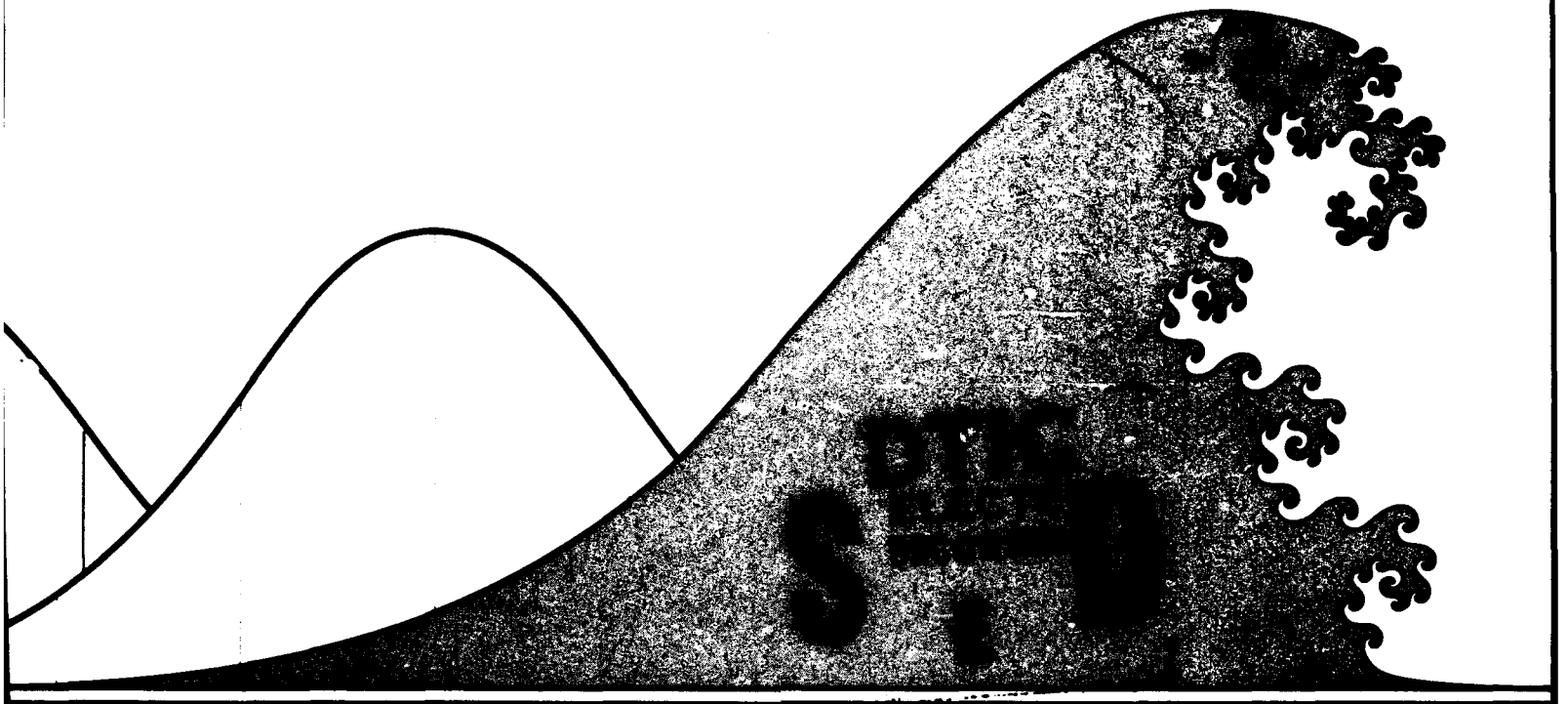
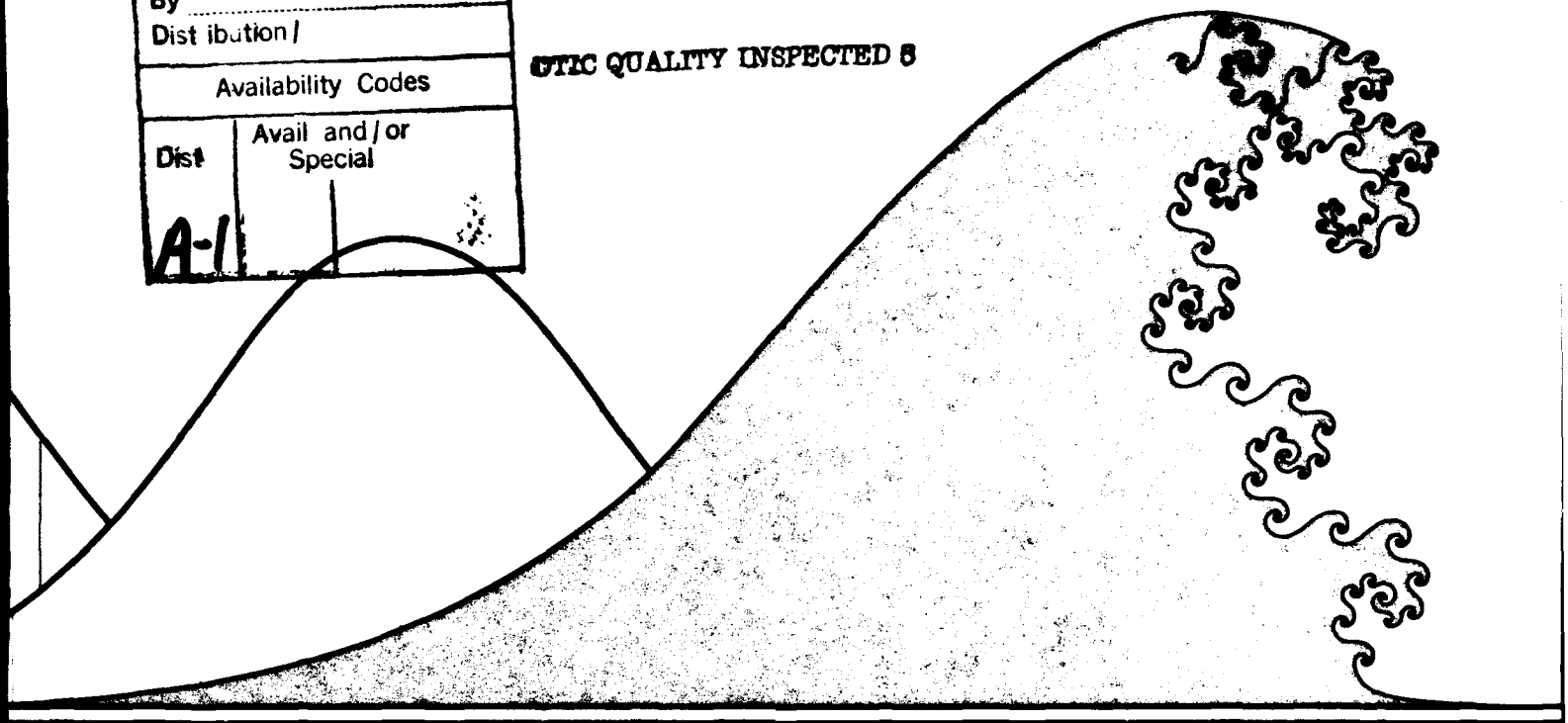
'Aha Huliko'a



STATISTICAL METHODS IN PHYSICAL OCEANOGRAPHY

Proceedings
Hawaiian Winter Workshop
University of Hawaii at Manoa
January 12–15, 1993

'Aha Huliko'a



STATISTICAL METHODS IN PHYSICAL OCEANOGRAPHY

PROCEEDINGS
'Aha Huliko'a
Hawaiian Winter Workshop
University of Hawaii at Manoa
January 12–15, 1993

PETER MÜLLER
DIANE HENDERSON
editors

Sponsored by the U.S. Office of Naval Research,
the School of Ocean and Earth Science and Technology,
and the Department of Oceanography, University of Hawaii

SOEST Special Publication • 1993

Cover, title page design by Brooks Bays, SOEST Publication Services

FOREWORD

The seventh 'Aha Huliko'a† Hawaiian Winter Workshop was held January 12–5, 1993 at the East-West Center in Honolulu, Hawaii. The topic was "Statistical Methods in Physical Oceanography."

Physical oceanographers deal with randomness and uncertainties when analyzing ocean data and formulating ocean models. They apply concepts and results from probability theory, statistical inference and stochastic processes. The size and complexity of oceanographic problems often prevent the application of standard methods, and physical oceanographers are faced with the task of inventing special methods that deal with the peculiarities of their problems in a sensible way. These special methods were the object of the workshop's lectures and discussions. The lectures are published in these proceeding. The order of the papers follows loosely the agenda of the workshop covering a variety of oceanographic observations, methods for efficient flow and data representation, frequentist versus Bayesian inference, data assimilation, and idealized dynamics. Also included is a summary of the meeting.

The workshop, made possible by a grant from the U.S. Office of Naval Research, was hosted by the Department of Oceanography of the School of Ocean and Earth Science and Technology of the University of Hawaii. The excellent facilities of the East-West Center and the capable staff directed by James McMahon contributed greatly to the success of the meeting. The local organization and logistical arrangements were expertly handled by Phyllis Haines. This proceedings volume came into existence through the creative and dedicated research of the scientists who gathered in Hawaii and provided the articles that follow. Barbara Jones and May Izumi provided skillful production assistance.

Peter Müller
Diane Henderson

Department of Oceanography
School of Ocean and Earth Science and Technology
1000 Pope Road
University of Hawaii
Honolulu, HI 96822

† 'Aha Huliko'a is a Hawaiian phrase meaning an assembly that seeks into the depth of a matter.

This work is related to Department of the Navy Grant N00014-93-1-0156 issued by the Office of Naval Research. The U.S. Government has a royalty-free license throughout the world in all copyrightable material herein.

PARTICIPANTS

(left to right) Peter Müller, Neil Frazer, Joerg Wolff, Boris Galperin, Dennis Moore, Andrew Lorenc, James O'Brien, Dudley Chelton, Jin von Storch, Rob Pinkel, Claude Frankignoul, Jens Schröter, Julio Ottino, Rudolf Kloosterziel, Hans von Storch, Skip Carter, Geoff Vallis, George Casella, Wojbor Woyczynski, Ann Gargett, Alfred Osborne, Joseph Kadane, Marie Farge, Michael Brown, Toshio Mike Chin, Greg Holloway, Yukio Kaneda, Doug Luther, and Phyllis Haines. Eric Firing and Gary Mitchum are not shown.

Table of Contents

Foreword iii

Participants v

Measurement and Analysis of the Energy-Containing Eddies of Turbulent Flows in the Coastal Ocean
Ann E. Gargett 1

Finescale Shear and Strain in the Thermocline
Robert Pinkel and Steve Anderson 17

Structure of the Upper Ocean Velocity Field on Scales Larger than 10 Kilometers
Eric Firing 37

Satellite Altimetry: Attempts to Progress Beyond Studies of the Statistics of Mesoscale Variability
Dudley B. Chelton and Michael G. Schlax 55

Observing "Integrating" Variables in the Ocean
Douglas S. Luther and Alan D. Chave 103

Wavelets and Wavelet Packets to Analyze, Filter, and Compress Two-Dimensional Turbulent Flows
Marie Farge, Eric Goirand, and Thierry Philipovitch 131

The Numerical Inverse Scattering Transform: Nonlinear Fourier Analysis and Nonlinear Filtering of Oceanic Surface Waves
A.R. Osborne 161

Principal Component Analysis: Basic Methods and Extensions
Gary T. Mitchum 185

Principal Oscillation Pattern Analysis of the Intraseasonal Variability in the Equatorial Pacific Ocean
Hans von Storch 201

Illustrating Frequentist and Bayesian Statistics in Oceanography
George Casella 229

Bayesian Methods: An Introduction for Physical Oceanographers
Joseph B. Kadane 241

A Bayesian Approach to Observation Quality Control in Variational and Statistical Assimilation
Andrew C. Lorenc 249

Optimal Space-Time Interpolation of Gappy Frontal Position Data
Toshio M. Chin and Arthur J. Mariano 265

Some Notes on Data Assimilation in Physical Oceanography
James J. O'Brien 291

Estimation of Free Parameters by Inverse Modeling
Jens Schröter 303

An Adaptive Inverse Method for Model Tuning and Testing
Claude Frankignoul, Nathalie Scoffier, and Mark A. Cane 331

New Developments in Stirring and Chaos: Possible Role in Ocean Science
Julio M. Ottino 351

Chaos in Ocean Physics
Michael G. Brown 373

Measurements of Chaos in the Ocean
Everett F. Carter, Jr. 381

Geometric Thermodynamics as a Tool for
Analysis and Prediction in Oceanography
Nessan Fitzmaurice, Wojbor Woyczynski, and Anatoly Odulo 397

Non-Eddy Resolving Modeling of β-Plane Turbulence
Boris Galperin, Semion Sukoriansky,
Steven A. Orszag, and Ilya Staroselsky 421

Frequency Shift of Rossby Waves in β-Plane Turbulence
Yukio Kaneda and Greg Holloway 453

Statistical Mechanics, Turbulence, and Ocean Currents
Geoffrey K. Vallis 473

Overview of Statistical Mechanics,
with Practical Application for Ocean Modeling
Greg Holloway 491

Experiments with a Hybrid Statistical Mechanics/
Ocean Circulation Model
Michael Eby and Greg Holloway 507

Statistical Methods in Physical Oceanography: Meeting Report
Peter Müller and Greg Holloway 519

MEASUREMENT AND ANALYSIS OF THE ENERGY-CONTAINING EDDIES OF TURBULENT FLOWS IN THE COASTAL OCEAN

Ann E. Gargett
Institute of Ocean Sciences, Sidney, B.C. Canada

ABSTRACT

Acoustic remote sensing techniques now allow measurement of the three-dimensional velocity field associated with the large-scale eddies of turbulent geophysical flows in the coastal ocean. Such techniques, continuous in time and requiring a minimum of technical supervision, are essential for assessment of turbulent coastal regimes, because of short space and time scales of variability. Algorithms under development should provide estimates of kinetic energy E, length scales, and kinetic energy dissipation rate ϵ of the turbulence, as well as the shear dU/dz of the mean flow. Recent addition of a towed CTD allows a direct measurement of buoyancy flux $\overline{\rho' w}$, a major goal of ocean microscale measurements over the last two decades. Preliminary data are available to compare this direct measurement with the widely used estimate $\overline{\rho' w} = 0.2\rho_0 g^{-1}\epsilon$, made from measurements of dissipation rate.

1. AN ACOUSTIC REMOTE SENSING TOOL FOR TURBULENCE RESEARCH

While shipborne acoustic Doppler current profilers (ADCPs) have been widely used for measuring "mean" currents in the surface layers of the ocean, use of a commercial ADCP for turbulence research required modification to both hardware and software. The hardware modification was to rotate one of the four beams of a standard Janus-configuration transducer head to vertical, leaving the other three beams at the normal (30°) slant angle from vertical. When mounted on a ship (Fig. 1), this beam (B3) is closely adjusted to vertical ($\pm 0.5°$), allowing a direct and unequivocal measurement of vertical velocity w. A combination of B4 and B3 provides an estimate of across-ship velocity component u, while a combination of B2 and B3 (or of B1 and B3) provides the along-ship component v. These horizontal velocity components can be affected by the slant-beam configuration, so this account will mostly use the straightforward measurement of w. Direct shipborne measurement of w is possible with incoherent Doppler systems because coastal turbulence is vigorous, and because the inner coastal waters of British Columbia, in which these data were taken, provide low levels of platform motion contamination. If a stable platform can be provided, however, recent development of more accurate coded-pulse Doppler systems suggests that the techniques discussed here will soon be extensible to the deep ocean.

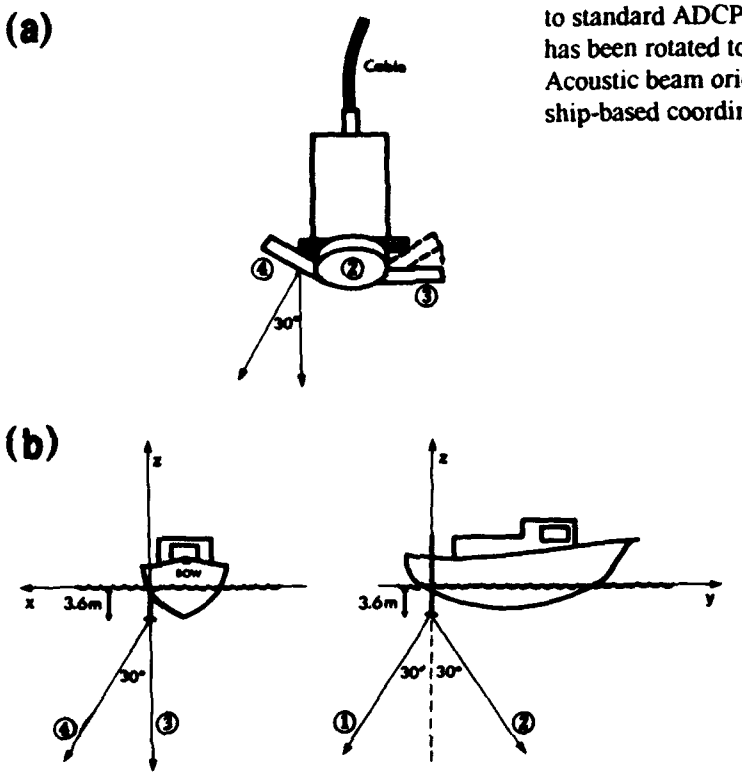


Figure 1. (a) Hardware modification to standard ADCP: one beam (B3) has been rotated to vertical. (b) Acoustic beam orientation relative to ship-based coordinates (x, y, z).

Special acquisition software was written to allow recording of raw (single-ping) beam velocities and acoustic amplitudes. After each ping, the processor associated with a single beam returns time (radial distance)-binned estimates of radial velocity, defined positive when the velocity is towards the transducer, and a measure of the strength of the return signal. In acquisition mode, both fields are recorded for all four beams, while up to four fields can be selected for colour-coding and real-time display. At present, we use amplitude signal only from the vertical beam, in order to locate the bottom (or lack of it) in the velocity records; subsequent processing uses only the water column velocities.

Single-ping velocity data are noisy. Figure 2a is a (poor) rendition of raw data from a turbulent tidal front. [An apology: Grey-scale rendering of *signed* quantities such as velocity is difficult, but must be attempted when colour graphics are not available. For presentation in this paper, I have chosen to bin the data very coarsely, effectively grey-scale 'contouring' the fields. With such coarse-binning, it is possible to use a symmetric grey-scale that differs only in the textures assigned to the bins nearest zero (center): thus in Figure 2, a maximum (black) that occurs as a progression through light grey (small circles) is a maximum downwards (upwards) w. While this presentation works reasonably well with smooth fields, it does a very poor job of the original noisy raw data in Figure 2a.] The standard technique for reducing the noise level of Doppler velocity estimates is to average values from consecutive pings: Figure 2b illustrates this technique, using an

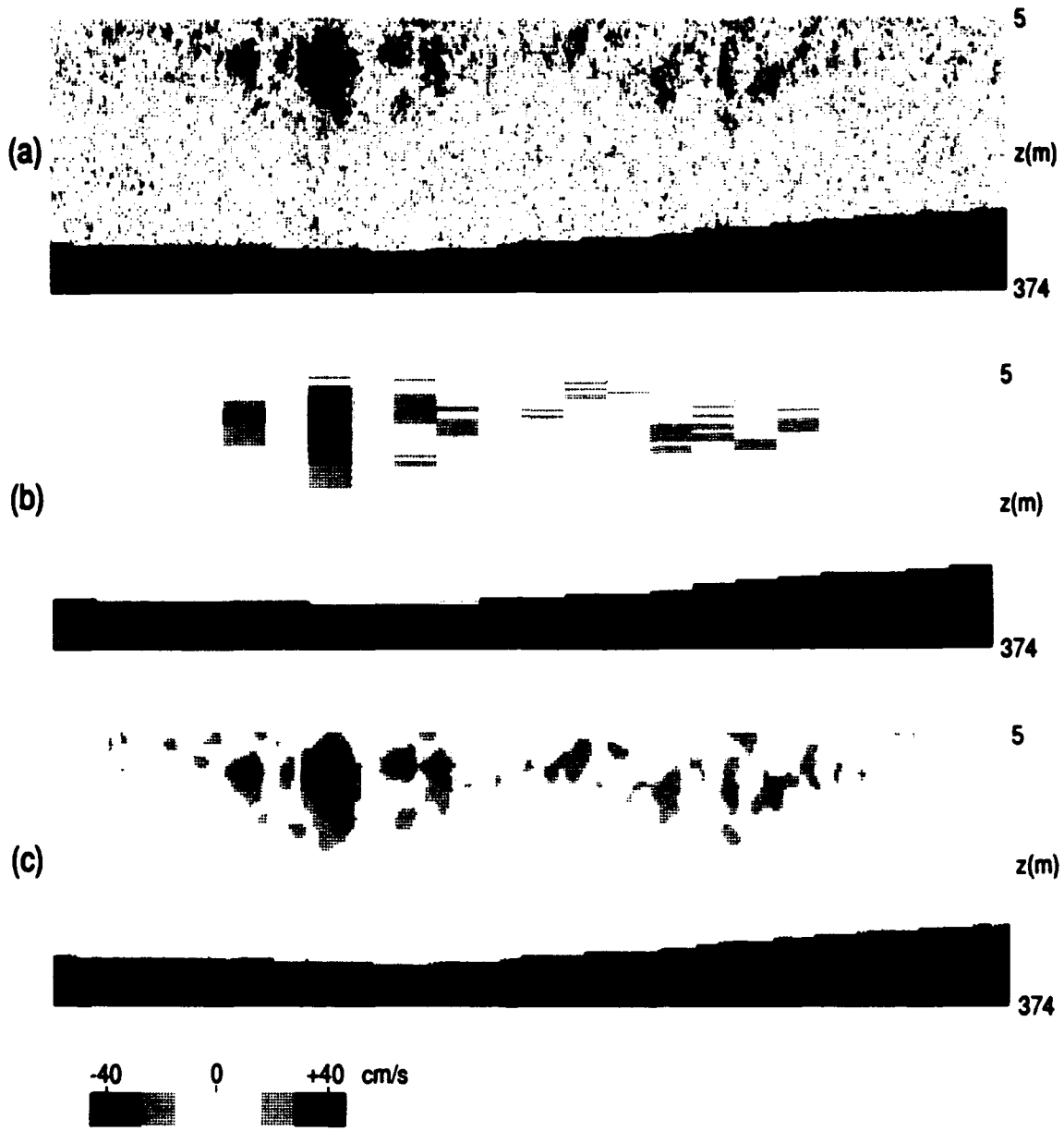


Figure 2. Grey-scale coded representation of the (signed) field of w as measured by the vertical beam: (a) Raw (single-ping) data, noise standard deviation $\sigma \simeq 10$ cm/s. (b) Standard ADCP processing required to produce $\sigma \simeq 2$ cm/s (boxcar average over 25 pings) fails to resolve the spatial structure of the turbulent flow in this tidal front. (c) Filtering with a sequential running mean filter yields $\sigma \simeq 2$ cm/s with spatial resolution of about 20 m.

average of 25 pings to reduce noise standard deviation from 10 cm/s to 2 cm/s. With post-processing, this brute strength technique, which severely degrades much of the spatial structure that is present, is easily replaced by a filter that produces the same 2 cm/s standard deviation, but retains spatial structure down to horizontal wavelengths of order 20 m (Fig. 2c).

2. WHY DO WE NEED THE VERTICAL BEAM?

While significant vertical velocities do not guarantee that a flow is turbulent, flows are *not* turbulent without significant vertical velocities. In survey mode, we may thus look for large vertical velocity as a necessary condition for turbulence. Having found this condition, such flows may be subject to more rigorous scrutiny with regard to characteristics—for example relative "eddy" and internal wave time scales, vertical buoyancy flux, phase between w and fluctuation density—which we associate with turbulence. Thus accurate measurement of the vertical velocity field is essential to turbulence measurement.

With a standard ADCP, velocity components are calculated under the assumption that the velocity field is uniform over the spread of slant beam pairs (Fig. 3a). If this *is* the case, the horizontal component v in the plane of B1 and B2 makes contributions of opposite sign to the beam velocities V1 and V2 in bin b; hence slant beam vertical velocity $ws = (V1+V2)/2 \cos 30°$. This slant-beam vertical velocity is shown in Figure 3c, below the field of w measured directly by the vertical beam (Fig. 3b) for a section of data from a tidal front. The obvious differences between w and ws are caused by the fact that the turbulent field has spatial scales that are comparable to the slant beam spread.

Scatter plots of ws vs w (Fig. 4) show that while $ws \simeq w$ at shallow depths (a), the correlation decreases with increasing depth (b); By the deepest bins (c), ws is essentially uncorrelated with w, although both remain significantly above the noise level, shown in (d). This must be expected to be a normal state of affairs in coastal waters, where the water depth H sets a maximum outer scale for turbulent eddies (the actual outer scale may be even smaller, because of conditions of shear or stratification). With the 30° angle of the standard slant beam pairs, slant beam separation at depth H is H, i.e., the scale at or below which we expect turbulent energy to reside. Accurate measurement of the vertical velocity field in coastal areas thus clearly requires the special vertical beam that is part of the DOppler Turbulence system (DOT).

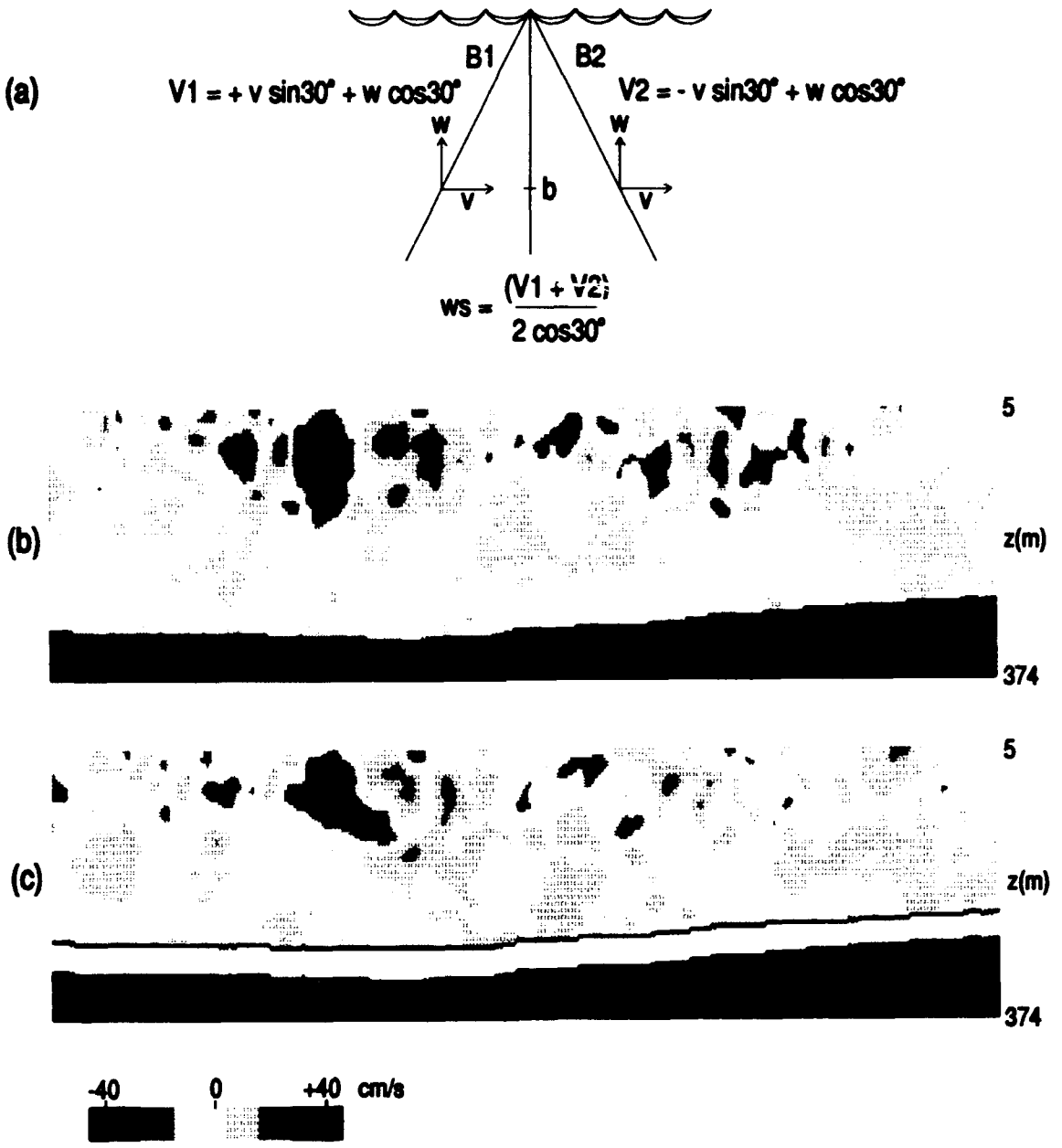


Figure 3. (a) Accurate determination of *w* from two slant beams (B1 and B2) requires that the velocity field be uniform over the (increasing with depth) horizontal spread between the beams. Fields of (b) *w* from B3 and (c) *ws* from B1 and B2 differ considerably in this tidal front, suggesting that this requirement is not met.

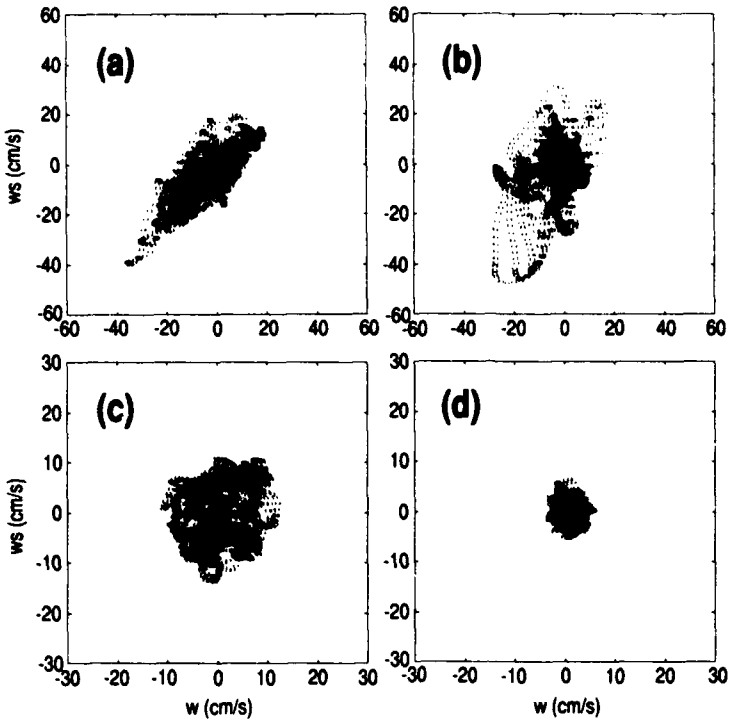


Figure 4. Scatter plots of *ws*, vertical velocity determined from the paired slant beams B1 and B2, versus the "true" *w* measured from B3, for various depths (a) 23 m, (b) 112 m, and (c) 201 m; (d) is noise level, taken at slack tide in a sheltered location. Near the transducer, the two variables are correlated, but as depth (slant beam separation) increases, *ws* and *w* become increasingly uncorrelated.

3. AN ALBUM OF COASTAL MIXING

With the shipborne, semi-automated system described above, it is possible to survey coastal waters for locations and processes that cause significant turbulence. Our experience is that most intense turbulence is associated in some way with flow geometry such as submarine sills, horizontal channel constriction, or sharp changes in channel direction. Coastal turbulence varies rapidly in time, since it is driven predominantly by the tides and is clearly modulated on the neap/spring cycle.

Figure 5 is a sampler of the kind of mixing regimes found in B.C. coastal waters. The depth range of the measurements vary, as marked; the horizontal scale is ~1100 m. In the upper panel (a) is a record taken in mid-winter at a time of minimum water column stratification. The tide floods from left to right over a sharp submarine sill that nearly blocks a tidal channel located in the southern Strait of Georgia. Water descends the downstream side of the sill with vertical velocity near 1 m/s; the subsequent flow exhibits intense fluctuations of vertical velocity far downstream. The centre panel (b) is another situation in which the tide floods from left to right across a sill; this however is a silled, fjord-type inlet, at a time of very strong near-surface density stratification. Whether because of this stratification "cap" or because of the gentler sill relief, dense water from outside the sill is found entering the inlet on the flood as a bottom boundary current, most visible in the vertical velocity field at those places where it accelerates downwards with increases in bottom slope. A final example in Figure 5c shows a turbulent surface jet

flowing (left to right) out of a narrow and shallow tidal passage. Water exiting the passage is well-mixed and lighter than the deeper water outside, hence flows out at the surface. Abrupt increase in channel width causes rapid shallowing of the jet just outside the channel mouth.

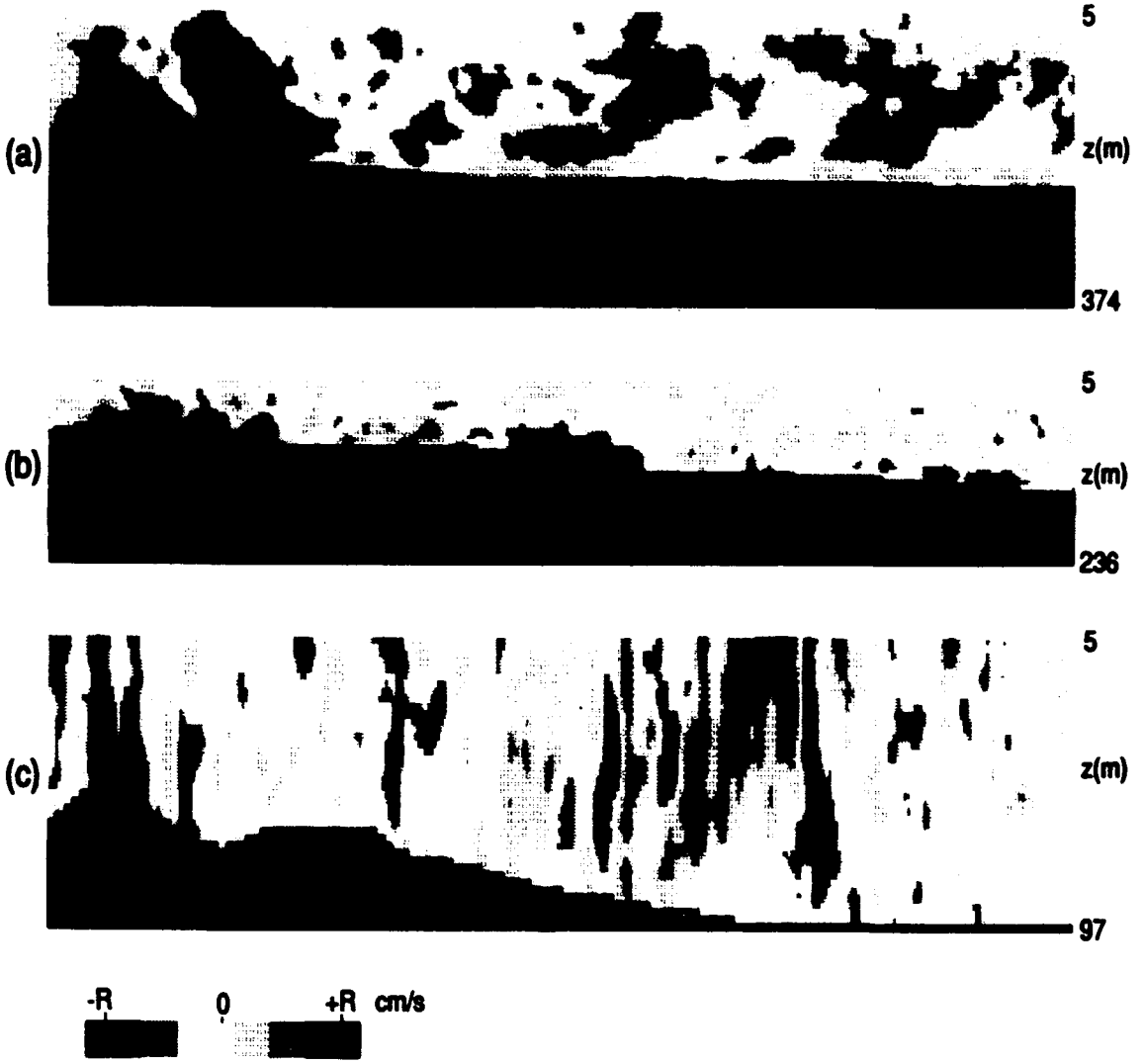


Figure 5. A variety of flows generate turbulence in the coastal ocean; the associated *w* fields are displayed in grey-scale. Recall that the sign of *w* maxima (black) may be determined by the surrounding pattern, light grey for downwards, circles for upwards vertical velocity. In all cases the mean horizontal tidal flow is from left to right. (a) weakly stratified flow over a sill: R=50 cm/s, (b) strongly stratified inflow to a coastal fjord: R=20 cm/s, (c) a "jet" of well-mixed fluid out of a narrow tidal channel: R=20 cm/s.

4. ESTIMATION OF TURBULENCE QUANTITIES

What properties of turbulence would we like to know? - turbulent kinetic energy E, the rate ϵ at which it is being dissipated, and the associated vertical fluxes of mass and momentum, are some that spring to mind. The DOT system, augmented by sporadic vertical profiles of density, should offer information in nearly all of these areas.

Turbulent kinetic energy:

The definition of turbulent kinetic energy per unit mass as $E = 1/2\ \overline{(u^2 + v^2 + w^2)}$ uses the components (u, v, w) of the turbulent velocity $\underline{u}$, itself defined as the (zero-mean) part left after removal of a "mean" velocity $\underline{U} = (U, V, W = 0)$ (where U and V are normally assumed to be functions of z only) from the total velocity $\underline{u}_T$. Inherent in this so-called Reynolds decomposition of the flow is an appropriate definition of the averaging process that defines the "mean" flow. While the assumption that $W = 0$ seems safe, it is difficult to decide how to form a "mean" horizontal component in situations where the flow is substantially inhomogeneous. The problem is illustrated in the record of Figure 6 which shows (a) the horizontal velocity component v (relative to the ship) along the axis of a tidal channel and (b) vb, the baroclinic part of this field, formed by removing the local depth-average of v. At the beginning (left) of this record, vb has a three-layer structure, with surface and bottom layers moving more rapidly than a mid-depth layer. By the end (right) of this section of record, the structure had changed to bottom-intensified two-layer flow. It is not at all clear what horizontal scale should be chosen for calculating a "mean" horizontal velocity component V, nor how that scale should change with time (horizontal distance).

Because of this uncertainty as to the appropriate averaging for the horizontal "mean" components, the cleanest definition of E would seem to be $E_i = 3/2\ \overline{(w^2)}$, where the overbar denotes an averaging length such that $\overline{w} = 0$, and the subscript is a reminder that this is an isotropic estimate, obtained from the vertical velocity component only.

Turbulent kinetic energy dissipation rate ϵ:

Also of interest is the rate at which mean flow energy is being removed to dissipation scales by the action of the turbulence. The possibility of remote measurement of this quantity has its roots in the work of Batchelor and Townsend (1948), who showed that the large scale eddies of turbulence lose their energy to the turbulent energy cascade (Kolmogoroff, 1941) within at most a few eddy turnover times. Since energy that enters the cascade is delivered to dissipation scales, this means that

$$\epsilon \sim \frac{rw^2}{\tau} \sim \frac{rw^3}{\ell} \qquad (1)$$

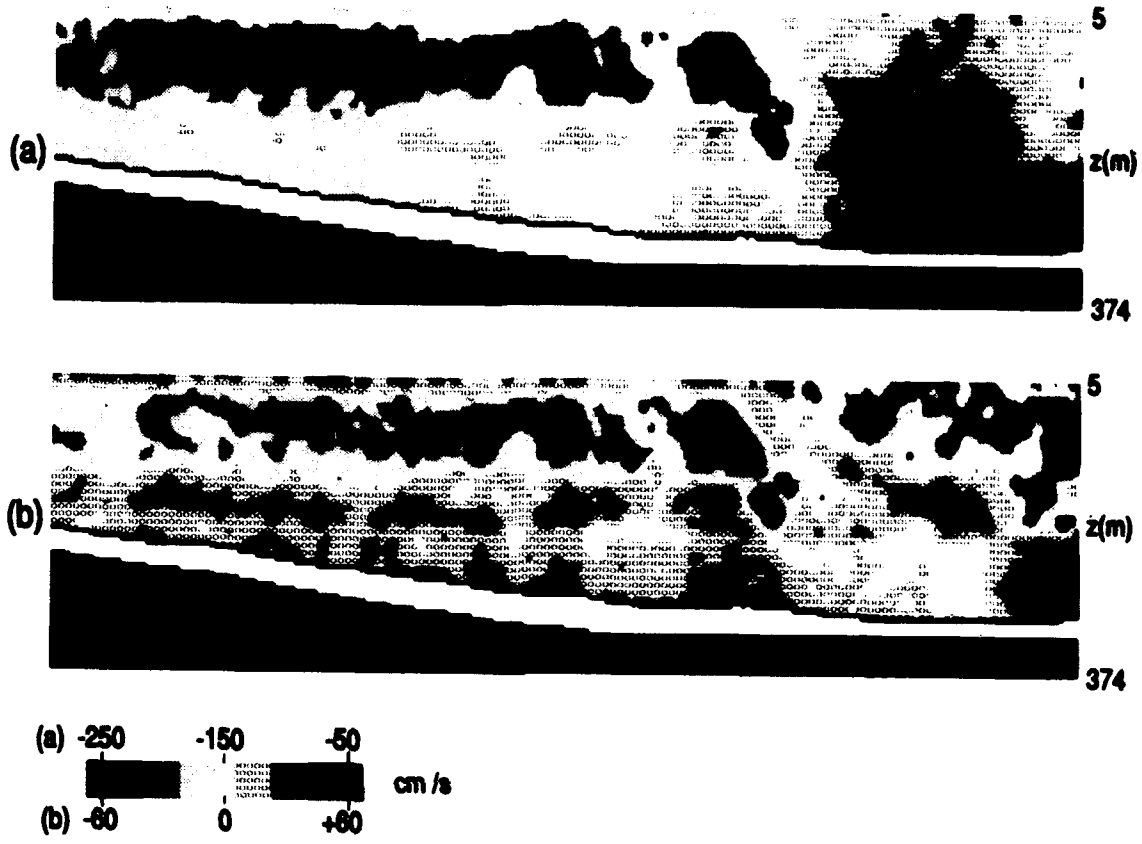


Figure 6. (a) Field of horizontal velocity v relative to the ship (determined from the fore-aft slant beam pair B1,B2) as the ship moves along the axis of a tidal channel. Variations in ship speed and/or the barotropic field are removed in (b) the baroclinic field $vb = v - \langle v \rangle$ where $\langle v \rangle$ is the (local) depth-averaged value. The strongly inhomogeneous nature of the horizontal flow makes calculation of horizontal turbulent velocity components difficult.

where $\tau \sim \ell/rw$ is the turnover time of an eddy of scale ℓ and rms turbulent vertical velocity rw. This is only a scale relationship, leaving an unknown constant to be determined. Direct measurements of ϵ, rw, and ℓ from the atmospheric boundary layer have confirmed the relation (1) above, and suggest that the constant involved is between 3 and 5 (Wamser and Müller, 1977).

Thus for both E_i and ϵ estimates, it is necessary to derive values for rw, an rms velocity typical of the energy-containing eddies of the turbulent field; for ϵ, we need in addition a value for the characteristic length scale of such eddies. Meteorologists identify the turbulent length scale ℓ as the location of the peak of a spectrum of vertical velocity as a function of horizontal wavenumber, the turbulent velocity scale rw as the square root of the spectral integral, a procedure that makes sense in view of the long and homogeneous records that can be obtained from meteorological towers. Unfortunately, the marked inhomogeneity of the turbulent fields in coastal waters means that "a" wavelength doesn't remain constant over the large number of wavelengths necessary for its determination by

remain constant over the large number of wavelengths necessary for its determination by such a Fourier technique. Wavelet analysis (Farge, this volume) may offer a more sophisticated means of determining local wavelength and energy values, but for now, I have used a very simple algorithm, shown schematically in Figure 7. The curve is that of vertical velocity w, measured at constant depth (bin), as a function of horizontal

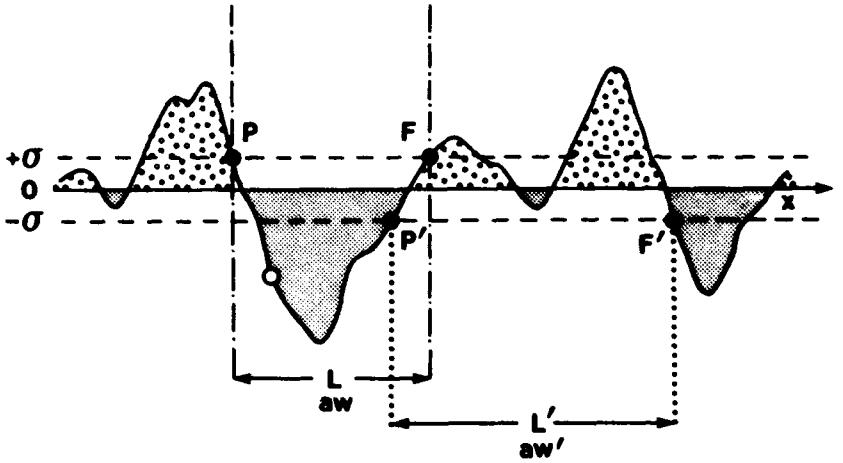


Figure 7. Schematic of a simple algorithm for determining *local* values of large-eddy turbulence parameters (half-wavelength average vertical velocity and length scale) needed for remote estimate of ϵ: for details, see text.

distance x: horizontal dashed lines denote $\pm\sigma$, one standard deviation of the measurement noise level about the zero mean. Starting with a point (say that marked by the open circle) where $|w| > \sigma$, the algorithm searches for locations of the nearest preceding and following points with $|w| > \sigma$ *but of the opposite sign* (respectively P and F in Fig. 7). The distance L between these points is taken as a local estimate of a half-wavelength. The average of w over L, denoted aw, is similarly considered to be the average of w over a half-wavelength. One then moves to point F and repeats the process, resulting in new estimates L' and aw'. These local estimates are assigned to the region over which they are calculated; in the (usually small) regions of overlap, the first (in space/time) estimates are arbitrarily chosen. Figure 8b shows the field of aw that results when this algorithm is applied to the tidal front data of Figure 8a.

Assuming that the other half-wavelength exists (although not necessarily in the plane of measurements), the values of aw are converted to a corresponding root-mean-square value (rw) by the scaling factor (1.11) appropriate for a pure sinusoid, then used with the length scale estimate $\ell = 2L$ (not shown) to form the estimate of ϵ, $e2 \equiv |1.11\ aw|^3 / 2L$, which is shown in logarithmic form in Figure 8c. Note that this estimate of rw can also be used in the estimate $E_i = 3/2(\overline{w^2}) = 3/2\,rw^2$ of turbulent kinetic energy.

How much one may trust such an estimate of ϵ can be determined by comparing it with values determined directly, by integration of the spectrum of small-scale shear measured in situ. Vertical profiles of such direct measurements of ϵ were taken at the two locations

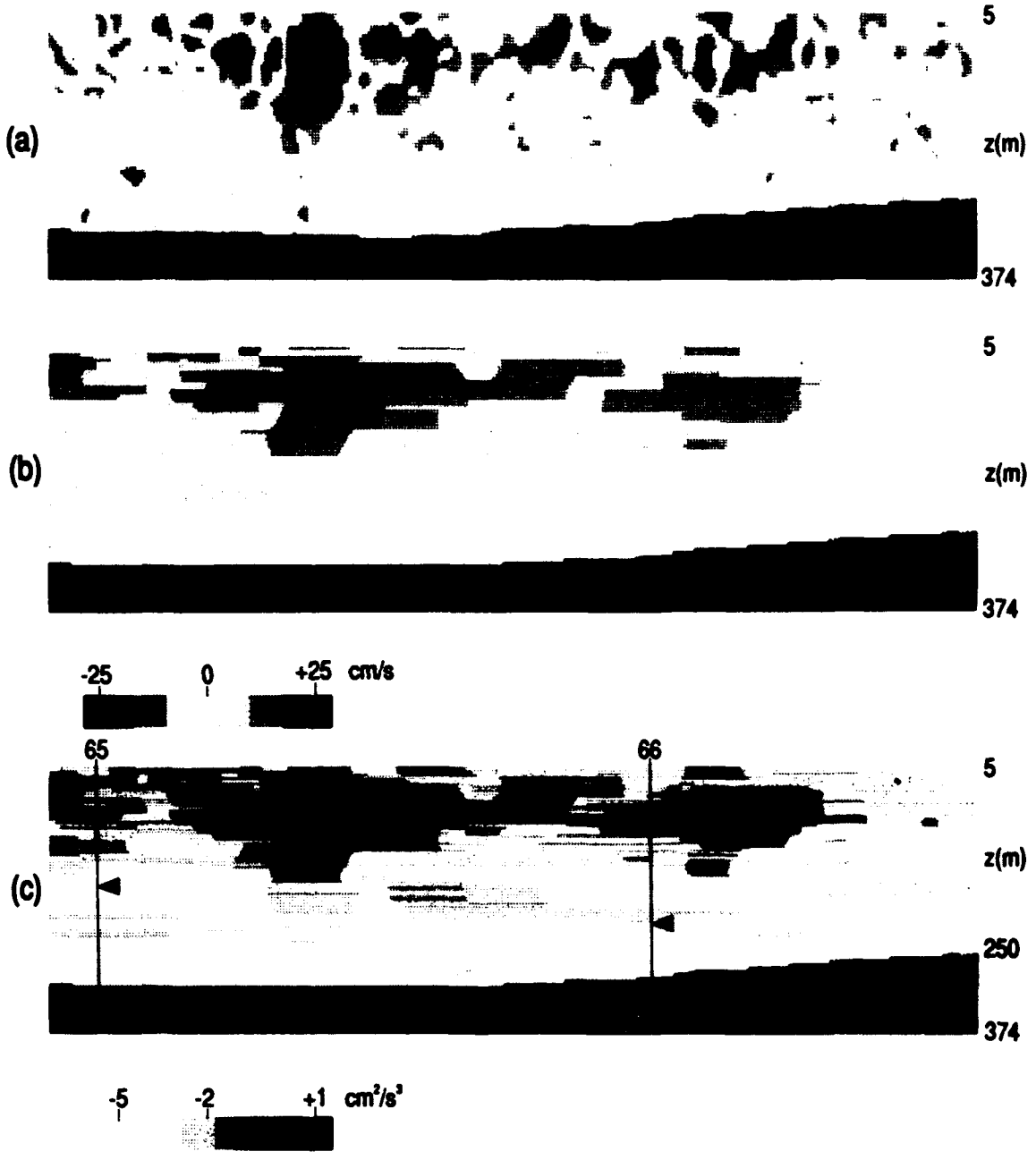


Figure 8. (a) Measured field of w in a tidal front. (b) Associated field of aw derived using the algorithm depicted in Figure 7; aw and L (not shown) can be used to form a field of $e2$, estimated turbulent kinetic energy dissipation rate, shown as $\log(e2)$ in the grey-scale presentation of (c). The vertical lines in (c) denote the launch times of a turbulence microprofiler (operated by Dr. J. Moum, Oregon State University) making direct measurements of ϵ: maximum profile depths are marked by arrows.

marked in Figure 8c by Jim Moum of Oregon State University. Figure 9(a,b) compares the direct profiler measurement (log ϵ, light line) with the indirect estimate log ($e2$) for each profile. The heavy line is the logarithm of the average value of $e2$ over ± 10 pings surrounding the launch of the profiler; the points give some idea of the spread of individual estimates within these 21 pings. The agreement between the two estimates is remarkably good for profile 65. In the subsequent profile, which went somewhat closer to the bottom (about 300 m at both profile locations), we see a defect which tends to recur in many such comparisons; namely a tendency for $e2$ to underestimate ϵ near both the surface and bottom boundaries of the flow. This may indicate the need to modify the definition of turbulent length scale ℓ. Hunt, Stretch and Britter (1988) suggest an alternate form, which tends toward the type of internal scale determined here when the flow is far from boundaries but toward the distance z to the nearest boundary when z is less than this inner scale. Indeed, in measurements taken in the ocean surface layer, Agrawal and Hwang (1991) demonstrate good correspondence between directly measured ϵ and $(rw)^3/\ell$, with $\ell = z$. Such a modification to ℓ, causing length scales to decrease, hence $e2$ to increase near boundaries, would act to correct the discrepancies seen in Figure 9b.

As shown in Figure 9c, however, there *are* profiles in which there remain very large and unsystematic differences between direct measurements and indirect estimates. Indeed, given the high turbulent intensities and spatial/temporal inhomogeneities characteristic of these flows, this seems scarcely surprising. Consider that the profiler is launched from the stern of the ship, at which time and location the w field is assumed "known" from the Doppler. Thereafter the profiler, falling vertically, can be advected horizontally by the local ambient flow, so does not necessarily remain at this geographic launch position. Even if it *were* to remain there, the flow field may change in the time taken for the profile (typically 4-5 minutes for a profile to 300 m). Various checks for the likelihood of time change can be devised, using the fact that the fore/aft slant beams allow two measurements of v that are separated in time, but this is merely an effort to avoid a statistical problem, that of estimating the degree of agreement (or disagreement) necessary before a remote measurement of a non-stationary and inhomogeneous field can be considered "proven" by a relatively sparse set of ground-truth measurements.

Vertical buoyancy (mass) flux:

Part of the reason one might like a remote technique for ϵ is because for the last decade, oceanographers have obtained what they often really *want*, the vertical buoyancy flux $\rho' w$, from what they are able to *get* from microstructure profiler measurements, namely ϵ, and a model (Osborn, 1980) which suggests that under certain assumptions,

$$\overline{\rho' w} = \frac{R_f}{1 - R_f} \rho_0 g^{-1} \epsilon \qquad (2)$$

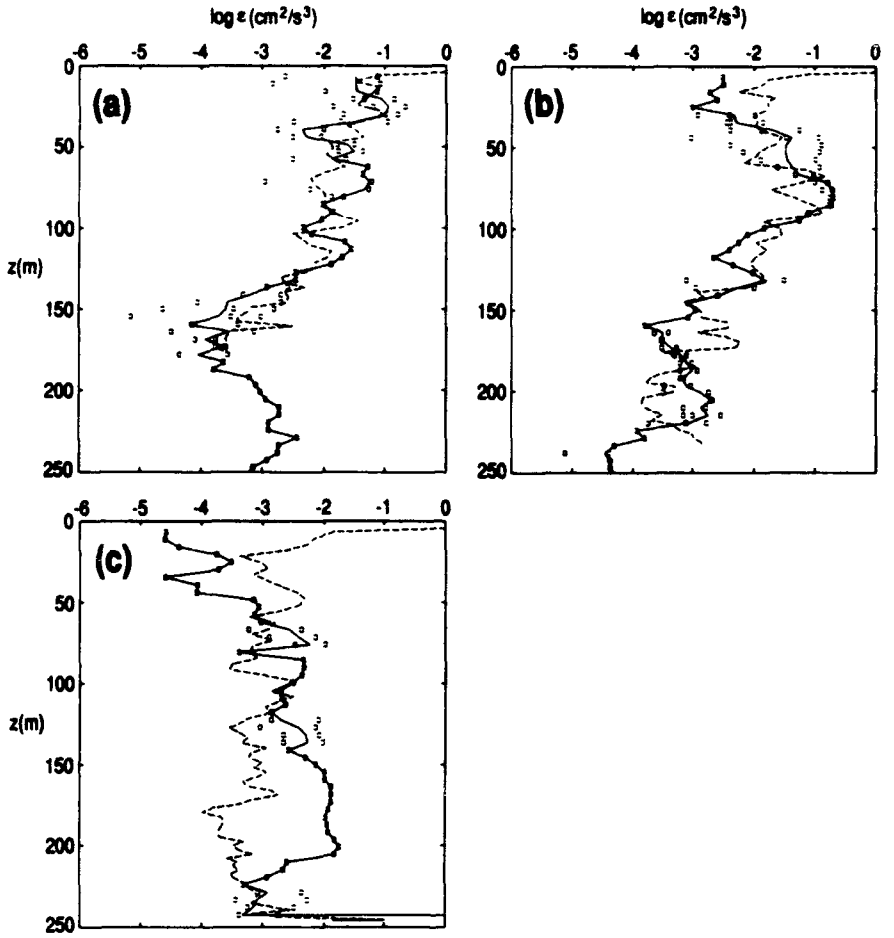


Figure 9. Comparison of log(ϵ) from direct profiler measurements (dashed lines) with the indirect estimates log($e2$) (solid lines) derived from the Doppler w field. The Doppler estimates are averaged over ± 20 pings about the launch position of the profiler (see Figure 8): the individual points give some idea of the variation in the estimates averaged. Profile 65 (a) shows remarkably good agreement, while Profile 66 (b) shows differences near top and bottom boundaries which suggest that the definition of ℓ may need modification in these regions. There remain profiles (c) in which agreement is low.

where R_f, the flux Richardson number, is the ratio of buoyancy sink to shear source terms in the turbulent kinetic energy equation. Oceanographers add the further assumption that $R_f \simeq 0.2$, resulting in an estimate of buoyancy flux as a constant fraction of the measured turbulent kinetic energy dissipation rate ϵ. If correct, this model means that a remote measurement of ϵ would correspond to a remote measurement of buoyancy flux. However, the model has rarely been checked by comparison with direct flux measurements, as these are extremely difficult to make in the ocean environment. The small amount of evidence which does exist (Yamazaki and Osborn, 1993) suggests that R_f is either not constant, or else considerably smaller than 0.2. Vertical turbulent fluxes (or equivalently, turbulent diffusivities) are important products of oceanic microstructure

measurements; it would be nice to know the circumstances (if any) under which such dissipation-based estimates are accurate, hence remote measurement of buoyancy flux would be possible.

With the addition of a towed CTD with finescale resolution (Ocean Sensors), it has proven possible to make statistically significant measurements of buoyancy flux using the DOT system. The CTD is towed at constant depth, just in front of the vertical beam of the Doppler, for long periods. Figure 10 shows the CTD measurement of density, and the associated time series of w measured in the Doppler bin that includes the CTD tow depth, over about three hours. Below is an enlargement of a small section of the record (taking care to preserve phase, the density field has been high-pass filtered to remove the very largest scales of variation in water properties). Buoyancy flux will be a positive quantity if on average downward(upward) vertical velocities carry lighter(heavier) water.

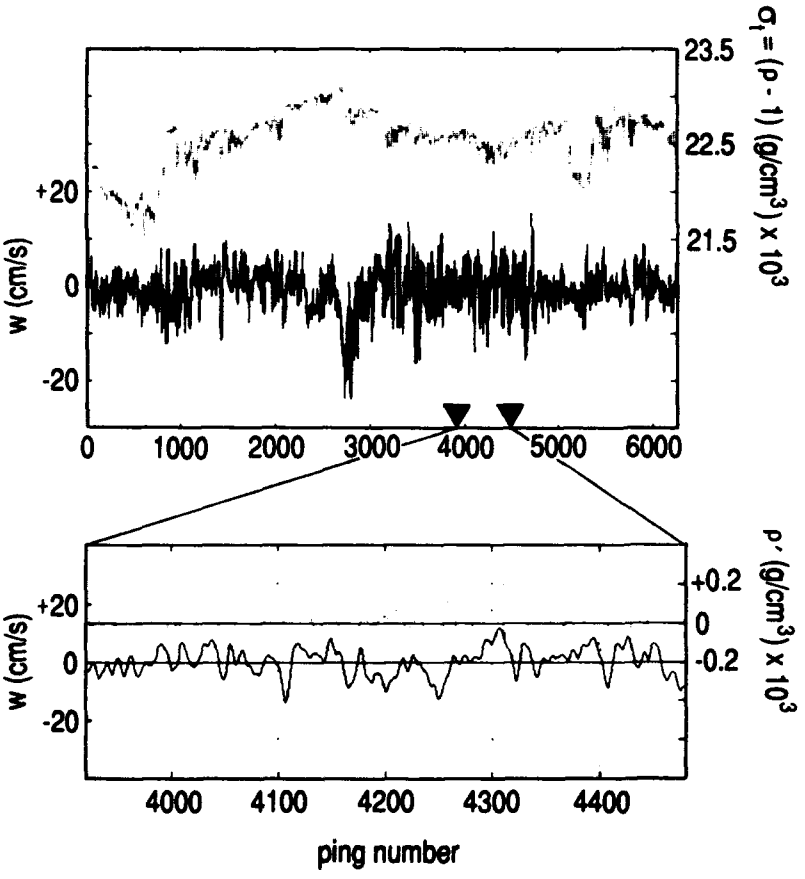


Figure 10. The top panel shows time series of CTD density (light line), along with w (dark line) from the Doppler bin within which the CTD was towed. Before calculating fluxes, the density time series is high-pass filtered (preserving phase) to remove the variance associated with large-scale water mass change: the enlargement shows filtered density and w over one of the interval lengths used in the flux calculation.

Figure 11 shows the direct flux estimates, formed by breaking the $\rho' w$ records into pieces of fixed length=spts, then forming $\overline{(\rho' - \overline{\rho'})(w - \overline{w})}$ where the average is over spts. Error bars are calculated from the variance of such estimates over the number(spts) of different starting points, and an estimate of the number of independent values determined from the

number of zero-crossings of w. The points in Figure 11 are the accompanying estimates of the buoyancy flux made using (2) above with $R_f/(1-R_f) = 0.2$ (ϵ values were taken from the Oregon State profiler measurements over a range of 6 m centered on the CTD tow depth: courtesy of Jim Moum). While there is encouraging general agreement, i.e., values tend to be high where the direct flux measurement is large and positive, low when the direct measurement is not statistically different form zero, we face (again) the problem of how best to average "point" estimates from the profiler for comparison with a more broadly based determination from the towed measurement.

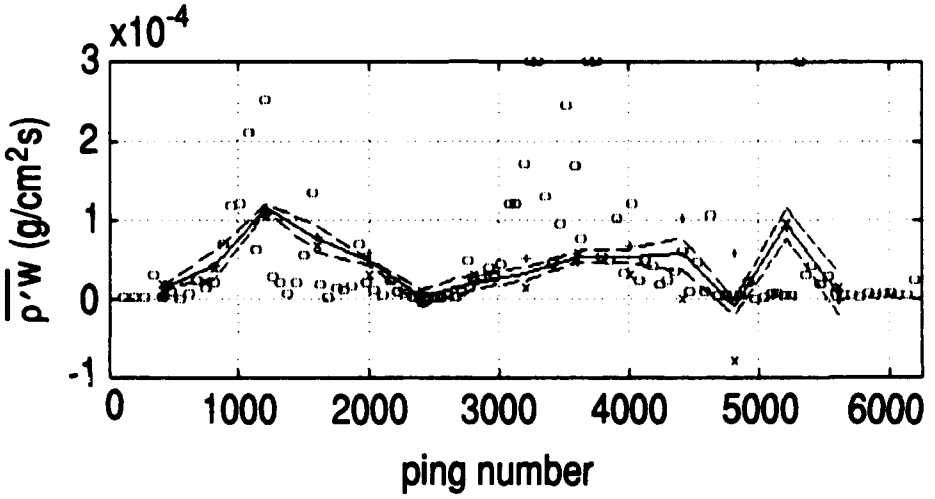


Figure 11. Direct calculation of buoyancy flux (solid line) with estimated error bars (dashed lines) over consecutive 400-point blocks of the time series shown in Figure 10. Circles are indirect estimates of the flux, using profiler "point" measurements of ϵ and the formula $0.2\rho_0 g^{-1}\epsilon$.

CONCLUSIONS

It is now possible to make measurements of the vertical velocity field in turbulent coastal flows, using a modified ADCP system. This allows us to site-survey for turbulence and, once found, to investigate its spatial and temporal variability. From the w field measurement, it will be possible to estimate turbulent kinetic energy E and possibly its dissipation rate ϵ. Addition of a towed CTD allows direct measurement of buoyancy flux: *if* the model (2) relating buoyancy flux to ϵ can be validated, remote measurement of ϵ would be equivalent to remote measurement of buoyancy flux, probably *the* feature of turbulent flows that is of the greatest importance to coastal applications.

Are results from the coastal ocean likely to be valid when translated to offshore oceans? From the data presented here, velocities characteristic of turbulence in the coastal ocean are clearly much higher than those we expect offshore. However, coastal stratification is

also much larger: the combination makes the coastal ocean less different from that offshore than one might think. The lower offshore signal level poses some challenges, but *if* a stable platform can be provided, the increased accuracy available with the newer coded-pulse sonars should allow this type of measurement to be made offshore as well: one foresees applications in studies of surface and bottom boundary layers in particular. It is my hope that the techniques discussed here will eventually prove as useful in the offshore environment as they are in the coastal ocean.

REFERENCES

Agrawal, Y.C., and P.A. Hwang, 1991: Measurements of dissipation rate in the upper surface layer. Quest Tech. Rep. No. 278, Quest Integrated Inc., 21414 68th Ave. S, Kent, WA 98037.

Batchelor, G.K., and A.A. Townsend, 1948: Decay of isotropic turbulence in the initial period. *Proc. Roy. Soc. A, 193*, 539-558.

Hunt, J.C.R., D.D. Stretch, and R.E. Britter, 1988: Length scales in stably stratified turbulent flows and their use in turbulence models. **In** *Stably Stratified Flow and Dense Gas Dispersion*, ed. J.S. Puttock. Academic Press. 285-321.

Kolmogoroff, A.N., 1941: The local structure of turbulence in an incompressible viscous fluid for very large Reynolds number. C.R. Acad. Sci., USSR, *30*, 301-305.

Osborn, T.R., 1980: Estimates of the local rate of vertical diffusion from dissipation measurements. *J. Phys. Oceanogr., 10*, 83-89.

Wamser, C., and H. Müller, 1977: On the spectral scale of wind fluctuations within and above the surface layer. *Quart. J. Roy. Met. Soc., 103*, 721-730.

Yamazaki, H., and T. Osborn, 1993: Direct estimation of heat flux in a seasonable thermocline. *J. Phys. Oceanogr., 23*, 505-516.

FINESCALE SHEAR AND STRAIN IN THE THERMOCLINE

Robert Pinkel and Steven Anderson*
Marine Physical Laboratory, Scripps Institution of Oceanography
La Jolla, California 92093-0213

Early studies of the temperature, density, and velocity fields in the sea were performed from a "hydrographic" perspective. The expectation was that one could "chart the oceans" structurally. The charts, once drawn, would remain valid. The tools of hydrography were the reversing thermometer and the Nansen bottle. These yielded a picture of the ocean interior on vertical scales of hundreds of meters, horizontal scales of tens of kilometers. From very early on it was appreciated that smaller scale phenomena were active in the ocean interior. Yet it was difficult to infer the role these small scale motions played in maintaining the hydrographic fields.

With contemporary sensors far clearer pictures of the small-scale oceanic fields are emerging. Yet the difficulty in quantifying the interaction with the hydrographic-scale ocean remains. In this work we concentrate on motions of vertical scale 3-50 m. Over this range, the scalar fields transition from highly skewed to nearly Gaussian behavior. The objective of this work is to quantify this transition in a statistical sense, with a particular focus on strain, shear and Richardson number, $R = N^2 / (\partial u / \partial z)^2$. Here, $N^2 = g / \rho \, \partial \rho / \partial z$ is the Vaisala frequency squared, where ρ is the potential density of the sea water.

Strain statistics were investigated in a previous work (Pinkel and Anderson 1992, henceforth PA 92) and are reviewed here in section 1. This previous study emphasized the utility of describing the finescale fields from the perspective of "reversible fine structure," a term introduced by Desaubies and Gregg (1981). They argued that the extremely intense finescale variability of passive scalars in the thermocline results from the simple straining of a smoother underlying field by the energetic internal wavefield. Irreversible processes such as turbulent mixing (Cox *et al.* 1969) and thermohaline intrusions (Stommel and Federov, 1967) typically play a secondary role. If one adopts the reversible finestructure hypothesis, it becomes attractive to describe variations in a coordinate system that is unaffected by the finescale straining. Using a repeatedly profiled CTD, we track the vertical motion of a set of isopycnal surfaces. The time evolution of both scalar and vector fields can be described in this isopycnal following frame (henceforth referred to as a semi-Lagrangian frame), as well as in a conventional Eulerian frame.

*Now at Woods Hole Oceanographic Institution
Woods Hole, MA 02543.

In section 2, shear and Richardson number statistics are presented. Shear data are obtained from a 161 kHz coded-pulse Doppler sonar mounted on the Research Platform FLIP. Sonar resolution is sufficient that the vertical advection of the shear field by the internal wavefield can be seen. This observation encourages the use of semi-Lagrangian coordinates to describe time evolution of the shear. The modeling of Richardson number takes on a different form in the semi-Lagrangian frame than in previous Eulerian studies, such as Desaubies and Smith (1982) or Munk (1981). In section 2 a simple model is derived and compared with the data. Agreement between model and data is encouraging. A brief discussion of results and implications concludes this work.

1. FINESCALE STRAIN IN THE THERMOCLINE

Strain Measurement

The data considered for the strain study are a set of 9000 CTD profiles, from the surface to 560 m. These were obtained during October 1986 from the Research Platform FLIP, when it was located at 34°N, 127°W, approximately 500 km west of Point Conception, California. Position was maintained to within 300 m by a two-point moor. Water depth at the site is 4 km.

The CTDs used are Seabird Instruments model SBE-9s. Two such instruments are profiled. The upper unit is cycled from the surface to 320 m. The lower system covers the depth range 250-560 m. Profiles are repeated at 3-min. intervals. The drop rate of the sensors is approximately 3.5 m s^{-1}. It is not necessary to pump water through the conductivity cell to achieve adequate spatial resolution at this drop rate. Following response corrections to the temperature and conductivity sensors (PA 92), density profiles are produced. A set of 560 isopycnals, of mean separation 1 m, is followed for the duration of the experiment.

The 3-hour record presented in Figure 1 represents a small portion of the 18.75-day data set. In it one sees a general trend toward decreasing isopycnal depth, associated with the baroclinic tide. Superimposed on this trend are higher-frequency (1-2 cph) internal waves. These are extremely coherent with depth. Against this large-scale background, the finescale straining of the density field is seen. Isopycnals converge to form "sheets" of high vertical gradient and diverge, forming low-gradient "layers." The typical time scale for the finescale variation appears to be from one-half to several hours, in this short record.

Protagonists in the present study are

$$\text{isopycnal displacement } \eta(t) = z(\rho,t) - \overline{z(\rho)},$$

$$\text{isopycnal separation } \Delta z_{ij}(t) = z(\rho_i,t) - z(\rho_j,t),$$

$$\text{the normalized separation } \gamma_{ij}(t) = \Delta z_{ij}(t) \,/\, \overline{\Delta z_{ij}},$$

$$\text{and the finite-difference strain } \hat{\gamma}_{ij}(t) = \gamma_{ij}(t) - 1.$$

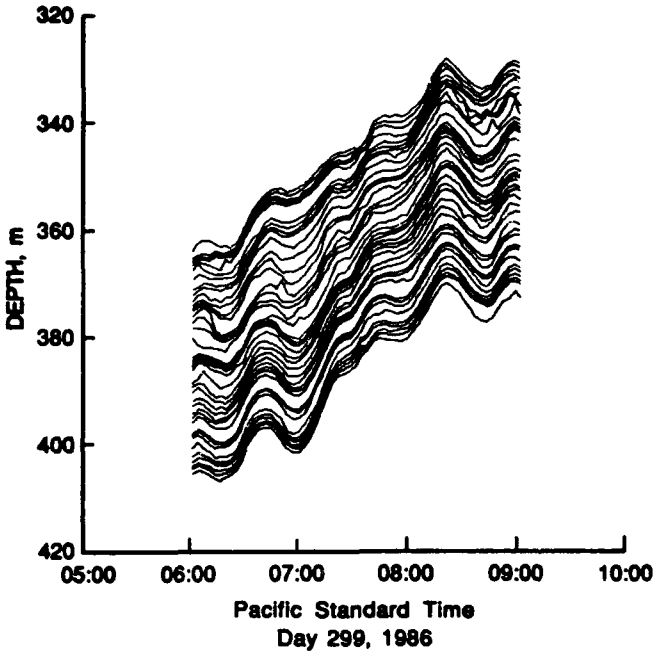


Figure 1. An example of isopycnal depth fluctuations as seen in the PATCHEX dataset. The statistics of isopycnal separation are the focus of the present study.

The Probability Density Functions of Strain

From the depth-time history of isopycnal displacement, strain statistics can be estimated in two distinct ways. One can simply calculate the probability density functions (pdfs) of separation between selected isopycnals pairs. This is the isopycnal following or "semi-Lagrangian" approach. One can also monitor the separation statistics of that pair of isopycnals that is bracketing a fixed reference depth. This provides an Eulerian view of the strain field. Both Eulerian and semi-Lagrangian pdfs have been calculated from the SWAPP data set. To investigate possible depth variability of the strain field, separate pdfs are formed for discrete 100-m depth regions: 100-200 through 400-500 m. Density functions for the 200-300 m region are presented in Figures 2 and 3, for mean isopycnal separations of 1-10 m. Each pdf is formed from 9×10^5 data, sorted into 100 bins. The data are not, however, mutually independent. Careful analysis (PA 92) suggests that there are between 50 and 90 independent estimates per bin in a typical histogram. Sample pdfs

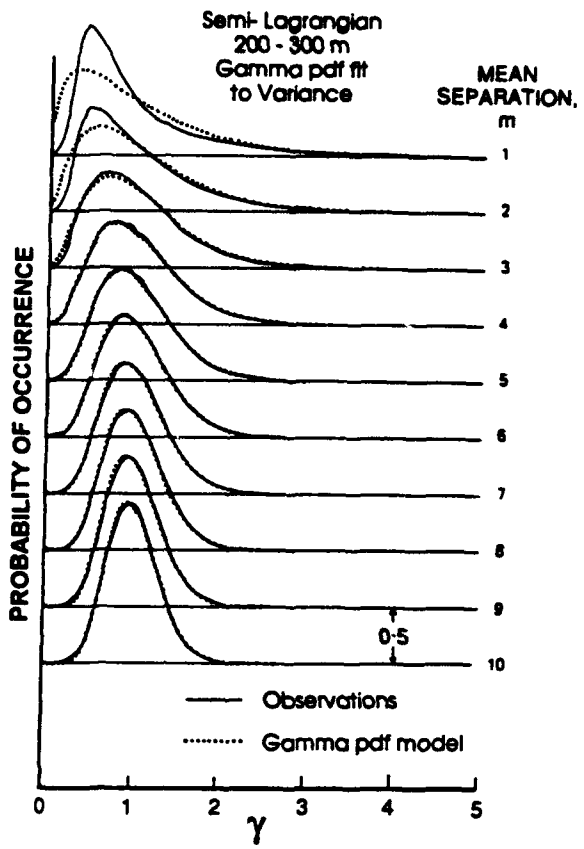


Figure 2. Probability density functions of normalized separations, γ, formed in a semi-Lagrangian frame, for mean isopycnal separations 1-10 m. Dotted lines give model Gamma pdfs, constrained to have unity mean and the observed variance. Data from 200-300 m depth are presented.

