

MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

ESL-TR-83-43

67

ARCHITECTURAL and FUNCTIONAL DESIGN of an ENVIRONMENTAL INFORMATION NETWORK

S. NAVATHE, and W. HUBER, et al.

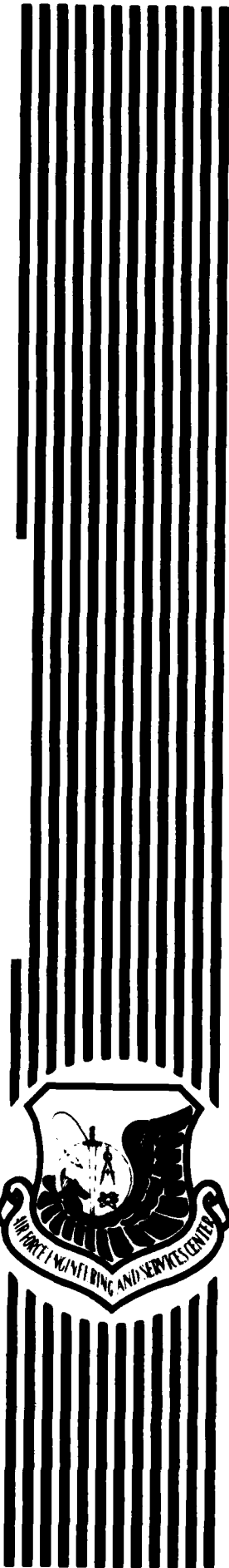
UNIVERSITY of FLORIDA
GAINESVILLE, FLORIDA 32611

APRIL 1984

FINAL REPORT
MAY 1983 - FEBRUARY 1984

DTIC
AUG 3 1984
A

APPROVED FOR PUBLIC RELEASE: DISTRIBUTION UNLIMITED



AD-A143 837



DTIC FILE COPY



ENGINEERING & SERVICES LABORATORY
AIR FORCE ENGINEERING & SERVICES CENTER
TYNDALL AIR FORCE BASE, FLORIDA 32403

84 08 02 005

NOTICE

Please do not request copies of this report from
HQ AFESC/RD (Engineering and Services Laboratory).

Additional copies may be purchased from:

National Technical Information Service
5285 Port Royal Road
Springfield, Virginia 22161

Federal Government agencies and their contractors
registered with Defense Technical Information Center
should direct requests for copies of this report to:

Defense Technical Information Center
Cameron Station
Alexandria, Virginia 22314

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE

REPORT DOCUMENTATION PAGE

1a. REPORT SECURITY CLASSIFICATION Unclassified			1b. RESTRICTIVE MARKINGS				
2a. SECURITY CLASSIFICATION AUTHORITY			3. DISTRIBUTION/AVAILABILITY OF REPORT Approved for public release; Distribution unlimited				
2b. DECLASSIFICATION/DOWNGRADING SCHEDULE			4. PERFORMING ORGANIZATION REPORT NUMBER(S)				
4. PERFORMING ORGANIZATION REPORT NUMBER(S)			5. MONITORING ORGANIZATION REPORT NUMBER(S) ESL-TR-83-43				
6a. NAME OF PERFORMING ORGANIZATION University of Florida		6b. OFFICE SYMBOL (If applicable)		7a. NAME OF MONITORING ORGANIZATION			
6c. ADDRESS (City, State and ZIP Code) Dept of Computer and Information Sciences Dept of Environmental Engineering Gainesville, Florida 32611				7b. ADDRESS (City, State and ZIP Code)			
8a. NAME OF FUNDING/SPONSORING ORGANIZATION HQ Air Force Engineering and Services Ctr.		8b. OFFICE SYMBOL (If applicable) RD		9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER F-08635-83-C-0136			
8c. ADDRESS (City, State and ZIP Code) Tyndall Air Force Base, Florida 32403				10. SOURCE OF FUNDING NOS.			
11. TITLE (Include Security Classification) Architectural and Functional Design of an Environmental Information Network				PROGRAM ELEMENT NO. 63723F	PROJECT NO. 2103	TASK NO. 90	WORK UNIT NO. 26
				12. PERSONAL AUTHOR(S) Navathe, Shamkant B.; Huber, Wayne C.; Su, Stanley Y.W.; Heaney, James P.; Sashidhar, T.			
13a. TYPE OF REPORT Final		13b. TIME COVERED FROM 83 May 01 to 84 Feb 01		14. DATE OF REPORT (Yr., Mo., Day) 84 Apr 30		15. PAGE COUNT	
16. SUPPLEMENTARY NOTATION							
17. COSATI CODES			18. SUBJECT TERMS (Continue on reverse if necessary and identify by block number)				
FIELD	GROUP	SUB. GR.	Computer Systems Information Networking				
09	02		Computer Interface Data Base Management				
12	01		Network Computer Operations Environmental Modeling				
19. ABSTRACT (Continue on reverse if necessary and identify by block number)							
<p>→ A review and evaluation of distributed data base management systems was undertaken to avail enhanced access to and use of computer resources for the environmental assessment process. A concept refinement was developed for computer-to-computer interfacing of environmental quality data with existing assessment models. Five developing data base management systems (CSIN, Distributed INGRES, MULTIBASE, R*, and SDD-1) were reviewed for applicability to a proposed environmental information network. Such a network would streamline the assessment process by meshing environmental quality data resources with evaluation programs and simulation models.</p> <p>A broad-scale approach, targeted for air quality applications, was used. Design and relational guidelines were developed for several system functions, including: model and data base selections, model and data base queries, text editing, menu-driven data entry, data reformatting, data transformation, and language translation between systems. A data dictionary, for referencing all relevant models and data bases, was</p>							
20. DISTRIBUTION/AVAILABILITY OF ABSTRACT UNCLASSIFIED/UNLIMITED <input checked="" type="checkbox"/> SAME AS RPT. <input type="checkbox"/> OTC USERS <input type="checkbox"/>				21. ABSTRACT SECURITY CLASSIFICATION Unclassified			
22a. NAME OF RESPONSIBLE INDIVIDUAL Capt Glenn E. Tapio				22b. TELEPHONE NUMBER (Include Area Code) (904) 283-4628		22c. OFFICE SYMBOL AFESC/RDWW	

DD FORM 1473, 83 APR

EDITION OF 1 JAN 73 IS OBSOLETE.

UNCLASSIFIED
SECURITY CLASSIFICATION OF THIS PAGE

#12 Proefke, Bonnie W.; Cornelio, Aloysius; MacIntyre, David F.; Miracle, David L.

#19 discussed in detail. Examples were developed for interactive access to an air quality data base (SAROAD) and batch-mode access to a hydrologic data base (HISARS).



PREFACE


This report was prepared by the University of Florida, Gainesville, Florida 32611, as a complementary effort from the Department of Computer and Information Sciences, and the Department of Environmental Engineering Sciences. The research study was accomplished under contract F08635-83-C-0130, Task 83-2 for Headquarters Air Force Engineering and Services Center, Engineering and Services Laboratory (HQ AFESC/RD), Tyndall Air Force Base, Florida, 32403.

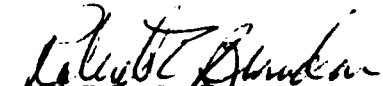
This report covers work performed between May 1983 and February 1984. AFESC/RDVW Project Officer was Captain Glenn E. Tapio.

The purpose of this development study was to assess functional data resources applicable to broad environmental quality management, and propose the best interface scheme for computerization of the assessment process. The rapid advances in computer technology in recent years, coupled with today's information explosion, provide an excellent target of opportunity for integrating specialty data resources with innovative computer operations. Due to the diverse applicability of such a scheme, a broad spectrum approach was used to evaluate architecture alternatives, with use targeted for air quality considerations. This work is a logical first step in applying computer resources for integrated environmental assessments.

This report has been reviewed by the Public Affairs Office (PA) and is releasable to the National Technical Information Service (NTIS). At NTIS, it will be available to the general public, including foreign nationals.

This technical report has been reviewed and is approved for publication.


GLENN E. TAPIO, Capt, USAF
Project Officer


ROBERT E. BRANDON
Deputy Director, Engineering
and Services Laboratory


JIMMY N. FULFORD, Lt Col, USAF
Chief, Environics Division



i
(The reverse of this page is blank.)

SEARCHED	INDEXED
SERIALIZED	FILED
APR 1984	
AFESC/RDVW	
TYNDALL AIR FORCE BASE	
FLORIDA	
A1	

TABLE OF CONTENTS

Section	Title	Page
I	INTRODUCTION.	1
	A. PROBLEM DEFINITION.....	1
	B. ROLE OF DATA BASE MANAGEMENT SYSTEMS.....	3
	C. PROJECT SCOPE.....	4
II	SURVEY OF OPERATIONAL ENVIRONMENTAL AND AIR QUALITY MODELS AND RELATED DATA BASES.....	5
	A. INTRODUCTION.....	5
	B. GENERAL CLASSIFICATION AND SELECTION SCHEME FOR ENVIRONMENTAL MODELS.....	6
	C. ATMOSPHERIC MODELS.....	10
	D. WATER MODELS.....	24
	E. SUMMARY.....	32
III	SURVEY OF DISTRIBUTED DATA BASE MANAGEMENT.....	33
	A. DATA BASES - GENERAL CONCEPTS.....	33
	B. DISTRIBUTED DATA BASE MANAGEMENT OVERVIEW.....	36
	C. SURVEY OF SOME DISTRIBUTED DATA BASE SYSTEMS.....	40
	D. APPLICATION OF DISTRIBUTED DATA BASE CONCEPTS TO THE CURRENT PROBLEM.....	43
	E. SUMMARY.....	48
IV	A PROPOSED ORGANIZATION OF THE ENVIRONMENTAL INFORMATION NETWORK.....	49
	A. INTRODUCTION.....	49
	B. A CLASSIFICATION OF THE INTERACTION AMONG MODELS AND DATA BASES.....	50
	C. OVERALL ARCHITECTURE AND FACILITIES.....	53
	D. DISCUSSION OF THE VARIOUS APPROACHES.....	55
	E. EXAMPLES OF TARGET SYSTEM MODELS AND DATA BASES....	56
	F. SUMMARY.....	57
V	GENERAL DESIGN CONSIDERATIONS.....	59
	A. STRUCTURE OF THE DIRECTORY.....	59
	B. DATA FORMATTING.....	64
	C. MODEL - DATA BASE TRANSFERS.....	64
	D. IMPLEMENTATION ISSUES.....	65
VI	SUMMARY AND RECOMMENDATIONS.....	67
	A. SUMMARY.....	67
	B. RECOMMENDATIONS FOR THE ENVIRONMENTAL INFORMATION NETWORK.....	68

TABLE OF CONTENTS (CONCLUDED)

Section	Title	Page
	REFERENCES.....	70
APPENDIX		
A	DISTRIBUTED DATA BASE MANAGEMENT SYSTEMS.....	73
	A. INTRODUCTION.....	73
	B. THE DISTRIBUTED DATA BASE SYSTEMS.....	73
	C. REFERENCES.....	90
B	EXAMPLE USE OF SAROAD INTERACTIVE-ORIENTED DATA BASE...	93
	A. INTRODUCTION.....	93
	B. SAROAD FILES.....	93
	C. SAROAD INTERACTIVE SYSTEM.....	98
	D. REFERENCES.....	98
C	EXAMPLE USE OF HISARS BATCH-ORIENTED DATA BASE.....	104
	A. DESCRIPTION OF HISARS.....	104
	B. HISARS LANGUAGE.....	104
	C. EXAMPLE OF USE.....	104
	D. REFERENCES.....	117

LIST OF FIGURES

Figure	Title	Page
1	Flow Chart for Environmental Systems Model Selection Procedure.....	11
2	General Architecture of Distributed Data Management System.....	37
3	Design Flow Chart of the Environmental Information Network.....	54
4	Structure of the Directory.....	60
A-1	Schema Architecture.....	74
A-2	MULTIBASE Component Architecture.....	77
A-3	R* Architecture.....	79
A-4	Architecture of SDD-1.....	82
A-5	CSIN Prototype Design Layers.....	85

LIST OF TABLES

Table	Title	Page
1	SUMMARY OF SEQUENTIAL STEPS FOR SETTING UP THE ENVIRONMENTAL SYSTEMS ANALYSIS.....	8
2	IDENTIFICATION AND DESCRIPTION OF CHARACTERISTICS FOR DESCRIBING ENVIRONMENTAL SYSTEMS MODELS.....	9
3	AIR QUALITY MODELS AND THEIR RELEVANCE TO SPECIFIC PROBLEMS.....	13-15
4	SUMMARY OF CLIMATOLOGICAL DISPERSION MODEL.....	16
5	POPULATION OF AIR QUALITY MODELS FOR POTENTIAL APPLICATION IN ENVIRONMENTAL SYSTEMS ANALYSIS.....	18-19
6	POSSIBLE AIR QUALITY MODEL SELECTION INTERACTIVE PROMPTS/QUESTIONS.....	20-21
7	APPLICABILITY OF RUNOFF MODELS TO VARIOUS PROBLEM CHARACTERISTICS.....	27
8	SUMMARY OF THE AGRICULTURAL RUNOFF MODEL.....	28
9	RECEIVING WATER MODELS AND THEIR RELEVANCE TO SPECIFIC PROBLEMS.....	30
10	SUMMARY OF THE REVISED DISSOLVED OXYGEN SAG MODEL.....	31
11	A COMPARATIVE REVIEW OF DISTRIBUTED DATA BASE MANAGEMENT SYSTEMS.....	46
B-1	RAW DATA LESS THAN 24-HOUR REPORT.....	94
B-2	METEOROLOGICAL RAW DATA REPORT.....	95
B-3	YEARLY FREQUENCY DISTRIBUTION REPORT.....	96
B-4	QUARTERLY FREQUENCY DISTRIBUTION REPORT.....	97
B-5	SAROAD COMMANDS.....	99
B-6	EXAMPLE INTERACTIVE SESSION.....	100-103
C-1	HISARS COMMAND LANGUAGE.....	105
C-2	PROGRAM TO SEARCH FOR APPROPRIATE STATION RECORDS.....	106
C-3	SUMMARY OF AVAILABLE RAIN DATA.....	108
C-4	SUMMARY OF AVAILABLE EVAPORATION DATA.....	109-110
C-5	SUMMARY OF AVAILABLE TEMPERATURE DATA.....	111
C-6	PROGRAM TO EXTRACT THE REQUIRED DATA.....	112
C-7	LISTING OF MONTHLY PRECIPITATION FROM 1970 THROUGH 1979.....	113
C-8	LISTING OF DAILY PRECIPITATION DURING 1979.....	114
C-9	LISTING OF MONTHLY EVAPORATION FROM 1970 THROUGH 1979.....	115
C-10	LISTING OF MONTHLY TEMPERATURES FROM 1970 THROUGH 1979.....	116

LIST OF ABBREVIATIONS AND ACRONYMS

- AEQ - Ambient Environmental Quality.
- AEROS - Aerometric and Emissions Reporting System, a collection of EPA air quality data bases.
- AFEMDEX - Air Force Environmental Model and Data Exchange. A proposed network (recommended by General Software Corporation) to integrate environmental models with data bases.
- ARPANET - A nationwide computer network developed by Advanced Research Projects Agency (ARPA) of the U.S. Department of Defense. This network allows a large number of dissimilar computers called hosts to communicate with each other.
- BACTLAER - Best Available Control Technology/Lowest Achievable Emission Rate.
- B-tree - A method of organizing indexes to files on secondary storage (typically disks) to facilitate insertions and deletions.
- B-tree index - The indexing part of a B-tree provides a quick means to locate a desired record in a file.
- CCA - The Computer Corporation of America, a computer software research organization.
- CDM - Climatological Dispersion Model, an air quality model from the UNAMAP series provided by EPA.
- CICS - Customer Information Control System which is a commercial data base/data communications system developed by IBM.
- CIT - A commercially available computer terminal.
- CODASYL - Conference on Data Systems Languages. It is the group that developed the COBOL language and proposed the Network data model.
- CPU - The Central Processing Unit of a computer. It performs arithmetic and logical operations on data. It contains the control unit which supervises the functioning of the machine as a whole, calling into operation the various units as required by a program.
- CSIN - Chemical Substances Information Network, designed by the Computer Corporation of America under EPA contract.
- DAPLEX - The Data Manipulation Language for the functional data model. This language is used in the MULTIBASE System.
- Datagram - A method of data transfer employed in packet-switched networks.

LIST OF ABBREVIATIONS AND ACRONYMS (CONTINUED)

- DBMS - Data Base Management System. It is a software system that allows the central definition of the data base to be stored, manages the various data files and provides a framework for storing, retrieving and updating data.
- DDBMS - Distributed Data Base Management System, a software system for managing DBMSs situated at different locations.
- DM - The Data Module in SDD-1.
- DML - Data Manipulation Language. Users retrieve, modify, or insert data from a data base using a data manipulation language. DML refers to a procedural language; i.e., the user must explicitly specify the operations in a sequential manner. A high-level nonprocedural DML is called a query language.
- EIN - Environmental Information Network, the proposed software system for Air Force application.
- EPA - Environmental Protection Agency.
- GDM - Global Data Manager of MULTIBASE. This module performs decomposition and recombination of a query. Any query in MULTIBASE is composed by the GDM into subqueries for the local data bases. The results from these subqueries are integrated by the GDM.
- GS - The Global Schema (GS) in MULTIBASE. It is a description of the entire data base containing all the host schemas mapped into a unified schema.
- GSC - General Software Corporation.
- HISARS - Hydrologic Information Storage and Retrieval System. A data base containing hydrologic information.
- IBM - International Business Machines Corporation.
- IFAS - Institute of Food and Agricultural Sciences at the University of Florida.
- IMS - Information Management System. It is a commercial data base management system based on the hierarchical data model offered by IBM.
- INGRES - Interactive Graphics and Retrieval System. It is a data base management system developed at the University of California and is based on the relational model.
- I/O - Input/Output of a computer.
- IS - The "Integration Schema" of MULTIBASE. Contains all the relevant information needed to integrate the local schemas.

LIST OF ABBREVIATIONS AND ACRONYMS (CONTINUED)

- JCL - Job Control Language. A type of language used to supply detailed information about data files, devices, libraries of routines, etc., so as to define the environment for the execution of one or more programs.
- LDI - Local data base interface module in MULTIBASE. The LDI is responsible for translating a subquery into the data manipulation language of the local system.
- leaf nodes - The lowest level of a B-tree index.
- LHS - Local host schema in MULTIBASE. Many different local data base schemas are mapped in the global schema in MULTIBASE. These local data base schemas are called local host schemas.
- MULTIBASE - A heterogenous distributed data base system developed at the Computer Corporation of America.
- MVS - Multiple Virtual Storage. A computer operating system running on the IBM 360/370 machines.
- NAAQS - National Ambient Air Quality Standards.
- NCC - National Climatic Center, the NOAA archival facility at Asheville, NC.
- NEDS - National Emissions Data System, an EPA air quality data base.
- NOAA - National Oceanic and Atmospheric Administration.
- NSM - Natural Systems Model.
- NSTAR - Night Stability Array Tabulation. A NOAA computer program that calculates the joint frequency distribution of 6 wind speeds, 16 wind directions and 6 stability classes.
- NWS - National Weather Service.
- PTMAX - Point Source, Maximum Concentration air quality model.
- QS - Quarterly Summary.
- QUEL - (A query language.) It is the query language for the INGRES data base system.
- R* - A distributed data base version of System R developed by IBM Research Laboratories.
- RDS - Relational Data System of System R.
- RELNET - Reliable Network in SDD-1.
- REQM - Residuals-Environmental Quality Management.

LIST OF ABBREVIATIONS AND ACRONYMS (CONCLUDED)

- residual - Generic term for water or air quality parameter, e.g., a pollutant.
- RFF - Resources for the Future, a nonprofit research firm in Washington, DC.
- RSS - Research storage system of System R.
- SAROAD - Storage and Retrieval of Aerometric Data, an EPA air quality data base.
- SAS - Statistical Analysis System, a package for general purpose statistical applications.
- SDD-1 - "System for Distributed Data Bases," a DDBMS developed by the Computer Corporation of America.
- SNA LU6 - System Network Architecture which is a protocol for long-haul networks proposed by IBM.
- SNOOP - A special machine function for deadlock detection in distributed INGRES system.
- SOT - Source Test Data System, an EPA air quality data base.
- SQL - The language for manipulating relations. Same as SQL/DS.
- SQL/DS - Structured Query Language. It is a high-level language for data manipulation on data bases using the relational model. Implemented in the SQL/DS system of IBM.
- STORET - Storage and Retrieval of Water Quality Control Information, the major EPA water quality data base.
- System R - A commercial data base management system developed by IBM, based on the relational model.
- TCM - Texas Climatological Model.
- TM - The transaction module in SDD-1.
- UNAMAP - User's Network for Applied Modeling of Air Pollution, a set of air quality models published by EPA.
- UNIX - A popular computer operating system.

SECTION I

INTRODUCTION

A. PROBLEM DEFINITION

1. Environmental Concerns

Like a small city, an Air Force Base can generate air and water pollution, noise, traffic, and other factors of concern to the surrounding area and the overall environment. Moreover, the Air Force is subject, in most instances, to the same federal, state, and local environmental regulations that apply to other governing bodies (e.g., cities). Thus, sewage treatment plants must obtain permits; air pollution from traffic and aircraft must be controlled; noise abatement procedures must be followed for both airborne and ground operations, etc. These factors relate to planning and design activities. At a more immediate level, accidents that might damage the environment must be rectified, e.g., hazardous waste spills on land and water, sewage releases, fires, explosions, etc. Reaction time is critical in such instances, and advanced, computer-based technology would greatly aid in instituting control measures. Such a technology would naturally be useful at all levels of environmental and engineering activities and is currently being implemented by the Air Force.

2. Environmental Models and Data

Within the past 25 years, all areas of engineering have received more powerful tools for routine application than were possible using "hand" calculations. In the environmental area, sophisticated mathematical models may now be used to address the kinds of problems previously mentioned. As an example, concentrations of air pollutants may be predicted, using any number of models for both short and long term assessments as a function of source and pollutant characteristics, terrain features, and meteorological conditions. Such models may be used to determine the impact of new construction on an Air Force Base or increased use (e.g., by additional traffic or more flights) of existing facilities. The Environmental Protection Agency (EPA) is an example of a federal agency that develops and maintains such computer technology.

Of equal importance is the availability of computerized data bases that contain information on all aspects of our society. In particular, environmental data bases contain time histories and spatial distributions of air and water quality data, as well as myriad meteorologic, hydrologic and demographic data. These resources now permit engineers to use statistical methods and other tools of analysis on a routine basis for planning, design, and operational purposes. Some models also require the kinds of data stored in such repositories. For example, certain air pollution models require historic meteorological data to predict long-term average and extreme pollutant concentrations (see Section II).

Clearly, the success of modeling, statistical analysis and other analytical procedures hinges upon the availability of appropriate computer technology for user interaction, linkage to data, and execution of numerical procedures. The Air Force is currently quite active in incorporating such technology into its environmental and other engineering services.

3. Software for Environmental Models and Data

This report focuses upon the architectural and functional design of software that will enable the Air Force to manage environmental models and data easily and accurately for application to recurring problems. This software, along with relevant models, data, computing network, new computer hardware and system software, etc., will constitute the basis of the Air Force Environmental Model and Data Exchange (AFEMDEX).

This developmental effort partially follows an exhaustive study (Reference 1) by the General Software Corporation (GSC) that outlines capabilities and deficiencies of current environmental engineering activities within the Air Force, including the areas of modeling, data bases, controller software, computer networks, etc. The GSC study noted deficiencies in data needed for model support, model-data format, compatibility, proximity of data to where they are needed, quality of modeling software and documentation, as well as other deficiencies. In general, limited facilities, noninteractive programs, and "non-user-friendly" computer procedures have limited past Air Force efforts to improve its engineering capabilities.

Major GSC recommendations include:

- a. Link functionally distinct Air Force centers with a computer network. Models and data collected and maintained at different locations could then be retrieved and transferred to any location on the network, at which point data could be merged and formatted properly for analysis.
- b. Coordinate data collection efforts.
- c. Identify facilities for integration into the network.
- d. Provide users with a friendly interface, e.g., menu-driven programs.
- e. Develop improved techniques for data collection, storage, retrieval, formatting, etc.
- f. Use experience gained in related data base activities, such as for the Chemical Substances Information Network (CSIN).

The GSC recommendations are general and do not address the technical complexity inherent in many of the proposals. For example, considerable developmental effort is required simply to construct a user-friendly and interactive environmental model. The effort is enormously magnified when user-friendly software must be developed to manage a large array of environmental models, data bases and network linkages. Nonetheless, the thrust of the research described in this report deals with the architectural design of this software. Through this means, considerable insight may be found, short of the much more massive effort required to construct the actual prototype software.

B. ROLE OF DATA BASE MANAGEMENT SYSTEMS

1. Problem Areas

The large diversity in computer systems and the concurrent growth in environmental models and data bases have led to severe problems of "matching" the models to the data bases. The data bases are managed by software of varying sophistication, most of which is batch-oriented, for example, the key repository of meteorological data at the Ashville National Climatic Center of the National Oceanic and Atmospheric Administration (NOAA), while a few have on-line capabilities. Moreover, some data bases have a simple sequential file structure (as on a magnetic tape), while others employ a complex random access structure.

Not only is the structure of data bases nonuniform, but the specific format requirements of various models may also lead to difficulties because of varying names and data types assigned to model parameters (e.g., real vs. integer, numerical vs. character). These are not standardized across models and may lead to different access requirements from different models to the same data base. A few environmental models avoid the model-data base incompatibilities by incorporating "default" data directly into the model. The user can then call up meteorological data nearest to the application site. However, a much more general solution is to eliminate the incompatibilities among useful models and data bases.

2. Data Base Management Systems

The incompatibility between models and data bases could be resolved if the data base-handling software were made independent of the model or application; an appropriately designed data base management system would allow models to run, using any data base. Such a system provides the user with an interface that will perform all functions of retrieval, creation, deletion, and modification of data. Its design would include three major components:

a. Common language interface. A common language allows a user to state his request independently of the structure and physical organization of the data base. All user queries are stated in the common language, using a standard query format, represented in the form of a table. A menu-driven query formulation must be designed for ease of use.

b. Transformation of data. All queries in standard format, using a common language, are decomposed to extract the relevant data from the data bases. The retrieved data are reformatted, and any conversion (data type, units, etc.) is done before the data are used by any model.

c. Common dictionary. A dictionary is a data base of "metadata"; i.e., it is a data base about various environmental data bases and models. The dictionary contains a properly classified list of all models and data bases included in the system, and contains information relevant to which data bases can be used by each model. Users are allowed to browse through the dictionary to decide upon a model and upon a data base with which to run the model.

These component needs require the development of a distributed data base system wherein different environmental data bases are accessible from any site on the Air Force computer network. In a conceivable scenario, a user at Site A could select a model at Site B and run it, using a data base at Site C. The state of the art of distributed data base management systems and a framework for the proposed architecture of such a system for the Air Force are described in Section III of this report.

C. PROJECT SCOPE

1. Project Goal

The project is to analyze existing distributed data base management systems and propose an architecture for an environmental information network on the basis of the following:

- a. The network will be in keeping with the goals of the AFEMDEX proposals.
- b. Environmental models to be considered initially as examples will be one or two of the EPA UNAMAP (User's Network for Applied Modeling of Air Pollution) air quality models.
- c. Environmental data bases to be considered initially will be the NOAA data bases from Asheville, plus EPA and hydrologic data bases.

2. Phase 1: Literature Review

Review and evaluate distributed data base management systems and techniques, including CSIN, as they relate to the needs of the project. Operational environmental models and data bases, particularly those related to air quality, will be similarly reviewed and evaluated.

3. Phase 2: General Design

- a. Design a data model to deal with environmental data bases. Examine some sample data bases.
- b. Categorize environmental models on the basis of their interaction with users and data bases.
- c. Investigate high-level languages that may be suitable for software construction and how they can relate to the low-level command languages used by current data base management systems for environmental data.
- d. Define the contents and structure of a data dictionary. Review environmental models and their input/output requirements in the dictionary for air quality data and models.

4. Reporting Format

Phase 1 is covered in detail in Sections II and III of this report; Phase 2 is considered in Sections IV and V. A detailed technical review of distributed data base management systems is given in Appendix A. Appendices B and C present examples of typical environmental data bases.

SECTION II

SURVEY OF OPERATIONAL ENVIRONMENTAL AND AIR QUALITY MODELS AND RELATED DATA BASES

A. INTRODUCTION

An important phase of environmental modeling is the application of theoretical or empirical knowledge of natural processes to predict the concentration and distribution of natural or manmade substances (e.g., pollutants) in the environment. Over the past two decades the development and use of environmental models have rapidly increased. This trend is mainly in response to mandated planning requirements of federal, state, and local environmental legislation. The various mandates require the partitioned environmental system analysis for evaluation of the human activity impacts on ambient environmental quality. Furthermore, analysis is required to assess the effectiveness of strategies and measures for achieving and maintaining an acceptable level of ambient quality (Reference 2).

Environmental modeling tools applied in such an analysis may range from the simple to the complex. Selection of an appropriate model is based on the time and budget available, questions to be answered, detail required to answer the questions, models available, computer requirements, and input data requirements.

This chapter contains material on environmental models and data bases with emphasis on model classification, selection, data requirements, and possible data base applications. The first section presents a general scheme for classifying and selecting an appropriate environmental model. The second section describes the scheme as applied to air quality models and includes a suggested interactive sequence for computer-aided selection of an appropriate air quality model. In addition, the section discusses general data requirements and possible data base applications for air models, along with an example of specific data requirements for a selected model. The third section presents a brief summary of a water model classification and selection scheme, as well as data requirements and data base applications.

Before proceeding further, several comments are appropriate regarding application of environmental data bases. In general, two types of input data are required for application of environmental models: environmental conditions data and pollutant discharge data. Environmental conditions data refer to variables such as air temperature, wind velocity, wind direction, streamflow, and water temperature. Pollutant discharge data relate to the quantities of various pollutants discharged from specific sources over specified areas and time periods. The user must enter most of these model input data manually, independent of available data bases. That is, most environmental models do not require input data from formal data bases. In addition, very few data bases are designed to support modeling activities. Therefore, in many instances, data from environmental data bases must be statistically processed to satisfy model requirements as noted in the GSC report (Reference 1). Although currently available data bases are not particularly useful for direct modeling applications, they do provide valuable background and statistical information necessary for complete analysis of a given environmental problem.

In this context, the word "model" is interpreted broadly within this report to mean any form of computer analysis of data. For instance, a statistical analysis of environmental data, or even a visual inspection of the data will still require most of the data base management facilities described in Sections IV and V of this report. Although the following discussion mainly describes models in terms of a set of programmed algorithms applied to specific environmental problems, this broader sense of the word should be borne in mind.

B. GENERAL CLASSIFICATION AND SELECTION SCHEME FOR ENVIRONMENTAL MODELS

1. Model Classification

Several classification schemes are commonly used for categorizing environmental models. In a recent Resources for the Future (RFF) Report, Basta and Moreau (Reference 3) propose the following broad delineation among environmental models: (1) models of residual (pollutant) generation and discharge from land surfaces, (2) models of surface receiving waters, (3) models of subsurface waters, (4) models of atmospheric systems, and (5) models of terrestrial ecological systems. This classification scheme is operational with respect to the ways in which environmental models are integrated into environmental quality management and the ways that different environmental models are linked. In addition, with the possible exception of noise models, these categories encompass specialized classes of models such as exposure (radiation), chemical spill, and waste disposal models (Reference 1).

Environmental models may be further classified according to any one, or a combination of three basic analytical approaches: (1) physical modeling, (2) conservation of mass and energy, and (3) statistical. Physical modeling involves construction of a small-scale physical replica of the environmental system of interest. The model is designed to simulate the behavior of the real system under controlled conditions. As this study focuses on mathematical models, physical models will not be further considered. Models based on the conservation of mass and energy approach attempt to simulate the material and energy transport and transformation processes in an environmental system, explicitly. In contrast, statistical models do not explicitly simulate system processes but simply provide estimated values of output variables for given values of other variables.

These analytical approaches to environmental modeling may be further classified on the basis of three important characteristics relating to all environmental models: (1) temporal variation, (2) averaging time, and (3) spatial dimensionality. Temporal variation refers to how model variables change over time. Analyses which involve no interconnections between time periods and no changes in variables for a given time period are considered steady-state analyses. On the other hand, non-steady-state analysis involves two or more interconnected time periods in which all input variables may differ. Intermediate between pure steady-state and pure non-steady-state models are "quasi-steady-state" models where only some, but not all, variables change from one time period to the next.

Non-steady-state models must be distinguished as stochastic or nonstochastic. In stochastic models, variations in each variable from one time period to another include a random component. Nonstochastic or

deterministic models include no random component in the time variations of each variable. Three types of models are defined in terms of temporal variations: (1) steady-state-nonstochastic, (2) non-steady-state nonstochastic, and (3) non-steady-state stochastic.

