data into several 32 point sub-intervals overlapping by 16 points. The electric and magnetic field spectra from each sub-interval are averaged before dividing to find $Z(f)$. Before calculating the individual spectra we subtract a linear least squares fit from the time-domain data and multiply by a Hanning window [*Press et al.*, 1986].

5.3 The Normalized Cross-Spectrum

Standing Alfvén waves can be distinguished from traveling waves or static fields not only by the magnitude of the complex impedance function as discussed above, but by its phase as well. In a pure standing electromagnetic wave (i.e. with $\Gamma = \pm 1$) the electric and magnetic fields are shifted by $\pm$ 90°, according to (5.3). *Dubinin et al.* [1985] measured this effect with the *Intercosmos-Bulgaria-1300* satellite by calculating the phase shift between each Fourier component of E and δB field fluctuations. If the standing wave ratio is finite, $|\arg(\tilde{Z}(f))|$ is less than 90° and varies along the propagation direction. For traveling Alfvén waves ($\Gamma = 0$), E_x and δB_y are in phase. This is also true for static fields as one can see by letting $\omega = 0$ in Equation 5.3.

For Alfvén waves reflecting from a complicated medium such as the ionosphere, we can use the numerical model in Chapter 4 to predict the phase angle of $\tilde{Z}(f)$ in order to compare with measurements. An experimental measure of the phase angle of $\tilde{Z}(f)$ is given by the normalized cross-spectrum defined in general by

$$\tilde{C}_{12} = \frac{\left\langle \tilde{S}_1(f)\,\tilde{S}_2^*(f) \right\rangle}{\left\langle |\tilde{S}_1(f)|^2 \right\rangle^{1/2} \left\langle |\tilde{S}_2(f)|^2 \right\rangle^{1/2}} \tag{5.5}$$

where $\tilde{S}_1$ and $\tilde{S}_2$ are the Fourier transforms of the two time series to be compared, brackets denote ensemble averages, and the asterisk signifies the complex conjugate. $\tilde{C}_{12}$ will provide us not only with a phase relation between E_x and δB_y but also with a measure of the validity of that estimate. General discussions of the cross-spectrum and applications can be found in signal processing textbooks, e.g. *Papoulis* [1965] or *Jenkins and Watts* [1968].

The cross-spectrum has been used in a variety of geophysical experiments. For example, in coherent backscatter radar work $\tilde{S}_1$ and $\tilde{S}_2$ are the Fourier transforms of the received signals from spatially separated antennas, thus the phase angle $\arg(\tilde{C}_{12}(f))$ (known as the "phase spectrum") can be used to estimate the source location of scatterers and the magnitude $|\tilde{C}_{12}(f)|$ (the "coherency spectrum") contains information both on the signal-to-noise ratio and the spatial extent of the scatterer [*Farley et al.*, 1981; *Kudeki*, 1983; *Providakes*, 1985; *Sahr*, 1990]. *Labelle* [1985] used the cross-spectrum to measure plasma wave vectors by correlating the signals from spatially separated density probes on board a sounding rocket.

In our case we will take $\tilde{S}_1$ and $\tilde{S}_2$ to be the Fourier transforms of the meridional electric and zonal magnetic fields E_x and δB_y respectively. The phase angle given by the cross-spectrum measures the difference of the phase angles of the Fourier components of E_x and δB_y , which of course is the phase angle of $\tilde{Z}(f)$. If a constant phase relation between E_x and δB_y is maintained throughout the time series and the signal is not noisy, then the coherency $|\tilde{C}_{12}(f)|$ will be close to unity. Smaller values

can be due either to a small signal-to-noise ratio or to an E_x-δB_y phase shift which changes throughout the time series. In choosing the duration of our time series we must balance these two effects. Increasing the length of the time series will increase the number of averages contributing to the cross-spectrum and therefore reduce the statistical fluctuations, but the probability of the measured process maintaining a coherent phase throughout the series decreases.

A typical data set which we will analyze in this chapter has two data points per second and is 100 s long. We separate the series into segments of 32 points which overlap by 16 points, yielding 11 segments. The data are then prepared in the same way as described in the previous section: we subtract a least-squares linear fit from each segment, multiply by a Hanning window, and Fourier transform. We then operate on the resulting transforms as indicated in (5.5).

Since our cross-spectra have a relatively small number of individual time series which contribute to the ensemble average, we can expect that statistical fluctuations might cause the coherency at a given frequency to be large even in the presence of uncorrelated data. Also, overlapping data segments by N/2 points can lead to an artificially enhanced coherency. To arrive at a criterion for selecting statistically significant data we processed time series of mock E and δB data made up of Gaussian white noise. We then took ensemble averages of 5, 10, and 15 individual cross-spectra formed from 32 data points each, overlapping the data used to create each spectrum by 16 points. We repeated the process 20 times using different random data each time to obtain an idea of the range of coherencies that a random signal source can generate. The results are given in Table 5.1. Thus if we choose to average 10

spectra, we expect noise-generated coherencies (averaged over frequency) to fall between 0.27 and 0.37.

Table 5.1. Ensemble average of the frequency-averaged coherency calculated from 20 different time series consisting of Gaussian white noise.

# of individual spectra averaged for each cross-spectrum	$\langle \lvert \bar{C}_{12} \rvert \rangle$	std. dev.
5	0.453	0.058
10	0.323	0.049
15	0.285	0.035

5.4 Greenland I Rocket Data

In Section 3.5 we found the DC Poynting flux for the upleg of a Black Brant X sounding rocket flight made during the Greenland I campaign. The rocket was launched eastward into the morning auroral oval from Sondrestrom, and it remained in the auroral oval during the entire upleg. When plotted at high resolution, coherent Alfvén waves can be identified in the time domain data [*Boehm et al.*, 1990]. Since we know Alfvén waves are present in the field data, this data set is ideal for testing the impedance spectrum and normalized cross-spectrum as diagnostic tools in low frequency electromagnetic field studies.

To form $Z(f)$ and $\tilde{C}_{12}(f)$ for the rocket data we split the entire E_x (northward) and δB_y (eastward) time series plotted in Figure 5.1 into 32 point segments overlapping by 16 points, giving a total of 21 power

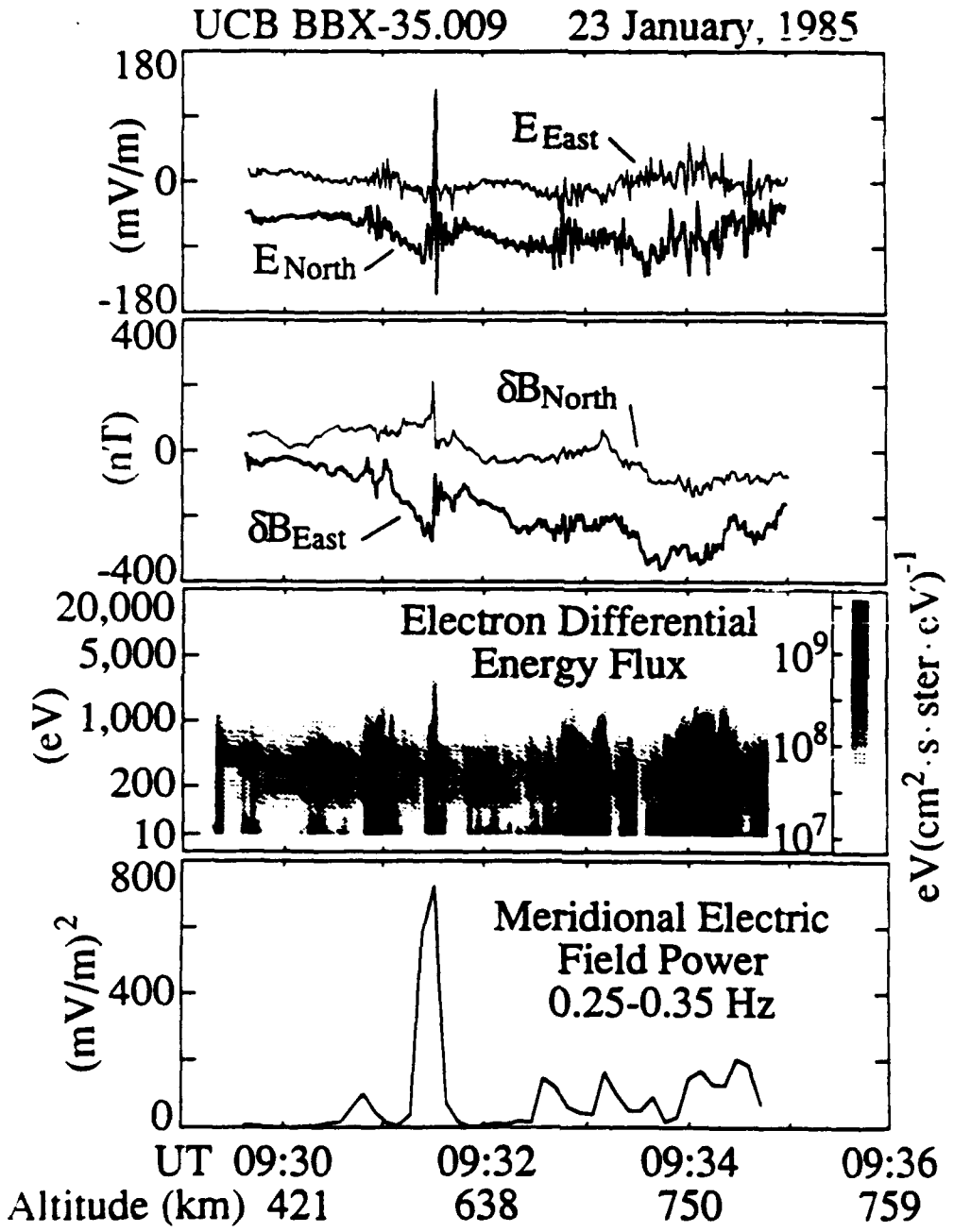


Figure 5.1 Data taken from a Black Brant X sounding rocket launched from Sondrestrom, Greenland on 23 January, 1985. The rocket traveled eastward along the auroral oval. (Data are courtesy of C. Carlson, B. McFadden, and M. Boehm at the University of California, Berkeley.)

spectra. The electric fields perpendicular to $\mathbf{B}_0$ were measured with perpendicular 3 m electric field booms, and magnetic measurements were taken with a fluxgate magnetometer. To obtain electric fields below the rocket spin frequency, electric field measurements were fit to a sine wave over time intervals of one spin period, resulting in a sampling period of 0.887 s. The impedance and phase spectra are shown in Figure 5.2. Also shown in Figure 5.2 are the predicted phase and impedance spectra from the numerical model (Chapter 4) assuming a constant measurement altitude of 600 km and a horizontal spatial scale λ_x of 1000 km. The density profile used as input to the model is shown in Figure 5.3 and was parameterized from a profile measured with a Langmuir probe on board a Terrier-Malemute rocket launched nearly simultaneously with the Black Brant X [*Earle*, 1988]. For reference we have plotted Σ_P^{-1} in Figure 5.2 as deduced from measurements made by the Sondrestrom radar during the rocket flight.

The data curve in the bottom panel of Figure 5.2 clearly shows that the phase between E_x and δB_y varies with frequency. In either the static model or the traveling Alfvén wave model, electric and perturbation magnetic fields are shifted by 0° or 180°. In the coordinate system we are using, a 180° phase indicates a downward-directed Poynting vector, which is the case in the zero frequency limit of Figure 5.2 for both theory and experiment. At higher frequencies the measured phase increases to a maximum of nearly 260° at 0.4 Hz. Fields such as these which are nearly out of phase are exactly what is expected in standing Alfvén waves, as shown by the theoretical curve plotted with the data. Thus with the phase spectrum in this case we are able not only to

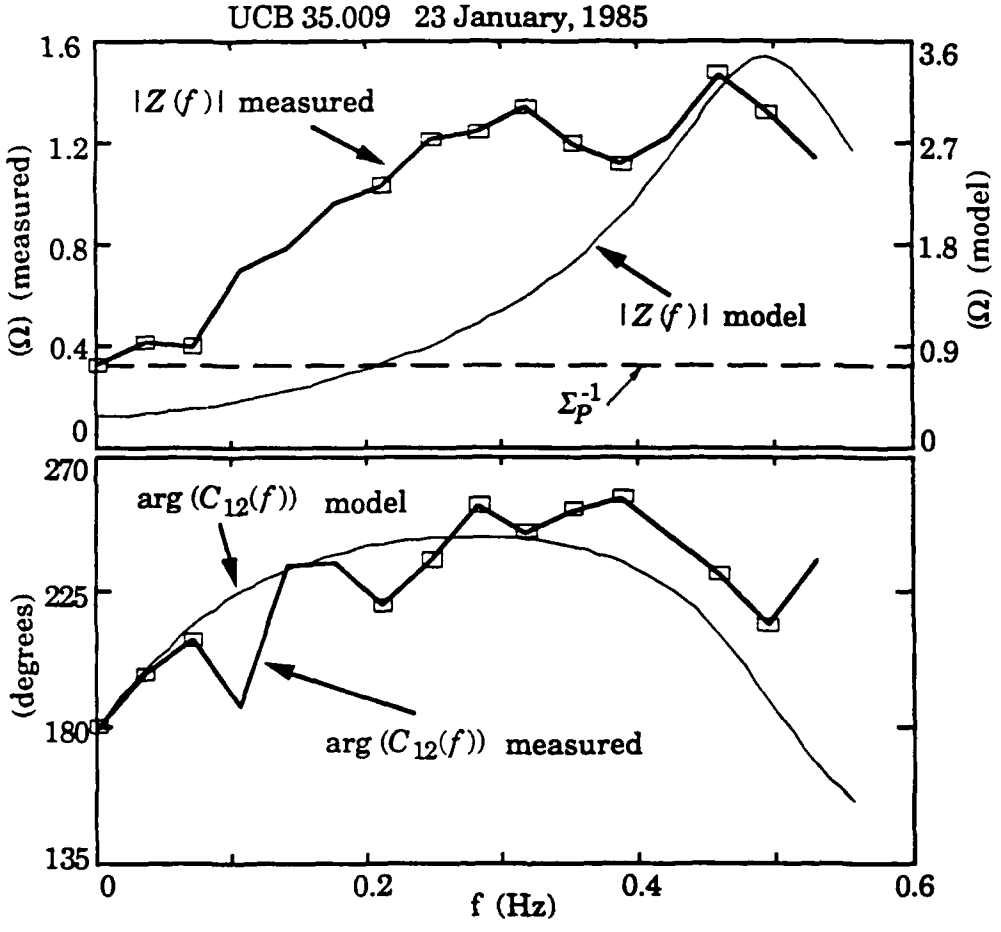


Figure 5.2 Numerical results compared with sounding rocket data averaged over the entire interval shown in Figure 5.1. Averages were formed from 21 sub-intervals of 32 points each.

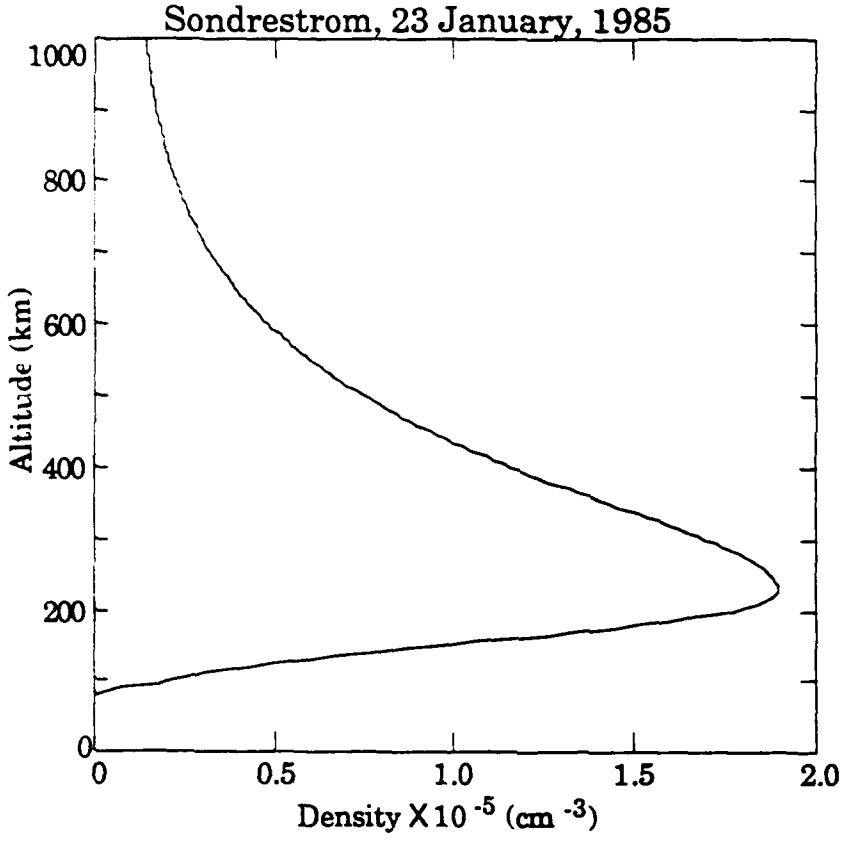


Figure 5.3 Electron density profile used as model input for the curves in Figure 5.2. This profile approximates the density profile taken on board a Terrier-Malemute rocket launched nearly simultaneously and in the same direction as the Black Brant [see *Earle*, 1988].

detect the presence of Alfvén waves, but we can determine that both incident and reflected components are present.

The impedance spectrum in Figure 5.2 is also consistent with the standing wave model. A trend toward increasing impedance at higher frequencies is clearly visible. The static field model predicts $Z(f) = \Sigma_P^{-1}$ for all Doppler-shifted frequencies. As shown in Figure 5.2, the standing Alfvén wave model predicts a peak in the impedance spectrum at 0.5 Hz.

Qualitatively, the theoretical and experimental impedance curves match well. However, the latter is less peaked, and the peak is much lower (1.5 vs. 3.5 Ω) than the theory curve. There are at least four possible reasons for this: 1) The model assumes uniform ionospheric density, collision frequencies, etc. in the horizontal direction, which is not the case in the auroral oval. Ionospheric structure will tend to smear the peaked nature of the electric and magnetic field spectra, which will in turn broaden the impedance spectrum and decrease its peak value. 2) If static fields are present in addition to Alfvén waves, the mixture of these two will tend to make the impedance fall somewhere between the pure standing wave impedance and Σ_P^{-1}. 3) The measured spectral power at a single frequency is actually an average of the power in a frequency range Δf. Thus any spectral peaks will be reduced by averaging with neighboring values. Finally, 4) we have assumed a constant measurement altitude for the numerical model's predictions, but in fact the rocket altitude varies between 400 and 770 km during the interval we have analyzed. The altitude dependence of the standing wave impedance will thus lead to smearing in $Z(f)$. However, while the rocket actually traverses about 350 km in altitude during the time interval we are interested in, over half of the data are taken in the upper 100 km due to

the parabolic time dependence of the rocket altitude. Running the numerical model for a 700 km rather 600 km measurement altitude decreases the frequency of the impedance peak only by about 5%, therefore we suspect that the altitude variation of the sounding rocket is not the primary cause of small measured impedance values.

The bottom two panels of Figure 5.1 allow us to compare the fluctuating electric field power integrated between 0.25-0.35 Hz with electron precipitation. The two quantities are well correlated and we may conclude that much of the Alfvén wave energy in this frequency band is spatially coincident with auroral arcs. The sounding rocket was traveling eastward, nearly parallel to auroral structures. This allowed the rocket to dwell in the vicinity of arcs and enhanced Alfvén wave activity. Polar-orbiting satellites fly perpendicular to the auroral oval, and as a consequence they traverse auroral arcs very quickly. We will see in the next section that this limits the amount of Alfvén wave energy one can measure from the HILAT satellite.

We have used the impedance and cross-spectrum to verify the presence of Alfvén waves in sounding rocket data, and also to show that incident and reflected wave components are interfering as a result of the highly conducting, and therefore reflecting, ionosphere. We will now turn to HILAT measurements of $Z(f)$ and $\tilde{C}_{12}(f)$ to search for evidence of Alfvén waves which are not obvious in the time domain data alone.

5.5 HILAT Satellite Data

In order to find HILAT data with a significant correlation between E_x and δB_y we studied the 26 different HILAT passes listed in Table 5.2. (In the HILAT coordinate system, x is the direction of the

Table 5.2. HILAT passes searched for 100 s intervals with a frequency-averaged E-δB coherency exceeding 0.5. Passes with an asterisk satisfy this criterion.

Year	Day	Start Time (UT)	MLT at 60°	Solar Zenith Angle (deg)
1983	344*	13:05	11:00	85
1984	019	09:19	05:45	115
1984	023	09:21	05:15	120
1984	047	18:55	16:30	70
1984	058	17:33	14:15	65
1984	063*	17:04	13:30	60
1984	067	02:05	00:00	130
1984	075	15:16	12:00	55
1984	096*	11:16	11:15	60
1984	122	18:45	18:45	60
1984	164*	12:23	11:45	35
1984	179*	10:58	10:00	45
1984	217*	04:37	05:15	90
1984	242*	03:56	01:45	110
1984	245	04:19	04:30	120
1984	261*	15:43	12:30	50
1984	318*	19:54	16:00	85
1984	329	16:48	17:00	80
1984	345	14:56	12:45	75
1984	346	14:30	12:45	75
1985	052	20:09	16:45	80
1985	089	00:59	23:00	115
1985	112	12:42	12:45	55
1985	148	17:30	17:45	45
1985	265	17:06	13:45	50
1985	277*	15:15	12:15	55

satellite's velocity, $\hat{z}$ is downward, and $\hat{y}$ completes the right-hand system.) Each pass contains roughly 10 minutes of data taken in the northern hemisphere auroral oval and polar cap. The HILAT magnetometer samples at 20 s^{-1} and the electric field is derived from ion drift measurements taken either at 16 or 32 s^{-1}. Since for our purposes we are only interested in the frequency components below 1 Hz we averaged the field quantities to 2 samples/s, or one measurement every 3.7 km. We did not filter out the lowest frequency magnetic field variations due to mechanical oscillations of the satellite as we did in Chapter 3 because those variations are well below the range of frequencies we will consider here. The magnetometer resolution is about ±6.7 nT, but averaging gives an effective resolution somewhat lower than this.

We separated the data from the 26 passes in Table 5.2 into 284 data segments of 100 s each and found the coherency spectrum from 11 sub-segments as described in Section 5.3. The resulting coherency from each 100 s interval we then averaged over frequency to obtain a single measure of the E_x-δB_y correlation. By averaging the coherency spectrum over all frequencies as a test to find meaningful correlations, we allow for the possibility of low correlation near DC but high correlation at higher frequencies.

Figure 5.4a shows the distribution of frequency-averaged coherencies $\overline{C_{12}}$ for the 288 100 s intervals obtained from the HILAT data survey. Figure 5.4b shows for comparison the distribution of coherencies for 284 100s intervals consisting of Gaussian white noise. The 3 intervals with values of $\overline{C_{12}}$ exceeding 0.8 are from Day 217, 1984 and occur during periods of extremely small electric and magnetic fields. Using straight

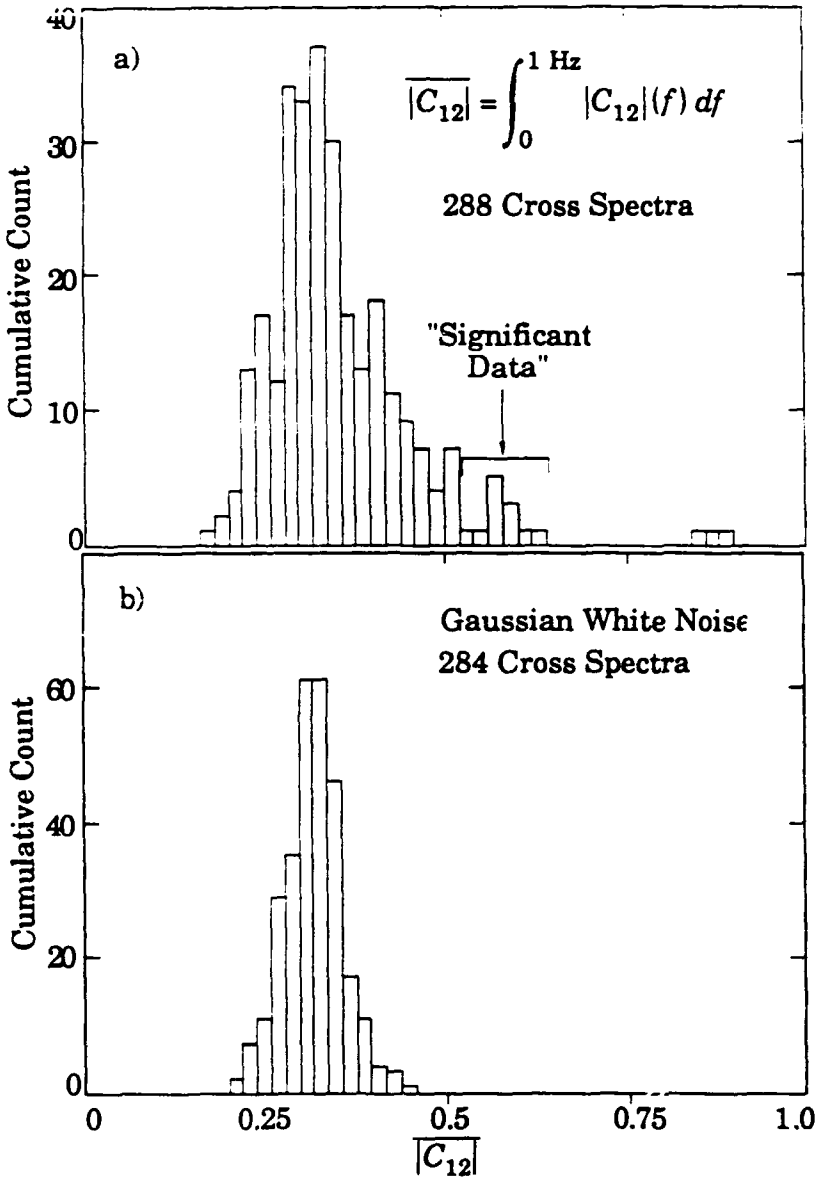


Figure 5.4 Distribution of frequency-averaged coherency spectra from a) HILAT data, and b) Gaussian white noise. Each coherency spectrum was formed from 11 electric and magnetic field spectra, which were in turn formed from 32 data points each.

lines (i.e. $E(t) = a_1 t + c_1$, $\delta B(t) = a_2 t + b_2$) alone as input to the analysis routines yields coherencies of unity, and we suspect that the Day 217 coherencies are high for this same reason. That is, any segment of the magnetic field data has a low-order trend due to the oscillation of the satellite, and when no geophysical signals are present, the time series resulting from this trend can generate anomalously high coherencies. The passes of most interest to us are those with a noticeable amount of signal energy present and with $\overline{C_{12}}$ above 0.5, indicated by the arrow in Figure 5.4a.

Notice from Figure 5.4b that even though the noise used to calculate the cross-spectra has a coherence of zero, the apparent coherence is near 0.3. This is due to the small number of data segments contributing to the ensemble average (i.e. N = 11). The average coherency in Figure 5.4b will decrease with increasing N, roughly as $N^{-1/2}$. In fact, it is easy to verify that the average coherencies in Table 5.1 fall quite close to $N^{-1/2}$ in magnitude.

Having now identified several promising data intervals we will analyze in detail six examples which were taken when the Sondrestrom Incoherent Scatter Radar was scanning more or less along the N-S meridian. The radar measured ionospheric density profiles up to 850 km altitude and over a 800 km range in latitude (depending on the measurement altitude). The density profiles were averaged over latitude to obtain a single altitude profile [*M. McCready and J. F. Vickrey*, personal communication, 1990]. Since the density profiles are averaged in latitude, some of the apparent structure in altitude may actually be due to horizontal structuring in plasma density.

Often the upper range gates of the radar data are corrupted either by a poor signal-to-noise ratio or by echoes from man-made satellites. When this appears to be the case we replace the top portion of the radar-measured profile with an exponentially decreasing density to use as input to our numerical model. We also smooth the radar profiles, and we extrapolate the profile below 95 km using a 6 km exponential scale height. The high-coherency HILAT data in each case were not necessarily taken in the same region that the radar measured.

In Figures 5.5-5.16 we show in chronological order the radar density profiles (both measured and smoothed), $E_x(f)$ and $\delta B_y(f)$ for the 100 s intervals of interest, and measured and modeled cross-spectra and impedance spectra for each of the six HILAT passes. While we present only the high-coherency data intervals, it is evident from the complete data sets that in each pass the large coherencies are found within the auroral oval, as indicated by large scale magnetic perturbations from the Region 1/Region 2 current systems. Unfortunately, unlike the sounding rocket example of the previous section, in all six HILAT passes it appears that the measured phase spectra have very large variances, and a comparison between measured and modeled curves is unconvincing. All cases do show a 180° phase shift between E_x and δB_y, at DC however, which indicates a downward-directed Poynting vector for quasi-static fields.

The measured and modeled impedance spectra are in better agreement than the measured and modeled phase spectra. On Day 096, 1984 (Figure 5.10), Day 179, 1984 (Figure 5.12), and Day 277, 1985 (Figure 5.16), the density profiles are somewhat similar and the numerical model predicts a single impedance peak in the 0.3-0.6 Hz range. In all three of

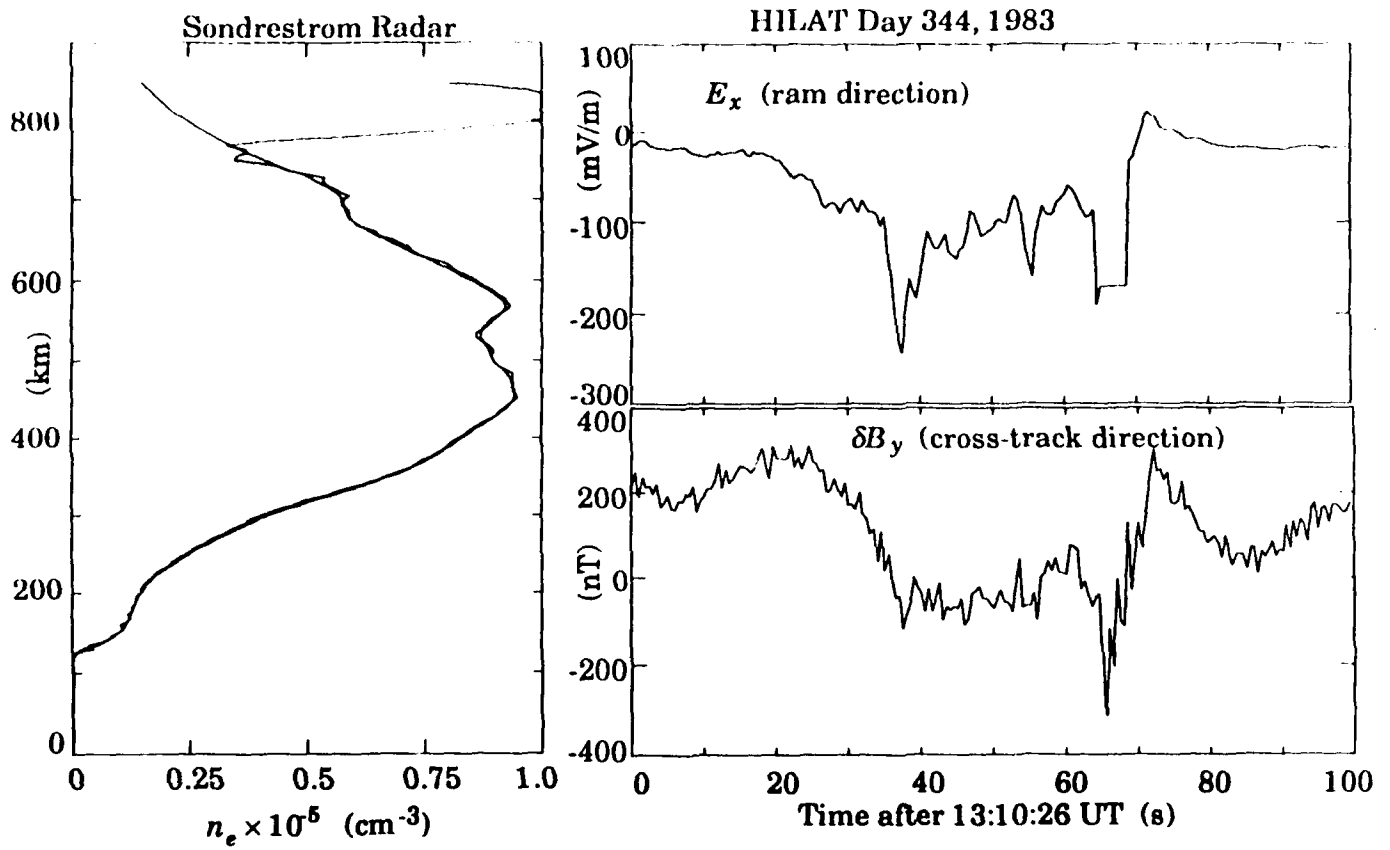


Figure 5.5 Smoothed and unsmoothed density profiles taken by the Sondrestrom radar on 10 December, 1983 and averaged in latitude (left), taken at the same time as the HILAT electric and perturbation magnetic field data shown at right.

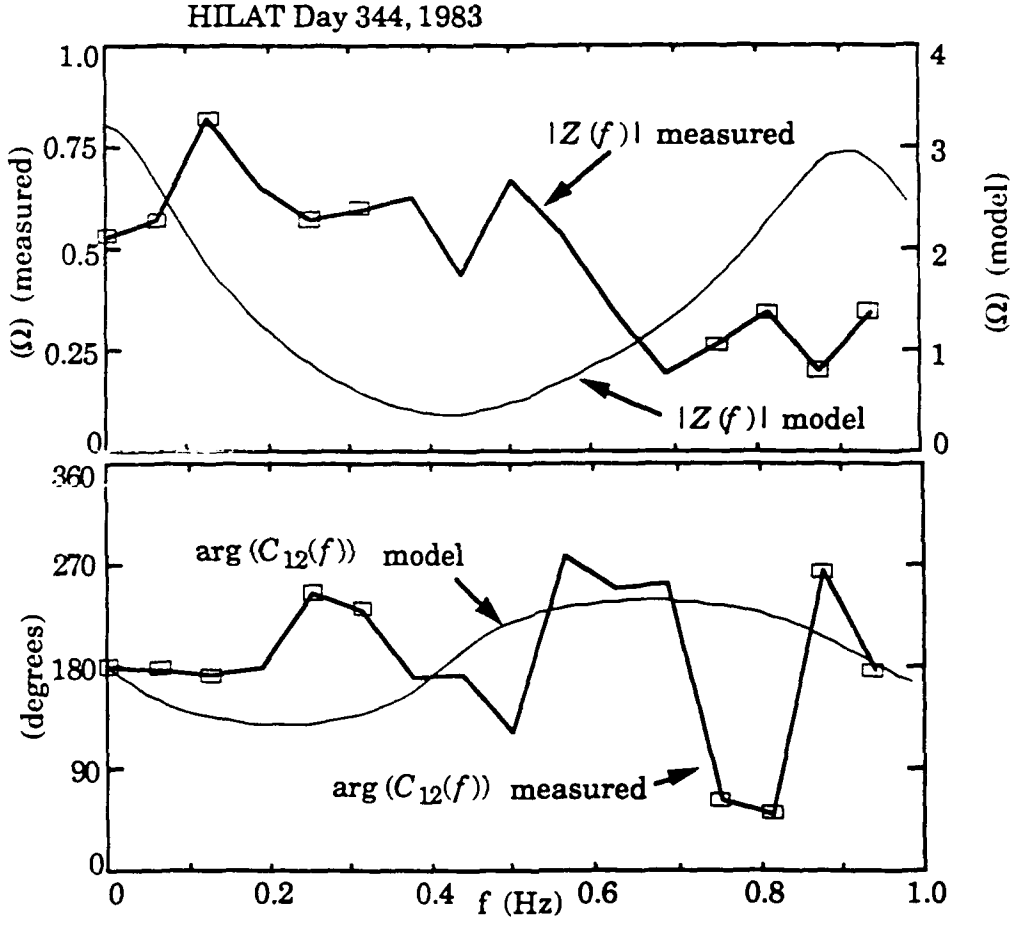


Figure 5.6 Comparison of numerical model and experimental results using the smoothed density profile and electric and magnetic fields shown in Figure 5.5. Ensemble averages were formed from 11 separate 32 point (16 s) intervals overlapping by 16 points each. Boxes indicate a coherency exceeding 0.5.

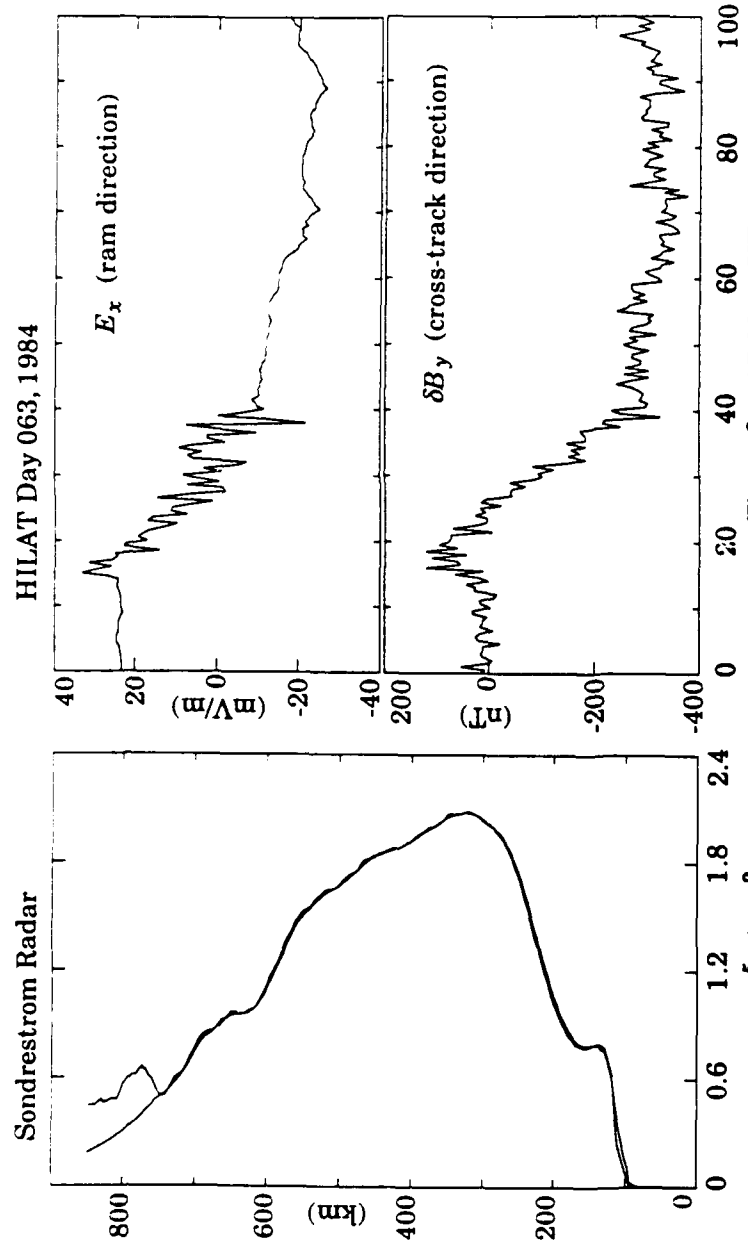


Figure 5.7 Smoothed and unsmoothed density profiles taken by the Sondrestrom radar on 3 March, 1984, and averaged in latitude (left), taken at the same time as the HILAT electric and perturbation magnetic field data shown at right.

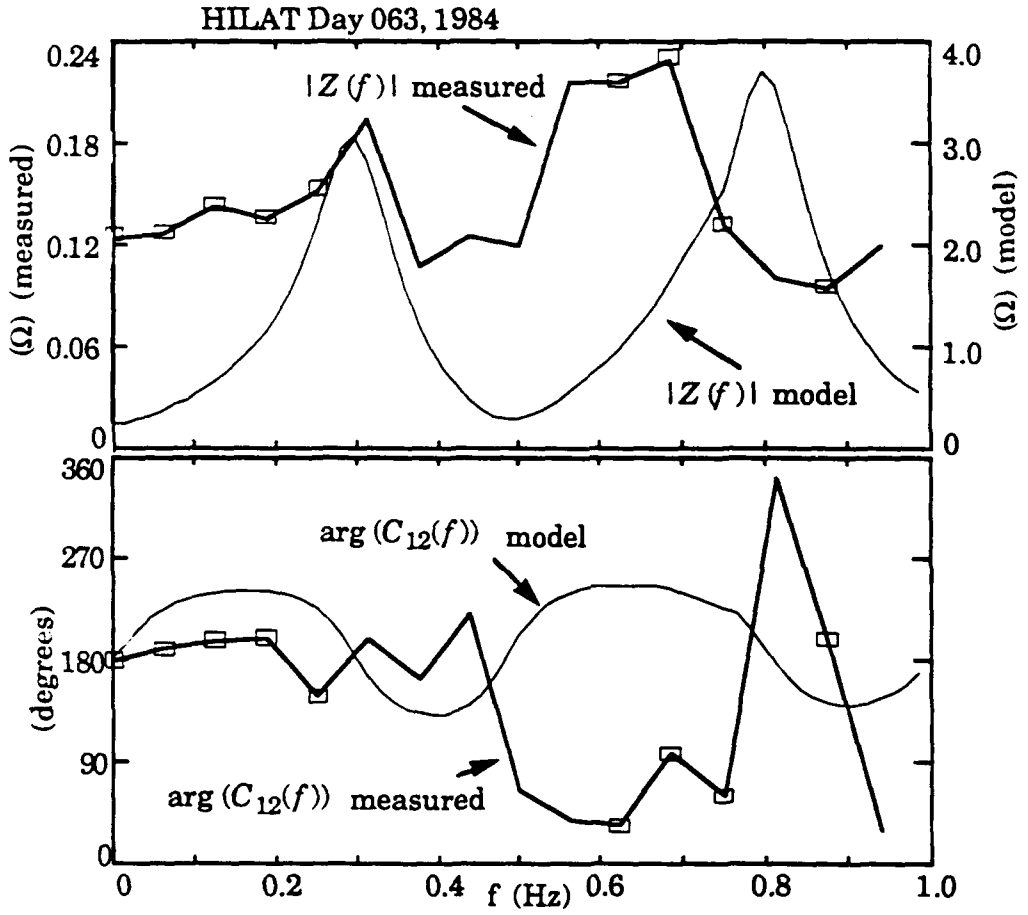


Figure 5.8 Comparison of numerical model and experimental results using the smoothed density profile and electric and magnetic fields shown in Figure 5.7. Ensemble averages were formed from 11 separate 32 point (16 s) intervals overlapping by 16 points each. Boxes indicate a coherency exceeding 0.5.

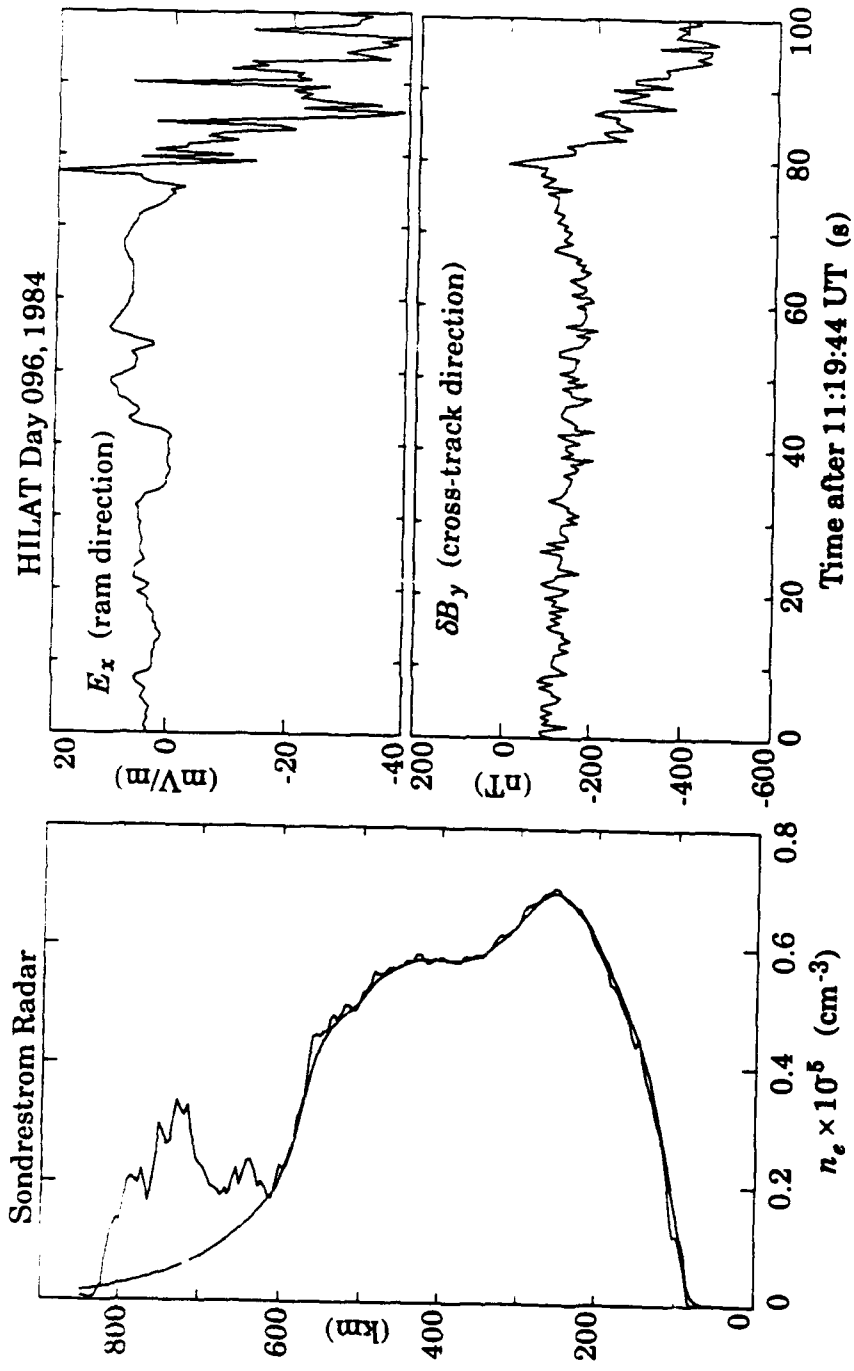


Figure 5.9 Smoothed and unsmoothed density profiles taken by the Sondrestrom radar on 5 April, 1984, and averaged in latitude (left), taken at the same time as the HILAT electric and perturbation magnetic field data shown at right.

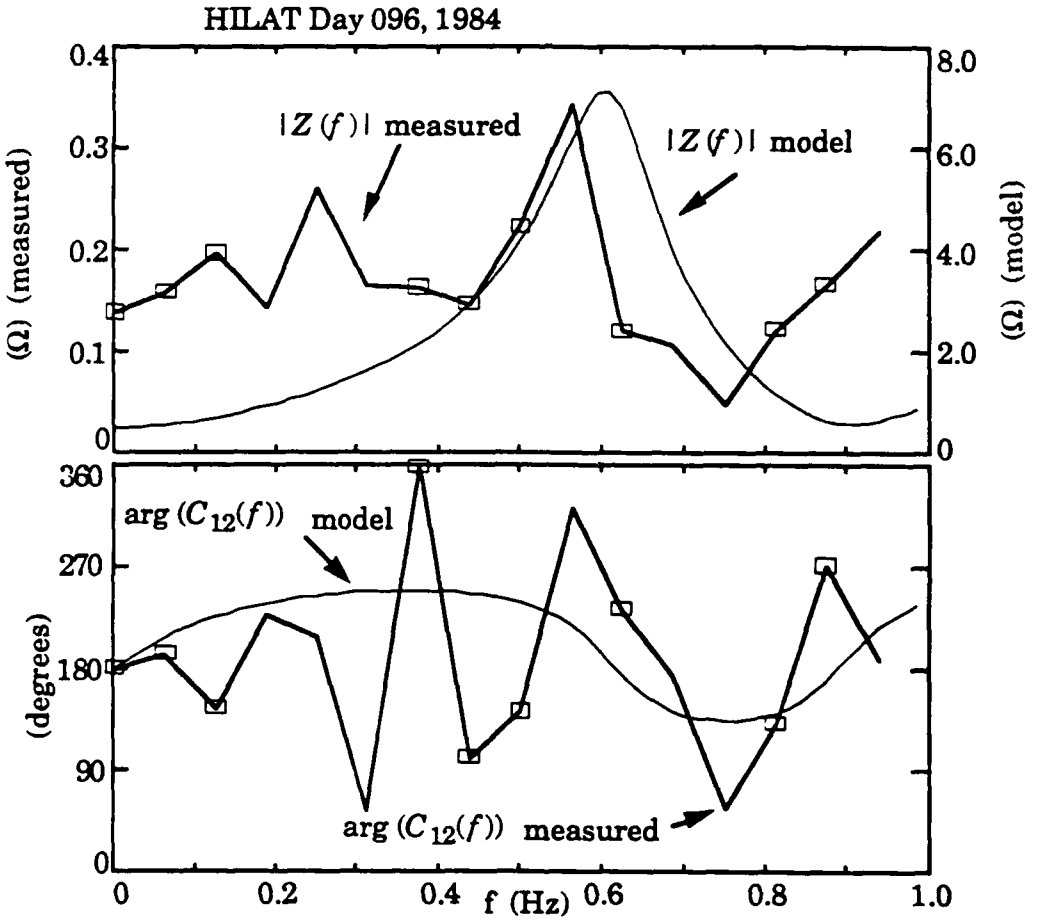


Figure 5.10 Comparison of numerical model and experimental results using the smoothed density profile and electric and magnetic fields shown in Figure 5.9. Ensemble averages were formed from 11 separate 32 point (16 s) intervals overlapping by 16 points each. Boxes indicate a coherency exceeding 0.5.

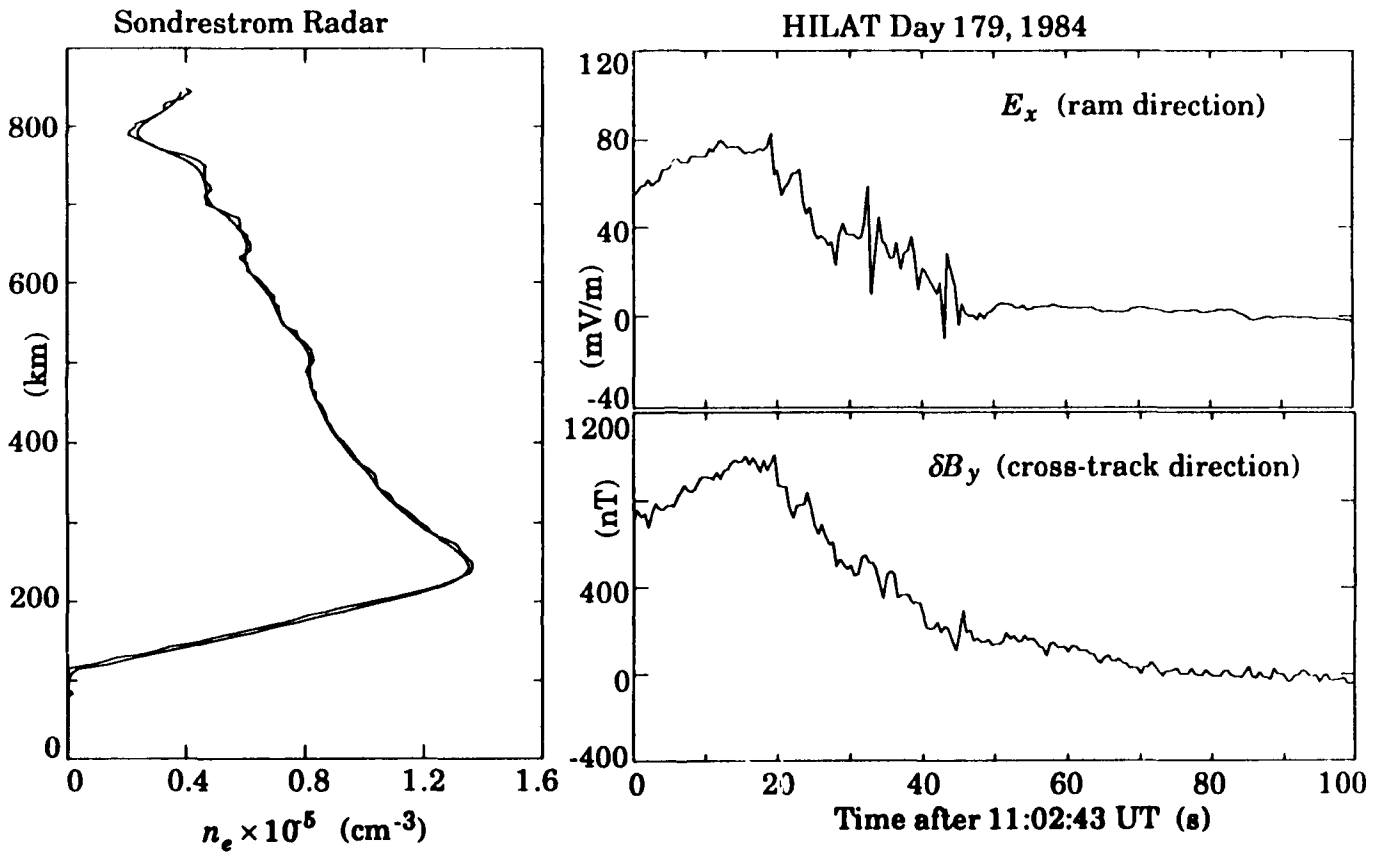


Figure 5.11 Smoothed and unsmoothed density profiles taken by the Sondrestrom radar on 27 June, 1984, and averaged in latitude (left), taken at the same time as the HILAT electric and perturbation magnetic field data shown at right.

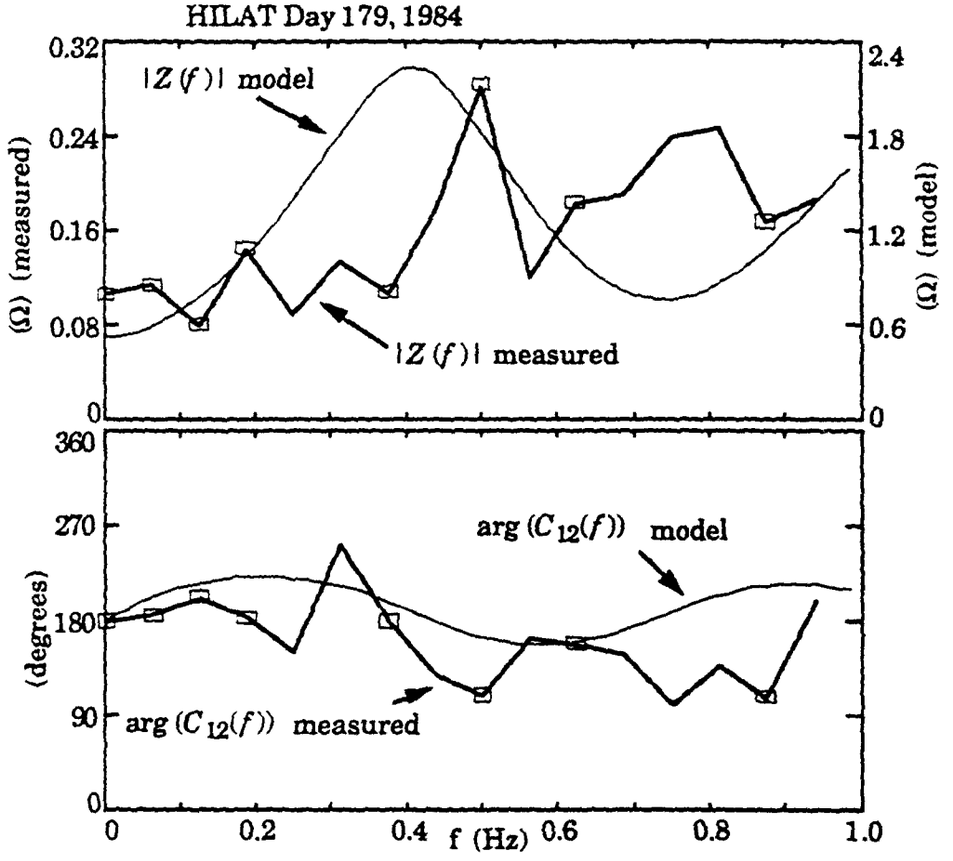


Figure 5.12 Comparison of numerical model and experimental results using the smoothed density profile and electric and magnetic fields shown in Figure 5.11. Ensemble averages were formed from 11 separate 32 point (16 s) intervals overlapping by 16 points each. Boxes indicate a coherency exceeding 0.5.

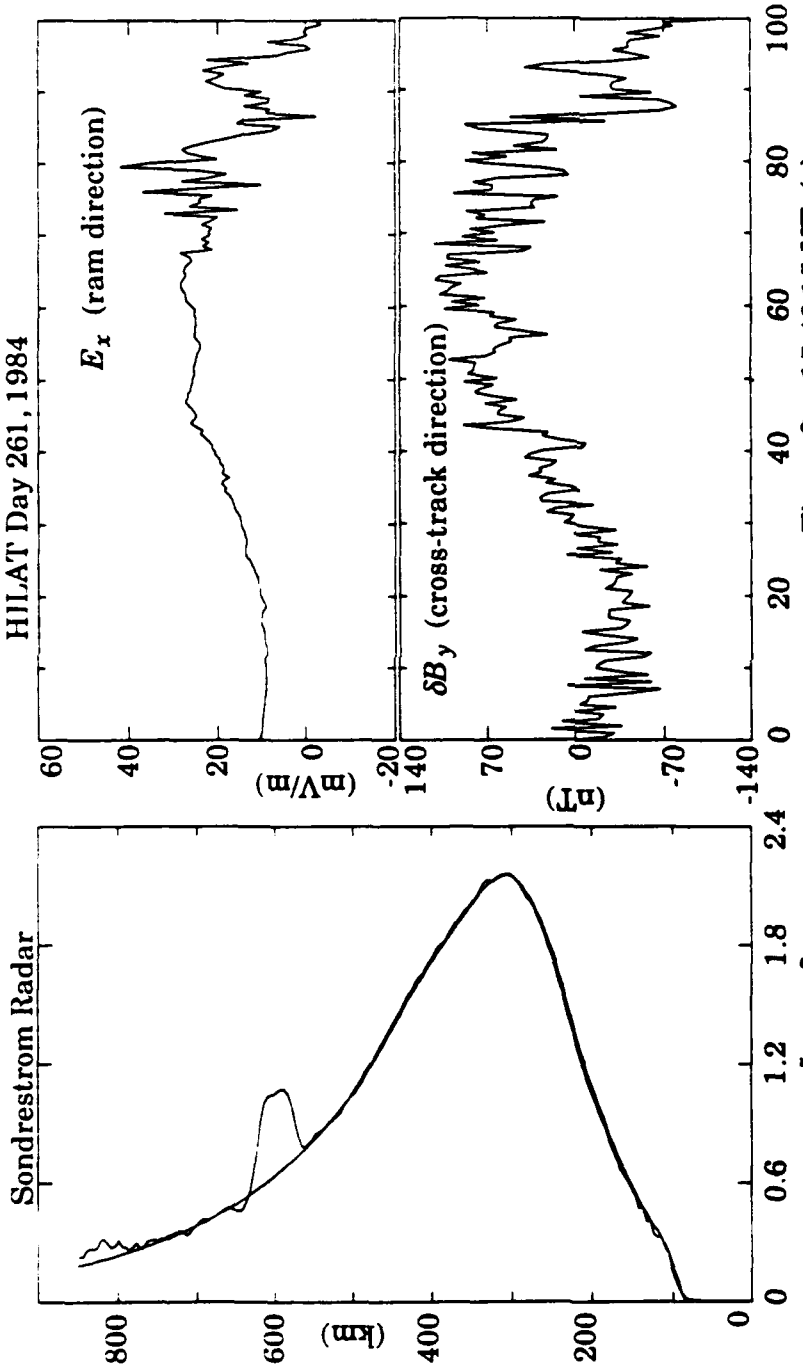


Figure 5.13 Smoothed and unsmoothed density profiles taken by the Sondrestrom radar on 17 September, 1984, and averaged in latitude (left), taken at the same time as the HILAT electric and perturbation magnetic field data shown at right.

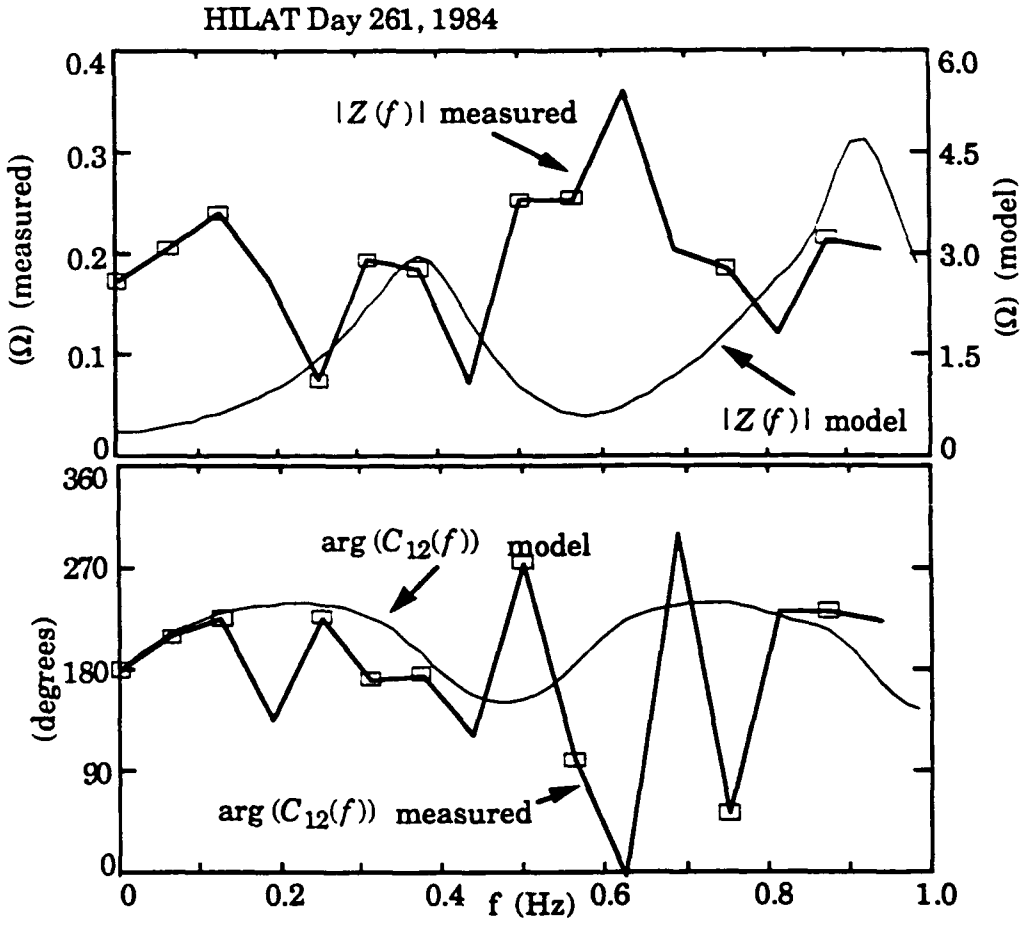


Figure 5.14 Comparison of numerical model and experimental results using the smoothed density profile and electric and magnetic fields shown in Figure 5.13. Ensemble averages were formed from 11 separate 32 point (16 s) intervals overlapping by 16 points each. Boxes indicate a coherency exceeding 0.5.

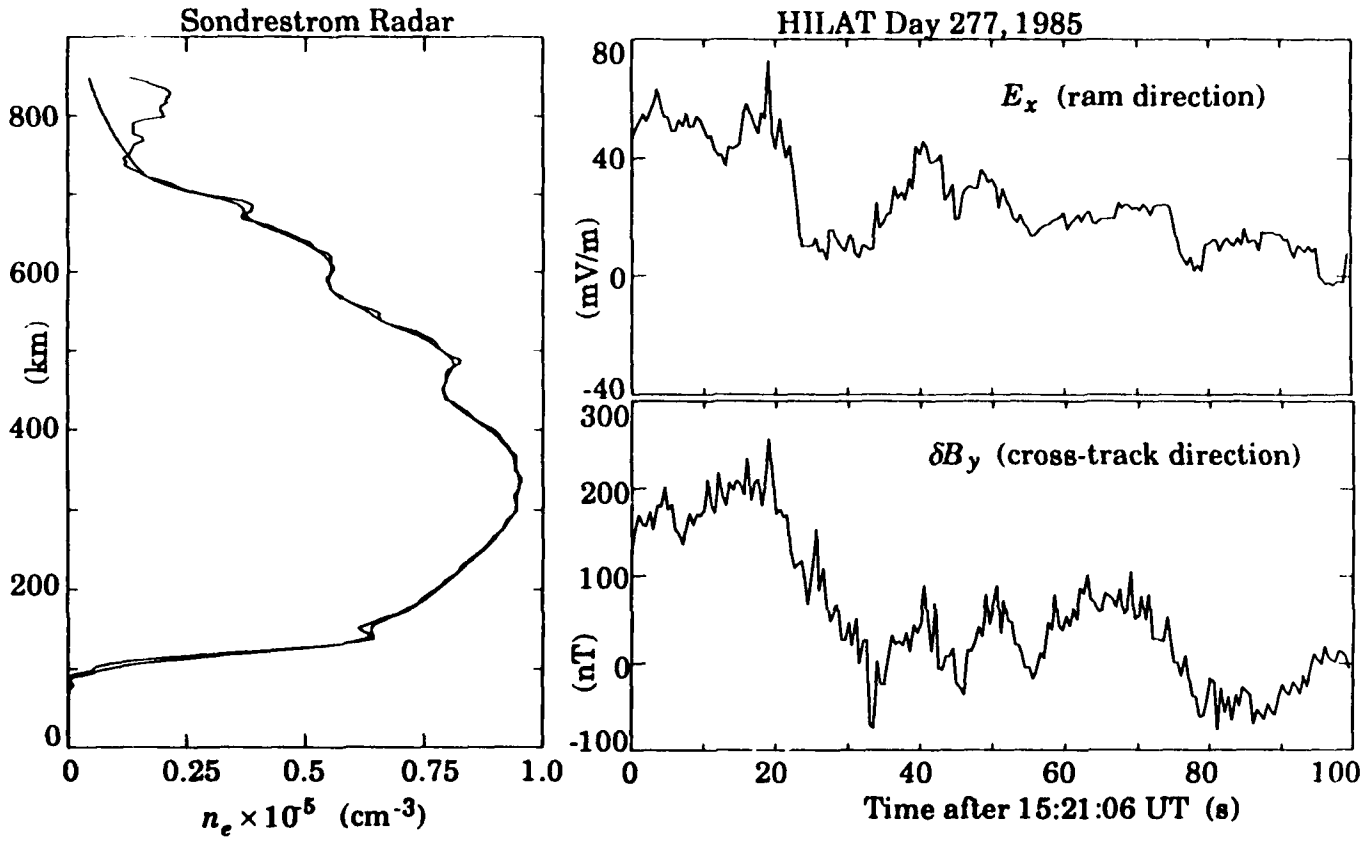


Figure 5.15 Smoothed and unsmoothed density profiles taken by the Sondrestrom radar on 4 October, 1985, and averaged in latitude (left), taken at the same time as the HILAT electric and perturbation magnetic field data shown at right.

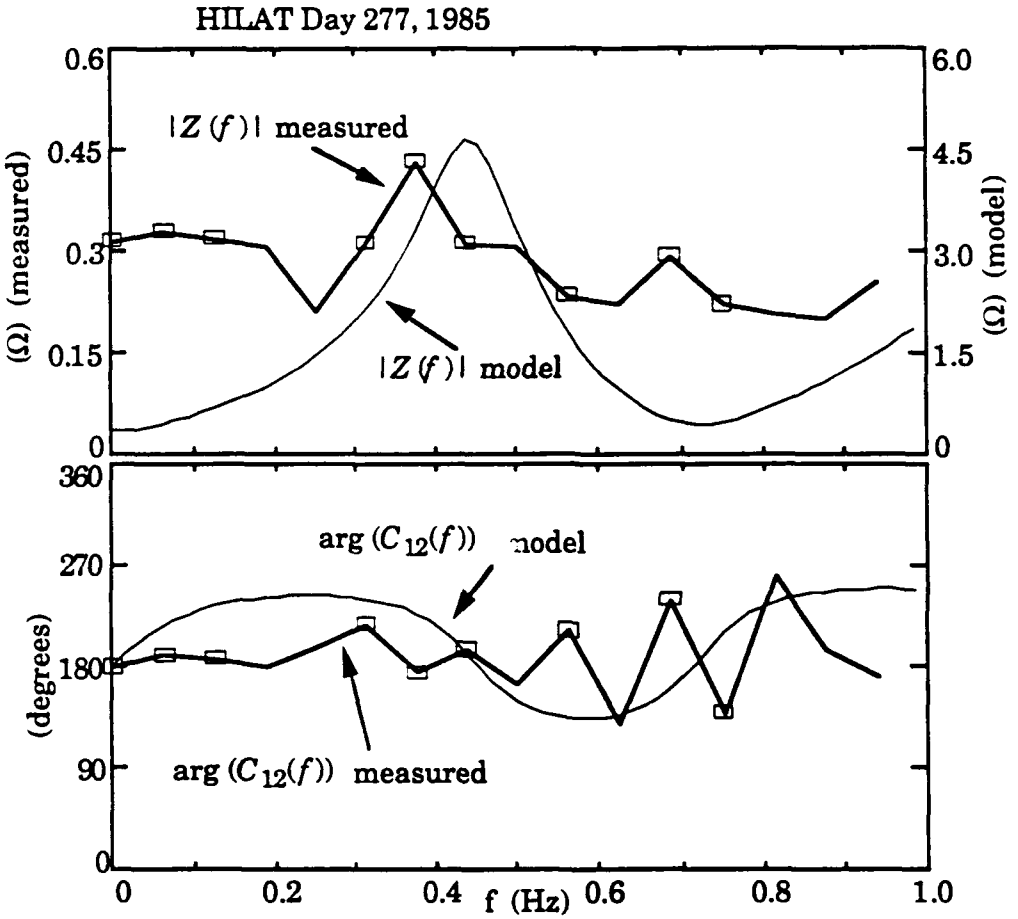


Figure 5.16 Comparison of numerical model and experimental results using the smoothed density profile and electric and magnetic fields shown in Figure 5.15. Ensemble averages were formed from 11 separate 32 point (16 s) intervals overlapping by 16 points each. Boxes indicate a coherency exceeding 0.5.

these cases, the maximum measured impedance value falls within about 15% of the model's prediction. Unfortunately, the peak measured impedance on Day 096, 1984 (Figure 5.10) does not have a coherence above 0.5 associated with it, and therefore the fact that it lies near the modeled peak may be coincidental. On Day 063, 1984 (Figure 5.8), the ionospheric density profile was such that there are two impedance peaks below 1 Hz, and two peaks are clearly visible in the data. However, only the second impedance peak has a coherency above 0.5.

On Day 261, 1984 (Figure 5.14), there is a poor match between the model and experiment, and in fact the peak measured impedance occurs at a frequency for which the modeled impedance is minimum. A possible reason for the poor match is that the average density profile measured by the radar is not representative of the actual ionosphere below HILAT when Alfvén wave energy, if any, was present.

Day 344, 1983 (Figure 5.6) is an especially interesting case. On other passes and in the sounding rocket data the measured impedance function starts near Σ_P^{-1} at zero frequency and tends to increase. This is a consequence of a highly conducting ionosphere with $\Sigma_P^{-1} < Z_A$ where Z_A is the Alfvén impedance at the measurement altitude. One can see from the Day 344 density profile in Figure 5.5 that there is not much density in the E region, which causes a Σ_P^{-1} (~3 Ω) greater than Z_A (~2 Ω). In this case the model standing wave impedance is maximum at 0 Hz and decreases with frequency. While the measured impedance values do not match the modeled values, the low order trend in the measured Z does decrease by a factor of 3 at higher frequencies, presumably because of the low E-region density.

Turning again to the example with two impedance peaks in both the measured and modeled data (Figure 5.8), we find that the frequency of the second measured impedance peak is .2 Hz smaller than predicted by the model. It is interesting to note that in the frequency range of the measured impedance maximum (0.6 - 0.8 Hz), the measured cross-spectrum has a high coherency and the phase is at a minimum, while the modeled phase maximizes. Comparing the measured and modeled phase spectra, one might expect that the input to the numerical model could be adjusted to bring them into agreement. One would have to find a parameter that affects only the higher frequency part of the curve while leaving the agreement between the impedance peaks at 0.3 Hz unaffected. Decreasing the electron density in the E and F regions will move both impedance peaks towards larger frequencies, thus an inaccurate density profile is probably not the cause for the disparity between the data and modeled curves.

In contrast to the electron density profile, the ion collision frequency ν_i near 150 km *can* affect only the high frequency part of the impedance and phase spectra. The reason for this is that, roughly speaking, Alfvén waves penetrate down into the ionosphere as long as the wave frequency ω exceeds ν_i. This effect is illustrated in Figure 5.17, where we have modeled the meridional electric field magnitude $|E_x|$ for two different frequencies. We have used the "EF" density profile in Figure 4.2 as input to the model. If one defines the Alfvén wave reflection altitude as the altitude near the E region where $|E_x|$ is minimum, then Figure 5.17 shows that a 1.2 Hz Alfvén wave penetrates to about 140 km, whereas a 0.5 Hz wave reflects at 180 km. Decreasing ν_i below 150 km could then lower the reflection altitude for waves with frequencies near 1

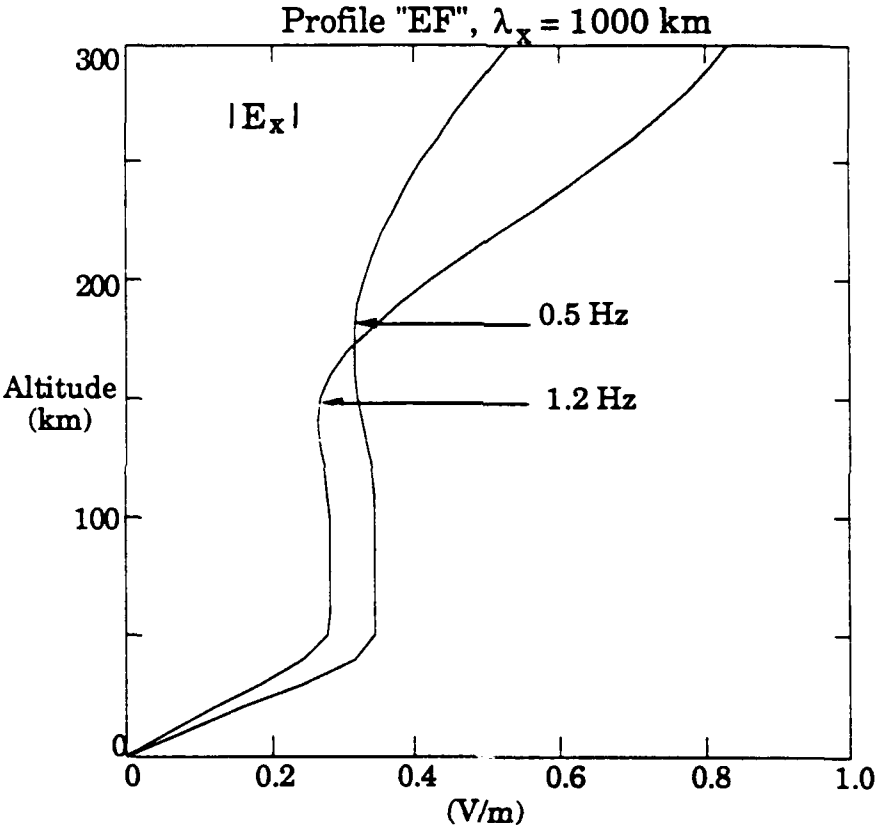


Figure 5.17 Model meridional electric field profiles for two Alfvén waves illustrating the fact that higher frequency waves reflect from lower altitudes.

Hz, which in turn would decrease the frequency of maximum $|Z|$ for those waves. Our modeled profiles of ν_i are based on a neutral atmosphere model with an assumed thermospheric temperature of 1000 K. The actual thermospheric temperature can vary widely, and as a result, the neutral atmosphere density and ion-neutral collision frequency can also vary. Thus the lack of agreement between our modeled and measured frequencies of peak $|Z(f)|$ might be explained in part by our poor knowledge of $\nu_i(z)$.

The frequency dependence of the reflection altitude also explains why the frequency of the second impedance peak in Figure 5.8 (0.8 Hz) is not three times the frequency of the first peak (0.3 Hz). This factor of three relation would hold for a constant reflection altitude because impedance peaks occur when an observer is $\lambda/4$ and $3\lambda/4$ above the reflector. But since higher frequency waves reflect from lower altitudes in the case of ionospheric reflections, the frequency of the second impedance peak is somewhat lower than three times the frequency of the first peak.

While we have found reasonable agreement between measured and modeled frequencies of the frequencies of maximum $|Z(f)|$, in all cases the measured impedance is much smaller than predicted by the standing Alfvén wave model, and in fact it is less than Σ_P^{-1} in most cases, although this also may indicate a problem with the neutral atmosphere or collision frequency models.

We pointed out that the sounding rocket discussed in the previous section traveled mostly parallel to auroral structures and therefore spent a significant amount of time in regions where Alfvén waves seem to occur the most. This could account for the good data/theory match in

that case, especially in the phase spectrum. HILAT flies mostly perpendicular to auroral structures and probably spends less time in regions with Alfvén waves. In the HILAT passes we have analyzed it appears much of the field energy is due to quasi-static fields with an associated impedance function $Z(f) = \Sigma_P^{-1}$. Even so, the impedance function in most cases shows a significant increase at frequencies for which it is predicted to be maximum by the standing wave model. Thus HILAT must have flown through regions of Alfvénic fluctuations, and most importantly *the frequencies of detectable Alfvén waves are determined by the standing wave pattern, not the wave source*. That is, in almost all cases there is at least some Alfvén wave energy present at peaks in the standing wave impedance function, thus the magnetosphere is apparently supplying a continuous spectrum of waves, but only at those frequencies corresponding to peaks in $Z(f)$ does the electric field from Alfvén waves constitute a significant fraction of the total electric field spectrum measured across the auroral oval.

5.6 A Quantitative Estimate of the Amount of Alfvén Wave Energy in Electromagnetic Field Data

We have found measured impedances in the sounding rocket and satellite data which fall somewhere between Σ_P^{-1} and the numerical model's prediction. Roughly speaking, a higher measured impedance means more spectral power due to Alfvén waves and less to Doppler-shifted static structures. We can attempt to state this relationship quantitatively by assuming that there is no smearing in the measured impedance spectrum due to variations in the Pedersen conductivity along the flight path, and that at any given time the measurements are due

completely either to Alfvén waves or to static fields. (The case in which both Alfvén waves and Doppler-shifted static fields are present at the same time almost certainly occurs but is difficult to treat since we have no way of knowing the relative phase between the Fourier components of the two contributions.) Under this last assumption we can conceptually separate the total power spectral density in a time series into 2 parts, i.e. $P_E = P_{E,A} + P_{E,S}$ and $P_B = P_{B,A} + P_{B,S}$,where $P_{E,A}$ and $P_{B,A}$ are the power spectral densities of electric and magnetic fields from sub-intervals containing only Alfvén waves, and $P_{E,S}$ and $P_{B,S}$ are from those sub-intervals containing static structures. The fraction of the electric field spectral power due to Alfvén waves at frequency f is

$$q_E(f) = P_{E,A}/(P_{E,A} + P_{E,S}) \tag{5.6a}$$

and for the magnetic field

$$q_B(f) = P_{B,A}/(P_{B,A} + P_{B,S}) \tag{5.6b}$$

Recognizing that $\mu_0^2 P_{E,S}/P_{B,S} = 1/\Sigma_P^2$ and $\mu_0^2 P_{E,A}/P_{B,A} = Z^2_{model}$ (Z_{model} is the numerical model's prediction of the standing wave impedance at a fixed altitude) leads to

$$q_E(f) = (\Sigma_P^2 - Z^{-2}_{measured}(f))/(\Sigma_P^2 - Z^{-2}_{model}(f)) \tag{5.7a}$$

$$q_B(f) = (\Sigma_P^2 Z^2_{measured}(f) - 1)/(\Sigma_P^2 Z^2_{model}(f) - 1). \tag{5.7b}$$

We can evaluate the above expressions for rocket data shown in Figure 5.2 by choosing values at the peak in the impedance function, near 0.5 Hz. In this case $\Sigma_P^{-1} = 0.3\ \Omega$, $Z_{measured} = 1.5\ \Omega$, and $Z_{model} = 3.5\ \Omega$. The result is $q_E = 0.96$ and $q_B = 0.18$, i.e. Alfvén waves are responsible for 96% of the measured electric field power and 18% of the magnetic field power at 0.5 Hz. We expect the Alfvén wave magnetic field contribution to

be small since the numbers were taken from the peak in $Z(f)$ which corresponds to a magnetic field node in the standing wave pattern. For the Day 063 HILAT data in Figure 5.8 the relevant numbers near the peak at 0.3 Hz are Σ_P^{-1} = 0.12 Ω, $Z_{measured}$ = 0.18 Ω, Z_{model} = 3.0 Ω, q_E = 0.55 and q_B = 0.002. Thus half of the electric field energy and almost none of the magnetic energy measured by HILAT during this particular pass is attributable to Alfvén waves.

At this point we must issue a note of caution concerning these estimates. The quantity q_E increases very rapidly with $Z_{measured}$, and quickly reaches a value above 90%.Thus any anomalous increases in the measured impedance will cause q_E to fall in the 90% range. For this reason, estimates of q_E may be biased towards large values. Another potential source of error is our assumption of horizontal spatial homogeneity in the ionosphere. However, at first guess it would seem that this effect would smear the impedance function and *decrease* the peak impedance measurement, causing a decrease in the q_E estimate.

5.7 Discussion

The impedance function measured with the Black Brant X sounding rocket flying nearly parallel to auroral structures (i.e. eastward) indicates that the electric and magnetic field fluctuations above 0.1 Hz in the spacecraft frame are due mainly to standing Alfvén waves rather than Doppler-shifted spatial structures. In contrast, spectral energy from static structures plays a more important role in the HILAT measurements, but some Alfvén wave electric field energy is clearly present in most passes at frequencies where the electric field standing wave pattern is predicted to be maximum. A plausible reason

for the fact that HILAT measured less Alfvén wave energy than the sounding rocket is that HILAT's velocity is perpendicular to most of the auroral structure, and most of the Alfvén wave energy seems to be localized at latitudes near auroral arcs. These findings argue for the importance of Alfvén waves in the electromagnetic structure of the disturbed auroral oval, and they lend credence to the idea that at the scale size of auroral arcs, magnetosphere-ionosphere coupling is influenced by Alfvén waves as suggested, for example, by *Hasegawa* [1976], *Goertz and Boswell* [1979], *Haerendel* [1983], *Lysak and Carlson* [1983], and *Seyler* [1988].

While we have concluded that the measured structure in the impedance function is due to standing Alfvén waves, we must also consider a few other explanations. For example, it is possible that the value of Σ_P associated with small scale static electric and magnetic fields is different than the Σ_P relating large scale fields, since small scale structures in the aurora, such as arcs, are associated with density enhancements. As a result, a spacecraft measuring Doppler-shifted static structures might find $Z(f)$ at low frequencies (i.e. large scales) to be larger than at high frequencies (small scales) since the Σ_P relating small scale fields would be higher. This mechanism is not the cause for the structure we observe in $Z(f)$ because 1) it predicts a decrease in Z rather than the observed increase, and 2) it cannot account for the phase spectrum measured by the sounding rocket.

In addition to standing waves, kinetic Alfvén waves (i.e. $k_x c/\omega_{pe} \sim 1$) can also increase the field impedance measured above the ionosphere above Σ_P^{-1}, as we can see from Faraday's Law, which tells us that $k_x E_z - k_z E_x = \omega B_y$. Eliminating E_z with Equation (2.17) and k_z with (2.15) gives

$$\mu_0 \frac{E_x}{B_y} = \mu_0 V_A \sqrt{1 + k_x^2 c^2 / \omega_{pe}^2} \tag{5.8}$$

The correction term $k_x^2 c^2/\omega_{pe}^2$ is small for the HILAT data because spatial scales on the order of c/ω_{pe} transform to several Hz in the satellite frame, but the peaks we measure in $Z(f)$ occur between 0 and 1 Hz. On the other hand, since the sounding rocket was traveling nearly parallel to auroral structures and in the same direction as the plasma flow, it is possible that the Doppler-shifted frequencies corresponding to auroral structures could fall into the tenths of Hz range, and kinetic Alfvén waves could possibly cause higher impedances in the rocket data. Again, (5.8) does not predict the phase spectrum shown in Figure 5.2, while the standing wave model does make such a prediction. We conclude that small horizontal wavelengths are not responsible for the structure in the rocket-measured $Z(f)$.

Another complication can arise if a spatially localized Alfvén wave in a drifting plasma reflects from the ionosphere, but the reflected part of the wave convects away from the incident part. This can happen if $d/l_x > V_A/V_{drift}$, where d is the distance of the measuring platform from the reflection point, l_x is the spatial scale of the Alfvén wave, and V_{drift} is the plasma drift velocity. A spacecraft above the ionosphere would then measure the local Alfvén wave impedance rather than the standing wave impedance. This scenario was discussed by *Mallinckrodt and Carlson* [1978], but is not relevant to the data discussed in this chapter since incident and reflected waves are clearly interfering in our data.

Throughout our analysis we have assumed that the field fluctuations we measure at any given time are either purely spatial or

purely temporal. Of course both types can and most likely do occur simultaneously, and incorporating this fact would complicate our analysis considerably. For example, if we measure a spatially modulated Alfvén wave, Doppler shifting of the spatial structure would cause the measured wave frequency to be "mixed" to different frequencies. Future work should attempt to include complications such as this and the ones listed above, but much more detailed information about the 3-dimensional density structure in the ionosphere will be needed, and therefore measurements from a single satellite or rocket alone will not suffice.

To compare our results here with previous work [*Sugiura et al.*, 1982], we performed a statistical analysis on the time-domain rocket data without filtering and we find that the correlation coefficient between the electric and magnetic fields ρ is 0.70, and the rms field fluctuations are related by $\mu_0 E_{rms} / B_{rms} = 0.33\ \Omega$. This value is in excellent agreement with the value of Σ_P^{-1} deduced from the Langmuir probe measurements. Since most of the spectral energy in the electric and magnetic fields is at low frequencies, this result gives the same impedance as found in the low frequency limit of $Z(f)$. It is interesting to note that at short time scales *Sugiura et al.* [1982] noticed fine structures in the electric field that were not present in the magnetic field. This corresponds to an increase in $Z(f)$ at higher spatial and/or temporal frequencies, as we have found to occur in the presence of standing Alfvén waves.

To reiterate the findings of this chapter we note that even though the HILAT-measured impedance spectra and the modeled spectra presented in Figures 5.6 - 5.16 do not match well in terms of numerical values, there is reasonable agreement between the measured and

modeled frequencies at which $Z(f)$ maximizes. This fact along with the excellent agreement between the sounding rocket data and the predictions of the numerical model allow us to conclude that Alfvén waves are an important part of the overall electrodynamic coupling between the magnetosphere and auroral ionosphere. Furthermore, the quantitative estimates in Section 5.6 indicate that the electromagnetic field energy carried by Alfvén waves exceeds the energy carried by structured Birkeland currents in the vicinity of auroral arcs.

CHAPTER 6

THE EFFECT OF ALFVEN WAVES ON INCOHERENT SCATTER RADAR MEASUREMENTS

6.1 Introduction

This thesis has helped to establish the importance of Alfvén waves in the high-latitude ionosphere, and we might now ask how the presence of Alfvén waves might affect the interpretation of radar measurements. High-latitude radars play a crucial role in the study of magnetosphere-ionosphere coupling, and rocket or satellite data are often analyzed in the context of supporting radar data since the radar measurements in many ways complement spacecraft measurements. Incoherent scatter radars are especially good at measuring density, temperature, and plasma drift velocity, among other parameters, both as a function of altitude and horizontal distance. Unfortunately the resolution of radar measurements is limited spatially by the antenna beam-width and temporally by the fact that many single measurements must be averaged due to the statistical nature of the returned signal. In practice, integration times are at least a few seconds and usually longer, depending on the ionospheric plasma density. Spacecraft, on the other hand, are able to take measurements of the electric field, for example, with a time resolution many orders of magnitude better than this, albeit only at one point in space per measurement.

The signal analysis which extracts plasma parameters from the returned radar signal usually relies on the assumption that all

macroscopic quantities are homogeneous within a volume defined by the radar beam width and pulse length, and that they are time-stationary during the integration period. However, high resolution satellite measurements have shown that spatial homogeneity at antenna beam-width scales is not always a safe assumption at high latitudes. *Swartz et al.* [1988] gave an example of a Millstone Hill radar measurement which suffered from a breakdown in the assumption of spatial homogeneity by using HILAT data to identify 2 km/s drift velocity variations within the 80 km width through which the antenna beam scanned during the 30 s integration period. By simulating the distortion in a theoretical spectrum which would result from the HILAT-measured velocity shears, the authors showed that a parameter fitting program would erroneously predict ion and electron temperatures of 2705 K and 990 K, while the undistorted spectrum would have indicated 1500 K. They suggested that velocity shears are an alternate explanation for distorted incoherent scatter spectra which had previously been attributed to non-Maxwellian plasmas [*Moorcroft and Schlegel*, 1988; *Lockwood et al.*, 1987; and *Løvhaug and Flå*, 1986] or ion hot spots [*Kofman and Lathuillere*, 1987].

There is evidence that time-stationarity during the radar integration period can also be violated at high latitudes. From our analysis in Chapter 5 of rocket and satellite data we have established that substantial time-varying electric fields can exist in the ionosphere with periods less than a typical radar integration period of 5 to 10 s. The event near 09:31 UT in Figure 5.1 shows an electric field pulse with an amplitude exceeding 100 mV/m and a characteristic frequency of about 0.3 Hz. The E×B drift velocity resulting from this electric field is over 2 km/s. Clearly

the assumption of time-stationarity would not be valid if radar measurements were taken in the vicinity of such a pulse.

6.2 Incoherent Scatter Spectra with Time-Varying Drifts

The theoretical spectrum received by an incoherent scatter radar was derived by *Dougherty and Farley* [1960, 1963], *Farley et al.* [1961], and *Farley* [1966]. Data processing in an ISR experiment usually involves measuring the spectrum or autocorrelation function at each range for each inter-pulse period (IPP), then averaging over several IPPs. The result is fit to the theoretical curve in a least squares sense.

The theoretical "ion line" or low Doppler shift portion of an incoherent scatter radar spectrum is shown in Figure 6.1 for 2 different plasma densities. The spectra were generated assuming that the transmitter frequency is 1290 MHz (the operating frequency of the Sondrestrom radar), the plasma has O^+ as its only ion constituent, $T_e = T_i$ = 1500 K, and the antenna is directed along $\mathbf{B}_0$. The spectrum for $n_e = 10^6$ cm^{-3} is typical for cases in which the radar wavelength is much longer than $4\pi\lambda_D$ where λ_D is the Debye length. With this set of parameters λ_D = 2.3 mm. The wavelengths of the Sondrestrom and EISCAT (933 MHz) radars are small enough so that λ_{radar} is comparable to $4\pi\lambda_D$ for lower densities, and the measured spectra lose their double-humped appearance as evidenced by the $n_e = 10^4$ cm^{-3} (λ_D = 2.3 cm) curve in Figure 6.1. The depth of the valley at zero Doppler shift depends not only on the Debye length but also the ratio T_e/T_i and the ion composition.

If the ionospheric plasma has a bulk drift, the entire spectrum is shifted by an amount $\omega_d = \mathbf{k}\cdot\mathbf{V}_d$ where $\mathbf{V}_d$ is the line-of-sight drift velocity and $|\mathbf{k}| = 4\pi/\lambda_{radar}$. If the drift is due to an Alfvén wave, $\mathbf{V}_d$ will change

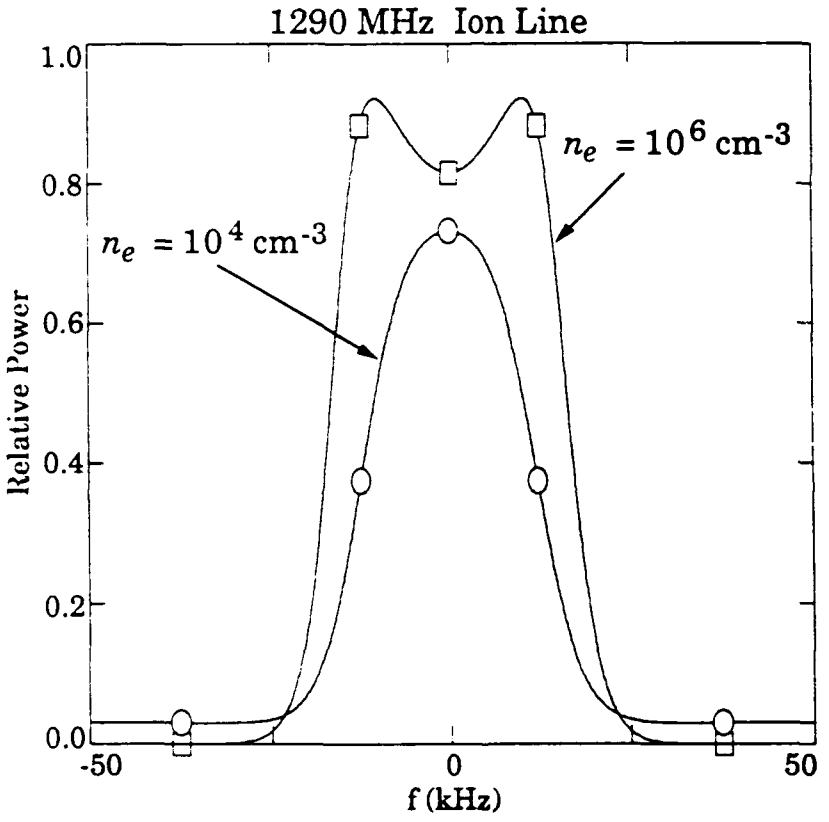


Figure 6.1 Theoretical ion-line spectra at 1290 MHz assuming an O+ plasma with two different densities.

from IPP to IPP and the spectrum $S'(\omega)$ resulting from the averaging process will be

$$S'(\omega) = \frac{1}{N} \sum_{i=1}^{N} S\left(\omega - \mathbf{k}\cdot\mathbf{V}_{d,i}\right) \tag{6.1}$$

where i is an IPP index and N is the number of IPPs included in an integration time. We have assumed that the drift velocity does not change substantially during a single IPP, which is typically on the order of 10 ms in duration.

In order to quantify the amount of distortion introduced by an Alfvén wave in a high latitude ISR experiment we distorted several theoretical spectra in the manner indicated by Equation (6.1) assuming that the drift velocity $\mathbf{V}_{d,i}$ varies sinusoidally with i, i.e.

$$|\mathbf{V}_{d,i}| = V_{d,max}\sin(2\pi i/N) \tag{6.2}$$

where we have used N = 1000. Figure 6.2 shows the effect of this operation on both spectra in Figure 6.1 for $V_{d,max}$ = 750 m/s. The associated electric field amplitude is about 38 mV/m, which is certainly possible at high latitudes. The $n_e = 10^6$ cm^{-3} spectrum appears to be most affected, with its double-humped structure nearly obliterated. Since the peak-to-valley ratio has changed drastically, one would expect a least-squares fitting program to underestimate the temperature ratio T_e/T_i. Distorting the spectrum in the smaller n_e case has widened it, but the fact that there was no pronounced valley in the spectrum to begin with suggests that the T_e/T_i estimate from a fitting program will not suffer from errors as large as those in the high density, small Debye length case.

Figures 6.3a and b show temperature estimates from a least-squares fitting program versus $V_{d,\,max}$ for the two spectra in Figure 6.1. The

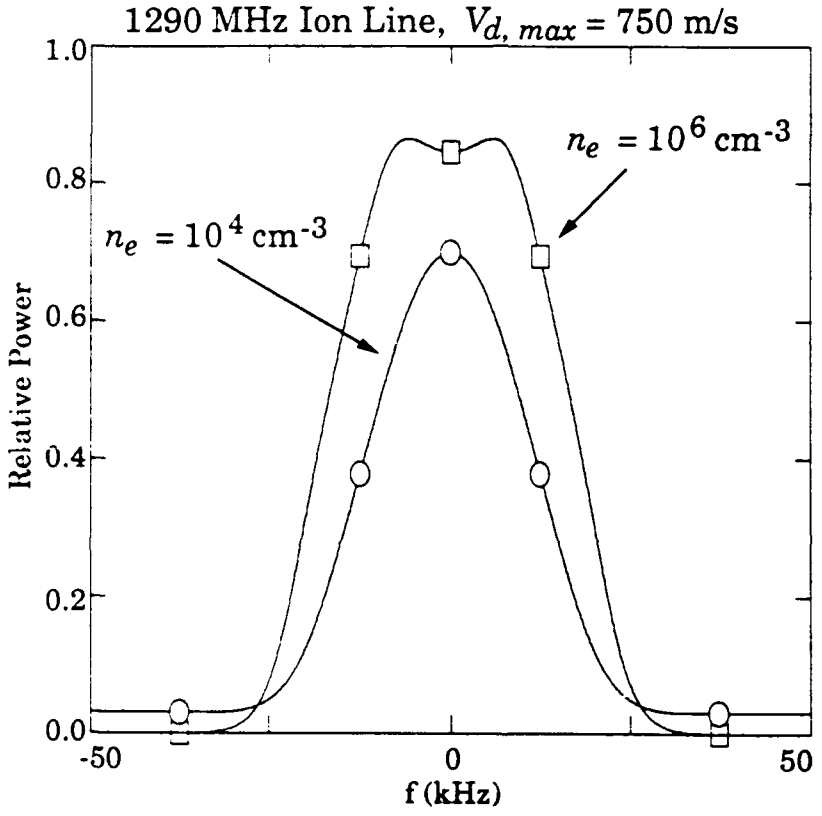


Figure 6.2 Spectra which would result from smearing the spectra in Figure 6.1 with a 750 m/s amplitude sinusoidal drift velocity which has a period less than the radar integration time.

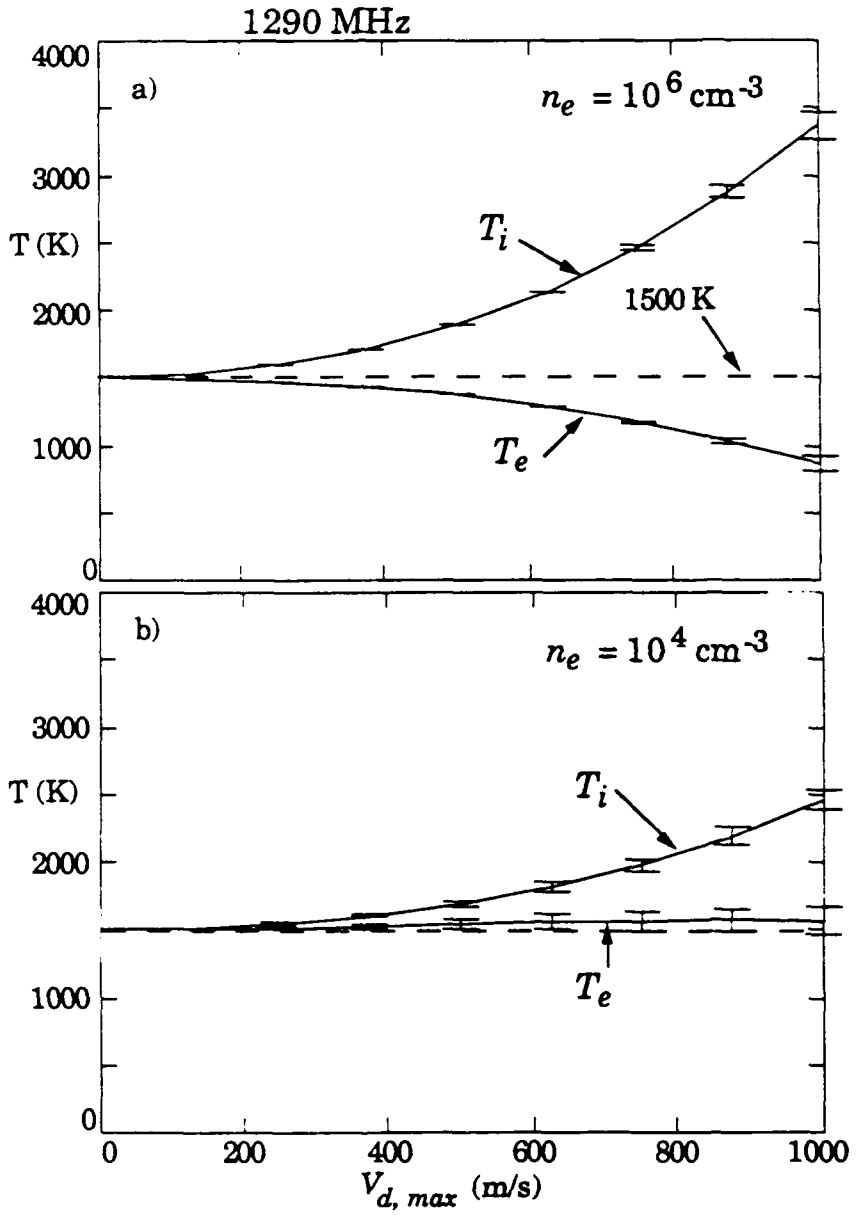


Figure 6.3 Electron an ion temperature fits to ideal 1290 MHz ion-line spectra which have been smeared with a drift velocity of the form $V_d = V_{d,max}\sin(2\pi t/T)$, where T is less than the radar integration time.

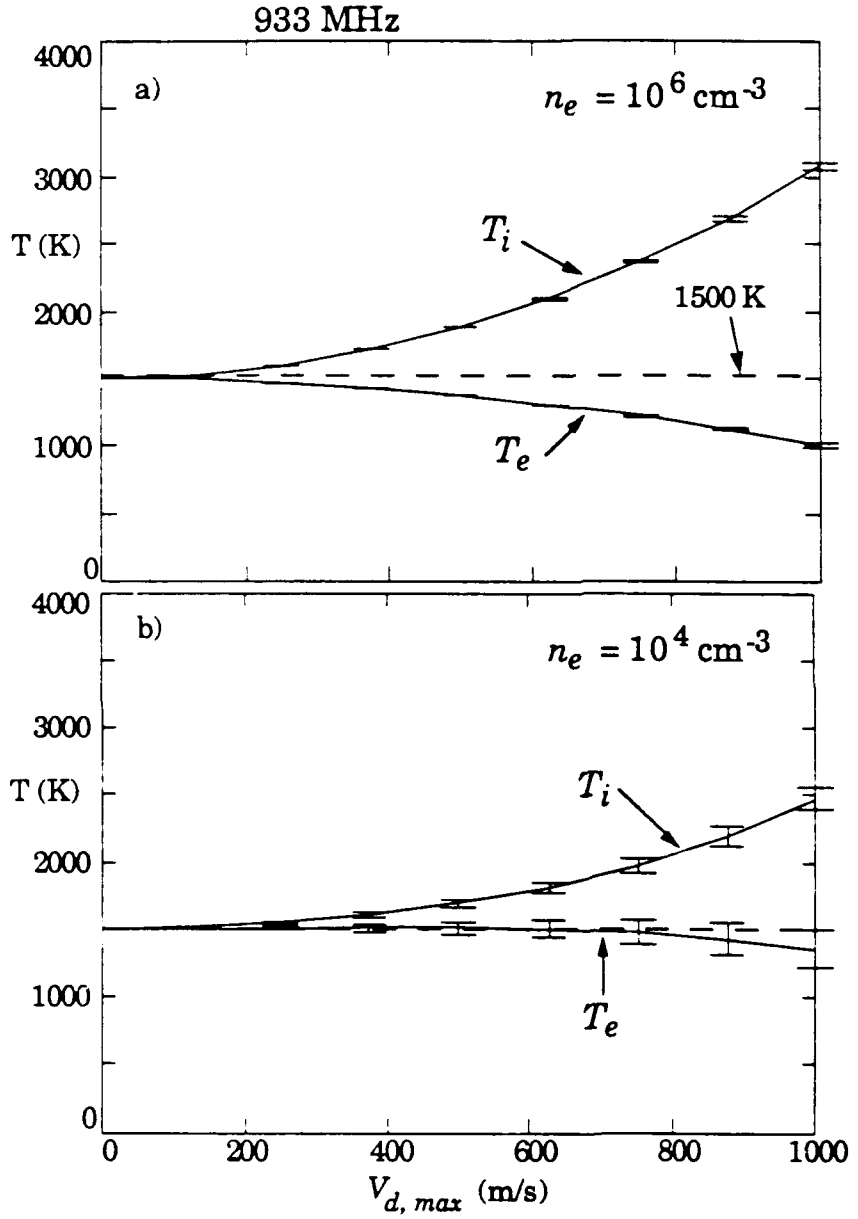


Figure 6.4 Electron an ion temperatures fits to ideal 933 MHz ion-line spectra which have been smeared with a drift velocity of the form $V_d = V_{d,\,max}\sin(2\pi t/T)$, where T is less than the radar integration time.

codes for generating theoretical spectra and performing the fits were supplied by *P. Erickson and J. Pingree* [personal communication, 1990]. Figures 6.4a and b show curves generated using parameters identical to those in Figure 6.3 except for the transmitter frequency, which has been changed to the EISCAT UHF frequency of 933 MHz.

According to Figure 6.3a, velocity fluctuations with amplitudes of 250 m/s, which are quite common in the auroral zone, can generate an anomalous increase in T_i of almost 100 K and an associated decrease in T_e of about 30 K. In the lower density case (Figure 6.3b) the temperature estimate errors are less than half of these values. Unfortunately, although the errors are smaller, lower density plasmas require longer radar integration times and thus it is more likely that time-stationarity will be violated. A 1 km/s sinusoidal drift can cause an erroneous T_i of nearly 3500 K and T_e of about 850 K in the $n_e = 10^6$ cm^{-3} case. But while fluctuations of this magnitude certainly occur in the auroral ionosphere, the event shown in the Black Brant data in Chapter 5 lasts only a few seconds. In this case a longer integration time might serve to reduce the error in the T_i estimate from what it would be if the Alfvén wave lasted the entire integration period.

At 933 MHz with $n_e = 10^6$ cm^{-3} (Figure 6.4) the temperature estimate errors are slightly less than in the 1290 MHz case. For $n_e = 10^4$ cm^{-3}, errors in the estimated T_i are about the same at both transmitter frequencies, while T_e estimates are in greater error at 933 MHz, although only by 150 K with a 1 km/s drift.

6.3 Conclusions

Swartz et al. [1988] have shown that spatial velocity shears at middle and high latitudes can violate the assumption of spatial homogeneity with ISR antenna beams, and can cause anomalously large ion temperature estimates. This phenomenon can mimic the effect of ion hot spots or non-Maxwellian plasma distributions.

We have shown that similar errors in temperature fits can occur in the presence of Alfvén waves. In the case where $\lambda_{radar} >> 4\pi\lambda_D$ the increased ion temperature estimates are accompanied by a marked decreases in T_e estimates. If observed over a period of several minutes, this signature might help to distinguish anomalously large ion temperatures due to Alfvén waves from actual occurrences of ion hot spots since the latter would not likely occur simultaneously with decreases in the electron temperature.

Ground-based magnetometers might also be used to identify ISR spectra distorted by Alfvén waves, although these instruments tend to average over a vast portion of the sky. An event like the one shown in Chapter 5 (Figure 5.1) near 09:31:30 UT might have a magnetic signature on the ground known as a "giant pulsation", and simultaneous measurement of such a magnetic pulsation with distorted ISR spectra would be a useful demonstration of the effect of Alfvén waves on radar measurements.

Optical data from image intensified TV images might prove more useful in identifying causes of distorted spectra since rapid motion of auroral forms would be a good indication of Alfvén wave activity. The amount of spectral distortion could be compared inside and outside of such regions, for example.

CHAPTER 7
CONCLUSIONS AND SUGGESTIONS FOR FUTURE RESEARCH

7.1 Summary of Results

The electric and magnetic fields which couple the magnetosphere and ionosphere carry a wealth of information regarding the amount and direction of energy flow, the characteristics of energy dissipation in the ionosphere, the temporal and spatial structure of magnetospheric energy sources, and the presence of neutral winds in the ionosphere. Since there are many physical processes which create and modify them, *in-situ* measurements of these fields can be difficult to interpret. In this thesis we have taken existing analysis techniques, e.g. spectral and cross-spectral analysis, and applied them in new ways to auroral electric and magnetic field data. We will now summarize the main results presented in this dissertation. Since Chapter 5 contains the bulk of the results, we begin there.

In-situ measurements of magnetosphere-ionosphere coupling via Alfvén waves. Chapter 5 contains data from two different experiments: 1) a sounding rocket launch into the dayside auroral oval, and 2) several oval crossings by the HILAT satellite. The results of these two experiments are similar in that at frequencies below about 0.1 Hz (as measured in the spacecraft frame) the meridional electric to zonal magnetic field ratio $Z(f) = \mu_0 E_x(f)/\delta B_y(f)$ is equal or nearly equal to the inverse of the height-integrated Pedersen conductivity of the ionosphere

Σ_P^{-1} , and the cross-product of the fields indicates a downward Poynting vector. These measurements are consistent with energy flow directed from the magnetosphere toward the ionosphere in the form of quasi-static electric fields and field-aligned currents. This energy is dissipated in the conducting part of the ionosphere. The measurements in Chapter 5 are also consistent with earlier findings from *Sugiura et al.* [1982], *Sugiura* [1984], and *Smiddy et al.* [1984].

For time scales shorter than 10 s the results from the rocket and satellite experiments differ. The rocket-measured $Z(f)$ increases smoothly to a peak value of about $4\Sigma_P^{-1}$ at 0.3 Hz and above. This value is near the Alfvén impedance $\mu_0 V_A$, which indicates that the fields are due to Alfvén waves and the fluctuations are therefore temporal rather than spatial structures which have been Doppler- shifted from the rocket motion. Furthermore, the shape of $Z(f)$ and the phase relation between $E_\perp$ and $\delta B_\perp$ indicate a standing Alfvén wave pattern due to reflections from the ionosphere.

The satellite data consist of six passes for which the frequency-averaged coherency spectra were significantly larger than could be expected from random noise. The measured impedance spectra fall near Σ_P^{-1} over most of the spectrum and do not vary as much as in the sounding rocket case. However, in most cases the frequency of the maximum measured impedance falls within 10-15% of the frequency predicted from the standing Alfvén wave model. A plausible explanation for this is that most of the field energy measured by HILAT is from structured quasi-static fields, but occasionally there are Alfvén waves present. Alfvén waves can be identified in the impedance spectra by their increased impedance, and we find that the frequency of the resulting

electric field enhancement is determined by the electrical length of the satellite above the reflecting part of the ionosphere. That is, HILAT measures increased impedances at frequencies for which its orbital altitude is at the peak of the electric field standing wave pattern. If the frequencies of detectable Alfvén waves were determined by the wave source, one would expect increased impedances at frequencies which bear no relation to those predicted by the standing wave model. This is apparently not the case. Unfortunately the variances of the satellite-measured phase spectra were too large to predict the phase shift between the meridional electric and zonal magnetic fields.

We found two pieces of evidence which suggest that the measured Alfvén wave energy is concentrated near auroral arcs. The first is from the rocket data in Figures 5.1c and d, which shows that increases in electric field energy in the 0.25-0.35 Hz range are correlated with enhancements in precipitating electron energy flux. The second is indirect, and follows from the observation that based on the measured impedance spectra a much larger fraction of the electromagnetic field energy measured from the sounding rocket is due to Alfvén waves than in any of the 6 HILAT passes. Since the sounding rocket velocity was eastward, it is likely that it spent much more time in the vicinity of auroral arcs than HILAT, which moves mostly perpendicular to auroral structures. Consequently, HILAT spends less time near individual arcs and this may explain the diminished evidence for Alfvén waves in HILAT data.

Again, the major points from Chapter 5 are:

- The frequency-dependent relations between amplitudes and phases of the meridional electric and zonal perturbation magnetic field

data from the Greenland II Black Brant sounding rocket are in excellent agreement with a standing Alfvén wave model.

- The values of the impedance function measured by HILAT are much lower than those predicted by the numerical Alfvén waves model, indicating that much of the spectral energy is dominated by quasi-static structures Doppler-shifted by the spacecraft velocity.

- HILAT does detect some Alfvén wave energy at frequencies for which the satellite is at a peak in the electric field standing wave pattern, indicating that the shape of the Alfvén wave frequency spectrum near the ionosphere is determined by wave interference, not the magnetospheric wave source.

- Alfvén waves appear to be localized in latitude near auroral arcs.

It is not clear from this study if the spatial coincidence of Alfvén waves and arcs is due to a causal link between the two, as suggested by previous authors (see Chapter 5). Hopefully, future studies can help to establish such a link by adding to the amount of low-frequency data taken by spacecraft traveling parallel to the auroral oval. These studies will have to be carried out either with sounding rockets or satellites in a somewhat lower inclination orbit than HILAT (81°), closer to 70°. Successful experiments in the future will need to include many different instruments, as in the upcoming Auroral Turbulence campaign. High time resolution ground-based auroral imaging will be especially helpful in determining the amount of spatial structuring and dynamic activity characterizing the environment in which *in-situ* measurements are taken.

Numerical model of Alfvén wave reflections in the ionosphere. A model similar to the one presented in Chapter 4 is a very useful tool in interpreting the data. Future models can be improved by relaxing the assumption of spatial homogeneity in horizontal directions, although this will greatly complicate the code. However, we have been able to explain several features in the data without taking into account auroral density structuring. Although we have used the numerical model mainly as an aid in interpreting experimental data, in Chapter 4 we showed that it is also a useful tool for understanding the general reflection and absorption properties of the ionosphere. A summary of these properties is as follows:

- The meridional electric field reflection coefficient $|\Gamma|$ for Alfvén waves with periods greater than 10 s is close to $(\Sigma_P^{-1} - \mu_0 V_A)/(\Sigma_P^{-1} + \mu_0 V_A)$ where the Alfvén velocity is taken above but close to the ionosphere. This can be interpreted as meaning that the ionosphere behaves as a thin conducting slab on these time scales.

- For time scales shorter than 10 s, $|\Gamma|$ decreases. In model ionospheres with an F region, $|\Gamma|$ experiences sharp nulls separated by a few tenths of Hz. These nulls correspond to resonances which increase the electric field amplitude (and thus the Joule heating) above the E region. Thus as a general rule, it appears that more E-region ionization increases $|\Gamma|$, and F-region ionization tends to decrease $|\Gamma|$.

- Electron collisions have little effect on the reflection coefficient, nor do the density and ionization scale heights in the lower atmosphere. Therefore the ion collision and plasma density profiles above 100 km are mainly responsible for the behavior of $|\Gamma(f)|$.

- The horizontal spatial scale of the Alfvén wave has a negligible effect on $|\Gamma|$ for $\lambda_x > 10$ km, and $|\Gamma|$ decreases somewhat as λ_x approaches 1 km.

Energy flow into and out of the upper atmosphere. Many parameters of the solar-terrestrial system (e. g. solar sunspot number and geomagnetic activity indices) are continuously monitored from ground-based and orbiting instruments. The total low-frequency electromagnetic energy flux into the polar cap and auroral oval is not currently monitored in this way, yet it is potentially an important factor in characterizing the energetics of the upper atmosphere.

In Chapter 3 we compare two quantities which are useful for measuring EM energy input into the ionosphere, Joule dissipation and Poynting flux. They have been used by previous authors who have made case studies of individual events, but they are not routinely monitored. The first quantity relies on electric field measurements and assumed or derived ionospheric density and collision frequency profiles. Satellites which can measure both electric and perturbation magnetic fields can determine the Poynting vector and thereby circumvent the potential errors in Joule dissipation measurements arising from incorrect ionospheric models. Another advantage of using Poynting flux is that it is that it is a signed quantity, thus upward Poynting flux can be used to indicate areas in which the neutral wind is acting as an electrical dynamo and supplying energy to the magnetosphere.

Effects of Alfvén waves on incoherent scatter radar spectra. Finally, in Chapter 6 we have argued that Alfvén waves can occur in the auroral oval with amplitudes and frequencies sufficient to severely distort

incoherent scatter radar spectra. The plasma drifts from a 20 mV/m wave with a period much less than the radar integration period tend to cause significant increases in apparent ion temperature as estimated from least-squares fitting programs. When the plasma Debye length is much less than the radar wavelength, the programs erroneously predict decreased electron temperatures as well.

7.2 Future Research: Quasi-Static Fields and Neutral Winds

In Chapter 3 we established that in many ways, satellite measurements of the DC Poynting vector $\mathbf{E}\times\mathbf{H}$ are superior to Joule heating estimates. Hopefully, satellite Poynting flux measurements will be part of future synoptic studies of the high-latitude ionosphere.