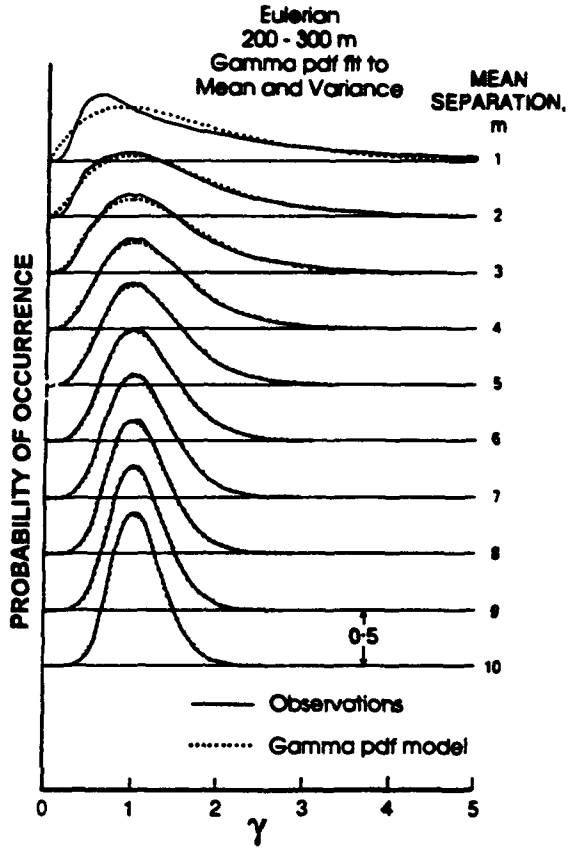


Figure 3. Probability density functions of normalized separation γ, as in Figure 2 except formed in an Eulerian frame. Dotted lines give model Gamma pdfs, constrained to have mean and variance identical to the observations. Data from 200-300 m depth are presented.

have been formed for mean separations as great as 50 m. While these appear nearly Gaussian at scales greater than 10 m, skewness and kurtosis estimates are significant to separations of order 30 m (Fig. 4).

The observed pdfs have been fit to a variety of classical forms, including Rayleigh, Weibull, Lognormal and Gamma. Significant discrepancies are subjectively apparent in all comparisons, with the notable exception of the Gamma pdf, which fits very well (Figs. 2,3). The Gamma pdf has the form

$$G(x) = \frac{\beta^{\alpha} x^{\alpha-1} e^{-\beta x}}{\Gamma(\alpha)} \qquad (1)$$

with mean <x> = α/β and variance $\langle x^2 \rangle - \langle x \rangle^2 = \alpha / \beta^2$ (Papoulis 1984).
The semi-Lagrangian data are constrained to have $\langle \gamma \rangle = 1, \langle \Delta z \rangle = \overline{\Delta z}$, by initial choice of isopycnals. Hence, $\alpha = \beta \overline{\Delta z}$. The fits presented in Figure 2 are thus one-parameter fits, with sample variance matched to the model variance. The Eulerian observations are not constrained to unity mean. These require two-parameter fits. The observed mean and variance are used to set model pdf parameters in Figure 3.

The Gamma pdf is seen to fit the observations well in the 200-300 m depth range, except at separations less than 4 m. The fits are comparable in the other depth ranges, with the exception of the 300-400 m interval, where the Lagrangian pdfs appear distorted at

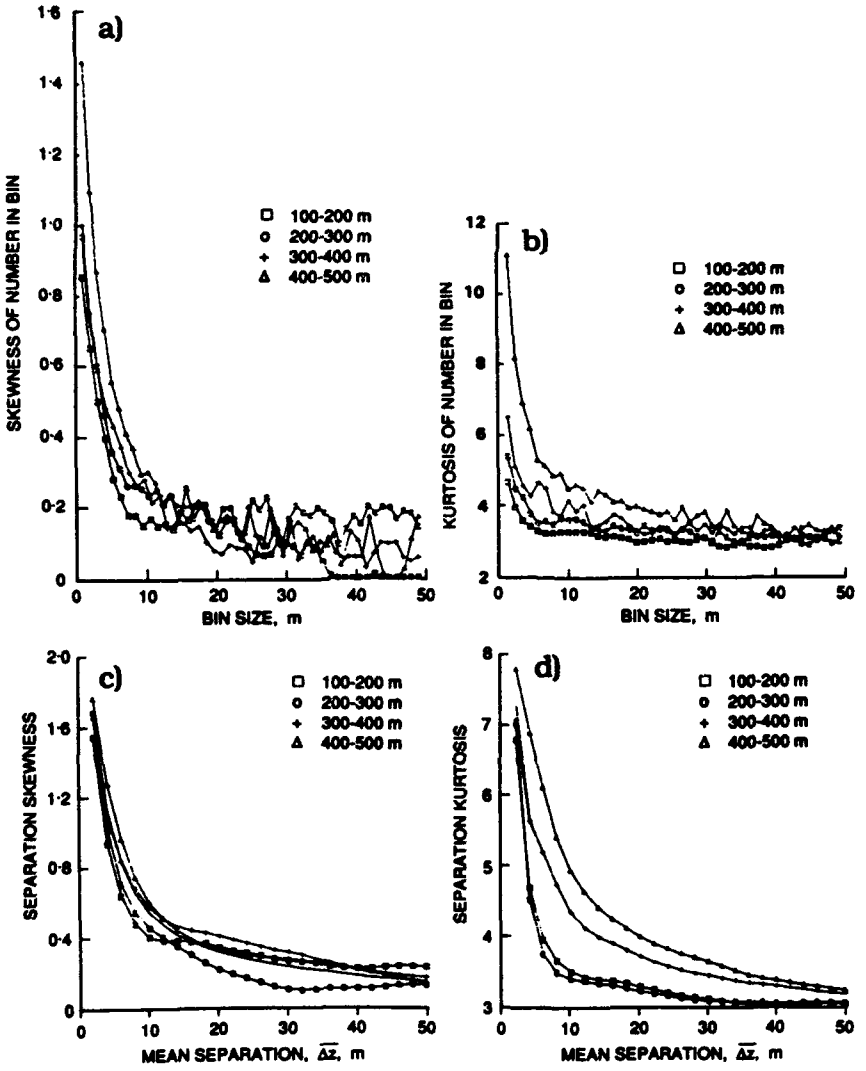


Figure 4. (a) Skewness and (b) kurtosis as a function of bin size for the Eulerian probability functions (Fig. 3). (c) Skewness and (d) kurtosis as a function of mean separation for the probability density functions of isopycnal separation (Fig. 2).

small γ, over a range of $\overline{\Delta z}$ = 3-7 m (PA 92). The fits could be improved by employing a least-squares fitting procedure. Optimizing the fit, however, is not the point of the present exercise.

A Statistical Model of Finescale Strain

Gamma pdfs are associated with the classical theory of Poisson processes. They describe the statistics of separation between the occurrence of Poisson "events" (Papoulis 1984). If, indeed, simple Poisson statistics describe the non-Gaussian behavior of the finescale field, the problem of modeling the motion field in this regime can be significantly advanced.

Considering the thermocline as a one-dimensional statistical process, we envision a set of "Poisson tracers," whose vertical position is tracked from one realization of the process to the next. Poisson statistics describe the occurrence of these tracers. The Poisson probability function gives the probability of occurrence of k tracers in a dimensional interval of length H:

$$P(n = k|H) = \frac{(\kappa_0 H)^k e^{-\kappa_0 H}}{k!} \tag{2}$$

The Poisson probability function has the interesting property that the mean number of "events" occurring in an interval $H, \kappa_0 H$, is equal to the variance of the number of events.

We define the normalized separation, γ, between two Poisson tracers to be the ratio of the instantaneous separation of the tracers to the mean separation, κ_0^{-1}. The strain, $\gamma - 1$, is assumed constant over the interval spanned by the tracers. Between adjacent tracer pairs, values of strain are uncorrelated. Thus an individual realization of the strain profile is discontinuous (Fig. 5b). However, the vertical profile of a passive scalar, θ, being strained in this Poisson field is continuous (Fig. 5a). The profile is composed of a series of constant gradient segments whose statistics are easily derived. The exponential distribution, $P(\Delta z) = \kappa_0 e^{-\kappa_0 \Delta z}$, governs the probability of separation between adjacent Poisson tracers, as well as the distance from arbitrary fixed points, Z_a, Z_b, to the adjacent tracers (Papoulis 1984). In a semi-Lagrangian study, the statistics of a specific pair of adjacent tracers are followed from one realization to the next. From the exponential distribution, one finds

$$\begin{aligned}
\langle \Delta z^2 \rangle_L &= 2/\kappa_0^2 \\
\langle \Delta z^2 \rangle_L &= 1/\kappa_0 \\
\langle \gamma^2 \rangle_L - \langle \gamma \rangle_L^2 &\equiv \langle \Delta z^2 \rangle_L / \langle \Delta z \rangle_L^2 - 1 = 1.
\end{aligned} \tag{3}$$

Thus, the strain variance as seen in a "tracer-following" frame is unity. While the tracer separation scale, κ_0, can be adjusted, the strain variance is fixed in this model.

In an Eulerian study one follows that pair, trio, or quartet of tracers that brackets the arbitrary fixed reference depths Z_a and Z_b. Different tracers may be involved from one realization to the next. In the event that a single pair of tracers brackets the reference depths, the separation between these tracers can be thought of as the sum of three terms:

$$\Delta z_{\text{bracket}} = (z_i - Z_a) + H + (Z_b - z_{i+1}).$$

Here $H = (Z_a - Z_b)$. Using the exponential distribution, it is easily shown that

$$\begin{aligned}
\langle \Delta z \rangle_E &= 2\kappa_0^{-1} + H \\
\langle \Delta z^2 \rangle_E &= 4\kappa_0^{-2} + 4\kappa_0^{-1} H + H^2 \\
\langle \gamma^2 \rangle_E - \langle \gamma \rangle_E^2 &= \frac{2}{(2 + \kappa_0 H)^2}.
\end{aligned} \tag{4}$$

In the limit of vanishing separation, H, the Eulerian strain variance has value 0.5. This is again an inherent aspect of the Poisson model, independent of the adjustable parameter κ_0.

An Eulerian covariance function for strain can be derived:

$$R_\gamma(Z_a, Z_b) = R_\gamma(H) = \left[\langle \Delta z_{ij} \Delta z_{kl} \rangle_E - \langle \Delta z \rangle_E^2 \right]. \tag{5}$$

Here Δz_{ij} gives the separation between those Poisson tracers that bracket depth Z_a while Δz_{kl} gives the separation between those tracers spanning depth Z_b. The brackets imply averaging over separate realizations of the profile. Given the hypotheses of the model, if one or more Poisson tracers occur between points Z_a and Z_b the strain will be uncorrelated: $R_\gamma(Z_a, Z_b) = 0$. If the points Z_a and Z_b fall between successive Poisson tracers, they will experience identical strain.

For this case the covariance is given by $R_\gamma(H) = \left[\langle\gamma^2\rangle_E - \langle\gamma\rangle_E^2\right] P_0$ (Papoulis 1984). Here $\langle\gamma^2\rangle_E - \langle\gamma\rangle_E^2$ is the strain variance of that tracer pair that is bracketing Z_a and Z_b, realization after realization. P_0 is the probability that Z_a and Z_b are spanned by a single pair of tracers. This is identically the probability that no Poisson tracers will be found in the interval, H, between the reference depths. From (2), $P_0 = e^{-\kappa_0 H}$. Combining (2) and (3), one has

$$R_\gamma(H) = \frac{2}{(2+\kappa_0 H)^2} e^{-\kappa_0 H}. \tag{6}$$

The corresponding vertical wavenumber spectrum of strain is given by

$$S(k) = 4\kappa_0^{-1}\,\mathrm{Re}\left[e^{2(1+2\pi ik/\kappa_0)} E_2(2(1+2\pi ik/\kappa_0))\right]. \tag{7}$$

Here E_2 is the exponential integral function (Abramowitz and Stegun 1970). The normalization is appropriate for a one-sided spectrum with k in cycles per meter. The covariance and wavenumber spectrum of strain are presented in Figures 5c,d.

This Poisson model of the thermocline is powerful by virtue of its primal simplicity. The single variable κ_0 describes all dimensional aspects of the model. The model successfully links the strain correlation scale κ_0^{-1}, of order 1 m in the open-ocean pycnocline, with the well-known cutoff of strain and shear that occurs near 10-m scale (Fig. 5d). There is no need to invoke a critical Richardson number criterion here, as in Munk (1981). The model predicts a spectral slope slightly steeper than the classical k^{-1} form, at scales shorter than 10 m.

The spectral level in the low wavenumber limit is $1.109\ \kappa_0^{-1}$, for a one-sided spectrum with wavenumber in units of cycles per meter. In the various Garrett-Munk models of the internal wavefield (e.g., Munk 1981), the strain spectral level is given by $S_{GM}(k) = \pi^2 E b j_*$ (Gregg and Kunze 1991), where $E=6.3 \times 10^{-5}$ is the dimensionless internal wave energy parameter, and $b = 1.3 \times 10^3$ m is the pycnocline scale depth. The G-M model best fits the Patchex strain spectrum for values of the bandwidth parameter j_* (Sherman 1989). For wavenumbers $0.01 < k < 0.1$, the Poisson and G-M model spectral levels are comparable if $\kappa_0 = 1.109/\pi^2 E b j_* = 1.37 m^{-1}: j_* = 1$. The Poisson approach differs from the G-M model in that the strain variance is fixed. The form of the wavenumber spectrum changes as the spectral level is altered, such that the variance is preserved. Also, unlike the G-M approach, the Poisson model relates variance to skewness, kurtosis, and higher-order quantities as a function of vertical scale.

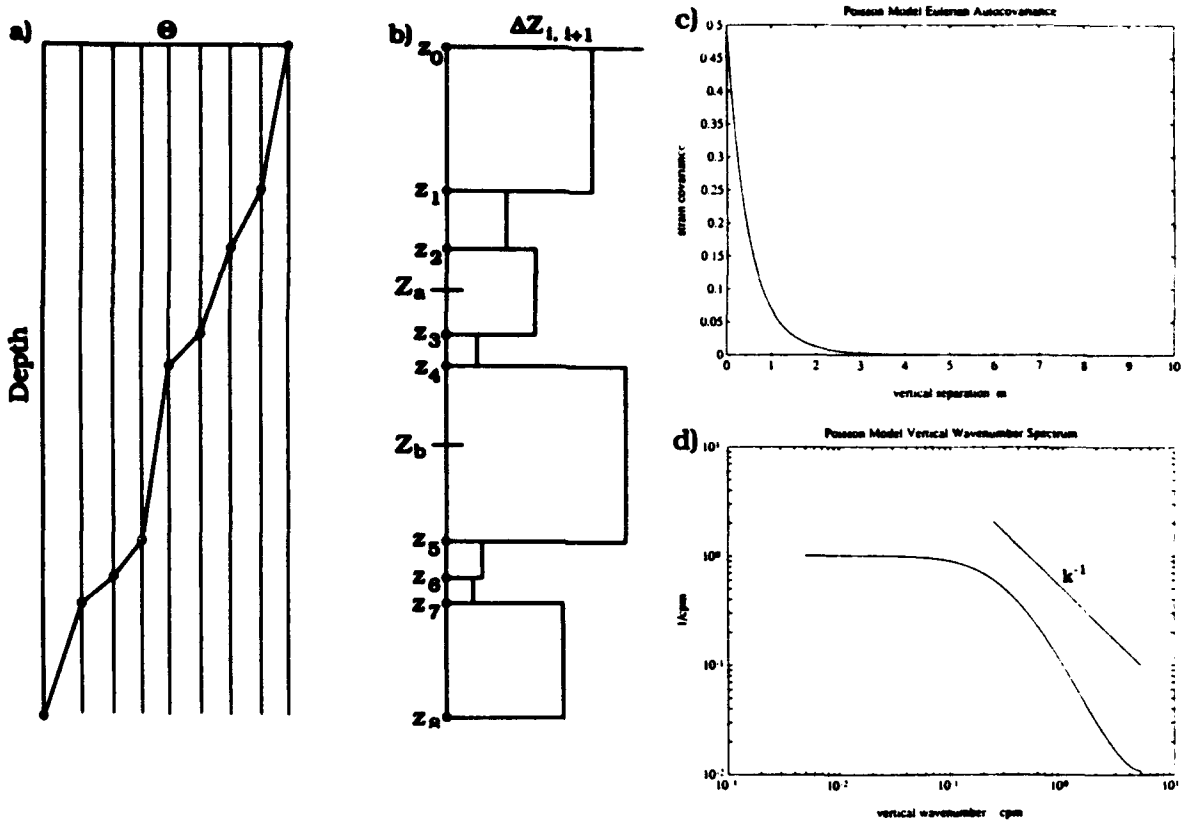


Figure 5. A model vertical profile of a passive scalar, θ. (a). The profile consists of a series of constant gradient regions. These correspond to the regions of constant strain (b), whose boundaries, $\{z_i\}$, vary from realization to realization as a Poisson process. The strain, $(z_j - z_{j+1}) / \overline{\Delta z}$ has a spatial auto covariance (c) and vertical wavenumber spectrum (d), here evaluated for $\kappa_0 = 1.1$ m^{-1}. Note that the Poisson scale κ_0, of order 1 m, is associated with a cutoff in the spectrum at a scale roughly 2π times larger.

2. FINESCALE SHEAR AND RICHARDSON NUMBER

Shear and Strain

The apparent success at modeling the strain field using a "reversible fine structure" approach prompted a similar investigation of fine-scale shear. An initial data set was collected during February and March 1990 in the surface waves processes experiment, SWAPP. A 155-kHz Doppler sonar was mounted on the Research Platform FLIP and operated in conjunction with the profiling CTD. During this period, FLIP was tri-moored at 35°N, 127°W. Water depth at the site was approximately 4 km.

The sonar obtained quality estimates of water velocity over the depth range 30-300 m, with 5.5 m vertical resolution. It operated continuously over a 19 day period. However, during the central period of the cruise, March 6 to 9, a large front passed under FLIP, significantly altering the qualitative nature of the velocity and shear fields. We restrict the subsequent study to the period before frontal passage, to avoid the atypical regime.

The profiling CTDs were similar to those used in the Patchex Experiment. In SWAPP, however, the profiling rate was increased to once per 130 s, rather than the previous 180 s. The increased rate was selected to improve CTD derived estimates of vertical velocity and strain rate. Profiles were achieved from the surface to a depth of 420 m.

SWAPP represents an evolutionary departure from previous FLIP-based examinations of the thermocline. Rather than using long range (~1.2 km) Doppler sonars of relatively low (15 m) vertical resolution (e. g., Pinkel *et al.*, 1987) here, a shorter range system with higher resolution is used. The development of a practical scheme for coding the sonar transmissions (Pinkel and Smith, 1992) enables the improved resolution and precision attained in SWAPP.

For the first time, the resolution scale of the sonar approaches the vertical displacement scale of the internal wavefield. When the shear field is closely examined the distortion due to the vertical displacement of the wavefield is clearly seen. In Figure 6, a representative 12-hour segment of the shear field is presented. Superimposed on the plot are the depths of a selected set of isopycnal surfaces. These illustrate the vertical displacement of the wavefield. Instances where the low frequency shear is being advected by high frequency waves are seen throughout the record.

While not a totally unexpected observation, the vertical advection of the low frequency shear by high frequency internal waves represents an interesting reversal of the typical view of wavefield kinematic behavior. It is more common to think of long wave-short wave interactions in terms of the short (high frequency) waves being advected/refracted by the long (low frequency) waves. In the oceanic thermocline, the "long" near inertial waves can have shorter *vertical* wavelengths than the "short" (horizontal wavelength) high frequency waves; hence, the opportunity to observe this distortion.

Modeling Shear and Richardson Number

The apparent displacement of the low frequency shear field by high frequency internal waves has both dynamic and kinematic consequences. Here, we focus on the purely descriptive problem, the appropriate modeling of shear and Richardson number statistics.

Figure 6 suggests that the statistics will be quite different, depending on whether the data are collected in an Eulerian or semi-Lagrangian frame.

The issue of the statistical independence of shear and strain is critical for understanding the evolution of the Richardson number. Desaubies and Smith (1982), in a previous attempt to model Ri, assumed the independence of these quantities. Munk (1981), in a separate study, assumed that strain was effectively constant; only shear fluctuations affected the Richardson number. Figure 6 suggests that the horizontal velocity and shear fields are being simply advected by the vertical velocity. We can avoid the kinematic aspects of the problem by shifting to an isopycnal following frame. However, a first order dilemma remains. In the semi-Lagrangian frame, is the shear,

$$\frac{\partial u}{\partial z}(\overline{\rho,t}) \equiv \left[\frac{u(\rho_1,t)-u(\rho_2,t)}{z(\rho_1,t)-z(\rho_2,t)}\right] = \left[\frac{u(\rho_1,t)-u(\rho_2,t)}{\overline{\Delta z}}\right]\cdot\gamma_{12}^{-1}(t) \tag{8}$$

truly independent of the strain, γ_{12}? If so, then the cross isopycnal velocity difference

$$\Delta u \equiv u(\rho_1,t)-u(\rho_2,t) \equiv \frac{\partial u}{\partial z}\gamma_{12}\overline{\Delta z} \tag{9}$$

must be dependent on strain. The converse also holds. It is not possible that both Δu AND $\partial u/\partial z$ be independent of strain.

To address this fundamental issue, a separate study was performed. The correlations between shear squared, velocity difference squared, and inverse strain (Vaisala frequency squared) were estimated. Non-zero correlation between quantities precludes the possibility of statistical independence. In both semi-Lagrangian and Eulerian frames, the shear squared – N^2 correlation coefficient,

$$R = \frac{\left\langle (\partial u/\partial z)^2\cdot\gamma^{-1}\right\rangle - \left\langle (\partial u/\partial z)^2\right\rangle\left\langle \gamma^{-1}\right\rangle}{\left[\left[\left\langle \left((\partial u/\partial z)^2\right)^2\right\rangle - \left\langle (\partial u/\partial z)^2\right\rangle^2\right]\left[\left\langle \gamma^2\right\rangle - \left\langle \gamma\right\rangle^2\right]\right]^{1/2}} \tag{10}$$

was of order 0.5, for average spatial separations between isopycnals (semi-Lagrangian) and vertical differencing intervals (Eulerian) of 4-20 m. (Desaubies and Smith (1982), assumed this correlation to be zero.)

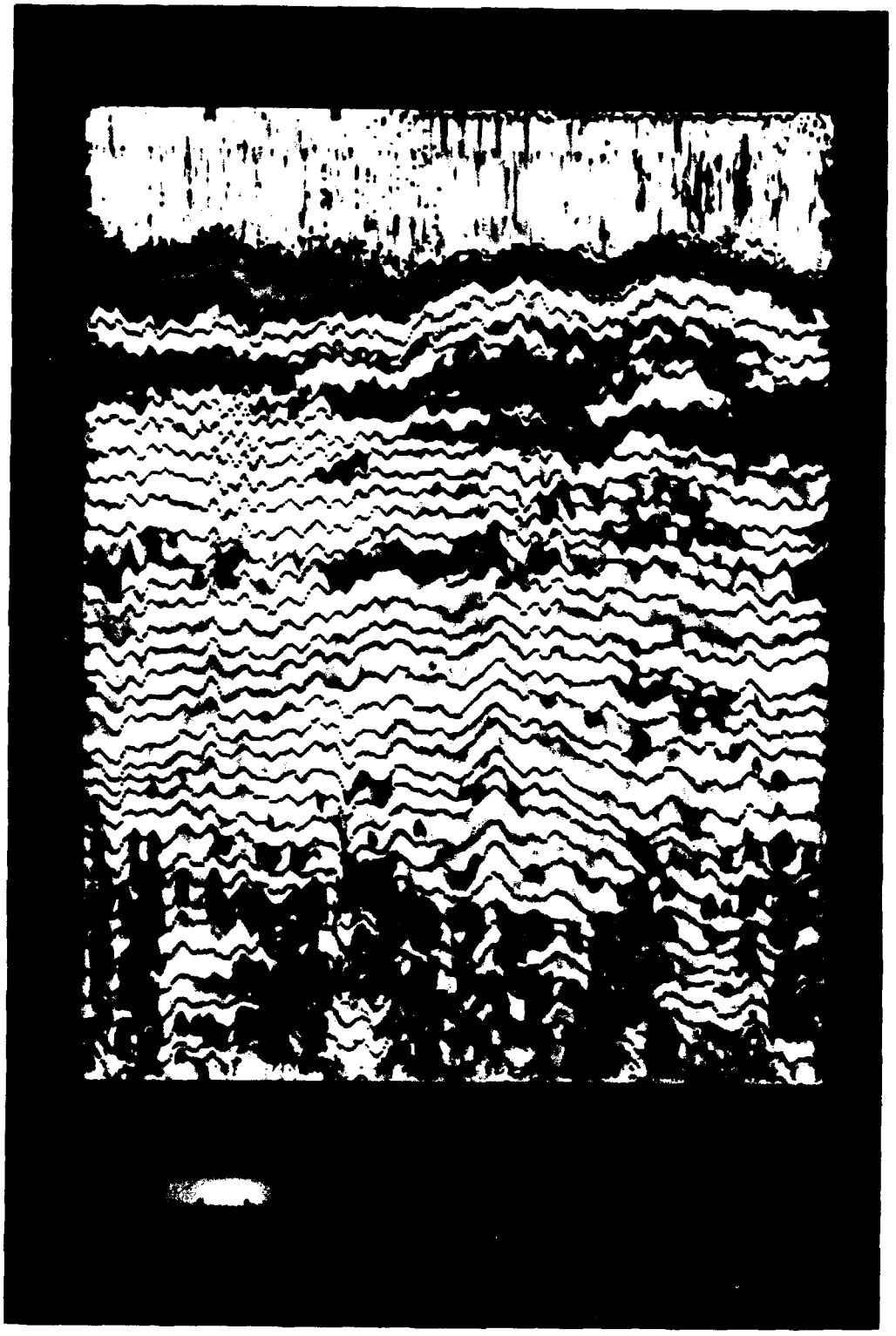


Figure 6. Shear magnitude, $[(\partial u/\partial z)^2 + (\partial v/\partial z)^2]^{1/2}$ plotted as a function of depth and time. Darker shading represents greater values of shear magnitude. Solid lines represent the depths of a selected set of isopycnals of uniform mean separation. There is evidence of the shear layers being vertically advected along with the density field by high frequency internal waves. This is seen most clearly at depths 80-200 m. Irregular shear variability below 300 m reflects imprecision in the sonar velocity measurement at great range.

$\Leftarrow$

In contrast, the correlation between velocity difference squared and N^2 was negative, of order –0.1 at 20 m scales, decreasing to –0.3 at 4 m mean isopycnal separation. The negative correlation indicates that larger values of Δu^2 were seen when isopycnals were far apart (small N^2), while smaller velocity differences are found when isopycnals are closely spaced (large N^2).

It is attractive to hypothesize that velocity difference and strain truly are uncorrelated. The observed correlation could result from the finite resolution of the Doppler sonar. Velocity differences along isopycnals are unbiased provided isopycnal separation is large compared to the sonar resolution scale. As isopycnals converge, Δu^2 estimates are biased low.

A model of the biasing effect was created, taking care to account for the non-Gaussian nature of the Δu^2 and N^2 fields. The model assumed the actual independence of these fields. The apparent correlation was then calculated, after modeling the effect of finite sonar resolution. The agreement between modeled and observed correlation was good, consistent with the hypothesis that Δu^2 and N^2 are indeed independent.

Toward a Statistical Model of Richardson Number

The indications of Figure 6 and the correlation studies referred to above suggest a particularly simple approach to the modeling of Richardson number. In a semi-Lagrangian frame, consider

$$\hat{R}i(t;\overline{\Delta z}) \equiv \frac{N^2(t;\overline{\Delta z})}{(\partial u/\partial z)^2(t;\overline{\Delta z})} = \frac{\langle N^2\rangle \cdot \gamma^{-1}(t,\Delta z)}{(\Delta u^2(t;\overline{\Delta z})/\Delta z^2(t;\overline{\Delta z^2}))} = \frac{\langle N^2\rangle \cdot \gamma(t)}{\Delta u^2(t)/\overline{\Delta z^2}}. \qquad (11)$$

It is convenient to define a scale Richardson number $Ri^* \equiv \dfrac{\langle N^2\rangle}{\langle \Delta u^2\rangle / \overline{\Delta z^2}}$ and to model the normalized Richardson number $R = \hat{R}i / Ri^* \equiv \gamma(t)/r(t)$ where

$$r(t) \equiv \Delta u^2(t) / \left\langle \Delta u^2 \right\rangle. \tag{12}$$

Note that the scale Richardson number Ri* is not, in general, equal to the expected value of the Richardson number $\left\langle \hat{R}i \right\rangle$. We proceed by recalling that the pdf of γ is given by the Gamma distribution (1) with one adjustable constant, $\kappa_0 = \beta$, which appears to have the near universal value of 1.1. The pdf of the velocity difference between two isopycnals has not been previously investigated. We hypothesize that the individual components of horizontal (along isopycnal) velocity difference are Gaussian. Thus Δu^2 represents the sum of the squares of two Gaussian quantities. Its associated pdf is presumably chi squared, with two degrees of freedom. At two degrees of freedom the chi squared pdf takes on exponential form:

$$P(\Delta u^2; \overline{\Delta z}) = \frac{1}{\left\langle \Delta u^2 \right\rangle} e^{-\Delta u^2 / \left\langle \Delta u^2 \right\rangle}; \qquad P(r) = e^{-r}. \tag{13}$$

In Figure 7 the probability density function of horizontal velocity difference squared is plotted for a variety of mean isopycnal separations. The velocities are normalized by the *mean* isopycnal separation to produce a quantity with units of shear, which actually represents the statistics of squared velocity difference between moving isopycnals. The pdfs are very nearly exponential in form with a velocity difference (shear) variance that increases (decreases) with increasing mean separation $\overline{\Delta z}$. In contrast to the pdfs of strain, which are highly skewed at small separation, becoming nearly Gaussian as $\overline{\Delta z}$ increases, the pdfs of squared velocity difference are of nearly unchanging form. Only the scale variance $\left\langle \Delta u^2 \right\rangle$ changes significantly with mean separation, $\overline{\Delta z}$.

Hypothesizing the independence of Δu^2 and γ, we can form the joint pdf of shear² and strain,

$$P(\Delta u^2, \gamma; \overline{\Delta z}) = \kappa \frac{(\kappa\gamma)^{\kappa-1} e^{-\kappa\gamma - \Delta u^2 / \left\langle \Delta u^2 \right\rangle}}{\Gamma(\kappa) \left\langle \Delta u^2 \right\rangle}. \tag{14}$$

Here $\kappa = \kappa_0 \overline{\Delta z}$ and $\left\langle \Delta u^2 \right\rangle$ are functions of $\overline{\Delta z}$. Identifying the normalized Richardson number $R(\overline{\Delta z}) = \hat{R}i / Ri^* = \gamma / r$, we can integrate (14) to obtain

$$P_{sL}(R; \left\langle \Delta u_{sL}^2 \right\rangle, \overline{\Delta z}) = \frac{\kappa}{R(\kappa R + 1)} \left[\frac{\kappa R}{\kappa R + 1} \right]^{\kappa}. \tag{15}$$

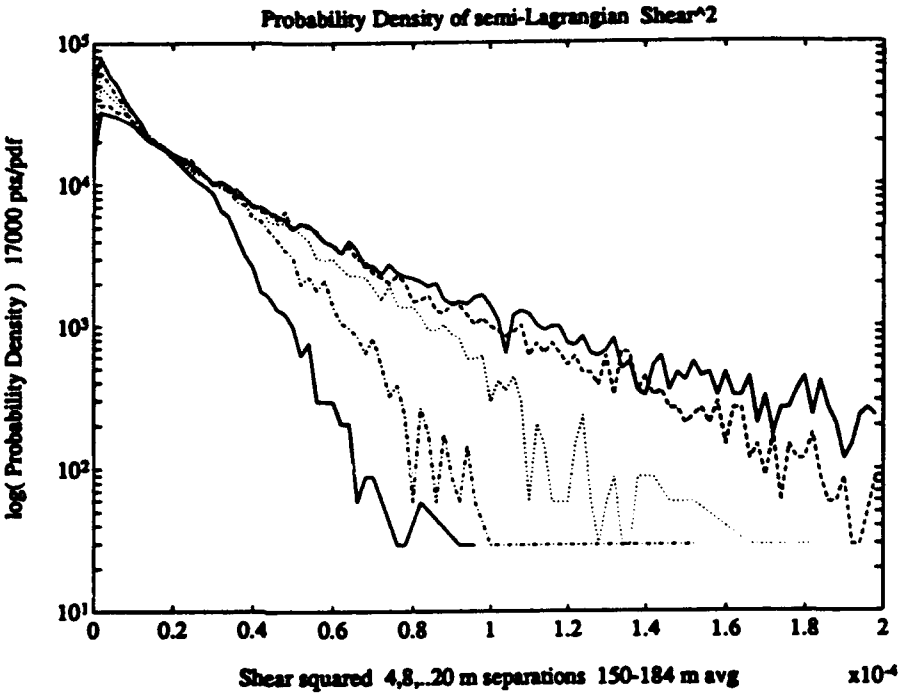


Figure 7. Histograms of incidence of occurrence as a function of velocity difference squared, formed in a semi-Lagrangian frame. Velocity differences are normalized by $\overline{\Delta z}^2$, to convert to shear-like units. Histograms are plotted for mean separations of 4, 8, 12, 16 and 20 m. An exponential model for the pdf of squared velocity difference implies a linear form for these semi-logarithmic histograms.

Similarly, in an Eulerian frame one has

$$P_{EU}(R;\langle \Delta u_{Eu}^2 \rangle, \overline{\Delta z}) = \frac{\kappa(\kappa+1)}{(\kappa R+1)^2}\left[\frac{\kappa R}{\kappa R+1}\right]^{\kappa}. \qquad (16)$$

Plots of the pdf of normalized Richardson number are presented in Figure 8. The semi-Lagrangian pdf is slightly more peaked than the Eulerian at small separations $\overline{\Delta z}$. This difference decreases with increasing mean separation. Again, in contrast to the strain, the skewness of these pdfs varies only weakly with increasing mean separation.

The initial comparison between the SWAPP observations and the model, while preliminary, is quite encouraging (Fig. 9).

3. DISCUSSION

To predict statistics of the actual (not normalized) Richardson number, at scales $\overline{\Delta z}$, or depths $\bar{z}$ beyond the resolution and reach of the SWAPP instruments, one must know the

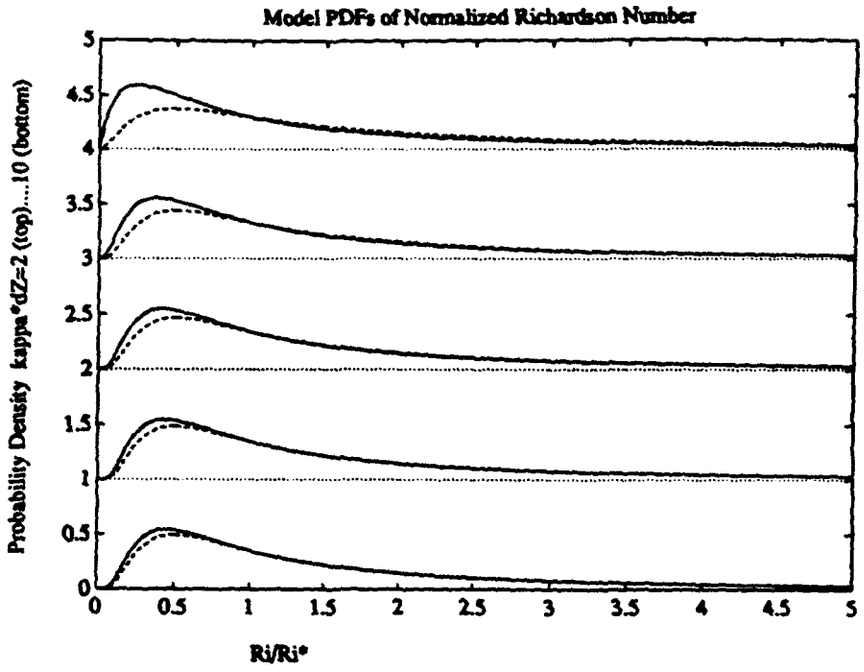


Figure 8. Model probability density functions of semi-Lagrangian (solid) and Eulerian (dashed) normalized Richardson number, for mean separations of 2-10 m. The difference between semi-Lagrangian and Eulerian observations decreases with increasing vertical separation. However, in contrast with the pdfs of strain, neither function approaches Gaussian form at large mean separation.

the behavior of $\langle \gamma^2 \rangle$ and $\langle \Delta u^2 \rangle$ as functions of depth and mean separation. Strain appears well modeled at scales $\overline{\Delta z}$ >3 m by the model presented above. The universality of the single adjustable parameter κ_0 is open to question. However, the behavior of $\langle \Delta u^2 \rangle$ is even less well known. In this study the estimates are influenced by instrument noise, which adds to the true variance, and instrument resolution, which detracts from it. Proper modeling of these effects is required for accurate estimates of $\hat{R}i$.

From vertical profiling measurements, there is a body of experience relating to the statistics of $\langle \Delta u^2 \rangle$, at least in an Eulerian frame. Gargett et al. (1981) were the first to synthesize a composite shear spectrum from a variety of profiling sensors. They concluded that the Eulerian shear spectrum has a form generally similar to the model strain spectrum presented in Figure 5d, being white at vertical wavenumbers less than 0.1 cpm and of k^{-1} slope at higher wavenumber. The shear spectral level scales as $\langle N^2 \rangle$, in contrast to the strain spectral level, which is independent of Vaisala frequency. The Gargett *et al.* (1981) empirical observation, sustained by more recent research, is that the spectral transition near 0.1 cpm is not a strong function of $\langle N^2 \rangle$. This behavior is inconsistent with linear dynamics in the WKB approximation. Vertical wavenumbers should vary as $\langle N^2 \rangle^{1/2}$ in a WKB pycnocline.

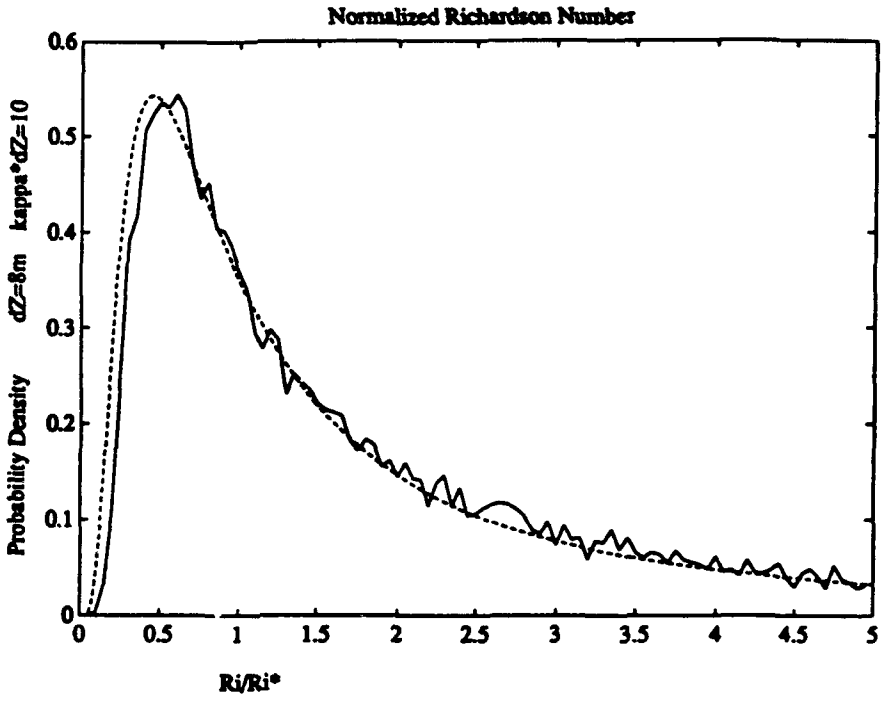


Figure 9. A comparison between the observed pdf of normalized semi-Lagrangian Richardson number and the corresponding model pdf. Agreement is generally good, although a clear systematic offset is seen. Observed values of low normalized Richardson number occur less often than predicted. Instrument noise and resolution affect both the instantaneous Richardson number observations and the value of the normalization factor Ri^*. Data from depths 150-184 m are used in this comparison.

If the Gargett *et al.* scalings are applied to the present statistical model of Richardson number, the scale Richardson number Ri^* is depth independent. The frequency of observance of instabilities should thus be independent of depth. This contrasts with the early internal wave breaking model of Garrett and Munk (1972). It is more consistent with the later view of Munk (1981).

There are several major concerns with the Richardson number modeling effort presented here. First and foremost, observations of overturning and instability in the thermocline typically indicate an overturning scale of a few meters or less. The model developed here is not supported by the observations at scales less than 3 m. In part this is due to the noise and resolution limits of the data. However, the Poisson strain model becomes internally inconsistent at scales smaller than κ_0^{-1}, the correlation scale of the strain field. The relevance of this "finescale" model of Richardson number variability to the occurrence of oceanic turbulence remains to be demonstrated.

A related concern is that the Richardson number might not at all be the parameter that is most sensitive to the occurrence of oceanic turbulence. Orlanski and Bryan (1969) suggested that a second form of instability, termed convective instability, was responsible for the bulk of the mixing in the sea. While Munk (1981) argued that both forms of instability were sensitive to the same aspects of the internal wave spectrum, the space/time distribution of convective mixing events might be far different than that of the events resulting from low Richardson number instability.

To investigate this concern, a microstructure probe was mounted on the CTD used in the SWAPP experiment. The sensor, a Seabird dual electrode microconductivity cell was capable of resolving overturns on scales as small as 10 cm. In the next phase of the analysis of this data set, we will attempt to correlate the occurrence of microstructure signals with the depth-time variation in finescale Richardson number. The degree of correlation will bear testimony to the relevance of the Richardson number as an indication of mixing in the thermocline.

REFERENCES

Abramowitz, M., and I. R. Stegun, 1970: *Handbook of Mathematical Functions*, U.S. Govt. Printing Office, Washington, D.C., 1046 pp.

Cox, C. S., Y. Nagata and T. Osborn, 1969: Oceanic fine structure and internal waves, *Bull. Jap. Soc. Fish. Oceanogr.*

Desaubies, Y. J. F., and M. C. Gregg, 1981: Reversible and irreversible fine structure. *J. Phys. Oceanogr.*, **11**, 541-556.

Desaubies, Y. J. F. and W. K. Smith, 1982: Statistics of Richardson number and instability in oceanic internal waves. *J. Phys. Oceanogr.*, **12**, 1245-1259.

Garrett, C. and W. H. Munk, 1972a: Oceanic mixing by breaking internal waves: *Deep Sea Res.*, **19**, 823-832.

Garrett C. and W. H. Munk, 1972b: Space-time scales of internal waves. *Geophys. Fluid Dyn.*, **3**, 225-264.

Gregg, M. and E. Kunze, 1991: Shear and strain in the Santa Monica Basin. *J. Geophys. Res.*, **96**, 16,709-16,719.

Munk, W. H., 1981: Internal waves and small scale processes. *Evolution of Physical Oceanography*, B. A. Warren and C. Wunsch, Eds., MIT Press, 264-290.

Orlanski, I. and K. Bryan, 1969: Formation of thermocline step structure by large amplitude internal gravity waves. *J. Geophys. Res.*, **74**, 6975-6993.

Papoulis, A., 1984: *Probability, Random Variables and Stochastic Processes.* McGraw-Hill, 576 pp.

Pinkel, R., and S. Anderson, 1992: Toward a Statistical Description of Finescale Strain in the Thermocline. *J. Phys. Oceanogr.* **22**, 773-795.

Pinkel, R., A. J. Plueddemann, and R. G. Williams, 1987: Internal wave observations from *FLIP* in MILDEX. *J. Phys. Oceanogr.*, **17**, 1737-1757.

Sherman, J. T., 1989: Observations of fine-scale vertical shear and strain in the upper-ocean. Ph.D. thesis, University of California, San Diego, 145 pp.

Stommel, H. and K. N. Federov, 1967. Small scale structure in temperature and salinity near Timor and Mindanao. *Tellus*, **19**, 306-325.

STRUCTURE OF THE UPPER OCEAN VELOCITY FIELD ON SCALES LARGER THAN 10 KILOMETERS

Eric Firing, Department of Oceanography,
School of Ocean and Earth Science and Technology
University of Hawaii, Honolulu, HI 96822

Abst act

Upper ocean currents, illustrated here by shipboard ADCP data, are a complex function of both space and time. Vertical shear is strong near the equator and decreases toward the poles. Particularly strong currents are found near the equator, in the southern ocean, and on western boundaries. High variability sometimes, but not always, coincides with strong mean currents. Inertial oscillations are ubiquitous and can dominate a dataset. Their spatial structure has not been well observed. An exploratory attempt to calculate horizontal wavenumber spectra from vertically averaged shipboard ADCP measurements shows potentially interesting differences between two sections, one at 35°N, the other near 18°N.

Introduction

Our knowledge of upper ocean currents is sketchy. The broad outlines come from statistical summaries of ship drift reports accumulated over more than a century. This global dataset shows the locations and typical speeds of the major surface currents, their average annual cycle, and a measure of their variability apart from the annual cycle (e.g., Wyrtki *et al.*, 1976; Richardson and Walsh, 1986; Richardson and McKee, 1989). The horizontal resolution of this dataset is coarse, typically 1–5°, and it indicates only currents averaged over the hull depth of the ships. There are many sources of error, such as the effects of wind and waves on the ship. Temporal resolution is also poor—the dataset is climatological, not synoptic. A second source of information about upper ocean currents is the hydrographic dataset, from which the geostrophic component of the currents may be calculated as a function of depth, not just at the surface (e.g., Toole *et al.*, 1988; Picaut and Tournier, 1990). When treated climatologically, this dataset has the same coarse resolution as ship drift data, but individual hydrographic sections can be inspected for a quasi-synoptic picture of the geostrophic current perpendicular to the ship track with a horizontal resolution of 0.5° or so. A third source of upper ocean current measurements is the surface drifter dataset (e.g., Hansen and Paul, 1984). It gives no information on vertical structure but gives a quasi-Lagrangian picture of

horizontal and temporal variations of currents at 10–15 m depth. It has recently been shown that time-averaged horizontal gradients of currents can be resolved on scales as small as 5 km by suitable averaging of a large drifter data set (Poulain, 1993). A fourth source of current measurements is the moored current meter dataset (e.g., McPhaden and Taft, 1988; Whitworth *et al.*, 1991). Temporal resolution is excellent, typically one hour or less. Horizontal resolution can be arbitrarily fine, but horizontal coverage is limited by the cost and availability of moorings. An array rarely includes more than 20 or so moorings.

During the last decade, a new source of upper ocean current measurements has been developed: the shipboard Acoustic Doppler Current Profiler (ADCP). An ADCP is now standard equipment on most research ships. The typical instrument (model VM-150 made by RD Instruments) can measure currents relative to the ship at 8-m depth intervals from about 20 m down to a maximum range of 200-450 m, depending on ambient noise and the density of acoustic scatterers. Individual profiles, measured once per second, are averaged into ensembles of a few minutes. The accuracy of these averages is usually of order 1 cm s^{-1}, although biases of order 10 cm s^{-1} can occur (Chereskin and Harding, 1993; Wilson and Firing, 1992). The velocity of the ship, measured by differencing position fixes, is added to the current profile relative to the ship to yield a profile of water velocity relative to the earth. With present Global Positioning System (GPS) navigation, 95% of fixes are within 100 m of the correct position. Fix errors are correlated over intervals of order 10 minutes. Velocity errors can be reduced to about 2 cm s^{-1} standard deviation in each component by differencing fixes 30 minutes apart. With a typical ship speed of 6 m s^{-1}, this means the effective horizontal resolution for absolute velocity profiles is about 10 km.

The purpose of this note is to show something of the character and complexity of upper ocean currents. We will use ADCP measurements from a few cruises in the Pacific to show how typical current speeds, horizontal scales, and vertical shears vary with latitude. We will illustrate, but not solve, the problem of combined temporal and spatial variability in the shipboard ADCP dataset. Examples of simple statistical analysis of this dataset will be given. They will show some features of ocean currents that have not been accessible previously and perhaps help motivate more extensive and sophisticated statistical analysis of shipboard ADCP data in the future.

The Central Pacific from 35°N to 60°S

Recent WOCE (World Ocean Circulation Experiment) Hydrographic Program (WHP) cruises provide high-quality current and hydrographic measurements

spanning the Pacific. Here we will look at the shipboard ADCP measurements from the central and southern portions of WHP lines P16, nominally along 150°W, and P17, nominally along 135°W. The cruises occurred in four legs on two ships: RV *Thomas Washington* in June through September 1991 (Talley and Swift, 1992) and RV *Knorr* in October and November 1992. Along most of the cruise track, 3–4-hour CTD stations were occupied at half-degree intervals.

A map of shallow current vectors shows how the character of the current field changes with latitude (Figure 1). Several distinct regimes can be distinguished. The region that probably catches the eye first is the band of strong, predominantly zonal currents within about 10° of the equator. Typical speeds are 50 cm s^{-1}, and the widths of the currents are 2–5°. The main currents seen here—the eastward North Equatorial Countercurrent (NECC), and the westward South Equatorial Current (SEC) split at the equator by a shallow fraction of the eastward Equatorial Undercurrent (EUC)—can be identified easily in most cross equatorial sections in the central or eastern Pacific. Still, as the difference between the sections on 135°W and 150°W suggests, their variability in time and space is substantial.

Poleward of the equatorial zone, in the tropics through mid latitudes, the typical currents are relatively weak, perhaps 20 cm s^{-1}, and their horizontal scale of variability is only 1–2° or less. What we see in these regions of Figure 1 is a field of eddies and other variability superimposed on a weak mean flow. The typical speeds and the horizontal length scales decrease with increasing distance from the equator until about 50°S, the northern edge of the Antarctic Circumpolar Current (ACC). There appears to be an abrupt decrease in eddy energy and scales at about 30°N and S, separating the ocean into a high-energy extra-equatorial region from 10–30° and a low-energy region from 30–50°. This tentative conclusion needs to be checked against additional datasets. High eddy energy can also be seen in Figure 1 near the Hawaiian Islands, as expected (Patzert, 1969).

On 150°W, the ACC apears to extend from 50°S to perhaps 62°S as a series of threads 1–2° wide. To the east, however, we see one or more large eddies or loops in the current. It appears that the ACC must turn north just east of 150°W and then loop south at about 140°W.

To see how the vertical structure of the currents varies with latitude, we turn to representative contoured sections (Figure 2). Along 35°N from the California coast to 135°W, most of the major current features are coherent in the vertical but decrease in amplitude with increasing depth. Typical vertical shears are of order 10^{-3} s^{-1}. In the equatorial band, by contrast, distinctly different currents are found at different depths in the upper 400 m. The eastward Northern Subsurface Countercurrent (NSCC; Tsuchiya, 1975), for example, has a maximum speed of 40 cm s^{-1} at 4.5°N, 230 m. Shears in the equatorial zone reach as high as

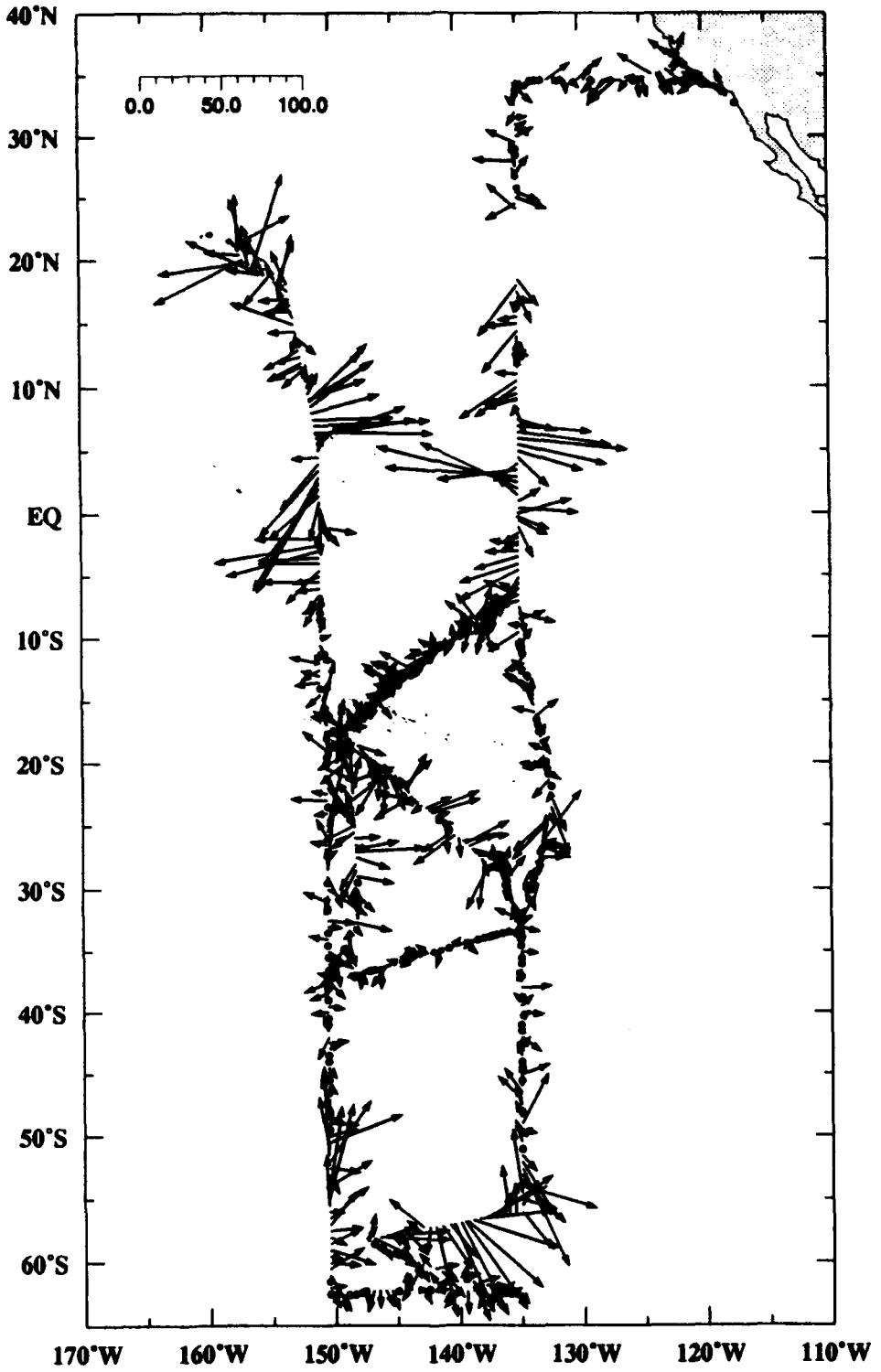


Figure 1. Currents averaged from 25 to 75 m on the central and southern portions of WHP lines P16 and P17, plus transits to and from port. These ADCP measurements were made on the *Thomas Washington* from May 31 to October 2, 1991, and on the *Knorr* from October 6 to November 27, 1992. The *Knorr* cruise went from Tahiti to 62.5°S and back.

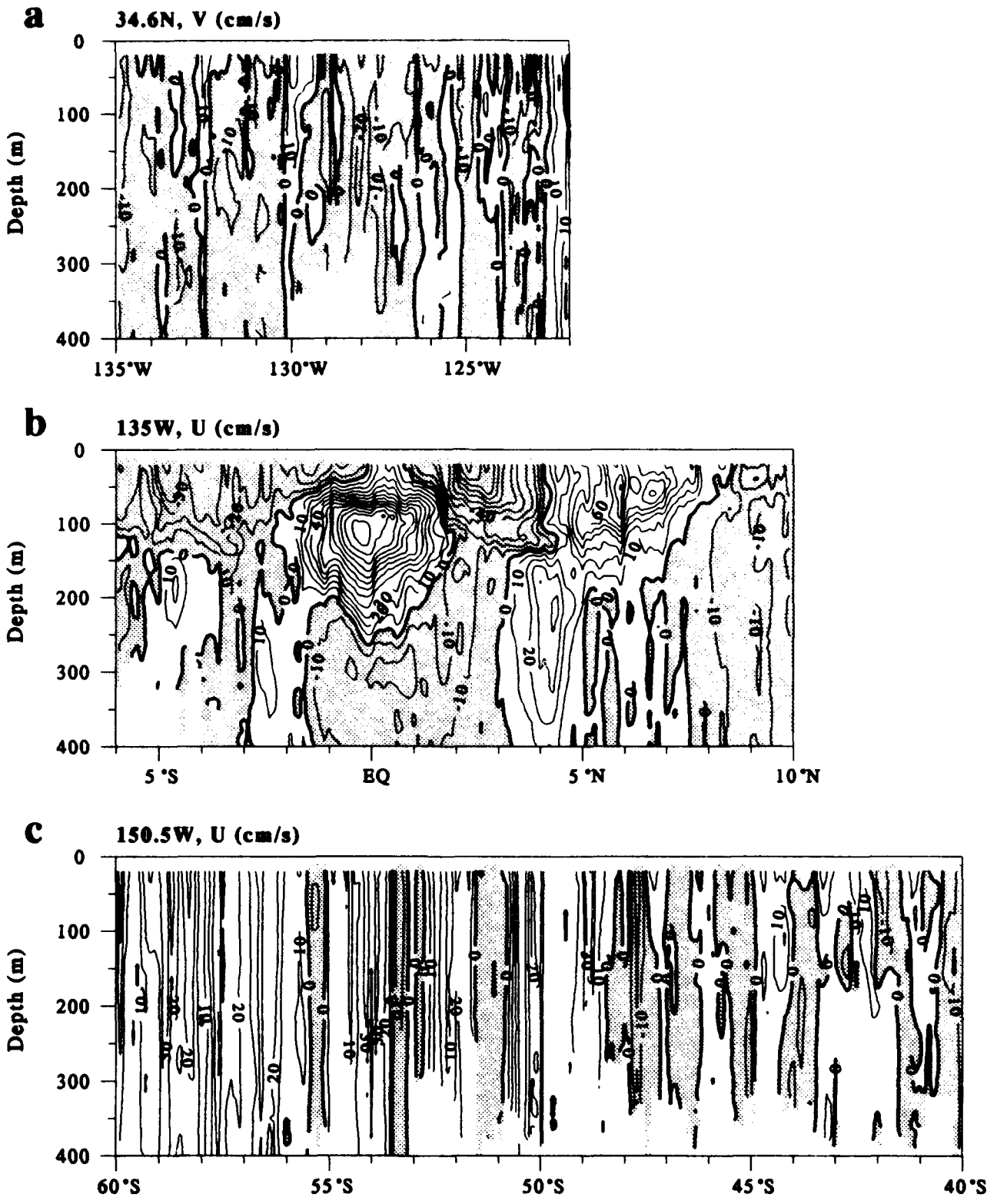


Figure 2. Upper ocean currents off the California coast (a: meridional component), near the equator (b: zonal component), and in the Southern Ocean (c: zonal component). Southward and westward flow is shaded. All contours are at 10-cm s^{-1} intervals. The axes are scaled uniformly in all three panels.

2×10^{-2} s^{-1}. In the ACC we find the opposite extreme: shears in the upper 400 m are typically less than 2×10^{-4} s^{-1}.

Across the Pacific at 10°N

From February through May 1989, the RV *Moana Wave* crossed the Pacific (Wijffels *et al.*, 1993). The cruise was run in three legs from west to east, mostly along 9.5°N, just north of the boundary between the NECC and the North Equatorial Current (NEC). CTD casts to the bottom were made every 2° of longitude, with closer spacing near the boundaries.

Along most of the section, the zonal component of current is westward, part of the NEC (Figure 3). Eddy-like variability is present everywhere, but is particularly strong near the western boundary and east of about 130°W. The dominant horizontal scales of this variability appear to vary from 1–5°. The signature of tropical instability waves (Hansen and Paul, 1984) is perhaps most evident in the strong currents near 120°W. These currents are very shallow; most of the energy in the eastern part of the section is found above 100 m.

The strongest currents of the section are found within 10° of the western boundary. The southward flow at the Philippine coast is the Mindanao Current, a permanent western boundary current (Lukas *et al.*, 1991). Fortunately, there are repeated sections across the Mindanao Current from several measurement programs; we will look here at data from cruises 3, 4, 5, 6, and 8 of the US/PRC TOGA Program (Delcroix *et al.*, 1992), from 1987 to 1990. Two of these cruises occurred in boreal fall, three in boreal spring. The mean meridional velocity component shows the Mindanao Current and little else; almost all of the region from 129°E to the end of the section at 141.5°E has a mean current below 10 cm s^{-1} (Figure 4). The mean Mindanao Current is less than 2° wide, has a maximum speed near the coast exceeding 80 cm s^{-1}, and extends below the 350-m depth range of these measurements. The pattern of variability differs greatly from the mean. The standard deviation is minimal, only 5–10 cm s^{-1}, at the coast, where the Mindanao Current is strongest. The standard deviation then increases eastward with maxima greater than 30 cm s^{-1} on the edge of the mean Mindanao Current and beyond the edge at 129°W. East of there, most of the variability is found in the upper 100 m, with typical standard deviations from 15–25 cm s^{-1}. Below 100 m the standard deviations are mostly 5–10 cm s^{-1}. There is no obvious seasonal difference in the currents in this dataset; variations among sections of the same season are as large as variations between the two seasons.

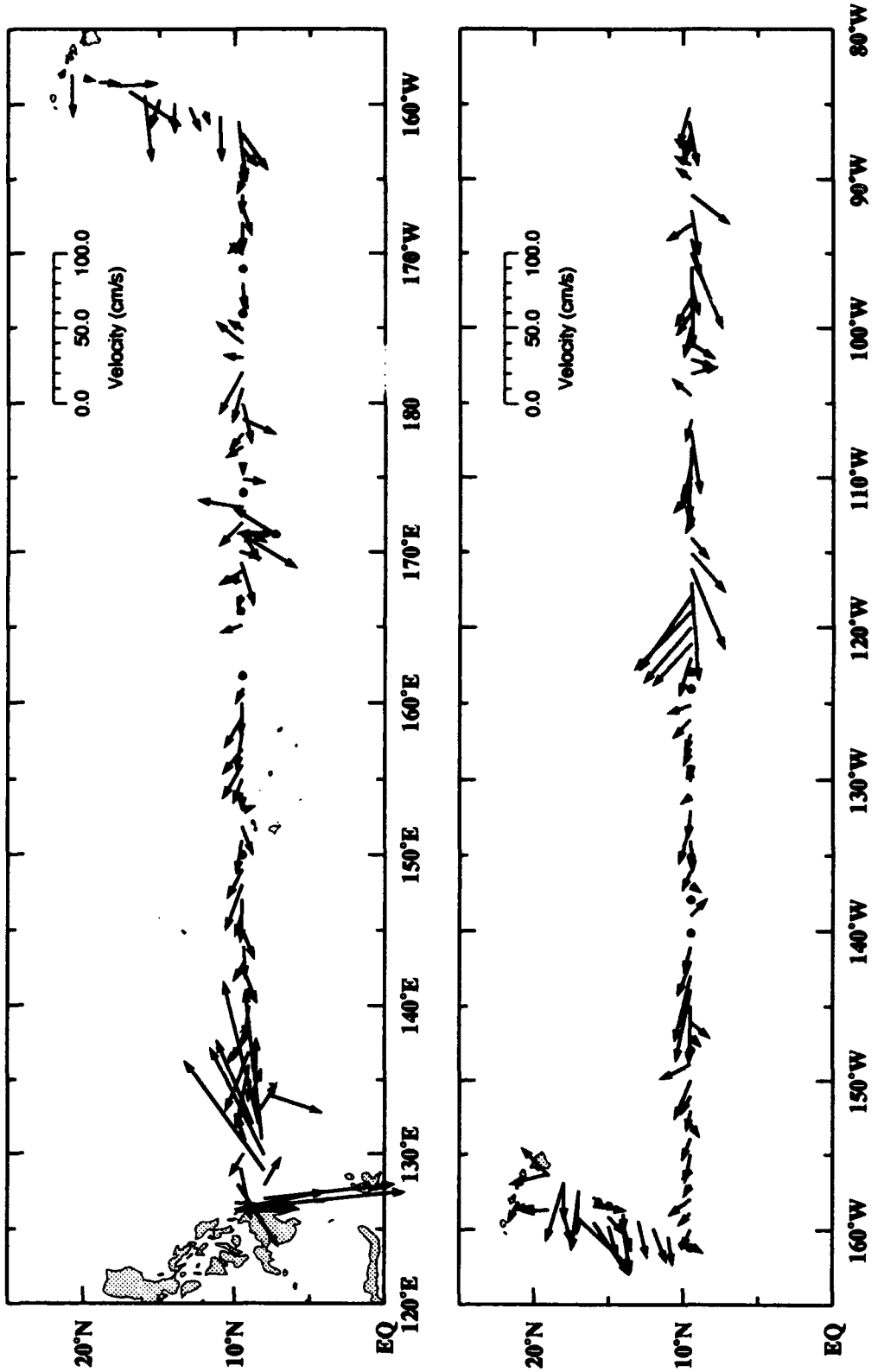


Figure 3. Currents averaged from 25 to 75 m on three cruise legs of the *Moana Wave*, February 2 to May 10, 1989.

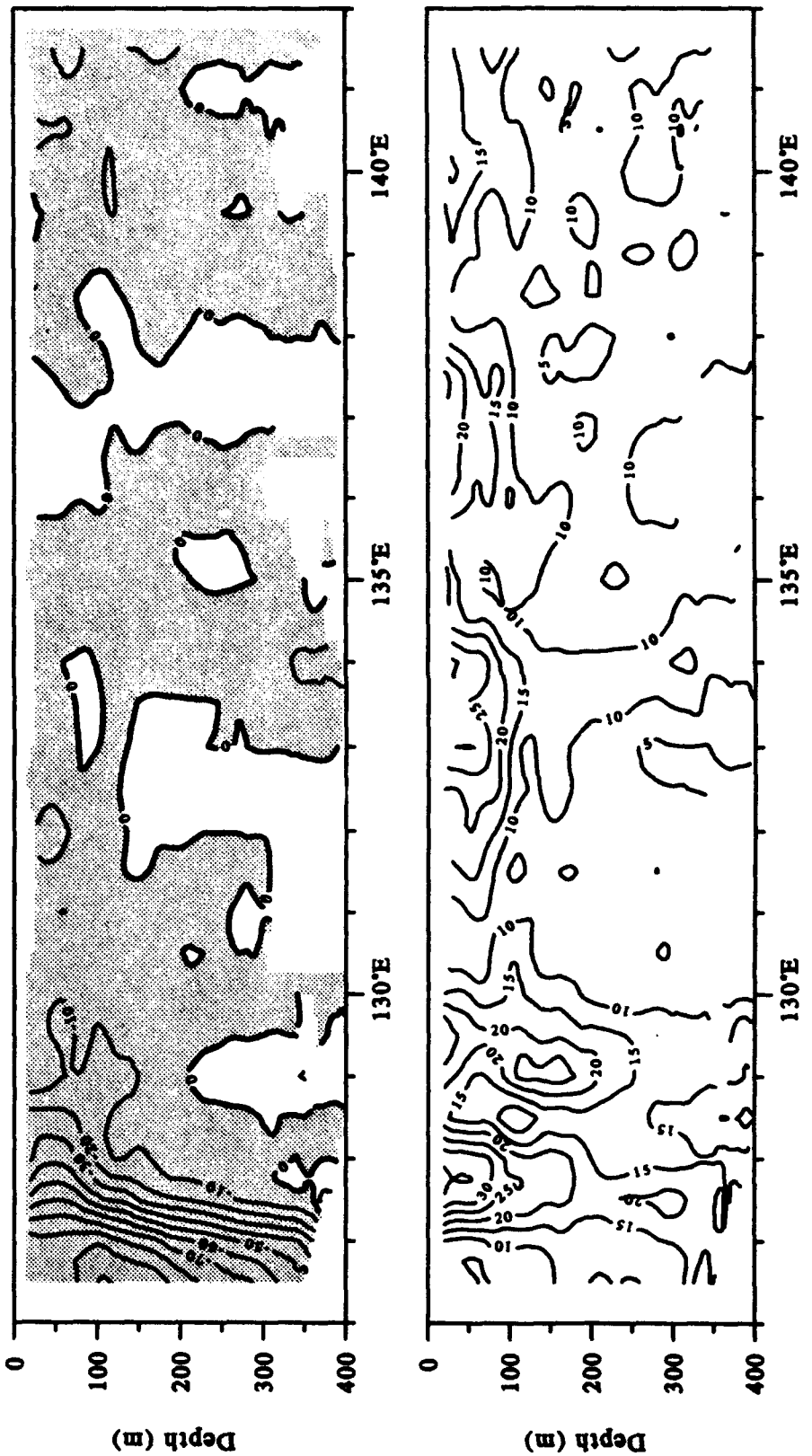


Figure 4. Mean (top panel) and standard deviation (bottom panel) meridional current component from five cruises in the US/PRC TOGA program, 1987 to 1990. Southward flow is shaded. The contours end at the Mindanao coast on the left side of each panel.

Typhoon-generated Currents

So far, we have interpreted shipboard ADCP measurements as showing primarily the spatial structure of currents along a section; we have inferred temporal variability only from cruise-to-cruise differences. Given this mindset, we would look at Figure 5 and conclude that there was an extroardinary series of eddies south of Samoa, with a wavelength of 3° and maximum speeds of nearly 100 cm s^{-1}. This conclusion would be wrong.

The *Moana Wave* left American Samoa for New Zealand on February 9, 1990, just six days after Typhoon Ofa passed 60 miles west of Savai'i in Western Samoa. On February 13 the *Moana Wave* track crossed the path of Ofa eight days before, at about 19°S. The strong currents in Figure 5 north of 20°S are near-inertial oscillations excited by Ofa's winds, mainly to the left of Ofa's path where the wind direction rotated anticyclonically (Lien *et al.*, 1993). The wavenumber vector for these oscillations is along Ofa's path, roughly perpendicular to the ship track. Therefore the currents measured from the moving ship can be treated as a time series. Looking at the time series of currents as a function of depth (Figure 6), we see that currents were uniform in the vertical above about 80 m, presumably the mixed layer depth. Substantial energy had propagated below the mixed layer by the time of these measurements; currents at 200 m were at times as strong as, or stronger than, those in the mixed layer. Phase propagation was upward, consistent with downward energy propagation, and there is a corresponding shift to higher frequencies (blue shift) with increasing depth (Price, 1983).

Apart from its interest as an ocean phenomenon, this instance of unusually strong inertial oscillations illustrates a general problem in determining the spatial structure of ocean currents: measurements almost always mix spatial with temporal variability. There is no measurement system in the ocean that can provide broad spatial coverage, high spatial resolution in more than one dimension, and good temporal resolution, all at the same time.

Near-inertial energy can be identified in some shipboard ADCP sections even without extraordinary forcing such as a typhoon. Wijffels *et al.* (1993) calculated frequency spectra of currents from the 10°N *Moana Wave* section (Figure 3). They found a prominent near-inertial peak in the clockwise spectrum of the shear between 20 m and 100 m, and concluded that the zonal wavenumber must be small compared to 2π divided by 650 km, the distance the ship travelled in one inertial period. This seems reasonable if the inertial oscillations are excited by wind fluctuations that also have much longer zonal scales than 650 km. The expected meridional wavenumber is larger than the zonal wavenumber because of the variation of inertial frequency with latitude (D'Asaro, 1989). At low latitudes, the tendency for wind fluctuations to have larger zonal than meridional scales should

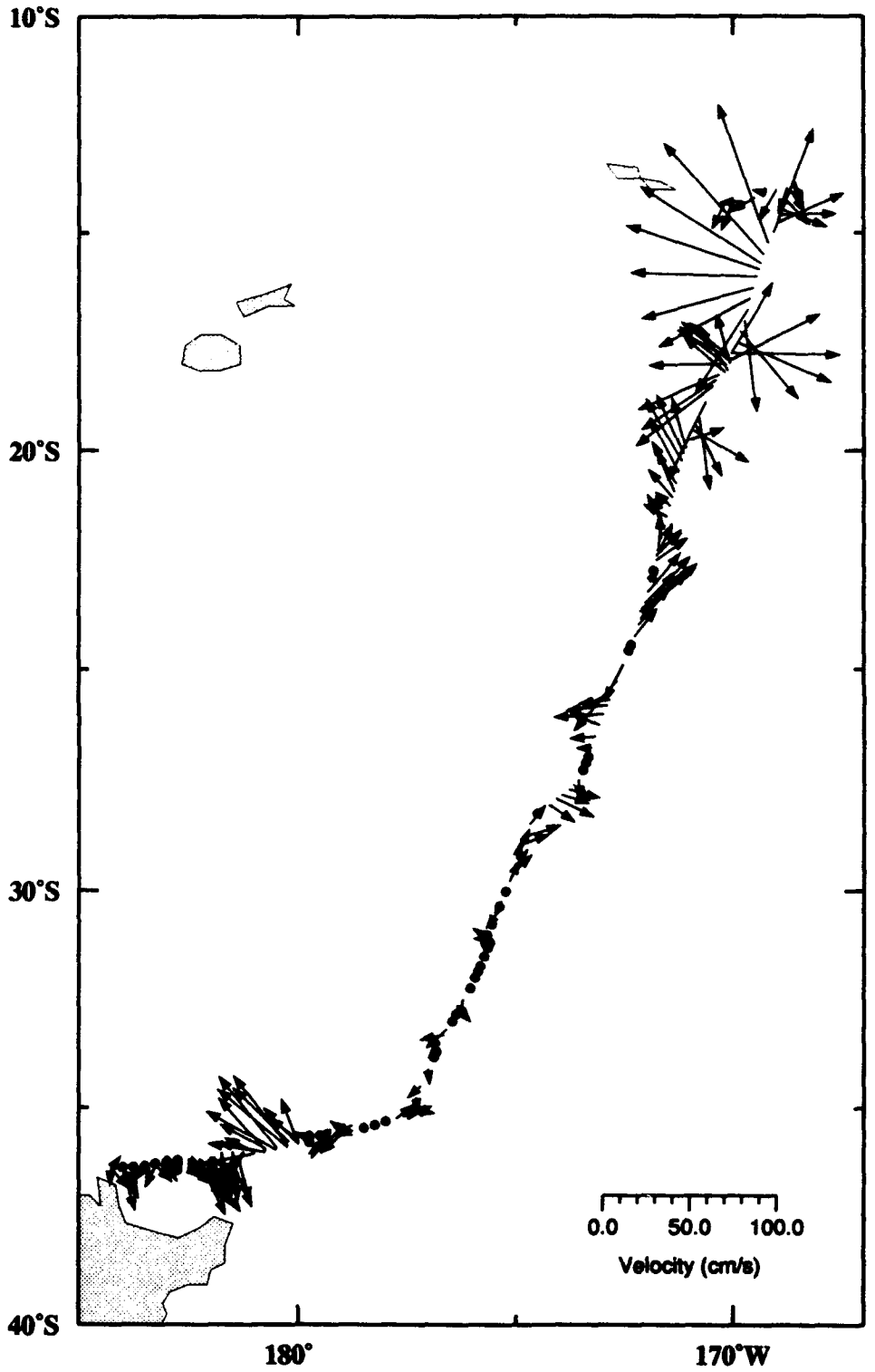


Figure 5. Currents averaged from 25 to 75 m on a *Moana Wave* cruise from Samoa to New Zealand, February 9-26, 1990, immediately following the passage of Typhoon Ofa south of Samoa.