Quantity and quality variables (e.g., flows and pollutant concentrations) vary with time and may often be represented as a time series. In order to simplify the presentation of time series data, an averaging time may be selected which is the unit of time over which single values for input and output variables are estimated. The averaging time chosen depends on management questions to be answered and the model used.

Similarly, variables may vary in up to three dimensions. Spatial dimensionality refers to the number of dimensions incorporated into the model and to the number and sizes of segments into which the environmental system is divided.

2. Model Selection

Basta and Moreau (Reference 3) present a sequence of steps that can be followed to determine what natural systems analysis (environmental model) to apply. Before considering the scope and types of appropriate models, preliminary specifications of certain analysis conditions must be determined. These specifications include: (1) residuals (i.e., pollutant or other quality parameter) of concern, (2) questions to be answered and translated into specific information concerning ambient environmental quality, (3) relevant time periods for the analysis, and (4) boundaries of the region to be studied.

For a given set of defined analysis conditions, the sequence of steps for choosing a suitable environmental modeling approach is summarized in Table 1. Steps (4) and (5) refer to a model selection process in which, for any given set of problem characteristics, a number of models which can address those aspects of a problem are identified. These various model and problem characteristic relationships are conveniently presented by Basta and Bower (Reference 2), who provide "problem tables" for runoff, surface receiving water, and atmospheric models in tabular form. Various characteristics which represent the most important attributes involving the applicability of a model are examined, to obtain specificity information (how and with what a model can analyze a given problem). Table 2 identifies and describes 15 characteristics applicable to the three model categories mentioned above. Time, space, and mathematical properties are applicable to all environmental models, as discussed previously. Characteristics, in addition to those in Table 2, may be defined for specific categories, such as those uniquely applicable to air quality models. Analysis of these characteristics provides a consistent means of summarizing environmental models for comparison purposes. In addition to model summaries presented in Basta and Bower (Reference 2), abstracts of 103 available EPA models are given in the EPA Environmental Model Catalogue (Reference 4). The catalogue describes each model in terms of problem and model characteristics. Next, available analytical resources to use the identified models are compared with the resources required. Finally, a model may be selected on the basis of evaluation criteria which include (1) accuracy of estimation, (2) capability of estimating ambient environmental quality responses to various physical measures for reducing residual (pollutant) discharges, (3) commonalities which a feasible environmental model has with

TABLE 1. SUMMARY OF SEQUENTIAL STEPS FOR SETTING UP THE ENVIRONMENTAL SYSTEMS ANALYSIS (FROM BASTA AND BOWER, REFERENCE 2).

Given: The "questions" to be addressed by the study have been identified and translated into specific information requirements to be generated on ambient environmental quality (AEQ), distributions of AEQ, and residuals-environmental quality management (REQM) costs. The residuals of concern, relevant time periods, and boundaries of the study area for which information is required have been specified.

Step 1: Make a preliminary identification of the environmental media and relevant characteristics thereof that are affected by the residuals of concern, e.g., by considering known transport pathways, impacts of residuals on relevant species, accumulation and/or degradation of residuals in environmental media, and a determination of the categories of natural systems analyses--residuals generation/discharge, surface receiving water, terrestrial flow routing, subsurface flow, terrestrial ecologic, and atmospheric ambient quality--that are necessary to undertake in relation to the questions to be answered.

Step 2: Based on the foregoing and on available data on environmental condition variables, residual discharge variables, and AEQ variables, make a preliminary analysis in order to determine: (1) if the right questions have been asked; (2) if the important residuals have been identified; (3) the relative importance of various point and aerial sources of residuals; (4) the potential importance of intermedia and interregional transfers; and (5) any additional characteristics of the problem which will help to determine the natural systems models (NSMs) to use.

Step 3: Based on the results of Step 2, specify the level of complexity desired for the natural systems analysis in terms of: (1) time intervals of analysis, i.e., hourly, daily, monthly, seasonal, mean annual; (2) spatial aggregations, i.e., number of grid areas, stream reaches, lake layers; (3) constituent/species aggregations to be considered, i.e., forms of nitrogen, number of fish species; and (4) specific physical, chemical, and biological processes to be incorporated.

Step 4: Determine the feasibility of carrying out the natural systems analysis specified in Step 3, in relation to the available analytical resources. To determine feasibility, see relevant references on natural systems categories for choosing among existing operational models utilizing: a model selection procedure, a model population table, a model problem table, model characteristics, and model summaries. Iteration may be necessary to make the choice of NSM(s) consistent with the available analytical resources.

Step 5: Select the NSM or NSMs to be used, based on the relevant evaluation criteria.

Step 6: Develop a work plan for the natural systems analysis which is compatible with other segments of the REQM analysis.

TABLE 2. IDENTIFICATION AND DESCRIPTION OF CHARACTERISTICS FOR DESCRIBING ENVIRONMENTAL SYSTEMS MODELS (FROM BASTA AND BOWER, REFERENCE 2).

Characteristic	Summary Description
Time Properties	Description of how temporal variations in flow and/or transport of residuals are represented in a model, both conceptually and computationally. Information includes: (1) time variability: steady state, quasi-nonsteady state, and nonsteady state; and (2) time units of application: input and output values and computational time step.
Space Properties	Identification of capability of model to represent spatial variations of residuals concentrations. Information includes: (1) model dimensionality: one-dimensional, horizontal or vertical plane; or multi-dimensional, longitudinal, latitudinal, and/or vertical plane; (2) spatial aggregation: size and number of segments, layers, and volumes possible; and (3) typical areal units of application.
Physical Properties	Identification of the physical processes considered in a model which account for the transport of residuals. Information includes: (1) the principal hydraulic and meteorologic driving forces, i.e., tidal action, wind currents, and stream currents; and (2) the individual physical processes involved: advection, diffusion, dilution, convection, heat budget-temperature, wind, Coriolis acceleration forces.
Chemical Processes	Description of the chemical transformations/interactions considered in a model which result in changes in concentration over time. Information includes: (1) basic chemical processes: thermochemical equilibrium and coupled or noncoupled chemical reactions; and (2) ambient quality indicators represented: conservative substances and nonconservative substances.
Ecological Processes	Identification of basic biological processes that affect interactions between AEQ indicators, constituents, and among the various organisms represented in a model. Included are: (1) biochemical processes, i.e., photosynthesis, respiration-biological decay; and (2) trophic dynamics, i.e., trophic levels, population growth dynamics, mortality, predator-prey interactions.
Mathematical Properties	Description of the theoretical basis for the mathematic representation in, and the solution methodology applied to, a model. Information includes: (1) theoretical aspects: deterministic, stochastic, and combinations; and (2) types of solution methodology: statistical, i.e., regression methods, and other mathematical techniques, i.e., analytically integrated or numerically integrated.
Computational Status	Identification of status of model in relation to manipulation on digital computers. Information includes: (1) whether model is coded or uncoded; (2) computer language used, e.g., FORTRAN; and (3) the various computer/accessory equipment required (hand calculator, analogue computer, digital computer, including needed storage capacity, compilers, magnetic tapes or disks for storage).
Input Data Requirements	Description of the various data required to set up, run, calibrate, and verify a model. Generalized historical data bases and/or site-specific data may be required, depending upon the extent of prior applications.
Ease of Application	Description of expected difficulties in obtaining, modifying, and applying a model. Information includes: (1) the availability of model and supporting documentation from various sources; and (2) identification of anticipated areas of difficulty in model application.
Output and Output Format	Identification of model output with respect to the types of information produced, spatial and temporal distributions possible, and the format in which the output is presented.
Linkages to Other Models	Description of linkages between a model and other types of models used in REQM analysis and the forms of such linkages. Information provided includes linkages to other natural systems models, damage/benefit models, and models of activities and the ways in which the linkages are structured.
Manpower Needs	Identification of types, number, and desired levels of experience of personnel required to apply a model. Requirements for additional model development and/or major internal modification may differ significantly from those for model application. Information includes: (1) position descriptions (engineer, programmer, systems analyst, ecologist, others); (2) need for specialized training or capabilities; and (3) the related experience of interdisciplinary team members.
Costs	Specification of all costs involved in applying a model, from initial acquisition of the model through analysis of model output. Model costs are a combination of: (1) man-hour expense required to complete certain tasks (including data collection); (2) expense of obtaining a model itself and/or available user's manuals/documentation reports; and (3) direct computation costs for the application. Where possible, cost information is provided under the categories of model acquisition, data preparation, actual computation expense, computer accessory costs, and output analysis. Complete cost information is not available for most NSMs.
Model Accuracy and Sensitivity	Description of the overall capability of a model to represent accurately a natural system and its essential processes. Information includes: (1) the representativeness in relation to the "real" system, especially the extent of the description and simplifying assumptions; (2) numerical accuracy-stability and dispersion (in numerically integrated models); and (3) sensitivity to input errors or rate coefficients, both known and estimated values.
Other Comments	Description of miscellaneous facts concerning a model, its history of development and modification, and informative observations as to its utility in REQM analysis. Information includes: (1) model limitations and restrictions; (2) special features and options; and (3) model originators or producer of current derivation (if modification of existing model).

^aOnly those characteristics which are relevant for more than one category of NSMs are included.

prior analyses in the region, and (4) contributions that the environmental model would make to long-term environmental quality management in the region. A summary of the model selection procedure is shown in the flow diagram of Figure 1.

C. ATMOSPHERIC MODELS

1. Model Classification

Six key characteristics can be used operationally to describe a given problem in terms of identifying suitable air quality models for use in analysis. These key characteristics include (1) type of pollutant, (2) type of source, (3) characteristics of receptor points, (4) type of terrain, (5) time properties, and (6) geographic scale. Different air quality models would be appropriate for different specifications of characteristics.

Since the physical and chemical processes of interest depend on the contributing pollutant(s), the type of pollutant is an important characteristic to identify. The choices between models also depend on how well each process is quantitatively characterized.

For modeling purposes, air discharge sources are characterized as (1) multiple or single stationary point sources, (2) stationary area sources, or (3) line sources. Sources are also characterized with respect to the elevation of discharge, either ground-level or elevated.

Receptor points are the points in a region at which estimates of ambient concentrations are desired. Important receptor characteristics include number, height, and spatial location (coordinates).

Type of terrain is particularly important since most operational models have been specifically developed for either smooth or gently rolling terrain, open terrain, or for "rough" urban areas on smooth terrain.

Time factors in air quality modeling are usually defined as either long-term (climatological) or short-term (episodic). Long-term modeling efforts are applicable to chronic air quality problems, which generally require estimates of annual or seasonal average pollutant concentrations. Short-term modeling efforts relate to acute air quality problems, which generally require estimates of ambient pollutant concentrations for averaging periods of a few hours to one day. Short-term modeling usually involves steady-state analysis (or analysis over a sequence of, say, hourly time steps), while long-term modeling involves quasi-steady-state analysis and usually produces long-term average concentrations over a season or year.

Geographic scale refers to the areal extent of the study site. For application of current air quality models, Muschett (Reference 5) cites four operationally defined geographic scales: (1) up to a few km, (2) from a few km to about 50 km, (3) from about 50 km to about 200 km, and (4) greater than 200 km. Note that models which analyze pollutant transport on a scale greater than 200 km are called long-range transport models. No models of this type are included in the 1982 RFF Report (Reference 2). The GSC report (Reference 1) divides air models into only two geographic scale categories. Regional transport models apply to scales greater than 50 km, while nonregional models apply to scales up to 50 km.

As discussed in the general environmental classification scheme, models in each broad category may be classified according to the analytical approach applied. Mathematical air quality models are based on either conservation of mass and energy or statistical approaches. Air quality models based on the principles of conservation of mass and energy can be operationally separated into three categories: volumetric, Gaussian, and numerical. The volumetric approach is the simplest and treats the atmospheric space above the area of interest as a single homogeneous volume. This approach is used when the actual conditions approximate the assumed steady-state conditions of instantaneous mixing of discharges in the volume upwind of a specified receptor point. The Gaussian approach is more complex and is based on analytical integrations of the three-dimensional advective-diffusion equation. These models are generally considered state-of-the-art techniques for estimating nonreactive pollutant concentrations. Numerical models are the most complex and are based on numerical integration of the advective-diffusion equation. These models are more appropriate than Gaussian models for multisource applications involving reactive pollutants.

The statistical approach for analyzing atmospheric systems is essentially the same as described for the general statistical approach. That is, statistical air models do not explicitly simulate system processes but provide estimated output variable values for given values of other variables. The principal statistical method applied is regression analysis. These techniques are generally used in situations where (1) a first, crude analysis is required, (2) the detailed data required for a more complex model are unavailable, or (3) incomplete scientific understanding of the physical and chemical processes makes other approaches impractical.

2. Model Selection

In the discussion of general environmental model classification and selection schemes, a sequence of steps was described (Table 1) which could be followed to determine what types of environmental systems analyses and models could be applied to any given problem. The first three steps involve specification of the required environmental analysis. Steps (4) and (5) involve selecting an appropriate model for the required analysis on the basis of problem and model characteristics. Therefore, for a given problem of interest, an appropriate air quality model may be selected on the basis of specified relevant problem characteristics discussed previously. Tables 3 and 4 show how air quality problems and model characteristics can be related for use in model selection.

Consistent with the model selection scheme outlined here is the procedure presented in the EPA Guideline (Reference 6) which classifies recommended air quality models on the basis of source characteristics and type of pollutant for various averaging times. Additional factors for consideration in determining the suitability of a particular model application include (1) the detail and accuracy of the data, (2) the meteorological and topographic complexities of the area, (3) the technical competence of those undertaking the simulation modeling, and (4) the resources available. The GSC report (Reference 1) classifies models according to geographic scale, type of terrain, type of pollutant, and purpose, e.g., assessment, spill, and heavy gas models. Other characteristics such as time properties, source characteristics, and analytical approaches, are described in individual model abstracts. While the

TABLE 3. AIR QUALITY MODELS AND THEIR RELEVANCE TO SPECIFIC PROBLEMS (FROM BASTA AND BOWER, REFERENCE 2, CONCLUDED).

MODEL TYPE/MODEL DESCRIPTION	PROBLEM CHARACTERISTICS																														
	REGIONAL	SOURCE	RECEPTOR	TERRAIN	TIME	SCALE																									
NUMERICALLY INTEGRATED Chemically Reactive Power Plant Plume Model IMPACT STIFF 3 ^b Westinghouse C.T. v.c.d. EDH 2 ^a DIFER ^a CDH SAI (1978 Version) ^a (Urban Air Shed Photochemicals) WFTDS	Sulfur	Acid Rain Deposition	Acid Rain Deposition (wet)	Surface Deposition (dry)	Surface Deposition (wet)	Acid Rain Deposition	Sulfur	Cooling Towers	Single Point Sources	Multiple Point Sources	Line Sources	Area Sources	Ground-Level Discharge	Elevated Discharge	Point Discharge	Area Receptor	Ground-Level Receptor	Elevated Receptor	Smooth or Chilly Rolling	Rough Urban	Shore of Coastal Area	Valley or Basin	Hilly or Mountainous	Forest	Steady-state	Dynamic	Long-Term	Short-Term	Up to a Few Km	100 Km to 200 Km or so	Greater than 200 Km
	PROPORTIONAL Ballbeck / Ballflower Appendix STATISTICAL Crooks - Roberts Bower Model																														

^a Frequently used for mobile sources ^b proprietary ^c C.T. denotes cooling tower ^d model used for cooling tower droplet deposition (drift description)
^e Air Quality Submodel of the Transportation and Air Shed Simulation Model

TABLE 4. SUMMARY OF CLIMATOLOGICAL DISPERSION MODEL (FROM BASTA AND BOWER, REFERENCE 2).

Characteristic	Summary Data
Terrain Characteristics	(1) "rough" urban areas
Physical Processes	(1) transport and diffusion; (2) dry deposition can be accounted for in the user-specification of a "half-life"
Chemical Processes	(1) none
Mathematical Properties	(1) type of relationship: (a) deterministic; (2) solution methodology: (a) analytically integrated, Gaussian model with some numerical approximations
Time Properties	(1) time variability of inputs: (a) steady-state; (2) time period of outputs: (a) long-term: seasonal, annual average
Space Properties	(1) dimensionality: (a) three; (2) spatial scale of application: (a) micro; (b) meso; (3) discretization: (a) one layer; (b) up to 50 x 50 grid areas and grid size is user-specified; (c) user-specified number of receptor points at user-specified locations; (4) resolution: (a) inputs: residuals discharge inputs as point or area sources, single point source meteorological data; (b) outputs: point concentrations
Input Data Requirements	(1) initial set-up and validation: (a) point and area residuals discharges; (b) stack parameters: height, diameter, temperature, exit velocity; (c) meteorological data: joint frequency of wind speed (six classes), wind direction (sixteen directions) and stability class (six classes), annual average or seasonal mixing height; (d) several ambient concentration measurements; (e) source coordinates; (f) receptor coordinates; (2) verification: (a) residuals discharges; (b) meteorological data (see above); (c) ambient concentration measurements
Computation Status	(1) model status: (a) coded; (2) equipment needed: (a) for use in interactive mode a terminal is required; (b) for use in batch mode a digital computer is required; (3) computer characteristics: (a) IBM 360 or IBM 370 or equivalent; (b) program language: FORTRAN IV (level C)
Ease of Application and Manpower Needs	(1) availability of operational model and supporting information: (a) public agency; (2) availability of model documentation: user guide available with: (a) theory; (b) subroutines; (c) data deck format; (d) test case; (3) expertise required: (a) computer programmer; (b) research assistant; (4) model user qualities: (a) mathematical ability through algebra and statistics; (b) understanding of model principles and assumptions and limitations; (c) experience in interpretation of model outputs or access to consultations
Output and Output Format	(1) types of output: (a) ambient concentration of SO ₂ and suspended particulates; (2) values given at: (a) user-specified points; (b) ground-level; (3) special features: (a) calibration subroutine; (b) receptor contribution analysis;* (c) statistical output subroutine*
Costs	(1) model acquisition: (a) public, "as is" model listing and user document available free or at a nominal cost; (2) model set-up: (a) program set-up and keypunch: 8 man-days (estimated); (b) test case and de-bug: 10 man-days (estimated); (3) data preparation: (a) analysis and reduction: 3 man-days (estimated); (b) keypunched or recorded: 2 man-days (estimated); (4) actual computation: dependent upon problem and number of sources and receptors, generally less than \$20/run; (5) output analysis: (a) graphics and data plotting costs not specified; (b) analysis of results varies with problem, but routine validation/verification can each be done in 5 man-days
Model Accuracy and Sensitivity	(1) representativeness to prototype system: (a) extent of description: fixed point meteorological data do not describe micrometeorological variations within city nor "urban heat island" air circulations, representation of wind speed as a "power law" function of height is an improvement upon AQDM, use of Briggs plume rise formula also is an improvement; (b) simplifying assumptions: Gaussian diffusion, homogenous diffusion; (2) model accuracy: (a) comparatively small absolute errors; (b) comparatively low standard errors; (3) model sensitivity: (a) known to be sensitive to effective stack height, wind speed, stability
Other Comments	(1) model limitations and restrictions: (a) intended for SO ₂ and suspended particulates; (2) special features: (a) designed to accept meteorological (airport) data summaries from National Climatic Center Star Program; (b) has modeling improvements in comparison to earlier AQDM, as noted above; (c) interactive model available; (d) part of EPA UNAMAP series; (e) new version has popular output features noted above; (3) application experience: (a) model is one of most widely used and best tested for state implementation plans

* This feature is found in updated version of model. Its COM9C subroutines as published in Addendum to User's Guide for Climatological Dispersion Model, EPA-450/3-77-015, May 1977.

RFF, EPA, and GSC reports all distinguish between various air quality models on the basis of the important model characteristics just described, the RFF approach provides an efficient, relatively simple means of evaluating all or almost all of the important characteristics involved with selection of an appropriate model. For additional insight to model evaluation criteria see Turner (Reference 7), who presents a comprehensive critical review of operational air quality models.

3. Interactive Computer-Aided Selection -- Example

a. Preliminary Analysis

The following is a suggested sequence of prompting and questioning for interactive computer-aided selection of an appropriate operational air quality model.

Before signing on to the computer, the user should have completed preliminary specification of the following conditions of the analysis as previously described: (1) pollutants of concern, (2) questions to be answered and translation of them into information to be generated on ambient air quality, (3) relevant time periods for the analysis, and (4) boundaries of regions to be studied. In addition, the user should have carefully considered and completed the analyses included in Steps 1, 2, and 3 (Table 1) in order to specify the required environmental analysis. At this stage in the analysis, the user is ready to test the feasibility of the analysis and compare and evaluate the various model characteristics in relation to the determined problem characteristics.

b. Problem Characteristics

Once the user has signed on to the computer, he/she should be given three options: (1) access the desired model and meteorological or calibration data directly, (2) access a menu listing available models and data for subsequent direct access, or (3) initiate the procedure for selecting an appropriate model. The menu of available models in option (2) might take a form similar to that given in Table 5 (Reference 2). Note that Table 5 includes nonoperational, as well as operational models. For this suggested application only available operational models would be included. If option (3) is selected, the user might receive a series of prompts/questions as shown in Table 6.

This interactive sequence would be accompanied by a program that screens the available models to identify those which satisfy the user's specifications. If none of the available models correspond to all the specified problem characteristics, the user should, at least, have the option to request a display of the models satisfying individual problem characteristics. More conveniently, the user should be able to see what models satisfy any single or combination of the specified problem characteristics. This is a necessary feature for deciding whether adjustments might be made to the problem specifications.

An alternate approach to this prompt/question procedure is to provide the user with a display table, similar to Table 3, relating problem characteristics to the available models. The user continues through the screening procedure without prompts. This approach avoids the problem of providing the user with intermediate results of the selection procedure.

TABLE 5. POPULATION OF AIR QUALITY MODELS FOR POTENTIAL APPLICATION IN ENVIRONMENTAL SYSTEMS ANALYSIS (FROM BASTA AND BOWER, REFERENCE 2).

KEY

* indicates model is operational

I indicates a pure air quality model

II indicates a combination air quality model, including one or more of:

A residuals discharge submodel

P plume rise submodel

W windfield submodel

Model Category/Model Name	I	II	Commonly Used Acronym	Originator /Contact Person
Conservation of Mass Models				
• Volumetric				
Clifford-Nanna ^a	X		---	Frank Clifford & Steve Nanna Atmospheric Diffusion Lab., Oak Ridge Nat'l Lab., Oak Ridge, Tenn.
Clifford-Nanna Photochemical ^a	X		---	" "
• Gaussian				
Air Quality Display Model ^a		P	AQDM	TW Systems Group under contract to National Air Pollution Control Admin. (presently EPA) Mr. Bruce Turner, Contract Officer
Climatological Display Model ^a		P	CDM	Meteorology Lab., EPA Research Triangle Park, N.C. (UAMAP Series)
Air Quality Submodel of the Transportation and Air Shed Simulation Model ^b		A,P		Norm Cooper, Office of Environmental Affairs U.S. Dept. of Transportation, Washington, D.C.
CO POLLUT		A	---	C.R. Gebhardt, Ohio Dept. of Transportation Columbus, Ohio
California Line Source Dispersion Model ^c	X		CALINE 2	Mr. Andy Banzieri, Calif. State Div. of Highways, Sacramento, Calif.
EPA RIMWAY ^a	X		RIMWAY	Meteorology Lab., EPA, Research Triangle Park, N.C. (UAMAP Series)
Point, Area, Line Source Model ^a		P	PAL	Bruce Turner & William R. Peterson, Meteorology Lab., EPA, Research Triangle Park, N.C.
Single Source Model ^a		P	CRSTER	Russell Lee, et al., Source Receptor Analysis Branch, Office of Air Quality Planning & Standards, EPA, Durham, N.C.
Air Pollution Research Advisory Committee LA		A	APRAC LA	R.C. Mamasco & P.L. Ludwig, Stanford Research Institute, Menlo Park, Calif., (UAMAP Series)
PTMAX ^a		P	PTMAX	Meteorology Lab., EPA, User Network of Available Models of Air Pollution (UAMAP Series)
PTDIS ^a		P	PTDIS	" "
PTMTP ^a		P	PTMTP	" "
Real-Time Air Pollution Model		P	RAM	Mr. Bruce Turner & Ms. Joan Novak, Meteorology Lab., EPA, Research Triangle Park, N.C. (to be included in UAMAP Series)
Gaussian Evaluation Model ^a		P	GM	A. Fabrick, R. Sklarow, J. Wilson, Science Applications, Inc., Westlake Village, Calif.
Argonne Puff ^a		P	---	E.J. Crone, J.J. Roberts, et al., Energy & Environmental Div., Argonne National Lab., Argonne, Ill.
Stanford Research Institute		P	SRI	Don Ruff, Stanford Research Institute, Palo Alto, Calif.
Great Lakes University of Wisconsin at Milwaukee Metro Meteorological Project ^h	X		GLUM ^h	Walter Ivone, University of Wisconsin at Milwaukee
VALLEY ^a		P	---	Edward Burt, Source Receptor and Analysis Branch, Office of Air Quality Planning & Standards, EPA, Durham, N.C.
Oak Ridge Fog and Drift ^a		P	ORFAD	John Wilson, Oak Ridge National Lab., Oak Ridge, Tenn.
Atmospheric Transport Model ^a		P	---	Walter Cukrowski and Malcolm Patterson, Oak Ridge National Lab., Oak Ridge, Tenn.
Drift Transport Cooling Tower	X		---	Custer Schecker, Environmental Systems Corp., Knoxville, Tenn.
The Research Corp. of New England Cooling Tower ^a		P	TBC	Norm Brown, The Research Corp. of New England, Uxbridgefield, Conn.

TABLE 5. POPULATION OF AIR QUALITY MODELS FOR POTENTIAL APPLICATION IN ENVIRONMENTAL SYSTEMS ANALYSIS (FROM BASTA AND BOWER, REFERENCE 2, CONCLUDED).

Model Category/Model Name	I	II	Commonly Used Acronym	Originator /Contact Person
• Gaussian				
Texas Episodic Model ^a		P	TEH	John Christiansen, Data Processing Div., Texas Air Control Board, Houston, Texas
Texas Climatological Model ^a		P	TCH	" "
• Numerically Integrated				
International Business Machines Air Quality I&2 ^a		W	IBMAQ I&2	Danny Shih, International Business Machines, Inc., San Jose, Calif.
INTERA		W	INTERA	Don Lantz, INTERA, Environmental Consultants, Ltd., Houston, Tx.
Atmospheric Diffusion Particle-in-Cell		W	ADPIC	Mike MacCracken & Rolf Lange, Lawrence Livermore Lab., Livermore, Calif.
Scavenging Model Incorporating Chemical Kinetics	X, J		SMICK	J.M. Hales, et al., Atmospheric Sciences Section, Battelle R.W. Lab., Richland, Wash.
Atmospheric Carbon Monoxide Simulation Program		W	ACROP	Robert Carlson & William Norton, Univ. of Alaska, Fairbanks, Alaska
Livermore Regional Air Quality I ^a		W	LIRAQ I	Mike MacCracken, Lawrence Livermore Lab., Livermore, Calif.
Livermore Regional Air Quality II		W	LIRAQ II	" "
Reactive Plume Model ^a		X	RPM	Wei-Kan-Liu, Systems Applications, Inc., San Rafael, Calif.
Chemically Reactive Powerplant Plume Model (not coded)			---	B.W. Miksad, K.E. DeBowers, and J.R. Brack, Austin, Texas
Integrated Model for Plume and Atmospheric in Complex Terrain ^a		W	IDFACT	A. Fabrick, R. Shlarow, J. Wilson, Science Applications, Inc., Westlake Village, Calif.
Brookhaven National Laboratory Long and Short Range Air Quality Model ^a		W	BRL	Sam Meyers, Brookhaven National Lab., Atmospheric Sciences Sec., Upton, N.Y.
Statistical Turbulent Incompressible Fluid Flow		X	STIFF-3	James Taft, Systems Science & Software, La Jolla, Calif.
Westinghouse Cooling Tower		X	---	Amran Roffman & Ralph Crumble, Environmental Systems Dept., Westinghouse Electric Corp., Monroeville, Pa.
Reactive Environmental Simulation Model ^a		A	REM & REM 2	Peter Drivas, Pacific Environmental Services, Inc., Santa Monica, Calif.
Diffusion Kinetic Model ^a		A	DIFKIN	Alan Eichenroeder, Environmental Research & Technology, Inc., Santa Barbara, Calif.
Center for the Environment and Man			CEN	Joseph Pandolfo, Center for the Environment & Man, Hartford, Conn.
Systems Application, Inc. Photochemical ^a		A	SAI	Steven Reynolds & Philip Ruth, Systems Applications, Inc., San Rafael, Calif.
Regional Transport Model of Atmospheric Sulfates ^a		X	---	K.S. Rao, I. Thompson, & B.A. Egan
Numerical Examination of Urban Smog		W	NEKUB	R. Shlarow, A. Fabrick, J. Prager, Systems Science & Software, Inc., La Jolla, Calif.
Other Models				
• Proprietary				
Rollback Appendix J ^a		X	APPENDIX J	U.S. Environmental Protection Agency Regulation (40CFR51)
• Statistical				
Crohe and Roberts ^a		X	---	F.J. Crohe and J.J. Roberts, Energy & Environmental Systems Div., Argonne, Ill.
Repro Model ^a		X	---	Technology Service Corp., Santa Monica, Calif., under contract to EPA, National Environmental Research Center, Research Triangle Park, N.C.

TABLE 6. POSSIBLE AIR QUALITY MODEL SELECTION INTERACTIVE PROMPTS/QUESTIONS.

Pollutant Characteristics

What pollutants are to be modeled?
(Select one or more)

- Sulfur dioxide
- Suspended particulates
- Nitrogen oxides
- Hydrocarbons
- Carbon monoxide
- Oxidants
- Toxic metals
- Nutrients or salts
- Water vapor
- Sensible heat
- Radioactivity
- Pesticides
- Aeroallergens
- Fugitive dust

Menu

What deposition processes are involved?
(Select one)

- Dry surface deposition
- Wet surface deposition
- Acid rain deposition
- Sulfates
- Cooling system fallout
- None

Source Characteristics

What type of discharge sources are involved?
(Select one or more)

- Single point source
- Multiple point sources
- Line source
- Area sources

Are any of the sources discharging at elevated levels?

- Yes
- No

TABLE 6. POSSIBLE AIR QUALITY MODEL SELECTION INTERACTIVE PROMPTS/QUESTIONS.
(CONCLUDED)

Receptor Characteristics

What types of receptors are involved?

Point receptors
Area receptors

Are any of the receptors located at elevated levels?

Yes
No

Terrain Characteristics

What type of terrain is the area of interest?
(Select one)

Smooth or gently rolling
Rough urban
Shore or coastal
Valley or basin
Hilly or mountains
Forest

Time Properties

What type of temporal variations are involved?
(Select one)

Steady-state (constant inputs)
Non-steady-state (time variable inputs and outputs)

What are the time-averaging requirements?
(Select one)

Long-term (climatological)
Short-term (episodic)

Geographic Scale

What is the areal extent of the problem?
(Select one)

Up to a few km
Few km to 50 km
50 km to 200 km
Greater than 200 km

c. Model Characteristics

Following specification of the problem characteristics, the user is provided with summary model characteristic information for all models satisfying the specified problem criteria. This information might be presented in tables similar to Table 4, or in parts selected by the user from a menu of model characteristic criteria. For example, if the EPA UNAMAP Climatological Dispersion (CDM) and Texas Climatological (TCM) models both satisfy a given set of problem criteria, the user is prompted to select a model summary table for each model for display and examination. Alternatively, the user selects parts of each summary for detailed comparison. Thus, if the user selected model cost information for each model, he/she would discover that computation time for TCM is about half of that for CDM. If more than one model can be applied within the operational constraints for the given problem, a model must be selected on the basis of the evaluation criteria discussed earlier. This stage of the selection procedure requires independent analysis by the user. However, a given model-data package or network design would not be likely to include such similar models.

4. General Data Requirements and Data Base Applications

For application to air quality models, pollutant discharge data may be conveniently categorized as source (point, line, and area) data and receptor data. Point sources usually refer to those that discharge a relatively large amount of pollutants, e.g., 50 tons per year, from a stack or group of stacks. Required point source data include pollutant types, discharge data and grid coordinates for each stack. In addition, stack parameters such as height, diameter, exit gas temperature, and exit gas velocity are required. Required discharge data for line sources such as highways include type and rate of discharge, road width, number of lanes, emissions from each lane, and discharge height. Also, the locations of the ends of the straight roadway segments must be specified in proper grid coordinates. Area sources often apply to multisource urban situations where detailed, individual source data are impractical. Required area source information includes type and average rate of pollutant discharges, size of the area, representative stack height for the area, and grid coordinates for the centroid or for the southwest corner of the source.