An especially interesting application of Poynting flux measurements is in the area of ionosphere-thermosphere interactions. In Section 3.3 we calculated electric fields and currents generated by neutral winds in the ionosphere for two different electrical loads, and we showed that neutral wind dynamos give rise to an upward Poynting vector above the ionosphere. While detections of upward Poynting flux by satellite can reveal much about the ionosphere and winds below, altitude profiles of $E_\perp(z)$ and $\delta B_\perp(z)$ (which must be measured with sounding rockets instead of satellites) would be of more use in a detailed study of wind-driven dynamos. A reason that altitude profiles are necessary is that, contrary to our assumption in the Section 3.3 examples, thermospheric neutral winds can vary in altitude as a result of tides and gravity waves. The height variation of the winds causes associated changes in perturbation magnetic fields, as we shall show later in this section.

At middle and low latitudes, global scale electric fields and currents in the ionosphere are controlled mainly by tidal modes in the thermosphere, especially during the daytime, as discussed by *Richmond et al.* [1976] and *Richmond and Roble* [1987]. The thermosphere-ionosphere interaction is important at smaller scales as well, where neutral atmosphere dynamics are driven by gravity waves. For example, *Röttger* [1973] and *Kelley et al.* [1981] showed that gravity waves can cause structuring of equatorial spread-F irregularities.

During magnetically quiet periods the neutral wind is an important source of electric fields and currents at high latitudes as well, as has been measured with the Chatanika radar by *Brekke et al.* [1974]. Even during magnetically active times the neutral wind can be important. As we discussed in Chapter 3, the effective conductivity of the ionosphere is modified by neutral winds, so even if the neutral winds are not driving dynamo fields, the load characteristics of the ionosphere are affected by winds. Another neutral wind effect was considered by *Forbes and Harel* [1989], who showed that a magnetospheric disturbance can accelerate the neutral wind in such a way that the net magnetic perturbation decreases after some time even though the driving electric field remains constant.

Vertical variation of wind velocities on scales of tens to hundreds of km in the thermosphere can occur as a result of upwardly propagating gravity waves and tides, and the correlation of these winds with ionospheric electric fields has been measured using chemical tracers released by sounding rockets [*Mikkelsen et al.*, 1981, 1987] and modeled numerically by *Pereira* [1979], among others. *Earle and Kelley* [1988] compared Chatanika-measured electric fields with mesospheric gravity waves and found similar spectral characteristics, suggesting that during

quiet times, thermospheric winds and electric fields are strongly coupled.

While gravity waves can create dynamo electric fields in the thermosphere, it is difficult to show experimentally that a particular spacecraft or radar electric field measurement is due to gravity waves. For example, one might try to show that the electric field **E** and neutral wind **U** vary together in time, but the minimum wave period for gravity waves is on the order of five minutes. Of course, spacecraft are unable to make measurements at a single point in space for this long. Ground-based radars can be used for the electric field measurement, but the problem of measuring the neutral wind above 100 km for tens of minutes remains.

A second way one might study the gravity wave-electric field interaction is to correlate the variations of **U** and **E** as a function of altitude. But, as we showed in Chapter 4, electric fields with time scales of over 10 s map along geomagnetic field lines, so **U** may vary but **E** will not. However, since the horizontal current in the ionosphere can be driven by both electric fields and winds, it so happens that $\delta\mathbf{B}_\perp$ *does* vary with altitude in the presence of gravity waves. We suggest that simultaneous rocket measurements of $\delta\mathbf{B}_\perp$ and **U** might be a useful way to study the interaction between gravity waves and the ionosphere.

To obtain some idea of the magnetic field magnitudes one can expect from gravity waves, we will now calculate some wind-driven magnetic field profiles for the simplified case in which $\mathbf{B}_0$ is vertical, $\partial/\partial y = 0$ (no variation in the zonal direction), and **U**, **E**, and $\delta\mathbf{B}$ vary as $\exp(ik_x)$. Furthermore, we will allow only a zonal wind U_y and a meridional electric field E_x.

In general, the current **J** is given by

$$\mathbf{J} = \sigma \cdot (\mathbf{E} + \mathbf{U} \times \mathbf{B}_0) \tag{7.1}$$

where σ is given by Equation (3.3). As we discussed in Chapter 3, (7.1) can be found by transforming $\mathbf{J}' = \sigma \cdot \mathbf{E}'$ from the neutral wind frame into the Earth-fixed frame. The $\hat{x}$ and $\hat{y}$ components of Ampere's law, $\nabla \times \mathbf{B} = \mu_0 \mathbf{J}$, can now be written:

$$\hat{x}: \frac{\partial B_y}{\partial z} = - \mu_0 \sigma_P (E_x + U_y B_0) \tag{7.2a}$$

$$\hat{y}: \frac{\partial B_x}{\partial z} - \frac{\partial B_z}{\partial x} = - \mu_0 \sigma_H (E_x + U_y B_0) \tag{7.2b}$$

Thus, given $E_x(z)$ (which is constant), $U_y(z)$, and a boundary condition for B_y, we can integrate (7.2a) to find $B_y(z)$. We can find both E_x and a boundary value for B_y above the ionosphere from current continuity and the fact that $J_x = \sigma_P (E_x + U_y B_0)$:

$$J_z = ik_x \int_{ionosphere} \sigma_P (E_x + U_y B_0)\, dz \tag{7.3}$$

We can eliminate J_z with Ampere's Law, $\mu_0 J_z = - ik_x B_y$. Since E_x is constant in altitude, (7.3) can be written

$$- B_y = \mu_0 \Sigma_P E_x + \mu_0 \int_{ionosphere} \sigma_P U_y B_0\, dz \tag{7.4}$$

Notice that we have divided all quantities by k_x. This is of course only valid for $k_x \neq 0$, and the physical reason for this is that we must have at least some variation in x to have electric and magnetic fields above the ionosphere. If there is no structure in the $\hat{x}$ direction, the neutral wind would still drive currents but there would be no divergence of currents, no charge buildup, and consequently no electric fields.

To completely solve for E_x and B_y above the ionosphere we need an additional relation between them, but this is dependent on the "load" which is receiving energy from the wind dynamo. In Chapter 3 we used the conjugate ionosphere with a neutral wind as a load. This assumes that the two ionospheres have been electrically connected for a long time compared to the time it takes an Alfvén wave to propagate between hemispheres so that a steady state has been reached, and the example is probably more useful as an illustrative tool than as a geophysical model.

A more appropriate load model in the auroral oval and especially in the polar cap is simply an outward-traveling Alfvén wave which does not reflect and never returns to the ionosphere. In the oval, field lines may be closed but they are very elongated, and an Alfvén wave would likely convect away from its region of origin even if it did reflect from the conjugate ionosphere. In the polar cap with southward IMF the field lines are open, so unless a neutral wind-driven Alfvén wave reflects from some magnetospheric turbulence or boundary, the neutral wind sees only an Alfvén wave load. This means that we relate the fields at the top of the ionosphere with the Alfvén impedance, $\mu_0 E_x/B_y = \mu_0 V_A = Z_A$. From (7.4), the neutral wind-driven electric field in and above the ionosphere is then

$$E_x = -\left(\frac{\mu_0 V_A}{1 + \mu_0 V_A \Sigma_P}\right)\int_{ionosphere} \sigma_P U_y B_0 \, dz \tag{7.5}$$

In most cases $\mu_0 V_A \Sigma_P > 1$. For a neutral wind which is constant in altitude we can put $U_y B_0$ outside of the integral in (7.4), and the resulting electric field will be slightly less than $U_y B_0$. This can be understood as follows. Currents in the $\hat{x}$ direction are driven by U_y, and charges build up where this is a divergence of current, creating electric

fields which oppose the current. If there were no load, the electric field would drive a current exactly opposing the wind-driven current and we would find the electric field $E_x = - U_y B_0$. Alfvén waves act as a high (but not infinite) impedance load and carry away some of the charge, making the electric field magnitude $|E_x| < |- U_y B_0|$. For any high impedance load, the electric field will give a fairly accurate measure of U_y in the ionosphere, although it is difficult to assure in any given high-latitude measurement that no electric fields applied in the magnetosphere are present. This should be less of a problem in the dayside mid-latitude zone and a study of such regions might be very interesting. Fields with an outward Poynting flux and which are related by the Alfvén impedance might be a useful indication of fields produced solely by neutral winds

We are now ready to investigate the altitude dependence of the zonal magnetic field B_y by integrating Equation (7.2a). We will not plot the electric field since it is constant in altitude, but we will note the value of E_x in each of the figures. Figure 7.1a shows $cB_y(z)$ for $U_y B_0(z) = 1$ with the "EF" model density profile (Figure 4.2a). There is not much difference between this and the field profile due to magnetospheric forcing shown in Figure 4.7c. The relation between E_x and B_y is quite different in the two cases, however, since in the wind-driven case the Poynting vector is away from the ionosphere and the field impedance is Z_A instead of Σ_P^{-1} .

Altitude-dependent winds complicate $B_y(z)$. Thermospheric winds with amplitudes of 100-200 m/s and wind shears with vertical wavelengths of tens to hundreds of km can be caused by the vertical propagation of gravity waves and tides [*Mikkelsen et al.*, 1987]. Figures 7.1b-d show $cB_y(z)$ for altitude-dependent winds of the form $U_y(z) =$

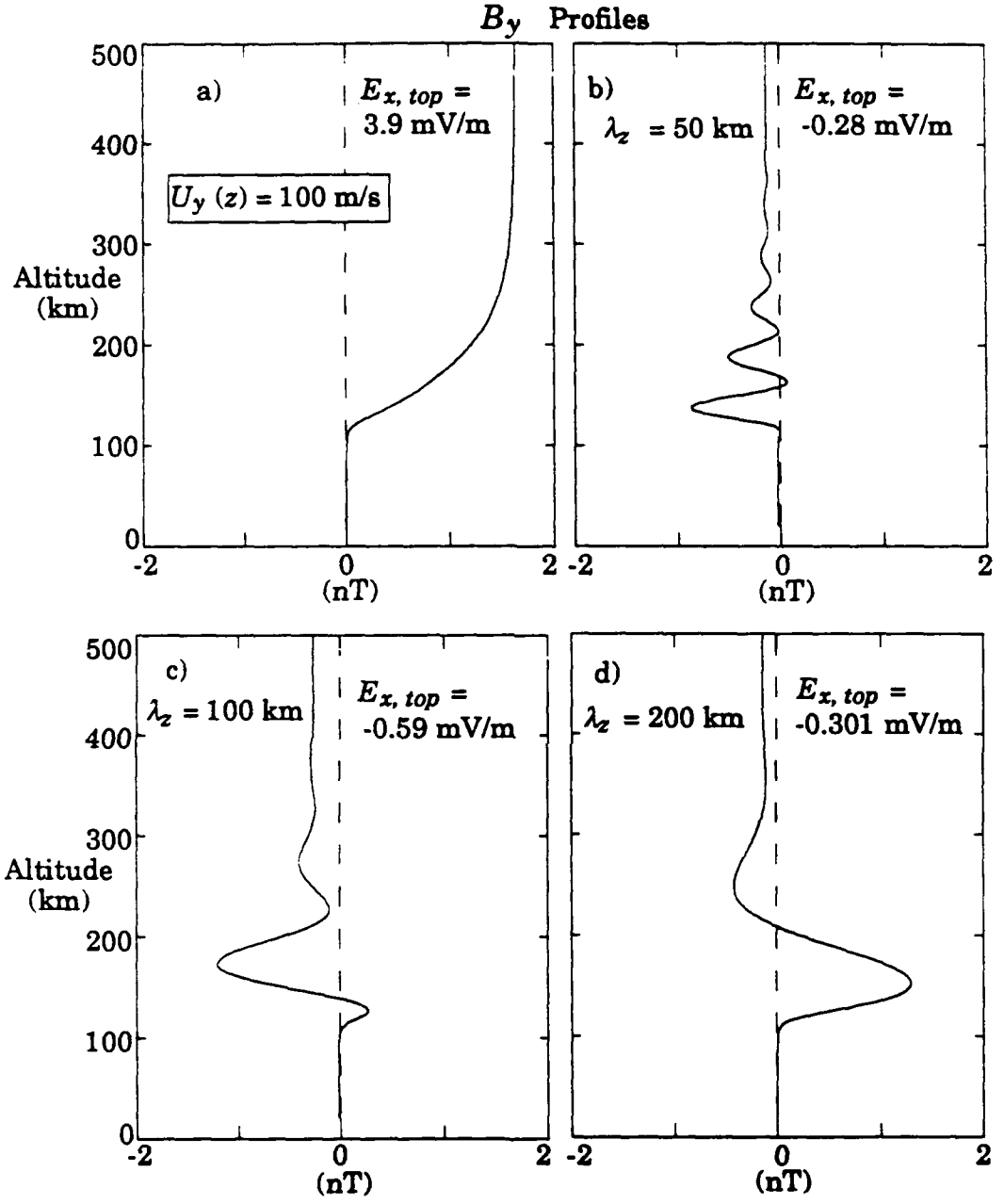


Figure 7.1 a) Zonal magnetic field perturbation due to a 100 m/s neutral wind which is constant in altitude. b-d) Magnetic perturbations due to zonal neutral winds of the form $U_y(z) = 100\cos(2\pi(z - 100\text{ km})/\lambda_z)$. All four profiles were calculated using Profile "EF" shown in Figure 4.2, and the upper boundary condition demands that $E_x/\delta B_y = V_A$.

U_{max} cos(2π(z - 100 km)/λ_z) with U_{max} = 100 m/s and λ_z = 50, 100, and 200 km.

At wavelengths for which the integral in Equation (7.5) is zero, E_x and B_y above the ionosphere vanish but there can still be horizontal currents and perturbation magnetic fields in the ionosphere. This suggests a way in which sounding rockets might identify gravity wave-driven magnetic fields: the zonal magnetic field due to gravity waves can be stronger in the E region than above, whereas static B-fields driven by the magnetosphere increase monotonically with height. This distinction is true only in the DC limit, thus such an experiment would have to be carried out during low magnetic activity. Although neutral winds at ionospheric heights are difficult to measure, an experiment which correlates the neutral wind altitude profile with $B_y(z)$ would be very useful in demonstrating the existence of a gravity wave-driven dynamo. From (7.5) we see that in the northern hemisphere one would look for magnetic perturbations in the same direction as Uy (since $B_0 < 0$) while in the southern hemisphere the two quantities would have opposite signs.

Unfortunately, the magnetic fields generated by gravity waves shown in Figure 7.1 are quite small, i.e. 1-2 nT. Winds on the order of 300 m/s could create magnetic fields of about 5 nT, but the measurement would still be difficult to make. One would probably have to perform the experiment at sub-auroral latitudes to minimize magnetospheric sources of electric and magnetic fields.

7.3 Future Research: Spacecraft Measurements of Alfvén Waves

The comparison of spacecraft measurements and numerical model predictions presented in Chapter 5 has proven to be a fruitful method for

studying Alfvén waves. Continued studies along these same lines may help to reveal the nature of the relationship we observed between Alfvén waves and auroral arcs. A two dimensional (i.e. vertical and meridional) model of the Alfvén wave-ionosphere interaction might be necessary to understand the latitudinal dependence of Alfvén wave occurrences.

Although we have searched roughly 25 satellite passes for evidence of Alfvén waves, we have only a few cases for which there is strong evidence for waves. Satellites with more sensitive instruments and lower inclination orbits than HILAT can possibly help to increase the number of observations of Alfvén wave associated with auroral arcs. *M. C. Kelley* [personal communication, 1990] has suggested that a statistical study of field fluctuations could help to quantify the relative importance of Alfvén waves in the auroral ionosphere. This study would be carried out with data from an extended satellite mission by calculating $E_\perp$ and $\delta B_\perp$ fluctuation amplitudes within a few wide frequency intervals between 0 and 1 Hz. Fluctuation amplitudes exceeding some minimum value (to ensure the presence of geophysical signals) would contribute to an overall average, and the resulting electric and magnetic field averages at each frequency would be divided to form impedance estimates. The amount of any increase in field impedances with increasing frequency could be used to make a numerical estimate, as described in Section 5.6, of the relative importance of Alfvén waves and quasi-static fields in auroral electrodynamics.

7.4 Future Research: Incoherent Scatter Radar Measurements of the Aurora

Spatially sheared and time-varying plasma drifts can hinder attempts to measure ionospheric plasma temperatures with incoherent scatter radars, as discussed in Chapter 6. But the same fluctuations responsible for the errors are worthy of study in themselves. A joint radar-optical experiment is presently being planned for the EISCAT radar which will seek to identify ISR spectra distorted by spatial and temporal electric field variations.

The EISCAT radar is a tri-static system with a transmitter in Trömso, Norway and receivers in Norway, Sweden, and Finland. The antenna beam widths for the UHF system are all 0.6°, thus the width of the Trömso beam at, say 100, 200 and 300 km above Kiruna, Sweden is 2.3, 3.0, and 3.3 km respectively. (The distance between Trömso and Kiruna is roughly 200 km.) The beam width of the Kiruna receiving antenna in the same regions is 1.1, 1.2, and 3.1 km. Thus velocity shears in the common volume of the two antenna beams are less likely to affect the received spectrum at Kiruna than at Trömso, especially at lower altitudes. An enhanced ion temperature or non-Maxwellian velocity distribution, on the other hand, would affect spectra at both receivers equally. Comparing the spectra from both locations is a good way to determine the relative importance of shears, ion hot spots, and non-Maxwellian plasmas.

The radar measurements will be taken with a high time resolution (tens of ms per frame) all-sky TV camera situated below the common volume in Kiruna. The optical data will provide valuable information concerning the spatial and temporal structure of electric fields in the

radar scattering region. Although the camera cannot measure electric fields directly, it can record the optical signature of auroral arcs, which are known to be associated with large velocity shears. Thus one would expect the velocity shears in a stable, quiescent arc within the common volume to broaden the backscattered spectrum in Trömso and have a smaller effect on the Kiruna measurement, due to the smaller receiving antenna beam width. When interpreting distorted spectra measured at both receiver sites and in the absence of auroral arcs, one could probably rule out spectral contamination from velocity shears.

The all-sky TV camera can also be useful for identifying conditions conducive to Alfvén waves. In at least one example, namely the Black Brant rocket flight we analyzed in Chapter 5, the very presence of auroral precipitation was an indication of Alfvén waves. At this point we do not know if all arcs have associated Alfvén waves, but optical evidence of fast time variations such as perturbations propagating along arcs or pulsating auroras would most likely be a telltale sign of Alfvén wave electric fields. As we suggested in Chapter 6, ground-based magnetometer data might also be used to verify the presence of temporal fluctuations. If, based on measurements from several instruments, one is fairly confident that Alfvén waves and spatial velocity shears are not present in the radar scattering volume, the ion hot spot or non-Maxwellian interpretation of distorted ISR spectra can be applied with some confidence.

An important part of the experiment we have outlined here is the fact that many instruments will be used simultaneously. There will be simultaneous data from the radar, an optical camera, a ground-based magnetometer, and possibly a satellite, if there happens to be a coincident

pass. Multiple diagnostics are necessary because although the visible part of the aurora lies in a relatively confined region, i.e. in the auroral oval between 100 and 1000 km in altitude, the keV electron energy source is thousands of km above the ionosphere, and the source of plasma is probably much farther away still. The structure of the visible aurora is thought to be imposed in the acceleration region, thus optical images provide information from a part of the auroral system which is quite removed from the E- and F-region radar measurements.

Unfortunately, the regions of the auroral system which lie beyond the acceleration zone are accessible only to satellites, and it is next to impossible to coordinate measurements in those regions with ionospheric measurements of the aurora, in large part because of the fact that one cannot know exactly how geomagnetic field lines map from the ionosphere to the magnetotail. Numerical simulations can help to piece together an understanding of the different parts of the auroral system, but of course simulations require accurate information concerning boundary conditions, and this information must be supplied with experimental data.

APPENDIX A

POYNTING'S THEOREM

A formal derivation of Poynting's theorem begins with consideration of the total magnetic energy in some volume,

$$\varepsilon_B = \left(\frac{1}{2\mu_0}\right)\iiint B^2 dV \tag{A1}$$

The time rate of change of this quantity can be written

$$\frac{\partial \varepsilon_B}{\partial t} = \left(\frac{1}{\mu_0}\right)\iiint \mathbf{B}\cdot\frac{\partial \mathbf{B}}{\partial t} dV \tag{A2}$$

Using $\partial \mathbf{B}/\partial t = -\nabla\times\mathbf{E}$ and the vector identity $\nabla\cdot(\mathbf{E}\times\mathbf{B}) = \mathbf{B}\cdot\nabla\times\mathbf{E} - \mathbf{E}\cdot\nabla\times\mathbf{B}$ we have

$$\frac{\partial \varepsilon_B}{\partial t} = -\frac{1}{\mu_0}\iiint \nabla\cdot(\mathbf{E}\times\mathbf{B})\, dV - \frac{1}{\mu_0}\iiint \mathbf{E}\cdot(\nabla\times\mathbf{B})\, dV \tag{A3}$$

If we consider the static case $\partial\varepsilon_B/\partial t = 0$ and furthermore, that $\nabla\times\mathbf{B} = \mu_0\mathbf{J}$, we can write

$$\frac{1}{\mu_0}\iiint \nabla\cdot(\mathbf{E}\times\mathbf{B})\, dV = -\iiint \mathbf{E}\cdot\mathbf{J}\, dV \tag{A4}$$

Finally, from Gauss' Theorem

$$\iint \mathbf{P}\cdot d\hat{s} = \iiint \mathbf{E}\cdot\mathbf{J}\, dV \tag{A5}$$

where $\mathbf{P} = (\mathbf{E}\times\mathbf{B})/\mu_0$ and the vector $d\hat{s}$ is pointed into the volume everywhere.

A classic example of this result is that of a long thin wire of resistance R carrying a current I across a voltage V. Since the magnetic field in this case is given by $B = \mu_0 I/2\pi a$ and $E = V/L$ where a is the wire radius and L is its length, the total energy flux into the wire is the surface integral of $\mathbf{P}$,

$$W = \iint \mathbf{P} \cdot d\hat{s} = \frac{1}{\mu_0}\left(\frac{V}{L}\right)\left(\frac{\mu_0 I}{2\pi a}\right)(2\pi a L) = VI \tag{A6}$$

which yields the total energy dissipated per unit time in the volume. Obviously, in deriving this result we have ignored the fringing fields and the contributions at the ends of the thin wire.

APPENDIX B
STATIC MAGNETIC FIELDS FROM AN IDEALIZED AURORAL ARC

Figure B1 shows an idealized auroral arc which is constructed of 3 infinite sheet currents with current density **K** (Amps/(unit length)). The field-aligned current sheet at $x = -d/2$ has a downward current in the $-\hat{z}$ direction, and the parallel sheet at $x = +d/2$ consists of upward current. Connecting the 2 sheets at $z = 0$ is a sheet of $\hat{x}$-directed current with width d which models the layer of Pedersen current in an auroral arc. To find the magnetic field vector **H** due to the current sheets we can use the Biot-Savart law:

$$\mathbf{H} = \frac{1}{4\pi} \int_{A'} \frac{\mathbf{K} \times \hat{a}_R}{R^2} \, dA' \, . \tag{B1}$$

$\hat{a}_R$ is a unit vector from the current sheets (at primed coordinates) to an observation point (at unprimed coordinates), and for the $\hat{z}$-directed sheet at $x = d/2$ it is given by

$$\hat{a}_R = \frac{(x - d/2)\hat{x} - y'\hat{y} + (z - z')\hat{z}}{\sqrt{(x - d/2)^2 + y'^2 + (z - z')^2}} \, . \tag{A2}$$

We have assumed that we are observing in the $y = 0$ plane. Using this with $\mathbf{K} = K\hat{z}$ in (B1) leads to

Ideal Auroral Arc

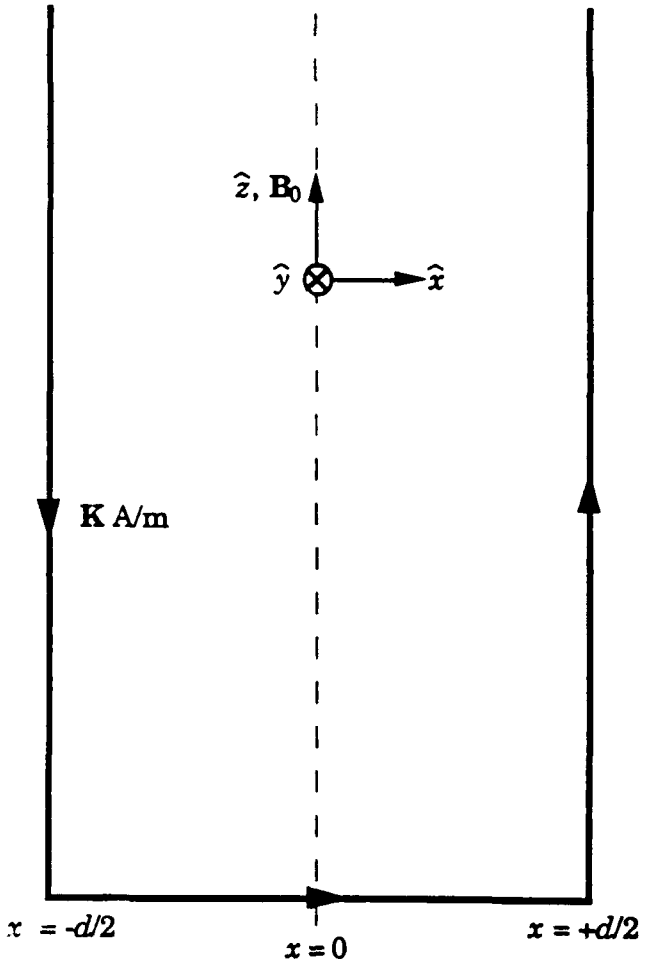


Fig. B1.

Figure B1 Geometry used to calculate magnetic fields due to an idea auroral arc which produces no Hall current and which has an infinitely thin Pedersen current layer at $z = 0$.

$$\mathbf{H}_1 = \frac{K}{4\pi}\int\int_{A'} \frac{y'\hat{x} + (x - d/2)\hat{y}}{\left((x - d/2)^2 + y'^2 + (z - z')^2\right)^{3/2}}\, dy'dz' \,. \tag{B3}$$

We use the subscript "*1*" to denote the magnetic field due only to the current sheet at $x = +d/2$. The $\hat{x}$ component in (B3) vanishes because the integrand is an odd function of y'. Integration of the remaining term over y' can be carried out with the aid of the following integral:

$$\int \frac{dw}{(w^2 + a^2)^{3/2}} = \frac{w}{a^2(w^2 + a^2)^{1/2}} \,. \tag{B4}$$

The result is

$$\mathbf{H}_1 = \frac{K}{2\pi}\int_0^{\infty} \frac{(x - d/2)\hat{y}}{(x - d/2)^2 + (z - z')^2}\, dz' \,. \tag{B5}$$

Another integral identity helps at this stage:

$$\int \frac{a\,dw}{w^2 + a^2} = \frac{1}{a}\tan^{-1}\left(\frac{w}{a}\right). \tag{B6}$$

Note that in applying (B6) to (B5) a sign change is introduced because $dw = -dz'$. The result of this integration gives the contribution to the magnetic field $\mathbf{H}_1$ from the field-aligned current sheet at $x = +d/2$:

$$\mathbf{H}_1 = \frac{-K\hat{y}}{2\pi}\left(\frac{\pi}{2}\mathrm{sgn}(d/2 - x) - \tan^{-1}\left(\frac{z}{x - d/2}\right)\right). \tag{B7}$$

The sgn(w) function is +1 for $w > 0$ and -1 for $w < 0$. The contribution $\mathbf{H}_2$ from the current sheet at $x = -d/2$ can be found by changing the sign of d and K in (B7):

$$\mathbf{H}_2 = \frac{-K\hat{y}}{2\pi}\left(\frac{\pi}{2}\mathrm{sgn}(x + d/2) + \tan^{-1}\left(\frac{z}{x + d/2}\right)\right). \tag{B8}$$

Next we will calculate the magnetic field $\mathbf{H}_3$ due to the $\hat{x}$-directed current sheet at $z = 0$. The unit vector from the current sheet to an observer is

$$\hat{a}_R = \frac{(x - x')\hat{x} - y'\hat{y} + z\hat{z}}{\sqrt{(x - x')^2 + y'^2 + z^2}} . \tag{B9}$$

This combined with the fact that $\mathbf{K} = K\hat{x}$ yields

$$\mathbf{H}_3 = \frac{K}{4\pi} \int\int_{A'} \frac{z\hat{y} + y'\hat{z}}{\left((x - x')^2 + y'^2 + z^2\right)^{3/2}} \, dy'dx' . \tag{B10}$$

With the aid of (B4) and (B6) the result of the integration is

$$\mathbf{H}_3 = \frac{-K\hat{y}}{2\pi} \left(\tan^{-1}\left(\frac{x + d/2}{z}\right) - \tan^{-1}\left(\frac{x - d/2}{z}\right) \right) . \tag{B11}$$

Finally we are ready to sum the 3 contributions $\mathbf{H}_1$, $\mathbf{H}_2$, and $\mathbf{H}_3$ to find the total magnetic field $\mathbf{H}$ with the aid of the identity

$$\tan^{-1}(w) + \tan^{-1}(1/w) = \frac{\pi}{2}\mathrm{sgn}(w) . \tag{B12}$$

The result is

$$\mathbf{H} = \frac{-K\hat{y}}{4} \left(\mathrm{sgn}(x + d/2) - \mathrm{sgn}(x - d/2) + \mathrm{sgn}\left(\frac{x + d/2}{z}\right) - \mathrm{sgn}\left(\frac{x - d/2}{z}\right) \right) . \tag{B13}$$

Thus for all points "inside" the ideal arc, i.e. $z > 0$ and $-d/2 < x < d/2$, $\mathbf{H} = -K\hat{y}$. Outside and below the arc, $\mathbf{H}$ is identically zero. The Pedersen current sheet exactly cancels the magnetic fields from field-aligned currents in an ideal arc, and therefore a ground-based magnetometer would not measure a zonal magnetic field under such an arc. However, we have neglected the Hall current associated with auroral arcs, and this current *will* produce a magnetic perturbation on the ground in the meridional (cross-arc) direction.

REFERENCES

Akasofu, S.-I., Auroral arcs and auroral potential structure, in *Physics of Auroral Arc Formation,* edited by S.-I. Akasofu and J. R. Kan, pp. 1-14, American Geophysical Union, Washington, D. C., 1981.

Alfven, H., *Cosmical Electrodynamics*, Oxford University Press, New York, 1950.

Allen, J., Sauer, H., Frank, L., and Reiff, P., Effects of the March 1989 solar activity, *EOS Trans., AGU, 70,* 1479, 1989.

Banks, P. M., and G. Kockarts, *Aeronomy, Part A,* Academic Press, New York, 1973.

Banks, P. M., and G. Kockarts, *Aeronomy, Part B,* Academic Press, New York, 1973.

Berthelier, A., J.-C. Cerisier, J.-J. Berthelier, J.-M. Bosqued, and R. A. Kovrazkhin, The electrodynamic signature of short scale field aligned currents, and associated turbulence in the cusp and dayside auroral zone, *Electromagnetic Coupling in the Polar Clefts and Caps, P. E. Sandholt and A. Egeland (eds.)*, 299, Kluwer Academic Publishers, 1989.

Birkeland, K., *Norwegian Aurora Polaris Expedition*, 1902-3 Part 1, H. Aschehoug and Company, Christiania, 1908.

Boehm, M. H., C. W. Carlson, J. P. McFadden, J. H. Clemmons, and F. S. Mozer, High resolution sounding rocket observations of large amplitude Alfvén waves, *J. Geophys. Res.,* in press, 1990.

Boström, R., G. Gustafsson, G. Holback, G. Holmgren, H. Koskinen, and P. Kintner, Characteristics of solitary waves and weak double layers in the magnetospheric plasma, *Phys. Rev. Lett., 61,* 82, 1988.

Brekke, A., J. R. Doupnik, and P. M. Banks, Incoherent scatter measurements of E region conductivities and currents in the auroral zone, *J. Geophys. Res., 79,* 3773, 1974.

Budden, K. G., *The Propagation of Radio Waves*, 669 pp., Cambridge University Press, 1985.

Chmyrev, V. M., V. N. Oraevsky, S. V. Bilichenko, N. V. Isaev, G. A. Stanev, D. K. Teodosiev, and S. I. Shkolnikova, The fine structure of intensive small-scale electric and magnetic fields in the high-latitude ionosphere as observed by *Intercosmos-Bulgaria* 1300 satellite, *Planet. Space Sci., 33,* 1383, 1985.

Cummings, W. D., and A. J. Dessler, Field-aligned currents in the magnetosphere, *J. Geophys. Res., 72,* 1007, 1967.

Dougherty, J. P., and D. T. Farley, A theory of incoherent scattering of radio waves by a plasma, *Proc. Roy. Soc., A259,* 79, 1960.

Dougherty, J. P., and D. T. Farley, A theory of incoherent scattering of radio waves by a plasma, 3, Scattering in a partly ionized gas, *J. Geophys. Res., 63,* 5473, 1963.

Dubinin, E. M., P. L. Israelevich, N. S. Nikolaeva, I. Kutiev, and I. M. Podgorny, Localized auroral disturbance in the morning sector of topside ionosphere as a standing electromagnetic wave, *Planet. Space Sci., 33,* 597, 1985.

Earle, G. D., Electrostatic plasma waves and turbulence near auroral arcs, Ph. D Thesis, Cornell University, Ithaca, New York, 1988.

Earle, G., and M. C. Kelley, Spectral studies of the sources of ionospheric electric fields, *J. Geophys. Res., 92,* 213, 1987.

Eather, R. H., *Majestic Lights, The Aurora in Science, History, and the Arts,* American Geophysical Union, Washington, D. C., 1980.

Erlandson, R. E., L. J. Zanetti, T. A. Potemra, L. P. Block, and G. Holmgren, Viking magnetic and electric field observations of Pc 1 waves at high latitudes, *J. Geophys. Res., 95,* 5941, 1990.

Farley, D. T., Jr., A theory of electrostatic fields in the ionosphere at nonpolar geomagnetic latitudes, *J. Geophys. Res., 65,* 869, 1960.

Farley, D. T., A theory of incoherent scattering of radio waves by a plasma, 4, The effect of unequal ion and electron temperatures, *J. Geophys. Res., 71,* 4091, 1966.

Farley, D. T., J. P. Dougherty, and D. W. Barron, A theory of incoherent scattering of radio waves by a plasma, 2, Scattering in a magnetic field, *Proc. Roy. Soc., A263,* 238, 1961.

Farley, D. T., Jr., H. M. Ierkic, and B. G. Fejer, Radar interferometry: a new technique for studying plasma turbulence in the ionosphere, *J. Geophys. Res., 86,* 1467, 1981.

Fejer, B. G., Larsen, M. F., and D. T. Farley, Equatorial disturbance dynamo electric fields, *Geophys. Res. Lett., 10,* 537, 1983.

Feynman, R. P., Leighton, R. B., and M. Sands, *The Feynman Lectures on Physics, Vol. II,* Addison-Wesley, Reading, Mass., 1964.

Foster, J. C., J. P. St. Maurice, and V. J. Abreu, Joule heating at high latitudes, *J. Geophys. Res., 88,* 4885, 1983.

Forbes, J. M., and M. Harel, Magnetosphere-thermosphere coupling: An experiment in interactive modeling, *J. Geophys. Res., 94,* 2631, 1989.

Francis, W. E. and R. Karplus, Hydromagnetic waves in the ionosphere, *J. Geophys. Res., 65,* 3593, 1960.

Goertz, C. K., and R. W. Boswell, Magnetosphere-ionosphere coupling, *J. Geophys. Res., 84,* 7239, 1979.

Greifinger, P., Ionospheric propagation of oblique hydromagnetic plane waves at micropulsation frequencies, *J. Geophys. Res., 77,* 2377, 1972.

Gurnett, D. A., R. L. Huff, J. D. Menietti, J. L. Burch, J. D. Winningham, and S. D. Shawhan, Correlated low-frequency electric and magnetic noise along the auroral field lines, *J. Geophys. Res., 89,* 8971, 1984.

Haerendel, G., An Alfvén wave model of auroral arcs, in *High-Latitude Space Plasma Physics,* edited by B. Hultqvist and T. Hagfors, 543 pp., Plenum Press, New York, 1983.

Hallinan, T. J., and T. N. Davis, Small-scale auroral arc distortions, *Planet. Space Sci., 18,* 1735, 1970.

Hargreaves, J. K., *The Upper Atmosphere and Solar-Terrestrial Relations*, 298 pp., Van Nostrand Reinhold Company, New York, 1979.

Hasegawa, A., Particle acceleration by MHD surface wave and formation of aurora, *J. Geophys. Res., 81,* 5083, 1976.

Hasegawa, A., Kinetic properties of Alfvén waves, *Proc. Indian Acad. Sci., 86A,* 151, 1977.

Hughes, W. J., The effect of the atmosphere and ionosphere on long period magnetospheric micropulsations, *Planet. Space Sci., 22,* 1157, 1974.

Hughes, W. J., Pulsation research during the IMS, *Rev. Geophys., 20,* 641, 1982.

Hughes, W. J., and Southwood, D. J., The screening of micropulsation signals by the atmosphere and ionosphere, *J. Geophys. Res., 81,* 3234, 1976.

Iijima, T. and T. A. Potemra, Field-aligned currents in the dayside cusp observed by Triad, *J. Geophys. Res., 81,* 5971-5979, 1976.

Inoue, Y., Wave polarizations of geomagnetic pulsations observed in high latitudes on the Earth's surface, *J. Geophys. Res., 78,* 2959, 1973.

Iyemori, T., and K. Hayashi, PC 1 micropulsations observed by MAGSAT in the ionospheric F region, *J. Geophys. Res., 94,* 93, 1989.

Jaccia, L. G., Static diffusion models of the upper ionosphere above the E-layer, *Res. Space Sci., Smith. Inst. Astrophys. Obs., Spec Rep.,* 332, 1971.

Jenkins, G. M., and D. G. Watts, *Spectral Analysis and Its Applications*, Holden-Day, San Francisco, 1968.

Kelley, M. C., *The Earth's Ionosphere, Plasma Physics and Electrodynamics*, Academic Press, Inc., San Diego, 1989.

Kelley, M. C., M. F. Larsen, C. LaHoz, and J. P. McClure, Gravity wave initiation of equatorial spread F: A case study, *J. Geophys. Res., 86,* 9087, 1981.

Kelley, M. C., D. J. Knudsen, and J. F. Vickrey, Quasi-DC Poynting flux measurements on a satellite: A diagnostic tool for space research, *J. Geophys. Res.* in press, 1990.

Knudsen, D. J., M. C. Kelley, G. D. Earle, C. Carlson, M. Boehm, and B. McFadden, Scale size dependence of auroral field impedances and Poynting flows, *EOS Trans. AGU, 69,* 432, 1988

Knudsen, D. J., Distinguishing Alfvén waves from quasi-static field structures associated with discrete aurora: Sounding rocket and HILAT satellite measurements, *Geophys. Res. Lett.,* in press, 1990.

Kofman, D., and C. Lathuillere, Observations by incoherent scatter technique of the hot spots in the auroral zone ionosphere, *Geophys. Res. Lett., 11,* 1158, 1987.

Koskinen, H., R. Bostrom, and B. Holback, Viking observations of solitary waves and weak double layers on auroral field lines, in *Ionosphere-Magnetosphere-Solar Wind Coupling Processes,* edited by T. Chang, G. B. Crew, and J. R. Jasperse, p. 147, Scientific, Cambridge, Mass., 1989.

Kudeki, E., Plasma turbulence in the equatorial electrojet, Ph. D Thesis, Cornell University, Ithaca, New York, 1983.

Labelle, J. W., Ionospheric turbulence: Case studies in equatorial spread F and development of a rocket-borne interfermeter, Ph. D Thesis, Cornell University, Ithaca, New York, 1985.

Lockwood, M., B. J. I. Bromage, R. B. Horne, J. P. St-Maurice, D. M. Willis, and S. W. H. Cowley, Non-Maxwellian ion velocity distributions observed using EISCAT, *Geophys. Res. Lett., 14,* 111, 1987.

Løvhaug, U. P., and T. Flå, Ion temperature anisotropy in the auroral F-region as measured with EISCAT, *J. Atmos. Terr. Phys., 48,* 959, 1986.

Lysak, R. L., and C. W. Carlson, The effect of microscopic turbulence on magnetosphere-ionosphere coupling, *Geophys. Res. Lett., 8,* 269, 1981.

Lysak, R. L., and C. T. Dum, Dynamics of magnetosphere-ionosphere coupling including turbulent transport, *J. Geophys. Res., 88,* 365, 1983.

Lysak, R. L., Auroral electrodynamics with current and voltage generators, *J. Geophys. Res., 90,* 4178, 1985.

Lysak, R. L., Coupling of the dynamic ionosphere to auroral flux tubes, *J. Geophys. Res., 91,* 7047, 1986.

Lysak, R. L., Theory of auroral zone PiB pulsation spectra, *J. Geophys. Res., 93,* 5942, 1988.

Mallinckrodt, A. J., and C. W. Carlson, Relations between transverse electric fields and field-aligned currents, *J. Geophys. Res., 83,* 1426, 1978.

McPherron, R. L., Magnetospheric substorms, *Reviews of Geophysics and Space Physics, 17,* 657, 1979.

Melrose, D. B., *Instabilities in Space and Laboratory Plasmas,* 280 pp., Cambridge University Press, Camridge, 1986.

Mikkelsen, I. S., T. S. Jørgensen, M. C. Kelley, M. F. Larsen, E. Pereira, and J. Vickrey, Neutral winds and electric fields in the dusk auroral oval 1. Measurements, *J. Geophys. Res., 86,* 1513, 1981.

Mikkelsen, I. S., M. F. Larsen, M. C. Kelley, J. Vickrey, E. Friis-Christensen, J. Meriwether, and P. Shih, Simultaneous measurements of the thermospheric wind profile at three separate positions in the dusk auroral oval, *J. Geophys. Res., 92,* 4639, 1987.

Moorcroft, D. R. , and K. Schlegel, Evidence for non-Maxwellian ion velocity distributions in the F-region, *J. Atmos. Terr. Phys., 48,* 455, 1988.

Mozer, F. S. and R. H. Manka, Magnetospheric electric field properites deduced from simultaneous balloon flights, *J. Geophys. Res., 76 (7),* 1697, 1971.

Mozer. F. S. and R. Serlin, Magnetospheric electric field measurements with balloons, *J. Geophys. Res., 74 (19),* 4739, 1969.

Nicholson, D. R., *Introduction to Plasma Theory,* 292 pp., John Wiley & Sons, New York, 1983.

Papoulis, A., *Probability, Random Variables, and Stochastic Processes,* McGraw-Hill, New York, 1965.

Paul, C. R., and S. A. Nasar, *Introduction to Electromagnetic Fields,* 742 pp., McGraw-Hill Book Company, New York, 1987.

Pereira, A. E. C., Numerical modeling of high latitude winds in the upper atmosphere, Ph.D Thesis, Cornell University, 1979.

Poole, A. W. V., P. R. Sutcliffe and A. D. M. Walker, The relationship between ULF geomagnetic pulsations and ionospheric Doppler oscillations: Derivation of a model, *J. Geophys. Res., 93,* 14656, 1988.

Potemra, T. A., Bythrow, P. F., Zanetti, L. J., Mobley, F. F., and W. L. Scheer, The HILAT magnetic field experiment, *Johns Hopkins APL Technical Digest, 5,* 120-124, 1984.

Providakes, J., Radar interferometer observations and theory of plasma irregularities in the auroral ionosphere, Ph. D Thesis, Cornell University, 1985.

Press, W. H., B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling, *Numerical Recipes: The Art of Scientific Computing,* 818 pp., Cambridge University Press, Cambridge, 1986.

Prince, C. E., Jr. and F. X. Bostick, Jr., Ionospheric transmission of transversely propagated plane waves at micropulsation frequencies and theoretical power spectrums, *J. Geophys. Res., 69,* 3213, 1964.

Ramo, S., J. R. Whinnery, and T. Van Duzer, *Fields and Waves in Communication Electronics*, John Wiley & Sons, New York, 1965.

Rich, F. J., Heelis, R. A., Hanson, W. B., Anderson, P. B., Holt, B. J., Harmon, L. L., Zuccaro, D. R., Lippincott, C. R., Girouard, D., and W. P. Sullivan, Cold plasma measurements on HILAT, *Johns Hopkins APL Technical Digest, 5*, 114-119, 1984.

Richmond, A. D., S. Matsushita, and J. D. Tarpley, On the production mechanism of electric currents and fields in the ionosphere, *J. Geophys. Res., 81,* 547, 1976.

Richmond, A. D., and R. G. Roble, Electrodynamic effects of thermospheric winds from the NCAR thermospheric general circulation model, *J. Geophys. Res., 92.* 12365, 1987.

Robinson, R. M., and R. R. Vondrak, Measurements of E region ionization and conductivity produced by solar illumination at high latitudes, *J. Geophys. Res., 89*, 3951, 1984.

Röttger, J., Wave-like structures of large-scale equatorial spread-F irregularities, *J. Atmos. and Terr. Phys., 35,* 1195, 1973.

Sagalyn, R. C., and H. K. Burke, Atmospheric electricity, in *Handbook of Geophysics and the Space Environment,* edited by A. S. Jursa, p. 20-1, National Technical Information Service, Springfield, VA, 1985.

Sahr, J. D., Observation and theory of the radar aurora, Ph. D Thesis, Cornell University, Ithaca, New York , 1990.

Seyler, C. E., Nonlinear 3-D evolution of bounded kinetic Alfvén waves due to shear flow and collisionless tearing instablility, *Geophys. Res. Lett., 15,* 756, 1988.

Schunk, R. W., and G. C. G. Walker, Theoretical ion densities in the lower thermosphere, *Planet. Space Sci., 21, pp.* 1875-1896, 1973.

Siefring, C. L., and M. C. Kelley, Analysis of standing wave patterns in VLF transmitter signals: Effects of sporadic-E layers and *in-situ* measurements of low electron densities, *J. Geophys. Res.*, in press, 1990.

Smiddy, M., Burke, W. J., Kelley, M. C., Saflekos, N. A., Gussenhoven, M. S., Hardy, D. A. and F. J. Rich, Effects of high-latitude conductivity on observed convection electric fields and Birkeland currents, *J. Geophys. Res., 85*, 6811-6818, 1980.

Stern, D. P., Large-scale electric fields in the Earth's magnetosphere, *Reviews of Geophysics and Space Physics, 15*, 156, 1977.

Stix, T. H., *The Theory of Plasma Waves*, 283 pp., McGraw-Hill, New York, 1962.

Sugiura, M., N. C. Maynard, W. H. Farthing, J. P. Heppner, and B. G. Ledley, Initial results on the correlation between the magnetic and electric fields observed from the DE-2 satellite in the field-aligned current regions, *Geophys. Res. Lett., 9*, 985, 1982.

Sugiura, M., A fundamental magnetosphere-ionosphere coupling mode involving field-aligned currents as deduced from DE-2 observations, *Geophys. Res. Lett., 11*, 877, 1984.

Swartz, W. E., J. F. Providakes, M. C. Kelley, and J. F. Vickrey, The effect of strong velocity shears on incoherent scatter spectra: a new interpretation of unusual high-latitude spectra, *Geophys. Res. Lett., 15*, 1341, 1988.

Takahashi, K. S. Kokubin, T. Sakurai, R. W. McEntire, T. A. Potemra, and R. E. Lopez, AMPTE/CCE observations of substorm-associated standing Alfvén waves in the midnight sector, *Geophys. Res. Lett., 15*, 1287, 1988.

Temerin, M., K. Cerny, W. Lotko, and F. S. Mozer, Observations of double layers and solitary waves in the auroral plasma, *Phys. Rev. Lett., 48*, 1175, 1982.

Vickrey, J. F., Vondrak, R. R., and S. J. Matthews, Energy deposition by precipitating particles and Joule dissipatioin in the auroral ionosphere, *J. Geophys. Res., 87*, 5184-5196, 1982.

Vickrey, J. F., R. C. Livingston, N. B. Walker, T. A. Potemra, R. A. Heelis, M. C. Kelley, and F. J. Rich, On the current-voltage relationship of the magnetospheric generator at intermediate spatial scales, *Geophys. Res. Lett., 13*, 495, 1986.

Weimer, D. R., C. K. Goertz, D. A. Gurnett, N. C. Maynard, and J. L. Burch, Auroral zone electric fields from DE 1 and 2 at magnetic conjuctions, *J. Geophys. Res., 90*, 7479, 1985.

Zmuda, A. J., J. H. Martin, and F. T. Heuring, Transverse magnetic disturbances at 1100 kilometers in the auroral region, *J. Geophys. Res., 71*, 5033, 1966.

Attachment 8

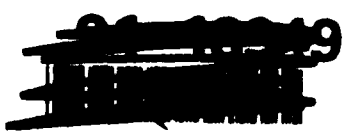
AFOSR-TR- 91 0982

THESIS BY:

THOMAS J. MULLEN

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

Subcontract No.# S-789-000-044

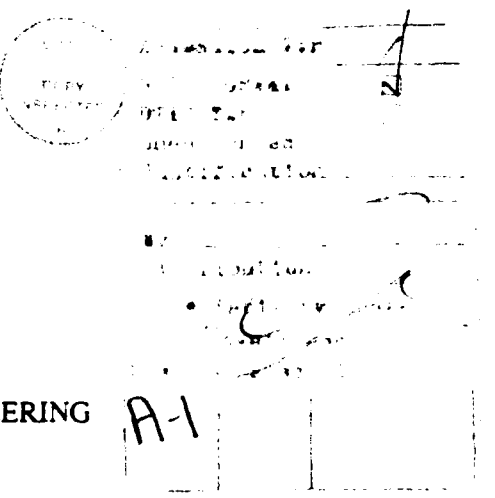


TRANSFER FUNCTION ANALYSIS OF AUTONOMIC ACTIVITY DURING MOTION SICKNESS

by

THOMAS JAMES MULLEN

B.S., Electrical Engineering,
Worcester Polytechnic Institute, 1987



Submitted to the
Department of Electrical Engineering
In Partial Fulfillment of the Requirements
for the Degree of

MASTER OF SCIENCE IN ELECTRICAL ENGINEERING

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
June, 1990

Signature of Author: Thomas J. Mullen
Department of Electrical Engineering
May, 1990

Certified by: Charles M. Oman
Charles M. Oman, Thesis Supervisor
Department of Aeronautics and Astronautics

Certified by: R. J. Cohen
Richard J. Cohen, Thesis Supervisor
Harvard-MIT Division of Health Sciences and Technology

Accepted by: _____
Arthur C. Smith, Chair
Department Committee on Graduate Students

TRANSFER FUNCTION ANALYSIS OF AUTONOMIC ACTIVITY DURING MOTION SICKNESS

by

THOMAS JAMES MULLEN

Submitted to the
Department of Electrical Engineering
in partial fulfillment of the requirements for the Degree of
Master of Science in Electrical Engineering

ABSTRACT

The physiological mechanisms underlying motion sickness are poorly understood. The role of the autonomic nervous system is controversial. This thesis describes a series of experiments on human subjects in which a new technique was applied to assess autonomic activity during motion sickness. The technique (Saul et al., Am J Physiol 256:H153-161, 1989) requires estimation of the transfer function between instantaneous lung volume (ILV) and instantaneous heart rate (IHR). Components of the transfer function provide information concerning relative levels of autonomic activity. In order to broaden the respiratory signal, so as to allow accurate transfer function estimation, subjects breathe in synchrony with a series of randomly spaced auditory tones. This process is termed random interval breathing.

Eighteen subjects (ages 18-30 yrs, 11 male, 7 female) participated. Control recordings of instantaneous lung volume (ILV, measured by inductance plethysmography) and electrocardiogram (ECG) were made during two fifteen minute random interval breathing segments. During the first segment, subjects were seated motionless and during the second they were seated rotating about an earth vertical axis. Each subject was then fitted with a pair of prism goggles which reverse the left-right visual field and was asked to perform a pre-specified series of manual tasks until moderate levels of motion sickness were attained. A relatively constant level of sickness was then maintained with periodic eye closure during rotation with the goggles. Lung volume and ECG were recorded during this motion sick condition as the subject completed a third random interval breathing sequence.

Comparisons of ILV to IHR transfer functions from the two non-sick conditions with each other and with known standards, indicate no change in autonomic control of heart rate due to rotation. Similar comparisons between the two rotating conditions indicate no change in transfer function due to motion sickness. These findings do not support the widely held notion that motion sickness can be classified as a generalized autonomic ("stress") response. A new functional model depicting a more discrete, organ specific role of the autonomic nervous system in the development of motion sickness is presented.

Thesis Supervisors:

Dr. Charles M. Oman,	Senior Research Engineer, Dept. of Aeronautics and Astronautics
Dr. Richard J. Cohen,	Hermann von Helmholtz Associate Professor, Harvard-MIT Div. of Health Sciences and Technology

Acknowledgements

This research was partially supported by NASA-JSC Grant NAG9-244 and NIH Grant 1RO1HL39291. Fellowship support was provided by an Air Force Laboratory Graduate Fellowship (Contract F49620-86-C-0127/SB5861-0426, Subcontract S-789-000-044).

I would first like to thank my thesis supervisors, Dr. Oman and Dr. Cohen, for their guidance and for their interest in pursuing a joint research effort despite the added complications associated with such projects.

Chris Eagon and Paul Albrecht guided my introduction to motion sickness research and the analysis of heart rate variability, respectively. I am very grateful for their contributions to my early pilot studies which were the precursors to this thesis research.

I would especially like to thank Ron Berger for his time and efforts on my behalf. Despite his busy schedule, he always found time to contribute to each phase of this research.

Dr. Alan Natapoff gave many hours toward the analysis of the experimental data.

The students in the Man-Vehicle lab are the best group with whom I've had the privilege to work. The healthy mix of work and recreation created an environment in which the group excelled at both. I am thankful for the friendship and professional interactions with them all. In particular, the laboratories' (sometimes) faithful leaders, Mark and Dan, and my officemates, Brad and Cheryl, have contributed to making the laboratory tolerable and the fun times more so.

I would also like to thank Patti for her companionship and moral support during those times when it was most needed.

Finally, I would like to express my love and gratitude to the two individuals who are most deserving: my mother, Vera, and my father, John.

Table of Contents

List of Figures and Tables

Figures

Tables

1 Introduction

1.1 Motivation

In modern society, most individuals have experienced motion sickness at one time or another. Whether on airplanes, automobiles, ships, amusement park rides or other modes of transportation, many have felt the discomforts associated with sickness. In fact, it is reasonable to believe that ever since humans began using vehicles for passive transport, motion sickness has been a concern. The first known written accounts of motion sickness were made by the ancient Greeks and, interestingly, the word "nausea" derives from the Greek word "naus", meaning ship (Reason and Brand, 1987). As modes of transportation have become more advanced and higher performance vehicles have evolved, the incidence of motion sickness has become more widespread. One of the newest forms of motion sickness, termed Space Motion Sickness (SMS), afflicts some astronauts during space flight. (Crampton, 1990) In most cases, motion sickness is merely an inconvenient and

unpleasant experience, but in space and military operations, it becomes a more costly and possibly life threatening occurrence.

The nature of motion sickness, its relationship with space motion sickness, and its impact on military operations have aroused significant research interest. (reviewed collectively by Tyler and Bard, 1949; Money, 1970; Reason and Brand, 1975; Crampton, 1990). As part of their research, many groups have induced sickness in laboratory subjects and have recorded their physiological responses. Numerous cardiovascular (Graybiel and Lackner, 1980), respiratory (Cowings et al., 1986), gastrointestinal (Stern et al, 1987; Rague, 1987; Eagon, 1988), biochemical (Eversmann et al., 1978; Habermann et al., 1978) and other physiological measures (Isu et al., 1987a, Isu et al., 1987b; Gaudreault, 1987; Drylie, 1987) have been monitored. Attempts have been made to correlate signs to symptoms, and theories have been proposed regarding systemic roles in the development of sickness.

The autonomic nervous system (ANS) is the division of the human nervous system which is generally responsible for subconscious control of bodily functions, maintenance of homeostasis, and mediation of an individual's physiological responses to stresses. As such, it has naturally been suspected to contribute to motion sickness. However, since no acceptable physiological definition of motion sickness is available, it is not clear how the ANS should be expected to respond during the syndrome, and some controversy exists. Some researchers speculate that motion sickness should be viewed as a generalized stress response and that the ANS, therefore, should be expected to respond in its classic "fight or flight" manner. The "fight or flight" response typically involves inhibition of the parasympathetic division of the ANS but more importantly, widespread activation of the sympathetic division. Evidence from pharmacological studies and the known effectiveness of certain drug therapies, however, provide clues which do not generally support this stress response view of motion sickness.

Furthermore, if one considers a hypothetical functional purpose of motion sickness, a generalized stress response seems inappropriate, perhaps. It has been proposed from an evolutionary standpoint that motion sickness could be a manifestation of an animal's early warning response to ingested toxins (Treisman, 1977). That is, the disorientation and sensory rearrangement typically associated with motion sickness are similar to those associated with ingestion of a toxin; therefore, the body responds by expelling the contents of the stomach and presumably the toxin. Under this hypothesis, it is expected that the parasympathetic system is inhibited to retard gastric motility and thus confine the toxin to the stomach for expulsion (Davis, 1986). However, under this hypothesis, the parasympathetic inhibition need not be accompanied by a widespread sympathetic activation as would occur in a generalized stress response.

In attempts to investigate the underlying physiology of motion sickness, many studies have focussed on observing trends in physiological parameters such as mean heart rate, skin potential, sweating, or skin pallor. Researchers have interpreted these parameters as autonomic manifestations and have extrapolated to draw conclusions concerning the ANS. Interpretations, however, are confounded by a number of issues. First, the autonomic nervous system consists of multiple control pathways which may interact in a complex way. In fact, at most organ sites, qualitatively similar effects can be induced by either division of the ANS. For example, increases in heart rate may be caused either by an increase in sympathetic or a decrease in parasympathetic activity at the sinoatrial node. Second, while observation of local effects may provide insight into local ANS activity, the broader integrated function of the ANS is not necessarily represented. Finally, trends in these physiological parameters during motion sickness have not been found to be consistent either within or between studies. These inconsistencies may be due in part to differences between subjects. However, they may also be due in part to a lack of controls implemented

in many studies. Activities such as changes in posture or exercise, which are known to have autonomic effects independent of motion sickness, have often been uncontrolled.

In order to better assess autonomic activity during motion sickness, it is desirable to use a well understood measure and a well established technique. Dr. R.J. Cohen of MIT and colleagues have developed such a technique using noninvasive measures of heart rate variability (Berger et al., 1989a; Berger et al., 1989b; Berger et al., 1986; Chen et al., 1987; Appel et al., 1989a). Through a number of studies, they have demonstrated that the transfer function from instantaneous lung volume (ILV) to instantaneous heart rate (IHR) may be used as sensitive probe of relative levels of autonomic control of heart rate. Further, they have developed an effective technique, termed Random Interval Breathing (RIB), to broaden the spectral content of the respiratory signal (input stimulus) and thus allow accurate estimation of the desired transfer function.

1.2 Purpose

The primary objective of this study was to apply the techniques developed by Cohen's group to determine whether or not autonomic changes, as detectable by these techniques, occur during motion sickness. As a prelude to this research, it was necessary to develop an experimental protocol which would allow controlled application of the technique. A protocol which limited confounding autonomic effects and permitted subjects to complete segments of random breathing was required.

II Background

2.1 Motion Sickness

2.1.1 General Characteristics

Motion sickness, as the name implies, is an illness which can be induced by certain motion environments. It is characterized by a collection of signs and symptoms, the most common of which are pallor, cold sweating, fatigue, nausea, and vomiting. However, many other signs and symptoms have been reported (Money, 1970), and the combination and relative severity of signs and symptoms varies between individuals. Generally, the first symptoms are mild ones such as fatigue, headache, or stomach awareness. These progress toward pallor, cold sweating, and nausea and eventually culminate in retching and vomiting if no preventive measures are taken. The dynamics of the time course of symptoms show four consistent characteristics (Bock and Oman, 1982; Gillingham, 1986). First, there is a

latency to the appearance of first symptoms. Second, there is a tendency for symptoms to avalanche as one nears the vomiting end-point. Third, symptom levels tend to overshoot upon the removal of provocative motion stimulation; and fourth, once symptoms are established, there is a period of hypersensitivity to stimulation. These dynamics have led some researchers to envision two pathways for the development of symptoms; a fast pathway to help explain the avalanching phenomena and a slow path to help explain latency and hypersensitivity (Oman, 1982; Oman 1990). As yet, however, the physiological mechanisms associated with this hypothesis have not been identified.

Many motion sickness signs and symptoms are qualitatively similar to those of other nausea and vomiting syndromes, such as radiation sickness or morning sickness (Grahamme-Smith, 1986). The characteristic which differentiates the various syndromes is their underlying cause. Unfortunately, the physiology of motion sickness, nausea, and vomiting remains poorly understood; therefore, theories concerning causation of sickness must rely heavily on knowledge of what types of stimuli are provocative and who is susceptible.

2.1.2 Incidence of Sickness

Not all types of motion cause motion sickness. People are generally able to participate in high motion activities such as running, dancing, or ball games without developing sickness. However, many situations in which individuals are subjected to passive motion induce motion sickness. The most common examples have the common names "sea sickness", "car sickness" and "air sickness". In each of these cases, an obvious motion is present. There are, however, other situations in which motion sickness occurs, and yet no real body motion is involved. One common example is "cinema sickness", in which a stationary individual develops symptoms while observing a moving visual scene. Other

conditions contrived for experimental studies have demonstrated that true subject motion is not required for the development of motion sickness (Reason and Brand, 1975).

Nearly everyone is susceptible to motion sickness. Most individuals, given a long enough exposure to the proper stimulation, will develop symptoms. It is difficult, however, to assess the incidence of motion sickness across the general population. The incidence is highly dependent on the type of motion environment. A number of studies provide rough estimates among sub-populations. The most recent of these studies suggest that 36 percent of 20,000 surveyed ferry passengers (Lawther and Griffin, 1988) and 67 percent of shuttle astronauts during their first flights (Davis J.R. et al., 1988) were afflicted. In general, susceptibility to motion sickness tends to peak as a function of age between ages 12 and 21, and women tend to report that they are more susceptible than men (Reason and Brand, 1975). However, susceptibility varies a great deal from person to person, dependent on the type of stimulation.

While very few individuals are believed to be completely immune, most are able to develop at least partial immunity through adaptation. During a long enough exposure to a particular environment, symptoms will eventually subside and individuals will become resistant to motion sickness during continued stimulation. However, upon return to the normal motion environment after extended time in an unusual one, individuals may experience symptoms as they re-adapt to the normal situation. For example, after extended periods aboard ship, many have reported symptoms upon return to land. This syndrome is termed "mal de debarquement". The rate at which adaptation can be attained is dependent on the individual and on the degree of stimulation provided by a particular environment. Typically, several days aboard ship or spacecraft are required to fully adapt and regain health. There is evidence that upon repeated exposure to the environment some aspects of adaptation may be preserved (Parker, 1972; Reason and Brand, 1975).

There is one group of individuals which does seem to be immune to motion sickness. A number of studies have demonstrated that those lacking vestibular function can endure motion situations which are normally highly provocative. In studies conducted in the Pensacola Slow Rotating Room and at sea, Graybiel and coworkers found that vestibular defectives not only reported no adverse symptoms, but typically enjoyed the experience (Graybiel, 1963).

2.1.3 Causation: The Conflict Theory

Before it was discovered that labyrinthine defectives seem immune to sickness, the most prevalent theories attributed motion sickness either to reduced blood flow to the brain or to mechanical stimulation of abdominal afferents caused by motion of the viscera (Reason and Brand, 1975). The discovery of the importance of the vestibular system led to the "vestibular overstimulation" theory which asserts that continual intense stimulation of the vestibular organs produces sickness. It purported to explain why travel in ships, planes, and automobiles is provocative, while activities like running and dancing are not. The theory, however, has lost support partially due to the realization that sickness can occur in the absence of true subject motion (Guedry, 1968; Oman, 1982).

The most popular current theory is the Conflict Theory. Claremont (1931) is generally cited as the first to note that motion sickness develops when two sensory modalities receive conflicting motion cues. Through a number of revisions and refinements, this idea has become known as the Conflict Theory. The first major revision to the theory was proposed by Reason (1978), who noted that the essential conflict is more likely between expected sensory input and the input actually received by the brain. This formulation was better able to explain adaptation and made more physiological sense since the brain would now

compare signals from the same sensory modality. Reason proposed the concept of a "Neural Store" or neural memory in which the brain maintains a sort of dictionary of paired sensory-motor memory traces.

A second significant revision, proposed by Oman (1982; 1990), eliminated the need for the "Neural Store" and provided additional insight into a number of different mechanisms by which adaptation could occur. Oman (1982; 1990), using an Observer Theory approach, developed a heuristic mathematical model of body motor control (refer to Figure 2.1, from Oman, 1990). In this development, the brain employs an internal model of the body and its sensors to calculate expected sensory signals. This internal model thus replaces the "Neural Store". Furthermore, in this development, a "conflict" signal has functional value

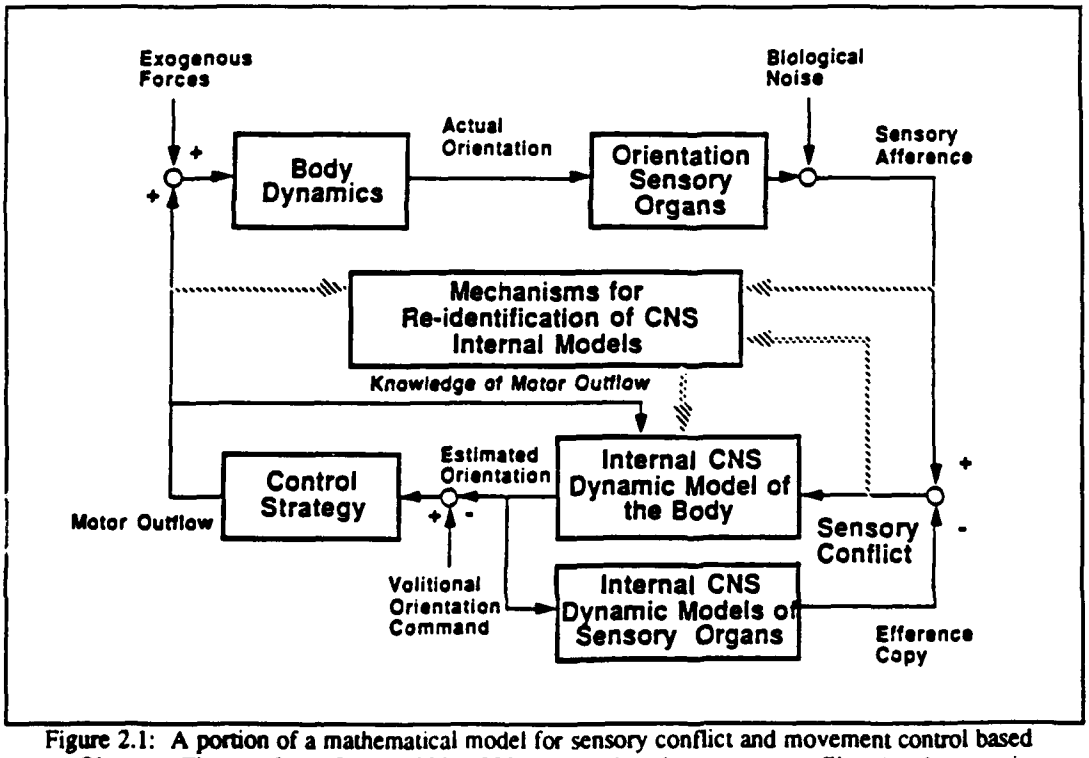


Figure 2.1: A portion of a mathematical model for sensory conflict and movement control based on Observer Theory (from Oman 1990, 1982). Note that the sensory conflict signal serves in feedback control.

other than to make one sick; it serves as an error signal generated from the feedback return of the control system.

In its present form, the Conflict Theory may be stated as follows (Oman, 1990): Motion sickness results when a conflict signal in the brain, normally used in posture and/or motor control, becomes large. This occurs when actual and anticipated sensory information are not in agreement. Further, since it is known that labyrinthine defective individuals are immune to motion sickness, the vestibular system must be implicated in the conflict.

2.1.4 Treatment and Prevention

Despite significant research and a better understanding of conditions leading to motion sickness, the best prevention still remains avoidance of provocative situations (Gillingham, 1986). The best treatments remain either removal of or adaptation to the provocative stimulation. Attempts to pre-adapt to prevent sickness in novel situations have generally failed due to the condition specificity of adaptation (Reason and Brand, 1975).

Pharmacological attempts at prevention and treatment have met with only moderate success. Many drugs and drug combinations have been tested as combatants to motion sickness (Wood and Graybiel, 1970, 1972; Kohl, 1985, 1987; Attias, 1989; Parrot, 1989). While no drug therapy has been found to confer immunity, some are effective in increasing resistance to sickness (refer to Figure 2.2 from Graybiel and Lackner, 1980). Presently, the most effective single drug seems to be Scopolamine (Wood and Graybiel, 1972; Gillingham, 1986). However, it has undesirable side effects such as dry mouth, drowsiness, pupillary dilation, and impaired visual accommodation. In efforts to alleviate

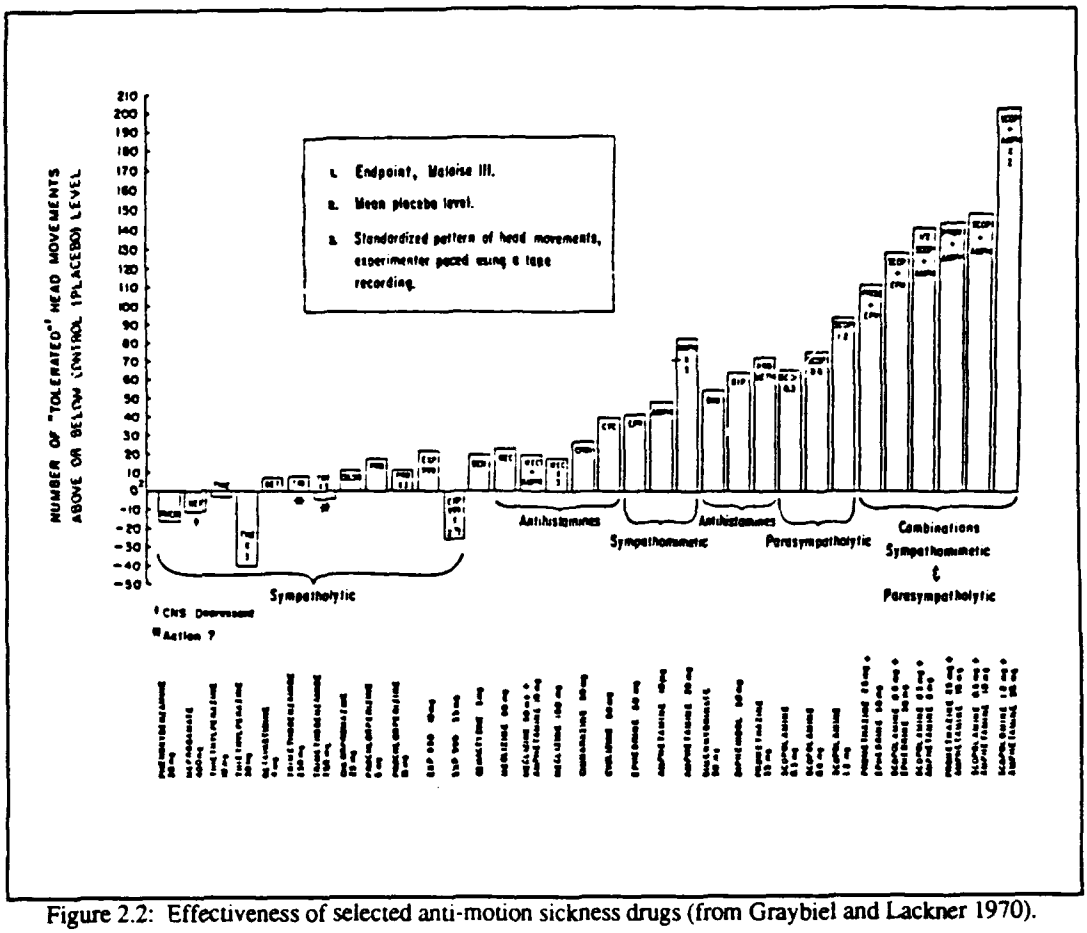


Figure 2.2: Effectiveness of selected anti-motion sickness drugs (from Graybiel and Lackner 1970).

these side effects two approaches have been taken. First, scopolamine is often administered in combination with dextro-amphetamine. The amphetamine alone is also somewhat effective in combating sickness. In combination, it serves to abate some of the side effects of scopolamine and in fact, the combination is more powerful than scopolamine alone (Wood and Graybiel, 1972). The second approach is to more accurately control the serum levels of scopolamine and avoid the peaking associated with oral administration. A transdermal application patch worn behind the ear has been shown to maintain effectiveness in combating sickness. However, while some have reported fewer side effects (Attias, 1989) using this application, others have found that the side effects remain a problem

(Parrot, 1989). Drug therapies have not yet been demonstrated to be effective in astronauts during space flight, as double blind, placebo controlled studies have not been attempted.

An alternative prevention or treatment proposed by some researchers is biofeedback and autogenic training. Individuals are trained to recognize their symptoms and through biofeedback and relaxation training are taught to control their autonomic responses to the motion stress. Accounts of the effectiveness of this treatment vary. (Cowings et al., 1990; Toscano and Cowings, 1982; Graybiel, 1980; Levy, 1981; Dobie, 1987) In Section 2.2.2, biofeedback and autogenic training will be discussed in greater detail.

2.2 The Autonomic Nervous System

2.2.1 Basic Anatomy and Physiology

The Autonomic Nervous System (ANS) is the motor division of the human nervous system which innervates smooth muscle, cardiac muscle and glands. (refer to Figure 2.3 modified from Tortora and Evans, 1986) It is generally responsible for integrating information from afferents* and exerting subconscious control of bodily functions. Its activities include regulation of digestion, heart rate and contractility, circulation, body temperature, breathing, and gland secretions (Tortora and Evans, 1987; Hockman, 1987; Guyton, 1986). Until early in this century, the ANS was considered to be functionally and anatomically distinct from the Central Nervous System (CNS) (Hockman, 1987). However, it is now recognized that autonomic control is achieved through reflexes at the level of the spinal cord or through central nervous mechanisms where control is ultimately

* By historical definition, stemming primarily from the writings of Gaskell and Langley (Guyton, 1986), the ANS includes only motor fibers; the forward loop of the control systems. Afferent fibers, which are by strict definition not part of the autonomic nervous system, provide regulatory feedback to close the control loops.

mediated by the hypothalamus (Tortora and Evans, 1987; Hockman, 1987; Van Toller, 1979; Guyton, 1986).

The system is composed of two divisions, the sympathetic and the parasympathetic, each of which encompasses multiple regulatory pathways.. Most visceral organ systems controlled by the ANS are innervated by both systems; a phenomenon termed dual innervation. Further, each division exerts tonic control on most organ systems. These

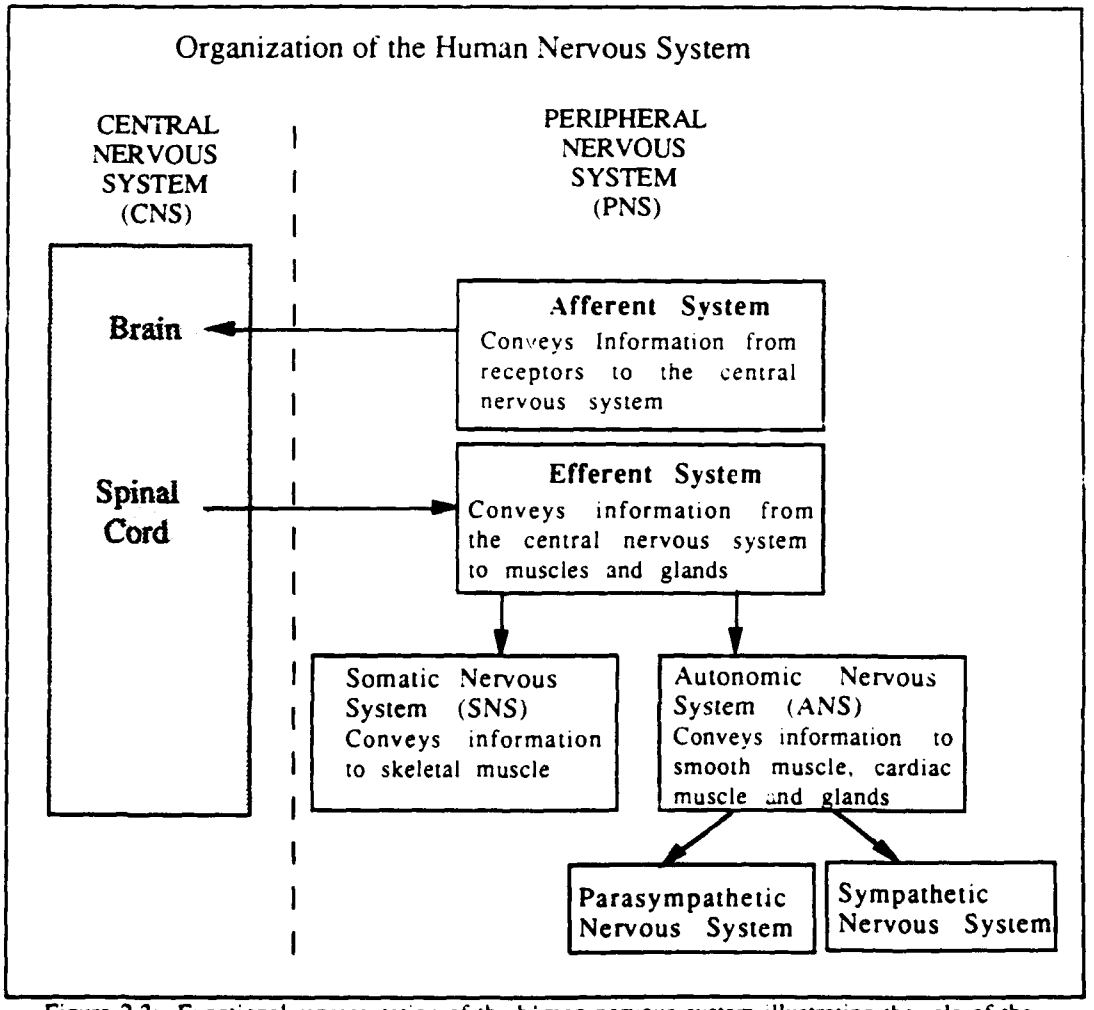


Figure 2.3: Functional representation of the human nervous system illustrating the role of the Autonomic Nervous System (ANS). (modified from Tortora and Evans, 1986)

basal rates of activity, termed parasympathetic and sympathetic "tone", allow each division to exert bi-directional control over the effector by either decreasing or increasing nervous activity. The two divisions most often act functionally in opposition to one another. That is, excitation of one division will generally have the opposite effect on most organ systems than excitation of the other division. This agonist/antagonist character of the two systems and the existence of tonic activity allows for precise control of effector organ systems.

The sympathetic division is also called the thorocolumbar division since its fibers extend from ganglia projecting from the thoracic and lumbar segments of the spinal cord. (refer to Figure 2.4 from Guyton, 1986) At ganglia and some effector organs, sympathetic fibers release acetylcholine. However, at most effector sites, the neurotransmitter is norepinephrine. The effects of sympathetic activation on particular organs are indicated in Table 2.1. Sympathetic effects are most prominently evident during mass activation of the system when it responds almost as a complete unit. This mass activation often occurs as the body responds to fear, pain or other emotional or physical stress, and therefore it has been termed the "stress response" or "fight or flight" response. Widespread sympathetic activation prepares the body to deal with the stresses by, for example, increasing heart rate and arterial pressure and diverting blood flow from visceral organs to those skeletal muscles which are needed for response.

In the past, the sympathetic system was thought to always respond via mass discharge, exerting similar influence on all controlled organs. This assumption has come into question. In human and animal studies, prominent rhythms at the respiratory and heart rates have been found in renal, cardiac splanchnic, and muscle sympathetic nerves (Cohen and Gootman, 1970; Eckberg et al., 1988; Ninomiya et al., 1976; Saul et al., 1990) indicating associations between the different outflows. However, there is also evidence

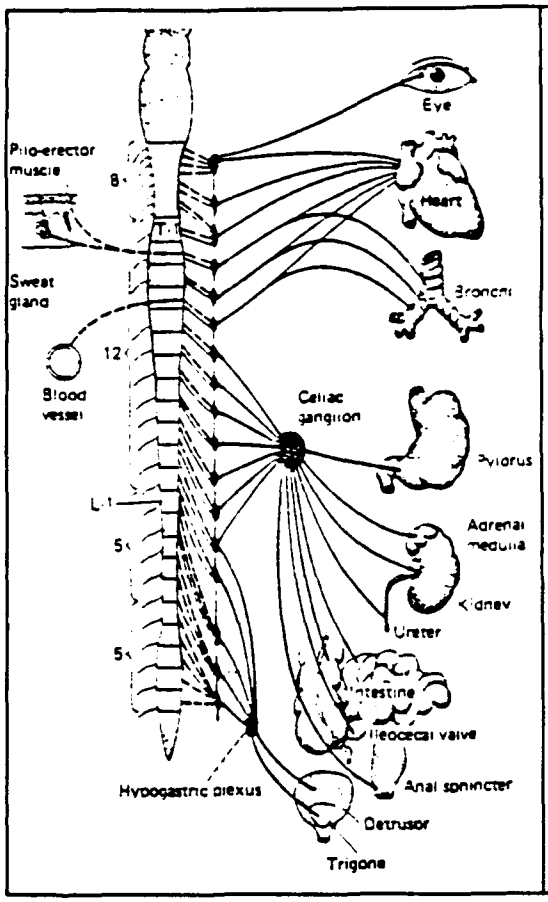


Figure 2.4: The Sympathetic Division of the Autonomic Nervous System (from Guyton, 1986)

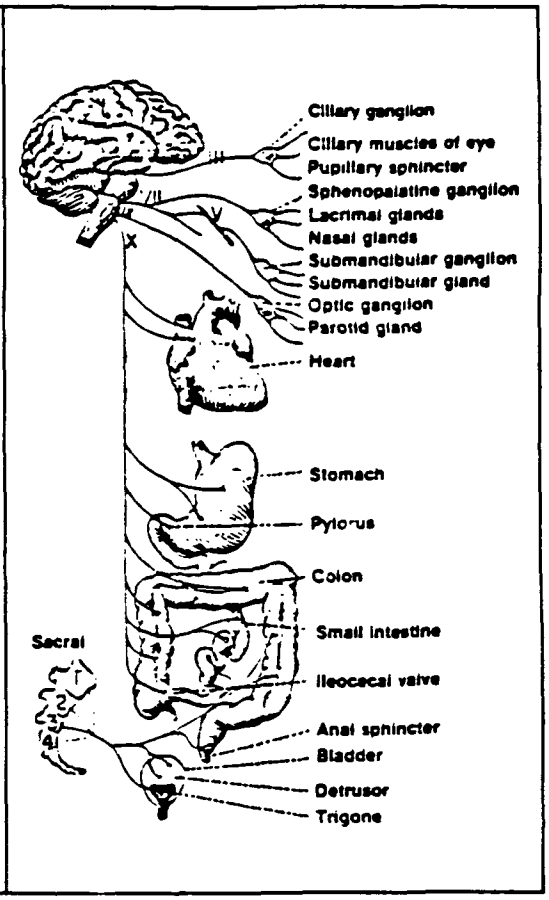


Figure 2.5: The Parasympathetic Division of the Autonomic Nervous System (from Guyton, 1986)

from direct sympathetic nerve recordings which demonstrates dissociation of sympathetic activity in different organs during mild stresses (Mark et al., 1986; Wallin, 1986; Karim et al., 1972; Simon and Riedel, 1975; Victor et al., 1986). Furthermore, there are a number of instances when the sympathetic system would be expected to exert very narrow isolated effects. For example, in control of body temperature, the system controls sweating and blood flow to the skin without affecting other organs. In fact, as body temperature rises, the sympathetic system must increase its influence over the sweat glands to induce sweating but decrease its activity in skin vessels to increase peripheral blood flow (Guyton, 1986).

Table 2.1 Autonomic Effects on Bodily Organs

(compiled from similar tables in Hockman (1987) and Guyton (1986))

ORGAN	SYMPATHETIC STIMULATION	PARASYMPATHETIC STIMULATION
Heart	Increased Rate Increased Contractility	Decreased Rate Decreased Contractility
Coronary Arteries	Constriction (Alpha) Dilation (Beta)	Dilation
Systemic Arterioles		
Muscle	Constriction (alpha) Dilation (Cholinergics & Beta)	No effect
Abdominal	Constriction	No effect
Skin	Constriction	No effect
Piloerector Muscles	Contraction	No effect
Small Intestine, Colon ¥Rectum	Decreased Secretions Decreased Peristalsis	Increased Secretion Increased Peristalsis
Adrenal Medulla	Increased Secretion	No effect
Glands: Lacrimal, Nasal, Parotid, Gastric, Submandibular	Slightly increased Secretion	Largely increased Secretion
Sweat Glands	Largely increased secretion (Cholinergics)	No effect

Thus, it now seems evident that while the sympathetic system often exerts widespread control activity, it is also capable of more localized control. This issue will be addressed further in conjunction with transfer function estimation in Section 2.3.3.

The parasympathetic division is also known as the craniosacral division since its fibers extend from ganglia in the brainstem and sacral segments of the spinal cord (refer to Figure 2.5 modified from Guyton, 1986). At ganglia and effector sites, the fibers release the

neurotransmitter acetylcholine. The effects of parasympathetic activation on particular organs are indicated in Table 2.1. In general, parasympathetic stimulation tends to bring the body toward a more relaxed state by, for example, decreasing heart rate and increasing digestive activity. In contrast to the sympathetic system, the parasympathetic system usually exerts very narrow organ-specific control. The system often affects cardiovascular activity without altering activity of other organ systems. (Guyton, 1986) On many occasions, however, there may be close association between parasympathetic activity in different effectors. For example, although on occasion salivation and gastric secretion may occur independently, these digestive secretions are often synchronized.

2.2.2 The Role of the ANS in Motion Sickness

The role the autonomic nervous system plays in the development of motion sickness remains undefined and a subject of much speculation. Generally, four categories of evidence are cited in support of autonomic contributions to sickness: (1) some success has been reported in applying biofeedback and autogenic training to alleviate sickness, (2) many signs and symptoms of sickness may be autonomic manifestations, (3) some "autonomically mediated" physiological parameters have been reported to change with sickness, and (4) the most effective drug therapies may be ANS effectors. While the evidence does seem to indicate that the ANS plays a significant role, it does not seem to support a consistent model for ANS contributions.

The first category of evidence, success in applying biofeedback and autogenic training in combating sickness, seems to support the notion that some role is played by the ANS in the development of sickness but does not imply a specific model. Biofeedback training is a process in which subjects are presented with augmented information about a particular

"autonomic" variable (ie. heart rate) and are taught to consciously affect the variable (typically through relaxation techniques). Autogenic training is also a self-regulatory technique. However, it generally does not involve augmented physiological feedback, but rather, is designed to teach subjects exercises by which they can induce specific bodily sensations.

Levy et al. (1981) and Jones et al. (1985) have applied biofeedback and relaxation training in Air Force fliers grounded for chronic severe motion sickness. The fliers were taught a number of relaxation and biofeedback techniques and were provided feedback as they were trained to control their symptoms during Coriolis stimulation. Levy et al. and Jones et al. have reported between 79 and 84 percent of affected fliers have been successful in overcoming sickness and returning to flight status.

Other researchers have reported similar successes in experiments at NASA Ames Research Center (Cowings et al., 1990; Toscano and Cowings, 1982; Cowings and Toscano, unpublished). However, they supplemented the biofeedback training with autogenic therapy. Toscano and Cowings (1982) have reported that trained individuals have significantly greater resistance to sickness than untrained individuals or individuals taught an alternative ("sham") cognitive task. Further, they found that resistance to sickness attained through biofeedback and autogenic training under one motion condition transfers to other conditions (Cowings and Toscano, unpublished). The successes of these researchers, however, have not been matched by others. Dobie et al. found that it was confidence building and desensitization training that provided a significant increase in resistance to sickness and that feedback training provided no significant additional benefit (Dobie et al., 1987).

The second and third categories of evidence are most often cited in support of a model involving increased sympathetic activity during motion sickness. In fact, some researchers have asserted that motion sickness can be viewed as a generalized stress response in which there is a marked, widespread increase in sympathetic activity (Cowings et al., 1986; Johnson and Jongkees, 1974). As evidence, they claim that symptoms such as pallor, cold sweating, and increased salivation, and trends in so called "autonomic manifestations" such as heart rate, blood pressure, respiration, gastrointestinal motility, or skin resistance, can be explained by the postulated sympathetic activation (Reason and Brand, 1975; Money, 1970; Dobie et al., 1987; Cowings et al., 1986; Johnson and Jongkees, 1974). However, this evidence is suspect since symptom patterns are known to vary between individuals and, as Money points out in his 1970 review (Money, 1970), reports of trends in most physiological recordings during sickness differ significantly (to the point of contradiction) from study to study.

As a more recent example of inconsistencies consider reports of heart rate and blood pressure recordings. In 1980, Graybiel and Lackner (1980) subjected 12 individuals to a repeated sudden stop paradigm in a rotating chair and monitored heart rate, blood pressure, and skin temperature as motion sickness developed. They found no significant correlation between any of these measurements and motion sickness symptoms. Cowings et al. (1986) on the other hand, utilized Coriolis stimulation in a rotating chair to induce sickness in 127 subjects. They monitored heart rate, blood pressure, basal skin resistance, and respiration and reported significant trends in all of these "autonomic responses." Similar controversies exist concerning other physiological measures (Stern et al., 1987; Eagon, 1988; Rague, 1987; Drylie, 1987; Gaudreault, 1987; Gordon, 1988, 1989). Thus, symtomatology and physiological data do not provide firm evidence upon which to base a model.

The fourth category of evidence, drug therapies, may be taken to indicate an opposite role for the ANS than that indicated by categories two and three. The most effective drug therapies (in terms of allaying peripheral motion sickness signs and symptoms) are either sympathomimetics, parasympatholytics (anticholinergics) or combinations of the two (Wood and Graybiel, 1970, 1972; Kohl, 1985). In addition, many sympatholytics have been shown to actually increase susceptibility to motion sickness. These findings could be taken to imply that increases in parasympathetic and decreases in sympathetic activities accompany sickness and the drugs are effective in combating these shifts. However, this model is confounded by evidence indicating that some sympatholytics are mildly effective in allaying sickness. Furthermore, it is not clear that the drugs are affecting peripheral nervous system autonomic centers. Although they are best known for their autonomic effects, it is quite possible that their success against motion sickness is due to other, possibly central nervous, mechanisms (Janowsky, 1985; Janowsky et al., 1984; Risch and Janowsky, 1985; Kohl and Homick, 1983). That is, while motion sickness signs may be mediated by autonomic outflow, the drugs may interfere with the development of sickness, not by affecting these peripheral autonomic pathways, but rather, by affecting the central mechanisms which are promoting the autonomic activation.

Cowings et al. have proposed an interesting model to explain the postulated trends of sympathetic activation in their subjects while at the same time accounting for the therapeutic effectiveness of sympathomimetics and parasympatholytics (Cowings, 1986). As indicated earlier in this section, the model is based on highly speculative assumptions, nevertheless it has interesting parallels to other syndromes. They have postulated that motion sickness is accompanied by a widespread sympathetic activation. This activation leads to the sympathetic manifestations. However, based on trends in recovering subjects and in order to explain drug effectiveness, they postulate that this prolonged sympathetic activation may

lead to a "parasympathetic rebound" associated with nausea and vomiting stages of sickness. As Cowings et al. point out, this model is quite similar to models of migraine headache (Sakai and Meyer, 1978) and vaso-vagal syncope (Graham et al., 1961) in which parasympathetic rebounds seem to follow intense sympathetic activations.

Ishii et al. (1987) and Igarashi et al. (1987) have studied heart rate variability in motion sick squirrel monkeys and based on their findings, they have supported a model in which an increase in parasympathetic activity leads to the vomiting of motion sickness. As will be discussed in Section 2.3.2, the variability in instantaneous heart rate is due in large part to the control actions of the autonomic nervous system. Ishii et al. (1987) monitored changes in the coefficient of variation (CV) of intervals between heart beats (RR intervals) defined as

$$CV = \frac{\text{Standard Deviation of RR Interval}}{\text{Mean RR Interval}} \times 100\%.$$

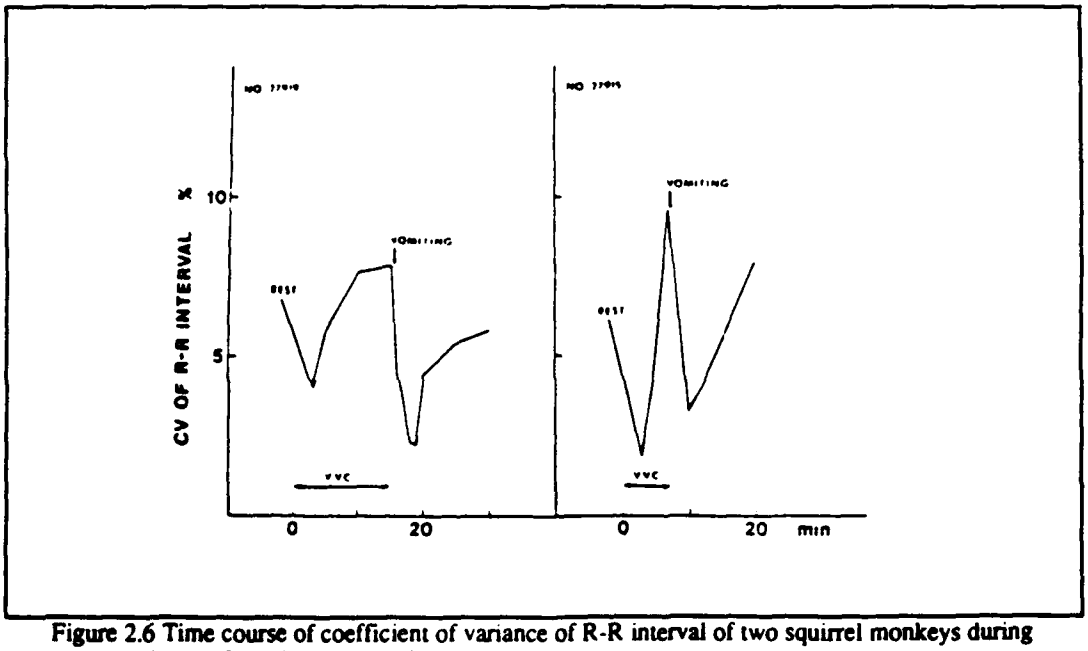


Figure 2.6 Time course of coefficient of variance of R-R interval of two squirrel monkeys during motion sickness (from Ishii et al., 1987)

An increase in the CV indicates an increase in heart rate variability which Ishii et al.interpreted as indicative of increased parasympathetic activity in monkeys. They reported consistent trends in the coefficient (Figure 2.6) throughout the course of the experiment. As vomiting became imminent, marked increases in the coefficient were noted and immediately following vomiting the coefficient decreased. Therefore, the data of Ishii et al, appear to support a model of motion sickness in which "reactions involved with vomiting" are parasympathetically mediated. However, Ishii et al. failed to control or monitor respiratory rates during their experiments. The CV is expected to be very sensitive to changes in respiration. If the monkeys were increasing their respiratory rates (ie. panting) as they became sick, the coefficient of variation might show increases which could be misinterpreted as parasympathetic increases. This issue will be discussed further in Chapter 4.

In proposing models for ANS activity, researchers have met a number of obstacles and it is evident that despite significant research, the role of the autonomic nervous system in motion sickness remains a speculative issue.

2.3 Transfer Function Estimation: A Probe to Autonomic Function

Cohen and colleagues have developed techniques to noninvasively assess autonomic activity (reviewed in Appel et al., 1989a). By applying linear system theory to analyze variability in cardiovascular parameters, they have shown that the characteristics of this variability provide information about relative levels of sympathetic and parasympathetic activity. Further, they have demonstrated that by perturbing parts of the cardiovascular control system in a known fashion, one can extend the amount of information attained in a

single experiment trial. The techniques are based on the theory of non-parametric transfer function estimation.

2.3.1 Transfer Function Estimation

Transfer function estimation is a linear system identification technique based on a Wiener filtering approach. Given a record of the input and output of a linear system, the Wiener filter is one which operates on the input and produces a minimum mean square error (MMSE) approximation to the output (Papoulis, 1984), under certain conditions on the random processes involved.

In discrete time (or for samples from a band-limited continuous time system), the convolution relation for a linear system is

$$y[n] = h[n] \otimes x[n] \equiv \sum_{m=-\infty}^{\infty} h[n]x[n-m] \tag{2.1}$$

where x[n] and y[n] are the input and output respectively, h[n] is the unit sample response of the system and $\otimes$ is the convolution operator defined by equation 2.1. In the time domain, the system identification problem is that of estimating h[n] given x[n] and y[n]. If the estimate of h[n] is $\hat{h}[n]$, the output, $\hat{y}[n]$, from the estimated system is

$$\hat{y}[n] = \hat{h}[n] \otimes x[n] \equiv \sum_{m=-\infty}^{\infty} \hat{h}[n]x[n-m] \tag{2.2}$$

The Wiener filter is that $\hat{h}[n]$ which minimizes the mean square error in the output given by

$$MSE = E\{(y[n] - \hat{y}[n])^2\} \tag{2.3}$$

where E{ } is the expectation operator (Papoulis, 1984; Bendat and Piersol, 1980, 1986).

The Orthogonality Principle and Projection Theorem (Papoulis, 1984) guarantee that MSE will be minimum if the error is orthogonal to the data. That is, for real signals,

$$E\{(y[n] - \hat{y}[n])\ x[n-m]\} = 0; \quad \text{all m} \tag{2.4}$$

or by expanding the sum

$$E\{(y[n]\ x[n-m] - \hat{y}[n]\ x[n-m]\} = 0; \quad \text{all m} \tag{2.5}$$

and substituting for $\hat{y}[n]$

$$E\{(y[n]\ x[n-m]\} - \sum_{p=-\infty}^{\infty} \hat{h}[p]\ E\{x[n-m]x[n-p]\} = 0; \quad \text{all m} \tag{2.6}$$

If $x[n]$ and $y[n]$ are stationary, (2.6) can be written

$$\mathbf{R}_{xy}[m] = \sum_{p=-\infty}^{\infty} \hat{h}[p]\ \mathbf{R}_{xx}[m-p] \quad ; \quad \text{all m} \tag{2.7}$$

where $$\mathbf{R}_{xy}[m] = E\{(y[n]\ x[n-m]\} \tag{2.7a}$$

and $$\mathbf{R}_{xx}[m] = E\{(x[n]\ x[n-m]\}. \tag{2.7b}$$

$\mathbf{R}_{xx}[m]$ is the autocorrelation function of the input and $\mathbf{R}_{xy}[m]$ is the cross-correlation between the input and output, at lag m.

Equation (2.7) defines the time domain constraints on the optimum unit sample response estimate. By transforming (2.7) to the frequency domain, however, a more useful formulation is attained. The power spectrum, $S_{xx}[f]$, of the input and the cross-spectrum, $S_{xy}[f]$, between the input and output, from the Weiner-Khinchin Theorem (Papoulis, 1984) are

$$S_{xx}[f] = T_s\ \text{DTFT}\{\mathbf{R}_{xx}[m]\} \tag{2.8a}$$

$$S_{xy}[f] = T_s\ \text{DTFT}\{\mathbf{R}_{xy}[m]\} \tag{2.8b}$$

where DTFT{} is the Discrete Time Fourier Transform operator and T_S is the sampling period of the discrete time signals.

Convolution in the time domain is equivalent to multiplication in the frequency domain, so (2.7) becomes

$$S_{xy}[f] = \hat{H}[f]\, S_{xx}[f] \tag{2.9}$$

and rearranging gives

$$\hat{H}[f] = \frac{S_{xy}[f]}{S_{xx}[f]} \tag{2.10}$$

where $\hat{H}[f]$ is the transfer function of the optimum system estimate.

The transfer function defined in (2.10) provides information about both magnitude and phase of the unknown system. Furthermore, a function termed coherence permits an assessment of the quality of the provided information. The coherence function, γ^2, is derived as follows. The power spectrum of the output from a system with transfer function $\hat{H}[f]$ given input $x[n]$ is

$$S_{\hat{y}\hat{y}}[f] = |\hat{H}[f]|^2 S_{xx}[f] = \left|\frac{S_{xy}[f]}{S_{xx}[f]}\right|^2 S_{xx}[f] \tag{2.11}$$

The fraction of the power in $S_{yy}[f]$, the power spectrum of the true system output, that is accounted for by $S_{\hat{y}\hat{y}}[f]$ as a function of frequency is the coherence,

$$\gamma^2 = \frac{S_{\hat{y}\hat{y}}[f]}{S_{yy}[f]} = \frac{|S_{xy}[f]|^2}{S_{xx}[f]\; S_{yy}[f]} \tag{2.12}$$

When the coherence is near unity, nearly all power in $y[n]$ is reproduced in $\hat{y}[n]$; as the coherence nears zero, very little power is retained in the output from the estimated system.

Equations (2.10) and (2.12) define the optimal transfer function estimate and its coherence function (Papoulis, 1984; Bendat and Piersol, 1980, 1986; Kay, 1988). In practice, infinite duration input-output records for a system are unavailable and estimates of $\hat{H}[f]$ and γ^2 must be made from finite duration records. Further, if the input signal does not contain significant power in a particular frequency band, the signal to noise ratio is low and the coherence of the estimated transfer function will typically fall (Papoulis, 1984; Bendat and Piersol, 1986). Therefore, for accurate transfer function estimation, the spectrum of the input signal, $S_{xx}[f]$, must be significantly non-zero over the frequency band of interest.

2.3.2 Cardiovascular Control

The cardiovascular system consists of two general component parts, the heart and the blood vessels. The system is responsible for the delivery of blood to all parts of the body. Since blood delivers nutrients to and removes wastes from bodily tissues, it is critical to provide sufficient perfusion to meet the tissues' metabolic needs. It is, however, inefficient to over-perfuse tissues during times of low metabolic need. Long range regulation of tissue perfusion inherently depends on other systems, such as the renal system, which regulate fluid balance and blood volume (Berne and Levy, 1982). Shorter term regulation and precision control, however, depend on variations in the performance of the two cardiovascular system components. Heart rate and contractility and vascular resistance and volume are continuously controlled to individually accommodate the varying metabolic needs of tissues.

There are a number of factors which contribute to the function of the heart and blood vessels (Berne and Levy, 1982; Tortora and Evans, 1986). For example, excess metabolites accumulated locally in tissues may affect vasodilation and blood bourne

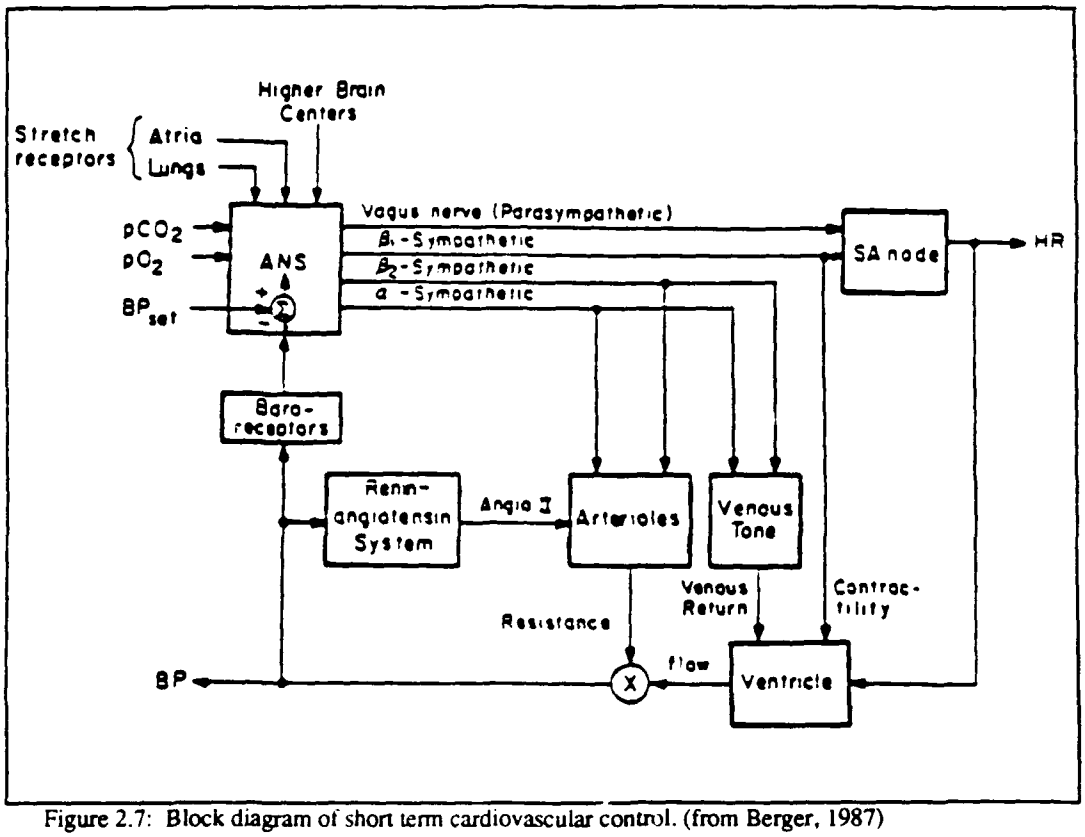


Figure 2.7: Block diagram of short term cardiovascular control. (from Berger, 1987)

hormones, ions or toxins can affect both vascular tone and cardiac function (Berne and Levy, 1982). Short term control on the order of seconds or minutes, however, is mediated primarily through neural mechanisms (Guyton, 1974; Berne and Levy, 1982). These mechanisms are precisely controlled through feedback loops involving the autonomic nervous system.

Figure 2.7 depicts a block diagram model of the autonomic components of short term cardiovascular control (from Berger, 1987). As metabolic needs in tissues such as skeletal muscle or digestive tract linings change, the system acts to tightly control blood pressure in critical bodily tissues while accommodating the new perfusion requirements. Baroreceptors located in the carotid sinus and the aortic arch constantly sense arterial blood

pressure and feed this information back to the ANS. In addition to the baroreceptor feedback, the ANS receives feedback regarding atrial and chest wall stretch and arterial partial pressures of oxygen and carbon dioxide. (Berger, 1987; Berne and Levy, 1982) The system is also affected by higher brain centers such as the respiratory center. Both branches of the ANS exert control over heart rate through innervation of the sinoatrial node. The sympathetic system also controls heart contractility (type β_1 fibers) and vascular parameters. Sympathetic type β_1 and α_2 fibers effect contraction of the smooth muscles in vein walls and thereby control venous return. The fibers also innervate sphincter muscles in arterioles and thus effect peripheral resistance.

As the system works to continuously meet bodily needs, its parameters continually vary. The variability in instantaneous heart rate or blood pressure is generally ignored by physicians who are most interested in the averages of these parameters over short time periods. One exception is the variability due to respiration. It has long been known that respiration modulates heart rate. (Hirsch and Bishop, 1981) Heart rate typically increases during inspiration and decreases during expiration. Kollai and Kazumi (1979), through direct neural recordings in dogs, have demonstrated that this modulation, termed Respiratory Sinus Arrhythmia (RSA), is mediated by the autonomic nervous system. In the model of Figure 2.7, it is the input from the respiratory center and feedback through arterial and chest wall receptors and the baroreceptors which most likely stimulate the autonomic modulation.

2.3.3 Transfer Function Estimation of Cardiovascular Control

Figure 2.8 (from Berger et al., 1988) presents a simplified and semi-quantitative model of cardiovascular control which relates three noninvasively measurable parameters:

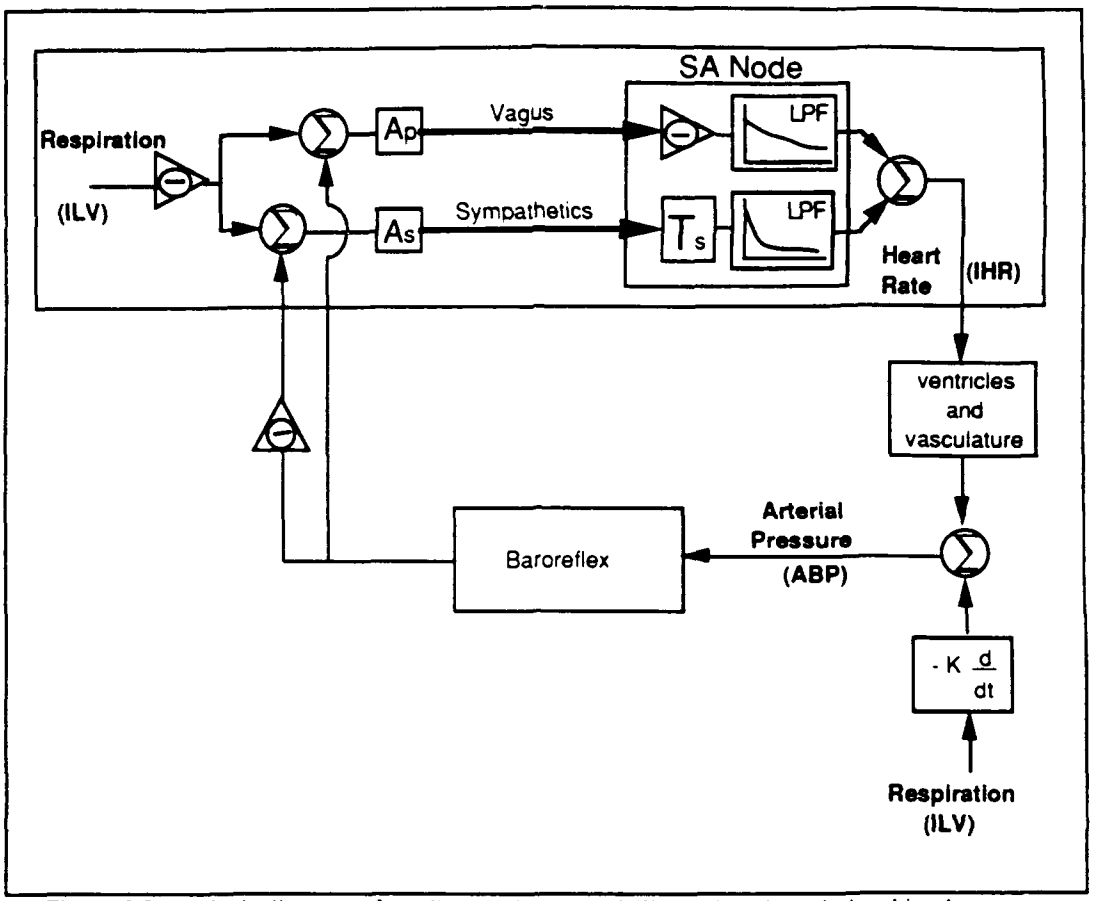


Figure 2.8: Block diagram of cardiovascular control illustrating the relationships between respiration, arterial blood pressure and heart rate. (modified from Berger et al., 1988)

respiration, heart rate and blood pressure. The blocks in this model mask some of the details of the model in Figure 2.7. Each block is representative of a lumped parameter system which relates its input and output signals. In large part, these systems are autonomically mediated. A number of recent studies have been conducted in attempts to characterize the signals or systems depicted in this model.

In 1981, Akselrod et al. (1981) monitored heart rate in trained, conscious, unanesthetized dogs. Using power spectral estimation of instantaneous heart rate prior to and during pharmacological blockade they identified frequency specific contributions of the parasympathetic and sympathetic divisions of the ANS. They found that during

parasympathetic blockade, high frequency components of variability, above approximately 0.1 Hz, were abolished. Upon coincident sympathetic blockade, nearly all variability in heart rate was abolished. Based on their data, they concluded that both divisions contribute to spontaneous heart rate variability at low frequencies but at higher frequencies only the parasympathetic system contributes.

In 1985, Pomeranz et al. (1985) and Akselrod et al. (1985) repeated similar studies in humans. Their results were similar. They reported that low frequency fluctuations in spontaneous heart rate between 0.02 and 0.1 Hz were mediated jointly by both ANS divisions, but that at higher frequencies (in the range 0.1-0.5 Hz) only the parasympathetic

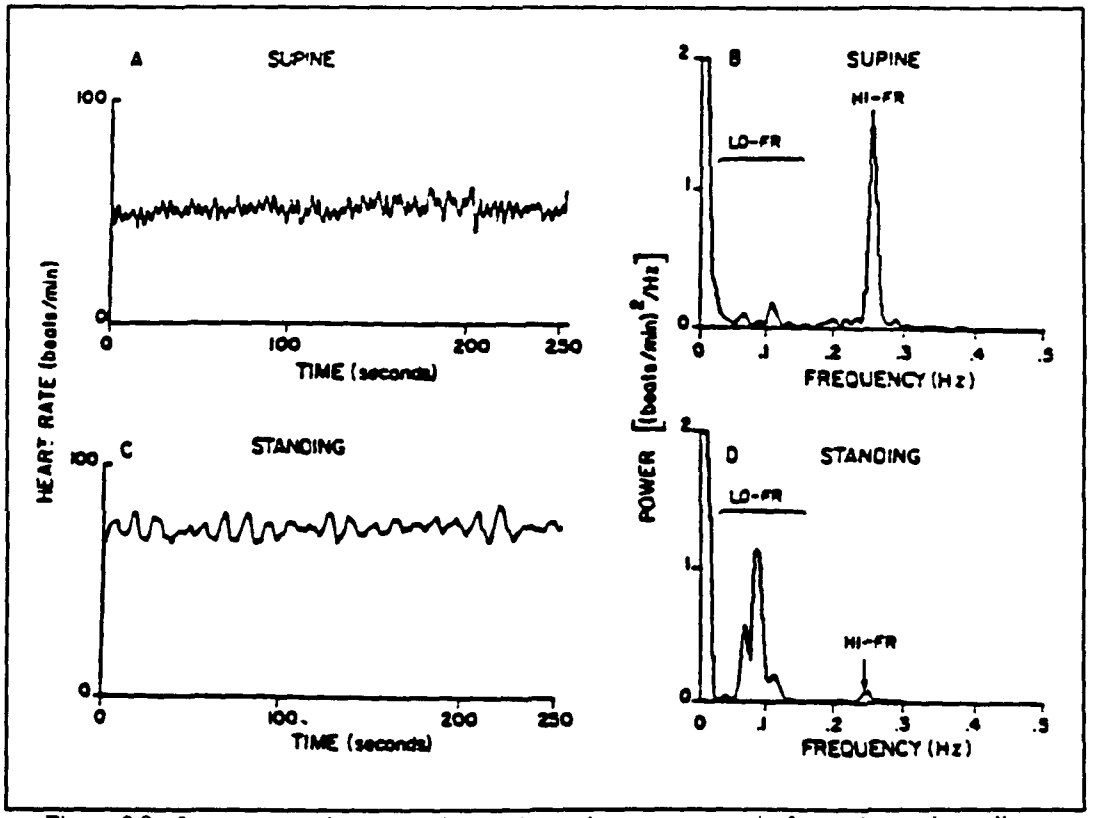


Figure 2.9: Instantaneous heart rate time series and power spectra in for supine and standing subject. Note shift from high frequency to low frequency oscillations in standing subject. (from Pomeranz et al., 1985)

division contributes. The Pomeranz study also demonstrated the sensitivity of the power spectral estimates to posture changes. Individuals while standing tend to have higher sympathetic levels and lower parasympathetic levels than while supine (Eckberg et al., 1976; Burke et al., 1977). This change occurs in large part to counteract the force of gravity which acts to impede venous return in erect individuals (and perhaps also in part to accommodate a generally less relaxed state). Pomeranz et al. demonstrated the expected decrease in magnitude of the high frequency peak and corresponding increase in magnitude of the low frequency peak upon transition from supine to standing (Figure 2.9).

Each of the aforementioned studies was designed to characterize the spontaneous heart rate signal. Akselrod et al. (1985) also characterized the spontaneous blood pressure signal. The spontaneous respiratory signal is generally a very narrow-band signal with power tightly contained near the mean respiratory rate. In addition to characterizing the spontaneous activity of the signals in the above model, researchers have attempted to characterize the relationships between the signals represented by the system blocks (Berger et al. 1986; Berger et al. 1988; Appel et al., 1989a; Appel et al., 1989b). The respiration to heart rate transfer block is of particular interest.