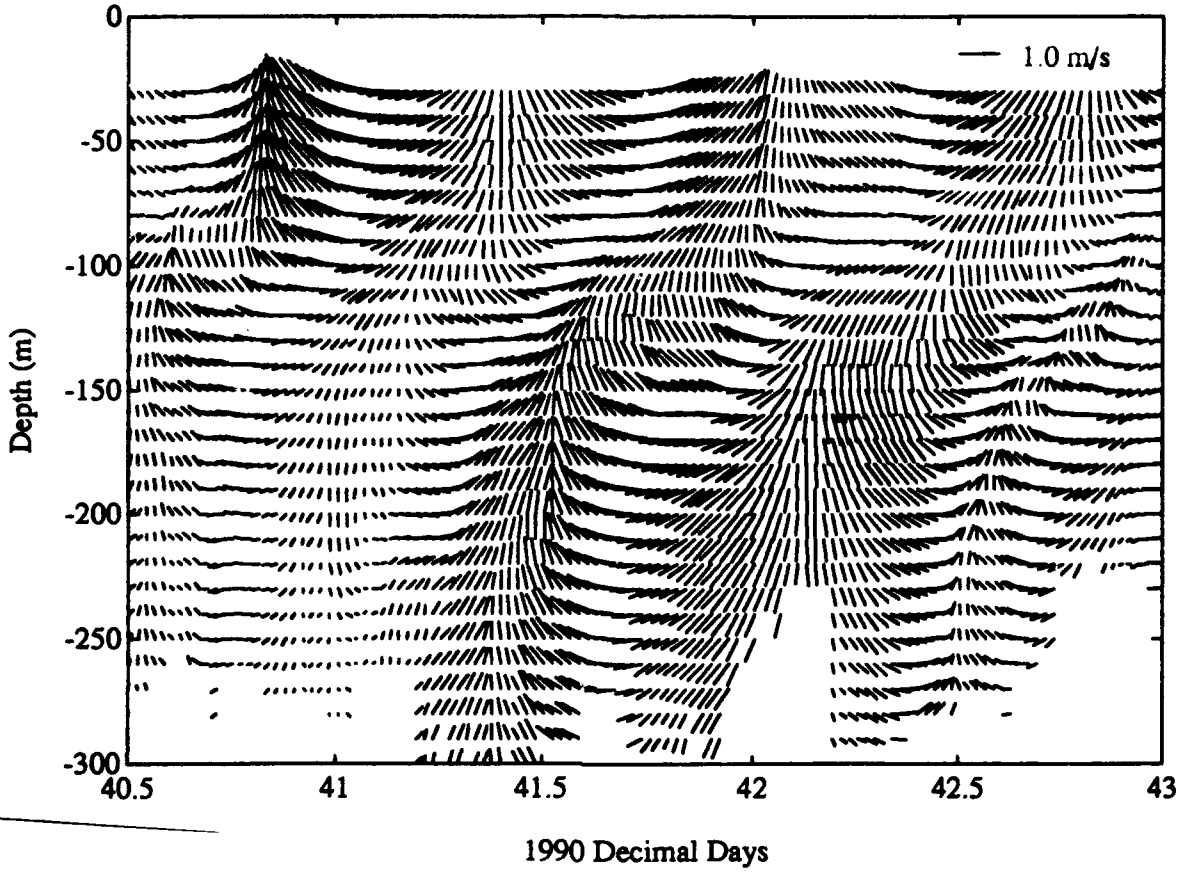


Figure 6. Current vectors (up is northward, to the right is eastward) as a function of time and depth, along the *Moana Wave* cruise track near where it crossed the path of Typhoon Ofa in February 1990.

increase the anisotropy in the near-inertial wavenumber spectrum. This spectrum has not yet been measured definitively, however.

Horizontal Wavenumber Spectra

Having just demonstrated the dangers of interpreting shipboard ADCP sections in terms of spatial rather than temporal variability, we will proceed to do just that—but gingerly, watching out for temporal signals. Two data sets will be used here: the WHP P17 cruise of the *Thompson* (Figure 1); and a cruise of the *Moana Wave* from Pohnpei to Hawaii in July 1990 (MW9009; Figure 7). These are chosen because they include fairly long, nearly zonal sections at different latitudes but within the mid-gyre current regime, where the mean flow is weaker than the eddies.

From the P17 cruise we will use the transect from the California coast to 135°W, nominally along 35°N. The large-scale flow is southward, comprising the general

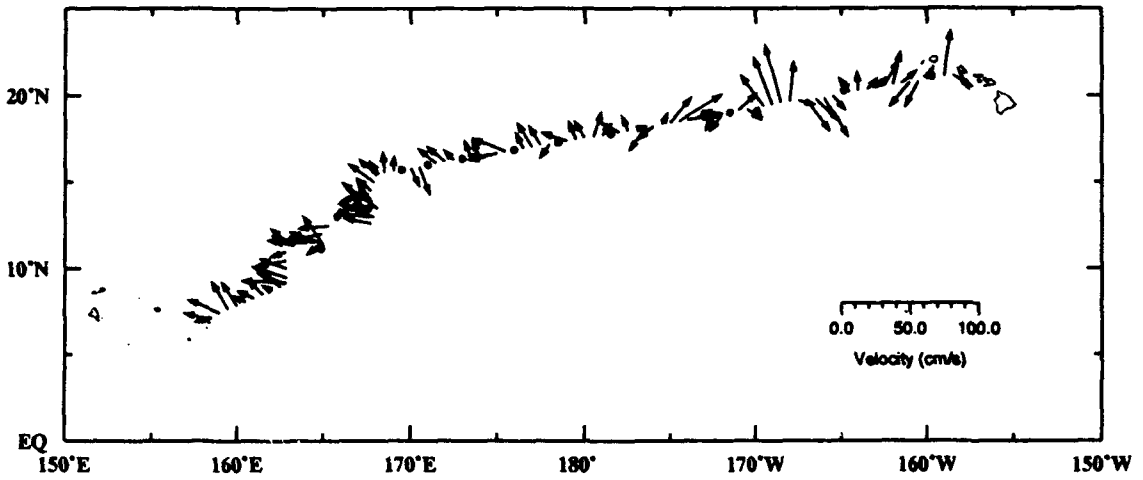


Figure 7. Currents averaged from 25 to 75 m on a *Moana Wave* cruise (MW9009) from Pohnpei to Hawaii, July 9-25, 1990.

southward Sverdrup flow of the gyre plus the California Current (Figure 2). The section is 1230 km long. It was sampled by block-averaging intervals of 0.05° (3.155 km). Including CTD station time, the ship covered the 13° in 6.1 days. The inertial period at 35°N, 20.9 hours, thus corresponds to about 2° wavelength along the track. If the zonal wavelength of the inertial oscillations was much larger than 2°, then the near-inertial spectral peak would appear at 1 cycle per 2° in the wavenumber spectrum of the currents measured from the ship.

From MW9009 we select the relatively straight transit eastward and slightly northward from 16°N 168°E to Oahu, 21°N 158°W. For convenience, we can assign this section a nominal latitude of 18°N. The westward flow of the North Equatorial Current is apparent only on the western half of the section. Like P17, this section was block-averaged in 0.05° longitude (5 km) bins. There were no pauses in the transit, so only eight days were required to cover the 33.8° (about 3570 km) in longitude. The inertial period ranges from 43.5 hours at 16°N to 33.5 hours at 21°N; a 40-hour period corresponds to about 7° along the track.

Horizontal wavenumber spectra were calculated from the Fourier transforms of 128-point segments, overlapping by 64-points. The segments were tapered with a parabolic window (Press *et al.*, 1986). The periodograms for each segment were averaged to yield spectral estimates and normalized so that integrating the single-sided spectral density gives the total variance in the record. There were four segments giving six degrees of freedom for the P17 spectral estimates; and 10 segments giving 16 degrees of freedom for MW9009.

To reduce contamination of the spectra by near-inertial oscillations, the velocity vectors were vertically averaged: from 20–310 m on the P17 section, and from

20–200 m on MW9009 where the ADCP depth range was less. Frequency spectral analysis of the P17 section (not shown) indicates that the 20–310 m vertical average suppresses the near-inertial peak but leaves a semi-diurnal peak. There seems to be no corresponding peak in the wavenumber domain, however, perhaps because the ship was stopped on station for more than half the time. In the time-domain spectra of the MW9009 section there are no clear near-inertial or semidiurnal peaks even in the shear (200 m relative to 20 m), but there is a peak near the diurnal period in both the shear and the vertically averaged velocity. This appears to be due to the eddy field traversed by the ship; if so, the diurnal period is simply a coincidence.

The horizontal wavenumber spectrum at 35°N falls off roughly as k^{-2}, apart from the range 20–40 cptkm (cycles per thousand kilometers) where it rises above the eyeball-fit k^{-2} line by about a factor of 3 (Figure 8). The energy is nearly evenly divided between zonal and meridional components, but over most of the range above 10 cptkm there is an excess of clockwise (moving westward along the track) over counterclockwise energy. Most of this energy is above 20 cptkm, well above the 10 cptkm wavenumber where we might expect the semidiurnal tide to appear in this dataset (unless it is Doppler-shifted substantially). Hence, the cause and significance of the excess in clockwise energy are unknown.

The wavenumber spectrum at 18°N is less energetic than the 35°N spectrum above 20 cptkm, and more energetic below 10 cptkm. Above 25 cptkm the spectral slope is about k^{-2}, but at lower frequencies there is no clear single slope. There is no disparity between the rotary components at high wavenumbers. Below 10 cptkm, meridional energy exceeds zonal energy, and clockwise (eastward along the track) energy exceeds counterclockwise energy.

The analysis given here is intended as no more than a first exploration of the possibility of studying the horizontal wavenumber structure of upper ocean currents with shipboard ADCP data. It shows that in regions of the ocean away from strong mean currents, there are indeed differences in the wavenumber spectra. We suspect that part of the difference shown here between sections at 35°N and 18°N reflects the difference in Rossby radius of deformation: the eddy energy is concentrated at wavelengths near the Rossby radius, which is larger at lower latitudes. Much of the difference, however, is found at shorter wavelengths, and this remains to be explained.

Discussion

The primary theme of this note has been the spatial variability of ocean currents. Vertical shear in the upper few hundred meters varies from almost nil at 60°S to 0.03 s^{-1} near the equator. Large-scale mean currents vary from near 1 m s^{-1} at the

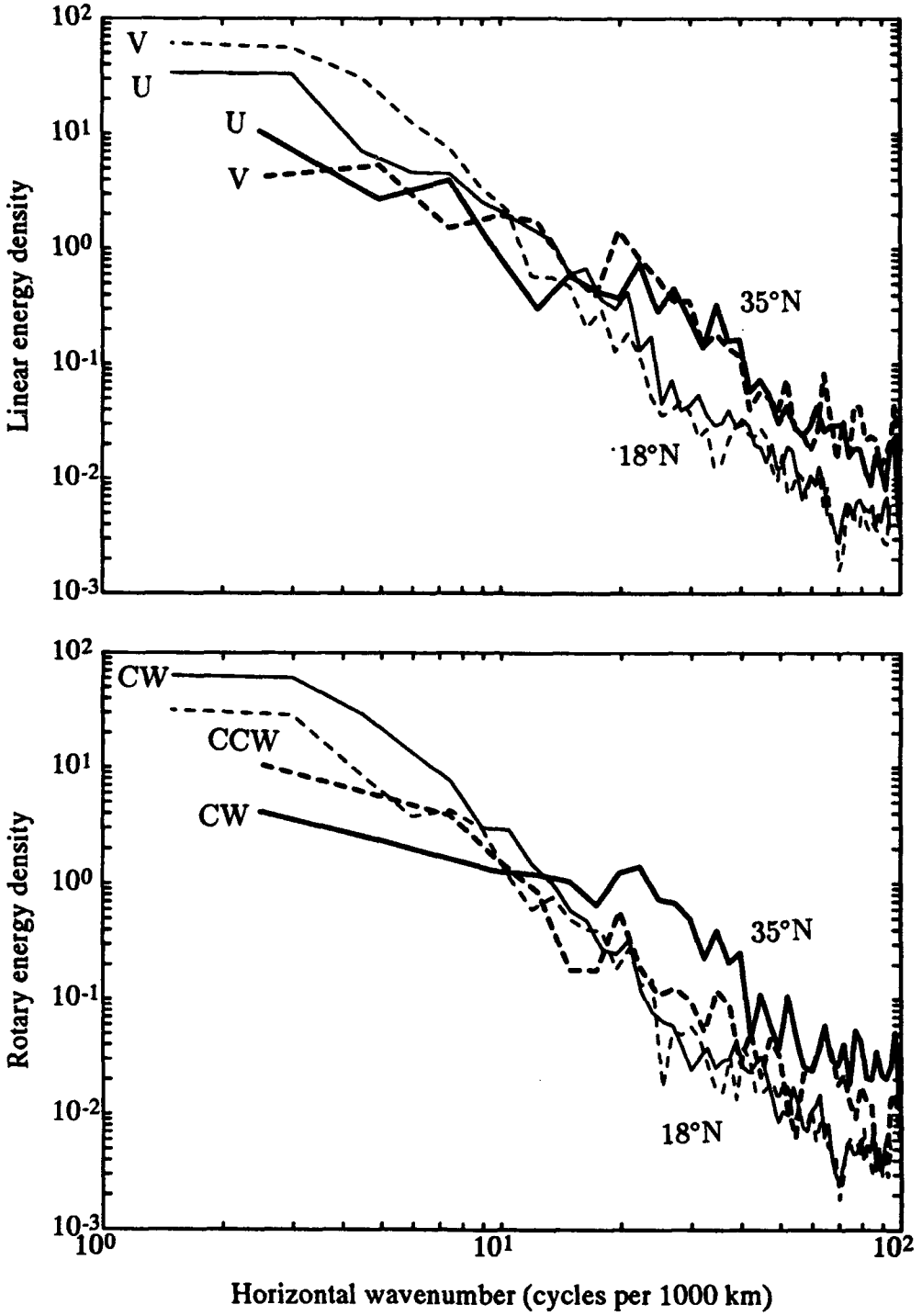


Figure 8. Horizontal wavenumber spectra of currents from a section along 35°N (WHP P17; heavy lines), and from 16-20°N (MW9009; finer lines). In the top panel, solid lines show the spectra of zonal velocity, dashed lines are for meridional velocity. In the bottom panel, solid lines show the clockwise spectra, dashed lines the counterclockwise spectra. Energy density units are m^2s^{-2} per cpm (cycle per meter).

Mindanao coast to less than 0.1 m s^{-1} 400 km offshore. Eddies are ubiquitous, but their typical amplitudes and length scales vary from place to place. Away from strong currents, both amplitude and length scale tend to vary inversely with latitude.

The secondary theme has been the complexity of temporal variability, and in particular the near-inertial oscillations and internal tides. We have shown near-inertial oscillations of nearly 1 m s^{-1}, albeit caused by extraordinary forcing: a typhoon. We have noted the potential danger in looking for an eddy horizontal wavenumber spectrum in shipboard current measurements, inevitably containing inertial and internal waves in addition to the eddies. The danger is reduced but not eliminated by vertical averaging.

Statistical analysis of upper ocean velocity measurements is clearly in its infancy. Even the simplest sorts of analysis—such as calculation of the mean and standard deviation of currents along a single section—have been done only in a very few places and with very few measurements. To my knowledge there has been no comprehensive attempt to characterize the spatial distribution of vertical shear variance. Horizontal wavenumber analysis of current measurements has been attempted rarely. There has been no systematic attempt to extract statistical information about eddies and the internal wave field from the rapidly growing shipboard ADCP data set. The size and quality of this data set are rapidly approaching the point where extensive statistical analysis will be feasible. I expect it will be fruitful as well, shedding light on the small and mesoscale phenomena that until recently have been almost impossible to observe in detail.

Acknowledgments

Thanks to Willa Zhu, Mei Zhou, Frank Bahr, and Julie Ranada for their data processing and plotting, and to Alex Orsi, Frank Bahr, and Sarah Gille for running the ADCP on the WOCE cruises. The cooperation of the technical assistance group at SOEST has allowed the collection of many interesting data sets on the Moana Wave, including two of those used in this paper. Helpful discussions and ideas have come from Susan Wijffels, Peter Muller, and Doug Luther. Peter Hacker has been my partner in all the WOCE work—I would not have even tried it without his help. Funding from the National Science Foundation under grant OCE-9015285 is gratefully acknowledged.

References

Chereskin, T. K., and A. J. Harding, 1993: Modeling the performance of an acoustic Doppler current profiler, *J. Atmos. and Oceanic Technol., 10,* 41–63.

D'Asaro, E. A., 1989: The decay of wind-forced mixed layer inertial oscillations due to the β effect, *J. Geophys. Res., 94,* 2045–2056.

Delcroix, T., G. Eldin, M-H. Radenac, J. Toole, and E. Firing, 1992: Variation of the western equatorial Pacific Ocean, 1986–1988, *J. Geophys. Res., 97,* 5423–5445.

Hansen, D. V., and C. A. Paul, 1984: Genesis and effects of long waves in the equatorial Pacific, *J. Geophys. Res., 89,* 10,431–10,440.

Lien, R. C., E. Firing, and P. Muller, 1993: Observations of strong inertial oscillations after the passage of typhoon Ofa, . in preparation.

Lukas, R., E. Firing, P. Hacker, P. L. Richardson, C. A. Collins, R. Fine, and R. Gammon, 1991: Observations of the Mindanao Current during the Western Equatorial Pacific Ocean Circulation Study, *J. Geophys. Res.. 96,* 7089–7104.

McPhaden, M. J., and B. A. Taft, 1988: Dynamics of seasonal and intraseasonal variability in the eastern equatorial Pacific, *J. Phys. Oceanogr., 18,* 1713–1732.

Patzert, W. C., 1969: Eddies in Hawaiian Waters, *Rep. HIG-69-8,* Hawaii Inst. of Geophysics, Honolulu, HI.

Picaut, J., and R. Tourni r 1990: Monitoring the 1979–85 equatorial Pacific current transports with XBT data, *J. Geophys. Res., 96 Suppl.,* 3263–3277.

Poulain, P.-M., 1993: Estimates of horizontal divergence and vertical velocity in the equatorial Pacific, *J. Phys. Oceanogr., 23,* 601–607.

Press, W. H., B. P. Flannery, Saul A. Teukolsky, and W. T. Vetterberg, 1986: Numerical Recipes , *Cambridge University Press,* Cambridge, 818 pp.

Price, J. F., 1983: Internal wave wake of a moving storm. Part I: Scales, energy budget and observations, *J. Phys. Oceanogr., 13,* 949–965.

Richardson, P. L., and T. K. McKee, 1989: Surface velocity in the equatorial oceans (20°N–20°S) calculated from historical ship drifts, *Woods Hole Oceanogr. Institution Tech. Report,* WHOI-89-9, 50 pp.

Richardson, P. L., and D. Walsh, 1986: Mapping climatological seasonal variations of surface currents in the tropical Atlantic using ship drifts, *J. Geophys. Res., 91,* 10,537–10,550.

Talley, L., and J. Swift, 1992: WHP sampling complete along P16C, P16S, and P17C, *Woce Notes, 4-1,* 1-4.

Toole, J. M., E. Zou and R. C. Millard, 1988: On the circulation of the upper waters in the western equatorial Pacific Ocean, *Deep-Sea Res., 35,* 1451-1482.

Tsuchiya, M., 1975: Subsurface countercurrents in the eastern equatorial Pacific Ocean, *J. Marine Res., 33 (Suppl.),* 145-175.

Whitworth, T., W. D. Nowlin, R. D. Pillsbury, M. I. Moore, R. F. Weiss, 1991: Observations of the Antarctic Circumpolar Current and deep boundary current in the southwest Atlantic, *J. Geophys. Res., 96,* 15,105-15,118.

Wijffels, S., E. Firing, and H. Bryden, 1993: Direct observations of the Ekman balance at 10°N in the Pacific, *J. Phys. Oceanogr.*, submitted.

Wilson, C. D., and E. Firing, 1992: Sunrise swimmers bias acoustic Doppler current profiles, *Deep-Sea Res., 39,* 885-892.

Wyrtki, K., L. Magaard, and J. Hager, 1976: Eddy energy in the ocean, *J. Geophys. Res., 81,* 2641-2646.

SATELLITE ALTIMETRY: ATTEMPTS TO PROGRESS BEYOND STUDIES OF THE STATISTICS OF MESOSCALE VARIABILITY

Dudley B. Chelton and Michael G. Schlax
College of Oceanic & Atmospheric Sciences, Oregon State University,
Corvallis, OR 97331

ABSTRACT

Because of uncertainties in the marine geoid and orbit height, most applications of altimetric data have focused on mapping the sea level variance statistic. These studies have been very successful at defining the geographical distribution of eddy variability and have highlighted the close relationship between transient eddies, the intensity of the mean flow and the bathymetry. Altimeter data have also been used to estimate surface geostrophic velocities and map the variance of geostrophic velocity (or, equivalently, the geostrophic Reynolds stresses). These studies have demonstrated the importance of the transport of horizontal momentum into the mean flow by transient eddies. Other obvious applications of altimeter data include mapping the time evolution of the sea level field for studies of wind and buoyancy forced ocean circulation and descriptive studies of mesoscale processes such as meandering and ring formation. Such applications are limited by a number of difficult technical challenges, mostly related to uncertainties about what space and time scales can be resolved by the complex space-time sampling characteristics of satellite data. A method is presented here for identifying aliasing patterns in an arbitrary sample design and for quantifying the resolution capability of the data set. Although the discussion emphasizes altimeter data, the method is applicable to any irregularly sampled data set. The maximum resolution capability of the GEOSAT orbit configuration (neglecting measurement errors and data dropouts) is found to be about 3° in latitude and longitude by 30 days.

1. INTRODUCTION

The TOPEX altimeter launched in August 1992 is the fifth in a series of altimeter satellites that have measured the global sea surface topography for studies of ocean circulation. The vast majority of applications of altimeter data have focused on the statistics of mesoscale variability. As discussed in section 2, this is because the effects of uncertainties in the orbit height and the marine geoid can be greatly mitigated if the interest is restricted to sea level variance statistics. In recent years, there has been an increasing interest in using altimeter data to map the time evolution of sea level in order to investigate the detailed spatial and temporal structure of sea level variations on a wide range of scales and relate them to wind and buoyancy forcing. Although examples can be cited from the literature of attempts to construct quasi-synoptic maps of mesoscale eddy fields with ~50 km spatial resolution from altimeter data, it should be obvious that the information content of altimeter data alone is not sufficient to do this; because of the asynoptic sampling and relatively coarse spacing (100–300 km) of the satellite ground tracks,

there are lower limits to the space and time scales that can be resolved by the data. To date, the choice of scales mapped has been rather ad hoc, with few attempts to assess the accuracy of the mapped fields.

The objective of this study is to present a technique for quantifying the space and time scales that can be resolved by an irregularly sampled data set. Although the particular interest here is to determine the resolution capability of altimeter data, the method is equally applicable to any irregularly sampled data set. The discussion in this paper is limited to the GEOSAT altimeter, which is the altimeter data set that has received the most attention because of its long (compared with other altimeter missions) 2-year duration.

It must be conceded at the outset that, because of asynoptic sampling and incomplete spatial coverage, some degree of smoothing is required to construct sea level maps from altimeter data. The technique presented here offers a method for deducing the minimum smoothing necessary to avoid undesirable error characteristics in the mapped fields. For multidimensional data sets such as altimeter data, the best choice of smoothing is complicated by the fact that there is a resolution tradeoff; high resolution in one of the dimensions can be achieved by reducing the resolution in the other dimensions. For example, high resolution in time can be obtained by sacrificing spatial resolution. Similarly, high resolution in space can be obtained at the cost of low temporal resolution. Alternatively, high spatial resolution in one dimension can be achieved at the cost of low spatial resolution in the other dimension. The best choice of the tradeoff between spatial and temporal resolution will depend on the intended application.

A brief summary of previous oceanographic applications of altimeter data is given in section 2. The section concludes with a statement of the need to quantify the resolution capability of altimeter data in order to construct meaningful maps of the time evolution of sea level fields. A method for quantifying the inherent wavenumber-frequency filtering characteristics of an irregularly sampled data set is given in section 3. The filter transfer function depends on the particular sampling characteristics of the data set and on the choice of smoothing parameters used to construct the maps. The utility of the transfer function is demonstrated by application to 1-dimensional examples and to the GEOSAT data. An expression for the errors of the smoothed fields is derived in section 4 in terms of the transfer function of the data set and the spectral characteristics of the field. The transfer function and error formalisms are applied in section 5 to determine the resolution capability of a 1-dimensional example and of the actual GEOSAT data.

2. SUMMARY OF PAST ALTIMETER STUDIES

2.1. The Measurement Technique

The measurement of sea surface topography by satellite altimetry is summarized schematically in Figure 1. Although the altimeter measurement of range h is straightforward in principle, it is very complex in practice, involving more than 50 computer algorithms to correct for instrumental effects, atmospheric refraction and biases introduced by the interaction between the electromagnetic radar pulse and the air-sea interface. It is remarkable that the accuracy of the range estimates after applying these corrections is better than one part in 10^7. The range measurements alone are not sufficient for oceano-

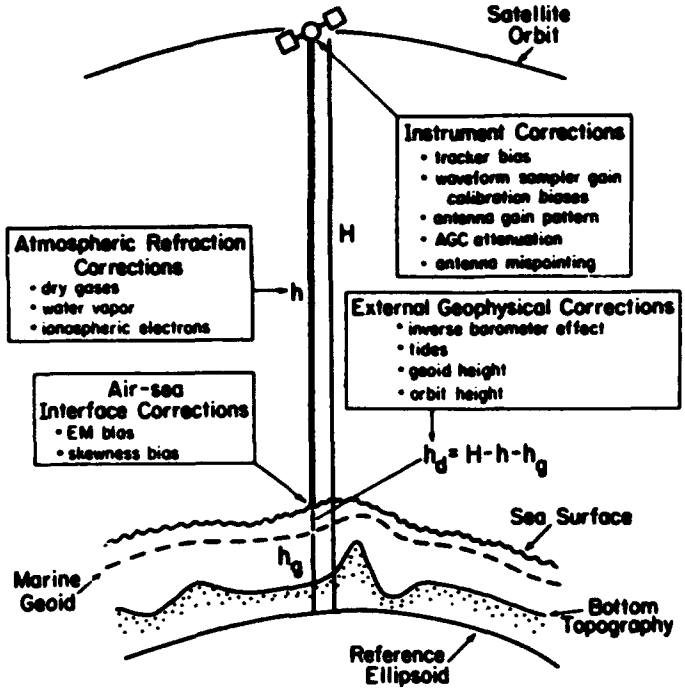


Figure 1. Schematic representation of altimeter measurements.

graphic applications; there are several contributions to the range measurements that are not part of the oceanographic signal of interest. It is therefore necessary to apply several additional external geophysical corrections to obtain the dynamic sea surface topography h_d associated with geostrophic ocean circulation. A detailed discussion of the range and external corrections is beyond the scope of this study; the interested reader is referred to Chelton (1988) and Chelton et al. (1989).

By far the largest source of error in altimeter estimates of sea surface topography is the correction for the geoid height h_g. The dynamic range of the geoid is almost 200 m globally, which is about two orders of magnitude larger than the global dynamic range of the oceanographic sea surface topography. Uncertainties in the geoid height are presently about 30 cm, which is comparable to the magnitude of the oceanographic signal. The geoid problem can be essentially eliminated if interest is restricted to studies of time-varying sea surface topography. Because temporal variations in the earth's gravity field are negligible over the duration of an altimeter mission, the geoid signal at each location along an exactly repeating ground track can be estimated as the time average of sea level over all repeat orbits by the so-called collinear analysis method (see, e.g., Appendix A of Cheney et al., 1983; sections 4.2 and 4.3 of Chelton et al., 1990). Regrettably, this time average also includes the time-invariant contribution to sea level from the mean ocean circulation but this signal must be sacrificed in order to eliminate the geoid problem. Ignoring the small errors introduced by the ± 1 km lateral variations of the repeating orbits, the geoid and mean ocean circulation contributions can be removed and the time-varying sea surface topography can be investigated from the residual sea level signal.

The second largest source of error in altimeter data is the correction for the satellite orbit height H. Until recently, uncertainties in the orbit height have been about 50 cm. Preliminary analyses of TOPEX data have shown that advances in precision orbit determination have reduced the orbit errors to less than 10 cm. As impressive as this accuracy is, there is still a need to estimate and remove these residual orbit errors from the altimeter data for most oceanographic studies. The spectral characteristics of orbit errors are dominated by variability at 1 cycle/rev (Wagner, 1989). If the interest is only in mesoscale variability (wavelengths shorter than about 1000 km), the very long wavelength orbit errors can be approximated and removed from the data by least squares polynomial fits over data arcs of 2000–3000 km (Zlotnicki et al., 1989; Tai, 1989; 1991). For studies of sea level variability on larger scales, the orbit errors are more appropriately modeled as sinusoids with a frequency of 1 cycle/rev (Chelton and Schlax, 1993).

The overall accuracy of altimeter estimates of the time-varying component of sea surface topography after applying all of the corrections and removing the geoid and orbit errors is probably 6–8 cm for the GEOSAT altimeter, although this is difficult to quantify. Because of significant improvements in the atmospheric refraction corrections and the orbit height estimates, the overall accuracy of the TOPEX data is likely to be smaller by about a factor of two.

While the estimation of sea level by satellite altimetry is much more technical than that by tide gauges, there are a number of problems that altimeter and tide gauge data share in common. All of the external geophysical corrections that must be applied to altimeter data must also be applied to tide gauge data. The primary distinction between the two methods of sea level estimation is that most of the unwanted contributions to the sea level measurements are easier to remove from tide gauge data. For example, nearly all of the tidal signal can be removed by low-pass filtering the tide gauge data, which are typically sampled at hourly intervals. Because the altimetric estimates of sea level at a given location are sampled at widely spaced intervals of 3–35 days (depending on the satellite orbital configuration), low-pass filtering is not possible. The tidal signal must therefore be removed from altimeter data on the basis of model estimates of the various tidal constituents. The present global accuracy of tidal models is believed to be 5–10 cm rms in the open ocean (Ray, 1993). The correction for atmospheric pressure loading (the "inverse-barometer effect") is also easier for tide gauge data since measurements of atmospheric pressure can usually be obtained from a nearby barometer. Here again, altimeter data require model estimates of sea level pressure since the altimeter observations are globally distributed but atmospheric pressure data are available only at discrete locations.

The correction for geoid contributions to the sea level signal are equally difficult for altimeter data and tide gauge data. There are few cases where tide gauges have been geodetically levelled to a common reference. Although levelling can now be done using astronomical techniques, it is a costly procedure and not likely to be done in the near future for the global tide gauge network. As with altimeter data, the geoid problem for tide gauge data can be avoided if interest is restricted to studies of the time variability of sea level; the time-averaged sea level can be removed from each tide gauge record.

Even the orbit error problem of altimetry has an analog in tide gauge data. The level of a tide gauge relative to a fixed reference can vary with time. Although there are examples of abrupt changes in the tide gauge datum level from catastrophic events such as

earthquakes, the sudden collapse of a pier, or relocation of the gauge, most of the vertical motion of the gauge is associated with very slow crustal uplift or subsidence. These secular signals are easily identified and removed from tide gauge data by simple statistical techniques.

2.2. Mean Sea Surface Topography

Determining the surface geostrophic general circulation of the ocean from the mean dynamic topography of the sea surface has long been an important objective of satellite altimetry (Wunsch and Gaposchkin, 1980). When combined with hydrographic data, knowledge of the absolute sea surface topography obtained from altimetry would solve the reference level problem of the dynamic method for computing geostrophic velocity from hydrographic data. This is one of the primary stated goals of the TOPEX mission. It is also the most challenging goal of the mission because it places the most stringent demands on the accuracy requirements of each of the many measurement components needed to determine the dynamic sea surface topography.

As noted in section 2.1, the two largest sources of error in altimeter data are uncertainties in the geoid height h_g and the orbit height H, both of which, until recently, have been known only to an accuracy of about 50 cm. This is comparable to the amplitude of the dynamic topography signal of interest. Orbit height errors have decreased to about 10 cm for the TOPEX data that are beginning to become available. Geoid errors have similarly decreased by constructing a global geoid from combined terrestrial gravity measurements and satellite tracking data using the method described by Rapp and Pavlis (1990) (see Rapp et al., 1991). The result of this analysis is a global model for the geoid height, expressed as an expansion of the spherical harmonic functions, with an estimated rms error of about 30 cm (Rapp, 1992). The geoid accuracy is not likely to improve much beyond this without a low-altitude dedicated gravity-mapping satellite mission. With present technology, it is possible to map the geoid with 100 km spatial resolution to an rms accuracy of about 3 cm by satellite. Several such missions have been proposed internationally over the past decade but none have yet reached approval for a new start. Until such a geoid model becomes available,, studies of the general ocean circulation will be limited to the very large scales that are known accurately in presently available gravity fields.

The approach that has been used most commonly to estimate the mean dynamic topography from altimeter data first subtracts the range measurements h from the estimated satellite orbit heights H (see Figure 1) to obtain an estimate of the total sea surface height at each measurement location. These sea surface height estimates are then adjusted to mitigate the effects of time-dependent orbit errors by a least squares procedure that approximates the predominantly 1 cycle/rev orbit errors as low-order polynomials or sinusoids as discussed in section 2.1. The adjusted sea surface heights are then interpolated to a regular grid along the satellite ground track and a gridded mean sea surface is computed as the arithmetic mean of all repeat estimates of the adjusted sea surface height at each grid location. The mean sea surface constructed in this way includes the geoid height, the mean dynamic topography, the geographically correlated orbit error (defined here to be the time-invariant component of orbit error that is the same for each repeat sample of a given ground track) and any time-invariant measurement errors.

Because the geoid is expressed as a global spherical harmonic expansion, the method generally used to estimate the mean dynamic topography has been to expand the adjusted altimetric mean sea surface as a spherical harmonic expansion of the same low degree and order to which the geoid is known accurately. The geoid height h_g expanded to this low degree and order is then subtracted from this low-pass filtered mean sea surface. The accuracy of the resulting estimate of the low-order spherical harmonic expansion of the mean dynamic topography h_d depends not only on the accuracy of the geoid estimate at these large scales but also on the magnitudes of the geographically correlated orbit errors and time-invariant measurement errors that are included in the sea surface height estimates.

An example of the application of this straightforward approach by Tai (1988) is shown in Figure 2a based on three months of SEASAT data expanded to degree and order 8. For comparison, the mean sea surface dynamic topography relative to 2250 db computed by Levitus (1982) from 80 years of historical hydrographic data is shown to the same degree and order in Figure 2b. It is immediately apparent from the hydrographic data that this low degree and order expansion, which corresponds to wavelengths longer than about 5000 km, provides only a crude representation of the true dynamic topography. Even the major gyre structures are only schematically present at these long wavelengths. Higher order terms of the spherical harmonic expansion are necessary to resolve the strong dynamic height gradients associated with intense currents such as the Gulf Stream.

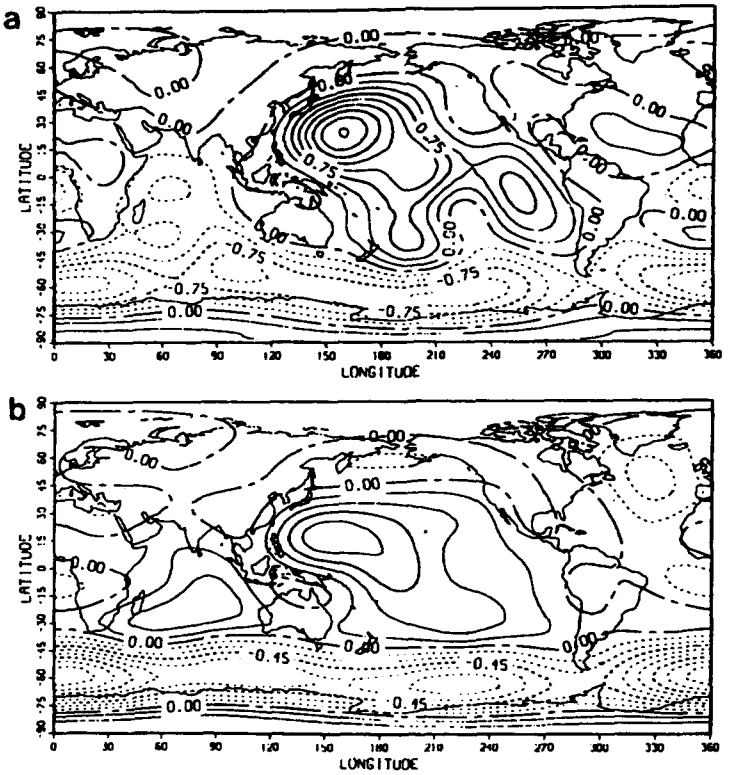


Figure 2. Spherical harmonic expansions to degree and order 8 of a) the mean sea level computed from 3 months of SEASAT data with the GEM-T1 geoid removed; and b) the Levitus (1982) surface dynamic height field (from Tai, 1988.)

It can also be seen from Figure 2 that there are large discrepancies between the altimetric and hydrographic estimates of the mean dynamic topography. Most notable is the region of high sea level in the altimeter data centered at about 15°S, 250°E in the eastern Pacific. There is also a region of low sea level in the altimeter data from the Indian Ocean. Although the differences in some regions such as the poorly sampled areas of the southern hemisphere may be attributable to errors in the hydrographic data, it is more likely that the large amplitude features in the eastern Pacific and Indian Ocean arise primarily from geoid errors and geographically correlated orbit errors. Tai (1988) argues that the accuracy of geoid models has improved to a point where orbit errors are now the dominant source of error in altimeter estimates of the mean dynamic topography. Large differences between the mean sea surface height estimates constructed separately from ascending and descending ground tracks at the crossover points attest to the presence of large geographically correlated orbit errors in the eastern Pacific and Indian Ocean. These orbit errors can be attributed to the poor ground-based tracking coverage along the ground tracks that pass over these regions.

Nerem et al. (1990) and others have attempted to reduce the effects of geoid and orbit errors on altimetric estimates of the dynamic topography (and at the same time improve estimates of the geoid height) by simultaneously estimating the mean dynamic topography, the geoid height and the orbit errors using a least squares inversion procedure first suggested by Wagner (1986). Compared with the earlier estimate by Tai (1988) shown in Figure 2a, the joint-solution estimate of mean dynamic topography to degree and order 10, computed from 51 days of GEOSAT data, is in closer agreement with the low-pass filtered dynamic topography from hydrographic data. The primary reason for the improvements in the GEOSAT-based mean dynamic topography when compared with the earlier estimates from SEASAT data is likely the explicit inclusion of orbit errors in the joint solution. Nonetheless, there are still large differences between the altimetric and hydrographic dynamic topographies. For example, the Atlantic Ocean gyre structure is much different in the two data sets and there are very large discrepancies in the Indian Ocean. In the Pacific Ocean, the gyre centers are displaced to the east in the altimetric data.

With an unprecedented orbit accuracy of better than 10 cm rms, TOPEX data have introduced a new era in absolute sea level determination by satellite altimetry. The dramatic improvement in the accuracy of the TOPEX orbits compared with previous altimeter satellites is primarily attributable to more complete ground-based tracking coverage and improved orbit modeling because of the reduced drag and gravitational effects on the satellite at the higher 1300 km TOPEX orbit altitude (compared with the 800 km SEASAT and CEOSAT orbit altitudes). Orbit errors are no longer the largest source of error in the mean dynamic topography constructed from altimeter data; errors in the geoid height are now the primary limitation. Because the geoid is still known most accurately at the largest scales, accurate altimetric estimates of the mean dynamic topography will continue to be limited to low degree and order terms in a spherical harmonic expansion. The challenge facing oceanographers is to develop data assimilation techniques that are able to utilize this large-scale information to constrain ocean models.

2.3. Variance Statistics

The geoid and geographically correlated orbit errors that limit the accuracy of absolute sea level determination by satellite altimetry are of relatively little concern for studies of sea level variability. As discussed previously, the time-invariant geoid and geographically correlated orbit errors (as well as any time-invariant measurement errors) are included in the mean sea level computed from repeat-track altimeter data. This mean sea level is removed for altimetric studies of sea level variability. After removing the mean sea level or as part of the mean sea level estimation (van Gysen et al., 1992; Chelton and Schlax, 1993), the time-dependent orbit errors are estimated and removed by one of the least squares techniques outlined in section 2.1. For exact repeat orbits, it is then a straightforward procedure to compute variance statistics from the residual sea level estimates; the sea level variance is computed as the arithmetic average of the squared sea level residuals at each grid location.

Global sea level variability has been calculated from 12 months of exact-repeat GEOSAT data by Zlotnicki et al. (1989) (Figure 3a). All of the major ocean currents are clearly delineated from the unique global perspective afforded by altimeter data. The regions of highest mesoscale sea level variability are coincident with the axes of the Gulf

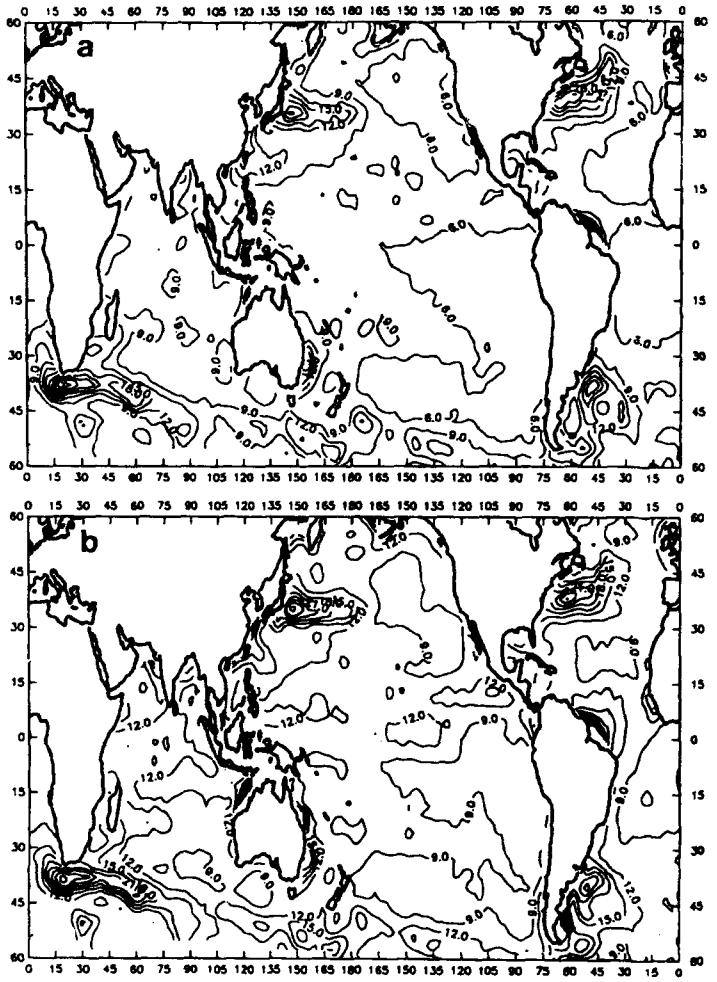


Figure 3. Global rms sea level variability computed from a) one year of GEOSAT data with quadratic orbit error corrections over 2500 km data arcs (courtesy of Zlotnicki); and b) two years of GEOSAT data with sinusoidal orbit error corrections over 4.5 consecutive orbits (see Chelton and Schlax, 1993).

Stream, the Kuroshio and the Antarctic Circumpolar Current. Sea level variability is also high in the southwest Atlantic at the confluence of the Brazil and Malvinas Currents, and in the East Australia Current. High mesoscale variability in these regions is not unexpected in view of the fact that they are all known to be regions of hydrodynamically unstable flow. The sea level variations are associated with transient eddies and meanders in the flow.

As shown by Zlotnicki et al. (1989), the amplitude of mesoscale variability deduced from altimeter data is sensitive to the method used to estimate the time-dependent orbit errors. The Zlotnicki et al. (1989) map of rms sea level variability in Figure 3a was obtained using second-order polynomial orbit error corrections over 2500 km data arcs. For comparison, the rms sea level variability derived from two years of GEOSAT data based on the long-arc (multiple orbital revolutions) sinusoidal orbit error corrections of Chelton and Schlax (1993) is shown in Figure 3b. The patterns of sea level variability are the same in both figures. However, the rms variability is larger nearly everywhere by a few centimeters in the long-arc data. While some of this additional energy is real ocean variability that has been removed by the short-arc polynomial orbit error approximations, some of it is likely attributable to the larger residual orbit errors and other measurement errors in the long-arc data discussed by Zlotnicki et al. (1989). More accurate orbit estimates and geophysical corrections such as those now available for TOPEX data will enable a partitioning of this variability between ocean signal and measurement errors.

Although sea level variance studies have been very useful for mapping the geographical distribution of mesoscale energy, they yield little insight into the detailed statistical characteristics of eddy variability. The spatial scales of mesoscale variability can be investigated from the wavenumber distribution of sea level variance. This is easily determined from 1-dimensional wavenumber spectra of altimeter data along the satellite ground track. Altimetry is the only observational technique that can provide such information because of the difficulty in obtaining synoptic profiles of sea level from in situ measurements.

Le Traon et al. (1990) analyzed the 2-year GEOSAT data set and computed wavenumber spectra of sea level variability for nineteen areas in the North Atlantic. The GEOSAT measurement errors of 3–5 cm allow the resolution of scales as short as about 50 km. The spectra for six regions along 35°N are shown in Figure 4. In the energetic western portion of the North Atlantic, the sea level wavenumber spectra are relatively flat at low wavenumbers with a broad peak centered at wavelengths of approximately twice the baroclinic Rossby radius of deformation. These peak wavelengths decrease with increasing latitude; peak wavelengths are about 500 km at 25°N, 400 km at 35°N, 300 km at 45°N and 200 km at 55°N. These values are consistent with the baroclinic Rossby radii estimated from historical hydrographic data by Emery et al. (1984). At wavelengths shorter than the Rossby radii, the spectra drop off as approximately k^{-4}, compared with the k^{-5} dependence expected from quasi-geostrophic turbulence theory (Charney, 1971). The weaker slopes in the GEOSAT data are not understood at present.

East of the Mid-Atlantic Ridge where eddy variability is much weaker, the wavenumber dependence of the sea level spectra ranged from about k^{-3} to k^{-1}. In contrast to the western region, the spectra in the eastern basin generally did not flatten at wavelengths shorter than the baroclinic Rossby radius of deformation. This implies that the energy

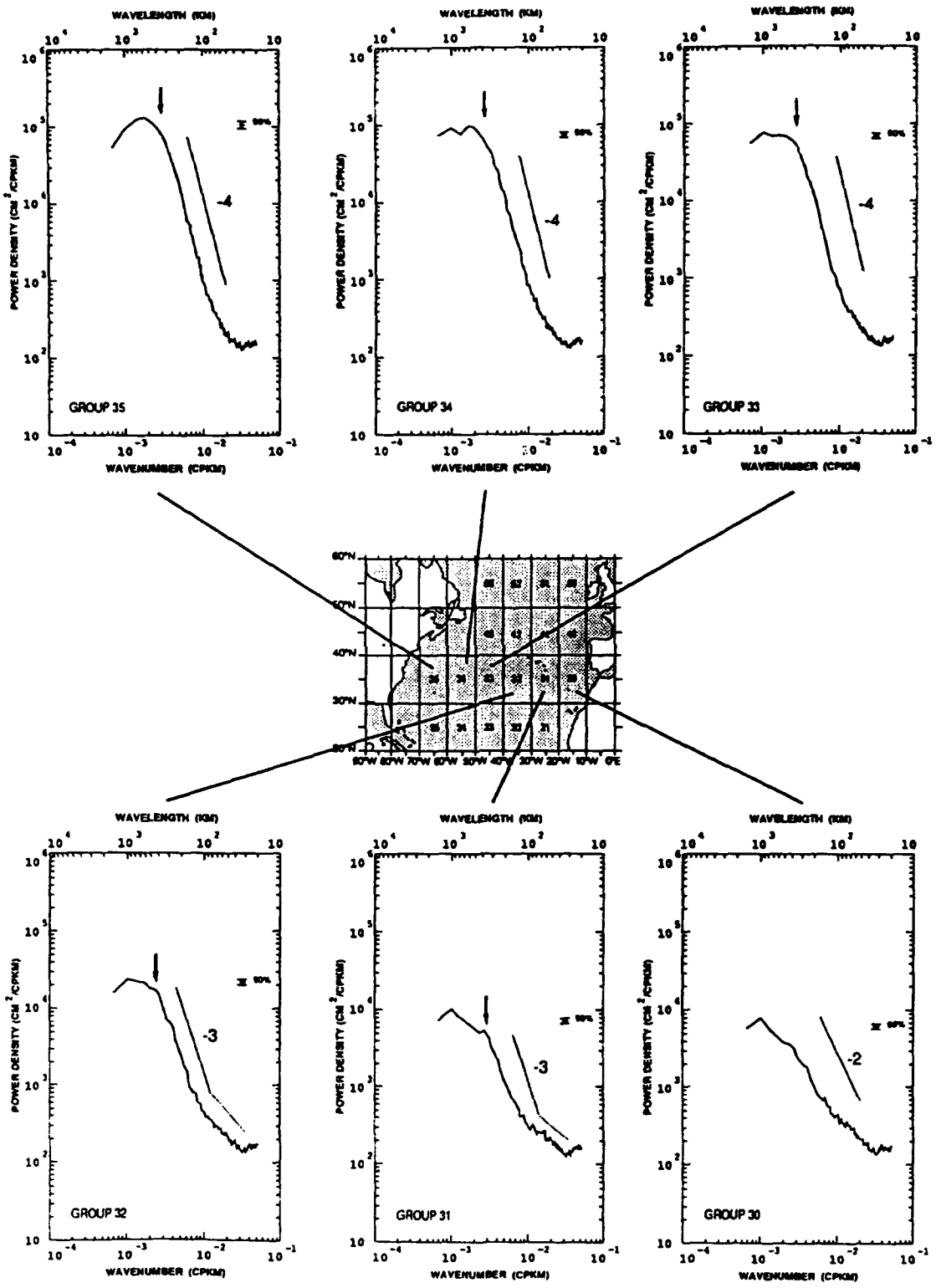


Figure 4. Wavenumber spectra of sea level from the GEOSAT altimeter for six midlatitude regions across the North Atlantic. (From LeTraon et al., 1990.)

source of turbulent variability in the eastern basin is at much longer wavelengths (smaller wavenumbers) than in the energetic western region. Le Traon et al. (1990) suggest that this may be an indication that the eddy variability in these regions of low energy is generated by fluctuating winds, as theorized by Frankignoul and Müller (1979) and Müller and Frankignoul (1981). The wind forcing occurs on much larger scales (order 1000 km) than the forcing by baroclinic instabilities, resulting in a downscale enstrophy cascade from smaller wavenumbers.

The energetics of eddy variability can be investigated from the geographical distribution of eddy kinetic energy. As described by Ménard (1983), this is easily estimated from cross-track geostrophic velocities derived from along-track sea level slopes computed from altimeter data if the eddy variability is assumed to be isotropic. The seasonal variability of eddy kinetic energy estimated in this manner has been investigated globally (with emphasis on the Gulf Stream, Kuroshio and Antarctic Circumpolar Current regions) from two years of GEOSAT data by Shum et al. (1990). From 3-month average estimates of eddy kinetic energy, they find a clear meridional migration of the position of the Gulf Stream extension east of 60°W. The location of maximum eddy kinetic energy shifts northward from the mean location by several degrees of latitude during the summer/autumn and then southward of the mean location by about the same distance during the winter/spring. The magnitudes of the eddy kinetic energy in the Gulf Stream region vary over the two-year record, but not with any clear seasonal cycle. The maps for the Kuroshio region are more difficult to interpret, perhaps because of the larger number of GEOSAT data dropouts in this region. Temporal variations of eddy kinetic energy are small throughout the Antarctic Circumpolar Current region over the 2-year GEOSAT data set.

An important limitation of altimetric studies of eddy kinetic energy from along-track sea level slopes as summarized above is the need to assume isotropic variability. Drifter data support this assumption in regions of low to moderate eddy energy. However, the eddy variability in energetic regions such as western boundary currents and the Antarctic Circumpolar Current is distinctly anisotropic (e.g., Richardson, 1983; Daniault and Ménard, 1985; Johnson, 1989). Morrow et al. (1992) developed a technique for determining the vector surface geostrophic velocity at the intersections of ascending and descending ground tracks. The method involves calculating cross-track velocity components along each of the ground tracks at the crossover locations. The two non-orthogonal components are then converted to orthogonal (north and east) geostrophic velocity components by a simple geometrical transformation first suggested by Parke et al. (1987). The resulting time series of north and east velocity components can then be used to calculate the variances and covariance of the two velocity components, from which velocity variance ellipses that define the principal axes of variability can be derived. A current ellipse with large eccentricity represents highly anisotropic variability with most of the velocity fluctuations aligned parallel to the major axis of the ellipse. Correspondingly, a circular variance ellipse represents isotropic variability with no preferred direction of the velocity fluctuations. The dense distribution of altimeter crossover locations provides a much higher spatial resolution of eddy variability than can practically be obtained from drifter data.

Application of the technique to two years of GEOSAT data in the Southern Ocean reveals energetic, anisotropic surface geostrophic velocity variability in the vicinity of all of the major currents (Figure 5). The orientations of the velocity variance ellipses rela-

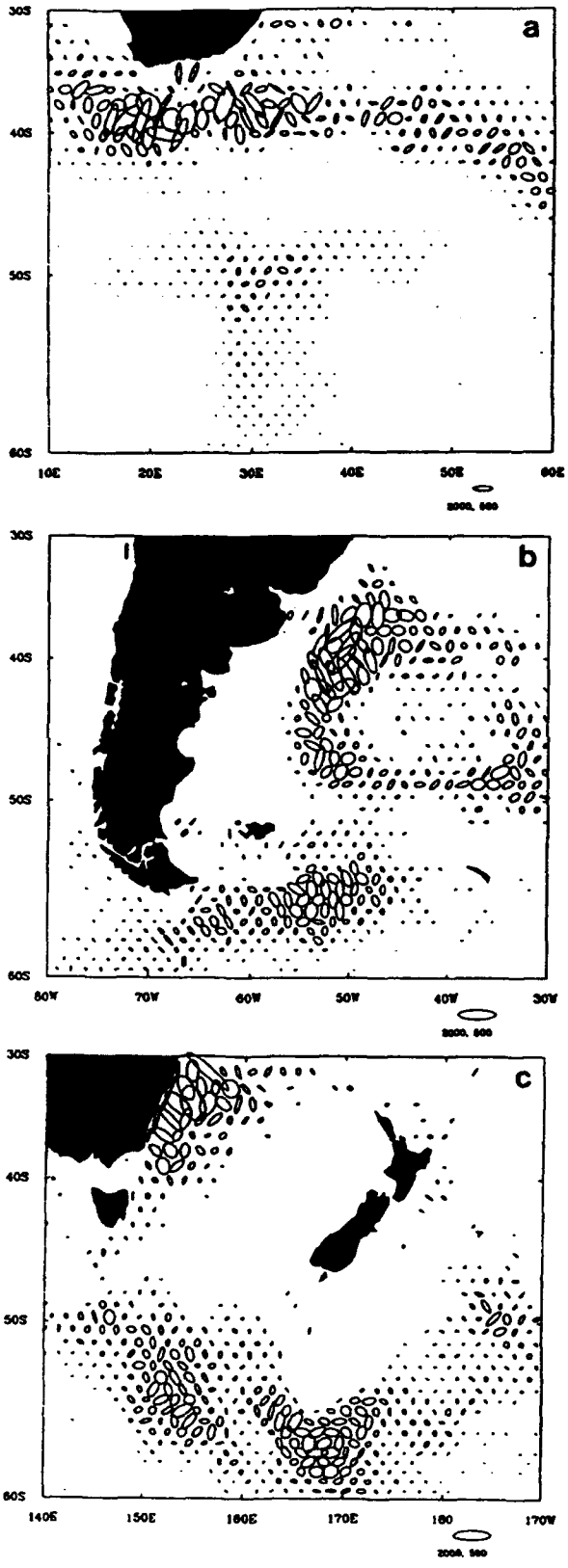


Figure 5. Surface geostrophic velocity variance ellipses from 2 years of GEOSAT data at the crossover locations of ascending and descending ground tracks for a) the Agulhas region; b) the southwest Atlantic; and c) the east Australia/New Zealand region. The scales of the current ellipses in cm^2/s^2 are shown at the lower right corner of each plot. (From Morrow et al., 1993.)

tive to the axis of the mean flow have important implications for eddy transport of horizontal momentum; where the ellipse axes are aligned perpendicular and parallel to the mean flow, there is no cross-stream transfer of along-stream momentum. Eddy momentum fluxes in the Southern Ocean have been quantified by Morrow et al. (1992; 1993) by estimating the gradients of the Reynolds stresses from the variances and covariances of surface geostrophic velocity components. They found a convergence of alongstream momentum in streamwise integrated Reynolds stresses along the mean axis of the Agulhas Return Current. This is an indication that eddy variability in this region tends to accelerate the mean flow, consistent with recent models of the Antarctic Circumpolar Current (Tregieur and McWilliams, 1990; Wolff et al., 1991). The GEOSAT data reveal a surprisingly complex geographical distribution of this Reynolds stress convergence.

The broad range of applications summarized in this section illustrate the significant contributions that altimetric studies of variance statistics for sea level, eddy kinetic energy and surface geostrophic velocity have made toward understanding the dynamics of mesoscale eddy variability. This information cannot be obtained by in situ observational techniques on the scales resolvable by altimeter data. To date, because of the short duration of the SEASAT data set, GEOSAT data have been most useful for these studies. It is an unfortunate fact that the non-scientific primary objective of the mission (high-resolution mapping of marine geoid for defense purposes) resulted in a number of inherent weaknesses in the GEOSAT mission design. Most importantly, there was no onboard microwave radiometer for the wet tropospheric correction, no active attitude control system (resulting in frequent data dropouts), estimates of the ionospheric range correction were inaccurate during the high solar activity that coincided with the period of the GEOSAT mission, and the geographical distribution of unclassified ground-based tracking stations for orbit determination was very sparse. Despite these shortcomings, GEOSAT data have provided very valuable experience with altimeter data, while at the same time yielding important new information about ocean variability. It must be kept in mind, however, that all of the results obtained to date are compromised to an unknown degree by measurement errors with a wide range of space and time scales (see, for example, Jourdan et al., 1990, and Figure 9 of Le Traon et al., 1990). Much improved estimates of mesoscale variability will be possible from the more accurate TOPEX data that are now becoming available.

2.4. Mapped Fields of Sea level Variability

The examples in section 2.3 demonstrate that it is relatively straightforward to compute variance statistics from altimeter data. For many applications, the statistics of the variability are not sufficient. For example, it is of interest to map the spatial and temporal evolution of the sea level field for studies of the dynamics of wind and buoyancy forced ocean circulation. This mapping poses a much more difficult problem than calculating variance statistics. As shown in Figure 6a, the GEOSAT ground tracks map out a diamond-shaped grid on the sea surface. The dimensions of the diamonds at middle latitudes are about 1.5° of longitude by 3° of latitude for the GEOSAT 17-day repeat orbit. These dimensions increase for shorter orbit repeat periods; the dimensions of the diamonds for the TOPEX 10-day repeat orbit, for example, are about 2.7° of longitude by 5.5° of latitude at middle latitudes. Clearly, the spatial structure of mesoscale variability cannot be resolved on all scales by altimeter sampling grids. Features with spatial

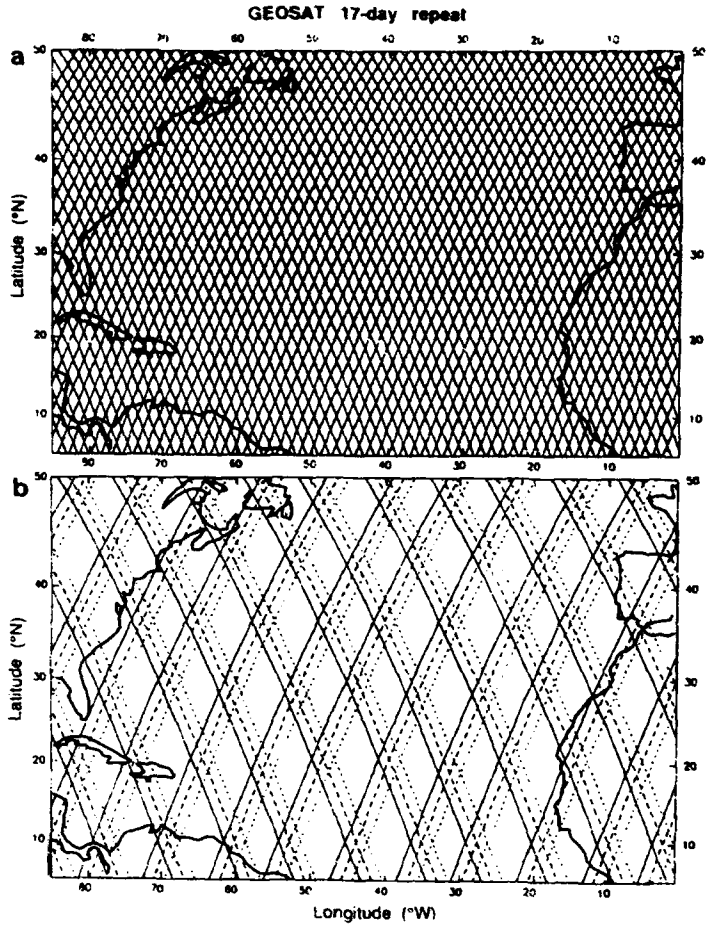


Figure 6. Ground track pattern for a) the full GEOSAT 17-day exact repeat orbit; and b) days 0-2 (solid lines), days 3-5 (dashed lines) and days 6-8 (dotted lines) of the 17-day repeat.

dimensions shorter than roughly a few hundred kilometers are aliased by the ground track pattern.

The aliasing problem is made even more complicated by the asynoptic sampling of the altimeter ground track pattern. As shown in Figure 6b, there is a 3-day subcycle in the GEOSAT sample grid; the ground track mapped out in a 3-day period consists of a coarse resolution diamond-shaped grid with dimensions of approximately 10° of longitude by 20° of latitude at middle latitudes. In each successive 3-day period, the same diamond-shaped pattern is mapped out, but shifted eastward by about 1.5° of longitude each period. The complete GEOSAT ground track pattern in Figure 6a is thus filled in over the 17-day repeat period. This systematic space-time coupling of the sampling characteristics introduces the possibility of the aliasing of propagating sea level features into the mean field as discussed in section 3.4.

A 3-day subcycle is a common characteristic of all exact-repeat altimeter orbit configurations. However, the direction and distance of the 3-day shifts of the coarse resolution grid depend on the details of the orbit configuration. For example, the 3-day subcycle of the TOPEX orbit also shifts eastward, but by about 2.7° of longitude because of the shorter 10-day repeat. In contrast, the 3-day subcycle of the ERS-1 35-day repeat orbit shifts westward by about 1.5° of longitude.

The effects of variability not resolved by the irregular sampling pattern can be mitigated by some degree of spatial and temporal smoothing. As described in section 2.3, removal of the geoid height and orbit errors is much easier for exact-repeat data than for a nonrepeating orbit configuration such as the first two months of the SEASAT mission and the GEOSAT 18-month Geodetic Mission. Mapping fields of sea level variability is therefore greatly simplified from exact-repeat altimeter data using the collinear analysis method described in section 2.1. In addition, the availability of two years of exact 17-repeat GEOSAT data (with a third year of partial coverage) has provided a long enough record of altimeter data to begin to investigate temporal variability of sea level with some (albeit still rather limited) statistical reliability. As a consequence of these two factors, there has been a great proliferation of altimetric studies of large-scale sea level variability.

Numerous studies have documented Kelvin and Rossby wave propagation in the tropical Pacific from collinear analyses of GEOSAT exact-repeat data. As these waves are characterized by much longer zonal than meridional scales, these studies have generally smoothed the data to a resolution of 8–10° of longitude by 1–3° of latitude by one month. An example from Delcroix et al. (1991) is shown in Figure 7. Data from the first year of the GEOSAT exact-repeat mission (November 1986–November 1987) were smoothed 300 km along track and then gridded and smoothed into approximate 10° × 2° × 1 month averages. An eastward propagating downwelling equatorial Kelvin wave characterized by a 15 cm positive sea level anomaly was observed beginning in December 1986, coincident with a strong westerly wind anomaly west of the dateline. Subsequently, an upwelling equatorial Kelvin wave with 10 cm negative sea level anomaly was generated in January–February 1987, coincident with an easterly wind stress anomaly. After arrival of this second Kelvin wave at the eastern boundary of the tropical Pacific in March 1987, a westward propagating baroclinic Rossby wave is evident as equatorially symmetric 12 cm negative sea level anomalies centered at 4°N and 4°S. The surprising result that the earlier downwelling Kelvin wave did not reflect as a Rossby wave is explained by the authors from a model simulation driven by observed winds. They show that the local response to wind forcing in the eastern part of the basin tends to weaken the reflected downwelling Rossby wave, but enhances the reflected upwelling Rossby wave. Owing to the short 1-year record length, the authors are not able to determine whether the observed Kelvin and Rossby waves are associated with the 1986–1987 El Niño or are part of the normal seasonal cycle.

Outside of the tropics, the scales of sea level variability are dominated by eddy dynamics, rather than the wave-like motions in the equatorial waveguide (Robinson, 1983). The appropriate spatial smoothing is thus less well defined than in the tropics. A wide variety of smoothing scales have been adopted in the literature, all of which are rather ad hoc. In some regions, the spatial scales of the eddies are large enough to be resolved by altimeter data. For example, the average diameter of eddies formed by pinching off of the Agulhas Retroflection is more than 300 km (Lutjeharms and Ballegooyen, 1988). Gordon and Haxby (1990) have tracked seven Agulhas eddies from one year of GEOSAT data. After detachment, these eddies drift northwestward into the South Atlantic at 5–8 cm/s (Figure 8). From the distribution of these large eddies, they estimate that about five eddies per year are shed from the retroflection and drift into the Atlantic. These eddies are important to the mass balance of the world oceans; Gordon and Haxby (1990) estimate that they carry as much as $10–15 \times 10^6$ m^3/s of Indian Ocean water into the Atlantic. In

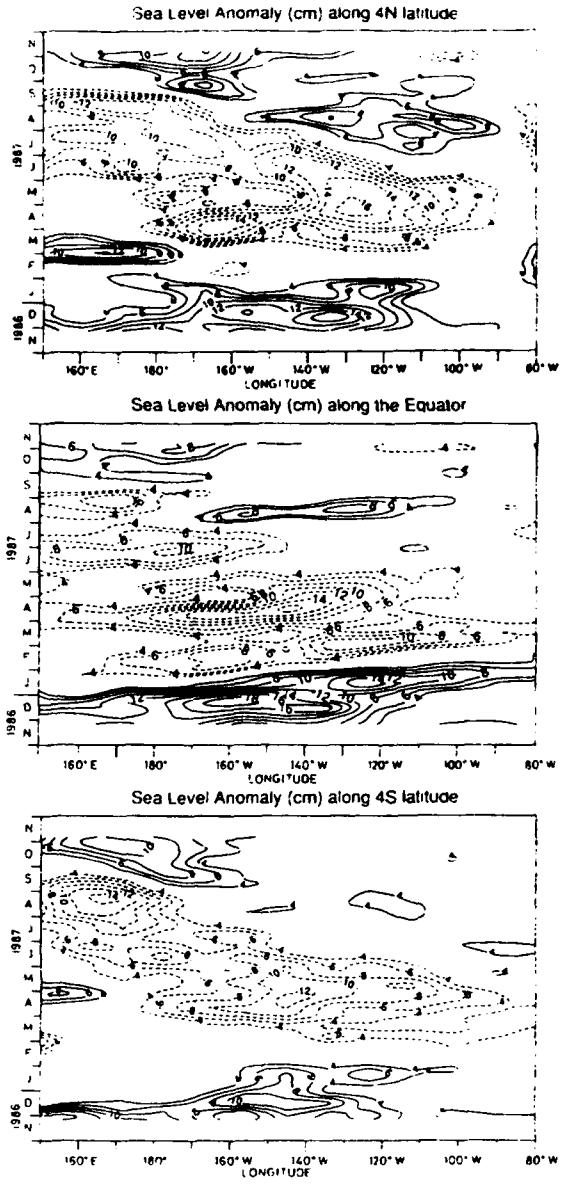


Figure 7. GEOSAT sea level anomalies (deviations from the one-year average) as a function of time and longitude along 4°N, the equator and 4°S (top to bottom). Contour intervals are 2 cm, and the 0 and 2 cm contours have been omitted to highlight the eastward and westward propagation. (From Delcroix et al., 1991.)

addition, Agulhas eddies support a large heat flux from the ocean to the atmosphere as the high sea surface temperatures of these features quickly cool by evaporation.

More generally, the spatial scales of mesoscale eddies are of order 100 km, which is too small to be resolved by altimeter sampling grids. An eddy that is detected as a localized bump in several successive profiles of sea level along a repeating ground track eventually drifts away from the ground track and disappears into a diamond-shaped region bounded by ascending and descending ground tracks. At some later time, the eddy is likely to reappear under a neighboring ground track. Cheney and Marsh (1981) present an example from exact-repeat SEASAT data illustrating the disappearance of an eddy over a 3-week period.

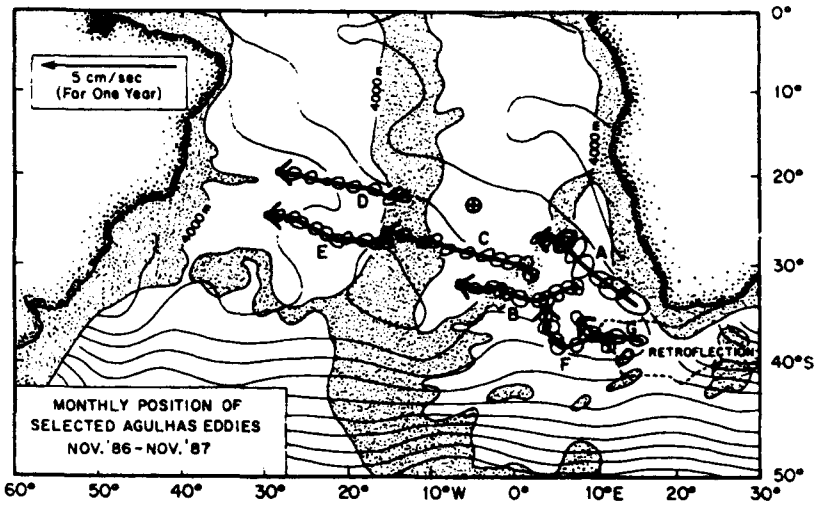


Figure 8. The trajectories of seven eddies in the South Atlantic as determined from one year of GEOSAT data. Eddy positions and approximate sizes are shown as the open symbols at approximately 1-month intervals. Solid dots represent the expected position when an eddy is not clearly evident in the GEOSAT data because of data dropouts or the location of the eddy relative to the ground track pattern. (From Gordon and Haxby, 1990.)

To reduce the geophysical noise introduced by the appearance and disappearance of unresolved eddies, sea level maps constructed from altimeter data must be smoothed over large enough scales to eliminate most of the mesoscale variability (a minimum of a few degrees of latitude and longitude by perhaps a month). Fields constructed in such a manner have been analyzed by time-longitude plots, correlation analysis and frequency-wavenumber spectral analysis to investigate westward propagation along selected latitude lines. Numerous such studies have found surprisingly clear evidence for westward propagation at approximately the annual cycle with phase speeds that lie very close to the dispersion curve for baroclinic Rossby waves (e.g., White et al., 1990; Matthews et al., 1992; Périgaud and Delecluse, 1992; Pares-Sierra et al., 1993; Tokmakian and Challenor, 1993). However, Jacobs et al. (1992) and Schlax and Chelton (1993) have cautioned that aliasing of the M_2 tidal period in the GEOSAT exact 17-day repeat data is manifested as westward propagating variability at near-annual period with a phase speed that very closely matches that of the first-mode baroclinic Rossby wave. Any errors in the model M_2 tidal constituent used to correct GEOSAT sea level data are therefore indistinguishable from Rossby waves. Presently available tide models are believed to be accurate generally to 4–5 cm, but are known to be uncertain by 10 cm or more over large areas of the ocean (Wagner, 1991; Ray, 1993). Consequently, all studies of Rossby wave propagation from GEOSAT data are compromised to an unknown degree by aliasing of M_2 tidal errors. The GEOSAT orbit configuration was thus a particularly poor choice for investigating Rossby wave dynamics. The TOPEX orbit has been carefully selected to avoid aliasing of this nature for any of the major tidal constituents.

The tidal aliasing problem can be reduced by smoothing the GEOSAT data zonally over length scales longer than the wavelength of the M_2 tidal alias (approximately 8° of longitude – see Jacobs et al., 1992). Chelton et al. (1990) examined large-scale sea level variability in the Southern Ocean from two years of GEOSAT data smoothed to a resolution of approximately 12° of longitude by 6° of latitude by 9 days. The variability was dominated by the seasonal cycle, with a zonally coherent annual component and a semiannual component with amplitude and phase that varied over the three major basins of the Southern Ocean. The variability in the South Atlantic has been investigated by

Matano et al. (1993) from sea level fields constructed from GEOSAT data with somewhat higher spatial resolution ($6° \times 3° \times 1$ month). The GEOSAT data show that the confluence of the Brazil and Malvinas Currents migrates seasonally by 2–3° of latitude from a most northerly location in austral winter to a most southerly location during austral summer (Figure 9). Numerical simulations of the wind-forced subtropical gyre of the South Atlantic and GEOSAT estimates of surface geostrophic velocity both indicate that the phase of the seasonal changes in the latitude of the confluence coincide with opposing seasonal variations in the alongshore transports of the Brazil and Malvinas Currents.

From the applications summarized in this section, the potential for altimeter data to contribute information unobtainable by any other means about the temporal evolution of sea level fields has been clearly demonstrated. Despite problems with measurement errors (particularly orbit errors and the wet tropospheric range correction) and tidal aliasing, GEOSAT data have provided new insight about equatorial wave dynamics, eddy propagation and large-scale sea level variability. An unsettling question that has arisen from most direct comparisons with in situ measurements from tide gauges and hydrographic data is why the amplitudes of variability inferred from GEOSAT data are generally somewhat small (e.g., Ménard, 1988; Cheney et al., 1989; Tai et al., 1989; Chelton et al., 1990; Arnault et al., 1990; 1992). This has variously been attributed to signal attenuation by the orbit error corrections applied or to excessive smoothing of the data. In order to assess the impact of the latter, some guidance is needed to determine the space and time scales that can be resolved by altimeter data. The objective of this study is to determine the minimum smoothing necessary so that the highest possible resolution is retained in the sea level fields constructed from altimeter data.

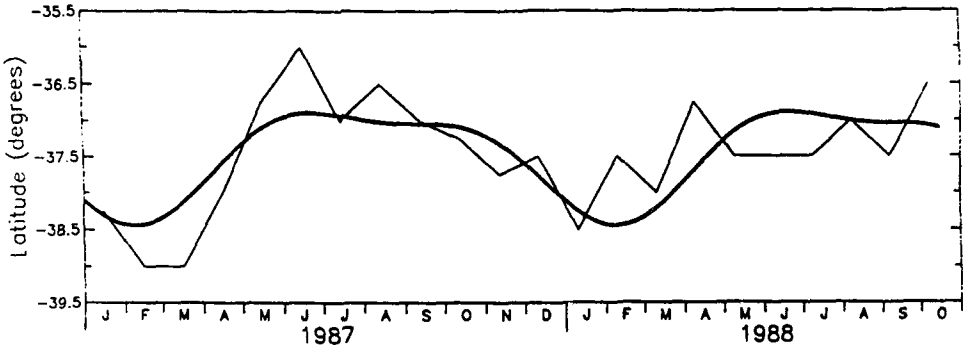


Figure 9. A time history of the latitude of the confluence of the Brazil and Malvinas Currents near the continental slope of South America as determined from two years of GEOSAT data (thin line). The smooth, heavy line represents a least-squares fit of annual and semiannual harmonics to the raw data. (From Matano et al., 1993.)

3. EQUIVALENT TRANSFER FUNCTION

3.1. Formalism

The question of the resolution capability of an irregularly sampled data set is investigated here by considering a simple approach to smoothing the data based on a linear estimate constructed from the N "nearest" (in space or time) observations. To simplify

the notation, the formalism is developed for a 1-dimensional case; extension to higher dimensions is straightforward. The jth observation of a stationary stochastic process $h(t)$ will be written as

$$g_j = h(t_j) + \epsilon_j, \qquad j = 1, \ldots, N, \tag{1}$$

where t is time and ϵ_j is the measurement error or unresolved geophysical "noise" in the jth observation. The general form for a linear estimate of h at an arbitrary time t_0 constructed from these N observations is

$$\hat{h}(t_0) = \sum_{j=1}^{N} \alpha_j(t_0) g_j . \tag{2}$$

Note that the α_j in general depend on the estimation time t_0. In the statistical literature, Eq. (2) is referred to as a smoother and the smoother weights α_j are referred to as the equivalent kernel. These weights can be specified by many methods (Buja et al., 1989). Examples include moving averages, Gaussian weighted averages, local least squares fits to a polynomial, local weighted least squares fits to a polynomial ("loess smoothers"), natural or smoothing spline estimates and Gauss-Markov estimates.