The user must usually supply these data independently. However, operational data base systems such as the EPA AEROS series (Reference 8), which includes the National Emissions Data System (NEDS), BACT/LAER (best available control technology/lowest achievable emission rate) determination (BACTLAER), and Source Test Data System (SOTDAT), may provide useful information. The three data bases are included in the GSC report (Reference 1) as data bases most suitable for Air Force modeling needs. The NEDS data base is a computerized data-handling system which accepts, stores and reports source information relating to five criteria pollutants. The system includes data describing annual emissions and operating characteristics for point sources (plants emitting more than 100 tons per year) and area sources (considered collectively on a county basis) in the United States. BACTLAER is a manual access system containing annual emissions data relating to 15 different pollutants for selected major point sources. SOTDAT is a computerized system for storage, retrieval, and analysis of stack test data and related engineering information necessary to calculate emission factors for 11 pollutants. None

of these data bases provides complete or suitable source data for direct input into operational air quality models. However, any one or a combination of the data bases, especially the NEDS system, may provide partial data applicable to some sources of interest or, more likely, provide valuable general information relevant to the given problem. In either case, the user must independently review and analyze the data before possible application to an air quality modeling problem.

In addition to source input data, air models require receptor input data. Receptor points are sites in a region at which estimates of ambient pollutant concentrations are desired. The height and grid coordinates must be specified for each receptor. The EPA Guideline (Reference 6) presents several techniques for choosing appropriate receptor sites.

For application to air quality models, environmental conditions data refer to representative meteorological data for the region of interest. Required meteorological data generally include hourly values of wind speed, wind direction, and atmospheric stability class. Long-term models generally require hourly data in the form of a joint occurrence frequency for 16 wind directions, 6 wind speed classes, and 6 stability classes. The National Climatic Center (NCC) computer program called NSTAR gives the proper form of the 576 values of the joint frequency function for direct input into several long-term air quality models. In addition, the NCC "CARD DECK 144" contains hourly meteorological data which can be processed and input to several short-term air quality models.

Some model applications may require additional types of data. First, for model calibration and verification procedures, measured ambient air quality data are required. The user may supply these data independently or obtain them through available data base systems. SAROAD (Storage and Retrieval of Aerometric Data -- Appendix B) is one such operational system for editing, storing, summarizing, and reporting ambient air quality data. As with other air data bases, application to most air quality models requires independent review and analysis by the user. In addition to the air quality measurements, data for reaction rates or decay factors are required for models which explicitly model chemical processes. Finally, to adequately assess the significance of the air quality impact of a pollutant source, background concentrations must often be considered. The EPA Guideline (Reference 6) describes several strategies for estimating background concentrations.

5. Data Requirements for the Climatological Dispersion Model (CDM)

a. Source Data

A brief description of specific input data for the EPA UNAMAP Climatological Dispersion Model (Reference 9) serves as an illustrative example of typical air model data requirements.

This model can simultaneously consider two nonreactive pollutants. The most frequently applied pollutants are sulfur dioxide and particulate matter. The source emission rate must be specified for each pollutant. The program allows specification of a constant factor relating day and night emissions. The computer program can accept as many as 2500 area sources and 200 point sources.

A rectangular grid array of uniform-sized squares is used to overlay the region of interest. The origin of the overlay grid is located in the lower left-hand corner of the array. The length of the side of a basic square is expressed in meters. The program will accept discharge information for squares with side lengths that are integer multiples of the length of a basic square.

To apply this model, it is necessary to estimate effective heights of pollutant emission for both point and area sources. For low-level area sources, an average height of emission may be estimated on the basis of building heights. Plume-rise corrections are generally not applied to these sources. The program procedure for numerical evaluation of area source concentrations requires specification of a constant effective stack height for the entire urban area of interest. For point sources, the effective stack height is determined from the physical stack height and the estimated plume rise. The model uses the Briggs' plume-rise equation which requires user input values for stack gas exit velocity, stack exit diameter, average gas temperature, and mean ambient air temperature.

b. Receptor Data

CDM performs computations for any number of ground-level receptor points. The user must specify appropriate grid coordinates for each receptor.

c. Meteorological Data

Computation of annual or seasonal average concentrations requires input of a joint frequency distribution of wind direction, windspeed, and stability for the same time period. The joint frequency function consists of 576 entries resulting from data for 16 different wind directions, 6 windspeed classes, and 6 stability classes. Joint frequency output from the NCC program, NSTAR, may be used directly as input into CDM. Finally, the model determines an effective mixing height by modifying the average maximum (afternoon) and average minimum (night) mixing heights, depending on stability category.

d. Additional Data

This model includes a calibration option specifying linear calibration coefficients. Alternatively, concentration data are entered into the program on receptor input cards for independent linear regression analysis by the modeler, to determine proper calibration coefficients. The model will also accept a single constant background concentration value for each pollutant. Finally, the model may incorporate exponential decay for a user-specified half-life for each pollutant.

D. WATER MODELS

1. Model Classification and Selection

a. Types of Water Models

In Section II-B, environmental models were classified into five broad categories. Three of these categories together compose a broader class of models commonly referred to as "water" models. The three types of water

models include (1) runoff models which are models of residual (pollutant) generation and discharge from land surfaces, (2) models of surface receiving waters, and (3) models of subsurface waters. As with general environmental model classification, water models may be classified on the basis of analytical approach, solution technique, temporal variation, averaging time, and spatial dimensionality. Commonly, water models are further characterized according to the emphasis on two interrelated problems: the water quality aspect and water quantity aspect. Some models address the transport and transformation of constituents in water systems (the quality aspect), while others are concerned with estimating the flow distribution of water volumes (the quantity aspect). Still other water models incorporate both water quality and quantity. The remainder of this subsection briefly describes general problem and model characteristics incorporated into the model selection process for runoff models and receiving water models. Further consideration of subsurface water (groundwater) models is not included in this discussion.

b. Runoff Models

As defined by Basta and Moreau (Reference 3), models of pollutant generation and discharge from land surfaces estimate the temporal and spatial distribution of water and materials that run off land surfaces and into receiving waters as a consequence of precipitation. Runoff models analyze pollutant generation and discharge from land surfaces with varying characteristics, including size of surface area, type of land use, soil characteristics, and frequency, duration, and type of precipitation. Several models also consider water flows in channels and pipeline networks prior to discharge into the receiving water.

Several characteristics can be used to describe a given problem operationally in terms of identifying an appropriate runoff model for use in analysis. The characteristics include (1) applicable land area, (2) time properties, (3) space properties, (4) hydrology, (5) hydraulics, (6) quality processes, and (7) pollutants. Applicable land area refers to the type of land use to which a runoff model is applicable. Runoff models are broadly classified as urban or nonurban. Urban land use may be further classified as residential, commercial, industrial, and open space. Nonurban land use is usually divided into natural, agricultural, silviculture, and mining. These subcategories of urban and nonurban land use are often further classified, according to more specific land uses and land use density.

Time properties refer to the different ways in which time is considered in runoff models: (1) seasonal or annual average, (2) continuous simulation, with a relatively short time step of typically 1 hour or 1 day, and (3) single-event simulation with time steps of seconds or minutes.

Space properties refer to the number of dimensions incorporated into the model and size and number of runoff surface areas (catchments). Runoff models are often multidimensional since they must consider subsurface flow, as well as surface flow. Huber and Heaney (Reference 10) cite three operationally defined catchment sizes: (1) small, less than 50 acres; (2) medium, between 50 and 500 acres; and (3) large, greater than 500 acres. An additional important spatial aspect relates to whether or not a model can analyze multiple catchments.

Hydrologic and hydraulic characteristics relate to the physical processes associated with the movement of water over the land surface and in conveyance systems.

Quality processes and pollutant characteristics relate to the physical, chemical, and ecological processes associated with the transport and transformation of materials in the water. In particular, pollutant type and loading data are important characteristics for consideration.

As with other environmental models, runoff models may be classified according to the analytical approach applied. Mathematical runoff models which analyze the quality aspect are based on either conservation of mass and energy, or statistical approaches. In addition to these two approaches, runoff models involving quantity aspects often apply two widely used approaches, the rational method and the hydrograph approach. For a discussion of these four analytical approaches and associated solution techniques for runoff modeling, see Huber and Heaney (Reference 10) or Viessman et al. (Reference 11).

For a given runoff problem of interest, an appropriate water model may be selected on the basis of relevant problem specification and model characteristics just discussed. Tables 7 and 8 are examples of how runoff problems and model characteristics can be related for use in the general model selection scheme summarized in Table 1.

c. Surface Receiving Waters

As defined by Basta and Moreau (Reference 3), models of surface receiving waters estimate the temporal and spatial distribution of ambient water quality which results from the discharge of pollutants into surface receiving waters. Many receiving water models include one form or another of a hydraulic model, that is, a model concerned solely with the quantity aspect. Surface receiving waters include streams and rivers, lakes, ponds, reservoirs, estuarine systems, and offshore marine systems. Several receiving water models include an ecological component which estimates the impact of pollutants on plant and animal life.

Hinson and Basta (Reference 12) identify seven major water quality problems that provide an operational basis for identifying the capabilities of various receiving water models. These major water quality problems include temperature, salinity, sedimentation, dissolved oxygen, eutrophication, toxic substances, and biological effects. For a discussion of associated water quality effects, see Hinson and Basta (Reference 12).

Other characteristics useful for describing a given problem in terms of identifying an appropriate receiving water model include (1) applicable receiving water body, (2) water body characteristics, (3) time properties, (4) space properties, and (5) mathematical properties. Applicable receiving water body refers to the type of water body included in the analysis. Some models can be applied to only one type of water body or set of water bodies with similar features. Other models are applicable to many different types of receiving waters.

TABLE 7. APPLICABILITY OF RUNOFF MODELS TO VARIOUS PROBLEM CHARACTERISTICS (FROM HUBER AND HEANEY, REFERENCE 10).

MODEL NAME	PROBLEM CHARACTERISTICS									
	Applicable Land Area	Temporal Properties	Spatial Properties	Hydrology	Hydraulics	Quality Processes	Residuals			
Hydroscapce										
MRI										
SWH-Level 2										
EPARS										
Simul. SWH										
ACTNO										
ADM										
HSP										
HPS										
QPS										
STODP										
ACRIIC										
CATEDAS										
SWH										
Urban										
Agriculture										
Pasture										
Wetlands										
Single Storm Events										
Continuous Simulation										
Annual or Seasonal Average										
Single Catchment										
Multiple Catchments										
Surface/Total Hydrograph Generation										
Subsurface Processes										
Shmueli										
Dry-weather/Zero Flow										
Flow Routing in Channels/Plains										
Backwater, Surcharging, Pressure Flow										
Flow Controls and Diversions										
Storage / Reservoir Routing										
Surface Generation										
Routing in Channels/Plains										
Scur / Deposition / Erosion										
Sediment in Channels/Plains										
Parameter Interaction										
Soil/Sediment-Parameter Interaction										
Routing through Storage										
Treatment Removal in Storage										
Treatment Processes										
Organics / BOD / COD										
Nitrogen Species										
Phosphorus										
Suspended Solids										
Coliforms										
Pesticides										
Arbitrary or other Conservative										
Arbitrary of other Non-Conservative										
Genotoxic Analysis										

TABLE 8. SUMMARY OF THE AGRICULTURAL RUNOFF MODEL (FROM HUBER AND HEANEY, REFERENCE 10).

Characteristic	Summary Data
Applicable Land Drainage Area	(1) Agriculture: (a) crops, (b) pasture
Time Properties	(1) Single runoff event simulation; (2) Inputs: precipitation data at 15 minute intervals, daily data for other parameters; (3) Outputs: simulation of time history results based on 15 minute intervals
Space Properties	(1) Small to large multiple catchments; (2) Dimensionality: (a) multiple dimensions: two-dimensional land surface, three-dimensional subsurface layers, one-dimensional stream channel; (3) Discretization limits: (a) watershed may be divided into as many as 200 subareas, (b) subareas may vary in size, (c) tributary drainage system may have up to 200 channels in dendritic pattern (lengths and cross sections can vary), (d) can handle up to 5 soil layers
Physical Processes	(1) Overland flow across land surface and in small stream channels; (2) Groundwater interflow; (3) Channelized tributary flow; (4) Sediment erosion estimated with Universal Soil Loss Equation
Chemical Processes	(1) Constituent transport limited to sediment-attached parameters; (2) All constituents considered conservative--no interaction or degradation; (3) Constituents represented: total suspended solids, BOD, and fecal coliforms are operational
Ecological Processes	None
Economic Analysis	None
Mathematical Properties	Using numerical integration of differential equation set
Computational Status	(1) Coded in Fortran IV; (2) Requires IBM 360/365 digital computer or equivalent; (3) Computer core storage requirements unspecified, but extensive
Input Data Requirements	(1) Precipitation data and drainage channel specifications; (2) Land use hydro-geometric data; (3) Watershed and soil characteristics
Ease of Application	(1) Nonproprietary model derived from SWMM Runoff Module; (2) Basic SWMM documentation available from NTIS or EPA Planning Assistance Branch, Washington, D.C.; (3) Documentation report from Iowa-Cedar River Basin Project available from Water Resources Engineers, Walnut Creek, Ca. or EPA Systems Development Branch, Washington, D.C.; (4) Extent of documentation: (a) SWMM documentation fairly complete for most applications, (b) AGRUN report presents brief information on model modifications, data deck design, sample input and output formats, but no program listings
Output and Output Format	(1) Print of model inputs; (2) Rainfall hyetograph (rate vs. time history) by catchment; (3) Time history of runoff in user-specified stream channels; (4) Time history of sediment concentration in user-specified stream channels; (5) Time history of constituent concentration in stream channels
Linkages to Other Models	(1) Designed to link with nonsteady state receiving water models; (2) Hydrologic linkages result from tributary channel inputs and groundwater inputs (interflow); (3) Final tributary channel output serves as input to receiving water model
Manpower Needs	(1) Environmental engineer experienced in water resources and water quality modeling; (2) Familiarity with Storm Water Management Model and its application useful
Costs	(1) Model acquisition costs are unknown; (2) Program listings could probably be obtained from WRE--Walnut Creek or EPA Planning Assistance Branch for a nominal reproduction charge; (3) Data preparation costs: actual computational costs are unknown
Model Accuracy and Sensitivity	(1) AGRUN can closely simulate watershed with numerous different fields and crop types; (2) Constituent transport limited to sediment-adsorbed substances; (3) No constituent interactions permitted--assumption probably valid if residence time in channels short; (4) No quality constituent (dissolved) transport permitted in ground-water interflow; (5) Rainfall intensity assumed to be uniform over each subcatchment
Other Comments	(1) Only sediment-attached residuals can be simulated; (2) Model originators: (a) AGRUN is a derivative of SWMM for nonurban applications, (b) AGRUN was developed by Water Resources Engineers, Walnut Creek, Ca. for EPA Systems Development Branch, Washington, D.C.

Water-body characteristics identify additional important features including scale, areal boundaries, flow conditions, and other hydrologic and geologic features. Scale relates to the size of water bodies applicable to a model. Areal boundaries relate to the latitudinal and longitudinal boundaries of water systems that a model can accept. Flow conditions include continuous steady flows, nonsteady flows, and intermittently variable flows. Other important hydrologic and geologic characteristics relate to depth, composition of bottom strata, watershed size, and water body gradient. These variables are closely related to flow conditions.

Time properties refer to the way in which receiving water models address time-dependent variables such as flow conditions, pollutant discharges, and meteorological conditions. Receiving water models are therefore categorized as steady-state, quasi-steady-state, or non-steady-state. Another important time property is the averaging time (time step) permitted in a model. For example, a non-steady-state model can provide analysis for a time period of several days with a time step of 1 hour.

Spatial properties of receiving water models include the dimensionality and discretization. Receiving water models can represent one, two, and three dimensions. As several models, including most two-dimensional non-steady-state models, have not been successfully verified (i.e., matched to ambient data different from those used to calibrate the model), the user should be aware of the limitations of these models. Thomann (Reference 13) reviews several quantitative measures of water quality model performance and credibility and how they may be integrated into the modeling analysis. As another important spatial property, discretization is the extent to which a model can divide a water body into functional volumes that can account for geological and topographical variations. Rivers and streams may be divided into reaches and junctions, deep stratified lakes or estuaries into layers, and well-mixed lakes or estuaries into equal-sized grid cells.

Two important mathematical properties of receiving water models are the theoretical mathematical basis of the model and the solution technique for solving the equations in the model. The theoretical mathematical basis of a receiving water model is either stochastic or nonstochastic (deterministic). However, most models combine both approaches. In addition, most of these models incorporate either analytical or numerical solution techniques.

As with many environmental models, receiving water models are based on either conservation of mass and energy or statistical approaches. The conservation of mass and energy approach has been widely used for many years during which time many and varied models have been developed. Statistical methods have provided and continue to provide a valuable and simpler means of analyzing water problems. With continuously increasing computational potential, use of statistical approaches has declined markedly. However, the approach continues to find extensive use in situations where a first-cut analysis is required and where the detailed data required to apply a more complex model based on conservation of mass and energy are unavailable.

For a given receiving water problem, an appropriate model may be selected on the basis of relevant problem specification and model characteristics. Tables 9 and 10 are examples of how runoff problem and model characteristics can be related for use in the general model selection scheme summarized in Table 1.

TABLE 10. SUMMARY OF THE REVISED DISSOLVED OXYGEN SAG MODEL (DOSAG-3)
(FROM HINSON AND BASTA, REFERENCE 12).

Characteristic	Summary Data
Applicable Water Body	(1) Streams; (2) Rivers; (3) Water conveyance canals
Water Body Characteristics	(1) Scale: (a) far field; (2) System boundaries: (a) longitudinal variation only; (3) Morphology: (a) DOSAG-3 permits no stratification (each reach well-mixed), (b) branched watershed network can be simulated
Time Properties	(1) Time variability: (a) steady state; (2) Time units of application: (a) variable depending upon total length of the system and prescribed flow
Space Properties	(1) Dimensionality: (a) one dimension, horizontal plane; (2) Discretization limits: (a) up to 49 stretches, (b) maximum of 80 reaches, (c) maximum of 20 headwaters, (d) up to 20 reaches per stretch
Physical Processes	(1) Principal driving force: (a) net downstream flow; (2) Processes represented: (a) advection, (b) dilution
Chemical Processes	(1) Processes represented: (a) chemical interactions, both coupled and uncoupled; (2) Constituents modeled: (a) dissolved oxygen, (b) carbonaceous BOD, (c) nitrogenous BOD, (d) phosphorus, (e) ammonia, (f) nitrite, (g) nitrate, (h) conservative materials--total N, chlorides, metal ions, e.g., iron
Ecological Processes	(1) Processes represented: (a) photosynthesis, represented by chlorophyll A, (b) coliform die-off, (c) first order bacterial decay, (d) algal respiration; (2) Trophic levels: (a) only decomposers and primary producers simulated; (3) Water quality-biota relationships represented: (a) algal (chlorophyll A)/nutrient dynamics
Mathematical Properties	(1) Theoretical basis: (a) deterministic assumptions; (2) Solution technique: (a) analytical integration (Laplace transforms), (b) utilizes LaGrangian approach where computations performed on reach by reach basis (progressing downstream)
Computation Status	(1) Model coded in FORTRAN IV; (2) Equipment requirements: (a) digital computer with 27,000 word storage capacity, (b) FORTRAN IV compiler, (c) no disks or tapes
Input Data Requirements	(1) Initial setup/calibration: (a) reach length, (b) mean discharge, (c) mean reach velocity, (d) mean reach depth, (e) average reach temperature, (f) residuals discharge inflows, (g) withdrawals and groundwater inflows, (h) constituent concentrations in all inflows, (i) initial constituent concentrations by reach; (2) Verification: (a) observed streamflow, (b) observed stream velocities, (c) observed constituent concentrations throughout the modeled area
Ease of Application	(1) Model documentation: (a) documentation report available from EPA, Systems Development Branch, Washington, D.C.; (2) Documentation contents: (a) complete documentation contains theory, subroutine descriptions, file and data deck design, program listings, and sample problem; (3) Modifications: (a) little recomputation time needed for data deck changes
Output and Output Format	(1) Output information: (a) constituent concentrations in each reach, (b) flow rate, (c) reach depth, (d) minimum dissolved oxygen value and river mile where it occurs, (e) mean velocity, (f) algae growth rate; (2) Output format: (a) tabular printout
Linkages to Other Models	(1) Model cannot handle nonpoint residuals discharges except as tributary inputs (incremental runoff inputs)
Manpower Needs	(1) Requires one junior-level engineer with some basic programming experience
Costs	(1) Model acquisition: (a) documentation reports may be obtained for a nominal charge from EPA, Washington, D.C. or Water Resources Engineers, Austin, Texas; (2) Model setup and data preparation: (a) requires 2-7 man-weeks; (3) Actual computation: (a) \$2-8 per simulation run; (4) Output analysis: (a) small, e.g., less than one hour
Model Accuracy and Sensitivity	(1) Representativeness: (a) assumes constant stream velocity throughout a reach, (b) assumes first order decay only; (2) Sensitivity: (a) high sensitivity to residuals inputs, stream velocities, (b) moderate sensitivity to flow and decay coefficients
Other Comments	(1) Limitations: (a) DOSAG-3 cannot compute nitrogenous BOD and other nitrogen species (ammonia, nitrite, nitrate) simultaneously; (2) Optional features: (a) flow augmentation option to reach prescribed DO standards, (b) model permits three different treatment efficiencies for residuals discharges; (3) DOSAG-3 derived from DOSAG-I by Water Resources Engineers, Austin, Texas

2. General Data Requirements and Data Base Applications

Input data required for water models are widely varied over the broad range of model classes and applications, and are, therefore, not as easily classified as air model data. Many data requirements are specific to an individual model or model application. General types of data that may be required for runoff model applications include climatologic, hydrologic, hydraulic, geographic, ecologic, and source discharge data, as well as initial pollutant concentrations and transport parameters. A given receiving water model may require a combination of receiving water geometry and hydraulics, watershed hydrology, climatological conditions, pollutant discharges and ambient concentrations, transport parameters, and ecological information.

The GSC report (Reference 1) presents abstracts of 11 water data bases for possible application to water models. Many of these data bases contain surface and groundwater quality data, site description, and pollutant discharge data. One of the most widely used computerized water data bases is the EPA Water Quality Information System (STORET). In addition to water quality and pollutant discharge data, this system contains geographic and descriptive station data and stream flow data. Data obtained from STORET have very little use as direct input to water models. The data must be statistically analyzed and reviewed to convert the raw data to more useful information. An example of a much simpler hydrologic data base is the Hydrologic Information Storage and Retrieval System (HISARS) which is presented in detail in Appendix C.

E. SUMMARY

Environmental models may be classified according to many schemes (e.g., deterministic vs. stochastic, temporal and spatial properties, etc.), and the schemes may vary, depending on whether air quality, water quality or another environmental parameter is under consideration. However, generalized schemes (e.g., Table 1 and Figure 1) exist that may be applied to the development of user-friendly, interactive software to aid in model selection.

Although several air and water quality data bases exist, their data are seldom required directly for model input. Rather, their main use is to provide background data for statistical analysis and/or model calibration and verification.

SECTION III

SURVEY OF DISTRIBUTED DATA BASE MANAGEMENT

A. DATA BASES - GENERAL CONCEPTS

1. What Are Data Bases?

In many organizations, computers are used to store and retrieve large amounts of data. Any large repository of data can be called a data base. Such a data base would be no different from a data file. The important characteristic of a data base which distinguishes it from a data file is that the data base should be integrated and sharable.

Integration means that the data base can be thought of as a unification of several distinct data files of an enterprise. Such a unification permits:

- a. expression of any relationship among the different data, and
- b. elimination of redundant data among the different files.

As a consequence of the unification, several users will have to share the same data base. By sharing, different users can access the same piece of data for different applications. The software that performs for integration and sharability is called a data base management system.

2. Definitions

a. Data Models

Data bases contain time-varying data about entities and their relationships in an enterprise. An entity is any object, tangible or intangible, such as an employee or event, etc., of interest to the enterprise. The entities of an enterprise are interrelated to each other. For instance, an EMPLOYEE is associated with a PROJECT, and every project has a PROJECT MANAGER, etc. A data base should be able to capture the relationship among various entities. A formalism of describing data and the relationships among data is called a data model. A data model serves two purposes. First, it serves as a vehicle to describe the various entities and the relationships among entities. The structures supported by the data model allow a straightforward translation of the entities and their relationship into physical data structures. Second, data models provide the operators to manipulate data. The operators are closely related to the structures supported by the model. The operators provide meaningful operations on the data base without the need to know any details about physical storage, job control language, etc.

b. Schema: (Also Called Conceptual View or Conceptual Schema)

The schema is an overall description of the data base, using the structures of any data model. It is the overall logical data base description, consisting of the descriptions of the various entities in the data base, along with the interrelationships among the entities. The schema does not contain any description of physical storage details. The schema is translated into physical storage structures by a mapping mechanism.

c. Subschema: (Also Called External View or View)

The data base is integrated to hold the data of an enterprise. The side effect of integration could be loss of privacy of data. A subschema is defined as the subset of schema of interest to the user. The subschema is stored in the dictionary. Any data base accessed by a user is checked against his subschema in the dictionary to prevent an unauthorized access to portions of the data base.

The subschema acts like a window enabling users to "view" only portions of the data base of interest. As a result, changes to the logical structure (schema changes), e.g., addition of a new entity or relationship, have no effect on the user.

d. Data Dictionary: (Also Called Directory)

A data dictionary is a collection of pertinent information about the data base. It contains data describing the files, programs, users, transactions, security, etc., of an organization. The schema and subschema of various users are stored in the data dictionary. A comprehensive dictionary will also include cross-reference information, showing which program uses which piece of data.

A data dictionary is usually integrated with the data base management system. Such an integrated dictionary provides numerous benefits. The directory can provide accurate and complete data definition for use by application programs. All data base access requests by a user are validated against his subschema.

An on-line dictionary browsing capability would be invaluable, since users will be able to browse through the dictionary. On-line updates to the dictionary allow dynamic changes to passwords, security levels, etc. A dictionary could also generate reports about access frequency by users, access frequency of files, etc., which would be helpful in determining the usage patterns of the data base. In distributed data base systems, the directory contains information about data and their location, traffic routing information, etc.

The directory of a traditional data base system contains the following information (Reference 14):

- (1) The logical data description (name of the relations or files, etc.).
- (2) Physical structure description (field formats, inverted lists, etc.).
- (3) File statistics (size, etc.).
- (4) Other details such as security restrictions, ownership, etc.

For a distributed data base to successfully operate, another category of information is needed, namely the location of data in the network. This information is required to execute global queries. Since a

global query could be issued at any node, all of the nodes require access to the directory. Various directory management schemes have been proposed (Reference 14).

- (1) The centralized approach, where the directory is stored only once at a central site.
- (2) The fully redundant approach, where the directory is stored in its entirety at every site.
- (3) The partitioned approach, in which each site maintains its own catalog for objects stored at that site.
- (4) Any combination of the above methods.

The factors that influence directory management are directory retrieval frequency, directory update frequency, and reliability. Frequent directory updates encourage a centralized directory, while frequent directory retrievals encourage a redundant directory. A fully redundant directory is more reliable than a centralized directory.

3. Centralized Versus Distributed Data Bases

In recent years, data base technology has been enhanced to manage geographically distributed data. The distributed approach fits rather naturally with the structure of certain organizations like the Air Force. These data management systems are called distributed data base management systems (DDBMSs). Distributed data bases are appropriate when several data bases already exist in an organization over geographically separated sites, and the necessity of performing global transactions is imminent.

A distributed data base is a collection of data files not stored in its entirety at a single location, but spread across a network of locations. The various locations in the network are interconnected by a communications link. A distributed management system provides location transparency and replication transparency to all users (Reference 15). Location transparency means that users need not know the location details of any data item. Instead, all information about data and locations is maintained by the distributed data base management system. A user request to operate on nonlocal data is handled by the system via (1) moving the data to the local site, (2) processing the request at a remote site and moving the results, or (3) a combination of (1) and (2).

A distributed data base may contain data replicated at many sites. Such replication is done to provide improved performance and availability. Application users need no longer communicate with remote sites for the data, thus ensuring better performance at the local site. Data updates at one site will have to be propagated to all copies of the data. Replication transparency implies that the DDBMS will support all details of maintaining and updating replicated data without the user's intervention. Replication and location transparency ensure that a distributed data base appears as a centralized data base to the users.

Distributed data base management systems offer four advantages:

- a. A distributed data base is more reliable since multiple computers are present. Failure of a single computer does not cause a total system breakdown.
- b. Replicated data at different sites could result in better performance of processing queries or transactions and improved availability.
- c. Distributed data bases allow local autonomy and yet provide sharing of global resources.
- d. A distributed data base could be easily upgraded. More user nodes could be added without a service disruption.

Distributed data bases could evolve from interconnecting various existing local data bases or from building a new system. Interconnecting various existing local data bases would generally result in a heterogeneous DDBMS since each of the existing local data bases would be different. A homogeneous DDBMS is one in which all local data bases use the same data model and are erected and processed by the same DBMS.

B. DISTRIBUTED DATA BASE MANAGEMENT SYSTEM (DDBMS) OVERVIEW

1. Architecture

The general architecture of a DDBMS is shown in Figure 2 (Reference 16). The user can state his requests, using a data manipulation language. The query is evaluated, using the global schema catalog, and based on the evaluation, the query is decomposed into subrequests involving the local data bases. The results from all the local data bases are recomposed to give the results to the user.

The distributed executive is responsible for interacting with the local DBMSs. It performs various functions like data format transformation, resources allocation, addressing and routing of data, etc. In the case of heterogeneous local DBMSs, the adaptation module handles problems due to the differences in data formats, data types, and representations.

The global schema is a superset of all the local schemata. In a homogeneous DDBMS, the local schemata are expressed in the same data model at all the nodes. The global schema is also expressed in the same data model. In a heterogeneous DDBMS, the local schemata are expressed in different local data models. In such a case, a global data model which can effectively express all the local schemata is chosen.

Distributed data bases are not just a distributed implementation of centralized data bases, since they possess many characteristics different from a centralized system. In distributed systems, the idea of centralized control is deemphasized. Each site could function independently of the others. This property is called site autonomy (Reference 17). Distributed data bases may differ very much in the degree of site autonomy, ranging from complete autonomy to almost complete centralized control.

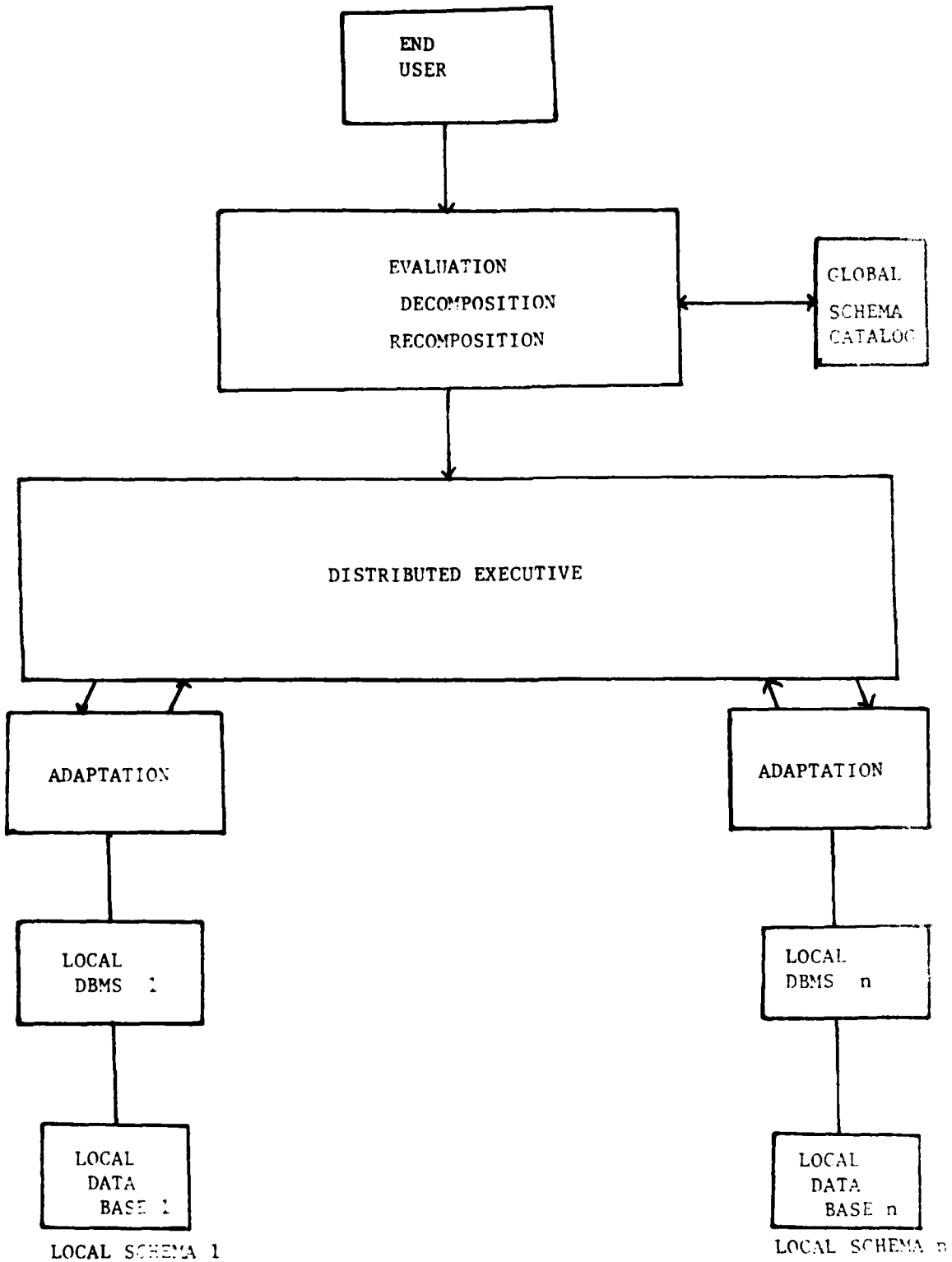


Figure 2. General Architecture of Distributed Data Management System

A data base user interacts with the data base using a data manipulation language (DML). The DML is used to retrieve, insert, and modify data to and from the data base. In a conventional DML, the user not only specifies what he needs from the data base, but also indicates how the data are to be retrieved. Such a language may not be efficient since the user specifies how the data are to be retrieved. A query language is a high-level language where the user specifies the needed data and the DBMS retrieves the data, usually in an optimum manner. Such a high-level interface could be a menu-driven, user-friendly interface.