Using transfer function analysis, as outlined in Section 2.3.1, Chen et al. (1987), Saul at al. (1989), Berger et al. (1989b), and Berger (1987) have studied the relationship between respiration and heart rate. Their studies have built on the results of Pomeranz et al. (1985) by characterizing the instantaneous lung volume (ILV) to instantaneous heart rate (IHR) transfer function in supine and standing subjects. As discussed in Section, 2.3.1, in order to extend the amount of useful information attained through transfer function estimation the system should be excited by an input with broad spectral content. Since the respiratory input is generally quite narrow band with power tightly contained around the mean

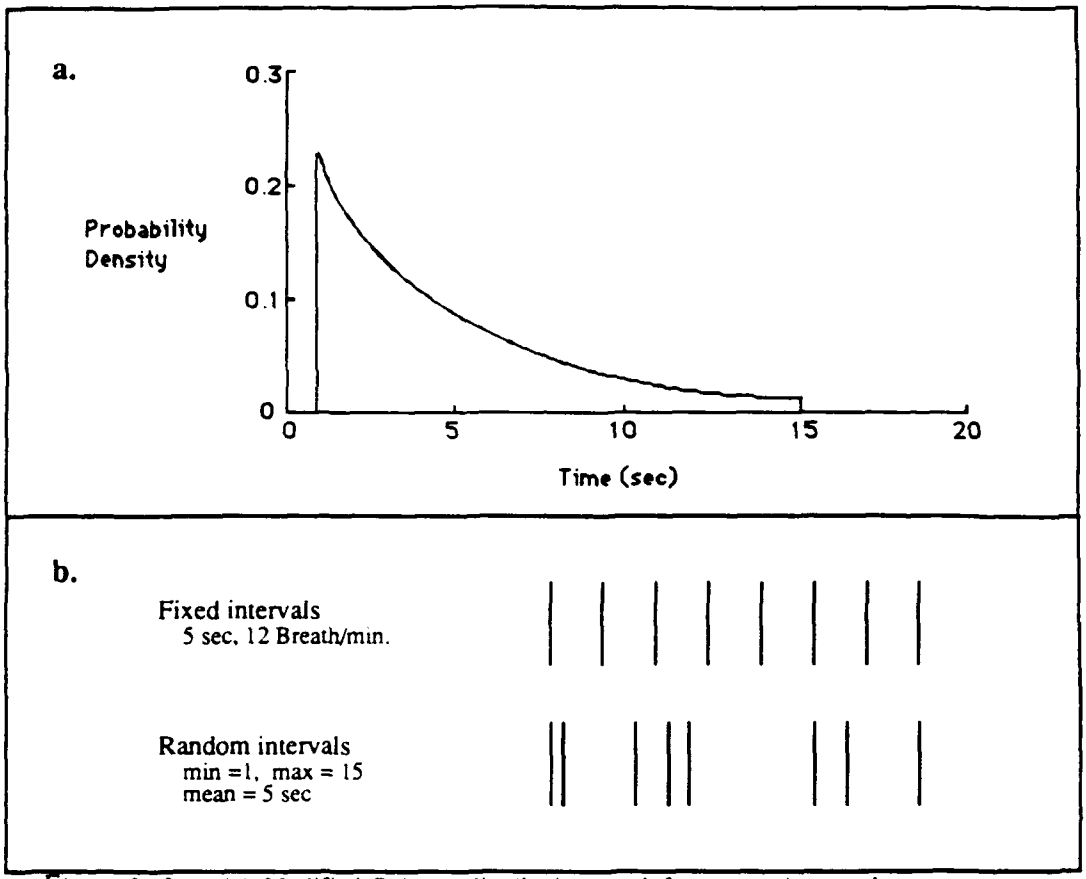


Figure 2.10: (a) Modified Poisson distribution used for generating random cue sequence. (b) Schematic diagram of fixed interval breathing and random interval breathing sequences

respiratory rate, a technique to broaden its spectral content was applied. The technique is termed Random Interval Breathing (RIB) and is described by Berger et al. (1989b). Under this paradigm, subjects are first asked to breathe in synchrony with a two minute segment of auditory tones equally spaced at the mean respiratory rate. During this time they are able to settle on a comfortable depth of inspiration for the given mean rate. After this constant interval segment, a segment of randomly spaced auditory tones is provided and subjects continue to breathe with each tone. The tones are generated by computer and occur at random according to a modified Poisson distribution with the same mean rate as the first segment (Berger et al., 1989b). (refer to Figure 2.10) This modified Poisson distribution disallows inter-breath intervals outside the range of one to fifteen seconds. This

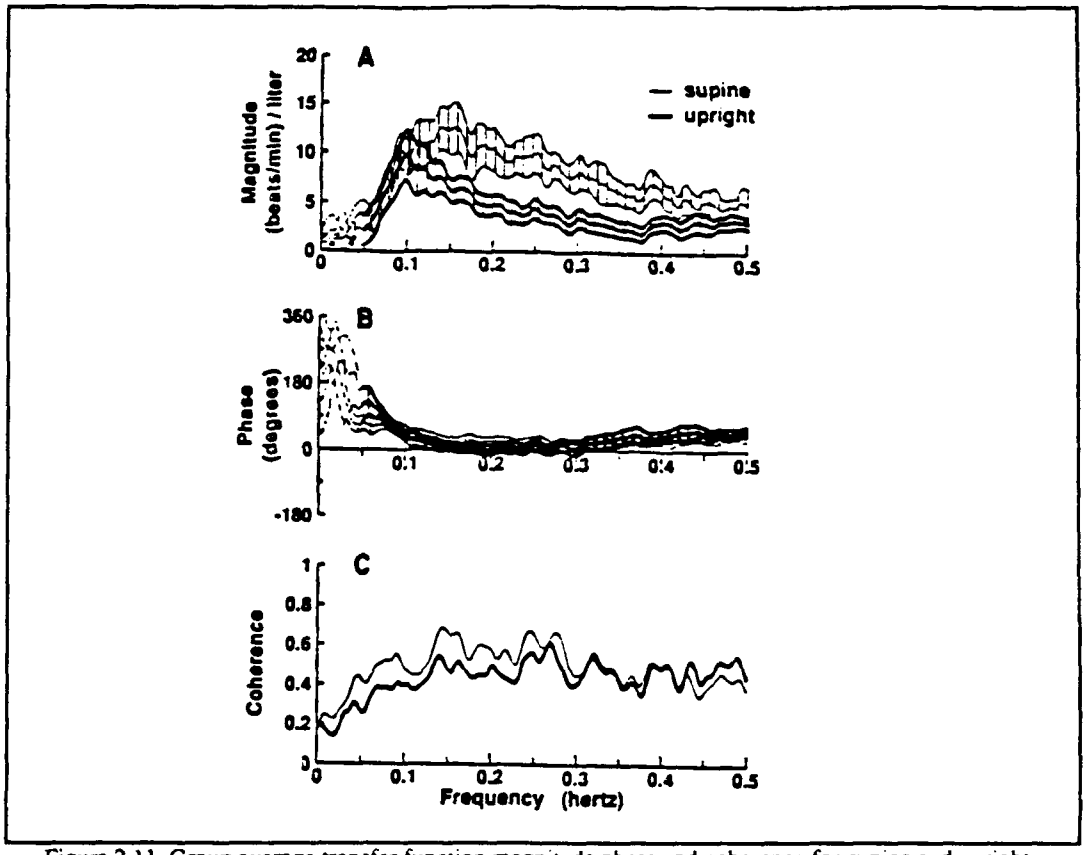


Figure 2.11 Group average transfer function magnitude phase and coherence for supine and upright postures (means ± standard errors). Below 0.05 Hz, average transfer function values are plotted with dotted lines because they are unreliable in that range. (from Saul et al.,1989)

distribution provides for a comfortable breathing pattern while effectively broadening the respiratory signal over the frequency band between 0.0 and 0.5 Hz. Each of the three studies found that group average transfer functions for standing subjects had significantly lower transfer magnitude at frequencies above approximately 0.1 Hz (Figure 2.11 from Saul et al., 1989). Thus, these studies demonstrated the sensitivity of the transfer function magnitude to shifts in autonomic balance as subtle as those associated with posture change (Chen et al., 1987).

Further studies have confirmed the sensitivity and accuracy of transfer function estimates obtained by this technique. Selective pharmacological blockades in conjunction with

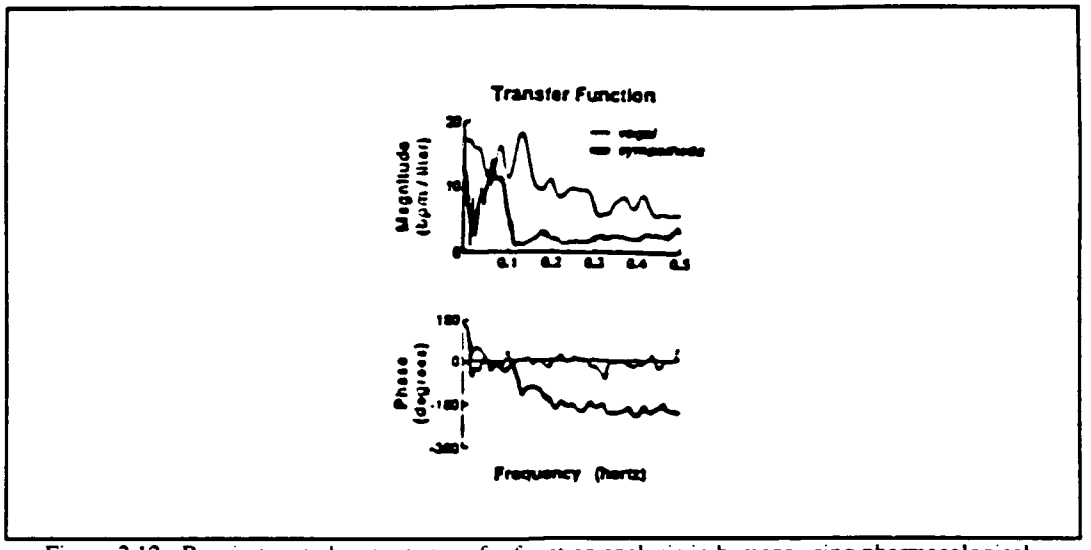


Figure 2.12: Respiration to heart rate transfer function analysis in humans using pharmacological blockades. Transfer function magnitudes and phases for 'vagal' and 'sympathetic' states. (modified from Appel et al., 1989a)

changes in posture have been used to obtain transfer function estimates from subjects in "pure" parasympathetic or "pure" sympathetic states. (Saul et al., 1988; Appel et al, 1989a) Parasympathetic states were achieved in supine human subjects with propanolol blockade of sympathetic activity. Sympathetic states were achieved in standing subjects with atropine blockade of parasympathetic activity. Figure 2.12 (from Appel et al., 1989a) illustrates the differences in transfer functions from these two states. In a "pure" sympathetic state, the system is characterized by a dramatic decrease in magnitude above 0.1 Hz and by a decreasing phase (time delay). The "pure" parasympathetic state is characterized by higher transfer gain and a near zero phase at all frequencies in the band between 0.0 and 0.5 Hz.

In anesthetized dogs, Berger et al. have investigated the transfer characteristics of the sinoatrial node (Berger 1987; Berger et al, 1989b) by applying broad band stimulation directly to either parasympathetic or sympathetic efferents and monitoring the resulting heart rate fluctuations. They report low pass filter characteristics for both parasympathetic

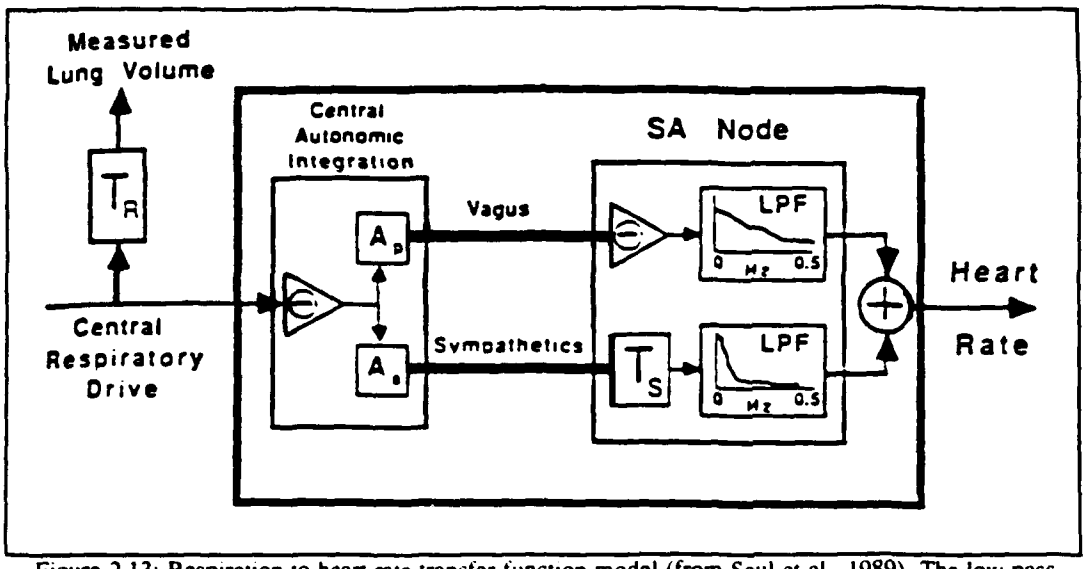


Figure 2.13: Respiration to heart rate transfer function model (from Saul et al., 1989) The low pass filter (LPF) characteristics of the sympathetic and parasympathetic pathways in the SA node are dependent on the mean neural firing rates of each division.

and sympathetic stimulation but found the sympathetic transfer function to have a much lower cutoff frequency and a pure time delay. The transfer characteristics, however, were found to depend significantly on mean sympathetic and parasympathetic firing rates. At low mean firing rates, the transfer functions typically had larger low frequency magnitudes and showed more rapid roll off than at high mean rates.

Saul et al. (1989) have incorporated these findings and those outlined above into a model relating respiration to heart rate. In their model (Figure 2.13 from Saul et al., 1989), a central respiratory drive is fed to a Central Autonomic Integrator to produce the sympathetic and parasympathetic (vagal) outflow to the sinoatrial node. This Central Integrator incorporates the many bodily reflexes and pathways through which respiration may effect autonomic outflow to the heart. The constants A_P and A_S represent the transfer relation between the respiratory drive and the modulation depth of the parasympathetic and sympathetic outflow respectively. At the sinoatrial (SA) node, the vagal outflow is inverted

(since increased vagal activity decreases heart rate) and passed through a characteristic low pass filter, the shape of which is dependent on mean parasympathetic firing rate. The sympathetic outflow is delayed by T_S seconds* and passed through its characteristic low pass filter whose shape is dependent on mean sympathetic firing rate. These two filtered signals sum to give instantaneous heart rate.

Four key model parameters, A_p, A_S, mean vagal firing rate and mean sympathetic firing rate, may be varied to independently manifest changes in the transfer function due to shifts in autonomic balance. (T_S and T_R are taken as constants.) In simulations, Saul et al. used simple one-pole low pass filters for each of the low pass filter (LPF) blocks at the sinoatrial node. In simulation of the supine condition, Saul et al. chose a mean vagal rate of 4 Hz and a mean sympathetic rate of 0.5 Hz. Using these mean rates and the experimental data of Berger et al. (1989b) they chose cutoff and gain factors of the two low pass filters. The weighting factors A_p and A_S were set to 2.5 and 0.4 Hz/liter, respectively. In simulation of the standing condition, mean vagal and sympathetic rates were each chosen as 1 Hz and again based on experimental data from Berger et al., two new low pass filter gains and cutoffs were selected. A_p and A_S were set to 1.6 and 2.0 Hz/liter respectively. Saul et al. note that while the model parameter choices are somewhat arbitrary, they are consistent with a shift from a generally vagal state when supine to a more sympathetic state when upright. Simulated transfer functions for supine and standing conditions are illustrated in Figure 2.14. Except at frequencies below 0.05 Hz, the simulation matches well the experimental data (compare to Figure 2.11). However, at these low frequencies the coherence of the experimental transfer function estimate is quite low and therefore confidence in its accuracy is low. Saul et al. explain this drop in coherence by postulating

* The lag between stimulation of sympathetic nerves at the sinoatrial node and the resulting change in heart rate is well recognized (Warner and Cox, 1962; Scher et al., 1972). It has been attributed to the slow diffusion rates of norepinephrine (Warner and Cox, 1962).

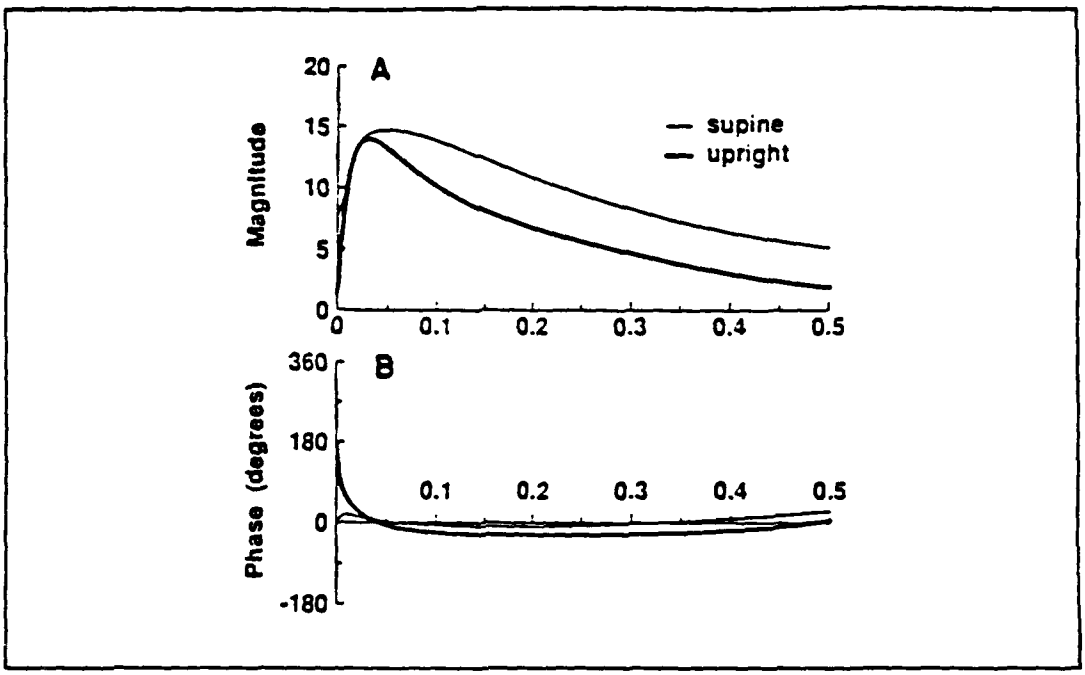


Figure 2.14: Simulations of transfer function magnitudes (a) and phases (b) for supine and upright postures.

that at these low frequencies either significant nonlinearities in the transfer relation or dominant effects on heart rate from other system inputs exist.

It is clear that estimates of the transfer function from respiration to heart rate demonstrate characteristic changes as autonomic control balance is altered and a successful model to qualitatively describe the alterations of the local control system has been developed. Furthermore, the model suggests that knowledge of the respiratory drive to the system is important when interpreting the resulting heart rate variability. Unknown changes in the respiratory drive to the system could easily be misinterpreted as changes in the parameters of the system representing autonomic balance. Therefore, to assess autonomic influences on heart rate variability, transfer function estimation is a desirable technique.

There is, however, some controversy over the issue of whether local autonomic control of heart rate may be interpreted as representative of more generalized autonomic activity. As indicated in Section 2.2.1, the function of the autonomic nervous system is to control subconscious functions in response to changing bodily needs. By the nature of their affects on particular organ systems, it is clear that in many instances (such as during bodily stress), generalized increases in the activity of one division with respect to the other are warranted. However, in other instances (especially thermoregulation) it is clear that a more localized control is desirable. The issue of when local activity can be expected to mirror that of other autonomic activity remains a speculative one.

Many animal studies have been conducted in which various sympathetic neural signals are simultaneously monitored (Cohen et al, 1970; Ninomiya et al., 1976, Simon and Riedel, 1975; Karim et al., 1972;, Victor et al., 1989). Due in part to the diversity of experimental conditions, the results of these studies do not provide a consistent representation of the extent of differentiation in sympathetic control. Many researchers report significant correlations between different sympathetic nerves, while others report more localized activity.

Recent studies in humans have tended to focus on recording sympathetic activity from either muscles or skin. (Visceral organs and parasympathetic nerves are less accessible.) Wallin summarizes the results well (Wallin, 1986). In general, skin and muscle sympathetic nerve activity have not been found to correlate well under normal conditions. For example, during a Valsalva maneuver, muscle sympathetic activity increases while no change is found in skin. Body cooling, sudden deep inspirations, and loud noises typically cause increases in skin sympathetic activity but no change in muscle. Interestingly, in resting subjects, muscle sympathetic activity shows activity synchronized to the cardiac

cycle while skin activity does not. However, when baroreceptor afferent activity is blocked (ie. by local anaesthesia of vagus and glossopharyngeal nerves), the cardiac rhythm in muscle sympathetic activity disappears and it more closely resembles that in skin. The diversity in response seen between skin and muscle is not generally seen between two different muscles. In most cases, activity recorded from different muscles is quite similar. These findings are best explained by considering the primary functions of the different sympathetic outflows. Wallin (1986) points out that skin sympathetic activity is primarily involved in thermoregulation, while muscle sympathetic activity exerts control primarily over blood pressure.

Based on these admittedly limited findings, Wallin (1986) proposes a simple functional model for sympathetic outflow which incorporates a number of central control pathways influencing distinct groups of effector organs. Some of these pathways may be influenced by particular feedback loops, such as baroreceptor feedback, while others may not. He proposes that the concept of "autonomic tone" is applicable only to these individual groups and not to the overall system during typical bodily activities. (Wallin does not discuss large scale "stress" response characteristics, but his model does not exclude the idea of global changes in sympathetic activity being associated with such instances.) It follows from the data and from Wallin's model that it may be reasonable to assume associated sympathetic activity in different autonomic effectors, if they contribute to similar control functions. For example, muscle sympathetic activity may be expected to correlate with cardiac sympathetic activity.

A recent study investigates the relationship between heart rate variability and sympathetic muscle activity. Saul et al. (1990) monitored muscle sympathetic nerve activity (peroneal nerve) and electrocardiogram during pharmacologically induced stepwise increases or

decreases in diastolic blood pressure. Muscle sympathetic activity decreased with each increase in pressure and increased with each decrease in pressure. Power spectra of instantaneous heart rate were calculated at each pressure change. During decreased blood pressures, the fraction of power in low frequency fluctuations (0.01-0.15 Hz) was loosely correlated with muscle sympathetic activity. However, at increased pressures no measures of low frequency variability correlated with muscle activity. These findings are in accord with models of heart rate control in which both sympathetic and parasympathetic divisions influence low frequency heart rate variations. When blood pressure is decreased, parasympathetic influence on cardiac rate is probably reduced and thus low frequency variations in rate are due mostly to sympathetic activity. Under these conditions correlations to muscle sympathetic activity would be more likely.

The transfer function from ILV to IHR has been demonstrated to be sensitive to shifts in the relative levels of autonomic activity involved in the control of cardiac rate. Extrapolating from this information concerning local activity to make assertions about more global autonomic activity involves some speculation regarding degrees of dissociation of ANS activity.

III Experiment Design

3.1 Experiment Design Issues

In designing a motion sickness experiment in which respiration to heart rate transfer function estimation is to be applied to assess autonomic activity, a number of special design issues arise.

A general paradigm for such an investigation must involve subjects random interval breathing during periods of typical health and during periods of motion sickness. It has previously been demonstrated that cooperative subjects are able to random interval breath while healthy. However, motion sick subjects had never been asked to do so. Therefore, the first design issue was that of determining how well subjects were able to follow the random breathing cues while motion sick.

A second design requirement was control of possibly confounding autonomic effects. Activities which could effect autonomic outflow independently of motion sickness had to be limited when possible. Since exercise and posture changes are known to alter autonomic activity, it was desirable that subjects in this experiment limit their endogenous movements and remain in one posture throughout the experiment. A technique for inducing motion sickness under these limiting conditions had to be developed.

A third design concern was the requirement of maintaining a relatively constant level of sickness throughout a random interval breathing segment. In order to obtain accurate transfer function estimates, the system being observed must remain the same over the estimation period (or more precisely the process must be stationary). Since random interval breathing segments are typically 15 minutes in duration, a relatively constant level of motion sickness had to be maintained for this duration.

A series of pilot experiments were conducted to test the feasibility of a proposed experimental protocol designed to meet these requirements.

3.2 Experiment Apparatus

Two key pieces of equipment, a rotating chair and a set of reversing prism goggles, were employed in pilot experiments and subsequently in the primary motion sickness experiments.

3.2.1 Rotating Chair Assembly

A Barany type rotating chair was used. The complete assembly is depicted in Figure 3.1. The chair is computer controlled and is capable of sinusoidal, trapezoidal or constant

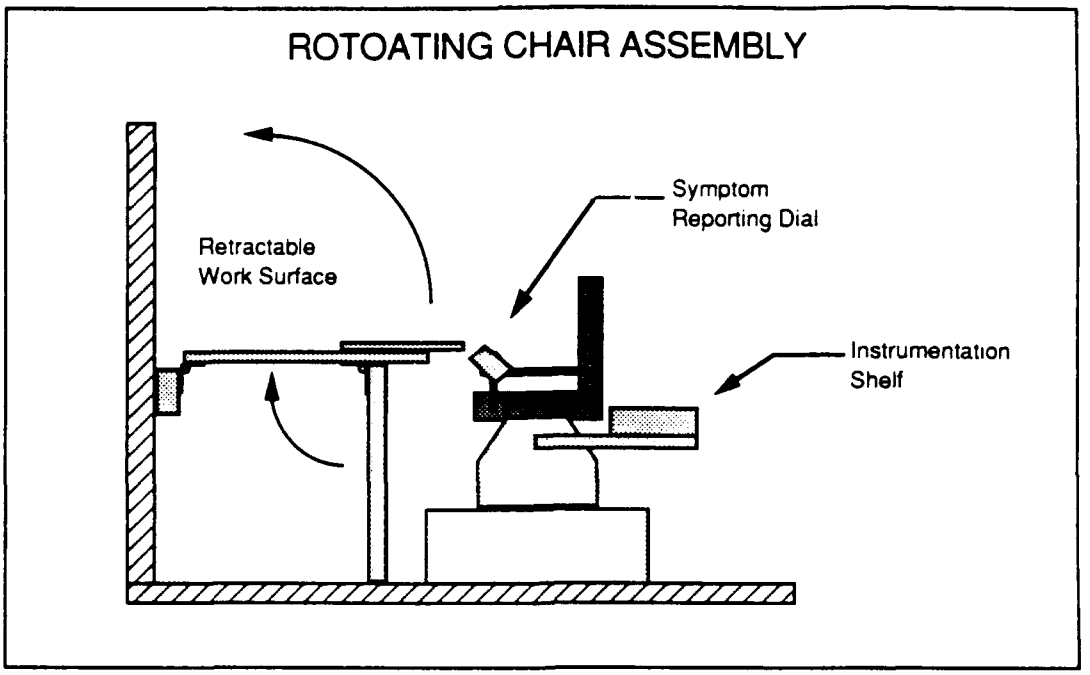


Figure 3.1: Rotating chair assembly illustrating illustrating retractable work surface, remote symptom reporting dial and instrumentation shelf.

velocity rotation about an earth vertical axis (Tole et al, 1981). A work surface folds down from an adjacent wall to provide a self supported, approximately two foot by four foot work area which is used by the subject during provocative tasking. An instrumentation shelf, mounted behind the subject's seat, houses necessary instrumentation and thus allows for amplification of signals prior to their passing through the chair's slip rings. (This is advantageous in increasing signal to noise ratios.) A remote symptom reporting dial was mounted on either the left or right armrest, as preferred by a particular subject. It consisted simply of a linearly stepped potentiometer which provided a voltage output proportional to dial position. The signal was passed through slip rings to a meter observable by the experimenter. The system provided for nonverbal symptom reporting. Details of symptom reporting are discussed in Section 4.1.3.

3.2.2 Reversing Prism Goggles

A pair of reversing prism goggles provided the "sensory rearrangement" necessary for inducing motion sickness. Employing two pairs of dove prisms (Figure 3.2a.), one pair in front of each eye, the goggles act to reverse the subject's left-right visual field. That is, objects to the right of a subject appear to the left and vice versa. Furthermore, when the subject's head moves rightward, the visual field is perceived to also move rightward. This is in direct opposition to the normal situation in which rightward head movements result in the visual scene passing leftward across the field. The goggles used in these experiments were the same as those used by Eagon (1988). A schematic of the goggles is provided in Figure 3.2b.

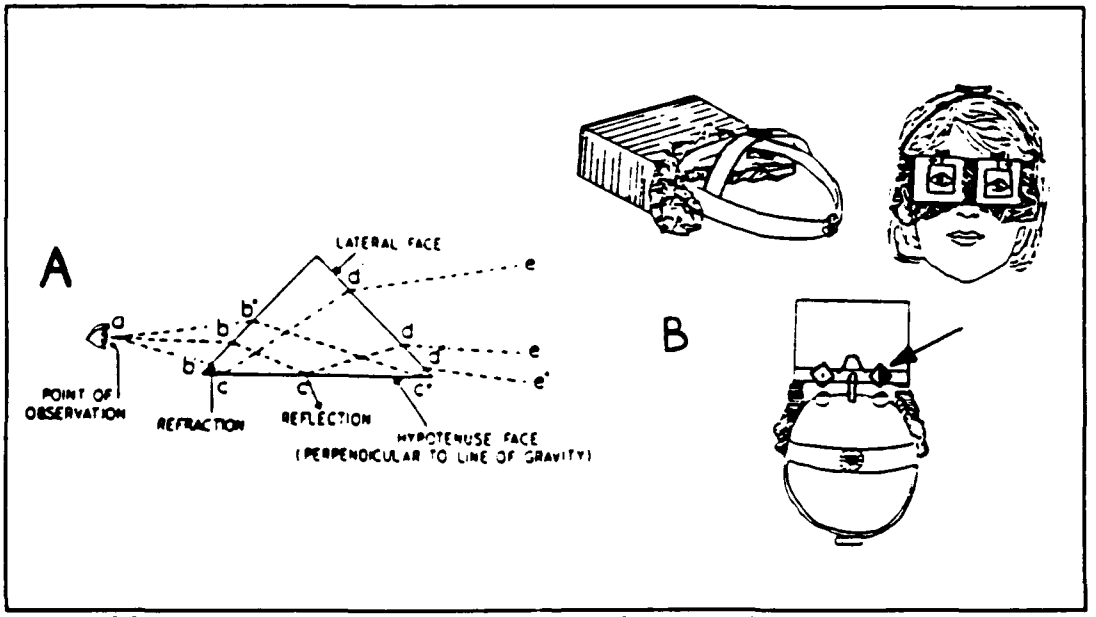


Figure 3.2: (a) Demonstration of the optical properties of the dove prism used in the reversing prism goggles (top view of one-half of the optic set used for each eye) and (b) Schematic diagram of the prism goggle headset (single prism indicated with arrow)

3.3 Pilot Experiments

Four subjects participated in various phases of pilot experimentation. They will be referred to by the letter codes A through D. Subjects A, B, and C were males aged 23, 30, and 45 years, respectively. Each had previous experience as a subject in motion sickness studies and in recognizing and reporting motion sickness symptoms. Subject D was female, aged 22, and had no previous experience as a subject in motion sickness experiments.

In a first pilot experiment, Subject C participated in an approximately 1 hour session designed to assess his ability to follow random breathing cues while motion sick. While wearing reversing prism goggles, the subject undertook coordinated tasks such as writing, drawing, and stacking cans until moderate motion sickness was attained. Upon maintaining these moderate symptom levels for ten minutes, the subject then undertook a fifteen minute segment of random interval breathing conducted in the same manner as described by Berger et al. (1989) (Section 2.3.3). During the breathing segment, the subject periodically continued tasking in order to maintain a relatively constant symptom level. The subject demonstrated an ability to follow the random breathing sequence while motion sick, however, completion of the manual tasks interfered with his ability to do so.

It was therefore decided that manual tasking was not a feasible means for maintaining symptoms during the random interval breathing segments. As an alternative, it was proposed that subjects undertake sinusoidal rotation (about an earth vertical axis) in the Barany chair while wearing the reversing prism goggles. When their symptoms reached a moderate level, by closing their eyes they could effectively "turn off" the provocative stimulation (by removing the sensory conflict) and their symptoms were expected to drop. By repeating the process of opening and closing their eyes, they could control their symptoms about a moderate level.

Subjects A, B, and C each participated in the second pilot experiment which was designed to test whether sinusoidal rotation while wearing the prism goggles was a provocative and controllable stimulus. Each subject was asked to wear the reversing prism goggles while rotating sinusoidally in the chair. Rotation velocity and frequency were varied during the experiments in order to assess differences in the provocative nature of the different conditions. Peak velocity ranged between 30 and 120 degrees per second and frequency was either 0.1 or 0.2 Hz. Subjects were asked to report symptoms and comment on the effectiveness of each condition in producing sickness. All subjects found the paradigm provocative and in general, subjects reported that larger peak angular velocities were more provocative than smaller peak velocities within the range tested. No consistent differences were noted between the two tested frequencies. However, the paradigm tended to produce symptoms with a variable latency but consistently rapid onset which made even these experienced subjects anxious about their ability to control their symptoms.

Pilot experiment number three was designed to solve the procedural problems discovered in pilot experiments 1 and 2. A manual tasking paradigm and brief rotation period were used to induce motion sickness symptoms in a more controllable fashion. Rotation with the goggles was employed as a means to maintain the desired symptom levels during random interval breathing. Subjects B, an experienced subject, and Subject D, a naive subject, participated in approximately 1.5 hr sessions. While seated, relaxed, and stationary, each subject undertook a twelve minute segment of random interval breathing. Subjects were then fitted with the prism goggles and undertook provocative tasking including writing, drawing, and stacking cans and a brief period of rotation. Upon reaching moderate levels of motion sickness and maintaining them for ten minutes, subjects were rotated in the chair at a frequency of 0.1 Hz and a peak velocity of 120 degrees/second. They were instructed to close and open their eyes as necessary to control and maintain symptom levels. When

they were comfortable in maintaining their symptoms in this fashion, subjects undertook a twelve minute sequence of random interval breathing. Subjects were monitored as to their ability to maintain symptoms and to follow the breathing sequence. On completion, subjects were asked to comment on the same. Neither subject reported or demonstrated difficulty in following the random breathing cues while not motion sick. Both subjects attained moderate symptom levels within 1 hour of tasking. Subject B demonstrated and reported no difficulty in completing a 12 minute segment of random breathing while rotating and maintaining symptoms. Subject D's symptoms increased slightly during rotation but she reported and demonstrated close attention to following the breathing pattern.

Based on pilot experiments 2 and 3, it was concluded that sinusoidal rotation while wearing the prism goggles was a sufficient stimulus for inducing and maintaining motion sickness during random interval breathing. However, it did not allow for as close control over the onset of symptoms as simple coordinated tasking with goggles did. Therefore, it was desirable to use rotation with goggles for maintaining symptoms once they were first developed but not for original development of symptoms.

The final experiment protocol described in Section 4.1.4 was designed to take advantage of these findings.

IV Methods

This chapter is divided into two major sections. The first section describes the motion sickness experiments, and the second outlines the analysis techniques applied to the data.

4.1 Primary Motion Sickness Experiment

4.1.1 Subjects

Potential subjects participated in a screening interview designed to identify any obvious vestibular, cardiovascular, or gastrointestinal disorders. The interviewers guideline is included in Appendix A. Three potential subjects were rejected on the basis of their responses to the interview. Accepted subjects were briefed on the details of the experiment and were given an information package including a Motion Sickness Questionnaire, Magnitude Estimation Instructions, Symptom Definitions List, and Subject Instruction Sheet. The motion sickness questionnaire was designed to provide greater detail in regard

to the subjects motion sickness history. The Magnitude Estimation Instructions outlined for the subject the technique to be used in symptom reporting. (Details of the technique are provided in Section 4.1.3.) The list of symptom definitions provided standard terminology for the subject to use in reporting symptoms. The Subject Instruction Sheet listed a number of requests for limiting subject activity on the day of the experiment. These instructions served to protect the subjects and to control a number of variables which could confound our measurements and analyses. They are similar to those used in previous studies (Eagon, 1988). Subjects were asked: (1) to eat their normal meal between 12:00 and 12:30 PM on the day of the experiment and to eat nothing thereafter, (2) to consume no medications or alcohol for 24 hours prior to the experiment, (3) to drink no coffee, tea or cola and to not smoke for twelve hours prior to the experiment, (4) to avoid heavy exercise six hours prior to the experiment, and (5) to get a normal night's sleep on the night before the experiment. Copies of all forms are provided in Appendix A.

On the day of the experiment, subjects were asked to arrive at 3:30 PM. They were again briefed on the nature of the experiment and were asked to sign an approved Informed Consent Document (Appendix A). A questionnaire was given to verify the subjects adherence to the instructions and the subject's general health. (Appendix A)

4.1.2 Physiological Recordings

Three physiological signals, electrocardiogram (ECG), instantaneous lung volume (ILV) and electrogastrogram (EGG), were recorded during the experiment. EGG, a measure of abdominal biopotentials, was recorded as part of a companion study (Blanford, 1990; Blanford and Oman, 1990). It required the placement of three electrodes on the subject's abdomen and did not interfere with the recordings of ECG and ILV important to this study.

Electrocardiogram was recorded from bipolar surface electrodes arranged to record approximately standard primary leads. A Hewlett-Packard Model 78203A ECG monitor, mounted on the instrumentation shelf, was used to filter and amplify the ECG signal. Of Leads I, II, and III, the lead with the qualitatively most distinguishable QRS complex was recorded (most often, Lead II). The output from the ECG monitor was routed through chair slip rings to one channel of an FM tape recorder (Hewlett-Packard Model 3964A).

Instantaneous lung volume was recorded via inductance plethysmography. A Respitrace (Ambulatory Monitoring Systems, Inc. model 10.9020 with transducer model 10.7000) was mounted on the instrumentation shelf and a single recording belt was positioned about the subjects chest. The ILV signal from the Respitrace was routed through slip rings to a second channel on the FM tape recorder. The respitrace signal was calibrated periodically using an 800 cc Spirobag (Ambulatory Monitoring Systems, Inc., model 10-4026). While

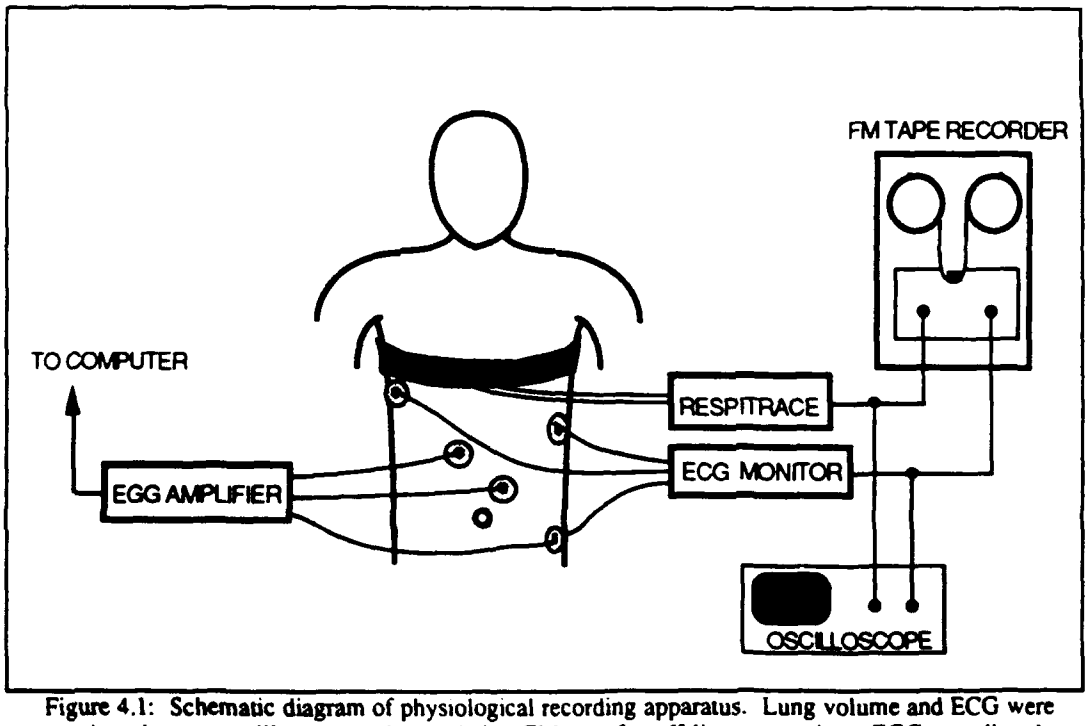


Figure 4.1: Schematic diagram of physiological recording apparatus. Lung volume and ECG were monitored on an oscilloscope and recorded to FM tape for off-line processing. EGG was directly filtered and digitized for a companion study.

ILV was recorded, subjects inhaled moderately, exhaled into the bag and held for five to ten seconds and then inhaled from the bag and held again. In this way, two constant levels of lung volume separated in magnitude by 800 cc were recorded. The process was typically repeated twice at each calibration.

Tape speed of the FM recorder was 1 7/8 inches per second which provided for recording bandwidth of 625 Hz for both signals. Both ECG and ILV were monitored in real-time on an oscilloscope (Tektronics Model 2225). A schematic of the recording apparatus is provided in Figure 4.1.

4.1.3 Symptom Monitoring

Throughout the experiments motion sickness signs were monitored by two experienced observers. Symptoms were monitored by subjects. They were reported verbally by use of standard definitions when possible and were continuously quantified by use of Magnitude Estimation. Magnitude Estimation is a numerical technique for reporting bodily sensations. It relies on subjects to report relative levels of sensations using a numerical ratio scale.

Stevens (1959) demonstrated that humans can reliably estimate magnitudes of the sensations produced by stimuli such as sound, electric shock or vibrations using ratio scales. When subjects were provided with a standard stimulus intensity, they were able to give consistent estimates of subsequent sensation intensities relative to the standard. Variations of the technique have been applied to motion sickness symptom reporting in studies by Bock and Oman (1982), Eagon (1988), and Rague (1987), among others. When applied to nausea, the technique calls for subjects to choose a level of nausea which they consider "halfway to vomiting" and assign it a numerical value. All reports of nausea are then to be made proportional to this standard. Thus, a doubling of the numerical report

is taken to indicate a doubling of nausea intensity. Although motion sickness involves a more complex stimulation and presumably more complex sensory modalities, than situations studied by Stevens, subjects are generally comfortable applying the technique and after some experience they give seemingly consistent estimates of their discomfort. (Bock and Oman, 1982; Eagon, 1988; Rague, 1987)

The technique applied in this thesis the same as that used by Eagon (1988) with the exception that subjects were instructed to rate only their 'nausea' levels and not their 'overall discomfort'. Thus, in the current experiments, milder motion sickness symptoms

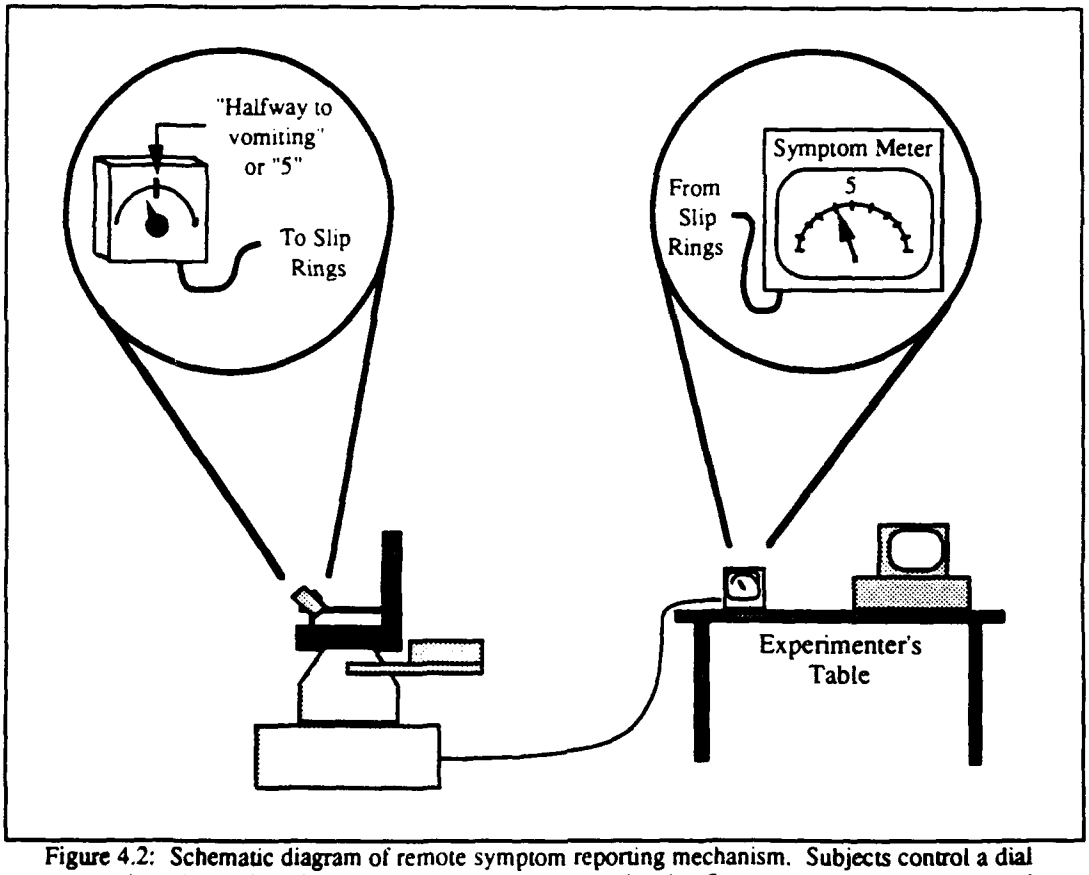


Figure 4.2: Schematic diagram of remote symptom reporting mechanism. Subjects control a dial mounted on the chair to indicate their present symptom level. Symptom reports are monitored remotely by the experimenter on a desk top meter.

such as headache or fatigue are not included in the magnitude estimates. Therefore, a numerical rating in this experiment presumably involves at least as much and probably more nausea than an equal rating in Eagon's experiment. The technique differs from that of Bock and Oman (1982), or Rague (1987) in that subjects were not exposed to motion induced nausea immediately prior to the experiment. Subjects' standard definitions of nausea level "halfway to vomiting" were derived from their memories of past experiences with nausea. This level was assigned the numerical value "five" and all subsequent reports were made relative to this standard.

During much of the experiment, subjects gave verbal reports of their magnitude estimates. However, during random interval breathing segments, the subjects were required not to speak. They reported their symptoms by turning a dial mounted on the arm of the chair. The level of 5 was indicated on the dial face plate. Each click of the dial changed their current magnitude estimate by one-half. Dial position was fed electrically to a meter on the experimenters table. The experimenter repeated the current estimates aloud periodically to insure the proper level was recorded. The system is illustrated schematically in Figure 4.2.

4.1.4 Experiment Protocol

The experiment protocol which follows was approved by the MIT Committee on the Use of Humans as Experimental Subjects (MIT-COUHES 1293). The rotating chair and reversing prism goggles described in Section 3.2.2 were used. A time line of the protocol is illustrated in Figure 4.3. Complete experimental sessions lasted between 3 and 4 hours.

Once subjects gave written informed consent, they were instrumented for physiological recordings and were seated in the Barany chair (Section 3.2.1). They remained in the chair

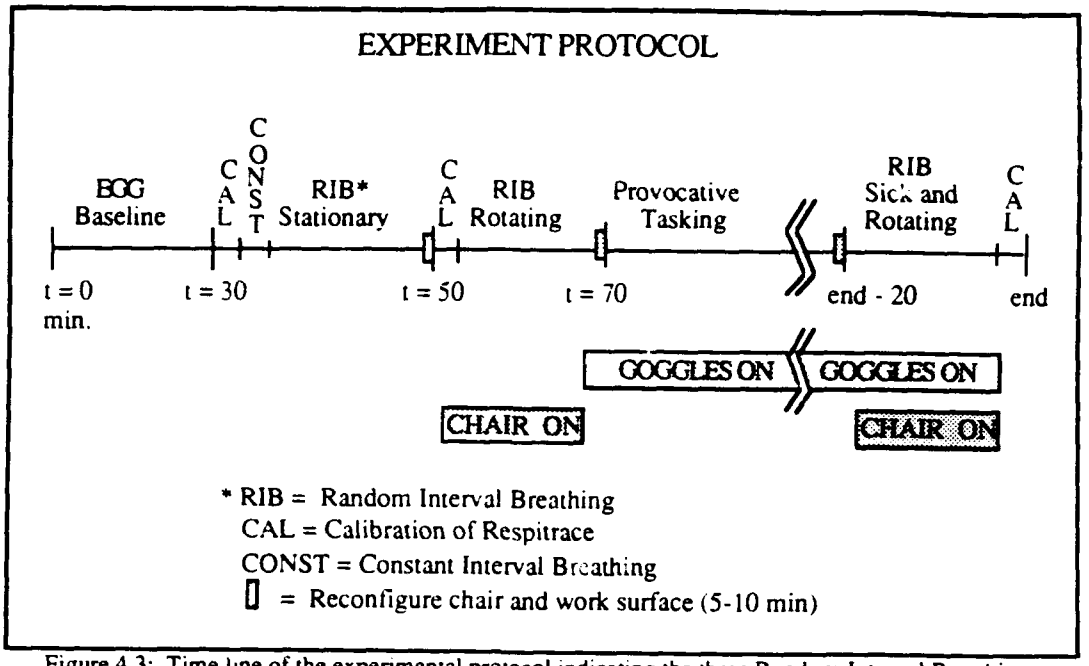


Figure 4.3: Time line of the experimental protocol indicating the three Random Interval Breathing (RIB) segments. Typical experiment duration was 3.5 hr.

throughout the experiment. The experiment protocol and symptom reporting were reviewed, and cued breathing was briefly practiced.

During the first one-half hour of each experiment, baseline EGG was recorded for the companion study (Blanford, 1990; Blanford and Oman, 1990) as subjects sat relaxed. EGG recording continued throughout the experiment. Instantaneous lung volume (ILV) and electrocardiogram (ECG) were recorded during three separate Random Interval Breathing (RIB) segments. The ILV signal was calibrated either prior to or immediately after each RIB segment.

RIB segments were completed in the same manner as that developed by Berger (1989b). A fifteen minute sequence of randomly spaced tones was prerecorded on cassette tape for playback during each experiment. The same sequence of tones was used during each RIB segment. The tones were generated by computer and were separated by intervals derived

from a modified Poisson distribution with a mean rate of 12 breaths/minute (refer to Section 2.3.3, Figure 2.10). The distribution was proposed by Berger et al. since the energy in a sequence of Poison intervals is constant over all frequencies. The modified distribution disallows intervals outside the range of 1 to 15 seconds but, as Berger et al. show; despite this modification the process remains quite broad band. During each RIB segment, subjects were instructed to initiate an inspiratory-expiratory cycle each time they were cued by a tone. The first RIB segment was preceded by a two minute sequence of tones spaced equally at the same mean rate of 12 breathes/minute. During this time subjects were able to settle on a comfortable depth of respiration. The second and third RIB segments were preceded by 20 to 30 seconds of constant interval breathing.

During the first RIB segment, subjects were seated and the chair was stationary. During the second segment, subjects remained seated but the chair was rotated sinusoidally about an earth vertical axis. The chair peak velocity was 120 degrees/second at a frequency of 0.1 Hz. At this rate, a full 360 degrees of rotation was completed in each half cycle.

After these two control recordings, subjects were fitted with the prism goggles (Section 3.2.2) and the retractable work surface was lowered in front of them. Subjects began a series of coordinated tasks designed to induce motion sickness in most subjects within approximately one-half hour. The tasks are similar to those described by Eagon (1988) with the exception of his head movement protocol.

The first task was to complete a one page questionnaire (Appendix A, identical to that used by Eagon, 1988). The questionnaire was printed mirror reversed so that it appeared normal to the subject. It required that the subject exercise eye-hand coordination in writing simple words, copying simple pictures, and solving simple math problems. This task was

performed for a maximum of fifteen minutes. Subjects were not required to complete the entire questionnaire in this time.

The second task was can structure building. A schematic drawing representing 15 soda cans arranged in a particular stacked structure was placed in front of the subject. (Photo-reduced copies of the schematics are provided in Appendix A.) Fifteen empty soda cans were placed on the right side of the work surface. Subjects were instructed to use one hand to move one can at a time from the right side of the work surface to the left side, building the structure represented in the drawing as they did so. When the first structure was complete or 5 minutes elapsed, any remaining cans were moved to the left side of the work station and a second schematic drawing was presented. The subject was asked to repeat the process except this time building the new structure by moving one can at a time from left to right. Again, the subject was allowed five minutes to complete the structure. A complete can structure protocol required repetition of this process until a total of ten structures were attempted.

The third tasking protocol involved copying drawings and solving simple mathematical problems on a white-board. A 2 by 3 foot white board marked with two drawings and four mathematical problems was positioned in front of the subject. The subject was provided with a marker and was asked to replicate the drawings and solve the problems.

Upon completion of the white board task, the subject completed a sequence of two more can structures and then repeated the white board task with new drawings and problems. The process of white board work followed by two can structures was repeated, as necessary, for the remainder of the tasking period.

Subjects who reached a level of nausea which they estimated to have a magnitude of between 3 and 4 (where 5 is "halfway to vomiting"), were asked to moderate their activity in order to maintain but not exceed these symptoms. They were instructed in the therapeutic value of closing their eyes and holding their heads still, and were asked to try to apply these techniques to control their symptom level. While their symptom levels remained between 3 and 4 they were instructed to continue tasking, but when their symptoms began to increase above this level they were to stop tasking, relax and close their eyes until symptoms began to subside.

The tasking period was ended under one of four conditions: (1) the subject reported and maintained for fifteen minutes, a sickness level between 3 and 4 on the magnitude estimate scale, (2) one and one-half hours of tasking was completed, (3) the experimenter opted to terminate the session in the best interest of the subject, and (4) the subject chose to withdraw from the experiment.

Subjects who were able to continue the experiment, participated in the final RIB segment. The table top was retracted and the same chair rotation profile used in RIB segment two was repeated. In this case however, subjects were motion sick and continued wearing the prism goggles. Subjects were instructed to hold their heads motionless against the headrest and to try to maintain symptom levels of 3 to 4 by opening and closing their eyes. During the first few cycles of rotation, subjects were allowed to again become comfortable with controlling their symptoms. Symptoms were reported non-verbally using the chair mounted system described in Section 4.1.3. The third and final RIB segment was completed as subjects rotated while maintaining the desired symptom level.

4.2 Analysis

The objective of the data analysis was to estimate and compare the transfer functions of ILV to IHR under the three different experimental conditions represented by the three RIB sessions. The first step in the analysis was to sample the data to computer and extract sequences of ILV and IHR representative of each condition (Section 4.2.1 and 4.2.2). The second step was to estimate the transfer and coherence functions (Section 4.2.3). The third and final step was to quantitatively compare the transfer functions (Section 4.2.4 and 4.2.5).

4.2.1 Digitization

Electrocardiogram and instantaneous lung volume signals were replayed from FM tape and were digitized through a 12 bit analog to digital (A/D) converter for computer analyses (Masscomp MC-500). Hardware constraints and timing considerations necessitated that both signals be sampled at the same rate. Neither signal contains significant power above 100 Hz, so in order to satisfy Nyquist's criterion and thus preserve signal integrity, a sampling rate of 200 Hz would suffice. However, in order to allow for accurate determination of temporal locations of heart beats without the requirement of interpolation, a sampling rate of 360 Hz was used (refer to Berger, 1987, for further discussion). Each signal was passed through a six pole Butterworth (anti-aliasing) filter with cutoff frequency of 180 Hz prior to sampling.

Instantaneous lung volume signals were digitally filtered and decimated down to an effective sampling rate of 3 Hz. (Decimation is completed using an efficient four stage algorithm described by Berger (1987)) Instantaneous heart rate signals were derived from the ECG.

4.2.2 Estimating Instantaneous Heart Rates

Many different algorithms have been proposed for estimating instantaneous heart rates. All of the algorithms require the detection and timing of individual heart beats. Since the QRS complex or, more specifically, the R peak is the most easily distinguishable component of the ECG of a single heart beat, most algorithms begin with detection of R peaks. The time between two consecutive R peaks is defined as an R-R interval (Figure 4.4). Perhaps the simplest algorithm for estimating instantaneous heart rate is to take the inverse of a series of R-R intervals. However, such a technique provides estimates of heart rate on a per beat basis and therefore provides a series of estimates unevenly spaced in time. It is desirable to work with a true time series in the sense that samples of heart rate are at constant time intervals (and, therefore, spectra and transfer functions have units per Hz).

Instantaneous heart rates were calculated as described by Berger et al. (1986) Their technique provides IHR estimates evenly spaced in time, at any chosen frequency, and avoids problems of bias and time delay accompanying other techniques (refer to Berger,

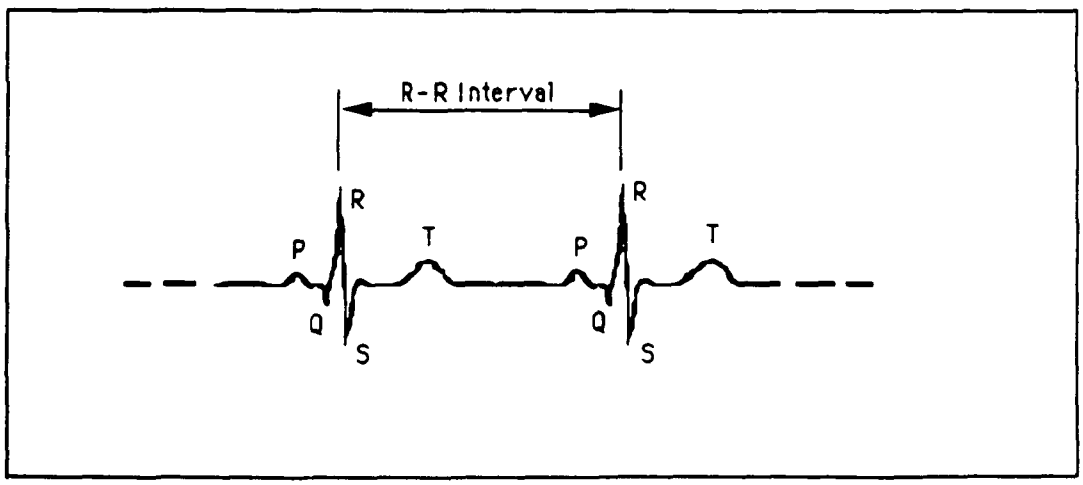


Figure 4.4: Sketch of an electrocardiogram of two heart beats indicating the P,Q,R,S, and T wave segments and the R-R interval.

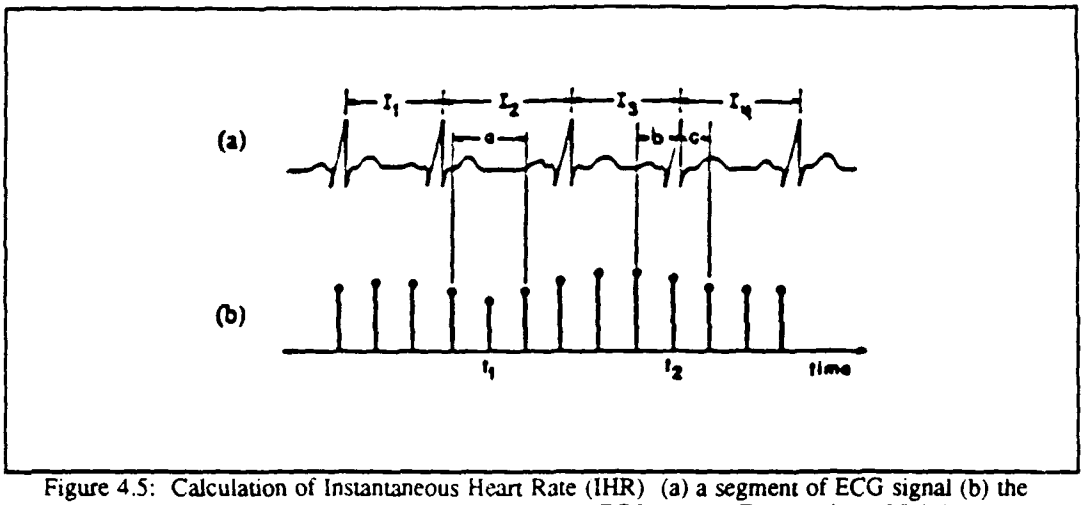


Figure 4.5: Calculation of Instantaneous Heart Rate (IHR) (a) a segment of ECG signal (b) the instantaneous heart rate samples corresponding to the ECG signal. The number of RR intervals within the local window centered at t_1 is given by: $RR[t_1] = a/I_2$ and at t_2: $RR[t_2] = b/I_3 + c/I_4$. (modified from Berger et al., 1986)

1986, for further discussion). In this study the sampling frequency, F_S (equivalently the inverse of sampling period, $1/T_S$) was chosen as 3 Hz for synchrony with the ILV signal. The estimate of IHR, IHR[n] , at the n^{th} sample point is given by:

$$IHR[n] = \frac{RR[n]}{2\,T_s} = \frac{F_s \times RR[n]}{2} \qquad (4.1)$$

where RR[n] is the number of RR intervals, including fractional intervals, which are contained in the time interval between the $(n-1)^{th}$ and $(n+1)^{th}$ sample points. Thus, IHR[n] is calculated as the number of RR intervals within a local window, divided by the duration of the local window. (for further discussion refer to Berger, 1986) Sample calculations are demonstrated in Figure 4.5 (from Berger, 1986).

Two 1024 point (5.69 min. at 3 Hz) paired sections of the ILV and IHR signals were extracted from each 15 minute random interval breathing segment for each subject. Each ILV[n] data sequence was calibrated using the calibration data most closely associated with each RIB segment. Figure 4.6 depicts a flow chart of the data processing involved in

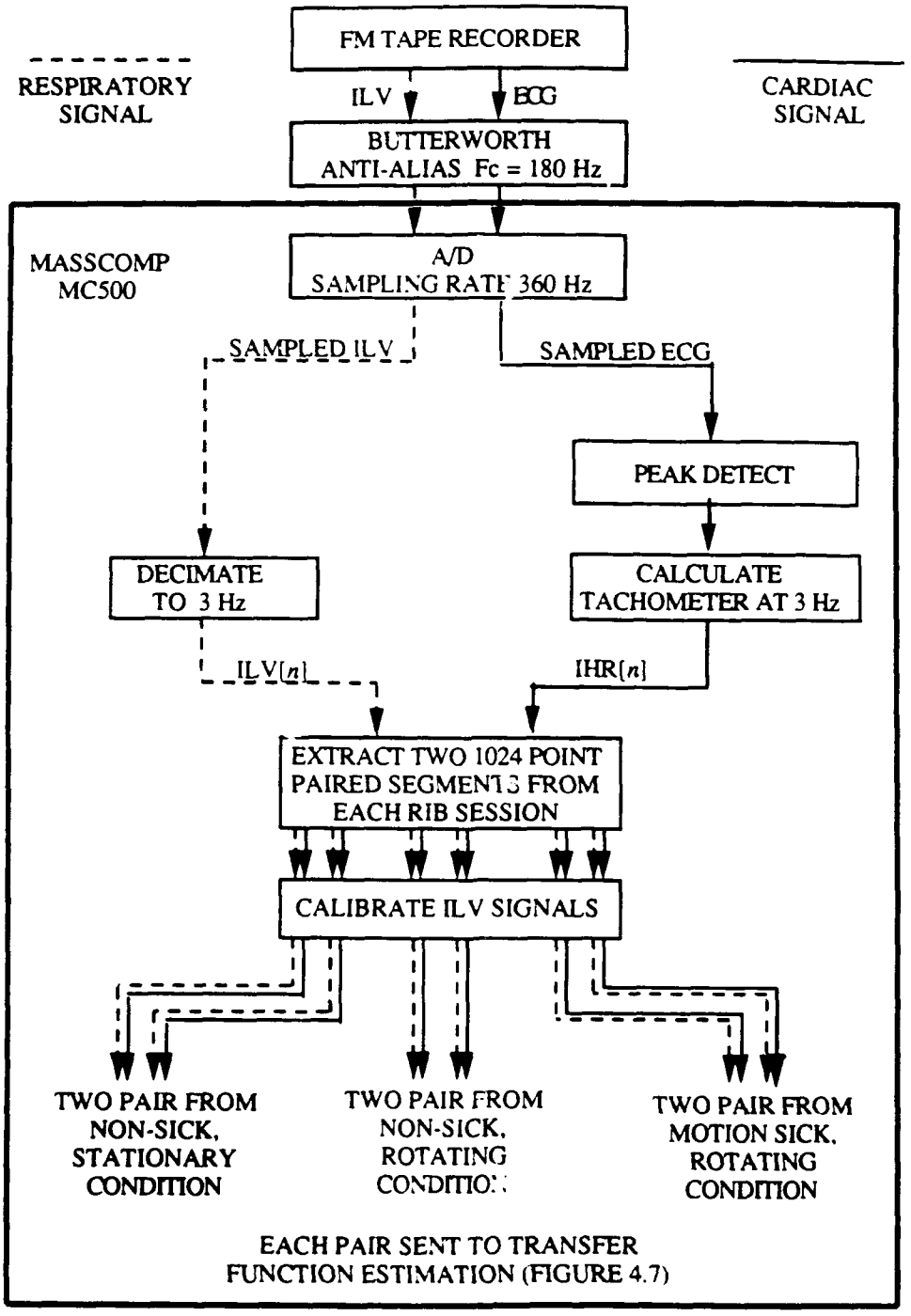


Figure 4.6: Flow graph indicating the steps applied to derive six paired ILV and IHR data segments for each subject

obtaining these paired sections. The six paired data sections will be referred to as

$\mathrm{ILV}_{c1}[n]$, $\mathrm{IHR}_{c1}[n]$:	First segments from stationary control segment
$\mathrm{ILV}_{c2}[n]$, $\mathrm{IHR}_{c2}[n]$:	Second segments from stationary control segment
$\mathrm{ILV}_{r1}[n]$, $\mathrm{IHR}_{r1}[n]$:	First segments from rotating control segment
$\mathrm{ILV}_{r2}[n]$, $\mathrm{IHR}_{r2}[n]$:	Second segments from rotating control segment
$\mathrm{ILV}_{s1}[n]$, $\mathrm{IHR}_{s1}[n]$:	First segments from rotating post stimulus segment
$\mathrm{ILV}_{s2}[n]$, $\mathrm{IHR}_{s2}[n]$:	Second segments from rotating post stimulus segment

Each of these pairs was used to estimate one ILV to IHR transfer function. Thus, a total of six transfer functions (two from each of three experimental conditions) were estimated for each subject.

4.2.3 Calculating Individual Transfer Functions

Each transfer function was estimated as follows (Figure 4.7). Autocorrelation functions, $\hat{R}_{LL}[k]$ and $\hat{R}_{HH}[k]$, of the paired finite data segments ILV[n] and IHR[n], respectively, were estimated by

$$\hat{R}_{HH}[k] = \frac{1}{N-|k|} \sum_{n=0}^{N-|k|-1} IHR[n]\ IHR[n+|k|] ; \qquad 1-N < k < N-1 \tag{4.2a}$$

and

$$\hat{R}_{LL}[k] = \frac{1}{N-|k|} \sum_{n=0}^{N-|k|-1} ILV[n]\ ILV[n+|k|] ; \qquad 1-N < k < N-1 \tag{4.2b}$$

where data points in each segment are indexed from one n=0 to n = N-1=1023.

Similarly, the crosscorrelation, $\hat{R}_{HH}[k]$ between ILV[n] and IHR[n] was estimated by

$$\hat{R}_{LH}[k] = \left\{ \begin{array}{ll} \frac{1}{N-|k|} \sum_{n=0}^{N-|k|-1} ILV[n]\ IHR[n+|k|]\ ; & 0 < k < N-1 \\ \frac{1}{N-|k|} \sum_{n=-k}^{N-1} ILV[n]\ IHR[n+|k|]\ ; & 1-N < k < 0 \end{array} \right\} \tag{4.3}$$

where, again, data points in each segment are indexed from n=0 to n=N-1=1023. In this case separate definitions for the cases $k > 0$ and $k < 0$ are required since the indices of ILV[n] and IHR[m] are not interchangeable. Note that the auto- and crosscorrelation functions are estimated at 2047 different lags and therefore these time series are 2047 (instead of 1024) points long.

Since the auto- and crosscorrelation function estimates are based on finite data records, there is significant variance in the estimates. Furthermore, as |k| approaches N, the correlations are based on fewer data points and therefore the variance of the estimates increase as |k| increases. Application of a window to the correlation functions prior to calculating power spectra or cross spectra serves to reduce the variance of the spectral estimates at the expense of spectral resolution and bias. A Gaussian window of the form

$$w[k] = e^{-(k\,T_s)^2/2\sigma_t^2}\ ; \qquad 1-N < k < N-1 \tag{4.4a}$$

was applied to each autocorrelation and crosscorrelation function. A value of $\sigma_t = NT_S/4\pi = 27.16$ was used. This value provides roughly a 14% decrease in the variance of the spectral estimates from that of their unwindowed levels (Berger, 1987). Jenkins and Watts (1968) have shown that the effective frequency resolution of the windowed spectra is decreased by a factor of 1/Q where

$$Q = \frac{1}{NT_S} \sum_{k=1-N}^{N-1} w^2[k]T_S \tag{4.4b}$$

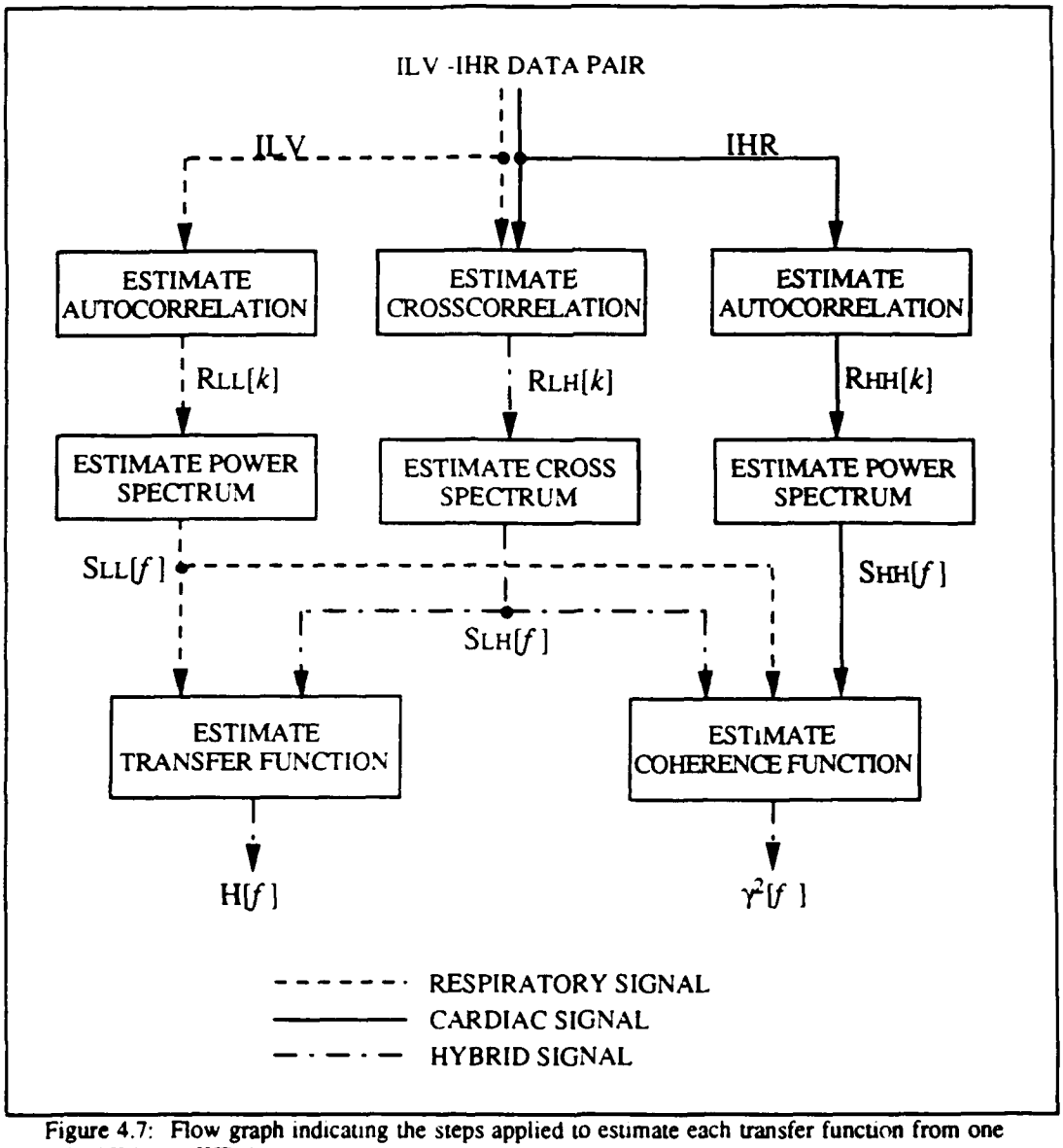


Figure 4.7: Flow graph indicating the steps applied to estimate each transfer function from one paired ILV and IHR data segment.

For the Gaussian window defined above, Q = 141 and therefore, effective spectral resolution is reduced by approximately seven times. As a result, independent frequency samples are separated by approximately 0.01 Hz. Windowing also introduces an estimation bias in the form of a tendency toward underestimation of spectral peaks and overestimation of spectral valleys.

Samples of the power spectra of IHR[n] and ILV[n] and the cross-spectrum between the two were calculated as

$$\hat{S}_{LL}[f] = T_s\,\mathrm{DFT}\left(\hat{R}_{LL}[k]\,w[k]\right) \tag{4.5a}$$

$$\hat{S}_{HH}[f] = T_s\,\mathrm{DFT}\left(\hat{R}_{HH}[k]\,w[k]\right) \tag{4.5b}$$

and

$$\hat{S}_{LH}[f] = T_s\,\mathrm{DFT}\left(\hat{R}_{LH}[k]\,w[k]\right) \tag{4.5c}$$

respectively, where DFT() is a 2048 point Discrete Fourier Transform operator.

In practice, efficient Fast Fourier Transform (FFT) algorithms are used in calculating both the correlation function estimates and the power spectrum estimates (refer to Berger, 1987).

Finally, the transfer function and coherence function are estimated as

$$\hat{H}[f] = \frac{\hat{S}_{LH}[f]}{\hat{S}_{HH}[f]} \tag{4.6a}$$

and

$$\hat{\gamma}^2[f] = \frac{|\hat{S}_{LH}[f]|^2}{\hat{S}_{HH}[f]\,\hat{S}_{LL}[f]} \tag{4.6b}$$

Following the method outlined above, six transfer function estimates per subject were calculated from the six ILV-IHR pairs. Henceforth they will be referred to as follows:

$H_{c1}[f]$: First estimate from stationary control segment
$H_{c2}[f]$: Second estimate from stationary control segment
$H_{r1}[f]$: First estimate from rotating control segment
$H_{r2}[f]$: Second estimate from rotating control segment
$H_{s1}[f]$: First estimate from rotating post stimulus segment
$H_{s2}[f]$: Second estimate from rotating post stimulus segment

Although 3 Hz sampling rate allows estimation of the transfer functions from 0 to 1.5 Hz, generally only the first one third (or 1024/3 ≈ 342 sample points) of the transfer function (0.0 to 0.5 Hz) was considered accurate and physiologically meaningful (Berger, 1987).

4.2.4 Calculating Group Average Transfer Functions and Confidence Intervals

In the interest of identifying consistent changes in the transfer functions between the three conditions, group average transfer functions with confidence intervals were calculated. The nonsick, stationary condition was represented by an average of all H_{c1} and H_{c2} transfer functions. The nonsick and rotating condition was represented by an average of all H_{r1} and H_{r2} transfer functions from the subject population. Finally, the motion sick and rotating condition was represented by an average of all H_{c1} and H_{c2} transfer functions. The algorithm for computing these pooled estimates was developed by Berger and colleagues (personal communication) and is similar to that described by Berger (1987). The development that follows closely parallels that given by Berger (1987) with the exception that magnitude and phase components (as opposed to real and imaginary parts) of the transfer functions are treated independently.

The maximum likelihood formulation for computing the group means weights each estimate by the reciprocal of its variance. The average magnitude, $\langle|\widehat{\mathbf{H}}[f]|\rangle$, and average phase, $\langle\hat{\vartheta}\rangle$, (assumed independent) are given by

$$\left\langle \left| \widehat{\mathbf{H}}[f] \right| \right\rangle = \frac{\sum_{i=1}^{M} \frac{\left|\widehat{\mathbf{H}}_i[f]\right|}{\sigma^2_{|\mathbf{H}|i}[f]}}{\sum_{i=1}^{M} \frac{1}{\sigma^2_{|\mathbf{H}|i}[f]}} \qquad (4.7a)$$

and

$$\left\langle \hat{\vartheta}[f] \right\rangle = \frac{\sum_{i=1}^{M} \frac{\left|\hat{\vartheta}_i[f]\right|}{\sigma^2_{\vartheta i}[f]}}{\sum_{i=1}^{M} \frac{1}{\sigma^2_{\vartheta i}[f]}} \qquad (4.7b)^*$$

where the index, i, iterates across the M individual estimates to be included in the average. The variance in each individual estimate of the magnitude and phase (as a function of frequency) are denoted $\sigma^2_{|H|i}[f]$ and $\sigma^2_{\vartheta i}[f]$, respectively. These variances are unknown and must themselves be estimated.

Berger distinguishes two components of these variances, the estimator variances, $\pi^2_{\vartheta i}[f]$ and $\pi^2_{|H|i}[f]$, and the population variances, $\rho^2_{\vartheta i}[f]$ and $\rho^2_{|H|i}[f]$. The estimator variances characterize the error inherent in the estimation procedure itself as a function of frequency for a particular transfer function. Berger (1987, personal communication), using results from Jenkins and Watts (1968), has demonstrated that the estimator variances can be approximated,through use of the coherence function, as

$$\pi^2_{|H|i}[f] = |\hat{H}_i[f]|^2 \left(\frac{2}{\mu - 2} f_{2,\mu-2}(0.68) \frac{1 - \gamma^2_i[f]}{\gamma^2_i[f]} \right) \qquad (4.8a)$$

and

$$\pi^2_{\vartheta i}[f] = \hat{\vartheta}^2_i[f] \sin^{-1}\left(\frac{2}{\mu - 2} f_{2,\mu-2}(0.68) \frac{1 - \gamma^2_i[f]}{\gamma^2_i[f]} \right) \qquad (4.8b)$$

where $f_{2,\mu-2}(0.68)$ is a number such that

$$\mathbf{Prob}\left[F_{2,\mu-2} \le f_{2,\mu-2}(0.68) \right] = 0.68 \qquad (4.9)$$

where $F_{2,\mu-2}$ is the F distribution with 2 degrees of freedom in the numerator and μ-2 degrees of freedom in the denominator. The value of the constant, μ, is the number of

* In practice, due to the periodic nature of phase, direct solution to equation 4.7b involves calculation of the roots of an Mth order equation and selecting the proper root as solution. A computationally less burdensome approach is typically used. The average phase is taken to be the phase of a weighted vector sum of unit vectors. Each unit vector has the same phase as one of the original vectors in the average. Each is weighted by the inverse of its total variance. This algorithm gives similar results to that of direct solution of 4.7b (Berger, personal communication).

degrees of freedom in the chi-squared distributions characterizing the spectral estimates. It is a function of the window used in the spectral estimation process and is equal to 14 in this analysis (Berger, 1987; Jenkins and Watts, 1968).

The population variances, $\rho^2_{\vartheta}[f]$ and $\rho^2_{|\mathbf{H}|}[f]$, characterize the variability in the transfer function estimates due to individual differences across the population. They can be approximated by

$$\rho^2_{|\mathbf{H}|}[f] = \frac{1}{M-1} \sum_{i=1}^{M} \left[|\mathbf{H}|_i[f] - \langle |\widehat{\mathbf{H}}[f]| \rangle \right]^2 \qquad (4.10a)$$

and

$$\rho^2_{\vartheta}[f] = \frac{1}{M-1} \sum_{i=1}^{M} \left[\vartheta_i[f] - \langle \widehat{\vartheta}[f] \rangle \right]^2 \qquad (4.10b)$$

However, equations 4.10a and 4.10b employ the group average magnitude and phase which are unknown and their estimation depends on knowledge of the population variances. An iterative approach to simultaneously solving equations 4.7 and 4.10 is possible, however, Berger (1987, personal communication) has successfully demonstrated a less computationally burdensome approach. For the purposes of estimating these population variances, the group averages were temporarily estimated using equations 4.7a and 4.7b with equal weights. That is, all variances were temporarily taken to be equal, to allow approximation of $\langle |\widehat{\mathbf{H}}[f]| \rangle$ and $\langle \widehat{\vartheta}[f] \rangle$. Using these approximations, $\rho^2_{\vartheta}[f]$ and $\rho^2_{|\mathbf{H}|}[f]$ were calculated.

The total variance was then taken as the sum of the estimator variance and the population variance. That is,

$$\sigma^2_{|\mathbf{H}|}[f] = \pi^2_{|\mathbf{H}|}[f] + \rho^2_{|\mathbf{H}|}[f] \qquad (4.11a)$$

and

$$\sigma^2_{\vartheta}[f] = \pi^2_{\vartheta}[f] + \rho^2_{\vartheta}[f] \qquad (4.11b)$$

Group average transfer magnitude and phase were then recalculated using these variances as the new weights in equations 4.7a and 4.7b.

The standard error of the group means were then given by

$$\sigma_{\hat{H}}[f] = \sqrt{\frac{1}{\sum_{i=1}^{M} 1 / \sigma^2_{Hi}[f]}} \qquad (4.12a)$$

and

$$\sigma_{\hat{\vartheta}}[f] = \sqrt{\frac{1}{\sum_{i=1}^{M} 1 / \sigma^2_{\vartheta i}[f]}} \qquad (4.12b)$$

4.2.5 Comparing Transfer Functions for Individuals

The method developed by Berger (1987, personal communication) and described in the previous section, allows quantitative comparisons between *group average* responses to the three experimental conditions. In order to make such comparisons for *individual* responses, a different approach was taken. Recall that for an individual, two transfer functions representing each of the three experimental conditions were calculated. Two distinct comparisons between these transfer functions were made. First, to assess changes due to rotation, the stationary control segments ($H_{c1}[f]$ and $H_{c2}[f]$) were compared with the rotating control segments ($H_{r1}[f]$ and $H_{r2}[f]$). Second, in order to assess changes due to motion sickness, the rotating control segments ($H_{r1}[f]$ and $H_{r2}[f]$) were compared to the rotating, motion sick segments ($H_{s1}[f]$ and $H_{s2}[f]$). Both of these comparisons were made in the same manner.

The first step was to obtain independent samples of each transfer function. As mentioned in Section 4.2.3, adjacent frequency samples are correlated due to the windowing procedure incorporated in the spectral estimation. Transfer function samples separated by 0.01 Hz (or, equivalently, by seven discrete samples) are mostly uncorrelated. Arithmetic (complex) averages across blocks of seven samples were calculated for each transfer function as follows:

$$\mathbf{H}[F_i] = \frac{1}{7}\sum_{j=-3}^{j=+3} \mathbf{H}[f_{i+j}]; \qquad i = 4,11,18,25....,333,340 \tag{4.13}$$

where the 1024 samples of the transfer function between 0 and 1.5 Hz are denoted sequentially by $\mathbf{H}[f_1]$, $\mathbf{H}[f_2]$, $\mathbf{H}[f_3]$, ..., $\mathbf{H}[f_{1024}]$.The $\mathbf{H}[F_i]$ are taken as independent estimates of the transfer function at frequencies, F_i. Only the first one-third of the transfer function is used. Thus, forty-nine independent samples of the transfer function between 0 and 0.5 Hz were obtained.

To test the null hypothesis that rotation has no effect on an individual's transfer function estimates, the following statistic was calculated at each independent frequency, F_i :

$$C_r[F_i] = \frac{\left(\frac{H_{r1}[F_i] - H_{c1}[F_i]}{\sqrt{2\sigma^2}}\right)^2 + \left(\frac{H_{r2}[F_i] - H_{c2}[F_i]}{\sqrt{2\sigma^2}}\right)^2}{\left(\frac{H_{r1}[F_i] - H_{r2}[F_i]}{\sqrt{2\sigma^2}}\right)^2 + \left(\frac{H_{c1}[F_i] - H_{c2}[F_i]}{\sqrt{2\sigma^2}}\right)^2} \tag{4.14}$$

Under the null hypothesis, the four transfer function estimates $H_{c1}[F_i]$, $H_{c2}[F_i]$, $H_{r1}[F_i]$ and $H_{r2}[F_i]$ are realizations of the same random process. This process has unknown variance σ^2. Under this hypothesis, each of the four squared differences in the ratio is

distributed as a chi^2 with one degree of freedom (chi$_1^2$).* The sum of two chi$_1^2$ statistics is distributed as chi$_2^2$, and the ratio of two chi$_2^2$ is distributed as F with 2 degrees of freedom in both the numerator and denominator ($F_{2,2}$). Thus, under the null hypothesis, the statistic, $C_r[F_i]$, is distributed as a $F_{2,2}$. (Note that in practice, the value of σ^2 is not needed to calculate $C_r[F_i]$ since the factors $\sqrt{2\sigma^2}$ cancel.)

A similar statistic was calculated to test for effects of motion sickness. The statistic, $C_s[F_i]$, was calculated as

$$C_s[F_i] = \frac{\left(\frac{H_{s1}[F_i] - H_{r1}[F_i]}{\sqrt{2\sigma^2}}\right)^2 + \left(\frac{H_{s2}[F_i] - H_{r2}[F_i]}{\sqrt{2\sigma^2}}\right)^2}{\left(\frac{H_{s1}[F_i] - H_{s2}[F_i]}{\sqrt{2\sigma^2}}\right)^2 + \left(\frac{H_{r1}[F_i] - H_{r2}[F_i]}{\sqrt{2\sigma^2}}\right)^2} \qquad (4.15)$$

Under the null hypothesis that motion sickness has no effect on the transfer function, $C_s[F_i]$ is also distributed as $F_{2,2}$.

---

* The distribution of each squared difference may be derived as follows. Each $H[F_i]$ under the null hypothesis is one realization of a random process. For simplicity the two realizations in each squared difference will be denoted generally as X_1 and X_2.

$$\left(\frac{X_1 - X_2}{\sqrt{2\sigma^2}}\right)^2 = \frac{((X_1 - \overline{X}) + (\overline{X} - X_2))^2}{2\sigma^2} \quad ; \text{ where } \overline{X} \text{ is the sample mean}$$

$$\frac{((X_1 - \overline{X}) + (\overline{X} - X_2))^2}{2\sigma^2} = \frac{(X_1 - \overline{X})^2 + 2(X_1 - \overline{X})(\overline{X} - X_2) + (\overline{X} - X_2)^2}{2\sigma^2}$$

$$\text{and } (X_1 - \overline{X}) = (\overline{X} - X_2) \text{ so}$$

$$\frac{(X_1 - \overline{X})^2 + 2(X_1 - \overline{X})(\overline{X} - X_2) + (\overline{X} - X_2)^2}{2\sigma^2} = \frac{4(X_1 - \overline{X})^2}{2\sigma^2} = \frac{2(X_1 - \overline{X})^2}{\sigma^2}$$

$$\text{Finally, } \frac{2(X_1 - \overline{X})^2}{\sigma^2} = \frac{1}{\sigma^2}\sum_{i=1}^{2}(X_i - \overline{X})^2 ,$$

which is distributed as chi$_1^2$. [Rice, 1988, pp. 172-3]

For each subject, $C_r[F_i]$ and $C_s[F_i]$, were plotted as functions of frequency, F_i, and the distributions of their magnitudes were plotted as histograms. Comparison to the $F_{2,2}$ distribution were made to assess significance of effects.

In order to assess the sensitivity of the analysis to the assumption that independent frequency samples of the transfer function are separated by 0.01 Hz, the process was repeated assuming a separation approximately half as large (0.0044 Hz or equivalently 3 discrete samples). The calculations were identical to those above except that the $H[F_i]$ were calculated as:

$$\mathbf{H}[F_i] = \frac{1}{3}\sum_{j=-1}^{j=+1} \mathbf{H}[f_{i+j}]; \qquad i = 2,5,8,11\ldots,338,341 \tag{4.16}$$

In this case 114 independent samples of the transfer functions between 0 and 0.5 Hz were obtained. Equations 4.14 and 4.15 were used to calculate the new $C_r[F_i]$ and $C_s[F_i]$. Again, $C_r[F_i]$ and $C_s[F_i]$ were plotted as functions of frequency and the distributions of their magnitudes were plotted as histograms.

V Results

5.1 Subject Information and Motion Sickness Levels

Eighteen subjects each participated in one experiment session. Seven subjects were female and eleven were male. The average subject age was 22.3 years (min 18 yr; max 30 yr; average female 22.29 years; average male 22.36 years). All subjects were non-smokers and reported no medical problems. All reported not using any medication or consuming any alcohol within 24 hours of the experiment session. All reported consuming no coffee, tea or cola within twelve hours of the experiment session and undertaking no exercise within 6 hours of the session. Each had a normal night's sleep on the evening prior to the experiment and had their usual meal between three and four hours prior to the session. All experiments began between 3:30 and 4:30 PM.

Subject genders, ages, experience, self ratings of susceptibility and experiment endpoints are provided in Table 5.1. Subjects were categorized as either experienced or not

experienced based on whether they had previously participated in motion sickness experiments. Self ratings of susceptibility were taken from Section C of the Motion Sickness Questionnaire completed by each subject prior to the experiment. Experiment endpoints were denoted as either 'Normal', 'Time', 'Abort' or 'Emesis'. 'Normal' experiment endpoint was reached if the subject attained symptoms which were estimated as magnitude 3 or greater and maintained symptoms near this level for 15 minutes prior to and during the final random interval breathing segment. 'Time' endpoint was reached if a subject competed 1.5 hours of tasking without developing significant symptoms. 'Abort' endpoint indicates that the experiment was aborted prior to completion of the tasking period. 'Emesis' endpoint was reached if a subject vomited prior to completion of the tasking period despite all efforts to prevent such occurrence. Two subjects (13 and 14) reached the Time' endpoint, two (subject 15 and 16) reached the 'Abort' endpoint and two

Table 5.1 Subject Information

SUBJECT NUMBER	GENDER	AGE	EXPERIENCE	SELF RATED SUSCEPTIBILITY	EXPT ENDPOINT
1	F	19	No	Immune	Normal
2	F	23	No	Immune/Less	Normal
3	M	19	No	Less	Normal
4	M	25	Yes	Less	Normal
5	F	24	No	Average	Normal
6	M	25	No	Less	Normal
7	M	20	No	More	Normal
8	F	24	No	Average	Normal
9	M	20	No	Less	Normal
10	F	19	No	Immune	Normal
11	M	18	No	Less	Normal
12	F	22	Yes	Average	Normal
13	M	28	Yes	Less	Time
14	M	30	Yes	More	Time
15	M	19	No	Average	Abort
16	F	25	No	Average/Less	Abort
17	M	19	No	Less	Emesis
18	M	23	Yes	Average	Emesis

(subject 17 and 18) reached the 'Emesis' experiment endpoint. The remaining twelve subjects (1 through 12) reached the 'Normal' endpoint. No consistent relationships are evident between the parameters in Table 5.1. In particular, neither age, gender, experience or self rated susceptibilities were predictive of experiment endpoint.

Analyses were conducted on the sub-population of twelve subjects who achieved a 'Normal' experiment endpoint (six male, six female; age 18-25 yr, avg. 21.5 yr). Magnitude estimates of nausea for these subjects are plotted in Figure 5.1. Statistics of the magnitude estimates are given in Table 5.2. Most subjects controlled their symptoms closely about a mean level between 3 and 4 on their magnitude estimate scale. Subjects 3 and 10 reported slightly higher levels and Subjects 2 and 7 reported slightly lower levels.

Table 5.2 Magnitude Estimation Statistics for 'Normal' Subjects during the final RIB Segment

SUBJECT NUMBER	MAGNITUDE ESTIMATE OF NAUSEA			
	MEAN	STANDARD DEVIATION	MINIMUM	MAXIMUM
1	3.8	0.5	3	4
2	2.5	0.3	1	3
3	4.3	0.6	4	5
4	3.0	0.9	2	4
5	3.0	0.0	3	3
6	3.3	0.4	1	4
7	2.0	0.7	1	3
8*	3.9	0.8	3	5
9	3.1	0.7	2	4
10	4.1	0.4	4	5
11	3.7	0.6	3	5
12	3.4	0.5	3	4

* Subject 8 completed only 8 minutes of the final RIB segment

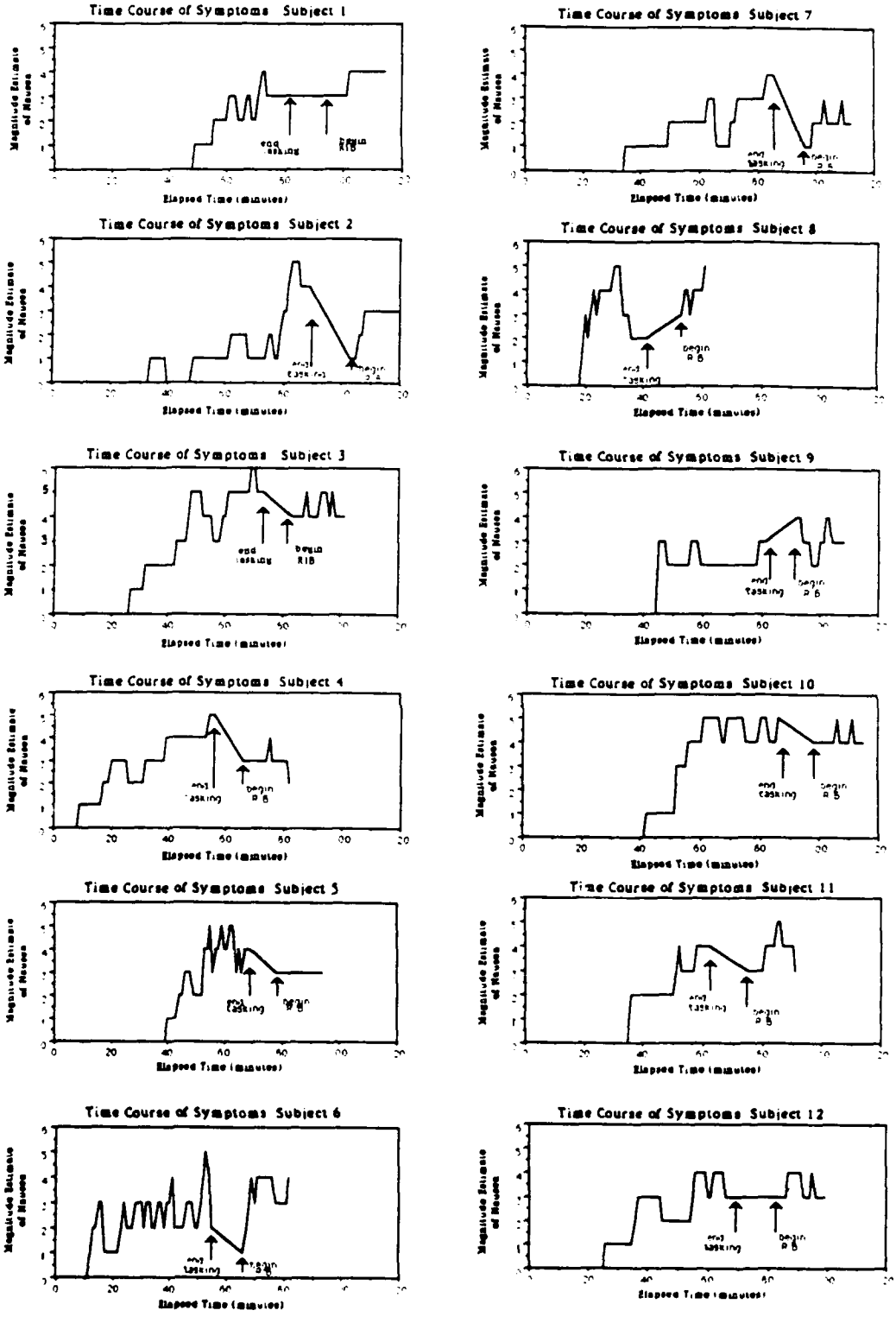


Figure 5.1: Time course of Magnitude Estimates of Nausea for 'Normal' Subjects (t =0 corresponds to the start of provocative tasking)

Subject 8 completed only 8 minutes of the final RIB segment. The group mean magnitude estimate level during the final RIB segment was 3.4 (standard error 0.65).

Each of the twelve subjects showed significant signs of sickness (evident to two experienced observers) such as pallor and sweating. During the tasking period, when verbal reports of symptoms were possible, or retrospectively, after the final RIB segment, each subject reported other symptoms of sickness accompanying their nausea. Seven of twelve subjects reported sweating, seven reported subjective feelings of warmth or cold, four reported increased salivation, four reported fatigue and all reported feelings of "fullness in the throat".

5.2 Sample Heart Rate and Lung Volume signals

For each of the twelve subjects except Subject 8, six paired ILV and IHR data segments (ILV_{c1}-IHR_{c1}, ILV_{c2}-IHR_{c2}, ILV_{r1}-IHR_{r1}, ILV_{r2}-IHR_{r2}, ILV_{s1}-IHR_{s1} and ILV_{s2}-IHR_{s2}) were extracted (Figure 4.6). Due to the limited duration of Subject 8's final RIB segment, ILV_{s2} and IHR_{s2} could not be calculated. Sample plots of ILV and IHR time series and their power spectra are given in Figure 5.2 (data from Subject 1). The random breathing pattern is evident in the plots of the ILV time series. The effective broadening of the frequency content of ILV is evident in the power spectra.

5.3 Individual Transfer Function Estimates

In Figures 5.3 through 5.14, plots of the magnitude, phase and coherence of the six transfer functions, H_{c1}, H_{c2}, H_{r1}, H_{r2}, H_{s1} and H_{s2}, are given for the twelve 'Normal' subjects. Note that H_{s2} was not calculated for subject 8 due to the limited duration of the final RIB segment.

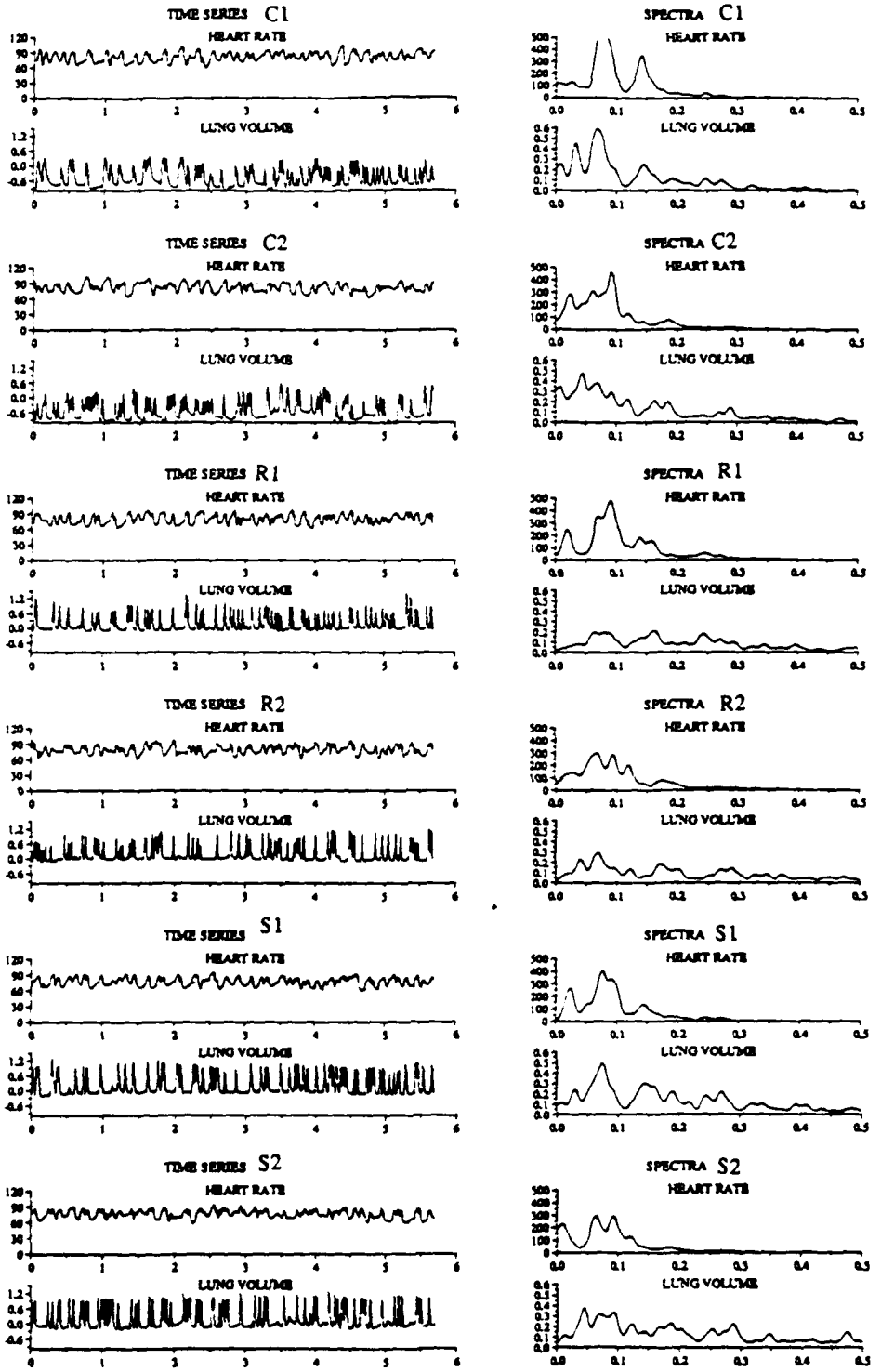


Figure 5.2: Six paired IHR-ILV data segments and their associated power spectra (Subject 1). Note random breathing pattern and effective broadening of ILV spectra. IHR time series are in units bpm vs minutes and spectra are bpm^2 vs Hz. ILV time series are in units liters vs minutes and spectra are $liters^2$ vs Hz.

5.4 Group Average Transfer Functions

Group average transfer function plots with error bars calculated from the data from the 'Normal' group are given in Figure 5.15. Figure 5.15a is the group average for the non-sick, stationary RIB segment. Figure 5.15b is the group average for the non-sick, rotating RIB segment. Figure 5.15c is the group average for the motion sick, rotating RIB segment. The transfer function magnitudes are not significantly different from one another over any band of frequencies between 0.0 and 0.5 Hz. The transfer function phases appear to differ only over the frequency band of 0.0 to 0.03 Hz. Specifically, the transfer function from the motion sick condition tends to drop off from 0.0 radians toward -2.0 radians while the non-sick cases tend to increase from -1.5 radians toward -1.0 radians and then drop back toward -2.0 radians. Note that these trends are simply approaches from different directions toward essentially zero phase. Furthermore, at these very low frequencies, the transfer function estimates are generally not reliable as they are associated with low coherence. Thus, the group average transfer function estimates are not significantly changed due to rotation or motion sickness.

5.5 Comparison of Transfer Functions for Individual Subjects

In Figures 5.16-5.26, C_r (A) and C_s (B) (Equations 4.14 and 4.15) are plotted as a function of frequency and their distributions are plotted as histograms for 11 of the 12 subjects in the 'Normal' group. (C_s and C_r were not calculated for Subject 8 due to the incomplete data set.) Plots are shown for two different assumptions: (1) that independent frequencies are separated by 7 discrete samples (0.01 Hz) and (2) that independent frequencies are separated by 3 discrete samples (0.044 Hz). No significant differences in the appearance of C_r or C_s are evident between the two cases. Thus, within reasonable limits the analysis is not sensitive to this choice.