The form of the linear estimate that is often preferred is the Gauss-Markov estimate in which the equivalent kernel is computed from a priori specified signal and noise covariance functions (see Appendix B). Gauss-Markov estimation is generally referred to as "objective analysis" in the oceanographic and meteorological literature (e.g., Gandin, 1965; Bretherton et al., 1976). Examples of objective analysis applied to altimeter data include De Mey and Robinson (1987) and Fu and Zlotnicki (1989). If the covariance function is the true covariance function for the process $h(t)$, the Gauss-Markov estimate is optimal in the sense that it has the lowest mean squared error of all linear estimates of the form Eq. (2). In practice, the optimal estimate generally differs little from other linear estimates. The primary advantages of the optimal estimate are that the formalism easily allows an explicit treatment of measurement errors and provides an expression for the expected error of the estimate.

The disadvantage of Gauss-Markov estimates is that they are computationally intensive when N is large. For this reason, we have found it more useful for applications to large altimeter data sets (see, for example, Chelton et al., 1990; Matano et al., 1993) to apply the quadratic loess smoother described in Appendix A. The computational requirements of the loess smoother are much lower than those of Gauss-Markov estimates. It is shown below that the filtering characteristics of the quadratic loess smoother are nearly as good as those of Gauss-Markov estimates (see Figures 10, 19 and 20).

When the observations are evenly spaced and the estimation times t_0 coincide with the observation times, the filtering properties of the smoother are the same at each t_0, except near the ends of the sample record, where edge effects become important. These end regions are usually discarded so that the frequency content is uniform throughout the smoothed time series. In this case, the filtering properties of the smoother are determined by expressing the linear estimate in the form of a convolution of the observations and then applying the convolution theorem to obtain the frequency transfer function of the smoother.

When the observations are irregularly spaced, the filtering properties can vary considerably with t_0 (see Figure 7 of Schlax and Chelton, 1992), and the convolution theorem is not easily applied. As shown in section 5.1, it is desirable to choose the smoothing parameter of the linear estimate so that the filtering characteristics are nearly the same at all t_0. Otherwise, the frequency content of the smoothed time series can be highly nonstationary (see Figure 14 below).

The filtering properties of the linear estimate are easy to quantify if the linear estimator Eq. (2) is expressed as an integral over t,

$$\hat{h}(t_0) = \int_{-\infty}^{\infty} \hat{p}(t;t_0) g(t)\, dt\,, \tag{3}$$

where

$$\hat{p}(t;t_0) = \sum_{j=1}^{N} \alpha_j(t_0)\delta(t - t_j) \tag{4}$$

is another way of expressing the equivalent kernel in terms of the Dirac delta function. The integral expression Eq. (3) can be expressed in the frequency domain using the Power Theorem (Bracewell, 1978) as

$$\begin{aligned} \hat{h}(t_0) &= \int_{-\infty}^{\infty} \hat{P}^*(f;t_0) G(f)\, df \\ &= \int_{-\infty}^{\infty} \hat{P}^*(f;t_0) H(f)\, df + \int_{-\infty}^{\infty} \hat{P}^*(f;t_0) N(f)\, df \end{aligned} \tag{5}$$

where f is frequency, $G(f)$ is the Fourier transform of the measurements $g(t)$, $N(f)$ is the Fourier transform of the measurement errors ϵ and $\hat{P}(f;t_0)$ is the Fourier transform of $\hat{p}(t;t_0)$, which reduces to

$$\hat{P}(f;t_0) = \sum_{j=1}^{N} \alpha_j(t_0) e^{-i2\pi f t_j}\,. \tag{6}$$

$\hat{P}$ is referred to as the equivalent transfer function (Schlax and Chelton, 1992), since it is closely related to the equivalent kernel weights α_j.

In three dimensions, the equivalent transfer function for an estimate of the field at location (x_0, y_0, t_0) is

$$\hat{P}(k,l,f;x_0,y_0,t_0) = \sum_{j=1}^{N} \alpha_j(x_0,y_0,t_0) e^{-i2\pi(kx_j + ly_j - ft_j)}\,, \tag{7}$$

where k and l are the zonal and meridional wavenumbers. The sign convention adopted in Eq. (7) defines the direction of propagating features that are aliased into the smoothed estimate. For example, positive k and f correspond to eastward propagation (see section 3.4).

The filtering characteristics of the smoother are clear from Eq. (5); the equivalent transfer function specifies how the frequency content of the measurements g_j (both the signal and noise components) are filtered by the linear estimate.

Determination of the equivalent transfer function $\hat{P}$ from the smoother weights α_j can be computationally intensive. This is especially true for large, multi-dimensional data sets. A method for computing $\hat{P}$ efficiently and with high wavenumber-frequency resolution by a fast Fourier transform technique is presented in the appendix of Schlax and Chelton (1992).

3.2. A 1-Dimensional Example

In one dimension, the quadratic loess smoother used here to investigate the resolution capability of irregularly sampled data sets is obtained by a weighted least squares fit of a quadratic function of t to observations within a distance d_t (referred to as the half-span of the smoother) of the estimation time t_0. A detailed description is given in Appendix A.

The equivalent transfer functions of the quadratic loess smoother for two different half spans are shown in Figure 10a for evenly spaced observations. The main feature of each transfer function is a low-pass band with near unit amplitude and a sharp cutoff at a frequency of $f_c \approx d_t^{-1}$ to near zero values at higher frequencies. This pass band defines the smoothing characteristics of the linear estimate; the frequency content of the observations g_j is rejected at frequencies where the transfer function has a magnitude of zero and is fully included where the transfer function has a magnitude of one. The cutoff frequency f_c can be decreased by increasing the span of the quadratic loess smoother, resulting in a smoother time series of estimates (see Figure 10a).

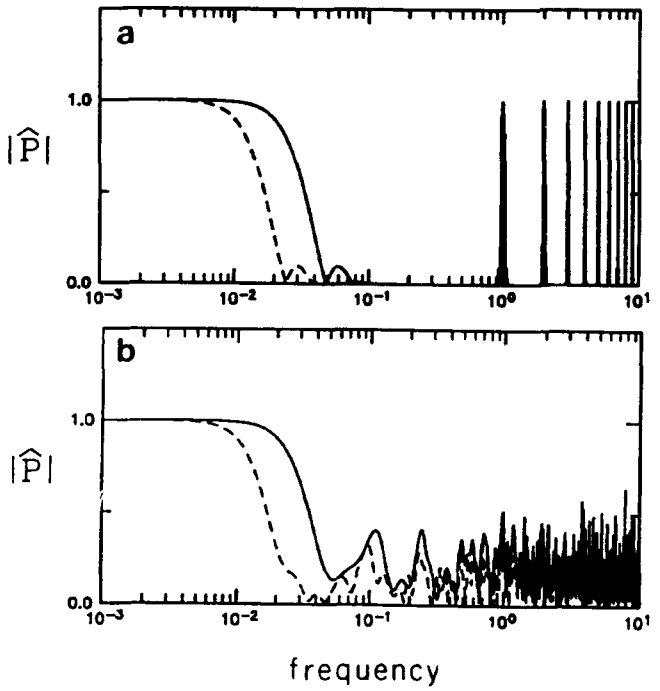


Figure 10. The 1-dimensional equivalent transfer function modulii of the quadratic loess smoother for a) an evenly spaced sample design with sample interval $\Delta = 1$ and smoothing parameters d_t=30 (solid line) and 60 (dashed line); and b) an irregularly spaced sample design with d_t =30. In both panels, the estimation point is at the midpoint of the data record. Note that the frequency axis is logarithmic.

The series of peaks with successively narrower width in the logarithmic plot of the equivalent transfer function are aliases of the low-pass band, folded about the Nyquist frequency $f_N = (2\Delta)^{-1}$, where $\Delta = 1$ is the sample interval. The aliasing peaks are cen-

tered at even multiples of f_N. In a linear plot, the widths of each of these alias peaks are the same as the $2f_c$ width of the central low-pass band that is symmetric about zero frequency. If there is any energy in the signal or noise at these higher frequencies, it will be aliased into the low-pass band and indistinguishable from actual low frequency variability.

For an ideal filter, the equivalent transfer function would drop abruptly from a magnitude of one to a magnitude of zero at the cutoff frequency f_c and would remain zero at all higher frequencies. The more gradual low-pass band edge rolloff and the alias peaks of real smoothers represent imperfections of the real filtering operation.

The equivalent transfer function for an example of irregularly spaced observations is shown in Figure 10b. The low-pass band of interest is very similar to that for the evenly spaced sample design with the same d_t shown in Figure 10a. The primary difference is the noisy continuum of energy in the transfer function for the uneven design at frequencies higher than f_c. The details of these high-frequency characteristics of the equivalent transfer function depend on the particular sample design and on the estimation time t_0. Just like aliasing for the case of evenly spaced observations, any energy in the signal or noise at these frequencies higher than f_c will contaminate the lower frequencies that are of interest in the smoothed estimates. The greater the amplitude of the equivalent transfer function at the higher frequencies, the less efficiently the smoothed estimates will isolate the low-frequency content of the signal of interest. Although aliasing loses its classical meaning when the observations are irregularly spaced, this high-frequency contamination in the equivalent transfer function will be referred to here as aliasing, for lack of a better term.

While the band-edge rolloff of the quadratic loess smoother is not quite as sharp as for Gauss-Markov estimates when the signal-to-noise ratio is high (compare Figure 10 with Figures 19a and 20a in Appendix B), it is sharper than those of other commonly used smoothers (see Schlax and Chelton, 1992), as well as Gauss-Markov estimates when the signal-to-noise ratio is small (Figures 19c and 20c). For most purposes, the slightly less efficient filtering characteristics are compensated for by the much greater computational efficiency of the quadratic loess smoother; in application to large 3-dimensional data sets such as altimeter data, Gauss-Markov estimates require about two orders of magnitude more computing effort and are therefore not practical for studies on basin scales.

3.3. The GEOSAT Ground Track Pattern Sampled Synoptically

The combined space and time characteristics of the satellite sampling pattern complicate interpretation of the equivalent transfer function. The separate effects of spatial and temporal sampling become clearer if time dependence is first neglected and synoptic sampling of the ground track pattern in Figure 6a is considered; the effects of asynoptic sampling of this grid are examined in section 3.4.

The 2-dimensional wavenumber equivalent transfer function for a quadratic loess estimate constructed from the GEOSAT 17-day sample grid is shown in Figure 11 for an estimation location at a point where ascending and descending ground tracks cross. For the purposes of this discussion, the GEOSAT data were subsampled at intervals of 50 km along the ground tracks. All of the information about the spatial regularity of the sample grid is contained in this figure. The transfer function is symmetric about both wavenum-

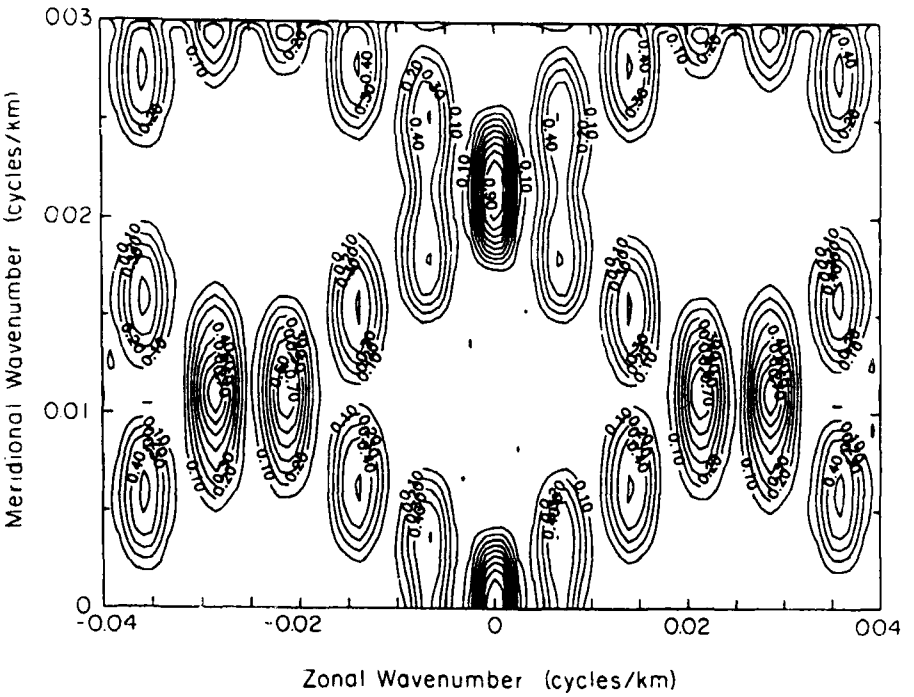


Figure 11. The 2-dimensional wavenumber equivalent transfer function modulus for the GEOSAT ground track pattern (see Figure 6a) sampled synoptically for estimation point (x_0,y_0)=(45°W°,30°N) and quadratic loess smoothing parameters (d_x,d_y)=(5°,3°).These smoothing parameters were chosen somewhat arbitrarily to illustrate the aliasing patterns inherent in the diamond shaped sample grid. Note that the wavenumbers axes are linear in this figure.

ber axes. The elliptical plateau centered at zero that drops off steeply to generally small values at higher wavenumbers is the low-wavenumber pass band of the smoother. The aspect ratio of this pass band (longer in the meridional wavenumber direction than in the zonal wavenumber direction) is the inverse of the ratio of smoothing spans $d_y/d_x = 3/5$. The other subsidiary peaks (with the same aspect ratio as the low-frequency pass band) are aliasing peaks that arise because of the very regular diamond-shaped grid of crossover points. At the 30° latitude of the estimation location, the dimensions of the diamond patterns mapped out by the ground tracks are approximately 1.5° of longitude by 3° of latitude (see Figure 6a). The corresponding Nyquist wavenumbers are about $k_N = 0.0036$ cycle/km (cycles per km) and $l_N = 0.0015$ cycle/km. The minima between the aliasing peaks and the maxima of the peaks are centered at odd and even multiples, respectively, of these Nyquist wavenumbers (compare with the 1-dimensional example in Figure 10a). The coarser ground track pattern of a shorter orbit repeat period would result in larger diamond patterns and, hence, lower Nyquist wavenumbers and more closely spaced aliasing peaks. For example, for the approximate 2.7° of longitude by 5.5° of latitude diamonds of the TOPEX 10-day repeat orbit, the series of aliasing peaks overlap because the smoothing parameter $d_y = 5°$ is too short for the TOPEX sample grid.

The diagonal patterns of regularly spaced aliasing peaks are thus an indication of the non-rectangular grid pattern of the crossover points. The tilting of the lines through the centers of these aliasing peaks are an indication that aliased features in the sea level

field are tilted parallel to the satellite ground tracks. The slopes of the lines through the aliasing peaks define the angles of the ground tracks in the spatial domain. In the extreme case of an orthogonal grid aligned east-west and north-south (which, of course, is not possible for a satellite orbit but is typical of sampling grids for other types of data), the aliasing peaks would lie along lines parallel to the wavenumber axes.

A second spatial scale is embedded in the regular pattern of the transfer function in Figure 11. At the 30° latitude of the estimate, the 50 km sample interval along the ground track represents zonal and meridional sample intervals of about 20 km and 46 km, respectively. The corresponding Nyquist wavenumbers are $k_N = 0.025$ cycle/km and $l_N = 0.011$ cycle/km. These Nyquist wavenumbers define the intersection points of the diagonal patterns of aliasing peaks; the intersections occur at odd multiples of the zonal and meridional Nyquist wavenumbers of the along-track sample interval. Sampling at closer intervals along the ground track would result in higher Nyquist wavenumbers and, hence, larger diamond patterns of the equivalent transfer function in wavenumber space.

3.4. The 3-Dimensional GEOSAT Data Set

When the asynoptic sampling of the satellite ground track is taken into consideration, visualization of the 3-dimensional equivalent transfer function is much more difficult than for the 2-dimensional sample grid considered in section 3.3. As an example of the ability of the equivalent transfer function to identify space-time structure in the satellite sampling pattern, a 2-dimensional slice through the transfer function along 90° azimuth (i.e., along the east axis with zero meridional wavenumber) is shown in Figure 12 for the GEOSAT data as actually sampled by the satellite. The location of the smoothed estimate for this example is a crossover point.

The low-frequency pass band of the smoother is evident as the plateau region centered at zero wavenumber and frequency. The interesting characteristic of the equivalent transfer function is the distortion of the usual elliptical pass band in the upper right quadrant. There is a series of aliasing peaks along a line of slope 1 in this log-log plot. It is easy to show that constant phase propagation at phase speed c_p is manifested in a log-log plot of the equivalent transfer function as a line with slope 1 that intercepts the $\log f = 0$ axis at $\log k = -\log c_p$. The -1.7 cycle/km intercept of the ridge of aliasing peaks in Figure 12 thus corresponds to a phase speed of about 48 km/day.

For the convention used here (see Eq. (7)), the positive k and f in the right half of Figure 12 represent eastward propagation. The propagation indicated by the ridge of aliasing peaks in Figure 12 is therefore eastward. An eastward propagation of 48 km/day corresponds to 144 km eastward propagation in three days. At this latitude of 30°N, this corresponds to the shift in the 3-day subcycle in the GEOSAT sampling pattern discussed in section 2.4 (see Figure 6b). The wavenumber-frequency transfer functions for the TOPEX and ERS-1 sampling patterns similarly show propagations of about 85 km/day eastward and 48 km/day westward, respectively. These are the zonal shifts of the 3-day subcycles for these other altimeter satellites.

The physical interpretation of the propagating aliasing pattern in the equivalent transfer function for the GEOSAT sampling pattern is that, if there is any eastward propagating sea level signal with spectral energy at any of the high-wavenumber, high-frequency peaks along the aliasing ridge, it will alias into the low-pass band of the loess

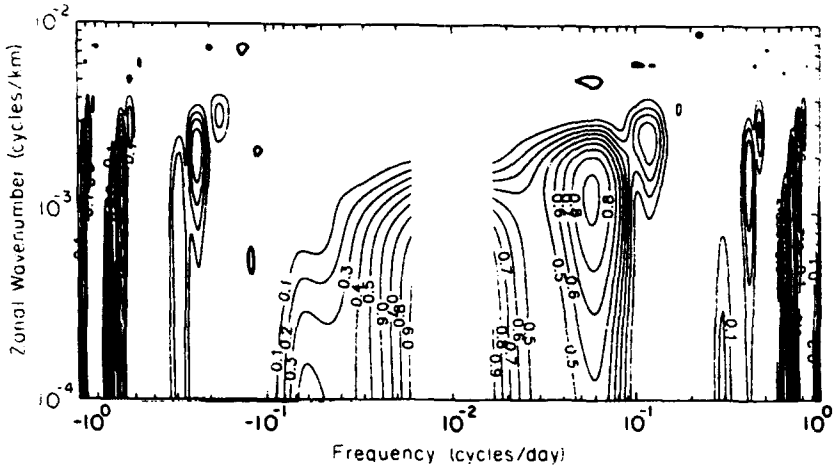


Figure 12. A slice through the 3-dimensional frequency-wavenumber equivalent transfer function modulus along the 90° azimuth (eastward) for the GEOSAT ground track pattern as actually sampled during each 17-day exact repeat period for estimation point (x_0, y_0, t_0) = (45°W,30°N, day 100) and quadratic loess smoothing parameters (d_x, d_y, d_t) = (8°,8°,35 days). These smoothing parameters were chosen somewhat arbitrarily to illustrate the eastward propagating aliasing pattern associated with the 3-day subcycle in the satellite orbit (see Figure 6b). Note that both axes are logarithmic.

smoothed estimate of the sea level field. That is, the aliased signal will be indistinguishable from the low-frequency, low-wavenumber variability of interest and is therefore undetectable in the smoothed sea level fields. On the other hand, if there is no sea level propagation at this phase speed, then this aliasing ridge is of little concern.

4. ERRORS OF THE SMOOTHED ESTIMATES

The equivalent transfer function only defines the filtering properties of the smoother for the specific smoothing parameters selected. Additional information about the signal characteristics is required to assess the quality of smoothed fields constructed from the irregularly sampled data. The degree to which imperfections in the filtering operation contaminate estimates of the large-scale, low-frequency signals of interest in the smoothed fields depends not only on the aliasing patterns in the equivalent transfer function, but also on the spectral energies of the signal and noise at the wavenumbers and frequencies of aliasing peaks in the transfer function. The combined effects of filtering properties and signal and noise characteristics on the accuracy of the smoothed estimates are quantified in this section.

The smoothed estimate $\hat{h}$ can be compared with an ideal low-pass filtered value, written as

$$\overline{h}(t_0) = \int_{-\infty}^{\infty} P^*(f; t_0, f_c) H(f)\, df\,, \tag{8}$$

where $H(f)$ is the Fourier transform of the unsmoothed signal $h(t)$ and $P^*(f; t_0, f_c)$ is the complex conjugate of the transfer function for the ideal smoothed estimate at time t_0. This ideal transfer function passes all of the signal at frequencies lower than the cutoff frequency f_c, and none of the signal at higher frequencies, i.e.,

$$P(f;t_0,f_c) = \begin{cases} e^{-i2\pi f t_0} & |f| < f_c \\ 0 & \text{otherwise.} \end{cases} \tag{9}$$

The complex transfer function P thus has unit modulus for frequencies less than f_c. We have found empirically that the cutoff frequency for the quadratic loess smoother used in this study is related to the half-span of the smoother by $f_c \approx d_t^{-1}$. This value of f_c is therefore used to define the ideal transfer function in Eq. (9).

Because the measurement errors have zero mean value, it can be seen from Eqs. (5) and (8) that the bias of the estimate $\hat{h}(t_0)$ is

$$\langle \hat{h}(t_0) \rangle - \bar{h}(t_0) = \int_{-\infty}^{\infty} \Delta\hat{P}^*(f : t_0, f_c) H(f)\, df\,, \tag{10}$$

where the angle brackets denote the mean value and

$$\Delta\hat{P}(f;t_0,f_c) = \hat{P}(f;t_0) - P(f;t_0,f_c) \tag{11}$$

represents the imperfection of the 1-dimensional equivalent transfer function at frequency f for an estimate at time t_0 with low-frequency cutoff f_c. The modulus of $\Delta\hat{P}$ is shown schematically by the hatched region in Figure 13. The bias given by Eq. (10) can be interpreted as the error of the estimate $\hat{h}(t_0)$ in the absence of any measurement errors. The bias thus focuses attention on errors introduced solely by the irregular sampling distribution.

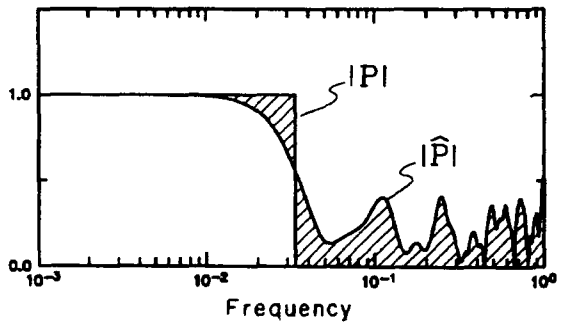


Figure 13. A schematic representation of the imperfections of the linear smoother given by Eq. (11) (hatched region).

In order to express the imperfections of the filtering operation in terms of the spectral characteristics of the signal $h(t)$, we write the integral in Eq. (10) in the limiting form

$$\langle \hat{h} \rangle - \bar{h} = \lim_{\delta f \to 0} \sum_{n=-\infty}^{\infty} \Delta\hat{P}^*(f_n) H(f_n)\, \delta f\,. \tag{12}$$

For convenience, the explicit dependencies on t_0 and f_c have been dropped in Eq. (12). The expected squared bias is

$$\left\langle \left[\langle \hat{h} \rangle - \bar{h} \right]^2 \right\rangle = \lim_{\substack{\delta f \to 0 \\ \delta s \to 0}} \sum_{n=-\infty}^{\infty} \sum_{m=-\infty}^{\infty} \Delta\hat{P}^*(f_n) \Delta\hat{P}(s_m) \left\langle H(f_n) H^*(s_m) \right\rangle \delta f \delta s\,. \tag{13}$$

Because $h(t)$ is assumed to be a stationary stochastic process, it is easy to show that

$$\langle H(f_n)H^*(s_m)\rangle = 0 \qquad \text{if } s_m \neq f_n \tag{14}$$

(see, for example, Priestley, 1992, p. 249). The expected squared bias then reduces to

$$\langle [\langle \hat{h} \rangle - \bar{h}\,]^2 \rangle = \lim_{\delta f \to 0} \sum_{n=-\infty}^{\infty} |\Delta \hat{P}(f_n)|^2 S_h(f_n)\,\delta f\,, \tag{15}$$

where

$$S_h(f_n) = \lim_{\delta s \to 0} \langle |H(f_n)|^2 \rangle\, \delta s \tag{16}$$

is the power spectral density of the random process $h(t)$ (Priestley, 1992, p. 208). In the limit, Eq. (15) becomes the integral

$$\langle [\langle \hat{h} \rangle - \bar{h}\,]^2 \rangle = \int_{-\infty}^{\infty} |\Delta \hat{P}(f)|^2 S_h(f)\,df\,. \tag{17}$$

The expected squared bias (ESB) given by Eq. (17) describes the combined effects of the signal spectral energy and the equivalent transfer function on the accuracy of the smoothed estimate $\hat{h}(t_0)$. At frequencies f where either the aliasing $|\Delta \hat{P}(f)|^2$ or the signal energy $S_h(f)$ are small, the integrand in Eq. (17) is small and consequently contributes little to the ESB. Aliasing at frequencies where $|\Delta \hat{P}(f)|$ is large is therefore of little concern if the corresponding signal spectral energy $S_h(f)$ is weak.

The ESB as a measure of the accuracy of the estimate $\hat{h}(t_0)$ can be compared with the mean squared error that is more traditionally used to assess the quality of an estimate. For a given realization of the stochastic process $h(t)$, the mean squared error can be decomposed into the sum of the squared bias and the variance. The expected value of the mean squared error over the ensemble of realizations of the process (the EMSE) is therefore the sum of the ESB given by Eq. (17) and the variance of the estimate. By the same method used to derive Eq. (17), it is easy to show from Eq. (5) that the variance of the estimate is

$$\langle [\langle \hat{h} \rangle - \hat{h}\,]^2 \rangle = \int_{-\infty}^{\infty} |\Delta \hat{P}(f)|^2 S_\epsilon(f)\,df\,, \tag{18}$$

where $S_\epsilon(f)$ is the power spectral density of the measurement errors. The variance of the smoothed estimate thus describes the combined effects of the spectral characteristics of the measurement errors and the equivalent transfer function on the accuracy of the smoothed estimate $\hat{h}(t_0)$.

The present study is primarily concerned with the limitations imposed by the sampling design, regardless of the measurement errors. In the extreme case of no measurement errors, the variance of the smoothed estimate is zero and the EMSE is just the ESB. Then all of the errors in the smoothed estimate arise from the sampling design. For a reasonably large signal-to-noise variance ratio (greater than 1) and a sufficiently dense sample design (i.e., a well behaved equivalent transfer function $\hat{P}$), the ESB is generally much larger than the variance. Then the EMSE can be approximated as just the ESB. For the

altimeter applications of interest in this study, the signal-to-noise variance ratio is large enough that the variance contribution to the EMSE can be neglected. We therefore restrict attention to the ESB as a measure of the accuracy of the smoothed estimates.

The mean squared error formalism is easily extended to three dimensions. The ESB for one dimension (Eq. (17)) then becomes

$$\left\langle \left[\left\langle \hat{h}(x_0, y_0, t_0) \right\rangle - \overline{h}(x_0, y_0, t_0) \right]^2 \right\rangle = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \left| \Delta \hat{P} \right|^2 S_h(k, l, f)\, dk\, dl\, df\,, \tag{19}$$

where

$$\Delta \hat{P} = \hat{P}(k, l, f; x_0, y_0, t_0) - P(k, l, f; x_0, y_0, t_0, k_c, l_c, f_c) \tag{20}$$

represents the imperfections of the 3-dimensional equivalent transfer function $\hat{P}$ compared with the 3-dimensional transfer function P of the ideal smoother. Determination of the ESB of a 3-dimensional smoothed estimate thus requires knowledge of the 3-dimensional wavenumber-frequency spectrum $S_h(k, l, f)$ of the signal.

In multiple dimensions, the smoother weights α_j for observations $g(x_j, y_j, t_j)$ on a sufficiently dense and regularly spaced sample grid depend only on the distance r from the estimation location (x_0, y_0, t_0). For the quadratic loess smoother with half spans d_x, d_y and d_t, this distance is defined by

$$r^2 = \left(\frac{x_j - x_0}{d_x} \right)^2 + \left(\frac{y_j - y_0}{d_y} \right)^2 + \left(\frac{t_j - t_0}{d_t} \right)^2 . \tag{21}$$

The Fourier transform of an elliptically symmetric function is also elliptically symmetric (Bracewell, 1978, p. 244). It is therefore appropriate to use an ellipitically symmetric ideal transfer function for the multidimensional bias calculation, i.e.,

$$P(k, l, f; x_0, y_0, t_0, k_c, l_c, f_c) = \begin{cases} e^{-i2\pi(kx_0 + ly_0 - ft_0)} & (k/k_c)^2 + (l/l_c)^2 + (f/f_c)^2 < 1 \\ 0 & \text{otherwise.} \end{cases} \tag{22}$$

As before, the low-pass wavenumber and frequency cutoffs k_c, l_c and f_c for the quadratic loess smoother are approximately the reciprocal of the half spans in each dimension.

In three dimensions, evaluation of the triple integral in Eq. (19) by the usual quadrature methods is computationally intensive. For this study, these integrals were estimated using a weighted Monte Carlo method that is based on sampling the region of integration at discrete sample points distributed with a probability density proportional to the signal spectral energy (Press et al., 1992, p. 306).

The power spectral density and the autocovariance function of the signal are Fourier transform pairs (Priestley, 1992, p. 211). The spectral properties of the signal can therefore be specified directly or can be computed from a specified autocovariance function (equivalent to specifying the signal variance and autocorrelation function). The need to specify the signal variance (which varies geographically for the sea level fields of interest in this study) can be sidestepped by considering the relative expected squared bias (RESB), defined to be the ESB given by Eq. (19) normalized by the signal variance σ_h^2. For the applications considered in section 5, the signal spectral shape $S_h(k, l, f)/\sigma_h^2$ needed to evaluate the relative accuracy of the smoothed estimate $\hat{h}(t_0)$ by the RESB was obtained from the Fourier transform of the specified signal autocorrelation function.

5. RESOLUTION CAPABILITY

5.1. A 1-Dimensional Example

The philosophy adopted here to define the resolution capability of an irregularly spaced data set is easily demonstrated by a simple 1-dimensional example. A densely sampled synthetic high-frequency time series with unit variance is shown in Figure 14a. The details of how this time series was generated are not important to this discussion. The effects of nonuniform sampling of this time series are illustrated by sampling the time series in Figure 14a with periodic bursts of closely spaced observations, separated by intervals of coarsely spaced observations. This sampling strategy is intended to be a 1-dimensional analog of the sampling characteristics of altimeter data, which are characterized by dense 2-dimensional sampling at crossover points and sparse coverage elsewhere. Two different loess smoothed time series were constructed from the unequally spaced observations to show how the ESB Eq. (17) can be used to select good smoothing parameters for the linear estimates.

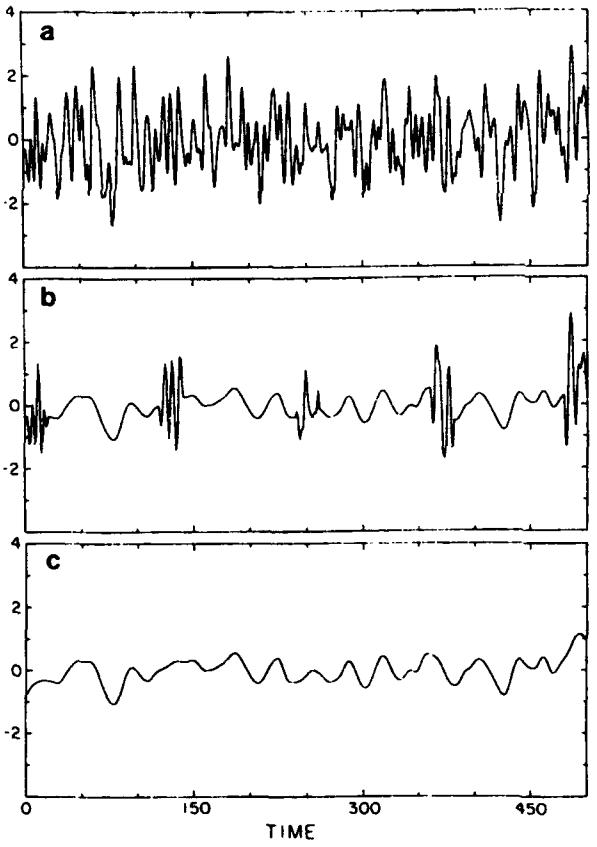


Figure 14. a) A high-frequency synthetic time series.This time series was observed in bursts of sample interval 0.2 separated by sparse observations at sample interval 2.0; b) a quadratic loess smoothed time series constructed at intervals of 0.2 using half spans of $d_t = 0.6$ during the bursts of closely spaced observations and $d_t = 30$ during the periods of coarsely spaced observations; c) a quadratic loess smoothed time series constructed at intervals of 0.2 using a fixed half span of $d_t = 30$ everywhere.

In the first example (Figure 14b), the smoothing parameters d_t of the loess estimates were chosen to maximize the information content of the observations. A small value of d_t was used during the bursts of closely spaced observations and a larger value of d_t was used during the periods of coarsely spaced observations. As noted in section 3.2, the low-pass frequency cutoff of the loess smoother is $f_c \approx d_t^{-1}$. Consequently, the spectral content of the loess estimates in the coarsely sampled periods is restricted to lower frequen-

cies than in the burst periods. The resulting nonstationarity of the smoothed time series is readily apparent from Figure 14b. Another undesirable characteristic of the smoothed time series is the nonstationary ESBs of the loess smoothed estimates, which vary from negligibly small in the burst periods to 0.02 in the coarsely sampled periods.

In the second example (Figure 14c), the smoothed time series was constructed by fixing the loess smoothing parameter throughout the record to the large value used in Figure 14b in the coarsely sampled periods. This is equivalent to sacrificing the higher resolution capability in the burst periods (i.e., "oversmoothing" the data). However, the benefits of this procedure are apparent from Figure 14c; the spectral content of the resulting smoothed time series is stationary. In addition, the ESBs of the loess estimates are uniform (0.02) throughout the record.

The need to degrade the higher resolution possible in the burst regions is disappointing. However, for analysis of the full record of unequally spaced observations, the homogeneously smooth time series in Figure 14c is much more desirable than the nonstationary time series in Figure 14b. If the interest is in the higher frequency variability that can be resolved in the burst periods, then the analysis must be restricted to just the burst periods. Then the longer-period information content of the full data set is lost by sacrificing the coarsely sampled periods of the data record.

The philosophy for choosing the appropriate smoothing parameter is therefore to smooth the data to the resolution that is possible in the sparsely sampled regions. This can be achieved by selecting a single smoothing parameter for the entire data set that yields a uniform ESB at every location at which a smoothed estimate is to be constructed. The spectral content of the resulting smoothed time series will be stationary.

5.2. The GEOSAT Ground Track Pattern Sampled Synoptically

Extension of the results of section 5.1 to two spatial dimensions further emphasizes the importance of degrading the resolution capability in densely sampled regions. As in section 3, the full 3-dimensional characteristics of altimeter sampling are more easily understood if time dependence is first neglected and synoptic sampling of the ground track pattern in Figure 6a is considered. Near the crossover points, this sample grid is capable of providing detailed maps of mesoscale variability. However, along the ground tracks connecting crossover points and in the unsampled diamond regions in between, only the larger scale variability can be resolved. A map constructed with the highest resolution possible at each location (analogous to the 1-dimensional case in Figure 14b) would consist of a patchwork quilt of eddies and meanders near the crossover points and smooth, large-scale variability elsewhere.

These effects can be quantified in terms of the RESB. The 2-dimensional wavenumber spectrum of sea level must be specified to obtain the RESB. Analyses of dynamic height fields from hydrographic data provide useful guidance. Shen et al. (1986), Carter and Robinson (1987) and other studies have found that the spatial structure of the sea level field can be approximated by an isotropic Gaussian autocorrelation function of the form

$$\rho(r) = e^{-(r/r_0)^2}, \tag{23}$$

where r is distance and the spatial scale r_0 is approximately 50 km. This spatial scale is

consistent with independent estimates of $\rho(r)$ computed directly from altimeter data (Le Traon et al., 1990). The normalized 1-dimensional wavenumber spectrum of sea level for computing the RESB from Eq. (17) was obtained analytically from the Fourier transform of this Gaussian autocorrelation function,

$$\frac{S_h(k)}{\sigma_h^2} = \sqrt{\pi} r_0 e^{-(\pi r_0 k)^2} \tag{24}$$

(Bracewell, 1978, p. 130), where σ_h^2 is the (unspecified) sea level variance.

The RESB was computed from the GEOSAT sample grid for a range of loess smoothing parameters d_x and d_y at three estimation locations: a crossover point, a diamond center, and a point along a ground track midway between two crossover points (referred to here as a midpoint). A contour plot of the RESBs for the midpoint is shown in Figure 15. It is evident from this figure that there is no unique choice of smoothing parameters for a particular RESB; a given RESB can be obtained with high meridional resolution and low zonal resolution, with low meridional resolution and high zonal resolution, or by compromising to obtain moderate resolution in both dimensions. The approximate 2-to-1 aspect ratio of the contours indicates that a greater degree of smoothing is required in the meridional direction than in the zonal direction to obtain a given RESB. This is because of the longer meridional dimensions of the diamonds formed by the ground track patterns.

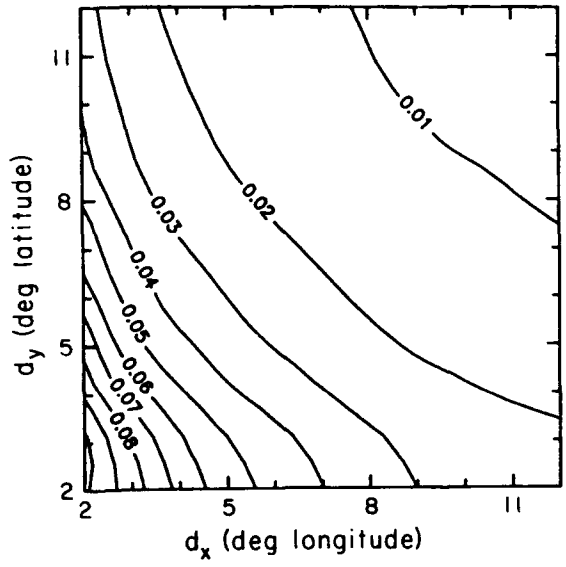


Figure 15. Contour plot of the relative expected squared bias as a function of the longitudinal and latitudinal quadratic loess smoothing parameters d_x and d_y for the GEOSAT ground track pattern (see Figure 6a) sampled synoptically for an estimate at a crossover point.

The simplest form of spatial smoothing is the isotropic smoother for which $d_x = d_y \equiv d_s$. Isotropic smoothing is used in Figure 16 to illustrate the geographical variability of the RESB. The three curves represent the RESB as a function of d_s for the three estimation points considered. At the shortest smoothing scale of $d_s = 2°$, the RESB is highest at the diamond center and lowest at the midpoint. At both of these locations, the RESB decreases monotonically as the smoothing parameter d_s increases, converging at about $d_s = 4°$. Curiously, the RESB at the crossover actually increases as d_s is increased from 2° to 2.5°and then decreases monotonically for larger d_s.

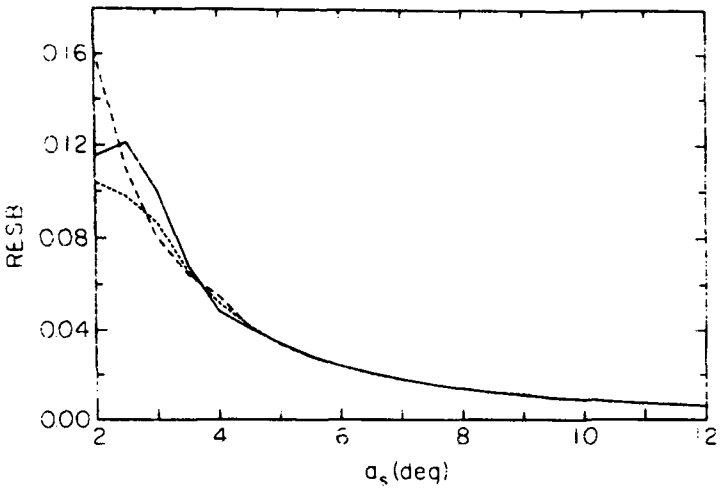


Figure 16. The relative expected squared bias as a function of isotropic quadratic loess smoothing parameter d_s for the GEOSAT ground track pattern sampled synoptically for estimates at a crossover location (solid line), a diamond center (dashed line) and along a ground track at the midpoint between two crossovers (dotted line).

The behavior of the RESB at the crossover is counter-intuitive. For very small d_s (not shown in Figure 16), the RESB is smaller at the crossover than at the midpoint. However, for $d_s = 2°$, the RESB is lower at the midpoint because the region of influence about the midpoint then includes observations on the neighboring ground tracks from a wide range of directions. In comparison, the region of influence about the crossover for $d_s = 2°$ includes data only from the two diagonal ground tracks passing through the crossover; observations are not available from the regions directly north, south, east and west of the crossover. As the span further increases to $d_s = 2.5°$, the lack of zonal and meridional constraints on the 2-dimensional smoothed estimate at the crossover becomes more significant, further increasing the RESB. When d_s exceeds 2.5°, the region of influence for the crossover estimate becomes large enough to include zonally adjacent crossovers and neighboring crossovers along the ground tracks that intersect at the estimation point. The 2-dimensional field is then well resolved in all directions and the RESB of the crossover estimate begins to decrease with increasing d_s.

The important point made by Figure 16 is that the RESB is not homogeneous over the map for small spatial smoothing parameter d_s. According to the criterion outlined in section 5.1, the best value of d_s for loess estimates constructed at an arbitrary location (x_0, y_0) is the smallest value that gives spatially homogeneous RESB. On the basis of Figure 16, this is about 5°, which is the value of d_s at which the RESB curves for the three estimation locations converge. Such a large degree of smoothing is somewhat overly pessimistic, however, since this is larger than the dimensions of the GEOSAT diamonds. The same RESB can be obtained at a somewhat higher resolution if estimates are constructed only at the crossover points. Moreover, when the asynopticity of the sampling of the GEOSAT grid is considered (see section 5.3), temporal smoothing can be used to further increase the spatial resolution capability.

5.3. The 3-Dimensional GEOSAT Data Set

The RESB is a complicated function of time and geographical location when the asynoptic sampling characteristics of the GEOSAT ground tracks are considered. The wavenumber-frequency spectral shape for determining the RESB from Eq. (19) was derived by assuming a Gaussian space-time autocorrelation function

$$\rho(r,\tau) = e^{-(r/r_0)^2} e^{-(\tau/\tau_0)^2}, \tag{25}$$

where r is the isotropic spatial lag as in section 5.2 and τ is the time lag. The spatial and temporal scales were chosen to be $r_0 = 50$ km and $\tau_0 = 30$ days. This form is consistent with the space-time autocorrelation function derived from dynamic height data (Shen et al., 1986; Carter and Robinson, 1987). The corresponding normalized power spectral density for computing the RESB is

$$\frac{S_h(k,l,f)}{\sigma_h^2} = \pi\, r_0 \tau_0\, e^{-(\pi r_0 k)^2} e^{-(\pi \tau_0 f)^2}. \tag{26}$$

The RESB at a crossover point is contoured in Figure 17 for a range of temporal and isotropic spatial smoothing parameters d_t and d_s at two different times during the

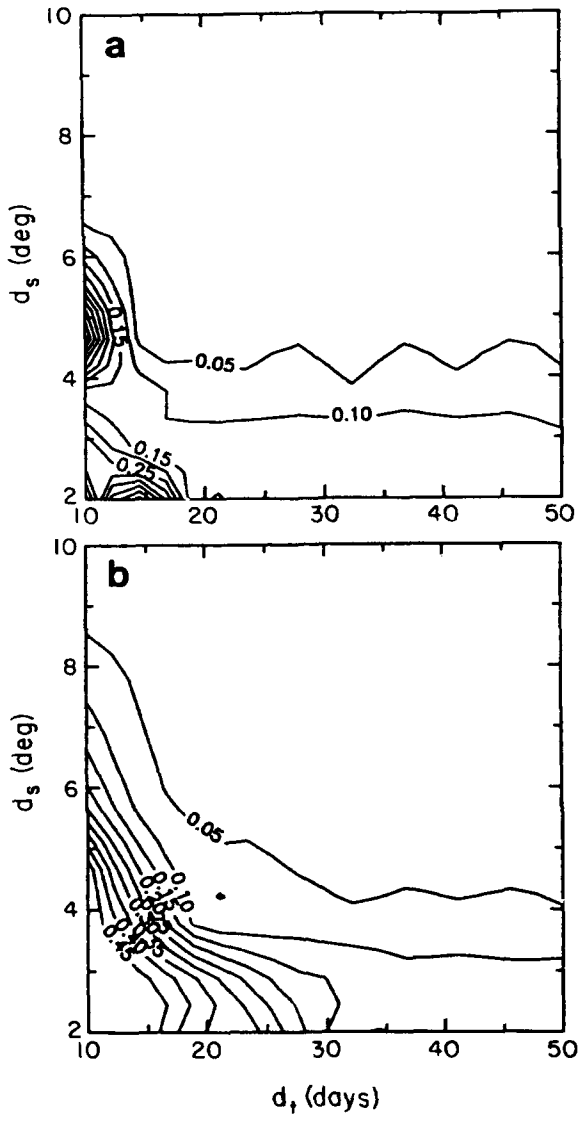


Figure 17. The relative expected squared bias as a function of the temporal and isotropic spatial loess smoothing parameters d_t and d_s for the GEOSAT ground track pattern as actually sampled during each 17-day exact repeat period. The two panels correspond to estimates at a crossover location on a) day 2; and b) day 11.

GEOSAT 17-day repeat. The plot for day 2 (Figure 17a) corresponds to a time when both ground tracks passed through this crossover within a 24-hour period. The RESB is therefore generally small. The peculiar behavior for small d_t and d_s is a more complex manifestation of the radius of influence problem discussed in section 5.2 (see Figure 16). When $d_s = 2°$, the RESB first increases with increasing d_t until $d_t \approx 15$ days and then decreases monotonically for larger d_t. This is because short temporal smoothing is well resolved near the time when both ground tracks sample the crossover point. The smoothed sea level field over longer 15-day periods is not as well resolved because of the long interval between GEOSAT sampling of neighboring ground tracks. The temporal half span must be increased to more than 15 days for the radius of influence of the 3-dimensional quadratic loess smoother to become large enough to include observations from neighboring ground tracks, thereby decreasing the RESB.

A similar effect occurs as a function of d_s when d_t is small. When $d_t = 10$ days, the RESB initially decreases with increasing d_s until $d_s \approx 3.8°$. For larger d_s, the RESB first increases until $d_s \approx 4.5°$ and then decreases monotonically for larger d_s. This effect is related to the complex temporal structure of the GEOSAT sampling of nearby ground tracks. For $d_t = 10$ days, the spatial structure of the low-pass filtered sea level field is not well resolved when $d_s = 4.5°$; the smoothed sea level field at this time and location is better resolved in the quadratic loess smoothed estimate by either decreasing or increasing the degree of spatial smoothing.

The contour plot of RESB for day 11 (Figure 17b) is much simpler than that for day 2. The RESB decreases monotonically with increasing d_t and d_s over the full ranges of these smoothing parameters. At day 11, this crossover point and the neighboring ground tracks are not sampled at nearby times. A much greater degree of smoothing (spatially or temporally) is therefore required than on day 2 to achieve a given value of RESB.

The spatial and temporal inhomogeneity of the RESB evident from Figure 17 complicates selection of a good combination of the smoothing parameters d_s and d_t. A given value of the RESB can be achieved at any particular estimation time t_0 by trading off d_s against d_t. However, the RESB for a specific choice of these smoothing parameters will, in general, differ for different estimation times t_0.

The temporal variability of the RESB at this crossover location during a 17-day GEOSAT repeat period is shown in Figure 18 for six different combinations of d_s and d_t. For small d_t, the RESB is rather erratic and varies by more than an order of magnitude over the 17-day repeat period unless d_s is very large (see the examples for $(d_s, d_t) = (4, 10)$ and $(8, 10)$); the RESB is generally large with localized decreases at times when there are GEOSAT ground tracks nearby. When the temporal span d_t is increased to values larger than the 17-day repeat period, the radius of influence of the quadratic loess smoother includes sufficient data to yield well-behaved time series of the RESB. For this particular crossover location, the RESB tends to have a minimum at day 2 with maxima at days 7 and 14 separated by a local minimum at day 11 (see the example for $(d_s, d_t) = (2, 20)$). These features of the RESB time series reflect the temporal distribution of ascending and descending ground tracks in the vicinity of the estimation location.

The time series of RESB at other crossover locations exhibit similar periodic variations over each 17-day repeat period. The timing of the minima and maxima vary, de-

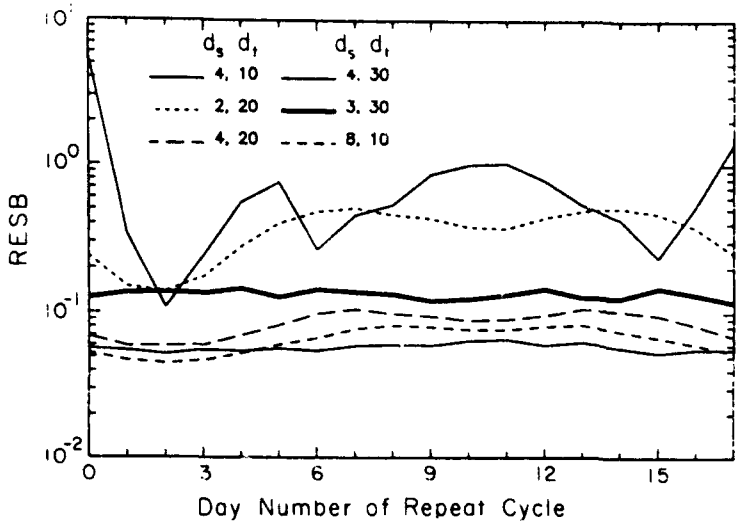


Figure 18. The relative expected squared bias as a function of time at a GEOSAT crossover location for six different choices of the temporal and isotropic spatial loess smoothing parameters d_t and d_s.

pending on the temporal distribution of GEOSAT data near the particular crossover location.

Two of the combinations of d_s and d_t in Figure 18 are of particular interest. When $(d_s, d_t) = (3,30)$ the RESB is approximately a constant value of 0.1 over the 17-day repeat period. When $(d_s, d_t) = (4,30)$, the RESB is about 0.05 and is also constant over the 17-day period. The time series of RESB at other crossover locations are similarly approximately constant over the 17-day repeat period. By the criterion outlined in section 5.1, either of these would therefore be good choices for d_s and d_t. The choice of $(d_s, d_t) = (4,30)$ is the more conservative of the two as it yields an RESB that is about half as large. Note that while the RESB for $(d_s, d_t) = (4,20)$ is everywhere smaller than that for (3,30), it varies by a factor of two over the 17-day period. As discussed in section 5.1, this temporally inhomogeneous RESB is less desirable than tolerating the somewhat higher RESB for $(d_s, d_t) = (3,30)$.

On the basis of Figure 18, we conclude that the GEOSAT sampling pattern is capable of resolving the spatial and temporal characteristics of sea level variability on monthly and longer time scales. The spatial resolution of monthly maps constructed from GEOSAT data is 3° or 4°, depending on how liberal one chooses to be about the degree of RESB that is tolerable. The effects of measurement errors and data dropouts have been neglected in this analysis.

It should be noted that the RESB for these choices of loess smoothing parameters are generally larger and not necessarily constant temporally over the 17-day repeat period at locations other than at crossover points. A greater degree of spatial or temporal smoothing would therefore be necessary for quadratic loess estimates at these other locations. However, the spatial smoothing parameters of $d_s = 3°$ or 4° are large enough that the spatial dimensions of the smoothed estimates at neighboring crossover locations already overlap. Consequently, estimates at only the crossover points are adequate for constructing maps of the sea level field and there is no need to estimate the smoothed field at any other locations.

6. DISCUSSION AND CONCLUSIONS

The summary of past altimeter studies in section 2 showed that determination of absolute sea surface topography by satellite altimetry is presently limited by uncertainties in the marine geoid and orbit height. As a result of significant improvements in precision orbit determination, orbit errors are now less than 10 cm, which has greatly extended the utility of altimeter data. The marine geoid will continue to be the limiting factor until a dedicated gravity-mapping satellite can be launched to map the global marine geoid with an accuracy of a few centimeters on scales of ~50 km and longer. The accuracy of presently available estimates of the marine geoid is ~30 cm overall. Because the longer scales of the marine geoid are known most accurately, accurate estimates of the mean sea surface topography are limited to only very large spatial scales.

Uncertainties in the marine geoid and orbit height become much less important if interest is restricted to studies of sea level variability: the marine geoid is eliminated because it is time invariant over the duration of a satellite mission, and time-dependent orbit errors can essentially be eliminated by simple statistical techniques. Numerous studies have shown that variance statistics can be reliably computed from altimeter data and analyzed to study ocean variability on geographical scales that cannot be addressed by other data sets. The global geographical distribution and dynamics of eddy variability have been investigated from sea level variances and wavenumber spectra derived from altimeter data. The anisotropy of surface velocity variability near topographic features and in the vicinity of intense currents has been investigated over the Southern Ocean from surface geostrophic velocity variances estimated from altimeter data. The velocity variances have also been used to investigate eddy transfer of momentum in strong, horizontally sheared mean flows.

The variance statistics that can be readily obtained from altimeter data do not fully exploit the information content of the data. For many applications, it is desirable to map the time evolution of the sea level field. This is a much more difficult problem as it requires an understanding of the resolution capability of altimeter data. To date, the scales considered in studies of mapped sea level variability vary widely and have been chosen rather arbitrarily.

In this paper, a method has been presented for quantifying the resolution capability of an arbitrarily sampled data set. The emphasis has been on altimeter data, but the method is applicable to any irregularly sampled data set. The focus here is on deriving sea level fields for applications such as descriptive studies of sea level variability and model validation. Ultimately, it may be possible to derive higher spatial and temporal resolution sea level fields by combining the data with a model through some form of sophisticated data assimilation. Before this is done, however, the information content of the data alone must be established. The method here identifies the scales at which reliable sea level fields can be derived from altimeter data.

The starting point for application of the method is to concede that a practical limitation of the coarse grid formed by the ground track pattern and asynoptic sampling of the grid is that altimeter data can only resolve large-scale, low-frequency variability. The data must therefore be smoothed to some degree to reduce the effects of aliasing of unresolved variability. The term aliasing is used here for irregularly spaced data in a more

general context than the classical meaning of the term, as discussed in section 3.2. The objective is to smooth the data to the minimum degree necessary, thereby preserving as much of the information content of the data as possible.

The methodology is based on the equivalent transfer function, which is easily computed as the Fourier transform of the weights of an arbitrary linear, smoothed estimate. The equivalent transfer function defines how the spectral content of the observations (signal plus noise) is filtered in a smoothed estimate of the sea level field at a specific location in space and time and for a specific choice of smoothing parameters. The equivalent transfer function also provides an efficient way to describe systematic patterns in the sampling characteristics that are often difficult to detect by other means. The 3-day subcycle in the ground track pattern of altimeter satellites shown in section 3.4 is a relatively simple example. A more complicated example where the equivalent transfer function has proven useful is in the determination of which tidal frequencies can significantly alias altimeter estimates of large-scale, low-frequency sea level variability for a specific orbit configuration (Schlax and Chelton, 1993).

The equivalent transfer function only identifies the wavenumbers and frequencies at which contamination of the low-frequency, low-wavenumber scales of interest is potentially a problem. As such, the equivalent transfer function is not sufficient to determine the resolution capability of the irregularly sampled data set. If there is no signal energy at these wavenumbers and frequencies then there is no contamination of the low-pass band. The mean squared error formalism in section 4 quantifies the degree of contamination by combining the equivalent transfer function and the signal spectrum to quantify the accuracy of the smoothed estimate of the field for a specific location and a specific degree of smoothing. In practice, the relative expected squared bias (RESB) contribution to the mean squared error is usually sufficient to determine the resolution capability of the data set.

The method thus requires that the shape of the signal spectrum be prescribed a priori in order to compute the RESB. For the sea level signal of interest here, the signal autocorrelation function was assumed to be Gaussian in space and time with spatial and temporal scales of 50 km and 30 days. This form was adopted on the basis of independent estimates from hydrographic data. The RESBs presented here are pessimistic if these decorrelation scales are too short.

The procedure for determining the resolution capability is straightforward but involves a large volume of information that must be examined to determine the degree of smoothing necessary to obtain sensible fields from the irregularly spaced observations. The wavenumber-frequency content of the linear estimate and the RESB in general vary with the time and location of the estimation point and with the specified degree of smoothing. The approach requires determination of the RESB at a large number of estimation points for a wide range of smoothing parameters. For the quadratic loess smoother used here (see Appendix A), the smoothing parameters are the half spans of the smoother in the three dimensions. For the Gauss-Markov smoothers discussed in Appendix B, the smoothing parameters are the correlation time scales in the three dimensions and the signal-to-noise variance ratio.

From this multidimensional array of RESB values (three dimensions for the estimation points plus three additional dimensions for the smoothing parameters), the recom-

mended approach is to find a fixed combination of smoothing parameters that yields a spatially and temporally homogeneous field of RESB. There is no unique solution for the "best" combination of smoothing parameters; the smoothing parameter in one dimension can be traded off against smoothing parameters in the other dimensions to obtain different resolutions with the same RESB.

By fixing the smoothing parameters to the same values at all estimation points, the wavenumber-frequency content of the estimated field is spatially and temporally homogeneous. This is a rather different philosophy than is usually adopted in the statistical literature. Statisticians generally select the smoothing parameter of a linear estimate according to the variance of the estimate (as opposed to the expected squared bias used here). The spans are then related to the number of observations in a linear estimate, rather than to the physical space spanned by the smoother. As shown in section 5.1, this approach causes the wavenumber-frequency content of the estimates to vary spatially and temporally, depending on the sampling distribution (see also Schlax and Chelton, 1992, section 2.3). Fixing the smoothing parameters to the same values everywhere yields estimates with essentially the same low-pass band at all estimation points.

In general, the RESB varies with estimation location when a fixed combination of smoothing parameters is used everywhere. This is why the recommended strategy is to seek a fixed combination of smoothing parameters that yields a spatially and temporally homogeneous field of RESB. The resolution capability in densely sampled areas of the data set is thus deliberately degraded by "oversmoothing" to the lower resolution that can be resolved in the sparsely sampled areas. The philosophy of this approach is that it is preferable to sacrifice the higher resolution that is possible at the densely sampled areas than to produce smoothed fields with spatially and temporally inhomogeneous spectral content and RESB that are purely an artifact of the data distribution and smoothing procedure.

If the interest is in short-scale variability, then low-pass filtering by the recommended approach is clearly undesirable. These shorter scales of variability can be retained as long as attention is restricted to the areas of the data record where they are adequately resolved. If the entire data set is to be analyzed as a single record, then the data must be low-pass filtered to retain only the long scales that are resolved everywhere in the data set.

Application of the method to the GEOSAT data set concludes that the spatial and temporal scales of sea level variability that can be resolved are about 3° or 4° of latitude and longitude by about 30 days for estimates constructed at the crossovers of ascending and descending ground tracks. At shorter spatial and temporal scales, the RESB of the smoothed estimates varies substantially over the GEOSAT 17-day repeat period and with location in the grid of crossovers.

It should be kept in mind that the estimates of sampling errors presented in section 5 neglect the effects of measurement errors. Residual orbit errors in GEOSAT data are likely to render the resolution capability deduced here somewhat optimistic. The estimates of sampling error are also based on the nominal GEOSAT sampling pattern and thus assume complete data coverage. Because of problems with GEOSAT attitude control, seasonally varying data dropouts at middle and high latitudes were common along descending ground tracks in the northern hemisphere and ascending ground tracks in the

southern hemisphere (see Cheney et al., 1988, Figure 2). This data loss also renders the GEOSAT resolution capability deduced here overly optimistic at locations and times of significant data dropouts.

The resolution capability of 3° or 4° by 30 days is adequate for studies of large-scale, low-frequency sea level variability. This is generally too large, however, for mapping mesoscale variability such as short-scale meanders in the flow and detachment and subsequent drift of rings. At the present time, there are two satellite altimeters simultaneously observing the global sea level variability. The ERS-1 altimeter launched in July 1991 and the TOPEX altimeter launched in August 1992 are expected to continue providing useful data for several years. By combining data from these two altimeters, it will be possible to map the sea level variability with higher spatial and temporal resolution than can be obtained from either altimeter individually. It is a straightforward application of the formalism presented in this paper to quantify the spatial and temporal resolution capability of the combined ERS-1 and TOPEX data sets.

APPENDIX A. QUADRATIC LOESS SMOOTHERS

Loess smoothers are discussed extensively by Cleveland and Devlin (1988) and Schlax and Chelton (1992). The quadratic loess estimate at time t_0 is defined to be a local weighted least squares fit of a quadratic function of t to the N observations nearest t_0,

$$\hat{h}(t) = a_1 + a_2 t + a_3 t^2 . \tag{A.1}$$

The smoothed estimate is the least-squares fit Eq. (A.1) evaluated at t_0. The coefficients a_1, a_2 and a_3 are determined by minimizing the function

$$\Phi = \frac{1}{W} \sum_{j=1}^{N} w_j^2 \left[\hat{h}(t_j) - h(t_j)\right]^2 , \tag{A.2}$$

where W is the sum of the weights w_j, defined by the bell-shaped function

$$w_j = \begin{cases} (1 - q_j^3)^3 & 0 \le q_j \le 1 \\ 0 & q_j > 1 \end{cases} \tag{A.3a}$$

$$q_j = \left(\frac{t_j - t_0}{d_t} \right)^2 . \tag{A.3b}$$

The parameter d_t is the half-span of the loess smoother.

The loess smoother formalism is easily extended to three dimensions, in which case there are ten least squares parameters a_i and the bell-shaped weighting function is ellipsoidal with half-spans d_x, d_y and d_t,

$$q_j = \left[\left(\frac{x_j - x_0}{d_x} \right)^2 + \left(\frac{y_j - y_0}{d_y} \right)^2 + \left(\frac{t_j - t_0}{d_t} \right)^2 \right] . \tag{A.4}$$

The quadratic loess estimate can be expressed in the standard form of a linear estimate Eq. (2) by the impulse response method. This is most easily seen from Eq. (3). Suppose that the only observation is $g_k = 1$. In this case, $g(t) = \delta(t - t_k)$, i.e., an impulse at time t_k. By the sifting property of the Dirac delta function (Bracewell, 1978, p. 75), the loess smoothed estimate Eq. (3) then reduces to

$$\begin{aligned}\hat{h}_k(t_0) &= \hat{p}(t_k; t_0) \\ &= \sum_{j=1}^{N} \alpha_j(t_0)\delta(t_k - t_j) \\ &= \alpha_k(t_0)\,.\end{aligned} \tag{A.5}$$

The smoother weight for the observation at time t_k is therefore the quadratic loess smoothed estimate Eq. (A.5) obtained by replacing the N observations with a single observation that has unit value at time t_k and values of zero at all other observation times. The N smoother weights α_j in Eq. (2) are thus derived by constructing N such quadratic loess estimates, one for an impulse function at each of the observation times t_j.

After obtaining the weights α_j for the particular smoothing parameter d_t by the impulse response method, it is straightforward to determine the filtering characteristics of the quadratic loess smoother from the equivalent transfer function Eq. (6). The equivalent transfer functions for 1-dimensional quadratic loess smoothers with evenly and irregularly spaced observations are discussed in section 3.2 (see Figure 10).

APPENDIX B. GAUSS-MARKOV SMOOTHERS

The formalism for Gauss-Markov estimation (also known as objective analysis) has been presented many times before (e.g., Gandin, 1965; Alaka and Elvander, 1972; Bretherton et al., 1976; Daley, 1991). The essential elements are reviewed here to establish a framework for investigating the filtering properties of Gauss-Markov estimates through the equivalent transfer function introduced in section 3.1. The smoother weights that minimize the mean squared error of the linear estimate Eq. (2) are given by

$$\alpha_j(t_0) = \sum_{i=1}^{N} \tilde{D}_{ij} A_i(t_0)\,, \tag{B.1}$$

a result known as the Gauss-Markov theorem. In Eq. (B.1),

$$\begin{aligned}A_i(t_0) &= \frac{\langle h(t_0)h(t_i)\rangle}{\langle h^2(t)\rangle} \\ &\equiv \rho(t_0 - t_i)\end{aligned} \tag{B.2}$$

is the signal autocorrelation at lag $(t_0 - t_i)$ and $\tilde{D}_{ij}$ is the i,jth element of the inverse of the $N \times N$ cross correlation matrix of the data observations g_j. The elements of this cross correlation matrix are

$$D_{ij} = P_{ij} + \lambda^{-1} N_{ij}\,, \tag{B.3}$$

where

$$\begin{aligned}P_{ij} &= \frac{\langle h(t_i)h(t_j)\rangle}{\langle h^2(t)\rangle} \\ &\equiv \rho(t_i - t_j)\end{aligned} \tag{B.4}$$

is the signal autocorrelation at lag $(t_i - t_j)$,

$$N_{ij} = \frac{\langle \epsilon_i \epsilon_j \rangle}{\langle \epsilon^2 \rangle} \equiv \eta(t_i - t_j) \tag{B.5}$$

is the autocorrelation of the measurement errors and λ is the signal-to-noise variance ratio.

The linear estimate constructed from smoother weights given by Eq. (B.1) is optimal (i.e., has the lowest mean square error of all linear estimates of the form Eq. (2)) only if the true autocorrelations $\rho(\tau)$, $\eta(\tau)$ (where τ is time lag) and signal-to-noise ratio λ are used. Moreover, the expected squared error of estimates computed by this formalism are valid only if the correct values for these parameters are used. If these three parameters are specified in a more arbitrary manner (perhaps because of ignorance of the true values or in order to filter the data as described below), the solution is more appropriately referred to as suboptimal or Gauss-Markov estimation. The latter term will be used here.

In order to investigate the filtering properties of Gauss-Markov estimates, the signal autocorrelation function $\rho(\tau)$ will be assumed to be a Gaussian function of time lag, $\rho(\tau) = e^{-(\tau/\tau_0)^2}$, and the measurement errors will be assumed to be uncorrelated, $\eta(t_i - t_j) = \delta_{ij}$. The equivalent transfer functions of the corresponding Gauss-Markov estimates for error-free measurements and signal correlation time scales of $\tau_0 = 30$ and 60 are shown in Figure 19a for evenly spaced observations at sample interval $\Delta = 1$. The transfer functions are characterized by a flat low-frequency pass band with unit amplitude and a very sharp cutoff at frequency $f_c \approx 1.2\tau_0^{-1}$. The filtering properties can thus be controlled by adjusting the signal correlation time scale τ_0, analogous to adjusting the half span d_t of the quadratic loess smoother considered in Appendix A and section 3.2. The series of high frequency peaks centered on even multiples of the Nyquist frequency $f_N = (2\Delta)^{-1}$ are the aliasing peaks discussed in section 3.2 for the loess smoother (see Figure 10a).

The effects of measurement errors are shown in Figures 19b and c, which are the equivalent transfer functions of Gauss-Markov estimates with signal correlation time scale $\tau_0 = 30$ and signal-to-noise ratios of $\lambda = 1$ and 0.1, respectively. With increasing measurement error variance (decreasing λ), the amplitude of the transfer function in the pass band decreases, the pass band cutoff frequency f_c shifts to lower frequencies and the sharpness of the band-edge rolloff decreases. In the limit of zero signal-to-noise ratio, the equivalent transfer function collapses to zero everywhere, corresponding to a linear estimate of zero. Note that the equivalent transfer function for $\lambda = 1$ is not significantly different from that of the quadratic loess smoother shown in Figure 10a, apart from the slightly less than unit value across the low-frequency pass band.

The direct incorporation of statistical information about measurement errors is an important advantage of Gauss-Markov estimates. The quadratic loess smoother and other commonly used linear smoothers have near unit amplitude across the entire low-frequency pass band, regardless of the measurement error characteristics. These other estimates therefore pass all of the low-frequency spectral energy of the measurement errors as well as of the signal of interest. As suggested by Press et al. (1992, section 13.3), the amplitudes of the transfer functions of the more traditional smoothers can be reduced in mag-

nitude to mitigate the effects of measurement errors on the variance of the linear estimates.

The equivalent transfer functions of Gauss-Markov estimates for an example of irregularly spaced observations are shown in Figure 20 for signal-to-noise variance ratios of $\lambda = \infty$, 1 and 0.1. The low-pass bands of interest are essentially the same as those of their counterpart equivalent transfer functions for uniformly spaced observations shown in Figure 19. The noisy continuums of energy in the transfer functions at frequencies higher

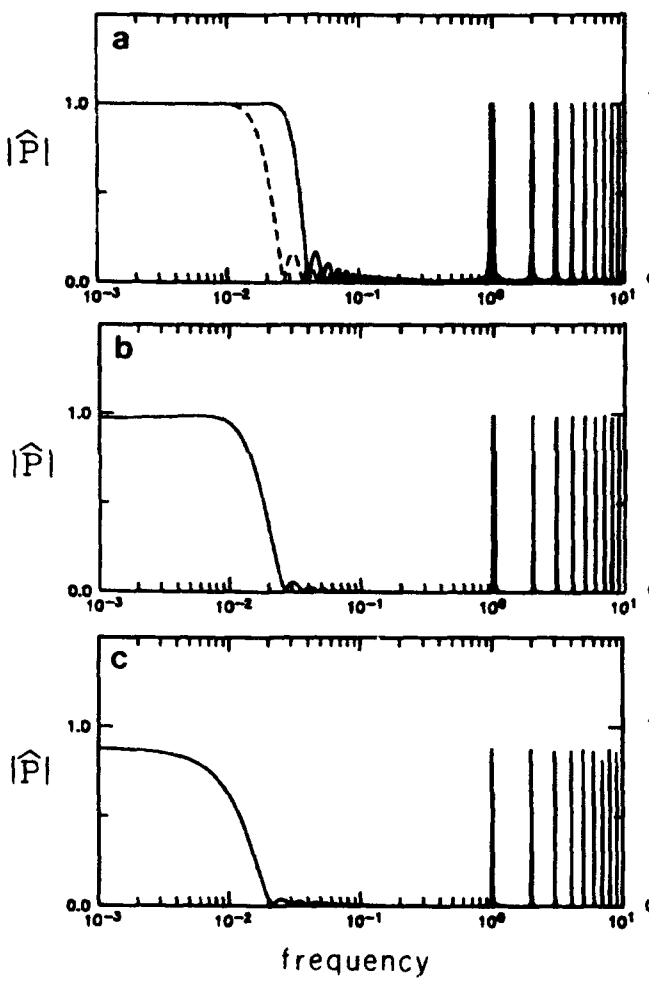


Figure 19. The 1-dimensional equivalent transfer function modulii of the Gauss-Markov smoother with Gaussian autocorrelation function and an evenly spaced sample design for a) signal-to-noise variance ratio $\lambda = 10{,}000$ and signal correlation scales $\tau_0 = 30$ (solid line) and $\tau_0 = 60$ (dashed line); b) $\lambda = 1$ and $\tau_0 = 30$; and c) $\lambda = 0.1$ and $\tau_0 = 30$. In all three panels, the estimation point is at the midpoint of the data record.

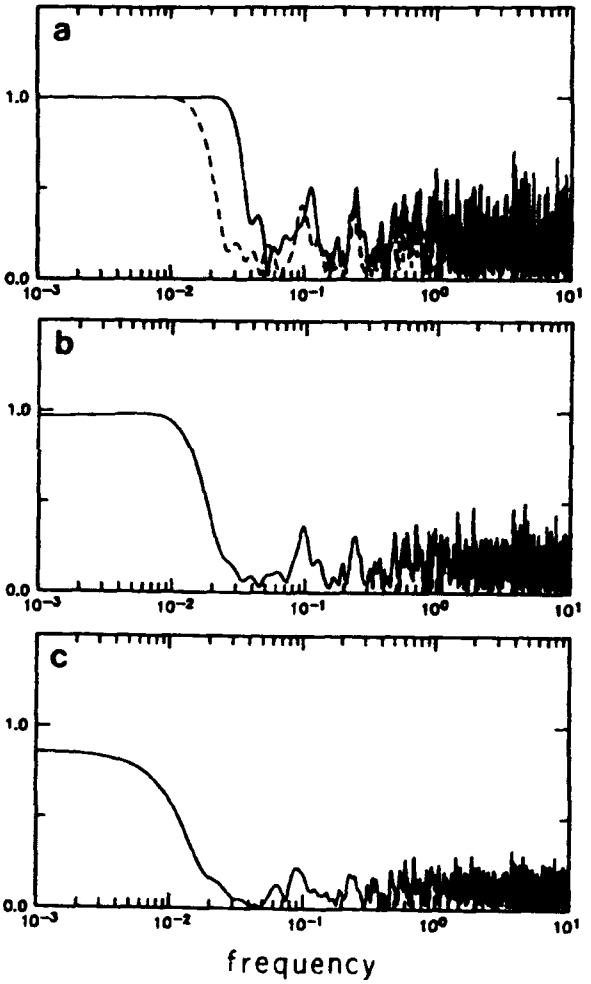


Figure 20. The same as Figure 19, except for an irregularly spaced sample design.

than f_c represent the effects of "aliasing" from the uneven sample design, as discussed in section 3.2. The details of this aliasing depend on the particular sample design and on the estimation time t_0.

Although it is not generally viewed in this context, it is apparent from the equivalent transfer functions in Figures 19 and 20 that Gauss-Markov estimation can be considered as a low-pass smoother. In this sense, it is just like any of the other smoothers that are commonly used to low-pass filter a noisy or irregularly sampled data set. The low-pass cutoff frequency and sharpness of the band-edge rolloff of the equivalent transfer function are controlled by appropriate choices of τ_0 and λ. It should be noted that Gauss-Markov estimates with arbitrarily prescribed τ_0 and λ are not the optimal estimate of low-pass filtered data. Such an optimal estimate can be constructed, however, by an extension of the Gauss-Markov formalism to find the minimum mean squared error estimate for the linear filtering operator applied to the data.

The disadvantage of Gauss-Markov estimates is the computational effort required to obtain the inverse of the $N \times N$ cross correlation matrix needed to determine the smoother weights by Eq. (B.1). If the primary interest is to obtain low-pass filtered estimates of $h(t)$ (as it is in this study), this computational effort is unwarranted; the filtering properties of the quadratic loess smoother considered in Appendix A and section 3.2 are very similar to those of the Gauss-Markov smoothers for realistic signal-to-noise ratios of $\lambda = 1$ to 10. The computational effort required for quadratic loess estimates is much smaller for large values of N.

ACKNOWLEDGMENTS

The authors wish to thank Peter Müller, Greg Holloway and Phyllis Haines for their kind hospitality organizing and hosting the 'Aha Huliko'a Hawaiian Winter Workshop and for their patience waiting for this (long) manuscript. We also thank Victor Zlotnicki for providing the data in Figure 3a in digitized form and Rosemary Morrow for permission to include her yet-unpublished results in Figure 5. Financial support to attend the workshop is also greatly appreciated. The manuscript was written while one of the authors (D.B.C.) was a visiting scientist at the CSIRO Marine Laboratories in Hobart, Tasmania. We wish to express gratitude to the CSIRO Division of Oceanography and to John Church in particular for their generous hospitality and the use of CSIRO facilities. This research was supported by contract 958127 from the Jet Propulsion Laboratory funded under the TOPEX Announcement of Opportunity.

REFERENCES

Alaka, M. A., and R. C. Elvander, 1972: Optimum interpolation from observations of mixed quality. *Mon. Wea. Rev.*, **100**, 612–624.

Arnault, S., Y. Menard, and J. Merle, 1990: Observing the tropical Atlantic Ocean in 1986–1987 from altimetry. *J. Geophys. Res.*, **95**, 17,921–17,945.

Arnault, S., A. Morlière, J. Merle, and Y. Ménard, 1992: Low-frequency variability of the tropical Atlantic surface topography: Altimetry and model comparison. *J. Geophys. Res.*, **97**, 14,259–14,288.