In a DDBMS, a user query may require several remote data bases to be accessed. In a replicated DDBMS there is usually more than one unique way of satisfying the query. The given global query is decomposed into a set of subqueries, where each subquery accesses one of the remote data bases. A global query can be decomposed into different sets of subqueries. Choosing the best strategy to satisfy the query is termed query optimization. The set of subqueries is chosen such that the cost factor in executing all subqueries is minimized. Some of the factors included in the cost formulation include I/O cost, CPU cost, communication cost, etc.

2. Basic Concepts and Issues

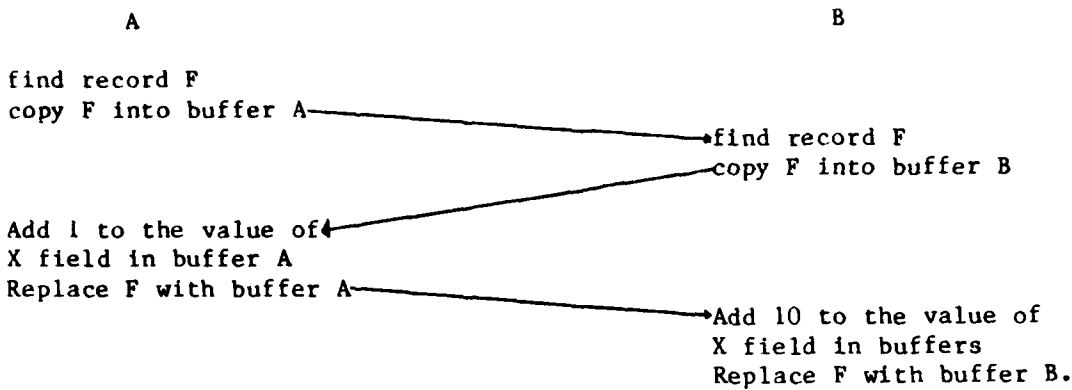
a. Consistency and Integrity

Integrity is the problem of ensuring that the data in the data base are accurate. Inconsistency between two entries representing the same data is an example of lack of integrity. Other integrity constraints could be specified, such as, "the age of an employee cannot be less than 10 years." Integrity can be enforced by defining validation procedures to be carried out whenever new data are inserted or an update operation is carried out.

In a DDBMS where data are geographically dispersed, integrity problems are more complex since data are replicated at many nodes. Update operations performed to any copy have to be propagated to all the copies. The strategy of propagating updates immediately to all copies may not work, however, since some of the sites may not be operational or various users could be accessing the data at that instant of time.

b. Concurrency

Integrity problems are compounded by the concurrent usage of replicated data by users. When multiple users access the same data precautions should be taken to mutually exclude the data from each other. Consider the example shown in Reference 15, where users A and B modify the same record F at different times.



In the above example, the update of A has been lost. When transactions are interleaved in time, care should be taken to see that all transactions operate exclusively on common data of interest. The above problem could have been avoided if B followed after A's completion or vice versa. When transactions are executed serially, the data base remains in a consistent state.

Two of the common techniques of concurrency control are locking and time-stamping.

(1) Locking

A lock is used to prevent access to data currently in use. In the locking technique of concurrency control, all portions of the data base being written or read by active transactions are locked (Reference 16). The locks are released only after the end of the transaction. This simple but inefficient strategy ensures correct results.

Locking could result in a deadlock. A deadlock is created when a transaction T_1 locks resource A and is concurrently waiting for resource B locked by ongoing transaction T_2 . In turn, transaction T_2 is waiting for T_1 to release resource A. Both transaction T_1 and T_2 are in indefinite wait state without each other's knowledge. Deadlocks are avoided either by preventing or detecting them. In the detection method, all transactions which do not end in a specified time-out period are aborted. In the prevention scheme, all resources are allocated only after checking for any deadlock possibilities.

(2) Time-stamping

In time-stamping, every transaction is assigned a globally unique identifier called a time-stamp, which can be thought of as the transaction's start time. The basic idea of time-stamping follows.

All updates in any transaction are applied only after the successful completion of a transaction. Every stored record in the data base carries a time-stamp of the last transaction that read it and a time-stamp of the last transaction that modified it. If a transaction T_1 requests a data base operation in conflict with a transaction already executed by a younger transaction T_2 (with a later time-stamp value), then the transaction T_1 is restarted. All restarted transactions are assigned new time-stamps.

C. SURVEY OF SOME DISTRIBUTED DATA BASE SYSTEMS

1. Survey Objectives

Some of the distributed data base systems available today are surveyed. The aim of the survey is to analyze each system and ascertain whether or not they could be applicable for AFEMDEX. The factors considered here are the properties of each system and the environment to which these systems would be best suited. The survey will also point out those properties that are desirable for AFEMDEX. Each system is briefly described here and a more detailed explanation is presented in Appendix A. The five systems most relevant to the project are surveyed. They are MULTIBASE, R*, SDD-1, distributed INGRES, and CSIN.

2. Summary of Features of Some Distributed Data Base Management Systems

a. The MULTIBASE System

MULTIBASE is a distributed data base management system (DDBMS) developed at Computer Corporation of America (CCA). The system was developed to integrate different data bases with different DBMSs, different languages and data models. The purpose of MULTIBASE was to suppress these differences and provide the end user a unified global schema and a single high-level query language. The integration of various heterogeneous data bases (each with its own DBMS) is a complicated problem, since all the different schemas have to be mapped to a single global schema. The other important feature in MULTIBASE is to provide autonomy to the local data bases; this is necessary because these different data bases may belong to different organizations who would not want their data base to be altered. This automatically implies that the data base can be altered only at the site where it is stored; all global requests can only read from it. This makes MULTIBASE a global "read only" data base which permits only local updates.

The above property makes MULTIBASE useless for applications where global updates are essential, e.g., in a banking environment. In spite of this restriction, MULTIBASE can be employed effectively in environments that do not require instantaneous updated information and where site autonomy is desired. The fact that updates are only locally managed greatly simplifies implementation, and problems like concurrency and consistency of the global data base are reduced to a local level.

The architecture of MULTIBASE is built in such a way that a single global request can make an integrated retrieval from the constituent data base. The global request has to be broken down to different subqueries, where each subquery is sent to different sites that contain relevant data. These subqueries have to be in the language used at the site. Moreover, inconsistencies in data, such as naming or different levels of data abstraction, have to be resolved before the data can be accessed. This incompatibility resolution involves a lot of software and is labor- and cost-intensive. MULTIBASE solves this problem by employing a global data model, called the functional data model, to define the global schema and by using DAPLEX as the high-level query language. All global requests are first in the global level after which they are mapped into the local level. The data, however, travel in the opposite direction. They are retrieved in the local level and then get mapped into the uniform global level.

The general environment for MULTIBASE is similar to that of AFEMDEX. However, MULTIBASE deals with a complicated set of issues such as integrating different schemas, different DBMSs and different languages. This is not the case with AFEMDEX, where a more uniform set of schemas is sufficient for its applications.

b. The R* System

R* is an ongoing research project at IBM. It is a homogenous distributed data base management system supporting the relational data model. It is built upon the System-R centralized system. It transparently supports access to data distributed at multiple processing and storage sites. It consists of individually controlled data base sites which are interconnected by the IBM communication product CICS. R* stresses onsite autonomy, and, in most of its implementations, this issue is of foremost importance. Each site is responsible for maintaining its data. No fixed central site supervises the action of the remaining sites in the network. The site at which the query originates becomes the 'master' as long as the query is being processed; after the query is finished, it ceases to be the master. Data are searched by referring to their birth site (site at which the data are created). The birth site contains all network information about the location and status of the data.

R* does not deal with data incompatibilities as in MULTIBASE, but it provides a greater degree of site autonomy. Global updates are permitted. Data may be stored in cache catalogs for faster processing. This information is not updated regularly; hence, the most recent update information may not be used. In such a case, the query is reprocessed with the latest version of the data. This is done automatically by the system, and the end user is guaranteed the latest updated information in the network.

R* provides instantaneous updated information at any site in the network. Although site autonomy is desirable, global updating facility is unnecessary for AFEMDEX, since there is no need for global updates to the data bases of AFEMDEX. Moreover, AFEMDEX may not have an environment as R* but it certainly will not have totally homogeneous data bases, and there will be a need to provide for some type of heterogeneous data base interfacing.

c. The SDD-1 System

SDD-1 is a prototype-distributed data base system developed at the Computer Corporation of America (CCA). It is one of the first distributed systems to be developed. The main goal of this system was to provide a truly transparent data base facility where the end users could assess any data at any site without knowing anything about the site at which the data are stored.

SDD-1 assumes the interconnection of homogeneous data bases and DBMSs and makes no effort to deal with heterogeneous data bases. The aim is to address specific data base issues such as concurrency, consistency, reliability, and distributed query processing. It has a set of highly developed mathematical protocols to prevent deadlocks that may occur in the data bases. This contrasts with other systems that adopt a simplified approach - allowing a deadlock to occur and then detecting and breaking it.

The functional architecture of SDD-1 can be divided into three classes: the first deals with the ordering and coordination of the transactions, the second deals with the actual manipulation and temporary storage of data, and the third is the network that holds the whole system together.

Data may be split up and stored at different sites to aid fast retrieval. This makes updating more difficult because of the overhead involved in searching for these bits of information scattered at different sites. The relational data model is employed, as in R* and remote updates are permitted. From all these properties mentioned, SDD-1 is suited for application in which R* is employed. The main difference is that there is no site autonomy; this disqualifies it to interconnect data bases from different organizations. It also leads to a more complicated centralized control system since all data base issues have to be addressed from a global viewpoint. So SDD-1 can be regarded as the least suited system for the AFEMDEX requirements.

d. The Distributed INGRES System

Distributed INGRES is an extension of the centralized INGRES system developed at the University of California. Most of the features of the centralized system were preserved, while designing the distributed version. The centralized version was built on the UNIX system which greatly influenced its design. Distributed INGRES deals with homogenous systems and like the SDD-1, it does not address the problem of interconnecting heterogeneous data bases.

Query processing is done with the master-slave technique used in R*. The site at which the query originates is the 'master' and all other sites are the "slaves". By employing a standard protocol called the "two phase commit protocol" the query is transferred between sites for retrieval of data. If the "master" fails during a transaction, one of the slaves takes over as master just long enough to safely complete or abort the transaction.

Each component machine in the network has its own concurrency control and it is similar to the centralized system version. Unlike SDD-1, there is no sophisticated deadlock detection technique; a global deadlock detection scheme is employed. This is done by a centralized machine called "SNOOP".

e. The CSIN System

The Chemical Substances Information Network (CSIN) is built by CCA. It is built upon several heterogenous data bases. It interfaces these data bases to form an integrated view for the user. Instead of providing a global- and local-level technique as in MULTIBASE, CSIN employs a layered architecture that transforms a global query into local requests and recombines the data from the different sites for the user. While a query language is used in MULTIBASE, CSIN uses only a menu-driven, user-friendly query technique.

Many organizations are tied together by CSIN, and site autonomy is mandatory. The data are highly varied and can range from bibliographic to numeric. The queries also vary considerably because the data are shared by

users from various disciplines with different interests. For all these users, CSIN has to provide a uniform interface. Many frequently repeated tasks such as log-on (to the various systems) and common queries are automated by the use of scripts, simplifying the handling of the system. There are two modes of operation: the script mode and the manual mode. In the script mode the system automatically does all the transformation of data between the sites, and the details are not important to the user. In the manual mode, the user must know all details about the other site to log-on and proceed with the transaction.

D. APPLICATION OF DISTRIBUTED DATA BASE CONCEPTS TO THE CURRENT PROBLEM

With different data bases located at different locations, the AFEMDEX project plans to interconnect these data bases by means of a loosely integrated network. This would facilitate the sharing of data which are stored at the different sites. Most of the data are in the form of tables with one entry per field.

1. Data Base Requirements for AFEMDEX

AFEMDEX is concerned with the linking together of heterogeneous data bases. These different data bases have various data formats and may not be compatible with all the component data bases connected to the system. The data retrieved must then be formatted to run on the environmental models that reside on some other node in the system. The basic requirements in this design are to interface the different systems and to provide a data reformatting capability.

a. The AFEMDEX High-Level Language Interface

The initial AFEMDEX design points to the development of a user-friendly, menu-driven interface. The development of any procedural high-level language is not necessary, since the users will not be expected to formulate queries conforming to syntactic rules. The user-friendly, menu-driven system will prompt and guide the user to model and data selection. The search for the model can be in the form of a hierarchical structured menu search or it can be a boolean expression of the different categories of interest. The search aids the user, who is not sure of the specific model required for a particular application. The system will query the user about the characteristics needed and then supply a choice of models that match these characteristics. In this way, the user can interact with the system without the need for any high-level query language.

All the environmental models (see Section II) are stored at specified sites called model sites. These sites should have the capability to run these models.

b. Query and Data Compatibility

The data stored in different data bases are in a general form. To extract these data for a particular application, data formats are required to be changed. This is done by providing an interface to transform data in a form compatible with the model. Besides changing the form of the data, there should be an interface to make the data compatible to the system on which the model resides. The interface transforms format-specific data for a particular site to a format that is uniform network-wide.

The interface will be component-specific and will depend upon the organization and capabilities of the component systems. This is necessary because the interface not only should carry out the formatting of the data, but should also bring up each component to an equal level of capability before interfacing to the network. The other way around this problem is to follow the approach provided in MULTIBASE. Here a filter removes all operations not supported at a particular data base site and transfers them to the "auxiliary" data base, which processes the query and transfers the result to the site from which it came. This approach calls for an independent data base that maintains only this mapping operation. The proposed methodology will have no auxiliary data base. This provides for modularity and implies easy expansion of the network. New component systems can be added onto the network without disturbing the rest of the system. The only significant software to be added between the new component and the network will be the interface unit. Other sites can extract data from this new component just like extraction from preexisting nodes.

The interface unit should be able to transform the global data addressed to it and convert them into the component specific form. Hence, a two-way conversion exists to provide for proper data transfer within the network. Data are transformed in the form of files between the component systems.

The query transformation problem resembles the data transformation problem. Here the user constructs a query after obtaining the necessary information from the system. All queries are of a standard format and the user just fills in the details of the search. The query is then analyzed at the originating site and is decomposed into a subset of queries, each destined to a particular site. According to the nature of the data base environment at AFEMDEX, there may not be more than two sites from which data must be extracted at the same time for a single global query. These subqueries preserve the format of the original query, the only difference being that each will contain only component specific information and not global query requests as the original query.

The subqueries are then transferred to the appropriate sites. Upon reaching their destinations, these queries are transformed into a form that is understood by that specific system. This is done by the interface layer. However, the interface does only one-way conversion; that is, it converts the subquery to a form compatible with the specific component system. Moreover, all the subqueries are in the standard format and no transformation has to be done. The site where the query originated should split the query into subqueries that only deal with local data and should not be sent to a second site for further operations. This does not imply that data extracted by a subquery are prevented from reaching another site during the execution of the query.

The method of transforming data to a standard network-wide format before releasing them into the network is essential, because the data can be edited manually at the user's site, if he or she desires. Editing capability should be provided by the system, so that the user can restrict the volume of data extracted if a great volume of data is retrieved. If this facility does not exist, the user must abort and restart the entire transaction just to correct a single error in the query that resulted in the

unexpected amount of retrieved data. It also facilitates the minor changes of data that may be necessary before being sent to the model site. This editing of data is not mandatory, and is bypassed if the data retrieved satisfy the user. The user is informed about the data status by system statistics concerning the number of data records retrieved, etc. The editing facility should include certain simple built-in system functions like delete, insert, restrict, and basic arithmetic operations.

c. Network Communication

The AFEMDEX network should provide fast and uncorrupted transfer of data between the participating sites of the network. A packet-switched network will be employed, such as the TELENET/ARPANET. The network will form an interconnection between three basic types of nodes. They are the nodes on which (a) the models reside; (b) the data reside; and (c) the model user resides. The user's request will originate at the user node; then the subqueries are transferred to the different nodes on which models and data reside. The final result is transferred back to the user node. Security measures can be incorporated into the network so that only a certain class of users can access restricted data. Provisions for time-outs for remote transactions can be incorporated into the network, so that the system will not wait unnecessarily for a reply from a remote site that has failed, or wait while an error has been detected at the remote end and the remote site is awaiting new commands from the original query site.

2. Evaluation of the CSIN Approach for AFEMDEX

The survey of the various DDBMSs has enabled a comparative study of these systems. It is also possible to evaluate their application to the AFEMDEX project. Each system has distinctive properties designed to serve a particular user environment. These systems are compared in Table 11. The first step in the selection of an appropriate design for AFEMDEX is to compare its users' environment with that of the five systems surveyed. The next step is to examine the properties of the system whose environment closely matches that of AFEMDEX and to determine the appropriateness of those properties.

SDD-1 was built from scratch and comprises homogeneous component systems. Distributed INGRES falls into the same class. These two systems do not deal effectively with the problem of heterogeneous data bases and systems. So these systems automatically have basic differences with the AFEMDEX system. MULTIBASE deals with heterogeneous systems in roughly the same environment as the AFEMDEX. However, one major difference between the MULTIBASE and AFEMDEX environment is that the systems underlying MULTIBASE support different data models. This is a very general case and is difficult to implement. These different models are mapped to a single model called the functional model; from this model the different views of the data are extracted. This problem does not exist in the AFEMDEX case, for all the data are assumed to be in the form of tables and more or less represent statistical data. The CSIN effort concentrates on the interface and reformatting problems of the component systems. It operates in a highly heterogeneous environment comprising various organizations and institutions. This is the same environment in which AFEMDEX will function, although the variety of the component systems may not be so large.

TABLE II. A COMPARATIVE REVIEW OF DISTRIBUTED DATA BASE MANAGEMENT SYSTEMS

Features	Interface Homogenous Data Bases	Built on Existing Centralized Systems	Data Language	Data Model	Site Autonomy	Remote Update	Access Planning Optimization Technique
SDD-1	Yes	No	DATA-LANGUAGE	relational	variable; depends on the designer	Yes	hill climbing
R*	Yes	Yes	SQL	relational	very high	No	breadth-first tree search; pruning heuristics
MULTIBASE	No	Yes	DAPLEX	functional	normal (only local updates)	No	-
Distributed INGRES	Yes	Yes	QUEL	relational	high	Yes	hill climbing
CSIN	No	Yes	Menu-Driven	-	high	No	-

The other aspect of a heterogeneous system is to protect each of the component systems from the other. This is done by providing autonomy for each of the sites. All these systems are strong in this area. In SDD-1 the data base designer can ascertain the degree of autonomy. MULTIBASE provides only local updates and hence maintains a normal level of autonomy. In R*, site autonomy is a central issue in the design and it has the highest degree of autonomy for the component sites. CSIN provides enough autonomy so that the data belonging to the various organizations are not disturbed by accident. AFEMDEX requires the same amount of site autonomy as CSIN. If the network of AFEMDEX evolves such that site autonomy is not a paramount concern (all nodes controlled by one organization), then the design could tilt more towards the distributed INGRES approach where there is some form of centralized control for managing the directories of the network. All systems except CSIN have a high-level query language as shown in Table 11. CSIN does not have a high-level query language because the research project is not yet completed. Most of these high-level languages compromise greatly on ease of use and are not user-friendly. A system that guides a user to the result is more straightforward and simple to use. A menu-driven user interface, with some prompts from the system, can replace a high-level language. This is also desirable for the AFEMDEX case because the data to be retrieved are more or less fixed, and there is not much semantic variation to the queries which would make a high-level language necessary.

One of the greatest advantages in a distributed data base environment is the capacity for parallel processing. The degree of parallel processing varies with the different systems. SDD-1 promotes parallel processing only if the data reside in numerous sites. This is a prerequisite for parallel processing. Distributed INGRES lifts this restriction by replicating the data at as many nodes as necessary. Valuable processing time is wasted when data are replicated in many nodes and the query requires only three (or fewer) relational tables to arrive at the result. The emphasis on site autonomy in R* and MULTIBASE severely restricts them from performing remote updates, although they provide for concurrent processing. The amount of parallel processing to be performed depends upon the type and nature of queries and data, i.e., it is desirable to build a system with a high degree of parallel processing if the queries often require data that are scattered around the network. AFEMDEX may typically require one site to retrieve the data. These data may be later shipped to another site for further processing. However, the number of sites involved at any one time will seldom be greater than two. This points to low amounts of parallel processing. The analysis of CSIN also points in the same direction; there is little or no software to enforce this type of processing - although for any one query more than one node may be necessary for the final result.

On examining the type of data to be processed, it is seen that CSIN deals with a large variety of users and data. AFEMDEX deals with only numeric data, so all the extra features to handle bibliographic data are not required for AFEMDEX. The layered architecture in CSIN is an extremely elegant form of design and can be employed for AFEMDEX.

C. SUMMARY

Of the systems surveyed, CSIN has the closest environment and properties to that of AFEMDEX. However, there are many differences in the nature of data handled. Both CSIN and AFEMDEX transform data between heterogeneous systems. In CSIN it is done in the transaction layer and in the data layer by using the component data interface. These interfaces are software units and comprise the bulk of the data layer architecture. They are tailored to a particular component and hence are not portable. So AFEMDEX would have to be concerned about developing these software packages that would interface these various heterogeneous systems onto the network. Therefore, although the CSIN approach may roughly coincide with that of AFEMDEX, it is not possible to transfer the CSIN system for direct AFEMDEX use.

SECTION IV

A PROPOSED ORGANIZATION OF THE ENVIRONMENTAL INFORMATION NETWORK

A. INTRODUCTION

In Section II, examples of various environmental models and data bases were presented, illustrating possible interaction with an Air Force user. Several distributed data base management systems (DDBMSs) were reviewed in Section III for potential AFEMDEX use, with Air Force needs in mind. The major requirements of an Air Force Environmental Information Network include:

1. A menu-driven user interface: This would make it easy for both novices and experienced users to operate on the network. Such an interface would provide the users with transparency against the complexity of the network operations.
2. A uniform query language: This would facilitate easy access and use of the network. This would also make software portable, i.e., software can be transferred from a node to another node and used without any further modifications.
3. Interface support to handle heterogeneous data bases: The data bases to be interconnected in the network are diverse, both in content and format. Commands issued at any node may need to be translated into a different sequence of actions depending on the local data base characteristics. All local data base-dependent information is maintained by the network.
4. Data formatting and editing capabilities: The system should allow for any retrieved data to be displayed on the terminal. The user should be able to modify or delete part of the displayed data before the data are used to run with any model.
5. Data dictionary browsing capability: The system should allow the user to browse through the data dictionary. The dictionary (directory) should provide information about the data (e.g., their storage and characteristics), as well as information about the model.

Some of the DBMSs surveyed (e.g., SDD-1, R* and INGRES) do not handle heterogeneous data bases, thus making them unsuitable for present needs. Although both MULTIBASE and CSIN do handle heterogeneous data bases, the former is designed to handle more complicated issues than will be needed by the AFEMDEX system, such as integration of different DBMSs, different languages, etc. Since it is anticipated that only a uniform set of schemas will be necessary for the Air Force environment, the excess capabilities of MULTIBASE are not required.

CSIN provides an interconnection among different data bases which is similar to the Air Force needs. But the cost associated with the development of interfaces to specific Air Force data bases would make CSIN unattractive. Besides, CSIN was developed only for data retrieval.

However, several functions present on CSIN are proposed for the Environmental Information Network:

1. A menu-oriented user interface would be useful in both selection of models and data. The menu displays the various options at every level and guides the user through the system.
2. During data retrieval, the system would give the user an estimate of the volume of data, asking the user how much he would like to receive. This would prevent large amounts of data from being displayed by an ill-formed query.
3. The CSIN concept of "file substitution" should be implemented, which implies that the amount of keying-in of commands can be minimized by letting the user substitute predefined files to supply run-time parameters. As an example, the user-supplied data could be merged with data retrieved from a data base in the proper format and run with a model. The above process could be done by "file substitution," where the user supplied data are stored in a file and the file is specified by the user during run time.
4. The network shall also provide site autonomy. Site autonomy refers to the local control of processing at each site of a network. This would enable different agencies to support and maintain their data bases.

In summary, the architecture for the proposed Air Force system can be designed more simply than the effort implied by adapting CSIN or another existing DBMS directly to the AFEMDEX computing environment. The capabilities are enumerated in the following subsections; components of the system will rely on knowledge of existing DBMSs wherever possible. Several design considerations relevant to the Environmental Information Network are discussed in Section V, e.g., design of directories and formatting/reformatting of data.

B. A CLASSIFICATION OF THE INTERACTION AMONG MODELS AND DATA BASES

1. Class 1: All Data Supplied Locally

Probably the majority of environmental models fall into this category. For such a model, the user supplies all input data without need to access a formal data base for some parameters. For instance, an air pollution model may require physical data about the source, such as height, diameter, emission rate, stack velocity and temperature, etc. Such data are particular to the problem at hand and must be tabulated without recourse to a formal computerized data base. (This is not to say that all such data will never be found in a data base. Rather, under ordinary circumstances, such data will be tabulated for model input on a case-by-case basis by an engineer familiar with the particular site and conditions being modeled.) Although most water and air quality models are still run in a batch computer mode, models within the Class 1 category are well-suited to an interactive mode since they tend to have fewer input data, and the input data may often be readily described in an interactive format (see Section II).

In general, models within this class should require fewer complex DBMS features, since there will not be a requirement to integrate model input from

the user and a separate data base. Rather, the system will only have to convert the data input by the user into the proper format for input to the model. However, another important feature of the model - data base interaction is one in which model input from users and separate data bases must indeed be integrated. This is discussed below.

2. Class 2: Use of Local Data and Remote (Data Base) Data

a. Class 2A: Interactive Target Data Base

Within this class, most of the model input data will still be supplied directly by the user, but a portion must be drawn from a separate data base. The separate data base may be either remote (at another location from the user) or located on the same computer on which the model is being run. For Class 2A, it is assumed that the target data base may be accessed in an interactive mode, using software developed by the developers of the data base itself. That is, it is proposed to use existing software where it already exists; except under extraordinary circumstances, it should not be necessary or desirable for the Air Force to redevelop interactive software where it already is functional due to the efforts of others. For instance, comprehensive and sophisticated interactive data bases presently exist for water (STORET) and air (SAROAD) data as described in Section II. It would be wasteful to reaccomplish these efforts. The penalty for this policy is that the user of the Air Force system must become familiar with the data bases accessed in this manner. However, this is appropriate since it is highly desirable that the user of the system know about the components used in the analysis of the problem at hand.

Most comprehensive air and water quality data bases (as well as ones devoted only to nonquality data) have the facility to present data in a self-describing format, e.g., one in which rows and columns are labeled, a station identification is given, etc. Such a format permits easy comprehension by the user. For use by the model, the particular self-describing format must be known and cataloged in advance by the system to reformat the data for model input. For instance, it must be possible to extract the required data array from the display of text and data. The description of such data, which is known a priori, is stored in the dictionary. Dictionaries are discussed in Section V.

b. Class 2B: Batch Target Data Base

This class has similar attributes to Class 2A, but in this case a program must be written to access data on the target data base system and submitted in a batch mode. The HISARS data base is an example of such a target. Since most data needs from separate data bases will be simple and common to several models, there should not be a large variety of types of queries to these data bases. Hence, it is proposed that the Air Force system be able to generate the appropriate target system batch commands to retrieve the data. The commands will be generated in response to a menu-driven dialogue with the user. The system will then either call upon a standard retrieval program from a library, or else construct such a program if the target system is simple enough. In most cases, it should be possible to catalog an exhaustive program library for the set of models and data bases currently in use by the Air Force system. Again, software must be available

to reformat the data returned by the target data base into a format suitable for the model being used.

With this system, there will almost certainly be a time delay during the execution of the batch job. The magnitude of the delay will depend upon the type and location of the target data base, and the current level of use. The present mode will be deemed acceptable in situations where this delay can be tolerated.

c. Some Special Cases

(1) Frequently Used Forms of Data

Certain classes of models may use a common form of input data. For instance, the long-term, climatological air pollution models discussed in Section II (e.g., CDM) all require input of a 576-element array of 6 wind speeds by 6 stability classes by 16 wind directions. Each element of the array contains the relative frequency of occurrence for that particular combination over the annual or other interval being modeled. The same frequency array is used by several air pollution models, since NOAA's National Climatic Center has standard programs to process raw meteorological data to produce that array in a standard format. One way of obtaining this array and inserting it into a model input data string would be to direct separate requests to NOAA each time such an array was needed. However, since Air Force use of such data will most likely be at a limited and predictable number of sites (e.g., the 113 air bases), it would be reasonable to obtain the 576-element arrays for all such locations, and store them in a data base, easily constructed for just this purpose, to be resident on the Air Force computer system.

As another example, several long-term (continuous simulation) hydrologic and water quality models use the time series of hourly precipitation at a rain gage near the basin being modeled. Again, NOAA has placed these data on magnetic tapes in a format that has been inserted into several water models. If the likely locations of model applications by the Air Force can be anticipated, the precipitation records for these areas may be obtained and stored on an Air Force system data base. Such an arrangement permits ready and inexpensive access to standard and frequently used data.

(2) Statistical Analysis and Review of Data

It is important to recall the ultimate objective of the use of the DBMS under consideration for the Air Force: to be able to aid in solving real engineering problems. Almost all such problems involve many steps, besides the execution of a mathematical model. One of the most important steps is often a review of relevant data, accompanied by simple analyses of the data. For instance, in an air pollution problem, it is important to assess background ambient concentrations prior to simulation of changes or control options, and similar concepts apply to water, noise, hazardous waste, or whatever environmental problem is under analysis. This means that there will often be a need to examine and analyze data themselves, apart from using them as input to models. Statistical analysis may be thought of as a sort of "model" in itself; one important function of the Air Force software will be to facilitate such analysis through use of available

statistical libraries of programs, such as the Statistical Analysis System (SAS). Depending on the routines ultimately employed in the Air Force system, the statistics software may itself be used to manipulate the data into the proper format for analysis, or unique software may be needed. In any event, the use of the system simply to retrieve and review the basic data relevant to the problem is likely to be one of the most important and must be provided for in its development.

C. OVERALL ARCHITECTURE AND FACILITIES

The overall flow chart for the sequence of actions is shown in Figure 3. After model and data selection, the network must gather the model and the data at a common node. This would involve shipping both the model and the data to the user's node, or shipping the data or model to a convenient node and getting the results back to the user.

The system provides a common user interface across the network. Any commands issued by the user to retrieve data will have to be translated into the sequence of actions appropriate to the local data bases. A standard format is proposed for the queries. The format may be designed in such a way that a given set of formats apply to several of the models. The query transformer module splits the user's query into components for each of the different sites involved in the transaction. The relevant instructions, such as JCL commands, log-on procedures, etc., are generated at the user's site and are transferred to the remote site. This would mean that when a user's transaction needs to access data from a remote site, the user's system generates the commands needed to log on at the remote site. Similarly, when the transaction at the site is over, an automatic log-out is performed.

The network employed will be a packet-switched network similar to TELENET/ARPANET. In packet-switched networks all data are transmitted in the form of packets which are self-contained blocks of information containing adequate supporting data for transmission/routing. All the standard issues of network communication are incorporated with some additional features. At the beginning of a transaction, a certain predefined time-out interval is specified. If the transaction exceeds this time limit, the transaction is aborted and tried at a later time. The user is informed about the status of the operation. Special considerations for security can be incorporated.

After receiving the query, the different sites will retrieve the data (after the appropriate projection and selection) and store them in a local temporary file. All the data are then transferred to the site where the model resides and the data are transformed back into a format compatible with the model. The data manipulator (editor) provides facilities to manipulate the data before they are run with the model. This is especially useful when there are obvious discrepancies in the retrieved data. Certain simple inbuilt functions like deletion, sorting, merging and selection can be employed at this stage.

The properly formatted data are now run with the model. It is assumed that the model's site supports all the necessary facilities to run the model. The final output, which is format-specific to the model, is captured on a file and transferred to the user's site where the final output format can be specified and hard copies requested.