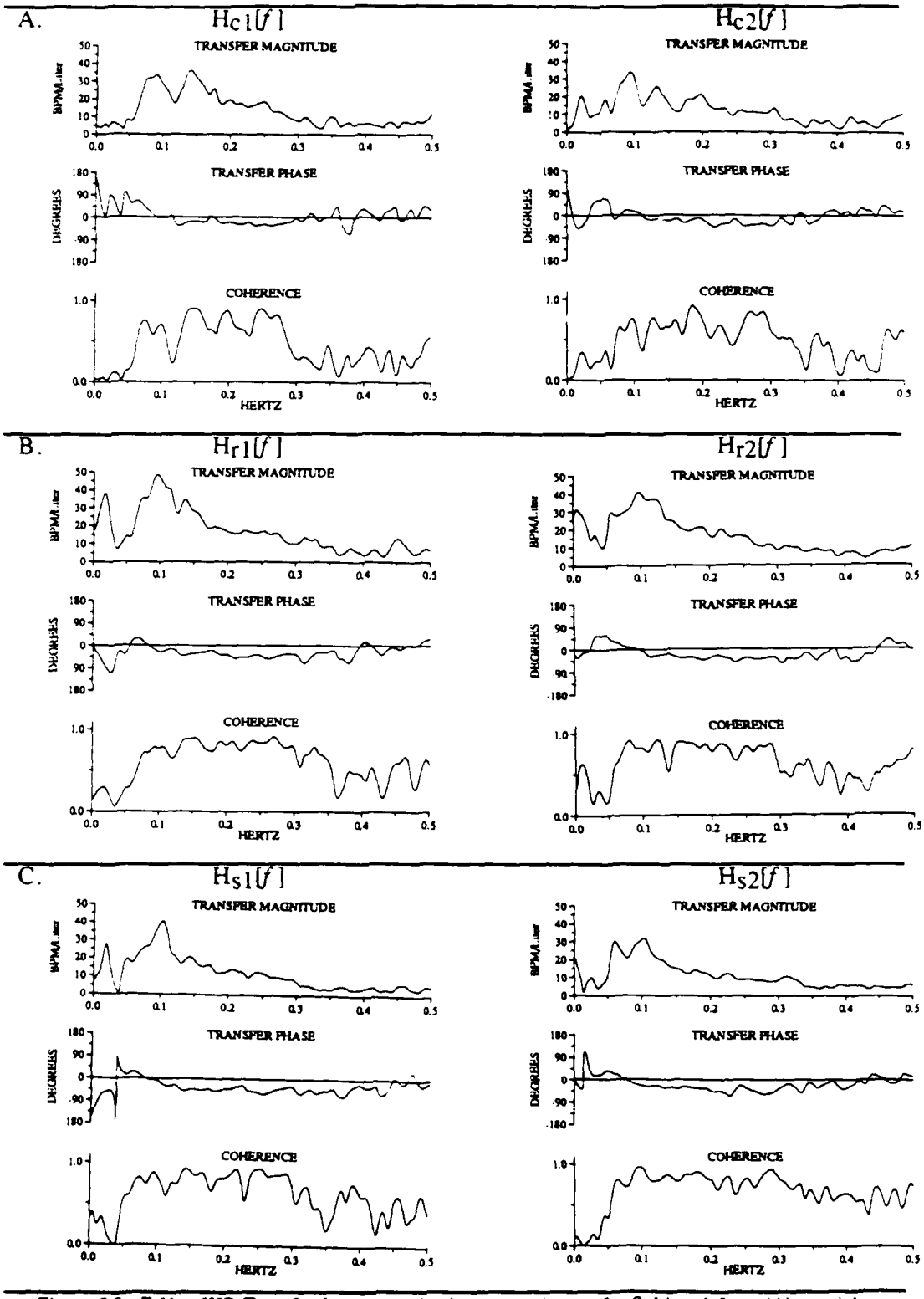


Figure 5.3: ILV to IHR Transfer function and coherence estimates for Subject 1 from (A) nonsick stationary RIB segment (B) nonsick rotating RIB segment and (C) motion sick rotating segment

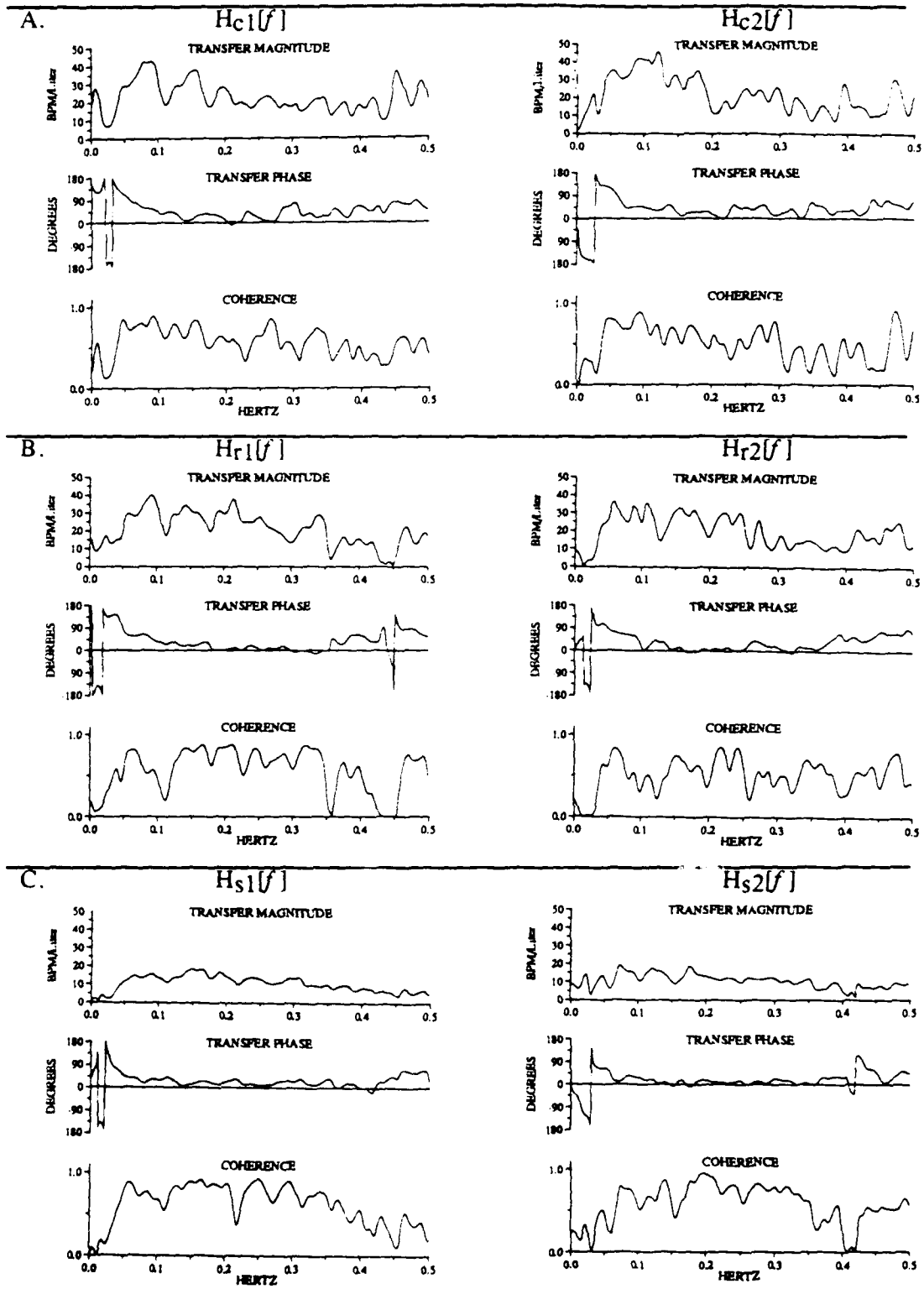


Figure 5.4: ILV to IHR Transfer function and coherence estimates for Subject 2 from (A) nonsick stationary RIB segment (B) nonsick rotating RIB segment and (C) motion sick rotating segment

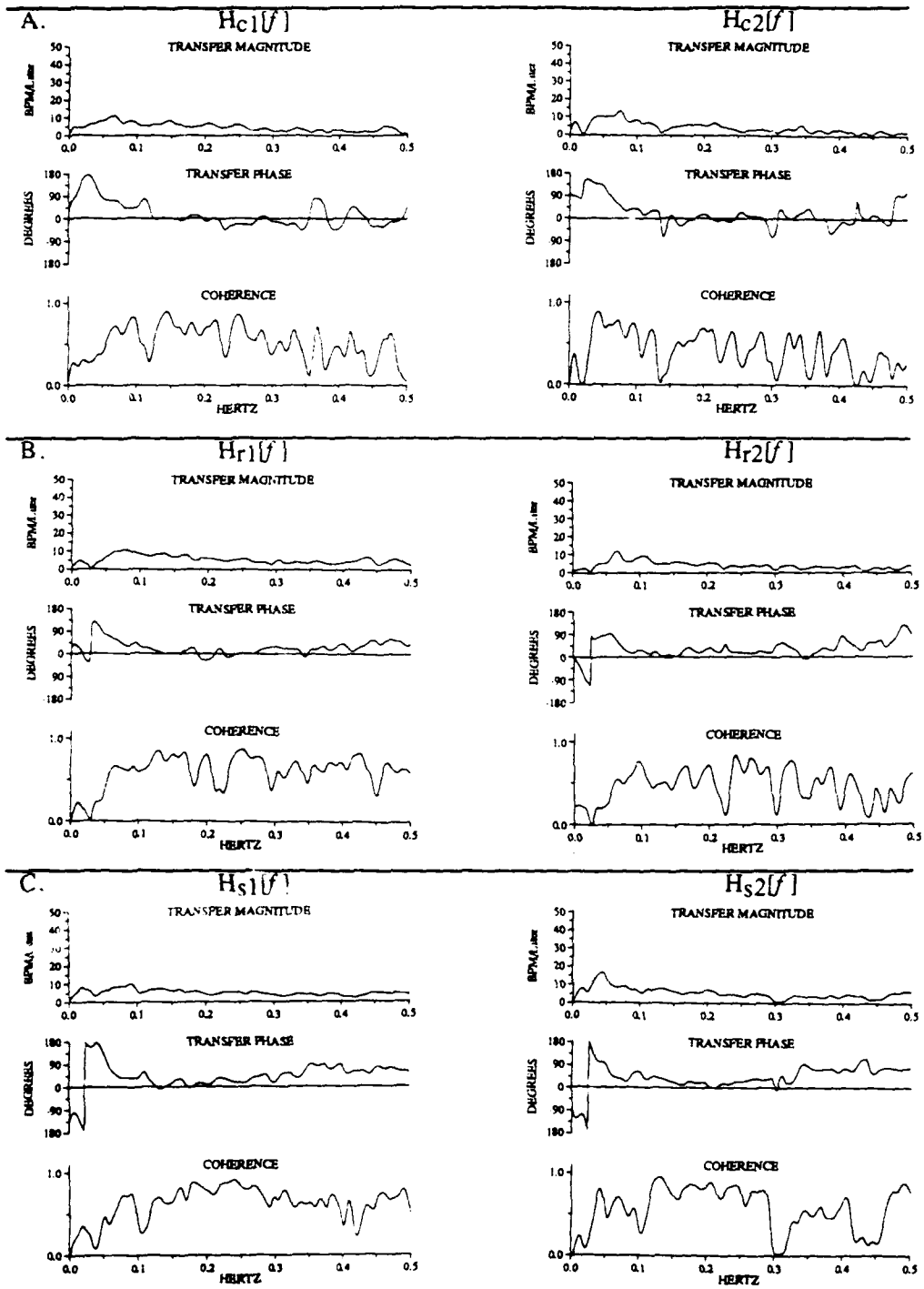


Figure 5.5: ILV to IHR Transfer function and coherence estimates for Subject 3 from (A) nonsick stationary RIB segment (B) nonsick rotating RIB segment and (C) motion sick rotating segment

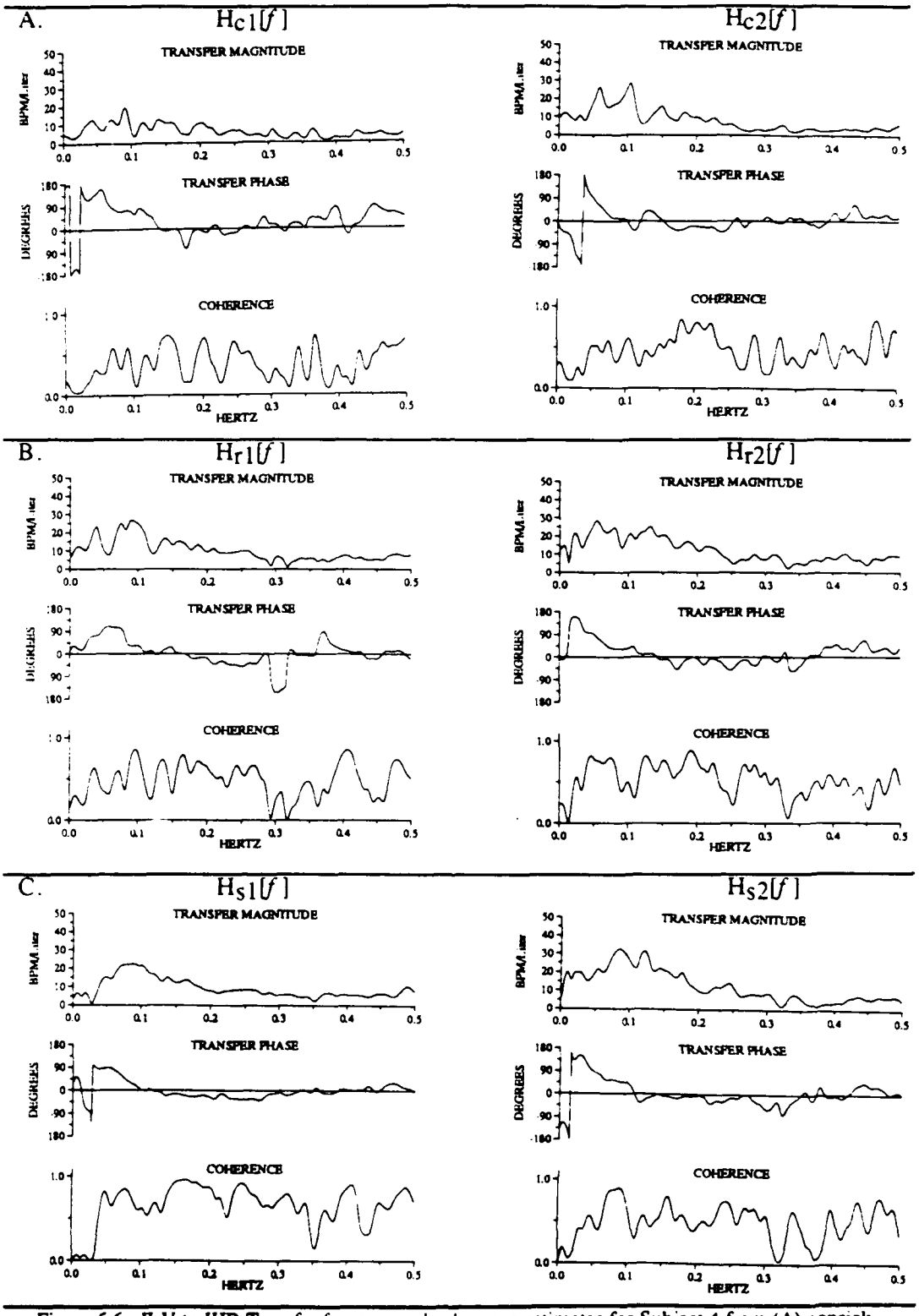


Figure 5.6: ILV to IHR Transfer function and coherence estimates for Subject 4 from (A) nonsick stationary RIB segment (B) nonsick rotating RIB segment and (C) motion sick rotating segment

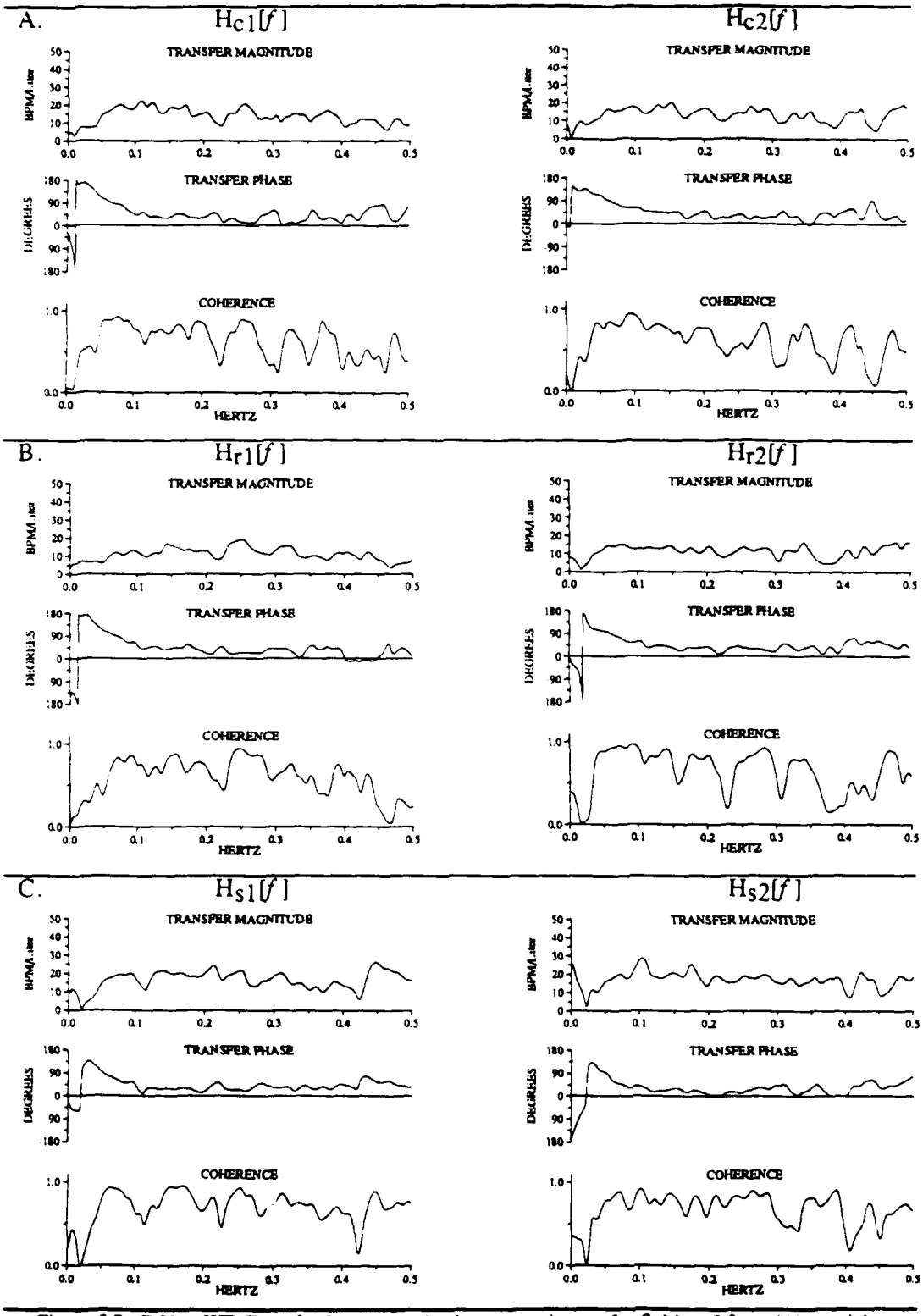


Figure 5.7: ILV to IHR Transfer function and coherence estimates for Subject 5 from (A) nonsick stationary RIB segment (B) nonsick rotating RIB segment and (C) motion sick rotating segment

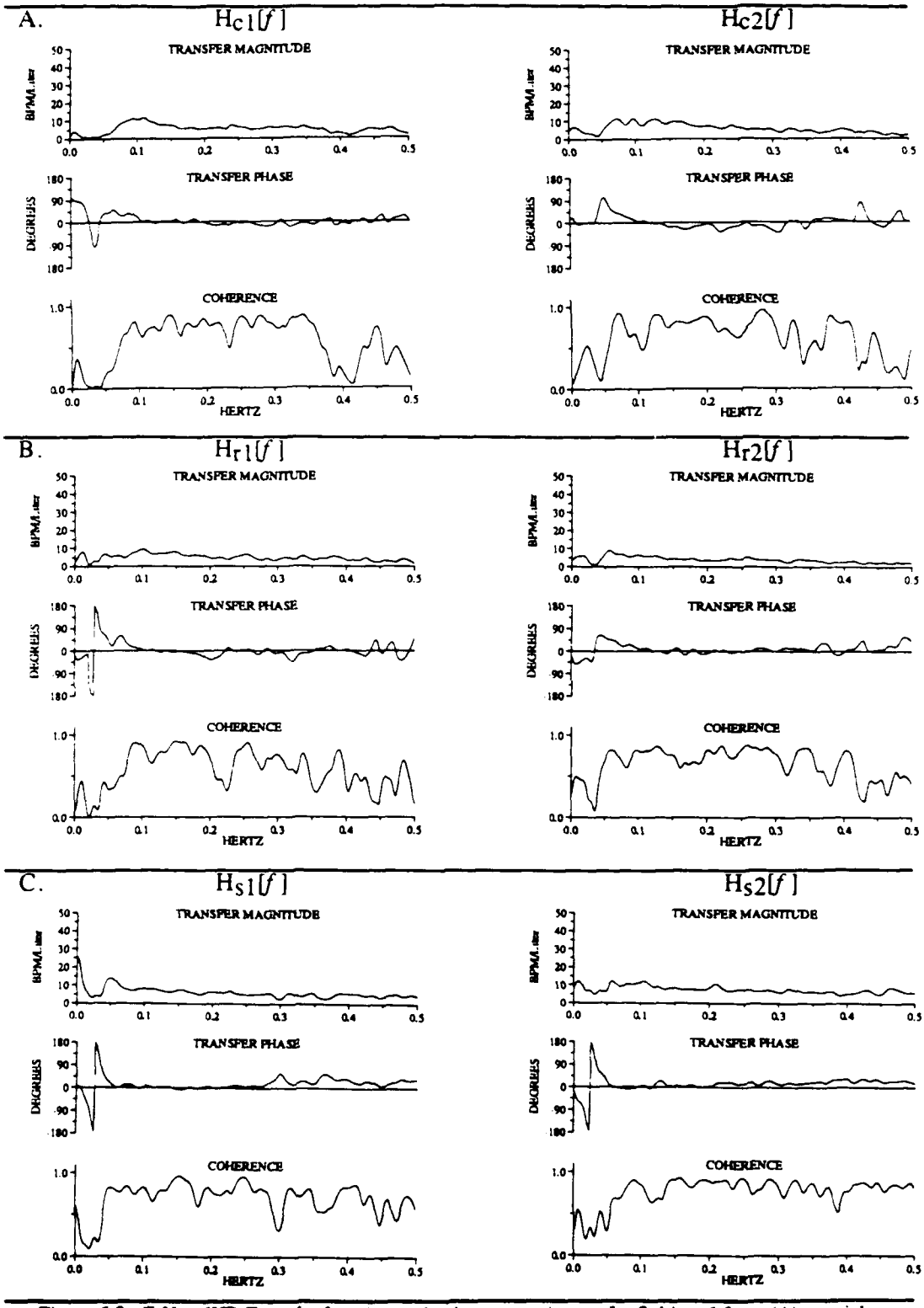


Figure 5.8: ILV to IHR Transfer function and coherence estimates for Subject 6 from (A) nonsick stationary RIB segment (B) nonsick rotating RIB segment and (C) motion sick rotating segment

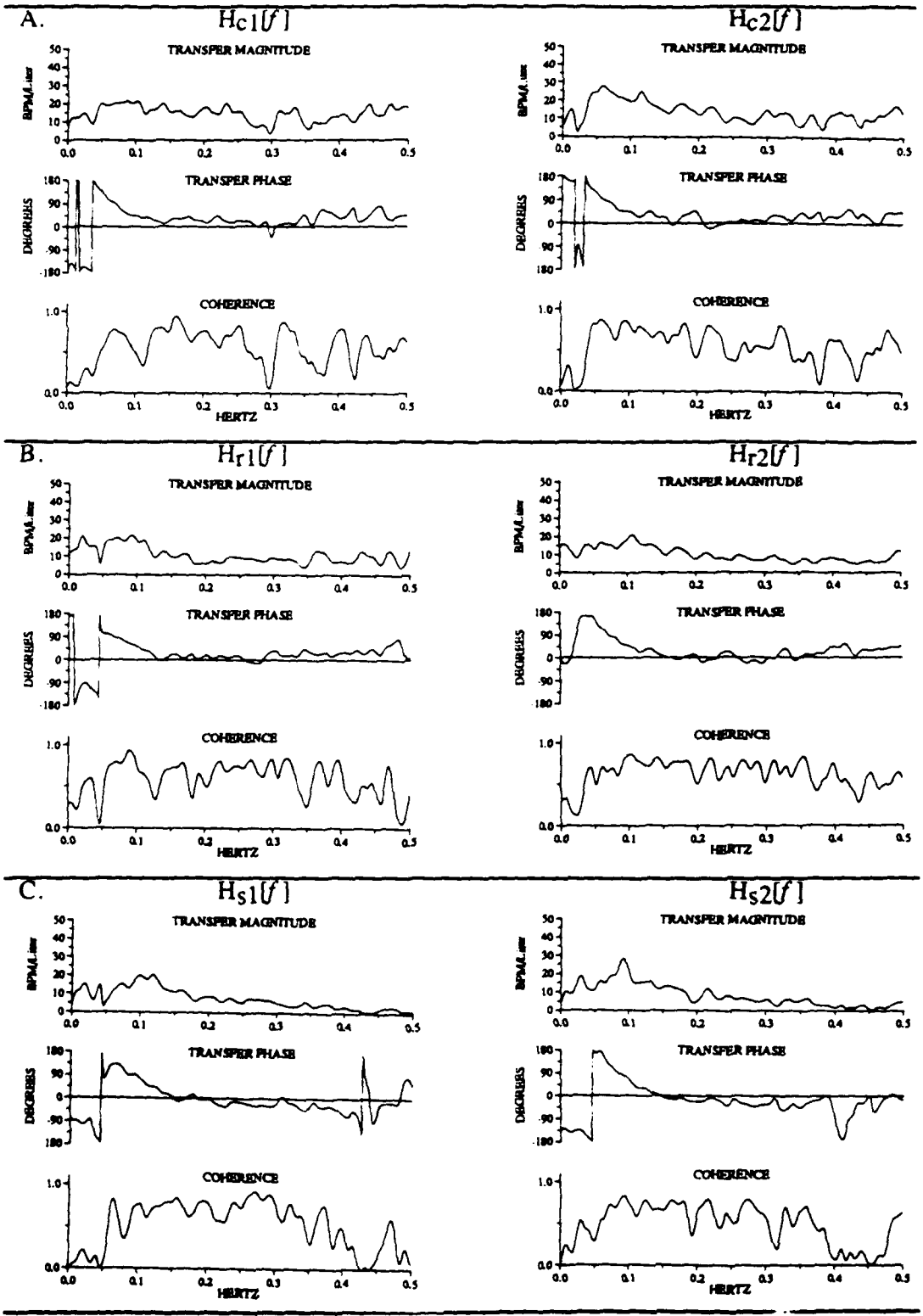


Figure 5.9: ILV to IHR Transfer function and coherence estimates for Subject 7 from (A) nonsick stationary RIB segment (B) nonsick rotating RIB segment and (C) motion sick rotating segment

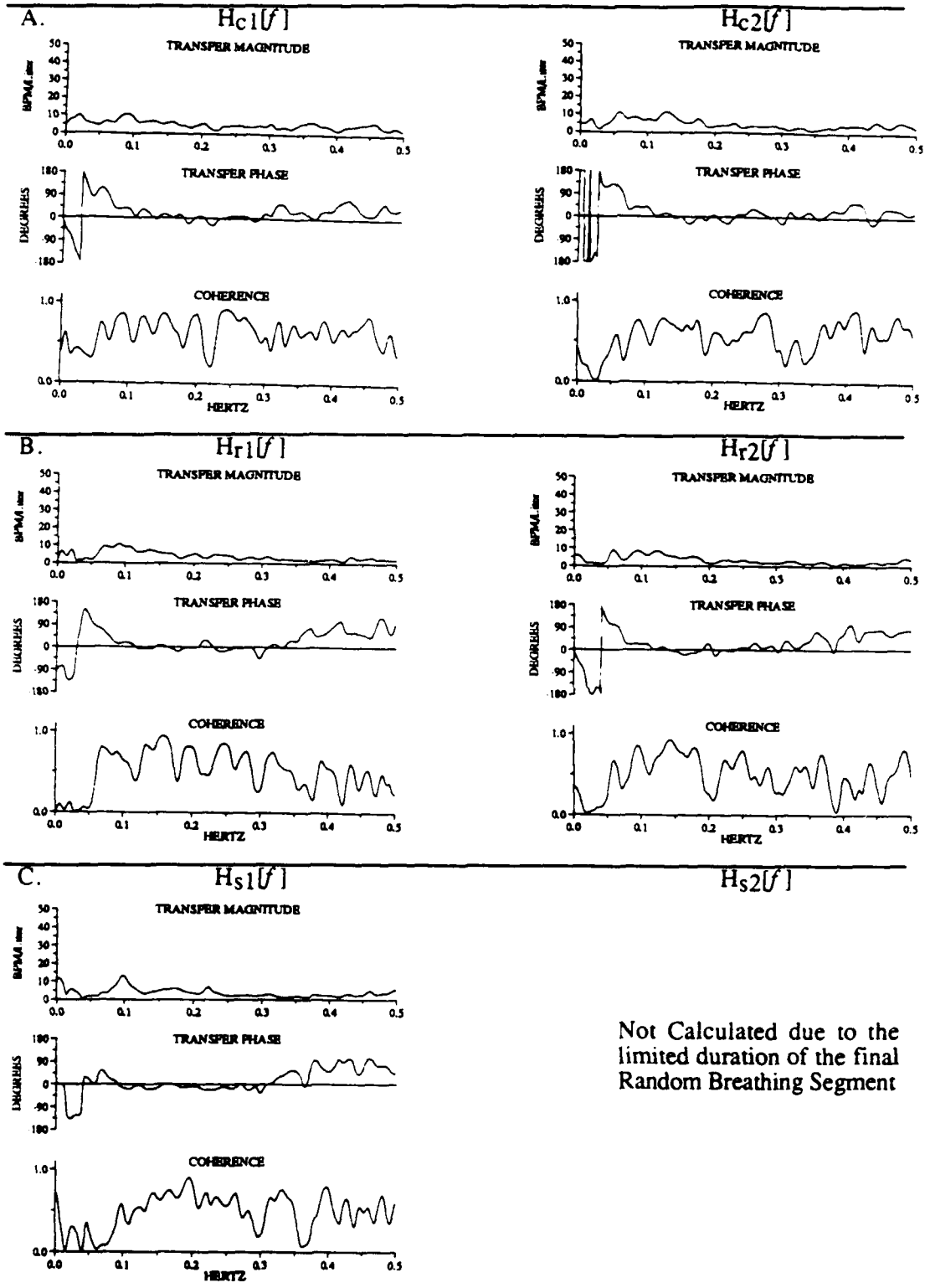


Figure 5.10: ILV to IHR Transfer function and coherence estimates for Subject 8 from (A) nonsick stationary RIB segment (B) nonsick rotating RIB segment and (C) motion sick rotating segment

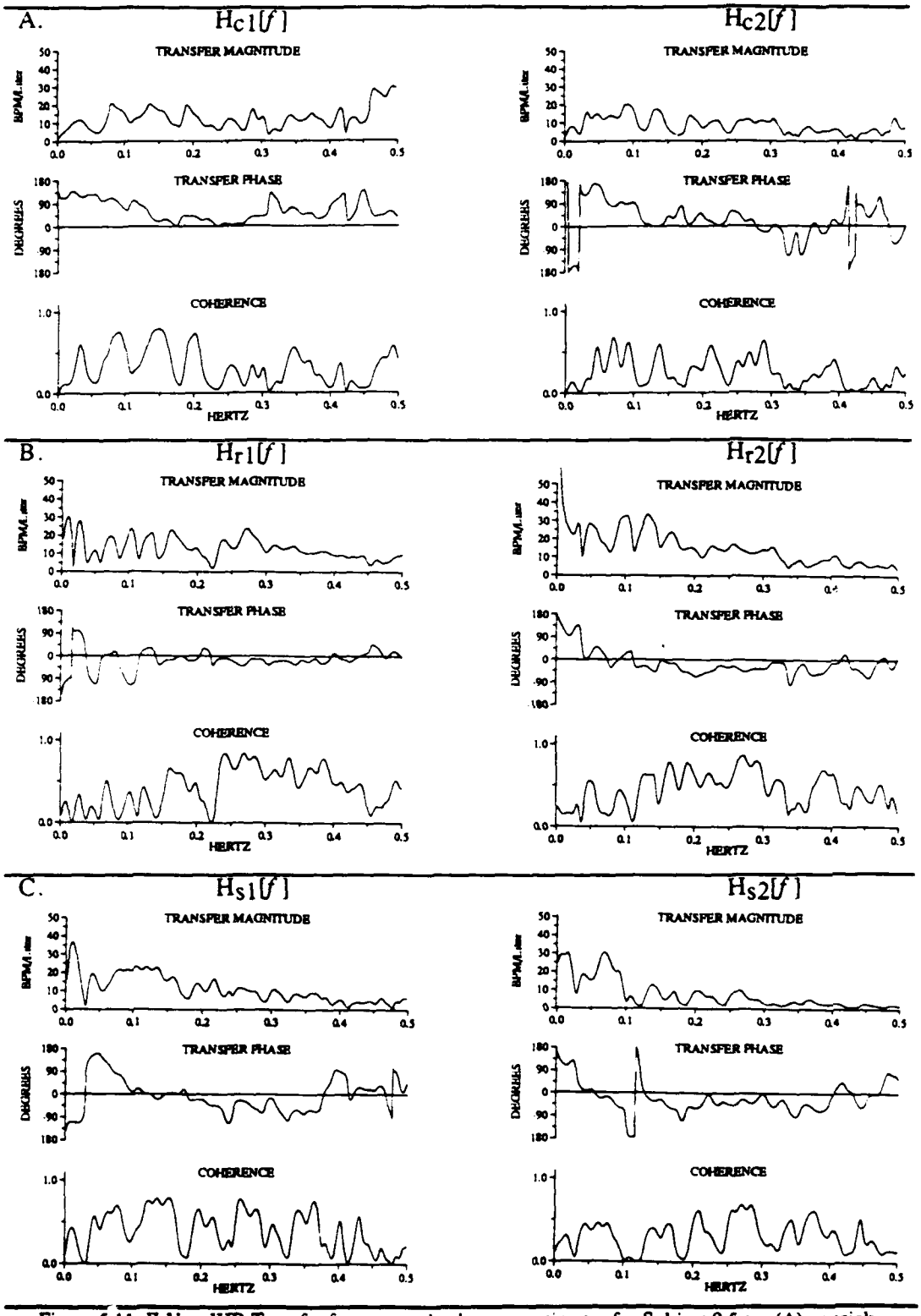


Figure 5.11: ILV to IHR Transfer function and coherence estimates for Subject 9 from (A) nonsick stationary RIB segment (B) nonsick rotating RIB segment and (C) motion sick rotating segment

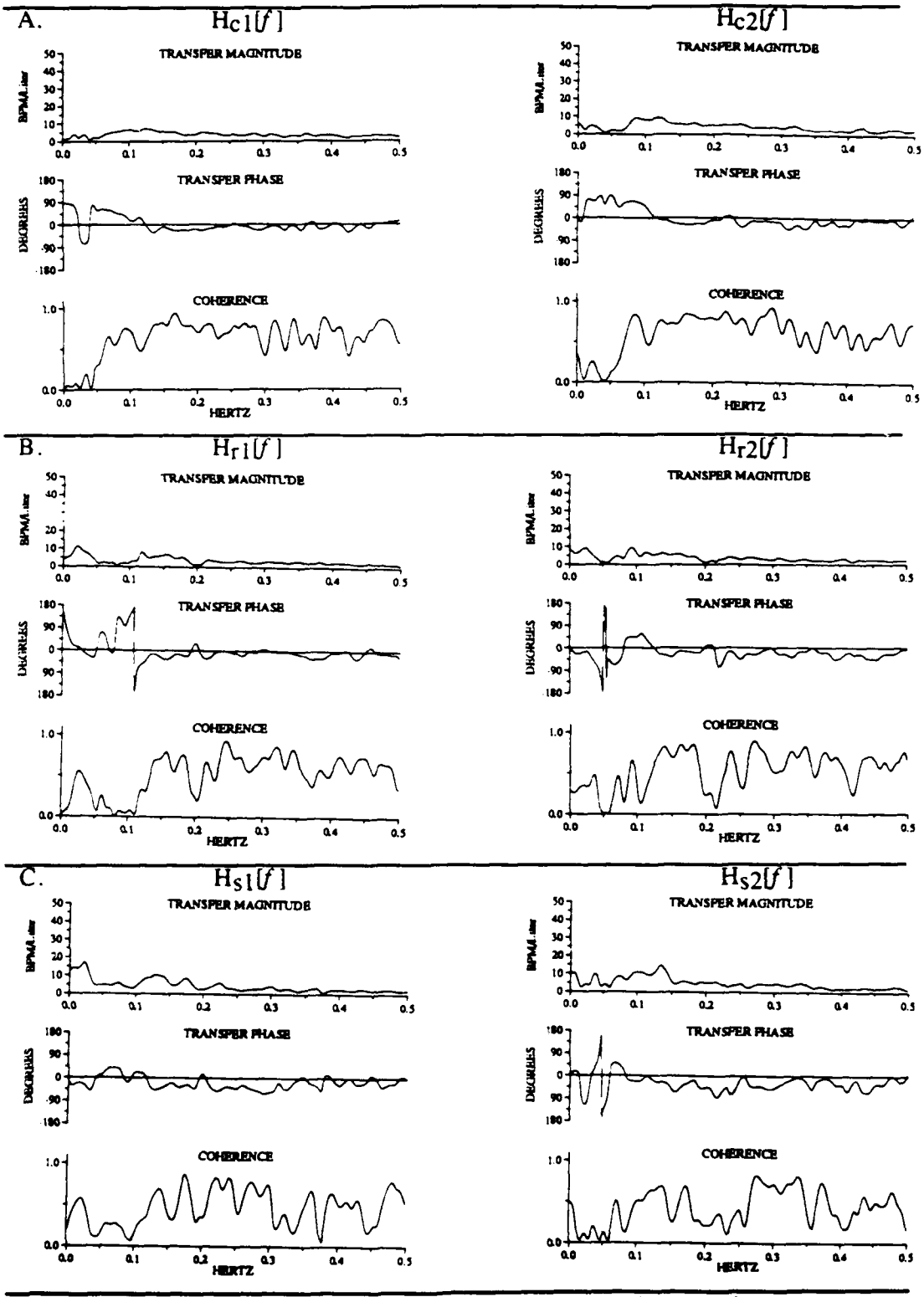


Figure 5.12: ILV to IHR Transfer function and coherence estimates for Subject 10 from (A) nonsick stationary RIB segment (B) nonsick rotating RIB segment and (C) motion sick rotating segment

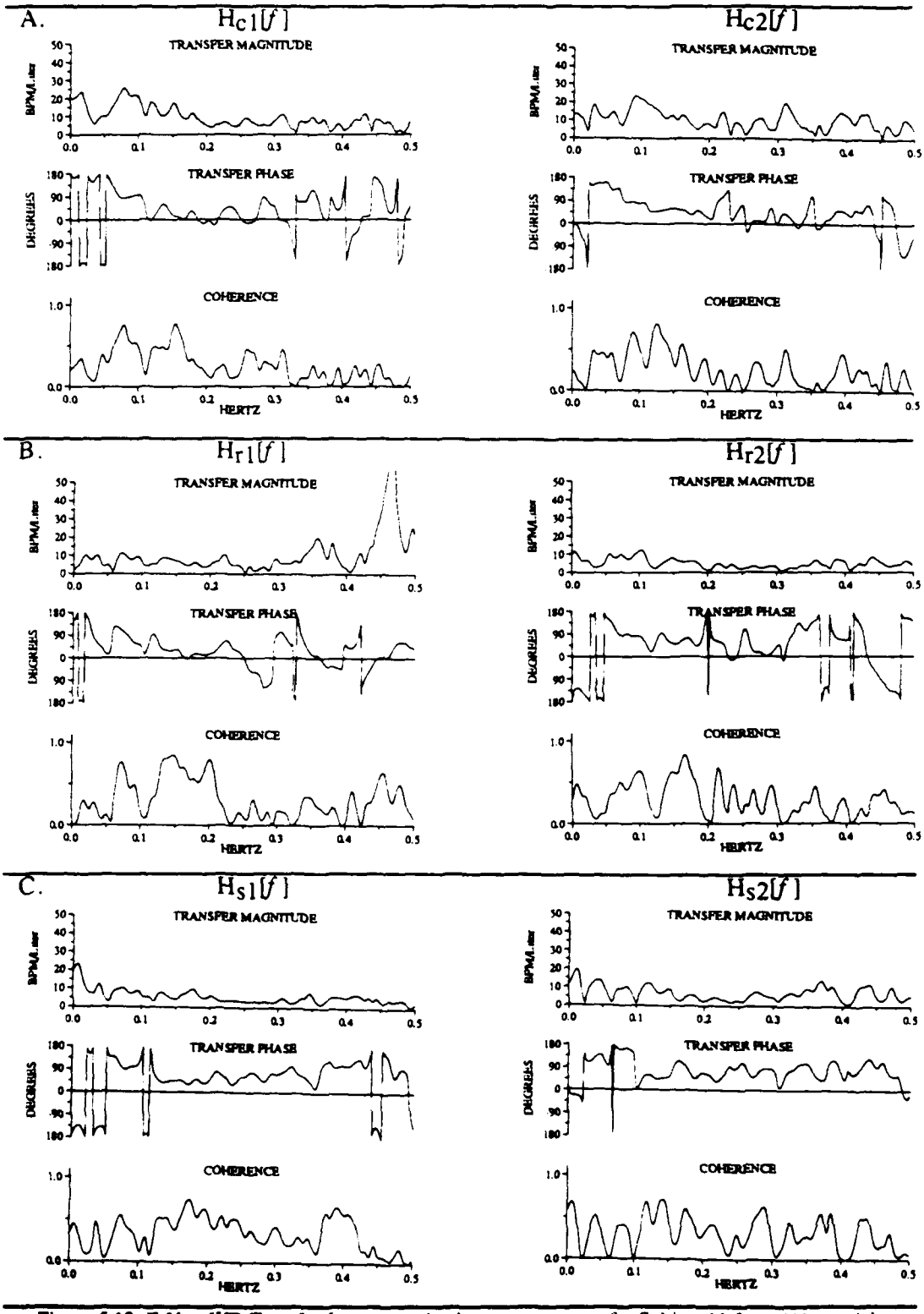


Figure 5.13: ILV to IHR Transfer function and coherence estimates for Subject 11 from (A) nonsick stationary RIB segment (B) nonsick rotating RIB segment and (C) motion sick rotating segment

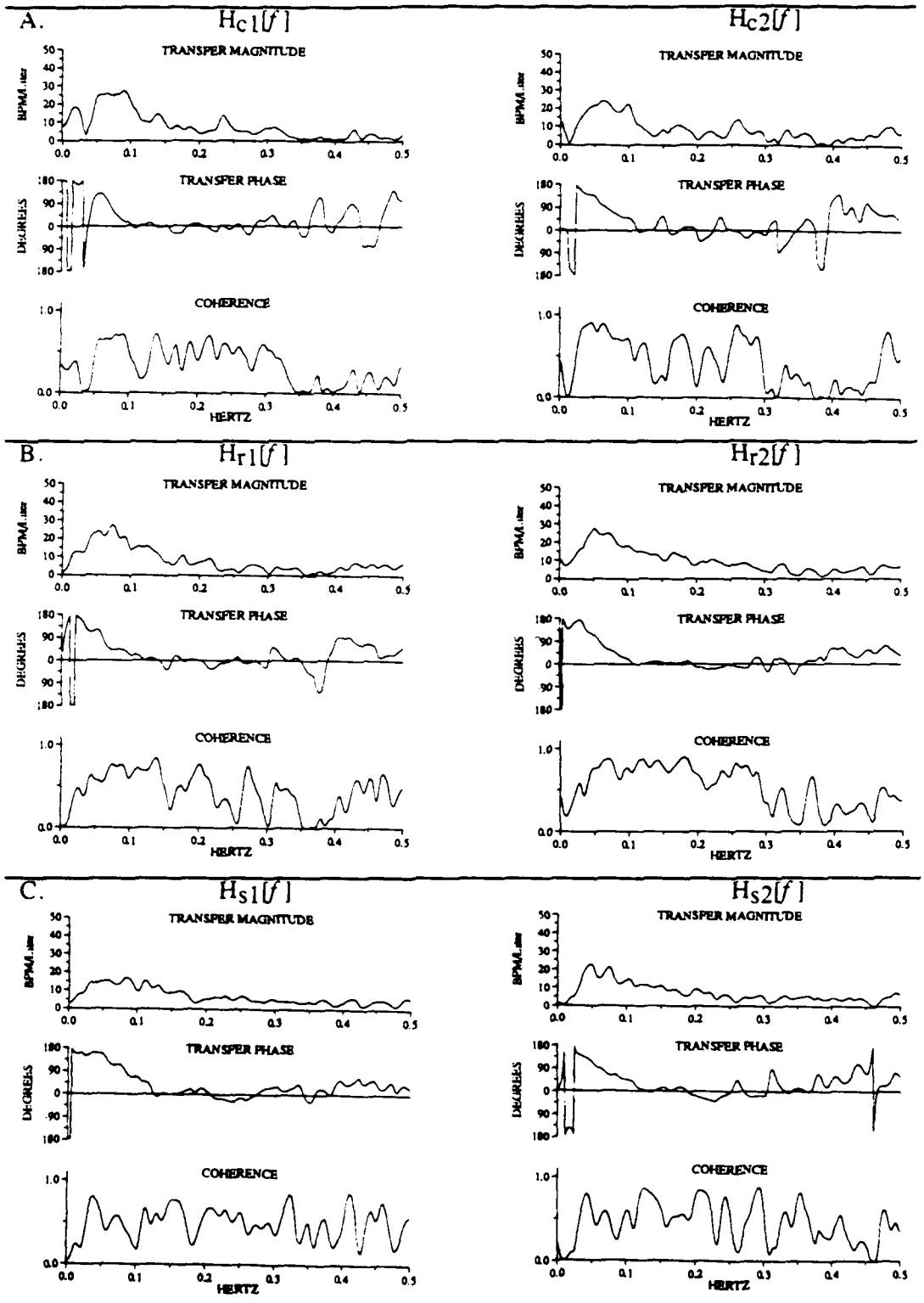
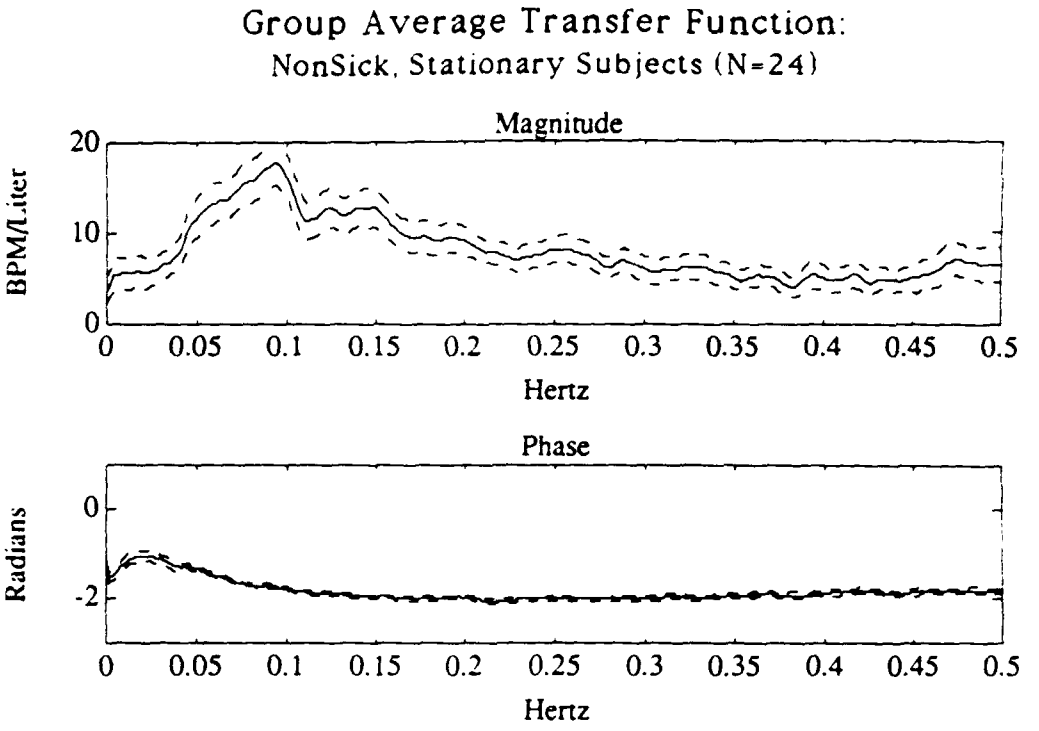
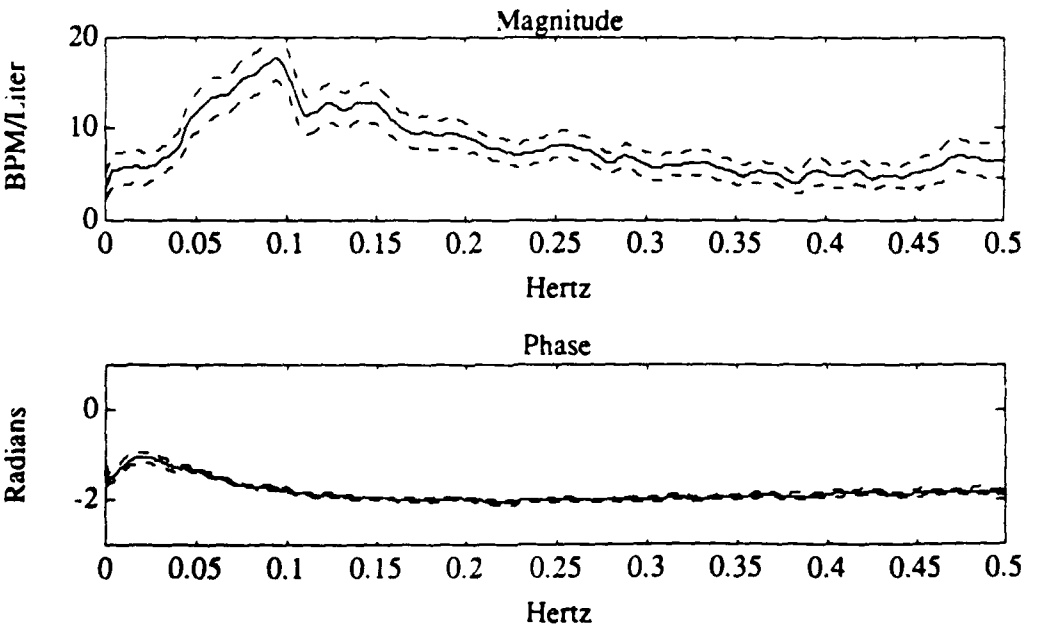


Figure 5.14: ILV to IHR Transfer function and coherence estimates for Subject 12 from (A) nonsick stationary RIB segment (B) nonsick rotating RIB segment and (C) motion sick rotating segment

A


Group Average Transfer Function:
NonSick, Stationary Subjects (N=24)
Magnitude
BPM/Liter
20
10
0
0 0.05 0.1 0.15 0.2 0.25 0.3 0.35 0.4 0.45 0.5
Hertz
Phase
Radians
0
-2
0 0.05 0.1 0.15 0.2 0.25 0.3 0.35 0.4 0.45 0.5
Hertz

B


Group Average Transfer Function:
NonSick, Rotating Subjects (N=24)
Magnitude
BPM/Liter
20
10
0
0 0.05 0.1 0.15 0.2 0.25 0.3 0.35 0.4 0.45 0.5
Hertz
Phase
Radians
0
-2
0 0.05 0.1 0.15 0.2 0.25 0.3 0.35 0.4 0.45 0.5
Hertz

C

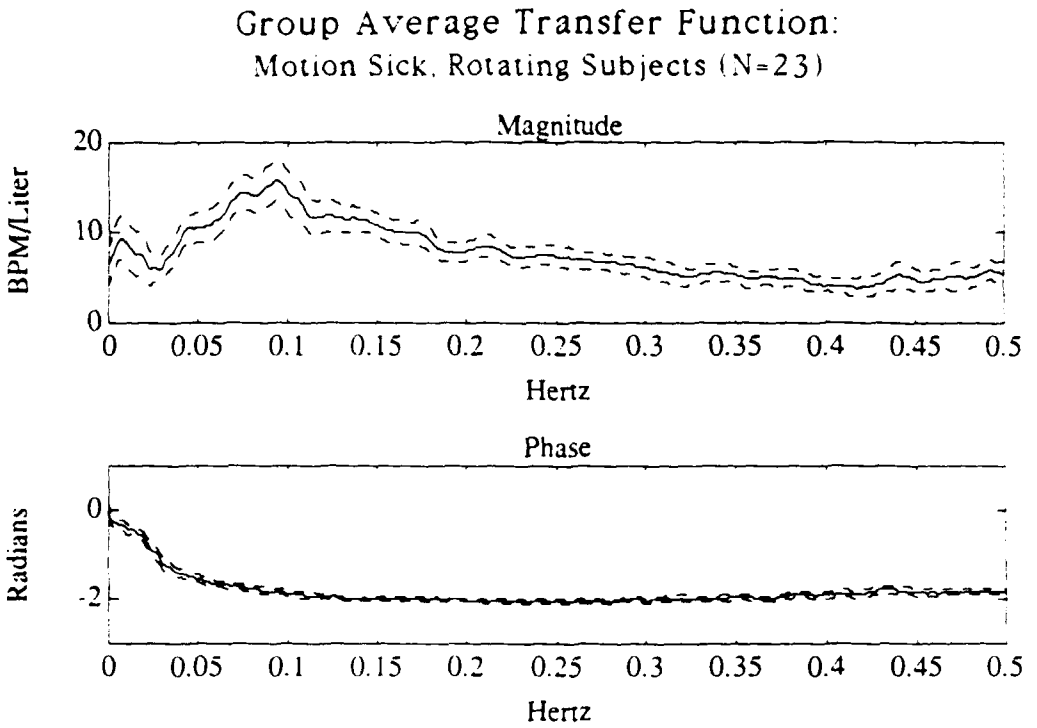


Figure 5.15: Group average transfer functions with error bars for Normal subjects during (A.) non sick stationary RIB segment (average of all Hc1[f] and Hc2[f]) (B.) non sick rotating RIB segment (average of all Hr1[f] and Hr2[f]) and (C.) motion sick rotating RIB segment (average of all Hs1[f] and Hs2[f]). (current and previous page)

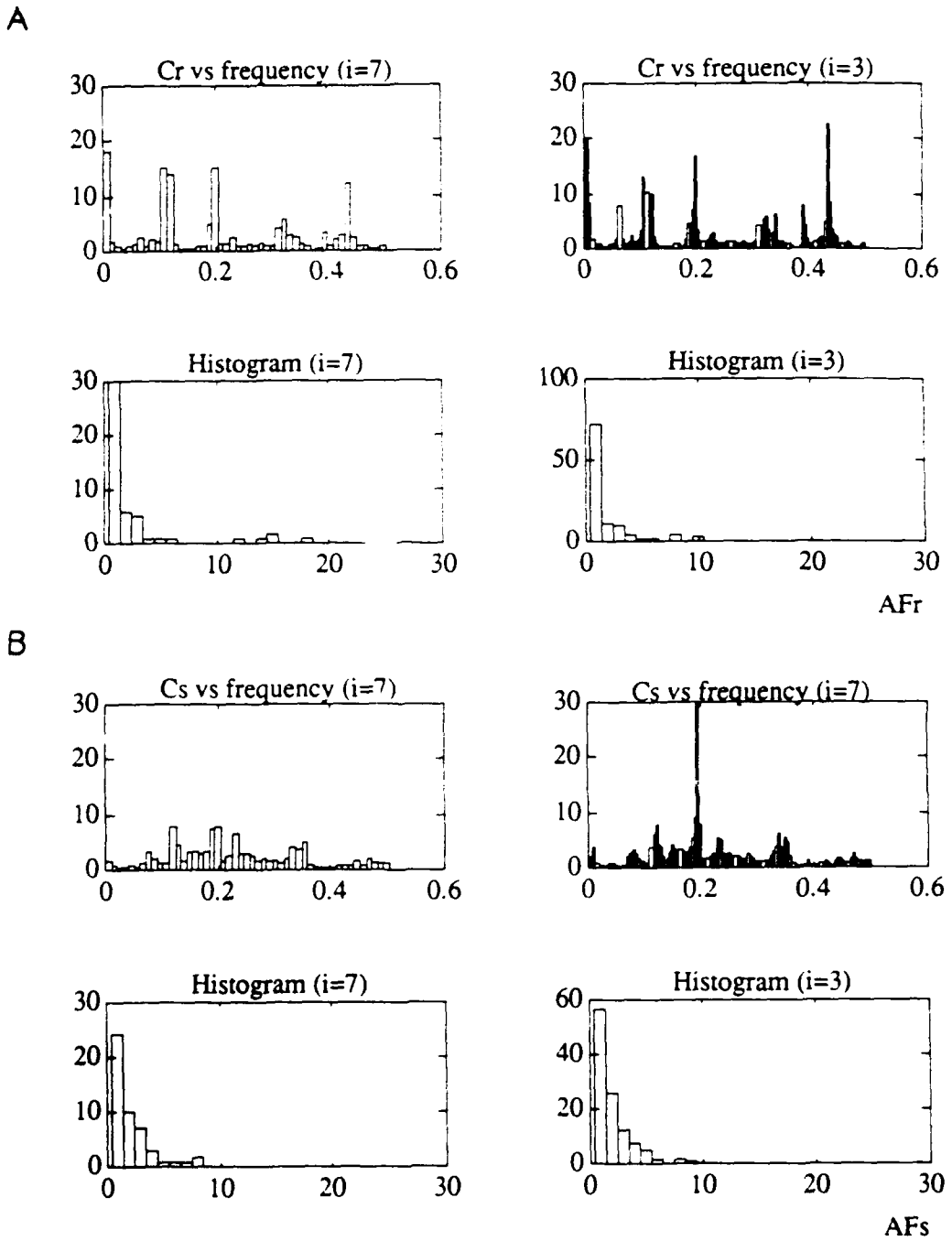


Figure 5.16: Subject 1 Statistics C_r (A) and C_s (B). plotted as functions of frequency and as histograms of magnitudes for two choices of independent frequency separations, i. Histograms have units of count per bin vs bin mean. Histograms can be normalized by dividing by the total number of values included. (49 for i=7, 114 for i=3).

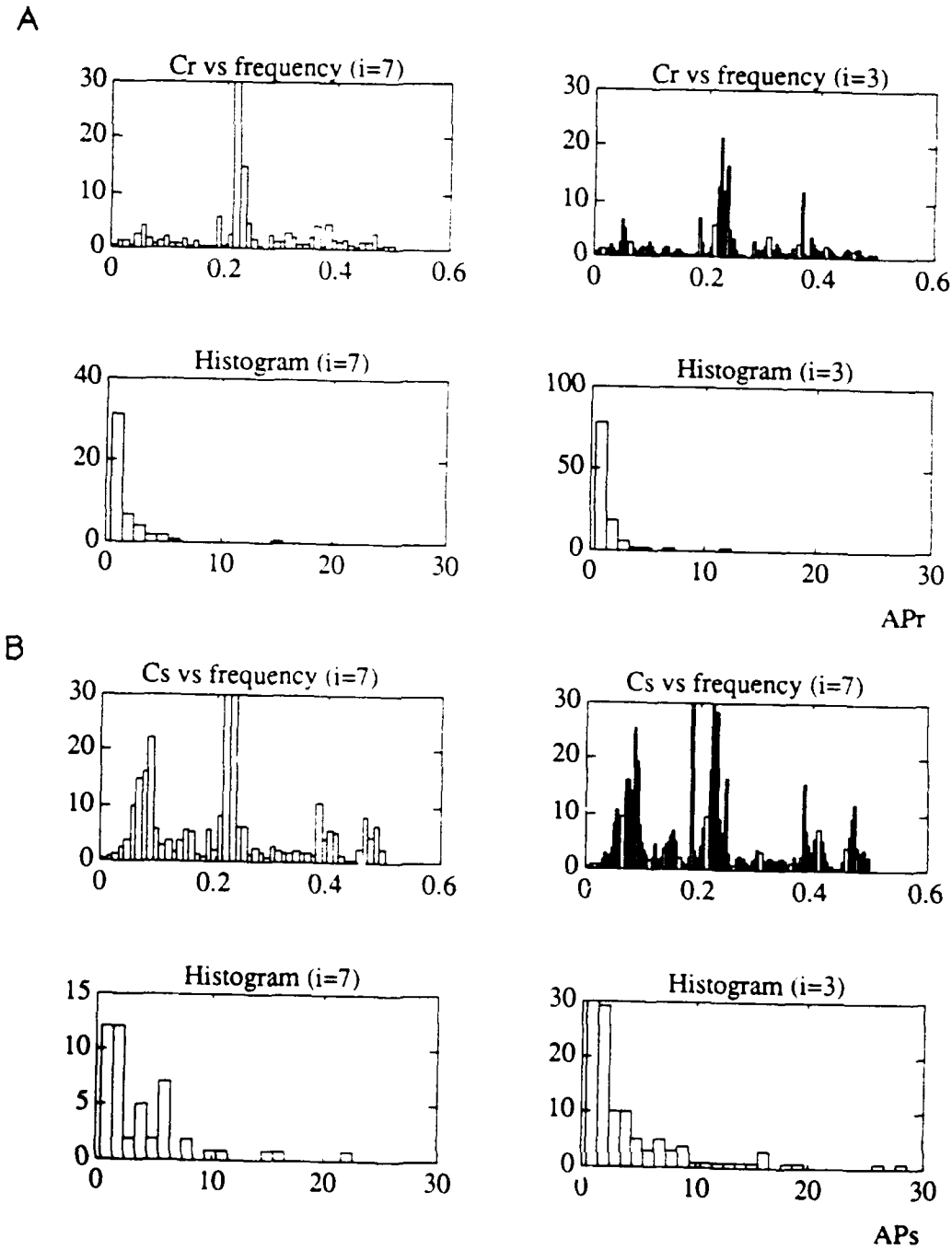


Figure 5.17: Subject 2 Statistics C_r (A) and C_s (B), plotted as functions of frequency and as histograms of magnitudes for two choices of independent frequency separations, i. Histograms have units of count per bin vs bin mean. Histograms can be normalized by dividing by the total number of values included. (49 for i=7, 114 for i=3).

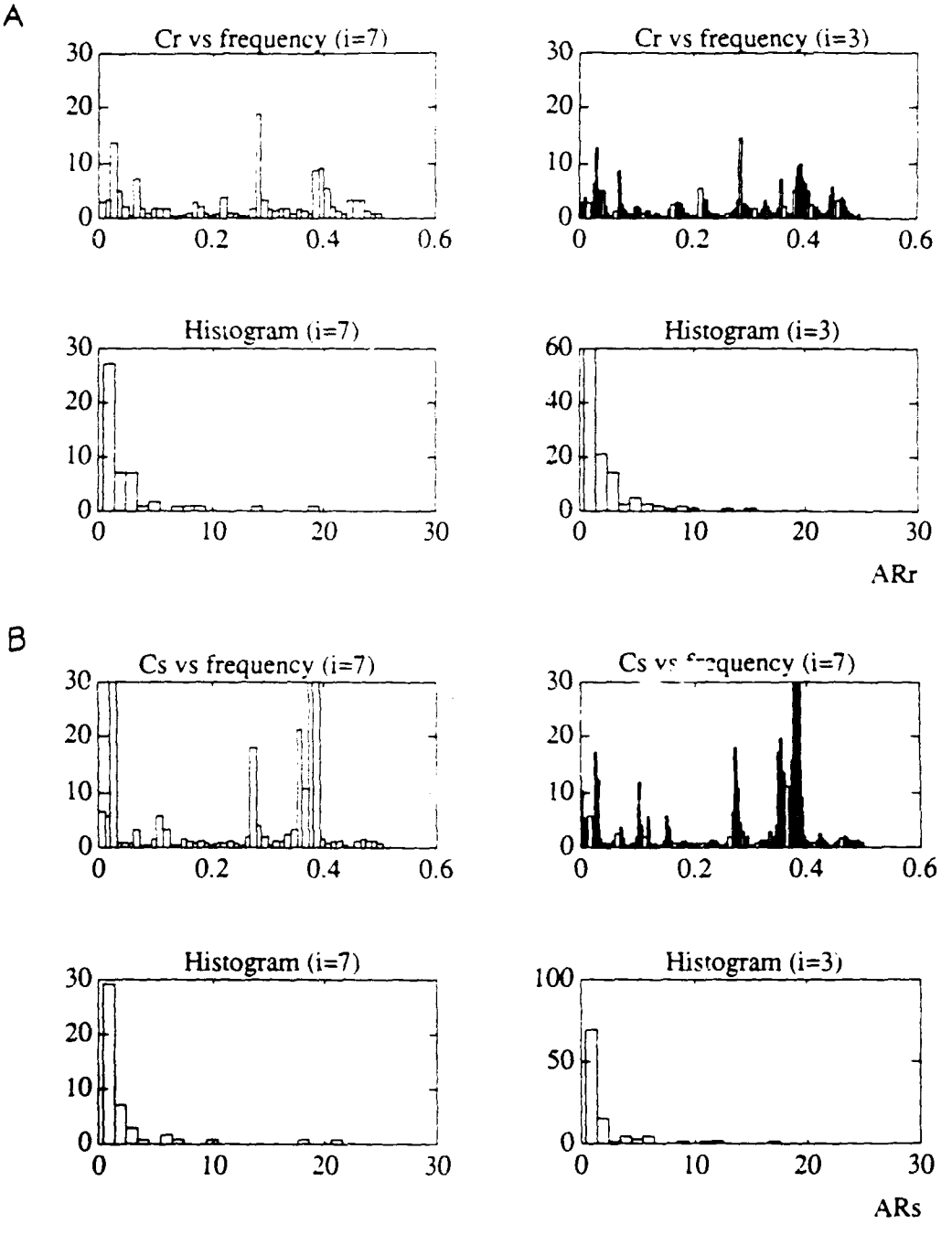


Figure 5.18: Subject 3 Statistics C_r (A) and C_s (B). plotted as functions of frequency and as histograms of magnitudes for two choices of independent frequency separations, i. Histograms have units of count per bin vs bin mean. Histograms can be normalized by dividing by the total number of values included. (49 for i=7, 114 for i=3).

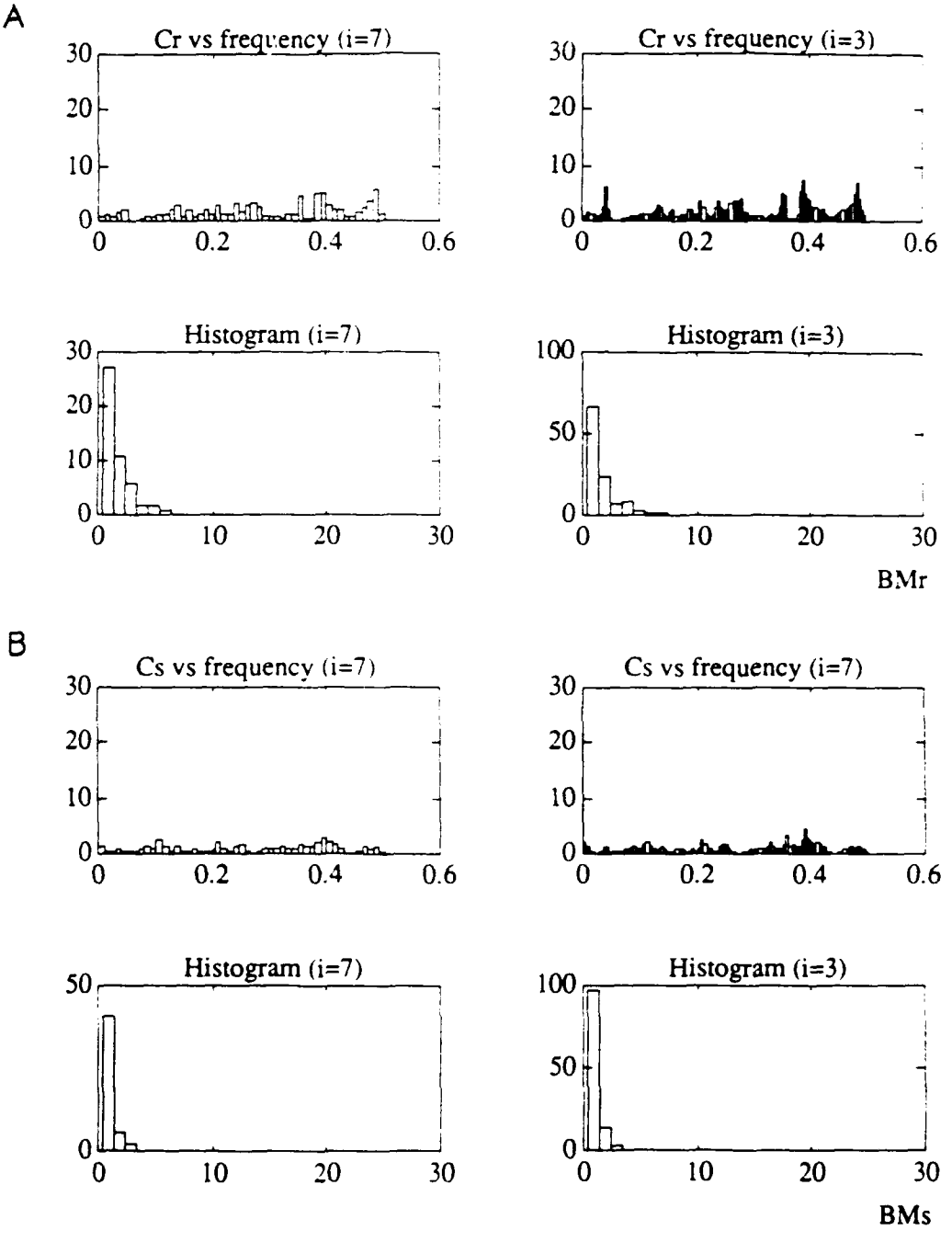


Figure 5.19: Subject 4 Statistics C_r (A) and C_s (B), plotted as functions of frequency and as histograms of magnitudes for two choices of independent frequency separations, i. Histograms have units of count per bin vs bin mean. Histograms can be normalized by dividing by the total number of values included. (49 for i=7, 114 for i=3).

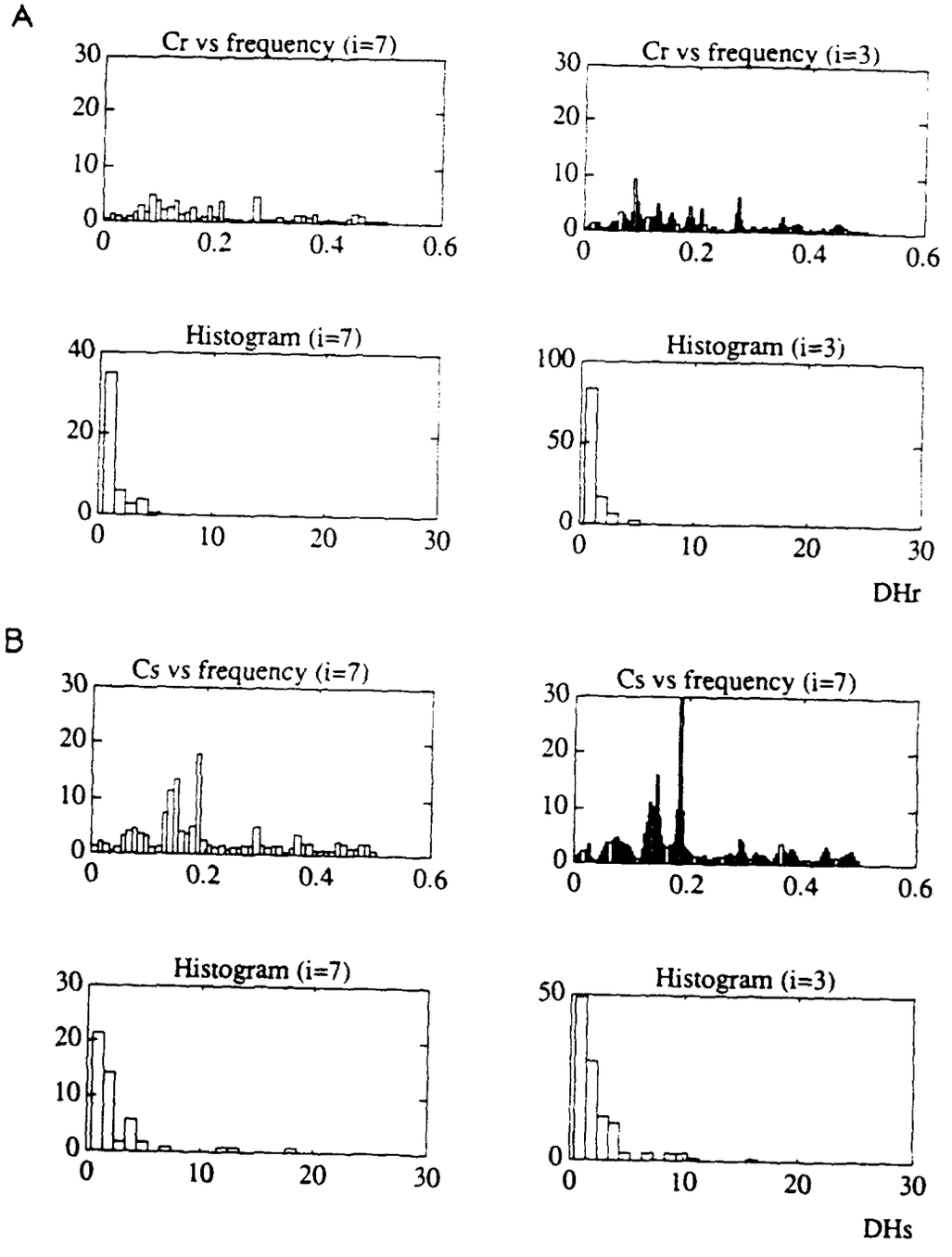


Figure 5.20: Subject 5 Statistics C_r (A) and C_s (B). plotted as functions of frequency and as histograms of magnitudes for two choices of independent frequency separations, i. Histograms have units of count per bin vs bin mean. Histograms can be normalized by dividing by the total number of values included. (49 for i=7, 114 for i=3).

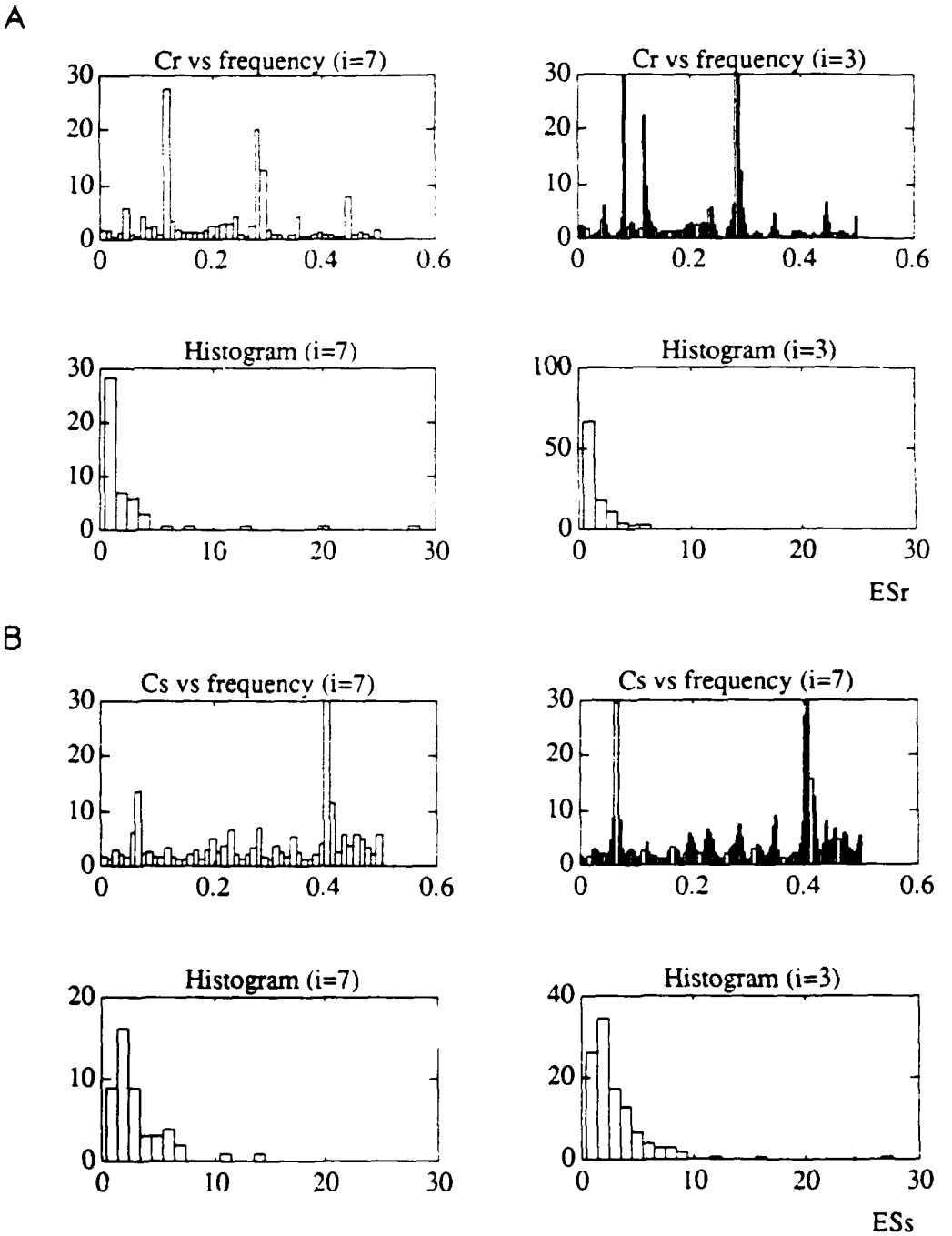


Figure 5.21: Subject 6 Statistics C_r (A) and C_s (B). plotted as functions of frequency and as histograms of magnitudes for two choices of independent frequency separations, i. Histograms have units of count per bin vs bin mean. Histograms can be normalized by dividing by the total number of values included. (49 for i=7, 114 for i=3).

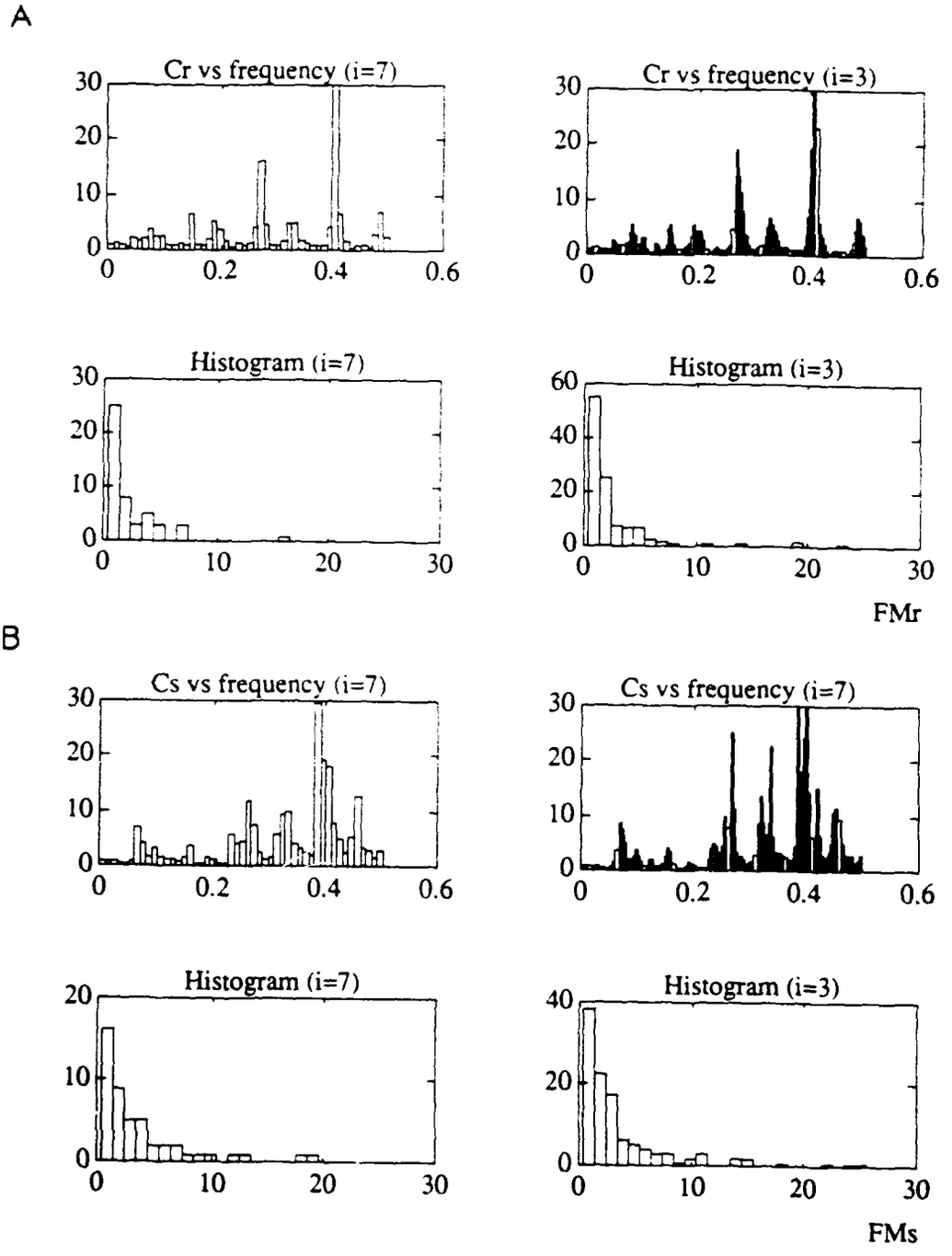


Figure 5.22: Subject 7 Statistics C_r (A) and C_s (B). plotted as functions of frequency and as histograms of magnitudes for two choices of independent frequency separations, i. Histograms have units of count per bin vs bin mean. Histograms can be normalized by dividing by the total number of values included. (49 for i=7, 114 for i=3).

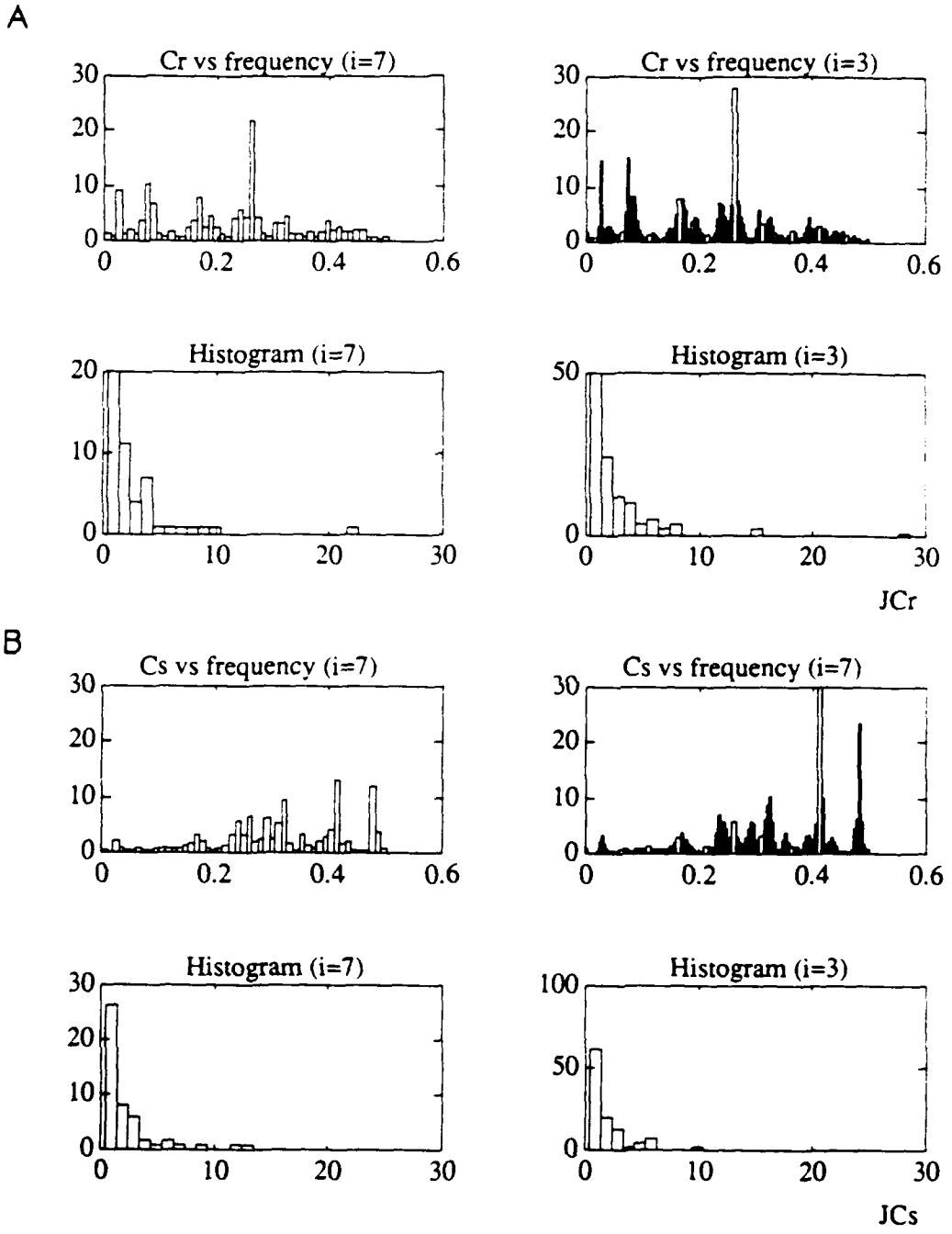


Figure 5.23: Subject 9 Statistics C_r (A) and C_s (B). plotted as functions of frequency and as histograms of magnitudes for two choices of independent frequency separations, i. Histograms have units of count per bin vs bin mean. Histograms can be normalized by dividing by the total number of values included. (49 for i=7, 114 for i=3).

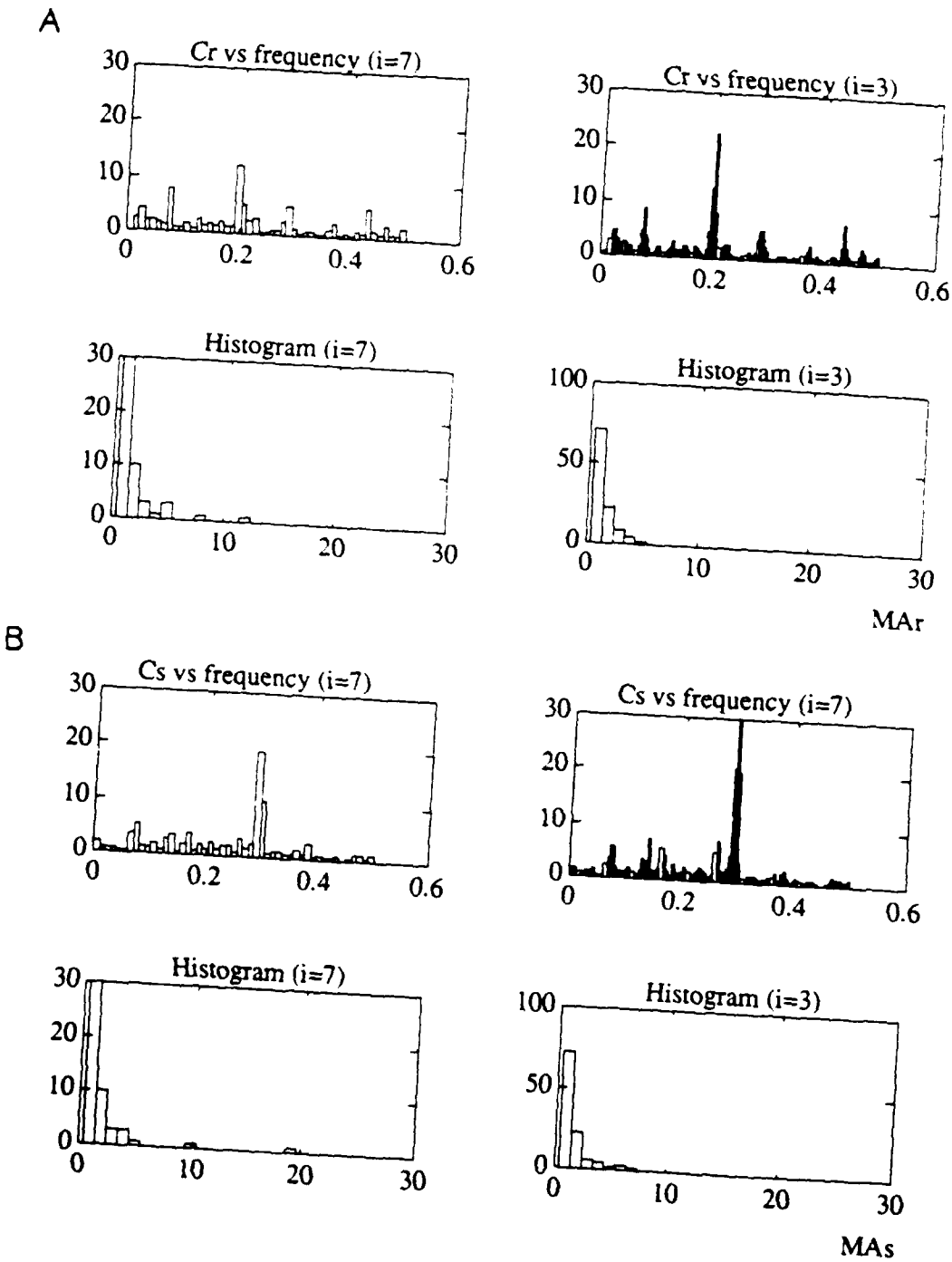


Figure 5.24: Subject 10 *Statistics* C_r (A) and C_s (B). plotted as functions of frequency and as histograms of magnitudes for two choices of independent frequency separations, i. Histograms have units of count per bin vs bin mean. Histograms can be normalized by dividing by the total number of values included. (49 for i=7, 114 for i=3).

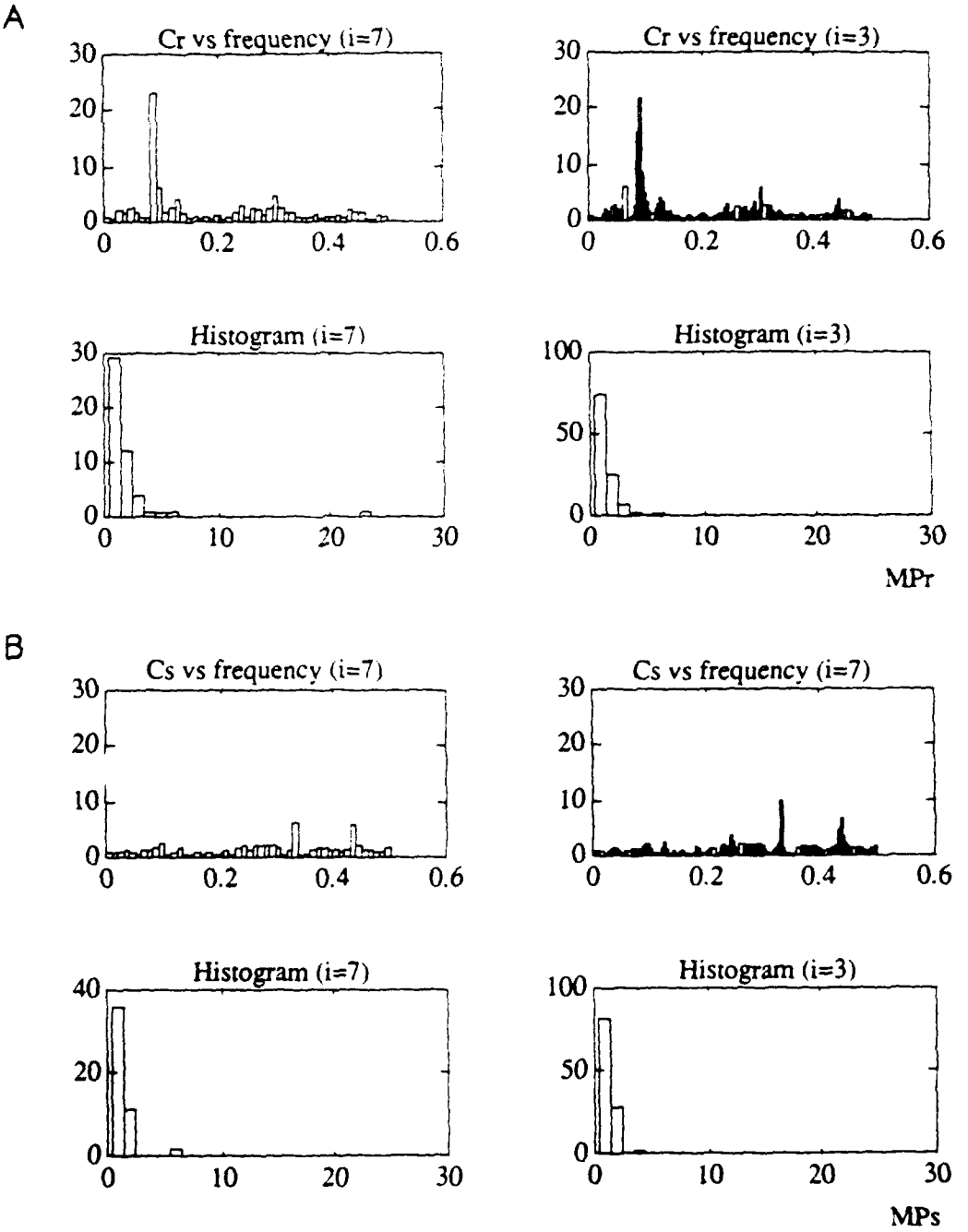


Figure 5.25: Subject 11 Statistics C_r (A) and C_s (B), plotted as functions of frequency and as histograms of magnitudes for two choices of independent frequency separations, i. Histograms have units of count per bin vs bin mean. Histograms can be normalized by dividing by the total number of values included. (49 for i=7, 114 for i=3).

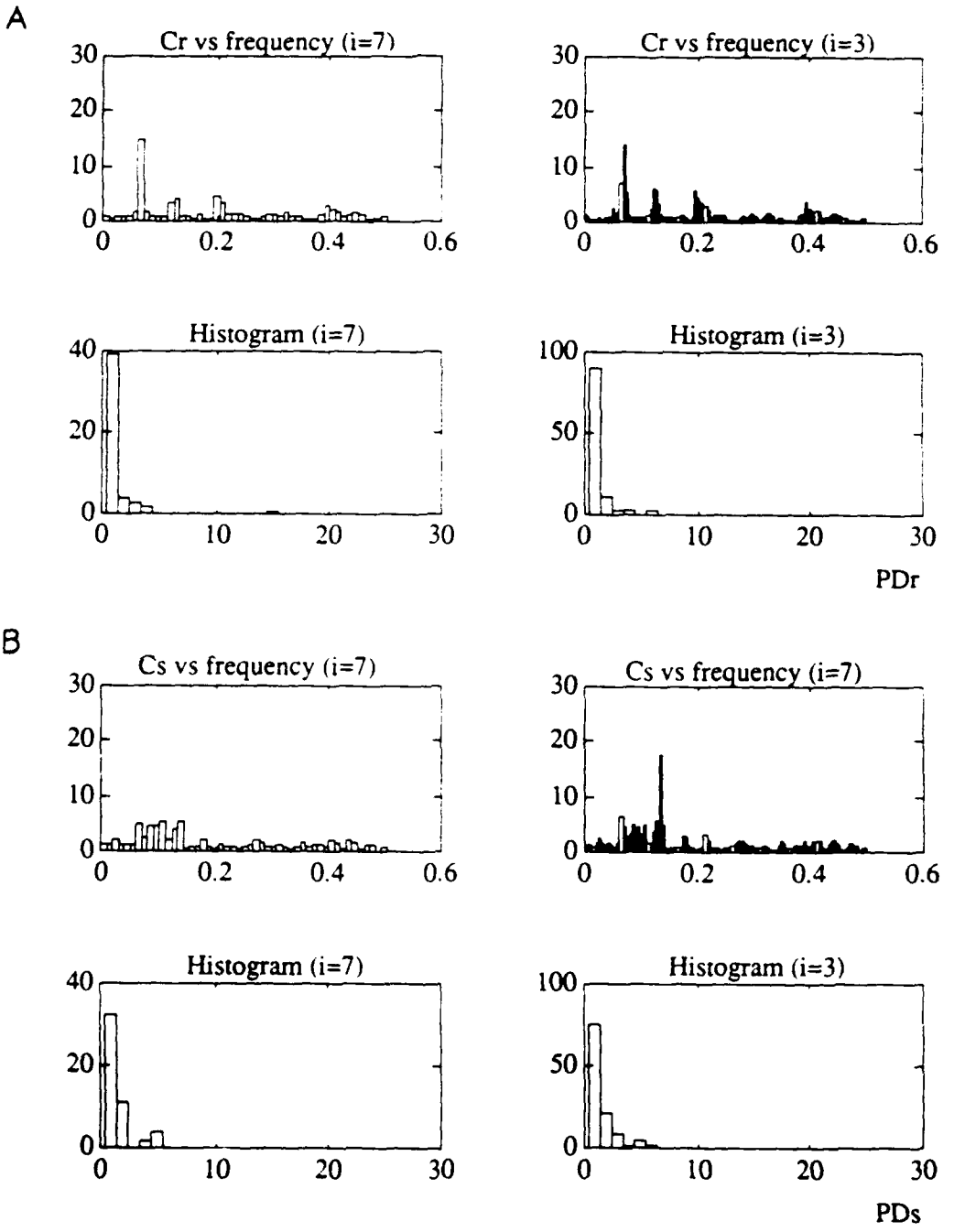


Figure 5.26: Subject 12 Statistics C_r (A) and C_s (B). plotted as functions of frequency and as histograms of magnitudes for two choices of independent frequency separations, i. Histograms have units of count per bin vs bin mean. Histograms can be normalized by dividing by the total number of values included. (49 for i=7, 114 for i=3).

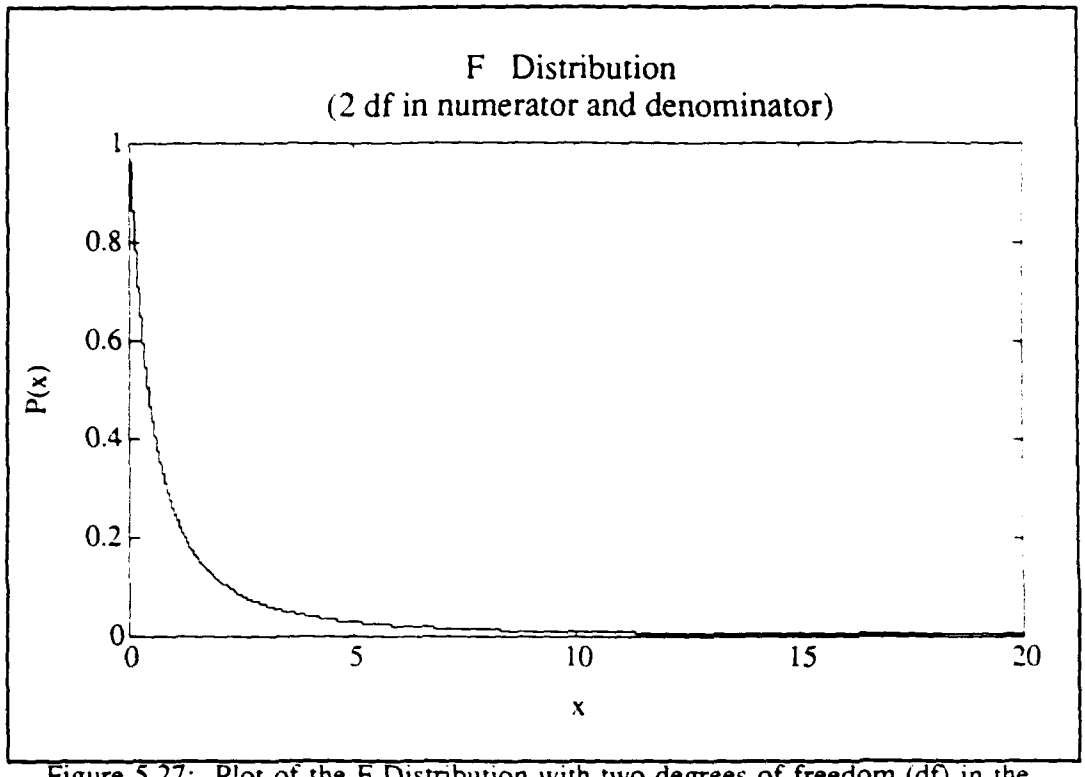


Figure 5.27: Plot of the F Distribution with two degrees of freedom (df) in the numerator and two degrees of freedom in the denominator, ($F_{2,2}$).

Under the null hypothesis that there is no change in the transfer function due to rotation, C_r is distributed as $F_{2,2}$. Similarly under the null hypothesis that there is no change in the transfer function due to motion sickness, C_s is distributed as $F_{2,2}$. The 95% significance level for the $F_{2,2}$ distribution is 19.0 the 90% significance level is 9.0 and the 75% significance level is 3.0. A plot of the $F_{2,2}$ distribution is given in Figure 5.27.

A wide band of frequencies with C_r or C_s values above 19.0 would indicate a significant difference (with 95 % confidence) in transfer functions due to rotation or motion sickness, respectively. In the plots versus frequency, a number of subjects exhibit large values (outliers) of C_s and C_r at a few discrete frequencies or over a few small frequency bands.

However, these outliers occur in limited number and there are no peaks in the C_r or C_s plots that reappear at the same frequency in different subjects*. Therefore, in general, the outliers are consistent with the null hypothesis. Further, the plots of C_r and C_s as histograms appear similar to the $F_{2,2}$ distribution (Figure 5.27) as they are expected to under the null hypothesis. Thus, the analyses of individual transfer functions indicate no significant changes due to rotation or motion sickness in any individual subject.

* If peaks in C_r or C_s were associated with the same frequencies in different subjects, they could not be interpreted as resulting from randomly occurring outliers and therefore would not be consistent with the null hypothesis. A frequency dependent effect would be indicated.

VI Discussion

6.1 The Development of Motion Sickness

The experiment protocol was successful in eliciting motion sickness symptoms in sixteen of eighteen subjects. Further, twelve of these sixteen subjects were able to control their symptoms around a moderate level during random interval breathing. The success of the paradigm was due, in part, to the moderate levels of symptoms induced in subjects. Although magnitude estimates presumably represent slightly different subjective feelings in different subjects, it has been our experience that when magnitude estimates above 5 are attained, avalanching of symptoms toward the vomiting endpoint becomes more likely. Rarely do symptoms avalanche from levels of 3 or 4 to culminate uncontrollably in vomiting. However, above levels of 5, a single nauseogenic stimulus (such as a single head movement) may lead rapidly to vomiting despite all subsequent efforts of the subject or experimenter. Ethical considerations and the desire to avoid inducing severe nausea,

retching or vomiting, therefore required that subjects attempt to maintain only moderate symptoms near 3 or 4 on their magnitude estimate scales.

The concern therefore arises over the severity of sickness experienced by the twelve 'Normal' subjects during the final RIB segment. The average level of sickness reported during the segment varied to some degree between subjects (Table 2.1). Furthermore, a magnitude estimate of 3 or 4 for one subject is probably not the same as that for other subjects. However, the level of symptoms maintained by each subject presumably did involve significant nausea. The presence of significant motion sickness is further supported by subjective reports and/or objective observations of signs and symptoms such as pallor, sweating and feelings of 'fullness in the throat' (Section 5.1). None of the twelve 'Normal' subjects experienced the retching or vomiting which is associated with severe motion sickness.

Thus, data from the final RIB segment is representative of subjects in moderate but perhaps not severe motion sick conditions. It is reasonable therefore to interpret the ILV to IHR transfer functions from the final RIB segment as representative of ANS cardiac control balance in moderately motion sick subjects. Comparisons between transfer functions from the three segments should allow identification of autonomic effects associated with rotation and with moderate motion sickness.

6.2 Analysis of Transfer Functions

Two techniques were used to compare transfer functions and each was designed to meet specific criteria.

The first technique, pooling transfer functions to generate group averages, was designed to identify changes in the transfer function magnitude and phase which were consistent over the population. It treats magnitude and phase components independently and accounts for differences in the coherence function both in calculation of the averages and in estimation of confidence intervals. It does not, however, take advantage of the paired nature of the data sets. That is, in averaging the transfer functions across different subjects information concerning trends characteristic of individual subjects is lost.

The second technique, calculation of C_r and C_s (Equations 4.14 and 4.15), was developed to identify differences in the complex transfer functions between two conditions for a single subject. It is a more powerful statistical test for two reasons. First, it takes advantage of the paired nature of the data sets and second, it does not assume independence of magnitude and phase. However, this technique treats all transfer function estimates (at all frequencies between 0.0 and 0.5 Hz and for all conditions) with equal confidence. That is, it does not explicitly account for differences in the coherence functions (the measure of confidence in the transfer function estimates). Therefore, the test results should be interpreted carefully with particular regard for not over-interpreting large values found in frequency bands associated with low coherence in one or more of the transfer functions. For example, in ILV to IHR transfer functions the frequency band between 0.0 and 0.05 Hz is typically associated with low coherence, and therefore values of C_r or C_s found in this frequency band may be unreliable. Although this second test is a more powerful statistical test, physiological interpretation of significant results may be more complicated. The pertinent physiological models are based on analyses of magnitude and phase differences in the transfer functions (Section 2.3.3). Therefore, if a significant difference is identified between the complex transfer functions, using C_r or C_s, the transfer functions must then be decomposed into their magnitude and phase components in order to relate the changes to their underlying physiological cause.

6.3 Physiological Interpretation

The experiment results indicate that neither rotation nor motion sickness was associated with changes in the group mean or in any individual's ILV to IHR transfer function. Physiological interpretation of these findings relies on the model of Saul et al. (1989) discussed in Section 2.3.3 and presented in Figure 2.13. The model is reproduced in Figure 6.1.

As indicated in Section 2.3.3, four model parameters, A_p, A_S, mean vagal rate and mean sympathetic rate may be varied to independently affect the transfer function. In order to generate their simulations of supine and standing transfer functions, Saul et al. chose the model parameters to be consistent with the current knowledge of autonomic responses associated with posture change. That is, mean vagal rate and depth of modulation, A_p, were set higher for the supine condition than for the standing condition. Mean sympathetic

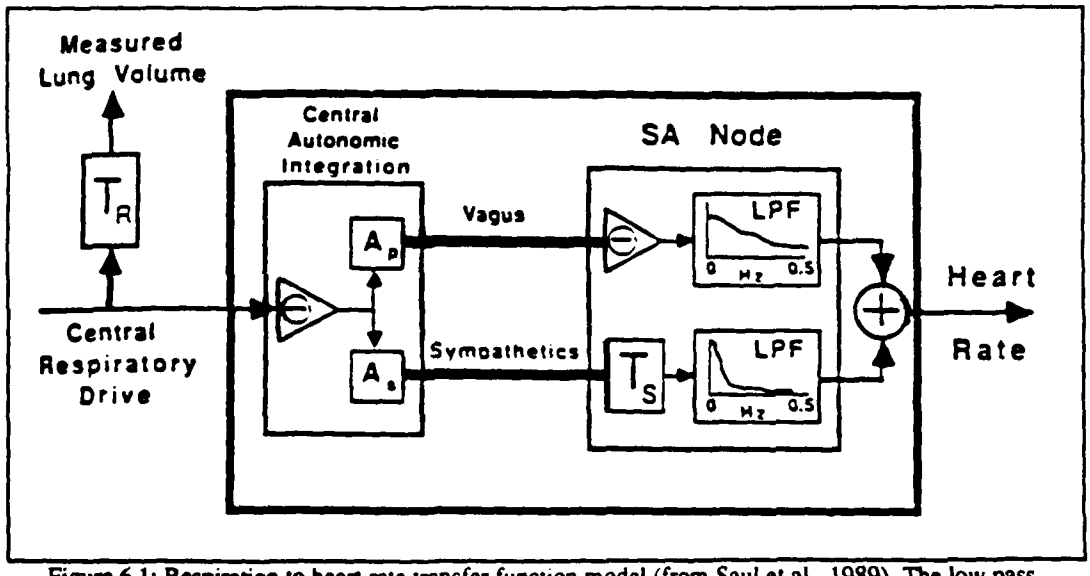


Figure 6.1: Respiration to heart rate transfer function model (from Saul et al., 1989) The low pass filter (LPF) characteristics of the sympathetic and parasympathetic pathways in the SA node are dependent on the mean neural firing rates of each division.

firing rate and depth of modulation, A_S, were set lower for the supine condition than for the standing condition. Within these restrictions, the parameters were then chosen to provide transfer functions which matched the experimental data.

In the exploratory analysis of autonomic responses to motion sickness, no a priori restrictions are placed on the model parameters. However, no changes were identified in the experimental transfer functions. Therefore, to simulate the three experimental conditions, either (1) no changes should be made in the model parameters between conditions or (2) changes in the model parameters between conditions must not alter the overall transfer relations. The question, then, is whether model parameters can be altered and yet effect no change in the overall transfer characteristics.

In Figure 6.2, the transfer characteristics of the parasympathetic and sympathetic pathways through the sinoatrial node are plotted for a number of mean firing rates (from Berger et al., 1989b). At all mean firing rates, the sympathetic pathway has non-zero transfer magnitude

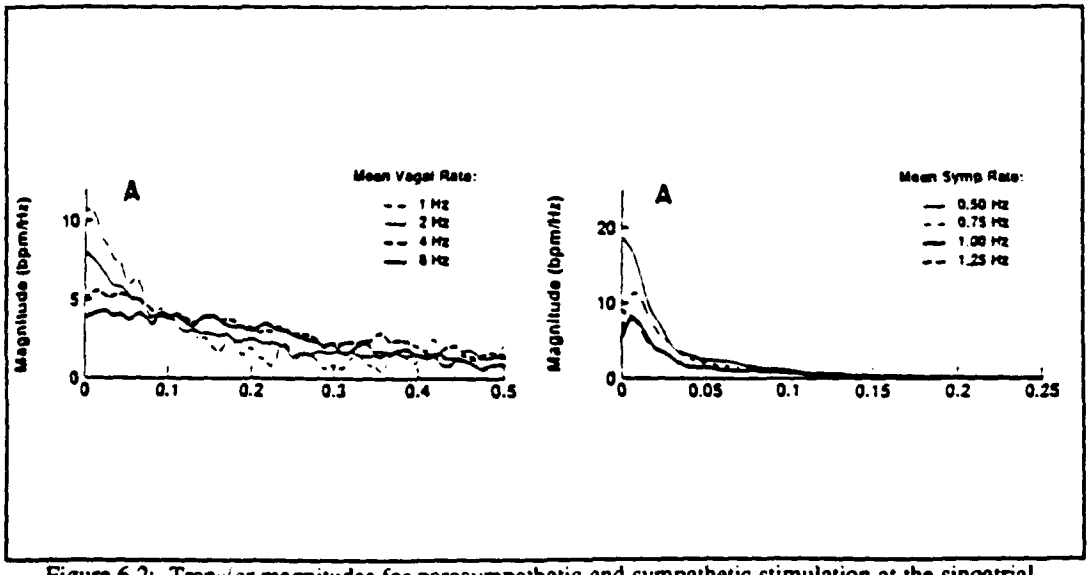


Figure 6.2: Transfer magnitudes for parasympathetic and sympathetic stimulation at the sinoatrial node. (Note differences in scale) (from Berger et al., 1989b)

only below 0.1 Hz. while the parasympathetic pathway has non-zero transfer magnitude throughout the range below 0.5 Hz. Therefore, it is clear that the depths of modulation A_p or A_s could not be modified in such a way as to exhibit no change in the overall ILV to IHR transfer function. If, for example, the depth of modulation of parasympathetic stimulation, A_p, were increased, no change in the other model parameters could compensate for increased gains which would be seen at frequencies above 0.1 Hz. However, it is also clear that changes in the shape of the transfer function due to changes in mean firing rates may be subtle. If such changes occurred they could effect little change in the overall ILV to IHR transfer characteristics. However, in most situations, increases in mean firing rates are expected to be associated with increases in depths of modulation (or variability) of the firing rates. Therefore the gain parameters A_p and A_s are taken as indicative of the relative levels of parasympathetic and sympathetic control of heart rate fluctuations (Berger, 1987; Chen et al., 1986; Saul et al., 1989; Appel., 1989a). Under this assumption, the parameters of the model could not be altered in such a way as to leave the overall ILV to IHR transfer function unchanged.

Therefore, no change in the ILV to IHR transfer functions between the three experimental conditions may be taken to indicate no change in the model parameters. Thus, it may be concluded, neither rotation nor motion sickness is associated with changes in the relative roles of the parasympathetic and sympathetic divisions in control of heart rate.

It is not particularly surprising that autonomic modulation of heart rate in seated subjects is unaffected by sinusoidal rotation about an earth vertical axis. Rotation rates were most likely not rapid enough to generate accelerations which would significantly affect the distribution of blood volume or blood flow in the seated subjects. The cardiovascular strain induced by the motion was presumably insignificant and autonomic counteractions were not warranted.

It may surprise many, however, that no significant shift in autonomic modulation of heart rate was detected in association with motion sickness. This result is in direct opposition to the findings of Ishii et al (1987). As described in Section 2.2.2, Ishii et al. reported trends in the coefficient of variation (CV) of RR interval as motion sickness was induced in squirrel monkeys. Three major distinctions must be drawn between the paradigm used in their experiment and the one used in this experiment. First, Ishii et al. explored the entire range of motion sickness symptoms up to and including the point of vomiting and they report that the most consistent and significant changes in CV occurred just prior to vomiting. Subjects 1 through 12 did not experience these severe symptoms and therefore this experiment provides no evidence concerning the role of the ANS in the severe phases of sickness. Activation patterns associated with severe sickness may differ. Second, Ishii et al. did not control or record respiration. If the monkeys tended to begin panting as vomiting became eminent, the change in CV of RR intervals could be due to the changing respiratory effect on heart rate. In other words it may be that what Ishii et al. identified were not changes in the parameters of the control system but rather changes in the input to the system. That is, the changes in CV may be due solely to changing respiratory patterns and not to changing levels of autonomic activation. The transfer function estimation technique applied in our study specifically eliminates this problem. Thirdly, the possibility of species dependent differences in cardiovascular responses must be considered.

The results are also not in accord with the models discussed in section 2.2.2. The three broad models discussed were (1) generalized sympathetic activation during sickness, (2) parasympathetic activation during sickness, and (3) sympathetic over stimulation leading to parasympathetic rebound as vomiting became eminent. If a generalized sympathetic activation accompanied the development of moderate motion sickness, the transfer function from the motion sick condition would be expected to exhibit greater magnitude at low

frequencies (< 0.1 Hz) and perhaps decreased magnitudes at higher frequencies. No such trends were evident either in individuals or in the population mean. If a generalized parasympathetic activation accompanied the development of motion sickness, a relative increase in the magnitude throughout the associated transfer function would be expected. Again no such increases were identified. It is more difficult to draw comparisons to the third model. In these experiments, subjects were generally in what would be considered the early to middle stages of sickness. Therefore, the model would seem to predict that subjects were experiencing effects of sympathetic over stimulation. The transfer function estimates do not support this assertion. All three of these models involves a generalized activation of the ANS. To the extent that no changes were identified in ANS control of heart rate, any model involving generalized ANS activity is not supported by the transfer function data.

The models described in Section 2.2.2 all assume that the ANS exerts diffuse, body wide effects as in a stress type response. This is a classic model for autonomic activity which is based on the idea that the sympathetic nervous system exerts body wide control. That is, the system is presumed to be described by a single parameter termed 'sympathetic tone' which is a measure of the global activity level of the system. A functional model representing this pattern of activation is given in Figure 6.3 (modified from Wallin, 1986). Autonomic outflow is generally viewed as arising from a central control (in large part, the hypothalamus) to produce similar shifts in autonomic outflow throughout the body. Supraspinal and spinal reflex responses are also assumed to be expressed in all effector organ systems. In Figure 6.3, these characteristics are functionally represented by a single control pathway which is affected by reflexive feedback and influences all organ systems with identical activation patterns.

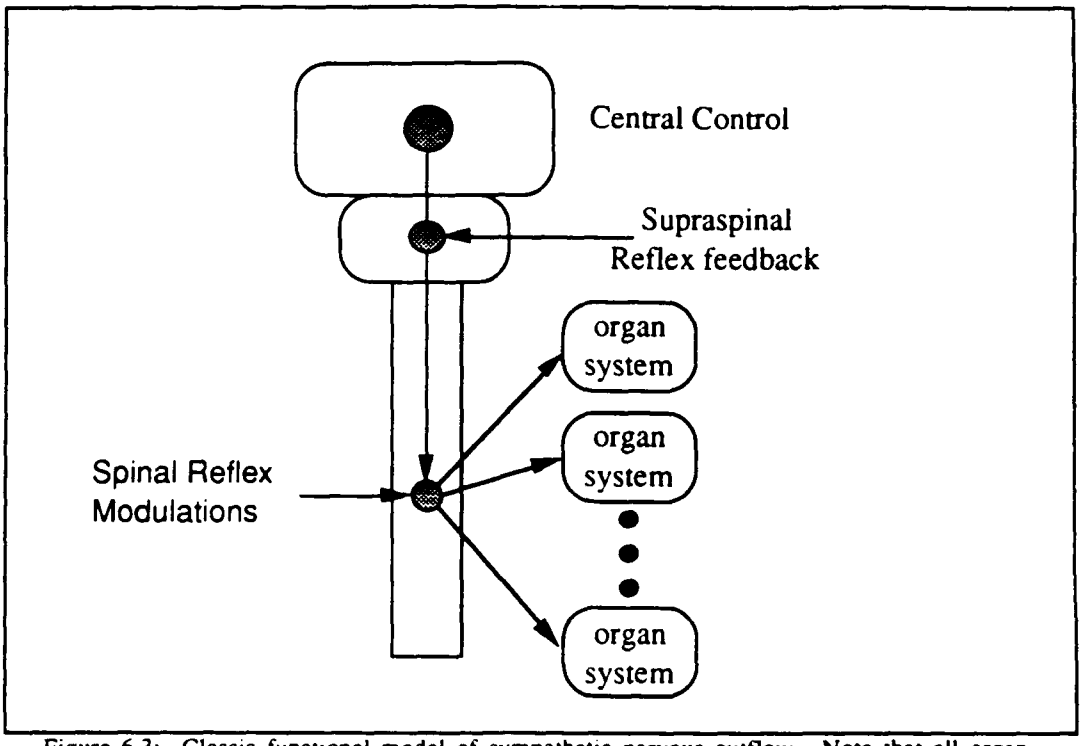


Figure 6.3: Classic functional model of sympathetic nervous outflow. Note that all organ systems receive the same central control outflow which is modulated by reflex feedback. (modified from Wallin, 1986)

If the control model in Figure 6.3 is accurate, then a lack of change in sympathetic outflow to one organ system would necessitate that all organ systems exhibit no change. This does not seem to be the case during motion sickness. No change was detected in the autonomic modulation of heart rate during these experiments. However, the presence of motion sickness symptoms such as pallor and sweating suggests that sympathetic outflow to the blood vessels and sweat glands in the skin may have increased.

As indicated in Section 2.3.3, microneurographic studies demonstrate that the sympathetic nervous system, in particular, can exert very narrow, isolated control (Wallin et al., 1986). For example, sympathetic nerves in the skin and in muscle demonstrate dissociated activity. The more recent knowledge of autonomic outflow characteristics and the current experimental results suggest a different model for the role of the ANS in the development

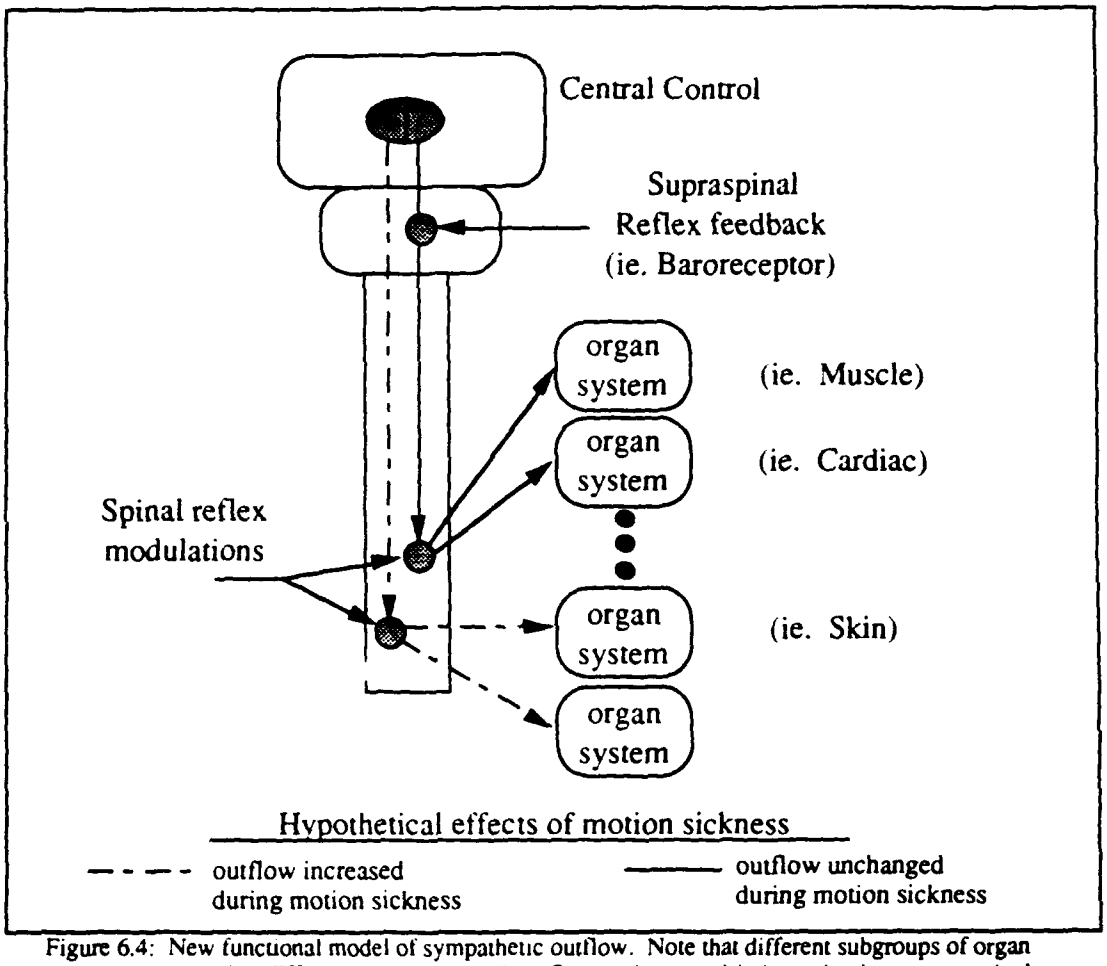


Figure 6.4: New functional model of sympathetic outflow. Note that different subgroups of organ systems may receive different central control outflow and respond independently to supraspinal reflexes. Spinal reflex modulations, however, may be closely associated in different subgroups. It is suggested that during motion sickness skin sympathetic activity may increase while muscle and cardiac sympathetic activity remain unchanged. (modified from Wallin, 1986)

of motion sickness. A new functional model of sympathetic outflow proposed by Wallin (1986) is illustrated in Figure 6.4 (modified from Wallin, 1986). In this model, different subgroups of organ systems receive independent control from central mechanisms. Subgroups of organ systems are comprised of organs involved with similar control functions. Organs within a subgroup receive similar sympathetic outflow. Different subgroups may receive different central control outflow and may be effected independently of one another by supraspinal reflex control loops such as the baroreceptor reflex. Finally, spinal reflex loops tend to effect a more global control which is apparent only in the

absence of more central control activity. These characteristics are functionally represented in Figure 6.4 by a number of control pathways, some of which involve specific supraspinal reflex feedback.

The current experiment results may be explained on the basis of the model in Figure 6.4. During motion sickness, increased sympathetic outflow to the skin may occur through one pathway independently of sympathetic activity in muscle and cardiac nerves. The increase in skin sympathetic activity would elicit the sweating and pallor seen in many subjects during sickness. Since skin sympathetic activity is known to be sensitive to arousal stresses, it may be that the skin response is evoked by an emotional arousal stress associated with motion sickness. Cardiac sympathetic activity, on the other hand, may not be affected during motion sickness. In the model of Figure 6.4, this is explained if cardiac sympathetic outflow is dissociated from skin sympathetic outflow. Since the two organ systems are involved in different primary control tasks and since cardiac outflow is influenced by baroreceptor feedback while skin sympathetic activity is not (Wallin, 1986), it may be reasonable to assume such a dissociation.

The model presented in Figure 6.4 is speculative but it is consistent with the current experiment results. The results suggest that motion sickness is not accompanied by changes in autonomic modulation of heart rate but that it is accompanied by other seemingly autonomic manifestations. Thus, it seems that the ANS does not exert global changes during the development of motion sickness but rather it effects more localized, organ specific manifestations.

VII Summary and Conclusions

A series of experiments were conducted on human volunteers to investigate the role of the autonomic nervous system in the development of motion sickness. A new technique exploiting the relationship between respiration and heart rate was applied to assess autonomic activity.

It is widely accepted that respiration influences heart rate and that its influence is mediated through autonomic mechanisms. In a series of studies, involving pharmacological blockades of the autonomic subsystems, Dr. Cohen and colleagues, at MIT, demonstrated that the transfer function from instantaneous lung volume (ILV) to instantaneous heart rate (IHR) provides information concerning relative levels of parasympathetic and sympathetic activity. Further, they developed an experimental technique which allows accurate estimation of the transfer function over the range of 0.0 to 0.5 Hz. The technique termed, Random Interval Breathing (RIB) requires that subjects breathe in sequence with a fifteen minute series of randomly occurring auditory cues.

Motion sickness was induced in the laboratory using a pair of reversing prism goggles and a rotating chair. Each of eighteen subjects (ages 18-30 yrs, 11 male, 7 female) participated in one four hour experimental session. Control recordings of instantaneous lung volume (ILV) and electrocardiogram (ECG) were made during two random interval breathing segments. During the first segment, subjects were seated motionless and during the second they were seated rotating about an earth vertical axis. Each subject was then fitted with a pair of prism goggles which reverse the left-right visual field and performed a series of coordinated tasks until pre-specified moderate levels of motion sickness were attained. When moderate symptoms were reached, subjects were asked to close their eyes as they were again rotated about an earth vertical axis. They were instructed to open their eyes if symptom levels dropped and re-close their eyes when the desired level was regained. Through repetition of this process, a relatively constant, moderate level of sickness was maintained. Lung volume and ECG were recorded during this motion sick condition as the subject completed a third random interval breathing sequence.

ILV to IHR transfer functions were calculated from segments of data collected during each of the three RIB segments. Comparisons of individual and group mean transfer functions from the two non-sick conditions with each other and with known standards, indicate no detectable shift in autonomic cardiac control due to rotation. Similar comparisons between the two rotating conditions indicate no consistent and significant shift in autonomic tone due to motion sickness. It was therefore concluded that moderate motion sickness is not accompanied by changes in the autonomic outflow controlling heart rate.

The lack of an identifiable shift in autonomic cardiac control is not in accord with models of motion sickness involving generalized autonomic activations. In particular, the results are at variance with the widely held notion that motion sickness can be viewed as a generalized

stress response. However, the presence of possibly autonomic manifestations such as pallor and sweating in bodily organ systems other than the heart indicates that the ANS may act independently at different organ systems during motion sickness. A new functional model of autonomic outflow during motion sickness was presented. Based on the work reviewed by Wallin et al. (1987), it was postulated that a number of sympathetic pathways act independently during the expression of moderate motion sickness. It was suggested that many of the outwardly visible symptoms of motion sickness, particularly pallor and cold sweating, may be due to increased sympathetic outflow to skin effector systems. Conversely, sympathetic outflow in pathways associated with muscle and cardiac organ systems was postulated not to change significantly during motion sickness.

The development of moderate motion sickness does not involve a significant change in autonomic control of heart rate. Therefore, motion sickness does not involve a widespread, generalized activation of the autonomic nervous system. Rather, it is postulated that the ANS plays a more discrete organ specific role in the development of sickness in which, for example, skin effector systems exhibit significant changes in ANS activity but cardiac systems do not.

VIII Recommendations

In this study, the ILV to IHR transfer function was applied to assess autonomic activity. Numerous studies have demonstrated the sensitivity of the transfer function to changes in autonomic activity (Section 2.3.3). However, the transfer function remains an investigational measure and has not yet been widely used. Therefore, a null result, such as that described in this thesis, is rendered somewhat questionable. One way to support the results of this study is to demonstrate a change in the transfer function, due to a well understood stress (ie. change from supine to standing posture), in the same subjects in which no change is found due to motion sickness. It is therefore recommended that future studies applying transfer function techniques to explore ANS activity during motion sickness or other "stresses" should include a supine vs standing comparison.