Bracewell, R. N., 1978: *The Fourier Transform and Its Applications.* McGraw-Hill, 444 pp.

Bretherton, F. P., R. E. Davis, and C. B. Fandry, 1976: A technique for objective analysis and design of oceanographic experiments applied to MODE-73. *Deep-Sea Res.*, **23**, 559-582.

Buja, A., T. Hastie, and R. Tibshirani, 1989: Linear smoothers and additive models. *Annals Stat.*, **17**, 453–555.

Carter, E. F., and A. R. Robinson, 1987: Analysis models for the estimation of oceanic fields. *J. Atmos. Oceanic Technol.*, **4**, 49–74.

Charney, J. G., 1971: Geostrophic turbulence. *J. Atmos. Sci.*, **28**, 1087–1095.

Chelton, D. B., 1988: WOCE/NASA Altimeter Algorithm Workshop. *U.S. WOCE Tech. Rep. 2*, 70 pp., U.S. Planning Office for WOCE, College Station, Tex.

Chelton, D. B., and M. G. Schlax, 1993: Spectral characteristics of time-dependent orbit errors in altimeter height measurements. *J. Geophys. Res.*, **98**, in press.

Chelton, D. B., E. J. Walsh, and J. L. MacArthur, 1989: Pulse compression and sea level tracking in satellite altimetry. *J. Atmos. Oceanic Technol.*, **6**, 407–438.

Chelton, D. B., M. G. Schlax, D. L. Witter, and J. G. Richman, 1990: Geosat altimeter observations of the surface circulation of the Southern Ocean. *J. Geophys. Res.*, **95**, 17,877–17,903.

Cheney, R. E., and J. G. Marsh, 1981: Seasat altimeter observations of dynamic topography in the Gulf Stream region. *J. Geophys. Res.*, **86**, 473–483.

Cheney, R. E., J. G. Marsh, and B. D. Beckley, 1983: Global mesoscale variability from repeat tracks of Seasat altimeter data. *J. Geophys. Res.*, **88**, 4343–4353.

Cheney, R. E., B. C. Douglas, R. W. Agreen, L. Miller, and N. S. Doyle, 1988: The NOAA Geosat geophysical data records: Summary of the first year of the exact repeat mission. *Tech. Mem. NOS NGS-48*, 20 pp., Natl. Oceanic and Atmos. Admin., Boulder, Colo.

Cheney, R. E., B. C. Douglas, and L. Miller, 1989: Evaluation of Geosat altimeter data with application to tropical Pacific sea level variability. *J. Geophys. Res.*, **94**, 4737–4747.

Cleveland, W. S., and S. J. Devlin, 1988: Locally weighted regression: An approach to regression analysis by local fitting. *J. Am. Stat. Assoc.*, **83**, 596–610.

Daley, R., 1991: *Atmospheric Data Analysis.* Cambridge University Press, 457 pp.

Daniault, N., and Y. Ménard, 1985: Eddy kinetic energy distribution in the Southern Ocean from altimetry and FGGE drifting buoys. *J. Geophys. Res.*, **90**, 11,877–11,889.

Delcroix, T., J. Picaut, and G. Eldin, 1991: Equatorial Kelvin and Rossby waves evidenced in the Pacific Ocean through Geosat sea level and surface current anomalies. *J. Geophys. Res.*, **96**, 3249–3262.

De Mey, P., and A. R. Robinson, 1987: Assimilation of altimeter eddy fields in a limited-area quasi-geostrophic model. *J. Phys. Oceanogr.*, **17**, 2280–2293.

Emery, W. J., W. G. Lee, and L. Magaard, 1984: Geographic and seasonal distributions of Brunt Väisälä frequency and Rossby radii in the North Pacific and North Atlantic. *J. Phys. Oceanogr.*, **14**, 294–317.

Frankignoul, C., and P. Müller, 1979: Quasi-geostrophic response of an infinite β-plane ocean by stochastic forcing by the atmosphere. *J. Phys. Oceanogr.*, **9**, 104–127.

Fu, L.-L., and V. Zlotnicki, 1989: Observing oceanic mesoscale eddies from Geosat altimetry: Preliminary results. *Geophys. Res. Lett., 16*, 457–460.

Gandin, L. S., 1965: *Objective Analysis of Meteorological Fields.* Israel Program for Scientific Translations, Jerusalem, 242 pp.

Gordon, A. L., and W. F. Haxby, 1990: Agulhas eddies invade the South Atlantic: Evidence from GEOSAT altimeter and shipboard conductivity- temperature-depth survey. J. Geophys. Res., 95, 3117-3125.

Jacobs, G. A., G. H. Born, M. E. Parke, and P. C. Allen, 1992: The global structure of the annual and semiannual sea surface height variability from Geosat altimeter data. *J. Geophys. Res.*, **97**, 17,813–17,828.

Johnson, M., 1989: Southern Ocean surface characteristics from FGGE buoys. *J. Phys. Oceanogr.*, **19**, 696–705.

Jourdan, D., C. Boissier, A. Braun, and J. F. Minster, 1990: Influence of wet tropospheric correction on mesoscale dynamic topography as derived from satellite altimetry. *J. Geophys. Res.*, **95**, 17,993–18,004.

Le Traon, P. Y., M. C. Rouquet, and C. Boissier, 1990: Spatial scales of mesoscale variability in the North Atlantic as deduced from Geosat data. *J. Geophys. Res.*, **95**, 20,267–20,285.

Levitus, S., 1982: Climatological atlas of the world ocean. *NOAA Prof. Pap. 13*, U. S. Dept. of Commerce.

Lutjeharms, J. R. E., and Van Ballegooyen, R. C., 1988: The retroflection of the Agulhas Current. *J. Phys. Oceanogr.*, **18**, 1570–1583.

Matano, R. P., M. G. Schlax, and D. B. Chelton, 1993: Seasonal variability in the southwestern Atlantic. *J. Geophys. Res.*, in press.

Matthews, P. E., M. A. Johnson, and J. J. O'Brien, 1992: Observation of mesoscale ocean features in the northeast Pacific using Geosat radar altimetry data. *J. Geophys. Res.*, **97**, 17,829–17,840.

Ménard, Y., 1983: Observations of eddy fields in the northwest Atlantic and northwest Pacific by SEASAT altimeter data. *J. Geophys. Res.*, **88**, 1853–1866.

Ménard, Y., 1988: Observing the seasonal variability in the tropical Atlantic from altimetry. *J. Geophys. Res.*, **93**, 13,967–13,978.

Morrow, R., R. Coleman, J. A. Church, and D. Chelton, 1993: Surface eddy momentum flux and velocity variances in the Southern Ocean from Geosat altimetry. *J. Phys. Oceanogr.*, submitted.

Morrow, R., J. Church, R. Coleman, D. Chelton, and N. White, 1992: Eddy momentum flux and its contribution to the Southern Ocean momentum balance. *Nature*, **357**, 482–484.

Müller, P., and C. Frankignoul, 1981: Direct atmospheric forcing of geostrophic eddies. *J. Phys. Oceanogr.*, **11**, 287–308.

Nerem, R. S., B. D. Tapley, and C. K. Shum, 1990: Determination of the ocean circulation using Geosat altimetry. *J. Geophys. Res.*, **95**, 3163–3179.

Pares-Sierra, A., W. B. White, and C. K. Tai, 1993: Wind-driven coastal generation of annual mesoscale eddy activity in the California Current. *J. Phys. Oceanogr.*, **23**, 1110–1121.

Parke, M. E., R. H. Stewart, D. L. Farless, and D. E. Cartwright, 1987: On the choice of orbits for an altimetric satellite to study ocean circulation and tides. *J. Geophys. Res.*, **92**, 11693–11707.

Périgaud, C., and P. Delecluse, 1992: Annual sea level variations in the southern tropical Indian Ocean from Geosat and shallow-water simulations. *J. Geophys. Res.*, **97**, 20,169–20,178.

Press, W. H., S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, 1992: *Numerical Recipes in Fortran: The Art of Scientific Computing.* Cambridge University Press, 963 pp.

Priestley, M. B., 1992: *Spectral Analysis and Time Series.* Academic Press, 890 pp.

Rapp, R. H., 1992: Computation and accuracy of global geoid undulation models. Sixth International Geodetic Symposium on Satellite Positioning, Columbus, OH.

Rapp, R. H., and N. K. Pavlis, 1990: The development and analysis of geopotential coefficient models to spherical harmonic degree 360. *J. Geophys. Res.*, **95**, 21,885–21,911.

Rapp, R. H., Y. M. Wang, and N. K. Pavlis, 1991: The Ohio State 1991 geopotential and sea surface topography harmonic coefficient models. *Rep. 410*, Dep. of Geod. Sci. and Surv., Ohio State Univ., Columbus, Aug. 1991.

Ray, R. D., 1993: Global ocean tide models on the eve of TOPEX/Poseidon. *Trans. on Geoscience and Remote Sensing*, **31**, 355–364.

Richardson, P. L., 1983: Eddy kinetic energy in the North Atlantic Ocean from surface drifters. *J. Geophys. Res.*, **88**, 4355–4367.

Robinson, A. R., 1983: *Eddies in Marine Science.* Springer-Verlag, 609 pp.

Schlax, M. G., and D. B. Chelton, 1992: Frequency domain diagnostics for linear smoothers. *J. Am. Stat. Assoc.*, **87**, 1070–1081.

Schlax, M. G., and D. B. Chelton, 1993: Detecting aliased tidal errors in altimeter height measurements. *J. Geophys. Res.*, (submitted).

Shen, C. Y., J. C. McWilliams, B. A. Taft, C. C. Ebbesmeyer, and E. J. Lindstrom, 1986: The mesoscale spatial structure and evolution of dynamical and scalar properties observed in the northwestern Atlantic Ocean during the POLYMODE Local Dynamics Experiment. *J. Phys. Oceanogr.*, **16**, 454–482.

Shum, C. K., R. A. Werner, D. T. Sandwell, B. H. Zhang, R. S. Nerem, and B. D. Tapley, 1990: Variations of global mesoscale eddy energy observed from Geosat. *J. Geophys. Res.*, **95**, 17,865–17,876.

Tai, C.-K., 1988: Estimating the basin-scale ocean circulation from satellite altimetry, I, Straightforward spherical harmonic expansion. *J. Phys. Oceanogr.*, **18**, 1398–1413.

Tai, C.-K., 1989: Accuracy assessment of widely used orbit error approximations in satellite altimetry. *J. Atmos. Oceanic Technol.*, **6**, 147–150.

Tai, C.-K., 1991: How to observe the gyre to global-scale variability in satellite altimetry: Signal attenuation by orbit error removal. *J. Atmos. Oceanic Technol.*, **8**, 272–288.

Tai, C.-K., W. B. White, and S. E. Pazan, 1989: Geosat crossover analysis in the tropical Pacific 2. Verification analysis of altimetric sea level maps with expendable bathythermograph and island sea level data. *J. Geophys. Res.*, **94**, 897–908.

Tokmakian, R. T., and P. G. Challenor, 1993: Observations in the Canary Basin and the Azores Frontal Region using Geosat data. *J. Geophys. Res.*, **98**, 4761–4773.

Treguier, A. M., and J. C. McWilliams, 1990: Topographic influences on wind-driven, stratified flow in a β-plane channel: An idealized model of the Antarctic Circumpolar Current. *J. Phys. Oceanogr.*, **20**, 321–343.

van Gysen, H., R. Coleman, R. Morrow, B. Hirsch, and C. Rizos, 1992: Analysis of collinear passes of satellite altimeter data. *J. Geophys. Res.*, **97**, 2265–2277.

Wagner, C. A., 1986: Accuracy estimate of geoid and ocean topography recovered jointly from satellite altimetry. *J. Geophys. Res.*, **91**, 453–461.

Wagner, C. A., 1989: Summer school lectures on satellite altimetry. *Theory of Satellite Geodesy and Gravity Field Determination*, F. Sansò and R. Rummel, *Lecture Notes in Earth Science*, **25**, Springer-Verlag, 285–334.

Wagner, C. A., 1991: How well do we know the deep ocean tides? An intercomparison of altimetric, hydrodynamic and gauge data. *Manuscr. Geod.*, **16**, 118–140.

White, W. B., C. K. Tai, and J. DiMento, 1990: Annual Rossby wave characteristics in the California Current Region from the Geosat Exact Repeat Mission. *J. Phys. Oceanogr.*, **20**, 1297–1311.

Wolff, J.-O., E. Maier-Reimer, and D. L. Olbers, 1991: Wind-driven flow over topography in a zonal β-plane channel: A quasi-geostrophic model of the Antarctic Circumpolar Current. *J. Phys. Oceanogr.*, **21**, 236–264.

Wunsch, C., and E. M. Gaposchkin, 1980: On using satellite altimetry to determine the general circulation of the oceans with application to geoid improvement. *Rev. Geophys. Space Phys.*, **18**, 725–745.

Zlotnicki, V., L.-L. Fu, and W. Patzert, 1989: Seasonal variability in global sea level observed with Geosat altimetry. *J. Geophys. Res.*, **94**, 17,959–17,969.

OBSERVING "INTEGRATING" VARIABLES IN THE OCEAN

Douglas S. Luther
Department of Oceanography, School of Ocean and Earth Science and Technology
University of Hawaii at Manoa, Honolulu, Hawaii 96822

Alan D. Chave
Department of Geology and Geophysics, Woods Hole Oceanographic Institution
Woods Hole, Massachusetts 02543

ABSTRACT

Some physical variables are natural spatial integrals of oceanic water motion or state properties. Observation of these variables permits isolation of physical processes that might otherwise be difficult to examine because of the superposition of many phenomena at one place. Independent of a particular physical model, observations of such integrating quantities frequently enable direct determination of relatedness between variables at different locations, and direct determination of causality, while more traditional point observations may fail to find such relationships. Furthermore, integral quantities such as volume and heat transport, which are now being studied with great fervor because of their climatic importance, are likely more accurately estimated using observations of "integrating" variables than using a set of point measurements. Examples of integrating types of variables, such as horizontal electric fields, vertical acoustic travel time and bottom pressure, are used to demonstrate the ideas above with examples drawn from the study of (a) atmospherically forced, mesoscale motions, and (b) the volume and heat transports of the Gulf Stream.

INTRODUCTION

At any particular location in the oceans, the sub-inertial water motions and fluctuations of state properties are likely to be due to a superposition (and, possibly, interaction) of a variety of phenomena that each have specific and different balances between acceleration, advection, Coriolis forces, pressure, dissipation, external forcing, and so on. Time-dependent boundary layers exist as a result of property fluxes to and from the atmosphere and earth. Semi-permanent meso- and gyre-scale currents (O(100 km) and O(1000 km), respectively) of the "general circulation" are forced by the winds and property fluxes, and, through instabilities, produce meso- and gyre-scale variability in the form of meandering currents, coherent vortices, radiated waves, and so on. Meso- and gyre-scale variability can also be directly driven by the atmosphere. Each of these, and many other unlisted phenomena, exist at a variety of space scales for each time scale, so that they overlap each other not only in physical space and time, but in frequency and wavenumber space as well.

To decipher the ocean's physics, it is often preferable to examine a single phenomenon at a time. Then one has to consider "contamination" from the other phenomena that would inhibit, for instance, direct detection of relatedness among oceanic variables and between oceanic and atmospheric variables.

There are of course a number of strategies for isolating particular phenomena in order to study their kinematics and dynamics. Sometimes, time series of variables are all that is needed to separate phenomena according to their characteristic frequencies. Other times, spatial information is needed, which raises the cost and difficulty of a field experiment, but which allows discrimination of wavenumbers or principal components and thereby possible discrimination of different processes. Frequently, experiments are designed so that there is a reasonable certainty that the phenomenon to be studied dominates all other processes. However, there are many instances when this cannot be accomplished. In these cases, observations are usually compared with model output visually, graphically, statistically, or through dynamical parameter estimation. Such comparisons can lead to the identification of the quality of the dynamical hypotheses as a function of frequency and/or wavenumber. It is not unusual for experiments to be designed to take advantage of most if not all of the strategies above.

The purpose of this note is to point out that there now exists an additional observational strategy, most components of which are rather new to oceanography, for isolating phenomena that are large scale in the vertical and/or horizontal. This strategy is based on the measurement of integrating variables. The spatial filtering inherent in these variables frequently enables statistical confirmation of important large scale kinematic and dynamic relationships which might otherwise go undetected except with a formidably large array of point measurements. Yet, in deference to the theme of this workshop, it must be acknowledged that isolating large scale phenomena does not imply that the phenomena observed, or statistics of these phenomena estimated from integrating variables, are homogeneous over large scales as well. This inhomogeneity complicates, if not invalidates, the application of many statistical procedures that assume homogeneity.

We define integrating variables as those that are natural spatial integrals of oceanic water motion or state properties. Table 1 lists a few of the more important integrating variables being observed today. These integrating variables are ones that by their very nature tend to filter out the shorter spatial scale variability. The techniques we'll discuss in this note are those whose usefulness is well-established and which offer the advantage of cost-effectiveness. In addition, these techniques may have greater accuracy in comparison to using a suite of point measurements when the ultimate goal of an investigation is the measurement of an integral quantity such as volume transport.

Table 1. Examples of Integrating Variables.

Variable	Principal component of integrand	References
Horizontal electric fields at a point	Conductivity-weighted horizontal water velocity, from seafloor to sea surface	Sanford (1971) Chave & Luther (1990) Luther et al. (1991)
Voltages across fixed horizontal distances (typically, using abandoned undersea telephone cables)	Conductivity-weighted horizontal water velocity (one component only), from seafloor to sea surface over a fixed horizontal distance	Larsen & Sanford (1985) Larsen (1992) Chave et al. (1992b)
Vertical acoustic travel time	Inverse sound speed from seafloor to sea surface	Watts & Rossby (1977) Pickart & Watts (1990)
Bottom pressure	Horizontal water velocity near the seafloor, over a fixed horizontal distance	Brown et al. (1975) Whitworth & Peterson (1985)
Horizontal acoustic travel time (acoustic thermometry)	Inverse sound speed along horizontal, depth-varying ray paths	Munk & Forbes (1989)
Reciprocal acoustic travel time	Water velocity along horizontal, depth-varying ray paths	Worcester (1977) Worcester et al. (1991)
Orientation of the earth's axis of rotation	Global mass distribution (especially in hydrologic reservoirs)	Chao (1988) Eubanks (1993)
Rotation rate of the earth	Global atmospheric angular momentum (principally, fluctuations in zonal winds)	Hide & Dickey (1991) Eubanks (1993)

Measurement of integrating variables allows the investigator to focus immediately on a specific region of wavenumber space, without the "contamination" of shorter scale variability that may depend on processes other than the one being sought. Furthermore, such restriction of the wavenumber space may enable the detection of properties (like spatial coherence or air-sea coherence) that tend to zero as the wavenumber bandwidth increases and may provide more useful constraints for numerical model simulations than do point measurements.

In the sections that follow, we will describe applications of three of the more underutilized, yet most cost-effective, integrating variables listed in Table 1, including point measurements of horizontal electric fields (HEFs), vertical acoustic travel time (VATT), and bottom pressure (P_b). We will show how observations of HEFs and P_b in the Barotropic, Electromagnetic and Pressure Experiment (BEMPEX) provided definitive evidence of the existence of gyre-scale motions that are directly forced by sea surface winds. Horizontal electric field data from the Synoptic Ocean Prediction (SYNOP) experiment will be shown that suggest the greater accuracy of these integrating variables in estimates of volume transport. And, we will outline the potential utility of combining HEF and VATT observations to obtain nearly direct estimates from the seafloor of heat transport and the gravest vertical structures of horizontal currents and temperature fluctuations.

HORIZONTAL ELECTRIC FIELDS

Motional electromagnetic induction is now theoretically well understood in certain idealized settings (e.g., Sanford, 1971; Chave and Luther, 1990). Assuming distant continental boundaries and a flat seafloor with laterally homogeneous conductivity, then for the low-frequency limit where the aspect ratio of ocean currents is small, where the effect of self induction is weak, and where the vertical velocity can be neglected in comparison with the horizontal velocity, it can be shown that the point HEFs are related to horizontal water velocity by

$$\vec{E}_h(x,y,t) = CF_z k \times < \vec{v}_h(x,y,t) >^* + \frac{\vec{J}^*}{\sigma} + \vec{N}(x,y,t), \tag{1}$$

where

$$< \vec{v}_h(x,y,t) >^* = \frac{\int_{-H}^{0} dz\, \sigma(x,y,z,t)\, \vec{v}_h(x,y,z,t)}{\int_{-H}^{0} dz\, \sigma(x,y,z,t)} \tag{2}$$

and is called the conductivity-weighted, vertically averaged (CWVA) water velocity; $\vec{v}_h(x,y,z,t)$ is horizontal water velocity; $\sigma(x,y,z,t)$ is seawater electrical conductivity; F_z

is the vertical component of the geomagnetic field; and $H(x,y)$ is ocean depth. The scale factor C depends on $\sigma(z \le -H)$; C can be estimated by intercomparisons, but extensive geophysical evidence suggests that $C = 0.95 \pm 0.05$ almost everywhere in the deep oceans (e.g., Chave and Luther, 1990). A noise term $\bar{N}(x,y,t)$ is composed of externally produced (i.e., in the ionosphere and magnetosphere) electromagnetic fields that dominate for periods shorter than a few days but are negligible at longer periods (e.g., Chave et al., 1989).

Locally and non-locally produced electric currents are represented by $\vec{J}^*$. Given usual oceanic scales of motion at sub-inertial periods (greater than half a pendulum day), locally produced electric currents are theoretically negligible if the bottom is flat (Chave and Luther, 1990) or the flow is aligned along isobaths (Stephenson and Bryan, 1992). Local generation of $\vec{J}^*$ may be sufficient to inhibit accurate estimation of ocean water currents with electric fields only where the currents cut across isobaths and then only if the underlying sediments are relatively non-conductive (Larsen, 1992; Stephenson and Bryan, 1992). Meandering of a narrow current like the Gulf Stream can theoretically produce non-zero $\vec{J}^*$ outside of the stream boundaries (the principal example of non-local generation of $\vec{J}^*$), which theoretically could be a large noise relative to the electric field signal induced by the smaller water currents there. However, Sanford (1986) has pointed out that observations have shown $\vec{J}^*/\sigma$ to be small [yielding errors of O(1 cm/s)] and generally negligible. And our own work with the SYNOP data has shown that the best agreement between the moored current meter data and the horizontal electrometer data occurs where the currents are moderate to weak, resulting in no detectable $\vec{J}^*$. Therefore, in the following, $\vec{J}^*$ is ignored.

Dropping the horizontal dependences and letting $\sigma(z,t)$ equal a vertical average part plus a residual, i.e.,

$$\sigma(z,t) = <\sigma(t)> + \hat{\sigma}(z,t), \text{ where } <\sigma(t)> = \frac{1}{H}\int_{-H}^{0} dz\, \sigma(z,t), \tag{3}$$

then Eq. 2 becomes

$$<\vec{v}_h(t)>^* = <\vec{v}_h(t)> + \frac{\int_{-H}^{0} dz\, \hat{\sigma}(x,t)\, \vec{v}_h(x,t)}{H <\sigma(t)>}. \tag{4}$$

The first term on the right hand side (RHS) of Eq. (4) is just the vertical average of horizontal water velocity (or depth-normalized transport per unit width). The second term on the RHS of Eq. (4) is a small contribution because seawater conductivity is a weak function of depth. Note that $|\hat{\sigma}(z,t)| \ll <\sigma(t)>$; $<\sigma(t)>$ is always greater than 3 Siemens m^{-1}; to a good approximation, $\sigma(z,t) = 3.0 + 0.09T(z,t)$ Siemens m^{-1}, where $T(z,t)$ is

unless the horizontal currents are strong and baroclinic (i.e., have large vertical shear). Assuming low-frequency motions so that $\bar{N}(t)$ can be ignored, and using Eq. (4), the components of Eq. (1) in the northern hemisphere become

$$\frac{E^u_{-y}(t)}{C|F_z|} = <u(t)> + \frac{1}{H}\int_{-H}^{0} \frac{\hat{\sigma}(z,t)}{<\sigma(t)>} u(z,t)\,dz, \tag{5a}$$

$$\frac{E^v_{x}(t)}{C|F_z|} = <v(t)> + \frac{1}{H}\int_{-H}^{0} \frac{\hat{\sigma}(z,t)}{<\sigma(t)>} v(z,t)\,dz, \tag{5b}$$

where the subscript on E denotes the direction in which that term is positive and the superscript indicates the horizontal water velocity component to which it is proportional. Clearly, the HEFs are integrating variables in the sense defined in the introduction, being proportional to the vertical average of horizontal water velocity plus a small "contamination" due to conductivity. The conductivity contribution is negligible if conductivity is independent of depth in the ocean (as it is at very high latitudes) or if the horizontal water velocity has little vertical shear (as frequently occurs at mid- to high-latitudes).

Normal Modes

To elucidate further the relationships in Eq. (5), it is useful to expand the various quantities in terms of the dynamical normal modes. While any complete basis set could be used, the dynamical normal modes have a vertical dependence that permits rapid convergence of the expansions of horizontal velocity and seawater conductivity and temperature. Despite the fact that the dynamical normal modes are obtained from the equations of motion by various simplifying assumptions such as no mean flow and linearity, in using these modes here we are not making any assumptions about the underlying dynamics of the quantities being observed. The modes are simply the most convenient basis set for our purposes.

The dynamical modes are obtained from the equations of motion by assuming no mean flow, mean stratification in hydrostatic balance, and a flat bottom. With horizontal velocity and pressure proportional to $\phi_i(z)$, and vertical velocity and density perturbations proportional to $\theta_i(z)$, the equations for Boussinesq linear waves (small perturbations) then separate into sets of equations for the horizontal/time dependence and vertical dependence. Specifically, with

$$\begin{Bmatrix} \vec{v}_h(z,t) \\ p(z,t)/\rho_o \end{Bmatrix} = \sum_{i=0}^{\infty} \begin{Bmatrix} \vec{a}_{h,i}(x,y,t) \\ d_i(x,y,t) \end{Bmatrix} \phi_i(z) \tag{6a}$$

and

$$\begin{Bmatrix} w(z,t) \\ \rho'(z,t)g/N^2(z) \end{Bmatrix} = \sum_{i=0}^{\infty} \begin{Bmatrix} f_i(x,y,t) \\ g_i(x,y,t) \end{Bmatrix} \theta_i(z) \tag{6b}$$

where $\rho = \rho_* + \bar{\rho}(z) + \rho'(z,t)$ and $N^2(z) = -\frac{g}{\rho_*}\frac{d\bar{\rho}}{dz}$, then ϕ_i and θ_i satisfy

$$\phi_i = \frac{d\theta_i}{dz} \text{ and } \frac{d\phi_i}{dz} = -\frac{N^2}{\gamma_i^2}\theta_i \tag{6c}$$

with the boundary conditions

$$\theta_i = 0 \text{ or } \frac{d\phi_i}{dz} = 0 \text{ at } z = -H, \tag{6d}$$

$$\frac{d\theta_i}{dz} - \frac{g}{\gamma_i^2}\theta_i = 0 \text{ or } \frac{d\phi_i}{dz} + \frac{N^2}{g}\phi_i = 0 \text{ at } z = 0. \tag{6e}$$

Equations (6c) are solved numerically with the appropriate boundary conditions to produce the vertical structure functions and eigenvalues, γ_i^2. The structure functions are orthogonal and are normalized so that

$$\frac{1}{H}\int_{-H}^{0} \phi_i \phi_j \, dz = \delta_{ij}. \tag{6f}$$

This normalization means that the ϕ_i are non-dimensional, while the θ_i have dimensions of length. We now have a complete orthonormal basis set for describing any quantity that varies with depth. The fact that these modes are "tuned" to the depth-dependences of real oceanographic variables makes them very useful, since it means expansions in terms of these modes should converge rapidly. For our purposes in this section, it is not important what assumptions were used to obtain the vertical structure equations in Eq. (6c).

Let's now expand conductivity in terms of the dynamical modes, viz.,

$$\frac{\hat{\sigma}(z,t)}{<\sigma(t)>} = \sum_{i=1}^{\infty} s_i(x,y,t)\phi_i(z), \tag{7}$$

where the i=0 (barotropic) mode has been dropped since the vertical average of $\hat{\sigma}$ is zero. Substituting this expression and those in Eq. (6a) into Eq. (5), and truncating after mode number 1, yields

$$\frac{E_{-y}^{u}(t)}{C|F_z|} \approx a_{x,0} + s_1 a_{x,1}, \tag{8a}$$

$$\frac{E_{x}^{v}(t)}{C|F_z|} \approx a_{y,0} + s_1 a_{y,1}, \tag{8b}$$

The truncation in Eq. (8) is quite reasonable given the expected decrease in modal current amplitudes with increasing mode number, and given the examples in Table 2 of mean s_i, calculated using Levitus (1982) data to compute structure functions and conductivity profiles. Table 2 suggests that in polar oceans and in the mid-latitude Pacific the conductivity-weighting contribution to the electric field is completely trivial. Using a year of electric field and current meter mooring data, Luther et al. (1991) have shown the validity of Eq. (5) in a region of the mid-latitude North Pacific with weak mean currents and weak baroclinic eddy activity. In that location, the conductivity-weighting contribution to the electric field was completely trivial.

Table 2. Expansion coefficients for conductivity from Eq. (7).

	s_i		
Mode (i)	32.5°N Atlantic	42.5°N Pacific	57.5°S Atlantic
1	0.119	0.017	-0.004
2	-0.014	0.021	-0.009
3	-0.012	-0.002	-0.004
4	0.008	0.008	-0.001

On the other hand, in the mid-latitude North Atlantic, equal amplitudes of the barotropic (i=0) and first baroclinic (i=1) modes of current imply a ~12% relative contribution to the electric field from the last term on the RHS of Eq. (5). Rossby (1987) found first baroclinic mode amplitudes up to 70% greater than barotropic mode amplitudes in the Gulf Stream at 73°W, with very small second and third baroclinic mode amplitudes. Consequently, in the Gulf Stream we can expect up to 20% conductivity-weighting contributions to the electric field due to the first baroclinic mode of current. In fact, our work with SYNOP data has shown occasional conductivity-weighting contributions up to 30%, although the mean contribution is <15%.

The variation of conductivity with depth in the oceans (e.g., Luther et al., 1991) suggests dominance of the first baroclinic mode in the conductivity contribution to Eq. (5), which allows the use of another integrating measurement, vertical acoustic travel time, to estimate first baroclinic mode current and temperature amplitudes in order to remove the conductivity contribution from the HEF. This procedure is outlined later.

Horizontal Electrometer (HEM) Versus Mooring Estimates of Transport

In deployments of seafloor HEMs to date, where comparison of HEM estimates of the vertically averaged horizontal water velocity, $< \vec{v}_h(t) >$, with current meter mooring

estimates of the same quantity have been possible, the HEM estimates have proven to be more accurate. These results provide an example of how measurement of an integrating variable provides a more accurate estimation of oceanic behavior than can be accomplished with a suite of conventional point measurements. In this specific case such accuracy has significant importance to climate studies that rely on estimates of transport (which is directly proportional to horizontal integrals of $H< \bar{v}_h(t) >$ for determining the world ocean's role in climate fluctuations.

The first HEM vs. mooring comparison of $< \bar{v}_h(t) >$ estimates was produced by Luther et al. (1991) from data collected during BEMPEX, an experiment that deployed a large number of HEMs and pressure gauges in the North Pacific to study direct atmospheric forcing of gyre-scale eddies (the results of which are discussed further below). The accuracy of the HEM estimates of $< \bar{v}_h(t) >$ was corroborated by current estimates made by reciprocal tomography, which is based on measuring reciprocal acoustic travel time differences (Table 1). The inaccuracy of the current meter mooring estimates was attributed primarily to stalling of the current meter rotors in the weak flows below 1000 meters depth. Another electrometer-mooring comparison presented below comes from the opposite extreme for oceanic flows, i.e., from the Gulf Stream which has strong currents at all depths so that rotor stalling is not expected to be a problem.

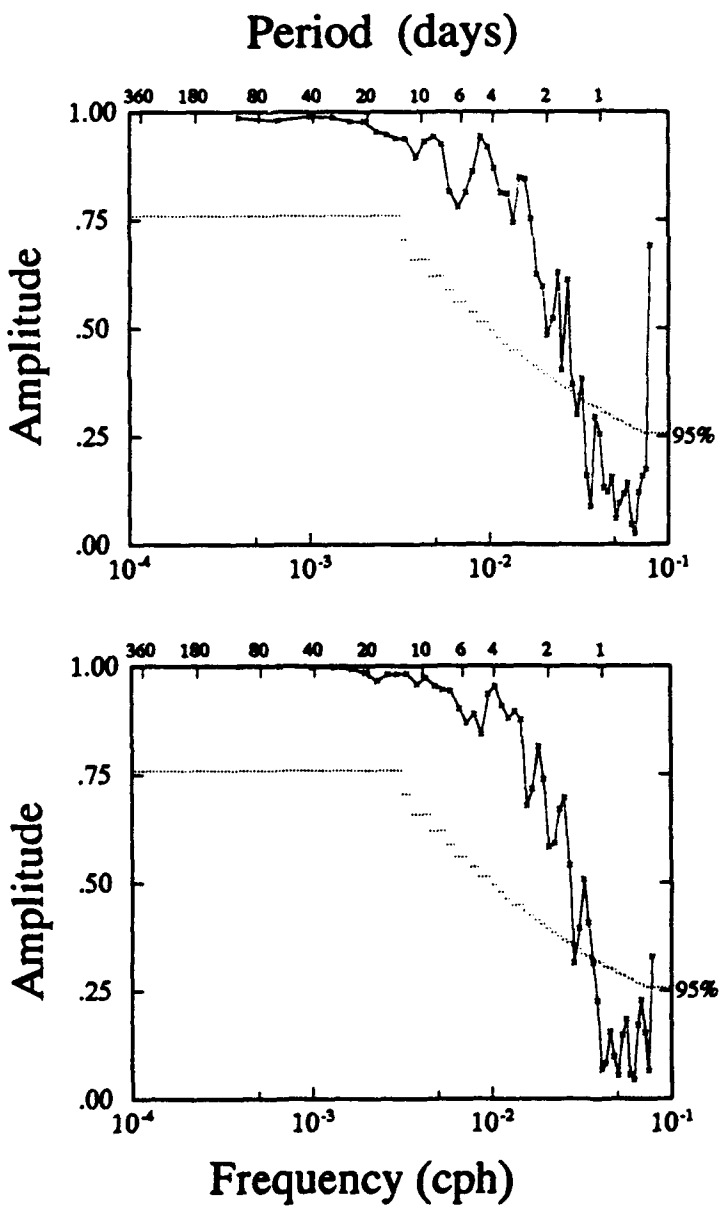
The Office of Naval Research provided funds for us (with Jean Filloux) to deploy four of Filloux' seafloor HEMs (Filloux, 1987) next to current meter moorings during the last year ('89-'90) of the SYNOP experiment in the Gulf Stream. The HEMs were deployed in an array centered around 37.5°N, 68.5°W, at depths near 4700 m. Near each HEM were sub-surface moorings deployed by J. Bane, T. Shay, R. Watts, and W. Johns, carrying current meters at nominal depths of 400 m, 700 m, 1000 m and 3500 m.

The LHSs of Eqs. (5a) and (5b) were obtained from the HEM data using C=0.95, as per estimates of C made by Sanford et al. (1985) in the western North Atlantic, and using an appropriate estimate of F_z for the time and location of the experiment, while the RHSs were estimated from the mooring data. The latter estimation included extrapolation of $\bar{v}_h$, temperature and pressure to a fictitious nominal 100-m depth, conversion of pressure to 'depth,' and estimation of conductivity using temperatures and a climatological temperature/salinity relation in an empirical formula. The currents and conductivities at the four real and one "fictitious" instruments were trapezoidally integrated, taking into account the time dependence of the depths of the instruments. The time series thus obtained, representing opposite sides of Eq. (5), are highly coherent, as shown in Figure 1.

While the coherence in Figure 1 is very encouraging, and the rms differences between the estimates of the LHS and RHS of Eq. (5) are no larger for instance than what has been considered very good agreement for testing schemes to remove the effects of mooring motion from current meter data (e.g., Hogg, 1991; Cronin, 1991), examination of the

individual time series (not shown) indicates that the LHS of Eq. (5) consistently has a greater magnitude than the RHS. That there is a systematic under-estimation of current by the mooring data, or an over-estimation by the HEM data, is most easily seen by casting the data in terms of a Gulf Stream coordinate system, rather than a geographic coordinate system, since the Gulf Stream position and direction vary with time.

Daily locations of the temperature front of the Gulf Stream (provided by R. Watts and W. Johns) were determined from an array of Inverted Echo Sounders (IESs) that measure VATT. These locations permitted estimation of the cross-stream positions of the moorings every day. Gulf Stream directions at each mooring were determined from the 400-1000 m shear. Daily estimates of Eq. (5) were then rotated into these Gulf Stream coordinates.



To put the problem in a more interesting context, the conductivity contribution term (the last term on the RHS of Eq. (5)) was moved to the LHS of Eq. (5). Now we can compare mooring estimates of vertically averaged water velocity, $< \bar{v}_h(t) >$, with HEM estimates of the same quantity (that incorporates a small

Figure 1. Coherence amplitudes between electrometer and mooring estimates of conductivity-weighted, vertically averaged water velocity (left and right hand sides of Eq. (5), as described in the text). Top is for zonal currents (Eq. (5a). The 95% level of no significance is indicated. Every other point plotted is independent due to a 50% overlap of frequency band-averaging.

mooring-derived conductivity correction), all cast in terms of Gulf Stream coordinates. The results for a single mooring are shown in Figure 2, along with the difference (error) between the two estimates. (Note that the error is not dependent upon which side of Eq. (5) we place the conductivity correction term.) The results in Figure 2 typify the comparisons made at other HEM locations. Integrating the estimates in Figure 2 across the stream results in a ~30% higher estimate of total transport from the HEM than from the mooring. This is certainly non-trivial.

Figure 2 shows good agreement between the estimates at distances farther "south" than -60 km and farther "north" than 30 km from the north wall of the Gulf Stream. The percentage difference between the estimates is not constant across the stream, implying that the difference is not due to a calibration error of the HEM. While there are many possible noises and small errors in the HEM data, none is known to result in an overestimate of velocity. We believe the error arises in the current meter data and/or its trapezoidal integration, but to date we have clearly identified only one source of error, which by itself, however, is insufficient to account for all of the error in Figure 2. Conductivity and temperature versus depth (CTD) data taken at this longitude by M. Hall indicate that the extrapolation of currents to the near-surface underestimates the upper ocean velocities (at and north of the maximum current) and the trapezoidal integration, which implies linear interpolation between 1000 and 3500 m, underestimates the transport

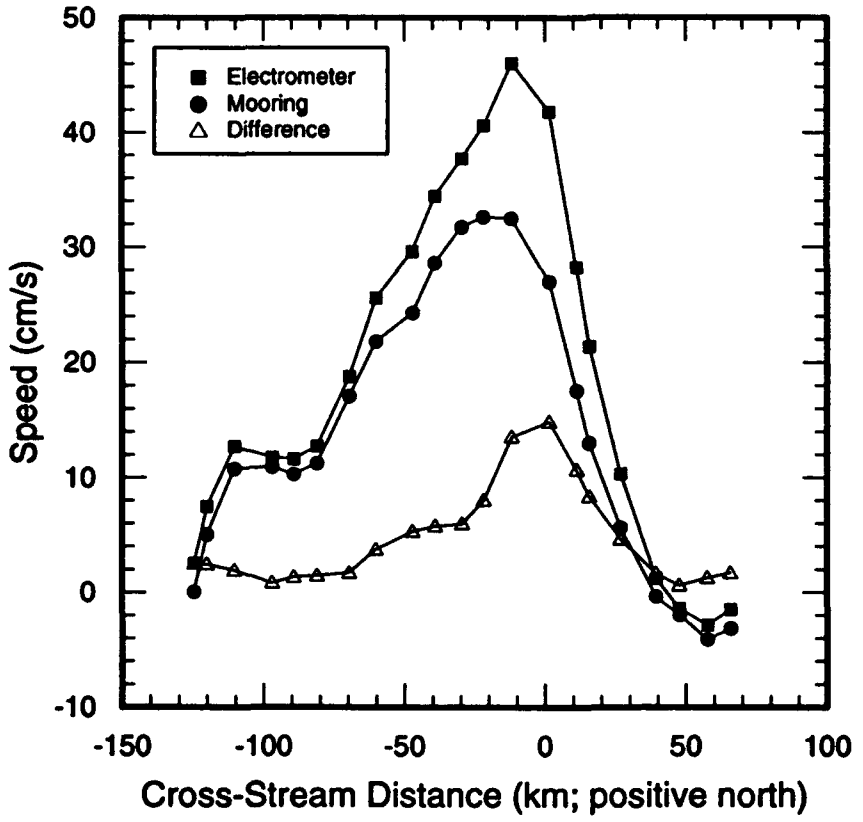


Figure 2. Two estimates of vertically averaged water velocity in the Gulf Stream at nominally 68°W, as a function of cross-stream distance, using electrometer and mooring data (see text). The differences between these estimates are also plotted.

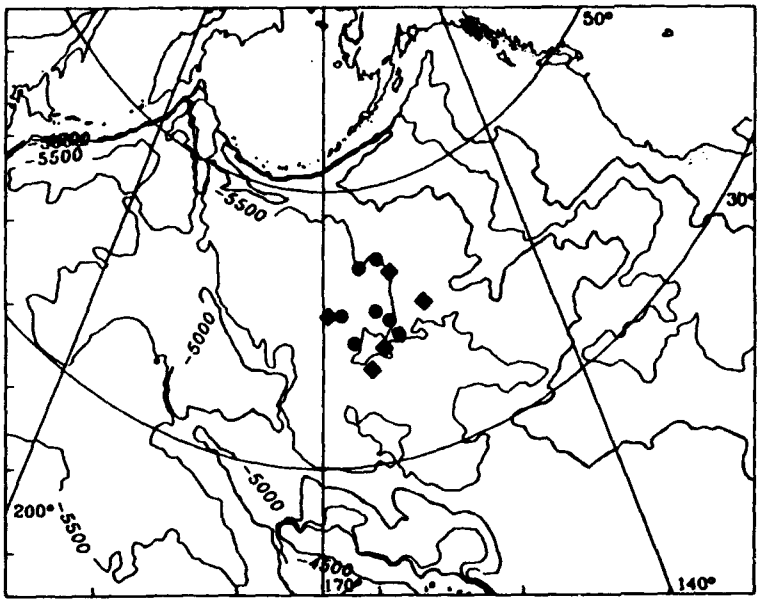


Figure 3. Polar stereographic projection of the North Pacific Ocean displaying seafloor isobaths from 4000 to 5500 m, and the locations of seafloor electrometers (solid circles) and pressure gauges during BEMPEX. Adjacent land masses are also shown.

between 1000 and 3500 m (again, at and north of the maximum current). Error from the trapezoidal integration is further suggested by the fact that the error time series is most highly coherent with currents measured at 1000 m.

Observation of Atmospheric Forcing of Sub-Inertial Gyre-Scale Eddies

A good example of the use of measurements of integrating variables to explore a phenomenon that defied unambiguous detection with traditional point measurements is the Barotropic, Electromagnetic and Pressure Experiment (BEMPEX). BEMPEX employed HEMs and bottom pressure gauges to specifically test theories (Frankignoul and Müller, 1979; Müller and Frankignoul, 1981) of stochastic forcing by the atmosphere of sub-inertial gyre-scale motions in the ocean. BEMPEX, fielded by ourselves with Jean Filloux and funded by the National Science Foundation, obtained seven HEF records and five bottom pressure records from a two-dimensional array spanning 1000 km centered around 40°N, 163°W (Figure 3) and lasting 11 months beginning in August, 1986. Luther et al. (1991) showed that the conductivity contribution (Eq. (5)) to the HEFs in BEMPEX was trivial, so that the HEFs were directly proportional to vertically averaged (barotropic) water velocity. In the following, we'll simply refer to the barotropic currents, rather than the HEFs, obtained from the HEMs.

Four of the observational strategies discussed in the introduction were employed in the design of BEMPEX: first, isolation, i.e., a region of the North Pacific was chosen for which it could be reasonably assumed that other sources of energy for gyre-scale motions (such as instabilities of strong "mean" currents) were weak; second, measurements of

integrating variables, HEFs and P_b, were planned, because the theories predicted that the oceanic response to atmospheric forcing would be essentially barotropic at the sub-inertial periods (i.e., a few days to a few months) that we could observe reasonably well with a one year record; third, a spatial array was planned for confirmation of theoretical predictions of frequency-wavenumber relations; and, fourth, visual and graphical comparisons with published model outputs of statistical parameters were planned. All these process discrimination strategies were employed because previous experiments had found that detection of atmospherically forced gyre-scale motions was difficult with traditional point measurements of currents and because the point measurements showed significant spatial inhomogeneities in what evidence of this phenomenon they did find. Figure 4 is presented as an example of how the integrating variable HEF readily provided evidence of atmospheric forcing, while at the same time measurements of currents in the surface mixed layer did not, probably because of the superposition of many phenomena in the mixed layer that have different, destructively interfering relationships with the surface atmospheric variables.

The Frankignoul and Müller papers listed above were the first papers to present the physics of atmospherically forced meso- and gyre-scale motions (which have the form of linear Rossby waves) in the realistic light of stochastic forcing; and, most important to empiricists, they presented testable hypotheses in the form of intervariable transfer and coherence functions. One example of the latter in flat-bottomed basins is the prediction of

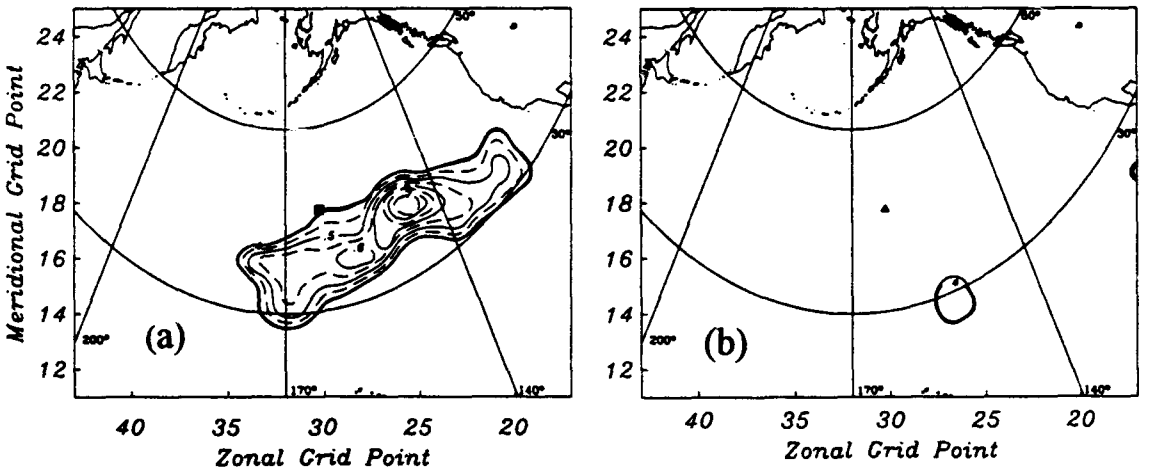


Figure 4. (a) Contour plot of squared coherence amplitude between BEMPEX zonal barotropic current (measured at the solid square) and surface zonal wind stress (at each grid point), in the 10-15 day period band. The wind stress was calculated (Chave et al., 1992b) from the Fleet Numerical Oceanography Center's surface wind product. Only squared coherence amplitudes greater than the 95% level of no significance are plotted. The large region of significant coherence indicates a strong relationship between oceanic barotropic (depth-independent) zonal current and atmospheric forcing. (b) As for (a), except with zonal current measured at nominally 73 m depth on a sub-surface mooring located near the electrometer in (a). The lack of significant coherence is interpreted as a null result, providing no information on the relatedness of oceanic near-surface zonal current and surface zonal winds.

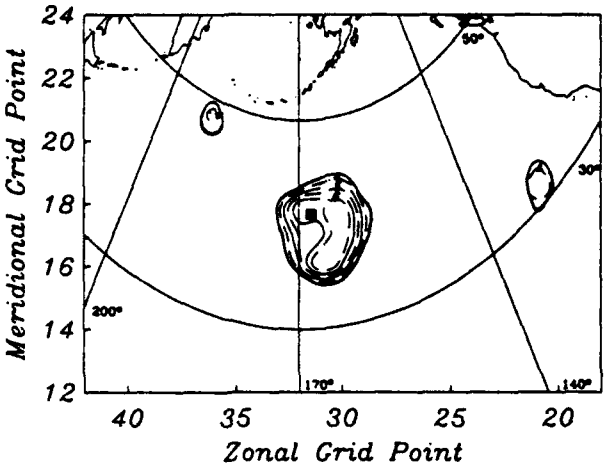


Figure 5. Contour plot of squared coherence amplitude between BEMPEX meridional barotropic current (measured at the solid square) and surface wind stress curl (at each grid point), in the 25-70 day band. Plotted as in Figure 4. The significant coherence surrounding the solid square suggests the oceanic meridional barotropic current is nearly in Sverdrup balance with the wind stress curl (see text). This is the only location (out of 7), and the only period band at this location, that exhibited a Sverdrup-like behavior.

strong coherence between meridional currents and local wind stress curl at periods greater than O(100 days). The coherence arises from the dominance of a "Sverdrup" balance in the vorticity conservation equation, in which the curl of the wind stress, which is a source of vorticity, is balanced by a meridional advection of planetary vorticity. The coherence does not occur at shorter periods due to destructive interference from many shorter scale waves with non-trivial relative vorticity. For basins with gently sloping bottoms, a "topographic Sverdrup" balance obtains between wind stress curl and oceanic currents that are perpendicular to isopleths of potential vorticity, f/H, where f is the Coriolis parameter.

Evidence for the flat-bottom Sverdrup relation was found in BEMPEX (Fig. 5), and evidence for the topographic Sverdrup relation was reported by Niiler and Koblinsky (1985). But, the coherence shown in Figure 5 did not occur at any other period for that instrument, nor was there Sverdrup-like coherence at any period for the other six HEMs. Furthermore, a systematic search of North Pacific current meter records by Koblinsky et al. (1989) produced no further examples of a topographic Sverdrup balance of oceanic currents. The problem lies with the generation of short-scale Rossby waves by short-scale topography as the wind stress curl drives the water across isopleths of f/H (Anderson and Corry, 1985; Cummins, 1991). The short-scale waves have substantial relative vorticity, so that a Sverdrup balance usually does not dominate vorticity conservation until very long periods. Cummins (1991) demonstrated, with a numerical model of the North Pacific having realistic topography, that by spatially filtering out the shorter scale waves the Sverdrup balance of the longer waves can be recovered. Following Cummins, we have averaged the meridional currents from the five HEMs that comprised a coherent sub-array in BEMPEX (Fig. 3). The resultant averaged meridional currents were coherent with wind stress curl at all periods >10 days; Figure 6 shows the coherences from two period bands.

BEMPEX yielded many significant statistics (Chave et al., 1992b) with which to determine the kinematics of the oceanic Rossby waves and with which to test Müller and

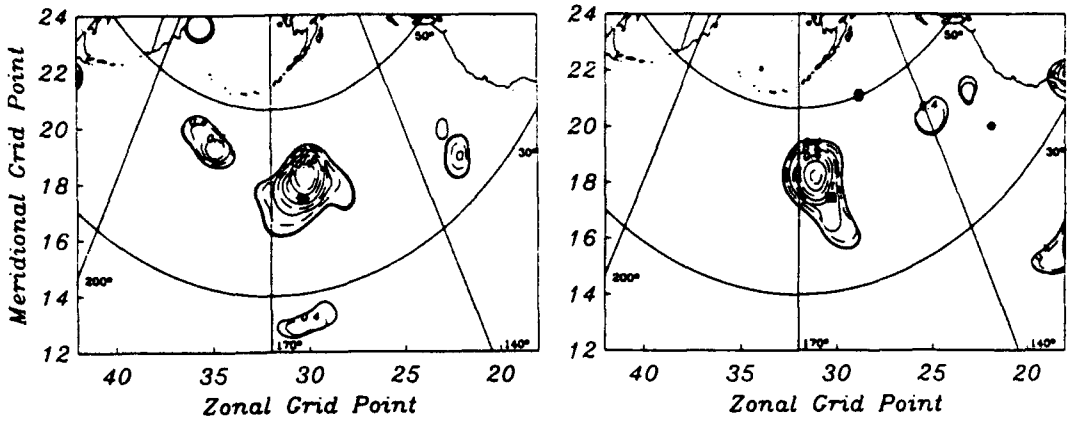


Figure 6. Contour plot of squared coherence amplitude between averaged BEMPEX meridional barotropic currents (measured at the 5 southernmost electrometers in Figure 3) and surface wind stress curl, in the (left) 25-70 day band, and (right) 13-19 day band. A Sverdrup-like relationship (see text) is evident in both period bands, and at all other periods greater than 10 days, for the averaged meridional barotropic current. The solid square in both plots locates the nominal center of mass of the five electrometers. Otherwise plotted as in Figure 4.

Frankignoul's (1981) predictions of frequency-dependent local coherence between various oceanic and atmospheric variables. Non-zero coherences between oceanic variables and non-local atmospheric variables, predicted by Brink (1989), were also unambiguously observed (Luther et al., 1990; Chave et al., 1992b). No point measurements of currents have yielded such clear evidence of direct atmospheric forcing of Rossby waves as has been obtained with measurements of the integrating variables, HEF and P_b (the latter to be discussed further below).

The example above, describing efforts to confirm the relatively simple physics inherent in the Sverdrup balance, emphasizes the non-homogeneity of even the larger scale barotropic motions in the ocean. Statistics estimated from observations of these phenomena are correspondingly inhomogeneous. Any observational program or statistical analysis technique, such as some of those highlighted at this workshop, must address these inhomogeneities or risk misdirected inferences.

BOTTOM PRESSURE (P_b)

The complete relationship between pressure and water velocity in the oceans is not easily represented by a simple integral. To lowest order, however, mid-latitude, sub-inertial motions are geostrophic, i.e.,

$$\vec{v}_h = \frac{1}{f\rho_s} k \times \nabla_h p, \tag{9a}$$

which permits the derivation of a simple relationship between pressure and the mass flux per unit vertical distance (Pedlosky, 1987), viz.,

$$p(\xi) = f \int_{\xi_0}^{\xi} k \cdot (\rho_s \vec{v}_h \times d\vec{r}) + p(\xi_0), \tag{9b}$$

where ξ and ξ_0 are two points in a horizontal plane; $d\vec{r}$ is an incremental vector parallel to an arbitary curve running from ξ_0 to ξ, so long as $p(\xi) > p(\xi_0)$; $\rho_s = \rho_* + \bar{\rho}(z)$; and k is the local upward unit vector.

Like HEFs, pressure is related to a spatial integral of horizontal velocity. Unlike HEFs, the spatial distance over which the integral operates is somewhat arbitrary for pressure. But, the greater the separation between members of a set of pressure measurements, the weaker the correlation between them due to the substantial wavenumber bandwidth of oceanic sub-inertial motions. Lack of coherence is usually fatal for process studies but is often considered irrelevant for basin-wide studies of transport, for instance. The integrating nature of pressure is in large part responsible for the successful mapping of the semi-permanent oceanic flows with hydrographic (temperature and salinity versus depth) data, from which pressure is calculated, because smaller scale variability tends to have a weaker impact on pressure (which can be argued from either Eq. (9a) or Eq. (9b)).

In addition to discriminating against smaller scales of motion, bottom pressure discriminates against baroclinic motions in favor of barotropic. This follows, for example, from the vertical structure functions, $\phi_i(z)$, that are determined from Eq. (6). The barotropic mode is independent of depth, while the baroclinic modes have their largest amplitudes near the sea surface. If the barotropic and baroclinic modes have identical total kinetic energy, integrated from the seafloor to the sea surface, then the barotropic mode will have a larger amplitude at the seafloor than any of the baroclinic modes. Since the barotropic and first baroclinic modes typically have similar kinetic energies (and the higher modes are weaker), P_b (but not pressure in the upper ocean) tends to be dominated by large-scale barotropic motions, even in regions of the oceans with energetic baroclinic mesoscale motions such as the western North Atlantic. This latter point accounts for the large horizontal correlation of sub-inertial P_b found over distances of hundreds of kilometers in the western North Atlantic by Brown et al. (1975) during the Mid-Ocean Dynamics Experiment, while horizontal correlations of currents and temperatures in the same general area tend to zero when separations of O(100 km) are attained (Owens, 1985).

Considering the tendency of P_b to be more or less dominated by the large-scale sub-inertial motions, we might expect that P_b in BEMPEX will be less affected by the short-scale waves that made detection of the Sverdrup balance, for instance, so difficult with point measurements of currents or even HEF-derived barotropic currents. In fact, we do find from BEMPEX that P_b is much more coherent with surface atmospheric variables (Fig. 7) than are the barotropic currents. And, the coherence between the pressure records is greater than found for the barotropic currents, despite the larger separation of the pressure gauges (Fig. 3). The extended regions of high squared coherence in Figure 7 are not so much evidence of waves reaching the instrument from all over the Pacific as they are evidence of high horizontal coherence in the surface atmospheric fields themselves. The non-local maximum of the squared coherence in Figure 7 is expected from the dominance of propagating waves over locally forced motions at these periods (Brink, 1989). As the period decreases, the maximum coherence between P_b and air pressure or wind stress curl becomes more local (Luther et al.,

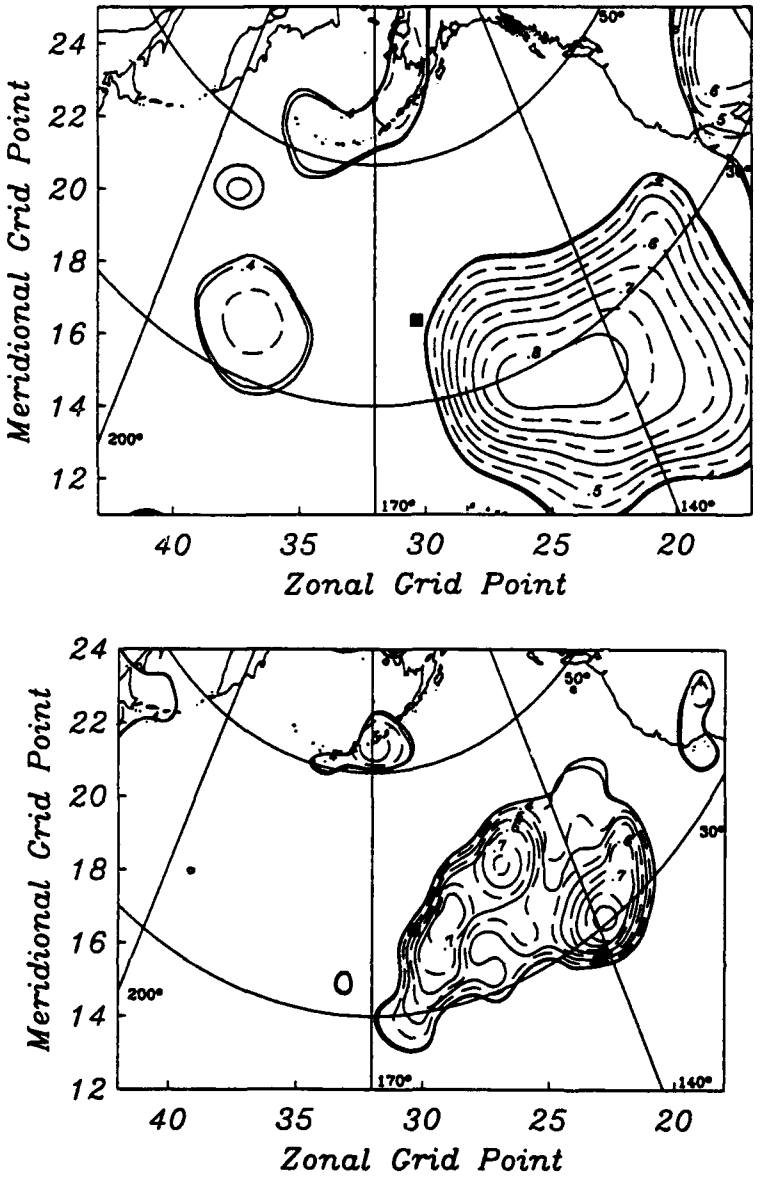


Figure 7. Contour plot of squared coherence amplitude between BEMPEX bottom pressure (measured at the solid square) and (top) surface air pressure, or (bottom) surface wind stress curl, both in the 19-38 day band. Plotted as in Figure 4. The extended regions of strong coherence are typical for the bottom pressure records at most periods, unlike the coherences found for barotropic currents which tended to be weaker, less extensive, more spatially inhomogeneous, and clearly significant in fewer period bands.

1990), in accordance with the disappearance of freely propagating Rossby waves (Müller and Frankignoul, 1981).

Bottom pressure P_b is so dominated by the larger scale barotropic motions that all the P_b records from BEMPEX display very similar coherence relationships with the atmospheric variables, unlike the situation for the barotropic currents which exhibit more inhomogeneities in their relationships with atmospheric variables. For P_b, atmospheric forcing is clear at all sub-inertial frequencies,as seen by the graphs of maximum coherence in Figure 8. The fact that the coherence of P_b with air pressure is frequently higher than its coherence with wind stress curl (Fig. 8) does not implicate a particular forcing mechanism, because the atmospheric variables are coherent among themselves, and there is more noise in wind stress curl than in air pressure. A simple scaling argument shows (Philander, 1978) that divergence of the surface (Ekman) boundary layer, produced by the curl of the wind stress, should dominate all other forcing mechanisms at the time and space scales observed in BEMPEX (Chave et al., 1992a).

COMBINING HEM AND IES MEASUREMENTS

The intent of this section is to demonstrate the great potential of combining measurements of two integrating variables listed in Table 1. The combination of measurements of horizontal electric fields (HEFs) and vertical acoustic travel times (VATTs) can provide estimates of (1) volume transport per unit width, (2) the gravest vertical structure (i.e., barotropic and first baroclinic modes) of the horizontal currents, and (3) the total heat flux (using the gravest vertical structures of the currents and temperature). Because seafloor HEMs and IESs are inexpensive to make and deploy compared to current meter moorings, it is not unreasonable to envision the deployment of large arrays of HEMs and IESs for both dynamical process studies and the accumulation of transport time series for climate studies. That most of the ocean's low frequency structure and variability can thereby be observed from the seafloor using integrating variables is quite remarkable.

The VATT measured by an IES is

$$\tau = 2\int_{-H}^{0} \frac{dz}{c}, \tag{10}$$

where H is the bottom profile, and $c = c(z,t)$ is the speed of sound. Potential small errors in the interpretation of VATT in terms of the simple relation in Eq. (10) have been enumerated by Watts and Rossby (1977).

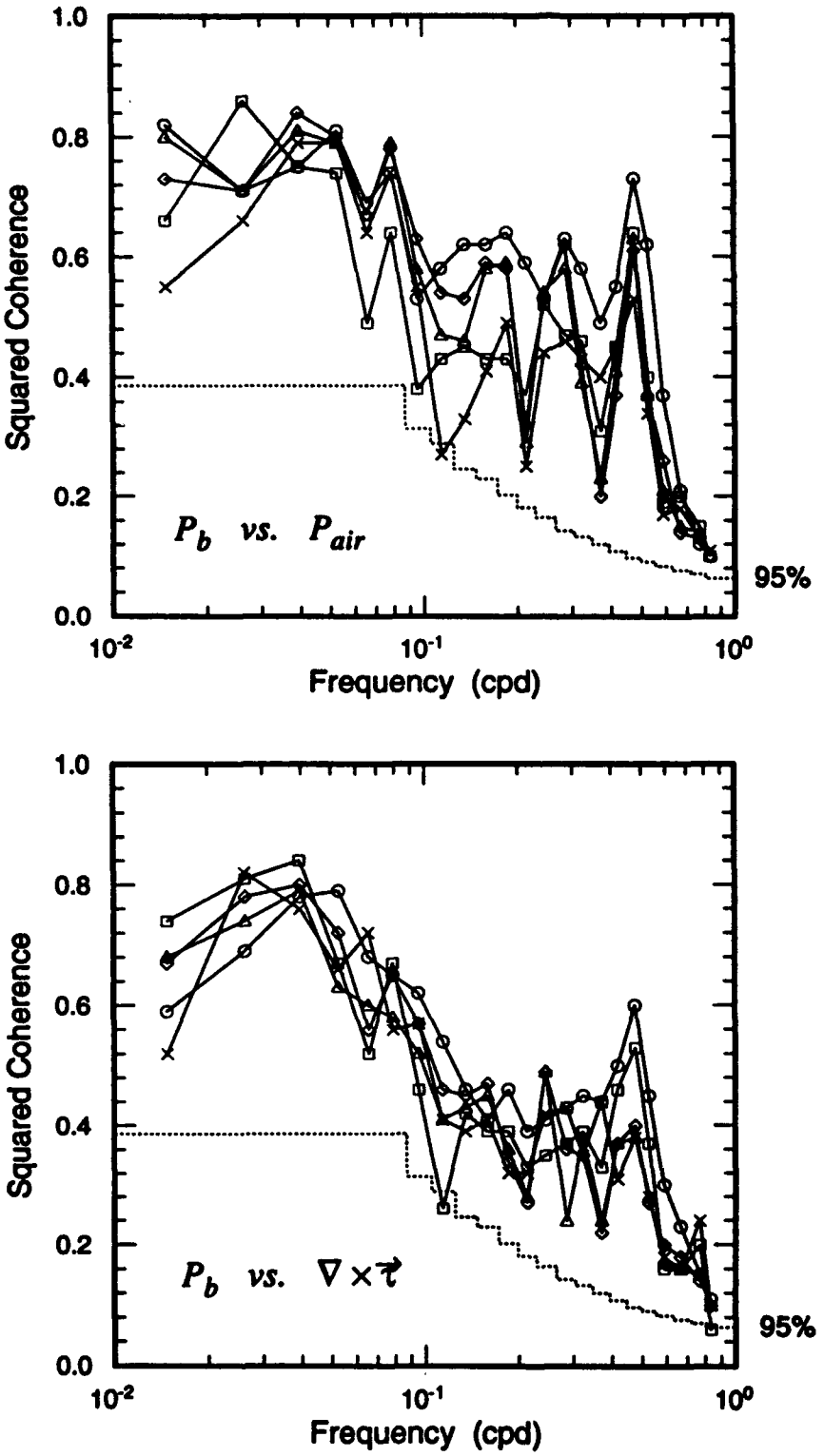


Figure 8. Maximum squared coherence amplitude, over the oceanic domain of Figure 3, between each BEMPEX bottom pressure record and (top) surface air pressure, or (bottom) surface wind stress curl, plotted as a function of frequency. The 95% level of no significance is indicated in each plot. For each station, every other point plotted is independent due to a 50% overlap of frequency band-averaging. The ubiquitously high coherence maxima indicate that bottom pressure, and hence the large-scale barotropic motions that it represents, is strongly related to atmospheric forcing in the central North Pacific.

First Baroclinic Displacement Mode Amplitude

Consider temperature, T, salinity, S, and pressure, P, as state variables, so $c(\bar{x},t) = c(T(\bar{x},t), S(\bar{x},t), P(\bar{x},t))$. Following Pickart and Watts (1990), we idealize variations in T and S as perturbations on a base profile which varies only with z (pressure is not perturbed since it is essentially the integration variable), therefore

$$T(z,t) = \overline{T}(z+\zeta(z,t)) \tag{11}$$

and similarly for S, where $|\zeta| << |z|$ by assumption. We now expand ζ in terms of displacement modes per Eq. (6), such that

$$\zeta(z,t) = \sum_{i=1}^{\infty} q_i(t)\,\theta_i(z), \tag{12}$$

where the q_i are non-dimensional since the θ_i have dimensions of length. Substituting Eq. (12) into the perturbation forms of T and S, and truncating after mode 1, the sound velocity can be written

$$c(z,t) = c\left[\overline{T}(z+q_1\theta_1), \overline{S}(z+q_1\theta_1), P(z)\right]. \tag{13}$$

After the basic state profiles are chosen, numerical evaluation of c based on its empirical dependence on T, S and P, using different values for q_1, leads to a functional relationship between τ and q_1 (Pickart and Watts, 1990), which can be inverted to yield the amplitude of the first baroclinic displacement mode for any measured VATT. In practice, since the depth is never known precisely enough, in situ profiles of T and S must be taken (by CTD or XBT) while the IES is deployed in order to calibrate the VATT. Pickart and Watts (1990) have shown evidence that the relationship between τ and q_1 in Eq. (13) is not sensitive to the choice of basic state profile of S (or buoyancy frequency, N, used in Eq. (6)), although they do note that the choice of basic state T profile is important, and a climatological mean T profile is inadequate in frontal regions such as the Gulf Stream.

The strong (weak) dependence of VATT on the first (other) baroclinic mode for mid-latitude hydrographic profiles has been documented by Watts and Rossby (1977) and Pickart and Watts (1990). (Also, Hall, 1986, and Pickart and Watts, 1990, have shown with current meter data that the first baroclinic mode dominates the vertical velocity, hence also the vertical displacement, in the Gulf Stream.) In the tropics, however, second baroclinic mode variability makes a non-trivial contribution to the VATT and cannot be ignored (Garzoli and Katz, 1981). In what follows, we are assuming the application is at mid- to high-latitudes.

First Baroclinic Current Mode Amplitude

Departing from previous authors, we develop an expression for the amplitude of the first baroclinic mode of current as follows. Under the hydrostatic and geostrophic approximations,

$$f\frac{\partial \bar{v}_h}{\partial z}=\frac{g}{\rho_*}k\times\nabla_h\rho. \tag{14}$$

Let there be small perturbations of ρ as per Eq. (11), so that

$$\frac{\partial \bar{v}_h}{\partial z}=\frac{g}{f\rho_*}\frac{d\bar{\rho}}{dz}k\times\nabla_h\zeta. \tag{15}$$

Substituting the modal expansions for $\bar{v}_h$ and ζ (see Eqs. (6a) and (12)) in Eq. (15), applying the second relation in Eq. (6c), and truncating after mode 1 yields an expression for the amplitudes of the first baroclinic current modes, viz.,

$$\bar{a}_{h,1}=\frac{\gamma_1^2}{f}k\times\nabla_h q_1, \tag{16}$$

where γ_1^2 is the first baroclinic mode eigenvalue determined from Eq. (6). Note that none of the physical assumptions leading to Eq. (16), except the modal truncation, is more severe than is typically used to estimate relative or absolute (β spiral) currents from hydrographic data or to estimate cross-Gulf Stream profiles of current (and transport, after upward extrapolations) from single moorings (e.g., Hogg, 1992).

Analysis of the combined HEF and VATT datasets from the SYNOP experiment is in its early stages, but we can show a simple preliminary comparison of two derivations of one horizontal component of $\bar{a}_{h,1}$ in Figure 9a. Rather than using observed VATTs to estimate first mode displacement from Eq. (16), we simply assumed that the difference of the measured VATTs from two IESs is proportional to the first mode current amplitude, then estimated the constant of proportionality by least squares. The result is the dotted curve in Figure 9a. The solid curve in Figure 9a is an average of the first mode current amplitudes from three moorings, two at the endpoints, and one close to, the line running between the two IESs. The agreement between the curves is certainly encouraging.

Volume Transport Per Unit Width

Our estimate of volume transport per unit width is simply $\bar{a}_{h,0}$ from Eq. (8) times the depth, H. To solve Eq. (8), we need estimates of s_1 and $\bar{a}_{h,1}$. The latter are obtained from the IESs by Eq. (16). The former are obtained by reconstructing a time-dependent conductivity profile (using IES-derived estimates of q_1 in an expression for conductivity similar to that for sound speed in Eq. (13)), which is then decomposed according to Eq. (7).

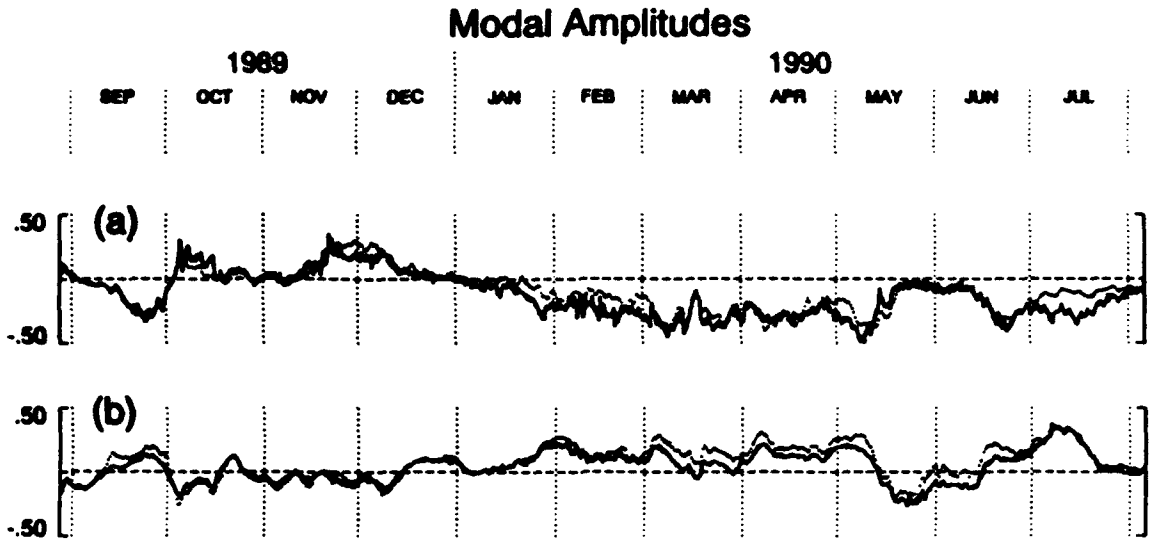


Figure 9. (a) IES (dotted curve) and mooring estimates (see text) of one component of the vector amplitude of the first baroclinic mode of horizontal current, $\vec{a}_{h,1}$. (b) HEM/IES (dotted curve) and mooring estimates (see text) of one component of the vector amplitude of the barotropic mode of horizontal current, $\vec{a}_{h,0}$. Data for both plots were taken during the SYNOP experiment in the Gulf Stream at nominally 68°W. Ordinate units are m/sec.

As in Figure 9a, a quick estimate of that component of $\vec{a}_{h,0}$ parallel to $\vec{a}_{h,1}$ shown in Figure 9a, is presented as the dotted curve in Figure 9b. The IES-derived estimate of $\vec{a}_{h,1}$ in Figure 9a was used in Eq. (8) with a climatological mean s_1. An average of the data from two HEMs (deployed near the IESs) was used in Eq. (8) as well. The only calibration employed was that for the first mode amplitude described above. The solid curve in Figure 9b is an average of barotropic mode current amplitudes from the same three moorings used in Figure 9a.

[Note that the comparison in Figure 9b is not directly relatable to the HEM-mooring comparison of transport estimates discussed previously, and evidenced by Figures 1 and 2, because Figure 9b only shows one of the two horizontal components of $\vec{a}_{h,0}$, and Figure 9b is necessarily derived from data spanning about 50 km, whereas the data for the prior comparison were all obtained at a single geographic location.]

Current Profiles and Heat Flux

The large vertical scale currents, $\vec{v}_h(z,t)$, are reconstructed by adding $\vec{a}_{h,0}$ and $\vec{a}_{h,1}\phi_1(z)$. The heat flux is readily obtained from this reconstructed current profile and a

reconstructed potential temperature profile, following Eqs. (11) and (12) truncated after mode 1.

Summary of Some Oceanic Variables Derivable from HEFs and VATTs

An ideal array would result in at least three IESs situated around each HEM. (Note that this does not mean that three times as many IESs are deployed as HEMs.) After choosing appropriate basic state temperature, $\overline{T}(z)$, and salinity, $\overline{S}(z)$, profiles, preferably from coincident CTD profiles rather than climatological mean profiles, the following are estimated:

$q_1(t)$, for each IES (see Eq. (13)) and subsequent discussion);

$\tilde{q}_1(t) = \frac{1}{m}\sum_{j=1}^{m} q_1^j(t)$, for the m IESs;

$\overline{T}_p(z)$, the basic state potential temperature profile, from the equations of state;

$T_p(z,t) = \overline{T}_p(z + \tilde{q}_1(t)\theta_1(z))$;

$\overline{\sigma}(z)$, the basic state conductivity profile, from the equations of state; and

$\sigma(z,t) = \overline{\sigma}(z + \tilde{q}_1(t)\theta_1(z))$.