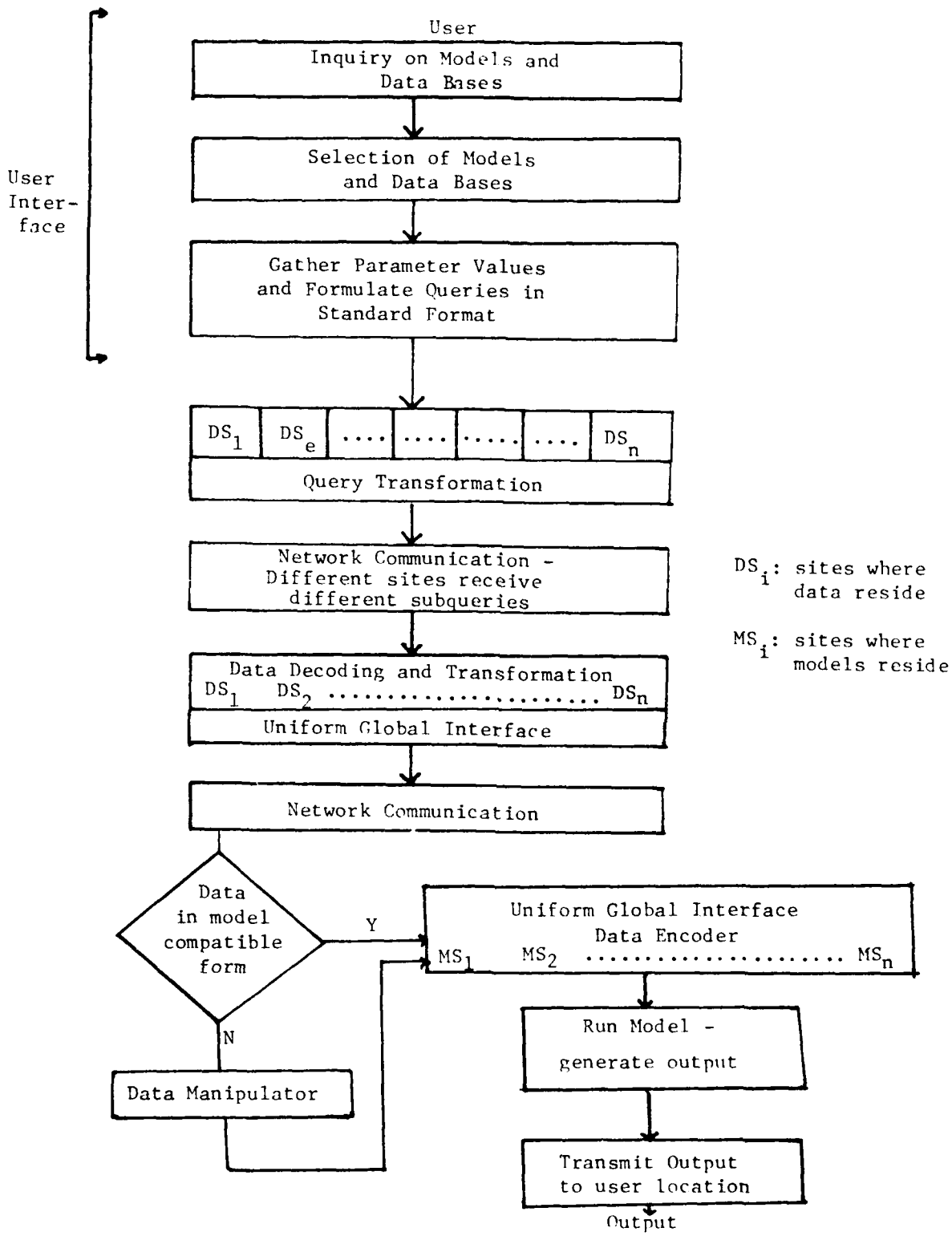


Figure 3. Design Flow Chart of the Environmental Information Network

From the users' point of view, the system should provide simple commands for model selection, data retrieval and gathering, and model execution. The data retrieval and data gathering are highly model-dependent, as shown in Subsection B. Two different functions (or sequences of actions) to perform data gathering are proposed: F1 and F2.

The function F1 is used only for models in Class 1 category, where all the data required to run a model are supplied locally. Depending on the model selected, the function prompts the user for all the relevant details.

The function F2 is used in cases where the data supplied from the user are insufficient to run a given model. In such cases, the system queries about the additional parameters that need to be obtained from a data base and performs the linkage. The querying for additional parameters could be done in a similar fashion to that of function F1.

The function F3 is used to provide on-line editing of data retrieved from a data base. This is useful in cases where the data retrieved from a data base need to be modified before using them to run with any model.

The function F4 is used during the model selection process. Given the area of interest of the user, i.e., the modeling needs of the user, the system browses through the dictionary to list all the models that can be used.

The function F5 is used to execute any model with the data collected using any of the data collection functions.

The data collection functions are stored in the dictionary along with the model data. When a particular model is selected, the functions that accompany the model are used by the system in querying the user for data. The system has some of the other common commands like SAVE, to store retrieved data. A DISPLAY command is used for the display of either data or model. A PRINT command is used for generating hard copy.

D. DISCUSSION OF THE VARIOUS APPROACHES

The proposed system will be easy to use both for novices and experienced users. A novice could browse through the dictionary to select the model. Depending on the modeling needs of the user, the data dictionary search will yield all the models that will satisfy the user's needs. After the model is selected, the site at which the model is located is displayed. At this point, the input and output parameters of the model are displayed, indicating to the user what data are required to run the model and what output values to expect.

Once the model has been selected, the system prompts the user for input. The input could be in different forms, depending on the model selected, as mentioned in Subsection B. If the data were to be provided locally, i.e., Class 1 models, the system would prompt the user to provide the data. The user could provide the data in a file name, which can be used by the system at run time against the model. In other cases, where data have to be retrieved from a data base, the user specifies the data in a standard query format. The system translates this query into subqueries to retrieve the data from the data base. If the data base is at a remote location, the data are shipped on the network. If the model chosen by the user is also at a remote site, the

data are shipped to the site of the model. The model and data are then run at the model site. The results are subsequently transferred to the user's site in a prescribed format. If the data base is not on-line, then a batch job is submitted at the site of the data base and the data are shipped to the site of the model after execution.

A further desirable feature of this proposed approach for the Air Force system is that implementation of various models and data bases into the system may be done in a modular sequence. That is, given the overall architecture of the system, it may be implemented initially with only a limited number of models and data bases deemed most necessary by the Air Force. As new capabilities are added to the system, only the initial user interface (Figure 3) and the data dictionary need be updated. The various system commands (F-functions of the previous subsection) can remain the same.

E. EXAMPLES OF TARGET SYSTEM MODELS AND DATA BASES

1. Class 1: PTMAX Model

The PTMAX (point source - maximum concentration) model is part of the UNAMAP collection of air quality models discussed in Section II. The model is very simple and well documented (Reference 18). The objective of the program is to analyze maximum ground-level concentrations of a pollutant downwind of an elevated point source. The concentrations may change because of variations in wind speed and stability, and a tabular output is provided for all combinations of speed and stability selected by the user. Input consists of the following information:

- a. Title
- b. Ambient air temperature
- c. Stability class
- d. Physical stack height
- e. Stack gas temperature
- f. Volume flow rate from stack
- g. Source strength
- h. Stack gas velocity
- i. Stack diameter

Computations are then performed for a range of wind speeds. Note that only 9 data entries are required, and all must be specified by the user. No calls to other data bases are necessary, and facility F1 discussed earlier may be designed to permit interactive data entry. (In fact, an interactive version of PTMAX already is available for certain computers from EPA.) Facilities F3, F4 and F5 may also be used with this model, as with all models and data bases.

2. Class 2A: CDM Model and SAROAD Data Base

The Climatological Dispersion Model (CDM) was discussed in detail in Section II and its potential for interactive use was mentioned; again, an interactive version is available for some computers from EPA. The CDM model requires input of the 576-element frequency array discussed earlier in Subsection B. Thus, CDM would utilize facility F2 to obtain the array, possibly from a data base containing such arrays constructed especially for the Air Force, as discussed earlier. The combined set of input data may be edited using facility F3 and executed using facility F5.

One provision of CDM calls for calibration of predicted and measured pollutant concentrations. The latter must be obtained from a data base; the EPA SAROAD facility, discussed briefly in Section II, is a likely candidate. Both interactive and batch access are possible. Since the Air Force system will use existing software wherever possible to facilitate ease of use, a detailed example of interactive access to SAROAD is presented in Appendix B. This also illustrates a sample for the type of menu-driven query system possible on the Air Force system.

3. Class 2B: HISARS Data Base

Virtually every hydrologic model requires input of rainfall data, of some kind, in addition to many other parameters. Sources of such data were discussed in Section II, e.g., NOAA hourly precipitation on magnetic tapes. A convenient source of rainfall and several other kinds of data that is available at some locations is the HISARS system (for Hydrologic Information Storage and Retrieval System). This data base may be queried about its content, followed by retrieval of desired information. In contrast to the SAROAD system, this data base may only be accessed in a batch mode. A detailed example of use of HISARS is presented in Appendix C. It is proposed that the relevant batch commands would be automatically generated using facility F2 in response to menu-driven user queries.

F. SUMMARY

Requirements for the Air Force Environmental Information Network include the following:

1. Menu-driven user interface
2. Uniform query language
3. Interface support to handle heterogenous data bases
4. Data formatting and editing capabilities
5. Data dictionary browsing capability

None of the data base management systems reviewed in Section III satisfy all the needs of the Air Force system. Rather, it is proposed that the system draw upon the best features of them all, most notably CSIN.

Models may be classified as to whether all input data are supplied directly by the user (Class 1) or both the user and from a separate, distinct data base (Class 2). The latter are further subdivided in terms of whether the target data base may be accessed interactively (Class 2A) or by a batch job (Class 2B). It is proposed that the Air Force system include interactive, menu-driven input of user-supplied data for all classes. This may be prepared in a modular fashion for each model and data base as they are included in the system. For interactive model and data bases that already exist, the existing software will be utilized, rather than engaging in an extensive redevelopment effort. In the case of batch-job access, new interactive software should be developed to generate the appropriate commands to retrieve data. Examples of models and data bases of the three classes are presented.

The system will contain facilities for accessing various models and data bases, editing of data, linking of models and data bases, and execution of programs. These include various language translation, formatting and file manipulation facilities. Updates and additions to the system may easily be handled through updates and additions to the data dictionary, described in Section V.

SECTION V

GENERAL DESIGN CONSIDERATIONS

A. STRUCTURE OF THE DIRECTORY

1. Introduction

The directory or data dictionary provides information about the data (i.e., their storage and nature) as well as information about the models. The models may be scattered at different locations or may be stored at one location. The directory should contain enough information to guide the user to the site at which the model is stored. It should also select a model for the user by evaluating certain characteristics particular to a given model.

In the extended case, the directory should be able to initiate the retrieval of the model after it has been selected. The other class of functions that the directory performs, is to locate the data required for a particular model and to guide the user to retrieve the data. Many models do not require data from a data base; all data are input to the model by the user (Class 1). In such cases, only the model identification and selection is of primary importance.

The actual reformatting of the data before running on a model could be done by a built-in program that automatically reads in the data and transfers them in a reformatted form suitable to run as input data for a given model.

The directory is in the form of a combination of tables and link structures. This data structure can be scanned to obtain information about the model and data.

2. Information About the Model

There are two approaches to this problem. The first uses only relations; each relation has a particular property against the model. This presentation is in a form that is easily readable; however, the system should have enough capability to conduct a relational search technique. For this, some type of relational language should be developed, to which the users' responses can be mapped. This involves an extra step. The second approach is to combine the features of relational tables with those of a link structure, thereby reducing the bookkeeping efforts of the system.

The proposed format of the directory to aid in model selection is shown in Figure 4. This structure consists of an index table that holds all the important classes of properties under which the models are grouped. Each entry of the index table then points to a list of features under the general properties of the other table. The list varies in length, depending upon the number of features that exist. For example, RESIDUAL (i.e., pollutant) may have up to 14 features under it.

Each of these features points to the next feature under the same class of properties. Once a property and a feature are identified by the user, the list of models corresponding to this particular feature is extracted. This process continues until all the relevant properties desired by the user are

(Properties) PROBLEM CLASS	(features) PROPERTIES OF CLASS	MODEL NAME
RESIDUAL	SO ₂ Toxic metals CO . . Fugitive dust	SR ₁ , atmospheric transport model EPA HIWAY, PAL PTMAX, RAM
SOURCE	Single point source . . Elevated discharge
RECEPTOR	Point receptor
TERRAIN	Forest
TIME	Steady-state
SCALE	Up to a few km.

Figure 4. Structure of the Directory

exhausted. The final model will be the intersection of all the lists of models retrieved.

After identifying the model to be used by the end-user, the system should retrieve information regarding the model. This is stored in the form of a relation where the key attribute (indicated with asterisk below) is the model name. The relation is given below:

MODEL PROPERTIES

NATURE (e.g., air, water ...)

*NAME (e.g., CDM ...)

ID

FLAGS (protection, integrity ...)

SIZE (in blocks)

LOCATION

NODE NUMBER

PARAMETERS REQD./DATA BASE REQD./BOTH

STORAGE TECHNIQUE

In case the model requires a data base to supply data as input, an appropriate table containing the nature and location of data should be supplied. This will be dealt with more carefully while discussing the data relation.

Although the properties shown in Figure 4 are specific to air-models, water models can be dealt with in a similar manner and the table can then be integrated. The level of user interaction is not considered here; it could be wholly automated by mapping the model name to the relation shown above and extracting the model features, or the user could manually enter the model selected and retrieve the information about the model.

The above relation is available for all models. Thus, its only function is to output all model characteristics, which are the different fields of a given tuple (row) once the model name has been determined.

3. Information About the Storage and Use of Data

Model search and selection have been discussed; data selection is more restricted than model selection because of three considerations:

- a. Some models do not require data from data bases.
- b. All models do not require data in the same form (with the exception of some air quality models).
- c. Some models may require data from data bases in some cases.

All these problems have to be addressed if data retrieval is expected to be automated. In the manual mode, however, the user has all the information required for data selection and instructs the system what to retrieve. In any case, the user must specify certain search parameters to retrieve the data.

The following relation gives the information required to locate and retrieve the data base:

DATA LOCATION

MODEL NAME (that requires the data) - optional
LOCATION
ID#
TYPE OF DATA (rain, flow, etc. . . .)
PROTECTION FLAGS

Once the location of the data is determined, the proper data selection is done by including more attributes in the above relation, e.g.,

PERIOD OF RECORD (e.g., 1/48 to 12/79)
TOTAL NUMBER OF MONTHLY VALUES

These two attributes could be appended to the table to supply adequate information about the data retrieval, since these values vary from query to query. By default, all values - or selected values - of the data base could be output to the user.

A stored relation may be necessary, depending on the type of implementation. This relation becomes useful if the data retrieved from the data base should be automatically input to the model. In this case, the nature of the data retrieved can be compared with the input data requirements of the receiving model. If the two formats (requested and retrieval) match in every aspect, then the transaction can proceed without any formatting; in the other case, the system will point out differences to the user so that he could make the necessary reformatting. Yet another approach is the building of a system routine that automatically does the reformatting without the user's intervention. Here the two data formats are taken as input along with the data, and the output is the reformatted file.

The nature of the following relation will contain mostly internal (system) information about the data:

DATA NATURE

NUMBER OF TUPLES - optional
NUMBER OF FIELDS
WIDTH OF FIELDS (offset/varying)
MAXIMUM AND MINIMUM VALUES
REAL/INTEGER/FLOATING
SIGNIFICANT PLACES

This relation will be constructed at the same time the data are being retrieved and reflects the nature of the data in the data file.

4. General Directories

Many system directories do not directly affect the user. These relations are essential for a network system that is interconnecting different users and accessing some common data base. Three are shown below:

USER INFORMATION

USER ID
PASSWORD
USER CLASS
JOB TITLE
NODE NAME, #

This relation holds general user information necessary for a multi-user system, including information required for validation of user identity.

ACCESS

NODE NAME, #
LOCATION
LOG ON

This, together with the previous relation can automatically log-on the user to a particular system.

NODE

NAME, #
LIST OF MODELS/DATA
STATUS (working, down)

In a multinode system, this relation keeps updated information about the status of a particular node. Every request first checks this relation to determine if the node is active, and then verifies if the model and data still reside at the node before sending off a request to it. It is important that this relation be kept up to date via some scheme (e.g., broadcasting).

Several factors affect the directory design:

- a. The updating frequency for models is very low, and data bases are updated principally by adding more data.
- b. The migration of model/data to a different node is rare.
- c. Factors a. and b. favor a replicated directory scheme at each node.
- d. Only a few directories (e.g., "NODE") need regular updating and this may be done by broadcasting.
- e. One centralized directory may hold information about the location of the replicated directories. It may also examine the status of the nodes periodically and post this information to all nodes in the network.
- f. Query variation is minimal, so the relations about models/data can be joined into one or two large tables (as has been suggested here) without serious performance degradation.

B. DATA FORMATTING

When programs are run on computers, data should be supplied in a particular order as needed by the program. Obviously, running an environmental model is like running a program. Hence, when running the models with data bases, the data should be rearranged in the form needed by the model. For example, if a model needs rainfall and snowfall data, the order of the snowfall or rainfall data to be supplied at run time is predetermined by the model. While all models read data in some predetermined order, some models may not accept free format data even though they may be in the right order, i.e., the data must be supplied in a particular format. In general, the reorganization of data, to make them suitable for a model, is called data formatting.

Data formatting is model-dependent. Two similar models could use the same data base, but may need the data in different formats. To perform any kind of automatic formatting, the format of data needed for every model should be maintained by the system. The format information is stored in the directory along with the model information. This information could be stored as:

1. the data elements needed by the model along with the order of intent, or
2. format information about each of the above data elements.

When a model is to be run, the system should perform both ordering and formatting of data. For Class 1 models (see Section IV), for which the user supplies all the data, the system prompts the user for data in the order required by the model. No further ordering of data is required.

To facilitate formatting of data, a reformatter routine is used. The ordered data along with the format declaration are passed to the reformatter. The output of this routine will be ready-to-use data for the model. A single formatting routine could be used for the models.

For other models, namely Class 2 models, for which data have to be retrieved from a data base, temporary files are used. The storage mode of data in temporary files is dependent on the source language used in creating the files. For example, a matrix in FORTRAN is stored in column major order, while a matrix created in PASCAL could be stored in a row major order. The system needs to be told about the order in which the data are arriving and the order in which a model expects them. Knowing these two, it should be able to perform the actual conversion. The data are then ordered and formatted using the information from the directory. The temporary file is then used at run time.

C. MODEL - DATA BASE TRANSFERS

A user on the Environmental Information Network has access to models and data bases present at all the nodes in the network. Therefore, a user could request a transaction that involves a model and data base residing at different remote sites. For such cases, either the model or the relevant data has to be shipped to the user's site.

Three strategies can be used to satisfy a user's request. (1) The model and data could be shipped to the user's site or an intermediate node. (2) The data could be shipped to the site of the model. (3) The model could be shipped to the site of the data.

The choice of approach depends on many factors, such as the volume of data to be shipped, communication costs, response time to satisfy the request, etc. Only relevant data should be shipped, i.e., the data should be selected from the data base before being sent on the network. When the model is shipped, many of the utilities, like the data manipulator, will have to be shipped to ensure that data are supplied in a correct format for the model.

To prevent excessive transfers on the network, a copy of the frequently used models could be stored at each of the sites. Since it is not feasible to replicate the data bases at each node, most of the transactions on the network would involve data transfers.

D. IMPLEMENTATION ISSUES

Two prototypes, MULTIBASE and CSIN, have accessed distributed heterogeneous distributed data bases. Neither of these is commercially available. MULTIBASE prototype implementation is scheduled to be completed by mid-1984. CSIN has been partially implemented, but no definitive plans are available regarding full implementation. The technology of distributed data bases is still in the experimentation stage. Nevertheless, some of the features of MULTIBASE and CSIN could be adapted to the Environmental Information Network.

To provide a uniform query language across the network involves selection of a suitable data model. Since environmental data are typically tabular in structure, the relational model appears to be a good choice. Implementation of the EIN would involve design and development of three components:

1. The host interface provides an interface between the local data base and the data manager. Every data base in the network has its own query/data manipulation language, and a schema in some data model. But all queries on the network are expressed in a standard query format. Therefore, the host interface needs to transform the query into the host data manipulation language. Since the host interface performs the query mapping, every type of data base that is connected on the network needs a corresponding host interface. The interface between the data manager and the host interface could be standardized. This would make the network easily expandable. Every time a new site is added in the system, only a new host interface would need to be designed.

2. The data manager is responsible for query processing. Since very few queries in the system would involve data from more than one data base and since the user is required to confirm the target data base(s) for running a model, the data manager would send the request to the relevant host interface and ask the data manipulator to format the retrieved data.

3. The implementation of the data dictionary has been described earlier in this section.

It is difficult to propose a suitable high-level language in developing the EIN. Typically, more than one language would be used. Specific languages may be used to implement specific functions of the network, if that language is ideally suited to perform that function. For example, CSIN uses LISP language to implement 'scripts' which would be very hard to implement in any other language. The language selection issue will be largely the implementor's choice.

At this time, it is difficult to estimate the cost and time needed to develop the EIN. The cost issue involves many factors, such as the number of data bases to be included in the network, global processing needs, etc. The cost of developing a system like CSIN (about \$4.5 million) is a good indication of the complexity of the task. It is even harder to estimate the time required to develop a system like EIN. The system could be built in an incremental way, integrating a site at a time. In this way, a running system linking a few sites can be operational in three to four years.

SECTION VI

SUMMARY AND RECOMMENDATIONS

A. SUMMARY

1. Project Scope

The Air Force is increasingly engaged in analysis of environmental problems of many kinds, for which it is necessary to access relevant environmental computer models and data. This report analyzes the Air Force's general requirements and provides architectural guidelines for the design of an information network. This software will enable the Air Force to manage environmental models and data easily, and accurately. Along with models, data, computer hardware, system software and computing network, it will constitute the basis of the Air Force Environmental Model and Data Exchange (AFEMDEX).

Operational environmental models and data bases have been reviewed for applicability for inclusion in the system, with special emphasis upon those related to air quality. Methods for model selection have been considered in detail. Existing distributed data base management systems (DDBMSs) have been similarly reviewed to determine the suitability of some of their features to the AFEMDEX. Architectural guidelines and a structure for the overall environmental information network have been proposed by incorporating some concepts present in these systems; however, the overall approach is kept as simple as possible.

2. Environmental Models and Data Bases

Environmental models may be classified according to many schemes (e.g., deterministic vs. stochastic, temporal and spatial properties, etc.), and the schemes may vary, depending on whether air quality, water quality or another environmental parameter is under consideration. However, some existing generalized schemes may be applied to the development of user-friendly, interactive software to aid in model selection. In fact, some interactive versions of air quality models (e.g., CDM, PTMAX) and data bases (e.g., SAROAD) already exist. It should be relatively easy to integrate these into the Air Force system.

Although there are many air and water quality data bases, their data are seldom required directly for model input. Their main uses are to provide background data for statistical analysis, and/or model calibration and verification. These are very important tasks, however, and are likely to represent the principal uses of the system.

Models may be classified as to whether all input data are supplied directly by the user (Class 1) or by both the user and a separate, distinct data base (Class 2). The latter may be further subdivided in terms of whether the target data base may be accessed interactively (Class 2A) or by a batch job (Class 2B). It is proposed that the Air Force system include interactive, menu-driven input of user-supplied data for all classes. This may be prepared in a modular fashion for all models and data bases as they are included in the system. For existing interactive models and data bases, the

existing software will be used, rather than engaging in an extensive redevelopment effort. New interactive software should be developed for batch-job access to generate the appropriate commands to retrieve data. Examples of models and data bases of the three classes have been presented in this report.

3. Distributed Data Base Management Systems

Five DDBMSs were reviewed for their applicability to AFEMDEX needs: SDD-1, R*, MULTIBASE, Distributed INGRES, and CSIN. Air Force requirements include:

- a. Menu-driven user interface
- b. Uniform query language
- c. Interface support to handle heterogenous data bases
- d. Data formatting and editing capabilities
- e. Data dictionary browsing capability

The system will contain facilities for accessing various models and data bases, editing of data, linking of models and data bases, and execution of programs. These include various language translations, formatting and file manipulation facilities. Updates and additions to the system must be easily handled through updates and additions to the data dictionary.

None of the data base management systems reviewed completely met all the needs of the Air Force system. Rather, it is proposed that the system draw upon the best features of them all, most notably CSIN. Of the systems surveyed, CSIN has the closest environment and properties to that of AFEMDEX. However, there are many differences in the nature of data handled. Both CSIN and AFEMDEX transform data between heterogeneous systems. In CSIN, it is done in the transaction and data layers by using the component data interface. These interfaces are software units and make up the bulk of the data layer architecture. These units are tailored to a particular component and hence are not portable. AFEMDEX would have to be concerned about development of software packages that would interface these various heterogeneous systems onto the network. Therefore, although the CSIN approach may roughly coincide with that of AFEMDEX, it is not possible to directly transfer the CSIN system for AFEMDEX use.

B. RECOMMENDATIONS FOR THE ENVIRONMENTAL INFORMATION NETWORK

In this report, an overall architecture has been proposed for the Environmental Information Network (EIN) (see Figure 3). It is a multilayer architecture wherein a variety of software layers will be interposed between the user and the data bases. They will serve the following functions:

1. Model inquiry and selection of an appropriate model by supplying requirements.
2. Selection of appropriate data base(s), if necessary, to supply data to the model.
3. Facilities for query formulation, parameter gathering, text editing, and menu-driven collection of external data.

4. Data transformation between the local formats of data and the format desired by the user.
5. Transformation of user requests from a given standard format into a specific query language at the target site.
6. A data dictionary/directory to keep all relevant information about models and data bases.

After a certain set of data bases is selected, along with an appropriate set of environmental models to be run using those data, the following sequence of actions will be necessary:

1. Define a common data model which is adequate to describe all data bases. In the authors' opinion, the relational model should be investigated first.
2. Define a standard query language which is powerful enough to express the selections/manipulation of needed data. The SQL language currently implemented on most relational systems should be investigated.
3. Define the "host interface" for every different data base system to be incorporated into the EIN.
4. Populate the dictionary with appropriate information about the models and the data bases. Currently, several data dictionary packages are commercially available; however, considering the limited needs of the proposed system, a specific dictionary should be designed, in the format indicated in Section V-A.

If the proposed EIN system is to be implemented by the Air Force, the next major task would be a detailed functional design of the system, particularly the data manager, data reformatter, data dictionary and the host interfaces. Additionally, basic research on the design of a standard query language, mappings among this language, and various target system languages will have to be carried out.

When the EIN is established, desired environmental models and data bases can be included in a modular fashion as needed. The Air Force may also benefit from establishing their own specific data bases for frequently used environmental data, such as those used in several air quality models.

REFERENCES

1. McKenzie, D., Milask, L. and Long, R., Feasibility Study for an Air Force Environmental Model and Data Exchange, Vols. I-IV, General Software Corporation, Report ESL-TR-82-13, U.S. Air Force, Tyndall AFB, FL, February 1983.
2. Basta, D.J. and Bower, B.T., eds., Analyzing Natural Systems: Analysis for Regional Residuals-Environmental Quality Management, Resources for the Future, Johns Hopkins University Press, Baltimore, (EPA-600/3-83-046, NTIS PB83-223321), June 1982.
3. Basta, D.J. and Moreau, D.H., "Introduction to Analyzing Natural Systems," in Basta, D.J. and Bower, B.T., eds., Analyzing Natural Systems: Analysis for Regional Residuals-Environmental Quality Management, Resources for the Future, Johns Hopkins University Press, Baltimore, (EPA-600/3-83-046, NTIS PB83-223321), June 1982.
4. U.S. Environmental Protection Agency, Environmental Modeling Catalogue: Abstracts of Environmental Models, EPA Information Clearinghouse (PM-211A), Washington, DC, August 1982.
5. Muschett, D.F., "Analyzing Atmospheric Systems," in Basta, D.J. and Bower, B.T., eds., Analyzing Natural Systems: Analysis for Regional Residuals-Environmental Quality Management, Resources for the Future, Johns Hopkins University Press, Baltimore, (EPA-600/3-83-046, NTIS PB83-223321), June 1982.
6. U.S. Environmental Protection Agency, Guideline on Air Quality Models, EPA-450/2-78-027, Research Triangle Park, NC, April 1978.
7. Turner, B.D., "Atmospheric Dispersion Modeling: A Critical Review," Journal of the Air Pollution Control Association, Vol. 29, No. 5, pp. 502-519, May 1979.
8. U.S. Environmental Protection Agency, AEROS Manual Series: Summary and Retrieval, Vol. III, EPA-450/2-76-009b, Research Triangle Park, NC, July 1981.
9. Busse, A.D. and Zimmerman, J.R., Users Guide for the Climatological Dispersion Model, EPA-R4-73-024 (NTIS PB 227346/AS), Environmental Protection Agency, Research Triangle Park, NC, December 1973.
10. Huber, W.C. and Heaney, J.P., "Analyzing Residuals Generation and Discharge from Urban and Nonurban Land Surfaces," in Basta, D.J. and Bower, B.T., eds., Analyzing Natural Systems: Analysis for Regional Residuals-Environmental Quality Management, Resources for the Future, Johns Hopkins University Press, Baltimore, (EPA-600/3-83-046, NTIS PB83-223321), June 1982.
11. Viessman, W., Jr., Knapp, J.W., Lewis, G.L. and Harbaugh, T.E., Introduction to Hydrology, 2nd Edition, Intext Education Publishers, New York, 1977.

12. Hinson, M.O. and Basta, D.J., "Analyzing Surface Receiving Water Bodies," in Basta, D.J. and Bower, B.T., eds., Analyzing Natural Systems: Analysis for Regional Residuals-Environmental Quality Management, Resources for the Future, Johns Hopkins University Press, Baltimore, (EPA-600/3-83-046, NTIS PB83-223321), June 1982.
13. Thomann, R.V., "Verification of Water Quality Models," ASCE Journal of the Environmental Engineering Division, Vol. 108, No. EE5, pp. 923-940, October 1982.
14. Goodman, N., Merrill, T. and Rothnie, J.B., "Database Management in Distributed Networks" in Protocols and Techniques for Data Communication Networks, Kuo, F.F. (ed.), Prentice Hall, New York, 1981.
15. Date, C.J., An Introduction to Database Systems, Vol. 2, Addison Wesley Publishing Co., Reading, MA, 1983.
16. Adiba, M., Chupin, J.C., Demolombe, R., Gardarin, G. and Le Bihan, J., "Issues in Distributed Database Management Systems: A Technical Overview," Database Engineering, Vol. 5, No. 4, December 1982.
17. Ceri, S. and Pelagatti, G., Distributed Databases: Principles and Systems, McGraw Hill, New York, 1984 (forthcoming).
18. Turner, D.B., User's Guide to PXXXX Air Quality Models: PTMAX, PTDIS, PTMTP, Environmental Protection Agency, Air Pollution Training Institute, Research Triangle Park, NC, (undated).

APPENDIX A

DISTRIBUTED DATA BASE MANAGEMENT SYSTEMS

A. INTRODUCTION

This appendix presents a survey of five DDBMSs presented in the report. The systems are MULTIBASE, R*, SDD-1, distributed INGRES and CSIN. This discussion covers, in detail, the architectures of these systems and the processing of queries by the individual systems in a distributed environment.

B. THE DISTRIBUTED DATA BASE SYSTEMS

1. The MULTIBASE System

a. Background

MULTIBASE is a DDBMS that provides the users integrated access to preexisting, heterogeneous distributed data bases (References A-1, A-2). It is also a software system that allows end users to query the data base in a common query language. Since all the component data bases are heterogeneous, the main goal of MULTIBASE is to provide a fast, easy and integrated access/retrieval system for the various heterogeneous data bases without changing the local data base systems or their application programs.

MULTIBASE is a read-only data base, and updates can be done only where the data object is stored. Therefore, each local site maintains autonomy for updates. Local application programs can operate using the existing local interfaces. The language provided to global users by MULTIBASE is called DAPLEX (Reference A-3), which is a data definition and manipulation language for data base systems. The model used is the functional data model (Reference A-3).

b. Architecture of MULTIBASE

MULTIBASE is particularly interesting since it deals with the problem of integrating heterogeneous data bases. The schema architecture and the component architecture are the most important elements to understand in the overall architecture of MULTIBASE.

(1) Schema Architecture

The task of providing the users with uniform data base management involves not only a homogenization of the various heterogeneous data bases, but also a resolution of data incompatibilities to produce integration.