A second recommendation also concerns future application of the transfer function estimation technique to assessing autonomic responses to stresses. The possibility of shortening random interval breathing segments should be investigated. If accurate and

meaningful transfer function estimates can be attained from shorter duration random breathing segments the technique may have broader application. Particularly in situations which are uncomfortable or dangerous to subjects (ie. motion sickness, increased gravitational stresses), the fifteen minute duration segments may be excessive.

The results of this work suggest a new direction for experiments in motion sickness physiology. In order to assess autonomic activity during motion sickness, direct recordings from muscle and skin sympathetic nerves could be made. A dissociation between muscle sympathetic and skin sympathetic activity during motion sickness would support the hypothesis that motion sickness involves dissociated, organ specific autonomic contributions. Furthermore, direct neural recordings would provide a running time course of sympathetic activity rather that the effectively discrete sampling provided by transfer function estimation.

Appendix A

SCREENING INTERVIEW QUESTIONS

TRANSFER FUNCTION ANALYSIS OF AUTONOMIC REGULATION DURING MOTION SICKNESS

SUBJECT:_____ TODAY'S DATE: _____

The following questions are to be asked of potential subjects in a telephone or in person interview. The answers to these questions will be used only for screening of subjects. Subjects are to be excluded from the study if the answers to these questions reveal a possible biasing of experimental results or an unusually high risk to the subject resulting from his or her participation. The interviewer should record the subjects responses by circling the appropriate response and noting any comments and explanations.

INSTRUCT SUBJECT: I will need to ask a number of questions regarding your medical and motion sickness history. Your answers will be treated in the strictest confidence.

1. Subjects Age: _____ GENDER: _____ LEFT OR RIGHT HANDED _____

2. Have you ever been diagnosed with a heart or lung disorder? NO YES
   If yes, EXCLUDE.

3. Do you experience frequent heart palpitations or abnormal beats? NO YES
   If yes, how often? _____
   If greater than daily, EXCLUDE.

4. Have you ever been diagnosed with Diabetes, Epilepsy or Aids? NO YES
   If yes, EXCLUDE.

5. Are you a smoker or ex-smoker? NO YES If yes, what and how frequently?
   If yes, Marginal Exclusion.

6. Have you ever been diagnosed with a gastrointestinal disorder such as an ulcer, hiatus hernia, carcinoma or recently diagnosed with gastritis? NO YES If yes, EXCLUDE.

7. Do you frequently have abdominal pain or discomfort which is relieved by antacids or food? NO YES If yes, EXCLUDE.

8. Have ever had an unexplained episode of nausea and/or vomiting? NO YES
   If yes, when? _____ If recently, EXCLUDE.

9. Have you recently suffered a loss of appetite or unusual weight loss? NO YES
   If yes, marginal exclusion.

10. Do you have a hearing defect? NO YES
    If yes suspect vestibular defect.

11. Have you ever experienced a persistent noises in your ears continuing for more than a few moments? NO YES If yes, suspect vestibular defect.

12. Have you ever experienced repeated episodes of disorientation or vertigo while not in a moving vehicle? NO YES If yes, suspect vestibular defect.

If suspect vestibular defect (from 10-12), pursue with:
Do you have trouble walking outside at night? NO YES
Do you ever have spells of dizziness? NO YES
Have you ever had surgery for otosclerosis? NO YES
Have you ever been diagnosed with a vestibular defect?
If yes, or if high suspicion of vestibular defect, EXCLUDE.

13. Have you ever experienced motion sickness before?

14. If you have experienced motion sickness, how long does it usually take for you to recover completely when the motion stimulus is removed? _____
If long, suspect high susceptibility and pursue by asking for examples of when subject experienced sickness. EXCLUDE those who describe very high susceptibility (ie. such that symptoms may be hard to control.)

15. Is your vision better than 20/50 after correction with contact lenses (if necessary) ? (Glasses may not be worn during the experiment.) YES NO If No, EXCLUDE.

16. Are you presently taking any medications? If yes, what type? _____
Exclude for drugs with central nervous effects(ie. antihistamines or anti-seizure medications).

17. Have you recently been under the care of a psychiatrist? _____ If yes, be concerned about paranoids or schizophrenics who may be problem subjects.

WHEN QUESTIONING IS COMPLETED:

If subject does not warrant exclusion, schedule a session.

Remind him/her of the following instructions and inform him that an instruction packet will be mailed. If a face to face interview, the subject can immediately be provided a packet

A. Please read the Motion Sickness Symptom Definitions.

B. Please read the Magnitude Estimation Instructions.

C. Please fill out the Motion Sickness Questionnaire.

D. Prior to the experimental session please try to do the following:

1. On the day of the experiment, please eat your normal meal between _____ and _____, and eat nothing thereafter prior to the experiment.

2. Please take no medications for 24 hours prior to the experiment.

3. Please consume no alcohol for 24 hours prior to the experiment.

4. Please drink no coffee, tea or cola, and do not smoke for 12 hours prior to the experiment.

5. Try to avoid heavy exercise for 6 hours prior to the experiment.

6. Try to get a normal night's sleep the night before the experiment.

TJM 4/18/89

SUBJECT NO. _____ 1

MOTION SICKNESS QUESTIONNAIRE

This questionnaire is designed to help us assess your susceptibility to motion sickness and the types of motion which have been most effective in causing your motion sickness. The form is divided into three parts. Section A is concerned with your childhood experiences of motion sickness (prior to the age of 12), Section B deals with your experience since the age of 12 and Section C asks you to estimate your present overall susceptibility to motion sickness.

Please try to accurately complete all sections. Your replies to all questions will be treated in the strictest confidence.

SUBJECTS NAME: _____ DATE: _____

TJM 4/18/89 adapted from (Eagon, 1987 and Reason and Brand, 1975)

SECTION A

All questions refer **ONLY** to your childhood experiences with motion sickness (if any) and travel, where childhood is defined as the period prior to 12 years of age. It is quite possible that you will have difficulty recalling childhood motion sickness. Nevertheless, please try to answer the questions to the best of your ability.

Put your answers to column 1 in column 1 of the table below; your answers to question 2 in column 2 and so on.

1. Indicate approximately how often you travelled as a passenger on each of the following vehicles (before age 12) by using the following numbers:

0 No experience
1 Less than 5 trips
2 Between 5 and 10 trips
3 More than 10 trips

Considering ONLY those types of transportation that you marked 1, 2 or 3 (i.e. those that you have travelled on as a passenger), go on to questions 2 and 3 below. Use the following letters to indicate the appropriate category of responses:

N Never
R Rarely
S Sometimes
F Frequently
A Always

2. How often did you <u>feel sick</u> (e.g. queasy or nauseated) while travelling?

3. How often did you <u>actually vomit</u> while travelling?

	1	2	3
CARS			
BUSES OR COACHES			
TRAINS			
AIRLINERS			
AEROBATIC AIRCRAFT			
LIGHT AIRCRAFT			
SMALL OPEN BOATS			
BOATS WITH CABINS			
SHIPS			
GYM SWINGS			
MERRY GO ROUND			
ROLLER COASTER			
OTHER SITUATIONS WHICH MADE YOU MOTION SICK? PLEASE SPECIFY TYPE. (USE REVERSE IF NECESSARY)			

SECTION B

This section is concerned with your experiences of motion sickness and travel SINCE the age of 12. Please try to answer the questions to the best of your ability. Put your answers to column 1 in column 1 of the table below; your answers to question 2 in column 2 and so on.

1. Indicate approximately how often you travelled as a passenger on each of the following vehicles (before age 12) by using the following numbers:

0 No experience
1 Less than 5 trips
2 Between 5 and 10 trips
3 More than 10 trips

Considering ONLY those types of transportation that you marked 1, 2 or 3 (i.e. those that you have travelled on as a passenger), go on to questions 2 and 3 below. Use the following letters to indicate the appropriate category of responses:

N Never
R Rarely
S Sometimes
F Frequently
A Always

2. How often did you feel sick (e.g. queasy or nauseated) while travelling?

3. How often did you actually vomit while travelling?

	1	2	3
CARS			
BUSES OR COACHES			
TRAINS			
AIRLINERS			
AEROBATIC AIRCRAFT			
LIGHT AIRCRAFT			
SMALL OPEN BOATS			
BOATS WITH CABINS			
SHIPS			
GYM SWINGS			
MERRY GO ROUND			
ROLLER COASTER			
OTHER SITUATIONS WHICH MADE YOU MOTION SICK? PLEASE SPECIFY TYPE. (USE REVERSE IF NECESSARY)			

SECTION C

In general, how would you grade your present susceptibility to motion sickness compared to others? (Circle One.)

TOTALLY IMMUNE LESS SUSCEPTIBLE THAN MOST AVERAGE

MORE SUSCEPTIBLE THAN MOST EXTREMELY SUSCEPTIBLE

MAGNITUDE ESTIMATION INSTRUCTIONS

We are interested in monitoring a number of physiological parameters during motion sickness. We will attach electrodes to monitor electrocardiogram and abdominal biopotentials and a belt to measure instantaneous lung volume as motion sickness symptoms develop. However, the most important symptoms of motion sickness are uniquely subjective and cannot be measured with an instrument. We therefore must rely on estimates which you make of the sensation intensity. We ask you to apply a technique which is called "magnitude estimation". We will ask you to judge the intensity of the sensation by comparing it to a "standard" intensity which you have previously experienced. You must then estimate the ratio between the current sensation and your memory of the "standard." Subjects usually find magnitude estimation to be an easy, natural method of reporting the intensity of sensation. Some subjects are at first skeptical of whether meaningful reports can be obtained with such a simple method until they try it and see how consistent their reports can be.

To give you the basic idea of magnitude estimation, try the following experiment which involves the length of lines:

Suppose we say the "standard" line is one inch long, and we call this line "10". You must now recall the image of a ruler or some such object which everyone in our culture has experienced.

Now suppose we present you with a line of unknown length:

_____ If the standard is "10", how long is this line?
How accurate do you think your estimate is?
Now how long is this line?

Finally, how long is this one?

-

On this last one, if you find the line length ratio so small that it is difficult to judge, it is better to say that the sensation (ie. the line) is present but is too small to judge.

We will ask you to use this same technique to report the intensity of your sensation in our motion sickness tests. The only real difference is in the type of sensation being judged -- nausea. You will have to rely on your memory of previous times that you have been nauseated in order to define your standard sensation level. We expect that it may take a little time before you feel your memory of the standard has stabilized and you believe your reports are consistent.

Once you begin to perform tasks while wearing the reversing prism goggles, after some time (depending on your susceptibility), you will begin to experience symptoms, which may include stomach awareness or discomfort, nausea, sweating, salivation, headache or dizziness. Most people are familiar with nausea which can be defined as an unpleasant sensation, usually referred to the stomach, chest or throat which at very high levels may eventually be associated with vomiting.

In this experiment, we want you to use the magnitude estimation technique to tell us about the intensity of your nausea. We want to work with only slight to moderate sensation levels, in order to minimize any chance that you will reach the point of vomiting, so do your best to tell the experimenter exactly how you feel at all times. early in the experiment, we will show you that if you stop moving your head and close your eyes, after a few moments, symptoms will rapidly subside. Once you have gained experience with this, you will gain confidence that symptoms can be limited to acceptable levels throughout the experiment with little difficulty. If at any time during the experiment, despite all precautions, you feel your symptoms are getting intolerably high, stop head movements and close your eyes immediately. Do not wait for the experimenter to so instruct you.

At the outset of the experiment, the experimenter will ask you to choose a sensation magnitude of nausea in the middle range of your experience or "halfway to vomiting." You should call this standard intensity a "5" and try to remember how it feels. Your task will be to estimate the magnitude of your subsequent sensation of nausea with respect to this standard. In other words, if you feel you sensation is half the standard you should report 2.5, if it is double the standard, report a 10, and so forth. If you are not experiencing the sensation say "absent."

We will review and practice the technique prior to the experimental session.

TJM 4/18/89 (adapted from Eagon, 1987, and Bock and Oman, 1982)

SUBJECT INSTRUCTIONS AND SESSION SCHEDULE

SUBJECT:_____ TODAY'S DATE: _____

Your Session has been scheduled for: SESSION DATE: _____ / / ____

SESSION TIME: _____ AT MIT, RM 37-146

The session should last between 3 and 4 hours. If a conflict arises which makes it impossible for you to attend this session, please call to reschedule. -Tom Mullen in 37-219 @ 253-7805
in 37-155 @ 253-7509

IN PREPARATION FOR YOUR SESSION:

Please skim over the motion sickness symptom definitions which are attached. Most people are familiar with the sensations experienced in motion sickness but terminology is rarely the same from person to person.

Also, please read the Magnitude Estimation Instructions which are attached. We will employ the technique as our primary measure of the time course of your symptoms. If you have questions, we can discuss the technique when you arrive for the experimental session.

Please complete the enclosed motion sickness questionnaire and bring it to the experimental session. Again, if you have questions, you can complete the questionnaire when you arrive for the session.

Motion sickness susceptibility may be affected by a variety of extraneous factors for which we want to control as much as possible. Because this control is important we ask that on the following:

1. On the day of the experiment, please eat your normal meal between _____ and _____, and eat nothing thereafter prior to the experiment.
2. Please take no medications for 24 hours prior to the experiment.
3. Please consume no alcohol for 24 hours prior to the experiment.
4. Please drink no coffee, tea or cola, and do not smoke for 12 hours prior to the experiment.
5. Try to avoid heavy exercise for 6 hours prior to the experiment.
6. Try to get a normal night's sleep the night before the experiment.
7. Please bring or wear a loose T shirt to the experimental session.

Thank you for your participation.

enclosures:
Motion Sickness Symptom Definitions
Magnitude Estimation Instructions
Motion Sickness Questionnaire

TJM 4/18/89

A BRIEF OUTLINE OF SYMPTOM DEFINITIONS

The following outline is meant to provide all subjects with the same definitions for common motion sickness symptoms. We expect that most subjects are familiar with most of the terms but we provide the outline to help insure that experimenter and subject are speaking the "same language" when symptom reports are given. You may find some of these terms handy in reporting your own symptoms. As was emphasized earlier, WE WISH YOU TO EXPERIENCE ONLY MODERATE SYMPTOMS. Early in the experiment, we will demonstrate for you how to control your symptoms and you should avoid reaching the point of severe nausea, retching or vomiting.

EPIGASTRIC AWARENESS - Sensation which draws attention to the epigastric area (stomach, throat etc.) but is not uncomfortable

EPIGASTRIC DISCOMFORT - Sensation in epigastric region which is just becoming uncomfortable. This is an intermediate report between epigastric awareness and nausea.

SLIGHT NAUSEA - Unpleasant sensation which can unequivocally be associated with vomiting but vomiting is not imminent.

MODERATE NAUSEA - Same sensation as above except more intense (an intermediate report)

SEVERE NAUSEA - Vomiting is imminent if stimulation continues "beginning to reach for the bag"

RETCHING OR VOMITING - unproductive "dry heaves" or actual emesis

FLUSHING - an increased reddening of the skin color

SUBJECTIVE FEELING OF WARMTH - a sudden sensation of warmth of the body surface

PALLOR - blanching or paling of the skin color

SWEATING
- MILD - small specks of perspiration skin feels cool
- MODERATE - an intermediate report beads of sweat are apparent
- PROFUSE - rivulets or sheets of sweat are apparent

OTHER SYMPTOMS WHICH ARE SELF EXPLANATORY:

INCREASED SALIVATION
HEADACHE
DIZZINESS
DROWSINESS

INFORMED CONSENT STATEMENT

TRANSFER FUNCTION ANALYSIS OF AUTONOMIC REGULATION DURING MOTION SICKNESS

I have been asked to participate as a subject in a quantitative study of the pattern of autonomic regulation during motion sickness. The stimuli to used are active head movements and passive rotation while wearing left/right vision reversing goggles. I will be asked not to drink alcohol or take medication for 24 hours prior to the experimental session, and not to drink coffee or other stimulants for 12 hours prior to the experiment. I understand that during the testing, I will probably experience mild to moderate motion sickness symptoms such as stomach discomfort, nausea, pallor, sweating or drowsiness and that these symptoms may persist for some time after the session. I will attempt to report my symptoms to the experimenter who will be simultaneously noting my objective symptoms. Conventional disposable surface electrodes may be applied to my chest and abdomen to record electrocardiogram and gastric potentials. The sites of the abdominal electrodes may be lightly scratched with a sterile hypodermic needle prior to the application of the electrode.

Although participation in a complete session is requested, I understand that I am free to withdraw from further participation at any time and for any reason. I realize that there is a slight chance that I may become nauseated to the point of vomiting, although every effort will be made to prevent this by limiting head movements and closing my eyes. I have no medical history such as heart or lung disease or chronic stomach trouble which would make an accidental vomiting episode medically undesirable. I am not diabetic or epileptic.

I understand that I should not operate a motor vehicle for three hours after the end of the experiment, and that I should report any persisting motion sickness symptoms to the experimenter.

I understand that my anonymity will be preserved when my questionnaires and experimental results are reported.

In the unlikely event of physical injury resulting from participation in this research, I understand that medical treatment will be available from the MIT Medical department, including first aid, emergency treatment and follow-up care as needed, and that my insurance carrier may be billed for the cost of such treatment. However, no compensation can be provided for medical care apart from the foregoing. I further understand that making such medical treatment available, or providing it, does not imply that such injury is the Investigator's fault. I also understand that by my participation in this study, I am not waiving any of my legal rights.*

* Further information may be obtained by telephoning the Institute's Insurance and Legal Affairs Office at 253-2882.

I understand that I may also contact the Chairman of the Committee on the Use of Humans as Experimental Subjects, Dr. Walter Jones (MIT E23-389, 253-1772), if I feel I have been treated unfairly as a subject.

I agree to participate in this experiment.

Signed:_____ Date:_____

Experimenter:_____ Date:_____

SESSION NO. _____

PRE-SESSION QUESTIONNAIRE

TRANSFER FUNCTION ANALYSIS OF AUTONOMIC REGULATION DURING MOTION SICKNESS

NAME: _____

DATE: _____

LEGAL ADDRESS: _____

SOCIAL SECURITY NUMBER: _____

Responses evaluated during this study may be directly or indirectly influenced by factors addressed in the following questions. Please answer each to the best of your ability. Your responses will be treated in the strictest confidence.

1. Are you in your usual state of physical fitness today?_____
   If No, Please explain. _____
2. Have you taken any medication (e.g. aspirin, cold preparations, prescriptions or "recreational" drugs) during the past 24 hours? _____
   If yes, what type and how much? _____
3. How much alcohol have you consumed during the past 24 hours?_____
   If any, what type and when? _____
4. How much coffee/tea/cola have you drunk during the past 24 hours?_____
   If any, what type and when? _____
5. How much tobacco have you used during the past 24 hours (# cigarettes, cigars or pipe-fulls)?_____
6. Hour many hours sleep did you get last night?_____
   How many hours would you estimate you get in a usual night?_____
7. How long has it been since your last meal?_____
   What did you eat/drink?_____
   Subjectively, how hungry do you feel now?
   Very Hungry Slightly Hungry Normal Slightly Overfed Very Overfed
8. Have you felt any stomach awareness, stomach discomfort or nausea during the past 24 hours?_____
   If yes, when and why?_____
9. Have you engaged in heavy exercise during the past 6 hours?_____
10. Please estimate: Your weight _____ lbs Your Height _____

TJM 4/17/89

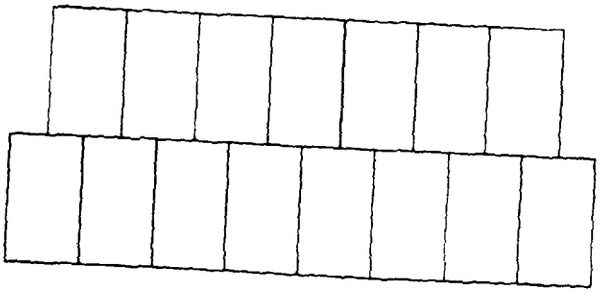
AD-A-243935

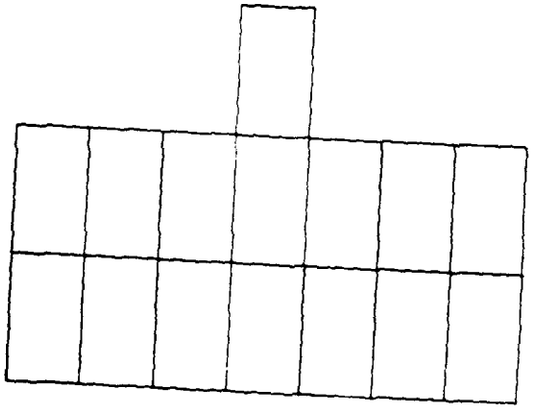
PAGE
#141

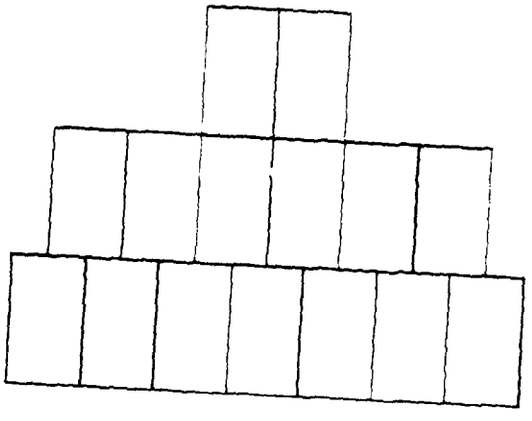
MISSING
FROM ORIGINAL
DOCUMENT
AS SENT TO
DTIC
FROM THE
ORIGINATOR

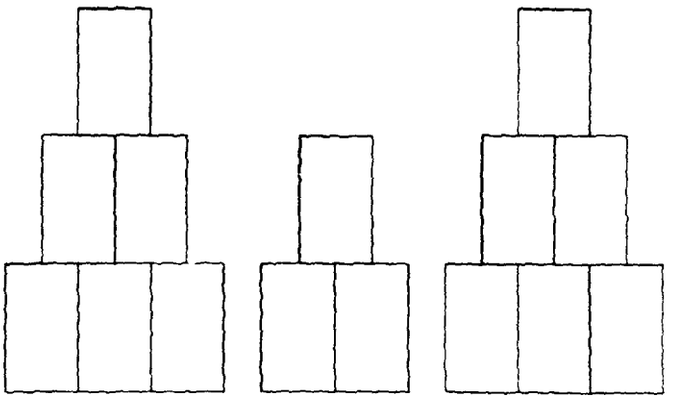
Schematic Diagrams of Can Structures

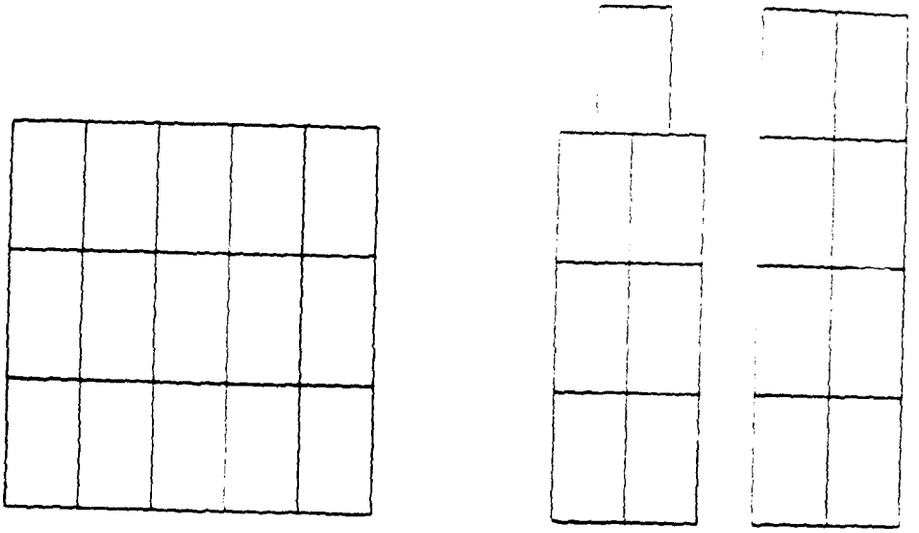
Photo-reduced from 8.5 by 11 inches each
Given in order of presentation to subjects

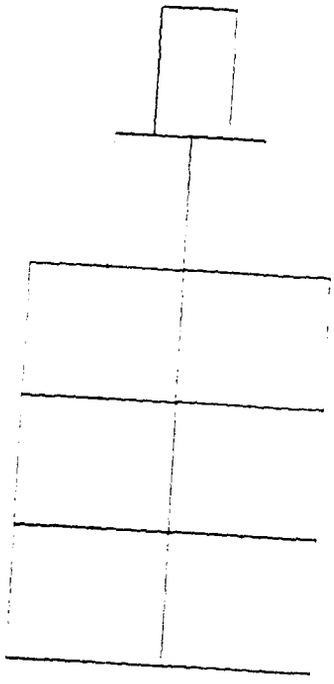


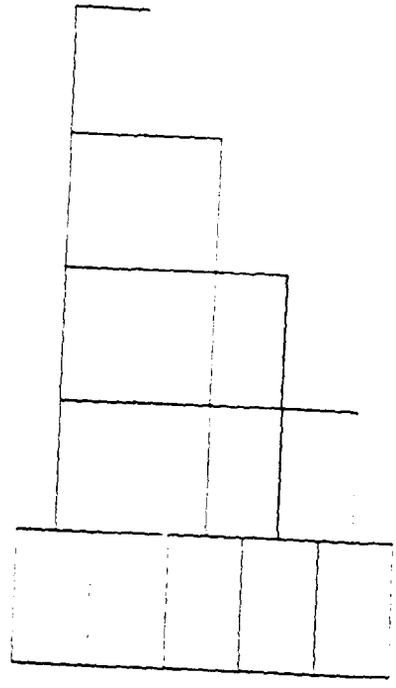


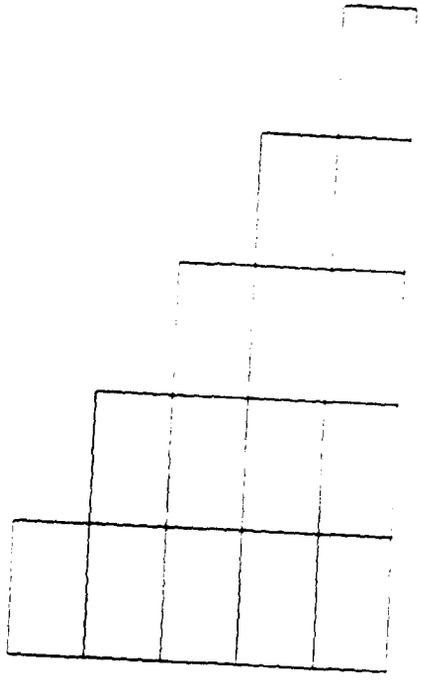


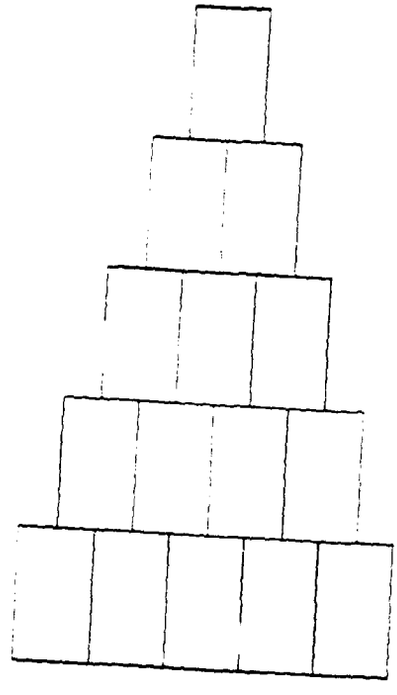












References

Akselrod, S., Gordon, D., Ubel, F.A., Shannon, D.C., Barger, A.C., Cohen, R.J. (1981),*Power Spectrum Analysis of Heart Ra·· Fluctuation: A Quantitative Probe of Beat-to Beat Cardiovascular Control.* **Science,** 213:220-222.

Akselrod, S., Gordon, D., Madwed, J., Snidman, N.C., Shannon, D.C., Cohen, R.J., (1985), *Hemodynamic regulation: In estigation by spectral analysis,* **Am J Physiol** 249(Heart Circ Physiol 18):H867-H875.

Appel, M.L., Berger, R.D., Saul, J.P., Smith, J.M., Cohen, R.J., (1989a), *Beat to Beat Variability in Cardiovascular Variables: Noise or Music?* **J Am Coll Cardiol,** 14(5):1139-1148.

Appel, M.L., Saul, J.P., Berger, R.D., Cohen, R.J., (1989b) *Closed-Loop Identification of Cardiovascular Regulatory Mechanisms,* **Comp in Cardiol,** In Press.

Appenzeller, O., (1976), **The Autonomic Nervous System: An Introduction to Basic and Clinical Concepts,** American Elsevier, New York.

Attias, J., Gordon, C., Ribak, J., Binah, O., Roznick, A., (1989), *Efficacy of Transdermal Scopolamine against Sea Sickness: A 3-Day Study at Sea,* **Aviat Space Environ Med,** 58:60-2.

Bendat, J.S., Piersol, A.G., (1980), **Engineering Applications of Correlation and Spectral Analysis,** John Wiley & Sons, New York.

Bendat, J.S., Piersol, A.G., (1986) **Random Data: Analysis and Measurement Procedures, 2nd ed.,** John Wiley & Sons, New York.

Berger, R.D., Akselrod, S., Gordon, D., Cohen, R.J., (1986), *An efficient Algorithm for Spectral Analysis of Heart Rate Variability,* **IEEE Trans Biomed Eng,** BME-33(9):900-904.

Berger, R.D., (1987), *Analysis of the Cardiovascular Control System using Broad-Band Stimulation,* (Ph.D. Thesis), MIT, Cambridge, MA.

Berger, R.D., Saul, J.P., Albrecht, P., Stein, S.P., Cohen, R.J., (1988), *Respiratory Effects on Arterial Pressure: A Novel Signal Analysis Approach,* Proc. IEEE EMBS 10th Annual International Conference, **CH566-8/88,** 533-544.

Berger, R.D., Saul, J.P., Cohen, R.J., (1989a), *Transfer function analysis of autonomic regulation I. Canine atrial rate response,* **Am J Physiol** 256(Heart Circ Physiol 25): H142-H152.

Berger, R.D., Saul, J.P., Cohen, R.J., (1989b), *Assessment of Autonomic Response by Broad Band Respiration,* **IEEE Trans Biomed Eng,** BME-36(11):1061-1065.

Berne, R.M., Levy, M.N., (1986), **Cardiovascular Physiology, 5th ed.,** C.V. Mosby Co., St. Louis.

Blanford, C., and Oman, C.M., (1990), *Diagnostic Classification of changes in the Human Electrogastrogram during Motion Sickness,* **Aviat Space Environ Med,** 65(5):490.

Blanford, C., (1990) *Diagnostic Classification of Changes in the Human Electrogastrogram during Motion Sickness,* (S.M. Thesis), MIT, Cambridge, MA.

Bock, O.L., Oman, C.M., (1982), *Dynamics of Subjective Discomfort in Motion sickness as Measured with a Magnitude Estimation Method,* **Aviat Space Environ Med,** 53(8):773-777.

Burke, D., Sundlof, G., Wallin, B.G., (1977), *Postural effects of Muscle Nerve Sympathetic Activity in Man,* **J Physiol,** 272:399-414.

Chen, M.H., Berger, R.D., Saul, J.P., Stevenson, K., Cohen, R.J., (1987), *Transfer Function Analysis of the Autonomic Response to Respiratory Activity During Random Interval Breathing,* **Computers in Cardiology,** 14:149-152.

Claremont, C.A., (1931), *The Psychology of Sea-Sickness,* **Psyche,** 11:86-90.

Cohen, M.I., Gootman, P.M., (1970), *Periodicities in efferent discharge of splanchnic nerve of the cat,* **Am J Physiol,** 218(4):1092-1101.

Cowings, P.S., Suter, S., Toscano, W.B., Kamiya, J., Naifeh, K., (1986), *General Autonomic Components of Motion Sickness,* **Psychophysiol,** 23(5): 542-551.

Cowings, P.S., Toscano, W.B., *A New Treatment for Space Motion Sickness: Autogenic-Feedback Training* **NASA Ames Research Center,** Unpublished Manuscript

Cowings, P.S., Naifeh, K.H., Toscano, W.B., (1990), *The Stability of Individual Patterns of Autonomic Responses to Motion Sickness Stimulation,* **Aviat Space Environ Med,** 61(5):399-405.

Crampton, G.H., (1990), **Motion and Space Sickness,** CRC Press, Boca Raton, FL.

Crampton, G.H., (1955), *Studies of Motion Sickness XVII: Physiological Changes Accompanying Motion Sickness in Man,* **J Applied Physiol,** 7:501-507.

Davis, C.J., Harding, R.K., Leslie, R.A., Andrews, P.L.R., (1986), *The Organization of Vomiting as a Protective Reflex: A commentary on the first day's discussions* in **Nausea and Vomiting: Mechanisms and Treatment,** ed. Davis, C.J., Lake-Bakaar, G.V., Grahame-Smith, D.G., Springer-Verlag, New York.

Davis, J.R., Vanderploeg, J.M., Santy, P.A., Jennings, R.T., Stewart, D.F., (1988), *Space motion Sickness during 24 Flights of the Space Shuttle,* **Aviat Space Environ Med** 59:1185-1189.

Dobie, T.G., May, J.G., Fischer, W.D., Elder, S.T., Kubitz, K.A., (1987), *A Comparison of two methods of Training resistance to Visually Induced Motion Sickness,* **Aviat Space Environ Med,** 58(9,Suppl):A34-A41.

Drylie, M.E., (1987), *An Analysis of Physiological Data related to Motion Sickness for use in a Real-time Motion Sickness Indicator,* **USAFIT-GE-ENG-87D-16,** USAF Institute of Technology, Wright Patterson AFB, Ohio.

Eagon, J.C., (1988) *Quantitative Frequency Analysis of the Electrogastrogram during Prolonged Motion Sickness,* (M.D. Thesis), Harvard-MIT Division of Health Sciences and Technology, Cambridge, MA.

Eckberg, D.L., Abboud, F.M., Mark, A.L., (1976), *Modulation of Carotid baroreflex responsiveness in man: effects of posture and propanolol,* **J Applied Physiol,** 41(3):383-387.

Eckberg, D.L., Rea, R.F., Anderson, O.,K., Hedner, T., Pernow, J., Lundberg, J.M., Wallin, B.G., (1988), *Baroreflex Modulation of Sympathetic Activity and Sympathetic Neurotransmitters in Humans,* **Acta Physiol Scand,** 133:221-231.

Eversmann, T., Gottsman, M., Uhlich, E., Ulbrecht, G., von Werder, K., Scriba, P.C., (1978), *Increased Secretion of Growth Hormone, Prolactin, Antidiuretic Hormone and Cortisol Induced by the stress of Motion Sickness,* **Aviat Space Environ Med,** 49(1):53-57.

Gaudreault, P.J., (1987), *Motion Sickness; A Study of its Effects on Human Physiology,* Thesis **USAF-GE-ENG-87D-20,** USAF Institute of Technology, Wright Patterson AFB, Ohio.

Gillingham, K.K., Wolfe, J.W., (1986), *Spatial Orientation in Flight,* **USAFSAM-TR-85-31,** USAF School of Aerospace Medicine, Brooks AFB, TX.

Gordon, C.R., Ben-Aryeh, H., Szargel, R., Attias, J., Rolnick, A., Laufer, D., (1989), *Salivary changes associated with seasickness,* **J Auton Nerv Syst,** 26:37-42.

Gordon, C.R., Ben-Aryeh, H., Szargel, R., Attias, J., Rolnick, A., Laufer, D., (1988), *Salivary changes associated with experimental motion sickness condition in man,* **J Auton Nerv Syst,** 22:91-96.

Graham, D.T., Kabler, J.D., Lunsford, L. Jr., (1961), *Vasovagal Fainting; A Diphasic Response*, **Psychosom Med,** 23:493-507.

Grahame-Smith, D.G., (1986), *The Multiple causes of Vomiting: Is there a common Mechanism?* in **Nausea and Vomiting: Mechanisms and Treatment,** ed. Davis, C.J., Lake-Bakaar, G.V., Grahame-Smith, D.G., Springer-Verlag, New York.

Graybiel, A., Johnson, W.H., (1963), *A Comparison of the Symptomatology Experienced by Healthy Persons and Subjects with Loss of Labyrinthine Function when Exposed to Unusual Patterns of Centripetal Force in a Counter-rotating Room,* **Oto Rhino Laryngol,** 72(2):357.

Graybiel, A., Lackner, J.R., (1980), *Evaluation of the Relationship Between Motion Sickness Symptomatology and Blood Pressure, Heart Rate, and Body Temperature,* **Aviat Space Environ Med,** 51(3):211-214.

Guedry, F., (1968), *Conflicting Sensory Orientation Cues as a Factor in Motion Sickness,* Fourth Symposium on the Role of the Vestibular Organs in Space Exploration, **NASA SP-187,** 45-51.

Guyton, A.C., Coleman, T.G., Cowley, A.W., Manning, R.D., Norman, R.A., Ferguson, J.D., (1974), *A Systems Analysis Approach to Understanding Long-range Arterial blood Pressure Control and Hypertension,* **Circ Res,** 35:159-176.

Guyton, A.C., (1986), **Textbook of Medical Physiology,** W.B. Sanders Co. Philadelphia.

Habermann, J., Eversmann, T., Erhardt, F., Gottsman, M., Ulbrecht, G., Scribe, P.C., (1978), *Increased Urinary Excretion of Triiodothyronine (T3) and Thyroxine (T4) and Decreased Serum Thyreotropic Hormone (TSH) Induced by Motion Sickness,* **Aviat Space Environ Med,** 49(1):58-61.

Hirsch, J.A., Bishop, B., (1981), *Respiratory Sinus Arrhythmia in Humans: How breathing Patterns Modulate Heart Rate,* **Am J Physiol,** 241(Heart Circ Physiol 10):H620-H629.

Hockman, C.H., (1987), **Essentials of Autonomic Function,** Charles C. Thomas, Springfield, IL.

Igarashi, M., Himi, T., Ishii., M., Patel, S., Kulecz, W.B., (1987), *The change in coefficient of variance of R-R interval and the susceptibility of sensory-conflict sickness (subhuman primate study)* **Space Life Sciences Symposium: Three Decades of Life Science Research in Space , June 1987,** pp. 208-210.

Ishii, M., Igarashi, M., Patel, S., Himi, T., Kulecz, W.B., (1987), *Autonomic Effects on R-R Variation of the Heart Rate in the Squirrel Monkey: An Indicator of Autonomic Imbalance in Conflict Sickness* **Am J Otolaryngol,** 3:144-148.

Isu, N., Takahashi, N., Koo, J., (1987a), *Skin Potential Reflex Corresponding to Transient Motion Sickness,* **Aviat Space Environ Med,** 58:576-80.

Isu, N., Koo, J., Takahashi, N., (1987b), *Changes of Skin Potential Level and of Skin Resistance Level Corresponding to Lasting Motion Discomfort,* **Aviat Space Environ Med,** 58:136-42.

Janowsky, D.S., (1985), *A Phystostigmine Model of Space Adaptation Syndrome* in **Proceedings of the Space Adaptation Syndrome Drug Workshop, July 1983,** ed. Kohl, R.L., pp.33-39, Space Biomedical Research Institute, USRA Division of Space Biomedicine, Houston, TX.

Janowsky, D.S., Risch, S.C., Ziegler, M., Kennedy, B.,Huey, L., (1984), *A Cholinomimetic Model of Motion Sickness and Space Adaptation Syndrome,* **Aviat Space Environ Med,** 55:692-696.

Jenkins, G.M., Watts, D.G., (1968), **Spectral analysis and its Applications,** Holden Day, San Francisco.

Johnson, W.H., Jongkees, L.B.N., (1974), *Motion Sickness,* Chapter VIII in **Handbook of Sensory Physiology VI/2: Vestibular System,** ed. H.H. Kornhuber, Springer Verlag, New York.

Jones, D.R., Levy, R.A., Gardner, L., Marsh, R.W., Patterson, J.C., (1985), *Self-Control of Psychophysiologic Response to Motion Stress: Using Biofeedback to treat Airsickness,* **Aviat Space Environ Med,** 56:1152-1157.

Karim, F., Kidd, C., Malpus, C.M., Penna, P.E., (1972), *The Effects of Stimulation on the Left Atrial Receptors on Sympathetic Nerve Activity,* **J Physiol.,** 227:243-260.

Kay, S.M., (1988), **Modern Spectral Estimation: Theory and Application,** Prentice Hall, Englewood Cliffs, NJ.

Kohl, R.L., Homick, J.L., (1983), *Motion Sickness: A modulatory role for the Central Cholinergic Nervous System,* **Neurosci Biobehav Rev,** 7(1):73-85.

Kohl, R.L., (1985), **Proceedings of the Space Adaptation Syndrome Workshop, July, July 1983,** Space Biomedical Research Institute, USRA Division of Biomedicine, Houston.

Kohl, R.L., (1987), *Failure af Metaclopromide to Control Emesis or nausea due to Stressful Angular or Linear Accelerations.,* **Aviat Space Environ Med,** 58:125-131.

Kollai, M., Koizumi, K., (1979), *Reciprocal and non reciprocal action of the vagal and sympathetic nerves innervating the heart,,* **J Auton Nerv Syst,** 1(1):33-52.

Lawther, A., Griffin, M.J., (1988), *A Survey of the Occurrence of Motion Sickness Amongst Passengers at Sea,* **Aviat Space Environ Med,** 59:399-406.

Levy, R.A., Jones, D.R., Carlson, E.H., (1981), *Biofeedback rehabilitation of airsick aircrew,* **Aviat Space Environ Med,** 52(2):118-121.

Mark, A.L., Victor, R.G., Nerhed, C., Seals, D.R., Wallin, B.G., (1986), *Mechanisms of Sympathetic Nerve Responses to Static and Rhythmic Exercise: New Insight from Direct Intraneural Recordings in Humans,* **The Sympathoadrenal System, Alfred**

Benzon Symposium 23, ed. Christensen, N.J., Henriksen, O., Lassen, N.A., Munksgaard, Copenhagen.

Money, K.E., (1970), *Motion Sickness,* **Physiol Rev** 50:1-39.

Ninomiya, I., Yonezawa, Y., Wilson, M.F., (1976), *Implantable electrode recording of nerve signals in awake animals,* **J Applied Physiol,** 41(1):111-114.

Oman, C.M., (1982), *A Heuristic Mathematical Model for the Dynamics of Sensory Conflict and Motion Sickness,* **Acta Otolaryngol Suppl.** 392.

Oman, C.M., (1989), *Motion Sickness: A Synthesis and Evaluation of the Sensory Conflict Theory,* :Proc. Emesis Symposium '88, **Can J Physiol Pharm,** 68(2):294-302.

Papoulis, A., (1984), **Probability, Random Variables and Stochastic Processes, 2nd ed.,** McGraw-Hill Inc, New York.

Parker, D.M., Schaffer, J., Cohen, E., (1972), *The Effect of Past Experience on Motion Sickness produced by Visual Stimuli,* **J General Psychol,** 87:65-68.

Parrot, A.C., (1989), *Transdermal Scopolamine: A Review of its Effects Upon Motion Sickness, Psychological Performance and Physiological Functioning,* **Aviat Space Environ Med,** 60:1-9.

Pomeranz, B., Macoulay, R.J.B., Caudill, M.A., Kutz, I., Adam, D., Gordon, D., Kilborn, K.M., Barger, C., Shannon, D.C., Cohen, R.J., Benson, H., (1985), *Assessment of Autonomic function in humans by heart rate spectral analysis,* **Am J Physiol** 248(Heart Circ Physiol 17): H151-H153.

Rague, B., (1987), *Analysis of Abdominal Slow Potentials during Motion Sickness,* (S.M. Thesis), MIT, Cambridge, MA.

Reason, J.T., Brand, J.J., (1975), **Motion Sickness,** Academic Press, New York.

Reason, J.T., (1978), *Motion Sickness Adaptation: A Neural Mismatch Model,* **J Roy Soc Med.** 71:819-829.

Rice, J.A., (1988), **Mathematical Statistics and Data Analysis,** Wadsworth and Brooks, Pacific Grove, CA.

Risch, S.C., Janowsky, D.S., (1985), *Possible Cholinergic Mechanisms in Space Adaptation Syndrome - Behavioral and Neuroendocrine Correlates* in **Proceedings of the Space Adaptation Syndrome Drug Workshop, July 1983,** ed. Kohl, R.L., pp.33-39, Space Biomedical Research Institute, USRA Division of Space Biomedicine, Houston, TX.

Sakai, F., Meyer, J.S., (1978), *Regional Cerebral Hemodynamics during Migraine and Cluster Headaches Measured by the ^{133}Xe Method,* **Headache,** 18:122-132.

Saul, J.P., Berger, R.D., Cohen, R.J., (1988), *Respiratory Sinus Arrhythmia: A Probe of Autonomic Control of the Heart,* **Am J Cardiol,** 62:500.

Saul, J.P., Berger, R.D., Chen, M.H., Cohen, R.J., (1989), *Transfer function analysis of autonomic regulation II. Respiratory Sinus Arrhythmia.*, **Am J Physiol** 256(Heart Circ Physiol 25): H153-H161.

Saul, J.P., Rea, R.F., Eckberg, D.L., Berger, R.D., Cohen, R.J., (1990), *Heart rate and muscle sympathetic nerve variability during reflex changes of autonomic activity*, **Am J Physiol,** 258(Heart Circ Physiol 27):H713-H721.

Scher, A.A., Ohm, W.W., Bumgarner, K., Boynton, R., Young, A.C., (1972), *Sympathetic and Parasympathetic control of heart rate in the dog, baboon and man*, **Fed Proc,** 31(4):1219-1225.

Simon, E., Riedel, W., (1975), *Diversity of Regional Sympathetic Outflow in Integrative Cardiovascular Control: Patterns and Mechanisms* **Brain Res,** 87:323-333.

Stern, R.M., Koch, K.L., Stewart, W.R., Lindblad, I.M., (1987), *Spectral Analysis of Tachygastria recorded during Motion Sickness*, **Gastroenterol,** 92:92-97.

Stevens, S.S., (1959), *Cross-modality validation of subjective scales for loudness, vibration and electric shock*, **Exp Psychol,** 57:201-209.

Tole, J.R., Yorker, J.G., Morrison, W.A., Renshaw, R.L., (1981), *A Microprocessor-Controlled Vestibular Examination Chair*, **IEEE Trans Biomed Eng,** 28(5):390-396.

Tortora, G.J., Evans, R.L., (1986), **Principles of human Physiology, 2nd ed.,** Harper & Row, New York.

Toscano, W.B., Cowings, P.S., (1982), *Reducing Motion Sickness: A Comparison of Autogenic-Feedback Training and an Alternative Cognitive Task*, **Aviat Space Environ Med,** 53(5):449-453.

Treisman, M., (1977), *Motion Sickness: An Evolutionary Hypothesis*, **Science,** 197:493-495.

Tyler, D.B., Bard, P., (1949), *Motion Sickness*, **Physiol Rev**, 29:311-369.

Van Toller, C., (1979), **The Nervous Body: An Introduction to the Autonomic Nervous System and Behavior,** John Wiley & Sons, New York.

Victor, R.G., Thoren, P., Morgan, D.A., Mark, A.L., (1989), *Differential Control of Adrenal and Renal Sympathetic Nerve Activity During Hemorrhages Hyptotension in Rats*, **Circ Res,** 64:686-694.

Wallin, B.G., (1986), *Functional Organization of Sympathetic Outflow in Man*, in **The Sympathoadrenal System, Alfred Benzon Symposium 23,** ed. Christensen, N.J., Henriksen, O., Lassen, N.A., Munksgaard, Copenhagen.

Warner, H.R., Cox, A., (1962), *A mathematical model of heart rate control by sympathetic and vagus efferent information*, **J Applied Physiol,** 17:349-355.

Wood, C.D., Kennedy, R.S., Graybiel, A., (1965), *Review of Antimotion Sickness Drugs from 1954-1964*, **Aerospace Med,** 36(1):1-4.

Wood, C.D., Graybiel, A., (1970), *A Theory of Motion Sickness based on Pharmacological Reactions*. **Clin Pharm Ther**, 11(5):621-629.

Wood, C.D., Graybiel, A., (1972), *Theory of Anti Motion Sickness Drug Mechanisms*. **Aerospace Med,** 43(3):249-25.

Attachment 6

AFOSR-TR- 91 0982

THESIS BY:

DAVID J. WARD

PRINCETON UNIVERSITY

Subcontract No.# S-789-000-023

STUDIES OF FEEDBACK STABILIZATION OF AXISYMMETRIC MODES IN DEFORMABLE TOKAMAK PLASMAS

David John Ward

A Dissertation

presented to the

faculty of Princeton University

in candidacy for the degree

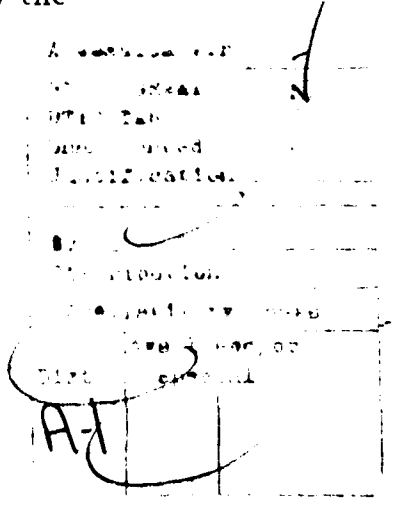
of Doctor of Philosophy

Recommended for acceptance by the

Department of

Astrophysical Sciences

January 1991



Abstract

A new linear MHD stability code, NOVA-W, is described and applied to the study of the feedback stabilization of the axisymmetric mode in deformable tokamak plasmas. The NOVA-W code is a modification of the non-variational MHD stability code NOVA[1] that includes the effects of resistive passive conductors and active feedback circuits. The vacuum calculation has been reformulated in terms of the perturbed poloidal flux to allow the inclusion of perturbed toroidal currents outside the plasma. The boundary condition at the plasma-vacuum interface relates the instability displacement to the perturbed poloidal flux. This allows a solution of the linear MHD stability equations with the feedback effects included.

The code has been tested for the case of passive stabilization against a simplified analytic model and against a different numerical calculation for a realistic tokamak configuration. The comparisons demonstrate the accuracy of the NOVA-W results. The utility and performance of the NOVA-W code are demonstrated for calculations of varying configurations of passive conductors. Active feedback calculations are performed for the CIT tokamak design demonstrating the effect of varying the position of the flux loops which provide the measurements of vertical displacement. The results compare well to those of earlier calculations using a less efficient nonlinear code.

The NOVA-W code is used to examine the effects of plasma deformability on feedback stabilization. It is seen that plasmas with shaped cross sections have unstable motion different from a rigid shift. Plasma equilibria with large triangularity show particularly significant deviations from a uniform rigid shift. Furthermore, the placement of passive conductors is shown to modify the non-rigid components of the motion in a way that reduces the stabilizing effects of these conductors. The eigenfunction is also modified under the effects of active feedback. This deformation is seen to depend strongly on the position of the flux loops. These non-rigid components of the eigenfunction always serve to reduce the stabilizing effect of the active feedback system by reducing the measurable poloidal flux at the flux-loop locations.

[1] C. Z. Cheng and M. S. Chance, J. Comp. Phys. 71 (1987) 124.

Contents

Figures

Tables

Acknowledgments

I would like to thank my advisor, Steve Jardin, for all his help and encouragement over the last several years. He is responsible for piquing my interest in computational MHD and for suggesting this topic to me. His knowledge, experience, and intuition of physics and computations seem boundless. I can only hope that I have gained some small fraction of his knowledge through the privilege of working with him. I would also like to thank Neil Pomphrey for all his advice and encouragement on everything from physics to squash. He never seemed to tire of helping me.

I thank Frank Cheng for giving me the NOVA code, without which none of this would be possible. His help with the intricacies of the NOVA code was indispensible. My thanks go also to Morrell Chance for extremely helpful discussions about the vacuum calculation.

A special thanks to my readers: Steve Jardin, Neil Pomphrey, and Hutch Neilson. Their comments have helped to make this dissertation readable and believable. Their patience and thorough examination of the text is greatly appreciated. Hutch Neilson's experience and knowledge of the experimental and practical aspects of feedback control were very helpful. His comments pointed out many things that would have otherwise gone unnoticed or not considered.

I thank Chuck Kessel for many helpful discussions about feedback control and about all the TSC simulations he has done. Tom Harley gave me much help and advice about NOVA.

For most of my graduate education I was supported under the Laboratory Graduate Fellowship Program of the U.S. Air Force Office of Scientific Research. My thanks to the people at Universal Energy Systems for capably administrating the fellowship program. This work was also supported under U.S. D.o.E. Contract No. DE-AC02-76CHO3073.

My thanks to all the grad students that I have known over the past 5 years for the lighter side of graduate life. My classmates and friends, Mark Bannister, John Bowman, Alain Brizard, John "Bert" Cuthbertson, Don Roberts, Anthony Chan, Bob Duvall, and Ed Powell, have helped me make it through grad school. We had some good times. Thanks to all my officemates over the years (Bob, Bert, Alain, and Billie Schulze-Berge) for putting up with me, and for the thousands of fun conversations we've had. Many thanks to Barbara Sarfaty and Dinah Larsen for their help with all the administrative details and their friendship.

I thank my family for all the encouragement and support that they have given me over the years.

Finally, I would like to thank my wife, Laura, for the happiness, the order, and the love she has brought to my life. Thank you for putting up with me (especially during the last few months); I hope that I can make it up to you. I never could have known that the most important thing I would receive in Princeton would have nothing to do with a Ph.D.

To Chan-ju

"Simple as the case is, there have been one or two very instructive details in connection with it."

Sherlock Holmes, *The Cardboard Box*

Chapter 1

Introduction

1.1 Elongated tokamak plasmas

The fusion of two light nuclei into a heavier nucleus releases large quantities of energy. In particular, the fusion of isotopes of hydrogen or He^3 could provide an incomparable energy source for the future. Controlled thermonuclear fusion in magnetic confinement devices, particularly tokamaks [1,2], gives the most promise of producing economical fusion energy in the near future despite recent unverified claims to the contrary [3].

In order to maintain an energy-producing fusion reaction in tokamaks one must make the Lawson parameter $n\tau_E$ [4] large enough ($\approx 3 \times 10^{20} m^{-3} s$), where n is the density and τ_E is the energy confinement time, at the appropriate plasma temperature $T \approx 15 - 20 keV$. When these conditions are met, there are enough fusion reactions to maintain the power balance in a self sustaining fusion reactor. Another critical parameter is $\beta \equiv 2\mu_0 nT / B^2$, where B is the total magnetic field strength. Specifically, this parameter is the ratio of the plasma thermal energy to the magnetic field energy. It is a figure of merit that is desirable to maximize, since it is related to the fusion power gain from a reactor and ultimately, therefore, to its economics. It has been shown that by increasing the total plasma current one can increase the tokamak β [5] and energy confinement time τ_E [6]. One way to increase the total plasma current without degrading the MHD stability is to modify the cross-sectional shape of the plasma by adding elongation and triangularity [7,8]. Even further cross-sectional shaping such as indentation to produce a bean-shaped plasma as in PBX-M has been shown to increase attainable β [9,10]. The increase in the β-limits that comes with

plasma shaping, for instance, in Doublet III has been shown [11] to be due to the increase in plasma current rather than to any intrinsic properties of the shape itself.

In addition, a divertor plasma (common in modern tokamak experiments) by its nature requires some cross-sectional shaping. A magnetic divertor guides a narrow layer of magnetic field lines near the plasma surface away from the plasma and into an external chamber. In this way, hot particles escaping from the plasma can be directed into a separate exhaust chamber, where the particles are neutralized and pumped away. Therefore, escaping particles are prevented from hitting the vacuum vessel wall and kept from re-entering the plasma along with wall impurities. The use of a divertor has been demonstrated to improve confinement. Axisymmetric, or poloidal-field, divertors are most effective, but necessarily perturb the geometry of the plasma surface. They are easily combined with D-shaped or elliptical cross sections.

Furthermore, a recent study [12] has demonstrated that plasma equilibria naturally become elongated and triangular with decreasing aspect ratio (the aspect ratio is defined by R_0/a, see Fig. 1.1; the ratio of major to minor radii). Unfortunately, plasmas with cross-sectional shaping are typically unstable to modes with zero toroidal mode number. Lortz [13] has explicitly demonstrated that for a large-aspect-ratio equilibrium an arbitrary deviation of the cross section from a circular cylinder is unstable. While toroidal effects are stabilizing for these modes, it remains true that for aspect ratios of interest, $2.5 \leq A \leq 4$, plasmas with elongations greater that about 1.1 or 1.2 will be unstable. These axisymmetric modes have been studied in some detail [14-20]. A basic explanation of the axisymmetric instability has been derived by several authors [21-23]; a summary of the derivation follows.

In a tokamak an equilibrium vertical field $\mathbf{B}_V$ is necessary to balance the hoop force and maintain equilibrium (see, for example, Freidberg [24], p. 77). Consider a plasma equilibrium symmetric about the midplane with a current distribution $J(\mathbf{r})$ (see Fig. 1.1). The vertical force (per unit length) on the plasma column due to the interaction between the plasma current and the equilibrium field is given by

$$F_Z = -I_\phi \times B_V^R = -\int_p J(\mathbf{r})\, B_V^R(\mathbf{r})\, d^2r. \tag{1.1}$$

where the integral is over the cross section of the equilibrium and where B_V^R and B_V^Z are the R- and Z-components of the equilibrium vertical field, respectively.

The direction of the vertical force depends on the direction of the radial component of $\mathbf{B}_V$. In other words, it depends on whether the vertical field is shaped such that it

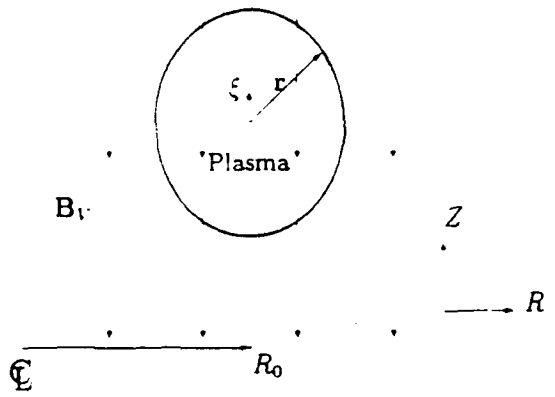


Figure 1.1: *Plasma in a vertical equilibrium magnetic field* $\mathbf{B}_V$ *required to balance the hoop force for equilibrium. The plasma current flows into the page. In this figure the vertical field has no radial component: therefore the plasma is marginally stable. The field lines can curve concave inward (corresponding to positive field index n and stability with respect to the vertical displacment* ξ*) or outward (corresponding to negative field index n and instability).*

curves concave inward or concave outward. One can see from the force law, Eq. (1.1), that a plasma displaced vertically from the midplane will experience a restoring force back toward the midplane if the vertical magnetic field lines are concave inward. However, if the field lines are concave outward the force will push the plasma further from the midplane and there will be instability. Straight vertical field lines (as shown in Fig. 1.1) lead to marginal stability since the restoring force is then zero.

Consider the infinitesimal displacement of a symmetric plasma from the midplane. We perturb the equilibrium by displacing it vertically by ξ. The radial component of the perturbed vertical magnetic field near the midplane can be expanded to first order:

$$B_V^R \simeq B_V^R(\mathbf{r}) - \xi \frac{\partial B_V^R}{\partial Z}. \tag{1.2}$$

Since the vertical field is produced by current-carrying coils external to the region we are considering, we have

$$\nabla \times \mathbf{B}_V = 0 \implies \hat{\phi} \cdot (\nabla \times \mathbf{B}_V) = 0 \implies \frac{\partial B_V^R}{\partial Z} = \frac{\partial B_V^Z}{\partial R}. \tag{1.3}$$

If we use the expanded form for B_V in Eq. (1.1) then we get

$$F_Z(\xi) = -\int_p J(\mathbf{r})\, B_V^R(\mathbf{r})\, d^2r - \int_p J(\mathbf{r})\, \xi\, \frac{\partial B_V^Z}{\partial R}\, d^2r. \tag{1.4}$$

The first integral vanishes because the equilibrium plasma current density $J(\mathbf{r})$ is symmetric about the midplane, and $B(\mathbf{r})$ is antisymmetric across the midplane. Therefore

$$F_Z(\xi) = -\xi \int_p J(\mathbf{r})\, \frac{\partial B_V^Z}{\partial R}\, d^2r \tag{1.5}$$

or

$$F_Z(\xi) = -\xi\, I_p \left\langle \frac{\partial B_V^Z}{\partial R} \right\rangle, \tag{1.6}$$

where

$$\langle \cdots \rangle = \frac{1}{I_p} \int_p J(\mathbf{r})\, (\cdots)\, d^2r \tag{1.7}$$

is the average over the cross section weighted by the current, and the total plasma current I_p is given by

$$I_p = \int_p J(\mathbf{r})\, d^2r. \tag{1.8}$$

Therefore we can define the vertical force as

$$F_Z(\xi) = -\langle n \rangle\, I_p\, \frac{B_V^Z(0)}{R_0}\, \xi, \tag{1.9}$$

where

$$\langle n \rangle = -\frac{R_0}{B_V^Z(0)} \left\langle \frac{\partial B_V^Z}{\partial R} \right\rangle \tag{1.10}$$

is the average of the "field index" n through the plasma cross section. The field index is a dimensionless parameter that measures the amount of curvature of the vertical field lines.

If there are no conductors near the plasma, then Eq. (1.9) represents the total vertical force per unit length on the plasma column. If we assume that the plasma moves rigidly, then the equation of motion for the plasma is

$$m \frac{\partial^2 \xi}{\partial t^2} = -2\pi R_0 \langle n \rangle\, I_p\, \frac{B_{Z0}}{R_0}\, \xi, \tag{1.11}$$

where $B_{Z0} = B_V^Z(0)$, and m is the plasma mass.

If we further assume that the motion is described by $\xi = \xi_0 \exp(\gamma t)$ the growth rate γ is given by

$$\gamma = \left[-\frac{2\pi^2 n \ I_p B_{Z0}}{m} \right]^{1/2}. \tag{1.12}$$

Since all the quantities in the square brackets of Eq. (1.12) are inherently positive except for $\langle n \rangle$, then the perturbation will grow and there will be instability if the average field index $\langle n \rangle$ is negative.

The condition for stability is given by $\langle n \rangle > 0$. The field index is closely related to the plasma cross-section. Zakharov [25] derived an approximate expression for the field decay index for an elliptical plasma of finite but large aspect ratio:

$$n = \frac{\frac{3}{4} \ln \frac{8R_0}{a} - \frac{17}{6} - (1 - \frac{l_z}{l_r}) \frac{R_0^2}{a^2}}{\ln \frac{8R_0}{a} - \frac{5}{4} - \beta}. \tag{1.13}$$

where R, a are the major and minor radii, $\beta = 2\mu_0 p / B^2$ is the plasma beta, and l_z, l_r are the vertical and horizontal radii of the ellipse. Thus, for any aspect ratio R_0/a, there is a critical elongation above which the field index n will be negative. When $n < 0$ the ellipse is unstable.

Therefore cross-sectional shaping leads to axisymmetric instability (at least in the large-aspect-ratio limit). Furthermore, it is clear from Eq. (1.12) that there is instability for any value of I_p if $\langle n \rangle$ is unfavorable—unlike, for example the external kink mode, where there is a threshold of stability, and one can stabilize the mode by decreasing I_p or by modifying the current profile.

The axisymmetric instability results in the gross plasma motion up or down away from the midplane on the ideal MHD time scale. This results in the rapid termination of the tokamak discharge as the plasma comes into contact with the surrounding vacuum vessel structure. Such a disruption of the plasma causes the rapid quenching of the plasma current, which in turn can induce tremendous forces on the tokamak vacuum vessel and support structure, which can be very damaging. The forces on the vacuum vessel resulting from a disruption following the loss of vertical control of a tokamak plasma have been calculated by Jensen and Chu [26].

The vertical instability was first considered in the simple infinite-aspect-ratio limit. Rutherford [15] examined the axisymmetric stability of incompressible elliptical plasmas with constant current-density in the infinite-aspect-ratio limit using the energy principle for perturbations of the form $\xi(r, \theta) = \sum_{m=1}^{\infty} \xi_m(r) \sin(m\theta)$. He found that

for any elongation greater than zero there is instability to an $m = 1$ rigid motion. For plasma elongations below approximately 4.5 the instability is composed of purely rigid ($m = 1$) motion, but at higher elongations an additional mode with $m = 1$ and $m = 3$ components arises. Rosen [16] examined rectangular plasmas in a similar limit. He found that an equilibrium with a square cross section was stable to a pure rigid-shift, but unstable to a rigid-shift combined with an additional $m = 3$ perturbation. (We will review this model in some detail in Sec. 2.2.) Rebhan [17] considers stability of the analytic Solov'ev equilibrium [27] with various cross-sectional shapes to a pure rigid displacement. Finite aspect ratio (toroidal effects) and strong triangular deformation were both found to be stabilizing for this case. At infinite aspect ratio any elongation leads to instability, but with finite aspect ratio stability can be achieved if the elongation is kept small enough.

Okabayashi and Sheffield [18] used a model in which an elongated toroidal plasma is represented as a collection of current filaments. They calculated the energy change due to a uniform rigid displacement of the filaments using a circuit equation analysis. They found stability for elongations ≤ 1.3 for an elliptic cross section, and stability for rectangular cross sections with elongations ≤ 3.0. They also found a strong dependence of the stability on the form of the current profile for the rectangular plasma, but not for the elliptical one.

Rebhan and Salat [19] expanded Rebhan's model for more general "slip" motions (so-called because the plasma "slips" vertically through the stationary toroidal field) in constant-pressure, surface-current equilibria. They found that in all cases the stability boundaries for the minimized general displacement are more pessimistic than the rigid displacement. The stability boundaries are close only in the case of nearly elliptical cross sections. They also explicitly show the increase of stability with decreasing aspect ratio, as well as the stabilizing effect of increasing β_p.

Bernard et al. [28] used the ERATO ideal MHD stability code [29] to examine stability of axisymmetric modes with and without a perfectly conducting wall. Plasma triangularity was found to be destabilizing, in contrast to the rigid model of Rebhan [17]. Others have also found that triangularity destabilizes [19,30]. Bernard et al. also found that for little or no triangular deformation the critical elongation is nearly the same for the ERATO results and the pure rigid-shift model. For a case with zero triangularity the displacement is vertical and rigid. For a D-shaped plasma the displacement is vertical and non-rigid. For high triangularity ERATO predicts lower

critical elongation (more unstable) than the rigid model.

1.2 Wall stabilization of the axisymmetric mode

The axisymmetric vertical instability can be stabilized by placing a superconducting wall close enough to the plasma surface [14,31,32]. Any displacement of the plasma will induce image currents in the conducting wall that oppose the motion of the plasma. The stabilizing effect of ideal conducting walls has been studied by several authors [28,30,33–35].

For plasmas stabilized by ideal walls the current profile effects are important [34,35], while pressure profile effects are small [32,34,36]. Both of these results are in contrast to the case with no wall. It is easy to see why current profile effects become important when one has a conducting wall near the plasma. If one peaks the current distribution (while leaving the total plasma current unchanged) it has the effect of moving the plasma current (inward) away from the conducting wall. This is equivalent to moving the conducting wall away from the plasma current, which is destabilizing. Therefore the critical wall distance for stability is smaller. On the other hand, a broader current profile puts more current near the wall and therefore allows a larger critical wall distance. Hoffman et al. [35] found that decreasing the width of the radial current profile of highly elongated racetrack equilibria by a certain amount has the same effect as increasing the distance between the ideal side walls by approximately twice that amount.

Pressure profile effects (the effects of increasing β_p) have been found [30,34] to be small and destabilizing. This is in contrast to the study by Rebhan and Salat [20] for skin-current equilibria with no surrounding conductors in which they find that increasing β_p is significantly stabilizing. Haas [32] found for elliptical plasmas given uniform vertical displacements that high-β/toroidal effects are important only when a conducting wall is present, but even then only when β is near the equilibrium β-limit.

As one might expect, the conducting wall needs to be brought closer to the plasma to obtain stability at higher elongations. It was demonstrated [28], however, that for the Solov'ev equilibrium with a small, fixed triangular deformation the critical wall distance decreases with increasing elongation to a minimum (at elongation of about 2.6) and then increases again. For high elongations the relevant critical parameter was

found to be the ratio between wall and plasma heights, which is a strictly decreasing function of elongation.

It has been shown [33,35] that for extremely elongated plasma equilibria the side walls are much more effective in stabilizing the mode than the top and bottom walls. The reason given was that the stabilizing eddy currents in the side walls are much closer to the plasma current channel than the top and bottom walls. Hofmann et al. [35] also demonstrated that the outboard side wall is more effective in stabilizing than the inboard side wall owing to the larger radial magnetic field per unit current induced by the outboard wall.

Passive stabilization with discrete conductors near the plasma, such as poloidal field coils, have also been considered [37,38]. The way in which these coils are connected to form a circuit was found to be critical when these coils provide the primary stabilizing influence.

It is well known, therefore, that when an ideally conducting wall is placed close enough to any plasma the axisymmetric modes will be stabilized. Unfortunately, ideally conducting walls do not exist in reality, and any conductors near the plasma almost certainly will have some resistivity. Since the walls have resistivity, the eddy currents induced in the conductors by the unstable plasma motion will decay on the characteristic L/R time of the conductors, and therefore the plasma cannot be stabilized indefinitely. While early tokamak experiments may have had plasma lifetimes that were short enough that the resistive wall acted like an ideal conductor on that time scale, modern experiments yield plasmas with durations many times the L/R time of the surrounding conductors.

It has been shown [39] that an unstable plasma that has been stabilized by being placed inside a perfectly conducting wall will again become unstable when resistivity is present in the wall. This instability will be much slower now, such that it is stable on the ideal time scale of the original motion, but it is now unstable on the resistive time of the wall. This is much shorter than the desired duration of modern tokamak plasmas; therefore, active feedback is required to stabilize the plasma over longer time scales. Fortunately, the growth times are long enough to be amenable to feedback stabilization by practical power supplies.

Passive stabilization of the axisymmetric instability by resistive conductors has also been considered [38–41]. Absolute stability is no longer a question in this case,

because the plasma is definitely unstable on the resistive time scale of the conductors. Rather, it is of interest to calculate the actual growth rate of the instability of a given plasma equilibrium with regard to the form of the passive conductors. This is particularly important so that one can see if the instability has been slowed enough to allow an active feedback system to further stabilize the plasma.

Most of the considerations for stability with the ideal wall apply to the resistive wall case as well. For instance, it was demonstrated [41] that peaking of the current density profile dramatically increases the resistive wall growth rate of the instability.

1.3 Active feedback stabilization of the axisymmetric mode

We now consider a vertically unstable plasma surrounded by passive conductors, which slow the instability from the ideal MHD time scale (typically 10–100 μs) to the resistive time scale of the surrounding conductors (typically 10–100 ms). On this much slower time scale an active feedback system [38,42] can stabilize and control the plasma motion over the duration of the discharge. An active feedback system consists of some means of measuring the vertical displacement of the plasma, and then a set of power supplies (controlled by this measurement) that drives appropriate currents in the active feedback coils that oppose the unstable plasma motion. Active feedback systems in tokamaks typically use measurements of the asymmetric poloidal flux (taken by pairs of flux loops symmetric about the midplane) to specify the plasma displacement and thereby control the power supplies. Other measurements such as the time rate of change of the asymmetric flux or perturbed magnetic field measurements can also be used to improve the feedback system performance.

Jardin and Larrabee [38] used a circuit model to examine active feedback with the active coils far away from the plasma. It was found that the plasma can be stabilized if the mutual inductance between the active coils and the vacuum vessel is not too great. This mutual inductance is destabilizing since it causes the passive conductors to effectively shield the plasma from the active elements. Furthermore, lowering the cutoff frequency (the frequency limit at which the active feedback system can respond) was found to be destabilizing.

Rebhan and Salat [42] use a $\delta W'$ treatment to examine skin-current equilibria with

an active feedback system. The fast time scale motion is assumed to be slowed by a conducting wall so that the feedback system has time to respond, but the interaction with the passive conductors is not explicitly considered. They address the problem of active-feedback-coil location with respect to the plasma, for different cross-sectional shapes, and the optimum coil currents and locations.

Small areas of ineffective feedback coil locations were found near the midplane. For large-aspect-ratio equilibria ($A = 10$) these ineffective regions are more symmetric, whereas for the lower-aspect-ratio equilibria ($A = 3$) the ineffective regions on the outboard side are much smaller, and the ineffective regions on the inboard side are significantly larger—therefore the toroidal effects are apparent. It was also shown that the pressure (β_p) effects on the optimum coil position and currents are very small. The optimum feedback currents are, however, affected by the distance from the plasma and strongly by the poloidal position of the coils.

The axisymmetric stability of the PBX tokamak has been examined [43] with the free-boundary, axisymmetric tokamak simulation code TSC [44]. The accuracy of the code was demonstrated through comparison with controlled experimental shots on PBX. It was demonstrated that PBX operates near the instability boundary (i.e., the stability boundary with ideal conductors). Small inward radial movements of the plasma (to regions of more unstable field index) can cause the loss of axisymmetric stability as the plasma becomes so unstable that the active feedback system is insufficient to control it. The rapid increase in the resistive wall growth rate as the ideal stability-limit is neared with these successive inward radial displacements was also demonstrated. The loss of stability is still within the limit of ideal MHD stability. It was therefore concluded that demonstration of ideal stability is necessary but far from sufficient to guarantee stability with an active feedback system.

The stabilization of the vertical motion is often considered in the greater context of position control [45–47]. A detailed analysis of the vertical stability control system in DIII-D has been given by Lazarus et al. [47]. A simple massless-plasma, single-filament model is used to study the control problem. The model is shown to be a second-order dynamical system. An eigenmode analysis of the vacuum vessel is used, and a critical index for the active feedback coils is defined above which the system is unstable in the absence of derivative gain. The inboard coils were shown to be much more effective in stabilization than the outboard coils owing to the interference of the latter with the passive stabilizing effects of the vacuum vessel wall.

An experimental study [48] of vertical stability on DIII-D verified the model in Ref. [47] at elongations up to about $\kappa = 2$, but at higher elongation with high triangularity it was found that the plasma is destabilized by the coupling of an $m/n = 3/0$ mode to the $m/n = 1/0$ mode. This non-rigid contribution led to loss of stability where the rigid body single-filament [47] and multi-filament [49] models predict stability.

It has been shown [50] that the non-rigid aspects of the unstable axisymmetric mode in highly shaped tokamak plasmas can be of critical importance in the ability of an active feedback system to stabilize the motion. Specifically, it was found that for certain locations of the flux-loop detectors (which would properly detect rigid motion) the plasma remains unstable regardless of the feedback gain. This is due to the plasma's ability to deform in such a way that the detected signal is insufficient to control the plasma. This article will be reviewed in detail in Chapter 2 because of its relevance to this dissertation.

1.4 A linear MHD stability code with active and passive feedback

The methods used in the various studies in the preceding sections all have some drawback. The current-filament, rigid-body models do not properly represent the true plasma motion when the motion has significant non-rigid components. Such a non-rigid contribution is typically present when there is strong cross-sectional shaping—particularly triangularity. Variational approaches suffer the same limitation when the trial function does not closely approximate the true eigenfunction. This limitation can be made worse if the feedback system actually modifies the eigenfunction in some way. This would be very difficult to represent with some standard trial function. Ideal MHD stability codes such as PEST and ERATO can accurately resolve the true eigenfunction of the instability, but are limited to using ideal walls with no active feedback. Transport time-scale simulation codes such as TSC can accurately compute the full nonlinear axisymmetric motion with all the realistic control aspects such as an active feedback system, resistive conductors, circuit and power supply dynamics. It is, however, very expensive computationally to obtain converged quantitative results with TSC, especially when tokamak parameter scans are needed.

It is evident, therefore, that there is a need for a linear MHD stability code that can calculate the eigenfunctions and growth rates for the axisymmetric mode of deformable tokamak plasmas with passive stabilization from arbitrary resistive conductors and an active feedback system. For this reason, we have developed the NOVA-W code. This is a modification of the ideal, linear MHD stability code NOVA [51], which includes all the nonideal aspects just mentioned. We are therefore particularly well suited to examine the effect of active and passive feedback on the vertical instability in tokamaks with a particular interest in the influence of the non-rigid components of the motion.

An outline of this dissertation is as follows. In the next chapter we review an article [50] that demonstrates the importance of the non-rigid motion of highly shaped tokamak plasmas and how these affect the performance of an active feedback system. The following chapter presents the basic formulation of the calculation involved in the NOVA-W code. In Chapter 4 some results of passive stabilization are presented. First the code is benchmarked against a simplified analytic model and then against the results from the Tokamak Simulation Code. The capabilities of the new code are demonstrated with results for two different tokamak design equilibria. In Chapter 5 we present some initial results for the active feedback stabilization. The TSC is again used to compare to the results from NOVA-W in a qualitative and semi-quantitative sense. Finally, in Chapter 6 we examine a highly triangular plasma equilibrium and study the interaction of the active and passive feedback with the non-rigid components of the motion. Chapter 7 will contain a brief summary and discussion as well as plans for future work.

Chapter 2

Feedback Stabilization of a Deformable Tokamak Plasma

A deformable plasma and its active feedback control system together form a coupled system. In this chapter we show that the stability of this coupled system can be substantially different than the stability of a similar system which allows only rigid plasma motion. In particular, for certain locations of the feedback sensors the deformable plasma motion can render the feedback system ineffective. This demonstrates the need to examine axisymmetric plasma stability of the plasma and its feedback system together. Of particular interest is the interaction of the feedback system with the non-rigid components of the plasma motion in highly shaped tokamak plasmas.

This chapter is primarily a review of the results presented in the article appearing in Nuclear Fusion entitled "Feedback Stabilization of the Axisymmetric Instability of a Deformable Tokamak Plasma" by Pomphrey, Jardin, and Ward [30]. The numerical results of this paper are reviewed in the following section, and the analytic model is presented in Section 2.2 in a somewhat expanded form with additional analysis and discussion of its relevance to this dissertation.

2.1 Numerical Model using TSC

Figure 2.1 shows a schematic of a bean-shaped plasma for which vertical position control is a necessary element of the design. The plasma carries 0.75 MA current, and has a cross section that is elongated and indented on the inboard side. Conducting

plates surround the plasma, which effect passive stabilization on the ideal MHD time scale. The L/R time for the passive conductors is 100 msec. The poloidal field coil system used for equilibrium, shaping, and feedback control is shown, as are two pairs of observation points. The flux difference between the top and bottom members of any one of these pairs is a measure of the displacement of the plasma from its equilibrium position on a midplane. The top and bottom points in each pair are symmetric about the midplane, and therefore this flux difference would provide an accurate measure of vertical displacement in the case of truly rigid plasma motion. Apart from the location of the pickup loops, the system shown in Fig. 2.1 is an accurate representation of the modified Princeton Beta Experiment (PBX-M) [52].

The Princeton Tokamak Simulation Code (TSC) [44] was used to analyze vertical position control issues for the configuration described above. The TSC accurately models the transport time-scale evolution of axisymmetric plasmas, including the plasma interaction with passive and active feedback systems. For the simulations described below, a simple feedback control law is employed in which an incremental current is requested from the vertical feedback coils in proportion to the measured flux imbalance between one or the other of the observation pairs, i.e.,

$$I_W(t) = \mathcal{G} \times \left[\psi_i^{top}(t) - \psi_i^{bot}(t)\right], \tag{2.1}$$

where $i = 1$ refers to the inboard observation pair, and $i = 2$ refers to the outboard observation pair, and where *top* refers to the top loop of the pair and *bot* refers to the bottom loop.

In the passive sense, both the inboard and the outboard observation pairs are equally good at detecting the vertical motion of the plasma. To see this, Fig. 2.2, Case A, shows the results of a simulation in which the active feedback system is turned off by setting the gain $\mathcal{G}$ equal to zero. The flux differences $\Delta\psi_i(t) = \psi_i^{top}(t) - \psi_i^{bot}(t)$ are plotted as a function of time for each observation pair. Note that the same growth rate for the instability is calculated by TSC using either observation pair, and that the amplitude of the flux detected by each pair of loops is essentially the same, which implies that the vertical displacement is nearly rigid. It will now be seen that although both observation pairs detect the unstabilized motion equally well, only the outboard pair can be successfully incorporated into the active feedback scheme defined by Eq. (2.1).

In Figure 2.2 two initial-value, time-dependent simulations with TSC are shown

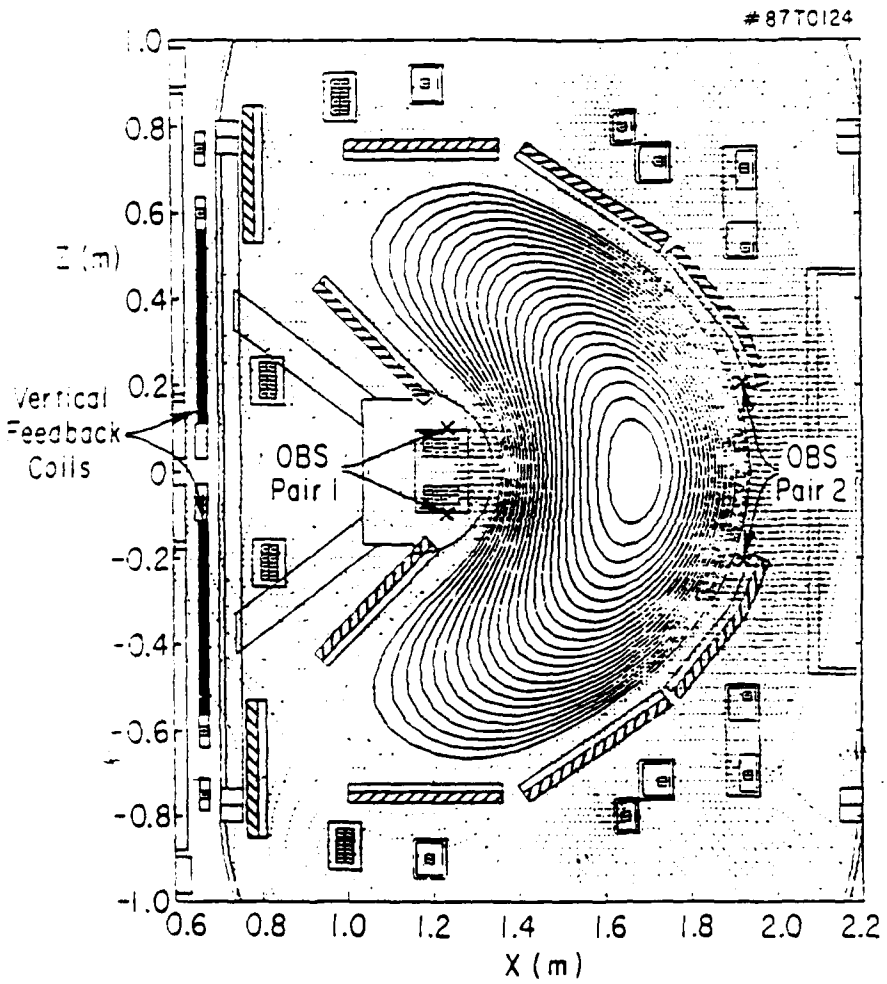


Figure 2.1: Schematic of the Modified Princeton Beta Experiment (PBX-M). The inboard and outboard observation pairs used for feedback control in the numerical simulations are denoted by crosses.

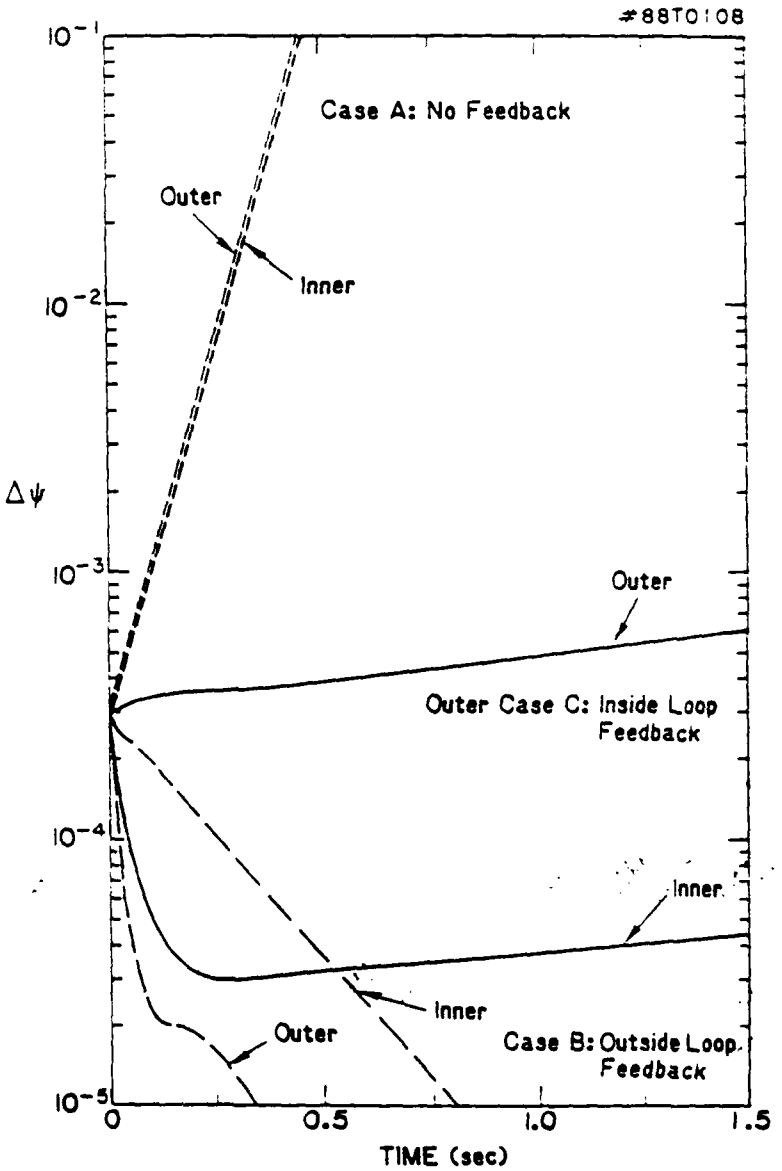


Figure 2.2: The flux differences $\Delta\psi_i(t) \quad \psi_i^a(t) - \psi_i^b(t)$ for the inner and outer observation pairs are plotted as a function of time. Case A: Simulation results when the feedback gain, β, equals zero. The same growth rate is obtained using either observation pair.

Case B: Active feedback with the feedback coils connected to the outside flux loops The plasma is stable.

Case C: Active feedback with the feedback coils connected to the inside flux loops The plasma is unstable.

The same feedback gain values were used in Cases B and C.

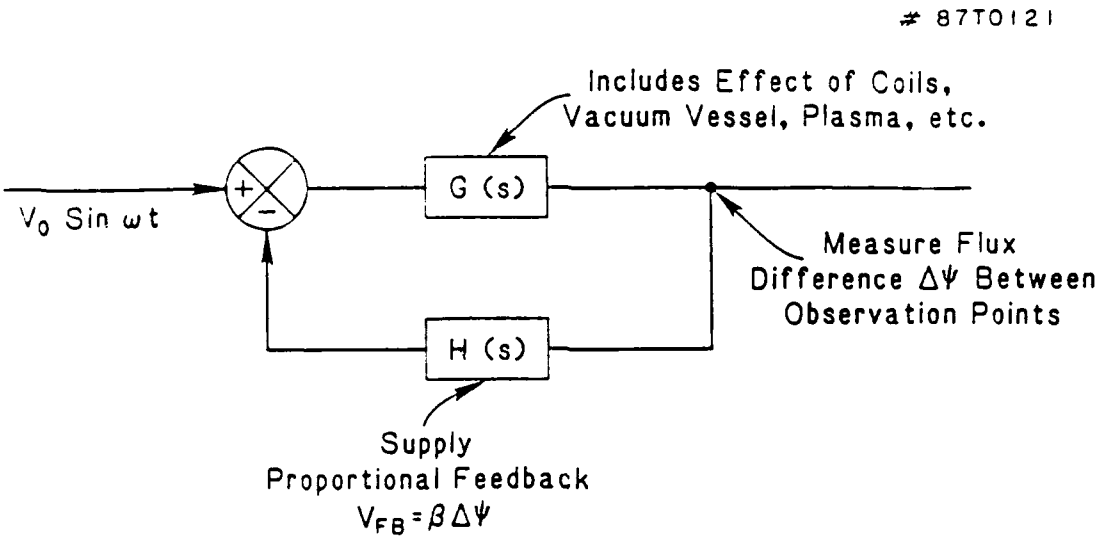


Figure 2.3: Block diagram for the frequency-response analysis of the control system.

using the same initial conditions for the plasma as in the passive calculation, Case A, but with the feedback system activated and connected first to the outside flux loops (Case B) and second the inside flux loops (Case C). For each simulation the feedback gain has the same value as in the frequency-domain analysis of Fig. 2.4.

For Case B, in which the feedback system is connected to the outer observation pair, the plasma motion is seen to be stabilized. The flux differences measured by the inner and the outer observation pairs both decrease with time, with the flux difference between the outer loops almost an order of magnitude less than the flux difference between the inner loops. This indicates some plasma distortion. For Case C, in which the feedback system is connected to the inboard observation pair, the plasma remains unstable, albeit with a much reduced growth rate. The flux difference between the inner observation loops is now an order of magnitude less than the flux difference between the outer loops, also indicating a plasma distortion, but one whose detailed form is different from that of with Case B.

In order to investigate the stability of the feedback system in more detail, some techniques of control engineering [53,54] are adopted. Figure 2.3 is a block diagram for the system; it shows the relationship between components, and the flow of signals

from input to output. A reference voltage signal is input to the vertical feedback coils, the plasma/conductor/vacuum MHD equations are advanced by TSC, and the measured flux difference between a pair of observation points is output. The feedback loop is closed by amplifying the output, and returning it to the feedback coils as a voltage correction to the reference signal. For stability, all poles of the closed-loop transfer function, $T(s) = G(s)/[1 - \mathcal{G}G(s)]$, must have negative real parts (i.e., must lie in the left half s-plane). Here $s = -i\omega$, therefore such an s implies stability for motion described by $\exp(-i\omega t)$. The location of the poles is determined by the Encirclement Theorem, which leads to the Nyquist criterion for stability:

$$\frac{1}{2\pi}\Delta \arg(1 - \mathcal{G}G(s))\Big|_{-i\infty}^{-i\infty} = N_p. \tag{2.2}$$

The Nyquist criterion for stability formally stated [54] reads: For a closed-loop system to be stable, the Nyquist plot of $\mathcal{G}G(s)$ must *encircle* the $(-1,0)$ point in the complex plane as many times as the number of poles of $\mathcal{G}G(s)$ that are in the right half of the s-plane, and the encirclement, if any, must be made in the counterclockwise direction for a clockwise trajectory of s around the right half-plane.

The phase change of the transfer function on the left-hand side of Eq. (2.2) can be interpreted as the number of counterclockwise encirclements of the point (-1,0) by the $\mathcal{G}G(s)$ locus as s is increased from $-i\infty$ to $-i\infty$, and N_p is the number of poles of $\mathcal{G}G(s)$ having positive real parts. For a vertically unstable plasma partially stabilized by resistive walls there is generally one unstable root in the MHD equations for the axisymmetric mode; therefore it is possible to show that $N_p = 1$.

The open-loop transfer function $G(s)$ is not expressed in closed form for this problem, and must be evaluated numerically. To do this, a test signal is input to the feedback coils in the form of a sinusoid with frequency ω. The steady-state response characteristics of a stable system are such that $|\mathcal{G}G(i\omega)|$ is equal to the amplitude ratio of the output and input sinusoids, and $\arg[\mathcal{G}G(i\omega)]$ is the phase shift of the output sinusoid with respect to the input sinusoid. The data are collected on opposite sides of the summing point (see Fig. 2.3).

In Fig. 2.4, the results obtained from running TSC are presented using first the inboard, and then the outboard observation pair for monitoring the flux. The sign and magnitude of the gain are the same for both cases. Results are shown as a Bode diagram, which consists of two graphs. One is a plot of $\log|\mathcal{G}G(i\omega)|$ versus ω; the

other, a plot of $\arg[\beta G(i\omega)]$ versus ω. Once the Bode diagram is constructed, the Nyquist plots follow readily. Figure 2.4 shows that use of the outboard observation pair (Pair =2) gives rise to a closed curve that meets the conditions required by the Nyquist stability criterion. On the other hand, the Nyquist curve obtained using the inboard observation pair (Pair =1) not only fails to enclose the point (-1.0), but it is also described in the wrong sense (clockwise). Since β is a constant, changing it simply scales the distance of each point on a curve to the origin, but leaves the sense of traversal unchanged. Therefore the feedback system which uses the outboard observation pair will be stable for a finite range of β, corresponding to the enclosure of (-1.0), whereas the feedback system which uses the inboard observation pair will be unstable for all values of β.

The essential difference in behavior of the two feedback systems is the response to low frequencies. At very high frequency the signal from the feedback coils is unable to affect the plasma motion because it cannot penetrate the intervening passive conductors. The feedback system is completely passive in this limit. If the frequency is lowered to become comparable to the inverse L/R time of the conductors, the signal has time to influence the plasma motion. The influence is seen in the Bode plots as a dramatic change in slope in the curves of amplitude and phase. When the frequency is lowered toward the zero-frequency limit, the contrast between using the different observation points is most clear. In this limit, the passive stabilizers are completely transparent to the feedback signal and therefore cannot affect the feedback response of the plasma. It is in this low-frequency limit that the placement of the flux detection loops can determine the overall stability of the system.

In summary, the unstable eigenfunction depends on the position of the flux loops used to detect the motion, even though the feedback coils, in which the feedback currents appear, are exactly the same in the two cases. The stability is now examined in the low-frequency limit with an analytic model.

2.2 Analytic Model

It was demonstrated by Rosen [16] that a constant current-density, square cross-section, straight plasma column is unstable to a non-rigid axisymmetric instability. In this section the analysis of Rosen will be modified to include an active feedback system, whose response is based on perturbed flux measurements, and it will be

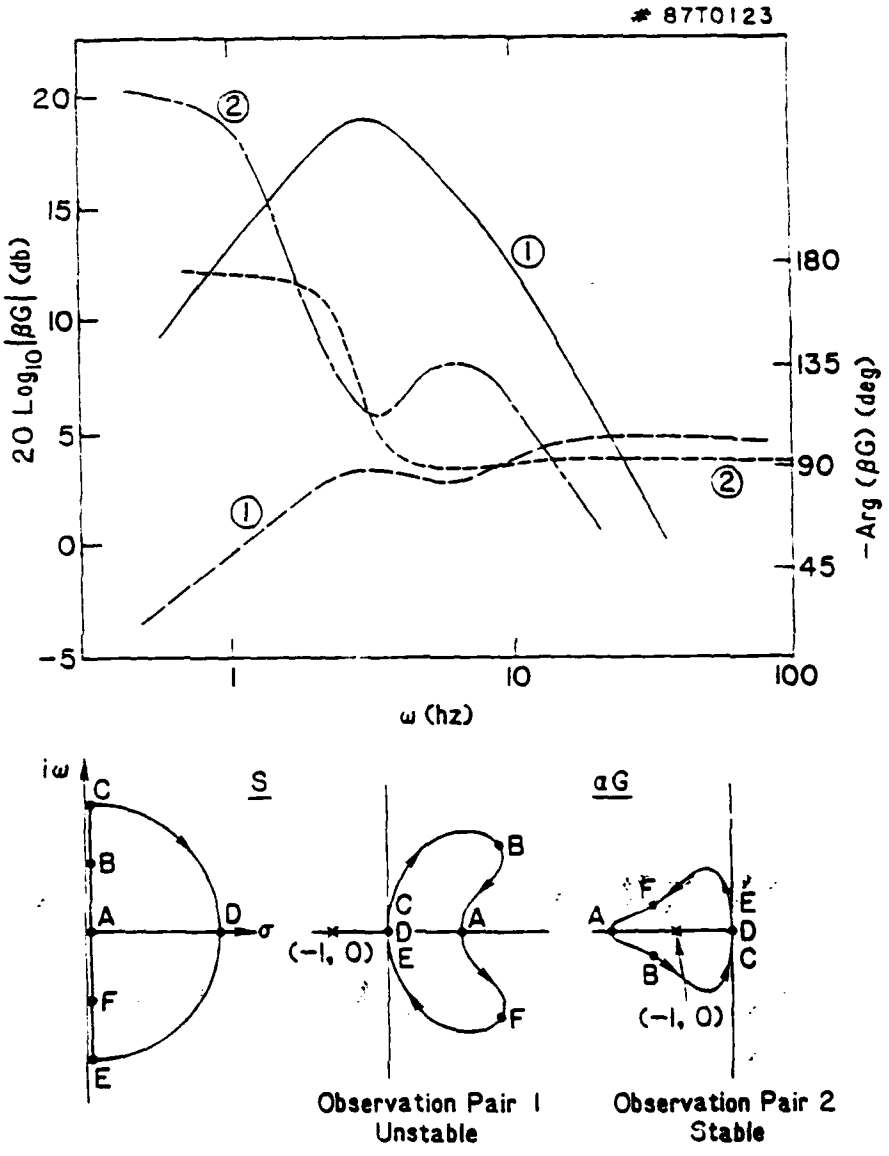


Figure 2.4: Top: Frequency dependence of the amplitude and phase transfer function G when observation pairs 1 and 2 are used to monitor the flux.
Bottom: Nyquist curves resulting from the mapping to the complex G plane.

shown that there are forbidden regions for the placement of the flux-loop detectors if stability is to be ensured.

First we review the results of Rosen [16] for a square cross-section plasma. (Rosen considers the more general case of a rectangular cross-section, but we will restrict the problem to that of a square plasma for simplicity.) We then consider an appropriate feedback system for this model. We will be using δW techniques [55] so we must adapt the model to satisfy the self-adjointness property of the force operator if the δW method is to be valid. The necessary constraints on the model feedback system to ensure this validity will be derived. Then the stability of the complete system, plasma – feedback, will be analyzed.

2.2.1 Review of the model with no feedback

We consider an infinitely long cylindrical plasma of constant current density j_z surrounded by a vacuum region. The longitudinal field B_z is uniform, and $B_z \gg B_\perp$, where $\mathbf{B}_\perp = \mathbf{z} \times \nabla\psi$. Ampere's Law then gives

$$\nabla^2\psi = j_z. \tag{2.3}$$

Given a constant current-density plasma, the equilibrium poloidal flux is given by

$$\psi = r^2 - \alpha r^4 \cos 4\theta \tag{2.4}$$

where α is the "squareness parameter," which is assumed to be small, and will be used as an expansion parameter in the calculation to follow. If $\alpha = 0$, then the plasma-vacuum interface ($\psi = 1$) is a circle. Even modest values of α, such as $\alpha = 0.2$, make the $\psi = 1$ surface nearly square. Figure 2.5 shows a schematic of the plasma, the poloidal field coils, and typical pairs of observation coils.

The most general form for δW of the energy principle can be written (see, for example, Freidberg)[24]

$$\delta W = \delta W_F - \delta W_S - \delta W_V \tag{2.5}$$

where

$$\delta W_F = \frac{1}{2}\int_P d\mathbf{r}\left[\mathbf{Q}_\perp^2 - \xi_\perp^* \cdot (\mathbf{j}\times\mathbf{Q}) - \gamma p|\nabla\cdot\xi_\perp|^2 - (\xi_\perp\cdot\nabla p)\nabla\cdot\xi_\perp^*\right] \tag{2.6}$$

$$\delta W_S = \frac{1}{2}\int_S dS\,\hat{\mathbf{n}}\cdot\xi_\perp\;\hat{\mathbf{n}}\cdot\left[\left[\nabla\left(p+\frac{B^2}{2}\right)\right]\right] \tag{2.7}$$

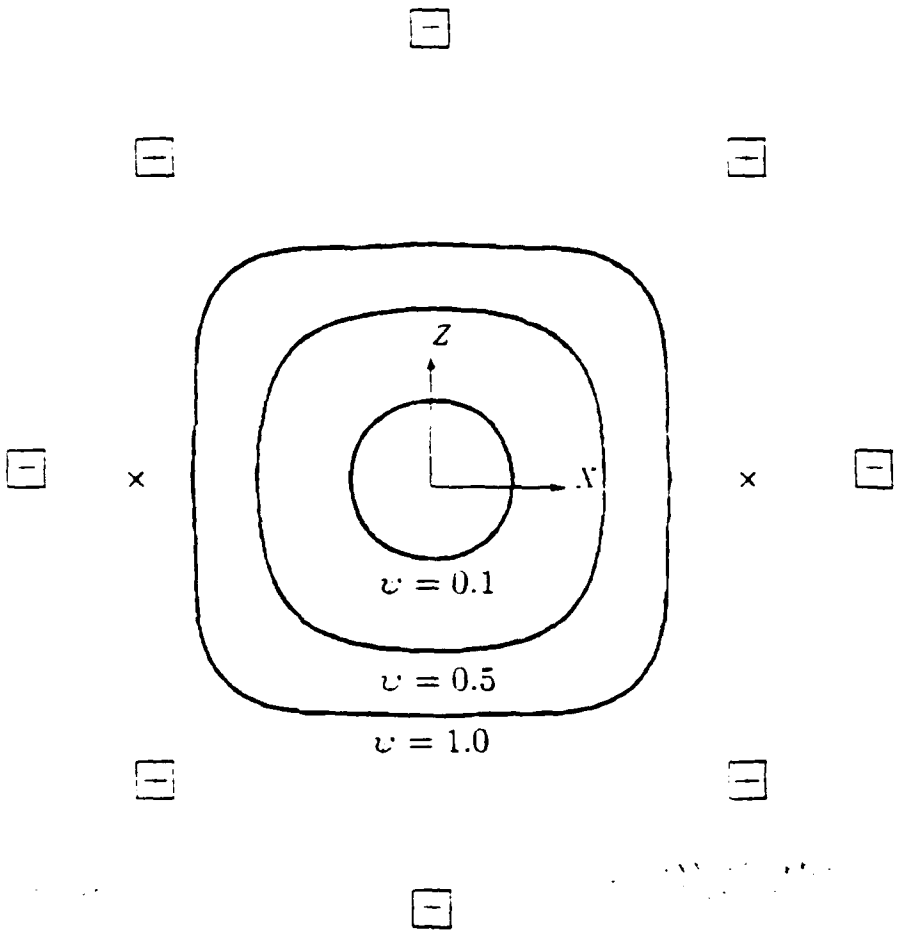


Figure 2.5: *Contours of constant flux for the square plasma analytic model. The equilibrium magnetic field is produced by coils which push on the sides and pull at the corners. The direction of the current flow in the field-shaping coils is denoted by $-$ or $-$ for current parallel or antiparallel, respectively, to the equilibrium plasma current. Typical observation points for perturbed flux measurements to determine plasma position are denoted by $\times$ symbols.*

$$\delta W_V = \frac{1}{2}\int_S dS(\hat{\mathbf{n}}\cdot\xi_-)(\mathbf{B}_-\cdot\mathbf{b}). \tag{2.8}$$

Here

$$\mathbf{Q} \equiv \nabla\times(\xi\times\mathbf{B}) \tag{2.9}$$

is the perturbed magnetic field in the plasma. ξ is the displacement. b is the perturbed magnetic field in the vacuum. and $[\ldots]$ denotes the jump from the plasma to the vacuum. δW_F is the contribution from the plasma. δW_S is the contribution from the plasma-vacuum interface. and δW_V is the contribution from the vacuum. In this case $\delta W_S = 0$ because there are no surface currents in the problem. The magnetic field perturbation in the plasma is defined by

$$\mathbf{Q} = \hat{\mathbf{z}}\times\nabla\phi_p \tag{2.10}$$

where ϕ_p is the perturbed flux in the plasma. Likewise the perturbed magnetic field in the vacuum b is defined in terms of ϕ_v:

$$\mathbf{b} = \hat{\mathbf{z}}\times\nabla\phi_v \tag{2.11}$$

We consider an incompressible plasma:

$$\nabla\cdot\xi = \nabla_-\cdot\xi_\perp = 0. \tag{2.12}$$

The displacement ξ is related to the perturbed flux as

$$\phi_p = -\xi\cdot\nabla\psi. \tag{2.13}$$

Therefore the form of δW can be expressed as separate contributions from the plasma and from the vacuum:

$$2\delta W = \int_p [(\nabla\phi_p)^2 - j_z\xi\cdot\nabla\phi_p]dA - \int_v [(\nabla\phi_v)^2 - \phi_v\nabla^2\phi_v]dA. \tag{2.14}$$

where the subscripts p and v refer to plasma and vacuum, respectively. To evaluate this expression for the given model, it is convenient to express δW in terms of the flux coordinates (ψ,θ,z). Thus we write

$$\delta W = \delta W_p - \delta W_j - \delta W_v \tag{2.15}$$

or

$$2\delta W = \int\int d\psi d\theta \mathcal{J}(\nabla\phi_p)^2 - j_z\oint d\theta\mathcal{J}\phi_p^2 - \oint d\theta\mathcal{J}\phi_v(\nabla\psi\cdot\nabla\phi_v), \tag{2.16}$$

where

$$|\nabla\phi_p|^2 = |\nabla\psi|^2\left(\frac{\partial\phi_p}{\partial\psi}\right)^2 - 2\nabla\psi\cdot\nabla\theta\left(\frac{\partial\phi_p}{\partial\psi}\right)\left(\frac{\partial\phi_p}{\partial\theta}\right) + |\nabla\theta|^2\left(\frac{\partial\phi_p}{\partial\theta}\right)^2. \tag{2.17}$$

and

$$\mathcal{J} = |\nabla\psi\cdot\nabla\theta\times\nabla z|^{-1} = \frac{1}{2}\frac{dr^2}{d\psi} \tag{2.18}$$

is the Jacobian. The integral denoted $\oint$ is evaluated around the $\psi = 1$ contour, and $J_z = 4$ in the units of Eq. (2.4). Since α is small, Eq. (2.4) can be inverted to obtain $r(\psi,\theta)$, from which the metric elements and the Jacobian can be evaluated. To order α^2 we expand the Jacobian and $r(\psi,\theta)$:

$$\mathcal{J} = 1 - 2\alpha\psi\cos 4\theta - 3\alpha^2\psi^2(1-\cos 8\theta) \tag{2.19}$$

$$r = \psi^{1/2}\left[1 - \frac{1}{2}\alpha\psi\cos 4\theta - \frac{1}{8}\alpha^2\psi^2\cos^2 4\theta\right]. \tag{2.20}$$

The incompressibility constraint can be expressed as

$$\oint\xi\cdot\hat{\mathbf{n}}dl = \oint(\xi\cdot\nabla\psi)\frac{dl}{|\nabla\psi|} = -\int\phi_p\,\mathcal{J}\,d\theta = 0. \tag{2.21}$$

This gives us the form of the solution to ϕ_p given here to order α^2:

$$\phi_p(\psi,\theta) = \sum_{m=0}^{\infty} D_m(\psi)\cos(m\theta) = D_0(\psi)\cos\theta \tag{2.22}$$

$$-\sum_m\left(\alpha D_m^{\alpha}(\psi) - \alpha^2 D_m^{\alpha\alpha}(\psi)\right)\cos(m\theta) - \alpha^2\psi D_m^{\alpha}(\psi),$$

where the α orderings are shown explicitly, so D_m^{α} is the first-order piece of D_m, etc.

We have not yet included the feedback currents in the vacuum region: therefore the vacuum perturbed flux is described by $\nabla^2\phi_v = 0$, and the solution for ϕ_v is

$$\phi_v(\psi,\theta) = \sum_{m=1}^{\infty} B_m r^{-m}\cos(m\theta) = B_1 r^{-1}\cos\theta - \sum_m(\alpha B_m^{\alpha} - \alpha^2 B_m^{\alpha\alpha})r^{-m}\cos(m\theta). \tag{2.23}$$

The boundary condition imposed at the plasma-vacuum interface is the continuity of the perturbed flux ϕ. This gives a relation between the plasma and vacuum coefficients D_m and B_m, where $D_m \equiv D_m(\psi = 1)$. The boundary conditions given order by order are

$$B_1 = D_1 \tag{2.24}$$

$$B_m^\alpha - \frac{1}{4}B_1\delta_{m,3} - \frac{1}{4}B_1\delta_{m,5} = D_m^\alpha \tag{2.25}$$

where $\delta_{m,m'}$ is the Kronecker delta, and

$$B_1^{\alpha\alpha} - \frac{5}{4}B_5^\alpha - \frac{3}{4}B_3^\alpha - \frac{5}{16}B_1 = D_1^{\alpha\alpha}. \tag{2.26}$$

(It turns out that we need only the $m = 1$ term for the boundary condition at order α^2.)

The expressions for ϕ_p and ϕ_v are substituted into Eq. (2.16), and the appropriate boundary conditions are applied. Furthermore, δW_p is minimized with respect to the $D_m(\psi)$ by applying Euler's equation at every order. The integrations over θ (and also ψ in the case of δW_p) are performed leaving,

$$\frac{2}{\pi}\delta W_v = -D_1^2 - \alpha 4 D_1 D_1^\alpha - \alpha^2\left[-6D_1^2 - 4D_1 D_1^{\alpha\alpha} - 4D_1 D_5^\alpha - 4D_1 D_3^\alpha - 2\sum_{m=1}(D_m^\alpha)^2\right] \tag{2.27}$$

$$\frac{2}{\pi}\delta W_p = D_1^2 - \alpha 2 D_1 D_1^\alpha - \alpha^2\left[\frac{3}{8}D_1^2 - 2D_1 D_1^{\alpha\alpha} - 3D_1 D_3^\alpha - \sum_{m=1} m(D_m^\alpha)^2\right] \tag{2.28}$$

and the minimized δW_p:

$$\frac{2}{\pi}\delta W_p = D_1^2 - \alpha 2 D_1 D_1^\alpha - \alpha^2\left[\frac{3}{8}D_1^2 - 2D_1 D_1^{\alpha\alpha} - 3D_1 D_3^\alpha - \sum_{m=1} m(D_m^\alpha)^2\right]. \tag{2.29}$$

Therefore the total δW is given by

$$\frac{2}{\pi}\delta W = \alpha^2\left[-\frac{21}{4}D_1^2 - 4D_1 D_3^\alpha - 4D_1 D_5^\alpha - 2\sum_{m=1}(m-1)(D_m^\alpha)^2\right]. \tag{2.30}$$

This is the result of Rosen [16] expressed in our notation. Note that δW is nonzero only at order α^2.

Through minimizing δW_p with respect to the various D_m the solution is found for the plasma eigenmodes:

$$D_1(\psi) = D_1\psi^{\frac{1}{2}} \tag{2.31}$$

$$D_5^\alpha(\psi) = (D_5^\alpha - \frac{1}{4}D_1)\psi^{\frac{5}{2}} - \frac{1}{4}D_1\psi^{\frac{3}{2}} \tag{2.32}$$

$$D_m^\alpha(\psi) = D_m^\alpha\psi^{\frac{m}{2}} \qquad m \neq 5 \tag{2.33}$$

This gives us the full solution for $\varphi_p(\psi,\theta)$, and from Eq. (2.12) we can solve for the displacement eigenfunction:

$$\xi_r^\alpha - 2\alpha r\xi_r^\alpha = -\frac{1}{2}D_1\cos\theta - \alpha\left[D_1^\alpha r\cos\theta + (D_3^\alpha - \frac{7}{4}D_1)r^3\cos 3\theta\right. \qquad (2.34)$$
$$\left. -(D_5^\alpha - \frac{1}{4}D_1)r^5\cos 5\theta - \sum_{m\neq 1,3,5}^{\infty} D_m^\alpha r^m\cos(m\theta)\right].$$

$$\xi_\theta^\alpha - 2\alpha r\xi_\theta^\alpha = -\frac{1}{2}D_1\sin\theta - \alpha\left[D_1^\alpha r\sin\theta - (D_3^\alpha - \frac{7}{4}D_1)r^3\sin 3\theta\right. \qquad (2.35)$$
$$\left. -(D_5^\alpha - \frac{1}{4}D_1)r^5\sin 5\theta - \sum_{m\neq 1,3,5}^{\infty} D_m^\alpha r^m\sin m\theta\right].$$

To zero order the eigenfunction is a rigid, horizontal shift. It is therefore appropriate to require the first-order rigid component to vanish ($D_m^\alpha = 0$) in the displacement eigenfunction, since it contributes nothing to the problem. (Note that it does not add anything to the $\delta W'$ energy in Eq. (2.30).) Therefore the first-order contributions to the eigenfunction are non-rigid. To test the stability of the square plasma with respect to the pure rigid shift, the first-order contribution must vanish. The components are therefore $D_3^\alpha = 7/4$, $D_5^\alpha = -1/4$, $D_m^\alpha = 0$ for $m \neq 3, 5$. When this solution is substituted into the expression for $\delta W'$ we get

$$\delta W' = \alpha^2\frac{\pi}{2}\frac{27}{2}D_1^2. \qquad (2.36)$$

The square plasma is stable to a pure rigid shift.

We minimize $\delta W'$, Eq. (2.30), with respect to the D_m^α in order to find the true $\delta W'$ and corresponding eigenfunction. The minimization yields $D_3^\alpha = -\frac{1}{2}D_1$, $D_5^\alpha = -\frac{1}{4}D_1$, and $D_m^\alpha = 0$ for $m \neq 3,5$. The first-order minimizing eigenfunction is

$$\xi_r^\alpha = \frac{9}{8}r^2D_1\cos 3\theta \quad , \quad \xi_\theta^\alpha = -\frac{9}{8}r^2D_1\sin 3\theta. \qquad (2.37)$$

which corresponds to an $m = 3$ "wrinkle" superimposed on the rigid shift [16]. With this minimizing eigenmode the $\delta W'$ becomes

$$\delta W' = -\alpha^2\frac{\pi}{2}\frac{27}{4}D_1^2. \qquad (2.38)$$

The minimizing perturbation of a rigid shift plus a lower-order $m = 3$ piece superimposed is clearly unstable. The additional non-rigid component makes the plasma unstable to the horizontal shift.

2.2.2 Feedback stabilization of the square plasma model

We will now consider the addition of a feedback system by modifying the Rosen analysis to include extra terms in the definition of the vacuum flux corresponding to feedback coils that respond to perturbed flux measurements. In principle, however, an active feedback system is a very non-ideal addition to this otherwise ideal plasma-vacuum system. It is not obvious that a feedback system would satisfy the necessary condition of self-adjointness [55] required to make the δW approach valid. In fact, any realistic active feedback system will not satisfy this condition, but a model feedback system under certain symmetry constraints will be self-adjoint. The necessary constraints are derived in Appendix A along with the boundary conditions required to define the feedback model.

It is interesting to note that Rebhan and Salat [42] used a δW analysis to analyze active feedback stabilization of tokamak plasmas. In particular, they studied the effect of coil position on the ability to stabilize the axisymmetric mode. They use a δW analysis that corresponds to Eq. (A.6) and satisfies the symmetry constraints of Eq. (A.7). However, they did not demonstrate that their formulation satisfied the necessary self-adjointness requirements. The choice of the formulation used, while being the most straightforward for their analysis, is not a realistic choice for an actual device, since it requires that the corresponding flux loop be located at the *same* position as the active feedback coil. This is an unrealistic configuration. It would therefore be difficult to use this δW method to examine active feedback stabilization for anything more complex than this kind of simplified model feedback system.