Then the amplitudes of the first baroclinic modes of horizontal current are estimated from Eq. (16), viz.,

$$\vec{a}_{h,1} = \frac{\gamma_1^2}{f} k \times \nabla_h q_1,$$

where the eigenvalue γ_1^2 is obtained from solving Eq. (6) with a basic state buoyancy profile, $N^2(z)$, derived from the equations of state using $\overline{T}(z)$ and $\overline{S}(z)$. The barotropic modal amplitudes follow from Eq. 8, viz.,

$$a_{x,0} = \frac{E_{-y}^{u}(t)}{C|F_z|} - s_1 a_{x,1},$$

$$a_{y,0} = \frac{E_{x}^{v}(t)}{C|F_z|} - s_1 a_{y,1},$$

where s_1 is obtained from Eq. (7), using $\sigma(z,t)$ from above.

Finally, we arrive at estimates of the following oceanic quantities:

- Volume transport per unit width = $H\vec{a}_{h,0}$;

- Horizontal current profiles, $\bar{v}_k(z,t) = \bar{a}_{k,0} + \bar{a}_{k,1}\phi_1(z)$; and

- Un-normalized heat transport per unit width = $\int_{-H}^{0} \rho_o C_p \bar{v}_k(z,t)\overline{T}_p(z,t)dz$,

where C_p is the specific heat of seawater at constant pressure.

CONCLUSIONS

The ability to observe variables (such as those listed in Table 1) that are natural spatial integrals of water motion or state properties in the oceans provides a useful, yet underutilized, strategy for process discrimination in field experiments. For those situations when observation of an integral quantity, like volume transport, is the desired end result, integrating variables are likely to yield more accurate results than point measurements of currents or state properties, as the one example presented above indicates. Integrating variables should also be more useful than point measurements for validation of numerical models of large-scale processes, because these variables in the ocean are not "contaminated" by short-scale processes that are not simulated in the models.

Specific estimation of statistics from integrating variables, examples of which were shown previously, demonstrate that even large-scale oceanic processes with the simplest physics exhibit significant spatial inhomogeneities. Any modelling effort, observational program, or statistical analysis technique, such as some of those highlighted at this workshop, must address these inhomogeneities or risk misdirected inferences.

Acknowledgments

The seventh 'Aha Huliko'a Hawaiian Winter Workshop on Statistical Methods in Physical Oceanography was an exceptionally stimulating meeting. We thank Peter Müller and Greg Holloway for designing and hosting the workshop, the Office of Naval Research for funding it, and Phyllis Haines for making sure it ran smoothly. Programming assistance from Jeff Bytof is gratefully acknowledged. Parts of the work presented here were supported by National Science Foundation Grant #OCE-8922948, Office of Naval Research Grant #N00014-90-J-1103, and National Oceanic and Atmospheric Administration Grant #NA16RC0545-01.

REFERENCES

Anderson, D.L.T. and R.A. Corry, 1985: Ocean response to low frequency wind forcing with application to the seasonal variation in the Florida Straits-Gulf Stream transport. *Prog. Oceanogr.*,14, 7-40.

Brink, K.H., 1989: Evidence for wind-driven current fluctuations in the western North Atlantic. *J. Geophys. Res.*, 94, 2029-2044.

Brown, W., W. Munk, F. Snodgrass, H. Mofjeld and B. Zetler, 1975: MODE bottom experiment. *J. Phys. Oceanogr.*, 5, 75-85.

Chao, B.F., 1988: Excitation of the Earth's polar motion due to mass variations in major hydrological reservoirs. *J. Geophys. Res.*, 93,13811-13819.

Chave, A.D., and D.S. Luther, 1990: Low-frequency, motionally induced electromagnetic fields in the ocean, 1, theory. *J. Geophys. Res.*, 95, 7185-7200.

Chave, A.D., D.S. Luther, and J.H. Filloux, 1992a: The Barotropic Electromagnetic and Pressure Experiment, 1. Atmospherically-forced barotropic currents, *J. Geophys. Res.*, 97, 9565-9593.

Chave, A.D., D.S. Luther, L.J. Lanzerotti and L.V. Medford, 1992b: Geoelectric field measurements on a planetary scale: Oceanic and geophysical applications. *Geophys. Res. Lett.*, 19, 1411-1414.

Chave, A.D., J.H. Filloux, D.S. Luther, L.K. Law, and A. White, 1989: Observations of motional electromagnetic fields during EMSLAB. *J. Geophys. Res.*, 94, 14153-14166.

Cronin, M., 1991: How good is the mooring motion correction? Tests using the Central Array current meter data. *SYNOPtician*, 2, 5-6 & 20-23.

Cummins, P.F., 1991: The barotropic response of the subpolar North Pacific to stochastic wind forcing. *J. Geophys. Res.*, 96, 8869-8880.

Eubanks, T.M., 1993: Variations in the orientation of the earth. American Geophysical Union Monograph: *Space Geodesy and Geodynamics*, in press.

Filloux, J.H., 1987: Instrumentation and experimental methods for oceanic studies. In: *New Volumes on Geomagnetism and Geoelectricity*, J. Jacobs (ed.), Academic Press, Chapter 3, pp. 143-247.

Frankignoul, C. and P. Müller, 1979: Quasi-geostrophic response of an infinite β-plane ocean to stochastic forcing by the atmosphere. *J. Phys. Oceanogr.*, 9, 104-127.

Garzoli, S., and E.J. Katz, 1981: Observations of inertia-gravity waves in the Atlantic from inverted echo sounders during FGGE. *J. Phys. Oceanogr.*, 11, 1463-1473.

Hall, M.M., 1986: Horizontal and vertical structure of the Gulf Stream velocity field at 68°W. *J. Phys. Oceanogr.*, 16, 1814-1828.

Hide, R. and J.O. Dickey, 1991: Earth's variable rotation. *Science*, 253, 629-637.

Hogg, N.G., 1991: Mooring motion revisited. *J. Atmos. Ocean. Technol.*, 8, 289-295.

Hogg, N.G., 1992: On the transport of the Gulf Stream between Cape Hatteras and the Grand Banks. *Deep-Sea Res.*, 39, 1231-1246.

Koblinsky, C.J., P.P. Niiler and W.J. Schmitz, Jr., 1989: Observations of wind-forced deep ocean currents in the North Pacific. *J. Geophys. Res.*, 94, 10773-10790.

Larsen, J.C., 1992: Transport and heat flux of the Florida Current at 27°N derived from cross-stream voltages and profiling data: theory and observations. *Phil. Trans. R. Soc. Lond. A*, 338, 169-236.

Larsen, J.C. and T.B. Sanford, 1985: Florida Current volume transport from voltage measurements. *Science*, 227, 302-304.

Levitus, S., 1982: *Climatological Atlas of the World Ocean*, NOAA Professional Paper #13, U.S. Government Printing Office, Washington, D.C.

Luther, D.S., Chave, A.D., J.H. Filloux, and P.F. Spain, 1990: Evidence for local and nonlocal barotropic responses to atmospheric forcing during BEMPEX. *Geophys. Res. Lett.*, 17, 949-952.

Luther, D.S., J.H. Filloux, and A.D. Chave, 1991: Low-frequency, motionally induced electromagnetic fields in the ocean, 2, Electric field and Eulerian current comparison from BEMPEX. *J. Geophys. Res.*, 96, 12797-12814.

Müller, P. and C. Frankignoul, 1981: Direct atmospheric forcing of geostrophic eddies. *J. Phys. Oceanogr.*, 11, 287-308.

Munk, W.H. and A.M.G. Forbes, 1989: Global ocean warming - An acoustic measure. *J. Phys. Oceanogr.*, 19, 1765-1778.

Niiler, P.P. and C.J. Koblinsky, 1985: A local time-dependent Sverdrup balance in the eastern North Pacific Ocean. *Science*, 229, 754-756.

Owens, B., 1985: A statistical description of the vertical and horizontal structure of eddy variability on the edge of the Gulf Stream recirculation. *J. Phys. Oceanogr.*, 15, 195-205.

Pedlosky, J., 1987: *Geophysical Fluid Dynamics*, 2nd Ed., Springer-Verlag, New York, 710 pp.

Pickart, R.S. and D.R. Watts, 1990: Using the Inverted Echo Sounder to measure vertical profiles of Gulf Stream temperature and geostrophic velocity. *J. Atmos. Ocean. Tech.*, 7, 146-156.

Philander, S.G.H., 1978: Forced oceanic waves. *Rev. Geophys. Space Phys.*, 16, 15-46.

Rossby, T., 1987: On the energetics of the Gulf Stream at 73W. *J. Mar. Res.*, 45, 59-82.

Sanford, T.B., 1971: Motionally-induced electric and magnetic fields in the sea. *J. Geophys. Res.*, 76, 3476-3492.

Sanford, T.B., 1986: Recent improvements in ocean current measurement from motional electric fields and currents. *Proc. IEEE Third Working Conf. on Current Measurement*, Airlie, Virginia, January 22-24, 1986, 65-76.

Sanford, T.B., R.G. Drever, and J.H. Dunlap, 1985: An acoustic Doppler and electromagnetic velocity profiler. *J. Atmos. Ocean. Technol.*, 2, 110-124.

Stephenson, D., and K. Bryan, 1992: Large-scale electric and magnetic fields generated by the oceans. *J. Geophys. Res.*, 97, 15467-15480.

Watts, D.R. and H.T. Rossby, 1977: Measuring dynamic heights with inverted echo sounders: Results from MODE. *J. Phys. Oceanogr.*, 7, 345-358.

Whitworth, T., III, and R.G. Peterson, 1985: The volume transport of the Antarctic Circumpolar Current from bottom pressure measurements. *J. Phys. Oceanogr.*, 15, 810-816.

Worcester, P.F., 1977: Reciprocal acoustic transmission in a midocean environment. *J. Acoust. Soc. Am.*, 62, 895-905.

Worcester, P.F., B. Dushaw and B. Howe, 1991: Gyre-scale reciprocal acoustic transmission. In: *Ocean Variability and Acoustic Propagation*, J. Potter and A. Warn-Varnas, Eds., Kluwer Academic Publishers, 119-134.

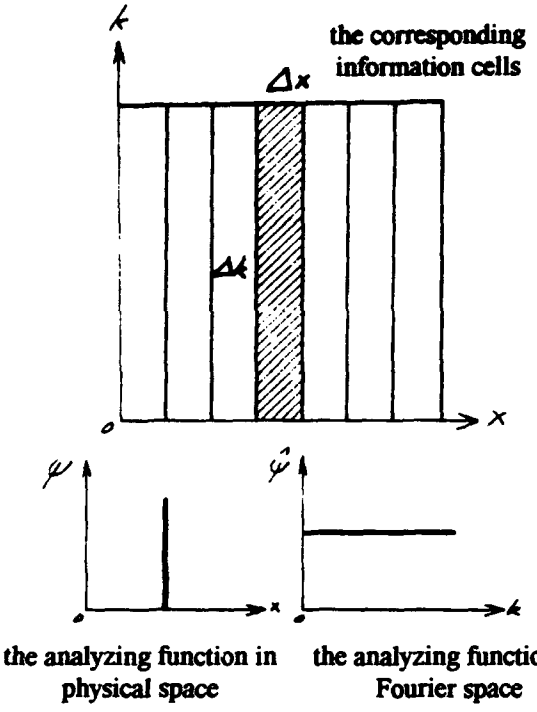
WAVELETS AND WAVELET PACKETS TO ANALYZE, FILTER, AND COMPRESS TWO-DIMENSIONAL TURBULENT FLOWS

Marie Farge, Eric Goirand, and Thierry Philipovitch
LMD-CNRS, Ecole Normale Supérieure, 24, rue Lhomond, 75231 Paris Cedex 5

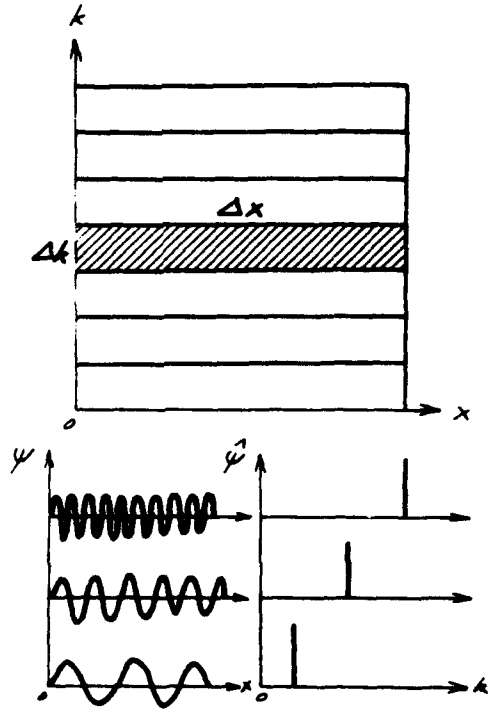
1. INTRODUCTION

The useful information in a signal is usually carried by both its frequency content and its time evolution. If we consider only the time representation, we do not know the spectrum, whereas the Fourier spectral representation does not give information on the time of occurrence of each frequency. A more appropriate representation should combine these two complementary descriptions. This is true in particular for turbulent signals, especially those presenting bursts or some intermittent, quasi-singular behaviours. The uncertainty principle precludes analysis of the signal from both sides of the Fourier transform at the same time because of the condition $\Delta t \cdot \Delta \nu \geq 1$ (normalized information cell). Therefore it is always a compromise: either good time resolution Δt but loss of spectral resolution $\Delta \nu$, which is the case when we sample a signal by convolving it with a Dirac comb (Fig. 1a), or good spectral resolution $\Delta \nu$ but loss of time resolution Δt, which is the case with the Fourier transform (Fig. 1b). These two transforms are the most commonly used in practice because they allow construction of orthogonal bases onto which the signal can be projected for analysis and eventual computation.

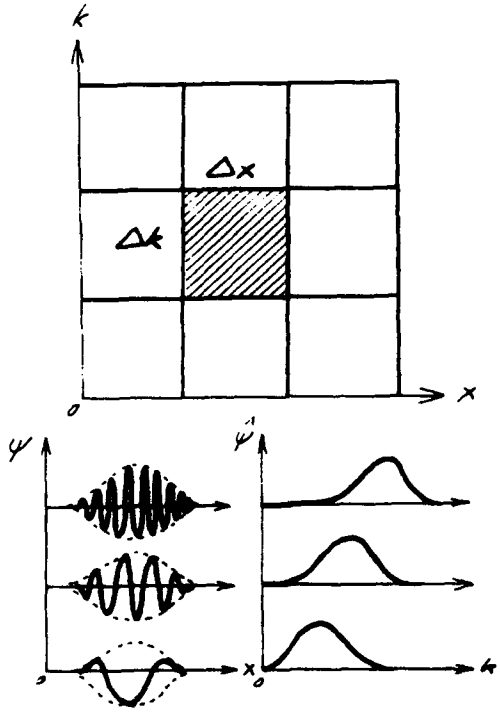
In order to improve time resolution while using the Fourier transform, Gabor (1946) has proposed the windowed Fourier transform, which consists of convolving the signal with a set of Fourier modes localized in a Gaussian envelope of constant width a_0 (Fig. 1c). This transform allows then a time-frequency decomposition of the signal at a given scale a_0, which is kept fixed. But unfortunately, as shown by Balian (1981), the bases constructed with such windowed Fourier modes cannot be orthogonal. More recently, Grossmann and Morlet (1984, 1985) have devised a new transform, the so-called wavelet transform, which consists of convolving the signal with a set of affine functions all presenting the same frequency ν_0; the family of analysing wavelets $\psi_{a,b}$ is obtained by dilation and translation of a given function ψ presenting at least one oscillation. The wavelet transform allows therefore a time-scale decomposition of the signal at a given frequency ν_0, which is kept fixed. Actually the wavelet transform realizes the best compromise of the uncertainty principle, because it adapts the time-frequency resolution $\Delta t \cdot \Delta \nu$ to each scale a. In fact it gives a good spectral resolution $\Delta \nu$ with a limited time resolution Δt in the large scales, but also gives a good time localization Δt with a limited spectral resolution $\Delta \nu$ in the small scales (Fig. 1d). The continuous wavelet transform has been extended to n dimensions by Murenzi (1989).



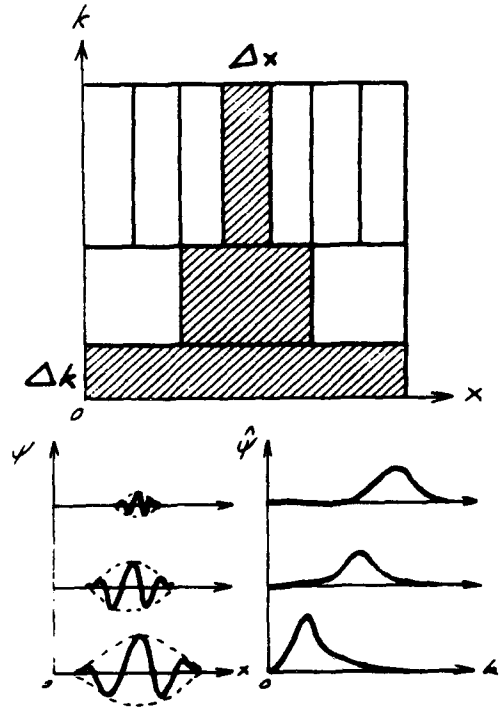
a. Sampling by a Dirac comb



b. Fourier transform



c. Windowed Fourier transform



d. Wavelet transform

Figure 1. Comparison between different types of transforms.

In 1985 Meyer, while trying to prove the impossibility of constructing orthogonal bases, as Balian had earlier done for the case of the windowed Fourier transform, was surprised to discover an orthogonal wavelet basis built with spline functions, now called the Meyer-Lemarié wavelets (Lemarié and Meyer, 1986). In fact the Haar orthogonal basis, which had been proposed in 1909, is now recognized as the first orthogonal wavelet basis known, but the functions it uses are not regular, which drastically limits its application. In practice one likes to build orthogonal wavelet bases using functions having a prescribed regularity to provide enough spectral decay depending on the application. In particular, following Meyer's work, Daubechies (1988) has proposed new orthogonal wavelet bases built with compactly supported functions of prescribed regularity defined by discrete quadratic mirror filters (QMF) of different lengths, the longer the filter, the more regular the associated functions. Mallat (1989) has devised a fast algorithm to compute the orthogonal wavelet transform using wavelets defined by QMF; it has been used in particular to devise more efficient techniques for numerical analysis (Beylkin, Coifman, and Rokhlin, 1992). Then, more recently, Malvar (1990), Coifman and Meyer (1991) found a new kind of window of variable width, which allows the construction of orthogonal adaptative local cosine bases. The elementary functions of such bases are then parametrized by their position b, their scale a (width of the window), and their wavenumber k (proportional to the number of oscillations inside each window). In the same spirit, Coifman et al. (1990), Wickerhauser (1990), and Coifman, Meyer, and Wickerhauser (1992) have proposed the so called wavelet packets which, similarly to compactly supported wavelets, are wavepackets of prescribed regularity defined by discrete QMF, from which one can construct orthogonal bases. A review of the different types of wavelet transforms and their applications to analysis and computation of turbulent flows in 2D and 3D is given in Farge (1992a,b).

2. THE CONTINUOUS WAVELET TRANSFORM

The only condition a function $\psi(x) \in \mathbb{L}^2(\Re)$, real or complex-valued, should satisfy to be called a wavelet is the admissibility condition:

$$C(\hat{\psi}) = 2\pi \int_{-\infty}^{\infty} |\psi(k)|^2 \frac{dk}{|k|} < \infty \tag{1}$$

with

$$\hat{\psi}(k) = \int_{-\infty}^{\infty} f(x) e^{-ikx} dx. \tag{2}$$

If ψ is integrable, this condition implies that the wavelet has a zero mean :

$$\int_{-\infty}^{\infty} \psi(x)\, dx = 0 \text{ or } \hat{\psi} = 0. \tag{3}$$

In practice one also wishes the wavelet to be as localized as possible on both sides in Fourier transform, namely that

$$|\psi(x)| < \frac{1}{1+|x|^n}, \tag{4}$$

and

$$|\hat{\psi}(k)| < \frac{1}{1+|k-k_0|^n}, \tag{5}$$

with k_0 being the frequency of the wavelet and n as large as possible.

Figure 2 shows examples of the most commonly used wavelets: the Marr wavelet (Fig. 2a), also called the Mexican hat, a real-valued function used for the isotropic continuous wavelet transform, the Morlet wavelet (Fig. 2b), a complex-valued function used for the non-isotropic continuous wavelet transform, the Meyer-Lemarié wavelet (Fig. 2c), and the Daubechies wavelet (Figs. 2d,2e), real-valued functions used to build orthogonal bases.

For several applications, in particular to study fractals, one also wishes the wavelet to have a good regularity, namely that $\hat{\psi}(k)$ decays rapidly near zero or, equivalently, that the wavelet has enough cancellations such as

$$\int_{-\infty}^{\infty} \psi(x)\, x^n dx = 0 \tag{6}$$

with n as large as possible.

Then, after having chosen the so-called 'mother wavelet' ψ, one generates the family of wavelets $\Psi_{b,a}(x)$, by continuously translating the 'mother wavelet' ψ along the signal b and continuously dilating it to all accessible scales a, which gives

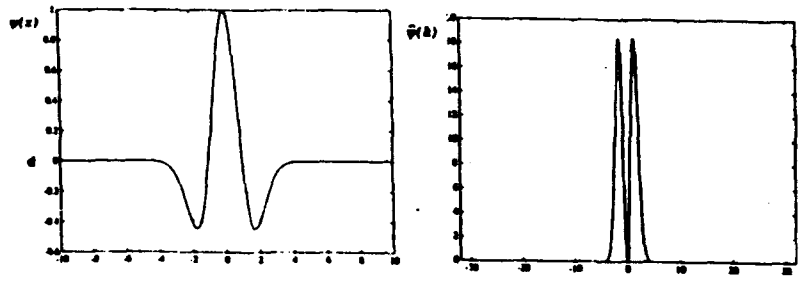
$$\Psi_{b,a}(x) = \frac{1}{N(a)} \psi\left(\frac{x-b}{a}\right) \tag{7}$$

with $N(a)$ a normalization coefficient equal, either to $a^{1/2}$ if one wishes the squared modulus of the wavelet coefficients to correspond to an energy density (L^2 norm), or to a if one uses the wavelet coefficients to analyze the local regularity of the signal (L^1 norm).

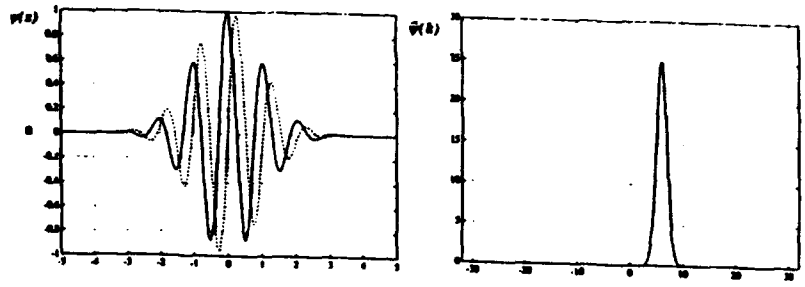
The continuous wavelet analysis of the function $f(x) \in L^2(\Re)$ is then the inner product between $f(x)$ and the set of all translated and dilated wavelets $\Psi_{b,a}(x)$, such as

$$\tilde{f}(b,a) = \int_{-\infty}^{\infty} f(x) \Psi^*_{b,a}\, dx, \tag{8}$$

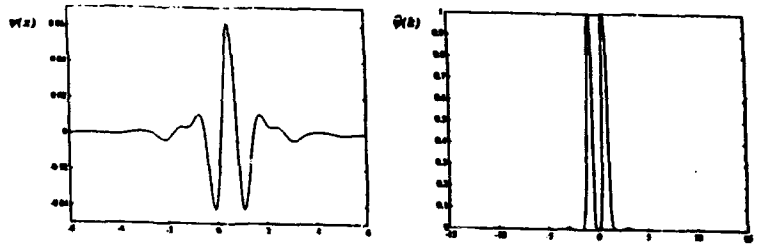
where * indicates the complex conjugate.



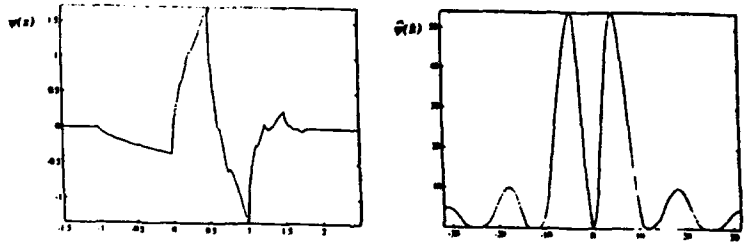
a. Marr wavelet (Mexican hat)



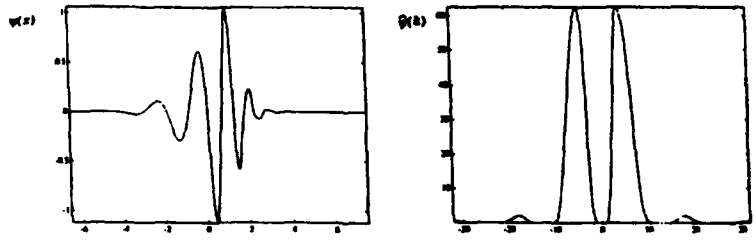
b. Morlet wavelet



c. Meyer-Lemarié wavelet



d. Daubechies wavelet (cancellation n=2)



e. Daubechies wavelet (cancellation n=7)

Figure 2. Most commonly used wavelets: (left) the functions and (right) their Fourier transforms.

The wavelet transform therefore projects the $L^2(\Re)$ space of finite energy functions into the $L^2(\Re\times\Re^+)$ space of wavelet coefficients having a measure $da\,db/a^2$, which is the Haar measure associated to the affine group. Figure 3 shows five examples of wavelet analysis of academic signals: a Dirac spike (Fig. 3a), the superposition of two cosine functions having different frequencies (Fig. 3b), the superposition of two cosine functions of very different amplitudes (Fig. 3c), a tchirp (Fig. 3d), a Gaussian white noise (Fig. 3e), and finally a tchirp in the presence of a strong noise (Fig. 3f).

From the wavelet coefficients $\tilde{f}(b,a)$, one is able to reconstruct the function $f(x)$ using the inverse wavelet transform, defined as

$$f(x)=\frac{1}{C(\hat{\psi})}\int_{-\infty}^{\infty}\int_{0}^{\infty}\tilde{f}(b,a)\Psi_{b,a}(x)\frac{da\,db}{a^2} \tag{9}$$

with

$$C(\hat{\psi})=2\pi\int_{-\infty}^{\infty}\left|\hat{\psi}(k)\right|^2\frac{dk}{|\vec{k}|},$$

a finite valued coefficient given by the admissibility condition (1).

One verifies that the wavelet transform conserves energy (as the Plancherel identity for the Fourier transform), namely that

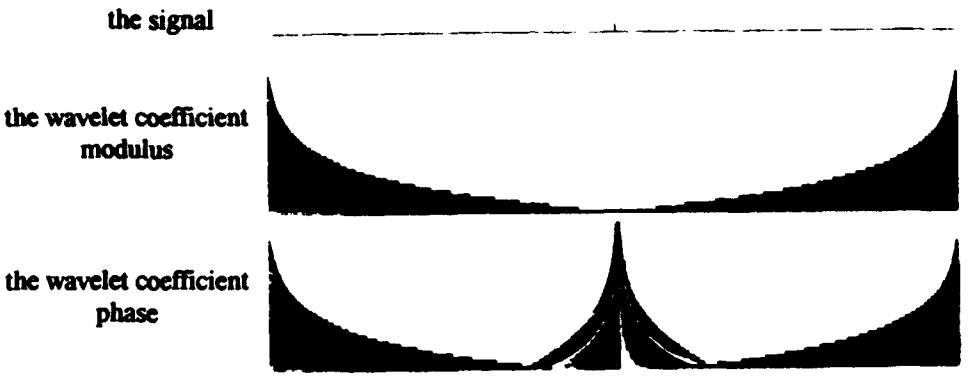
$$\int_{-\infty}^{\infty}|f(x)|^2dx=\frac{1}{C(\hat{\psi})}\int_{-\infty}^{\infty}\int_{0}^{\infty}\left|\tilde{f}(b,a)\right|^2\frac{da\,db}{a^2}. \tag{10}$$

If the function $f(x)$ belongs to the functional space $L^2(\Re)$, and if the wavelet is regular enough and therefore well localized in Fourier space (5), the wavelet analysis may be interpreted as a pass-band filter with dk/k being constant (Fig. 1d) :

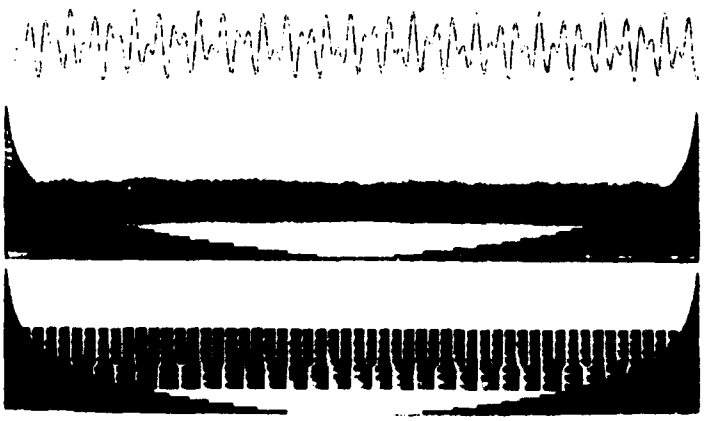
$$\tilde{f}(b,a)=\frac{1}{2\pi N(a)}\int_{-\infty}^{\infty}\hat{f}(k)\hat{\psi}^*(ak)e^{ibk}dk. \tag{11}$$

The extension of the continuous wavelet transform to analyze signals in n dimensions has been done by Murenzi (1989), considering in this case the Euclidean group with dilations. The generation of the wavelet family $\Psi_{a,r,\mathbf{b}}(x)$ is obtained by translation (vector $\mathbf{b}$), dilation (parameter a) and rotation (corresponding to the operator r defined in $\Re^n$), such as

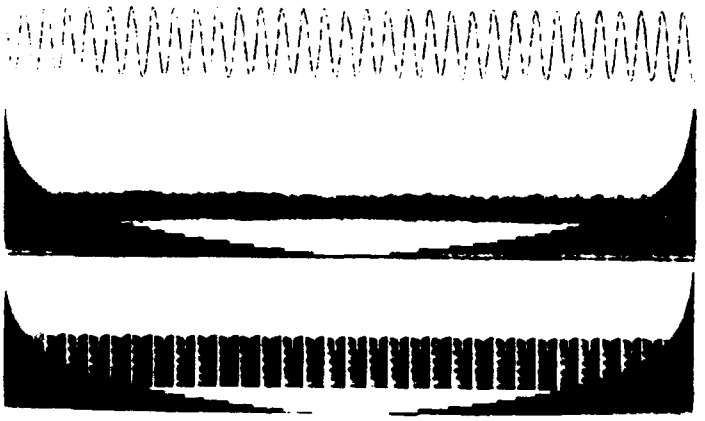
$$\Psi_{a,r,\vec{b}}(\vec{x})=N(a)^{-n}\psi(a^{-1}r^{-1}(\vec{x}-\vec{b})). \tag{12}$$



a. Dirac spike

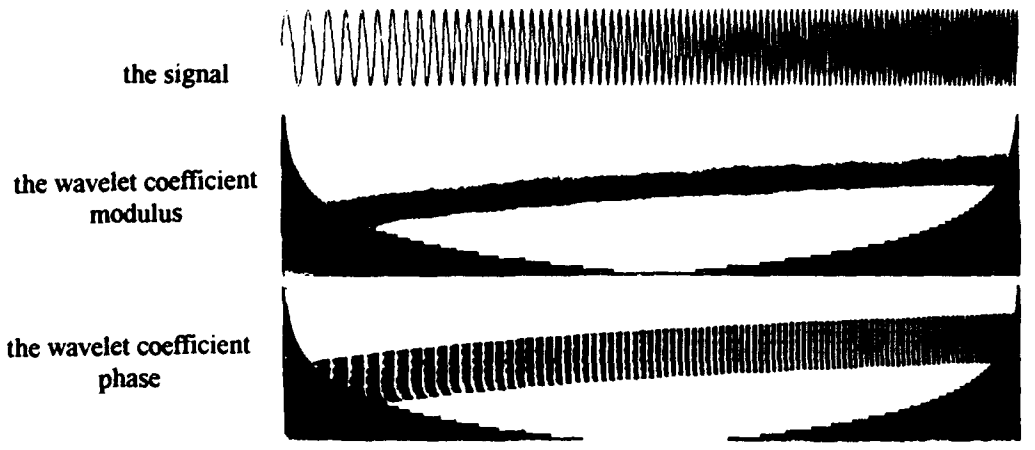


b. The superposition of two cosine functions having different frequencies
$\cos(t) + \cos(1.68t)$

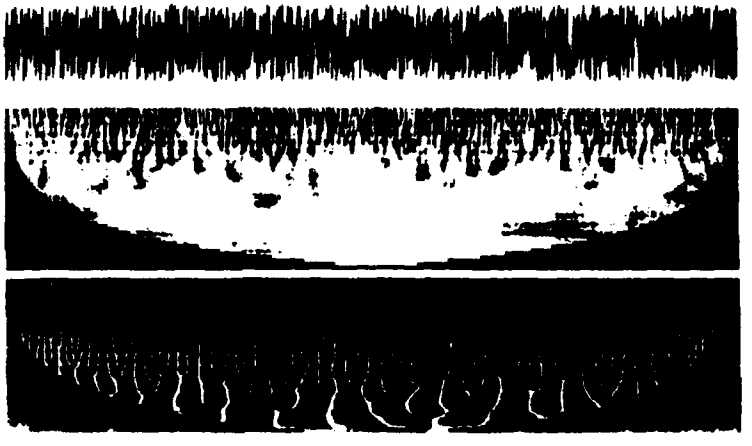


c. The superposition of two cosine functions of very different amplitudes
$\cos(t) + 0.02\cos(2t)$

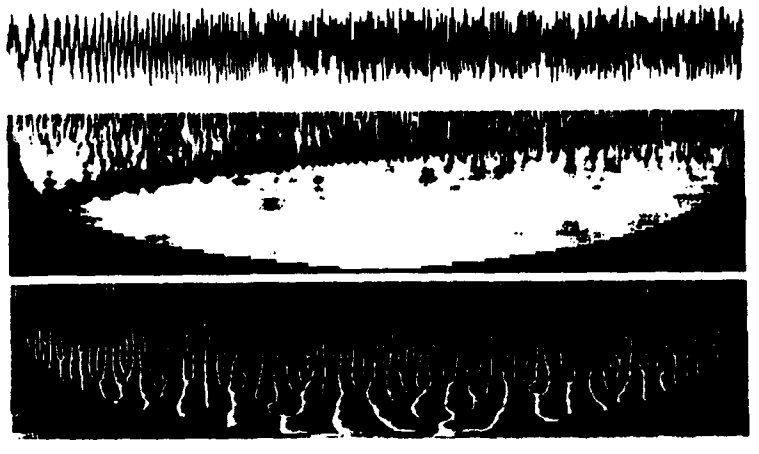
Figure 3. Wavelet transforms of several academic signals using a Morlet wavelet (continued next page).



d. A tchirp $\sin(t^2)$



e. Gaussian white noise



f. A tchirp $\sin(t^2)$ in presence of a strong noise

Figure 3. (continued) Wavelet transforms of several academic signals using a Morlet wavelet. [We have used the code TecLet 1D (copyright Science & Tec.)]

For $\Re^2$, r is the rotation matrix:

$$r = \begin{vmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{vmatrix} \tag{13}$$

with θ the rotation angle.

In n dimensions the admissibility condition becomes

$$C(\hat{\psi}) = (2\pi)^n \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} |\psi(\vec{k})|^2 \frac{d^n\vec{k}}{|\vec{k}|^n} < \infty. \tag{14}$$

The analysis and synthesis are then

$$\tilde{f}(a,r,\vec{b}) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(\vec{x}) \Psi^*_{a,r,\vec{b}}(\vec{x}) d^n\vec{x} \tag{15}$$

$$f(\vec{x}) = \frac{1}{C(\hat{\psi})} \int_0^{\infty} \int_0^{2\pi} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \tilde{f}(a,r,\vec{b}) \Psi_{a,r,\vec{b}}(\vec{x}) \frac{da\,dr\,d^n\vec{b}}{a^{n+1}}. \tag{16}$$

The energy conservation still holds:

$$\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} |f(\vec{x})|^2 d^n\vec{x} = \frac{1}{C(\hat{\psi})} \int_0^{\infty} \int_0^{2\pi} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} |\tilde{f}(a,r,\vec{b})|^2 \frac{da\,dr\,d^n\vec{b}}{a^{n+1}} \tag{17}$$

Holschneider (1988) has shown that one can reconstruct the function $f(x)$ from its wavelet coefficients $\tilde{f}(b,a)$ by using any other function $\phi(x)$, which verifies a modified admissibility condition such as

$$\int_{-\infty}^{\infty} \hat{\psi}(k) \hat{\phi}^*(k) \frac{dk}{|k|} < \infty. \tag{18}$$

This, for instance, allows us to reconstruct $f(x)$ by a simple summation of all wavelet coefficients along the verticals b = constant. This in fact corresponds to using a Dirac function as the function $\phi(x)$ to reconstruct the signal, which gives

$$f(x) = \frac{1}{C(\hat{\psi})} \int_{-\infty}^{\infty} \tilde{f}(x,a) \frac{da}{a} \tag{19}$$

with

$$C(\hat{\psi}) = \sqrt{2\pi} \int_{-\infty}^{\infty} \hat{\psi}(k) \frac{dk}{|\vec{k}|} < \infty.$$

3. PROPERTIES OF THE CONTINUOUS WAVELET TRANSFORM

3.1 Covariance by Translation and Dilation

One property of the continuous wavelet transform, which is lost in the case of the orthogonal wavelet transform, is its covariance, by both translation, i.e., shift by x_0

$$W[f(x-x_0)] = \tilde{f}(b-x_0,a) \tag{20}$$

with W the continuous wavelet transform operator, and dilation, i.e., under scale changes by a factor λ

$$W\left[f(\frac{x}{\lambda})\right] = \frac{1}{\lambda}\tilde{f}\left(\frac{b-b_0}{\lambda},\frac{a}{\lambda}\right) \tag{21}$$

3.2 Linearity

The continuous wavelet transform is a linear transform; therefore we have the following superposition principle:

$$W[\alpha f_1(x)+\beta f_2(x)] = \alpha\tilde{f}_1(b,a)+\beta\tilde{f}_2(b,a) \tag{22}$$

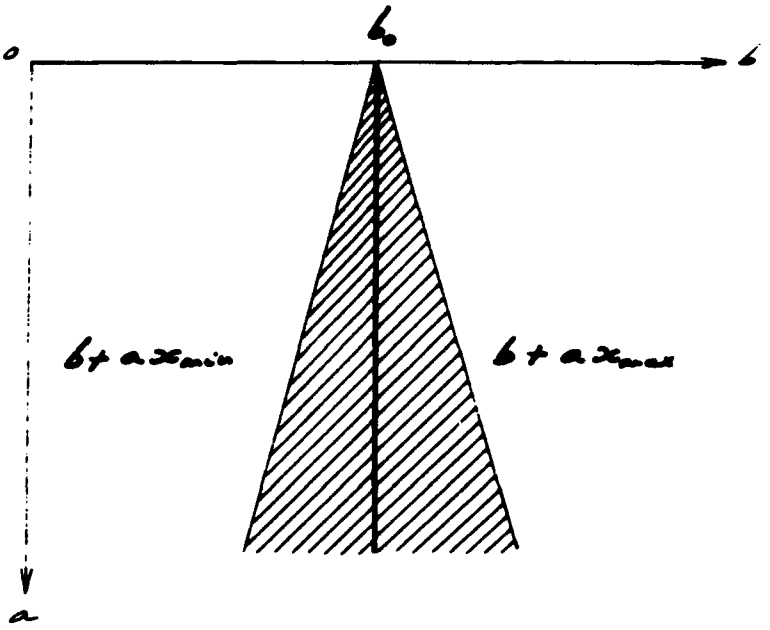
with a and b two arbitrary constants.

3.3 Locality in Both Space and Scale

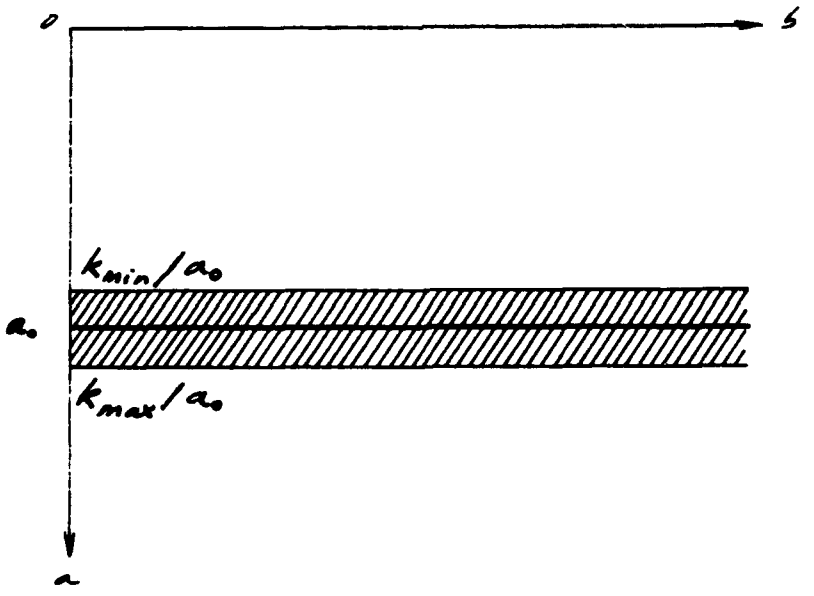
The localization of wavelets by both position b and scale a yields both values from the wavelet coefficients. This is not the case with the Fourier coefficients because the basis functions are nonlocal: a given Fourier coefficient therefore depends on the behaviour of the whole signal. On the contrary a given wavelet coefficient $\tilde{f}(b_0,a_0)$ does not depend on the value of the signal outside the so called 'influence cone' localized in $b_0 + \Delta b/a$, with Δb depending on the support of the wavelet (Fig. 4a). Likewise the wavelet coefficients at a given scale a_0 depend only on the spectral behaviour of the signal in the bandwidth $[k_{min}/a_0, k_{max}/a_0]$ with k_{min} and k_{max} given by the support of $\hat{\psi}$ (Fig. 4b). The support of $\hat{\psi}$ is defined as the region where ψ is larger than a given value, because wavelet ψ has at least an exponential decay.

3.4 Characterization of the Local Regularity of a Function

One of the most useful properties of the wavelet transform in analyzing turbulent flows is the fact that the local scaling of the wavelet coefficients computed in L^1 norm, i.e., with the normalization $N(a)=a$ in (7), allows us to characterize the regularity of the signal



a. the influence cone attached to x_0



b. the spectral band attached to wavenumber k_0

Figure 4. Locality in wavelet coefficient space.

(Holschneider 1988) and (Jaffard 1989). Thus, if $d^m f / dx^m$ exists, i.e., if f is m times continuously differentiable in x_0, then

$$\left\| \tilde{f}(x_0, a) \right\|_{L^1} \sim a^m \tag{23}$$

when a tends to 0.

If $f \in \Lambda^\alpha(x_0)$, the space of Lipschitz functions having exponent $-1<\alpha<1$, which are continuous functions non differentiable in x_0, such that

$$f(x) - f(x_0) \leq C|x - x_0|^\alpha \tag{24}$$

with constant $C>0$. Then

$$\tilde{f}(x_0, a) \sim a^\alpha \tag{25}$$

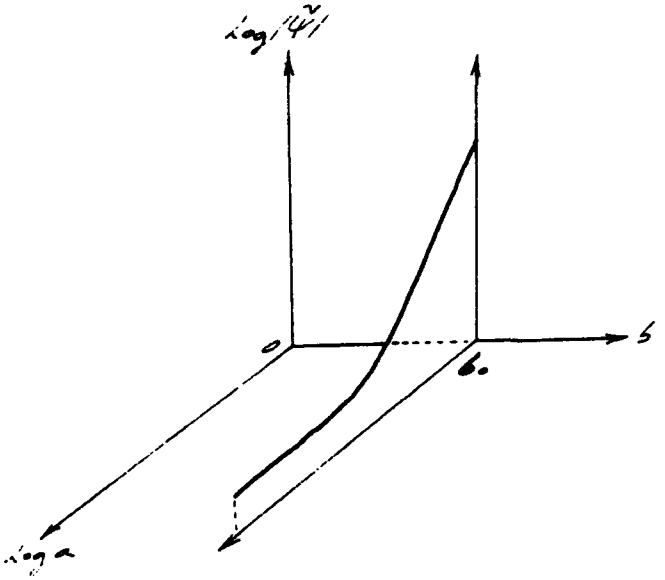
when a tends to 0.

Thus the behaviour of the wavelet coefficients $\tilde{f}(x_0, a)$ at x_0 in the limit $a \to 0$ measures the local regularity of the function f in x_0, which is given by the slope of the modulus of (x_0, a) represented in log-log coordinates. For instance, the wavelet coefficients computed in norm L^1 of a function presenting a Lipschitz singularity a in x_0 will diverge in the very small scale limit (Fig. 5a), while those of a function which is regular in x_0 will tend to zero in the same limit (Fig. 5b).

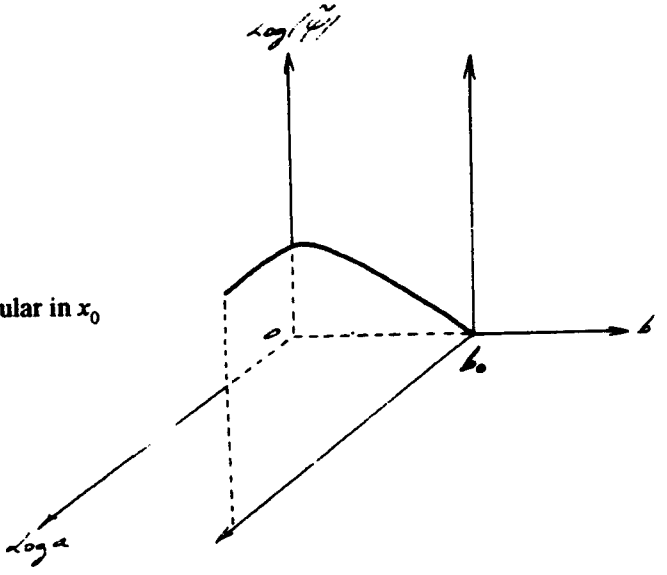
4. ANALYSIS OF TWO-DIMENSIONAL TURBULENT FLOWS

"In the last decade we have experienced a conceptual shift in our view of turbulence. For flows with strong velocity shear... or other organizing characteristics, many now feel that the spectral description has inhibited fundamental progress. The next "El Dorado" lies in the mathematical understanding of coherent structures in weakly dissipative fluids: the formation, evolution and interaction of metastable vortex-like solutions of nonlinear partial differential equations..." Norman Zabusky (1984).

As Norman Zabusky stated, it is essential before modelling turbulent flows to understand the dynamical role of coherent structures and analyze their contribution to the different nonlinear interactions. Because the Fourier modes contain nonlocal information, we are unable to discriminate the role of coherent structures and we cannot separate the coherent structures from the rest of the flow. However, this local spectral analysis becomes possible



a. f is a function presenting a Lipschitz singularity α in x_0



b. f is a function that is regular in x_0

Figure 5. Analysis of the local regularity of a function f in x_0 (given by the slope of the modulus of $\tilde{f}(x_0, a)$ represented in log-log coordinates).

when using the wavelet transform and with it we can devise new types of diagnostics. After defining them, we will apply them to analyze some vorticity fields corresponding to long-time evolution of a forced two-dimensional flow, computed with a resolution 128^2.

4.1 The Wavelet Coefficients

If we denote the position as b, the scale as a, and the angle as θ, the wavelet coefficients computed in LP norm are

$$\tilde{f}(a,\theta,\vec{b}) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(\vec{x})\Psi^{*}_{a,r,\vec{b}}(\vec{x})d^2\vec{x} \tag{26}$$

with

$$\Psi_{a,r,\vec{b}}(\vec{x}) = N(a)^{-n}\,\psi(a^{-1}r^{-1}(\vec{x}-\vec{b})), \text{ and } r = \begin{vmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{vmatrix} \tag{27}$$

If $N(a) = a^{1/2}$, the wavelet coefficients are in L^2 norm and the squared wavelet coefficients correspond to the local energy density of the signal at location b, scale a and direction θ. If $N(a) = a$, the wavelet coefficients are in L^1 norm and in this case the local scaling of the wavelet coefficients gives information on the local regularity, or the Lipschitz exponent in the case of discontinuities, of the signal at location b, scale a and direction θ.

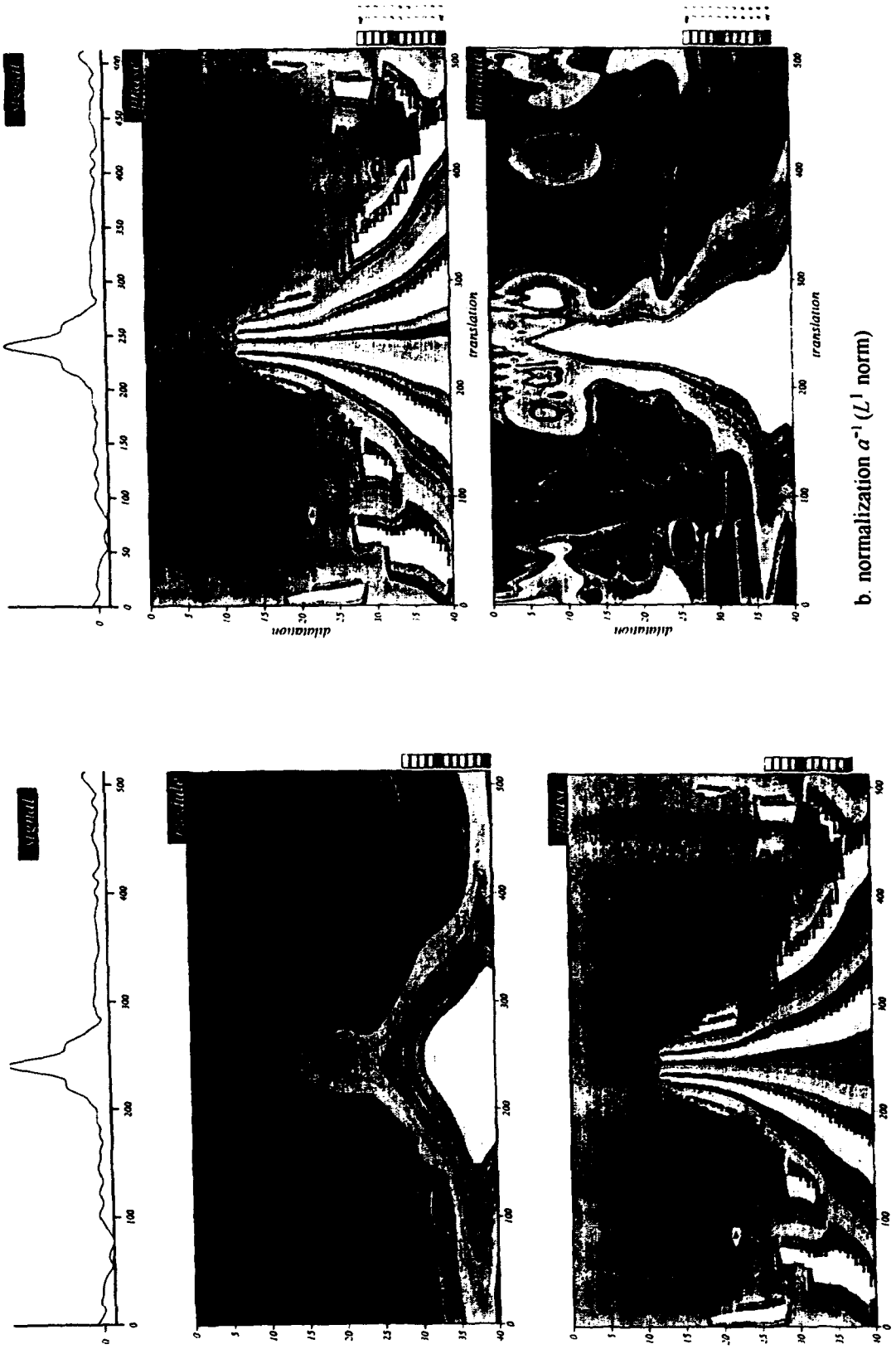
In Figure 6 we show the 1D continuous wavelet analysis along a cut done in a two-dimensional turbulent vorticity field. The wavelet coefficients are computed, either in L^2 norm (Fig. 6a), or in L^1 norm (Fig. 6b), using the Morlet wavelet with k_0=5.

In Figure 7 we show the 2D continuous wavelet analysis of a two-dimensional turbulent vorticity field. The wavelet coefficients are computed in L^2 norm at three different scales, namely 32 pixels (Fig. 7b), 16 pixels (Fig. 7c), and 2 pixels (Fig. 7d), using the isotropic Marr wavelet (in this case, there is no angular dependence of the wavelet coefficients resulting from to the wavelet isotropy).

4.2 The Intermittency Factor

The intermittency factor is given by the wavelet coefficients renormalized by the space averaged energy at each scale, such that

$$I(a,\vec{b}) = \frac{\left|\tilde{f}(a,\theta,\vec{b})\right|^2}{\int_0^{2\pi}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\left|\tilde{f}(a,\theta,\vec{b})\psi_{a,r,\vec{b}}(\vec{x})\right|^2 d\theta d^2b/a^3}. \tag{28}$$



a. normalization $a^{-1/2}$ (L^2 norm)

b. normalization a^{-1} (L^1 norm)

Figure 6. Continuous wavelet analysis of a one-dimensional cut done in a two-dimensional turbulent vorticity field.

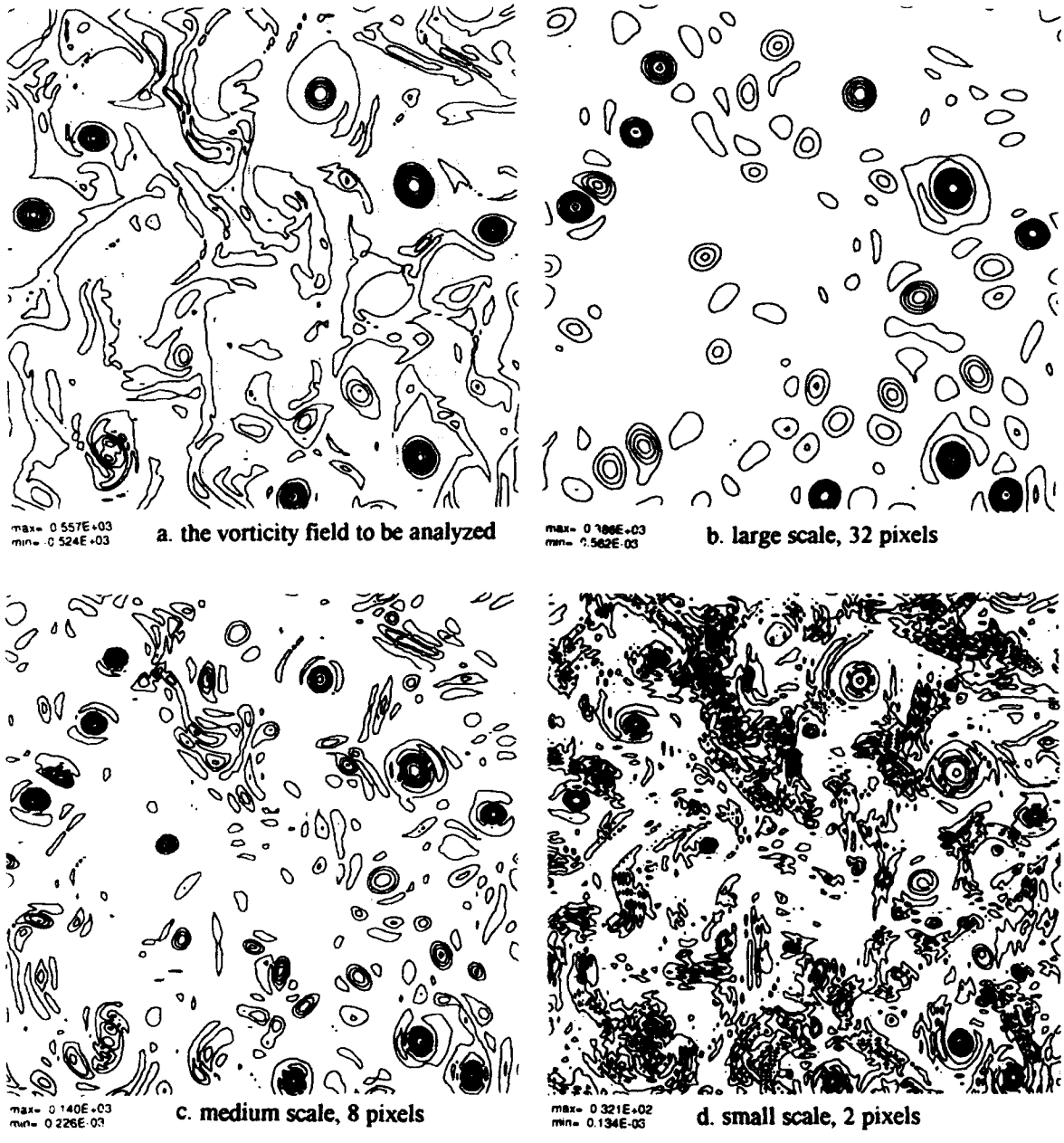


Figure 7. The wavelet coefficients in L^2 norm computed using the Marr wavelet.

It gives information on the space variance of the energy spectrum, namely if $I(a,\vec{b})=1$ the field is homogeneous and there is no space variance of the energy at scale a. If $I(a,\vec{b})$ is large, the field is intermittent, namely all the energy contribution at scale a comes from a few very excited regions, while the rest of the field has little energy at this scale.

Figure 8 shows the intermittency factor computed at three different scales, namely 32 pixels (Fig. 8b), 8 pixels (Fig. 8c), and 2 pixels (Fig. 8d) using the isotropic Marr wavelet (there is no angular dependence of the wavelet coefficients resulting from the wavelet isotropy in this case).

4.3 The Local Energy Spectrum

The local energy spectrum is defined from the wavelet coefficients, such that

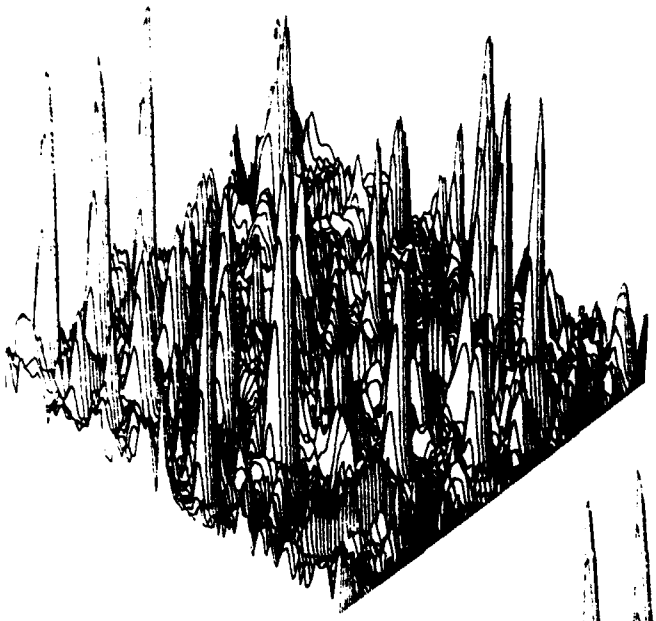
$$E(a,\vec{b}_0)=\frac{\int_0^{2\pi}\left|\tilde{f}(a,\theta,\vec{b}_0)\,d\theta\right|^2}{a^2}. \tag{29}$$

Figure 9 shows the local energy spectra (Fig. 9d) computed by integrating in space the Marr wavelet coefficients after segmenting the vorticity field (Fig. 9a) into three different regions using the Weiss criterium (Weiss 1981): the elliptical region corresponding to the cores of the coherent structures (Fig. 9b), the parabolic region corresponding to the shear layers at the periphery of the coherent structures (Fig. 9c), and the hyperbolic region corresponding to the vorticity filaments of the incoherent background flow. We observe that the elliptic region scales as k^{-6}, the parabolic region as k^{-4}, while the hyperbolic region scales as k^{-3} . Therefore the more coherent the region is, the steeper its spectrum, whereas an incoherent region, such as the background flow, is much more homogeneous and has a flatter spectrum—similar to noise.

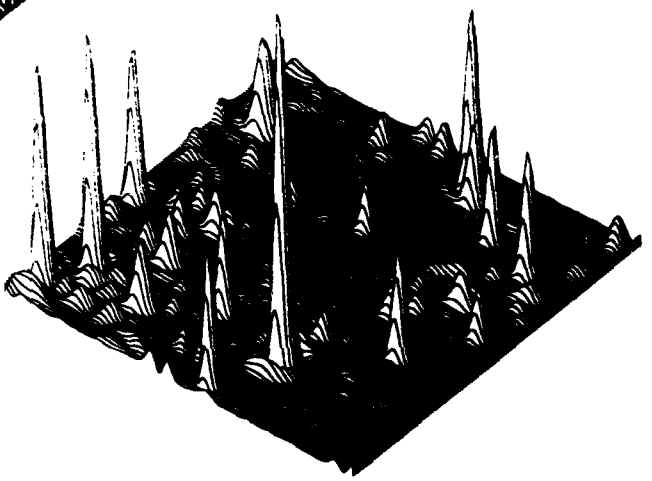
5. FILTERING OF TWO-DIMENSIONAL TURBULENT FLOWS USING CONTINUOUS WAVELETS

Because the wavelet transform is invertible it is always possible to select a subset of the coefficients and reconstruct a filtered version of the field from them. We propose several filtering techniques to extract coherent structures from the background vorticity in two-dimensional turbulent flows. The first one consists of discarding all wavelet coefficients outside the influence cones (Fig. 4a) attached to the local maxima of the vorticity field that corresponds to the coherent structures' cores. The second method consists of discarding all wavelet coefficients smaller than a given threshold that depends on the quantity of enstrophy we want to retain in the filtered vorticity field.

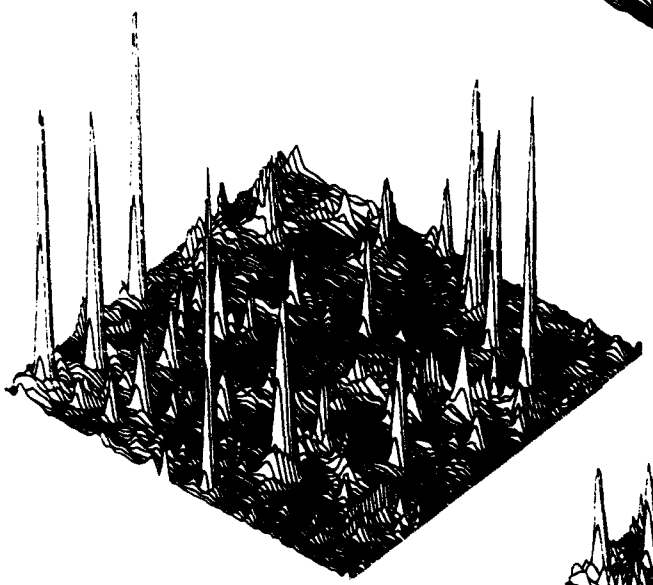
Figure 10 shows the extraction of one coherent structure, done by filtering all wavelet coefficients outside the influence cone attached to the center of this coherent structure,



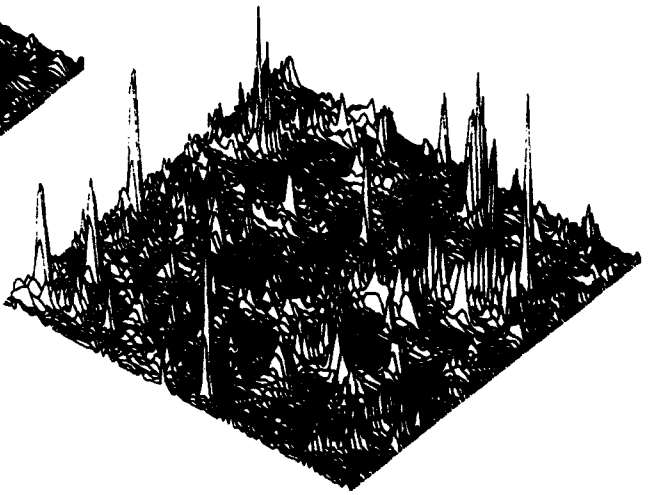
a. the vorticity field to be analyzed



b. large scale, 32 pixels (min 0, max 79)

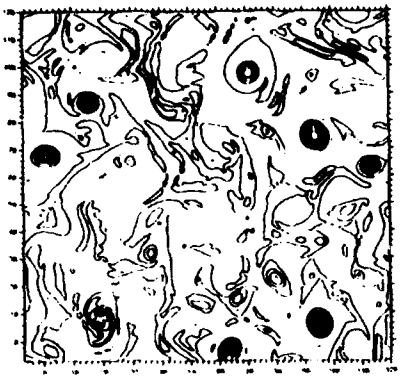


c. medium scale, 8 pixels (min 0, max 71)

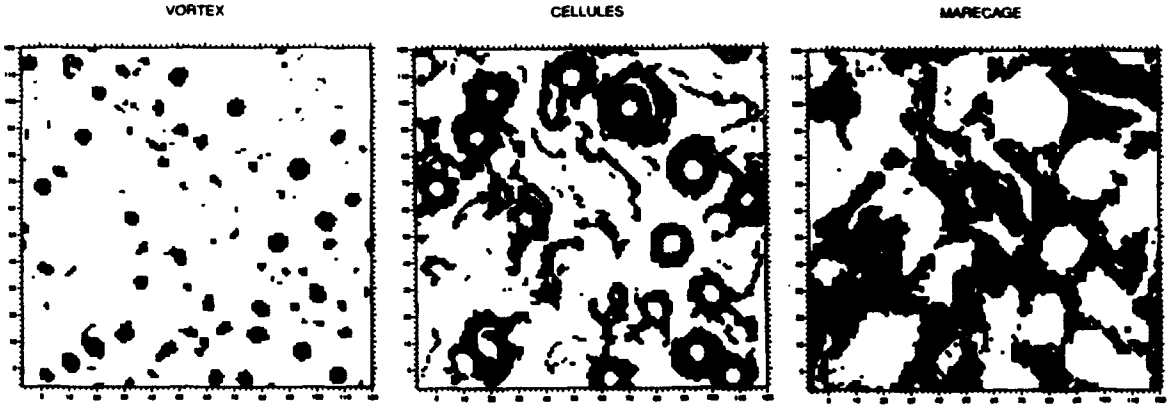


d. small scale, 2 pixels (min 0, max 44)

Figure 8. The intermittency factor computed using the Marr wavelet.



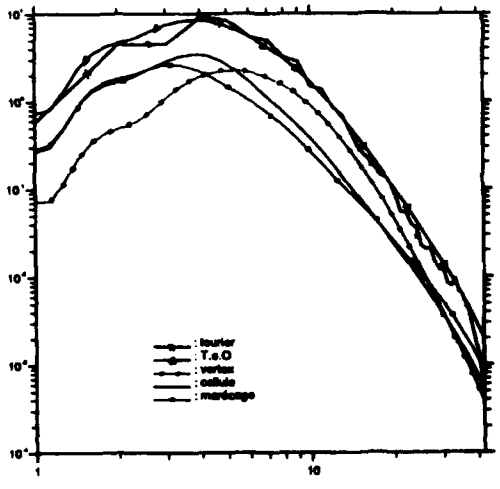
a. The complete vorticity field



b. The elliptical region corresponding to the coherent structures

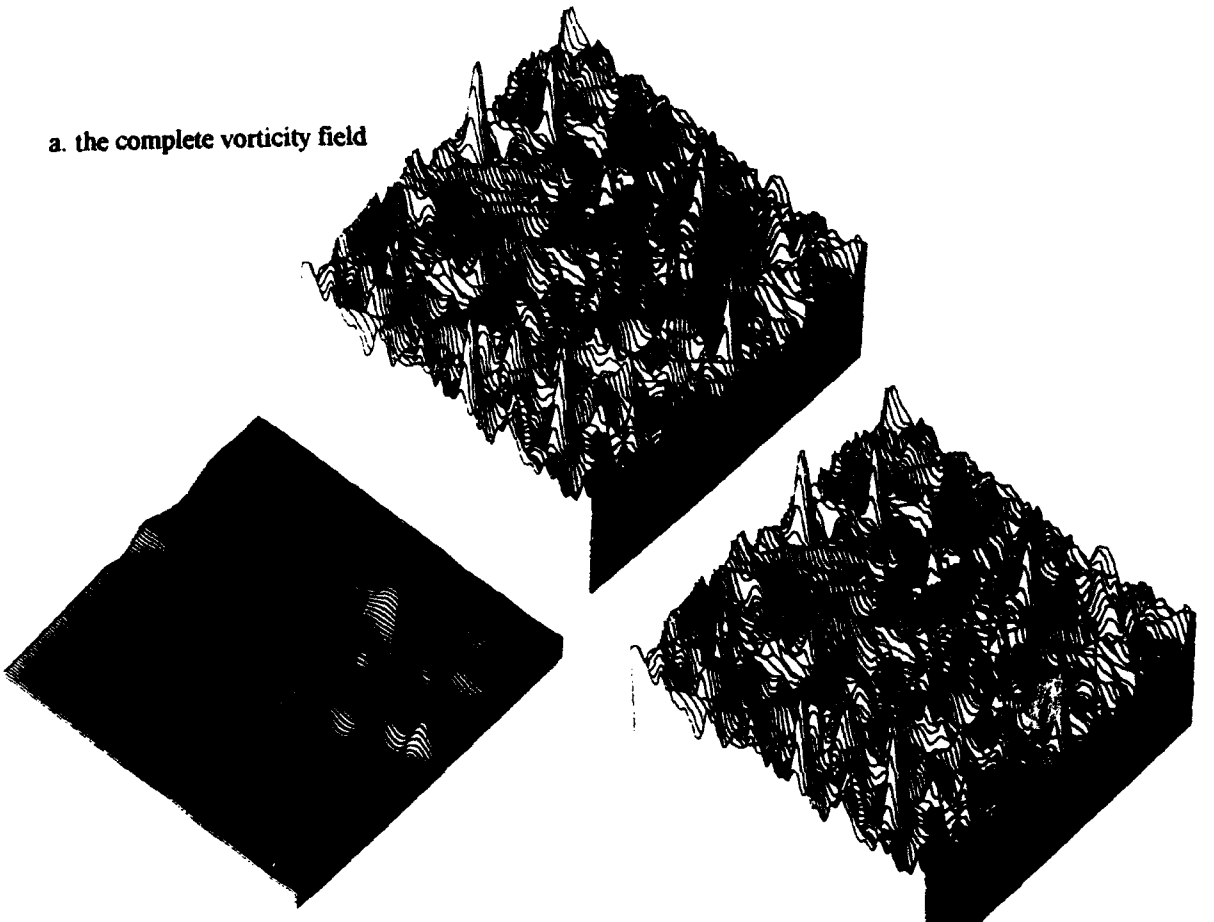
c. The parabolic region corresponding to the shear layers at the periphery of the coherent structures

d. The hyperbolic region corresponding to the vorticity filaments of the incoherent background flow



e. The corresponding energy spectra

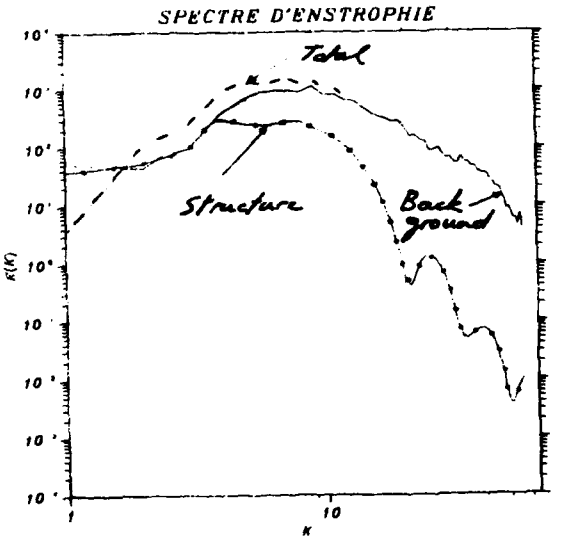
Figure 9. Local energy spectra computed from the wavelet coefficients after segmenting the vorticity field into three different regions.



a. the complete vorticity field

b. the coherent structure alone

c. the vorticity field without the coherent structure



d. the energy spectra of the three previous fields

Figure 10. Extraction of one coherent structure, done by filtering all wavelet coefficients outside the influence cone attached to the center of this coherent structure, before computing the inverse wavelet transform.

before computing the inverse wavelet transform. We display the complete vorticity field (Fig. 10a), the coherent structure alone (Fig. 10b), the vorticity field without the coherent structure (Fig. 10c), and the energy spectra of the three previous fields (Fig. 10d).

Figure 11 shows the extraction of the 40 most excited coherent structures, done by filtering all wavelet coefficients outside the influence cones attached to the centers of these coherent structures, before computing the inverse wavelet transform. We display the complete vorticity field (Fig. 11a), the 40 coherent structures alone (Fig. 11b), the vorticity field without the coherent structures (Fig. 11c), and the energy spectra of the three previous fields (Fig. 11d).

Figure 12 shows the extraction of all excited coherent structures, done by filtering all wavelet coefficients smaller than a given threshold and then computing the inverse wavelet transform. We display the complete vorticity field (Fig. 12a), the coherent structures alone (Fig. 12b), the vorticity field without the coherent structures (Fig. 12c), and the energy spectra of the three previous fields (Fig. 12d).

As seen with the local energy spectra,these filtering techniques show again that the spectral behaviour depends on the region of the flow, with a tendency to scale around k^{-6} near the cores of the coherent structures, between k^{-4} and k^{-5} at their periphery, and around k^{-3} in the background.

6. COMPRESSION OF TWO-DIMENSIONAL TURBULENT FLOWS USING WAVELET PACKETS

Wavelet packets represent a family of orthogonal bases that unifies wavelets with Dirac, Fourier and wavepacket functions, affording increased flexibility in tiling the information plane, because now each element of the basis is parametrized independently in position b, scale a and wavenumber k (cf. Coifman et al., 1992). For a given signal sampled on N points the wavelet packet algorithm generates 2^N possible orthogonal bases and then selects the one that minimizes the number of coefficients having significant contributions to the total signal. In this sense, the wavelet packet algorithm defines the most efficient basis, so called the Best Basis, upon which to expand a given signal. If the flow is dominated by point vortices, then it is optimally represented using the Dirac grid point basis, and the output of the wavelet packet algorithm will reflect this. On the contrary, if the flow is dominated by waves, then it is optimally represented using the Fourier basis, and the output of the wavelet packet algorithm will again reflect this. If the flow behaviour is in between these two extreme situations, other bases will be more appropriate and the wavelet packet algorithm will give us the Best Basis in which the vorticity field can be represented with the smallest number of significant coefficients. The computation of the Best Basis for a signal sampled on N points is done in $N.\log_2 N$ operations, while the reconstruction of the signal from its projection onto the Best Basis is done in N operations.