There are three levels of schemata in MULTIBASE: a global schema (GS) at the top, an integration schema (IS), one local schema (LS) per local data base at the middle level, and one local host schema (LHS) per local data base at the bottom level (see Figure A-1). The LHSs are the preexisting local schemas which may be defined by a variety of data models. The LHS is mapped into the LS which is defined by the functional data model. Therefore, the LSs are expressed in only one common data model. The IS together with the LSs provide the integration of the various data bases and resolves data

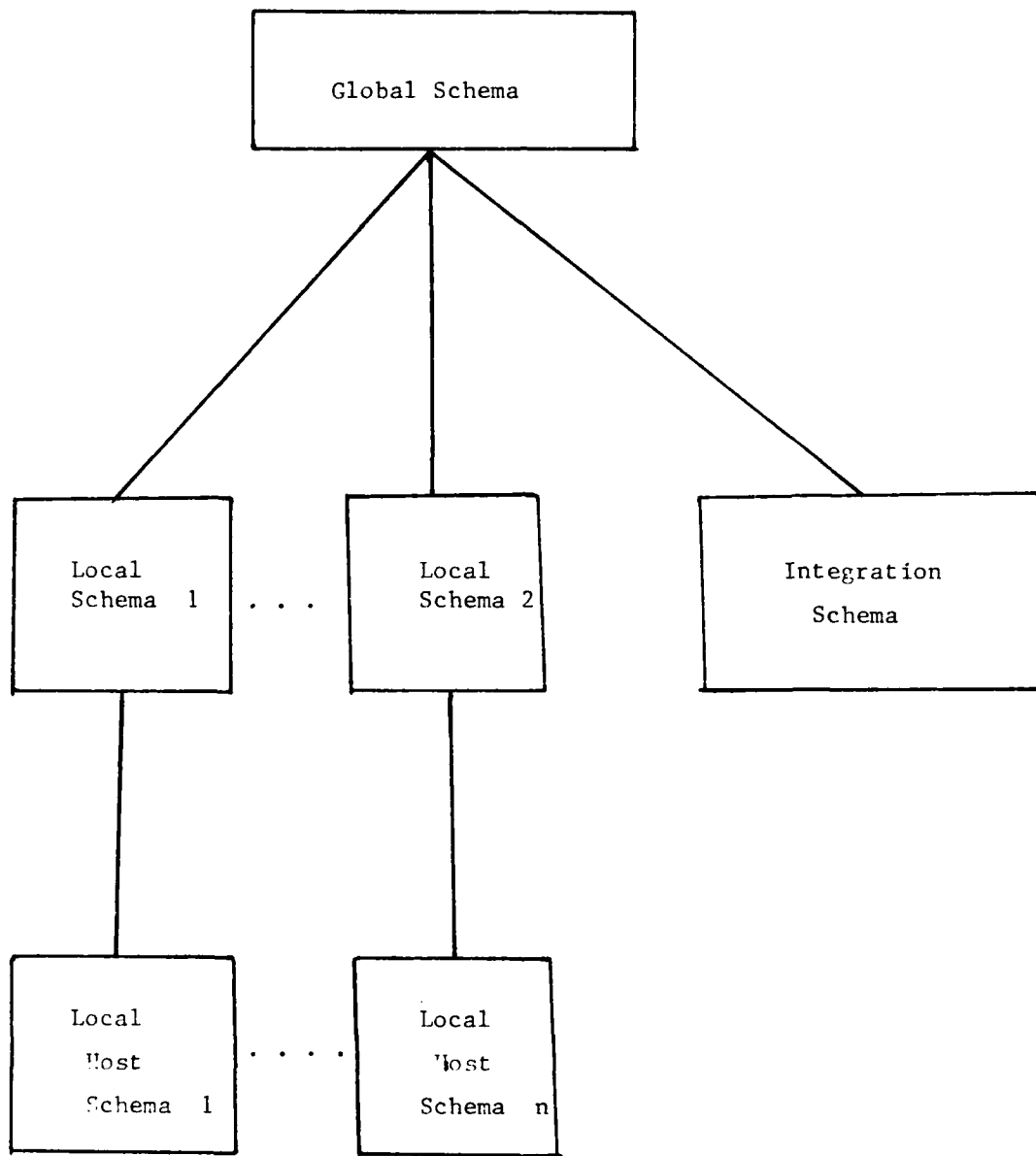


Figure A-1. Schema Architecture

incompatibilities (scale, name, etc.) and maps them to the GS. The whole schema architecture can be visualized as a LHS-to-LS mapping that provides homogeneity and a LS-and-IS-to-GS mapping that provides integration. All the schemas except the LHSs are in the DAPLEX form.

(2) Component Architecture

The MULTIBASE system has two types of components: a global data manager (GDM) and a local data base interface (LDI). The GDM handles all global aspects of a query. It transforms the DAPLEX global query into a set of DAPLEX single-site queries, each accessing only a single site. The LDI receives this (single-site) query and transforms it into a local language query. This query is executed against the DBMS and the results are reformatted by the LDI into a standard format for the GDM. The GDM recombines data from different LDIs and presents the user with a consistent result.

Since future expansion of the network is desired, the interface between GDM and any LDI is standardized so that when a new DBMS is added to the MULTIBASE configuration, only a new interface between the host system and the LDI needs to be designed.

c. Query Processing

Query processing will be discussed in terms of the GDM and the LDI. The GDM has the following logical components that process the query: transformer, optimizer, decomposer, filter, monitor, internal DBMS, and translator.

The transformer converts a global DAPLEX query that references the global schema into a set of queries which reference the local schema and the auxiliary data base schema, whose function will be described later. The transformer consists of techniques for resolving incompatibilities among local data bases through renaming, restricting, logically restructuring and generalizing the objects from the underlying data bases. Once the view for the local data base is formed, the user can query any of the data objects without regard to the underlying data structures or locations.

The "query execution strategy" is globally optimized by the optimizer. The optimizer's main function is to reduce the response time (Reference A-4). This strategy involves the combination of local processing and data movements. It has roughly the same function as the access planner in the R* and SDD-1 systems, to be described later.

The decomposer, as the name implies, breaks down the global DAPLEX query into single site queries, and the filter reduces the decomposed queries by removing from them operations that are not supported by the local DBMS. A query must be fully processed at the site to which it is sent, but certain operations like arithmetic operators may not be supported at that site. In such a case, the filter gets all the necessary additional information needed from the local system. The original query is modified, and this information is passed to the internal DBMS (defined below) which performs the arithmetic operations. This is done to overcome some of the deficiencies of the preexisting data base systems.

The monitor carries out the overall execution of the strategy developed by the optimizer until the result is finally integrated by the GDM for the user.

The internal data base management system includes the auxiliary data base which contains necessary information to resolve incompatibilities and inconsistencies among data stored at different sites. The internal DBMS may contain (1) statistics that are used to determine which data values should be used in case of a conflict, (2) conversion tables for performing data transformation, and (3) synonyms if more than one word is used to describe a data object.

After the GDM has split the query, the local data base interface (LDI) takes over. The relationship between the GDM and LDI is shown in Figure A-2. The LDI has a module called the optimizer. This module examines the DAPLEX query and determines a processing strategy that will speed up the query execution. The translator module provides a uniform interface to all the data bases. This component can translate a DAPLEX query into two widely differing local queries depending upon the local sites involved.

The LDIs are simple processors and not general purpose DBMSs. If the LDIs are made more powerful by supplying them the functions absent in local DBMSs, then excessive data movement by the GDM can be avoided. The filter can be spared from performing all the extra data retrieval and internal query generation.

The functional model used in MULTIBASE implies the use of techniques to handle generalized objects. MULTIBASE has to interface with a wide variety of DBMSs, some with procedural, navigational data manipulation languages (e.g., IMS and CODASYL) and others with high level query languages (e.g., INGRES and SQL/DS). Global optimization must deal with an additional problem because these systems differ greatly in the level of direct control over access path selection given to programmers. This forces the LDI to give different amounts of optimization when dealing with different systems.

In summary, MULTIBASE is a read-only system. There is no provision for synchronized read capabilities across different sites. Hence, global concurrency and integrity problems are reduced to the ones faced in a centralized data base environment.

2. The R* System

a. Background

R* (Reference A-5) is a homogeneous DDBMS, in which the central design goal is to provide site autonomy. Each of the supported sites has total control of its data and can perform local data base operations without the need to consult any other site. R* uses the relational data base model and is a project developed at IBM. It is based on System-R. The SQL language used in System-R is extended to deal with the various distributed features encountered in R*. The communication is via CICS, an IBM software product. CICS delivers the correct message without duplications or replications; however, there is no guarantee that the messages will ever be delivered.

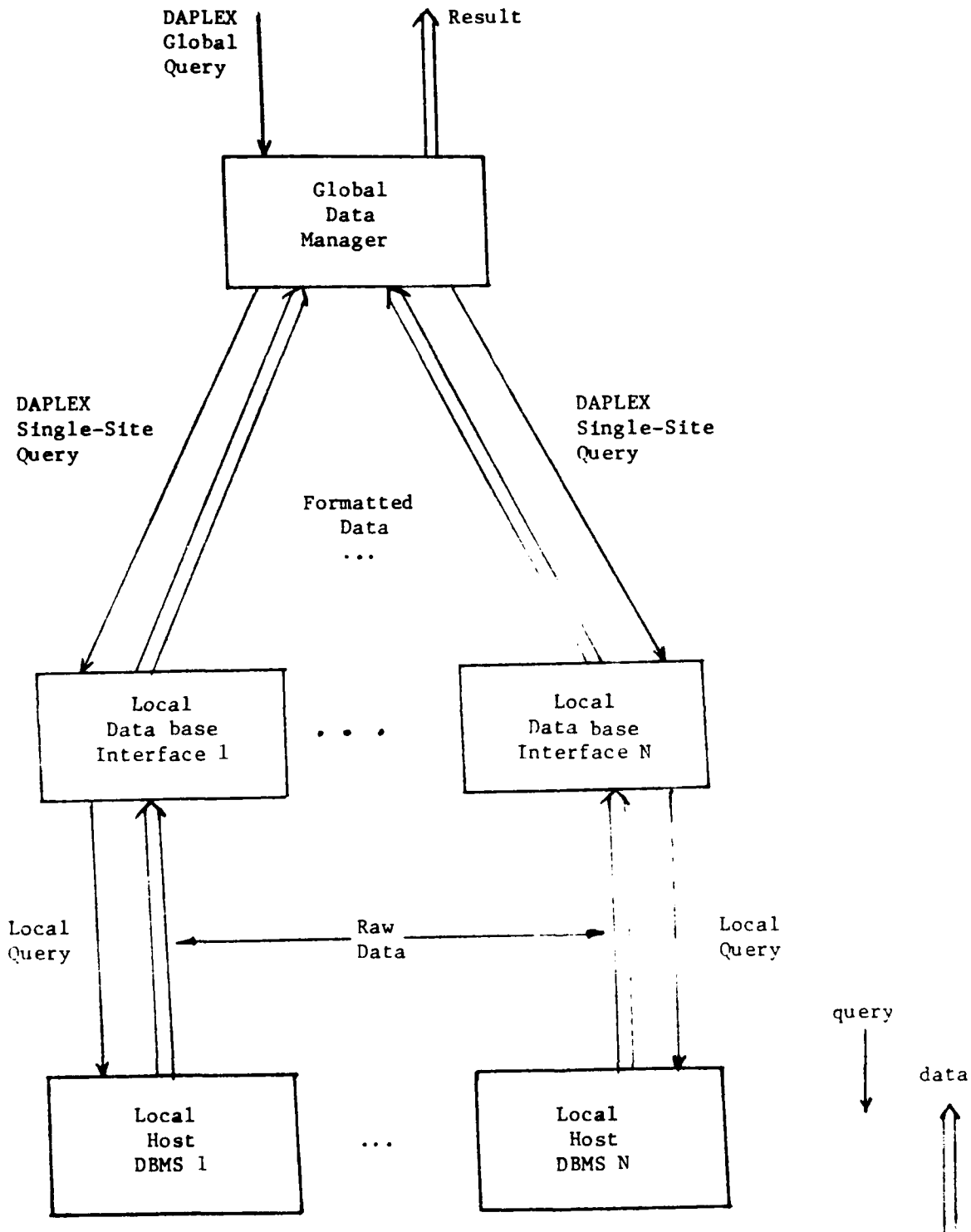


Figure A-2. MULTIBASE Component Architecture.

Data are stored in tables and may be dispersed, replicated and partitioned horizontally or vertically. The horizontal partition could be based on some separation criteria. In vertical partitioning, the different partitions should have a few common columns to enable a one-to-one match of the fragmented records during reconstruction. All the accesses and reconstruction of tables are transparent to the end user. Another feature in R* is the concept of snapshots, which is a copy of relations where the data are consistent but not up to date. It provides a read-only facility and offers a static view of the required data base, saving on retrieval costs if these snapshots are used instead of the current relations.

b. Architecture of R*

At each site, the R* can be subdivided into three principal subsystems (Reference A-6): a transaction manager, a communication manager and a data base manager. The R* architecture is shown in Figure A-3.

The transaction manager is responsible for committing and aborting a transaction. It is also responsible for the detection and prevention of deadlocks in a distributed global environment.

The communication manager is an extension to CICS (Reference A-7), and routes incoming messages to both the transaction manager and the data base manager by establishing virtual circuits. Currently, communications are done using SNA LU6 protocol. The virtual circuit is a half-duplex circuit in which the direction of message flow is controlled by the process using the circuit. The R* also uses datagram protocol for distributed global deadlock detection and to resolve transactions that failed during the commit process. Both these functions are time-triggered, periodic activities.

The data base manager has two components, research storage system (RSS) and relational data system (RDS). The local data base manager provides access to stored relations, concurrency control, and local recovery. These local recovery techniques are the same as that of System-R from which they are derived. The upper level of the data manager is responsible for processing statements in the data definition and manipulation language and is also responsible for sending and processing the messages used to implement distributed query planning and execution.

RSS provides B-tree indexes which are stored on pages separated from those used for data storage. The leaf nodes are chained together and have key values and record identifiers which point to the data records.

c. Query Processing

Query processing in R* is greatly affected by the design goal of maintaining the maximum possible site autonomy. This led the designers to avoid any type of centralized supervision. Here, one node, namely the node at which the query originated, will assume the role as the coordinator for the query processing and is called the "master." All other nodes that take part in the processing of the query are called "apprentices" (Reference A-8).

Due to the absence of any centralized catalog system, R* had to resolve ambiguities of names by mapping every user's name to internal system

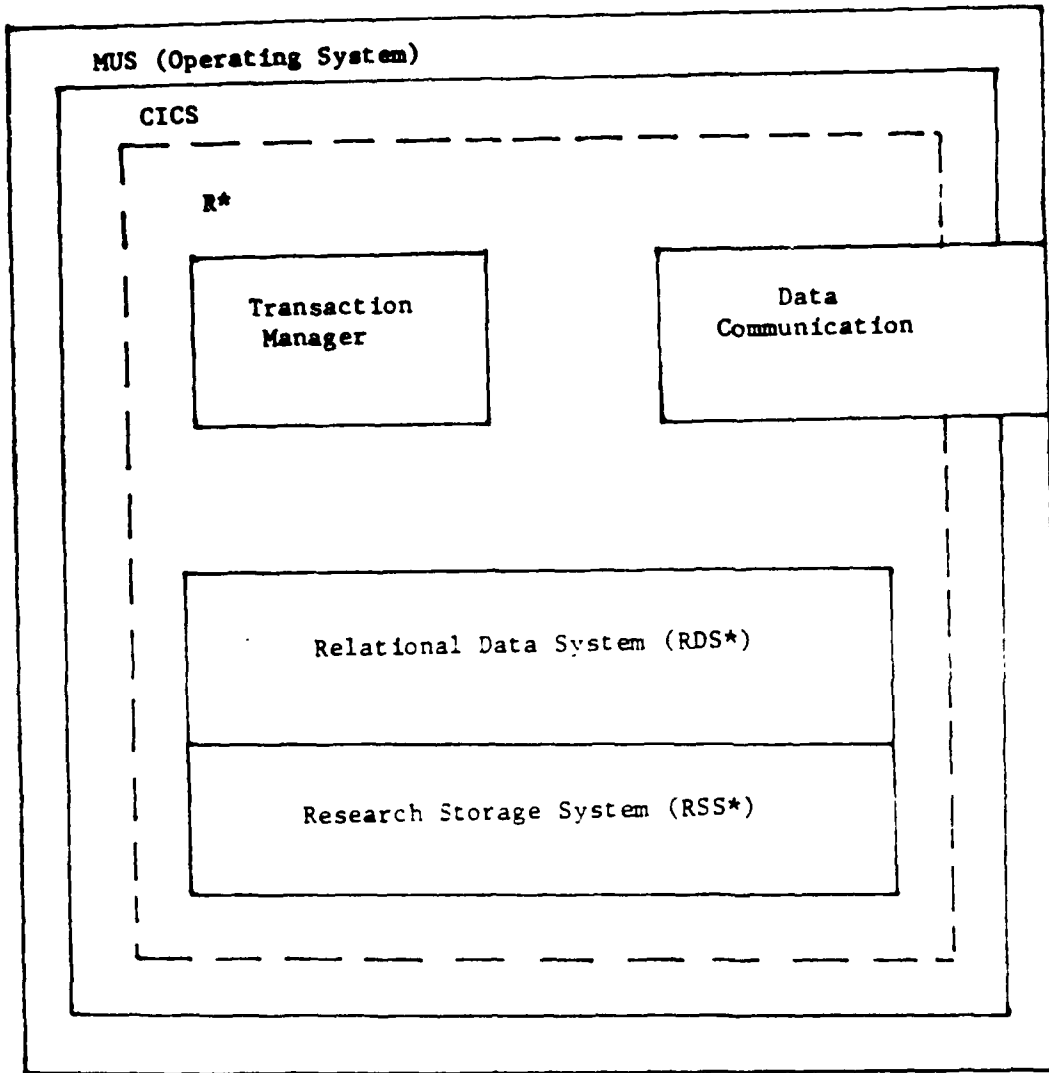


Figure A-3. R* Architecture.

wide names (SWN) that were unique to the entire system. A SWM has the form

USER-NAME @ USER-SITE • OBJECT-NAME @ BIRTH-SITE

where BIRTH-SITE is the place where the object was first created. When processing a query, the SQL statements at the master site are passed and their syntax verified; then the name referenced in the query is resolved and catalog data for each of the names are retrieved. At this stage, rights of the user to access the local data referenced by the query are checked by the RDS. Now the access planner constructs an overall query-processing strategy which is based on approximate cache information (since the cache catalog information about other sites may not be updated) obtained from the apprentices; this plan is called the 'skeleton global plan'. The plan is then broadcasted to the different sites involved, and each site will check the validity of the global plan by verifying whether or not the version number of the information used to generate the global plan is current. This verification is done by performing a catalog look-up. In case outdated information was used, the plan is reconstructed by using the new version of the data.

The R* compiles SQL data manipulation operations into machine language programs called access modules. Each of the apprentices is responsible for compiling parts of the access plan that are sent to them. Thus, portions of an access module are stored at each site involved with the execution of a query. These modules can be run repeatedly without incurring the overhead of recompiling until the module is invalidated. Invalidation can result if any of the access paths or authorizations used in creating the access module are revoked. The apprentice sites perform the same set of operations as the master site, except that access planning is limited to local sites only.

One very important feature in access planning is optimization. Here the global access path selection algorithm minimizes a cost function which includes input/output cost, CPU cost, and message cost. The evaluation is different from other distributed systems since it not only considers communication, but deals with the various CPU costs resulting from considering the number of pages fetched from storage. The CPU costs are affected by choosing the correct type of join-operator. Global access planning uses a breadth-first tree-search technique to explore the possible access plans. Different join methods, order of table accesses and access paths to each table are factors considered by the global access path selector.

d. Catalog

The catalog architecture adheres to the rule of providing site autonomy. All catalogs are stored in the form of tables, similar to the data. This facilitates the use of the same SQL statements as used on the data. Each site maintains a catalog of all data present there. Information about data frequently referenced by this site may also be stored. To aid performance, cache copies of catalog data from remote sites are also stored; however, no attempt is made to have any form of central copy of all catalog data at one site, and no synchronous broadcast is done to update the catalog replicas that are stored at each site. The cached information is not kept fully updated since this involves a large amount of code to perform this operation. Instead, a version number is tagged to each catalog entry. If a

plan is generated using an outdated version number, the cached information is updated and the plan regenerated, causing the user to wait slightly longer than usual for the result.

System-wide names take the place of a centralized catalog system. In case the data object migrates because of reorganization, the catalog maintains the name of the site to which it has migrated. This facilitates data recovery since the user has only to access the birth site of the object to retrieve it.

The greatest disadvantage in the R* catalog system is that once the birth-site of an entry is nonfunctional, the data cannot be retrieved even if they have migrated to another site. This, however, is not expected to introduce serious service degradation due to the cached information.

The catalog architecture has to be extended for distributed objects (horizontal or vertical partitioning). A distributed relation is represented by a tree of fragments. Only the leaves of the tree correspond to the stored fragments; the inner nodes represent the distribution. Replication can also be represented. Maintaining a consistent distributed object catalog requires synchronizing catalog updates for all interior nodes at all storage sites of the distributed object. Here, there is a compromise on site autonomy, and it is necessary to maintain a distributed relation. Each site which contains a partition has a copy of the distribution tree.

e. Concurrency

Concurrency control and deadlock detection is based on locking mechanisms. Deadlocks are detected by wait-for graphs. All transactions are numbered and any deadlock is broken by pulling out the youngest transaction, that is, the one with the largest number. Transactions use the two-phase commit protocol. During the generation of a plan, each site creates transaction save points. These reference points help the transaction to undo the plan up to the save point when a site rejects the plan.

f. Reliability

The R* logs all changes to data records (Reference A-9). Both the old and new values of the record are logged. When the system crashes, the shadow pages of the secondary memory can bring the system up to a consistent state.

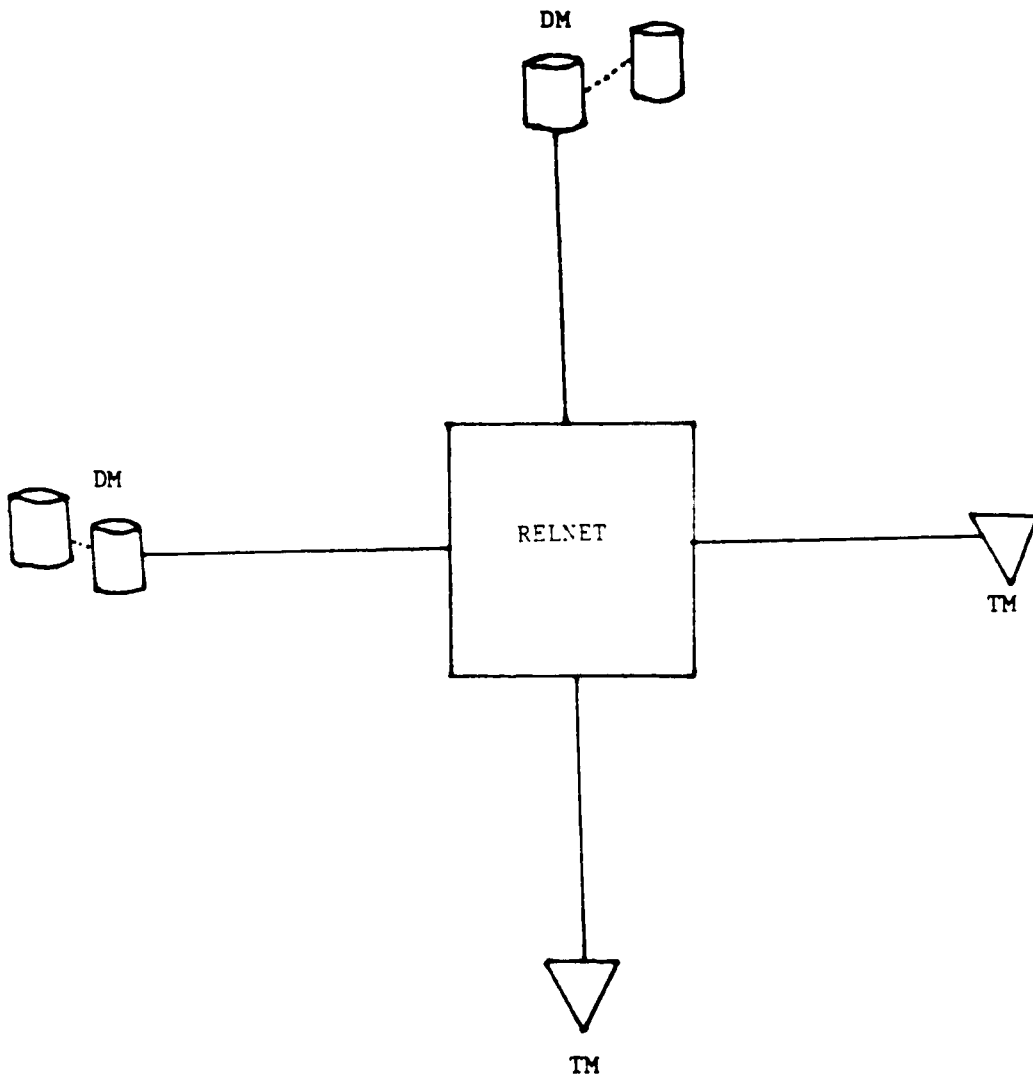
3. The SDD-1 System

a. Background

SDD-1 is a prototype distributed data base system developed by Computer Corporation of America (Reference A-10).

b. Architecture of SDD-1

Logically, SDD-1 comprises three virtual machines: transaction module (TM), data module (DM), and a reliable network (RELNET). The architecture is shown in Figure A-4.



DM - Data Module
TM - Transaction Module

Figure A-4. Architecture of SDD-1.

TM provides the execution of the user transaction and also performs access planning, fragmentation, and concurrency control. The DM stores the data and has work areas where data can be cached and worked upon if required. RELNET is the communication system for the network. RELNET allows a reliable delivery of messages and deals with cases of node failure or network partition. Its function also involves site monitoring and maintaining the system clock.

c. Query Processing

SDD-1 uses a query language called "datalanguage." While processing a query, the data are treated in the form of 'envelopes' (Reference A-11). An envelope contains relational calculus type of expressions that maps the data base to a sub-data base. There may be more than one envelope for each query. The objective is to find the best "envelope" that reduces the search. To do this, "reducers" are employed.

Reducers restrict the data by employing local operators like PROJECT, RESTRICT and SEMIJOIN. The SEMIJOIN operator is very powerful and monotonically reduces the size of the data base to work with while processing a query. In a SEMIJOIN, the projection of one table on the join columns is transferred to the other table's site, and then only records from the second table which match one of the join keys from the first table are returned and joined, either at the first table's site or at another location. Although SEMIJOIN is a powerful operator to reduce data, two scans have to be performed on one of the tables: first to form the projection and then the join. This is of no consequence if only communication costs are taken into account for the optimal access strategy, as in system INGRES and SDD-1. In R*, this would be a major disadvantage because CPU costs are also included. While performing multisite joins, there are many more messages transferred in the R* system than in SDD-1.

The SDD-1 transaction module's access planner tries to minimize the communication costs by transferring the least amount of messages over the communication network. The overall strategy is refined by using a hill-climbing technique that relies on finding the most cost-effective nonlocal SEMIJOIN and appending it to a sequential program which is made up of the local operators that reduce the data base. This technique is called the greedy algorithm since it seeks to optimize at every step without looking ahead.

d. Catalog

The SDD-1 catalog is in the form of one single logical table. The catalog may be replicated or fragmented at the different sites. To improve performance, SDD-1 also caches catalog entries. Unlike the R*, SDD-1 maintains all the cache information up to date. This automatically implies a great burden of bookkeeping and global operations to invalidate caches at many sites. In order to locate the catalog fragments, SDD-1 maintains a 'directory locator' data structure at each site. Reconfiguration of a catalog has global consequences. The degree of site autonomy is determined by the data base designer, as the extent of the replicated directories is varied.

e. Concurrency

Transactions in SDD-1 involve three steps (Reference A-12). The first is the read phase where the TM analyzes what part of the data base has to be read and instructs the DM to read the necessary data. The second is the compute phase where all the access planning and materialization is done. The output is a list of data items ready to be written into the data base. This is done in the third step, namely the write phase. Here, the TM broadcasts the results to the DMs that are to write the new data items into the data base.

Transactions are grouped into predefined classes. Two transactions conflict only if their classes conflict. A conflict graph is used to determine the extent of conflict between two transactions. The graph helps to serialize the transactions for the technique of time-stamping. All data and transactions are time-stamped. The rule is that a transaction will never update data with a lesser time-stamp. Transactions within a class are serialized by a class pipelining rule, i.e., the transactions are done in order of their time-stamps. To synchronize across transaction classes, SDD-1 employs four sets of protocols.

f. Reliability

RELNET is essential for the reliable functioning of SDD-1 (Reference A-13). It monitors all sites and maintains a list of all sites that are active. SDD-1 uses the ARPANET communicating system, which checks for any duplicated messages and corrects any transmission errors. During a site failure, a spooler mechanism records the updates for the particular nodes. When the node recovers, it automatically updates itself from the spooler. The two-phase commit protocol is used by SDD-1 to ensure that the transaction is properly completed.

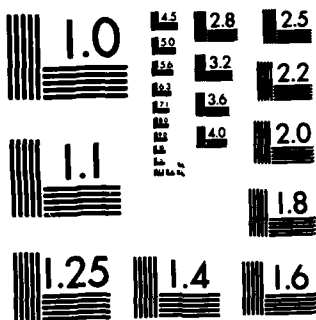
4. The CSIN System

a. Background

The Chemical System Information Network (CSIN) is built upon several heterogeneous data bases so that there can be sharing of data among the different sites (References A-14, A-15). The data handled by CSIN are widely varied; they can be bibliographic, text data, or numeric data. These data are shared by engineers, scientists, government agencies, academic institutions, and special interest groups. Due to the varied range of users, the queries will also be very different. Although the network intends to link different heterogeneous systems, special emphasis is on maintaining site autonomy. This is mandatory since different independent organizations support and maintain these data bases. While preserving the integrity of the information resources, CSIN will present its users with a uniform command interface.

b. Architecture of CSIN

The basic structural architecture of CSIN is that of processing layers (Figure A-5). There are several layers, each insulating the higher layer from the tasks of the lower layers (Reference A-16). The top two layers



MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

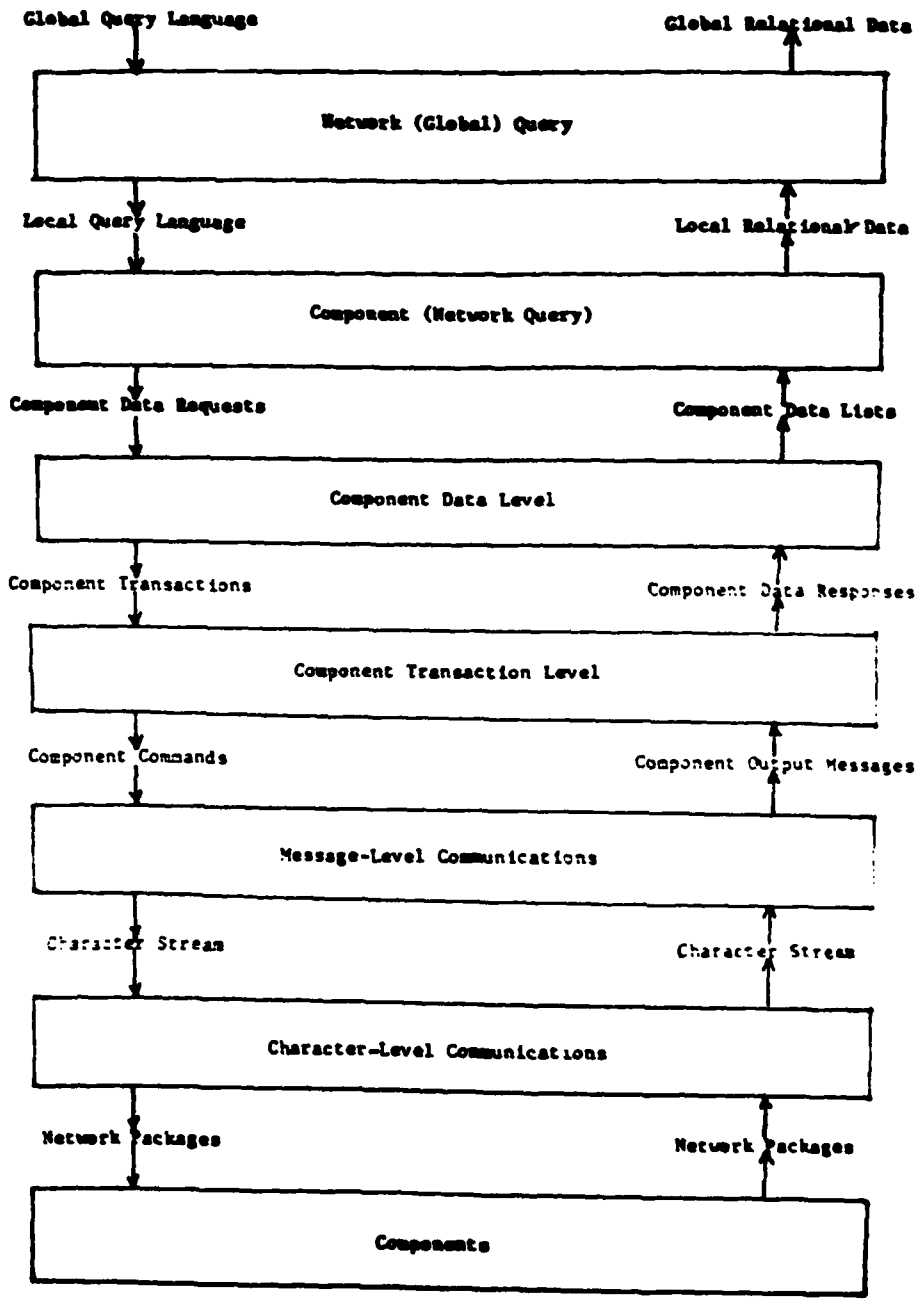


Figure A-5. CSIN Prototype Design Layers

are the local query layer and global query layer. The global query layer has the responsibility of decomposing the global users' queries into an efficient sequence of requests to the local query layer for each of the components. The query decomposition algorithm is similar to the one used in SDD-1. The local query layer transforms the queries issued by the global query layer, which is expressed in a subset of the relational query language, into a series of requests to the layer just below it, i.e., the component data layer. The resulting data-lists returned by the component data layer are then converted to relational form, and the individual data items are converted into the format specified by the component local data base schema. These top two layers are not yet implemented, so at present there is no formal high-level relational query language supported by CSIN. The user now generates queries by responding to a menu-driven interface.