The procedure for evaluating δW follows the description given above for the square plasma model but with the perturbed vacuum flux modified by the presence of current-carrying feedback coils. Figure 2.6 shows a schematic of the plasma and vacuum regions for the analytic calculation. The perturbed vacuum flux is defined in a fashion similar to that in the previous section:

In region I

$$\phi_v^I(r,\theta) = \sum_{m=1}^{\infty} B_m^I r^{-m} \cos m\theta - \sum_{m=1}^{\infty} L_m r^m \cos m\theta, \tag{2.39}$$

and in region II

$$\phi_v^{II}(r,\theta) = \sum_{m=1}^{\infty} B_m^{II} r^{-m} \cos m\theta. \tag{2.40}$$

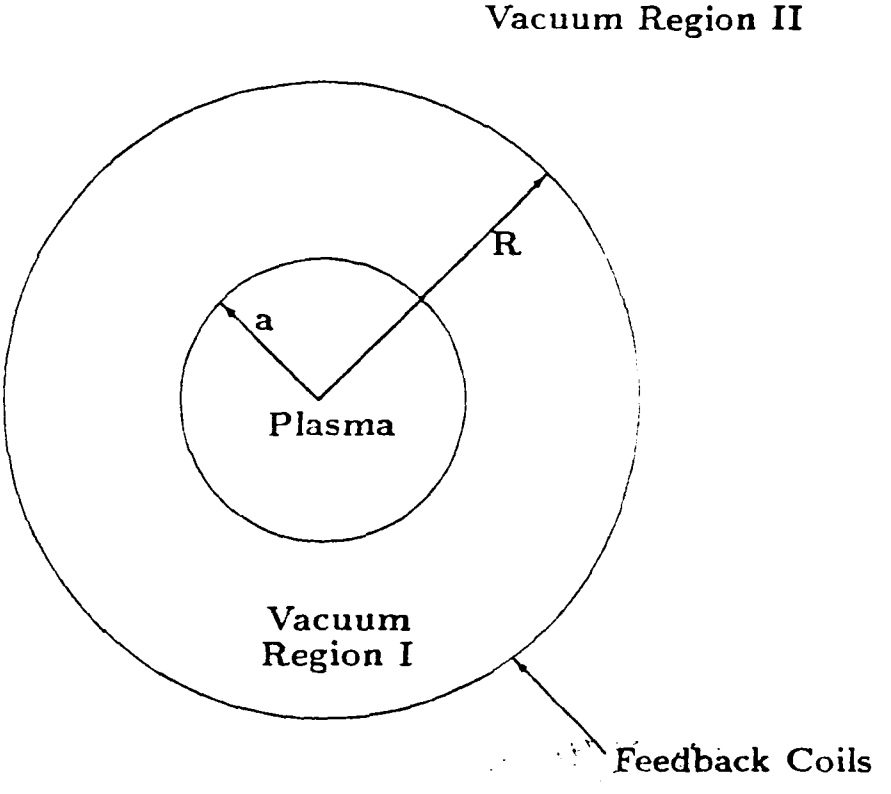


Figure 2.6: *Plasma and vacuum regions for analytic calculation. The feedback system consists of an "equivalent feedback system" made up of a multipolar decomposition representing m identical coils equally spaced on the circular contour of radius R. The two separate vacuum regions are shown. The appropriate boundary conditions must be applied at the interface between Region I and Region II as well as on the plasma surface.*

In this case, however, we have an extra sum representing the effect of the feedback coils in the region within the feedback coils. The two vacuum regions are separated by a circular contour, of radius R, upon which the feedback coils lie. The second term in Eq. (2.39) is the multipolar decomposition of the flux from these coils. Each of the terms in the sum is interpreted as the flux from an "equivalent feedback system" made up of m identical coils, equally spaced on the circular contour. The $m = 0$ terms in Eqs. (2.39) and (2.40) are absent, since they have zero gradient. In the plasma, the perturbed flux is again given the form

$$\psi_p(\psi, \theta) = \sum_{m=0}^{\infty} D_m(\psi) \cos m\theta. \tag{2.41}$$

We now repeat the δW analysis done in Sec. 2.2.1 including active feedback terms in the perturbed vacuum flux ψ_v. The δW terms for the plasma and plasma-current contributions remain unchanged and are still defined by Eqs. (2.29) and (2.27), respectively. The calculation for δW_v must be repeated, however, using the new form for the vacuum flux, Eq. (2.39), with the feedback terms included. The new δW_v is given by

$$\frac{2}{\pi}\delta W_v = D_1^2 - \alpha\left[2D_1 D_3^\alpha - 2D_1 L_1^\alpha\right] - \alpha^2\left[\frac{3}{8}D_1^2 - 2D_1 D_3^{\alpha\alpha} - 2D_1 L_1^{\alpha\alpha}\right. \tag{2.42}$$
$$\left. - 3D_1 D_3^\alpha - \frac{5}{2}D_1 L_5^\alpha - \frac{3}{2}D_1 L_3^\alpha - \sum_{m=1} m(D_m^\alpha)^2 - 2\sum_{m=1}^{\infty} m D_m^\alpha L_m^\alpha\right].$$

Therefore the new total δW with the feedback terms included becomes

$$\frac{2\delta W}{\pi} = \alpha\left[-2D_1 L_1^\alpha\right] - \alpha^2\left[-\frac{21}{4}D_1^2 - 4D_1 D_3^\alpha - 4D_1 D_5^\alpha\right. \tag{2.43}$$
$$\left. - 2\sum_{m=1}^{\infty}(m-1)(D_m^\alpha)^2 - 2D_1 L_1^{\alpha\alpha} - \frac{3}{2}D_1 L_3^\alpha - \frac{5}{2}D_1 L_5^\alpha - 2\sum_{m=1}^{\infty} m D_m^\alpha L_m^\alpha\right].$$

So far we have not specified any details of the feedback system. By analogy with the numerical experiment described in Section 2.1 and also corresponding to the generic form of the feedback systems defined in Chapter 3, we let each feedback system respond to the plasma motion by generating coil currents in proportion to some linear combination of perturbed flux. Schematically,

$$L_m = G_{mn}\Delta\psi_v(D_n). \tag{2.44}$$

The G_{mn} element of this gain matrix relates the m^{th} term of the multipolar decomposition of the feedback-system response to the n^{th} harmonic of the perturbed

plasma flux. If the gain matrix G has the correct symmetry properties (derived in Appendix A) this form of feedback law can be shown to leave the stability operator self-adjoint, so the energy principle will still apply. The gain elements G_{mn} should be at least of order α so that the feedback system has no effect when α is zero. (A circular plasma is motionally stable.) Hence, $L_m^\alpha = G_{mn}^\alpha(\Delta\phi_v^0)$. Equation (2.39) for ϕ_v shows that $\Delta\phi_v^0$ is proportional to D_1, with a constant of proportionality that depends on the location of the observation points. If we absorb the constant of proportionality into the G-symbol for the gain, we have

$$L_1^\alpha = G_{11}^\alpha D_1. \tag{2.45}$$

This feedback gain term relates the $m = 1$ component response of the feedback system to the rigid, lowest-order component of the displacement eigenfunction. The first-order contribution to the energy becomes

$$\frac{2\delta W^{(\alpha)}}{\pi} = -2G_{11}^\alpha (D_1)^2. \tag{2.46}$$

Recall that in the case with no feedback the $\delta W^{(\alpha)}$ was zero. The plasma was neutrally stable to order α. By adding active feedback, the stability to order α is now dependent only on the factor G_{11}^α. It can be made very stable, or very unstable, depending on the sign and magnitude of G_{11}^α.

Suppose now that $L_1^\alpha = 0$. Then $\delta W^{(\alpha)}$ vanishes, and stability is determined by the second-order terms in δW. To simplify the analysis of $\delta W^{(\alpha\alpha)}$, we choose a simplified feedback model motivated by the form of the eigenfunction without feedback. Recall the minimized energy eigenfunction for the model without feedback: $D_3^\alpha = -\frac{1}{2}D_1$, $D_5^\alpha = -\frac{1}{4}D_1$, and $D_m^\alpha = 0$ for $m \neq 3,5$. This gives a displacement eigenfunction corresponding to a first-order (in α) $m = 3$ wrinkle superimposed on a zero-order rigid shift.

We are therefore interested in a feedback system that can respond to $m = 1$ and $m = 3$ perturbations with $m = 1, 3$ reaction from the multipolar feedback system. Thus the feedback currents defined by Eq. (A.5) are given by

$$\begin{bmatrix} J_0^1 \\ \\ J_0^3 \end{bmatrix} = \begin{bmatrix} g_{11} & g_{13} \\ \\ g_{31} & g_{33} \end{bmatrix} \cdot \begin{bmatrix} \phi_v^1(R) \\ \\ \phi_v^3(R) \end{bmatrix} \tag{2.47}$$

We found in the previous section that we must have $g_{13} = g_{31}$ in order to satisfy the constraints on the δW formalism.

From Eqs. (A.16) and (A.17) we can define the feedback coefficients L_m in terms of the Region I perturbed flux coefficients B_n^I. These, in turn, are related to the plasma perturbation coefficients D_n through the boundary condition at the plasma-vacuum interface. Therefore we find

$$\begin{bmatrix} L_1 \\ \\ L_3 \end{bmatrix} = \begin{bmatrix} \alpha^2 G_{11} & \alpha G_{13} \\ \\ \frac{1}{3}\alpha G_{13} & \alpha^2 G_{33} \end{bmatrix} \cdot \begin{bmatrix} D_1 \\ \\ D_3 \end{bmatrix} \tag{2.48}$$

We keep the terms through order α^2 and find that the feedback coefficients are defined as

$$L_1^{\alpha\alpha} = G_{11}^{\alpha\alpha} D_1 - G_{13}^{\alpha} D_3^{\alpha}, \tag{2.49}$$

$$L_3^{\alpha} = G_{31}^{\alpha} D_1 = \frac{1}{3} G_{13}^{\alpha} D_1.$$

$$L_m^{\alpha} = 0 \qquad m > 3.$$

This corresponds to coils with an ability to respond (on account of the flux measurements) to both the rigid shift and the $m = 3$ perturbations.

We substitute these feedback terms into the expression for δW, Eq. (2.44). For a trial displacement we use the eigenfunction found for the system without feedback, given in Eq. (2.37). The calculated δW with the feedback terms included is

$$\frac{2\delta W}{\pi} = \alpha^2 D_1^2 \left[-\frac{27}{4} - 2G_{11}^{\alpha\alpha} - \frac{5}{2} G_{13}^{\alpha} \right]. \tag{2.50}$$

Without feedback ($G_{11}^{\alpha\alpha} = G_{13}^{\alpha} = 0$), the plasma is seen to be unstable. With feedback, δW can be made positive for a range of values of the gain coefficients. Specifically, for any choice of G_{13}^{α}, a $G_{11}^{\alpha\alpha}$ can be found that is stabilizing. The converse is also true. This can be seen in Fig. 2.7, which gives the stability boundary (long-dashed line) for this trial displacement in $(G_{11}^{\alpha\alpha}, G_{13}^{\alpha})$ parameter space.

Now let us consider the effect of the feedback system on the eigenfunction. We take our expression for δW, Eq. (2.44), with the feedback terms included, and minimize with respect to the D_m^{α}. This yields $D_3^{\alpha} = -\frac{1}{2} - \frac{1}{2} G_{13}^{\alpha} D_1$, $D_5^{\alpha} = -\frac{1}{4} D_1$, and $D_m^{\alpha} = 0$ for $m \neq 3, 5$.

The minimized δW is

$$\frac{2\delta W}{\pi} = \alpha^2 D_1^2 \left[-\frac{27}{4} - 2G_{11}^{\alpha\alpha} - \frac{5}{2} G_{13}^{\alpha} - (G_{13}^{\alpha})^2 \right], \tag{2.51}$$

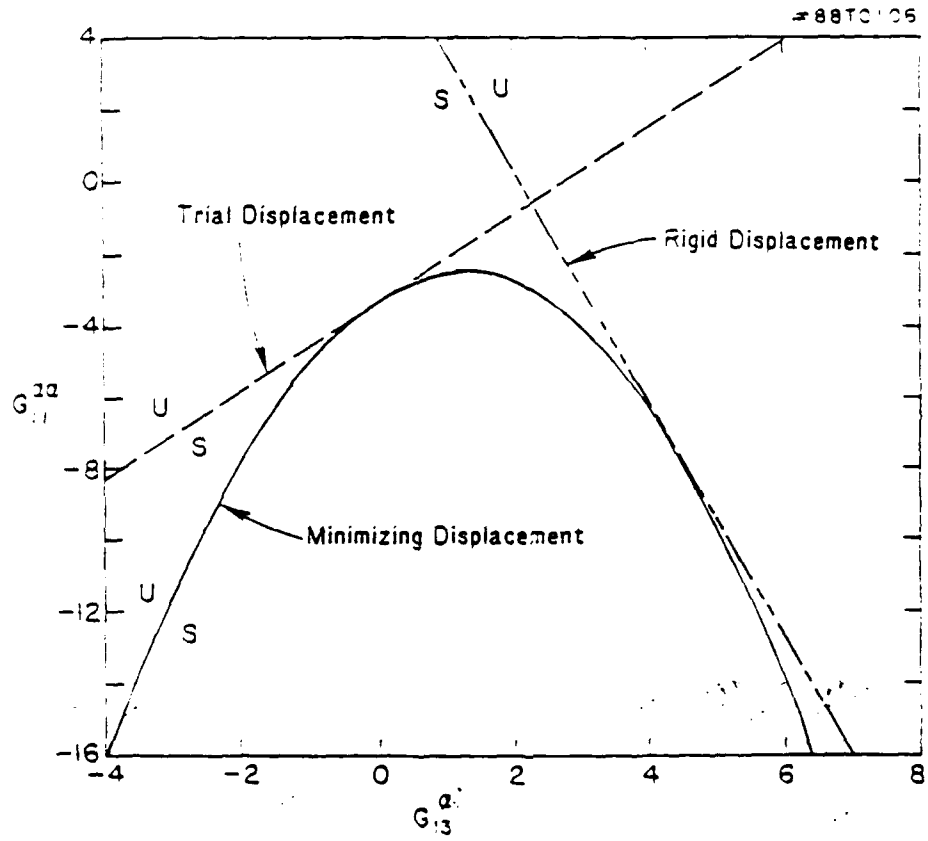


Figure 2.7: Stable and unstable regions for a range of values of the gains $G_{11}^{\alpha\alpha}$ and G_{13}^{α}. The long-dashed line represents the stability boundary for the trial displacement given in Eq. (2.37). The short-dashed line represents the stability boundary for the rigid displacement. The solid curve denotes the stability boundary for the minimized eigenfunction that is deformed by the feedback system. The regions in the $(G_{11}^{\alpha\alpha}, G_{13}^{\alpha})$ parameter space where the plasma is stable or unstable are marked on either side of the stability boundary by S and U, respectively.

with the corresponding (first-order) eigenfunction

$$\xi_r^\alpha = D_1 r^2 \left(\frac{9}{8} - \frac{G_{13}^\alpha}{4} \right) \cos 3\theta \,, \quad \xi_\theta^\alpha = -D_1 r^2 \left(\frac{9}{8} - \frac{G_{13}^\alpha}{4} \right) \sin 3\theta. \tag{2.52}$$

The feedback system clearly modifies the form of the minimized-energy eigenfunction. In particular, it is the feedback component G_{13} that couples the rigid motion/response to the non-rigid response/motion, and that modifies the eigenfunction.

The form of δW is no longer a simple linear relationship between $G_{11}^{\alpha\alpha}$ and G_{13}^α. The last two terms in δW are seen to be stabilizing for $0 < G_{13}^\alpha < 5/2$, and optimally stabilizing for $G_{13}^\alpha = 5/4$. For this optimal value, the mode is stabilized if $G_{11}^{\alpha\alpha} < 83/32$. This is summarized in Fig. 2.7, which shows the stable and unstable boundaries for this *minimized* δW in the $(G_{11}^{\alpha\alpha}, G_{13}^\alpha)$ parameter space. It is seen that for any choice of G_{13}^α, a $G_{11}^{\alpha\alpha}$ can be found that stabilizes the mode. However, the converse is not true, and the choice of $G_{11}^{\alpha\alpha}$ is critical. Since the actual values of the G-gains depend on the location of the observation points, we see that this result translates into a criticality for the placement of the flux pickup loops.

It is also interesting to consider the conclusions that result from restricting the instability to take the form of a rigid displacement. These indicate that the above behavior is due to the ability of the feedback systems to distort the eigenfunction. For a rigid displacement, we must choose (see Eq. (2.34)) $D_3^\alpha = \frac{5}{4} D_1$, $D_5^\alpha = -\frac{1}{4} D_1$, and $D_m^\alpha = 0$ for $m = 3, 5$. Then the expression for δW becomes

$$\frac{2\delta W}{\pi} = \alpha^2 D_1^2 \left[\frac{27}{2} - 2G_{11}^{\alpha\alpha} - \frac{13}{2} G_{13}^\alpha \right]. \tag{2.53}$$

The stable and unstable regions for this rigid instability are also shown in Fig. 2.7. Because of the linearity of δW on the Gs, it follows that a $G_{11}^{\alpha\alpha}$ can be found to stabilize the mode for any value of G_{13}^α. The converse is also true. Therefore, for a rigid instability, the detailed placement of flux pickup loops for monitoring the motion is not critical. Note that the stable/unstable boundary for the rigid instability encloses the stable region for the deformable instability except for a common point at $G_{13}^\alpha = 9/2$. Equation (2.52) shows that for this value of the gain, G_{13}^α, the eigenfunction no longer supports an $m = 3$ deformation, so the eigenfunctions for the rigid and deformable instability coincide.

The stable/unstable boundary for the trial displacement, given by Eq. (2.37), also encloses the stable region for the deformable instability. There is a single point, at

$G_{13}^{\alpha} = 0$, at which these two stability boundaries coincide. At this point the feedback system is not affecting the instability eigenfunction and therefore the form of the eigenfunction and the energy is the same as in the case of the trial displacement with no feedback. It should be noted from Fig. 2.7 that the stability region enclosed by the minimizing eigenfunction encloses a smaller area in gain space than the common stability regions for the trial and rigid displacements. Particularly at large magnitudes of the feedback gain coefficients, the stability boundaries for the trial and rigid displacement eigenfunctions lie far within the unstable region for the minimizing eigenfunction. It is clear that the minimizing eigenfunction, Eq. (2.52) is affected by the feedback, and this results in a plasma that can be stabilized only by feedback in which the gain $G_{11}^{\alpha\alpha}$ lies within a restricted region. Thus we see that the feedback system with certain placements of the observation points can allow the unstable eigenfunction to deform so that the plasma will remain unstable.

2.3 Conclusions

It is clear from the analysis of both the numerical simulations and the analytic model that there is a significant interaction between the feedback system and the plasma instability. From the numerical simulations it is seen that in the PBX-M tokamak, certain placements of the magnetic pickup coils, on the inboard side, lead to an unstable system, regardless of the gain, whereas other placements, on the outboard side, will give a stable system for sufficiently large values of the gain. A simplified analytical model suggests that this behavior results from the nonrigid deformable nature of the plasma cross section and the plasma's ability to modify its unstable eigenfunction according to the particular feedback system. We therefore see that the feedback system can change and *enhance* the non-rigid character of the unstable eigenfunction. Furthermore, it is seen that the feedback system can, in certain cases, modify the non-rigid components of the instability in such a way that the plasma remains unstable regardless of the gain. This is not the case if the instability takes a rigid form, or even if it takes its original form in the absence of feedback.

We can interpret this phenomenon qualitatively in terms of a simple physical picture. As part of the plasma instability, a perturbed magnetic field is produced in the vacuum region, and this is sensed by the magnetic pickup coils. If the plasma is unstable enough, it can modify its eigenfunction to deform its cross section so

that a null in the perturbed vacuum magnetic field will appear at the position of the observation loops. Since these loops will then be unable to detect the plasma instability, the feedback system will be rendered inoperative.

It is therefore clear that stability methods that rely on rigid plasma models, current filament models, or simplified trial function analysis are insufficient to examine the problem of feedback stabilization of highly shaped tokamak plasmas with complicated active feedback systems. What is called for is a way to calculate the true eigenfunction of the unstable mode in the presence of an arbitrary active feedback system and additional passive resistive conductors. The analysis must allow one to see how the feedback system interacts with the plasma instability—in particular, the enhancement of the non-rigid components of the eigenfunction.

We have, therefore, developed the NOVA-W code. This is a linear MHD stability code that includes realistic active feedback and arbitrary configurations of resistive conductors in the region surrounding the plasma. NOVA-W solves for the *true* unstable eigenfunction of the linear MHD equations. In the next chapter we present the formulation of the vacuum calculation, which includes the feedback, and the connection of the vacuum to the plasma stability calculation.

Chapter 3

The NOVA-W Formulation

In this chapter we present the basic formulation for the NOVA-W code. NOVA is an ideal MHD stability code that is nonvariational, i.e., it does not use the δW formulation. For this reason, it can be extended to include nonideal effects such as resistive conductors and active feedback systems. NOVA-W is the modification of the original NOVA code [51] of Cheng and Chance that includes these new effects.

The vacuum calculation of NOVA has been modified to include resistive conductors and an active feedback system in the vacuum region outside the plasma. In order to accommodate these changes, the vacuum calculation was modified by basing it on a perturbed poloidal flux representation instead of a scalar potential representation for the perturbed magnetic field in the vacuum. A perturbed poloidal flux representation is the natural method to choose, since it allows perturbed toroidal currents to exist in the vacuum region, and also permits a simple "thin-wall" approximation to model the effects of a resistive wall. Furthermore, perturbed flux measurements are used as input to the feedback control, as in the numerical model of Section 2.1.

We begin by introducing the ideal MHD stability equations for the plasma and the formulation and techniques used by the NOVA code to solve them, thereby describing the plasma calculation. Next we consider the details of the vacuum calculation. It is the goal of the vacuum calculation to provide the necessary boundary condition at the plasma-vacuum interface so that the plasma stability eigenvalue equations can be solved with the feedback effects included. We describe the perturbed flux formulation, the boundary conditions that connect the vacuum calculation to the plasma eigenvalue calculation, and the method used to evaluate the surface integrals.

Then we present the method used to deal with the resistive walls in the vacuum region. This "thin-wall approximation" leads to a jump condition in the normal derivative of the perturbed flux between the vacuum region outside the resistive wall and the region between the wall and the plasma. An active feedback system with feedback currents controlled by flux-loop measurements in the vacuum region is then included in the calculation, and the necessary feedback matrices are derived. Finally, we add the necessary circuit equations needed to describe a realistic feedback system and additional discrete conductors that may be present in the vacuum region.

3.1 The NOVA Code

The NOVA stability code solves the linear ideal MHD stability eigenvalue equations

$$\rho\omega^2\xi = \nabla p_1 - \mathbf{b} \times (\nabla \times \mathbf{B}) - \mathbf{B} \times (\nabla \times \mathbf{b}) \tag{3.1}$$

$$p_1 - \xi \cdot \nabla P - \gamma P \nabla \cdot \xi = 0, \tag{3.2}$$

where

$$\mathbf{b} = \nabla \times (\xi \times \mathbf{B}) \tag{3.3}$$

is the perturbed magnetic field in the plasma. $\mathbf{B}$ is the equilibrium magnetic field, p_1 and P are the perturbed and equilibrium particle pressures, respectively, ρ is the plasma mass density, $\gamma = \frac{5}{3}$ is the ratio of specific heats, ξ is the displacement vector, and ω is the eigenvalue (normalized growth rate). The equilibrium magnetic field is represented by

$$\mathbf{B} = \nabla\zeta \times \nabla\psi - q(\psi)\nabla\psi \times \nabla\Theta \tag{3.4}$$

or

$$\mathbf{B} = \nabla\phi \times \nabla\psi - \mathbf{g}(\psi)\nabla\phi. \tag{3.5}$$

where $2\pi\psi$ is the poloidal flux contained within a surface, Θ is the generalized poloidal angle, ζ is the generalized toroidal angle, ϕ is the standard toroidal angle from (X, ϕ, Z) cylindrical coordinates, $q(\psi)$ is the safety factor, and $\mathbf{g}(\psi)$ is the toroidal field function. The second definition for $\mathbf{B}$, Eq. (3.5), follows for an axisymmetric equilibrium. The generalized angle coordinates (Θ, ζ) are chosen to make the magnetic field lines appear straight in this coordinate system. This representation greatly

simplifies the form of the operator:

$$\mathbf{B}\cdot\nabla = \mathcal{J}^{-1}\left(\frac{\partial}{\partial\Theta} - q\frac{\partial}{\partial\zeta}\right), \tag{3.6}$$

where the Jacobian $\mathcal{J}$ is defined by

$$\mathcal{J}^{-1} = \nabla w \times \nabla\Theta\cdot\nabla\zeta. \tag{3.7}$$

There are several straight-field-line coordinate systems that can be used. This category includes the PEST coordinates, equal-arc-length coordinates, Hamada coordinates, and Boozer magnetic coordinates. All use the poloidal flux function w as the radial variable, since it is natural to separate the motion across flux surfaces from that along the surfaces, given the large difference in the time scales of the motion. They differ in their definitions for poloidal angle Θ and toroidal angle ζ.

The PEST coordinates [56] have the advantage that the generalized toroidal variable ζ is equal to the standard toroidal variable ϕ. Therefore $\nabla\phi\cdot\nabla v = \nabla\phi\cdot\nabla\Theta = 0$ and $\nabla\phi^2 = 1/R^2$, which simplifies the metric. The Jacobian of this coordinate system is proportional to R^2. The Θ points are defined such that equal-area increments of the cross section on a given flux surface, $\mathcal{J}dwd\Theta$, enclose equal amounts of toroidal flux. This does not lead to a very evenly distributed Θ-grid, however. In fact, the points in the Θ-grid are heavily weighted toward the inboard side of the cross section, especially in low-aspect-ratio equilibria.

The equal-arc-length coordinates, as the name suggests, divide up equally incremented points in Θ on a flux surface so that they are equally separated in arc length. This has the advantage of evenly distributing points in the Θ-grid over each flux surface. With this definition of the poloidal angle Θ, the toroidal coordinate ζ can no longer be equal to ϕ if the field lines are to appear straight.

Several other straight-field-line coordinate systems have other useful properties. In the Hamada [57] coordinates, the Jacobian is equal to unity. The straight-field-line coordinate system [58] with the Jacobian proportional to B^{-2} is useful for its Hamiltonian representation of a three-dimensional equilibrium magnetic field. It was introduced by Boozer as a convenient coordinate system with which to evaluate guiding center drift equations.

A complete description of the derivation of magnetic coordinates can be found in the book on tokamak physics by White [59]. Also, a discussion of the importance of

using a straight-field-line coordinate system in the MHD stability problem as well as a description of the calculation used here that maps an arbitrary equilibrium into the chosen coordinate system is presented in the Ph.D. thesis of Harley [60].

The straight-field-line coordinate system that best represents our problem is the equal-arc coordinate system. The points describing the conducting wall in the vacuum region are also defined in equal-arc-length fashion. In the equal-arc coordinate system the Jacobian is defined by

$$\mathcal{J}(X, Z) = \frac{X}{\alpha(\psi)|\nabla\psi|}, \tag{3.8}$$

where $\alpha(\psi)$ is a flux surface function given by the necessity that Θ increase by 2π during one poloidal circumference.[51] Thus

$$\alpha(\psi) = 2\pi / \oint \frac{\mathcal{J}|\nabla\psi}{X} d\Theta. \tag{3.9}$$

The NOVA formulation decomposes the displacement vector as

$$\xi = \frac{\xi_\psi}{|\nabla\psi|^2}\nabla\psi - \frac{\xi_s}{B^2}(\mathbf{B}\times\nabla\psi) - \frac{\xi_b}{B^2}\mathbf{B}. \tag{3.10}$$

and the perturbed magnetic field is decomposed as

$$\mathbf{b} = \frac{Q_\psi}{|\nabla\psi|^2}\nabla\psi - \frac{Q_s}{|\nabla\psi|^2}(\mathbf{B}\times\nabla\psi) - \frac{Q_b}{B^2}\mathbf{B}. \tag{3.11}$$

The momentum equation (3.1) can be divided into three component equations (see Appendix B) by taking projections along the basis vector directions of ξ and $\mathbf{b}$. The b components, Q_ψ, Q_s, and Q_b, can be eliminated from these three equations in terms of the ξ components, ξ_ψ, ξ_s, and ξ_b, by using the definition of b, Eq. (3.3). Also, the third component momentum equation (in ξ_b) can be eliminated owing to the relation of ξ_b to p_1 and ξ_ψ. This reduces the linear ideal MHD eigenvalue equations (3.1)–(3.2) to four component equations which can be written in matrix form:

$$\nabla\psi\cdot\nabla\begin{pmatrix} P_1 \\ \xi_\psi \end{pmatrix} = \mathbf{C}\begin{pmatrix} P_1 \\ \xi_\psi \end{pmatrix} - \mathbf{D}\begin{pmatrix} \xi_s \\ \nabla\cdot\xi \end{pmatrix} \tag{3.12}$$

and

$$\mathbf{E}\begin{pmatrix} \xi_s \\ \nabla\cdot\xi \end{pmatrix} = \mathbf{F}\begin{pmatrix} P_1 \\ \xi_\psi \end{pmatrix}, \tag{3.13}$$

where P_1 is the total perturbed pressure $P_1 = p_1 - \mathbf{b}\cdot\mathbf{B}$ and **C**, **D**, **E**, and **F** are 2×2 matrix operators involving surface derivatives $\mathbf{B}\cdot\nabla$ and $(\mathbf{B}\times\nabla\psi)\cdot\nabla$. These matrices

are fully defined in Appendix B. A more detailed description of the derivation of the above matrix equations can be found in the original NOVA paper [51].

The Galerkin method is used to solve the eigenvalue equations. The perturbed quantities are represented in terms of a finite Fourier series

$$\xi_\psi(\psi,\Theta,\zeta) = \sum_{m=m_1}^{m_2} \xi_m(\psi)\exp i(m\Theta - n\zeta), \tag{3.14}$$

where we have a truncated series over the $M = m_2 - m_1 - 1$ poloidal harmonics and n is the toroidal mode number, which in the case of axisymmetric modes is equal to zero.

The components ξ_s and $\nabla\cdot\xi$ are eliminated in favor of P_1 and ξ_ψ by inverting the matrix $\mathbf{E}$ in Eq. (3.13). The matrices have at this point been decomposed into a truncated algebraic Fourier matrix representation. The equations are now over the Fourier components of the remaining perturbed quantities. This leaves a pair of equations written in the form of a single matrix differential equation in the Fourier components $\xi_{\psi m}$ and P_{1m}.

Finally, the P_{1m} terms are found in terms of ξ_m—provided there is a proper boundary condition relating P_{1m} to ξ_m at the plasma-vacuum boundary—and only a matrix differential equation in ξ_m remains. The radial dependence of the Fourier modes of the perturbation $\xi_m(\psi)$ is in turn represented in terms of cubic B-spline finite elements.

For details on the reduction of the ideal MHD stability equations (3.1)–(3.2) into a single equation in ξ_m, on the use of the cubic B-spline finite elements and their convergence properties, and on the eigenvalue determination methods of the NOVA code, the reader is referred to the original NOVA paper by Cheng and Chance [51].

The necessary boundary condition at the plasma surface for the final eigenvalue equation is given in the form of the value of the perturbed total pressure at the plasma edge, $P_1 = \mathbf{b}\cdot\mathbf{B}$, in terms of the radial displacement ξ_ψ at the plasma edge. This goes into Eq. (3.12) and gives the proper boundary condition when P_{1m} is eliminated in favor of ξ_m. This boundary condition contains all the information about the vacuum region and its effects on the plasma stability. This condition includes the influence of the geometry and the presence of resistive conductors and active feedback currents in the vacuum region surrounding the plasma. The details of this vacuum calculation are provided in the next section.

3.2 The Vacuum Calculation

3.2.1 The perturbed poloidal flux formulation

We are dealing with axisymmetric modes: thus we can use the axisymmetric flux formulation to represent the perturbed magnetic field in the vacuum

$$\mathbf{b} = \frac{1}{2\pi}\nabla\phi \times \nabla\chi - a_t \nabla\phi, \tag{3.15}$$

where χ is the perturbed poloidal flux in the vacuum and a_t is related to the perturbed toroidal field.

The perturbed currents in the vacuum region will be the currents induced in the axisymmetric coils or generated by the active feedback system in the active coils. Ampere's law therefore gives us

$$\nabla \times \mathbf{b} = \mu_0 j_\phi \hat{\phi}. \tag{3.16}$$

where $\phi = \nabla\phi$, $\nabla\phi = \hat{X}\nabla\phi$. From this, and given that $\nabla \cdot \mathbf{b} = 0$, we find that

$$\frac{\partial a_t}{\partial\varphi} = \frac{\partial a_t}{\partial\Theta} = 0. \tag{3.17}$$

Therefore a_t = constant in space.

The perturbed poloidal flux χ is related to the perturbed toroidal currents in the vacuum region:

$$X\nabla \cdot \left(\frac{1}{X^2}\nabla\chi\right) = 2\pi\mu_0 j_\phi, \tag{3.18}$$

where we use the standard (X, ϕ, Z) cylindrical coordinates and j_ϕ represents the perturbed toroidal currents from active feedback coils in the vacuum region and the eddy currents present in resistive passive conductors (although we will deal with the latter by using a jump condition for the normal derivative in the flux for the limit of a thin resistive wall). The current density in the active feedback coils is represented by

$$j_{A.F.} = \sum_{i=1}^{N} I_m \, \delta(\mathbf{r} - \mathbf{r}_i). \tag{3.19}$$

Here, the I_m are the perturbed currents in axisymmetric coils that are part of an active feedback system in the vacuum region located at $\mathbf{r}_i = (X_i, Z_i)$. The sum is over all the feedback coils.

We solve Eq. (3.18) using Green's function techniques. The corresponding Green's function $G(\mathbf{r};\mathbf{r}_S)$ is defined by

$$X_S\nabla\cdot\left(\frac{1}{X^2}\nabla G(X,Z;X_S,Z_S)\right) = 4\pi\delta(X-X_S)\delta(Z-Z_S). \tag{3.20}$$

Since it can be shown that

$$X\nabla\cdot\left(\frac{1}{X^2}\nabla G\right) = \sigma\cdot\nabla^2(G\nabla\sigma),$$

we can rewrite Eq. (3.20) as

$$\nabla^2 G(\mathbf{r};\mathbf{r}_S)\nabla\sigma = 4\pi\,\delta(\mathbf{r}-\mathbf{r}_S)\,\hat{\sigma}. \tag{3.21}$$

This is of the same form as the vector Poisson's equation for the magnetic vector potential where the current density is given by a single axisymmetric current loop at $\mathbf{r}_s = (X_s, Z_s)$ [61]. The solution is [61]

$$G(\mathbf{r};\mathbf{r}_S) = -\int\frac{\delta(\mathbf{r}_s-\mathbf{r}')}{\mathbf{r}-\mathbf{r}'}d^3\mathbf{r}'. \tag{3.22}$$

This integral can be cast into a form that defines the Legendre function. Therefore the Green's function is given [56] by

$$G(\mathbf{r}_T;\mathbf{r}_S) = G(X_T,Z_T;X_S,Z_S) = 4\pi\,\frac{X_TX_S}{r}\,P^1_{-1/2}(w). \tag{3.23}$$

where

$$w = \frac{X_S^2 - X_T^2 - (Z_S-Z_T)^2}{r^2}$$

$$r = [(X_S^2-X_T^2)^2 + (Z_S-Z_T)^4 - 2(X_S^2-X_T^2)(Z_S-Z_T)^2]^{1/4}$$

and where $P^1_{-1/2}$ is the associated Legendre function of order $-\frac{1}{2}$ (also known as a toroidal function or ring function) [62].

By using identities for Legendre functions and for the complete elliptic integrals, we can also write the Green's function as

$$G(\mathbf{r}_T;\mathbf{r}_S) = \frac{-4X_TX_S}{[(X_T-X_S)^2+(Z_T-Z_S)^2]^{1/2}}\left[\frac{(2-k^2)K(k^2)-2E(k^2)}{k^2}\right], \tag{3.24}$$

where

$$k^2 = \frac{4X_TX_S}{(X_T+X_S)^2+(Z_T-Z_S)^2}$$

and where K and E are the complete elliptic integrals of the first and second kind. When we evaluate $G(\mathbf{r};\mathbf{r}')$ the elliptic integrals are used [56] and are calculated using Hastings's formula [63].

By integrating Eqs. (3.18) and (3.20) over the volume of the vacuum region and by applying Green's theorem, we obtain a Green's equation that relates the perturbed vacuum flux on the boundary surface to the currents in the vacuum region and to the sum of the integrals over all the boundary surfaces:

$$\chi(\mathbf{r}) = \sum_{i=1}^{N} \mu_0 I_i G(\mathbf{r};\mathbf{r}_i) - \sum_S \frac{1}{2\pi} \oint_S \frac{dl_S}{X_S} [\chi \nabla_n G(\mathbf{r};\mathbf{r}_S) - G(\mathbf{r};\mathbf{r}_S) \nabla_n \chi]. \qquad (3.25)$$

The three boundary surfaces consist of the plasma-vacuum interface, the inside surface of the resistive wall, and the outside surface of the resistive wall. Here $\nabla_n = \hat{n} \cdot \nabla$, and the feedback currents are defined as linear combinations of the perturbed flux and the corresponding time-derivative terms at prescribed observation points. Perturbed magnetic field measurements (magnetic probe measurements) can also be included in the feedback law. As an example we can consider a simple feedback law in which the desired current for a given feedback coil is proportional to the difference in the perturbed flux (and its time derivative) at two observation points symmetric about the midplane (equivalent to the form used in Sec. 2.1). This flux difference serves as a measure of the vertical displacement of the plasma, and is very accurate in the case of rigid plasma motion. In this case the feedback currents are defined as

$$I_m = \alpha_m [\chi(X_{o1}, Z_{o1}) - \chi(X_{o2}, Z_{o2})] - \beta_m [\dot{\chi}(X_{o1}, Z_{o1}) - \dot{\chi}(X_{o2}, Z_{o2})]. \qquad (3.26)$$

The integrals in Eq. (3.25) are over the surfaces that are the boundaries to the vacuum region, i.e., the plasma-vacuum surface and the surface of the resistive wall surrounding the plasma. We take the principal part of the integral over the contour that is the interface or wall for a poloidal cross section of the torus. The incremental arc length dl on the contour is given by

$$dl = \frac{\mathcal{J} |\nabla \psi|}{X} d\Theta. \qquad (3.27)$$

Separate Green's equations are obtained for the vacuum region between the plasma and the resistive wall, and for the region outside the resistive wall. By discretizing the quantities on the contour into a finite grid, the contour integrals in Eq. (3.25) are expressed as sums over the grid-points on the contour. The collocation method [64] is

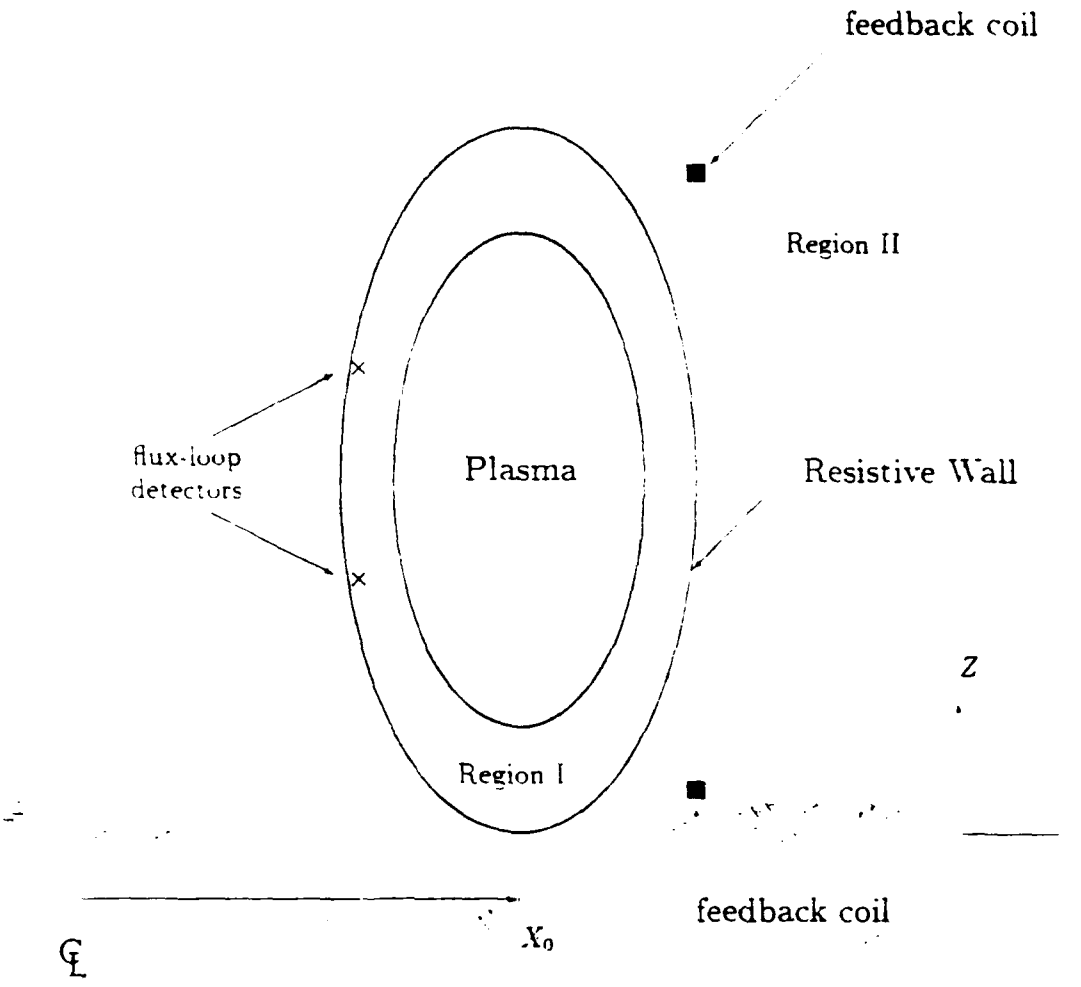


Figure 3.1: *Plasma and Resistive Wall, showing definitions of Regions I and II. The active feedback coils are outside the resistive wall in Region II. The flux-loop detectors are within the resistive wall in Region I (Case A for the feedback matrices derivation in Sec. 3.4). They can also be placed in Region II (Case B).*

used to solve the integral equations. Therefore Eq. (3.25) is written M_Θ times, where M_Θ is the number of theta grid points, with every grid point on the contour serving as the observation point (R_S, Z_S). This series of equations can be combined into a single matrix equation in which the integrals are expressed as matrices multiplying column vectors labeled as χ_p, $\nabla_n\chi_p$, χ_w, and $\nabla_n\chi_w$, which contain the values of the perturbed flux and its normal derivative at every grid point on the plasma surface and wall surface, respectively. A total of 3 matrix equations result, corresponding to the number of surfaces that serve as boundaries to the two regions. (Region I is the space between the plasma surface and the inner wall surface, and Region II is the space outside the resistive wall extending to infinity—see Fig. 3.1.) The Green's equations in matrix form for a configuration with feedback coils present in the region outside the resistive wall are as follows:

$$(\underline{1} - \mathbf{M}_{pp}) \cdot \chi_p - \mathbf{M}_{pw} \cdot \chi_w^- = \mathbf{G}_{pp} \cdot \nabla_n\chi_p - \mathbf{G}_{pw} \cdot \nabla_n\chi_w^- \tag{3.28}$$

$$\mathbf{M}_{wp} \cdot \chi_p - (\underline{1} - \mathbf{M}_{ww}) \cdot \chi_w^- = \mathbf{G}_{wp} \cdot \nabla_n\chi_p - \mathbf{G}_{ww} \cdot \nabla_n\chi_w^- \tag{3.29}$$

$$(\underline{1} - \mathbf{M}_{ww}^-) \cdot \chi_w^- = \mathbf{G}_{ww}^- \cdot \nabla_n\chi_w^- - \mathbf{P}_w \cdot \chi_w^- - \mathbf{N}_w \cdot \nabla_n\chi_w^- - \mathbf{P}_p \cdot \chi_p - \mathbf{N}_p \cdot \nabla_n\chi_p \tag{3.30}$$

where the matrices $\mathbf{M}$ and $\mathbf{G}$ are defined such that the i^{th} components of the matrix-vector products are given by

$$\mathbf{M}_{wp}^i \cdot \chi_p = -\frac{1}{2\pi}\oint_p \frac{dl_p}{X_p}(\hat{\mathbf{n}} \cdot \nabla G(\mathbf{r}_p; \mathbf{r}_w^i))\chi(\mathbf{r}_p) \tag{3.31}$$

$$\mathbf{G}_{pw}^i \cdot \nabla_n\chi_w = -\frac{1}{2\pi}\oint_w \frac{dl_w}{X_w} G(\mathbf{r}_w; \mathbf{r}_p^i)(\hat{\mathbf{n}} \cdot \nabla\chi(\mathbf{r}_w)) \tag{3.32}$$

$$\mathbf{M}_{pp}^i \cdot \chi_p = -\frac{1}{2\pi}\oint_p \frac{dl_p}{X_p}(\hat{\mathbf{n}} \cdot \nabla G(\mathbf{r}_p; \mathbf{r}_p^i))\chi(\mathbf{r}_p) \tag{3.33}$$

$$\mathbf{G}_{ww}^i \cdot \nabla_n\chi_w = -\frac{1}{2\pi}\oint_w \frac{dl_w}{X_w} G(\mathbf{r}_w; \mathbf{r}_w^i)(\hat{\mathbf{n}} \cdot \nabla\chi(\mathbf{r}_w)) \tag{3.34}$$

and so on, with the definitions for the other matrices following the subscripts. The identity matrix is expressed by $\underline{1}$, and the normal vector is defined by $\hat{\mathbf{n}} = \nabla\psi / \nabla\psi$. Also, $\mathbf{r}_p^i$ is the i^{th} point on the grid of the plasma surface, and $\mathbf{r}_w^i$ is the i^{th} point of the wall grid. The '$-$' superscripts denote quantities on the outer surface of the thin resistive wall (Region II). The w subscript without the '$-$' superscript denotes the inner surface of the wall (Region I), and the p subscript denotes the plasma surface.

The integrals are over the plasma or wall contours. The matrix Green's equations (3.28)–(3.29) are the Green's equations for Region I. The final equation, Eq. (3.30), is that for Region II.

Special care must be exercised when evaluating the singular regions (the regions in which the observation points and the integration contour coincide) of the integrals $\mathbf{G}_{pp} \cdot \nabla_n \chi_p$, $\mathbf{G}_{ww} \cdot \nabla_n \chi_w$, $\mathbf{M}_{pp} \cdot \chi_p$, and $\mathbf{M}_{ww} \cdot \chi_w$. In particular, the integrals $\mathbf{M}_{pp} \cdot \chi_p$, and $\mathbf{M}_{ww} \cdot \chi_w$ are highly singular due to the derivative of the Green's function. A method similar to that used by ERATO [65] is used to deal with this logarithmic singularity. The singular contribution is removed before the numerical integration and then added back in analytic form. This leaves the integrals well behaved.

The matrices $\mathbf{P}$ and $\mathbf{N}$ in Eq. (3.30) contain the terms representing the active feedback system. The forms of the feedback matrices $\mathbf{P}$ and $\mathbf{N}$ are quite different depending on whether the observation flux loops controlling the feedback system are inside the resistive wall (Region I) or exterior to the wall (Region II: see Fig. 3.1). Since the active feedback currents [defined in Eq. (3.26)] are a linear function of the perturbed flux at the observation points, they can, after some algebra, be expressed in terms of matrix relations involving χ and $\nabla_n \chi$ on the wall and plasma surfaces. The derivations of the active feedback matrices will be given in the next section.

3.2.2 Matching conditions at the resistive wall

We need to relate the terms in the region outside the wall to the region inside the wall. For this we make use of a "thin-wall approximation." In this approximation the perturbed flux does not change across the resistive wall boundary: $\chi_w \equiv \chi_w^- = \chi^-$. However, the normal derivative of the flux changes at the boundary. We need to calculate this jump condition. The jump in the tangential magnetic field across a boundary is given by [61]

$$\hat{\mathbf{n}} \times (\mathbf{b}_2 - \mathbf{b}_1) = \mu_0 \mathbf{K}, \tag{3.35}$$

where $\mathbf{K}$ is the current flowing on the boundary surface, and in our case $\mathbf{b}$ is the perturbed poloidal magnetic field

$$\mathbf{b} = \frac{1}{2\pi} \nabla\phi \times \nabla\chi.$$

Therefore

$$\hat{\mathbf{n}} \times (\mathbf{b}_2 - \mathbf{b}_1) = \frac{1}{2\pi}\nabla\phi\,(\hat{\mathbf{n}}\cdot\nabla\chi)_2 - (\hat{\mathbf{n}}\cdot\nabla\chi)_1 = \frac{1}{2\pi}\nabla\phi\,(\hat{\mathbf{n}}\cdot\nabla\chi)_{\omega}, \tag{3.36}$$

and we get

$$(\mathbf{n}\cdot\nabla\chi)_{\omega} = 2\pi\mathcal{N}^2\mu_0\mathbf{K}\cdot\nabla\phi, \tag{3.37}$$

where $(\ldots)_{\omega}$ denotes the jump across the resistive wall. We need to find a relation for the surface-current term. By definition, $\mathbf{K}\cdot\nabla\phi = \delta_{\omega}\mathbf{J}\cdot\nabla\phi$, where δ_{ω} is the thickness of the resistive wall. By taking the poloidal component of Faraday's law and substituting Ohm's law, we can find an expression relating $\mathbf{J}\cdot\nabla\phi$ and the time rate of change of the perturbed poloidal flux at the resistive wall. This gives

$$\frac{\partial\chi}{\partial t} = 2\pi\mathcal{N}^2\mathbf{E}\cdot\nabla\phi = 2\pi\mathcal{N}^2\eta\mathbf{J}\cdot\nabla\phi. \tag{3.38}$$

Therefore the thin-wall approximation gives an expression that relates the jump in the normal derivative across the thin resistive wall boundary to the time derivative of the flux on the boundary:

$$(\hat{\mathbf{n}}\cdot\nabla\chi)_{\omega} = \delta_{\omega}\frac{\mu_0}{\eta}\frac{\partial\chi}{\partial t} = -i\omega\mu_0\frac{\delta_{\omega}}{\eta}\chi_{\omega} \equiv K\chi_{\omega}, \tag{3.39}$$

where δ_{ω} is the thickness of the resistive wall and η is the resistivity of the wall. This relation is valid for a wall whose thickness is less than the classical skin depth, $\delta_{1,40}$, as this is the limitation of the validity of Eq. (3.35). This is normally a good approximation, as we will see. If it were not a good approximation, then flux from equilibrium field coils, control coils, and even from the ohmic field coils would have great difficulty penetrating the resistive vacuum vessel wall fast enough to be effective.

By using a variable jump coefficient $K(\Theta)$ at the various theta points, we can simulate a wall with variable thickness and material resistivity, and even a wall with toroidally axisymmetric gaps by using the limit $K \to 0$.

Consider the case with a resistive wall and no active feedback system (i.e., the active feedback matrices $\mathbf{P}$ and $\mathbf{N}$ are identically zero). Equations (3.28)–(3.30) may be solved to obtain $\nabla_n\chi_p$ in terms of χ_p. First χ_{ω} is eliminated in favor of $\nabla_n\chi_{\omega}^-$ in Eq. (3.30). The jump condition, Eq. (3.39), is provides the necessary relation between $\nabla_n\chi_{\omega}^-$ and $\nabla_n\chi_{\omega}^-$. Then from Eq. (3.29) χ_{ω} is calculated in terms of χ_p and $\nabla_n\chi_p$. Finally Eq. (3.28) gives $\nabla_n\chi_p$ in terms of χ_p. Therefore we find that

$$\nabla_n\chi_p = \mathbf{D}^{-1}\cdot\mathbf{C}\cdot\chi_p, \tag{3.40}$$

where

$$\mathbf{D} = \mathbf{G}_{pp} - \mathbf{M}_{pw} \cdot \mathbf{A} \cdot \mathbf{B}^{-1} \cdot \mathbf{G}_{wp} - \mathbf{G}_{pw} \cdot \mathbf{B}^{-1} \cdot \mathbf{G}_{wp} \tag{3.41}$$

$$\mathbf{C} = (\underline{\underline{1}} - \mathbf{M}_{pp}) - \mathbf{G}_{pw} - \mathbf{M}_{pw} \cdot \mathbf{A} \cdot \mathbf{B}^{-1} \cdot \mathbf{M}_{wp}$$

$$\mathbf{B} = \mathbf{G}_{ww} - (\underline{\underline{1}} - \mathbf{M}_{ww}) \cdot \mathbf{A}$$

$$\mathbf{A} = -(\underline{\underline{1}} - \mathbf{M}_{ww} - K\mathbf{G}_{ww})^{-1} \cdot \mathbf{G}_{ww}.$$

For the case with active feedback the procedure for reducing the equations is similar, but the active feedback matrices are included and the definitions of the matrices through D are different. We give the derivations for those cases in the next chapter.

In the limit of an ideally conducting wall we have the boundary condition that $\chi(\mathbf{r}_w) = constant$ on the ideal wall. Since this constant is arbitrary we choose $\chi(\mathbf{r}_w) = 0$. There are only the two Green's equations for Region I, Eqs. (3.28) and (3.29). The equation for Region II, Eq. (3.30), does not exist in this case because anything outside the ideal wall has no effect in the region within the wall and therefore contributes nothing to the problem.

3.2.3 Boundary conditions at the plasma-vacuum interface

The boundary condition required by NOVA to solve the eigenvalue equations is the Fourier components P_{1m} of the total perturbed pressure P_1 in terms of the components ξ_m of the radial displacement ξ_ψ at the plasma-vacuum interface. It is the goal of the vacuum-region calculation to provide this boundary condition. The perturbed pressure P_1 at the P-V interface is found in terms of the normal derivative of the flux at the boundary:

$$P_1 = \mathbf{B} \cdot \mathbf{b} = (\nabla\phi \times \nabla\psi) \cdot (\nabla\phi \times \frac{\nabla\chi}{2\pi}) - g(\psi_{edge})\nabla\phi \cdot \mathbf{a}_t \nabla\phi, \tag{3.42}$$

$$= \frac{\nabla\psi}{2\pi \cdot Y^2} \nabla_n \chi - \mathbf{a}_t g(\psi_{edge}) \frac{1}{Y^2}$$

where ψ_{edge} is the value of ψ at the plasma-vacuum boundary.

We have calculated the normal derivative of the flux in terms of the perturbed flux at the interface in Eq. (3.40). Therefore it remains to find how the perturbed flux χ and the toroidal perturbation $\mathbf{a}_t$ are related to the modes of the displacement ξ_m at the interface.

Recall Eq. (3.15):

$$\mathbf{b} = \frac{1}{2\pi}\nabla\phi \times \nabla\chi - a_v\nabla\phi.$$

Since we know that $\mathbf{b} = \nabla \times \delta\mathbf{A}$ we can find $\delta\mathbf{A}$:

$$\delta\mathbf{A} = -\frac{1}{2\pi}\chi\nabla\phi - a_v \ln R\frac{\partial Z}{\partial w}\nabla w - \frac{\partial Z}{\partial\Theta}\nabla\Theta - \nabla F. \tag{3.43}$$

where F is some arbitrary scalar function. From Eq. (3.3) we also know the form of δA on the plasma surface:

$$\mathbf{b} = \nabla \times (\xi \times \mathbf{B}) \implies \delta\mathbf{A}_p = \xi \times \mathbf{B}. \tag{3.44}$$

If we equate Eqs. (3.43) and (3.44) and take the projection along $\nabla\phi$ we find

$$\nabla\phi \cdot \xi \times \mathbf{B} = -\frac{1}{2\pi}\chi\,\nabla\phi^2. \tag{3.45}$$

where the other terms have vanished owing to axisymmetry. Therefore we have

$$\nabla\phi \cdot \{\xi \times (\nabla\phi \times \nabla v) - g(v)\nabla\phi\} = -\frac{1}{2\pi}\chi\,\nabla\phi^2. \tag{3.46}$$

This reduces to

$$\xi_w = \xi \cdot \nabla v = \sum_m \xi_m \exp(im\Theta) = -\frac{1}{2\pi}\chi. \tag{3.47}$$

This gives us the required relation between the perturbed poloidal flux χ at the interface and the modes of the displacement ξ_m.

To find a_v, we begin by taking the definition for $\mathbf{b}$ in the vacuum, Eq. (3.15), and apply Faraday's law.

$$\frac{\partial\mathbf{b}}{\partial t} = -\nabla \times \delta\mathbf{E}. \tag{3.48}$$

We then take the toroidal projection of this equation and integrate over the surface defined by a plane at constant ϕ between the plasma and the conducting wall. This gives

$$\frac{\partial\Phi}{\partial t} = \frac{\partial}{\partial t}\int \mathbf{b}\cdot d\mathbf{S} = -\int(\nabla \times \delta\mathbf{E})\cdot d\mathbf{S}. \tag{3.49}$$

where Φ is the perturbed toroidal flux through the area bounded by the plasma-vacuum interface and the conducting wall:

$$\frac{\partial\Phi}{\partial t} = -i\omega\int \nabla\phi|a_v\,dS = -\int(\nabla \times \delta\mathbf{E})\cdot d\mathbf{S} \tag{3.50}$$

or

$$-i\omega \oint \frac{\mathbf{a}_t}{X} dS = -\oint_w \delta\mathbf{E}\cdot d\mathbf{l} - \int_p \delta\mathbf{E}\cdot d\mathbf{l}. \tag{3.51}$$

The perturbed toroidal flux Φ is defined by

$$\Phi = \int \mathbf{b}_t \cdot d\mathbf{S} = \int \frac{\mu_0 I_w}{2\pi X} dS = L_w I_w. \tag{3.52}$$

where I_w is the poloidal current induced in the surrounding wall owing to the toroidal part of the magnetic field perturbation $\mathbf{b}$, and L_w is the corresponding self-inductance term. We also find that

$$\oint_w \delta\mathbf{E}\cdot d\mathbf{l} = \eta \oint_w \mathbf{j}\cdot d\mathbf{l} = R_w I_w. \tag{3.53}$$

where R_w is the poloidal resistance of the resistive wall. Therefore we get a "circuit" equation for the poloidal currents in the resistive wall

$$L_w \frac{\partial I_w}{\partial t} - R_w I_w = (R_w - i\omega L_w) I_w = -\oint_p \delta\mathbf{E}\cdot d\mathbf{l}. \tag{3.54}$$

We can determine I_w from this circuit equation, and from that we can calculate $\mathbf{a}_t$, but first we must evaluate the integral on the right-hand side of Eq. (3.54),

$$-\oint_p \delta\mathbf{E}\cdot d\mathbf{l} = i\omega \oint_p \delta\mathbf{A}\cdot d\mathbf{l}. \tag{3.55}$$

because $\delta\mathbf{E} = -\partial\delta\mathbf{A}/\partial t = i\omega\delta\mathbf{A}$. Note that the gradient of an arbitrary gauge-dependent scalar function in this definition will vanish in the integration of Eq. (3.55). We recall that on the plasma surface $\delta\mathbf{A} = \xi \times \mathbf{B}$; thus

$$i\omega \oint_p \delta\mathbf{A}\cdot d\mathbf{l} = i\omega \oint_p (\xi \times \mathbf{B})\cdot \tau dl. \tag{3.56}$$

where τ is the surface unit tangent vector $\tau = \varphi \times \hat{\mathbf{n}}$.

We find

$$\tau \cdot \xi \times \mathbf{B} = \tau \cdot \xi \times \mathbf{B}_t = -\mathbf{g}\cdot\nabla\phi\,(\xi\cdot\hat{\mathbf{n}}) = -\frac{g(\psi)}{X\,|\nabla\psi|}\xi_\psi. \tag{3.57}$$

This gives

$$i\omega \oint_p (\xi \times \mathbf{B})\cdot d\mathbf{l} = -i\omega \mathbf{g}(\psi_{edge}) \oint_p \frac{\xi_\psi}{X\,|\nabla\psi|} dl = -i\omega \mathbf{g}(\psi_{edge}) \oint_p \mathcal{J}\frac{\xi_\psi}{X^2} d\Theta. \tag{3.58}$$

where this last part follows from Eq. (3.27), which defines the arc length dl. Using Eq. (3.47) we finally get

$$-\oint_p \delta\mathbf{E}\cdot d\mathbf{l} = -i\omega g(v_{edge})\sum_m \xi_m \oint_p \frac{\mathcal{J}}{X^2}\exp(im\Theta)d\Theta. \tag{3.59}$$

It is interesting to note that if one is using PEST coordinates ($\mathcal{J} = X^2$), then only the $m = 0$ term will contribute to the sum (corresponding to a pure compression of the plasma cross section). We can therefore complete the circuit equation (3.54) and solve for I_w:

$$I_w = \frac{-i\omega g(v_{edge})}{R_w - i\omega L_w}\sum_m \xi_m \oint_p \frac{\mathcal{J}}{X^2}\exp(im\Theta)d\Theta. \tag{3.60}$$

Then, recalling the definition for the perturbed toroidal flux Φ from Eq. (3.52),

$$\Phi = L_w I_w = a_r \int \frac{1}{X} dS, \tag{3.61}$$

we get an expression for a_r:

$$a_r = \frac{-i\omega g(v_{edge})L_w}{I_s(R_w - i\omega L_w)}\sum_m \xi_m \int_0^{2\pi} \frac{\mathcal{J}}{X^2}\exp(im\Theta)d\Theta. \tag{3.62}$$

where

$$I_s = \int \frac{1}{X} dS = \int_0^{2\pi} \frac{Z}{X}\frac{\partial X}{\partial \Theta}d\Theta. \tag{3.63}$$

Also, we know from Eq. (3.52) the definition for L_w,

$$L_w = \frac{\mu_0}{2\pi}\int \frac{dS}{X} = \frac{\mu_0}{2\pi}\int_0^{2\pi} \frac{Z}{X}\frac{\partial X}{\partial \Theta}d\Theta. \tag{3.64}$$

and the poloidal resistance R_w is given by

$$R_w = \oint_w \frac{\eta}{\delta}\frac{dl}{2\pi X}. \tag{3.65}$$

It is also interesting to note that the result we get for a_r in Eq. (3.62) (which comes from our circuit model) in the limit of an ideal conducting wall ($R_w = 0$) is the same result obtained using the method of Lüst and Martenson [66] (which is used in the ideal-vacuum version of NOVA [51]). The expression that we derive for the resistive wall using the circuit model in fact differs from the expression of the ideal wall of Lüst and Martenson only by a factor of $-i\omega L_w/(R_w - i\omega L_w)$.

Therefore using Eq. (3.62) in Eq. (3.42) for the toroidal component and using Eqs. (3.40) and (3.47) in Eq. (3.42), we can completely determine the total perturbed pressure P_1 in terms of the components of the radial displacement ξ_m at the plasma-vacuum interface. We then integrate Eq. (3.42) over Θ to get P_1 in terms of the Fourier modes:

$$P_1 = \mathbf{b} \cdot \mathbf{B} = \sum_m p_m \exp(im\Theta) = \sum_{m,m'} \tilde{M}_{m,m'} \xi_{m'} \exp(im\Theta). \tag{3.66}$$

The matrix $\tilde{M}$ is the result of the vacuum calculation. This matrix relates P_1 at the boundary to the Fourier components of ξ_ν and includes the effects of the resistive conductors, the active feedback currents, and the geometry of the vacuum region.

3.3 The Active Feedback Matrices

3.3.1 Current-control feedback matrices

Here we derive the form of the feedback matrices P and N. As mentioned above, the form of the feedback matrices depends on the region in which the flux observation detection loops are located. We begin with the basic Green's equation defining the perturbed poloidal flux at an observation point $\mathbf{r}_o = (R_o, Z_o)$ *not* on the plasma surface or the wall surface:

$$\chi(\mathbf{r}_o) = \frac{1}{2} \sum_{m=1}^{M} \mu_0 I_m G(\mathbf{r}_m; \mathbf{r}_o) - \frac{1}{4\pi} \oint \frac{dl_T}{R_T} \chi_T (\hat{n} \cdot \nabla_T G(\mathbf{r}_T; \mathbf{r}_o)) \tag{3.67}$$

$$- \frac{1}{4\pi} \oint \frac{dl_T}{R_T} G(\mathbf{r}_T; \mathbf{r}_o)(\hat{n} \cdot \nabla_T \chi_T).$$

The magnitudes of the feedback-coil currents are proportional to the difference between the perturbed poloidal fluxes at the two observation points as given in Eq. (3.26). Therefore we must calculate the value of the perturbed flux χ at these points in terms of surface integrals over χ and $\nabla_n \chi$ from Eq. (3.67). Here we see that the form of the feedback matrices depends upon the region in which the flux loop detectors are located. The value of the perturbed flux at some point in Region I as calculated from Eq. (3.67) will clearly depend on integrals over the surface of the plasma as well as the inner surface of the resistive wall. On the other hand, if the detectors are in Region II, the value of χ at the observation points will depend on the

surface integrals only over the outer wall (not over the plasma surface). In addition, however, there will be a direct contribution from the feedback coils themselves, as opposed to the former case, in which the detectors in Region I "feel" the active feedback coils only through the boundary condition that connects Region I and Region II.

In this section we are considering only current-control feedback, in which the actual feedback currents are a function only of the flux-loop measurements. The currents are axisymmetric current loops with none of the characteristics of a real active feedback circuit. The case of voltage feedback will be considered in the next section, in which the circuit equations of the active feedback coils are included in the feedback matrices.

Case A: Detector loops in Region I

We consider first the case in which the observation points are located in Region I, while the active feedback coils are in Region II. We use Eq. (3.67) to evaluate the perturbed flux at the observation points. The region is bounded by the plasma surface and the wall surface; therefore the flux at the observation point $\mathbf{r}_{o1} = (X_{o1}, Z_{o1})$ is given by

$$\chi(X_{o1}, Z_{o1}) = \frac{1}{4\pi}\oint_p \frac{dl_p}{X_p}(\hat{\mathbf{n}}\cdot\nabla G(\mathbf{r}_p;\mathbf{r}_{o1}))\chi(\mathbf{r}_p) - \frac{1}{4\pi}\oint_p \frac{dl_p}{X_p}G(\mathbf{r}_p;\mathbf{r}_{o1})(\hat{\mathbf{n}}\cdot\nabla\chi(\mathbf{r}_p)) \tag{3.68}$$
$$-\frac{1}{4\pi}\oint_w \frac{dl_w}{X_w}(\hat{\mathbf{n}}\cdot\nabla G(\mathbf{r}_w;\mathbf{r}_{o1}))\chi(\mathbf{r}_w) - \frac{1}{4\pi}\oint_w \frac{dl_w}{X_w}G(\mathbf{r}_w;\mathbf{r}_{o1})(\hat{\mathbf{n}}\cdot\nabla\chi(\mathbf{r}_w))$$

or, equivalently,

$$\chi(\mathbf{r}_{o1}) = -\mathbf{M}_{o1,p}\cdot\chi_p - \mathbf{M}_{o1,w}\cdot\chi_w - \mathbf{G}_{o1,p}\cdot\nabla_n\chi_p - \mathbf{G}_{o1,w}\cdot\nabla_n\chi_w, \tag{3.69}$$

where

$$\mathbf{M}_{o1,w}\cdot\chi_w = -\frac{1}{4\pi}\oint_w \frac{dl_w}{X_w}(\hat{\mathbf{n}}\cdot\nabla G(\mathbf{r}_w,\mathbf{r}_{o1}))\chi(\mathbf{r}_w) \tag{3.70}$$

$$\mathbf{G}_{o1,w}\cdot\nabla_n\chi_w = -\frac{1}{4\pi}\oint_w \frac{dl_w}{X_w}G(\mathbf{r}_w;\mathbf{r}_{o1})(\hat{\mathbf{n}}\cdot\nabla\chi(\mathbf{r}_w))$$

and similarly for $\mathbf{G}_{o1,p}\cdot\nabla_n\chi_p$ and $\mathbf{M}_{o1,p}\cdot\chi_p$. Note that the arrays $\mathbf{G}_{o1,p}$, $\mathbf{M}_{o1,p}$, and so on, are row vectors (as opposed to matrices as in the case of $\mathbf{M}_{pp}$, $\mathbf{G}_{wp}$, and so on, from the previous section) because we are evaluating χ on the left-hand side of Eq. (3.67) at only one observation point at a time, instead of over all points of the plasma surface or wall surface grid. The same equation gives the flux $\chi(\mathbf{r}_{o2})$ at the observation point $\mathbf{r}_{o2} = (R_{o2}, Z_{o2})$.

We combine the expressions for the perturbed fluxes at the observation flux-loop positions in Eq. (3.26) to give the currents in the active feedback coils:

$$\begin{aligned} I_m &= \alpha_m(\chi(\mathbf{r}_{o1}) - \chi(\mathbf{r}_{o2})) - \beta_m(\dot{\chi}(\mathbf{r}_{o1}) - \dot{\chi}(\mathbf{r}_{o2})) = (\alpha_m - i\omega\beta_m)(\chi(\mathbf{r}_{o1}) - \chi(\mathbf{r}_{o2})) \\ &= (\alpha_m - i\omega\beta_m)\{-\mathbf{M}_{o1,p}\cdot\chi_p - \mathbf{M}_{o1,w}\cdot\chi_w - \mathbf{M}_{o2,p}\cdot\chi_p - \mathbf{M}_{o2,w}\cdot\chi_w - \mathbf{G}_{o1,p}\cdot\nabla_n\chi_p \\ &\qquad -\mathbf{G}_{o1,w}\cdot\nabla_n\chi_w - \mathbf{G}_{o2,p}\cdot\nabla_n\chi_p - \mathbf{G}_{o2,w}\cdot\nabla_n\chi_w\}. \end{aligned} \tag{3.71}$$

Now we include this expression in the sum over the feedback coils in Eq. (3.67). Since the Green's function in the summation is evaluated at every grid point on the wall surface in order to derive the matrix equation (3.30), our feedback sum becomes a matrix in which the rows span the grid points on the wall. The elements (k,l) of $\mathbf{P}_w$ are defined by

$$\mathbf{P}_w^{k,l} = -\sum_m (\alpha_m - i\omega\beta_m)\, G(\mathbf{r}_w^k; \mathbf{r}_c^m)\,[\mathbf{M}_{o1,w}^l - \mathbf{M}_{o2,w}^l]. \tag{3.72}$$

The summation here is over all the active feedback coils, and $\mathbf{r}_c^m$ is the position of the m^{th} feedback coil. Likewise $\mathbf{r}_w^k$ is the position of the k^{th} grid point on the wall surface. Each may have a different set of gain coefficients (α_m, β_m), and there may be any number of coils. This will not affect the size of the feedback matrix, as its row and column dimensions are defined by the number of grid points on the wall surface or plasma surface, as the case may be. The elements (k,l) of $\mathbf{N}_w$ are given by

$$\mathbf{N}_w^{k,l} = \sum_m (\alpha_m - i\omega\beta_m)\, G(\mathbf{r}_w^k; \mathbf{r}_c^m)\,[\mathbf{G}_{o1,w}^l - \mathbf{G}_{o2,w}^l]. \tag{3.73}$$

Likewise the elements (k,l) of $\mathbf{P}_p$ are defined by

$$\mathbf{P}_p^{k,l} = -\sum_m (\alpha_m - i\omega\beta_m)\, G(\mathbf{r}_w^k; \mathbf{r}_c^m)\,[\mathbf{M}_{o1,p}^l - \mathbf{M}_{o2,p}^l], \tag{3.74}$$

and, similarly, the elements (k,l) of $\mathbf{N}_p$ are given by

$$\mathbf{N}_p^{k,l} = \sum_m (\alpha_m - i\omega\beta_m)\, G(\mathbf{r}_w^k; \mathbf{r}_c^m)\,[\mathbf{G}_{o1,p}^l - \mathbf{G}_{o2,p}^l]. \tag{3.75}$$

The matrix equations to be solved are now in the form of Eqs. (3.28)–(3.30) with the active feedback matrices nonzero. We can now rewrite Eq. (3.40) to include the feedback matrices. Now, however, the equation is a bit more complicated:

$$\nabla_n \chi_p = \mathbf{D}^{-1} \cdot \{(\underline{\underline{1}} - \mathbf{M}_{pp}) - \mathbf{C}_2 \cdot \mathbf{M}_{wp} - (\mathbf{M}_{pw} \cdot \mathbf{A} - \mathbf{C}_1) \cdot \mathbf{P}_p\} \cdot \chi_p, \tag{3.76}$$

where

$$\mathbf{D} = \mathbf{G}_{pp} - \mathbf{M}_{pw} \cdot \mathbf{A} \cdot \mathbf{N}_p - \mathbf{C}_2 \cdot \mathbf{G}_{wp} - \mathbf{C}_1 \cdot \mathbf{N}_p \tag{3.77}$$

$$\mathbf{C}_1 = \mathbf{C}_2 \cdot (\underline{\underline{1}} - \mathbf{M}_{ww}) \cdot \mathbf{A}$$

$$\mathbf{C}_2 = [\mathbf{M}_{pw} \cdot \mathbf{A} \cdot (\mathbf{N}_w - \mathbf{G}_{ww}) - \mathbf{G}_{pw}] \cdot \mathbf{B}$$

$$\mathbf{B} = [\mathbf{G}_{ww} - (\underline{\underline{1}} - \mathbf{M}_{ww}) \cdot \mathbf{A} \cdot (\mathbf{N}_w - \mathbf{G}_{ww})]^{-1}$$

$$\mathbf{A} = [\underline{\underline{1}} - \mathbf{M}_{ww} - K\mathbf{G}_{ww} - \mathbf{P}_w]^{-1}.$$

The additional complexity compared to the case with no feedback in Eq. (3.40) arises because of the additional matrices $\mathbf{P}$ and $\mathbf{N}$, and also because we have new terms that multiply χ_p and $\nabla_n \chi_p$, which appear in the *third* matrix equation (3.30). Previously, of course, these terms appeared only in the first two matrix equations, Eqs. (3.28)–(3.29).

Case B: Detector loops in Region II

In this case the observation points are outside the resistive wall. This region is bounded by only one surface (and extends to infinity). However, the feedback-coil currents are in this region, so they make a direct contribution to the value of the perturbed flux at the observation points as well. The perturbed flux χ at the observation point $\mathbf{r}_{o1} = (X_{o1}, Z_{o1})$ in the vacuum region, not on a boundary surface, is given by

$$\chi(\mathbf{r}_{o1}) = \frac{1}{4\pi} \oint_{w-} \frac{dl_w}{X_w} (\hat{\mathbf{n}} \cdot \nabla G(\mathbf{r}_w; \mathbf{r}_{o1})) \chi(\mathbf{r}_w) \tag{3.78}$$

$$- \frac{1}{4\pi} \oint_{w-} \frac{dl_w}{X_w} G(\mathbf{r}_w; \mathbf{r}_{o1}) (\hat{\mathbf{n}} \cdot \nabla \chi(\mathbf{r}_w)) + \frac{1}{2} \sum_{m=1}^{M} \mu_0 I_m G(\mathbf{r}_{o1}; \mathbf{r}_c^m).$$

The feedback currents are defined as before:

$$I_m = (\alpha_m - i\omega\beta_m)(\chi(\mathbf{r}_{o1}) - \chi(\mathbf{r}_{o2})). \tag{3.79}$$

We therefore get

$$\chi(\mathbf{r}_{o1})\Big[2 - \sum_m (\alpha_m - i\omega\beta_m) G(\mathbf{r}_{o1};\mathbf{r}_c^m)\Big] = \frac{1}{4\pi}\oint_{w-} \frac{dl_w}{N_w}(\hat{\mathbf{n}}\cdot\nabla G(\mathbf{r}_w;\mathbf{r}_{o1}))\chi(\mathbf{r}_w) \quad (3.80)$$

$$-\frac{1}{4\pi}\oint_{w-}\frac{dl_w}{N_w} G(\mathbf{r}_w;\mathbf{r}_{o1})(\hat{\mathbf{n}}\cdot\nabla\chi(\mathbf{r}_w)) - \chi(\mathbf{r}_{o2})\sum_{m=1}(\alpha_m - i\omega\beta_m) G(\mathbf{r}_{o1};\mathbf{r}_c^m).$$

Combining this with the equivalent equation for $\chi(\mathbf{r}_{o2})$, we get two linear equations defining the perturbed fluxes at the observation points:

$$\chi(\mathbf{r}_{o1}) = f_1[-\mathbf{M}_{o1,w}\cdot\chi_w - \mathbf{G}_{o1,w}\cdot\nabla_n^+\chi_w - S_1\chi(\mathbf{r}_{o2})] \quad (3.81)$$

$$\chi(\mathbf{r}_{o2}) = f_2[-\mathbf{M}_{o2,w}\cdot\chi_w - \mathbf{G}_{o1,w}\cdot\nabla_n^+\chi_w - S_2\chi(\mathbf{r}_{o1})], \quad (3.82)$$

where $\mathbf{M}_{o1,w}$, $\mathbf{G}_{o1,w}$, and so on, are defined as before, and where

$$S_1 = \sum_m (\alpha_m - i\omega\beta_m) G(\mathbf{r}_{o1};\mathbf{r}_c^m) \quad ; \quad S_2 = \sum_m (\alpha_m - i\omega\beta_m) G(\mathbf{r}_{o2};\mathbf{r}_c^m) \quad (3.83)$$

$$f_1 = \frac{1}{(2 - S_1)} \quad ; \quad f_2 = \frac{1}{(2 + S_2)}. \quad (3.84)$$

To solve for $\chi(\mathbf{r}_{o2})$ we substitute Eq. (3.81) into Eq. (3.82) to find

$$\chi(\mathbf{r}_{o2})[1 - S_1S_2f_1f_2] = f_2\{-\mathbf{M}_{o2,w}\cdot\chi_w^- - \mathbf{G}_{o2,w}\cdot\nabla_n^+\chi_w \quad (3.85)$$
$$-S_2f_1[-\mathbf{M}_{o1,w}\cdot\chi_w^- - \mathbf{G}_{o1,w}\cdot\nabla_n^+\chi_w]\},$$

and likewise for $\chi(\mathbf{r}_{o1})$ we get

$$\chi(\mathbf{r}_{o1})[1 - S_1S_2f_1f_2] = f_1\{-\mathbf{M}_{o1,w}\cdot\chi_w^- - \mathbf{G}_{o1,w}\cdot\nabla_n^+\chi_w \quad (3.86)$$
$$-S_1f_2[-\mathbf{M}_{o2,w}\cdot\chi_w^- - \mathbf{G}_{o2,w}\cdot\nabla_n^+\chi_w]\}.$$

These expressions are now substituted into Eq. (3.79) to give the expression defining the active feedback currents I_m:

$$I_m = \alpha_m(\chi(\mathbf{r}_{o1}) - \chi(\mathbf{r}_{o2})) + \beta_m(\dot{\chi}(\mathbf{r}_{o1}) - \dot{\chi}(\mathbf{r}_{o2})) = (\alpha_m - i\omega\beta_m)(\chi(\mathbf{r}_{o1}) - \chi(\mathbf{r}_{o2}))$$

$$= \frac{(\alpha_m - i\omega\beta_m)}{F}\{-f_1(1 - f_2S_2)\mathbf{M}_{o1,w}\cdot\chi_w - f_2(1 - f_1S_1)\mathbf{M}_{o2,w}\cdot\chi_w$$
$$- f_1(1 - f_2S_2)\mathbf{G}_{o1,w}\cdot\nabla_n^+\chi_w - f_2(1 - f_1S_1)\mathbf{G}_{o2,w}\cdot\nabla_n^+\chi_w\}, \quad (3.87)$$

where

$$F = 1 - S_1 S_2 f_1 f_2. \tag{3.88}$$

Again we substitute this expression into the sum over the feedback coils in Eq. (3.67) to get our feedback equations. This defines the feedback matrices, and thus the elements (k,l) of $\mathbf{P}_w$ are defined by

$$\mathbf{P}_w^{k,l} = \frac{1-K}{F} \sum_m (\alpha_m - i\omega\beta_m)\, G(\mathbf{r}_w^k; \mathbf{r}_c^m) [-f_1(1 - f_2 S_2)\mathbf{M}_{o1,w}^l - f_2(1 - f_1 S_1)\mathbf{M}_{o2,w}^l, \tag{3.89}$$

and the elements (k,l) of $\mathbf{N}_w$ are given by

$$\mathbf{N}_w^{k,l} = \frac{1-K}{F} \sum_m (\alpha_m - i\omega\beta_m)\, G(\mathbf{r}_w^k; \mathbf{r}_c^m) [f_1(1 - f_2 S_2)\mathbf{G}_{o1,w}^l - f_2(1 - f_1 S_1)\mathbf{G}_{o2,w}^l, \tag{3.90}$$

where K is the jump coefficient for the resistive wall as defined in Eq. (3.39). In this case $\mathbf{P}_p$ and $\mathbf{N}_p$ are identically zero because the plasma surface is in Region I, and therefore the plasma affects the perturbed flux at the observation points in Region II only through the boundary condition at the resistive wall.

The matrix equations (3.28)–(3.30) are again solved for $\nabla_n \chi_p$ in terms of χ_p. This case is a little more straightforward than the previous case because $\mathbf{P}_p$ and $\mathbf{N}_p$ are both zero matrices.

$$\nabla_n \chi_p = \mathbf{D}^{-1} \cdot [(\underline{\underline{1}} - \mathbf{M}_{pp}) - \mathbf{C}_2 \cdot \mathbf{M}_{wp}] \cdot \chi_p, \tag{3.91}$$

where

$$\mathbf{D} = \mathbf{G}_{pp} - \mathbf{C}_2 \cdot \mathbf{G}_{wp} \tag{3.92}$$

$$\mathbf{C}_1 = \mathbf{C}_2 \cdot (\underline{\underline{1}} - \mathbf{M}_{ww}) \cdot \mathbf{A}$$

$$\mathbf{C}_2 = [\mathbf{M}_{pw} \cdot \mathbf{A} \cdot (\mathbf{N}_w - \mathbf{G}_{ww}) - \mathbf{G}_{pw}] \cdot \mathbf{B}$$

$$\mathbf{B} = [\mathbf{G}_{ww} - (\underline{\underline{1}} - \mathbf{M}_{ww}) \cdot \mathbf{A} \cdot (\mathbf{N}_w - \mathbf{G}_{ww})]^{-1}$$

$$\mathbf{A} = [\underline{\underline{1}} - \mathbf{M}_{ww} - K\mathbf{G}_{ww} - \mathbf{P}_w]^{-1}.$$

3.3.2 Active and passive feedback circuit equations

In the previous section we considered only the case of a "perfect" feedback system in which the feedback currents are a function only of the flux-loop measurements. In

a realistic control system. of course. one would have the active feedback coils driven by a power supply. which is in turn controlled by the flux-loop measurements. The true dynamics of the current trajectories in the active feedback coils depend on the characteristics of the active feedback circuits. The currents are driven in the active feedback coils on the characteristic L/R time of the circuit. and there is coupling between the active feedback coils. between the coils and the vacuum vessel wall. and indeed between the coils and the plasma itself.

In order to make our model more realistic. we include the proper circuit equations in the feedback derivations and therefore in the feedback matrices. This will give us the additional benefit of accounting for the additional passive stabilization of discrete conducting elements in the vacuum region that are not part of the resistive wall. Therefore we will be able to include the passive effects of the active coils themselves or whatever other conducting elements lie outside the vacuum vessel wall. In addition. we can include the effect of a simple time delay in the power-supply response. We note that our model does not include power-supply characteristics such as a voltage limit. however. as this does not fit within the framework of our simple linear model.

We define the voltage applied to a feedback coil to be some linear combination of the perturbed fluxes at prescribed observation points. By analogy with the feedback system of Section 2.1 we can define the voltage to be proportional to the flux difference (and corresponding time derivative) between two observation points symmetric about the midplane. This is also analogous to the definition of the desired current for the ideal feedback-current model of the previous section. The feedback coil voltages are thus defined as

$$V_i = \tilde{\alpha}_i(\chi(\mathbf{r}_{o1}) - \chi(\mathbf{r}_{o2})) - \tilde{\beta}_i(\dot{\chi}(\mathbf{r}_{o1}) - \dot{\chi}(\mathbf{r}_{o2})). \tag{3.93}$$

The new gain coefficients $\tilde{\alpha}$ and $\tilde{\beta}$ differ from the α and β of Eq. (3.26) and must satisfy the units of this equation. One simple definition would be to specify $\tilde{\alpha}_i$ and $\tilde{\beta}_i$ as simply the corresponding coil resistance r_i multiplied by the current gain coefficients α_i and β_i.

We are. of course. free to define V_i in any manner we choose. A more efficient feedback law is to define

$$V_i = V_{gi}(I_i^{want} - I_i), \tag{3.94}$$

where I_i^{want} is the feedback current we "want" in the coil. defined by Eq. (3.26).

I_i is the actual coil current at a given moment in time, and V_g is the voltage gain coefficient.

We therefore have a circuit equation for the i^{th} coil:

$$r_i I_i - L_i \frac{dI_i}{dt} - \sum_{j \neq i} M_{i,j} \frac{dI_j}{dt} - \sum_w M_{i,w} \frac{dI_w}{dt} - \frac{d}{dt}(M_{i,p} I_p) = V_i. \tag{3.95}$$

This accounts for the resistance of the coil r_i, its self-inductance L_i, and its mutual inductance due to coupling with the other coils $M_{i,j}$, the resistive wall $M_{i,w}$, and the plasma $M_{i,p}$. The inductance terms can all be expressed in terms of the perturbed poloidal flux at the i^{th} coil:

$$-\frac{d}{dt}\chi(\mathbf{r}_c^i) = L_i \frac{dI_i}{dt} - \sum_{j \neq i} M_{i,j} \frac{dI_j}{dt} - \sum_{v,v.} M_{i,v} \frac{dI_v}{dt} - \frac{d}{dt}(M_{i,p} I_p). \tag{3.96}$$

Therefore the circuit equation (3.95) becomes

$$r_i I_i = V_i - \frac{d}{dt}\chi(\mathbf{r}_c^i) = V_i - i\omega\chi(\mathbf{r}_c^i). \tag{3.97}$$

We make use of the equation specifying the flux at an "observation point" in the vacuum region (this observation point is simply the position of the coil in Eq. (3.96)) and we use the definition for V_i from Eq. (3.93). Then the circuit equation for the i^{th} coil is

$$(r_i - i\omega L_i) I_i - \frac{i}{2}\omega\mu_0 \sum_{j \neq i} G(\mathbf{r}_c^i; \mathbf{r}_c^j) I_j = (\tilde{\alpha}_i - i\omega\tilde{\beta}_i)(\chi(\mathbf{r}_{o1}) - \chi(\mathbf{r}_{o2})) \tag{3.98}$$

$$-\frac{i\omega\mu_0}{4\pi}\oint_{w+} \frac{dl_w}{X_w}(\hat{\mathbf{n}} \cdot \nabla G(\mathbf{r}_w; \mathbf{r}_{o1}))\chi(\mathbf{r}_w) - \frac{i\omega\mu_0}{4\pi}\oint_{w+} \frac{dl_w}{X_w} G(\mathbf{r}_w; \mathbf{r}_{o1})(\hat{\mathbf{n}} \cdot \nabla\chi(\mathbf{r}_w)),$$

where the mutual inductance between two coils, $M_{i,j}$, is just given by $\frac{1}{2}\mu_0 G(\mathbf{r}_c^i; \mathbf{r}_c^j)$, and where the self-inductance of a square coil of dimension Δx located at $\mathbf{r}_c^i = (X_i, Z_i)$ is given by [67]

$$L_i = \mu_0 X_i[\ln(\frac{8X_i}{\Delta x}) - (\frac{7}{2} - \frac{1}{2}\ln 2)] - O[(\frac{\Delta x}{2X})^2]. \tag{3.99}$$

When we sum this over all the coils we can rewrite the equations in matrix form that emulates the Ohm's circuit law:

$$\mathbf{R}_{i,j} \cdot \mathbf{I}_j = \mathbf{V}_i, \tag{3.100}$$

where $\mathbf{R}$ is the impedance matrix, $\mathbf{L}$ is the array of feedback coil currents, and $\mathbf{V}$ is the array of "voltage" expressions. The latter contains all the Green's function surface integrals that come from Eq. (3.98). The elements of the matrix $\mathbf{R}$ are given by

$$\mathbf{R}_{i,j} = (r_i - i\omega L_i)\,\delta_{i,j} - \frac{i}{2}\omega\mu_0 G(\mathbf{r}_c^i;\mathbf{r}_c^j)(1-\delta_{i,j}), \tag{3.101}$$

where $\delta_{i,j}$ is the Kronecker delta. The elements of $\mathbf{V}$ are defined by the right-hand side of Eq. (3.98). The expressions derived in the previous section for the flux-loop measurements for a pair of detectors either inside or outside the resistive wall may be substituted into the expression for the elements $\mathbf{V}_i$. In order to determine the true feedback currents to be included in Eq. (3.67) so that we can derive the form of the feedback matrices, we simply invert the matrix $\mathbf{R}$ in Eq. (3.100)

$$\mathbf{I}_j = (\mathbf{R}^{-1})_{j,i}\cdot\mathbf{V}_i. \tag{3.102}$$

Using the results for the flux-loop detectors located in Region I, we derive the active feedback matrices with the circuit equations included, so that the elements (k,l) of $\mathbf{P}_p$ are defined by

$$\mathbf{P}_p^{k,l} = -\sum_j G(\mathbf{r}_w^k;\mathbf{r}_c^j)\sum_i (\mathbf{R}^{-1})_{i,j}\,(\tilde{\alpha}_i - i\omega\tilde{\beta}_i)\,\mathbf{M}_{o1,p}^l - \mathbf{M}_{o2,p}^l. \tag{3.103}$$

The elements (k,l) of $\mathbf{P}_w$ are defined by

$$\mathbf{P}_w^{k,l} = -\sum_j G(\mathbf{r}_w^k;\mathbf{r}_c^j)\sum_i (\mathbf{R}^{-1})_{i,j}\,(\tilde{\alpha}_i - i\omega\tilde{\beta}_i)\,\mathbf{M}_{o1,w}^l - \mathbf{M}_{o2,w}^l \tag{3.104}$$

$$-i\omega\mu_0\sum_j G(\mathbf{r}_w^k;\mathbf{r}_c^j)\sum_i (\mathbf{R}^{-1})_{i,j}\,\mathbf{M}_{i,w}^l - K\mathbf{G}_{i,w}^l.$$

The elements (k,l) of $\mathbf{N}_p$ are defined by

$$\mathbf{N}_p^{k,l} = \sum_j G(\mathbf{r}_w^k;\mathbf{r}_c^j)\sum_i (\mathbf{R}^{-1})_{i,j}\,(\tilde{\alpha}_i - i\omega\tilde{\beta}_i)\,\mathbf{G}_{o1,p}^l - \mathbf{G}_{o2,p}^l, \tag{3.105}$$

and similarly the elements (k,l) of $\mathbf{N}_w$ are given by

$$\mathbf{N}_w^{k,l} = \sum_j G(\mathbf{r}_w^k;\mathbf{r}_c^j)\sum_i (\mathbf{R}^{-1})_{i,j}\,(\tilde{\alpha}_i - i\omega\tilde{\beta}_i)\,\mathbf{G}_{o1,w}^l - \mathbf{G}_{o2,w}^l \tag{3.106}$$

$$-i\omega\mu_0\sum_j G(\mathbf{r}_w^k;\mathbf{r}_c^j)\sum_i (\mathbf{R}^{-1})_{i,j}\,\mathbf{G}_{i,w}^l.$$

Equivalent forms for the feedback matrices with the flux-loop detectors in Region II are easily derived. The elements (k,l) of $\mathbf{P}_w$ are given by

$$P^{kl}_w = \frac{(1-K)}{F}\sum_j G(\mathbf{r}^k_w;\mathbf{r}^j_c)\sum_i (\mathbf{R}^{-1})_{i,j}\,(\tilde{\alpha}_i - i\omega\tilde{\beta}_i)\cdot -f_1(1-f_2S_2)\mathbf{M}^l_{o1,w} \quad (3.107)$$

$$-f_2(1-f_1S_1)\mathbf{M}^l_{o2,w} - i\omega\mu_0\sum_j G(\mathbf{r}^k_w;\mathbf{r}^j_c)\sum_i (\mathbf{R}^{-1})_{i,j}\,(\mathbf{M}^l_{i,w} - K\mathbf{G}^l_{i,w}),$$

and the elements (k,l) of $\mathbf{N}_w$ are given by

$$N^{kl}_w = \frac{(1-K)}{F}\sum_j G(\mathbf{r}^k_w;\mathbf{r}^j_c)\sum_i (\mathbf{R}^{-1})_{i,j}\,(\tilde{\alpha}_i - i\omega\tilde{\beta}_i)\cdot -f_1(1-f_2S_2)\mathbf{G}^l_{o1,w} \quad (3.108)$$

$$-f_2(1-f_1S_1)\mathbf{G}^l_{o2,w} - i\omega\mu_0\sum_j G(\mathbf{r}^k_w;\mathbf{r}^j_c)\sum_i (\mathbf{R}^{-1})_{i,j}\,\mathbf{G}^l_{i,w}$$

The quantities f_1, f_2, S_1, S_2, F, and the jump coefficient K are the same as in Section 3.3. The matrices $M_{i,w}$ and $G_{i,w}$ are defined in the same fashion as $M_{o1,w}$ and $G_{o1,w}$, except that the observation point is now the i^{th} feedback coil:

$$\mathbf{M}_{i,w}\cdot\chi_w = -\frac{1}{4\pi}\oint_w \frac{dl_w}{N_w}(\hat{\mathbf{n}}\cdot\nabla G(\mathbf{r}_w,\mathbf{r}^i_c))\chi(\mathbf{r}_w) \quad (3.109)$$

$$\mathbf{G}_{i,w}\cdot\nabla_n\chi_w = -\frac{1}{4\pi}\oint_w \frac{dl_w}{N_w}G(\mathbf{r}_w;\mathbf{r}^i_c)(\hat{\mathbf{n}}\cdot\nabla\chi(\mathbf{r}_w)). \quad (3.110)$$

This completes the formulation of the vacuum calculation of NOVA-W including the effects of the active and passive feedback.

3.4 Summary

We have presented the basic formulation for the vacuum calculation performed in the NOVA-W code. We have a set of feedback matrices for the cases in which the flux loops are either inside or outside the resistive wall. In both of these cases the active feedback coils are outside the resistive wall in Region II. A slightly different set of matrices is required for the case with active coils inside the resistive wall. This case has not been considered here since it is not common, but is only a slight modification of the previous cases. In addition, we have included circuit equations in the formulation

so that the circuit dynamics of the active feedback coils are accounted for, as well as the effects due to any additional discrete passive conductors in Region II. In the following chapters we will present the results from the application of this formulation to realistic problems with passive and active feedback stabilization.

The NOVA-W code as formulated above allows the examination of axisymmetric stability in two ways. First, one may consider an equilibrium, which is unstable to the axisymmetric mode, with or without any surrounding passive conductors. The growth rate of the (partially stabilized) instability may be calculated with all the surrounding wall elements and additional conductors present in order to quantify the strength of the instability in the absence of active feedback.

Secondly, one may consider a particular model for an active feedback system and calculate how effective it is in stabilizing the vertical instability. Various feedback control laws can be considered, as can various positions of the active feedback coils and flux observation loops. In addition, the role of the circuit dynamics of the feedback system, including any power-supply delay times, can be included in the calculation.

The natural output of NOVA-W is the Fourier components of the ξ_ψ term of the displacement. This output can be used to determine the rigid and non-rigid contributions to the instability. In particular, it allows one to see how the non-rigid contributions vary (with respect to feedback parameters) by measuring the changes in the $m > 1$ components with respect to the $m = 1$ components. One can also examine the form of the perturbed flux in the vacuum region in order to assist in determining effective locations for the flux-loop detectors that control the feedback response.

Chapter 4

Passive Stabilization Results

In this chapter we present the results for passive stabilization calculations performed with the NOVA-W code. These results include tests of the code against independent models for calculating growth rates of the axisymmetric mode with passive stabilization. The first test we perform is against an analytic model [40] for the simplified case of an elliptical-cross-section plasma with constant current density in the large-aspect-ratio cylindrical limit. In addition, we compare NOVA-W results to those obtained from a sophisticated MHD transport code (TSC) [41] for realistic tokamak design configurations such as CIT.

We will then demonstrate the utility of the NOVA-W code for determining passive growth rates of tokamak designs and the greatly improved performance of the NOVA-W code over TSC in calculating these growth rates. This is especially true for calculating changes in the growth rate due to modification of the surrounding passive conductor configuration for a given equilibrium.

In this chapter we will also describe in detail the numerical procedure involved in calculating the growth rates from the NOVA-W code. This includes generating an accurate equilibrium, mapping the equilibrium into the stability coordinates, calculating the normalization, and determining the proper convergence of the growth rate with respect to the extrapolation of the plasma surface to the separatrix surface.

4.1 The numerical method

In this section we describe the method of calculating the passive growth rate for a realistic tokamak design equilibrium. One begins with a 2-D equilibrium code (free-boundary or fixed-boundary) that generates a equilibrium file compatible with the PEST format. In principle, any equilibrium code can be used to generate the equilibria, but for the calculations done in this dissertation using realistic tokamak designs the Tokamak Simulation Code is chosen. The TSC calculates the free-boundary MHD equilibrium and transport for realistic tokamak configurations and can itself simulate the axisymmetric motion of a vertically unstable plasma. The basic equilibrium information (profiles and plasma surface definition) provided by the TSC code is then used as input to a 2-D equilibrium fixed-boundary code that calculates the MHD equilibrium to the desired accuracy.

The accurately resolved equilibrium is then mapped into equal-arc stability coordinates. Special care must be given to diverted equilibria. Since the gradient of the poloidal flux vanishes at the x-point, the Jacobian is singular and the coordinate transformation becomes ill-defined. Therefore, defining the plasma-vacuum equilibrium boundary too close to the separatrix surface can adversely affect the accuracy of the mapping to stability coordinates and thereby affect the entire calculation. It should be noted here that the vacuum calculation depends on the metric quantities at the plasma surface: therefore these quantities must be well resolved on and near the surface. The metric quantities that result from the mapping calculation must be examined to make sure that there are no discontinuities or errors near the plasma edge. If so, then an equilibrium that is better resolved with more grid points is calculated.

In these cases we must use equilibria limited by a surface interior to the separatrix surface for the calculation. The growth rate is calculated for several surfaces limited within the separatrix surface and then these values are extrapolated to the separatrix in order to get an accurate converged result. The boundary surface is labeled by the parameter w_{rat}, which is defined as the ratio of the poloidal flux contained within the given surface to the poloidal flux contained within the separatrix surface. (Therefore $w_{rat} = 1.0$ labels the separatrix surface.) For equilibria whose boundary surface is close to the separatrix ($w_{rat} \geq .97$) one must find the correct growth rate by performing a convergence in the number of surfaces used in the equilibrium calculation (as we shall see in Fig. 4.7). The shape of the boundary surface changes rapidly as

ψ_{rat} approaches 1, and more equilibrium surfaces are needed to properly resolve the equilibrium. It is seen that the growth rates as a function of ψ_{rat} (near $\psi_{rat} = 1$) fit a straight line which gives us a converged growth rate by extrapolating to $\psi_{rat} = 1$. It is more efficient to use equilibria in the range $.94 \leq \psi_{rat} \leq .96$ to calculate the growth rate convergence since it is much easier to get the converged growth rate of these individual equilibria. This is slightly less accurate than carrying out the convergence through to $\psi_{rat} = .99$, but it is much less time consuming, and the resulting growth rate is within 5% of the properly converged growth rate.

The NOVA input code takes the equilibrium mapped into stability coordinates and generates the disk files needed by the eigenvalue solver. Finally, the NOVA-W code itself is executed, which performs the vacuum region calculation (including the feedback system and resistive conductors) and solves for the resulting eigenvalues of the linear MHD stability equations. The vacuum calculation must be performed at every iteration when searching for the root ω of the dispersion relation, since the boundary condition, Eq. (3.66), is a function of ω [see, for example, Eq. (3.39)]. The calculation for the case of an ideal vacuum region (no wall or an ideally conducting wall—with no active feedback system) differs in that the vacuum boundary condition does not depend on ω and therefore need be calculated only once prior to the eigenvalue search.

The NOVA code calculates a eigenvalue (growth rate) that is normalized to the poloidal Alfvén time. Therefore to find the actual growth rate we must include the normalization factor, given by

$$\tau_0 = \frac{B_T(0)}{\sqrt{\mu_0 \rho_0}\; q(1)\; R_{mag}}, \tag{4.1}$$

where B_T is the toroidal magnetic field strength at the magnetic axis, ρ_0 is the mass density at the axis, $q(1)$ is the safety factor at the plasma edge, and R_{mag} is the major radius of the magnetic axis. This normalization enters into Eq. (3.39) where the correct frequency, ω in sec^{-1}, is needed for the jump condition. It also enters into the feedback current definition when there is a derivative gain term as in Eq. (3.26). For a plasma well stabilized by the wall, the growth time is determined by the resistivity and geometry of the wall. Therefore the surface integrals over the jump condition, Eq. (3.39), ultimately determine the growth rate of the instability. In this case, then, the accuracy of the normalization is not important, since the normalization factor is effectively canceled out of the calculation. However, when the wall is distant from

the plasma and the plasma is approaching ideal instability (instability with an ideally conducting wall) the normalization is important, since the growth rate is now affected by plasma inertia, and the mass density factor ρ_0 would then be an important part of the calculation.

The resistive wall can be defined in terms of some geometric function of the plasma surface for simplified cases (as in the case to be examined in Section 4.2). For realistic wall definitions (as we shall use in, for example, Section 4.3) the coordinates defining the wall surface are entered through an input file. The wall resistivity and thickness are defined at every point. The input format is similar to the format for the "wires" used in the TSC calculations. This makes it straightforward to take the wall conductor definitions for TSC and enter the equivalent input for NOVA-W. The code takes this set of points defining the wall and selects a subset corresponding to the number of points M_Θ in the Θ-grid defining the wall (and plasma) surface. This subset is then smoothed to generate a continuous wall contour. Then this set of points is redefined so that adjacent points are equidistant on the wall contour. We then have a smooth wall contour defining a surface in the equal-arc-length stability coordinates. Equal-arc-length stability coordinates are used for a majority of the calculations.

4.2 Code test: analytic model

In this section we introduce an analytic model derived by Dobrott and Chang [40] that calculates growth rates of the vertical instability for a simplified plasma model with a resistive wall and discrete resistive conductors surrounding the plasma. The model uses a straight constant current-density plasma equilibrium with a simple elliptical cross section. This plasma model was first examined with regard to axisymmetric instability without any wall or conductors in the vacuum region by Rutherford [15]. Dobrott and Chang added to the model a thin resistive wall and discrete conductors that satisfy certain geometric constraints and derived a dispersion relation for the growth rate of the instability that is partially stabilized by the resistive wall. We begin by briefly reviewing the analytic model from the paper by Dobrott and Chang [40]. We will use the model with a thin resistive wall, but without any additional conductors in the vacuum region; this simplifies their results somewhat. We then compare results from a numerical approximation of this simplified model to the results predicted by the dispersion relation.

4.2.1 Introduction of the analytic model

We begin with the straight plasma model of Rutherford [15] in which the plasma cross section is elliptical with width a and height b, with $b > a$. The plasma has a uniform equilibrium current density j_z so that the equilibrium poloidal flux $2\pi w$ is defined as

$$w = w_0(\frac{x^2}{a^2} - \frac{y^2}{b^2}). \tag{4.2}$$

where w_0 the defines the poloidal flux value at the plasma surface with

$$w_0 = \frac{a^2 b^2 j_z}{2(a^2 - b^2)}. \tag{4.3}$$

The δW for this simplified plasma model is given by

$$\delta W = \frac{1}{2}\int_p dA\, |\nabla w_1|^2 - \frac{1}{2}\oint_p dl \frac{j_z w_1^2}{\nabla w} - \frac{1}{2}\oint_p dl \frac{w_1}{\nabla w}(\nabla w \cdot \nabla w_1). \tag{4.4}$$

where the perturbed poloidal flux function is given by

$$w_1 = -\xi \cdot \nabla w. \tag{4.5}$$

and ξ is the incompressible plasma displacement. The first integral in Eq. (4.4) is the contribution of the perturbed plasma flux. The second integral in this δW expression is the so-called "current drive" term. The final term gives the contribution from the vacuum region.

In the plasma the natural flux coordinates (w, θ) are used, but in the vacuum region the confocal-ellipsoidal coordinate system (μ, θ) is used, where

$$x = (b^2 - a^2)^{1/2} \sinh\mu \cos\theta \tag{4.6}$$

$$y = (b^2 - a^2)^{1/2} \cosh\mu \sin\theta.$$

The solution to the perturbed flux in the plasma is given by

$$w_1 = \sum_{m=1}^{\infty} w_m(X) \begin{cases} \cos(m\theta) \\ \sin(m\theta) \end{cases} \tag{4.7}$$

where $X \equiv w/w_0$. Rutherford [15] found that only the sine (antisymmetric) perturbations (corresponding to the vertical instability) are unstable, and only the $m = 1$ perturbations are unstable below an elongation of approximately 4.5. This corresponds

to a uniform, rigid vertical motion. Furthermore, only the odd-m perturbations contribute to the problem at any elongation.

The solution to the perturbed flux in the vacuum is given by

$$\psi_1 = \sum_{m=1}^{\infty} \psi_m(1) \exp[m(\mu_0 - \mu)] \begin{Bmatrix} \cos(m\theta) \\ \sin(m\theta) \end{Bmatrix} \tag{4.8}$$

The plasma surface is defined by $\mu = \mu_0$ and the wall surface is defined by [illegible]. The wall is constrained to fit the shape of the confocal-ellipsoidal coordinate surface.

A thin-wall approximation is used to derive the jump condition across the resistive wall for the perturbed flux:

$$\left[\left[\frac{\partial \psi_1}{\partial \mu}\right]\right] = -i\omega\sigma_w A\psi_1. \tag{4.9}$$

where A is a geometric factor defining the thickness of the wall, σ_w is the wall conductivity, and $[\![\ldots]\!]$ is the jump across the resistive wall. This is equivalent to the jump condition, Eq. (3.39). For a solid conductor $2\pi A$ is the annular area of the wall between two elliptical surfaces given by $\mu_w^-(\theta)$ and $\mu_w^-(\theta)$. The quantity A is defined as

$$A \equiv \int_0^{2\pi} d\theta \int_{\mu_w^-}^{\mu_w^-} d\mu(\cos^2\theta - \sinh^2\mu), \tag{4.10}$$

and the average wall thickness δ_w is defined such that

$$\frac{\delta_w^2}{A} = \frac{\int_0^{2\pi} d\theta \int_{\mu_w^-}^{\mu_w^-} d\mu(\cos^2\theta - \sinh^2\mu)}{2\pi^2 \left[\int_0^{2\pi} d\theta(\sinh^2\mu_w - \cos^2\theta)^{1/2}\right]^2}. \tag{4.11}$$

The solution for the perturbed flux in the vacuum is used in the vacuum term of δW^- the final integral in Eq. (4.4), along with the effect of the jump condition, Eq. (4.9), at the resistive wall to give the total vacuum contribution. This is combined with the plasma current term of δW^- to give

$$\delta W_1 = -\frac{1}{2}\oint_p dl \frac{j_z \psi_1^2}{|\nabla\psi_1|} - \frac{1}{2}\oint_p dl \frac{\psi_1}{|\nabla\psi_1|}(\nabla w \cdot \nabla\psi_1) \tag{4.12}$$

$$= \frac{\pi}{2}\sum_{m=1}^{\infty} \left[m\frac{N_m(\sigma_w,\Omega)}{D_m(\sigma_w,\Omega)} - \frac{a^2-b^2}{ab}\right]\psi_m^2(1),$$

where the resistive wall dispersion terms N_m and D_m are defined by

$$N_m(\sigma_w, \Omega) = 1 - \frac{\dot{\sigma}_w \Omega}{m} \exp(m\Delta_{ow}) \cosh(m\Delta_{wo}) \tag{4.13}$$

$$D_m(\sigma_w, \Omega) = 1 + \frac{\dot{\sigma}_w \Omega}{m} \exp(m\Delta_{ow}) \sinh(m\Delta_{wo}). \tag{4.14}$$

The normalized growth rate is $\Omega = -i\omega/\omega_A$, and

$$\dot{\sigma}_w = \sigma_w A \omega_A$$

$$\Delta_{\alpha\beta} = \mu_\alpha - \mu_\beta.$$

The normalized Alfvén frequency ω_A is given by

$$\omega_A^2 \equiv \left(\frac{a}{b}\right)^2 \frac{2\langle B^2 \rangle}{\rho(a^2 - b^2)}, \tag{4.15}$$

where $\langle \ldots \rangle$ is the spatial average over the plasma volume, and ρ is the mass density.

The variation of the plasma-perturbation part of δW is taken with respect to $\xi_m(X)$ in order to minimize $\delta W_p \equiv \frac{1}{2}\int_p dA\, \xi_1^{\,2}$. The equation $K - \delta W_1 - \min \delta W_p = 0$ provides the desired dispersion relation from Dobrott and Chang's model for the unstable $m = 1$ mode:

$$\Omega^2 - \frac{b^2}{a^2} + \frac{b}{a}\frac{N_1(\sigma_w, \Omega)}{D_1(\sigma_w, \Omega)} = 0, \tag{4.16}$$

where

$$K = \frac{1}{2}\rho\omega^2 \int_p dA |\xi|^2 \tag{4.17}$$

is the plasma kinetic energy of the displacement.

When there is no wall in the region surrounding the plasma, Eq. (4.16) reduces to

$$\Omega^2 - \frac{b^2}{a^2} + \frac{b}{a} = 0.$$

This gives the normalized growth rates for this elliptical plasma with no wall as

$$\Omega_{nw} = \pm\sqrt{\frac{b}{a}\left(\frac{b}{a} - 1\right)} \tag{4.18}$$

When the resistive wall is present one must solve Eq. (4.16) to obtain the resistive growth rates. The resulting cubic equation is a function of the wall conductivity σ_w and the wall separation from the plasma Δ_{wo} as well as the elongation ratio b/a. The

three roots correspond to the resistive wall roots discussed on page 314 in Freidberg [24]. The first two roots correspond to the two stable roots when the wall is ideal. These roots are on the $Re\,\omega$ axis when the wall is ideal, and then move into the lower half of the complex plane when the wall resistivity is increased from zero. These roots are therefore stable and oscillatory. The third root moves up the $Im\,\omega$ axis from the origin when the resistivity becomes nonzero. This root does not exist in the case of the ideal wall, and it is the root we are interested in. It corresponds to the vertical instability that grows on the resistive time scale of the wall.

4.2.2 Comparison of numerical results to analytic model

We approximate the constant current-density straight elliptical plasma equilibrium of the analytic model by generating a numerical equilibrium of elliptical cross section with a very large aspect ratio ($A = 100$) and with a nearly constant current density. The q-profile for this numerical equilibrium increases from $q = 1.001$ at the magnetic axis to only $q = 1.011$ at the plasma edge. The resistive wall is constructed to follow the constant-μ contour of the (μ, θ) confocal-ellipsoidal coordinate system defined in Eq. (4.6). For simplicity a wall of constant thickness is used in the numerical calculation, and the resulting growth rate is normalized by the average wall thickness defined in Eq. (4.11). The analytically calculated normalized growth rate must be multiplied by the Alfvén frequency, Eq. (4.15), to compare to the actual growth rate calculated numerically.

We calculate the growth rate analytically and numerically for several equilibria with elongations ranging from $\kappa = 1.2$ to $\kappa = 2.0$. The numerical growth rates are calculated as described in the previous section. We do not need to do the convergence in the w_{rat} parameter discussed in Section 4.1 because we are using a fixed-boundary, non-separatrix equilibrium. The numerical calculation uses 50 radial finite elements, 128 Θ-points around the circumference, a total of 31 poloidal harmonics ($-15 \rightarrow +15$, although it can be seen in Fig. 4.4 that fewer harmonics can be used), and 50 radial surfaces in the equilibrium calculation. The analytic growth rates are calculated from the dispersion relation, Eq. (4.16), and renormalized as discussed above. The results of this comparison are shown in Fig. 4.1a. The comparison is excellent for the range of elongations tested here.

We also compare the growth rates calculated by both methods with no surrounding

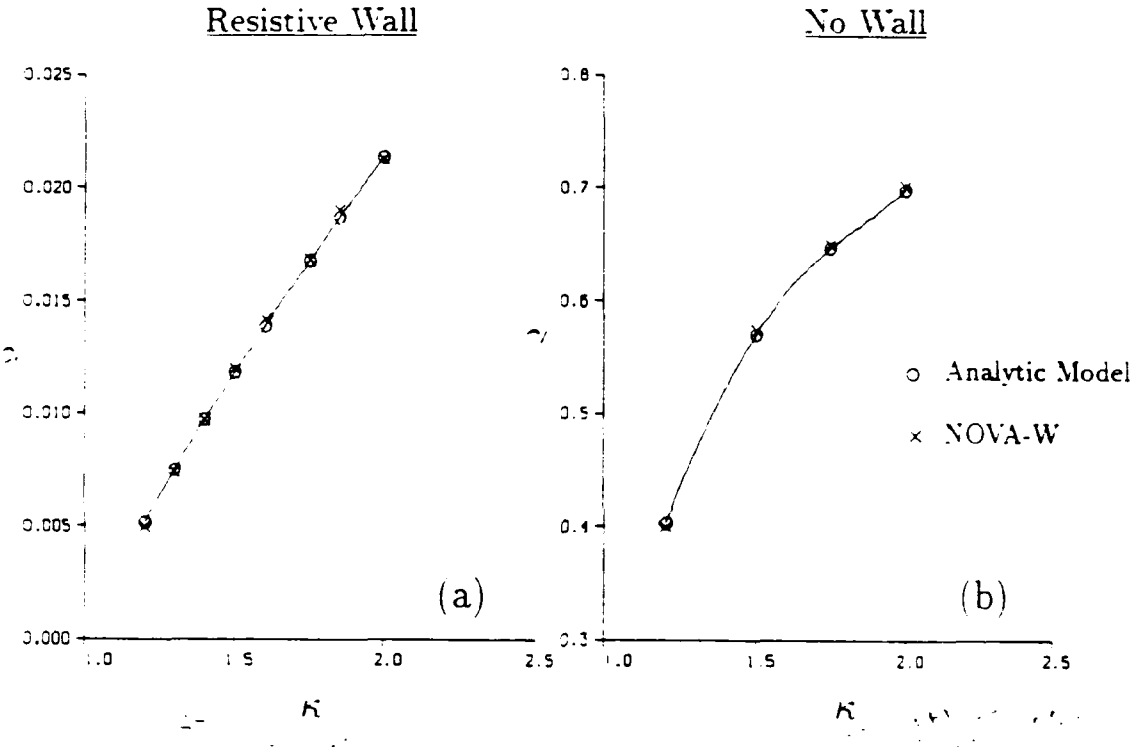


Figure 4.1: (a) *Comparison of resistive wall growth rates* γ *from the NOVA-W code to those of the analytic model with respect to ellipticity* κ *of the plasma. The growth rates* γ *are normalized here to the factor given in Eq. (4.1). For these model equilibria we have* $B_T = 1$, $q_{edge} = 1.011$, $X_{mag} = 10$.

(b) *Comparison of growth rates* γ *from the NOVA-W code to those of the analytic model with respect to ellipticity* κ *of the plasma. This is for the case of a plasma with no wall.*

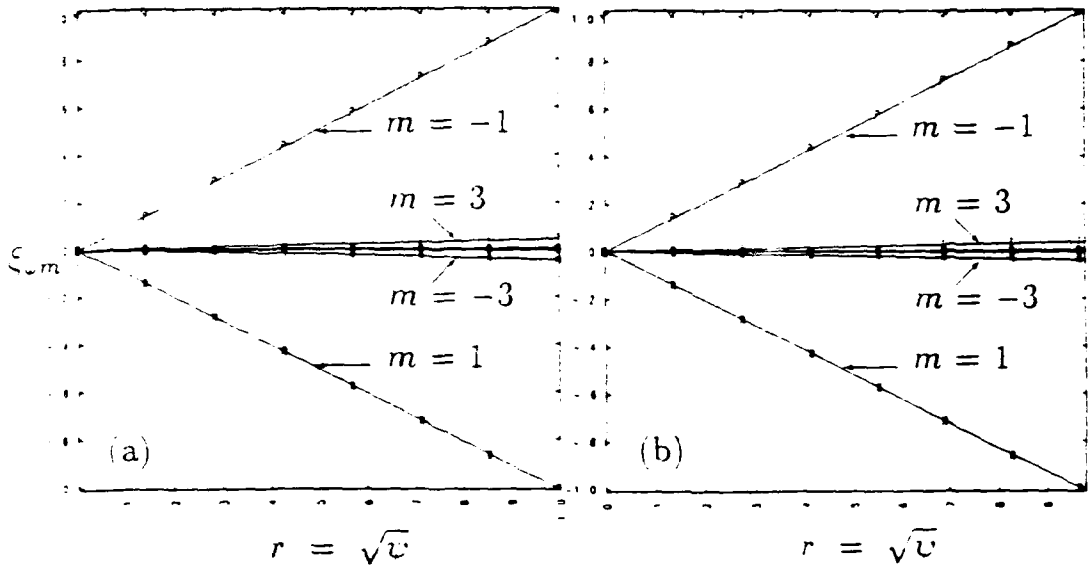


Figure 4.2: *Fourier components of the radial displacement ξ_ω vs. $\sqrt{\upsilon}$ for the $\kappa = 1.4$ case for the true eigenfunction (a) and for a uniform vertical rigid displacement (b). The eigenfunction has primarily $m = \pm 1$ components with much smaller $m = \pm 3$ components. Note that the q-profile is plotted as well. In this case, however, because of the constraints of the analytic model the q-profile is very nearly flat and is plotted at the top of the graph, since the magnitude of the $m = \pm 1$ components at the edge are always normalized to the value of q at the edge.*

resistive wall. The analytic growth rates are given by Eq. (4.18). We must still make sure to use the same normalization, Eq. (4.15), but in this case there is no need to normalize with respect to average thickness and wall conductivity. The results for the case with no resistive wall are presented in Fig. 4.1b for the same range of elongations. In this case as well, we see an excellent comparison between the numerical results and those predicted by the analytic model.