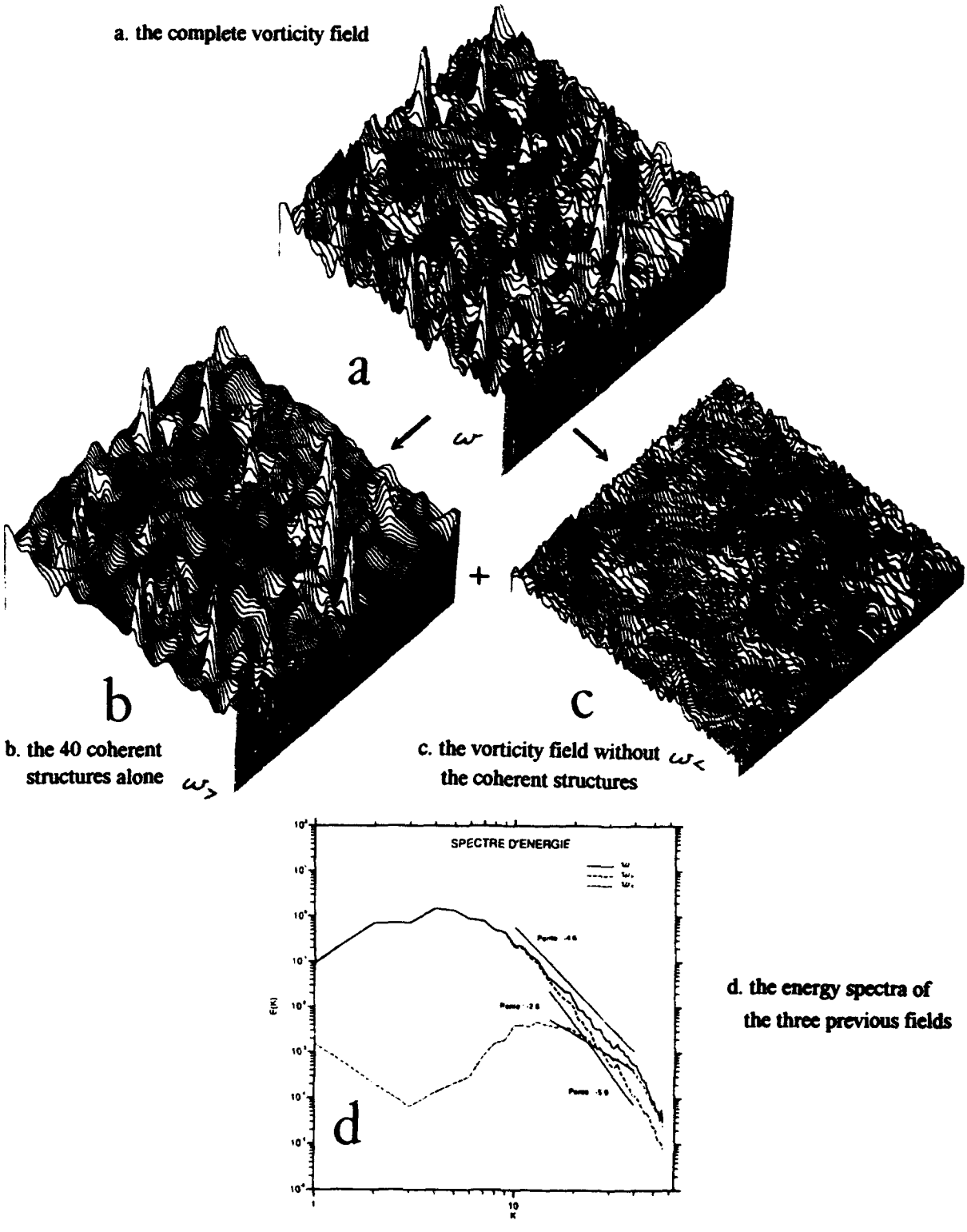


Figure 11. Extraction of the 40 most excited coherent structures, done by filtering all wavelet coefficients outside the influence cone attached to the center of these coherent structures before computing the inverse wavelet transform.

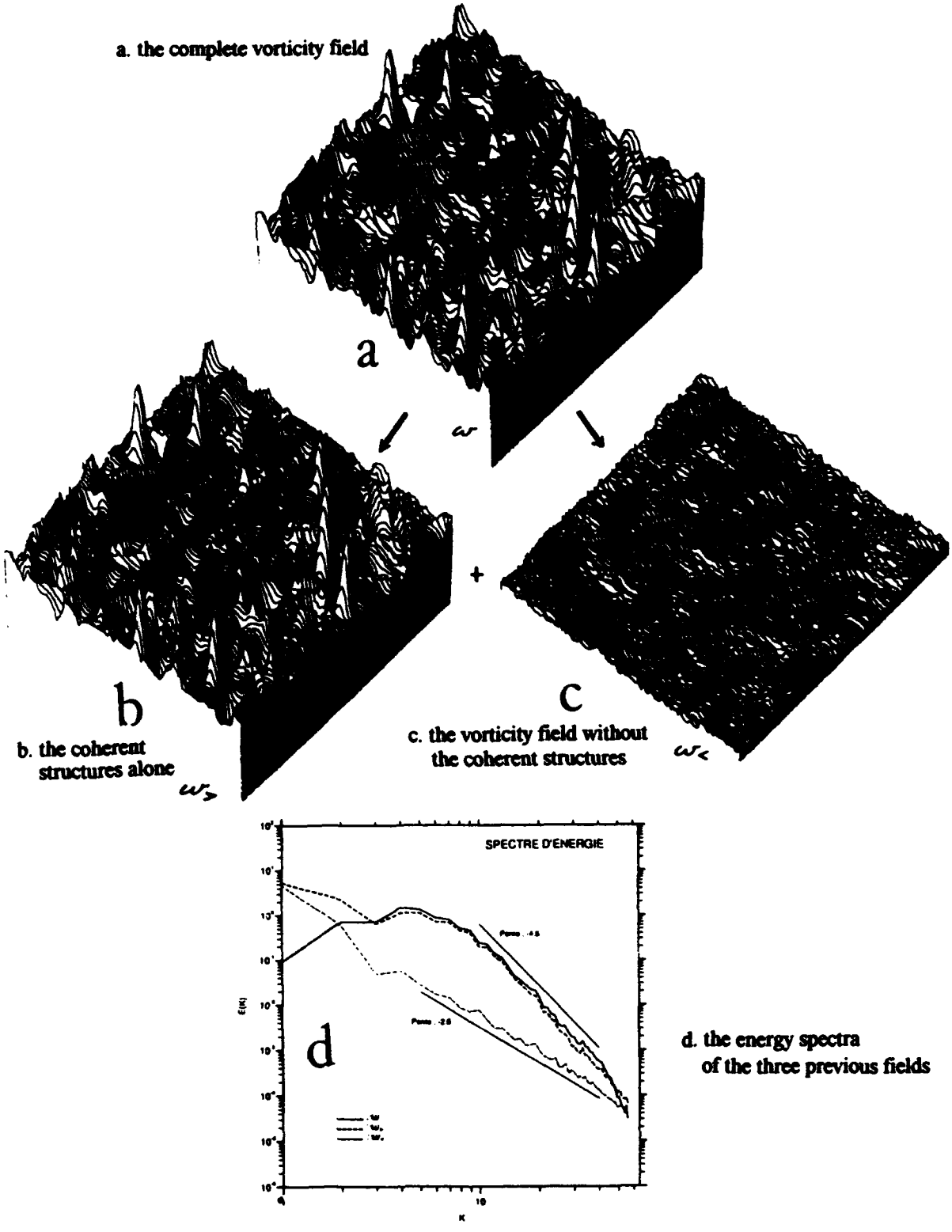


Figure 12. Extraction of all excited coherent structures, done by filtering all wavelet coefficients smaller than a given threshold and then computing the inverse wavelet transform.

Figure 13 shows the compression of a two-dimensional vorticity field using its wavelet packet coefficients with three different compression ratios. For a compression by 2 (Fig. 13a) we split the field into the 50% strongest wavelet packet coefficients and the 50% weakest wavelet packet coefficients. Then for a compression by 20 (Fig. 13b) we split the field into the 5% strongest wavelet packet coefficients and the 95% weakest wavelet packet coefficients, and for a compression by 200 (Fig. 13c) we split the field into the 0.5% strongest wavelet packet coefficients and the 99.5% weakest wavelet packet coefficients. For each of the three compression ratios we display the uncompressed field with its energy spectrum, the compressed field with its energy spectrum, and the discarded field with its energy spectrum. These results have been obtained in collaboration with Meyer, Pascal and Wickerhauser and are extensively discussed in Farge et al. (1992).

With these compression techniques we find as before that the spectral behaviour depends on the region of the flow we analyze, with a tendency to scale around k^{-6} near the cores of the coherent structures, around k^{-4} at their periphery, and around k^{-3} in the background.

7. CONCLUSION

Nowadays turbulence is commonly viewed from one of two alternative perspectives, depending upon which side of the Fourier transform one looks from. In physical space, we observe coherent vortices and wonder if there is universality in their structure and interactions. In Fourier space, we see transfers of energy and enstrophy between different scales of motion and ask, for example, if the slope of the energy spectrum is universal. The selection of bases in which turbulence may be examined must be extended if these perspectives are to be effectively reconciled. Through the use of wavelets and wavelet packets, we have constructed a class of bases, which includes grid point and Fourier representations as special cases, from which we select the basis which is optimal for a given flow, namely the one which compresses the most the information while keeping track of the behaviour of the flow in both space and scale.

With such a wavelet or wavelet packet representation we can compute a local energy spectrum. Using the continuous wavelet transform, we have shown that different regions of the flow present different slopes for the local energy spectrum. Clearly the Fourier transform is unable to detect these different spectral behaviours which vary in space, while the wavelet transform is here the appropriate tool. Typically we have observed that the cores of the coherent structures, which correspond to the elliptic regions, scale as k^{-6}, the shear layers around the coherent structures, which correspond to the parabolic regions, scale as k^{-4}, while the vorticity filaments in the background, which correspond to the hyperbolic regions, scale as k^{-3}. From this result we infer that the variation of the Fourier spectral slope we commonly observe for two-dimensional flows may be related to the density of coherent structures which varies depending on the initial conditions and

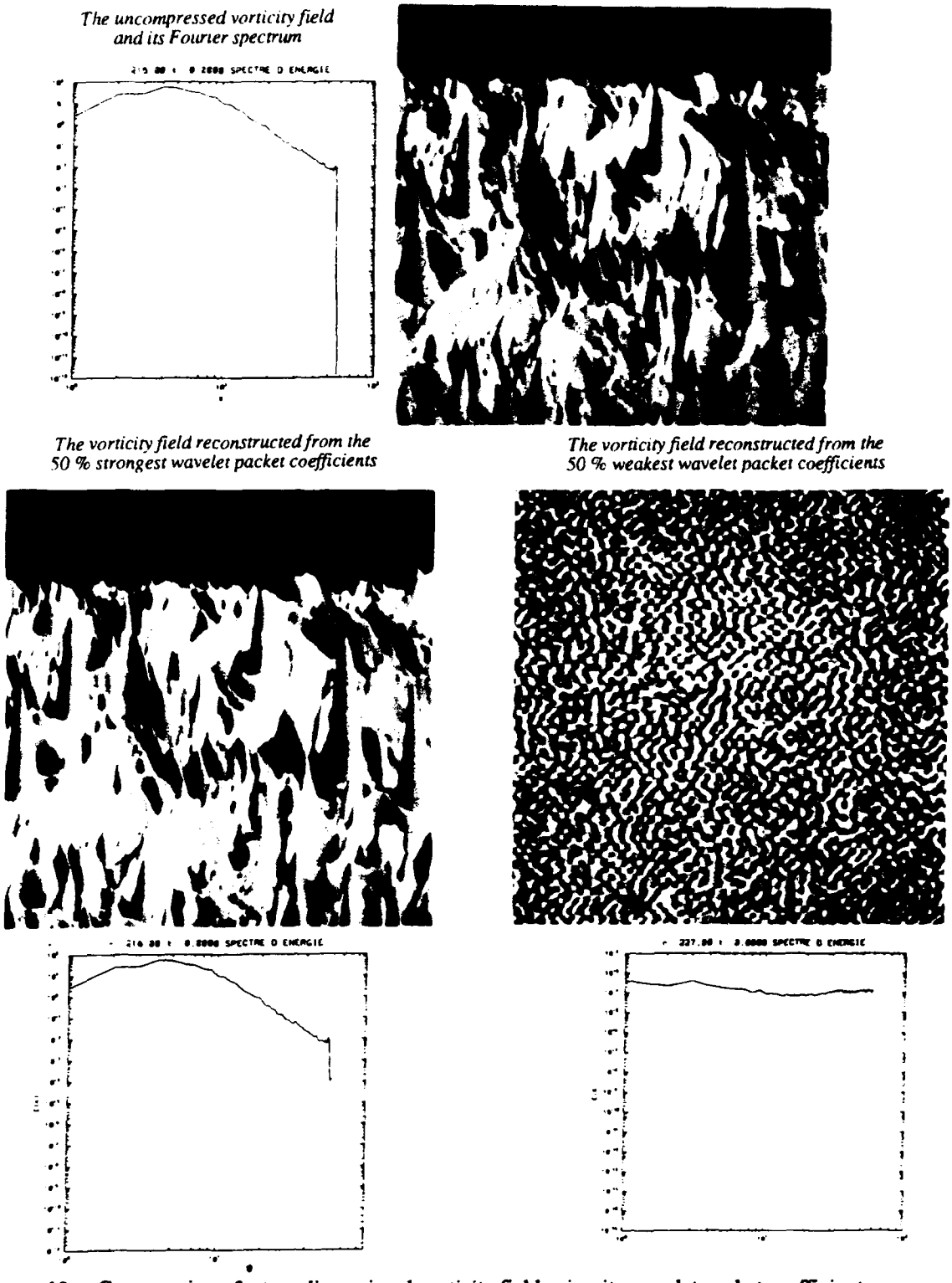


Figure 13a. Compression of a two-dimensional vorticity field using its wavelet packet coefficients, compression by a factor 2; (top) the uncompressed field and its energy spectrum, (center left and lower left) the compressed field and its energy spectrum, (center right and lower right) the discarded field and its energy spectrum. The visualisation was done in collaboration with Jean-Francois Colonna.

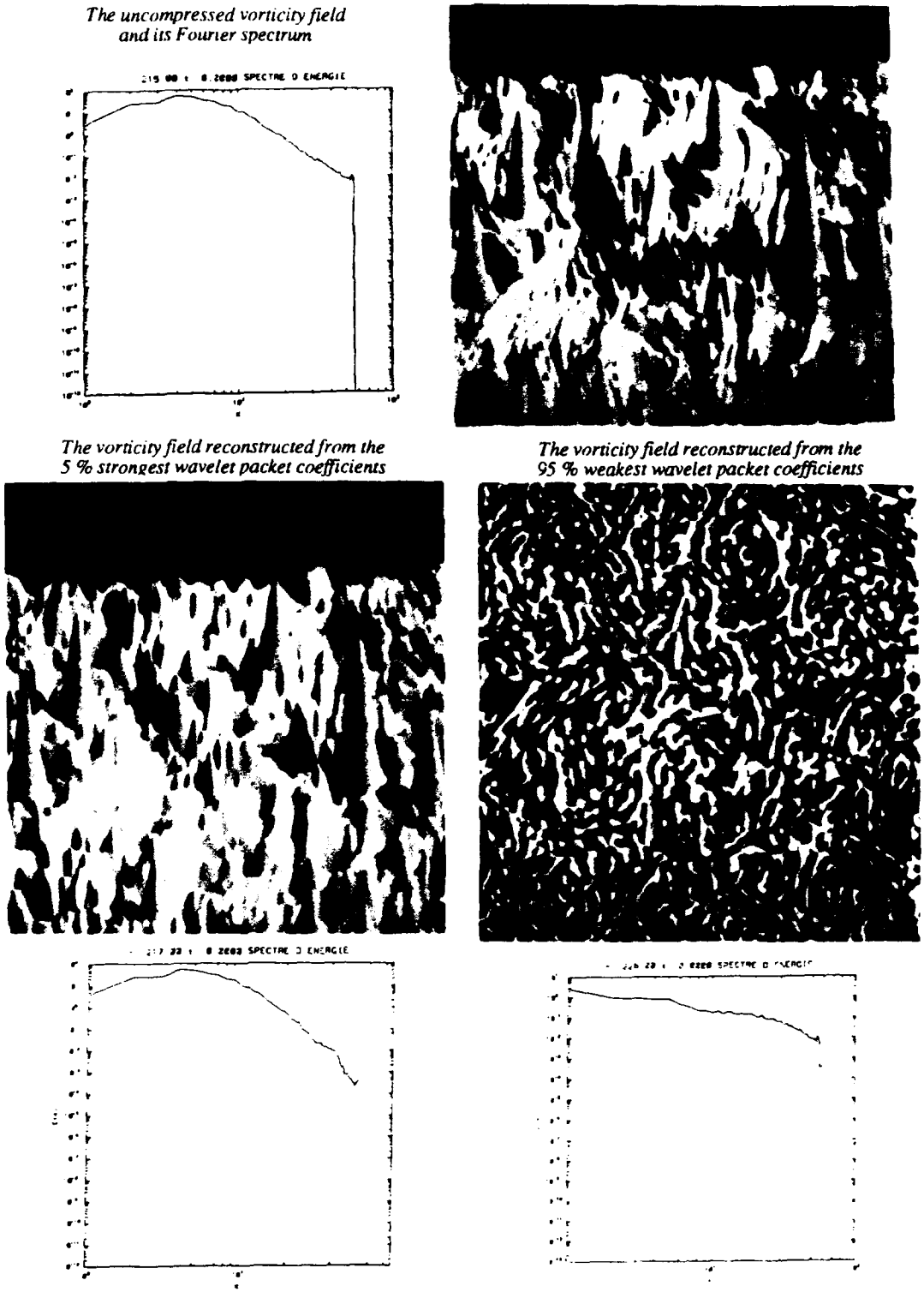


Figure 13b. Compression of a two-dimensional vorticity field using its wavelet packet coefficients, compression by a factor 20; (top) the uncompressed field and its energy spectrum, (center left and lower left) the compressed field and its energy spectrum, (center right and lower right) the discarded field and its energy spectrum. The visualisation was done in collaboration with Jean-Francois Colonna.

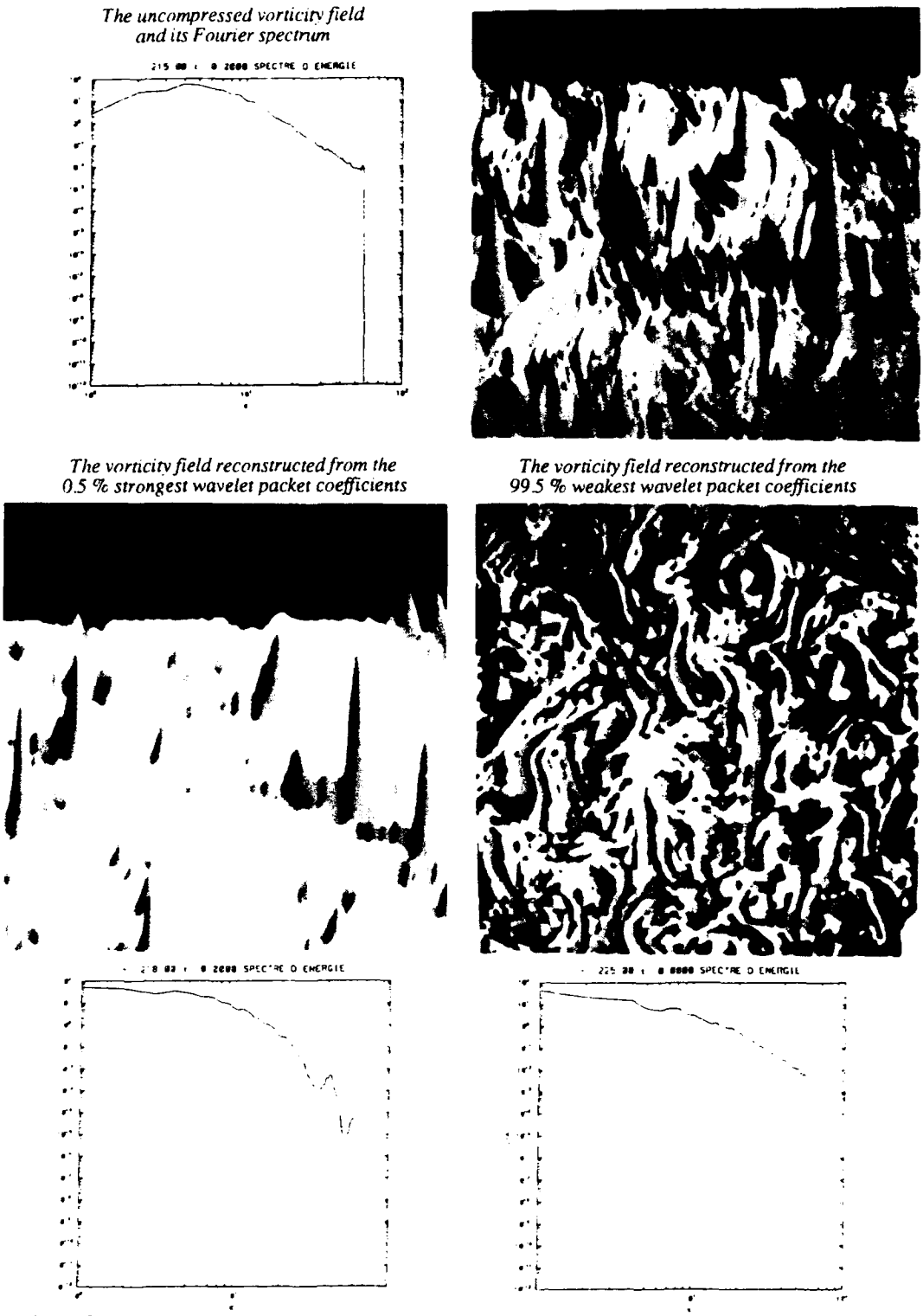


Figure 13c. Compression of a two-dimensional vorticity field using its wavelet packet coefficients, compression by a factor 200; (top) the uncompressed field and its energy spectrum, (center left and lower left) the compressed field and its energy spectrum, (center right and lower right) the discarded field and its energy spectrum. The visualisation was done in collaboration with Jean-Francois Colonna.

on the forcing. If this is true we may hope that the local scaling of the different regions may be universal enough in order to be able to model their behaviour, each region then having its own parametrization.

Using the orthogonal wavelet packet transform, we have shown that the significant coefficients correspond to the coherent structures, while the weak coefficients correspond to the vorticity filaments which are only passively advected by the coherent structures. One possible application of the wavelet packet algorithm is to apply it from time to time during a numerical simulation, in order to separate regions with highly active small scales, which need a better grid resolution, from regions with inactive small scales, which do not contribute much to the dynamics and can either be discarded or modelled. Indeed the wavelet packet Best Basis seems to distinguish the low-dimensional, dynamically active part of the flow from the high-dimensional, passive components. It gives us some hope of drastically reducing the number of degrees of freedom necessary to the computation of two-dimensional turbulent flows.

Acknowledgments. The authors are supported by the CEE 'Human Capital and Mobility Program' (contract n° ERB-CHRX-CT92-0001) and the NATO 'Collaborative Research' program (contract n° CRG-930456). The computations were carried out on the Cray-2 of the Centre de Calcul Vectoriel pour la Recherche, Palaiseau, France, and granting of the computing time is gratefully acknowledged.

REFERENCES

Balian R., 1981. Un principe d'incertitude fort en théorie du signal ou en mécanique quantique.*C. R. Acad. Sci. Paris, 292, II,* 1357-1361.

Beylkin G., Coifman R., and Rokhlin V., 1992, Wavelets in Numerical Analysis, *Wavelets and their Applications,* ed. Ruskai M. B. et al., Jones and Bartlett, 181-210.

Coifman R., Meyer Y., Quake S. and Wickerhauser M. V., 1990, *Signal Processing and Compression with Wavelet Packets,* Numerical Algorithms Research Group, Yale University.

Coifman R., Meyer Y. and Wickerhauser M. V., 1992, Wavelet Analysis and Signal Processing, in *Wavelets and their Applications,* ed. Ruskai M. B. et al., Jones and Bartlett, 153-178.

Coifman R. and Wickerhauser M. V., 1990, *Best-adaptated Wave Packet Bases,* Preprint, Department, Yale University.

Coifman R. and Meyer Y., 1991, Remarques sur l'analyze de Fourier à fenêtre, *C. R. Acad. Sci. Paris,* **312**, serie I, 259-261.

Daubechies I., 1988, Orthonormal Bases of Compactly Supported Wavelets, Comm. in Pure and Applied Math., **49**, 909-996.

Farge M., 1992a, Wavelet Transforms and their Applications to Turbulence, *Annu. Rev. Fluid Mech.*, ***24***, 395-457.

Farge M., 1992b, The Continuous Wavelet Transform of Two-dimensional Turbulent Flows, *Wavelets and their Applications*, ed. Ruskai M. B. et al., Jones and Bartlett, 99.

Farge M., Goirand E., Meyer Y., Pascal F. and Wickerhauser M. V., 1992, Improved Predictability of Two-dimensional Turbulent Flows using Wavelet Packet Compression, *Fluid Dynamics Research, **10**, 229-250.*

Gabor D., 1946, Theory of Communication, *J. Inst. Electr. Engin., **93**, III, 429-457.*

Grossmann A. and Morlet J., 1984, Decomposition of Hardy Functions into Square Integrable Wavelets of Constant Shape, *S.I.A.M., J. Math. An., **15**, 723-736.*

Grossmann A. and Morlet J., 1985, Decomposition of Functions into Wavelets of Constant Shape, and Related Transforms, *Mathematics and Physics, Lectures on Recent Results, World Scientific Publishing.*

Holschneider M., 1988, On the wavelet transform of fractal objects, *Stat. Phys., **50**, no. 5/6.*

Jaffard S., 1989, Construction of wavelets on open sets, *Proceedings of the 1st International Conference on Wavelets, Marseille, 14-18 December 1987*, ed. Combes et al., Springer, 247-252.

Lemarie P.G. and Meyer Y., 1986, Ondelettes et bases Hilbertiennes, *Rev. Mat. Ibero-americana, **2**, 1.*

Mallat S.G., 1989, A Theory for Multiresolution Signal Decomposition: the Wavelet Decomposition, *IEEE Trans. on Pattern Analysis and Machine Intelligence, **11**, 674-693.*

Malvar H., 1990, Lapped Transforms for Efficient Transforms/Subband Coding, *IEEE Trans. on Acoustics, Speech and Signal Processing, **38**, 969-978.*

Murenzi R., 1989, Wavelet Transforms Associated to the N-dimensional Euclidean Group with Dilatations: Signal in More than One Dimension, *Proceedings of the 1st International Conference on Wavelets, Marseille, 14-18 December 1987*, ed. Combes et al., Springer, 239-246.

Weiss J., 1981, The Dynamics of Enstrophy Transfer in Two-dimensional Hydrodynamics, *Report LJI-TN-121, La Jolla Institute, San Diego.*

Wickerhauser M. V., 1990, Picture Compression by Best-Basis Subband Coding, *Preprint, Mathematic Department, Yale University.*

Zabusky N., 1984, Computational Synergetics, *Physics Today, July 1984, 2-11.*

THE NUMERICAL INVERSE SCATTERING TRANSFORM: NONLINEAR FOURIER ANALYSIS AND NONLINEAR FILTERING OF OCEANIC SURFACE WAVES

A. R. Osborne
Istituto di Fisica Generale dell'Università, Via Pietro Giuria 1, 10125 Torino, Italy

ABSTRACT

Nonlinear Fourier analysis is discussed as it arises from the exact spectral solution to large classes of nonlinear wave equations which are integrable by the *inverse scattering transform* (IST). The approach may be viewed as a generalization of the ordinary, linear Fourier transform or Fourier series. Numerical methods are discussed which allow for implementation of the approach as a tool for the time series analysis of oceanic wave data. I specifically consider the case for shallow water, where integrable nonlinear wave motion is governed by the Korteweg-deVries equation with periodic/quasi-periodic boundary conditions. Numerical procedures given herein allow the computation of a nonlinear Fourier series for a measured time series. The nonlinear oscillation modes of KdV obey a linear superposition law, just as do the sine waves of a linear Fourier series. However, the KdV basis functions themselves are highly nonlinear, undergo nonlinear interactions with each other and are distinctly non sinusoidal. I analyze surface wave data from the Adriatic Sea and apply the concept of nonlinear filtering to enhance understanding of nonlinear interactions.

INTRODUCTION

This paper summarizes a new numerical approach for the *nonlinear Fourier analysis of space and time series* of complex, nonlinear wave trains. The method, based upon the (periodic/quasi-periodic) *inverse scattering transform* (IST), is a kind of nonlinear generalization of the ordinary, linear Fourier transform. I focus on nonlinear wave motion for shallow-water waves as governed by the Korteweg-deVries (KdV) equation. IST may be exploited to determine the *numerical inverse scattering transform* (NIST) spectrum of a measured or computed wave train which is assumed to be periodic (or quasi-periodic) in space or in time. The approach may also be applied to *numerically construct complex solutions* to the KdV equation. I build on previous successes in the application of the periodic scattering transform to the analysis of computer generated or experimentally measured data [Bishop et al., 1986; Osborne and Bergamasco, 1985, 1986; Osborne and Segre, 1990; Terrones et al., 1990; Flesch et al., 1991; Osborne et al., 1991; Osborne,

1991a, 1991b; McLaughlin and Schober, 1992; Osborne, 1993]. In particular I analyze measured wave data obtained in the Adriatic Sea on a fixed offshore platform in 16.5 m of water, about 10 km from Venice, Italy. This paper describes some of the recent work done in collaboration with L. Cavaleri [Osborne et al., 1991; Osborne and Cavaleri, 1993]. It is hoped that the results of this paper will complement other recent work in the propagation of nonlinear shallow water waves [Elgar and Guza, 1986].

THE KdV EQUATION AND PERIODIC INVERSE SCATTERING THEORY

The Kortweg-deVries equation describes (among many other physical applications) the motion of small, finite-amplitude nonlinear wave trains in shallow water. KdV was the first of many nonlinear wave equations to be completely integrated by what is now called the *inverse scattering transform* [Zakharov et al., 1980; Ablowitz and Segur, 1981; Dodd et al., 1982; Newell, 1985; Degasperis, 1991].

The dimensional form for the (space-like) KdV equation is given by [Whitham, 1974; Miles, 1980]:

$$\eta_t + c_0 \eta_x + \alpha \eta \eta_x + \beta \eta_{xxx} = 0 \tag{1}$$

where $\eta(x,t)$ is the wave amplitude as a function of space x and time t. For shallow water wave motion the constant coefficients of KdV are given by $c_0 = (gh)^{1/2}$, $\alpha = 3c_0 / 2h$ and $\beta = c_0 h^2 / 6$. Eq. (1) has the linearized dispersion relation $\omega = c_0 k - \beta k^3$; g is the acceleration of gravity, c_0 is the linear phase speed, and h is the water depth. Subscripts with respect to x and t refer to partial derivatives. KdV solves the *Cauchy problem*: given the spatial behavior of the wave train at $t = 0$, $\eta(x,0)$, (1) determines the motion for all space and time thereafter, $\eta(x,t)$. Here we use periodic boundary conditions so that $\eta(x,t) = \eta(x + L,t)$, for L the spatial period of the wave train.

The most common experimental situation is to record data as a function of time at a single spatial location. The reasons are often economical, e.g., the measurement of time series requires a single wave staff or pressure recorder; the measurement of space series requires remote sensing capability. These considerations motivate the need to determine the scattering transform of a *time series*, $\eta(0,t)$. To this end one may apply the *time-like* KdV equation (TKdV) [Karpman, 1974; Ablowitz and Segur, 1981]:

$$\eta_x + c_0' \eta_t + \alpha' \eta \eta_t + \beta' \eta_{ttt} = 0 \tag{2}$$

where $c_0' = 1/c_0$, $\alpha' = -\alpha / c_0^2$ and $\beta' = -\beta / c_0$; (2) has the linearized dispersion relation $k = \omega / c_0 + (\beta / c_0^4)\omega^3$. TKdV solves a *boundary value problem*: given the temporal

evolution $\eta(0,t)$ at a fixed spatial location $x = 0$, (2) determines the wave motion over all space as a function of time, $\eta(x,t)$. Periodic boundary conditions ($\eta(x,t) = \eta(x,t+T)$) are assumed herein in order to be consistent with linear Fourier algorithms (discrete and fast Fourier transforms). Due to recent advances in numerical methods TKdV may now be routinely applied to the time series analysis of experimental data [Osborne, 1991a; Osborne et al., 1991; Osborne and Segre, 1990].

All solutions of (1) may be easily converted to all solutions of (2) by simple transformations given elsewhere [Osborne, 1983; Osborne, 1993]. Hence the scattering transform of (2) may be easily expressed in terms of the scattering transform of (1). For present purposes it is only necessary to note that given the IST for (1), the IST for (2) may be easily determined. Therefore, I give herein only the mathematical development of IST for (1).

According to the periodic inverse scattering transform the solution to the periodic KdV equation (1) may be written as a *linear superposition of nonlinearly interacting, nonlinear waves* called hyperelliptic functions, $\mu_j(x;x_o,0)$:

$$\lambda\eta(x,t) = -E_1 + \sum_{j=1}^{N}[2\mu_j(x;x_o,t) - E_{2j} - E_{2j+1}] \tag{3}$$

The constant parameter $\lambda = \alpha / 6\beta$. This is the first of the so-called trace formulae for the KdV equation [Dubrovin and Novikov, 1974; Flaschka and McLaughlin, 1976] and may be interpreted as a kind of *nonlinear Fourier series*. The constant parameters E_{2j}, E_{2j+1} are eigenvalues of the "main spectrum" of periodic theory as discussed in the next section; x_o is an arbitrary *base point* in the interval $0 \le x \le L$. The μ_j are the *nonlinear oscillation modes* of periodic KdV, i.e., they are analogous to the sine waves of linear Fourier analysis. The μ_j spatially evolve according to the following system of coupled, nonlinear, ordinary differential equations:

$$\frac{d\mu_j}{dx} = \frac{2i\sigma_j R^{1/2}(\mu_j)}{\prod_{\substack{k=1 \\ j\neq k}}^{N}(\mu_j - \mu_k)} \tag{4}$$

where

$$R(\mu_j) = \prod_{k=1}^{2N+1}(\mu_j - E_k). \tag{5}$$

The $\sigma_j = \pm 1$ are the signs of the square root of $R(\mu_j)$. The μ_j dynamically evolve on two-sheeted Riemann surfaces; the branch points connecting the surfaces are referred to

as "band edges" and are denoted by the E_{2j} and E_{2j+1}. The spatially and temporally varying μ_j evolve inside an "open band," e.g., in the interval $E_{2j} \le \mu_j \le E_{2j+1}$, and oscillate between these limits as a function of x and t, as will be demonstrated graphically below. When a μ_j reaches a band edge (either E_{2j} or E_{2j+1}) the sign σ_j changes and the motion leaps to the other Riemann sheet. This fact, together with the strong nonlinear coupling occurring among the μ_j presented considerable difficulties for Osborne and Segre [1990] in numerical integrations of (4). These difficulties have been largely circumvented by the methods given herein for the time series analysis of nonlinear wave trains.

The temporal evolution of the μ_j is given by the following differential equations:

$$\frac{d\mu_j}{dt} = -2[\lambda\eta(x,t) - 2\mu_j]\frac{d\mu_j}{dx} \tag{6}$$

where $\lambda\eta(x,t)$ is given by (3). The space (4) and time (6) ODEs evolve the $\mu_j(x,t)$ (the nonlinear oscillation modes of KdV) and the *nonlinear Fourier series* (3) allows one to construct general solutions to the KdV equation. In what follows I describe methods for numerically computing the oscillation modes $\mu_j(x,0)$ at a particular instant of time, $t = 0$. The requisite numerical methods are then christened *nonlinear Fourier analysis procedures for space or time series* [Osborne, 1991a].

Generally speaking I refer to the numerical determination of the *main spectrum* (E_i; $1 \le i \le 2N+1$) and the *auxiliary spectrum* ($\mu_j(0,0), \sigma_j = \pm 1; 1 \le j \le N$) as the *direct scattering transform* (see details in the Section below). The computation of the hyperelliptic functions $\mu_j(x,t)$ as solutions of the nonlinear ODEs (4)-(6) and the construction of solutions of the KdV equation by the trace formula (3) constitutes the *inverse scattering transform.* Herein I (a) discuss new numerical procedures for obtaining the direct scattering transform and (b) show that the inverse scattering transform as obtained by numerical integration of (4)-(6) (e.g. as considered by Osborne and Segre [1990]) can be replaced by a much simpler, more precise and faster algorithm.

THE PERIODIC INVERSE SCATTERING TRANSFORM

The spectral problem (the direct scattering transform) for KdV (1) is the Schroedinger eigenvalue problem of quantum mechanics:

$$\psi_{xx} + [\lambda\eta(x) + k^2]\psi = 0 \qquad (k^2 = E) \tag{7}$$

where $\eta(x) = \eta(x,0)$ is the solution to the KdV equation (1) at an arbitrary time $t = 0$; k is the spectral wavenumber. Periodic boundary conditions are assumed so that we take $\eta(x,t) = \eta(x+L,t)$ for L the period.

Details of the inverse scattering theory will not be given here, but may be found elsewhere [Dubrovin and Novikov, 1974; Dubrovin, Matveev and Novikov 1976; Flaschka and McLaughlin, 1976; McKean and Trubowitz, 1976]. For numerical purposes it is appropriate to consider a basis of solutions (c, s) of (7) such that

$$\begin{pmatrix} c(x_o) & c'(x_o) \\ s(x_o) & s'(x_o) \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}. \tag{8}$$

The wronskian $W(c, s) = 1$ so that (c, s) is a basis set of (1). The matrix α carries the solution of (1) from the point x to $x + L$:

$$\begin{pmatrix} c(x+L) & c'(x+L) \\ s(x+L) & s'(x+L) \end{pmatrix} = \begin{pmatrix} \alpha_{11} & \alpha_{12} \\ \alpha_{21} & \alpha_{22} \end{pmatrix} \begin{pmatrix} c(x) & c'(x) \\ s(x) & s'(x) \end{pmatrix} \tag{9}$$

α is often referred to as the *monodromy matrix*. This is the fundamental matrix of periodic spectral theory for KdV; α contains *all* spectral information about KdV in the wavenumber domain.

The so called *main spectrum* of KdV consists of eigenvalues E_i that correspond to the Bloch eigenfunctions of the Schroedinger equation (7) for a particular period L. The *auxiliary spectrum* is defined as the eigenvalues for which the eigenfunctions $s(x)$ have the fixed boundary conditions $s(x_0+L) = s(x_0) = 0$. To this end one has these specific spectral definitions:

$$\begin{aligned} &\text{main spectrum } \{E_i;\ 1 \le i \le 2N+1\}: && \tfrac{1}{2}(\alpha_{11} + \alpha_{12})(E) = \pm 1 \\ &\text{auxiliary spectrum } \{\mu_j; 1 \le j \le N\}: && \alpha_{21}(\mu) = 0 \\ &\{\sigma_j\} = \{sgn[\alpha_{11}(E) - \alpha_{22}(E)]_{E=\mu_j};\ 1 \le j \le N\}. \end{aligned} \tag{10}$$

The eigenvalues $\{E_i; \mu_j; \sigma_j\}$ constitute the *direct scattering transform* of a wave train of N degrees of freedom, $1 \le i \le 2N+1$; $1 \le j \le N$. The *inverse scattering transform*, (3)-(6), then allows for the construction of complex wave train solutions of the KdV equation.

THE NUMERICAL ALGORITHM

The numerical search for the scattering eigenvalues $\{E_i;\mu_j;\sigma_j\}$ suggests the need for computing the derivatives of the matrix α_{ij} with respect to the energy E. This is because one normally uses a Newtonian numerical root-finding algorithm to determine the eigenvalues. To achieve this goal, a matrix method for obtaining the evolution of the eigenfunction ψ as a function of x and E for a particular wave train $\eta(x,0)$ has been developed. The key to this approach is the analytical estimation of derivatives of the matrix elements with respect to E.

To this end the spectral equations are

$$\begin{aligned} \psi_{xx} &= -q\,\psi \\ \psi_{xxE} &= -q\,\psi_E - \psi \end{aligned} \tag{11}$$

where the subscripts refer to differentiation with respect to x and E; $q(x) = \lambda\eta(x) + E$. Writing (11) in four-vector notation and using a Taylor series expansion for the solution to the scattering equations (11) one obtains

$$\begin{pmatrix} \psi(x+\Delta x) \\ \psi_x(x+\Delta x) \\ \psi_E(x+\Delta x) \\ \psi_{xE}(x+\Delta x) \end{pmatrix} = \mathbf{H} \begin{pmatrix} \psi(x) \\ \psi_x(x) \\ \psi_E(x) \\ \psi_{xE}(x) \end{pmatrix} \tag{12}$$

where

$$\mathbf{H} = \begin{pmatrix} \mathbf{T} & \mathbf{0} \\ \mathbf{T}_E & \mathbf{T} \end{pmatrix} \tag{13}$$

Each element of $\mathbf{H}$ is a two-by-two matrix. The matrix $\mathbf{0}$ has zero for all its elements and the other matrices are given by:

$$\mathbf{T} = \begin{pmatrix} \cos(\kappa\Delta x) & \dfrac{\sin(\kappa\Delta x)}{\kappa} \\ -\kappa\sin(\kappa\Delta x) & \cos(\kappa\Delta x) \end{pmatrix} \tag{14}$$

and

$$\mathbf{T}_E = \frac{\partial \mathbf{T}}{\partial E} = \begin{pmatrix} -\dfrac{\Delta x \sin(\kappa \Delta x)}{2\kappa} & \dfrac{\Delta x \cos(\kappa \Delta x)}{2\kappa^2} - \dfrac{\sin(\kappa \Delta x)}{2\kappa^3} \\ -\dfrac{\Delta x \cos(\kappa \Delta x)}{2} + \dfrac{\sin(\kappa \Delta x)}{2\kappa} & -\dfrac{\Delta x \sin(\kappa \Delta x)}{2\kappa} \end{pmatrix} \tag{15}$$

for $\kappa = (q)^{1/2} = (\lambda \eta(x) + E)^{1/2}$. While κ may be either real or imaginary, the matrix $\mathbf{T}_E$ is always real with determinant 1. This property is exploited in the numerical algorithm below.

As in previous numerical problems of this type I assume the wave train $\eta(x)$ has the form of a *piece wise constant function* with $2M$ partitions on the periodic interval $(0, L)$, where the discretization interval is $\Delta x = L / 2M$ [Osborne, 1991a]. Each partition has wave amplitude $\eta_n (1 \le n \le 2M)$ which is associated with a discrete value of the spatial variable $x_n = n\Delta x$. The four-by-four scattering matrix $\mathbf{M}$ can then be defined:

$$\mathbf{M} = \prod_{n=M-1}^{-M} \mathbf{H}(\eta_n, \Delta x) \tag{16}$$

The initial conditions of the basis (c, s) at the base point x_0 are given by

$$\begin{pmatrix} c(x_0) \\ c'(x_0) \\ c_E(x_0) \\ c'_E(x_0) \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \end{pmatrix} \quad \begin{pmatrix} s(x_0) \\ s'(x_0) \\ s_E(x_0) \\ s'_E(x_0) \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \\ 0 \\ 0 \end{pmatrix}. \tag{17}$$

From the definition of the matrix α one has

$$\{\alpha_{ij}\} = \begin{pmatrix} c(x+L) & c'(x+L) \\ s(x+L) & s'(x+L) \end{pmatrix} \begin{pmatrix} c(x) & c'(x) \\ s(x) & s'(x) \end{pmatrix}^{-1}. \tag{18}$$

Thus at x_0 one finds

$$\frac{1}{2}(\alpha_{11} + \alpha_{22}) = \frac{1}{2}(M_{11} + M_{22}) \tag{19}$$

$$\alpha_{21} = M_{12} \tag{20}$$

while the derivatives are given by

$$\frac{\partial}{\partial E}\tfrac{1}{2}(\alpha_{11}+\alpha_{22})=\tfrac{1}{2}(M_{31}+M_{42}) \tag{21}$$

$$\frac{\partial \alpha_{21}}{\partial E}=M_{32}. \tag{22}$$

Implementation of the Numerical Algorithm

Because $\kappa=(\lambda\eta(x,0)+k^2)^{1/2}$ can be either real or imaginary, but not complex, the matrix $\boldsymbol{H}$ is always real. This result allows implementation of an algorithm which is entirely real. The following relations have been used in the computer code:

$$T_{11}=T_{22}=\begin{cases}\cos(k'\,\Delta x) & \text{if } \kappa^2\geq 0\\ \cosh(k'\,\Delta x) & \text{if } \kappa^2<0\end{cases} \tag{23}$$

$$T_{12}=\begin{cases}\dfrac{\sin(\kappa'\,\Delta x)}{\kappa'} & \text{if } \kappa^2\geq 0\\ \dfrac{\sinh(\kappa'\,\Delta x)}{p'} & \text{if } \kappa^2<0\end{cases} \tag{24}$$

$$T_{21}=\begin{cases}-\kappa'\sin(\kappa'\,\Delta x) & \text{if } \kappa^2\geq 0\\ \kappa'\sinh(\kappa'\,\Delta x) & \text{if } \kappa^2<0\end{cases} \tag{25}$$

where

$$\kappa'=\sqrt{|\lambda\eta+k^2|}=\sqrt{|\kappa^2|} \tag{26}$$

and analogously for the matrix $\boldsymbol{T_E}$.

The reconstruction of complex solutions of the KdV equation by (3) (as well as *nonlinear filtering*) are carried out by computing the auxiliary spectra $\mu_j(x_o=x_n)$ for the $2M$ different base points $x_0=x_{-M}\ldots x_1, x_2, \ldots x_{M-1}$. The approach is formally called *base point iteration* and is carried out by computing $2M$ different monodromy matrices (16) which differ from each other by a horizontal shift Δx in the wave train η_n. This procedure arises from the following similarity transformation which is easily seen from (16):

$$M(x_{n+1},E)=H(\eta_n,E)M(x_n,E)H(\eta_n,E)^{-1} \tag{27}$$

The latter expression relates the matrix $M(x_{n+1}, E)$ at a base point x_{n+1} to the previously computed matrix $M(x_n, E)$ at the base point x_n for a particular value of $E = k^2$. Values of the auxiliary spectra $\{\mu_j(x_n)\}$ for each x_n are computed from the matrices $M(x_n, E)$. Knowledge of the auxiliary spectra at every point x_n allows reconstruction of the wave train $\eta(x_n)$ via a *discrete version* of (3):

$$\lambda\eta(x_n) = -E_1 + \sum_{j=1}^{N}[2\mu_j(x_n) - E_{2j} - E_{2j+1}] \tag{28}$$

for $n = -M...1,2,...M-1$. This is a finite-term *nonlinear generalization of a Fourier series* for the discrete wave train $\eta(x_n)$. As indicated by the notation, each nonlinear oscillation mode $\{\mu_j\}$ implicitly depends upon the associated wavenumber k_j of the mode. The k_j are theoretically given by the simple relation $k_j = j\Delta k, \Delta k = 2\pi / L$; surprisingly these are exactly the same as for the linear Fourier transform, provided that periodicity is assumed. The IST spectrum then consists of the widths of the open bands of the Floquet discriminant, $a_j = (E_{2j+1} - E_{2j}) / 2\lambda$, graphed as a function of k_j (or f_j for a time series).

EXAMPLE OF NONLINEAR FOURIER ANALYSIS

To illustrate the numerical inverse scattering transform in the analysis of nonlinear wave trains, in Figure 1 I give the numerical construction of a three degree-of-freedom wave train. In panel (a) are the hyperelliptic functions $\mu_j, j = 6, 9, 11$; in the present case the μ_j are constructed from a rather arbitrary selection of the eigenvalues E_{2j}, E_{2j+1}. The linear superposition of the three oscillation modes gives the solution to KdV as shown in the upper part of panel (a). Note that the hyperelliptic oscillation modes are highly non-sinusoidal in appearance due to nonlinear effects. In panel (b) are shown the amplitudes of the *linear* Fourier modes (solid line) and of the three hyperelliptic modes (vertical lines). Comparing these results one concludes that only three nonlinear oscillation modes (three $\mu_j(x)$) are required to describe the motion, while instead the number of linear Fourier modes is quite large (~ 50) for this example.

ANALYSIS OF MEASURED ADRIATIC SEA WAVETRAINS

I extend results recently discussed by Osborne et al. [1991] and Osborne and Cavaleri [1993] with regard to the analysis of nonlinear wave data obtained in a measurement program in the Adriatic Sea about 10 km from Venice, Italy. The data were recorded in 16.5 m of water on the offshore research platform of the Italian National Research Council (Consiglio Nazionale delle Ricerche) in a region where the bottom slope is rather small, e.g., ~ 1/1000. A typical measured wave train, a 500 point time series with

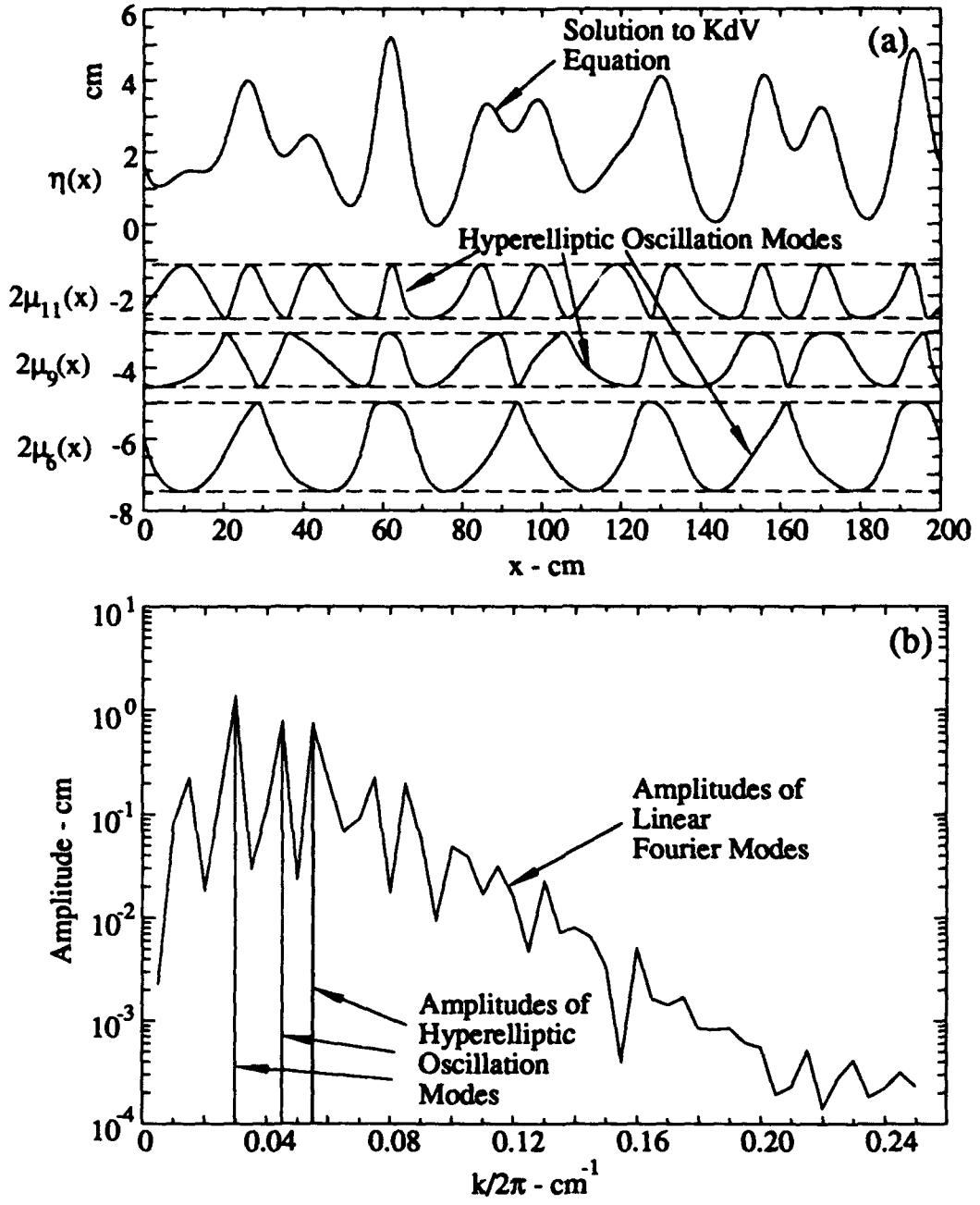


Figure 1. Synthesis of a wave train solution to the KdV equation. In (a) three hyperelliptic function oscillation modes are linearly superposed to give the solution to KdV. In (b) are graphed the linear Fourier transform of the wave train (solid line) and the three nonlinear Fourier amplitudes (the a_j, vertical lines).

temporal discretization $\Delta t = 1$ sec, is shown in Figure 2(a). A data set was selected for which most of the wave energy was in the dominant direction of propagation; only 3% of the wave energy was perpendicular to this direction. This insured that the waves were essentially unidirectional, a requirement of the KdV equation and consequently of the inverse scattering transform analysis given herein. The significant wave height (average of the highest one third waves) is $H_s = 2.9$ m and the dominant period is $T_d = 10.2$ sec. The linear Fourier spectrum is shown in Figure 2(b); the results are quite typical of measured ocean wave spectra, e.g., a central peak (around the dominant period) decays rapidly at low frequency and has a power law spectrum at high frequencies.

It is worthwhile briefly indicating how one determines whether the KdV inverse scattering transform is appropriate for analyzing a particular measured wave train. Clearly if the physics of the wave motion is not that of the KdV equation, then the results of an IST analysis are of dubious value. Three of the more important tests for ascertaining the applicability of KdV for a particular data set are [Osborne and Cavaleri, 1993] (1) Determine whether the data lie in the KdV region of the *Ursell number diagram* [Osborne 1993]. (2) Determine if most of the wave energy lies to the left of $f_{KdV} = 1.36c_o / 2\pi h$ in the frequency domain. (3) Determine if there is little directional spreading in the wave field. For the data analyzed herein all three criteria are met rather well. The results of the first test are discussed in detail in Osborne and Cavaleri [1993], e.g. the (time-like) Ursell number, $\mathrm{Ur} = 2gH_sT_d^2/4h^2 \sim 8$; hence, the Adriatic Sea waves may be judged to be mildly nonlinear. The second test is verified in Figures 3 and 4. Since only three percent of the wave energy is normal to the dominant wave direction, the last criterion is also satisfied to good accuracy. The above constraints on the selection of experimental data given herein may be considered to be rather conservative; efforts are underway to extend the applicability of the present approach to less severely restricted data sets [Osborne, 1993].

I now discuss the nonlinear Fourier analysis of the measured wave train; the Floquet discriminant is shown in Figure 3(a); this constitutes a graph of the half-trace of the monodromy matrix ($\Delta = (\alpha_{11} + \alpha_{22})/2$, the first of equations (10)) as a function of frequency squared, $E = (\pi \mathrm{f})^2$. Note that the fluctuations in $\Delta(E)$ are quite large so that a logarithmic scale has been used to graph the function outside the vertical range $(-1 \le \Delta \le 1)$ (the graph is instead linear inside this interval). The spectrum is seen to divide itself into two widely-separated regions of activity corresponding to solitons (on the left) and radiation components (on the right). Since the soliton part of the Floquet diagram is not easily visible (it is too dense in the domain $\mathrm{E} \sim \mathrm{f}^2$), this part of the spectrum has been graphed separately in Figure 3(b). Here the large oscillations to the left represent the soliton modes in the spectrum. The vertical dotted line is the so-called *reference level* [Osborne and Bergamasco, 1986], which represents the level upon which the solitons propagate in physical space.

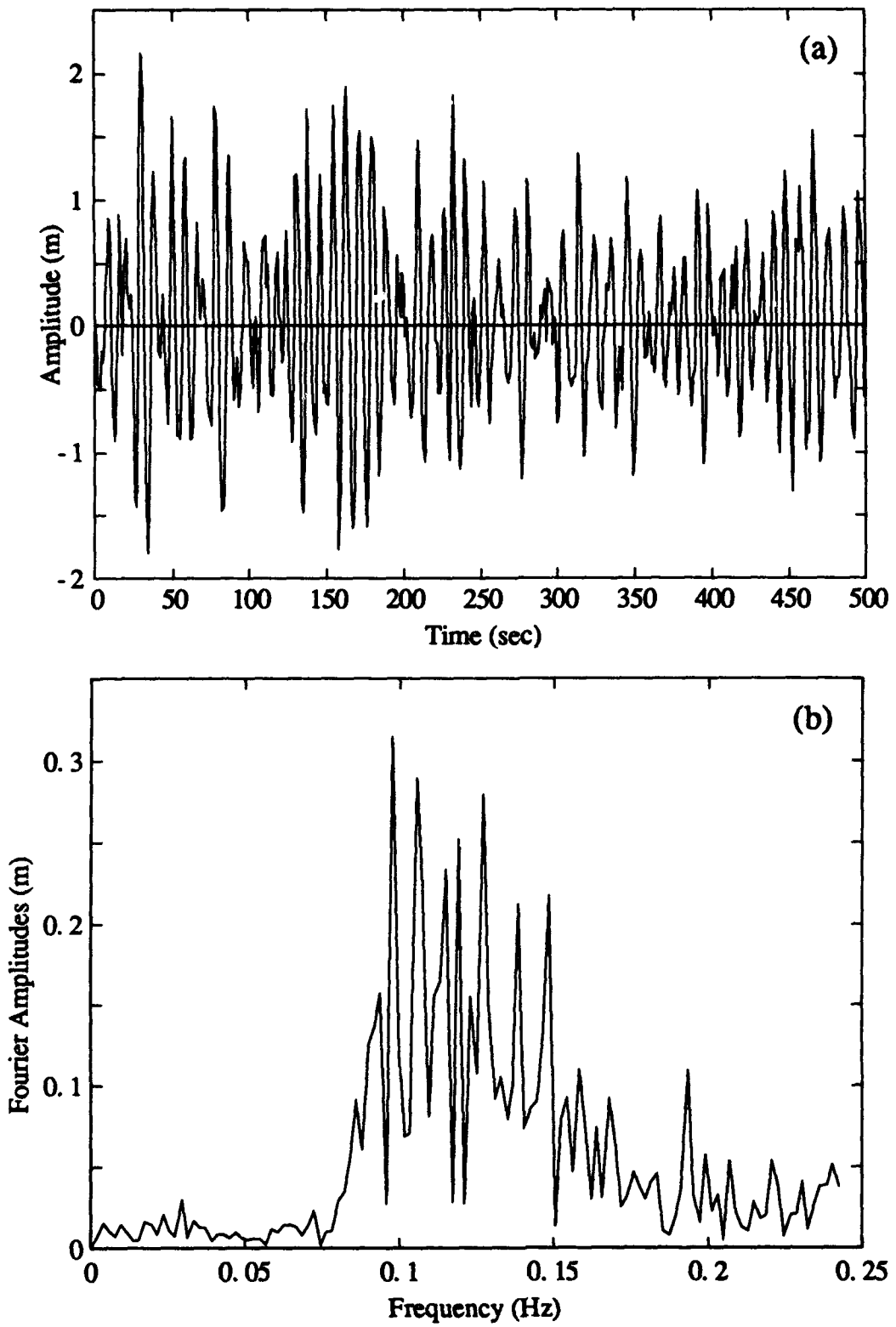


Figure 2. Time series measured in the Adriatic Sea in 16.5 m water depth (a). In (b) the linear Fourier transform of the measured time series is shown.

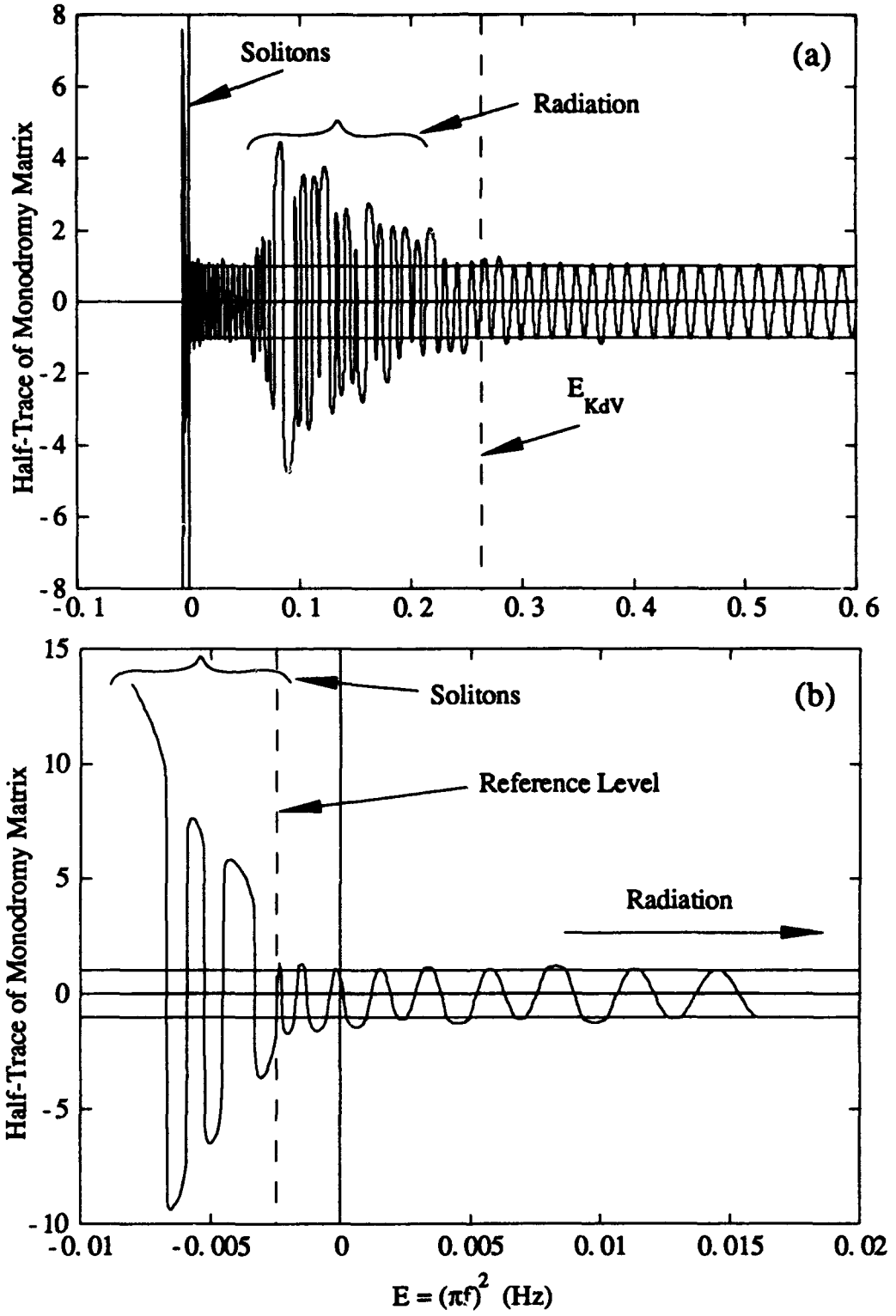


Figure 3. Floquet diagram of measured time series in Figure 2(a) is shown in (a). The Floquet diagram in (a) has been expanded in the soliton (low frequency) part of the spectrum in (b) to reveal the presence of the reference level and the solitons themselves.

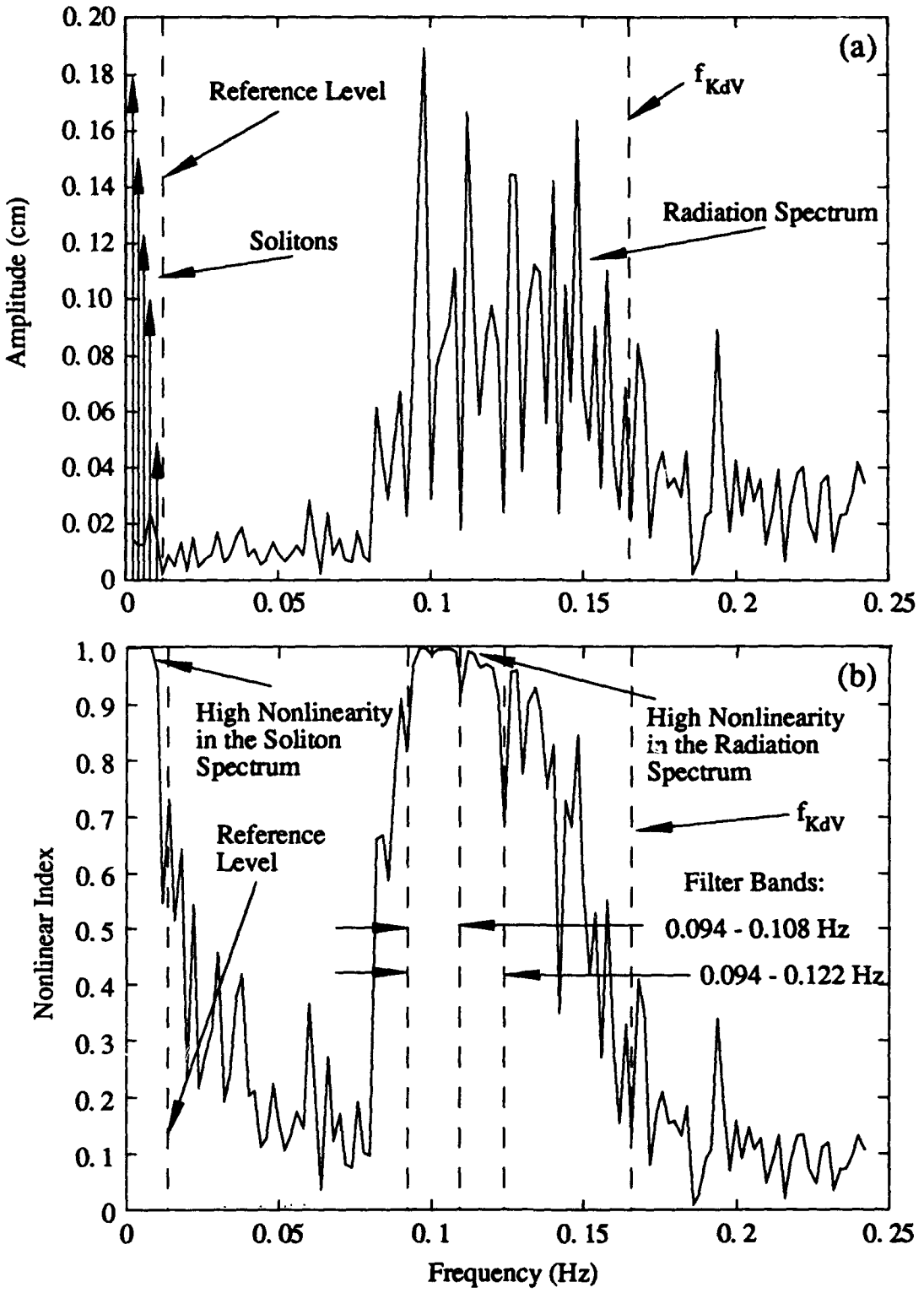


Figure 4. The scattering transform spectrum of the measured time series in Figure 2(a). Shown are the solitons (vertical arrows) and the radiation spectrum (solid line). In (b) is the nonlinear spectral index. Values of the index near 1 indicate strong nonlinear behavior.

The IST spectrum is given in Figure 4(a) where the spectral components are graphed as a function of frequency, just as for the linear Fourier transform. The nonlinear Fourier amplitudes, $a_j = (E_{2j+1} - E_{2j})/2\lambda$, are the amplitudes of the open bands in the Floquet spectrum of Figure 3(b). The radiation spectrum is shown as a solid curve on the right, while the solitons are displayed on the left as vertical arrows. About 7% of the wave energy lies in the soliton part of the spectrum. It is useful to compare the amplitudes of the nonlinear spectrum in Figure 4(a) with those of the linear spectrum in Figure 2(b). Note that the radiation components in the scattering transform spectrum are smaller than those for the linear Fourier spectrum. Physically this occurs because part of the energy has been transferred from the radiation spectrum to the soliton spectrum, due to the presence of nonlinear effects, by the inverse scattering transform.

The *nonlinear spectral index* is shown in Figure 4(b). This parameter indicates just how nonlinear the spectral components are at a particular frequency [Osborne and Bergamasco, 1986]. Since the index indicates strong nonlinear behavior for values near 1, two frequency ranges are of interest in this analysis. The first is at low frequency, signaling the presence of solitons in the spectrum. The second is near the peak of the radiative part of the wave train. Nonlinear interactions are quite strong in these two regions. It is of interest to explore these particular cases using nonlinear filtering, as discussed below.

The next step in the analysis is to compute the hyperelliptic functions (nonlinear oscillation modes) of the data by *base point iteration*. The first 100 nonlinear modes are given in Figure 5(a). The horizontal lines separate each mode from its neighbor on the vertical scale, which has units of squared frequency (these are the units of the horizontal coordinate of the Floquet diagram in Figure 3). While the scale of the nonlinear modes is rather small in this figure, it is still easily seen that they are distinctly non sinusoidal, especially near the larger radiation modes. The solitons are not easily observable at the scale of this figure, but these will be graphed below in such a way as to render them visible. In order to illustrate IST and its associated linear superposition law, I now show how the linear superposition of the nonlinear oscillation modes reconstructs the wave train in Figure 5(b). I have summed nonlinear components only out to 0.2 Hz (the Nyquist frequency is 0.5 Hz), but a comparison of Figure 5(b) with the measured wave train Figure 2(a), indicates that most of the spectral energy has been included. High frequencies have been essentially *filtered* out (above 0.2 Hz) in the reconstruction of Figure 5(b). This example constitutes my first application of the concept of *nonlinear filtering*.

I now consider two further applications of filtering using the nonlinear oscillation modes. The first is with regard to the soliton part of the spectrum, the second is with regard to the most nonlinear part of the radiation spectrum. In Figure 6 I show the hyperelliptic modes in the *soliton* part of the spectrum; the vertical scale has been expanded to allow easy visualization of the soliton m-functions (this corrects the situation

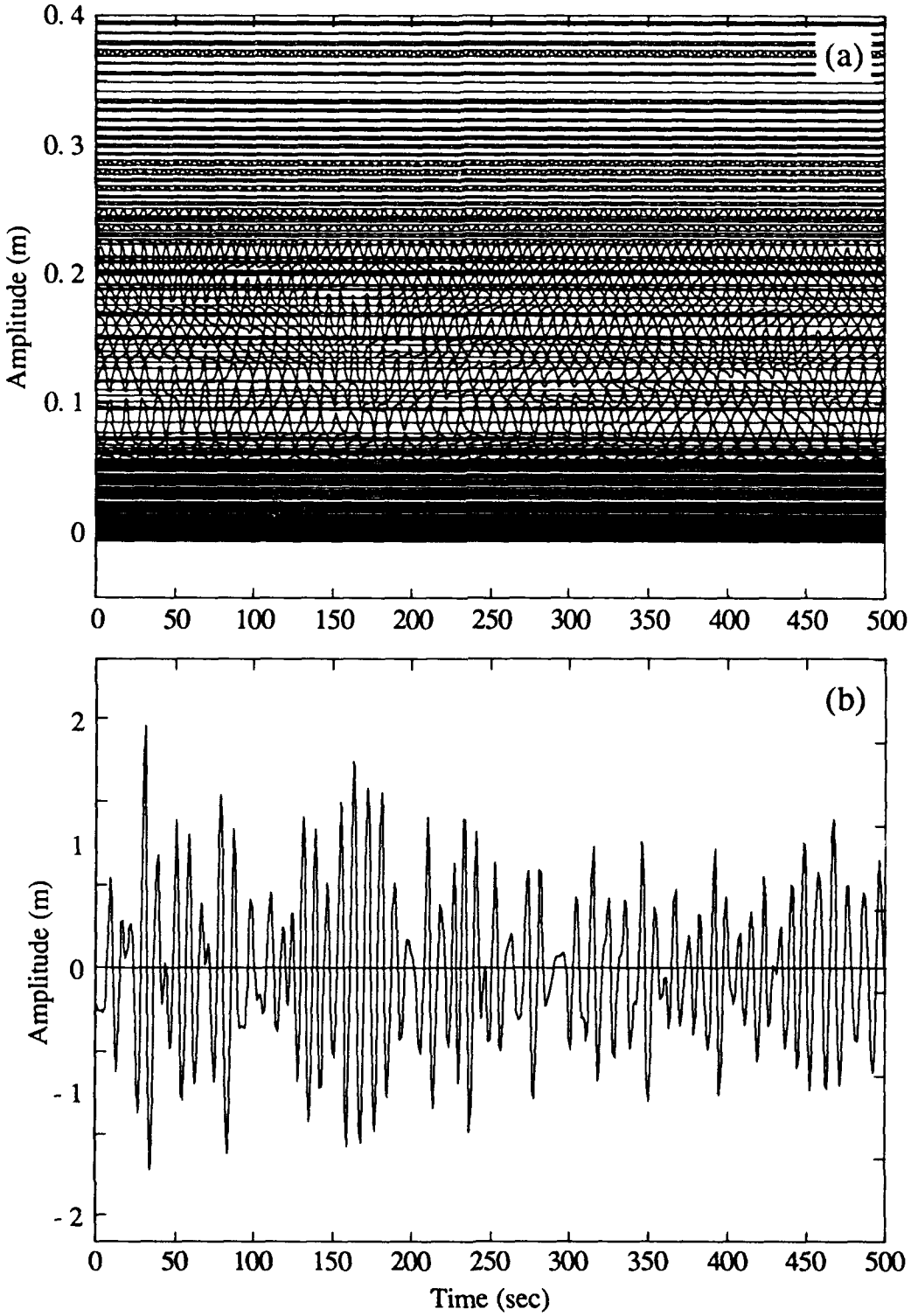


Figure 5. (a) The hyperelliptic oscillation modes for the measured wave train in Figure 2(a). The latter are computed in the frequency range 0.0-0.2 Hz. The linear superposition of these modes gives the wave train shown in (b), which results by low pass filtering the measured wave train. This is the first example of a nonlinearly filtered wave train using the periodic inverse scattering transform.