The lowest two levels are the communication layers. The lowest one is called the character level communication and the one above that is known as the message level communication. The character level communication defines the interface between the hardware and software available for network communication and the internal communication. It carries out the buffering, multiplexing, and demultiplexing of the character stream. This layer is tightly coupled with the operating system and consists mostly of device drives operating at the interrupt level. The message level communication serves to start and maintain communication between the CSIN prototype and other communicating processors. This layer protects the higher layer from the details of a packet switching line or a dial-up line. In addition to serving as a connection manager, it also does message aggregation, buffering, and isolates higher layers from the overhead of character-by-character communication.

Just above the communication layers is the transaction layer. It deals with transactions by issuing appropriate commands, testing the results and returning results along with condition codes about the status of the transaction. Just above this layer and under the local query layer is the component data layer. The input and output of this layer is component-specific, containing an interface called the component data interface. This interface supports a set of global commands and captures the common functions provided by the CSIN component systems. However, due to the differences between the component systems, this interface will vary with each system. All output from the component data layer will be in the form of a common data organization called data-lists. Each component of the data-lists will be in a standard network-wide format. The local query layer will later use the component data schema to convert these data to the appropriate relational format for global processing.

c. Scripts

Frequently repeated transactions are automated by the use of scripts. Scripts simplify the execution of repetitive transactions by speeding up its execution and making queries more reliable. The goal of script capability is to create a single operation that can handle all the steps needed to carry out common and basic routine transactions. Scripts are written in the form of programs and can be invoked by name. There can be different levels in script implementation.

The levels of the scripts can be distinguished by the layer they interface; it can be the transaction layer, the message layer, data layer, or local query layer. Users can also create scripts in the form of files by using the CSIN editor. These files are checked for syntactic and semantic correctness before being used on the system. Script use in the role of automating certain transactions is akin to that of the transaction layer and the data layer, i.e., executing an operation, testing the result, transferring control, and processing the output. To perform these operations, the script language used should have special properties. It should be able to recognize messages sent from different component systems for menu I/O and I/O capability.

d. Development of the System

The preprototype system, implemented in 1979, was designed for a computer intelligent terminal (CIT) that automates the connection to a remote system (Reference A-17). After the user identifies the system to which he desires connection the terminal is automatically logged into the remote system. The second feature was to assist the users in generating routine queries. A terminal inputs "query-lists" (terms used to specify on-line searches) to remote terminals, for conducting a search. Using the CIT, users can capture data output from a remote system and store it in the form of files on the terminal disk. These locally stored data can be subsequently modified and printed. There is a general editing capability that enables users to add or delete from the query-list. The last feature can automatically transform data to be used by the query list.

These functions, although basic, greatly improved the information retrieval capability of CSIN, and the average speed of retrieval was 10 to 15 times that of the manual speed. Later, an improved form was developed and called Version I Prototype. It provides three levels of use: (1) direct use, (2) enhanced direct use, and (3) script use. The direct use of components provides the users with all the facilities of the preprototype version and permits complete control of the details of interacting with the remote systems. Here, the user should be knowledgeable of the language and instruction set of the remote system. With enhanced direct use, in addition to the services described above, the system can automate the sending of lengthy and tedious messages to remote systems, thus eliminating some user errors. In the script use mode, a prestored script (program) mediates between the user and any number of remote systems. In addition to all the capabilities provided by enhanced direct use, scripts can compose and send variable messages to remote systems and make decisions based on the response from remote systems. Specific scripts are planned to handle the most common cases of multisystem queries, while requiring minimum user input.

A number of additional utility functions were specified to manage and edit files. All of these CSIN prototype capabilities are available to the user via an easy-to-use menu-oriented interface. The user can shift between direct-use mode and script-use mode in case he runs into trouble while using the direct-use mode. If he needs to return to the direct-use mode, he should log-out from the script-use mode by typing in a special character.

At present, the preprototype and Version I have almost reached completion. The Version II which involves the construction of the top two

query layers (global and local) has yet to be developed. The fate of the Version II Prototype will depend on the development funding granted for the continuation of the CSIN effort.

5. Distributed INGRES

a. Background

The distributed INGRES system is being developed at the University of California, Berkeley (Reference A-18). It is built atop a centralized INGRES system, and many of the centralized features are preserved (Reference A-19).

b. Architecture Distributed INGRES

The Distributed INGRES is an extension to the centralized INGRES, to manage data bases distributed over multiple machines in a computer network. Distributed INGRES is a homogeneous DDBMS and all the data at various sites should be in relational form. The DDBMS used unix-to-unix communication facility for inter-machine communication. This requires that all sites to be running on the UNIX system. It is also possible to configure the DDBMS with varying amount of centralized control in a single machine (Reference A-18).

c. Query Processing

The query language used in INGRES is QUEL. The query is processed by decomposing a multivariable query into irreducible components which are one-variable queries (Reference A-20). It is the same method employed in the centralized system (References A-21, A-22). Record substitution is also done here; however, the problem is more complicated because an extra parameter, "location of data" comes into the algorithm for deciding the best overall query processing strategy.

The process of breaking the query into irreducible components is done at the "master" node, that is, the node where the query originated. All other participating nodes are known as "slaves." The whole query-splitting operation can be considered as a tree, with the global query as the root. The operation is a sequence of "move" and "run" commands issued by the master. The master, after getting the irreducible components forms a deferred update list, which is shuffled across the network to the appropriate slave nodes. The slave nodes and the master node follow a protocol that ensures a consistent data base at the end of the transaction. After each slave has completed its job on the subqueries, the results are stored in temporary relations that can be transferred to the master or to some other slave for further processing. The master then returns the unified result to the user.

The whole query process algorithm considers two factors: (1) the cost factor that involves data transfer, and (2) the parallel processing factor which, in turn, is a function of the maximum time required to process a subquery.

To determine these factors the following considerations are taken into account in the design:

- (1) model of the query (semantics and structures),
- (2) location of the fragments and the cardinalities of the relations involved, and
- (3) network type, e.g., point to point or broadcast.

d. Catalog

The CSIN designers assumed infrequent system updates, so the directory is replicated at every node. This makes retrieval very fast, but updates become costly and time-consuming. The INGRES directory is implemented as a set of eight relations, six of which are common to the centralized INGRES system. The directories contain local and global relations. The global relations are replicated at all sites, while the local relations are stored at only one site. The index relation (one of the eight) is a local relation; if it is required by other sites, this information is sent to them and is valid for a specific length of time, after which it becomes void. The designers, however, have not dealt with the problem of using outdated directories for query processing.

e. Concurrency

Each individual machine has its own concurrency controller which is almost identical to the one used in the centralized system (References A-23, A-24, A-25). Here, all locking is done locally, thus, site autonomy exists for locking of data. However, due to the unpredictability of the data retrieval pattern in a distributed environment, deadlocks will occur. The problem is complicated further by the retrieval strategy employed by INGRES. In a read-only transaction, the data are retrieved at the local site if a copy of the data is present locally. Otherwise the primary copy, stored at some other remote site is retrieved. For updates, the primary copy alone is accessed. This variation between global and local access methods has a great effect on ordering the transactions. Since local locking is employed, deadlock prevention is impossible. To overcome this, a centralized control detects the deadlock and breaks it. This is done by a special machine function called SNOOP which detects global deadlock by means of a wait-for graph.

A deadlock is detected if a cycle exists in a wait-for graph. In case SNOOP crashes, special software automatically configures another machine as SNOOP. The atomic transaction (action which when performed does not make the data base inconsistent) is defined as a QUEL statement.

f. Reliability

There is a two-phase commit protocol to ensure a consistent data base in case of system failures. The master has a commit point at which it sets a commit flag; if a failure occurs before this point then the whole transaction is reset. However, if a fault occurs after this point the data base is updated. If the node to receive the commit message has crashed, then the message is queued for the down node, which updates itself when it recovers. The sender of the message relies on a time-out strategy to decide if a node is alive; if no message is returned then the node is presumed to have crashed.

Every node maintains an "uplist" of all the nodes that are functioning in the system. The primary copy of data is the one stored at a site that has the lowest position in the "uplist." This uplist maintains a certain ordering of the nodes that it contains. In case a node fails, the uplist is updated and the query strategy is regenerated with the new version of the uplist and a new primary copy. In the event of a network partition, there will be two uplists. This creates a serious problem because two primary copies will exist for most of the stored data, and updating them will permanently corrupt the data. To avoid this, the primary copy is assigned to be on the side of the network that has the maximum number of copies of data. This is done by a simple counting of the copies on each side of the network. All nodes on the side of the network without primary copy cannot update their data. To implement this scheme, a very important point must be considered. The system must distinguish between a site failure and a network partition. In the former case, the uplists can be reconfigured and the primary copy determined and updated. In the latter case, determining the primary copy is not so straightforward and may corrupt the data if an improper primary copy is updated.

C. REFERENCES

- A-1. Landers, T.A. and Rosenberg, R.L., "An Overview of Multibase," in Distributed Databases, H.J. Schneider (editor), North Holland Publishing Company, Amsterdam, 1982.
- A-2. Smith, J.M., Bernstein, P.A., Dayal, U., Goodman, N., Landers, T., Lin, K.W. and Wong, E., "Multibase - Integrating Heterogeneous Distributed Database Systems," Proceedings of the National Computer Conference, 1981.
- A-3. Shipman, D., "The Functional Data Model and the Data Language DAPLEX," ACM Transactions on Database Systems, Vol. 6, No. 1, March 1981.
- A-4. Dayal, U., Landers, T. and Yedwab, L., Global Query Optimization in Multibase: a System for Heterogeneous Distributed Databases, Computer Corporation of America, Technical Report, Cambridge, October 1982.
- A-5. Haas, L. M., Selinger, P.G., Bertino, E., Daniels, D., Lindsay, B., Loham, G., Masernaga, Y., Mohan, C., Ng, P., Wilms, P. and Yost, R., "R*: A Research Project on Distributed Relational DBMS," Database Engineering, Vol. 5, No. 4, December 1982.
- A-6. Williams, R., Daniels, D., Haas, L., Lopis, G., Lindsay, B., Ng, P., Obermarck, R., Selinger, P., Walker, A., Wilms, P. and Yost, R., "R*: An Overview of the Architecture," Proceeding of the International Conference on Database Systems, Jerusalem, Israel, June 1982. Also published as IBM research report RJ3325, San Jose, California, December 1981.
- A-7. CICS/VS System/Application Design Guide, IBM form number SC33-0067-1, Chapter No. 13, June 1978.
- A-8. Ng, P., Distributed Compilation and Recompile of Database Queries, IBM research report RJ3375, San Jose, California, January 1982.

- A-9. Gray, J., McJones, P., Blasgen, M., Lindsay, B., Lorie, R., Price, T., Putzolu, F. and Traiger, I., "The Recovery Manager of the System R Database Manager," ACM Computing Surveys, Vol. 13, No. 2, June 1981.
- A-10. Rothnie, J.B., Bernstein, P.A., Fox, S., Goodman, N., Hammer, M., Landers, T.A., Reeve, C., Shipman, D.W. and Wong, E., "Introduction to a System for Distributed Databases (SDD-1)," ACM Transactions on Database Systems, Vol. 5, No. 1, March 1980.
- A-11. Bernstein, P.A., Goodman, N., Wong, E., Reeve, C.L. and Rothnie, J.B., "Query Processing in a System for Distributed Databases (SDD-1)," ACM Transactions on Database Systems, Vol. 6, No. 4, December 1981.
- A-12. Bernstein, P.A., Shipman, D.W. and Rothnie, J.B., "Concurrency Control in a System for Distributed Databases (SDD-1)," ACM Transactions on Database Systems, Vol. 5, No. 1, March 1980.
- A-13. Hammer, M. and Shipman, D., "Reliability Mechanisms for SDD-1: A System for Distributed Databases," ACM Transactions on Database Systems, Vol. 5, No. 4, December 1980.
- A-14. Eastlake III, D.E. and Leslie, J.T., CSIN Final Report 1978-1982, Computer Corporation of America, Technical report CCA-82-06, Cambridge, December 31, 1982.
- A-15. An Overview of the Chemical Substances Information Network, Computer Corporation of America, Draft, Cambridge, April 20, 1979.
- A-16. Eastlake III, D.E., Lozaro-Perez, T. and David, A.L., Design of the Version I Prototype Chemical Substance Information Network, Draft, Computer Corporation of America, Technical report no. CCA-80-6, Cambridge, September 12, 1980.
- A-17. Bolt, Beranek and Newmann Inc., "CSIN User Support: Intelligent Terminal Study," BBN report no. 4838, Boston, February 1982.
- A-18. Stonebraker, M. and Neuhold, E., "A Distributed Database Version of INGRES," Proceedings of the 1977 Berkeley Workshop on Distributed Data Management and Computer Networks, University of California, Berkeley, 1977.
- A-19. Stonebraker, M., Wong, E. and Kreps, P., "Design and Implementation of INGRES," ACM Transactions on Database Systems, Vol. 1, No. 3, September 1976.
- A-20. Epstein, R., Stonebraker, M. and Wong, E., Distributed Query Processing in a Relational Database System, Electronics Research Laboratory, University of California, Berkeley, (Internal Report), April 1978.
- A-21. Youssefi, K. and Wong, E., Query Processing in a Relational Database Management System, Electronics Research Laboratory, University of California, Berkeley, (Internal report), March 1978.

- A-22. Wong, E. and Youssefi, K., "Decomposition - A Strategy for Query Processing," ACM Transactions on Database Systems, Vol. 1, No. 3, September 1976.
- A-23. Stonebraker, M., Concurrency Control and Consistency of Multiple Copies of Data in Distributed INGRES, Electronics Research Laboratory, University of California, Berkeley, (Internal Report), May 1978.
- A-24. Stonebraker, M., "Concurrency Control, Crash Recovery and Consistency of Multiple Copies of Data in a Distributed Database System," Proceedings 3rd Berkeley Workshop on Distributed Databases and Computer Networks, San Francisco, California, August 1978.
- A-25. Stonebraker, M., "Concurrency Control and Consistency of Data in Distributed INGRES," IEEE Transactions on Software Engineering, Vol. SE-5, No. 3, May 1978.

APPENDIX B

EXAMPLE USE OF SAROAD INTERACTIVE-ORIENTED DATA BASE

A. INTRODUCTION

The Storage and Retrieval of Aerometric Data (SAROAD) file contains air quality data from throughout the United States and is maintained by the Environmental Protection Agency in Research Triangle Park, North Carolina. SAROAD may be viewed as a complete system which is divided into two major subsystems: (1) the SAROAD site file with its associated information, and (2) the SAROAD air quality data, consisting of several varieties. The following discussion is derived from the SAROAD User's Manual (Reference B-1).

B. SAROAD FILES

The SAROAD site file contains descriptive information on the sampling site. This information includes the address, latitude, longitude, and elevation of the sampling device. Data for approximately 15000 sites are stored in the SAROAD site file.

The SAROAD air quality data consist of ambient air quality and meteorological data collected at monitoring sites. Included in this file are:

1. Raw data less than 24 hours. A report from this file lists the hourly observations of pollutant concentrations in 24 columns, one line per day, one month per page (see Table B-1). Below the row for the last day of the month are three other rows, one for the average values for each hour, one for the number of observations made at each hour throughout the month, and one for the maximum value occurring at each hour. In the lower right-hand corner are the average values for the entire month, the total number of observations for the month, and the maximum value that occurred in the month.
2. Meteorological raw data. This report contains meteorological data such as wind speed (see Table B-2).
3. Year frequency distribution report. This report lists the site description and summary information for the data (see Table B-3). The site description gives the geographical information necessary to locate the site and is separated from the data by a data heading. Below the data heading, the codes and names are given for the pollutant, the sampling and analysis methods, the sampling interval, and the standard units in which the data are printed. The data items include: the year of the data; the percent of observations for continuous data; the number of observations; the number of primary and secondary violations of the NAAQSs (National Ambient Air Quality Standard); the minimum detectable for the sampling method; the minimum, maximum and second maximum observations; the 10, 30, 50, 70, 90, 95, and 99 percentiles; the arithmetic mean; the geometric mean; and the geometric standard deviation.
4. Quarterly frequency distribution report. This report contains quarterly summary information for the data (see Table B-4). The format is the same as for the yearly frequency distribution report.

TABLE B-1. RAW DATA LESS THAN 24-HOUR REPORT

MAY 21, 1981		NATIONAL AEROMETRIC DATA BANK:												PAGE 1													
MAY 19, 1981 DATA BASE VERSION		ENVIRONMENTAL PROTECTION AGENCY												SELECT CARD 1													
		OHIO																									
		TOLEDO																									
		NO 1 FIRE STATION 545 M MURON																									
		(36660007H09)																									
		REPORTING ORG:																									
		JAN 1977																									
		CARBON MONOXIDE																									
		#210111																									
		CONCENTRATION IN PARTS PER HILLION																									
		INSTRUMENTAL NONDISPERSIVE INFRARED																									
		01-HOUR DATA LISTING																									
DAY	12	1	2	3	4	5	6	7	8	9	10	11	12	1	2	3	4	5	6	7	8	9	10	11	DAILY MEAN NO.		
01	0.8	0.8	0.8	0.7	0.7	0.7	0.7	0.7	0.7	0.7	0.7	0.7	0.7	0.7	0.7	0.7	0.7	0.7	0.7	0.7	0.7	0.7	0.7	0.7	0.7	0.7	
02	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	
03	1.2	1.3	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	
04	3.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
05	1.7	0.7	1.0	0.6	0.6	0.6	0.6	0.6	0.6	0.6	0.6	0.6	0.6	0.6	0.6	0.6	0.6	0.6	0.6	0.6	0.6	0.6	0.6	0.6	0.6	0.6	
06	5.0	7.5	0.5	6.0	2.0	1.5	1.8	3.0	3.5	2.7	2.0	1.7	2.0	2.0	2.3	3.0	4.0	4.0	4.0	2.5	1.7	1.5	1.5	1.5	1.5	1.5	3.0
07	1.5	7.3	3.5	3.5	3.5	3.5	3.5	3.5	3.5	3.5	3.5	3.5	3.5	3.5	3.5	3.5	3.5	3.5	3.5	3.5	3.5	3.5	3.5	3.5	3.5	3.5	3.5
08	1.1	0.5	1.0	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
09	5.5	5.3	0.0	5.0	0.5	0.0	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
10	5.5	5.5	5.4	3.7	1.5	1.0	1.1	1.1	1.1	1.1	1.1	1.1	1.1	1.1	1.1	1.1	1.1	1.1	1.1	1.1	1.1	1.1	1.1	1.1	1.1	1.1	1.1
11	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8
12	1.5	7.7	5.5	5.5	5.5	5.5	5.5	5.5	5.5	5.5	5.5	5.5	5.5	5.5	5.5	5.5	5.5	5.5	5.5	5.5	5.5	5.5	5.5	5.5	5.5	5.5	5.5
13	1.5	7.7	5.5	5.5	5.5	5.5	5.5	5.5	5.5	5.5	5.5	5.5	5.5	5.5	5.5	5.5	5.5	5.5	5.5	5.5	5.5	5.5	5.5	5.5	5.5	5.5	5.5
14	0.9	7.7	6.6	6.6	6.6	6.6	6.6	6.6	6.6	6.6	6.6	6.6	6.6	6.6	6.6	6.6	6.6	6.6	6.6	6.6	6.6	6.6	6.6	6.6	6.6	6.6	6.6
15	4.1	1.1	0.3	0.7	1.0	0.7	0.6	0.7	0.7	0.7	0.7	0.7	0.7	0.7	0.7	0.7	0.7	0.7	0.7	0.7	0.7	0.7	0.7	0.7	0.7	0.7	0.7
16	6.6	6.6	6.6	6.6	6.6	6.6	6.6	6.6	6.6	6.6	6.6	6.6	6.6	6.6	6.6	6.6	6.6	6.6	6.6	6.6	6.6	6.6	6.6	6.6	6.6	6.6	6.6
17	4.4	3.3	3.3	3.3	3.3	3.3	3.3	3.3	3.3	3.3	3.3	3.3	3.3	3.3	3.3	3.3	3.3	3.3	3.3	3.3	3.3	3.3	3.3	3.3	3.3	3.3	3.3
18	7.5	7.3	3.3	6.1	2.4	2.4	2.4	2.4	2.4	2.4	2.4	2.4	2.4	2.4	2.4	2.4	2.4	2.4	2.4	2.4	2.4	2.4	2.4	2.4	2.4	2.4	2.4
19	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0
20	1.2	0.8	0.7	0.6	0.6	0.6	0.6	0.6	0.6	0.6	0.6	0.6	0.6	0.6	0.6	0.6	0.6	0.6	0.6	0.6	0.6	0.6	0.6	0.6	0.6	0.6	0.6
21	1.5	0.9	0.8	1.0	1.5	2.0	2.0	3.5	2.6	2.2	1.9	1.5	1.9	1.0	1.8	1.8	1.9	1.7	1.5	1.3	1.2	1.2	1.2	1.2	1.2	1.2	1.2
22	1.1	1.0	1.2	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
23	3.2	3.5	2.8	1.8	1.6	1.7	1.7	1.6	1.6	1.6	1.6	1.6	1.6	1.6	1.6	1.6	1.6	1.6	1.6	1.6	1.6	1.6	1.6	1.6	1.6	1.6	1.6
24	1.7	1.5	1.5	1.5	1.5	1.5	1.5	1.5	1.5	1.5	1.5	1.5	1.5	1.5	1.5	1.5	1.5	1.5	1.5	1.5	1.5	1.5	1.5	1.5	1.5	1.5	1.5
25	5.5	5.5	5.5	5.5	5.5	5.5	5.5	5.5	5.5	5.5	5.5	5.5	5.5	5.5	5.5	5.5	5.5	5.5	5.5	5.5	5.5	5.5	5.5	5.5	5.5	5.5	5.5
26	5.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
27	5.3	3.2	3.3	3.3	3.3	3.3	3.3	3.3	3.3	3.3	3.3	3.3	3.3	3.3	3.3	3.3	3.3	3.3	3.3	3.3	3.3	3.3	3.3	3.3	3.3	3.3	3.3
28	3.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
29	5.5	5.5	5.5	5.5	5.5	5.5	5.5	5.5	5.5	5.5	5.5	5.5	5.5	5.5	5.5	5.5	5.5	5.5	5.5	5.5	5.5	5.5	5.5	5.5	5.5	5.5	5.5
30	7.7	7.7	7.7	7.7	7.7	7.7	7.7	7.7	7.7	7.7	7.7	7.7	7.7	7.7	7.7	7.7	7.7	7.7	7.7	7.7	7.7	7.7	7.7	7.7	7.7	7.7	7.7
31	7.7	7.7	7.7	7.7	7.7	7.7	7.7	7.7	7.7	7.7	7.7	7.7	7.7	7.7	7.7	7.7	7.7	7.7	7.7	7.7	7.7	7.7	7.7	7.7	7.7	7.7	7.7
12	1	2	3	4	5	6	7	8	9	10	11	12	1	2	3	4	5	6	7	8	9	10	11	12			
AVG	1.0	0.9	0.9	0.9	0.9	0.9	0.9	0.9	0.9	0.9	0.9	0.9	0.9	0.9	0.9	0.9	0.9	0.9	0.9	0.9	0.9	0.9	0.9	0.9	0.9	0.9	
NO	31	31	31	31	31	31	31	31	31	31	31	31	31	31	31	31	31	31	31	31	31	31	31	31	31	31	
MAX	5.0	7.5	6.5	6.0	2.0	1.0	2.5	3.5	4.6	3.2	2.7	2.0	3.0	3.0	5.0	5.0	4.0	5.0	2.5	4.0	2.5	2.7	4.0	4.0	8.5	743	

TABLE B-2. METEOROLOGICAL RAW DATA REPORT

PAGE: 1

NATIONAL AERONAUTIC DATA BANK
MONTHLY METEOROLOGICAL DATA REPORT
FOR JAN 1973

RUN DATE: APR 09, 1981

LATITUDE: 30 D. 43 N. 11 S. W
LONGITUDE: 088 D. 03 W. 32 S. W
MTN ZONE: 16
MTN HOBTMNG: 3390070
MTN EASTING: 0020010
ELEVATION ABOVE GROUND: 010 FT.
ELEVATION ABOVE MSL: 6616 FT.
DIFF. GHI: WEST 06 HOURS

LOCATION: MOBILE
COUNTY (240C): MOBILE CO
SITE ADDR: WING TRANSMITTING SITE TELEGRAPH ROAD
STATION TYPE (111): CENTER CITY - INDUSTRIAL
ASCC (605): MOBILE-PENACOLA-PANAMA CITY-SOUTHERN MISSISSIPPI
SNEZ (5160): MOBILE, ALABAMA

SITCODE: 01230000
AGENCY/PROJECT: 001
AGENCY TYPE: COUNTY
CITY POPULATION: 199,076
SACC POPULATION: 2,108,546
FPA REGION: 6
SUPERVISING AGENCY: MOBILE COUNTY BOARD OF HEALTH
COMMENTS: SHIPPERS ON PLATFORM FOR UNIFORM WIND DISTRIBUTION NEAR
HEAVILY INDUSTRIALIZED AREA

UNITS: MILES/HOUR

METHOD: INSTRUMENTAL
SPOT READING

PARAMETER: WIND SPEED

DAY	00	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16	17	18	19	20	21	22	23	MEAN	NUM	
TIME	*****																										
****	*****																										
01	3.0	1.0	3.5	4.5	5.0	4.0	4.0	4.0	4.0	4.5	4.0	4.5	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.0
02	3.0	2.5	3.0	4.0	3.5	4.0	4.0	3.5	4.0	3.5	4.0	4.0	4.0	3.5	3.5	3.5	3.5	3.5	3.5	3.5	3.5	3.5	3.5	3.5	3.5	3.5	3.5
03	2.0	1.5	1.5	2.0	2.5	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0
04	2.0	1.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0
05	2.0	2.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0
06	2.0	2.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0
07	3.5	4.5	4.5	3.5	3.5	3.5	3.5	3.5	3.5	3.5	3.5	3.5	3.5	3.5	3.5	3.5	3.5	3.5	3.5	3.5	3.5	3.5	3.5	3.5	3.5	3.5	3.5
08	2.5	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0
09	4.0	4.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0
10	3.0	2.5	2.5	2.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0
11	3.0	2.5	2.5	2.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0
12	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0
13	3.5	1.5	1.0	1.5	1.5	1.5	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0
14	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0
15	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0
16	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0
17	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0
18	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0
19	2.5	2.0	2.5	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0
20	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0
21	3.5	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.0
22	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
23	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
24	3.5	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.0
25	4.0	4.0	3.0	3.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0
26	4.0	4.0	3.0	3.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0
27	4.0	4.0	3.0	3.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0
28	3.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0
29	3.0	3.5	3.5	3.5	3.5	3.5	3.5	3.5	3.5	3.5	3.5	3.5	3.5	3.5	3.5	3.5	3.5	3.5	3.5	3.5	3.5	3.5	3.5	3.5	3.5	3.5	3.5
30	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0
31	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0
MEAN	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0
MAX	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0
AVG	1.7	1.7	1.7	1.7	1.7	1.7	1.7	1.7	1.7	1.7	1.7	1.7	1.7	1.7	1.7	1.7	1.7	1.7	1.7	1.7	1.7	1.7	1.7	1.7	1.7	1.7	1.7
08Z	31	31	31	31	31	31	31	31	31	31	31	31	31	31	31	31	31	31	31	31	31	31	31	31	31	31	31

TABLE B-3. YEARLY FREQUENCY DISTRIBUTION REPORT

NATIONAL AEROMETRIC DATA BANK
YEARLY FREQUENCY DISTRIBUTION
STATE (41): RHODE ISLAND

LATITUDE: 41 D. 49 M. 54 S. N
LONGITUDE: 71 D. 24 M. 56 S. W
UTM ZONE: 19
UTM NORTHING: 4633700
UTM EASTING: 00299400
ELEVATION ABOVE GROUND: 050 FT.
ELEVATION ABOVE MSL: 0100 FT.
DIFF. GMT: MEET 05 HOURS

05-21-81
SITECODE: 41030005F01
AGENCY/PROJECT: F01
AGENCY TYPE: STATE
CITY POPULATION: 179,213
RUR POPULATION: 1,645,380
EPA REGION: 1
SUPPORTING AGENCY: RHODE ISLAND DEPARTMENT OF HEALTH
COMMENTS: PROVIDENCE STATION B

LOCATION: PROVIDENCE
COUNTY (0320): PROVIDENCE CO
SITE ADDR: STATE OFFICE BUILDING
STATION TYPE (11): CENTER CITY - INDUSTRIAL
ABCR (120): METROPOLITAN PROVIDENCE
SMSA (6480): PROVIDENCE-PANTUCKET-WARRICK, R. I.-MASS
SUPPORTING AGENCY: RHODE ISLAND DEPARTMENT OF HEALTH
COMMENTS: PROVIDENCE STATION B

POLLUTANT NAME	METHOD OF COLLECTION AND ANALYSIS										INTERVAL			STANDARD UNITS							
	POLLUTANT-METHOD-INTERVAL-UNITS CODE		MIN		SEC		PERCENTILES		90	95	99	MAX	2ND	ARIT	GEOM						
YR	ORG	OBS	PRI	SEC	BETEC	MIN	MAX	10	30	50	70	90	95	99	MAX	OBS	MEAN	MEAN	STD	DEV	
PARTICULATE																					
11101-91-7-01																					
73		51	0	0	1	21	33	40	46	64	107	130	138	137	136	137	58	52	1.6	1.6	
74		41	0	2	1	23	37	51	61	82	103	142	157	155	157	155	70	63	1.6	1.6	
75		54	0	0	1	32	35	43	57	67	97	102	121	119	121	119	60	57	1.4	1.4	
76		50	0	0	1	27	37	47	53	64	85	117	129	129	129	129	60	60	1.4	1.4	
HI-VOL GRAVIMETRIC																					
24-HOUR UG/CU METER (25 C)																					
73		47	4126	0	0	0.6	0.3	0.6	1.2	1.7	2.9	4.0	5.2	6.9	19.6	13.6	2.3*	1.9*	1.9*	1.9*	
CARBON MONOXIDE																					
42101-11-1-05																					
73		4098	0	0	0.6	0.6	0.6	0.9	1.4	2.0	2.7	3.9	4.7	6.1	9.1	8.2					
1-HOUR MG/CU METER (25 C)																					
8-HR-AVG MG/CU METER (25 C)																					
INSTRUMENTAL MONDISPERSIVE INFRA-RED																					
INSTRUMENTAL MONDISPERSIVE INFRA-RED																					
SULFUR DIOXIDE																					
42401-11-1-01																					
73		85	7471	0	0	26	13	39	66	118	216	288	524	956	956	956	98	59	2.9	2.9	
74		84	7349	0	0	26	13	13	39	74	170	282	406	968	939	939	68	38	2.9	2.9	
75		24	2134	0	0	26	13	13	45	70	108	168	212	311	417	390	66.4	65.4	2.3*	2.3*	
1-HOUR UG/CU METER (25 C)																					
24-HR-AVG UG/CU METER (25 C)																					
SULFUR DIOXIDE																					
42401-11-X-01																					
73		6960	0	0	26	13	28	53	78	115	195	266	401	634	538						
74		7111	1	0	26	13	14	26	44	77	150	212	322	414	358						
75		2117	0	0	26	13	34	61	86	106	150	167	195	222	204						

** DENOTES A VALUE DERIVED FROM DATA WHICH DO NOT MEET SAROD SUMMARIZATION CRITERIA OF OARPS GUIDELINE 1.2-340, VOL 3, SEC 2.3.0

TABLE B-4. QUARTERLY FREQUENCY DISTRIBUTION REPORT

PAGE 41-0001

NATIONAL AEROMETRIC DATA BANK
QUARTERLY FREQUENCY DISTRIBUTION
STATE (41), RHODE ISLAND

05-21-81

SITICODE: 41030005701
AGENCY/PROJECT: F01
AGENCY TYPE: STATE
CITY POPULATION: 179,213
AREA POPULATION: 1,645,300
EPA REGION: 1
SUPPORTING AGENCY: RHODE ISLAND DEPARTMENT OF HEALTH
COMMENTS: PROVIDENCE STATION B

LOCATION: PROVIDENCE
COUNTY (0120): PROVIDENCE CO
SITE ADDR: STATE OFFICE BUILDING
STATION TYPE (11): CENTER CITY - INDUSTRIAL
ASCR (128): METROPOLITAN PROVIDENCE
SMSA (640): PROVIDENCE-PANTUCKET-HARRICK, R.I.-MASS
LATITUDE: 41 D. 49 W. 54 S. N
LONGITUDE: 71 D. 24 W. 56 S. W
UTM ZONE: 19
UTM EASTING: 643700
UTM NORTHING: 0029400
ELEVATION ABOVE GROUND: 950 FT.
ELEVATION ABOVE MSL: 0100 FT.
DIFF. GMT: WEST 05 HOURS

POLLUTANT NAME	METHOD OF COLLECTION AND ANALYSIS												STANDARD UNITS					
	POLLUTANT-METHOD-INTERVAL-UNITS CODE				PERCENTILES				INTERVAL				24-HOUR		1-HOUR		8-HR-AVG	
REP X	0	EXCURSIONS	MIN	MIN	10	30	50	70	90	95	99	MAX	2ND	ARIT	GEOM	MEAN	STD DEV	
YR-RT	ORG	OBS	PRI	SEC	BTEC	OBS	OBS	OBS	OBS	OBS	OBS	OBS	MAX	MEAN	MEAN	MEAN	STD DEV	
PARTICULATE																		
MI-VOL GRAVIMETRIC																		
1181-91-7-01	16	0	0	1	21	42	48	51	98	122	137	137	137	122	72	64	1.7	
73-01	14	0	0	1	28	37	37	43	46	69	107	107	107	69	48	45	1.4	
73-02	13	0	0	1	21	33	39	43	68	130	138	138	138	130	58	51	1.7	
73-03	8	0	0	1	32	32	37	38	59	71	71	71	71	64	47	46	1.4	
CARBON MONOXIDE																		
INSTRUMENTAL NONDISPERSIVE INFRA-RED																		
42101-11-1-85	93	1998	0	0	0.6	0.6	1.7	2.3	2.9	4.6	5.8	7.5	19.6	13.8	2.5	2.1	1.94	
73-01	97	2128	0	0	0.6	0.3	0.6	1.2	1.7	2.3	3.5	5.8	10.4	9.2	2.0	1.7	1.03	
CARBON MONOXIDE																		
INSTRUMENTAL NONDISPERSIVE INFRA-RED																		
42101-11-2-85	1982	0	0	0.6	0.6	1.0	1.7	2.4	3.1	4.3	5.2	6.6	9.1	8.2	6.1	6.1	0.0	
73-01	2114	0	0	0.6	0.6	0.9	1.3	1.7	2.3	3.6	4.2	5.5	6.9	6.1	6.1	6.1	0.0	
73-02	2	0	0	0.6	2.3	2.3	2.3	2.3	2.3	2.3	2.3	2.3	2.3	2.3	2.3	2.3	0.0	

C. SAROAD INTERACTIVE SYSTEM

The SAROAD Terminal System enables users to access air quality data files for retrieval via demand (interactive) terminals. The system is designed so that users who are not familiar with computers as well as users who are computer-oriented can access the system and submit their requests.