Figure 4.2 shows the Fourier components of the radial displacement ξ_ω as a function of minor radius for the $\kappa = 1.4$ plasma for the true eigenfunction (a) and for the uniform vertical rigid shift (b). The eigenfunction is composed primarily of $m = \pm 1$ components with a small contribution from the $m = \pm 3$ components. The true eigenfunction appears to be identical to the displacement for the rigid motion. In Fig. 4.3 we see ξ_ω for the true eigenfunction and for the rigid shift for the $\kappa = 2.0$ plasma. In this case we see a larger contribution from the $m = 3, 5$ components. It also appears that the true eigenfunction is slightly different from the rigid vertical shift. This suggests that the simplified analytic model may be breaking down for these higher

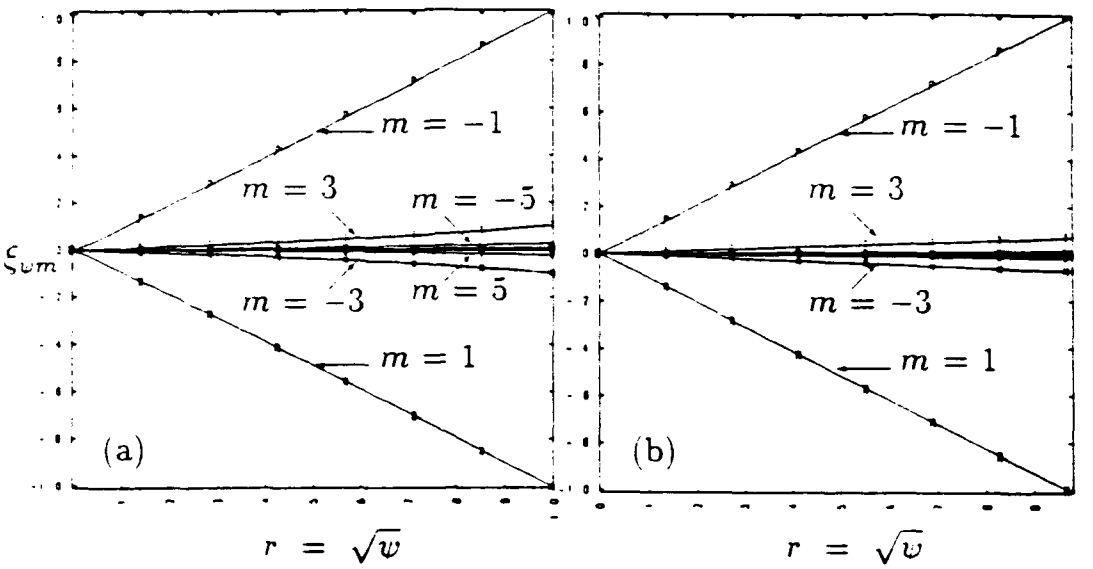


Figure 4.3: *Fourier components of the radial displacement ξ_ψ vs. $\sqrt{\psi}$ for the $\kappa = 2.0$ case for the true eigenfunction (a) and for a uniform vertical rigid shift (b). This true eigenfunction differs slightly from the rigid displacement unlike the $\kappa = 1.4$ case. This suggest a possible breakdown of the simplified model for higher elongations at this aspect ratio ($A = 100$).*

elongations for the finite aspect ratio ($A = 100$) equilibria used in this calculation.

Numerical convergence properties

It is important to demonstrate the convergence properties of the code with respect to the numerical parameters used in the calculation. We consider first the convergence properties with respect to the total number of poloidal harmonics M defining ξ_ψ in Eq. (3.14).

The theory of spectral methods [68,69] tells us that with a Fourier series expansion one can expect the k^{th} coefficient of the expansion to decay faster than any inverse power of k (exponential convergence) when the function and all its derivatives are infinitely smooth and periodic. There must be enough terms in the expansion to represent all the structure of the function, but beyond that one should see an exponential convergence, also termed "spectral accuracy."

In a problem where such infinite smoothness is not present, but some sort of discontinuity exists, one no longer sees exponential convergence, but rather obtains

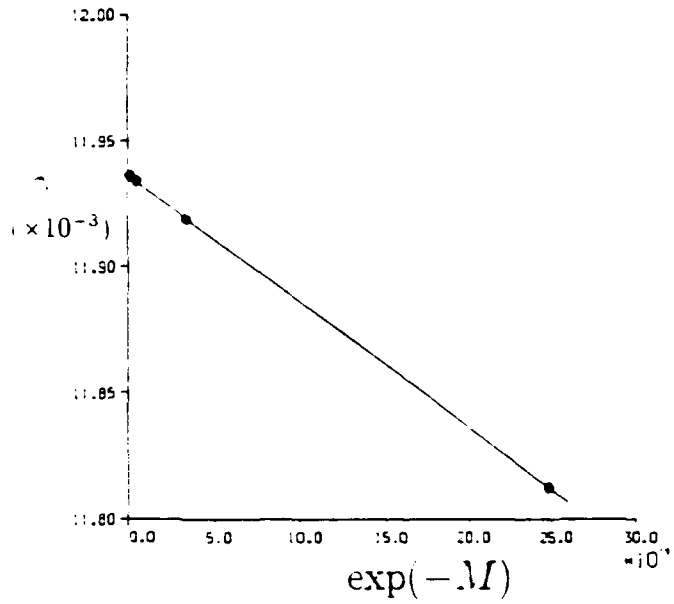


Figure 4.4: *Growth rate convergence in Fourier harmonics for the elliptical plasma* $\kappa = 1.5$ *equilibrium. Note the exponential convergence in the poloidal harmonics for this simple equilibrium.*

a global convergence of order $1/M^2$. We shall see this behavior in the next section when we consider equilibria with separatrices.

In the case of the simple large-aspect-ratio elliptical equilibrium the smoothness and periodicity constraints are satisfied, and in fact we see a definite exponential convergence in the poloidal harmonics in Fig. 4.4. Furthermore, the analytic theory of Rutherford [15] tells us that only the odd-m harmonics contribute to the unstable eigenfunction, and in fact we see only odd-m contributions to the eigenfunction (see Figs. 4.2 and 4.3), and there is absolutely no change in the eigenvalue when the calculation is performed with m_{odd} harmonics or with $m_{\text{odd}} - 1$ harmonics.

It is also necessary to demonstrate that the equilibrium used in the stability calculation is well resolved in the number of surfaces N_ψ used in the equilibrium calculation. This is especially important with regard to the discussion in Section 4.1 about calculating the growth rates of diverted equilibria. We defer a demonstration of the convergence properties with respect to N_ψ until we consider diverted plasmas in Section 4.3.1. The convergence of γ with respect to N_ψ for two CIT equilibria are shown in Fig. 4.7b.

4.3 Code test: realistic numerical model

In this section we perform a vertical stability calculation for a CIT equilibrium with a surrounding resistive vacuum vessel wall. We will compare our results with those obtained using the Tokamak Simulation Code. Vertical instability growth rates are calculated using TSC by perturbing an up-down symmetric equilibrium and then tracking the vertical motion of the plasma by observing the time development of the flux difference between pairs of up-down symmetric flux observation points. Several pairs are typically used. These flux differences are fit to an exponential to obtain a growth rate.

In order to get an accurate growth rate by this method, one must do a convergence in the mass-enhancement factor (FFAC) that is used by TSC [44] to deal with the difference between the resistive and ideal MHD time scales. Several runs must be performed at different values of FFAC, and an extrapolation to FFAC= 1 is made to get a converged growth rate. The run time of each TSC simulation is proportional to 1/FFAC, and although accurate growth rates are obtained, the method is computationally expensive. In addition, TSC advances the *nonlinear* MHD transport equations in time, so calculating the *linear* growth rates can be difficult if the arbitrary initial perturbation (which is clearly not the true eigenfunction) produces transients that are slow to decay. One can calculate the linear growth rate only after these transients have died away. However, if these transients decay slowly, then the displacement may get large and introduce new nonlinearities.

4.3.1 CIT plasma with vacuum vessel wall

The CIT equilibrium used here is a $\kappa = 2$ (at the 95% flux surface) diverted plasma with relatively low triangularity ($\delta = 0.26$) that is in the current-ramp stage (just prior to flat-top) of the CIT evolution. The parameters describing the CIT equilibrium are given in Table 4.1. Figure 4.5a shows the CIT equilibrium with the surrounding vacuum vessel structure as used in the TSC model. The CIT vacuum vessel structure consists of Inconel 625 (resistivity $\eta = 1.35\times10^{-6}$ Ω-m) on the inboard region and Inconel 600 ($\eta = 1.08\times10^{-6}$ Ω-m) on the outboard region. The thickness varies from 4 cm on the inboard region to 8.75 cm on the thickest outboard section.

The resistive wall used in the NOVA-W calculation is shown in Fig. 4.5b. The

Plasma Current I_p	12.30 MA
Major Radius R_0	2.182 m
Minor Radius a	0.660 m
Elongation $\kappa(95\%)$	1.996
Triangularity $\delta(95\%)$	0.258
Toroidal Field $B_T(0)$	11.0 T
$q(95\%)$	4.5
β	0.0092
$n_e(0)$	1.08×10^{21} m^{-3}

Table 4.1: *Equilibrium parameters of CIT plasma used in the passive stabilization study.*

wall definition was entered using the same points defining the vacuum vessel wall as in the TSC calculation, Fig. 4.5a. It is specified to have the same total resistance as the sum of all the conductors that make up the wall in the TSC calculation.

The wall contour has been smoothed in the NOVA-W calculation to obtain a continuous curve appropriate for the surface integrations. Note that the port extension on the far outboard side has been reduced and smoothed over in the NOVA-W version. It will be seen in Fig. 4.11 that the eddy currents are very small in this section of the wall (where the conductors are very thin) and contribute very little to the problem. Therefore we can modify this section of the wall with little effect on our growth rates. This wall section was modified in order to make a smooth, continuous wall contour appropriate to the equal-arc coordinates used in the calculation. The Θ-grid is shown connecting the wall to the plasma surface. Note also that the direction of increasing Θ is clockwise in this coordinate system.

The NOVA-W growth rate is calculated following the numerical procedure outlined in Section 4.1. The original equilibrium information is obtained from the TSC code. A well-resolved equilibrium is calculated from this information and then mapped into equal-arc stability coordinates.

The CIT equilibrium has a separatrix surface, and we must therefore perform a convergence study as discussed in Section 4.1. The convergence of the growth rate γ

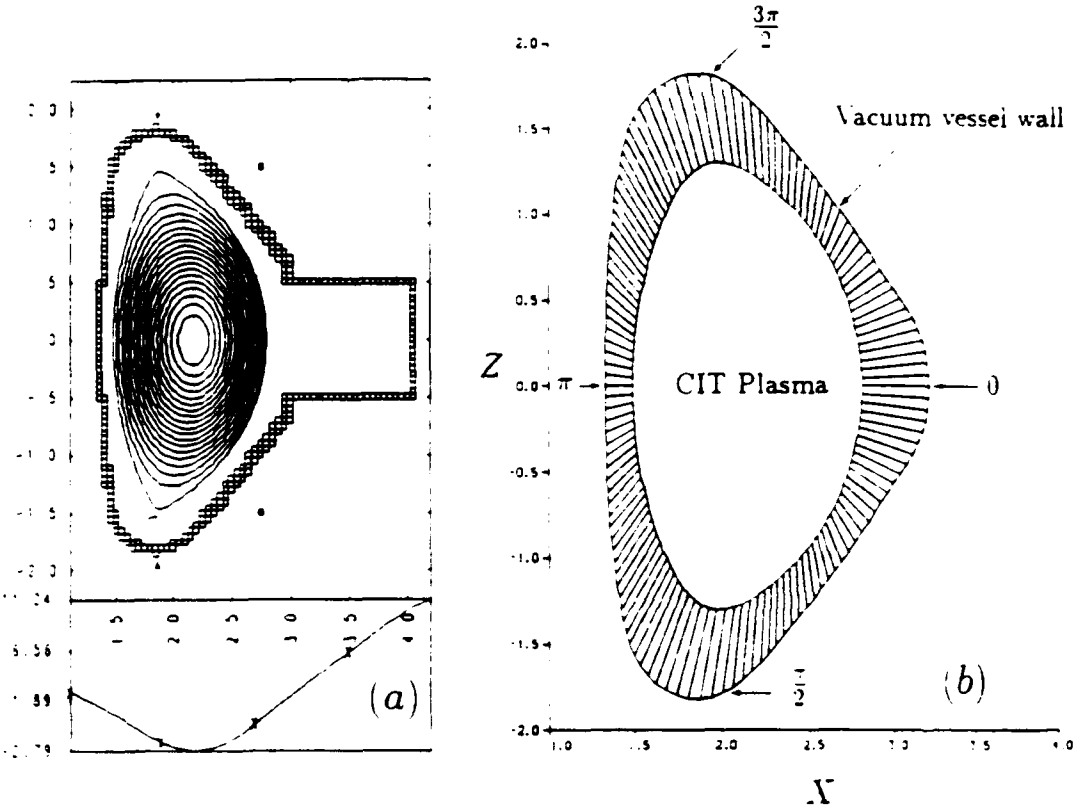


Figure 4.5: (a) *Equilibrium poloidal flux contours (from TSC) of the* $\kappa = 2$ *equilibrium used in this calculation. The TSC representation of the vacuum vessel wall is shown as well as the two active feedback coils outside the vacuum vessel. The poloidal flux through the plasma along the midplane is shown at the bottom. The arrows indicate the point of radial extension for the study presented in Section 4.4.1.*

(b) *CIT wall contour used by NOVA-W. The points of the* Θ*-grid are shown on the wall and plasma surfaces, and the corresponding points are connected with line segments to show the relation between the wall points and the points on the plasma surface. The points designated by* $\Theta = 0, \pi/2, \pi, 3\pi/2$ *are indicated on the figure. Note the comparison to the TSC representation of the wall shown in (a).*

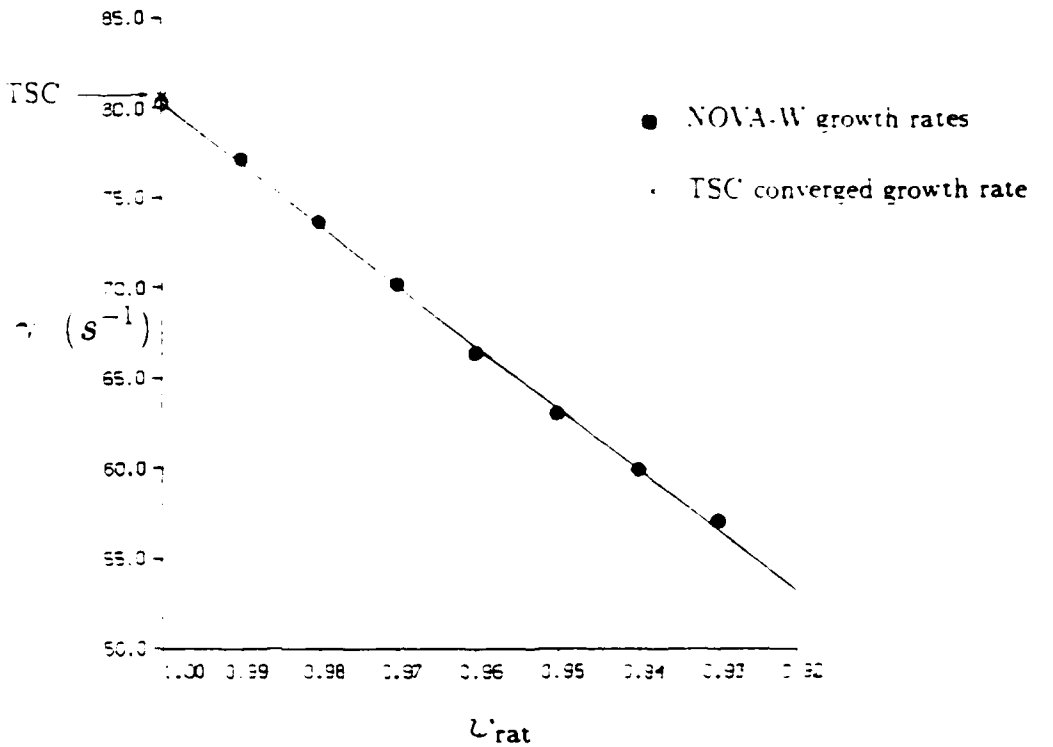


Figure 4.6: *Convergence of NOVA-W growth rate with respect to* ω_{rat}. *Also shown is the converged value of the TSC calculated growth rate. The growth rate* γ *is in inverse seconds. The comparison is seen to be quite good.*

as a function of ω_{rat} is shown in Fig. 4.6. We obtain the converged growth time of $\gamma = 80.31$ sec^{-1} ($\tau = 12.45$ ms). This compares well with the result obtained from TSC of $\gamma = 80.65$ sec^{-1} ($\tau = 12.4$ ms).

Now consider the convergence of the growth rate in poloidal harmonics for the CIT ($\omega_{rat} = 0.95$) equilibrium. Following the discussion in Section 4.2.2 regarding the convergence properties of the spectral method, we do not expect an exponential convergence in this case because of discontinuities in higher-order derivatives of some of the equilibrium metric quantities at the plasma boundary due to the existence of a separatrix. In fact, the results shown in Fig. 4.7a indicate a $1/M^2$ convergence for this equilibrium.

In Fig. 4.7b we see the convergence with respect to the number of radial surfaces needed to resolve the equilibrium. Whereas for the $\omega_{rat} = .96$ equilibrium we see that there are sufficient surfaces ($N_\psi > 100$) to obtain a converged growth rate, the

growth rate for the $v_{rat} = .99$ equilibrium keeps increasing with larger N_ψ. Therefore one must obtain a convergence in N_ψ for the high v_{rat} equilibrium.

It is less expensive to obtain a converged growth rate by extrapolating from lower values of v_{rat}. For instance, one could use the results from equilibria in the range $.94 \leq v_{rat} \leq .96$ to obtain a converged value. The result may not be as precise as if one used equilibria right up to $v_{rat} = .99$, but the difference will within about 5%.

Figure 4.8a shows the Fourier modes of the displacement eigenfunction ξ_ψ for the CIT equilibrium used in this study. Figure 4.8b shows ξ_ψ for a uniform vertical rigid shift of the CIT equilibrium. It is clear that these two forms of ξ_ψ are quite different, and therefore the eigenfunction differs significantly from a rigid vertical shift. In particular, the $m = 2,3$ components are nearly zero for the true eigenfunction, whereas there are significant $m = 2,3$ components needed to represent the rigid shift for this equilibrium. Figure 4.9 shows the projection of the displacement ξ onto the poloidal plane and indicates the motion of the unstable plasma. We can see how it differs from a vertical shift. The plasma displacement has a significant radial component superimposed on the vertical motion. The consequence is that the unstable plasma motion is toward the x-point in the lower half-plane. It is clear from Fig. 4.8 and 4.9, therefore, that the unstable motion of the CIT equilibrium is not particularly well represented by a uniform rigid shift. We shall examine these non-rigid components of the eigenfunction and how they interact with the passive conductors and active feedback in more detail in Chapter 6.

4.4 Additional studies in passive stabilization

4.4.1 Radial and vertical extension of CIT vacuum vessel wall

In this section we present results of a study of the vertical stability of CIT with respect to variations in the surrounding vacuum vessel. In particular, it is of interest to see how the growth rate changes with respect to moving the outboard section of the vacuum vessel outward, away from the plasma, and also how it changes with respect to an extension of the vacuum vessel vertically. This will give an indication of how well the given vacuum vessel design stabilizes the plasma, and how detrimental it would be to put the wall any farther from the plasma.

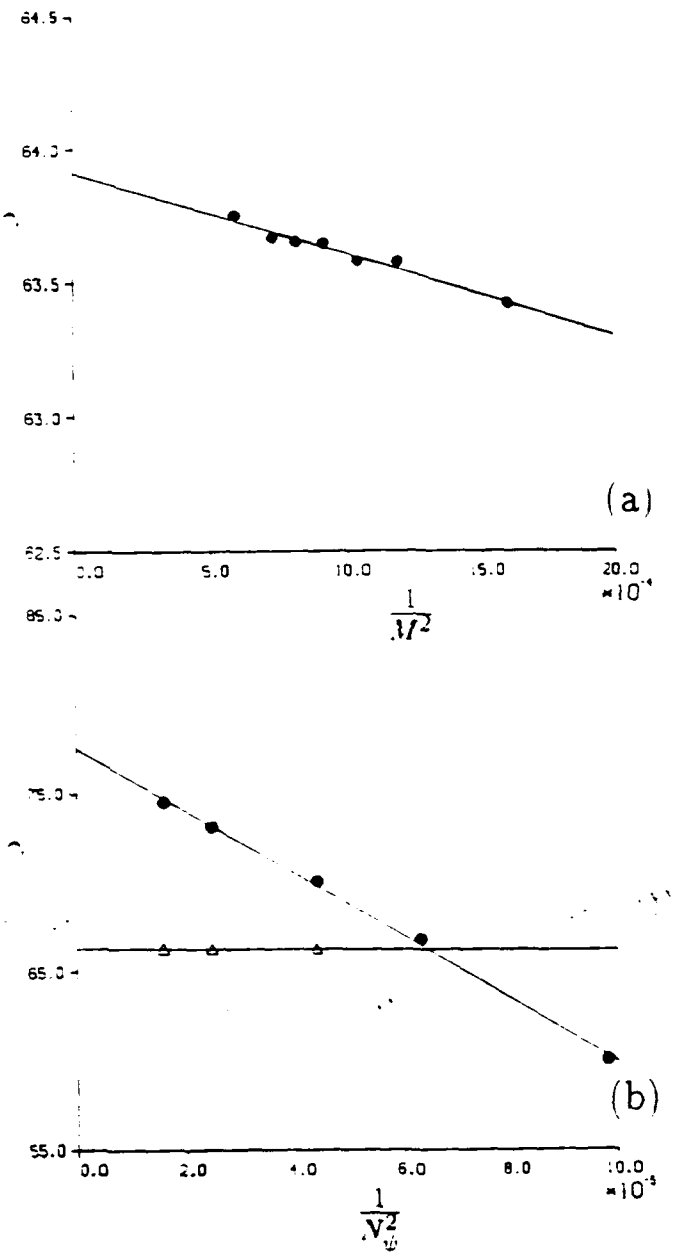


Figure 4.7: (a) *Growth rate convergence in Fourier harmonics for the CI ($\omega_{rat} = 0.95$) equilibrium.*

(b) *Growth rate convergence in the number of equilibrium surfaces N_ψ^2 for th $\omega_{rat} = 0.99$ (circles) and $\dot{\omega}_{rat} = 0.96$ (triangles) equilibria. Note that for this numb of surfaces ($N_\psi > 150$) the $\dot{\omega}_{rat} = .96$ is well converged, while the $\omega_{rat} = .99$ equili rium requires a far greater number of surfaces in the equilibrium calculation to obta*

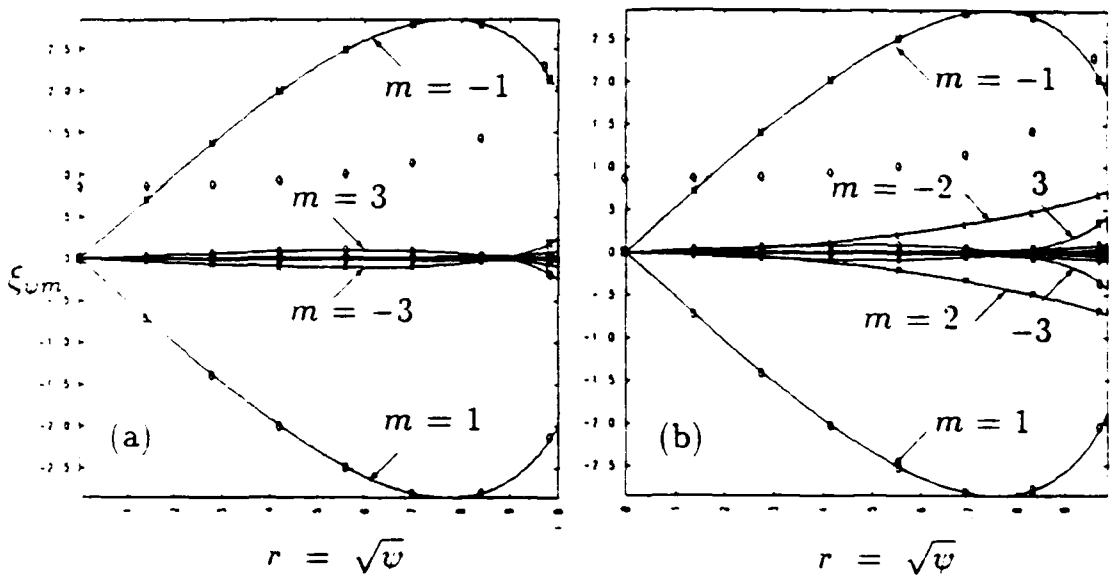


Figure 4.8: (a) *Fourier components of the radial displacement of the eigenfucntion* ξ_w *vs.* $\sqrt{\bar{w}}$ *for the CIT* ($v_{rot} = 0.95$) *equilibrium. The eigenfunction is dominated by the* $m = \pm 1$ *components. There is also a small contribution from the* $m = \pm 3$ *components, which have considerable variation in structure as a function of* r. (b) *Fourier components of a uniform rigid shift. The form of the* $m = 2, 3$ *components is clearly much different from that of the true eigenfunction.*

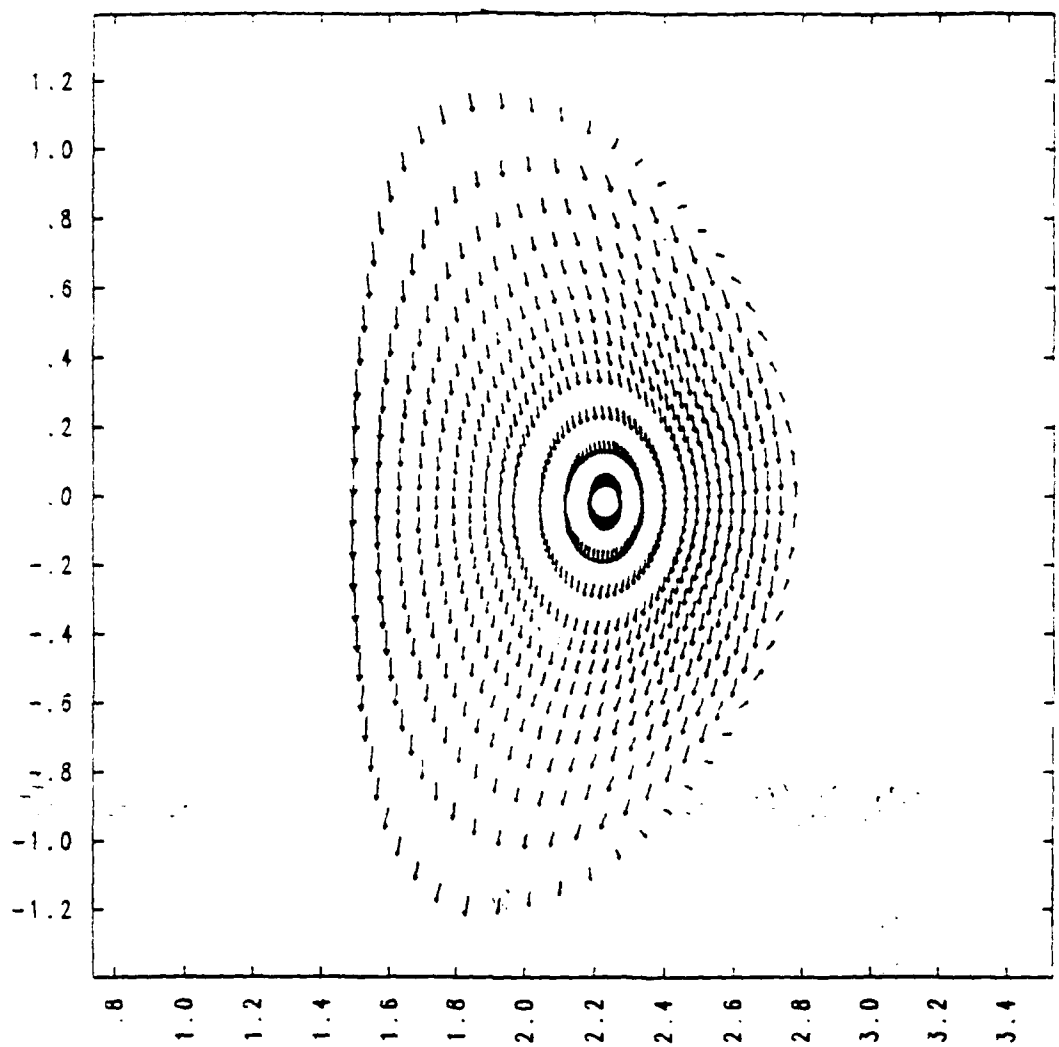


Figure 4.9: This plot shows the motion of the instability for the CIT equilibrium. Note how this motion varies from a rigid vertical shift. The true motion appears to be primarily vertical, but with some inward motion superimposed. The overall effect is that the motion is in the direction of the x-point, or the "corner" of the D-shaped plasma. This variation from the rigid vertical shift is shown in terms of the Fourier components in Fig. 4.8.

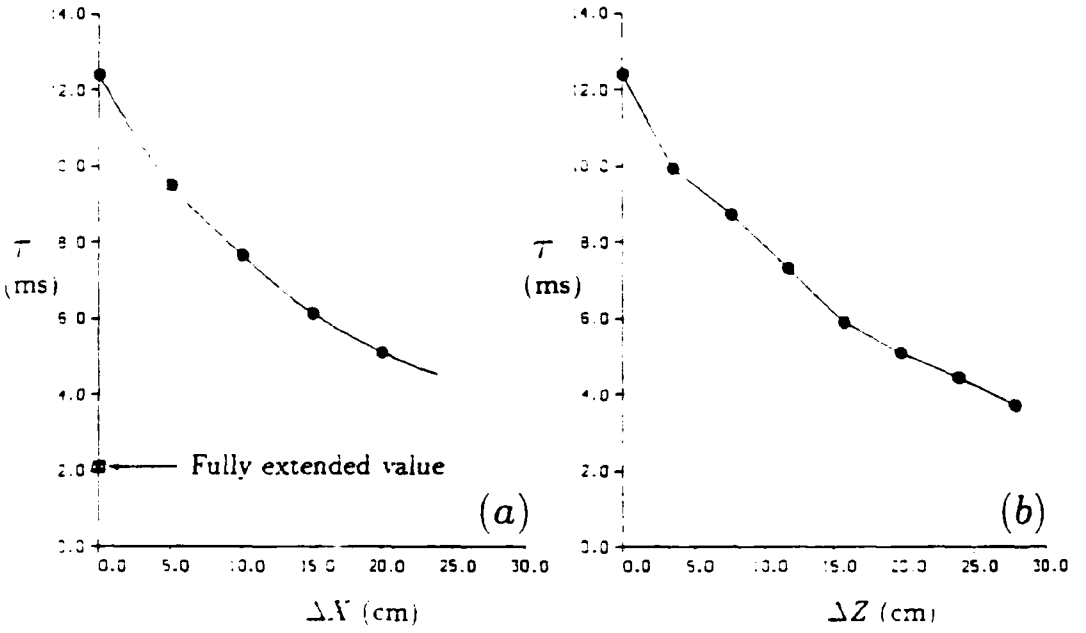


Figure 4.10: (a) *Growth time τ of the vertical instability for the CIT equilibrium with respect to the radial extension ΔX of the vacuum vessel structure at the points indicated by the arrows in Fig. 4.5. In addition, the limiting value for full extension is shown, whereby the top and bottom sections are extended and the outboard section is effectively removed.*

(b) *Growth time τ of the vertical instability for the CIT equilibrium with respect to the vertical extension ΔZ of the vacuum vessel structure at the midplane. The growth time τ is in milliseconds and the distance of extension ΔX, ΔZ is in cm.*

In the first case we take the CIT vacuum vessel and move the outboard region away from the plasma by adding wall elements in 5 cm increments at the points indicated by the arrows in Fig. 4.5a. The inboard section (everything to the right of the arrows) remains in its original position while the outboard section (to the left of the arrows) is extended outward by 5 cm increments.

The variation of the growth times τ with respect to the distance of radial extension ΔX is shown in Fig. 4.10a. The growth time decreases rapidly in the first 10 cm of extension, but beyond that, the relative decrease in τ is less. It appears that τ will approach a limit as the outboard section is extended horizontally. In order to

estimate this limiting value, we performed a calculation in which we extended the vacuum vessel horizontally outward from the indicated points in Fig. 4.5 by a very large distance ($\sim$ 1 m) and remove the completely. This yielded the indicated limiting value of $\sim$ 2 ms.

Next we consider a similar extension of the vacuum vessel, except this time in the vertical direction. We "break" the original vacuum vessel at the midplane and move both the upper and lower halves away from the midplane in increments of 3.5 cm. This extends both the inboard and outboard sections vertically, but the relative position of the inboard wall to the left of the plasma will change very little. However, the top, bottom, and outboard sections will move away from the plasma.

The results are shown by plotting the variation of the growth time τ versus the distance of vertical extension ΔZ in Fig. 4.10b. Again the growth time drops off rapidly for the first 5-10 cm of extension, but the relative change in τ vs. ΔZ decreases at larger values of ΔZ.

It appears that in both cases the plasma is stabilized well enough by the modified wall to keep from going ideally unstable. In Figure 4.11 we plot the current density in the wall J_ϕ and the wall surface current K_ϕ (which is equal to the current density J_ϕ multiplied by the wall thickness δ_w) versus the poloidal angle Θ around the wall circumference. The latter quantity is directly related to the jump in the normal derivative of the perturbed flux as in Eq. (3.37), and is therefore a measure of the stabilizing effect of the wall eddy currents. This calculation was done for the standard CIT vacuum vessel ($\Delta X = \Delta Z = 0$). The wall contour and the definition of the Θ-grid are shown in Fig. 4.5b. It is easy to see that there are large eddy currents in the outboard diagonal regions of the wall and that this will be destabilizing when ΔX or ΔZ is increased. But there are also large stabilizing currents (with nearly the same magnitude of K_ϕ) in the inboard regions of the wall at $\Theta \simeq 3\pi/4,\ 5\pi/4$. These currents will be somewhat less effective in stabilizing the plasma, because the radial field induced by currents on the inboard wall are smaller than those on the outboard wall, but the eddy currents are large enough to provide significant stabilization. This section of the wall remains close to the plasma in both of the wall extension scenarios described above. The eddy currents at the top and bottom of the wall are relatively small, as can be seen by the relative minimum in K_ϕ between the inboard and outboard regions. Therefore the effect of extending the wall radially and vertically as outlined above will increase the growth rate of the vertical instability,

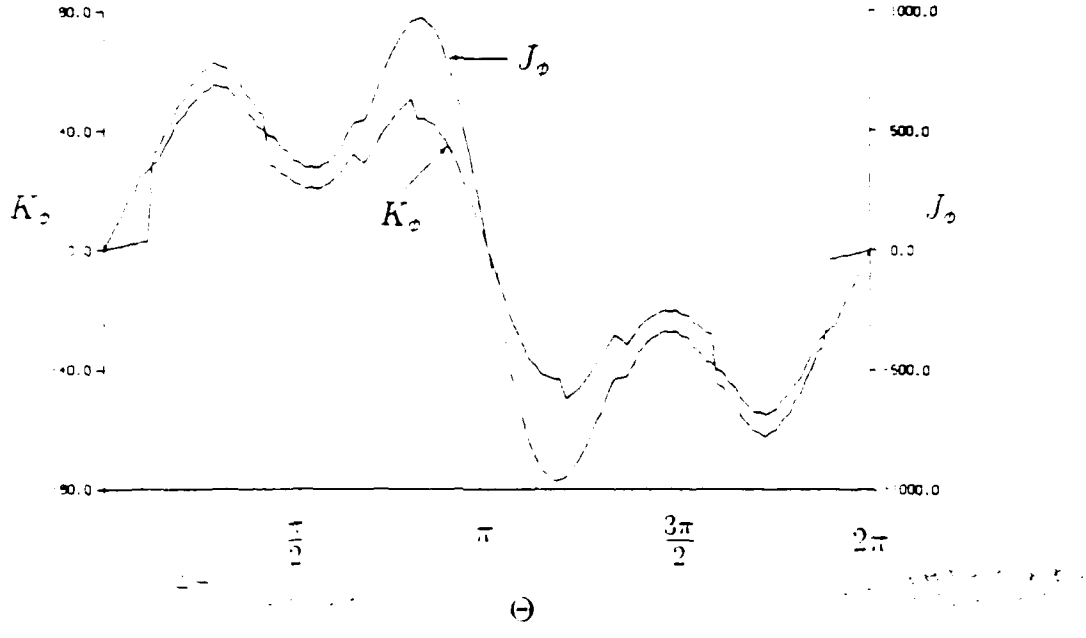


Figure 4.11: *Current density J_ϕ and Surface current density K_ϕ, where $K_\phi = \delta_w J_\phi$, in the CIT vacuum vessel wall. The wall contour and Θ-grid are shown in Fig. 4.5b. The relatively sharp variations in K_ϕ are due to sudden changes in the wall thickness δ_w at those points. The current density J_ϕ itself varies more smoothly. The diagonal outboard and the upper and lower inboard regions carry most of the eddy currents.*

Plasma Current I_p	11.30 MA
Major Radius R_0	7.245 m
Minor Radius a	1.602 m
Elongation $\kappa(95\%)$	1.613
Triangularity $\delta(95\%)$	0.430
Toroidal Field $B_T(0)$	11.6 T
$q(95\%)$	4.5
β	0.02
$n_e(0)$	3.51×10^{20} m^{-3}

Table 4.2: *Equilibrium parameters of ARIES-I plasma used in the passive stabilization study.*

but not rapidly enough to reach the ideal time scale because there are still large stabilizing wall eddy currents near the plasma. However, the growth rates are fast enough for these larger separations that active feedback control may be difficult.

It is interesting to note in Fig. 4.11 that the surface eddy currents in the far outboard region of the wall contour ($\Theta \simeq 0, 2\pi$) are very small compared with the magnitudes in the other wall regions. Therefore, we were justified in modifying this extended outboard region as described earlier, because this region has very little effect on the plasma.

4.4.2 Passive stabilization of the ARIES-I reactor design

The NOVA-W code has been used in an analysis of the passive vertical stability characteristics of the ARIES-I reactor [70] design. The ARIES-I tokamak reactor is a low-β, low-current, high-field, moderate-aspect-ratio, steady-state design intended to minimize the cost of electricity by minimizing plasma current and associated current-drive cost. Because of the relatively high shaping of the cross section and the large amount of separation between the plasma and the conducting wall required to include the blanket and first wall in such a reactor design, the vertical stability is an important consideration. The parameters of the ARIES-I equilibrium used in this calculation are listed in Table 4.2, and the poloidal flux contours are shown in Fig. 4.12a.

A comparison between the calculated vertical instability growth rate from NOVA-W

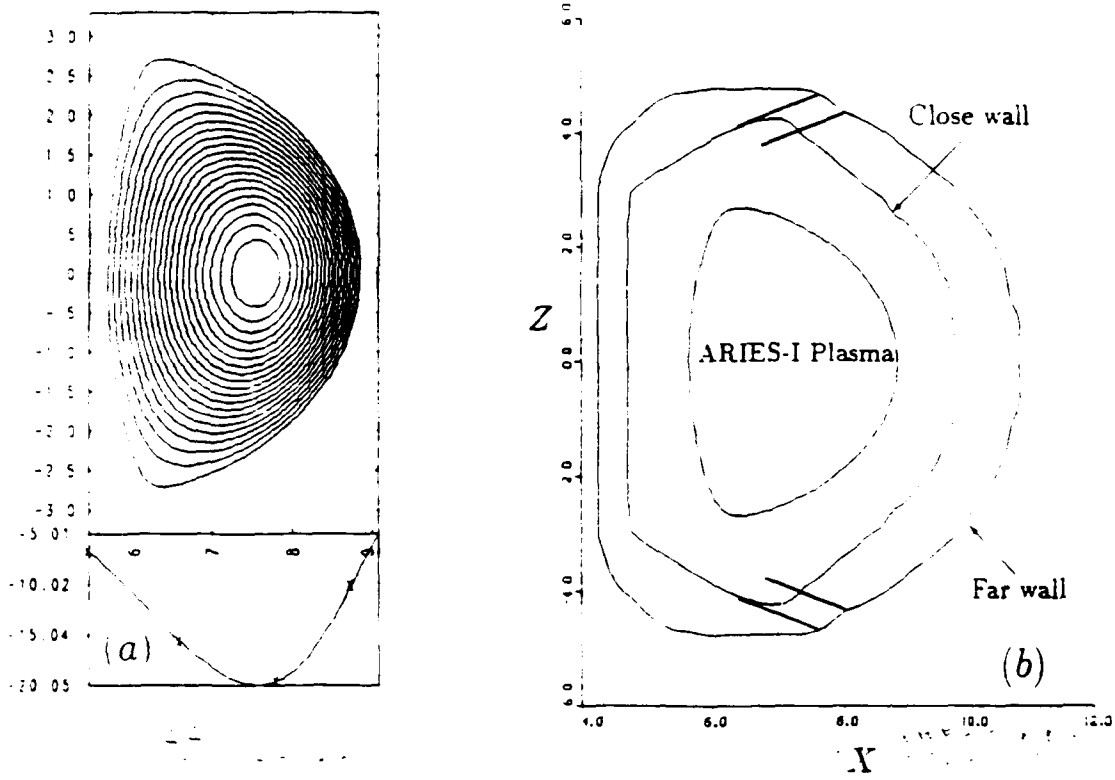


Figure 4.12: (a) *Equilibrium poloidal flux contours (from TSC) of the ARIES-I equilibrium used in this calculation. The poloidal flux through the midplane is shown at the bottom.*

(b) *The ARIES-I plasma surface and the two basis wall contours used for the interpolations described in Sec. 4.4.3. The "close" wall lies between the blanket and the neutron shield, while the "far" vacuum vessel wall would be outside both the blanket and the shield.*

and the converged growth rate from TSC has been performed for a preliminary version of the ARIES-I equilibrium and vacuum vessel [71], and the results agree to within 2%. We consider here the final version of the equilibrium and the two vacuum vessel designs that were under consideration. The first vacuum vessel (close) wall design provides better stabilization since it is relatively close to the plasma. This vessel wall would lie between the blanket and the neutron shield. The second wall version (far) would lie outside both the blanket and the shield. The second is much farther away and allows room for the blanket. The growth time for the inner wall version is found to be 330 ± 5 ms, whereas for the outer wall version the growth time is 0.67 ± 0.05 ms. The outer wall version is distant enough from the plasma that the plasma has a very fast growth rate and, in fact, is near ideal instability (instability with an ideally conducting wall). In the next section we will consider the stability of the ARIES-I plasma as the wall is moved progressively farther away from the plasma.

Figure 4.13 shows the form of the displacement eigenfunction ξ_v compared to the form of a uniform rigid vertical shift for the ARIES-I equilibrium ($\omega_{rat} = 0.95$) used in this study. The differences between Figures 4.13a and 4.13b show that this highly shaped plasma equilibrium has an unstable eigenfunction with considerable non-rigid structure. As in the case of the CIT equilibrium in Section 4.3.1, the true eigenfunction of the ARIES equilibrium has $m = 2,3$ components that are much different from those required to represent a rigid vertical shift. As in that case, the true plasma motion is in the direction of the x-point. Clearly, then, the vertically unstable motion of the ARIES-I plasma is not well represented by a rigid shift.

4.4.3 Extension of the vacuum vessel wall for the ARIES-I plasma

As an additional study we consider the variation of the vertical instability growth rate of the ARIES-I plasma with respect to the uniform radial extension of the vacuum vessel wall. We select modifications of the close and far wall versions of the previous section and interpolate linearly between these contours. The wall contours have been simplified by removing the extensions for the divertor structure (as shown in Fig. 4.12b), thereby making the contours much more continuous. The two basis contours, between which we will interpolate, are shown in Fig. 4.12b.

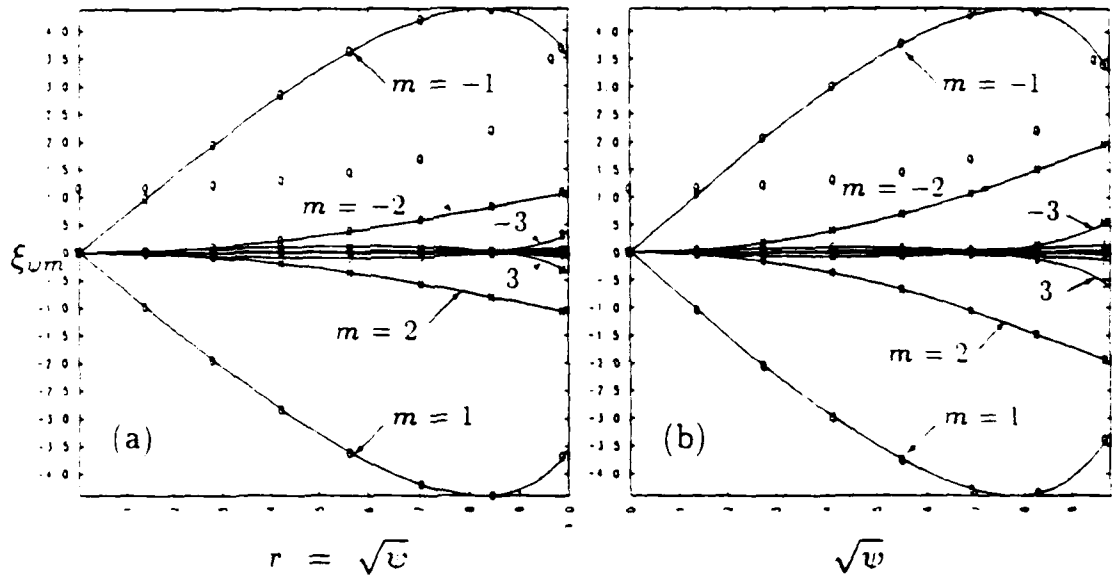


Figure 4.13: *Fourier components of the radial displacement ξ_v vs. $\sqrt{v}$ for the ARIES-I equilibrium ($v_{rat} = 0.95$) used in Sections 4.4.2 and 4.4.3. There are considerable non-rigid contributions to the eigenfunction for this highly shaped plasma. The magnitudes of the $m = \pm 2$ components at the plasma edge are almost 1/3 of the magnitude of the $m = \pm 1$ components.*

If the radial position of the wall coordinates is described by

$$r_{close} = \sqrt{(R_{wall}^{close} - R_{plasma})^2 + (Z_{wall}^{close} - Z_{plasma})^2} \tag{4.19}$$

and

$$r_{far} = \sqrt{(R_{wall}^{far} - R_{plasma})^2 + (Z_{wall}^{far} - Z_{plasma})^2} \tag{4.20}$$

for the modified close and far wall versions, respectively, the newly interpolated wall contour can be described by

$$r_{new} = r_{close} - f(r_{far} - r_{close}) \tag{4.21}$$

where f is the interpolation parameter, which varies between 0 and 1.

The variation in the resistive wall growth rate with respect to the interpolation parameter f is shown in Fig. 4.14. The growth rate is seen to increase rapidly as the wall approaches the ideal stability limit (limit of stability with an ideally conducting wall). The plasma surrounded by the wall calculated with $f = 0.7$ is found to be ideally unstable. This is a good demonstration of the unique ability of the NOVA-W code to calculate accurately the vertical instability growth rate of a deformable tokamak plasma right up to the ideal stability limit.

4.5 Summary

In this chapter we have seen that the NOVA-W code can provide accurate growth rates for vertically unstable equilibria partially stabilized by passive conductors. Comparisons against a simple analytic model and numerical results for a realistic configuration prove that NOVA-W produces reliable results. The eigenfunctions for the CIT and ARIES-I equilibria demonstrate that for these highly shaped plasmas, the unstable motion is not well represented by a rigid vertical shift. Analysis of such non-rigid motion is ideally suited to the NOVA-W code.

The performance of NOVA-W for producing these results far exceeds that of TSC. To obtain an accurate converged result from TSC one must perform several (typically three) runs at progressively smaller time steps. Each successive run, therefore, takes longer to complete by a factor equal to the time step reduction. The longest runs require 150–200 CRAY-2 minutes. The NOVA-W code, on the other hand, requires at most 20–25 minutes of CRAY-2 time for the first point in the convergence in ψ_{res}.

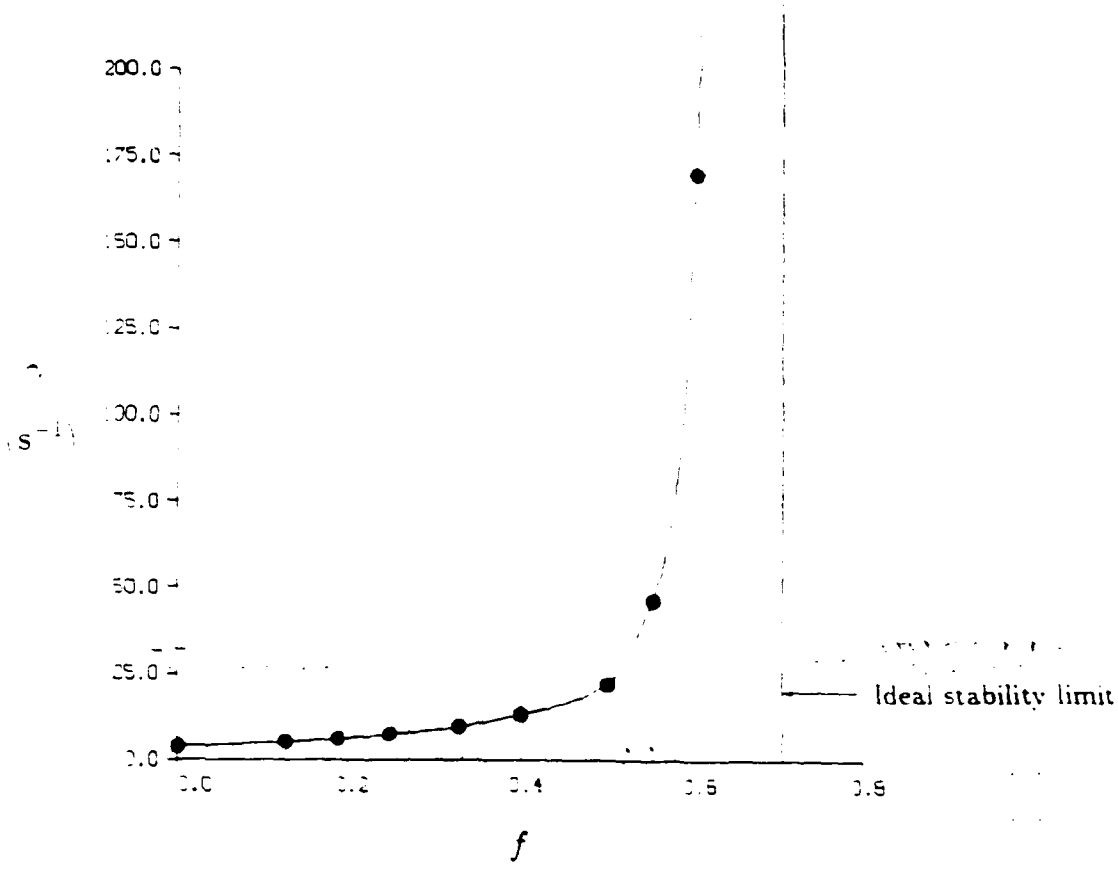


Figure 4.14: *Growth rates for ARIES-I equilibrium vs. the wall interpolation (radial separation) parameter f. The growth rates increase rapidly as f approaches the ideal stability limit.*

Each additional point in the convergence (typically three points are needed) requires a *shorter* run (10–15 minutes) because one has a good initial guess for the eigenvalue from the previous run.

The performance of NOVA-W truly excels for a parameter scan such as the variations in the vacuum vessel wall in Sections 4.4.1 and 4.4.3. In such a scan, the results from the previous point allow one to make a good initial guess for the next point, which speeds up the process considerably. If one is using TSC, on the other hand, the same time consuming convergence procedure must be followed for each point in the scan, because the calculation procedure does not change based on previous results. This leads to an improvement in performance of NOVA-W over TSC of as much as a factor of 10 or more in such calculations.

Chapter 5

Active feedback calculations

In this chapter we present results of active feedback calculations using the NOVA-W code. Following a description of the numerical procedure involved in the calculation, we apply the NOVA-W formulation to the active feedback of the vertically unstable CIT plasma.

We consider the stability of the feedback stabilized CIT plasma, with regard to various placements of the flux loops around the plasma within the vacuum vessel. A similar study [72] has been performed using TSC, allowing us to compare our results, at least qualitatively, with those obtained by TSC. In this chapter we focus on the basics of calculating stability with an active feedback system. We also point out the improved performance of the NOVA-W method over other numerical methods. In the following chapter we examine how the feedback system can affect the non-rigid components of the eigenfunction and thereby degrade the effectiveness of the feedback in certain situations.

5.1 The numerical procedure

The addition of active feedback to the stability calculation is a non-ideal effect, destroying the self-adjointness of the system of equations. With feedback included, we no longer expect an eigenvalue that is purely real, corresponding to purely oscillatory motion, or purely imaginary, corresponding to purely growing or decaying exponential motion.

For the case of a wall stabilized plasma in the presence of passive resistive con-

ductors, we found that the solution of interest has a purely imaginary eigenvalue that corresponds to exponential growth of the instability on the resistive time scale of the surrounding conductors. This root does not exist in the absence of resistivity. With a perfectly conducting wall that stabilizes the ideal instability, there are two stable roots which are purely real (oscillatory). With the addition of resistivity these roots remain oscillatory but become damped. However, a third root, the unstable root of interest, arises from the origin and moves along the imaginary axis. With the addition of feedback, however, we cannot expect this root to remain purely imaginary.

The contribution of the active feedback is calculated through constructing the feedback matrices derived in Sec. 3.3. The form of these matrices depends on the location of the flux loops—inside the resistive wall (Case A of Sec. 3.3.1) or outside the wall (Case B of Sec. 3.3.1). The effect of the dynamics of the active feedback circuit can be included as derived in Sec. 3.3.2.

The procedure for the calculation of stability with active feedback begins by determining the eigenvalues of the system with the feedback gain set to zero, as described in Sec. 4.1. One then increases the feedback gain from zero and tracks the motion of the eigenvalue in the complex plane. The eigenvalue is calculated using Muller's method (see, for example, Ref. [73]), for determining complex zeros of analytic functions in the complex plane. Since Muller's method assumes the function is a smoothly varying function of ω and that one begins with a good initial guess, it works most reliably when we increment the gain in relatively small steps so that the initial guess is always close to the actual solution. In practice, however, the dispersion function $D(\omega)$ (resulting from Eqs. 3.12 and 3.13 after being decomposed into poloidal harmonics and finite elements), is often a well behaved and smoothly varying function for this problem, and therefore one can often make reasonably large increments of gain and successfully find the root, even with a poor initial guess. However, this is not always the case.

It is difficult to track the root after it moves into the lower half of the complex plane (region of stability). Whereas there is usually only one eigenvalue of interest on the unstable side of the axis, there are always many stable roots that exist which have no significance for the vertical stability problem. Therefore, one must follow the procedure of making small increments in gain and carefully tracking the root in this region of the complex plane. Furthermore, one must check the form of the eigenfunction to be sure that it corresponds to the correct root.

While we must, in general, look for a complex root, in practice the root is often purely imaginary, as in the case with no active feedback. In particular, when the gain is relatively low and the root is far from the stability boundary (on the unstable side) the root is most often purely imaginary, corresponding to a (partially stabilized) exponentially growing instability.

In many cases, however, a real part to the eigenvalue does develop. This corresponds to some oscillation in the vertical motion which can be detrimental to the effectiveness of the feedback system. Even if the imaginary part of the eigenvalue is negative (stable), a large real (oscillatory) piece indicates poor control. This oscillation can usually be reduced by introducing or increasing the derivative gain $\mathcal{J}_g$ in the gain law for the feedback currents:

$$I_{fb} = \alpha_g(\chi_{o1} - \chi_{o2}) - \mathcal{J}_g(\dot{\chi}_{o1} - \dot{\chi}_{o2}). \tag{5.1}$$

The oscillation is due to an "overshoot" in the plasma motion. In a linear system, if the slope of the response curve to a step input is large, a large overshoot will be the result [54]. The feedback system pulls strongly, trying to bring the vertical position to the desired value. The motion goes beyond the desired point, and the feedback reverses to try to bring it back. An oscillatory motion about the desired position is the result, which may be damped (stable) or growing. The derivative term measures the instantaneous slope, and predicts and corrects for the overshoot before it happens. In the following section we shall see an example of a configuration with large oscillations in the feedback response and how increasing the derivative gain improves the feedback control.

Consider a simple example of a model feedback system to illustrate the effect of derivative gain on the feedback response. We will use the straightforward circuit analysis considered by Jardin and Larrabee [38]. The equation of motion for a plasma under the force from an external field gradient (which drives the instability) and the active feedback coils is given by

$$m\ddot{z} = I_p \sum_{i=1}^{N} M'_{ip} I_i + I_p 2\pi R \frac{\partial B_r}{\partial z} z, \tag{5.2}$$

where m is the plasma mass, I_p is the plasma current, $M_{ip}(z)$ is the inductive coupling between the plasma and the feedback coil, I_i is the feedback current in the i^{th} coil, B_r is the radial component of the equilibrium poloidal field, the prime denotes the

derivative with respect to z, and $\dot{z}$ represents the first time derivative of z, and so on. The sum is over the N coils. The circuit equations for the active feedback coils are given by

$$L_i \frac{dI_i}{dt} - r_i I_i - \sum_{i \neq j}^{N} M_{ij} \frac{dI_j}{dt} - M'_{ip} I_p \dot{z} - \alpha_v z - \beta_v \dot{z} = 0. \tag{5.3}$$

where L_i is the self-inductance of the i^{th} coil, r_i is its resistance, and M_{ij} is the mutual inductance between coils i and j. The last two terms in Eq. 5.3 correspond to the feedback voltage applied for a displacement in z from the desired value: α_v is the proportional derivative term (voltage applied to the active feedback coils per unit length displacement of the plasma from the midplane), and β_v is the derivative gain term. Note that this is a circuit equation, so the feedback terms are now voltages.

For simplicity we consider one active feedback coil, and a single passive conductor. The latter represents the effects of the vacuum vessel, etc. We assume a time response that goes as $z(t) = z_0 \exp(\gamma t)$. This gives

$$\gamma L_a I_a - r_a I_a - \gamma M_{aw} I_w - \gamma M'_{ap} I_p z - (\alpha_v - \gamma \beta_v) z = 0 \tag{5.4}$$

$$\gamma L_w I_w - r_w I_w - \gamma M_{aw} I_a - \gamma M'_{wp} I_p z = 0. \tag{5.5}$$

Here the 'a' subscript denotes the active coil, the 'w' subscript denotes the passive (wall) conductor, and the 'p' denotes the plasma. We substitute these equations into the equation of motion, Eq. (5.2), and ignore the inertia term $m\ddot{z}$ because we are interested in the roots that correspond to a plasma stabilized (or nearly stabilized) with active feedback. This eliminates the roots that correspond to the ideal time scale. These roots are stable (damped and oscillatory) as we discussed before. We also rewrite our equation in terms of the following parameters:

$$\tau_a = \frac{L_a}{r_a}, \qquad \tau_w = \frac{L_w}{r_w}, \qquad \delta_{aw} = \frac{M_{aw}}{\sqrt{L_a L_w}}. \tag{5.6a}$$

$$\Omega_a^2 = \frac{(M'_{ap})^2 I_p}{k L_a}, \qquad \Omega_w^2 = \frac{(M'_{wp})^2 I_p}{k L_w}. \tag{5.6b}$$

$$G_\alpha^2 = \frac{\alpha_v^2 L_a}{r_a^2 I_p k}, \qquad G_\beta^2 = \frac{\beta_v^2 L_a}{r_a^2 I_p k}, \qquad k = 2\pi R \frac{\partial B_z}{\partial z}. \tag{5.6c}$$

Upon substituting the exponential time behavior, we obtain a quadratic equation for γ:

$$A\gamma^2 - B\gamma - C = 0. \tag{5.7}$$

where

$$A = (\delta_{aw}^2 - 1) + (\Omega_w^2 - \Omega_a^2) - 2\delta_{aw}\Omega_a\Omega_w - G_\beta\frac{\Omega_a}{\tau_a}(\frac{\Omega_w}{\Omega_a}\delta_{aw} - 1) \tag{5.8}$$

$$B = \frac{\tau_a}{\tau_w}(\Omega_a^2 - 1) - (\Omega_w^2 - 1) - G_\alpha(\Omega_a - \Omega_w\delta_{aw}) - \frac{\Omega_a G_\beta}{\tau_w} \tag{5.9}$$

$$C = \frac{\tau_a}{\tau_w}(\Omega_a G_\alpha - 1). \tag{5.10}$$

The solution is $\gamma = (-B \pm \sqrt{B^2 - 4AC})/2A$. To have a stable solution we require $A > 0$, $B > 0$, and $C > 0$. The latter gives us the constraint

$$G_\alpha > \frac{1}{\Omega_a} \tag{5.11}$$

or,

$$\alpha_v > r_a\frac{k}{M'_{ap}}. \tag{5.12}$$

This just translates to the necessity that the proportional feedback gain must be large enough to exceed a certain critical quantity in order for the active feedback to successfully stabilize the motion. This critical quantity consists of three terms. The necessary gain α_v is proportional to k which is the destabilizing force on the plasma normalized to the plasma current (see Eq. 5.2). It is inversely proportional to M'_{ap}, which is a measure of the force (normalized to the plasma current) exerted on the plasma by a unit current in the active feedback coil (see Eq. 5.2). Finally, the proportional gain for the coil voltage must multiply the required current by r_a.

The requirement that $B > 0$ for a stable solution gives a constraint on the value of the derivative gain

$$G_\beta > \frac{\tau_a}{\Omega_a}(1 - \Omega_a^2) - \frac{\tau_w}{\Omega_a}(1 - \Omega_w^2) - G_\alpha\tau_w(\frac{\Omega_w}{\Omega_a}\delta_{aw} - 1). \tag{5.13}$$

We can make some sense of this requirement if we consider the meaning of the various terms. The term Ω_w (defined in Eq. (5.6b)) is essentially the ratio of the stabilizing force of the passive conductor (wall) to the driving force of the instability k. If the passive conductor is at all effective in stabilizing the plasma then $\Omega_w > 1$. Therefore the term $1 - \Omega_w^2$ will be negative. The term Ω_a is the same ratio for the active feedback coil. Since the active coil may or may not have any significant passive stabilization effect, the value of Ω_a could be greater or less than 1.0, respectively. The terms τ_w and τ_a are the L/r times for the passive and active feedback conductors, respectively,

as defined in Eq. (5.6a). The term G_a is the proportional gain term defined in Eq. (5.6c), δ_{aw} is the mutual interaction between the active and passive conductors. It represents the shielding of the active feedback by the passive conductor (wall).

The first term on the right hand side of the inequality in Eq. (5.13) will be small and negative if the active coil has any significant passive stabilization effect, or else small and positive if it does not. The second term will be negative. Its magnitude represents how effective the wall is in passively stabilizing the plasma. The third term, which is normally positive since we can expect $\Omega_w/\Omega_a \gg 1$, is the term. ... represents the effects that may require a nonzero derivative gain. If there is significant shielding of the active feedback by the passive conductor, then δ_{aw} will be large and that will increase the need for derivative gain. If the passive effect of the conductors is small, then the first two terms will have a smaller negative magnitude, and that will also increase the need for a derivative gain. Also the third term is multiplied by the proportional gain term, so that the higher the proportional gain used, the higher the derivative gain needed. In fact, we usually quantify the derivative gain by the ratio of the derivative gain to the proportional gain (this ratio has units of s^{-1}). These terms combine to show how much derivative gain is required. If the right hand side of Eq. (5.13) is less than zero, then the system is stable with no derivative gain, given sufficient proportional gain.

The final constraint is that $A > 0$. We can divide this constraint into two parts. First, for there to be sufficient passive stabilization so that the mode is stable on the ideal time scale, we have the constraint

$$(\delta_{aw}^2 - 1) - (\Omega_w^2 - \Omega_a^2) - 2\delta_{aw}\Omega_a\Omega_w > 0. \tag{5.14}$$

The derivative gain must also satisfy an additional constraint in order to satisfy $A > 0$:

$$G_\beta \frac{\Omega_a}{\tau_a}(1 - \frac{\Omega_w}{\Omega_a}\delta_{aw}) > (\delta_{aw}^2 - 1) - (\Omega_w^2 - \Omega_a^2) - 2\delta_{aw}\Omega_a\Omega_w \tag{5.15}$$

When discussing the minimum allowable derivative gain G_β, we assumed that $\delta_{aw}\Omega_w/\Omega_a$ 1. If it turns out that $\delta_{aw}\Omega_w/\Omega_a < 1$, then Eq. 5.15 sets a limit on the maximum derivative gain allowable for stability. In that case there would be no minimum (positive) derivative gain required for stability according to Eq. 5.13.

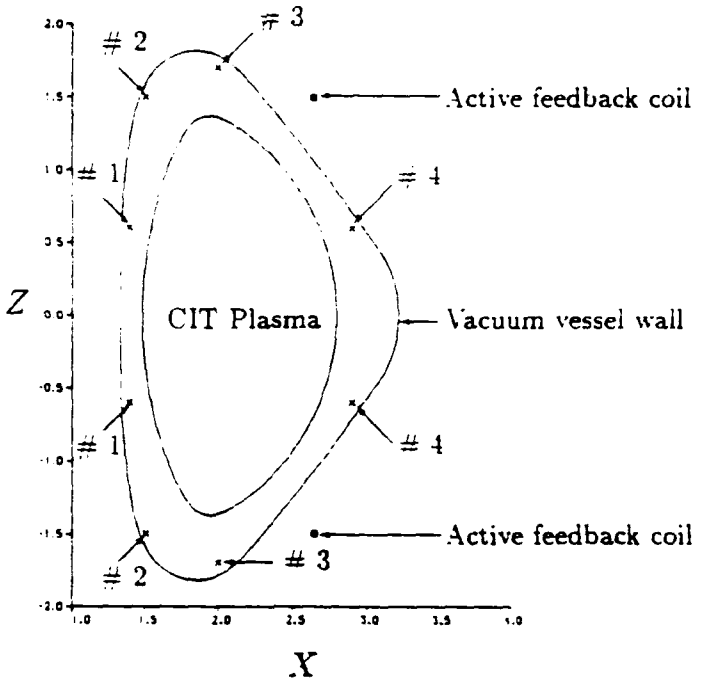


Figure 5.1: *CIT plasma, resistive wall, and flux-loop pair locations.*

5.2 Feedback Stabilization of the CIT plasma

We use the CIT equilibrium, vacuum vessel, and feedback coils introduced in Sec. 4.3. We compare the stability of this system when different locations of the flux pickup loops are used. Figure 5.1 shows the CIT plasma, resistive wall, active coils and the four locations for the flux-loop observation pairs used in this study.

Figure 5.2 shows the results of the active feedback system using flux measurements from the four flux-loop pair positions shown in Fig. 5.1. The growth rate is plotted against proportional gain for the positions and values of derivative gain as labeled. The curve labeled (a) shows the γ vs. α_g for flux-loop pair #1 located at $(X_o, Z_o) = (1.295, \pm 0.65)$. The derivative gain is $\beta_g/\alpha_g = 0.01 s^{-1}$. This is clearly the most effective flux-loop pair considered here. Indeed, it corresponds to the most effective region for flux measurements in the TSC calculations performed for the same problem [72].

The perturbed poloidal flux is the difference between the total poloidal flux of the displaced equilibrium and the equilibrium poloidal flux (which is symmetric about the midplane for symmetric plasmas and conductor configurations). The perturbed

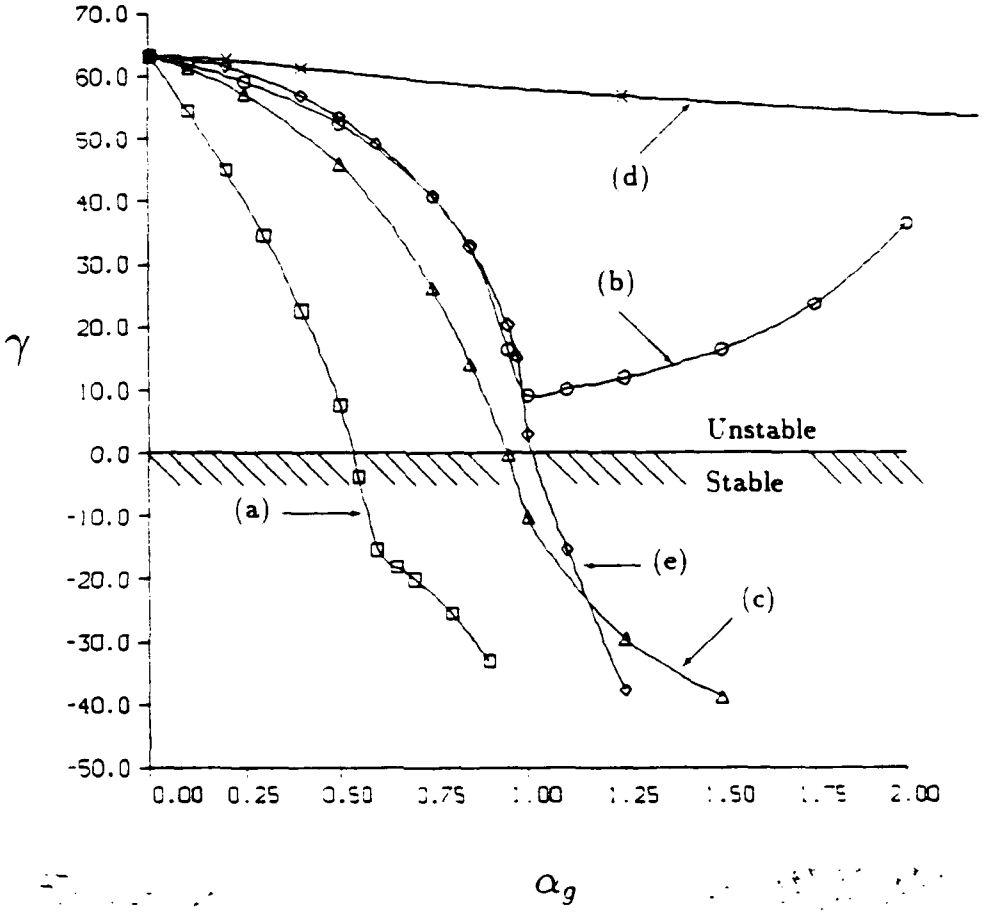


Figure 5.2: *Growth rates vs. gain for various flux loop locations on CIT. The flux looop locations refers to the labels shown on Fig. 5.1.*

(a) squares, flux-loop position #1. $\beta_g/\alpha_g = 0.01s^{-1}$.
(b) circles, flux-loop position #2. $\beta_g/\alpha_g = 0.01s^{-1}$.
(c) triangles, flux-loop position #2, $\beta_g/\alpha_g = 0.02s^{-1}$,
(d) crosses. flux-loop position #3. $\beta_g/\alpha_g = 0.02s^{-1}$.
(e) diamonds, flux-loop position #4, $\beta_g/\alpha_g = 0.01s^{-1}$.

flux then is essentially the asymmetric flux with respect to the midplane. This is composed of three parts: the part due to the displaced plasma, the component from the currents in the feedback coils, and that from the eddy currents in the passive conductors.

We use the difference of the perturbed flux between two observation points (flux observation loops) symmetric about the midplane. This is equal to the difference in the total poloidal flux between those two points, and it is used as measure of the vertical position of the plasma. This is a standard way of determining position from magnetics measurements in experiments. An effective location is one that is sensitive to the perturbed flux contribution from the plasma with respect to a vertical displacement. Therefore a plot of perturbed flux contours in the vacuum region is a useful device with which to see how effective is a particular pair of flux loops in determining plasma vertical position. From that we see how effective a pair a flux loops would be as part of the overall feedback system.

Figure 5.3 shows the perturbed flux contours in the vacuum region for the eigenfunction of the active feedback stabilization of the CIT equilibrium with $\alpha_g = 0.5$ using flux measurements at flux-loop position $\#1$. One can see that these flux loops lie adjacent to perturbed plasma flux contours of relatively high flux. In terms of the ability of the flux loops to provide vertical displacement good measurements for the feedback system. Fig. 5.3 shows that the region including flux-loop position $\#1$ and extending slightly higher along the inboard wall has the highest perturbed flux values, and is therefore probably the best region to place the flux loops. The perturbed flux along the wall then decreases in magnitude as one moves the flux loops further from the midplane and toward the outboard side on the wall. One would expect, therefore, that by placing the flux loops at position $\#2$ (see Fig. 5.1) the performance would be degraded somewhat, the reason being that the flux loops are now in a region with much smaller perturbed flux for a given vertical displacement. Therefore, the flux loops are less sensitive to the vertical motion of the plasma. These loops lie just outside the outermost (lowest magnitude) plasma flux contour, although they are still far from the zero-contour.

The curves marked (b) and (c) in Fig. 5.2 show the results for active feedback at flux-loop position #2, $(X_o, Z_o) = (1.51, \pm 1.50)$. Figure 5.4 shows these two curves separately. Figure 5.4a shows γ vs. α_g, and Fig. 5.4b shows γ vs. the frequency of oscillation ω_r. It is seen that the case with lower derivative gain ($\beta_g/\alpha_g = 0.01 s^{-1}$)

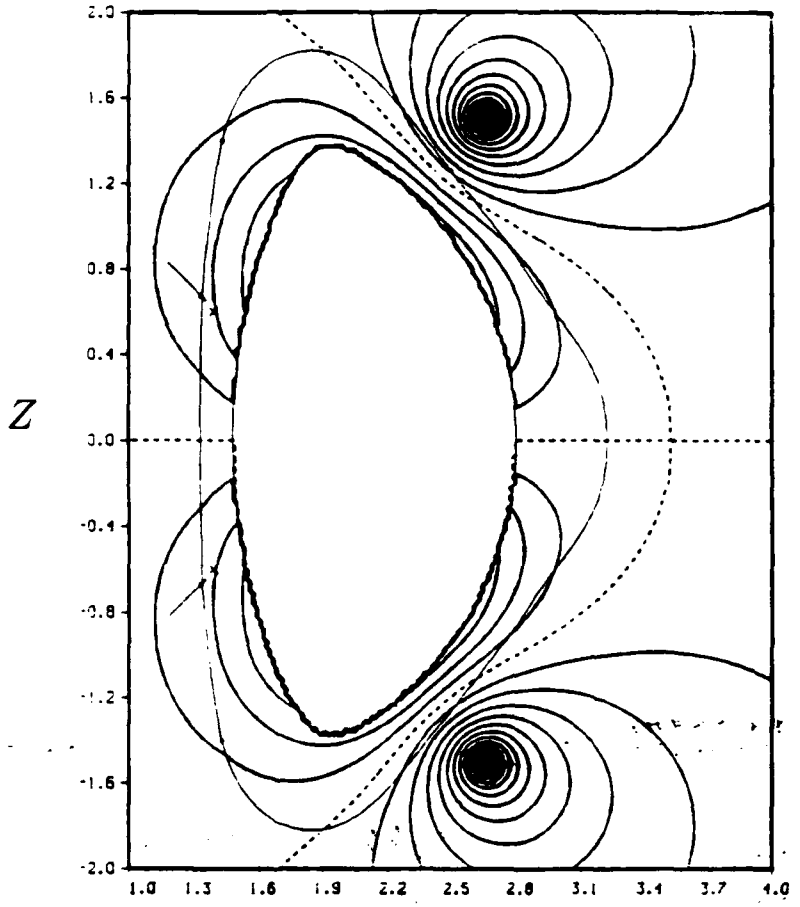


Figure 5.3: *Perturbed flux contours in the vacuum region for active feedback in which the active feedback systems uses flux measurements from flux-loop pair #1. The flux loops of pair #1 are represented by ' × ' symbols which are indicated by the arrows. The zero-flux contour is shown as a dashed line. The flux loops in this case lie on a contour of large plasma flux, and are far from the zero-flux contour.*

is always unstable. The growth rate decreases with increasing gain until a turning point is reached at $\alpha_g \approx 1.0$, at which point an increase in the gain no longer reduces the growth rate, but actually *increases* the growth rate at still higher gain. It is seen from Fig. 5.4 that at this turning point the value of ω_r increases rapidly from zero with increasing feedback gain. There is, therefore, an overshoot that begins approximately at the gain value where there is a turning point in γ. In this case, once the gain value reaches a certain threshold, the feedback system drives this overshoot instead of reducing the growth rate. A further increase in gain increases the oscillation frequency caused by the overshoot and actually has the effect of increasing the growth rate as well. The second curve—(c) from Fig. 5.2—shows the results of doubling the derivative gain to $\beta_g/\alpha_g = 0.02s^{-1}$. In this case we see that the growth rate continues to decrease smoothly and becomes stable. This case with, larger derivative gain, shows virtually no oscillation until well after the mode has been completely stabilized. The last point shows a small oscillation frequency with a large damping rate (negative growth rate) for the displacement. This is a good example of how increasing the derivative gain will reduce the oscillations and improve the overall performance of the feedback system. The reduction in the efficacy of the feedback system as the flux loops are moved to a less sensitive position higher on the inboard wall from position =1 to position =2 agrees well with TSC results [72].

Figure 5.2, curve (d), shows the results for active feedback with measurements taken at flux-loop position =3, $(X_o, Z_o) = (2.0, \pm 1.7)$. It can be seen that the plasma is far from being stabilized regardless of the value of the gain. There is only a small decrease in the growth rate with increasing gain. The curve shown is for $\beta_g/\alpha_g = 0.02s^{-1}$ derivative gain, but the curve is nearly identical for $\beta_g/\alpha_g = 0.05s^{-1}$. In fact, the oscillation frequency ω_r for both cases is nearly zero. Therefore the problem in stabilizing the mode is not due to overshoot and oscillation, and increased derivative gain will not help. This agrees with results from TSC simulations [72], in which there were no gain combinations α_g, β_g found that could come anywhere close to stabilizing the mode. Using this flux-loop location is apparently completely ineffective.

Figure 5.5 gives an explanation of why this position for the flux loops is unable to provide adequate feedback stabilization. This figure shows the perturbed flux contours for the active feedback with the flux loops at this position. It shows the contours of zero flux lying very close to the flux loops used for position measurement. The value of perturbed poloidal flux at these flux loops is very nearly zero. These

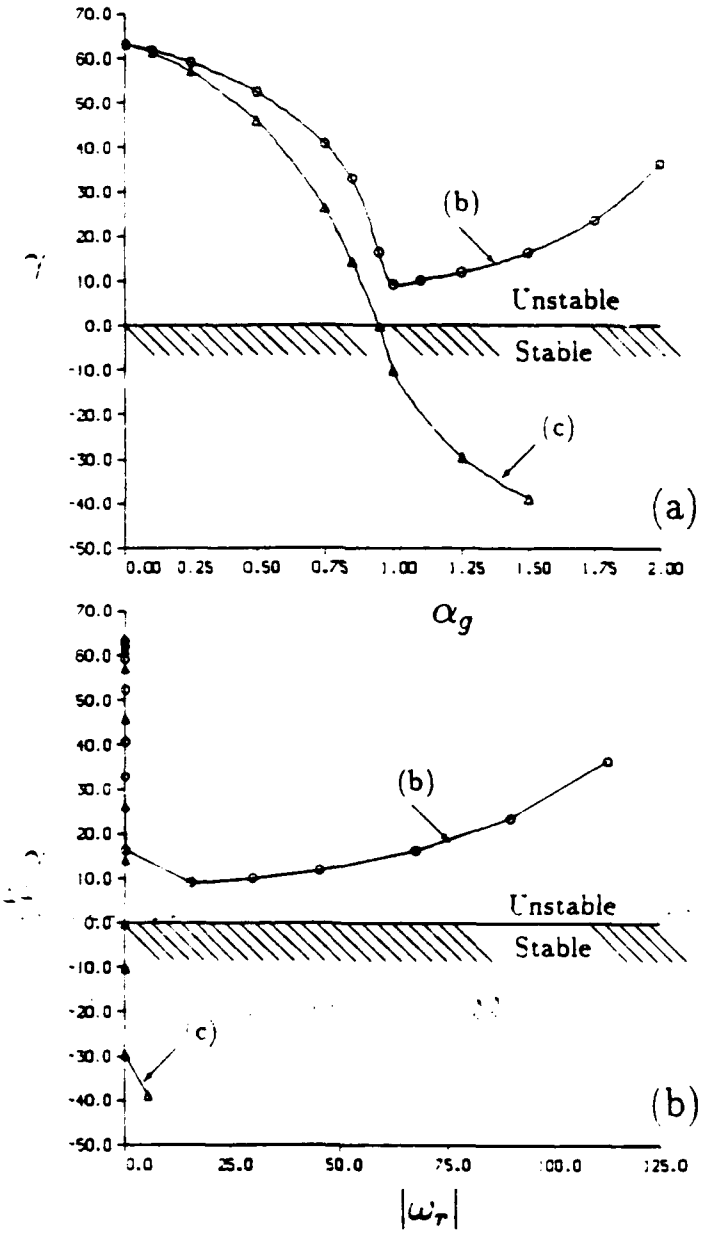


Figure 5.4: (a) *Growth rate γ vs. proportional gain α_g for the two cases using flux loop position #2. The circles correspond to the case with derivative gain $\beta_g/\alpha_g = 0.01s^{-1}$. The triangles correspond to $\beta_g/\alpha_g = 0.02s^{-1}$.*

(b) *Growth rate γ vs. oscillation frequency ω_r for the two cases with flux loop position #2. Doubling the derivative gain to $\beta_g/\alpha_g = 0.02s^{-1}$ virtually eliminates the oscillation and stabilizes the plasma.*

flux-loop locations are extremely insensitive to changes in the vertical position during active feedback.

It is the interaction of the perturbed plasma flux with the flux from the active feedback coils and the eddy currents in the passive conductors that creates this region of nearly zero flux. This effect demonstrates the importance of careful placement of the flux loops with consideration toward the interaction of the feedback system with the plasma flux. In this case, it is a result of the plasma–feedback system geometry, and not the result of any significant deformation of the eigenfunction. In Chapter 6 we examine the effect of plasma deformability on this interaction.

If one could subtract the active coil and passive conductor contributions to χ so that only the plasma contribution is measured, then any pair of up/down symmetric flux loops could effectively measure the vertical displacement and therefore control the plasma. The effects of the active feedback coils could be easily subtracted from the signal by redefining the perturbed flux measurement to be

$$\tilde{\chi}_o = \chi_o - \sum_i M_{o,i} I_i \,. \tag{5.16}$$

where the sum is over the active feedback coils. The effects of the eddy currents in the vacuum vessel wall could, in principle, also be subtracted out. However, this would require a detailed knowledge of the eddy current distribution in the wall for the given plasma displacement and active feedback response. This could easily be done in our calculation, but it might prove difficult in an experiment. Subtracting out only the active feedback coil contribution would probably improve the sensitivity of the flux loops, but it is not clear how much this would be improved without accounting for the eddy current effects. This will be considered in future work.

Finally, curve (e) of Fig. 5.2 shows the results of using the flux loops at position ± 4 of Fig. 5.1, $(X_o, Z_o) = (2.9, \pm 0.6)$. This places the flux loops on the outboard side of the plasma at a relative position with respect to the plasma similar to that of the flux loops at position $\#1$ but, on the opposite side of the plasma. It is seen that the plasma can be stabilized using flux measurements at these points, although it takes a higher gain α_g than when using flux loops at position #1. Therefore, this flux-loop pair position is less sensitive than pair #1, but is still sensitive enough to successfully control the plasma. There is no need to increase the derivative gain beyond the 1% used in the first case.

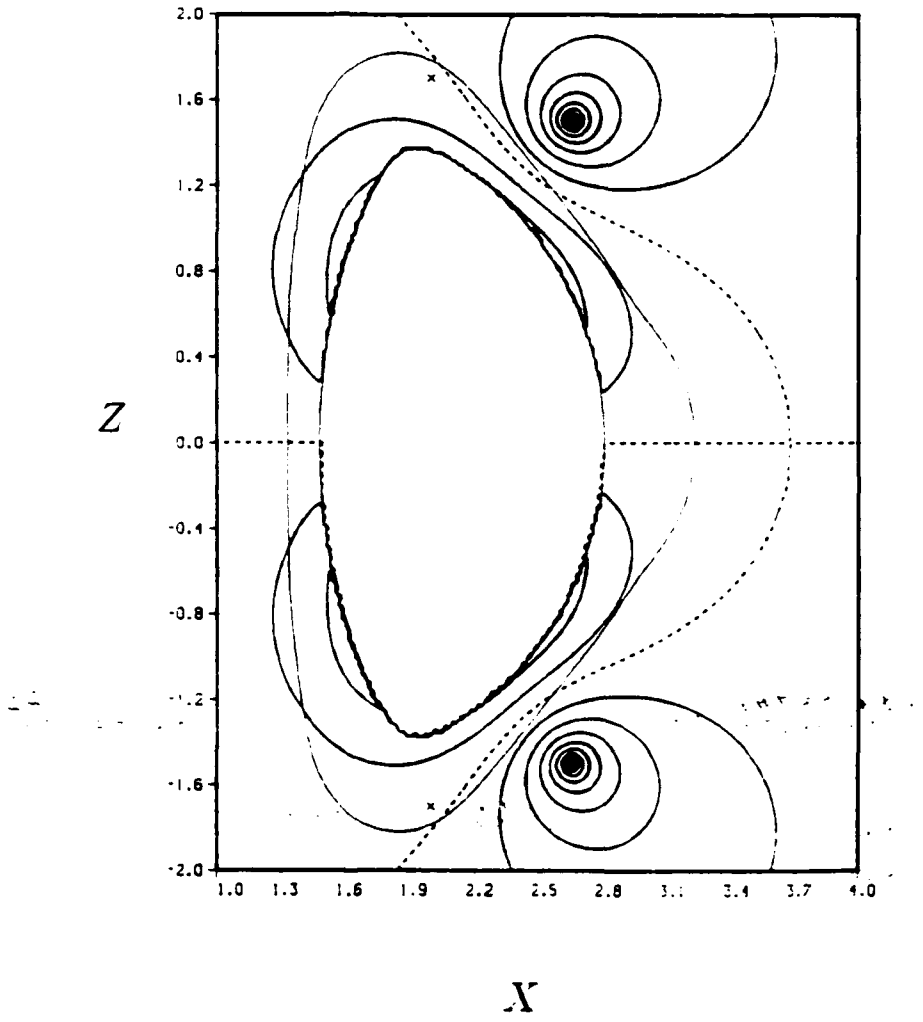


Figure 5.5: *Perturbed flux contours in the vacuum region for active feedback in which the active feedback systems uses flux measurements from flux-loop pair #3. The flux loops of pair #3 are shown by ' × ' symbols. The zero-flux contour is shown as a dashed line. The flux loops in this case lie very close to the zero-flux contour.*

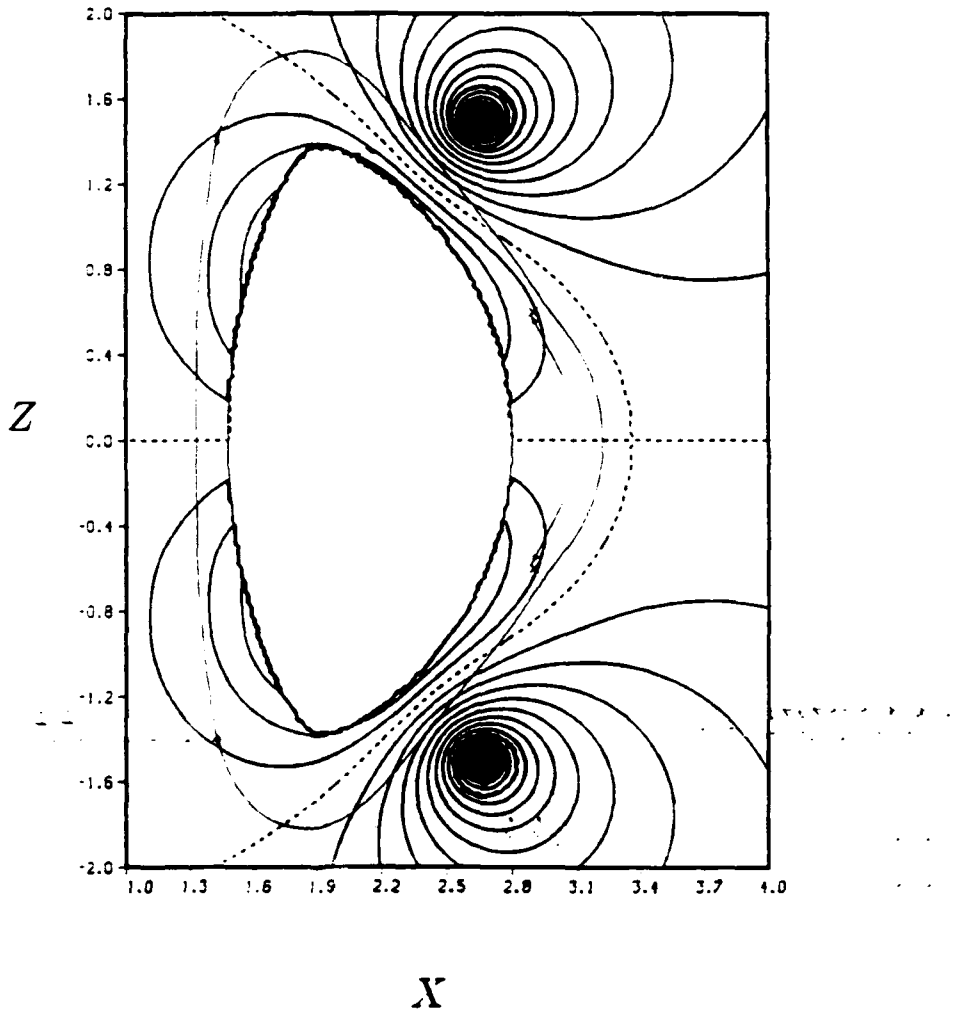


Figure 5.6: *Perturbed flux contours in the vacuum region for active feedback in which which the active feedback systems uses flux measurements from flux-loop pair #4. The flux loops of pair #4 are represented by ' × ' symbols which are indicated by the arrows. The zero-flux contour is shown as a dashed line.*

The perturbed flux contours for the active feedback using the flux loops at position ± 4 is shown in Fig. 5.6. Even though the flux loops at position ± 4 are almost as close to the active feedback coils as the flux loops at position ± 3, the geometry is such that the zero-flux contour does not closely approach the flux loops at position ± 4. These loops lie within a region where the perturbed flux is large enough to provide a flux difference measurement that can stabilize the plasma. The ability of the active feedback to stabilize the plasma using these flux loops also agrees with TSC results [72].

5.3 Discussion

We have used the NOVA-W code to examine the active feedback of the CIT design plasma. The study focused on the effectiveness of the active feedback system using measurements from various flux-loop pair locations. Qualitatively, the results compared well with those from TSC for the same locations for the flux loops.

The direct comparison of quantitative results and performance between the NOVA-W and TSC codes is, at best, very difficult, owing to the great difference in the calculations performed. TSC performs a full nonlinear time evolution of the plasma and feedback systems, and shows the behavior in time for a particular feedback gain law. The NOVA-W code on the other hand calculates the eigenvalue (growth rate) for the linear stability problem. The method of using NOVA-W dictates that one must perform a scan in feedback gain in order to find the stabilization due to a particular feedback gain. In practice, one can usually use widely spaced increments in gain in the scan while the mode is still unstable. Once the gain becomes high enough to stabilize the mode, however, one must carefully follow the root in the complex plane to ensure that the solution is not some unrelated stable mode. This procedure allows one to see the performance of the feedback system with respect to gain. It is also ideally suited for calculating the effect of feedback due to small changes in various parameters: gain law (α_g, β_g), plasma equilibrium, vacuum vessel, feedback coil positions, and so on. This is because it is much faster to find the solution for a particular system if one has a good initial guess—such as the solution for a slightly different case. If one is using TSC, on the other hand, a full simulation must be performed in each case and the time spent will be the same.

A full TSC calculation used to obtain these results [72] uses approximately 220–250

Cray-2 minutes. The time required by NOVA-W to calculate each growth rate for a particular gain value with the parameters used in the calculations of this chapter (200 radial finite elements, ± 13 poloidal Fourier harmonics, 128 Θ-grid points) is about 10–15 Cray-2 minutes. The improved performance of NOVA-W is apparent, especially in the case of a parameter scan. Moreover, the NOVA-W time can be significantly improved by a reformulation of the routines that calculate the eigenvalue, as we shall discuss in a later section.

We have examined the basic effect of a feedback system on the CIT plasma in this chapter. We have not considered the effect of non-rigid vs. rigid motion although the NOVA-W code is ideally suited for such an analysis. In the following chapter we consider the effects of both passive and active feedback systems on the non-rigid components on the plasma eigenfunctions.

Chapter 6

The effects of plasma deformability

Iste bombus aliquid significat!

6.1 Non-rigid effects on the passive stabilization

6.1.1 The ARIES-I equilibrium

The ARIES-I plasma (introduced in Sec. 4.4.2) has high triangularity ($\delta = 0.43$; see Table 4.2 for the equilibrium parameters) and moderate elongation ($\kappa = 1.61$). As mentioned in Chapter 1, theoretical [28,30] and experimental [48] studies have suggested that triangularity is destabilizing and can cause a significant non-rigid contribution to the vertical instability.

An analysis equivalent to that done by Rosen for the square plasma [16] (reviewed in detail in Chapter 2) demonstrates that the minimized eigenfunction for the axisymmetric mode in a triangular plasma consists of a $m = 2$ non-rigid perturbation superimposed on a rigid shift (compare to the $m = 3$ "wrinkle" of the square plasma). As in the case of the square plasma, the plasma is stable to a pure rigid shift, but becomes unstable to a non-rigid component superimposed on the rigid shift. Therefore we would expect significant non-rigid components in the unstable eigenfunction for the dee-shaped ARIES-I plasma.

Figure 6.1a shows the unstable eigenfunctions for the ARIES-I equilibrium in the absence of any conducting wall. The overall normalization of the eigenfunction

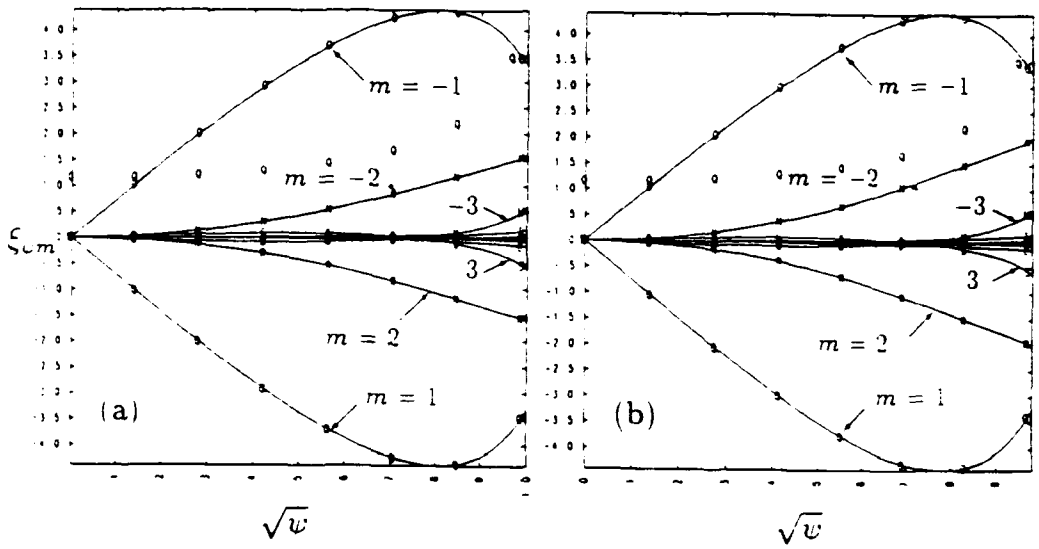


Figure 6.1: (a) *Fourier components of the radial displacement eigenfunction* ξ_ψ *vs.* $\sqrt{\psi}$ *for the ARIES-I equilibrium* ($\psi_{wall} = 0.95$) *with no surrounding wall. The q-profile is shown by the dotted curve with the "Q" labels. Note that the value of the* $m = 1$ *component at the edge is normalized to* q_{edge}.

(b) *Fourier components of the displacement* ξ_v *for a uniform vertical rigid shift.*

is chosen to allow display of the q-profile on the same plot. The $\pm m$ components of the eigenfunction plots are antisymmetric since the unperturbed plasma-vacuum configuration is symmetric about the midplane. Therefore, we will refer to the $\pm m$ component pairs simply as the $m = 1, 2, 3$ etc. components. These are, in effect, the coefficients of $\sin(m\Theta)$.

For an infinite-aspect-ratio plasma with a circular cross-section, a rigid shift would be represented by a pure $m = 1$ component in the eigenfunction. For the more complicated cross-sectional shapes at finite aspect ratio that we consider here, a rigid shift is still primarily composed of an $m = 1$ contribution to the eigenfunction, but there are also contributions from the higher m components. Figure 6.1b shows the purely rigid form of ξ_ψ for the ARIES-I equilibrium. For this equilibrium, the Fourier decomposition of the uniform rigid shift in the equal-arc-length magnetic coordinate system is quite different from a pure $m = 1$ component: there is a large $m = 2$ contribution as well as a significant $m = 3$ piece at the edge. The actual eigenfunction (Fig. 6.1a) is somewhat different, in particular the $m = 2$ contribution is decreased by approximately 20%. Therefore there is a non-rigid contribution to the true eigenfunction that is effectively a negative $m = 2$ component added to to the rigid shift.

Figure 6.2a shows the corresponding eigenfunction for the ARIES-I equilibrium with a completely surrounding resistive wall. It is clear that there is an additional non-rigid contribution that further reduces the $m = 2$ component, as well as reducing the $m = 3$ component with respect to the rigid shift. This non-rigid motion amounts to reducing the upward shift on the outboard side of the plasma and turning the upward motion inward so that the plasma is drawn into the x-point region. The conducting wall serves to modify the eigenfunction in such a way as to enhance this non-rigid motion. We will return to this point later.

6.1.2 ARIES-I stability with respect to conducting plates

It was discovered that when only small segments of the wall (conducting plates) were present, there is a strong dependence of the relative magnitude of the $m = 2, 3$ components on the poloidal position of these conducting plates. Figure 6.2 shows the eigenfunction for the case with a pair of conducting plates on the outboard side. It can be seen that this eigenfunction differs from a rigid shift even more than the eigenfunction with a complete wall. This is evident in that the $m = 2$ component

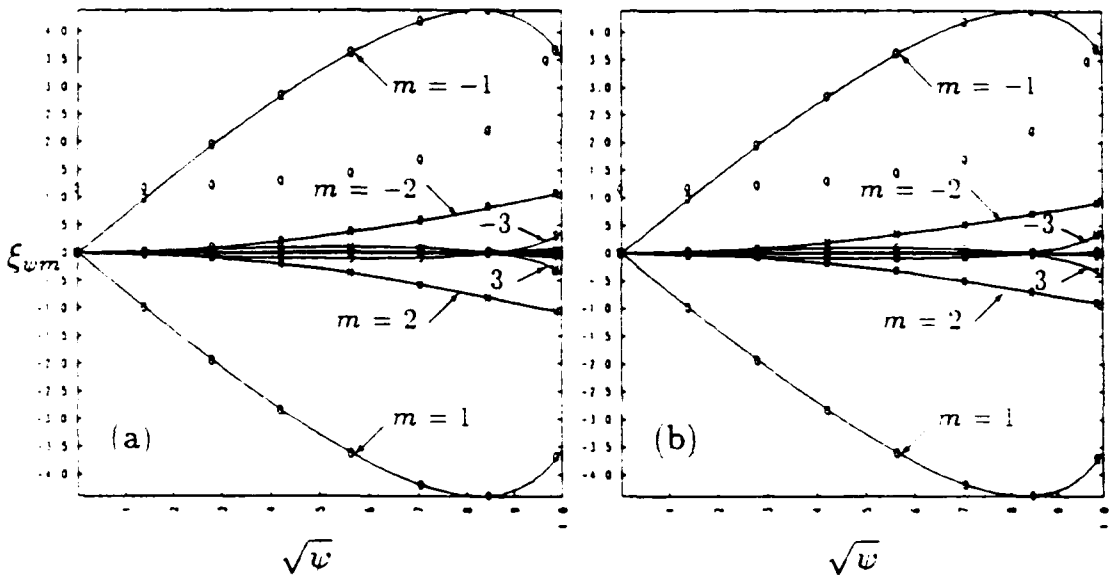


Figure 6.2: (a) *Fourier components of the radial displacement* ξ_w *vs.* $\sqrt{\psi}$ *for the ARIES-I equilibrium* ($\omega_{rot} = 0.95$) *with a surrounding resistive wall. This is the same as Fig. 4.13, and the wall contour is shown in Fig. 4.12b.*

(b) *Fourier components of the displacement for the case with a pair of symmetric conducting plates on the outboard side at about* 80° *off the midplane. Note a small reduction of the* $m = 2$ *component in relation to the eigenfunction with the full resistive wall, (a).*

for the partial wall case is somewhat smaller than for the complete wall case. It is of interest to understand where is the most effective location for discrete conductors for passive stabilization. In a reactor design, for example, the space near the plasma is very valuable, and it would be desirable therefore to reduce the amount of passive conductors that must lie very close to the plasma.

In order to study this question in detail, we performed a series of calculations in which there is a single pair of up-down symmetric plates. The length of each plate is approximately 1/8 of the circumference of the wall used above, and has a resistance equal to 1 cm of aluminum. In each calculation the plates are centered at a different poloidal location, Θ. The value of Θ is changed, between calculations, in 14° increments from $\Theta = 0$ at the outboard midplane to $\Theta = \pi$ at the inboard midplane. We are particularly interested in how the non-rigid components of the motion change with regard to the variation of the passive conductor position. To characterize these changes we calculate the changes in the the ratios of the $m = 1$ components to the $m = 1$ component, ξ_{wm}/ξ_{w1}, measured at the plasma edge.

The results are shown in Fig. 6.3. The figure shows the computed values of ξ_{w2}/ξ_{w1} and ξ_{w3}/ξ_{w1} as a function of Θ. Also shown is the growth rate γ as a function of Θ. The values of γ, ξ_{w2}/ξ_{w1}, and ξ_{w3}/ξ_{w1} for a continuous wall and with no wall are also shown, as are the ξ_{w2}/ξ_{w1} and ξ_{w3}/ξ_{w1} ratios for the Fourier representation of the uniform vertical rigid shift.

The results show that the most effective stabilizing position for the plates is on the outboard side between $\Theta = \pi/4$ and $\Theta = \pi/2$. Specifically, the smallest growth rate is when the plates are centered at $\Theta \approx 5\pi/12$ and $\Theta \approx 19\pi/12$. The growth rate varies by a factor of about 120 as the poloidal plates traverse the half-circle. There is a variation of nearly a factor of 30 just as the plates move through the outboard region. It is interesting to note that the highest growth rate (corresponding to the case in which the plates are adjacent at the inboard midplane) is over 600 times faster than the growth rate with the complete continuous resistive wall, but only about 70 times slower than the growth rate with no passive conductors at all. The plasma is in fact ideally unstable with the conductors at this location. It is clear that conductors in this region provide very little stabilization. The minimum growth rate for the plates, on the other hand, is only about 5 times the growth rate for a completely enclosing wall. Therefore, optimally placed conductors can provide much stabilization.

The relative contribution of the $m = 2, 3$ components to the eigenfunction is seen

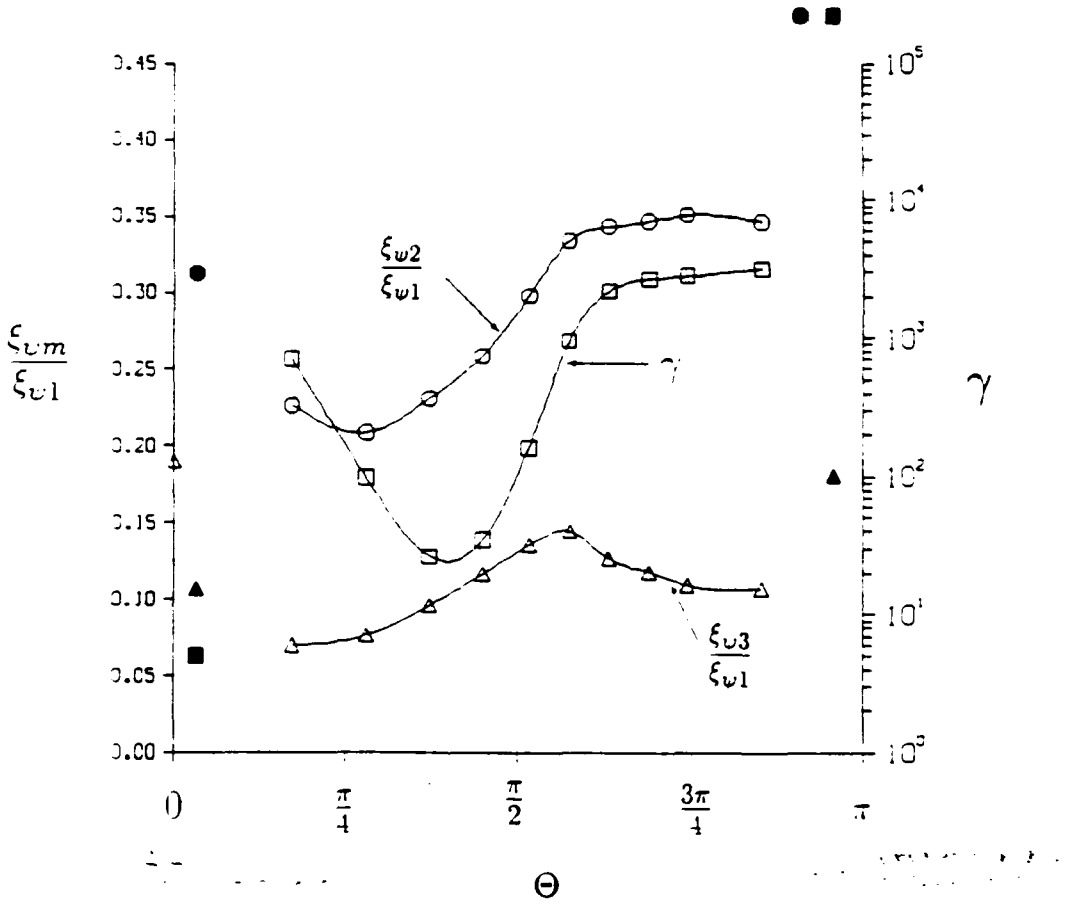


Figure 6.3: *Effect of poloidal position of conducting plates on eigenfunction and growth rate of the ARIES-I equilibrium. The ratios of the $m = 2, 3$ components of the eigenfunction to the $m = 1$ component at the plasma edge are graphed as a function of the poloidal position of the conducting plates. In particular, the $m = 2$ (circles) and the $m = 3$ (triangles) contributions are shown here. Also shown is the growth rate γ (squares) as a function of the plate position. The corresponding values for a continuous complete wall are shown in solid on the left, and the values for the case with no wall whatsoever are shown in solid on the right. The values of the ratio ξ_{wm}/ξ_{w1} for the Fourier representation of the rigid shift are shown on the left axis, however, the ratio ξ_{w2}/ξ_{w1} is off the scale at about 0.613. This study was done for the $w_{rat} = 0.95$ equilibrium.*

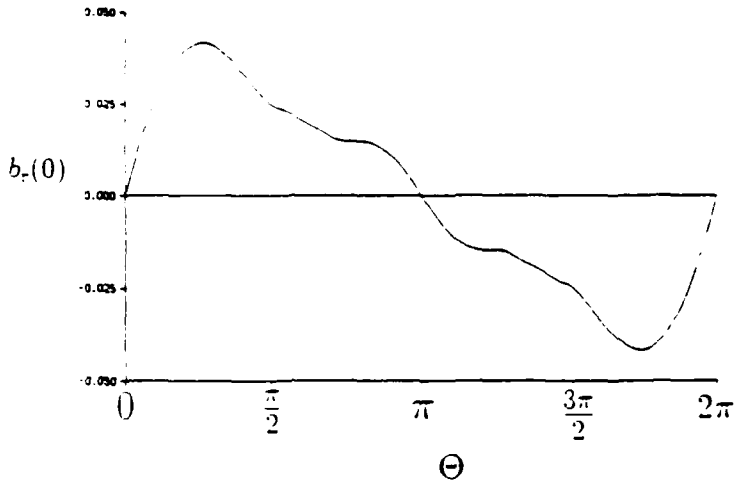


Figure 6.4: *Magnitude of the perturbed radial magnetic field at the magnetic axis, $b_r(0)$, due to a unit (positive) current at every point along the wall contour used in the ARIES-I calculation.*

to vary significantly as the plate position is varied, with the $m = 2$ contribution dominating the $m = 3$ contribution by a factor of 3–4. The ratio ξ_{w2}/ξ_{w1} shows a minimum at Θ just above $\pi/4$. This region corresponds to a rapid decrease in γ as Θ increases to the minimum in γ near $\Theta = 5\pi/12$. As Θ increases further, a steep rise in ξ_{w2}/ξ_{w1} follows a very steep rise in γ until $\Theta \approx 5\pi/8$, where both ξ_{w2}/ξ_{w1} and γ begin to level off. The variation of ξ_{w3}/ξ_{w1} with respect to Θ is characterized by a slow increase from its minimum (at the smallest Θ value used) to a maximum at about $\Theta = 9\pi/16$, followed by a gentle decrease with increasing Θ.

Clearly the eigenfunction is non-rigid, with the details changing with respect to the position of the plates. The deformations in the eigenfunction affect the growth rate, but there is another important effect here which must be considered first. Figure 6.4 shows the magnitude of the perturbed radial field at the magnetic axis, $b_r(0)$, due to a unit (positive) current that is placed at different positions along the wall contour. During a vertical instability the plasma motion induces eddy currents in the surrounding conductors which in turn produce a stabilizing radial field at the plasma. The magnitude of $b_r(0)$ is the true measure of the stabilizing effect (at the magnetic axis) of the eddy currents. The toroidal effects are apparent in Fig. 6.4. The currents

on the outboard section for $\pi/4 \lesssim \Theta \lesssim \pi/2$ are seen to be much more stabilizing than the corresponding currents on the inboard side. Furthermore, an axisymmetric current loop which is displaced vertically from the magnetic axis will induce a current pattern with the same shape as in Fig. 6.4 thereby, compounding the effect. This will likely make the outboard conductors more effective for passive stabilization by about a factor of 5. Figure 6.3, however, shows a factor of over 100 difference between the most effective plate position and the least effective plate position. This large difference cannot therefore be explained by the toroidal effects on b_r.

Much insight into the role of non-rigid components of the eigenfunction in so far as they affect the growth rate can be gained by considering plots of the perturbed poloidal flux. Recall from Eq. 3.38 that the eddy currents induced at a given point on the wall are proportional to the (time derivative of the) perturbed poloidal flux at that point on the wall. Therefore the region with the highest induced eddy currents should correspond to the region of the wall that intersects the contour of highest flux χ.

Figure 6.5 shows the contours of perturbed poloidal flux χ for the ARIES-I equilibrium for the case of a completely enclosing resistive wall. It is seen that the perturbed flux contours are weighted toward the outboard region. This is a consequence of the triangular shape of the plasma. The wall region on the outboard side between about $R = 8.2$ m and $R = 9.2$ m intersects the flux contour of highest χ at the wall. One would expect, therefore, that this would be the best place to locate a conducting plate. This is confirmed in Fig. 6.3 which shows that this is roughly the location with the smallest growth rate.

Figure 6.6a shows the perturbed flux contours when only the $m = 1$ component of the eigenfunction is used in the calculation of the flux contours. Figure 6.6b shows the perturbed flux contours due to the $m > 1$ components of the motion only, i.e., the difference in the flux between Fig. 6.5 and Fig. 6.6a. The signs of the perturbed flux are indicated (corresponding to positive perturbed flux in the upper half-plane for the $m = 1$ component: see Fig. 6.6a). Also calculated are the perturbed wall currents due to the components of the perturbed flux all along the wall contour. Figure 6.7 shows the induced current $J_\phi(\Theta)$ distribution in the resistive wall. The total current, the current due to the $m = 1$ component of the motion, and the current due to the $m > 1$ components of the motion are shown together. By examining this figure we

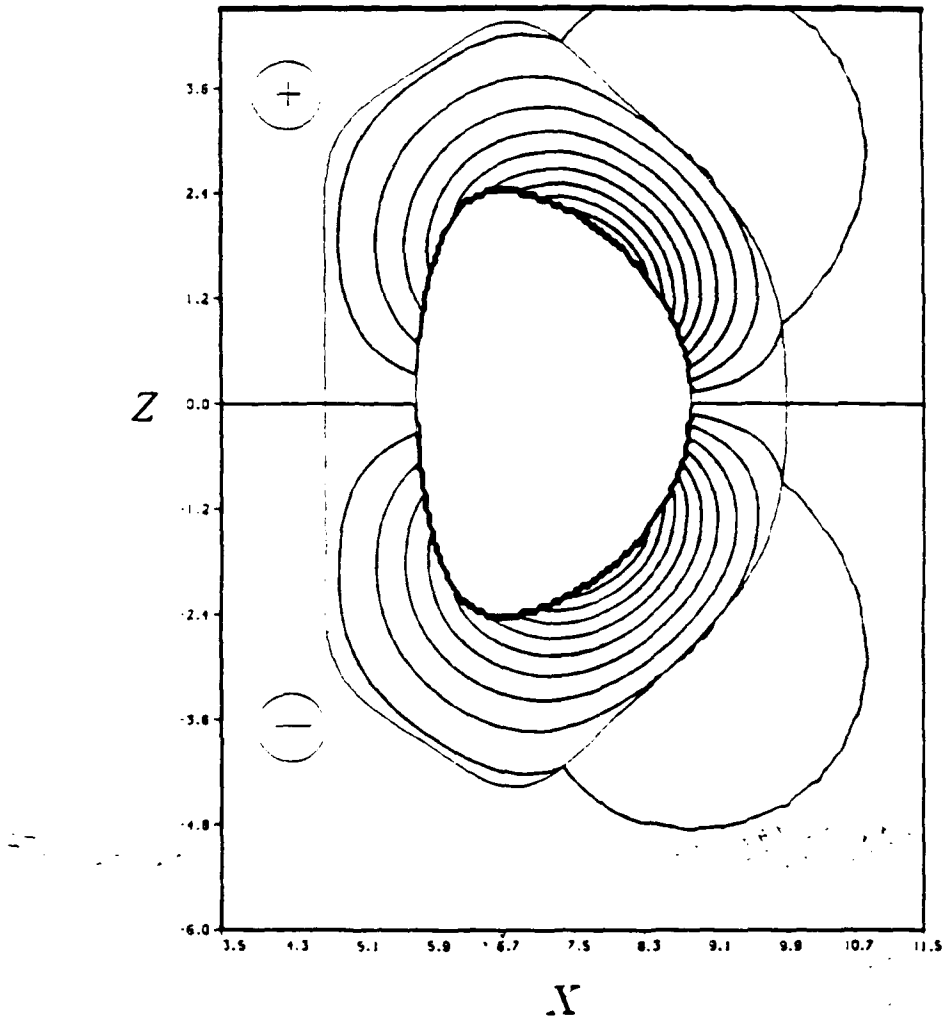


Figure 6.5: *Perturbed flux contour plots for ARIES-I—full motion (rigid and non-rigid). Notice how the perturbed flux χ is weighted toward the outboard region. The outermost flux contour (contour of lowest χ magnitude shown in this plot) lies entirely within the wall on the inboard side. Only on the outboard side does it extend beyond the wall. The sign of the perturbed flux changes across the midplane as this is a pure antisymmetric mode. The sign of the flux is shown in the circles. The midplane is itself a zero-flux (null) contour.*

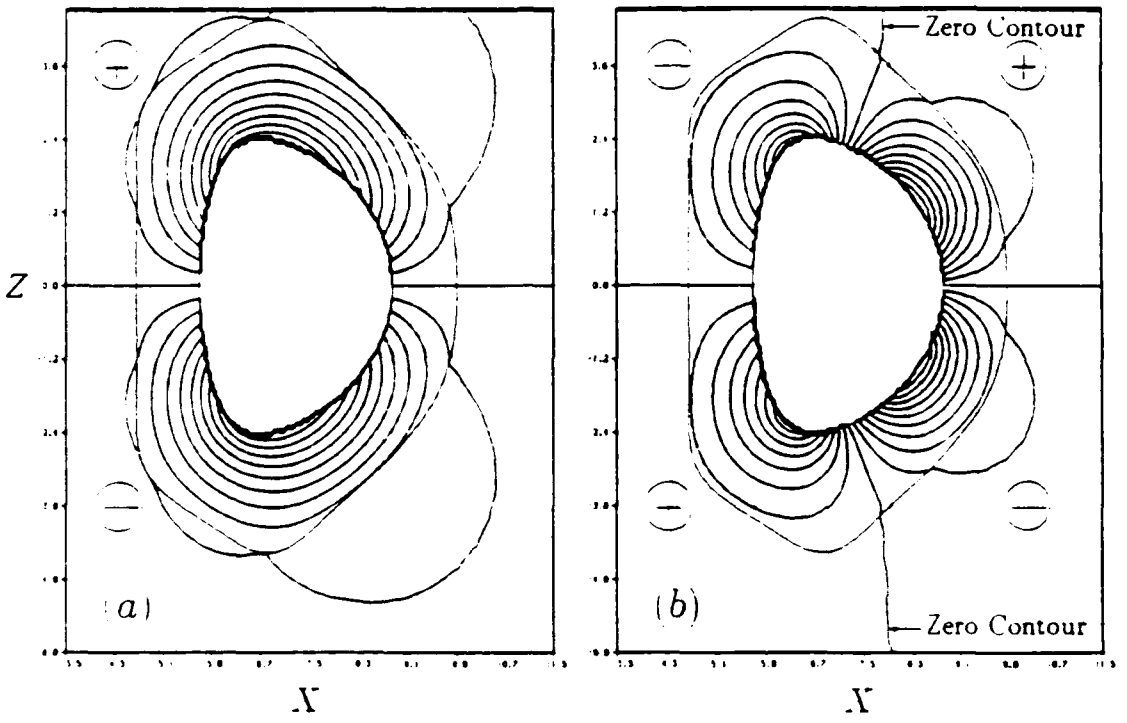


Figure 6.6: (a) *Perturbed flux contour plots for ARIES-I corresponding to the* $m = 1$ *component of the motion only. The outermost flux contour extends beyond the wall over most of the inboard side. The sign of the flux is shown and changes through the midplane. The midplane itself is therefore a zero-flux contour.*

(b) *Perturbed flux contour plots for ARIES-I corresponding to the non-rigid components of the motion only. There are two additional zero-flux contours (in addition to the midplane contour, which is always present for antisymmetric modes). The additional zero-flux contours are shown along with the signs (the signs are shown in the circles) of the flux contours.*

can see why there is so much variation in the relative non-rigid contributions to the eigenfunction.