When the computer is ready to accept commands, it will display 'COMMANDS'. The user should select and enter a command word from the list shown in Table B-5. The command word can be entered in its entirety or can be abbreviated to the first two characters of the word. Where '=' is present in the command, it should appear after the first two characters if the abbreviated mode is being used; otherwise, the '=' should follow the full name of the command.

Throughout the program the site-pollutant commands are listed as follows:

```
'state=??'  
'site=???'
```

The number of question marks following the equal sign defines the number of characters in that code. In the above examples, the state command requires two characters following the equal sign and site requires three. Asterisks are used in the 'do not care' fields.

An example interactive session is shown in Table B-6. The user is requesting a quarterly summary report. The SAROAD "quarterly summaries" command is an interactive function which allows the user to retrieve selected summary information for any site for a given pollutant. When the quarterly summaries (QS) function is entered, the user may ask for a listing of the available commands by responding 'Yes' to the display 'COMMAND LIST - - - YES OR NO?'. The final output for this example consists of quarterly reports of carbon monoxide concentrations at Fairbanks, Alaska. Data such as these could be used for calibration of air quality models such as CDM, as discussed in Section IV.

D. REFERENCES

- B-1. U.S. Environmental Protection Agency, AEROS Manual Series Volume III: Summary and Retrieval, Third Edition, EPA-450/2-76-009b, Office of Air, Noise and Radiation, Research Triangle Park, NC, July 1981.

TABLE B-5. SAROAD COMMANDS

THE VALID REPORT TYPES ARE:

KR KEYS QUALIFYING REPORT
QR QUARTERLY SUMMARIES REPORT

THE VALID SITE-POLLUTANT COMMANDS ARE:

STATE=?? THE STATE DESIRED
AREA=???? THE CODE NUMBER FOR THE CITY OR COUNTY
SITE=??? THE SITE NUMBER WITHIN THE AREA
AGENCY=? THE CODE FOR THE SPONSORING AGENCY
PROJECT=? THE CODE FOR THE TYPE OF SAMPLING
POLLUTANT=????? THE CODE FOR THE DESIRED POLLUTANT
METHOD=? THE CODE FOR THE SAMPLING METHOD
INTERVAL=? THE CODE FOR THE SAMPLING INTERVAL
BYEAR=? BEGINNING YEAR DESIRED
BQUARTER=? BEGINNING QUARTER DESIRED
EYEAR=? ENDING YEAR DESIRED
EQUARTER=? ENDING QUARTER DESIRED
KEY=???... THE COMPLETE KEY IN THE ORDER GIVEN ABOVE
(32 CHARACTERS)

THE VALID SUMMARY INFORMATION COMMANDS ARE:

ALL ALL STATISTICS REQUIRED

THE VALID COMMANDS FOR PROGRAM DIRECTION ARE:

FIND (QR) GETS THE FIRST SEQUENTIAL RECORD OF THE
QUARTERLY SUMMARY INFORMATION CONTAINING
THE USER INPUT KEY
(KR) GETS ALL THE RECORD KEYS OF THE QUARTERLY
SUMMARY FILE CONTAINING THE USER INPUT KEY
NEXT ?? (QR) GETS THE NEXT SEQUENTIAL RECORD OF THE
QUARTERLY SUMMARY INFORMATION OF THE SUBSET
CONTAINING THE USER INPUT KEY. ?? IS THE
NUMBER OF RECORDS WANTED.
HELP LISTS ALL VALID COMMANDS
END SIGNALS END OF SESSION. PROGRAM ENDS.

TABLE B-6. EXAMPLE INTERACTIVE SESSION

GNA206

SAROAD TERMINAL SYSTEM (V.2) DATE: 08/10/78

TIME: 10:42:12

ARE YOU USING A CRT TERMINAL? (YES OR NO)

YES

DO YOU WISH FOR INSTRUCTIONS? (YES OR NO)

NO

FUNCTION?

QS

AIR POLLUTION QUARTERLY DATA SUMMARY REQUESTS

COMMAND LIST---YES OR NO?

YES

THE VALID REPORT TYPES ARE:

'KR' KEYS QUALIFYING REPORT
'QR' QUARTERLY SUMMARIES REPORT

THE VALID SITE-POLLUTANT COMMANDS ARE:

'STATE=??' THE STATE DESIRED
'AREA=????' THE CODE NUMBER FOR THE CITY OR COUNTY
'SITE=???' THE SITE NUMBER WITHIN THE AREA
'AGENCY=?' THE CODE FOR THE SPONSORING AGENCY
'PROJECT=??' THE CODE FOR THE TYPE OF SAMPLING
'POLLUTANT=?????' THE CODE FOR THE DESIRED POLLUTANT
'INTERVAL=?' THE CODE FOR THE SAMPLING INTERVAL
'BYEAR=??' BEGINNING YEAR DESIRED
'BQUARTER=???' BEGINNING QUARTER DESIRED
'BDAY=???' NOT USED, ENTER **.
'EYEAR=??' ENDING YEAR DESIRED
'EQUARTER=???' ENDING QUARTER DESIRED
'EDAY=???' NOT USED, ENTER **.
'KEY=???...' THE COMPLETE OR PARTIAL KEY IN THE ORDER GIVEN ABOVE

THE VALID SUMMARY INFORMATION COMMANDS ARE:

'ALL' ALL STATISTICS ARE REQUIRED
'OBSERV' THE NUMBER OF OBSERVATIONS
'MIN' THE MINIMUM VALUE ENCOUNTERED
'MAX' THE MAXIMUM VALUE ENCOUNTERED
'MTIME' TIME MAXIMUM OCCURRED (MM:DA:HR)
'SMAX' SECOND MAXIMUM VALUE
'TS' TIME SECOND MAXIMUM OCCURRED (MM:DA:HR)
'%OBSERV' THE PERCENT OF POSSIBLE OBSERVATIONS THAT ARE PRESENT
'ZSUB' ONE HALF THE MINIMUM DETECTABLE
'ZCNT' THE NUMBER OF TIMES VALUES OCCURRED THAT WERE LESS THAN THE MINIMUM DETECTABLE

FOR INTERVALS OTHER THAN 3, 8 AND 24 HOUR RUNNING AVERAGES--

'AMEAN' THE ARITHMETIC MEAN
'GMEAN' THE GEOMETRIC MEAN
'ASTDEV' THE ARITHMETIC STANDARD DEVIATION

TABLE B-6. EXAMPLE INTERACTIVE SESSION (CONTINUED)

'GSTDEV' THE GEOMETRIC STANDARD DEVIATION
'PVIOL' NUMBER OF PRIMARY VIOLATIONS
(BLANK IF NO STANDARD)
'SVIOL' NUMBER OF SECONDARY VIOLATIONS
(BLANK IF NO STANDARD)

FOR 3, 8 AND 24 HOUR RUNNING AVERAGES--

'NPRI' NUMBER OF NON-OVERLAPPING PRIMARY VIOLATIONS
'NSEC' NUMBER OF NON-OVERLAPPING SECONDARY VIOLATIONS
'OP' NUMBER OF OVERLAPPING VALUES THAT EXCEEDED
THE PRIMARY STANDARD
'OS' NUMBER OF OVERLAPPING VALUES THAT EXCEEDED
THE SECONDARY STANDARD

VALID SUMMARIZATION CRITERIA COMMANDS ARE:

'CR' SELECT ONLY CRITERIA DATA
'NC' SELECT ONLY NON-CRITERIA DATA
'BB' SELECT BOTH CRITERIA AND NON-CRITERIA DATA.
BLANK STATISTICS FOR NON-CRITERIA DATA.
'BF' SELECT BOTH CRITERIA AND NON-CRITERIA DATA.
FLAG NON-CRITERIA DATA.

THE VALID COMMANDS FOR PROGRAM DIRECTION ARE:

'FIND' (QR) GETS THE FIRST SEQUENTIAL RECORD OF THE
QUARTERLY SUMMARY INFORMATION CONTAINING
THE USER INPUT KEY.
(KR) GETS ALL THE RECORD KEYS OF THE QUARTERLY
SUMMARY FILE CONTAINING THE USER INPUT KEY
'NEXT ??' (QR) 1. GETS THE NEXT SEQUENTIAL RECORD OF THE QUARTERLY
SUMMARY INFORMATION OF THE SUBSET CONTAINING THE USER
INPUT KEY. ?? IS THE NUMBER OF RECORDS WANTED.
(QR) 2. IF NO MATCH ON 'FIND' COMMAND, A NEW KEY MUST
BE INPUT.
'?KEY' ALLOWS USER TO LOOK AT CURRENT USER KEY
'XKEY' FILLS USER KEY WITH ALL ASTERISKS
'HELP' LISTS ALL COMMANDS.
'END' SIGNALS END OF SESSION. PROGRAM ENDS.

START ENTERING COMMANDS

COMMAND?

KR

KR = KEYS QUALIFYING REPORT

COMMAND?

KE=020160*F**42101***750300760200**

COMPLETE KEY=020160***F**42101***750300760200

COMMAND?

F1

TABLE B-6. EXAMPLE INTERACTIVE SESSION (CONTINUED)

KEYS QUALIFYING REPORT

KEY ENTERED:	ST	AREA	SITE A	PR	POLL	ME	I	BEG DATE	END DATE
	02	0160	*** F	**	42101	**	*	7503**	7602**

KEYS MATCHING:	ST	AREA	SITE A	PR	POLL	ME	I	DATE
	02	0160	013 F	01	42101	11	1	75/03
	02	0160	013 F	01	42101	11	1	75/04
	02	0160	013 F	01	42101	11	1	76/01
	02	0160	013 F	01	42101	11	1	76/02
	02	0160	013 F	01	42101	11	Z	75/03
	02	0160	013 F	01	42101	11	Z	75/04
	02	0160	013 F	01	42101	11	Z	76/01
	02	0160	013 F	01	42101	11	Z	76/02

NO. MATCHING KEYS: 8

COMMAND?

GR

GR = QUARTERLY SUMMARIES REPORT

COMMAND?

KE=020160013F014210111*750400750400

COMPLETE KEY=020160013F014210111*750400750400

COMMAND?

FI

SITECODE=020160013F01 FAIRBANKS ,ALASKA
 POLL/ME=4210111 UNITS=05 INT=1 YR=75 GR=04
 CARBON MONOXIDE INSTRUMENTAL NONDISPERSIVE INFRA-RED

COMMAND?

ALL

ALL

COMMAND?

NE

NEXT HIGHER SEQUENTIAL RECORD AFTER USER INPUT KEY

TABLE B-6. EXAMPLE INTERACTIVE SESSION (CONCLUDED)

```

SITECODE=020160013F01    FAIRBANKS          ,ALASKA
POLL/ME=4210111          UNITS=05    INT=Z    YR=75    GR=01
CARBON MONOXIDE          INSTRUMENTAL NONDISPERSIVE INFRA-RED
# OBSERVATIONS=                1807
MAXIMUM=                      30.6
MINIMUM=                       .3
SUBSTITUTE VALUE=              .3
# SUBSTITUTIONS MADE=          0
# SECONDARY VIOLATIONS=        153
# NON-OVLP PRI-VIOLATIONS =    26
# NON-OVLP SEC-VIOLATIONS =    26
TIME MAXIMUM OCCURRED=        01:22:18
TIME SECOND MAXIMUM OCCURRED=  01:22:22
SECOND MAXIMUM VALUE=         23.7
# OVLP VALS EXCEEDING PRIMARY= 153
# OVLP VALS EXCEEDING SECONDARY= 153
    
```

```

SITECODE=020160013F01    FAIRBANKS          ,ALASKA
POLL/ME=4210111          UNITS=05    INT=Z    YR=75    GR=02
CARBON MONOXIDE          INSTRUMENTAL NONDISPERSIVE INFRA-RED
# OBSERVATIONS=                1688
MAXIMUM=                      4.2
MINIMUM=                       .3
SUBSTITUTE VALUE=              .3
# SUBSTITUTIONS MADE=          0
# SECONDARY VIOLATIONS=        0
# NON-OVLP PRI-VIOLATIONS =    0
# NON-OVLP SEC-VIOLATIONS =    0
TIME MAXIMUM OCCURRED=        04:04:13
TIME SECOND MAXIMUM OCCURRED=  04:03:12
SECOND MAXIMUM VALUE=         4.0
# OVLP VALS EXCEEDING PRIMARY= 0
# OVLP VALS EXCEEDING SECONDARY= 0
    
```

```

SITECODE=020160013F01    FAIRBANKS          ,ALASKA
POLL/ME=4210111          UNITS=05    INT=Z    YR=75    GR=03
CARBON MONOXIDE          INSTRUMENTAL NONDISPERSIVE INFRA-RED
# OBSERVATIONS=                1946
MAXIMUM=                      3.7
MINIMUM=                       .3
SUBSTITUTE VALUE=              .3
# SUBSTITUTIONS MADE=          0
# SECONDARY VIOLATIONS=        0
# NON-OVLP PRI-VIOLATIONS =    0
# NON-OVLP SEC-VIOLATIONS =    0
TIME MAXIMUM OCCURRED=        07:23:22
TIME SECOND MAXIMUM OCCURRED=  09:17:12
SECOND MAXIMUM VALUE=         3.7
# OVLP VALS EXCEEDING PRIMARY= 0
    
```

APPENDIX C

EXAMPLE OF USE OF HISARS BATCH-ORIENTED DATA BASE

A. DESCRIPTION OF HISARS

The acronym "HISARS" stands for "Hydrologic Information Storage and Retrieval System." This computer facility for manipulating a hydrologic data base was written by Dr. E.H. Wiser of North Carolina State University (Reference C-1). It is provided for use at the University of Florida through the Institute of Food and Agricultural Science (IFAS) Department of Statistics. The system's command language is so simple that students can make good use of HISARS after an hour of studying the user's manual (Reference C-2).

The full version of HISARS contains data on rainfall, air temperature, pan evaporation, evaporation pan temperature, occurrences of weather events (e.g., fog, thunder, tornado, etc.), snowfall, stream flow, stream hydrograph information, stream dam characteristics, and a descriptive file on stream environmental conditions. However, the version available at the University of Florida contains only National Weather Service data from stations in Florida (i.e., no stream data). The data are stored on a computer disk pack, so that users must run batch jobs which are executed once the disk pack has been accessed. Typical turnaround time for a high-priority job is 5-10 minutes. However, the system could readily be made interactive if the data were stored in a form immediately accessible to the computer. In practice, the use of batch mode is frequently preferable since many typical information requests produce an amount of data which is easier to read and evaluate when seen on a print-out rather than a few lines at a time on a computer terminal screen. This is particularly true if the user is scanning the data files to see what available data might be relevant to his problem, so that he can then submit a more detailed and limited data request.

B. HISARS LANGUAGE

HISARS contains its data on different data files which are listed in the user's manual. Once an appropriate data file has been chosen, the computer can be directed to read that file with the IBM Job Control Language (JCL) command "// EXEC filename." Data can then be extracted using HISARS command language. The HISARS commands are shown in Table C-1; most of them are in the form "COMMAND...operand." The "COMMAND" tells the computer what type of operation to perform, while the "operand" gives the details of the operation. For instance, all requests for a particular section of the data record start with the command "PERIOD," while the operand specifies the starting and ending dates of the period of interest. Each set of commands specifying a particular data retrieval starts with the command "ACCESS," which is then followed by a string of commands which specify the data type, measurement location, period of interest and type of output required.

C. EXAMPLE OF USE

A hypothetical hydrology problem provides a good example of how HISARS might be used. A number of lakes are situated in and around Gainesville, Florida. Suppose that a monthly water budget must be calculated for one of

TABLE C-1. HISARS COMMAND LANGUAGE

Command	Operand(s)	Command Use
ACCESS	None	First command in a string specifying a particular data retrieval; indicates beginning of command string.
ELEMENT	One word description of data type (e.g., RAINFALL, TEMPERATURE, ETC.)	Specifies type of data required - data must be available on the file specified in the job control language written before the "ACCESS" command.
STATION	One or more 6-digit numbers specifying station(s) at which data were measured.	Can be used to retrieve data from one or more stations whose code numbers are known to the user.
ALTERNATIVE	One or more code numbers.	Can be used to specify a "flag code" on stations (e.g., a particular code number could be attached to all stations tended by a particular agency).
LOCATION	One or more 4 to 10 digit numbers.	Requests data from all stations within the specified area(s). Areas defined by latitude and longitude. Areas must be a quadrangle, and may be up to one degree or down to 60/64 minutes on each side.
BASIN	One or more code numbers.	Requests data from all stations within the specified river basin(s) or water use area(s).
REGION	One or more 2-digit code numbers.	Requests data from all stations within the specified NWS climatological region(s).
COUNTY	One or more county names	Requests data from all stations within the named counties.
ELEVATION	Minimum elevation, maximum elevation	Requests data from all stations within the specified minimum and maximum elevations.
AREA	Minimum area, maximum area	Requests data from all stations within a river basin having an area between the specified minimum and maximum drainage areas.
PERIOD	"Starting date" to "ending date"	Specifies period of interest for which data are required. Dates are specified numerically (e.g. January 1976 = "1/1976"). Default value is complete period of record.
LIST	"INDEX", "MONTHLY" or "DAILY"	Specifies printed output, "INDEX" gives a summary of the periods of complete record available at the specified stations. "MONTHLY" and "DAILY" specify monthly or daily measured values in standard formats. (E.g., total daily or monthly rainfall, maximum and minimum daily temperatures, average maximum and minimum monthly temperatures, etc.)
PROCESS	Various	Performs statistical analyses on requested data, prints results.
COPY	Several	Writes requested data onto disk file or computer tape in standardized formats. Data can then be read as input data for another computer program; it is often necessary to write a program to manipulate the data into a suitable format for input into another program.
AND	None	Logical "and" used to link two or more data request commands.
OR	None	Logical "or" used to link two or more data request-commands.

TABLE C-2. PROGRAM TO SEARCH FOR APPROPRIATE STATION RECORDS

```
0000 //DFM JOB (2006,3401,2,2,0), 'D.F.MACINTYRE', CLASS=A
0001 //♦PASSWORD
0002 //♦ROUTE PRINT REMOTE1
0003 //♦SETUP          3330,CROP01
0004 // EXEC RAIN
0005 ACCESS
0006 ELEMENT          RAINFALL
0007 COUNTY           ALACHUA
0008 LIST             INDEX
0009 // EXEC EVAP
0010 ACCESS
0011 ELEMENT          EVAPORATION
0012 COUNTY           ALACHUA
0013 LIST             INDEX
0014 // EXEC TEMP
0015 ACCESS
0016 ELEMENT          TEMPERATURE
0017 COUNTY           ALACHUA
0018 LIST             INDEX
0019 //♦ EOJ
END OF WORK FILE
```

these lakes for the years 1970 through 1979. In addition, a rainfall-runoff model is to be calibrated for an intermittent creek that drains into the lake, for which daily discharge measurements during 1979 are available. The lake is typical of many Florida lakes in that it has no permanent inflows or outflows; it may be thought of as a large depression. The only inputs are rainfall and runoff from rain falling on its surrounding watershed, while the only outputs are evaporation and groundwater seepage. In order to estimate inputs and outputs for the lake some knowledge of rainfall, evaporation and potential interaction with groundwater is therefore required. In calculating runoff over a long term it is sometimes useful to estimate the amount of water lost to evapotranspiration; and for this purpose information on temperatures is needed.

HISARS can help with the need for rainfall, evaporation, and temperature data, but it does not contain information on groundwater. The first use of HISARS will be to see if any stations can be found near the hypothetical lake which contain rainfall, evaporation and temperature data measured during the period 1970 through 1979. The rainfall, evaporation and temperature files must therefore be accessed. In addition to specifying the source file, candidate stations and type of output desired must also be specified. For this example, it is assumed that specific stations near the lake are not known; hence, one of the commands that instruct the program to search all stations within a geographical area around the lake must be used. The two most appropriate such commands are LOCATION and COUNTY. Using LOCATION all stations within a quadrangle centered on the lake could be specified, while using COUNTY requests HISARS to specify all stations in the surrounding counties. Since Gainesville is located very centrally in Alachua County, the simpler command is "COUNTY ALACHUA." For output, a summary of the available data is desired; output is specified as "LIST INDEX." The program which does this is shown in Table C-2, and the output is shown in Tables C-3 through C-5.

From these tables, it can be seen that three stations are close enough to Gainesville to be worth considering as being representative of conditions at the lake. Of these three stations, only Gainesville 2 WSW (code number 08-3321) has records covering the years 1970 through 1979. A second program can now be written to extract the required data. Again, the rainfall, evaporation and temperature files must be accessed. The specific station from which data are required (number 08-3321) can now be specified for the water budget. Monthly data on rainfall, evaporation and temperature for the period 1970 through 1979 are required, so "PERIOD 1/1970 to 12/1979" and "LIST MONTHLY" for these elements is specified. However, daily rainfall data are also required to compare with daily discharge measurements during 1979 so that the rainfall-runoff model can be calibrated. Hence, "PERIOD 1/1979 to 12/1979" and "LIST DAILY" for rainfall is also specified. This program is shown in Table C-6, and the output from it is shown in Tables C-7 through C-10.

The data in Tables C-7, C-8 and C-10 are complete, but many of the values in the table of monthly evaporation values (Table C-9) are followed by " - " symbols. This means that these months contain missing data for one or more days. It would be advisable to look at the daily listings for these months so that some estimating technique could be applied to fill in the missing data. To use the data as input to computer programs for calculating the water budget and rainfall runoff relationship, the user could manually keypunch the data from the written output tables. Alternatively, the user could submit another

TABLE C-3. SUMMARY OF AVAILABLE RAIN DATA

RAINFALL STATIONS		LONGITUDE		REGION		STATION NO.	
LATITUDE	ELEVATION	PERIOD OF RECORDS	FT MBL RECORDS	LENGTH, MONTHS	PERIOD OF RECORDS	FT MBL RECORDS	LENGTH, MONTHS
GAINESVILLE UNIVERSITY OF FLA							
29-39-00	17	01/1903 - 04/1905	277	28	82-21-00	2982-424-224	08-3316
		06/1905 - 06/1928	425	425			
		08/1928 - 12/1963					
GAINESVILLE 2 MBL							
29-38-00	92	10/1953 - 10/1954	13	4	82-22-00	2982-424-222	08-3321
		12/1954 - 03/1955	4	4			
		05/1955 - 10/1955	6	6			
		12/1955 - 12/1955	1	1			
		02/1956 - 05/1956	4	4			
		07/1956 - 10/1956	4	4			
		12/1956 - 02/1957	3	3			
		05/1957 - 12/1979	272	272			
GAINESVILLE FAA AP							
29-42-00	14	05/1960 - 12/1969	116	116	82-16-00	2982-424-412	08-3326
HIGH SPRINGS							
29-50-00	65	07/1948 - 03/1958	117	117	82-36-00	2982-341-324	08-3956
		05/1958 - 02/1971	154	154			
		04/1971 - 10/1972	19	19			
		12/1972 - 12/1979	83	83			
ISLAND GROVE							
29-27-00	74	10/1955 - 06/1963	93	93	82-06-00	2982-144-321	08-4327
		08/1963 - 12/1979	197	197			
MELROSE							
29-43-00	15	11/1929 - 03/1961	19	19	82-03-00	2982-414-423	08-5622
		08/1961 - 02/1969	91	91			

TABLE C-4. SUMMARY OF AVAILABLE EVAPORATION DATA (CONCLUDED)

EVAPORATION STATIONS

MELROSE	LATITUDE 29-43-00	LONGITUDE 82-03-00	STATION NO. 08-5622
	ELEVATION 15 FT MSL	REGION LENGTH, MONTHS	2982-414-423
	PERIOD OF RECORDS 02/1963 - 02/1963	1	ALACHUA GEOGRAPHIC LOCATION BLOCK
			BASIN

TABLE C-6. PROGRAM TO EXTRACT THE REQUIRED DATA

```

0000 //DFM JOB (2006,3401,2,2,0), 'D.F.MACINTYRE', CLASS=2
0001 //♦PASSWORD
0002 //♦ROUTE PRINT REMOTE1
0003 //♦SETUP          3330,CROP01
0004 // EXEC RAIN
0005 ACCESS
0006 ELEMENT          RAIN
0007 STATION          083321
0008 PERIOD           1/1970 TO 12/1979
0009 LIST             MONTHLY
0010 ACCESS
0011 ELEMENT          RAIN
0012 STATION          083321
0013 PERIOD           1/1979 TO 12/1979
0014 LIST             DAILY
0015 // EXEC EVAP
0016 ACCESS
0017 ELEMENT          EVAPORATION
0018 STATION          083321
0019 PERIOD           1/1970 TO 12/1979
0020 LIST             MONTHLY
0021 // EXEC TEMP
0022 ACCESS
0023 ELEMENT          TEMPERATURE
0024 STATION          083321
0025 PERIOD           1/1970 TO 12/1979
0026 LIST             MONTHLY
0027 //♦ EDJ
END OF WORK FILE

```


TABLE C-7. LISTING OF MONTHLY PRECIPITATION FROM 1970 THROUGH 1979

GAINESVILLE 2 WSW	ALACHUA												ANNUAL	Σ AVE
	JANUARY	FEBRUARY	MARCH	APRIL	MAY	JUNE	JULY	AUGUST	SEPTEMBER	OCTOBER	NOVEMBER	DECEMBER		
1970	4.32	7.92	7.55	2.98	6.93	6.53	7.24	10.80	1.87	1.97	0.13	2.13	40.53	114
1971	3.08	3.81	7.91	2.89	8.78	3.52	5.27	13.49	1.12	4.49	7.27	3.10	50.24	96
1972	4.16	4.23	4.14	2.74	3.97	10.72	3.14	13.49	1.12	0.99	7.27	4.15	47.78	130
1973	0.21	2.38	2.29	1.17	3.48	10.05	7.95	7.20	2.84	0.91	0.84	3.13	50.60	97
1974	3.11	4.23	0.99	2.21	3.01	4.87	4.45	4.41	12.55	2.01	1.03	4.03	50.51	99
1975	1.20	1.49	1.42	3.19	4.45	1.57	4.94	2.84	3.73	2.21	2.78	3.97	51.40	92
1976	3.20	4.16	1.22	0.83	0.45	3.26	1.44	7.10	3.73	0.12	1.95	4.94	48.11	84
1977	4.20	4.98	4.22	0.64	3.45	3.90	10.24	9.14	5.73	0.47	1.00	4.73	33.36	84
1978	8.59	2.34	1.17	8.18	3.36	4.55	4.29	7.59	12.23	0.11	1.32	4.73	47.82	115
1979	3.93	3.92	3.90	3.08	4.85	6.41	5.75	7.80	5.33	1.48	1.84	4.30		
MEAN OF MONTHS	10	10	10	10	10	10	10	10	10	10	10	10		
PERCENT ANNUAL	7.6	7.6	6.8	9.9	9.3	12.3	11.1	19.0	10.3	2.9	3.6	8.3		
	MEAN ANNUAL PRECIPITATION 52.21 INCHES													

PAGE 1
STATION NO. 08-3321

TABLE C-8. LISTING OF DAILY PRECIPITATION DURING 1979

GAINESVILLE 2 MB		ALACHUA											
		DAILY PRECIPITATION IN INCHES											
1979		JANUARY	FEBRUARY	MARCH	APRIL	MAY	JUNE	JULY	AUGUST	SEPTEMBER	OCTOBER	NOVEMBER	DECEMBER
1	1.35			T		1.15	0.81	0.24	0.40	0.12	0.08	0.53	
2					3.23			0.01	T	0.36			
3				0.85				0.09		1.89			1.05
4	0.21	0.04	0.04	T					0.36	0.44			4.26
5	0.81	0.88	0.04			0.25		0.12	0.89	0.67		0.01	
6					0.98	0.26		T	1.95	0.03		0.43	
7				0.11	0.17		0.41	0.34	1.00	0.31	0.02	0.05	
8								0.06	0.35	0.44			
9	1.12				0.92	0.12		0.10		0.22			0.01
10	0.86					0.12	0.07	0.49		0.31			0.01
11						0.12	0.01	0.49		0.22			0.01
12								0.08		1.09			0.01
13								0.08	0.01	T			0.01
14								0.49	0.01	T			0.01
15								0.01	0.01	T			0.01
16								0.01	0.01	T			0.01
17								0.01	0.01	T			0.01
18								0.01	0.01	T			0.01
19								0.01	0.01	T			0.01
20	2.32		0.04					0.01	0.01	T			0.01
21								0.01	0.01	T			0.01
22								0.01	0.01	T			0.01
23	0.94			0.17				0.01	0.01	T			0.01
24								0.01	0.01	T			0.01
25								0.01	0.01	T			0.01
26	0.60							0.01	0.01	T			0.01
27								0.01	0.01	T			0.01
28								0.01	0.01	T			0.01
29	0.48							0.01	0.01	T			0.01
30	8.69		2.34	1.17	8.18	3.36	4.55	4.39	7.39	12.23	0.11	1.32	6.09
31													
TOTAL													

program similar to that in Table C-6, using the "COPY" command to extract the required data and transfer them onto a disk file or computer tape. The disk file or computer tape could then be used as an input file for the other computer programs. As described in Section IV, it is proposed that the Air Force system automatically perform the extraction and data reformatting; the above remarks pertain to the effort currently necessary on the University of Florida HISARS system.

D. REFERENCES

- C-1. Wiser, H.E., HISARS, Hydrologic Information Storage and Retrieval System Reference Manual, North Carolina Agricultural Experiment Station Bull. 215, North Carolina State University, Raleigh, NC, 1975.
- C-2. Portier, K.M., A Guide to the Use of HISARS, a Hydrologic Information Storage and Retrieval System, Department of Statistics, IFAS, University of Florida, Gainesville, FL, October 1981.

FILME