On the outboard side of the zero contour in Fig. 6.6b the perturbed flux of the $m > 1$ components is of the same sign as the perturbed flux due to the $m = 1$ motion. Therefore the induced current in this region of the wall will combine to stabilize all of these components. Since the perturbation due to the $m > 1$ components is more localized than for the $m = 1$ component, a conducting plate in this region will stabilize

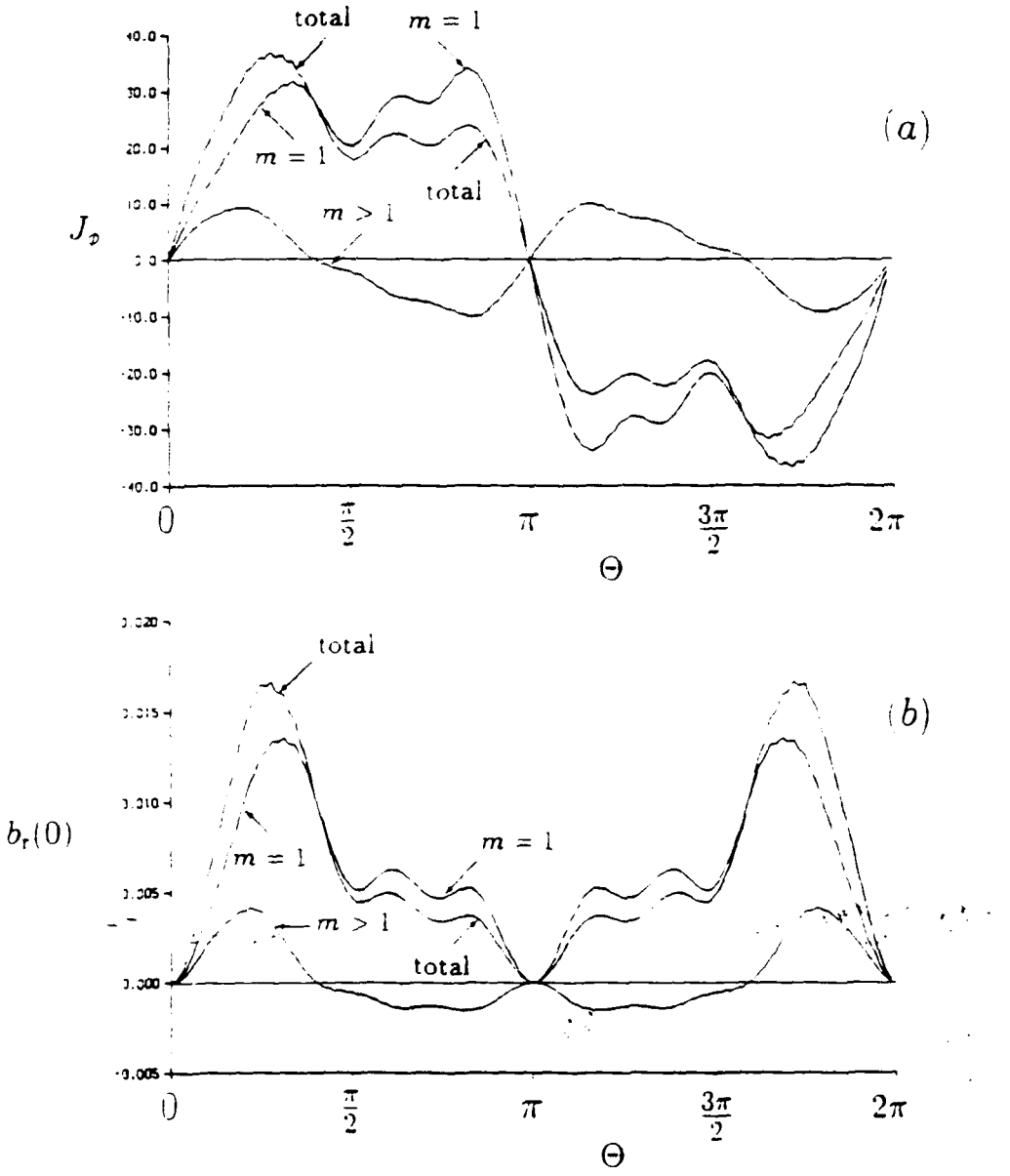


Figure 6.7: (a) *Perturbed eddy currents in the wall induced by the unstable displacement as a function of poloidal position on the wall contour. Shown are the total current from the complete eigenfunction, the current induced by the* $m = 1$ *component of the eigenfunction, and the current from the* $m > 1$ *components of the eigenfunction.* (b) *Perturbed radial field at the magnetic axis due to the perturbed eddy currents in the resistive wall as a function of the poloidal position of the eddy currents in the wall. The value of* $b_r(0)$ *due to the complete eigenfunction is shown, as well as the contributions from the* $m = 1$ *and* $m > 1$ *components of the eigenfunction.*

these components more effectively compared with the $m = 1$ component. If there is a conducting plate in this region, we would expect the eigenfunction to be modified so that the $m > 1$ components are reduced, since this would be energetically favorable and thus would lower the stabilizing influence of the plates and enhance the instability. This is seen in Fig. 6.3 to be the case.

On the inboard side of the zero-flux contour, however, the perturbation from the $m > 1$ components changes sign with respect to the $m = 1$ component perturbation. The value of the perturbed flux (and therefore induced current) is less on the inboard section of the wall than on the outboard section (and less than the induced current due to the $m = 1$ motion), but the sign is now opposite to that due to the $m = 1$ component. Therefore the $m = 1$ and $m > 1$ components are working against each other on this inboard wall region, significantly reducing the overall stabilizing effect. Figure 6.7 shows how the eddy currents induced by the $m > 1$ motion are opposite in sign to the stabilizing total current. If there is a conducting plate in this region, we expect a deformable plasma to be able to modify its eigenfunction so that the $m > 1$ components are increased in magnitude compared with the $m = 1$ component in order to minimize the energy, thereby maximizing the growth rate. This is demonstrated in Fig. 6.3 by the increase in the $m > 1$ components (particularly $m = 2$) and by the rapid increase in the growth rate γ as the plates move from the outboard region toward the inboard region.

Now that we understand how the different parts of the wall affect the components of the eigenfunction, we can see how the eigenfunction is modified in order to minimize its energy for a particular configuration of the stabilizing plates. The plasma motion is modified in such a way that it can "slip" around the passive conductors in order to reduce the resistance to its motion. An equivalent way to express this is to say that the plasma eigenfunction is modified in such a way that the stabilizing eddy currents in the surrounding conductors are reduced as much as possible.

This analysis is not only important for the case of individual discrete conducting plates, but also for the case of a complete conducting shell. The analysis provides insight on where one might increase the wall thickness to significantly increase the passive stabilization without everywhere increasing the thickness which might have otherwise detrimental effects. The eigenfunction for a completely surrounding wall (Fig. 6.2a) reflects the sum of the effects described above. In particular, the stabilizing influence of the outboard wall section is reflected by the fact that the $m = 2,3$

components are reduced from the eigenfunction with no wall (making the eigenfunction further from a rigid shift). This works to enhance the non-rigid effect of drawing the plasma into the separatrix region. Notice, however, that the $m = 2, 3$ components are larger than for the case of individual plates on the outboard region. This is because the complete wall includes conductors in the separatrix region which oppose this non-rigid motion.

6.1.3 The CIT equilibrium

We now perform the same study using the CIT equilibrium of Section 4.3. This equilibrium has a smaller triangularity ($\delta = 0.26$: see Table 4.1 for the equilibrium parameters), but higher elongation ($\kappa = 2.0$). The eigenfunction for this equilibrium is shown in Fig. 4.8. With a completely surrounding conducting wall the $m > 1$ components are small compared with to the $m = 1$ component.

We study the variations in the growth rate and the relative non-rigid component contributions to the eigenfunction with respect to the position of the symmetric poloidal plates. The results are shown in Fig. 6.8, in the same fashion as the ARIES results of Fig. 6.3.

Qualitatively the curves look somewhat similar to the ARIES results: however, there are significant quantitative differences. First consider the similarities. The growth rate γ decreases to a minimum between $\Theta = \pi/4$ and $\Theta = \pi/2$, and then increases rapidly as the plates are moved to the inboard side ($\Theta > \pi/2$). This gives a minimum in γ at about the same position in Θ of the plates as in the ARIES case. The relative $m = 2$ contribution ξ_{w2}/ξ_{w1} also decreases slightly to a minimum in this same region and then increases rapidly as the plates move inboard. The relative $m = 3$ contribution ξ_{w3}/ξ_{w1} increases as Θ increases to a maximum at Θ slightly above $\pi/2$ and then decreases. Clearly the outboard region with $\pi/4 \lesssim \Theta \lesssim \pi/2$ is the best place to put a conducting plate in this case as well.

There are some noticeable differences, however, when compared to Fig. 6.3. While the ξ_{w2}/ξ_{w1} curve retains the same general shape, it is shifted toward the negative, with $\xi_{w2}/\xi_{w1} < 0$ for all $\Theta \lesssim \pi/2$ (corresponding to the plates being in the outboard region). Furthermore, the $m = 3$ component is generally larger in this case (dominating the $m = 2$ component for some parts), with a very large drop in ξ_{w3}/ξ_{w1} for $\Theta > \pi/2$. (There is a similar, but much smaller, drop in Fig. 6.3.) Also there is a

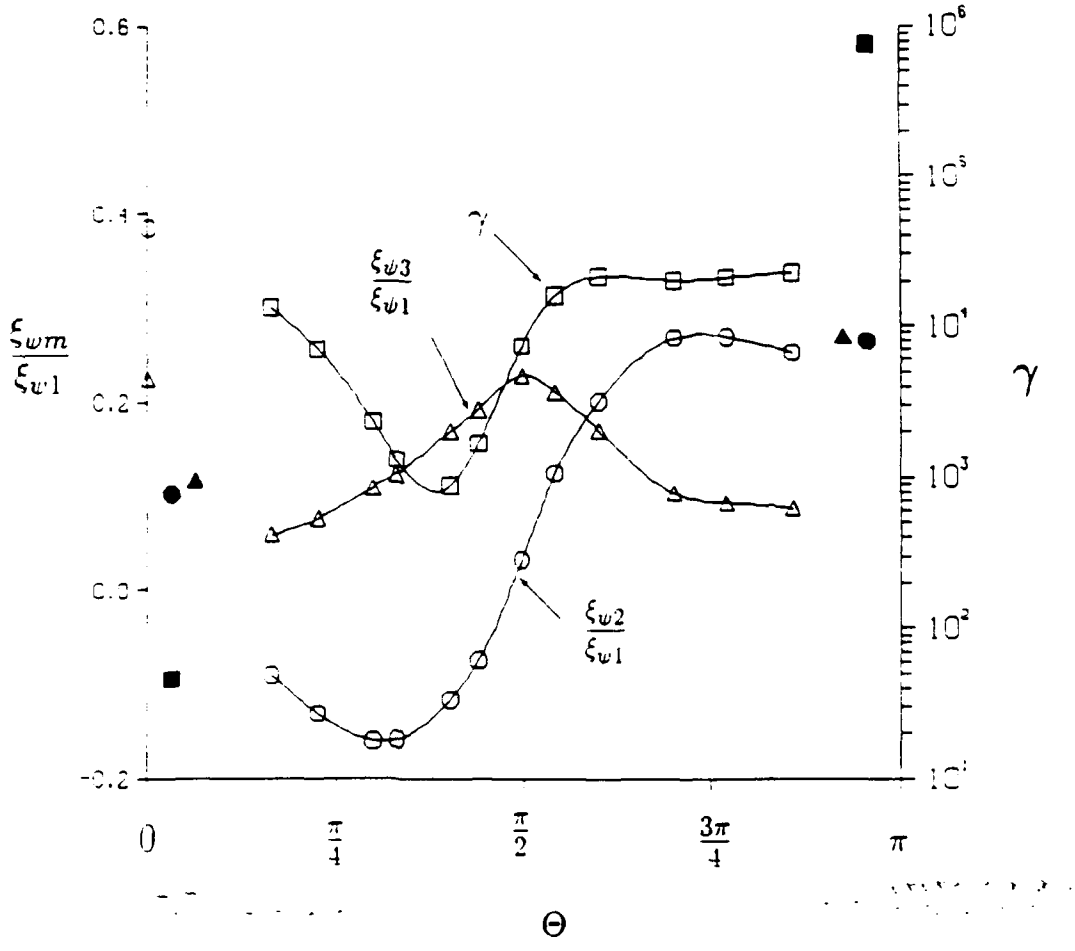


Figure 6.8: *Effect of poloidal position of conducting plates on eigenfunction and growth rate for CIT equilibrium. The ratios of the $m = 2, 3$ components of the eigenfunction to the $m = 1$ component at the plasma edge are graphed as a function of the poloidal position of the conducting plates. In particular, the $m = 2$ (circles) and the $m = 3$ (triangles) contributions are shown here. Also shown is the growth rate γ (squares) as a function of the plate position. The corresponding values for a continuous complete wall are shown in solid on the left, and the values for the case with no wall whatsoever are shown in solid on the right. The values of $\xi_{\psi m}/\xi_{\psi 1}$ for the Fourier representation of the uniform vertical rigid shift are shown on the left axis. This study was done for the $w_{rat} = 0.95$ equilibrium.*

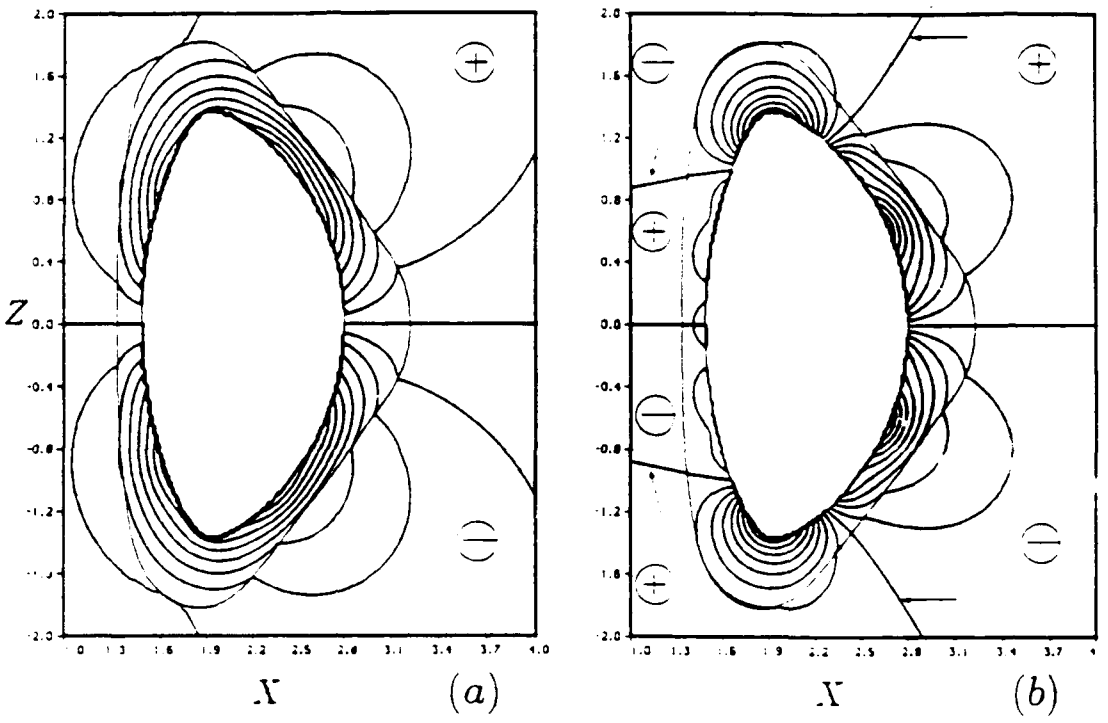


Figure 6.9: (a) *Perturbed flux contour plots for the CIT equilibrium—full motion (all components). The signs are shown in the circles. The midplane is itself a zero-flux (null) contour. This is a* $v_{rat} = 0.99$ *equilibrium.*

(b) *Perturbed flux contours for CIT due to the* $m > 1$ *components of the eigenfunction only. Note the 3-part structure of the contours in each half-plane. The zero-flux contours are denoted with an arrow (except for the midplane null contour).*

small decrease in γ between $\Theta = \pi/2$ and $\Theta = 3\pi/2$ (this is difficult to detect in the logarithmic plot since the drop is small, about 6.5%). There is no such drop in the ARIES case.

We can, again, gain some insight by considering the perturbed flux plots. Figure 6.9a shows the perturbed flux contours for the complete eigenfunction. Figure 6.9b shows the contours calculated using only $m > 1$ components. Since the $m = 3$ contribution is larger in this case, we see more of a 3-fold symmetry in Fig. 6.9b. There is one extra pair of null contours, so the upper and lower half-planes are each divided into three regions with varying sign in the perturbed flux.

With this equilibrium we see a strong variation in both the $m = 2$ and $m = 3$ components plate position is varied. A positive $m = 2$ component to the eigenfunction will result in a stabilizing effect from the conductors in the outboard region as we saw in the previous section. A positive $m = 2$ component will, however, result in eddy currents in the inboard wall conductors that are of opposite sign to the stabilizing currents induced by the dominant $m = 1$ component. We see in Fig. 6.8 that for conducting plates on the outboard region ($\Theta \lesssim \pi/2$) the eigenfunction has been modified from the form it takes with a completely surrounding wall, so that the $m = 2$ component is negative. The eddy currents due to this component of the motion oppose those stabilizing the primary $m = 1$ component and therefore weaken the stabilizing influence of the conducting plates. If the plates are moved to the inboard region ($\Theta \gtrsim \pi/2$), on the other hand, the $m = 2$ component again becomes positive, which is destabilizing for plates in this region. Therefore we see that no matter where the conducting plates are located, the eigenfunction will be modified so that the $m = 2$ component weakens the stabilizing influence of the plates.

The $m = 3$ component is somewhat larger for this equilibrium than for the ARIES equilibrium examined in the last section, and therefore plays a larger role in the overall stability. Although the $m = 3$ component never changes sign in order to minimize the stabilizing influence of the conducting plates, it does vary greatly in magnitude in order to do so. For the $m = 3$ component the stabilizing regions for the plates (where the induced eddy currents add to the currents from the $m = 1$ motion) are $0 \lesssim \Theta \lesssim \pi/3$ and $2\pi/3 \lesssim \Theta \lesssim \pi$, whereas the $m = 3$ induced current in conductors located within the region $\pi/3 \lesssim \Theta \lesssim 2\pi/3$ will be destabilizing. We see in Fig. 6.8 that when the plates are located in the two stabilizing regions, the $m = 3$ component become small thereby, minimizing its stabilizing contribution. On the other hand, when the plates are in the $m = 3$ destabilizing region ($\pi/3 \lesssim \Theta \lesssim 2\pi/3$), the $m = 3$ component is much larger, again minimizing the stabilizing influence of the plates.

Finally, we notice a small drop in the growth rate γ for $\pi/2 \lesssim \Theta \lesssim 3\pi/4$. In this region a positive contribution from the $m > 1$ components (see Fig. 6.9b) produces stabilizing eddy currents in the conductors on the inboard region. The effect is stabilizing, but this is limited by the modification of the eigenfunction to reduce the $m = 3$ component as much as possible. We see a rapid decrease in ξ_{w3}/ξ_{w1} in the same region in Θ where there is a slight drop in γ. Once again, a modified eigenfunction will keep the plasma as unstable as possible.

6.1.4 A purely elliptical equilibrium

We perform this study once again but with a purely elliptical plasma. The equilibrium parameters such as physical dimension, elongation, and toroidal field are the same as the for ARIES plasma (see Table 4.2) but the triangularity is zero. Shown in Fig. 6.11 are the Fourier components for the true eigenfunction and for the minimum vertical rigid shift. The eigenfunction is for the case of an enclosing shell. The wall contour is a concentric ellipse of the same elongation with the distance between the wall and the plasma on the midplane equal to 1.2 of the plasma minor radius. The form of the components for the actual eigenfunction and the rigid shift are very similar, indicating that the most unstable displacement is, in fact, very nearly a uniform rigid shift. Nevertheless, the form of the eigenfunction changes quite significantly with respect to the position of the plates, and thus shows a great degree of deformability and difference from the rigid shift. Figure 6.10 shows how the growth rate γ and the $m > 1$ component ratios ξ_{wm}/ξ_{w1} vary with respect to poloidal plate position.

The curves all are much more symmetric with respect to $\Theta \approx \pi/2$ than we saw in the other cases since there is no triangularity here. The toroidal effects are therefore apparent. The optimal plate position for passive stability is still on the outboard side at about $\Theta \approx 7\pi/16$ because of toroidal effects (see Fig. 6.4), but this is not as heavily weighted toward the outboard side as in the cases of the triangular equilibria that we considered. Also, the difference between the inboard and outboard position growth rates is not as large as in the other cases.

The discussion given in Section 6.1.3 for the CIT plasma regarding the changes in ξ_{w2}/ξ_{w1} and ξ_{w3}/ξ_{w1} can be applied to this equilibrium. When the poloidal plates are in the outboard region the eigenfunction is such that the $m = 2$ component is negative and therefore destabilizing. As the plates are positioned in the inboard region the $m = 2$ component becomes positive, which is destabilizing for the plates in this region. Also, while the $m = 3$ component never becomes negative, it does change in magnitude with respect to the plate position in order to weaken the stabilizing influence of the plates as much as possible.

An explicit demonstration of the plasma deformation is presented in Figs. 6.12 and 6.13. The figures show are the Fourier decomposition of the eigenfunction ξ_w and the projection of the displacement ξ onto the poloidal plane for the case of the plates on the inboard side $\Theta \approx \pi/4, 7\pi/4$ (Fig. 6.12) and on the outboard side at

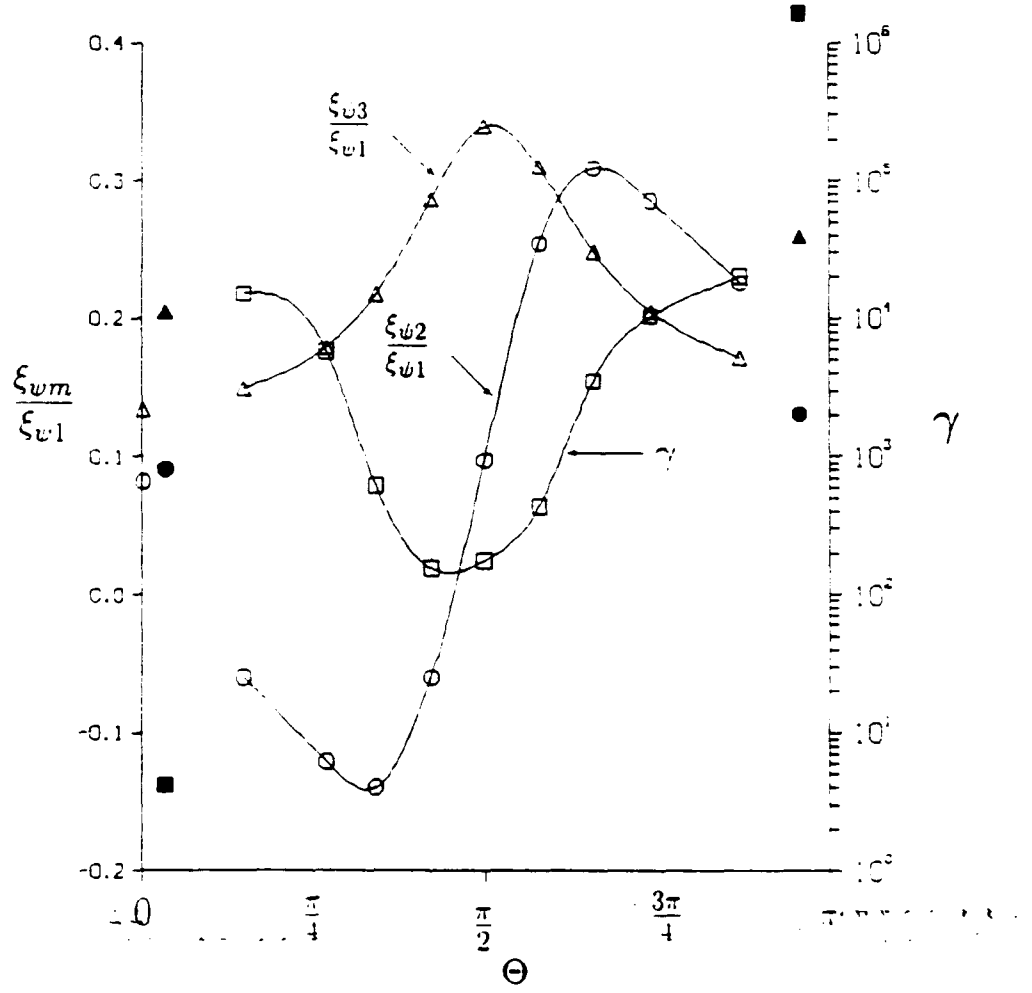


Figure 6.10: *Effect of poloidal position of conducting plates on eigenfunction and growth rate of the elliptical equilibrium. The ratios of the non-rigid components of the eigenfunction to the rigid component at the plasma edge are graphed as a function of the poloidal position of the conducting plates. In particular the* $m = 2$ *(circles) and the* $m = 3$ *(triangles) contributions are shown here. Also shown is the growth rate* γ *(squares) as a function of the plate position. The corresponding values for a continuous complete wall are shown in solid on the left, and the values for the case with no wall whatsoever are shown in solid on the right. The values of* ξ_{wm}/ξ_{w1} *for the Fourier decomposition of the uniform vertical rigid shift are shown on the left axis. The* $m > 1$ *component contributions are actually fairly small (see Figs. 6.11–6.13), but appear larger here because the ratio* ξ_{wm}/ξ_{w1} *is taken at the plasma edge, and the* $m = 1$ *component is much less at the edge than at its maximum at* $\sqrt{w} = 0.7$.

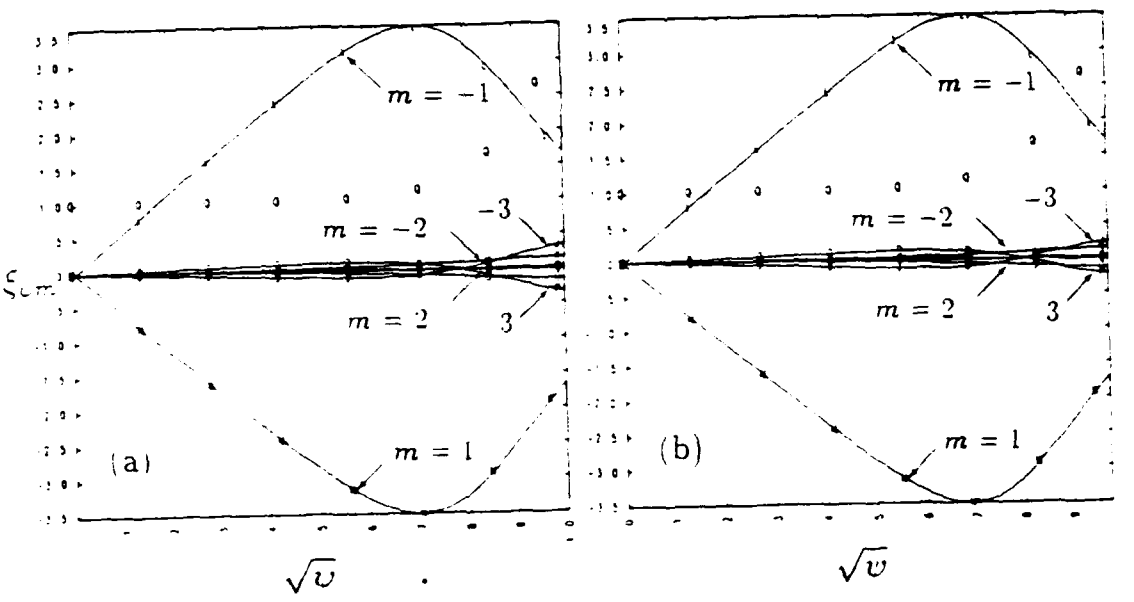


Figure 6.11: a) *Fourier components of the radial displacement* ξ_w *vs.* $\sqrt{w}$ *for the elliptical equilibrium with a complete continuous resistive wall.*

b) *Fourier components of* ξ_w *for a uniform vertical rigid shift displacement of the elliptical plasma.*

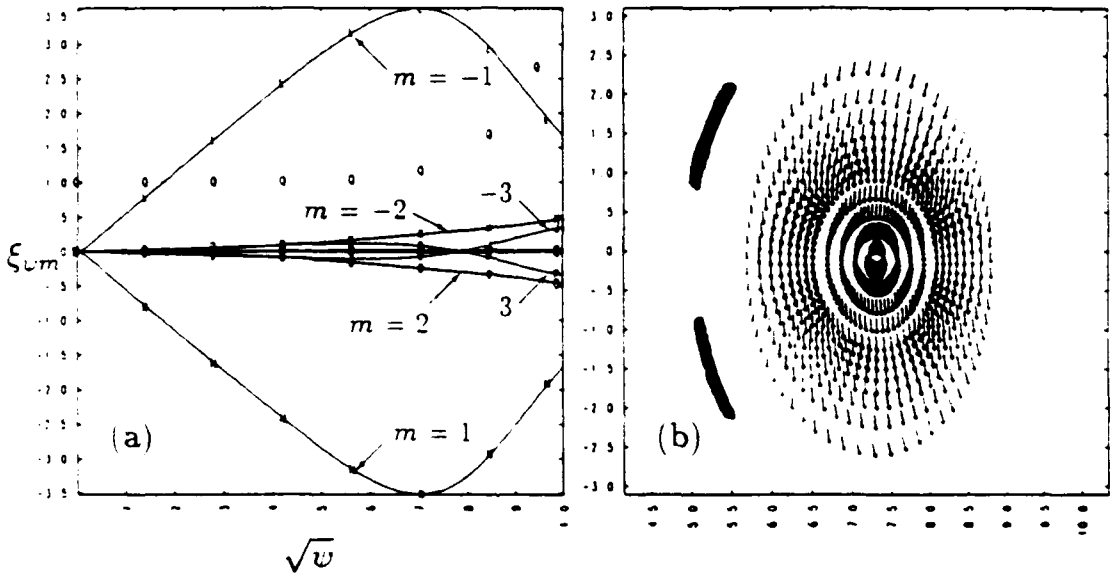


Figure 6.12: (a) *Fourier components of the radial displacement* ξ_w *vs.* $\sqrt{w}$ *for the elliptical equilibrium with conducting plates at about* 45° *off the midplane on the inboard side.*

(b) *Displacement vectors showing the motion of the unstable plasma. The plasma is partially stabilized by conducting plates on the inboard side of the plasma. Note the deformation of the plasma motion as it tries to move around the plates.*

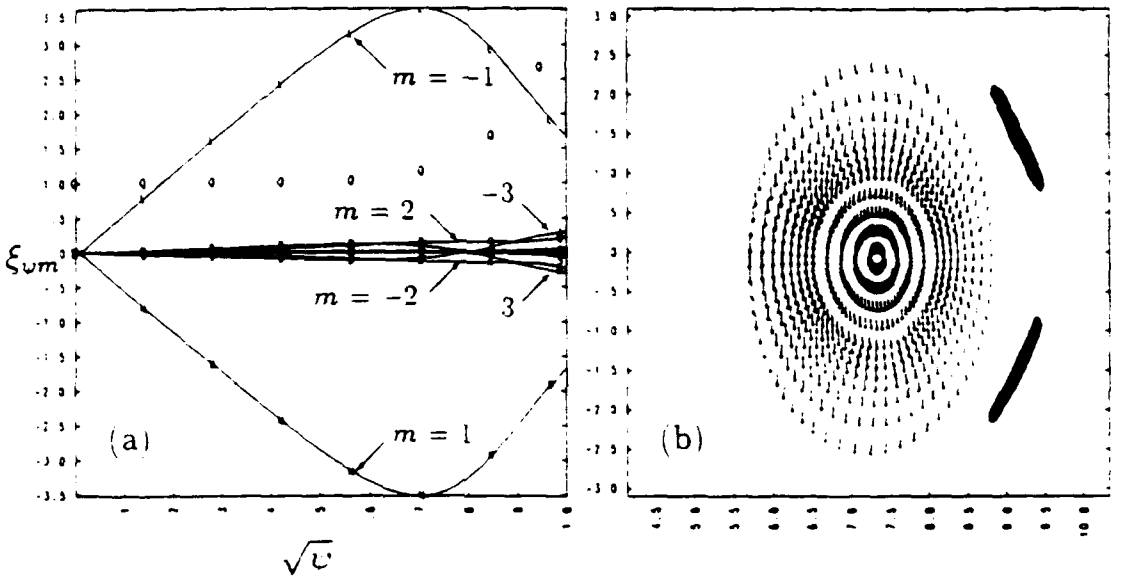


Figure 6.13: (a) *Fourier components of the radial displacement* ξ_{ω} *-vs-,* $\sqrt{v}$ *for the the elliptical plasma with conducting plates at about 45° off the midplane on the outboard side.*

(b) *Displacement vectors showing the motion of the unstable plasma. The plasma is partially stabilized by conducting plates on the outboard side of the plasma. Note the deformation of the plasma motion as it tries to move around the plates. Note also the difference in the deformation compared with that of Fig. 6.12.*

$\Theta \approx 3\pi/4, 5\pi/4$ (Fig. 6.13). It is clear that the plasma is moving with respect to the plates in such a way as to reduce the stabilizing effect of the plates. The magnitude of the displacement seen in Fig. 6.12 is small on the inboard side near the plates and can be seen to be moving away from the plates.

Figure 6.13 shows the instability with the passive plates on the outboard side at $\Theta \approx -\pi/4$ (second point from the left on the curves in Fig. 6.10). The deformation of the eigenfunction is clearly different from that of Fig. ??. This difference can be traced to a change in magnitude of the $m = 3$ component and a change in *sign* and magnitude of the $m = 2$ component. The displacement vector plot shows the plasma is again displaced in such a way that the stabilizing effect of the plate is reduced the *left*). The magnitude of the displacement is also seen to be very small on the outboard side near the plates while it is larger elsewhere.

In summary, we see that different m-components of the eigenfunction induce different poloidal current distributions in a surrounding wall. Conducting elements at different locations around the plasma will induce differing modifications of the non-rigid components of the eigenfunction. These modifications serve to reduce the stabilizing effect of that section of passive conductor. This is done by reducing the stabilizing components and enhancing the destabilizing components, for the given plate position, as much as is energetically favorable. Therefore, the placement of the discrete conducting plates can be of critical importance for a shaped tokamak plasma over and above the relative effect of the outboard side due to toroidal effects.

6.2 Non-rigid effects on the active feedback stabilization of PBX-M

In this section we consider the effects of an active feedback system on the form of the eigenfunction and how this affects the overall stability. For this we consider a PBX-M equilibrium similar to that used in Chapter 2.

In Chapter 2 a review was given of a numerical TSC calculation of the active feedback stabilization of the axisymmetric instability in the PBX-M tokamak. It was demonstrated that different flux-loop locations—which measure equally well the plasma displacement in the passive sense—would not work equally well in stabilizing the axisymmetric motion given the same active feedback coils and gain law. In par-

Plasma Current I_p	567.4 kA
Major Radius R_0	1.635 m
Minor Radius a	0.308 m
Elongation $\kappa(95\%)$	1.951
Toroidal Field $B_T(0)$	1.20 T
$q(95\%)$	2.51
β	0.02
$n_e(0)$	3.35×10^{19} m^{-3}

Table 6.1: *Equilibrium parameters of PBX-M plasma used in the active feedback stabilization study. This equilibrium corresponds to a slight modification of the equilibrium of experimental shot #226879.*

ticular, it was shown that the flux loop pair on the inboard side was ineffective in stabilizing the vertical instability regardless of the value of the gain. The outboard pair, however, could be used successfully to stabilize the plasma.

This was demonstrated using Nyquist techniques, and the proposed explanation was that if the plasma were unstable enough, it would be able to deform under the influence of the active feedback (i.e., modify the eigenfunction) in such a way that the flux difference measurement at the flux loops could be made so close to zero that the active feedback system would be rendered ineffective. The feedback system would operate normally, but the flux-loop measurements would be useless owing to the plasma deformation. It was impossible, however, to *explicitly* demonstrate this conjecture using TSC and the other analysis methods used. The NOVA-W code, on the other hand, is ideally suited for examining this problem, and we will show how the active feedback system can induce a modification of the PBX-M eigenfunction in such a way as to make the active feedback system ineffective for certain flux-loop locations, and how it will minimize the stabilizing effect of the feedback system for any flux-loop configuration.

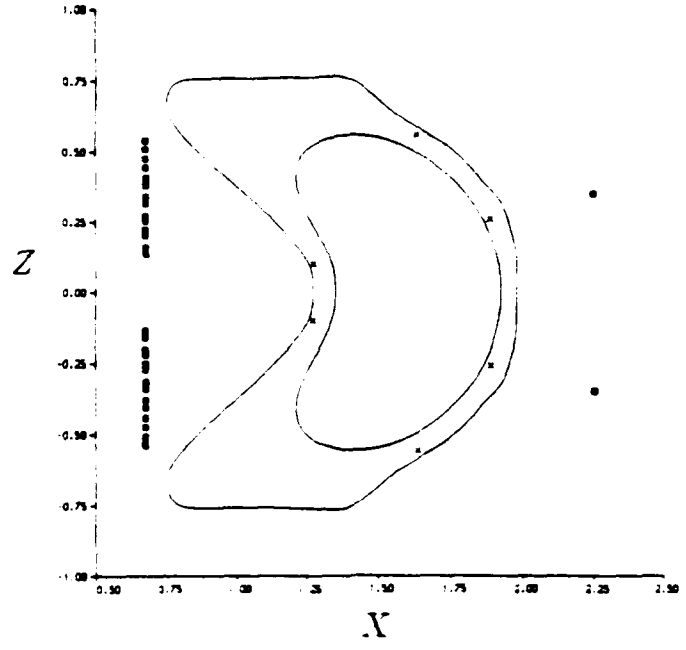


Figure 6.14: *The PBX-M plasma boundary, resistive wall contour, active feedback coils, and three sets of flux observation loops (inboard pair, centered-outboard pair, and far-outboard pair) to be used in these calculations.*

6.2.1 Active feedback stabilization of PBX-M using the inboard flux loops

We use a PBX-M equilibrium with parameters that correspond to experimental shot #226879. The equilibrium parameters are listed in Table 6.1. The equilibrium used here was taken from a time-dependent TSC simulation in which the actual experimental coil currents from this shot were used for the simulation, and the TSC results were compared with the actual magnetics data [74]. To produce the modified equilibrium used in these calculations, the vertical field was increased in order to move the plasma inward, away from the outboard stabilizing plates and toward higher negative field index. This makes the equilibrium much more vertically unstable than the original experimental equilibrium. Such an inward radial shift could be caused in an experiment by a loss of thermal energy or redistribution of current [43]. It should be noted that this PBX-M equilibrium is different from the one used in the TSC calculations of Ref. [50] and reviewed in Chap. 2.

Figure 6.14 shows the equilibrium plasma boundary, the PBX-M wall contour (the wall is composed of the series of passive stabilizing plates of high conductivity used

in the PBX-M device with connecting regions of very high resistivity which represent the axisymmetric gaps between the passive stabilizers, the active feedback coils, and the three sets of flux loops to be used. We consider first the case of the inboard flux loops, which were found to be ineffective for active feedback stabilization. The inboard set of flux loops will be used in the first calculation. A proportional gain law was used in this calculation. Figure 6.15 shows the instability growth rate vs. normalized feedback gain, and selected component ratios (ξ_{w2}/ξ_{w1} and ξ_{w4}/ξ_{w1}) vs. gain.

As the feedback gain is increased from zero, the growth rate drops rapidly, indicating that the feedback system is operating properly. The components of the eigenfunction remain fairly constant with respect to the gain. At higher gain (approximately $\alpha \geq 1.5$), however, we see that the $m = 2$ component of the eigenfunction changes significantly. It becomes less negative, then positive, and then rapidly increases in magnitude with increasing gain. In the same region of gain space where we see the sudden a rapid increase of ξ_{w2}/ξ_{w1} we also notice that the growth rate γ starts to level off. The growth rate has approached marginal stability and does not appear to become more stable at high values of feedback gain. At $\alpha_g = 6.0$, double the maximum gain shown in Fig. 6.15, the growth rate is still at marginal stability. It is virtually unchanged (only very slightly smaller) from the gain at $\alpha_g = 3.0$. At these high values of gain the active feedback is no longer effective in providing additional stabilization to the plasma. Instead, the eigenfunction is changing in form, thus maintaining the instability. The ratio ξ_{w4}/ξ_{w1} also changes significantly as the feedback gain is increased—it roughly quadruples in magnitude in this range of the feedback gain.

Figures 6.16–6.18 show the perturbed flux contour plots at three different values of normalized gain ($\alpha_g = 1.0, 2.0, 3.0$) spanning the range of Fig. 6.15. Examining these perturbed flux plots gives us some insight into how the eigenfunction modification changes the effectiveness of the feedback system.

Figure 6.16 shows the perturbed flux contours when the normalized gain $\alpha_g = 1.0$. Referring to Fig. 6.15 we see that the eigenfunction is nearly identical to the form it takes with no active feedback—the ratio ξ_{w2}/ξ_{w1} is virtually unchanged, and ξ_{w4}/ξ_{w1} is only slightly more negative. We see that the perturbed flux contours from the plasma are fairly equally weighted on both sides of the plasma. The zero contour is distant from the flux loops, and the value of the perturbed flux at the flux loops is

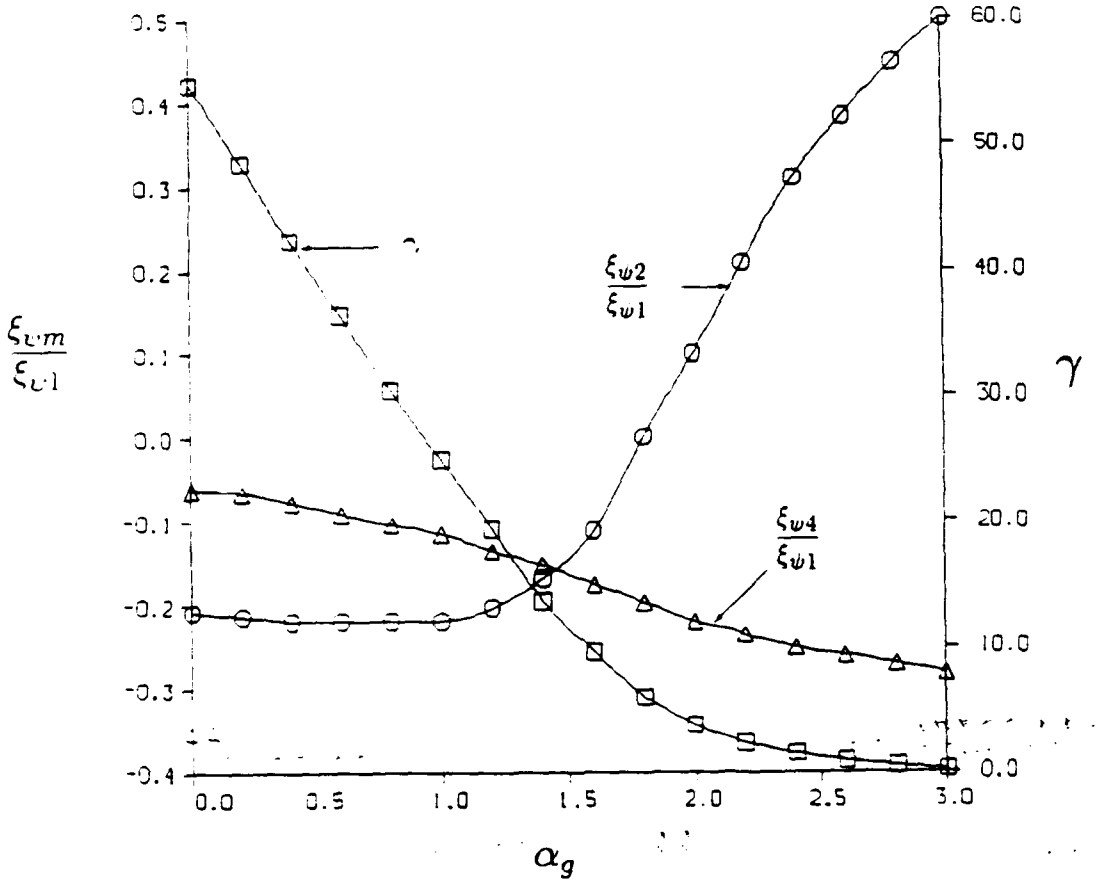


Figure 6.15: *Growth rate γ and variation in $m > 1$ components vs. feedback gain for flux loops on the inboard side. The ratios of the $m = 2, 4$ components of the eigenfunction to the $m = 1$ component at the plasma edge are graphed as a function of the feedback gain. In particular, the $m = 2$ (circles) and the $m = 4$ (triangles) contributions are shown here. Also shown is the growth rate γ (squares) as a function of the gain.*

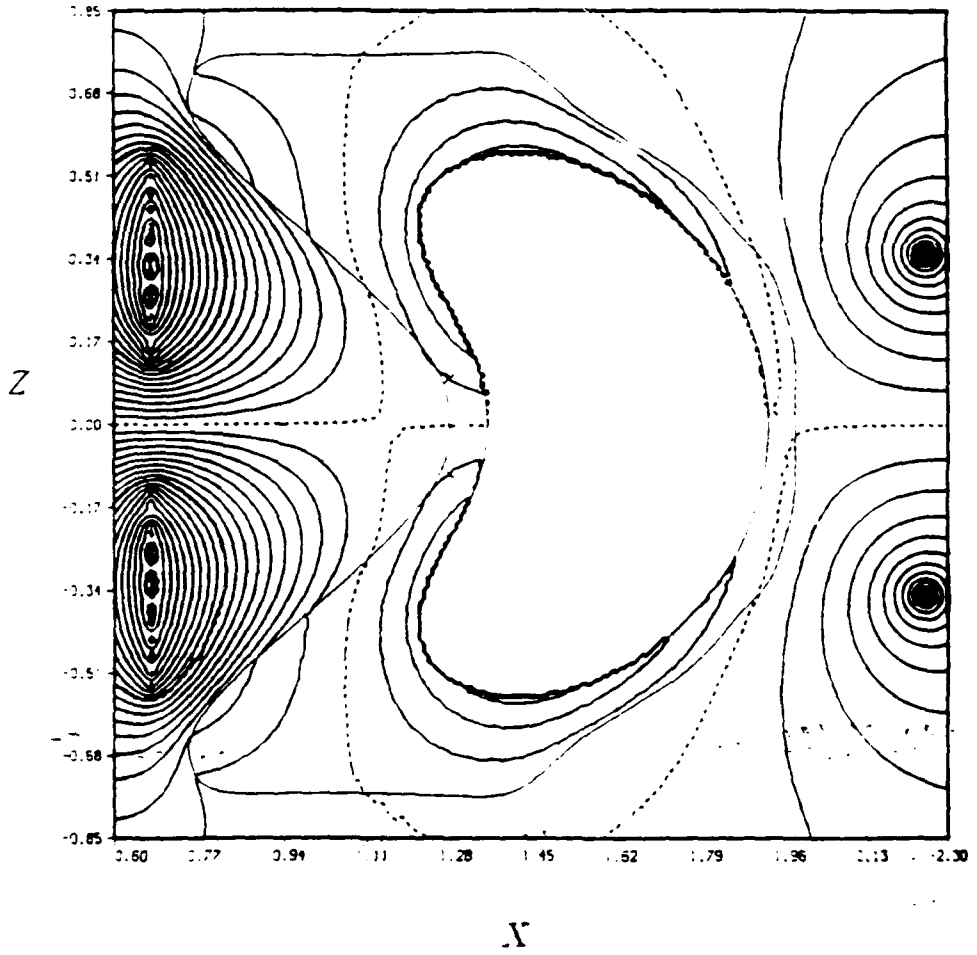


Figure 6.16: *Perturbed flux contour plots for PBX-M with active feedback using the inboard flux loops, and normalized feedback gain* $\alpha_g = 1.0$. *The zero-flux contours are shown as dashed lines. The flux loops are shown by ' × ' symbols.*

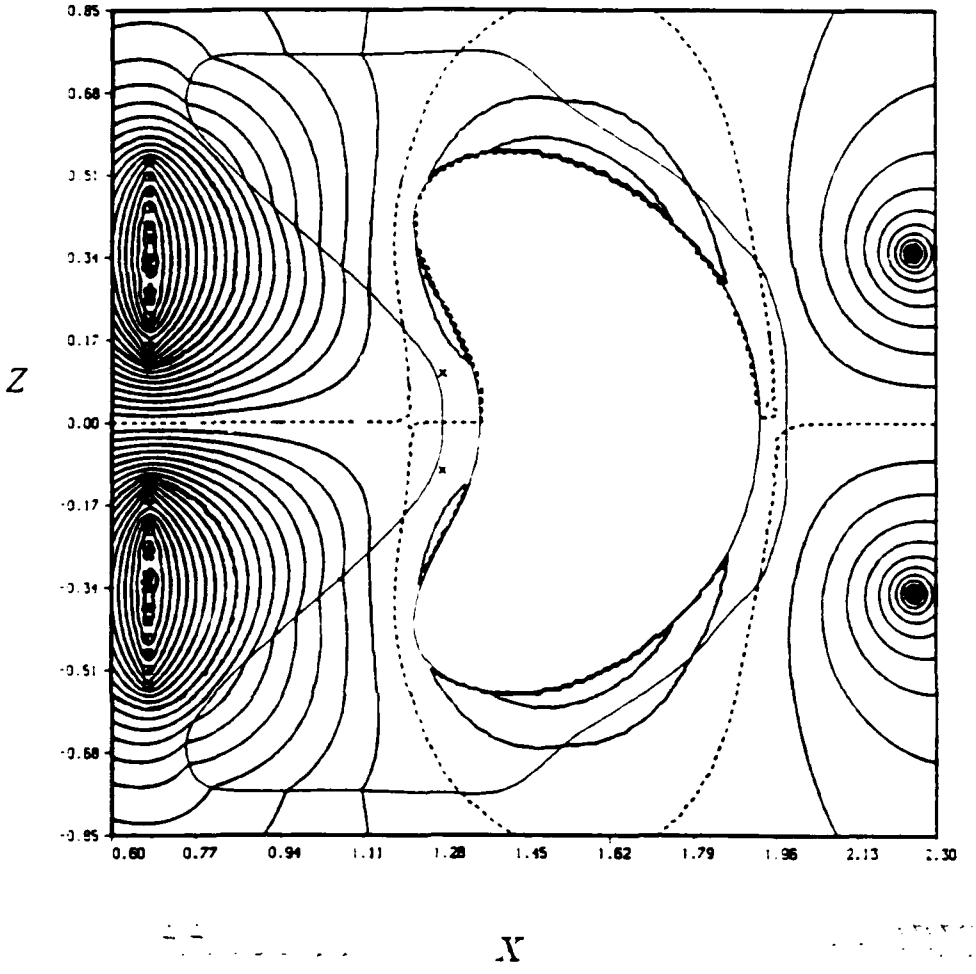


Figure 6.17: *Perturbed flux contour plots for PBX-M with active feedback using the inboard flux loops, and normalized feedback gain* $\alpha_g = 2.0$. *The zero-flux contours are shown as dashed lines. The flux loops are shown by ' × ' symbols.*

relatively large.

Figure 6.17 shows the perturbed flux contours when the normalized gain $\alpha_g = 2.0$. Referring to Fig. 6.15 we see that the eigenfunction has undergone considerable modification at this value of gain. In particular, the $m = 2$ component is quite different—and has even changed sign. One can see from the plot of the perturbed flux contours that the contours of plasma flux have become shifted toward the outboard side, and the value of perturbed flux on the inboard side near the flux loops has been greatly reduced. The null contour has moved closer to the flux loops. We see

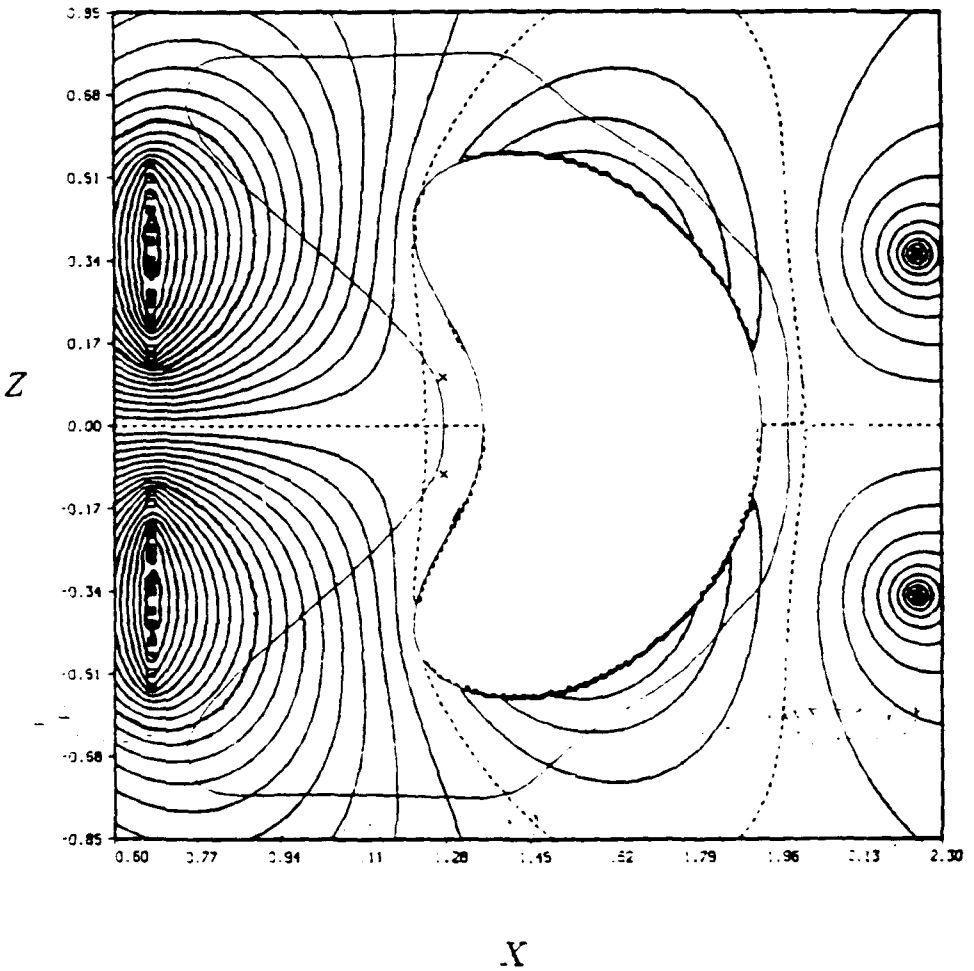


Figure 6.18: *Perturbed flux contour plots for PBX-M with active feedback using the inboard flux loops, and normalized feedback gain* $\alpha_g = 3.0$. *The zero-flux contours are shown as dashed lines. The flux loops are shown by* ' × ' *symbols.*

from Fig. 6.15 that at this point the growth rate has already begun to level off with respect to increasing feedback gain. The change in the form of the perturbed flux contours from the plasma indicates a change in the unstable plasma motion toward the outboard side.

Figure 6.18 shows the perturbed flux contours when the normalized gain $\alpha_g = 3.0$. This is the highest gain value shown in Fig. 6.15. The plasma deformation is quite large at this point, especially with regard to the $m = 2$ component. The perturbed flux contours also reflect the considerable deformation of the plasma. The contours from the plasma are heavily weighted toward the outboard side. The perturbed flux indicates that the unstable motion is now more of a vertical motion instead of the motion toward the separatrix (in the direction of the tip of the bean) that is characteristic of the instability with little or no active feedback—see Fig. 6.16. Notice also that the perturbed flux on the inboard side, near the flux loops, is almost zero. There are no contours of plasma flux seen on the inboard side, and the deformation of the eigenfunction has allowed the null contour to move closer to the flux loops. The measured flux difference at the flux loops is now very small, therefore the flux loops are insensitive to further vertical displacement. The feedback system is rendered ineffective because it can no longer measure and feed-back on the deformed vertical motion.

It should be noted that although the case we examined in this section had zero derivative gain, an increase in the derivative gain was found to have no significant effect on the stabilization. The growth rate at large gain α_g is still at the marginal stability limit. This is not surprising, since the results show the oscillation frequency ω_r to be zero. Therefore there is no overshoot and no oscillation. The ineffectiveness of the feedback using these flux loops is due solely to the eigenfunction deformation.

6.2.2 Active feedback stabilization using the centered-outboard flux loops

We next consider feedback stabilization using the centered-outboard pair of flux loops of Fig. 6.14, $(X_o, Z_o) = (1.64, \pm.56)$. This flux-loop pair location corresponds most closely to the actual flux loops used for vertical control in the PBX-M experiment. Figure 6.19a shows the growth rates vs. proportional gain α_g for three different values of derivative gain β_g. Figure 6.19b shows the growth rate γ vs. oscillation fre-

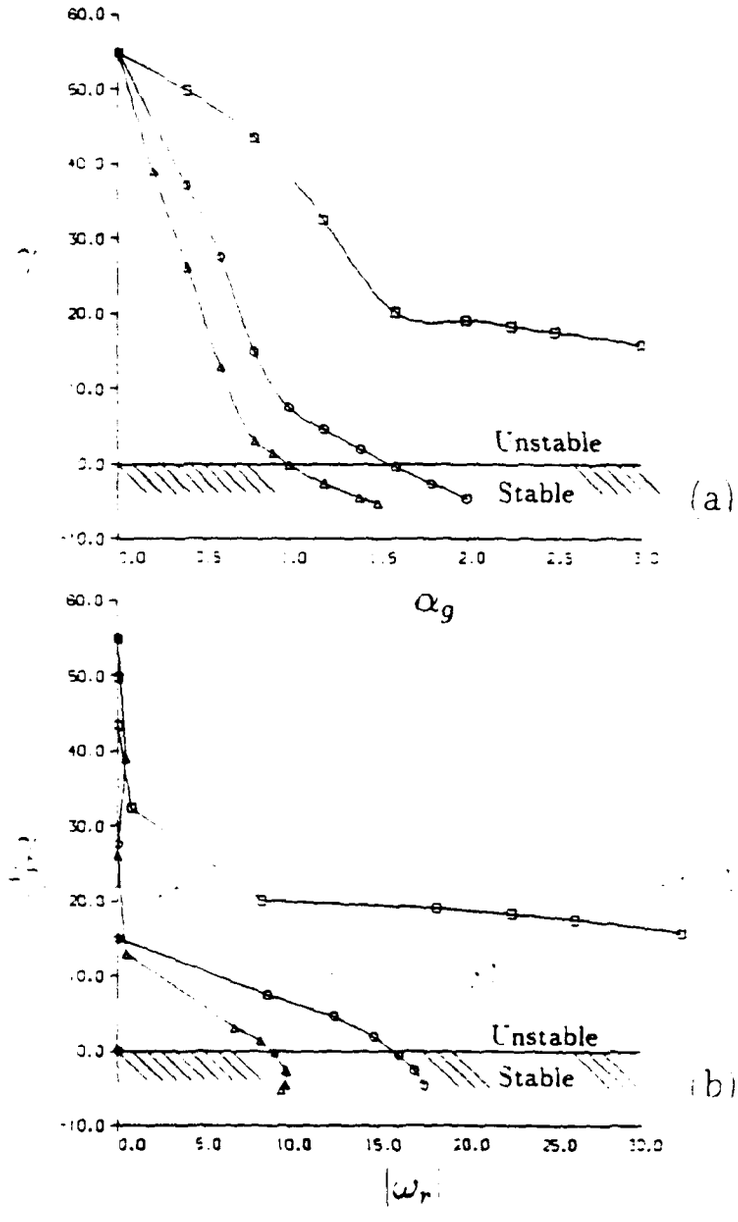


Figure 6.19: (a) *Growth rate vs. gain* α_g *for* $\beta_g/\alpha_g = 0$. *(squares),* $\beta_g/\alpha_g = 0.05s^{-1}$ *(circles),* $\beta_g/\alpha_g = 0.10s^{-1}$ *(triangles).*

(b) *Growth rate vs.* ω_r *for the same three values of* β_g.

quency ω_r for the same three values of derivative gain. It can be seen that with zero derivative gain the axisymmetric mode cannot be stabilized. The growth rate can be reduced to about 20 s^{-1} at $\alpha_g = 1.5$, but further increases in gain do not appreciably lower the growth rate while they do however, significantly increase the oscillation frequency ω_r. Large proportional gain is driving a large overshoot that leads to oscillations. Clearly some derivative gain is necessary. By increasing the derivative gain to $\beta_g/\alpha_g = 0.05s^{-1}$ the plasma can be stabilized and the oscillations significantly reduced. A further increase in derivative gain to $\beta_g/\alpha_g = 0.10s^{-1}$ decreases the oscillation frequency even more and allows stabilization to occur at a lower value of proportional feedback. Unlike the inboard flux-loop pair, this pair of flux loops does allow for adequate stabilization of the plasma provided the derivative gain is large enough. This agrees with results obtained with TSC for this same equilibrium [74]. Figures 6.20 and 6.21 show the variation of the $m = 2,4$ components of the eigenfunction with respect to feedback gain α_g for the three values of derivative gain.

Figure 6.20a shows these results for the case with no derivative gain. We see a strong reduction in the growth rate γ with increasing gain α_g until the gain reaches $\alpha_g \approx 1.5$. At this point the γ curve levels off and does not stabilize much with further increase in gain. Furthermore, we see ω_r increase rapidly from zero beginning at $\alpha_g \approx 1.25$. The oscillation frequency ω_r increases steadily with increasing gain, while γ no longer decreases by any significant amount. This demonstrates that the restoring force from the feedback system is driving the oscillation instead of stabilizing the plasma. Clearly derivative gain is required.

We also see some modification of the $m = 2,4$ components of the eigenfunction in Fig. 6.20a. The $m = 2$ component ratio $\xi_{\psi 2}/\xi_{\psi 1}$ begins to decrease (increase in negative magnitude) as the gain is increased above $\alpha_g \approx 1.5$. Note that this change is opposite to the change in $\xi_{\psi 2}/\xi_{\psi 1}$ from Fig. 6.15 for the inboard flux loops. We found that when with the inboard flux loops were used, $\xi_{\psi 2}/\xi_{\psi 1}$ changed sign and grew to a large (positive) value with increasing feedback gain. This caused the perturbed flux contours from the plasma to be shifted toward the outboard side away from the flux loops. This left a relatively small value of perturbed plasma flux at the inboard flux loops. In the present case, when the centered-outboard flux loops are used to control the feedback system, we see the opposite effect. This implies that the perturbed flux on the outboard region near these flux loops is somewhat reduced.

Figure 6.22 shows the perturbed flux contours for the active feedback at gain $\alpha_g =$

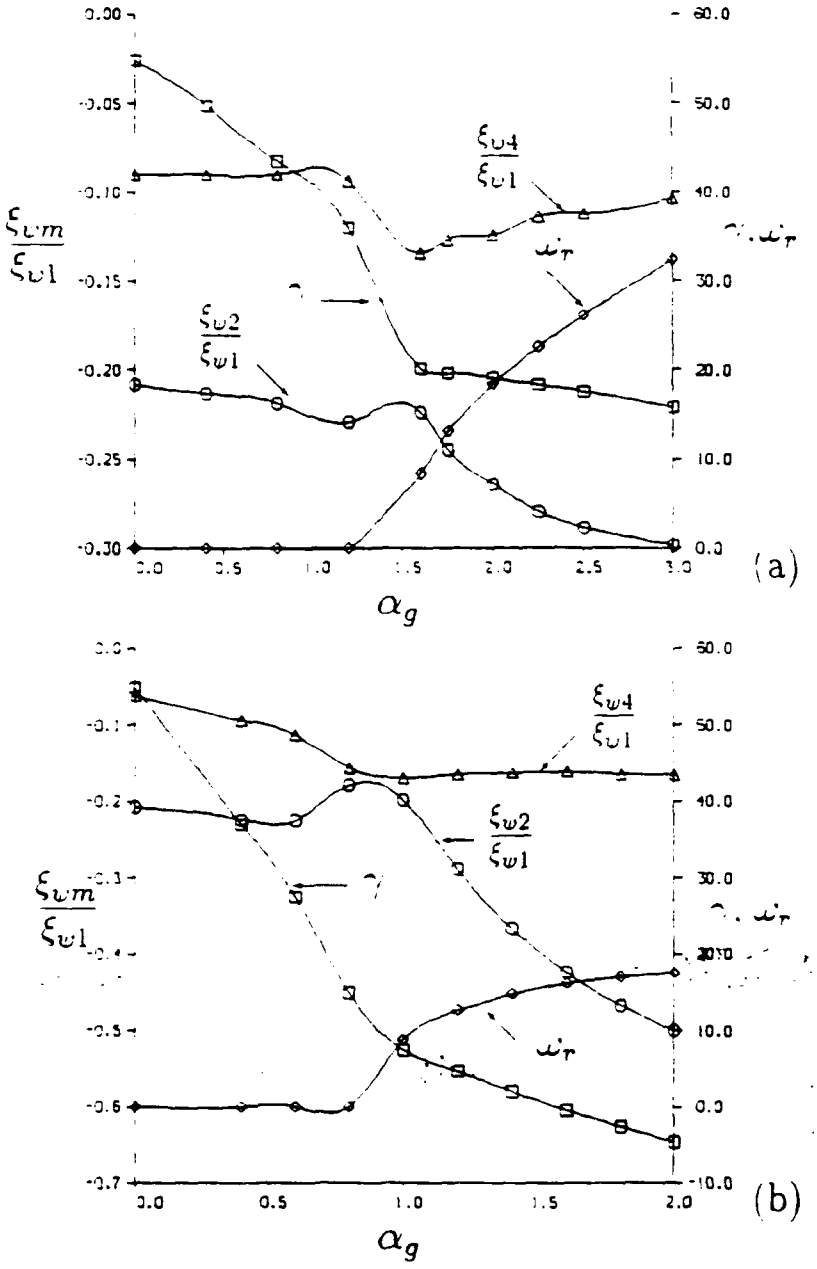


Figure 6.20: *Growth rate γ, oscillation frequency ω_r, and variation in $m > 1$ components vs. feedback gain for centered flux loops. The ratios of the $m = 2, 4$ components of the eigenfunction to the $m = 1$ component at the plasma edge are graphed as a function of the feedback gain. In particular, the $m = 2$ (circles) and the $m = 4$ (triangles) contributions are shown here. Also shown is the growth rate γ (squares) and oscillation frequency ω_r as a function of the proportional gain for:*
(a) $\beta_g/\alpha_g = 0$.
(b) $\beta_g/\alpha_g = 0.05$

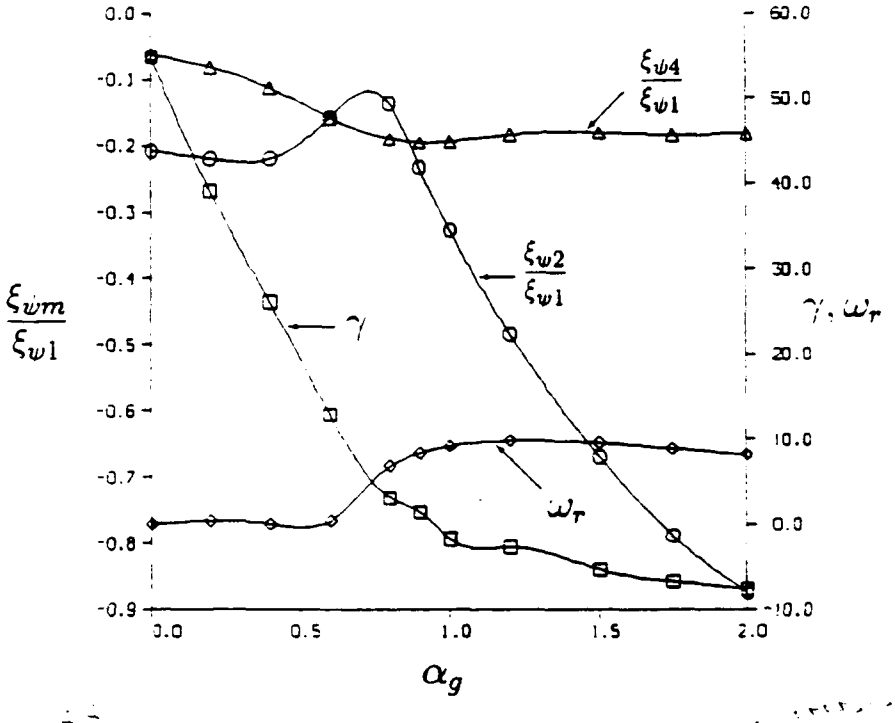


Figure 6.21: *Growth rate γ, oscillation frequency ω_r, and variation in $m > 1$ components vs. feedback gain for centered flux loops. The ratios of the $m = 2, 4$ components of the eigenfunction to the $m = 1$ component at the plasma edge are graphed as a function of the feedback gain. In particular, the $m = 2$, (circles) and the $m = 4$ (triangles) contributions are shown here. Also shown is the growth rate γ (squares) and oscillation frequency ω_r as a function of the proportional gain for $\beta_g/\alpha_g = 0.10\mathrm{s}^{-1}$.*

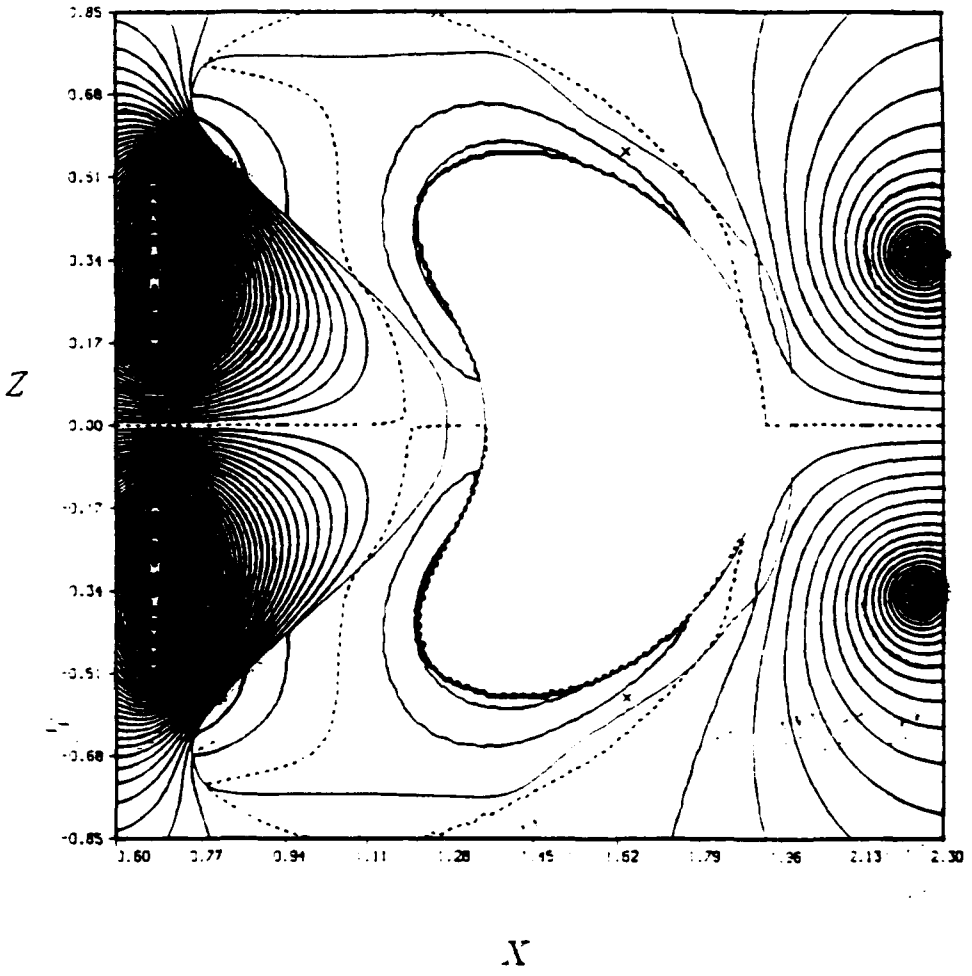


Figure 6.22: *Perturbed flux contour plots for PBX-M with active feedback using the centered flux loops, and normalized feedback gain* $\alpha_g = 2.25$. *The zero-flux contours are shown as dashed lines. The flux loops are shown by '×' symbols.*

2.25 using the centered-outboard flux loops. Careful examination and comparison with Fig. 6.16 shows that the perturbed flux near the flux loops is reduced from the case with little or no feedback gain. The plasma eigenfunction is again deformed in such a way to reduce the effectiveness of the feedback system using this particular pair of flux loops. However, the deformation is clearly very different from the inboard flux-loop case—as can be seen by comparing Fig. 6.22 and Figs. 6.16–6.18. The only difference between these two cases is the location of the flux loops, and this difference leads to vastly different plasma deformations.

Figure 6.20b shows the results for the case using the centered-outboard flux loops with $\mathcal{J}_g/\alpha_g = 0.05s^{-1}$ derivative gain. In this case the plasma can be stabilized with large enough feedback gains. The oscillation frequency is less than in the zero derivative gain case, although it is still a significant fraction of that case and is increasing with increasing gain.

We also see a larger change in $\xi_{\psi 2}/\xi_{\psi 1}$ with respect to increasing gain α_g than in the zero derivative gain case. After a small decrease in negative magnitude at $\alpha_g \approx 0.75$, the value of $\xi_{\psi 2}/\xi_{\psi 1}$ begins to get more negative with increasing gain. This change in $\xi_{\psi 2}/\xi_{\psi 1}$ corresponds to a change in the slope of the γ vs. α_g curve. As $\xi_{\psi 2}/\xi_{\psi 1}$ begins to change rapidly, the decrease in γ with respect to α_g lessens. Again, there is a plasma deformation that is reducing the effectiveness of the feedback system with the flux loops at this location. This modification is not enough to keep the feedback system from stabilizing the vertical instability, but it does reduce the effectiveness as shown by the change in slope of the γ vs. α_g curve.

Finally, Figure 6.21 shows the results for the case with $\mathcal{J}_g/\alpha_g = 0.10s^{-1}$ derivative gain. In this case the oscillations have been much reduced, and the plasma is stabilized at a lower value of α_g. We again see a large change in $\xi_{\psi 2}/\xi_{\psi 1}$ with respect to the gain. There is a rapid decrease in $\xi_{\psi 2}/\xi_{\psi 1}$ with respect to α_g starting at $\alpha_g \approx 0.75$. This, again, is strongly correlated with the sudden change in the slope of the γ vs. α_g curve. The γ vs. α_g curve levels off in the region where the eigenfunction is significantly deformed. This corresponds to about the point where the mode becomes stable, so the deformation is not enough to keep the plasma unstable when these flux loops are used, but it does keep the feedback system from stabilizing the motion any further, as witnessed by the sudden change in slope of the γ vs. α_g curve.

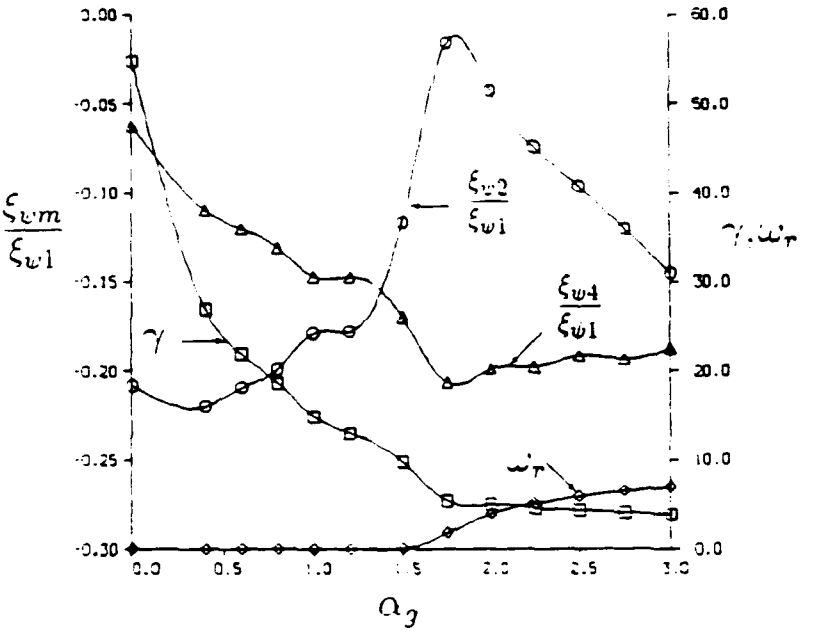


Figure 6.23: *Growth rate γ, oscillation frequency ω_r, and variation in $m > 1$ components vs. feedback gain for centered flux loops. The ratios of the $m = 2, 4$ components of the eigenfunction to the $m = 1$ component at the plasma edge are graphed as a function of the feedback gain. In particular, the $m = 2$ (circles) and the $m = 4$ (triangles) contributions are shown here. Also shown is the growth rate γ (squares) and oscillation frequency ω_r as a function of the proportional gain for $\beta_g/\alpha_g = 0.10 s^{-1}$.*

6.2.3 Active feedback stabilization using the far-outboard flux loops

Next we consider active feedback using the far-outboard flux loops shown in Fig. 6.14. Figure 6.23 shows the results using this pair of flux loops with a gain law that includes derivative gain of $\beta_g/\alpha_g = 0.10 s^{-1}$. We see that even with this derivative gain and the correspondingly low values of ω_r shown in Fig. 6.23, the plasma cannot be stabilized beyond a certain point ($\gamma \approx 4\ s^{-1}$). Figure 6.23 shows a large initial drop in γ with increasing gain. This stabilization begins to level off, however, at $\alpha_g \approx 1.75$.

The component ratios $\xi_{\psi 2}/\xi_{\psi 1}$ and $\xi_{\psi 4}/\xi_{\psi 1}$ show a significant and varied deformation of the eigenfunction with respect to increasing gain. Initially there is a sharp rise

in ξ_{w2}/ξ_{w1} toward less negative values. This mimics the rapid rise in ξ_{w2}/ξ_{w1} shown in Fig. 6.15 for the case using the inboard flux loops. In that case, ξ_{w2}/ξ_{w1} moves to positive values and continues to increase in magnitude. In Fig. 6.23 we see such an initial rise in ξ_{w2}/ξ_{w1}, but then a sharp reversal occurs at $\alpha_g \approx 1.75$, followed by a rapid decrease of ξ_{w2}/ξ_{w1} (toward more negative values) at higher gain. This is paralleled by a similar, but much less dramatic, change in ξ_{w4}/ξ_{w1} at about the same point. The γ vs. α_g curve levels off at $\alpha_g \approx 1.75$, near the point where we see the sudden changes in the eigenfunction.

Figure 6.24 shows the perturbed flux contours for this case with the gain $\alpha_g = 2.0$ and the derivative gain $\beta_g/\alpha_g = 0.10s^{-1}$. We see from this figure that the null contour lies almost directly on the flux loops. This seems to be a natural consequence of the geometry of the feedback system and this particular inward-shifted equilibrium. Note from Fig. 6.16 (inboard flux loops, gain $\alpha_g = 1.0$) that the null contour on the outboard side is in about the same place even though the flux loops are on the inboard side in this case. However, in the case using the inboard flux loops, we see that at higher values of gain the eigenfunction deformation is such that the perturbed flux contours are shifted strongly toward the outboard side. This, in turn, pushes the null contour away from the plasma.

If this same deformation of the eigenfunction were to occur when the flux loops are on the outboard side of the plasma, then the feedback system would be very effective owing to the large measurable perturbed plasma flux at the flux-loop locations (see Figs. 6.17–6.18). What we do see when the flux loops are on the outboard side is a deformation initially similar to the previous case, but then there is a sudden reversal in the plasma deformation at the point where continued deformation would move the null contour away from the flux loops. Instead, the eigenfunction is modified so as to keep the null contour on the flux loops and reduce the stabilizing effect of the feedback system.

6.3 Summary

We have seen in the preceding section that the eigenfunction of the axisymmetric mode for the PBX-M plasma will modify itself under the influence of an active feedback system to provide the weakest signal possible to the flux loops that measure the plasma displacement. This compromises the stabilizing effect of the active feedback

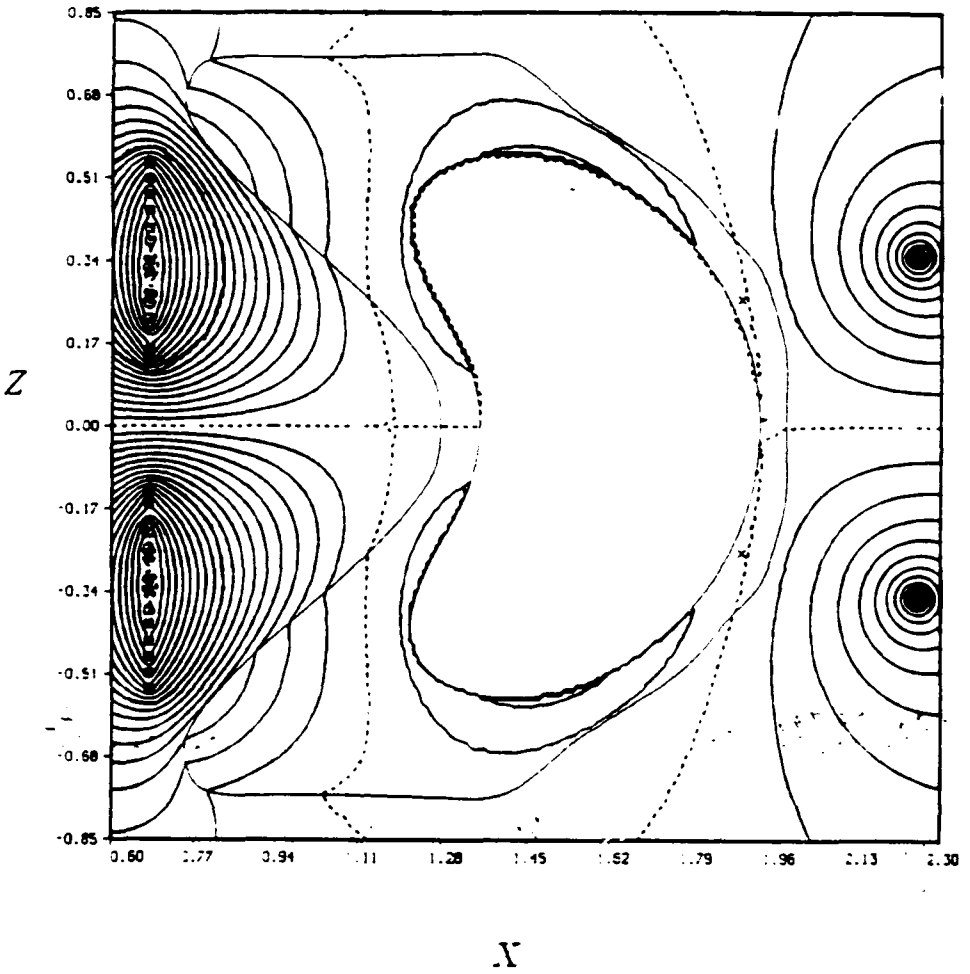


Figure 6.24: *Perturbed flux contour plots for PBX-M with active feedback using the outboard flux loops, and normalized feedback gain* $\alpha_g = 2$. *The zero-flux contours are shown as dashed lines. The flux loops are shown by* ' × ' *symbols.*

system, and can in some cases leave the feedback system so ineffective that stability cannot be achieved regardless of the strength of the feedback gain. We examined, in particular, three possible locations for placement of the flux loops that control the feedback, and in each case we see a *different* plasma deformation that leads to a reduced flux signal at the flux loops. These eigenfunction deformations do not leave the plasma unstable in every case, but they do reduce the stabilizing effect of the particular feedback configuration. We have not considered feedback on a flux difference signal with the active coil contributions to the flux subtracted out. This might improve the performance of some flux-loop combinations, but given the strong variation in the deformations of the eigenfunctions observed in the three cases considered here, it seems very likely that the plasma deformation will change in such a way as to minimize the sensitivity of this new measurement of vertical displacement. This will be an interesting case to consider in the future.

In Sec. 6.1 we examined plasma deformations with passive feedback only. Here we found that for various positions of discrete passive conducting plates, the eigenfunction will be modified in such a way as to minimize the stabilizing eddy currents in those plates. Different plate orientations lead to varying modifications of the eigenfunction, but each seems tailored to reducing the stabilizing effect of the particular passive conductors.

The plasma equilibria with significant triangularity (CIT and particularly ARIES-I equilibria) show the strongest deformation. However, even the finite-aspect-ratio ellipse examined in Sec. 6.1.4 showed significant deformation. In the case of the ellipse, the deformations reflected a certain symmetry with respect to conductor position relative to a vertical line through the plasma center (except for toroidal effects). The triangular (dee-shaped) plasmas, on the other hand, show a strong asymmetry, which favors plate position in the outboard region above and beyond what would be expected based on toroidal effects alone. An inverse-dee shaped plasma would undoubtedly show quite different behavior.

The effects due to the deformability of the plasma can therefore play an important role in determining the effects of passive and active feedback stabilization of highly shaped tokamak plasmas. This should be an important consideration in the design and operation of such tokamaks in the future.

Chapter 7

Summary and discussion

7.1 Summary

We have developed a linear MHD stability code to examine the feedback stabilization of deformable tokamak plasmas by passive resistive conductors and active feedback currents in the vacuum region surrounding the plasma. This code, NOVA-W, is a modification of the linear, ideal MHD stability code NOVA. The NOVA code uses a non-variational approach to solve the linear MHD stability equations. Because it does not use the energy principle, we are not constrained by self-adjointness properties. Therefore we can add non-ideal effects such as resistive conductors and an active feedback system.

The vacuum calculation has been modified to a formulation based on perturbed poloidal flux. This allows the representation of active feedback currents in the vacuum region, and the effects of eddy currents in the resistive wall can be represented by a jump condition in the normal derivative of the perturbed flux in accordance with a thin wall approximation. The flux formulation also makes it straightforward to use flux difference measurements to represent the vertical displacement in agreement with experimental methods. A Green's function formulation relates the perturbed flux and the normal derivative of the flux on the plasma and resistive wall surfaces through a series of surface integral equations. The boundary condition at the plasma-vacuum interface relates the perturbed pressure at the surface to the normal derivative of the flux. The perturbed flux, in turn, is related to the radial component of the instability displacement. This provides the necessary boundary condition for the solution of the linear MHD stability equations.

The effects of a resistive wall are calculated using a thin wall approximation. Variations in the wall thickness and/or resistivity are easily accounted for. Even toroidally symmetric gaps can be included by increasing the resistivity in the calculation by several orders of magnitude.

The effects of active feedback currents are calculated with a set of active feedback matrices which represent the relation of the currents to the Green's function integrals over the boundary surfaces. The magnitude of the feedback currents are proportional to the difference between the flux measured at two points symmetrically above and below the midplane. This flux difference is a measure of the vertical displacement of the plasma. Additional magnetics measurements such as the poloidal field at various points can be added to the feedback law, if desired. The feedback gain relates these measurements to the desired currents. An additional term proportional to the time derivative of the flux difference is usually included. The form of the feedback matrices differs depending on whether the flux loops are inside or outside the resistive wall. Equations for the active feedback circuits can be included in the formulation of the feedback matrices. A time delay can be included in the feedback law to represent the time response of power supplies. The last two features have not been used to any extent in the results presented here.

The code has been tested in the case of passive stabilization against an analytic model in the limit of an infinite aspect ratio for a simple elliptical plasma. The comparison is excellent. Another comparison is performed for a realistic tokamak configuration using the Tokamak Simulation Code (TSC). A convergence in the ψ_{rat} parameter of TSC must be performed to obtain an accurate growth rate when using TSC-generated, diverted equilibria. An equilibrium with ψ_{rat} too close to 1.0 yields numerical problems because of the behavior of the magnetic coordinates near the x-point. However, calculating an equilibrium with the plasma surface defined with a smaller ψ_{rat} changes the equilibrium metric quantities at the plasma edge in such a way as to give results reflecting a less unstable equilibrium. Therefore a convergence in ψ_{rat} is needed.

A convergence using ψ_{rat} in the range $0.94 \leq \psi_{rat} \leq 0.96$ will usually give sufficiently accurate results at reasonable numerical cost. Equilibria with higher ψ_{rat} can be used, but the equilibrium calculation must include many surfaces to sufficiently resolve the metric terms near the plasma edge, making it computationally more expensive. It is therefore preferable to perform the convergence at smaller ψ_{rat}. The

comparison to the TSC results is excellent. The improved performance of NOVA-W over TSC allows one to obtain convergence results using up to as 10–20 times less computational time.

The performance and utility of the NOVA-W code is demonstrated in two studies in which the passive growth rate is calculated with respect to incremental variations in the surrounding vacuum vessel structure. The NOVA-W code is well suited for such parameter scans because a good guess of the growth rate (based on results from a neighboring point in the parameter study) reduces the computational effort. A code such as TSC, on the other hand, must perform the same time consuming calculation at each point in the parameter scan regardless of results at neighboring points. Furthermore, NOVA-W can accommodate calculations for plasma-conductor configurations in which the plasma is very poorly stabilized and the instability growth rate is approaching the ideal time scale. Obtaining converged results for such a configuration would prove very expensive using TSC because of the necessity of reducing the time step to the ideal time scale.

The NOVA-W code has been applied to the study of active feedback of the CIT tokamak design. The study focused on the comparison of effectiveness of the feedback system with regard to flux-loop location within the vacuum vessel. The results compare favorably to a similar study previously undertaken using the TSC code. It was seen that the sensitivity of the flux measurements varied with respect to the location at various points along the inside wall of the vacuum vessel. This greatly changes the efficacy of the feedback system. At some locations with apparently reduced sensitivity of the flux measurements to plasma position, an increase in derivative gain can improve performance. At another location it was found that no combination of gains could stabilize the plasma. It may be possible to improve the performance of the feedback system by subtracting the calculated effect of the active feedback coils at the flux-loop position, but this analysis was not done here.

We have used the NOVA-W code to examine the deformable aspects of the plasma on the feedback stabilization (passive and active). For the case of passive feedback stabilization using discrete conducting plates it was found that the growth rate was sensitive to the poloidal position of the plate. It was also seen that as the conducting plates are moved to different positions around the plasma, the unstable eigenfunction is modified. This modification (or deformation) of the non-rigid components of the motion serves to reduce the stabilizing effect of the conductors at the given position.

The higher m components of the eigenfunction are modified such that the stabilizing eddy currents in the conductors are reduced. For a given poloidal position of the conductor it was found that if an $m > 1$ component induced stabilizing eddy currents in that conductor, then the eigenfunction would be modified so that the magnitude of this component is reduced with respect to the $m = 1$ component. Conversely, if the particular $m > 1$ component induces destabilizing eddy currents, then the eigenfunction will be modified to enhance the magnitude of this component with respect to the $m = 1$ component.

We also used the NOVA-W code to study the active feedback stabilization of the PBX-M plasma. Three different locations for the flux-loop pairs were considered while keeping all other aspects of the feedback system the same. It was found that under the influence of active feedback the plasma would modify its eigenfunction in such a way as to reduce the sensitivity of the flux-loop pair to changes in the vertical displacement. For some flux-loop positions the feedback system was made ineffective owing to the deformation. One pair of flux loops was sensitive enough that the plasma deformation could not keep the feedback system from stabilizing the plasma, but the deformation did reduce the stabilizing effect.

7.2 Future work

We will continue to use the NOVA-W code to examine axisymmetric stability of tokamaks with particular interest in the deformable nature of the eigenfunction. In particular, the TCV tokamak would prove to be an interesting subject owing to its strong cross-sectional shaping. The variety of equilibrium configurations and the variety of possible deformations that might therefore arise will prove to make an interesting control problem. The NOVA-W code is ideally suited for such analysis.

An addition to the formulation of the active feedback matrices will allow us to be able to subtract the effects of the coupling of the active feedback coils to the flux loops. This should improve performance of the feedback system for many flux-loop locations. Then the effect of the wall eddy currents alone on the flux-loop measurements can be studied.

Finally, we have seen that for all the equilibria and vacuum configurations considered here (fully symmetric about the midplane), the eigenfunction is purely anti-

symmetric. If the eigenvalue solver can be reformulated so that the poloidal angular dependence is expanded in terms of $\sin(m\Theta)$ instead of $\exp(im\Theta)$, then the overall number of poloidal harmonics M can be reduced by a factor of 2. Since the computation required in the calculation of the eigenvalue and eigenfunction goes as M^3, this section of the code can be speeded up by a factor of 8. And since this part of the calculation typically takes 50% of the total CPU time (depending on the various parameters of the calculation), it would be possible to improve the performance of NOVA-W by a factor of 4. Optimization and improved vectorization will improve the performance even further.

"A very sweet little problem, and I would not have missed it for a good deal."

Sherlock Holmes, *The Adventure of the Beryl Coro*

Appendix A

Self-adjointness considerations for an active feedback system

Since the first two terms in Eq. (2.16) for δW are manifestly self-adjoint, our discussion can be limited to the vacuum contribution.

$$\delta W_v = -\oint d\theta \mathcal{J} \phi_{v1} (\nabla w \cdot \nabla \phi_v) = \int_v (\nabla \phi_v)^2 - \phi_v \nabla^2 \phi_v \, dA. \tag{A.1}$$

The first term on the right-hand side of Eq. (A.1) is again clearly self-adjoint. The second term, which comes from the active feedback currents and which was zero in the model without feedback, must be made explicitly self-adjoint. We define the vacuum flux $\phi_v(r, \theta)$ in a manner similar to that in the previous section, Eq. (2.23),

$$\phi_v(r, \theta) = \sum_{m=0}^{\infty} \phi_v^m(r) \cos m\theta. \tag{A.2}$$

As we see in Fig. 2.6, the feedback coils lie evenly spaced along a circular contour of radius $r = R$; therefore

$$\nabla^2 \phi_v = J_{f.b.} = \sum_{m=0}^{\infty} \frac{1}{\pi R} \delta(r - R) J_0^m \cos m\theta. \tag{A.3}$$

Using Eqs. (A.3) and (A.2), the second term of (A.1) becomes

$$\delta W_v^{(2)} = \sum_{m=1}^{\infty} \phi_v^m(R) \, J_0^m. \tag{A.4}$$

Consider now a feedback law that relates the coil currents to the perturbed flux at the coils according to

$$J_0^m = \sum_{n=1}^{\infty} g_{mn} \, \phi_v^n(R). \tag{A.5}$$

That is, the current in the m^{th} term of the multipolar decomposition is proportional to the n^{th} harmonic of the perturbed vacuum flux at the feedback-coil contour through the constant of proportionality g_{mn}.

Therefore we see that $\delta W_v^{(2)}$ becomes

$$\delta W_v^{(2)} = \sum_{m=1} \sum_{n=1} \phi_v^m(R)\, g_{mn}\, \phi_v^n(R). \tag{A.6}$$

We also see that if the gain matrix g_{mn} is symmetric, i.e.,

$$g_{mn} = g_{nm}, \tag{A.7}$$

then $\delta W_v^{(2)}$, and hence δW_v, is self-adjoint.

The symmetry restriction on g_{mn} constrains the relationship between the feedback-current coefficients L_m and the Fourier coefficients D_m of the perturbed flux on the plasma-vacuum boundary. This relationship is obtained in two stages. In the first stage, the matching conditions for the vacuum flux are applied at the boundary between regions I and II in the vacuum (see Fig. 2.6). In the second stage the matching conditions are applied at the plasma-vacuum interface.

Stage 1

Continuity of ϕ_v at $r = R$ from Eqs. (2.39) and (2.40) gives

$$B_m^I - L_m R^{2m} = B_m^{II} \tag{A.8}$$

There is a jump in the normal derivative of the flux between Region I and Region II due to the feedback-coil currents that are on the interface contour. If we integrate Eq. (A.18) over a cylindrical volume with radius just greater than that of the feedback coil contour $r = R^+$, and subtract the volume integral with radius just inside the contour $r = R^-$, we can calculate the jump condition.

$$\int \nabla^2 \phi_v dA = \oint (\hat{\mathbf{n}} \cdot \nabla \phi_v) dl = \int dA \sum_{m=0}^{\infty} \frac{1}{\pi R} \delta(r - R) J_0^m \cos(m\theta) \tag{A.9}$$

or

$$\oint \frac{1}{r} \frac{\partial \phi_v}{\partial r} d\theta = \int dr \int \frac{1}{r} d\theta \sum_{m=0}^{\infty} \frac{1}{\pi R} \delta(r - R) J_0^m \cos(m\theta) \tag{A.10}$$

$$= \int d\theta \sum_{m=0}^{\infty} \frac{1}{\pi R^2} J_0^m \cos(m\theta).$$

This gives us the jump condition

$$\left[\left[\frac{\partial \phi_v^m}{\partial r}\right]\right] = \frac{\partial \phi_v^m}{\partial r}\bigg|_{r=R^-} - \frac{\partial \phi_v^m}{\partial r}\bigg|_{r=R^-} = \frac{1}{\tau R} J_0^m. \tag{A.11}$$

Therefore Eqs. (2.39) and (2.40), along with this jump condition, give us a second boundary condition:

$$B_m^I - L_m R^{2m} = B_m^{II} - \frac{J_0^m R^m}{m\pi}. \tag{A.12}$$

Thus

$$L_m = -\frac{R^{-m}}{2\pi m} J_0^m. \tag{A.13}$$

Inserting Eq. (A.13) into Eq. (2.39) and using definition Eq. (A.5) gives us a new definition for ϕ_v^I:

$$\phi_v^I(r,\theta) = \sum_{m=1}^{\infty} \left[B_m^I r^{-m} - \left(\frac{r}{R}\right)^m \frac{1}{2\pi m} \sum_{n=1}^{\infty} g_{mn}\, \phi_v^n(R) \right] \cos(m\theta). \tag{A.14}$$

Then solving for $\phi_v^m(r)$ at $r = R$ obtains

$$\phi_v^m(R) = \left[\delta_{mn} - \frac{1}{2\pi m} g_{mn}\right]^{-1} B_n^I R^{-n}. \tag{A.15}$$

Substituting Eq. (A.15) into Eq. (A.5) and then that result into Eq. (A.13) gives us

$$L_m = C_{mn} B_n^I, \tag{A.16}$$

where the matrix C_{mn} is defined by

$$C_{mn} = -\frac{R^{-m}}{2\pi m} g_{mn} R^{-n} \left[\delta_{mn} - \frac{1}{2\pi m} g_{mn}\right]^{-1}. \tag{A.17}$$

Therefore Eq. (A.16) relates the Region I vacuum coefficients B_m^I to the feedback coefficients L_m via the matrix C_{mn}.

Stage 2

The other boundary condition that we must satisfy is the continuity of perturbed flux across the plasma-vacuum interface. This implies that

$$\sum_{m=0}^{\infty} D_m \cos m\theta = \sum_{m=1}^{\infty} B_m^I r^{-m} \cos m\theta - \sum_{m=1}^{\infty} L_m r^m \cos m\theta, \tag{A.18}$$

where

$$r^m = 1 - \frac{m}{2}\alpha \cos 4\theta - O(\alpha^2) \quad \text{at } \psi = 1. \tag{A.19}$$

Equating Fourier coefficients on each side of Eq. (A.18) at each order, solving for B_m^I in terms of D_m and L_m, and substituting into Eq. (A.16) gives the matrix equation

$$L_m = G_{mn} D_m. \tag{A.20}$$

Appendix B

NOVA Matrix Operators

In this appendix we present the explicit form for the matrix operators used in Eqs. (3.12) and (3.13) in Section 3.1. In addition, the three projections of the momentum equation in the NOVA formulation are shown, as well as the definitions for the three components of the perturbed magnetic field, Q_w, Q_s, and Q_b.

The matrices C, D, E, and F are defined by

$$C = \begin{bmatrix} 2K_w & G \\ 0 & -\nabla w^2 \nabla \cdot \left(\frac{\nabla \psi}{\nabla w^2} \right) \end{bmatrix} \tag{B.1}$$

$$D = \begin{bmatrix} \left(\nabla w^2 S - \mathbf{B} \cdot \mathbf{J} \right) \frac{\nabla \psi^2}{B^2} \mathbf{B} \cdot \nabla & 2\gamma P K_w \\ \nabla w^2 \left(2K_s - \frac{\mathbf{B} \times \nabla \psi}{B^2} \cdot \nabla \right) & \nabla w^2 \left\{ 1 - \frac{\gamma P}{\omega^2 \rho} \mathbf{B} \cdot \nabla \left(\frac{\mathbf{B} \cdot \nabla}{B^2} \right) \right\} \end{bmatrix} \tag{B.2}$$

$$E = \begin{bmatrix} \frac{\omega^2 \rho \nabla \psi^2}{B^2} - \mathbf{B} \cdot \nabla \left(\nabla w^2 \frac{\mathbf{B} \cdot \nabla}{B^2} \right) & 2\gamma P K_s \\ 2K_s & \frac{\gamma P - B^2}{B^2} - \frac{\gamma P}{\omega^2 \rho} \mathbf{B} \cdot \nabla \left(\frac{\mathbf{B} \cdot \nabla}{B^2} \right) \end{bmatrix} \tag{B.3}$$

$$F = \begin{bmatrix} -2K_s - \frac{\mathbf{B} \times \nabla w}{B^2} \cdot \nabla & \mathbf{B} \cdot \nabla \frac{\nabla \psi^2}{B^2} S - \frac{\mathbf{J} \cdot \mathbf{B}}{B^2} \mathbf{B} \cdot \nabla - 2P' K_s \\ -\frac{1}{B^2} & \frac{-2K_\psi}{\nabla \psi^2} \end{bmatrix} \tag{B.4}$$

Here we have

$$G = \omega^2\rho - 2P'K_\psi - |\nabla\psi|^2\mathbf{B}\cdot\nabla\left(\frac{\mathbf{B}\cdot\nabla}{|\nabla\psi|^2}\right) - (\mathbf{B}\cdot\mathbf{J} - S|\nabla\psi|^2\frac{S|\nabla\psi|^2}{B^2} \tag{B.5}$$

also

$$K_\psi = \mathbf{K}\cdot\nabla\psi \quad \text{and} \quad K_s = \mathbf{K}\cdot\frac{\mathbf{B}\times\nabla\psi}{B^2} \tag{B.6}$$

where

$$\mathbf{K} = \frac{\mathbf{B}}{B}\cdot\nabla\left(\frac{\mathbf{B}}{B}\right) \tag{B.7}$$

is the magnetic field curvature, and

$$S = \frac{(\mathbf{B}\times\nabla\psi)}{|\nabla\psi|^2}\cdot\nabla\times\left(\frac{\mathbf{B}\times\nabla\psi}{|\nabla\psi|^2}\right) \tag{B.8}$$

is the negative local magnetic shear. Furthermore, ρ is the mass density, γ is the ratio of specific heats, P is the equilibrium pressure, and $P' = \partial P/\partial\psi$.

The three components of the momentum equation come from taking projections of the momentum equation, Eq. (3.1), along $\nabla\psi$, $\mathbf{B}\times\nabla\psi$, and $\mathbf{B}$:

$$\nabla\psi\cdot\nabla P_1 = \omega^2\rho\xi_\psi - |\nabla\psi|^2\mathbf{B}\cdot\nabla\left(\frac{\mathbf{B}\cdot\nabla\xi_\psi}{|\nabla\psi|^2}\right) \tag{B.9}$$

$$-(|\nabla\psi|^2 S - \mathbf{B}\cdot\mathbf{J})\frac{|\nabla\psi|^2}{B^2}(\mathbf{B}\cdot\xi_s - S\xi_\psi) - 2(\mathbf{K}\cdot\nabla\psi)Q_b,$$

$$(\mathbf{B}\times\nabla\psi)\cdot\nabla P_1 = \omega^2\rho|\nabla\psi|^2\xi_s - (\mathbf{B}\cdot\mathbf{J})\mathbf{B}\cdot\nabla\xi_\psi \tag{B.10}$$

$$-B^2\mathbf{B}\cdot\nabla\left[\frac{|\nabla\psi|^2}{B^2}(\mathbf{B}\cdot\nabla\xi_s - S\xi_\psi)\right] - 2\mathbf{K}\cdot(\mathbf{B}\times\nabla\psi)Q_b,$$

and

$$\omega^2\rho\xi_b = \mathbf{B}\cdot\nabla(p_1 - P'\xi_\psi). \tag{B.11}$$

Here we have

$$Q_\psi = \mathbf{B}\cdot\nabla\xi_\psi, \tag{B.12}$$

$$Q_s = \left(\frac{|\nabla\psi|}{B}\right)^2(\mathbf{B}\cdot\nabla\xi_s - S\xi_\psi), \tag{B.13}$$

and

$$Q_b = B^2\mathbf{B}\cdot\nabla\frac{\xi_b}{B^2} - B^2\nabla\cdot\xi - 2\mathbf{K}\cdot(\mathbf{B}\times\nabla\psi)\xi_s - 2(\mathbf{K}\cdot\nabla\psi)\frac{B^2}{|\nabla\psi|^2}\xi_\psi + P'\xi_\psi. \tag{B.14}$$

Here $\nabla \cdot \xi$ is defined by

$$\nabla \cdot \xi = \frac{\nabla w \cdot \nabla \xi_w}{\nabla w^2} - \left[\nabla \cdot \left(\frac{\nabla w}{|\nabla w|^2}\right)\right] - \frac{\mathbf{B} \times \nabla w \cdot \nabla \xi_s}{B^2} \tag{B.15}$$

$$-2\mathbf{K} \cdot (\mathbf{B} \times \nabla w)\xi_s - \mathbf{B} \cdot \nabla \left(\frac{\xi_b}{B^2}\right).$$

"I have some few references to make."

Sherlock Holmes. *The Sign of Four*

Bibliography

[1] H. P. Furth. Nucl. Fusion **15** (1975) 487.

[2] J. Wesson. *Tokamaks*. The Oxford Engineering Science Series. Oxford University Press. Oxford. 1987.

[3] F. Close. *Too Hot to Handle: The Story of the Race for Cold Fusion*. Princeton University Press. Princeton. 1991.

[4] J. D. Lawson. Proc. Phys. Soc. London. Sec. B **70** (1957) 6.

[5] F. Troyon. R. Gruber. H. Saurenmann. S. Semenzato. and S. Succi. Plasma Physics and Controlled Fusion **26** (1984) 209.

[6] R. J. Goldston. Plasma Physics and Controlled Fusion **26** (1984) 87.

[7] B. Coppi. R. Dagazian. and R. Gajewski. Phys. Fluids **15** (1972) 2405.

[8] T. Ohkawa and H. G. Voorhies. Phys. Rev. Lett. **22** (1969) 1275.

[9] M. Okabayashi et al., Studies of bean-shaped tokamaks and beta limits for reactor design, in *Plasma Physics and Controlled Nuclear Fusion Research 1984 (Proc. 10th Int. Conf. London, 1984), Vol. 1*, pp. 229–238. Vienna. 1985. IAEA. IAEA.

[10] K. Bol et al.. Phys. Rev. Lett. **57** (1986) 1891.

[11] R. D. Stambaugh and et al., Test of beta limits as a function of plasma shape in Doublet III. in *Plasma Physics and Controlled Nuclear Fusion Research 1984 (Proc. 10th Int. Conf. London. 1984), Vol. 1*. pp. 217–227. Vienna. 1985. IAEA, IAEA.

12 S. P. Hakkarainen, R. Betti, J. P. Freidberg, and R. Gormley, Physics of Fluids B **2** (1990) 1565.

13 D. Lortz, Plasma Physics and Controlled Fusion **32** (1990) 117.

14 G. Laval, R. Pellat, and J. Soule, Phys. Fluids **17** (1974) 835.

15 P. H. Rutherford, On the stability of the elliptical cross section for free-boundary tokamak, MATT Report 976, Princeton University Plasma Physics Laboratory, Princeton, NJ 08543, 1973.

16 M. D. Rosen, Phys. Fluids **18** (1975) 482.

17 E. Rebhan, Nucl. Fusion **15** (1975) 277.

18 M. Okabayashi and G. Sheffield, Nucl. Fusion **14** (1974) 263.

19 E. Rebhan and A. Salat, Nucl. Fusion **16** (1976) 805.

20 E. Rebhan and A. Salat, Nucl. Fusion **17** (1977) 251.

21 S. M. Osovets, *Plasma Physics and the Problems of Controlled Thermonuclear Reactions*, volume 2, pp. 322–?, Pergamon, Oxford, 1959.

22 S. Yoshikawa, Phys. Fluids **7** (1964) 278.

23 V. S. Mukhovatov and V. D. Shafranov, Nucl. Fusion **11** (1971) 605.

24 J. P. Freidberg, *Ideal Magnetohydrodynamics*, Plenum Press, New York, 1987.

25 L. E. Zakharov, Sov. Phys.–Tech. Phys. **16** (1971) 645.

26 T. K. Jensen and M. S. Chu, Phys. Fluids B **1** (1989) 1545.

27 L. S. Solov'ev, Sov. Phys.–JETP **26** (1968) 400.

28 L. C. Bernard, D. Berger, R. Gruber, and F. Troyon, Nuclear Fusion **18** (1978) 1331.

29 R. Gruber et al., Comput. Phys. Comm. **21** (1981) 323.

30 M. S. Chu and R. L. Miller, Phys. Fluids **21** (1978) 817.

31 K. Lackner and A. B. MacMahon. Nucl. Fusion **14** (1974) 575.

32 F. A. Haas. Nucl. Fusion **15** (1975) 407.

33 J. K. Lee. Nuclear Fusion **26** (1986) 955.

34 F. Hofmann, F. B. Marcus, and A. D. Turnbull. Plasma Phys. and Contr. Fusion **28** (1986) 705.

35 F. Hoffmann, A. D. Turnbull, and F. B. Marcus. Nucl. Fusion **27** (1987) 743.

36 M. Chu, R. Miller, and T. Ohkawa. Nucl. Fusion **17** (1977) 465.

37 S. C. Jardin. Phys. Fluids **21** (1978) 1851.

38 S. Jardin and D. Larrabee. Nucl. Fusion **22** (1982) 1095.

39 D. Pfirsch and H. Tasso. Nucl. Fusion **11** (1971) 259.

40 D. Dobrott and C. Chang. Nucl. Fusion **21** (1981) 1573.

41 S. W. Haney and J. P. Freidberg. Physics of Fluids B **1** (1989) 1637.

42 E. Rebhan and A. Salat. Nucl. Fusion **18** (1978) 1431.

43 S. Jardin et al., Nucl. Fusion **27** (1987) 569.

44 S. Jardin, N. Pomphrey, and J. DeLucia, J. Comp. Phys. **66** (1986) 481.

45 L. A. Charlton, D. W. Swain, and G. H. Neilson, IEEE Transactions on Plasma Science (1979) 190.

46 F. B. Marcus, S. C. Jardin, and F. Hofmann, Physical Review Letters 55 (1985) 2289.

47 E. A. Lazarus, J. B. Lister, and G. H. Neilson. Nucl. Fusion **30** (1990) 111.

48 J. B. Lister et al., Experimental study of the vertical stability of high decay index plasmas in the DIII-D tokamak, LRP 382/89, C.R.P.P.-E.F.P.L., Lausanne, 1989.

49 J. A. Leuer, Fusion Technology **15** (1989) 489.

50 N. Pomphrey, S. C. Jardin, and D. J. Ward, Nucl. Fusion **29** (1989) 465.

[51] C. Z. Cheng and M. S. Chance. J. Comp. Phys. **71** (1987) 124.

[52] P. Materna, J. Chrzanowski, and F. Homan, in *Proc. of 12th Symposium on Fusion Engineering*, p. 346, Monterey, California, 1987, IEEE =87CH2507-2.

[53] K. Ogata. *Modern Control Engineering*. Prentice-Hall, Inc., Englewood Cliffs, N.J., 1970.

[54] B. C. Kuo. *Automatic Control Systems*. Prentice-Hall, Inc., Englewood Cliffs, N.J., 5th edition, 1987.

[55] I. B. Bernstein, E. A. Frieman, M. D. Kruskal, and R. M. Kulsrud, Proc. R. Soc. (1958) A244.

[56] R. Grimm, J. Greene, and J. Johnson, Computation of the magnetohydrodynamic spectrum in axisymmetric toroidal confinement systems, in J. Killeen, editor, *Methods in Computational Physics*, *Vol. 16*, pp. 253–280, Academic Press, New York, 1976.

[57] S. Hamada, Nuclear Fusion **2** (1962) 23.

[58] A. H. Boozer, Phys. Fluids **23** (1980) 904.

[59] R. B. White. *Theory of Tokamak Plasmas*, North-Holland, Amsterdam; New York, 1989.

[60] T. R. Harley, *The Computation of Resistive MHD Instabilities in Axisymmetric Toroidal Plasmas*, PhD thesis, Princeton University, Princeton, NJ, 1990.

[61] J. D. Jackson, *Classical Electrodynamics*, Wiley, New York, second edition, 1975.

[62] M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions*, Dover Publications, New York, 1965.

[63] C. Hastings Jr., J. T. Hayward, and J. P. Wong, *Approximations for Digital Computers*, Princeton University Press, Princeton, NJ, 1955.

[64] M. S. Chance et al., Comparison of different methods of calculating the vacuum energy integral in the PEST program, in *Proceedings of the 8th International Conference on Numerical Simulation of Plasmas*, pp. PC-5, Monterey, CA, 1978.

65 F. Troyon, L. C. Bernard, and R. Gruber, Comput. Phys. Comm. **19** (1980) 161.

66 R. Lüst and E. Martenson, Zeitschrift für Naturforschung **15** (1960) 706.

67 F. W. Grover, *Inductance Calculations: Working Formulas and Tables*, Dover Press, New York, 1946.

68 C. Canuto, M. Y. Hussaini, A. Quarteroni, and T. A. Zang, *Spectral Methods in Fluid Dynamics*, Springer Series in Computational Physics, Springer-Verlag, New York, 1988.

69 C. A. J. Fletcher, *Computational Galerkin Methods*, Springer Series in Computational Physics, Springer-Verlag, New York, 1984.

70 F. Najmabadi, R. W. Conn, and The ARIES Team, The ARIES tokamak fusion reactor study, in *IEEE 13th Symposium on Fusion Engineering*, IEEE Press, 1989.

71 C. G. Bathke, S. C. Jardin, J. A. Leuer, D. J. Ward, and The ARIES Team, Vertical stability requirements for ARIES-I reactor, in *IEEE 13th Symposium on Fusion Engineering*, IEEE Press, 1989.

72 C. Kessel, CIT PF Design group meeting, Feb. 13, 1990; to be published.

73 W. H. Press, B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling, *Numerical Recipes: The Art of Scientific Computing*, Cambridge University Press, Cambridge, 1986.

74 D. J. Ward et al., Bull. Am. Phys. Soc. **33** (1988) 2036.

Where is the wise man? Where is the scholar?
Where is the philosopher of this age?
Has not God made foolish the wisdom of the world?

For the foolishness of God is wiser than man's wisdom.
and the weakness of God is stronger than man's strength.

1 Corinthians 1: 20, 25