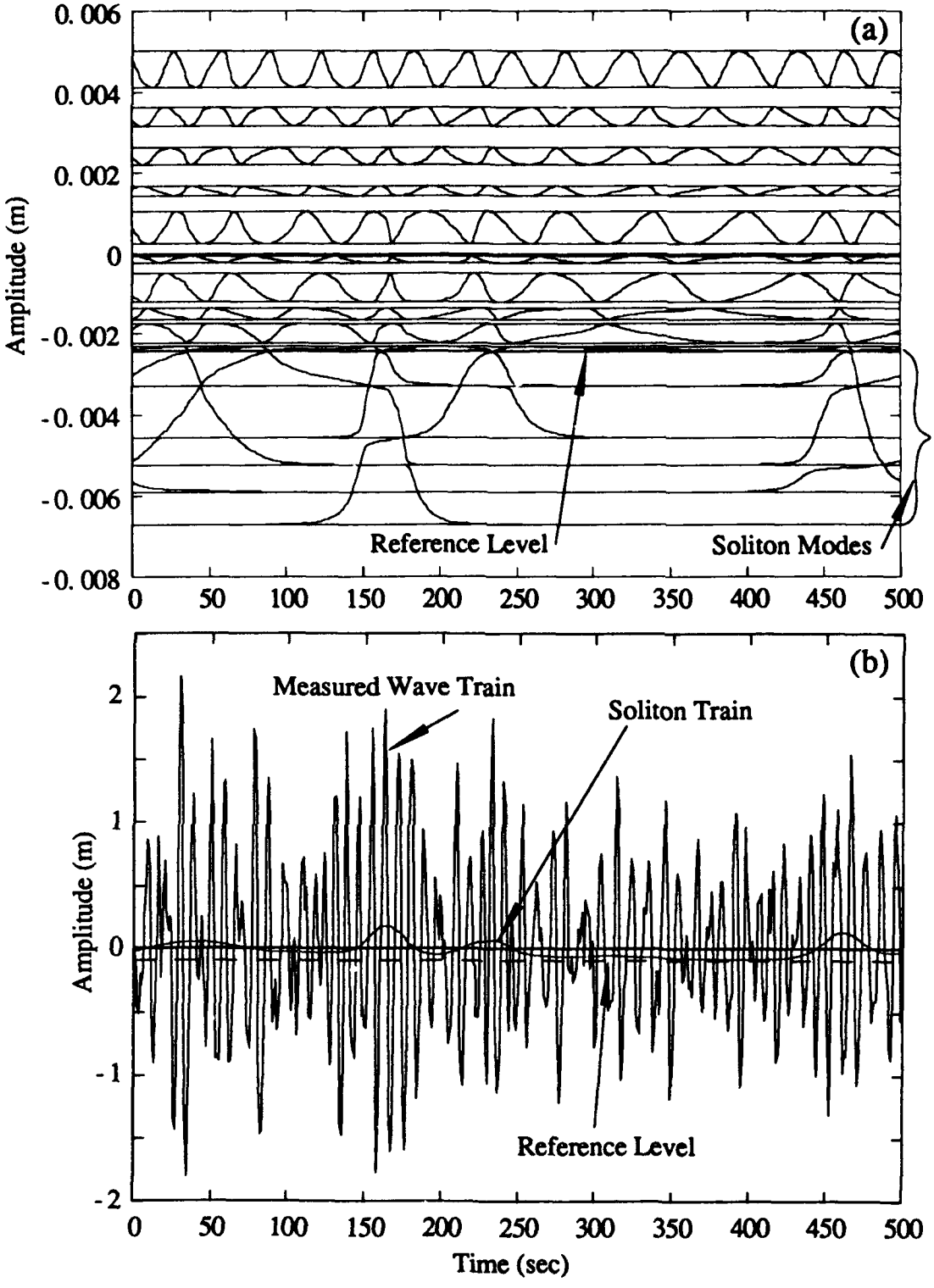


Figure 6. The nonlinear oscillation modes in the soliton part of the spectrum (a). In (b) the solitons are constructed by a linear superposition of these nonlinear modes. The solitons are seen to lie beneath the measured wave train and to propagate on a 'reference level' which lies below the mean water level.

soliton train is shown in Figure 6(b), where the original wave train is also superposed on the figure. As noted previously [Osborne et al., 1991] the soliton contribution to the wave train consists of a long, low-amplitude signal lying beneath the overlying, narrow-banded wave train, which is dominated by the radiation modes. Again I find that the solitons tend to lie beneath the maxima of the local wave groups; this topic is discussed in detail elsewhere [Osborne and Cavaleri, 1993]. It is impossible to stress how important the nonlinear filtering process is to the understanding of the soliton dynamics; I know of no other method for extracting them from an arbitrary oceanic wave train of the type studied here.

The most nonlinear of the *radiation* modes have also been filtered from the measured wave train. These results are shown in Figures 7 and 8. Figure 7(a) shows the hyper-elliptic modes centered near the peak of the spectrum, where the nonlinear spectral index is nearly one, in the frequency range 0.094–0.108 Hz. Figure 7(b) gives the modes over a somewhat larger frequency range extending from 0.094–0.122 Hz. These ranges are indicated on the nonlinear spectral index graphed in Figure 4(b). Scrutiny of the nonlinear modes in Figure 7 reveals that they are clearly not sinusoidal and that *phase locking* plays an important role in their dynamics (details are discussed in [Osborne and Cavaleri, 1993]). It is important to note the main differences in the nonlinear filtering process applied in the present paper and the usual one for linear Fourier analysis: (1) here I use the spectral index to select the most nonlinear parts of the spectrum to study and (2) the filtering process is fully nonlinear and often requires an iterative process [Osborne, 1993]. The regions that have a large spectral index are inverted to allow reconstruction and study of the wave trains in the most nonlinear parts of the IST spectrum; linear superposition of these modes give the wave trains shown in Figure 8(a, b). These wave trains are highly nonlinear and are not generally represented by the linear Fourier transform. For reference I also show the soliton part of the wave train, superposed on the nonlinear radiation modes in Figures 8(a, b). Figure 8(b) therefore represents the most nonlinear contributions (as seen in the time domain) to the measured wave train in Figure 2(a).

SUMMARY AND CONCLUSIONS

It is worth pointing out that in the originally measured wave train (Figure 2(a)), the soliton components are obscured by the radiation modes, e.g. solitons reside in the nonlinear spectrum, but they are not directly visible due to the presence of the energetic radiation components. The soliton dynamics are physically significant, but *not directly visible by an observer of the measured wave train.* Nevertheless, using the numerical methods described herein, we are able to locate the solitons and to explore their dynamics. This constitutes an exercise in nonlinear filtering. Returning to the spectrum in

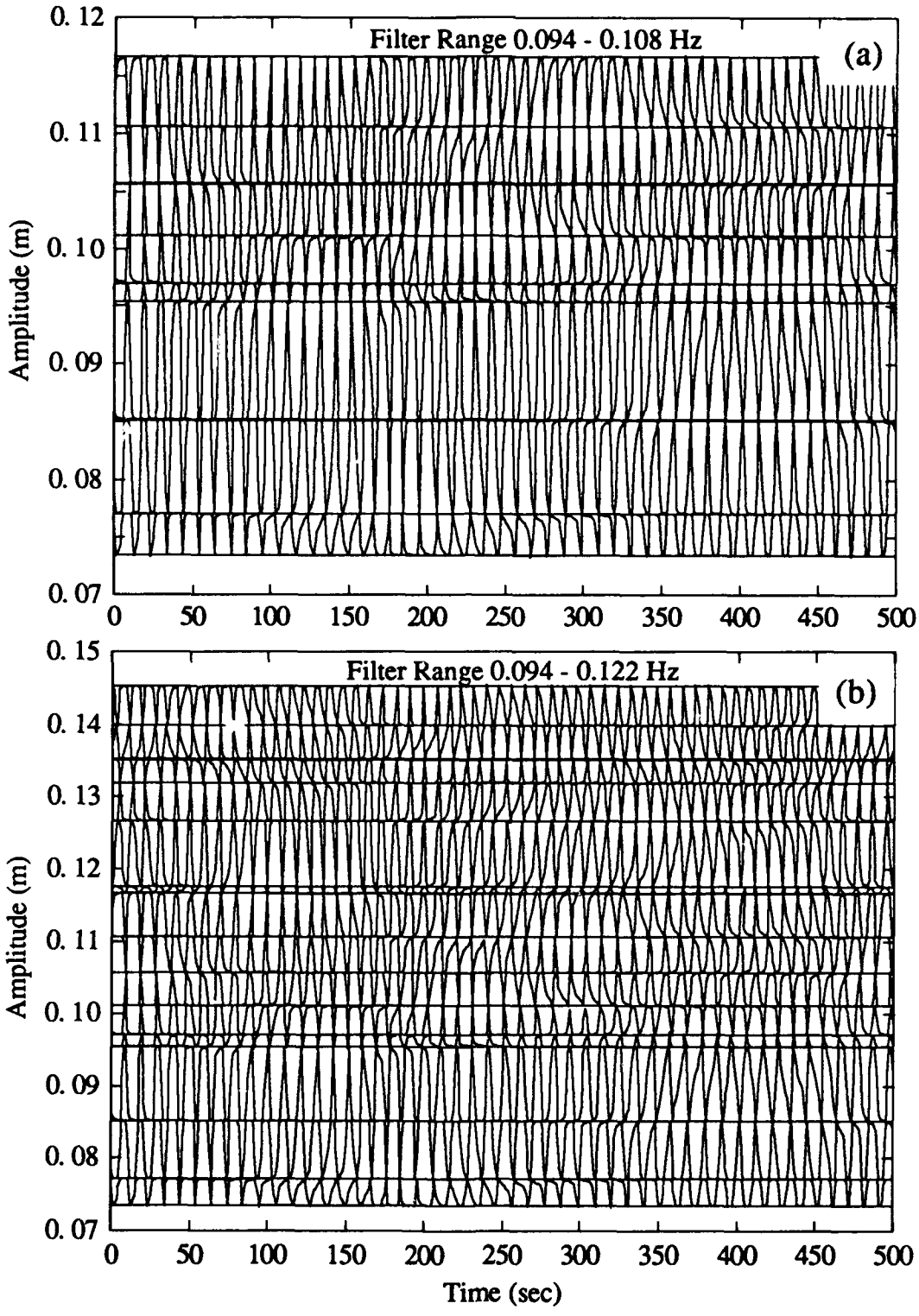


Figure 7. Hyperelliptic functions in the most nonlinear part of the radiation spectrum as indicated on Figure 4(b) and in the text (a). In (b) are the modes for an expanded region of the radiation spectrum, again defined in Figure 4(b).

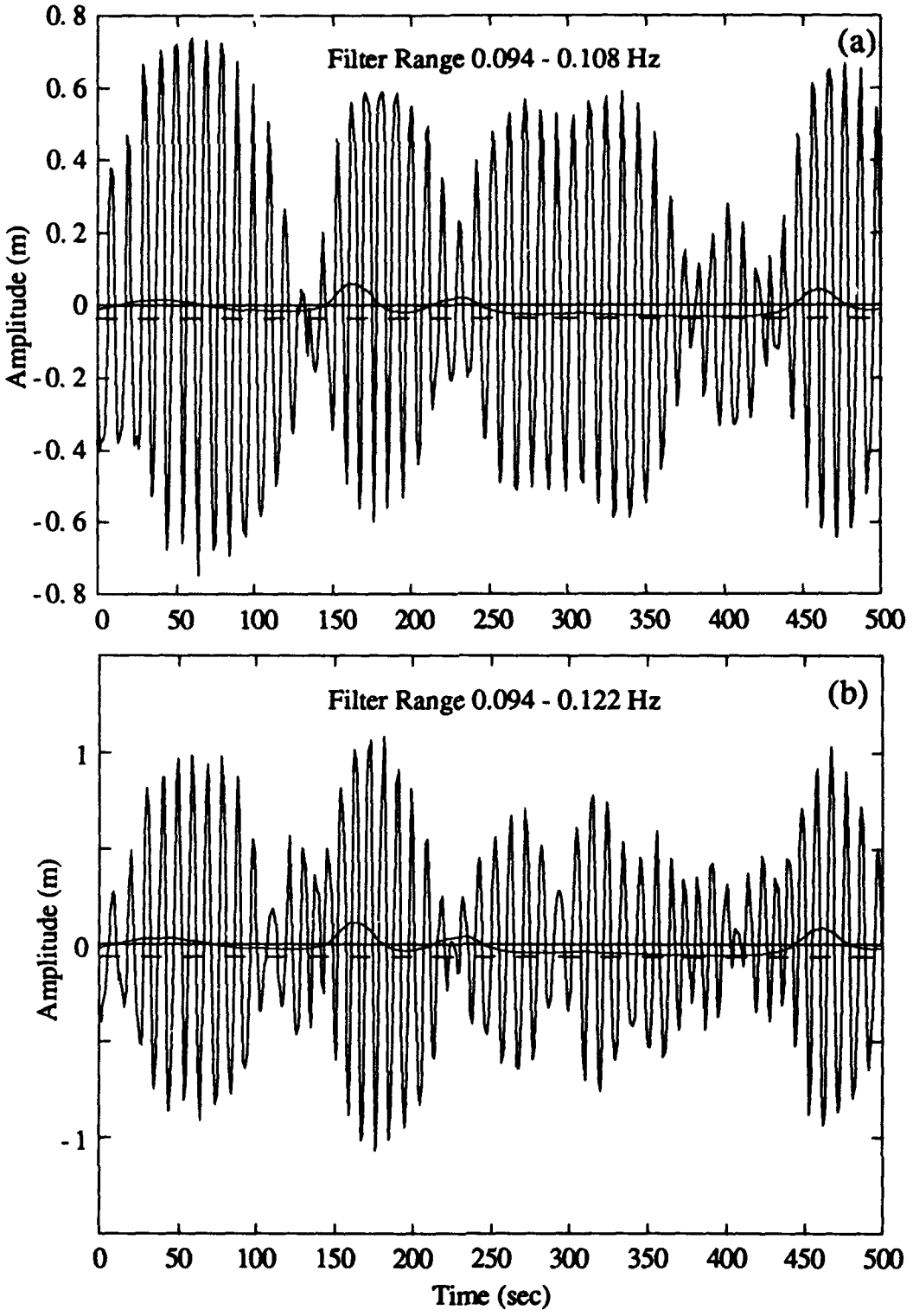


Figure 8. (a) Sum of the nonlinear oscillation modes in Figure 7(a). (b) The sum of the nonlinear modes shown in Figure 7(b). Both of these are examples of the application of nonlinear filtering by the inverse scattering transform as developed in this paper.

Figure 4(a) one can think of each component (as a function of frequency) as contributing to the nonlinear Fourier series (28). By deleting the terms corresponding to the radiation modes, and then summing the remaining terms for the soliton part of the spectrum, one obtains *only* the contributions that the solitons make to the measured nonlinear wave train. One finds a long, low-amplitude train, consisting of five nonlinearly interacting solitons. We have therefore, using the numerical inverse scattering transform as a tool for nonlinear filtering, found the solitons hidden in a sea of background radiation. An important physical result is that the solitons tend to be *phase locked* beneath the maxima of the wave packets. I am personally convinced that this fact provides an important clue to the eventual understanding of the behavior of nonlinear wave dynamics in the Ursell number regime under investigation. Theoretical understanding of these results is, however, still lacking.

Another result of the application of nonlinear filtering to the analysis of the Adriatic Sea data is that related to the construction of the nonlinear, narrow-banded wave trains in Figure 8. Since the nonlinear modes are clearly not sinusoidal (see Fig. 7), the effect of nonlinear interactions amongst these closely separated components is evidently rather important. These results are given here for the first time and are entirely new. A further surprising result is that the nonlinear spectral index can be near 1 for the radiation spectrum as well as for the solitons (Fig. 4(b)). Large nonlinear interactions in the radiation components is evidently another new result, yet to be fully explored. Complete understanding of the influence of these nonlinear effects on the physics of narrow-band wave motions, particularly with regard to phase locking, is a topic of future research.

Acknowledgments

L. Bergamasco is thanked for continued valuable encouragement and support. L. Cavaleri has provided friendship and consultation for over a decade. This work was supported in part by the Office of Naval Research of the United States of America (Grant N00014-92-J-1330) and by the Marine and Science Technology Program of the European Economic Community. We thank Alan Brandt and Peter Müller for the opportunity to participate in the Aha Huliko'a Hawaiian Winter Workshop. Phyllis Haines is thanked for invaluable aid and assistance.

REFERENCES

Ablowitz, M. J., and H. Segur, 1981: *Solitons and the Inverse Scattering Transform*, SIAM, Philadelphia.,

Bishop, A. R., M. G. Forest, D. W. McLaughlin and E. A. Overman II, 1986: A quasi-periodic route to chaos in a near integrable PDE, *Physica D, 18*,293-312.

Degasperis, A., 1991: Nonlinear Wave Equations Solvable by the Spectral Transform, in: *Nonlinear Topics in Ocean Physics*, A. R. Osborne ed., Elsevier, Amsterdam.

Dodd, R. K., J. E. Eilbeck, J. D. Gibbon and H. C. Morris, 1982: *Solitons and Nonlinear Wave Equations*, Academic Press, London.

Dubrovin, B. A., and S. P. Novikov, 1974: Periodic and conditionally periodic analogues of the many-soliton solutions of the Korteweg-deVries equation, *Sov. Phys. JETP 40*, 1058-1063.

Dubrovin, B. A., V. B. Matveev, and S. P. Novikov, 1976: Nonlinear equation of Korteweg-deVries type, finite-zone linear operators, and ABelian varieties, *Russian Math. Surv., 31*, 59-146.

Elgar, S. and R. T. Guza, 1986: *J. Fluid Mech. 167*, 1-26.

Flaschka, H., and D. W. McLaughlin 1976: Canonically conjugate variables for KdV and Toda lattice under periodic boundary conditions, *Prog. Theoret. Phys., 55*, 438-456.

Flesch, R., M. G. Forest, and A. Sinha, 1991: *Physica D 48*, 169-208.

Karpman, V. I., 1974: *Non-Linear Waves in Dispersive Media*, Pergamon, Oxford.

McKean, H. P. and E. Trubowitz, 1976: Hill's operator and hyperelliptic function theory in the presence of infinitely many branch points, *Comm. Pure Appl. Math. 29*, 143-226.

McLaughlin, D. W., C. M. Schober, 1992: Chaotic and homoclinic behavior for numerical discretizations of the nonlinear Schroedinger equation, *Physica D 57*, 447-465.

Miles, J. W., 1980: Solitary waves, *Ann. Rev. Fluid Mech. 12*, 11-43.

Newell, A. C., 1985: *Solitons in Mathematics and Physics*, SIAM, Philadelphia.

Osborne A. R., 1983: The Spectral Transform: Methods for the Fourier analysis of nonlinear wave data, *in* "Statics and Dynamics of Nonlinear Systems," G. Benedek, H. Bilz and R. Zeyher eds., Springer-Verlag, Berlin, 121-133.

Osborne, A. R., 1991a: Nonlinear Fourier analysis, *in*: "Nonlinear Topics in Ocean Physics," A. R. Osborne ed., North-Holland, Amsterdam, 669-700.

Osborne, A. R., 1991b: Nonlinear Fourier analysis for the infinit-interval Korteweg-deVries Equation I: An algorithm for the direct scattering transform, *J. Comp. Phys., 94*, No. 2, 284-313.

Osborne, A. R., 1993: in preparation.

Osborne, A. R., and L. Bergamasco, 1985: The small-amplitude limit of the spectral transform for the periodic Korteweg-deVries equation, *Nuovo Cimento B, 85*, 229-243.

Osborne, A. R., and L. Bergamasco, 1986: The small-amplitude limit of the spectral transform for the periodic Korteweg-deVries equation, *Physica D, 18*, 243.

Osborne, A. R. and L. Cavaleri, 1993: in preparation.

Osborne, A. R., and E. Segre, 1990: Numerical solutions of the Korteweg-deVries equation using the periodic scattering transform m-representation, *Physica D, 44*, 575-604.

Osborne, A. R., E. Segre, G. Boffetta, and L. Cavaleri, 1991: Soliton basis states in shallow-water ocean surface waves, *Phys. Rev. Lett. 64*, No. 15, 1733-595.

Terrones, G., D. W. McLaughlin, E. A. Overman II and A. Pearlstein, 1990: Stability and bifurcation of spatially coherent solutions of the damped driven nonlinear Schroedinger equation, *SIAM J. Appl. Math. 50*, 791-818.

Whitham, G. B., 1974: *Linear and Nonlinear Waves*, Wiley-Interscience, New York.

Zakharov, V. E., S. V. Manakov, S. P. Novikov, and M. P. Pitayevsky, 1980: *Theory of Solitons. The Method of the Inverse Scattering Problem*, Nauka, Moscow (in Russian).

PRINCIPAL COMPONENT ANALYSIS: BASIC METHODS AND EXTENSIONS

Gary T. Mitchum

Department of Oceanography, University of Hawaii at Manoa
1000 Pope Road, MSB 307, Honolulu, HI 96822

ABSTRACT

The basic method of principal component analysis is relatively well understood by physical oceanographers. Some less generally understood ideas involve significance testing and rotation of the basis functions. Also, a number of other analysis techniques related to principal component analysis can be more easily understood by using it as a starting point. Such techniques include factor analysis, extended empirical orthogonal function analysis, canonical correlation analysis, and complex empirical orthogonal function analysis, for example.

The basic calculations comprising principal component analysis are presented, and significance testing and rotation are discussed. Pacific sea level data are used to illustrate these techniques. The paper concludes with a discussion of various extensions to the basic technique and an evaluation of the usefulness of the extensions.

INTRODUCTION

It is important to establish at the outset that this paper is not intended to be a comprehensive review of principal component analysis (PCA) or its applications. It is similarly not intended as a detailed review of the other techniques that will be discussed as related to PCA, or as extensions of it. Rather, the intention of this paper is to briefly review the PCA technique in order to establish a common frame of reference and to then point out how several other commonly used techniques can be viewed as applications of PCA to a more general dataset. The reason for doing this is to place all these techniques in a sensible framework, to point out where they overlap, and to give viewpoints, my own and others, as to the relative merits of the various techniques. I have not avoided giving the results of my own experience with the various techniques, but I have tried to clearly identify my opinions in order that the reader can decide what should be ignored.

For the reader interested in a more comprehensive treatment of PCA, and of factor analysis (FA), which is much more commonly used in some fields other than oceanog-

raphy and meteorology, several books are recommended. Preisendorfer (1988) provides an extensive bibliography of applications and source material, and also gives additional detail on nearly everything discussed in this paper. The book by Harman (1976) on FA, while somewhat dated and primarily written from the point of view of workers in psychology, is an excellent source for insight about rotation methods. A book written for geologists (Joreskog *et al.*, 1976) is also well done and appears to be commonly used by oceanographers. Finally, a very recent book by Reyment and Joreskog (1993), which I have not yet seen, is noteworthy because of the inclusion of an appendix that includes a set of electronically available routines for the MATLAB programming environment.

The organization of the paper is as follows. The first section describes the basic formulation of PCA. This section includes a comparison, due to Preisendorfer (1988), of FA and linear regression analysis (LRA) that helps to illustrate why the technique is so powerful and widely useful. This section continues with a discussion of the problem of significance testing and the technique of rotation. Both of these latter topics should be understood by any user of PCA. Throughout this section, examples are given using a Pacific monthly mean sea level anomaly dataset. All of the discussion in this section deals with the PCA of a scalar-valued dataset consisting of time series at a set of stations. The following section treats the extension of PCA to the analysis of vector-valued data and to the analysis of propagating signals. Finally, I will examine the relationship of PCA to canonical correlation analysis (CCA), which is used for the simultaneous analysis of more than one data field.

THE BASICS OF PRINCIPAL COMPONENT ANALYSIS

Basic Computations

We will consider first a very straightforward application of PCA to a set of time series collected at a set of stations. As an example of this I will use sea level time series collected at 46 stations in the Pacific Ocean (Figure 1). The temporal means are removed from the time series at each station, which are also corrected for atmospheric pressure and the mean seasonal cycle. The PCA model for this dataset can be written

$$h(s,t) = \sum_{k=1}^{K} a_k(t) e_k(s) \tag{1}$$

where s is the station index, t is the time, and K is equal to the number of stations.

The variance-covariance matrix for the dataset $h(s,t)$, which I will refer to simply as the variance matrix, is a special case of what Preisendorfer (1988) calls the scatter matrix, and is written

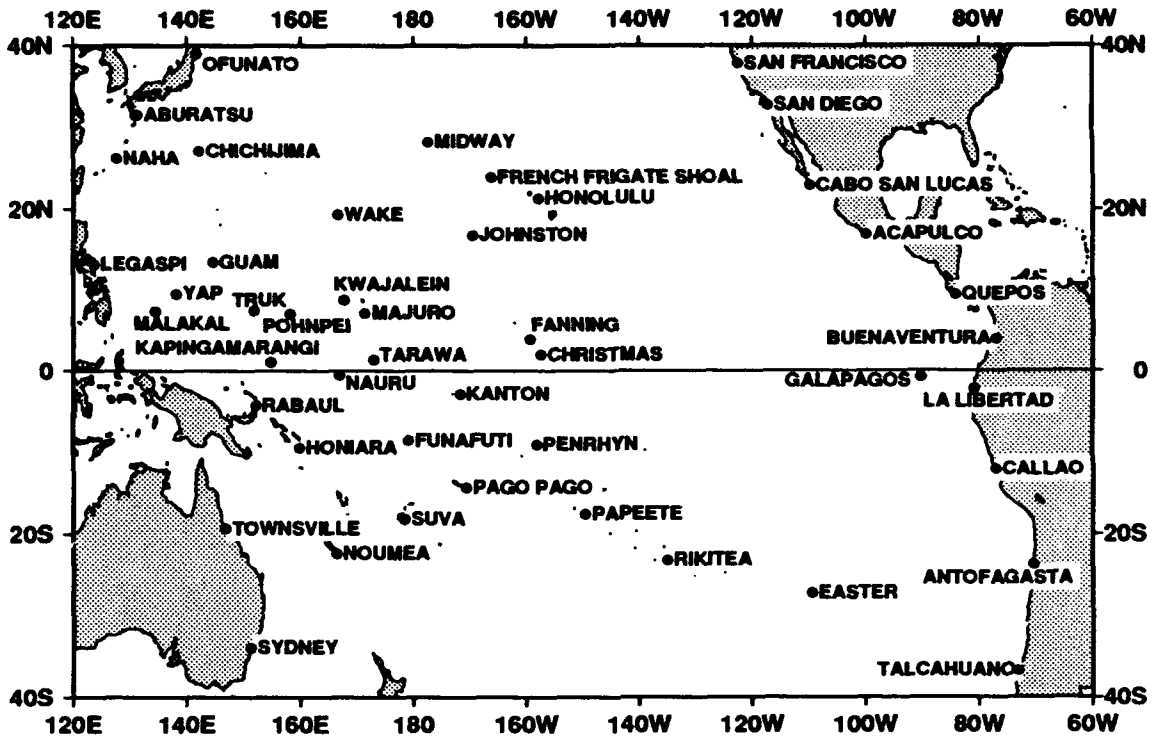


Figure 1: Pacific sea level stations used in the PCA example. There are 46 stations, each with a monthly mean time series spanning 1975 to 1990. The monthly mean values are corrected for atmospheric pressure and the mean seasonal cycle before the PCA is performed. Data gaps are interpolated.

$$S(s,s') = \sum_t h(s,t)h(s',t) \tag{2}$$

Since the variance matrix is real-valued and symmetric, it has real eigenvalues and eigenvectors. The eigenvalues, which are generally sorted into decreasing order, give the amount of variance in the original dataset that is accounted for by the associated eigenvector and its time history function. The eigenvectors are mutually orthogonal and form the basis set for the expansion shown in Eq. (1). The associated time history functions, which are also mutually orthogonal, are computed from the original data and the eigenvectors as

$$a_k(t) = \sum_{m=1}^{K} h(s,t)e_m(s) \tag{3}$$

where I have assumed that the eigenvectors, $e_m(s)$, are normalized to unit variance, and the time history functions, $a_k(t)$, are allowed to carry the variance.

Figure 2 shows the results of applying the PCA technique to the Pacific sea level dataset. The eigenvectors associated with the two largest eigenvalues are interpreted as space maps, and the analogous time history functions can be interpreted as modulating the space maps and indicating when that particular space map's pattern is strongly expressed in the original dataset. In this particular case the two functions are both associated with the El Niño/Southern Oscillation (ENSO) events in the tropical Pacific. The space maps show this clearly. The time history functions, however, are somewhat nondescript, but this will be discussed more later.

Relationship to Linear Regression Analysis and Factor Analysis

In order to better understand what the PCA expansion defined in Eq. (1) does, it is instructive to compare it to a linear regression analysis (LRA) and a factor analysis (FA). If we truncate the PCA and FA expansions (criteria for doing this are discussed in the next section), then these various expansions can be written

$$h(s,t) = \sum_{m=1}^{M} a_k(t)e_k(s) + \delta(s,t) \qquad \text{PCA} \qquad (4a)$$

$$h(s,t) = \sum_{m=1}^{M} \varphi_k(t)\beta_k(s) + \varepsilon(s,t) \qquad \text{LRA} \qquad (4b)$$

$$h(s,t) = \sum_{m=1}^{M} f_k(t)\lambda_k(s) + \nu(s,t) \qquad \text{FA} \qquad (4c)$$

In these expansions, e_k, β_k, and λ_k are thought of in the present context as the basis functions; a_k, φ_k, and f_k are the amplitude functions; and δ_k, ε_k, and ν_k are the residual noise terms.

These expansions look very similar, but actually there are quite different underlying assumptions. For the LRA, the basis functions are specified *a priori*, and the time history functions are fit, typically by a least squares criterion, in order that the defined basis functions account for the maximum amount of variance. In the PCA analysis, the data are allowed to choose their own basis set under the criterion that each function must explain the maximum variance, subject to the additional constraint that each succeeding function be orthogonal to all the preceding ones. The truncated PCA expansion is therefore maximally efficient in accounting for the variance of the original dataset with the fewest basis functions, but there is a cost. Namely, there is no guarantee that the eigenfunctions correspond to any physically meaningful modes of variability of the original data. LRA, on the other hand, can be constructed using basis functions that are defined from *a priori* knowledge of the dynamics of the system being studied.

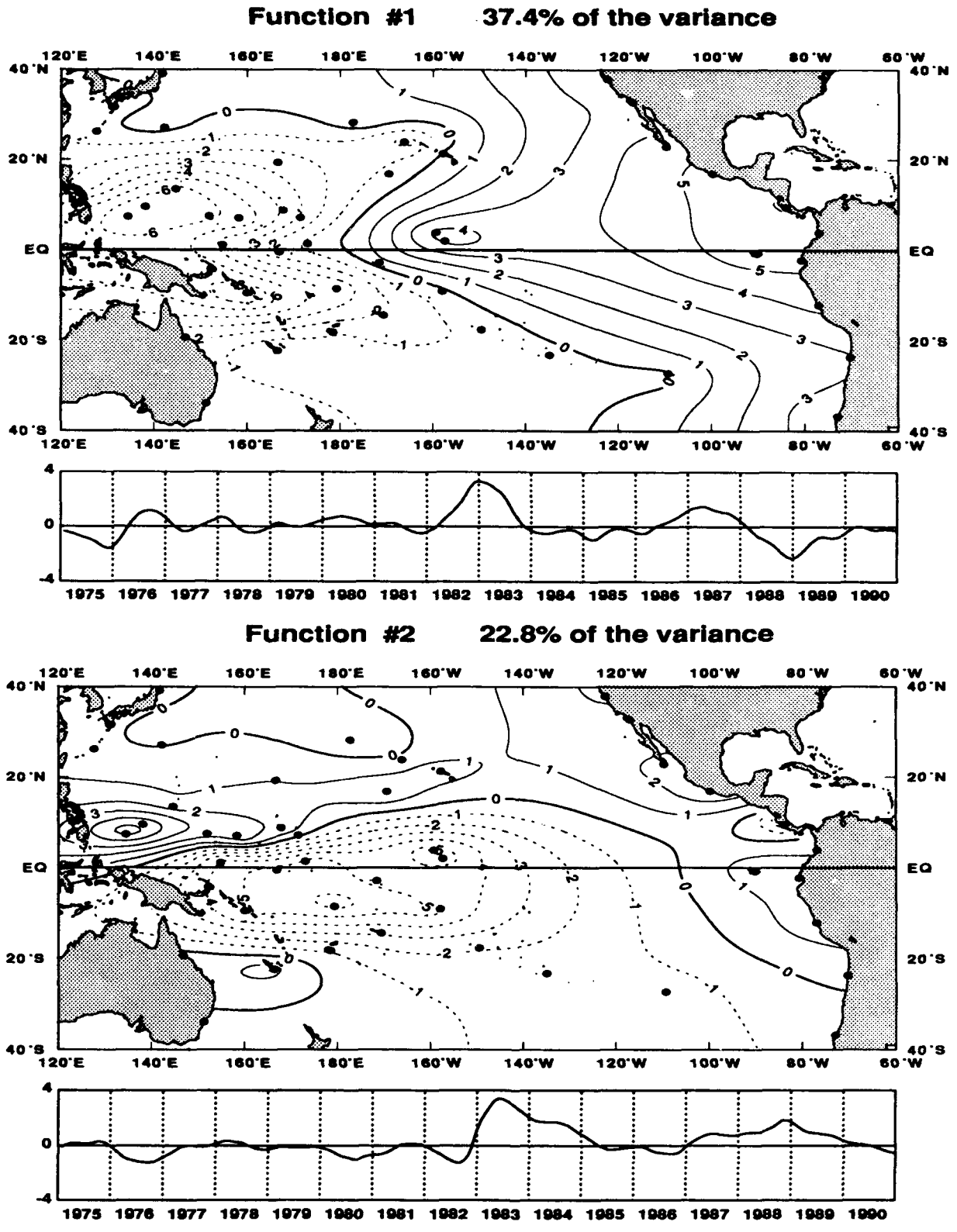


Figure 2: Results of PCA of Pacific sea level dataset. Space map contour units are 1 cm, and negative contours are dashed. The space map values must be multipled by the time function plotted below it in order to obtain values comparable to the observations.

The strengths and weaknesses of FA as compared to LRA are similar to those of PCA. However, when compared to PCA, the FA case is more subtle. Basically, while PCA is a completely objective technique that needs only the original dataset to proceed, the FA treats the number of factors used in the expansion and the residual noise term, $v_k(s,t)$, as unknowns, and thus is an underdetermined system. Many suggestions exist for ways to close the system (Harman, 1976), but the technique remains somewhat subjective. It requires specification of *a priori* information that is usually not trivial to provide. Preisendorfer (1988) discusses these issues at some length and claims that FA is the "conceptually deeper" of the two techniques. I have never been able to convince myself that the additional subjectivity associated with FA provides much advantage over PCA.

Significance testing

The full PCA expansion defined by Eq. (1) has the exact same information content as the original dataset. It is rare, however, that the original data are free of noise, and one must therefore assume that many, if not most, of the PCA functions simply represent noise. The question naturally arises, then, of how to select the functions that may represent signal, in order that they can be further analyzed. Preisendorfer (1988) is particularly good on this topic of selection rules for PCA, and the interested reader should consult that text on this topic. Harman (1976) provides an interesting historical perspective on the development of older selection rules that have largely been superseded.

The basic idea behind all of the selection rules is quite simple. The functions are compared to those that would result from data drawn from a specific noise model, and those that are not consistent with such noise data are deemed worthy of further study as signals with possible geophysical significance. The selection rules discussed by Preisendorfer (1988) fall into three broad categories: variance dominant, time history, and space map rules. *N.B.*, the use of the words "time" and "space" are simply convenient and do not restrict the application of these techniques to time series data at stations, such as that used in my Pacific sea level example.

The variance dominant rules are probably the most commonly used selection rules in oceanography and meteorology. Recall that the eigenvalues of the scatter matrix are equal to the amount of variance of the original dataset accounted for by the associated eigenvector (space map), and amplitude function (time history). Figure 3 shows the 46 eigenvalues obtained for the Pacific sea level dataset after placing them in decreasing order. Also shown on this figure are the eigenvalue curves obtained by the application of three different variance dominant selection rules. The first, labeled Rule N, is based on the assumption that the original time series are simply noise, which is uncorrelated from station to station. To apply this rule, 100 such datasets are generated and the eigenvalues are computed and placed in descending order. Then the 95th percentile is found and plotted. Eigenvalues from the actual dataset that fall above this curve are deemed as unlikely to result from a dataset consisting of only noise. In the case of the Pacific sea level dataset, only the first two functions are thus selected.

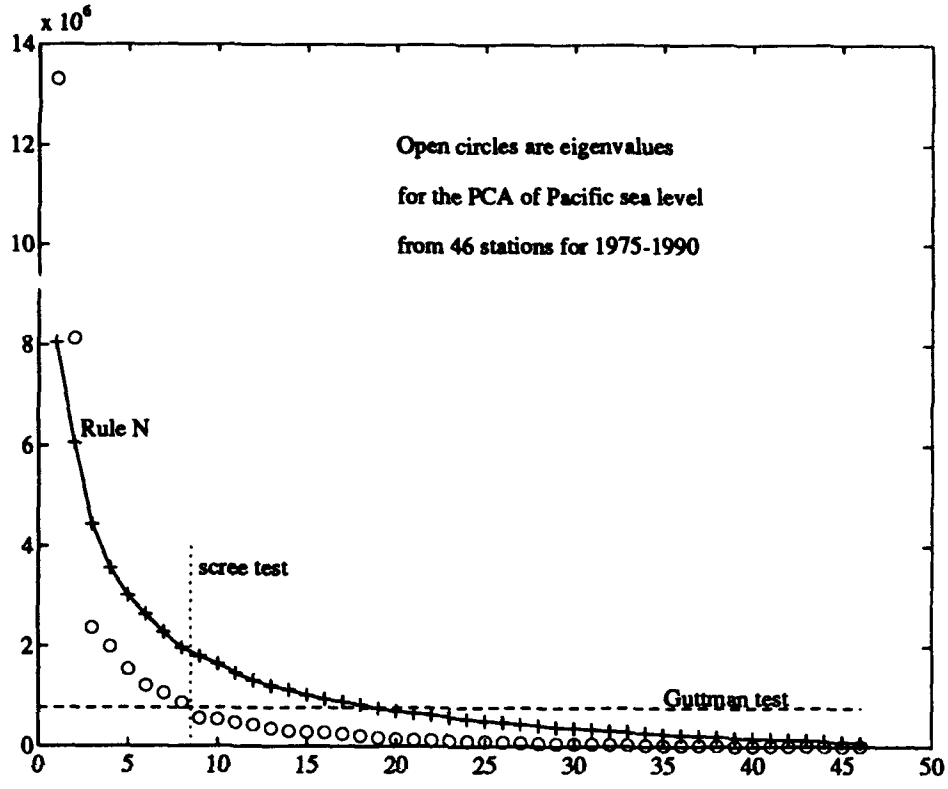


Figure 3: The 46 eigenvalues from the PCA of the Pacific sea level dataset are shown as open circles. Also shown are the results of applying three difference selection rules (see text for details).

One complication worth considering in the application of Rule N to geophysical data is the fact that most time series of such data are not well-modeled by a white noise model, but have significant serial correlation due to oversampling. One method that I have used to deal with this problem is to define a noise model that consists of "red" noise. I characterize the model according to the approximate spectral slope obtained from a Fourier analysis of the noise series. Figure 4 shows the result of applying Rule N to the Pacific sea level dataset using several values for the spectral slope. Clearly, quite different conclusions about the significance of the low order functions would be reached depending on which model is chosen. An f^{-1} noise model is appropriate for the Pacific sea level dataset, and this was in fact the noise model used in generating the Rule N curve in Figure 3.

Variance dominant criteria other than Rule N are discussed by Preisendorfer (1988), but Rule N appears to be the most widely used selection rule of this class. Two other selection rules are shown on Figure 3 that are of historical interest. The scree test is a subjective test that requires the analyst to select the point on the eigenvalue curve where the curve begins to tend up more sharply at lower orders. This is an experientially based test.

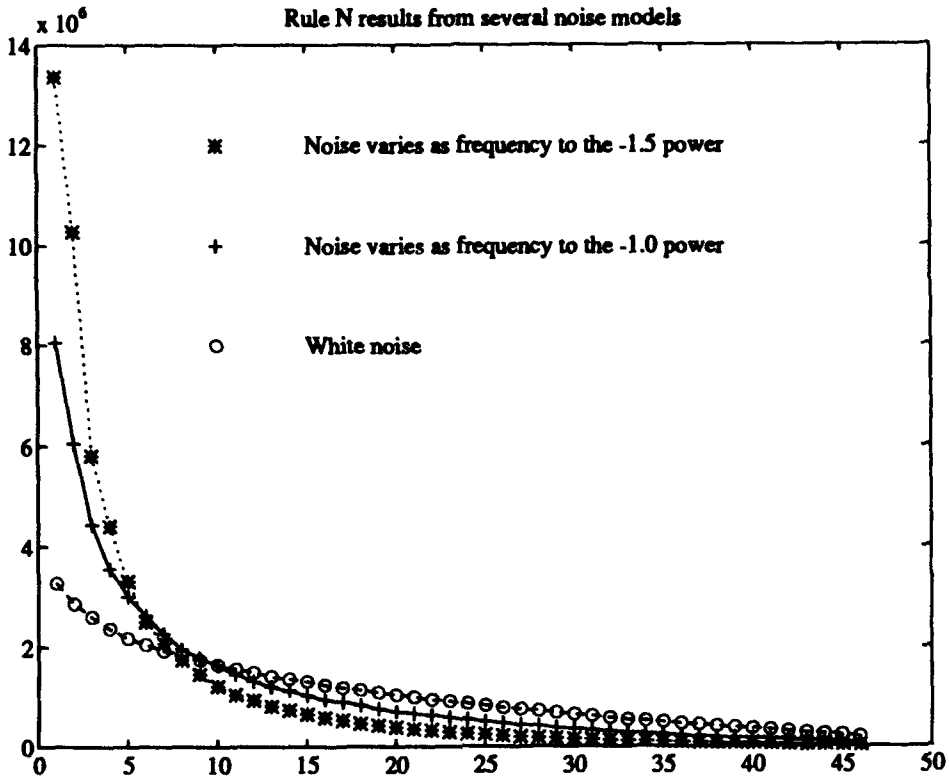


Figure 4: Rule N applied to Pacific sea level data using various models for the noise.

The second test, labeled the "Guttman" test, is objective. This test computes the average eigenvalue and defines any that are larger than this value to be of possible interest. The rationale here is simply that the selected functions explain more than an average amount of variance. These latter two tests tend to be less conservative than Rule N, at least when the 95th percentile is used in that test. These less conservative tests have some value in the rotation problem that is discussed in the next section.

Time history and space map rules are less commonly used than variance dominant rules, but this is probably due to the simplicity of the the latter rather than to any inherent advantage in them. The time history rules work by examining the time history function and testing it for low frequency variability. Several ways of doing this are described by Preisendorfer (1988). The space map rules are similar, but work on the eigenfunctions (space maps) and look for coherent "spatial" patterns. These rules should probably be more widely used, particularly since they should be able to detect functions that map to a geophysical signal that does not dominate the variance of the dataset but can be identified by coherent temporal or spatial patterns. An example is seen in Barnett (1977).

Another area of concern for significance testing is that of placing confidence limits on the time series and space maps themselves. For example, how large do signals in the time series or space maps shown in Figure 2 have to be before there is reasonable confidence that they represent real modes of variation in the actual data? Unfortunately, the variance dominant rules used to select the functions cannot answer such questions. To the best of my knowledge, this is an unsolved problem, although there has been some progress in this area using bootstrap techniques (D. Chelton, pers. comm.).

Rotation

If one thinks of the eigenvectors as a set of orthogonal basis vectors for expanding the original dataset, then it is easy to imagine geometrically "rotating" this basis set. Another way to view this rotation is to imagine replacing the original set of basis vectors with a set that consists of linear combinations of the original set. In the rotated frame, the basis vectors can still be orthogonal, but the amplitude functions will now be correlated.

But why would one want to rotate a perfectly good set of basis functions anyway? If the primary purpose of the PCA is to compress the original dataset into a few functions that still capture a large portion of the variability of the data, then there is no need for rotation. The PCA frame is, by construction, the most efficient description of the variance possible. If one desires to interpret the individual functions in physical terms, however, then this efficiency can be a problem.

Imagine that the original data consist of a number of distinct, but not necessarily orthogonal and unrelated, modes of variation. In order to explain the most variance possible, the PCA technique will return linear combinations of these modes - not the modes themselves. The hope of the rotation technique is that by relaxing the requirement that the basis functions are maximally efficient at explaining variance, then it may be possible to obtain modes that more closely resemble the natural modes of the dataset. In fact, it is often claimed that non-rotated PCA frames should not be interpreted physically at all.

Harman (1976) is the best reference I have found for discussion of rotations, although it is written in the context of FA. There are a large number of rotation techniques, which can be separated into orthogonal and oblique rotations. Both sets of rotations relax the requirement that the functions be maximally efficient at explaining variance, but orthogonal rotations preserve the orthogonality between the basis vectors while oblique rotations do not require even this. Orthogonal rotations in general, and varimax specifically, are most common and are much simpler to perform, and also to interpret, than oblique rotations. My experience has been that rotation should always be done before attempting to interpret the functions, but that little is gained by going beyond simple varimax rotation. A contrary opinion and example is given by Richman (1981).

Orthogonal rotations in general work by searching for a rotated frame that minimizes the number of basis functions that any particular time series in the original dataset projects to. To say this another way, the technique seeks a rotated frame where any

station's time series from the original time series projects to the new basis functions in such a way that the projections are near 0 or 1. Preisendorfer (1988; pg. 271-274) gives an excellent graphical explanation of why this criterion is appropriate; the purpose here is simply to document how the basic algorithm operates. In doing the optimization, there is a penalty function that increases when a projection is "far" from 0 or 1, and the exact form of this penalty function defines numerous rotation schemes, of which varimax is probably the most common. A review of many others is given by Harman (1976).

Figure 5 shows the result of applying a varimax rotation to the first 10 functions from the PCA of the Pacific sea level dataset. The reason for using 10 functions is that results for low order rotated functions are unstable if too few functions are rotated, but are relatively insensitive to adding in a few extra functions that represent only noise. This conclusion is stated by Harman (1976), and my experience bears it out. The scree and Guttman tests shown in Figure 2 are often useful indicators of the maximum number of possibly interesting functions, and I have found them useful as a guide to choosing the number of functions to include in the rotation procedure.

The maps and time series shown in Figure 5 account for almost exactly the same amount of variance (50%) from the total dataset as the first two unrotated functions. These functions are much simpler to interpret, however. Examination of the maps and time series shows that the first and second functions, respectively, map to a western Pacific ENSO response that is primarily north and south of the equator. The associated time series show that the northern pattern occurs in the late part of the calendar year, while the southern pattern is associated with a timing that is several months later. Particularly interesting is the fact that the various events in the records map on to these two modes differently; only the 1982-83 event shows a strong expression of both types of events. Thus, it seems that the ENSO events tend to be one of the two types: a "northern" type that sees mass lost primarily from north of the equator in the western Pacific late in the calendar year, or a "southern" type where the mass comes from south of the equator during the early part of the calendar year. Interpreting these two signals as simply the beginning and ending stages of the same event is not quite satisfying, since in that view most events either do not have a beginning or do not have an end.

EXTENSIONS TO PRINCIPAL COMPONENT ANALYSIS

Vector data

As developed in the preceding section, PCA works for scalar data via the variance, or scatter, matrix. In fact, this restriction to scalar data is unnecessary, and vector data can be treated as long as an appropriate scatter matrix can be defined. An appropriate scatter matrix is one that properly represents the variability characteristics of the dataset being studied, and one that has a full set of eigenvalues and eigenvectors. Note that there can be

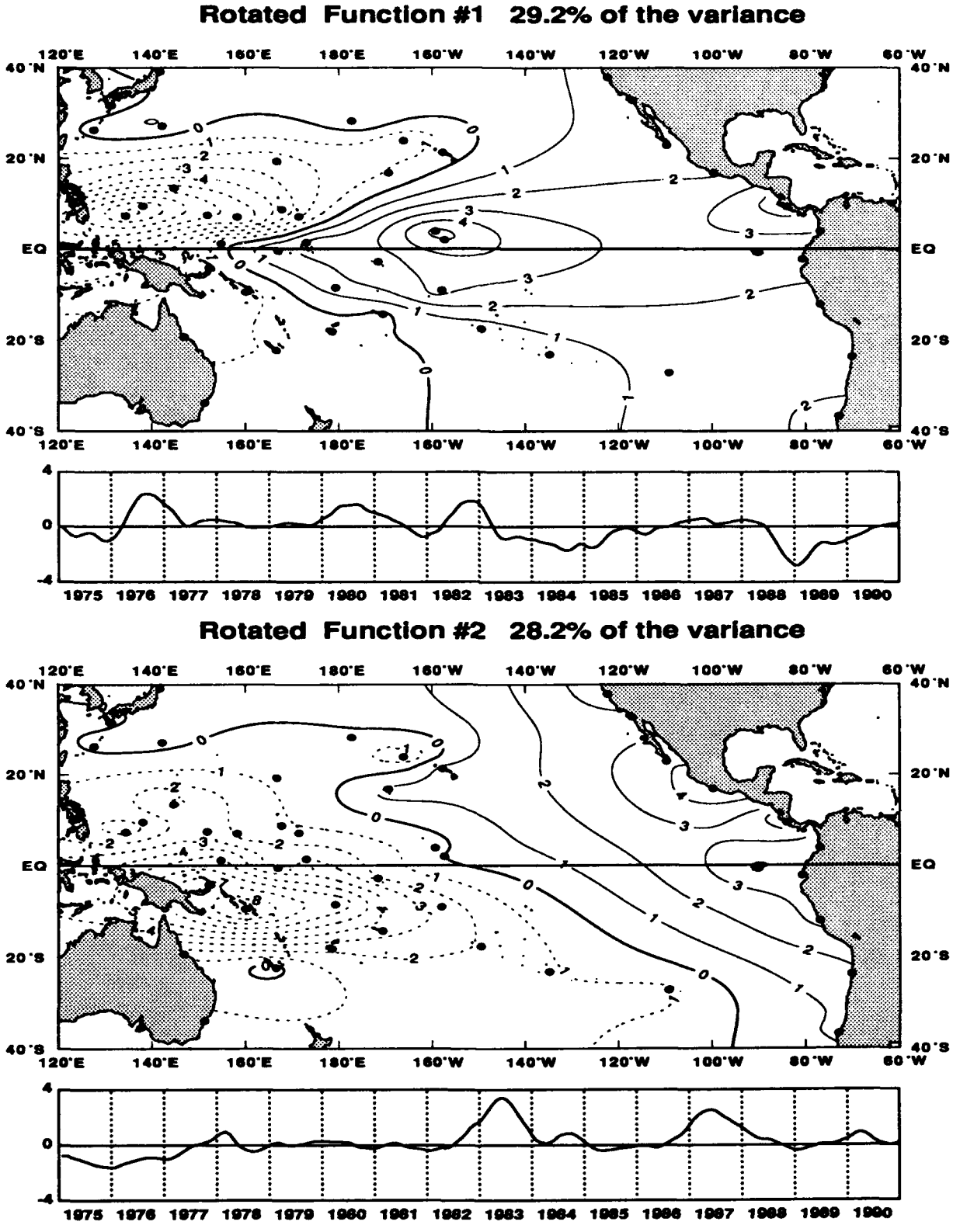


Figure 5: As in Figure 2, but for the rotated functions.

eigenvalues with multiplicity greater than one, although this does not generally happen in datasets containing noise.

An early application of PCA to a vector-valued field was made by Barnett (1977), who analyzed monthly mean surface wind vectors over the Pacific Ocean. His analysis was not truly a vector analysis, however, in the sense that he separately analyzed the zonal and meridional wind components without taking the vector nature of the data into account. Legler (1983) treated basically the same dataset in a truly vectorial way by defining the wind vectors as a complex quantity and then defining the scatter matrix by using complex conjugates to form the variance matrix.

Preisendorfer (1988) points out that the vector analysis can also be done by simply defining each of the wind components at each station as a separate variable in a normal (scalar) PCA. In this case, if there are M stations with N time points, then the scatter matrix is 2M by 2M, and there are 2M eigenvectors, each of which has a real-valued N-point time history function associated with it. He goes further to argue that this analysis is *not* equivalent to the complex method, but contains some additional information on the complex phase that is lost in forming the scatter matrix with complex conjugation. In fact, I find that this technique is also simpler in practice than the complex one, in that the same routines developed for scalar analysis apply to this problem as well.

Propagating signals

One problem with the basic PCA of space-time data is that the resulting functions do not properly represent propagating signals. This is because the dataset is described by a set of space maps modulated by separate time series. This is a serious drawback in oceanography and meteorology where propagating signals are usually present, and are often the features of primary interest to the analyst. A number of techniques have been developed that extend the basic PCA to the case of signals that cannot be represented by separable functions of space and time.

An early development was the method usually referred to as frequency domain empirical orthogonal functions (FDEOFs). The basic references for this technique are Wallace and Dickinson (1972) and Wallace (1972). Basically, this procedure starts by transforming the time series at each point in space into the frequency domain. The resulting complex spectrum are averaged over a frequency band of interest, and the resulting space map of complex numbers is analyzed via a complex form of PCA. The result of this analysis is a map of amplitude and phase that can be analyzed for phase propagation signatures. This method has not been widely used and in my experience is not overly successful at identifying signals that are not readily apparent in the original data.

An improved technique, referred to as complex empirical orthogonal functions (CEOFs), was described by Barnett (1983). The results of this analysis are somewhat similar to the output of FDEOFs, but the calculations are simplified by the use of a Hilbert transform on the original time series, which builds in the phase information neces-

sary to identify propagating signals. The CEOF technique is more general than the FDEOFs, and should be superior to that method at identifying features that are distinctly non-sinusoidal in nature. Although computationally simpler than FDEOFs, I have found in my own work that the CEOFs are similarly difficult to interpret in most applications.

Probably the most widely used technique for describing propagating signals is the extended empirical orthogonal function method (Barnett and Hasselmann, 1979; Weare and Nasstrom, 1982). This technique builds in the phase information by "extending" the analysis to include not only the original dataset, but also the same dataset at a variety of temporal lags. These lagged time series are simply input as additional variables, and the normal machinery for basic PCA therefore applies. With the extended dataset it is possible to identify patterns at one time that have high correlations with patterns at a later time. A good example of the application of EEOFs is given by White and Tai (1992). One advantage of this method is that the signals can deform in space and time in fairly general ways without being lost to the technique. I have found this technique to be very useful in a number of different contexts.

Another technique that can identify propagating disturbances is the principal oscillation pattern analysis. I will not discuss this technique, but will refer the interested reader to the paper by von Storch in this same volume.

Canonical Correlation Analysis

In all of the discussion preceding I have dealt only with datasets consisting of one data type; e.g., sea level or wind vectors. In fact, if the data are appropriately non-dimensionalized, then there is nothing to prevent data with different units from being included in the PCA. This procedure is often useful, but only identifies the major modes of variability of the datasets. It does not identify the patterns of variability in the different datasets that are related, or co-varying. There is, however, a technique related to PCA that looks for these types of relationships; this technique is called canonical correlation analysis (CCA), and in the past few years it is being more widely used in oceanography.

Preisendorfer (1988) shows how the PCA description of two different datasets can be used to derive CCA, although the original derivation of CCA, which he attributes to Hotelling (1936), did not actually make use of this machinery. To drastically oversimply, the time history functions from the PCA for each of the datasets can be used to form a correlation matrix. This matrix is then used to form an eigenvalue problem that leads to the canonical correlation functions. The first of these functions can be interpreted as the pattern in one dataset that is maximally correlated with the corresponding pattern in the other dataset. Then the second function reveals the patterns that give the highest correlation between the datasets after removing the correlation due to the first canonical correlate, and so on. As with PCA, there are selection rules to be applied to determine whether the CCA results could arise from data consisting simply of noise.

Aside from the fact that CCA can be derived via PCA, it is considered a related technique because it can be viewed as a natural extension of PCA. PCA describes the variance structure of a dataset, or datasets, while CCA describes the covariance between two datasets. Some useful examples of oceanographic and meteorological applications of this technique are given by Barnett and Preisendorfer (1987) and Graham *et al.* (1987).

CONCLUSIONS

The basic calculations involved in PCA, which are probably familar to most oceanographers, were reviewed. Methods for testing the significance of the PCA functions were discussed, and it was suggested that, in addition to the common use of variance dominant selection rules (e.g., Rule N), more use should probably be made of time history and space map selection rules. Also, the technique of factor rotation was discussed briefly, with the conclusion that rotation should be an important part of any attempt to physically interpret the results of a PCA. Orthogonal rotations, such as varimax, are likely sufficient for most applications.

Extensions of the basic PCA technique were discussed that allow the analysis of vector-valued datasets, as well as datasets containing signals that propagate in space-time. My experience is that the most general procedure for dealing with vector data, described by Preisendorfer (1988), is also the simplest to apply. Similarly, for the analysis of propagating signals, the EEOF method is also the simplest to use and manages to perform at least as well as the more complicated frequency domain and complex techniques. Finally, it was pointed out that CCA, which identifies patterns of covariance between different datasets, is a natural extension of PCA that is gradually finding more widespread use in oceanography.

REFERENCES

Barnett, T., 1977: The principal time and space scales of the Pacific trade wind field. *J. Atm. Sci., 34*, 221-236.

Barnett, T., 1983: Interaction of the monsoon and Pacific trade wind system at interannual time scales, Part I: The equatorial Zone. *Mon. Weather Rev., 111*, 756-773.

Barnett, T., and K. Hasselmann, 1979: Techniques of linear prediction, with application to oceanic and atmospheric fields in the tropical Pacific. *Rev. of Geophysics and Space Physics, 17*, 949-968.

Barnett, T., and R. Preisendorfer, 1987: Origins and levels of monthly and seasonal forecast skill for United States surface air temperatures determined by Canonical Correlation Analysis. *Mon. Weather Rev., 115*, 1825-1850.

Graham, N., J. Michaelson, and T. Barnett, 1987: An investigation of the El Niño-Southern Oscillation cycle with statistical models, 1. Predictor field characteristics. *J. Geophys. Res., 92*, 14251-14270.

Harman, H., 1976: *Modern Factor Analysis*, 3rd Ed. Univ. of Chicago Press, Chicago.

Hotelling, H., 1936: Relations between two sets of variates. *Biometrika, 28*, 321-377.

Joreskog, K., J. Klovan, and R. Reyment, 1976: *Geological Factor Analysis*. Elsevier Pub. Co., N.Y.

Legler, D., 1983: Empirical orthogonal function analysis of wind vectors over the tropical Pacific region. *Bull. Am. Meteorol. Soc., 64*, 234-241.

Preisendorfer, R., 1988: *Principal Component Analysis in Meteorology and Oceanography*. Elsevier Pub. Co., N.Y.

Reyment, R., and Joreskog, K., 1993: *Applied Factor Analysis in the Natural Sciences*. Cambridge Univ. Press.

Richman, M., 1981: Obliquely rotated principal components: An improved meteorological map typing technique? *J. Appl. Met., 20*, 1145-1159.

Wallace, J., 1972: Empirical orthogonal representation of time series in the frequency domain, II. Application to the study of tropical wave disturbances. *J. Appl. Met., 11*, 893-900.

Wallace, J., and R. Dickinson, 1972: Empirical orthogonal representation of time series in the frequency domain, I. Theoretical considerations. *J. Appl. Met., 11*, 887-892.

Weare, B., and J. Nasstrom, 1982: Examples of extended empirical orthogonal function analysis. *Mon. Weather Rev., 110*, 481-485.

White, W., and C.-K. Tai, 1992: Reflection of interannual Rossby waves at the maritime western boundary of the tropical Pacific. *J. Geophys. Res., 97*, 14305-14322.

PRINCIPAL OSCILLATION PATTERN ANALYSIS OF THE INTRASEASONAL VARIABILITY IN THE EQUATORIAL PACIFIC OCEAN

Hans von Storch
Max-Planck-Institut für Meteorologie, Hamburg, Germany

ABSTRACT

In the present paper the concept of the *principal oscillation pattern* (POP) analysis is reviewed. This technique is used to simultaneously infer the characteristic patterns and time scales of a vector time series. The POPs may be seen as the normal modes of a linearized system whose system matrix is estimated from data. As a demonstration, the POP technique is used for the analysis of the intraseasonal variability in the equatorial Pacific Ocean; first results are presented. Daily observations of temperature and currents in the upper 500 m of the equatorial Pacific, recorded by moored buoys, are analyzed with respect to intraseasonal (40-180 day band) variations. Two oscillatory highly coherent modes are found with periods between 65 and 120 days. Both modes propagate eastward along the equator. The modes are clearly reflected in both the zonal currents and the temperatures, which trail behind the zonal currents by 45°. In the slower of the two modes, the temperature signal propagates more slowly than the zonal current signal, and no signal occurs in the meridional current. The mode's activity is enhanced during warm events of the Southern Oscillation. In the faster mode a signal also appears in the meridional current. Its amplitude exhibits an annual cycle, with variance on the annual and on the semiannual period. The slower mode might be an equatorial Kelvin wave but the faster mode, which has a significant meridional current component, is inconsistent with the concept of an equatorial Kelvin wave.

1. INTRODUCTION

Principal oscillation pattern analysis. In the present paper the *principal oscillation pattern* (POP) technique is reviewed (Section 2) and its usefulness is demonstrated by an analysis of the intraseasonal variability in the equatorial Pacific (Section 3). The POP analysis is a multivariate technique to empirically infer the characteristics of the space-time variations of a complex system in a high-dimensional space (Hasselmann, 1988; von Storch et al., 1988). The basic ansatz is to identify a low-order system with a few free parameters fitted to the data. Then, the space-time characteristics of the low-order system are regarded as being the same as those of the full system.

Applications of POP analysis. The POP analysis is now a routinely used tool[1] to diagnose the space-time variability of the climate system. Processes analysed with POPs are

- The low-frequency variability of the thermohaline circulation in the global ocean (Mikolajewicz and Maier-Reimer, 1991; Weisse et al., in press),
- The low-frequency variability in the coupled atmosphere-ocean system (Xu, 1993),
- The El Niño / Southern Oscillation ENSO (Xu and von Storch, 1990; Xu, 1990; Blumenthal, 1991; Latif and Villwock, 1989; Latif and Flügel,1990; Bürger, 1993; Xu, 1992; Latif et al., 1993),
- The Madden and Julian Oscillation (MJO), also named the tropical 30- to 60-day oscillation (von Storch et al., 1988; von Storch and Xu,1990; von Storch and Baumhefner, 1991; and von Storch and Smallegange,1991),
- The stratospheric Quasi-Biennial Oscillation (Xu, 1992),
- Tropospheric baroclinic waves (Schnur et al., 1993).

Generalizations of the POP analysis. There is a series of generalizations of the basic POP approach which we will not detail in the present paper. The predictive potential of the POP method has been tested with the Southern Oscillation (Xu and von Storch, 1990) and with the Madden and Julian Oscillation (von Storch and Xu, 1991). In the *cyclo-stationary* POP analysis, the estimated system matrix is allowed to vary deterministically with an externally forced cycle (Blumenthal, 1991). In the *complex* POP analysis not only the state of the system but also its "momentum" is modeled (Bürger, 1993).

Organization. In Section 2, the POPs are introduced in two conceptually different ways. One way is to define POPs as normal modes of a linear system in which parameters are inferred from a vector time series. The other way is to regard POPs as a simplified version of principal interaction patterns (PIPs). The PIP ansatz (Hasselmann, 1988) is a fairly general approach which allows for a large variety of complex scenarios. In Section 3 a POP analysis of daily hydrographic reports (temperature, zonal and meridional currents, as well as surface wind) from moored buoys in the tropical Pacific Ocean is presented. Two eastward propagating modes, both similar to the mode described by Johnson and McPhaden (1993), are identified and their spatial signatures are described. The paper is concluded in Section 4 with some remarks on the general merits and limitations of the POP technique.

[1] A FORTRAN code with a manual (Gallagher et al., 1991) of the regular POP analysis is free of charge available at the Deutsches Klimarechenzentrum, Bundesstrassse 55, 2000 Hamburg 13, Germany.

2. PRINCIPAL OSCILLATION PATTERNS

The following notations are used: Vectors are given as **bold** letters and matrices as calligraphic letters like $\mathcal{A}$ or $\mathcal{X}$. If $\mathcal{A}$ is a matrix then $\mathcal{A}^T$ is the transposed matrix. If x is any complex quantity then x^* is its conjugate complex. It should be noted that the POP formalism—conventional, cyclostationary, and complex POP analysis—may be applied to linear systems whose system matrices are estimated from data or whose system matrices are derived from theoretical dynamical considerations (Schnur et al., 1993).

2.1 POPs and Normal Modes

Normal modes. The normal modes of a linear discretized real system

$$\mathbf{x}(t+1) = \mathcal{A} \cdot \mathbf{x}(t) \tag{1}$$

are the eigenvectors $\mathbf{p}$ of the matrix $\mathcal{A}$. In general, $\mathcal{A}$ is not symmetric and some or all of its eigenvalues λ and eigenvectors $\mathbf{p}$ are complex. However, since $\mathcal{A}$ is a real matrix, the conjugate complex quantities λ^* and $\mathbf{p}^*$ satisfy also the eigen-equation $\mathcal{A} \cdot \mathbf{p}^* = \lambda^* \mathbf{p}^*$. In most cases, all eigenvalues are different and the eigenvectors form a linear basis. So each state $\mathbf{x}$ may be uniquely expressed in terms of the eigenvectors

$$\mathbf{x} = \sum_j z_j \cdot \mathbf{p}_j . \tag{2}$$

The coefficients of the pairs of conjugate complex eigenvectors are conjugate complex, too. Inserting (2) into (1) we find that the coupled system (1) becomes uncoupled, yielding n single equations, where n is the dimension of the process $\mathbf{x}$,

$$z(t+1) \cdot \mathbf{p} = \lambda \cdot z(t) \cdot \mathbf{p} \tag{3}$$

so that if $z(0) = 1$

$$z(t) \cdot \mathbf{p} = \lambda^t \cdot \mathbf{p} . \tag{4}$$

The contribution $\mathbf{P}(t)$ of the complex conjugate pair $\mathbf{p}$, $\mathbf{p}^*$ to the process $\mathbf{x}(t)$ is given by

$$\mathbf{P}(t) = z(t) \cdot \mathbf{p} + [z(t) \cdot \mathbf{p}]^* . \tag{5}$$

Writing $\mathbf{p} = \mathbf{p}^1 + i \cdot \mathbf{p}^2$ and $2z(t) = z^1(t) - i \cdot z^2(t)$, this reads

$$\begin{aligned} \mathbf{P}(t) &= z^1(t) \cdot \mathbf{p}^1 + z^2(t) \cdot \mathbf{p}^2 \\ &= \rho^t \cdot (\cos(\eta t) \cdot \mathbf{p}^1 - \sin(\eta t) \cdot \mathbf{p}^2) \end{aligned} \tag{6}$$

with $\lambda = \rho \cdot \exp(-i\eta)$ and if $z(0) = 1$. The geometric and physical meaning of (6) is that between the spatial patterns $\mathbf{p}^1$ and $\mathbf{p}^2$ the trajectory $\mathbf{P}(t)$ performs a spiral (Figure 1) with period $T = 2\pi/\omega$ and e-folding time $\tau = -1/\ln(\rho)$, in the consecutive order

$$\cdots \rightarrow \mathbf{p}^1 \rightarrow -\mathbf{p}^2 \rightarrow -\mathbf{p}^1 \rightarrow \mathbf{p}^2 \rightarrow \mathbf{p}^1 \rightarrow \cdots \tag{7}$$

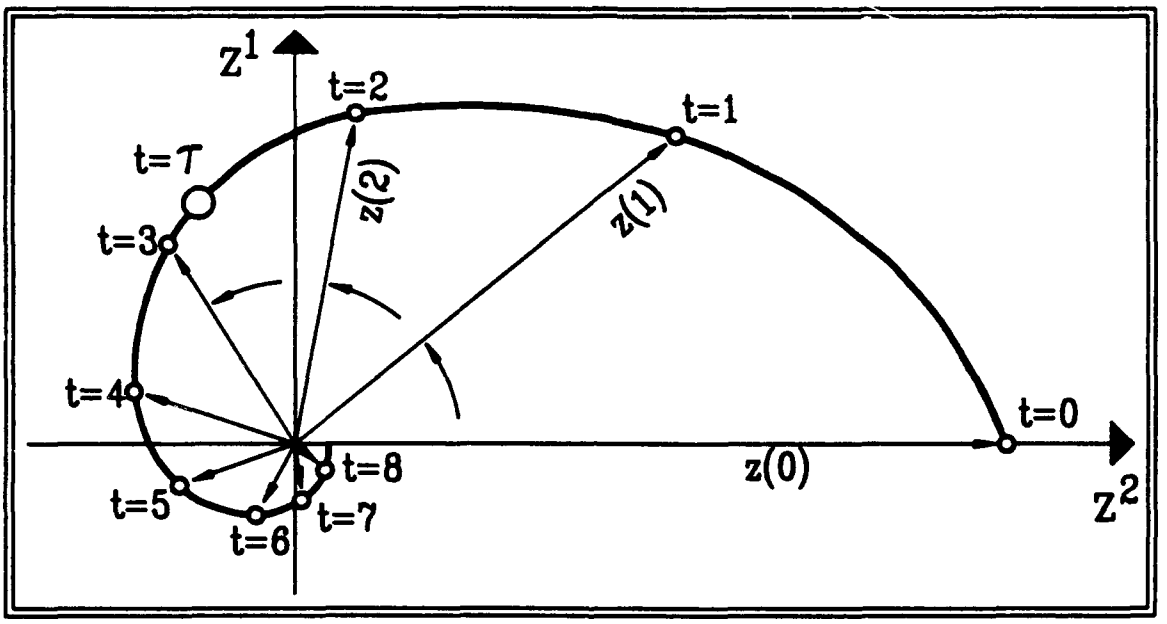


Figure 1. Typical evolution of a POP signal, given by Eq. (6), if $z^1(0) = 0$ and $z^2(0) = 1$. In this demonstration the period is $T \approx 9$ and the e-folding time is $\tau \approx 2.8$.

The e-folding time. The e-folding time has to be considered with some caution. It represents formally the average time for an amplitude of strength one to reduce to $1/e$. But in the POP context this time is a statistic of the entire time interval, i.e., it is derived not only from the episodes when the signal is active but also from those times when the signal is weak or even absent. As such, the mode will be dampened less quickly as indicated by the e-folding time when the mode is active. The other limitation refers to the presence or absence of high-frequency variations. If these are filtered out, as in Section 3, the e-folding time is lengthened.

Representation of normal modes. The modes may be represented either by the two patterns $\mathbf{p}^1$ and $\mathbf{p}^2$, or by plots of the local wave amplitude $A^2(\mathbf{r}) = [\mathbf{p}^1(\mathbf{r})]^2 + [\mathbf{p}^2(\mathbf{r})]^2$ and relative phase $\psi(\mathbf{r}) = \tan^{-1}[\mathbf{p}^2(\mathbf{r})/\mathbf{p}^1(\mathbf{r})]$ (Figure 2). The transformation (7) between the patterns $\mathbf{p}^1$ and $\mathbf{p}^2$ can assume various geometric wave forms. If $\mathbf{p}^2(\mathbf{r}) = \mathbf{p}^1(\mathbf{r} - \mathbf{r}_0)$ with a location vector $\mathbf{r}$ and a fixed vector $\mathbf{r}_0$, the signal appears as a parallel crested wave of

wavelength $4\mathbf{r}_0$, propagating in the $\mathbf{r}_0$-direction (Figure 2a). In Figure 2b an amphidromal (rotational) wave is shown.

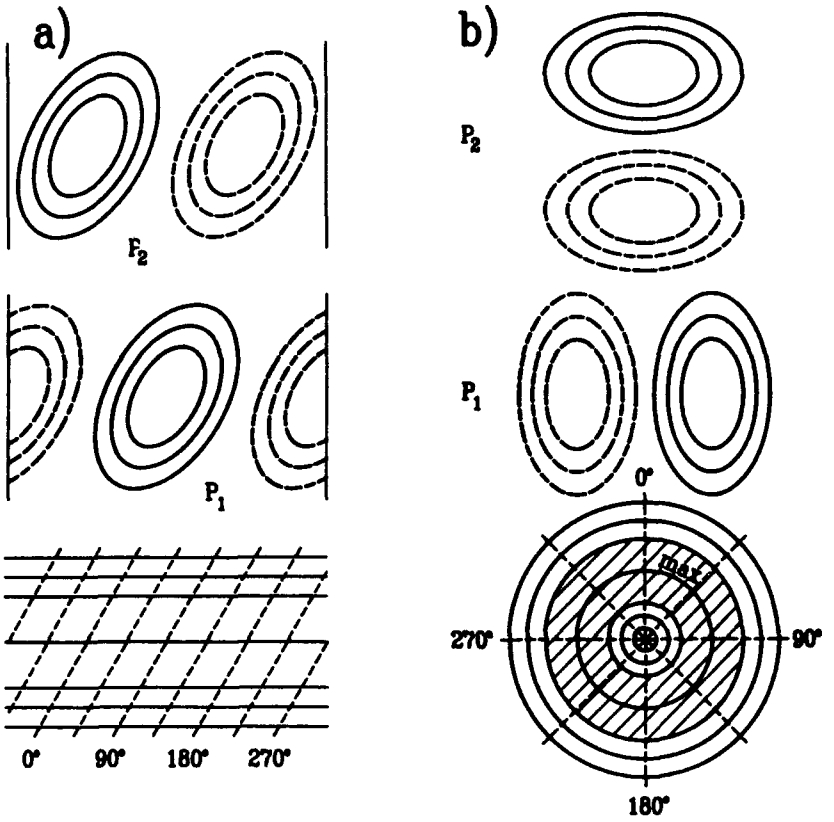


Figure 2. Examples of (a) a propagating wave and (b) an amphidromal wave and their representation in terms of POPs. Top two panels: representation by $\mathbf{p}^1$ and $\mathbf{p}^2$. Bottom panel: representation by phase ψ (dashed) and amplitude A (solid). From von Storch et al. (1988).

Time coefficients. The pattern coefficients z_j are given as the dot product of $\mathbf{x}$ with the *adjoint patterns* $\mathbf{p}_j^A$, which are the normalized eigenvectors of $\mathcal{A}^T$:

$$(\mathbf{p}_j^A)^T \mathbf{x} = \sum_k z_k (\mathbf{p}_j^A)^T \mathbf{p}_k = z_j \tag{8}$$

POPs. All information used so far is the existence of a linear equation Eq. (1) with some matrix $\mathcal{A}$. No assumption was made about the origin of this matrix. In dynamical theory, the origins of Eq. (1) are linearized and discretized differential equations. In case of the POP analysis, the relationship

$$\mathbf{x}(t+1) = \mathcal{A} \cdot \mathbf{x}(t) + \text{noise} \tag{9}$$

is hypothesized. Multiplication of Eq. (9) from the right hand side by the transposed $\mathbf{x}^T(t)$ and taking expectations, E, leads to

$$\mathcal{A} = E[\mathbf{x}(t+1)\mathbf{x}^T(t)] \cdot \left[E[\mathbf{x}(t)\mathbf{x}^T(t)]\right]^{-1}. \tag{10}$$

The eigenvectors of Eq. (10) or the normal modes of Eq. (9) are called *principal oscillation patterns*. The coefficients z are called *POP coefficients*. Their time evolution is given by Eq. (3), superimposed by noise

$$z(t+1) = \lambda \cdot z(t) + \text{noise}. \tag{11}$$

The stationarity of Eq. (11) requires $\rho < 1$. In practical situations, when only a finite time series $\mathbf{x}(t)$ is available, $\mathcal{A}$ is estimated by first deriving the sample lag-1 covariance matrix $\mathcal{X}_1 = \sum_t \mathbf{x}(t+1)\mathbf{x}^T(t)$ and the sample covariance matrix $\mathcal{X}_0 = \sum_t \mathbf{x}(t)\mathbf{x}^T(t)$ and then forming $\mathcal{A} = \mathcal{X}_1\mathcal{X}_0^{-1}$. The eigenvalues of this matrix always satisfy $\rho < 1$.

To reduce the number of spatial degrees of freedom in some applications, the data are subjected to a truncated empirical orthogonal function (EOF) expansion, and the POP analysis is applied to the vector of the first EOF coefficients. A positive by-product of this procedure is that noisy components can be excluded from the analysis. Then, the covariance matrix $\mathcal{X}_0$ has a diagonal form.

If there is a priori information that the expected signal is located in a certain frequency band, it is often advisable to time-filter the data prior to the POP analysis. A somewhat milder form of focusing on selected time scales is to derive the EOFs from time-filtered data and then to project the unfiltered data on these EOFs.

Criteria to decide whether a POP contains useful information or if it should be regarded as reflecting mostly sample properties are given by von Storch et al. (1988). The most important rule-of-thumb is related to the cross spectrum of the POP coefficients z^1 and z^2: at the POP period T, or at least in the neighborhood of T, the two time series should be significantly coherent and 90° out of phase, according to Eq. (6).

Invariance against coordinate transformations. If the original time series $\mathbf{x}(t)$ is transformed into another time series $\mathbf{y}(t)$ by means of $\mathbf{y}(t) = \mathcal{L} \cdot \mathbf{x}(t)$ with an invertible matrix $\mathcal{L}$, (i.e., $\mathcal{L}^{-1}$ exists), then the eigenvalues are unchanged and the eigenvectors transform as x:

$$\mathcal{A}_X = \mathcal{X}_1\mathcal{X}_0^{-1};\ \mathcal{A}_Y = \mathcal{Y}_1\mathcal{Y}_0^{-1}$$

with $\mathcal{Y}_1 = E(\mathbf{y}(t+1)\mathbf{y}^T(t)) = \mathcal{L}\mathcal{X}_1\mathcal{L}^T$ and $\mathcal{Y}_0 = \mathcal{L}\mathcal{X}_0\mathcal{L}^T$. Thus $\mathcal{A}_Y = \mathcal{L}\mathcal{A}_X\mathcal{L}^{-1}$. If $\mathbf{p}_X$ is an eigenvector of $\mathcal{A}_X$ with eigenvalue λ, i.e., $\mathcal{A}_X\mathbf{p}_X = \lambda\mathbf{p}_X$ then $\mathcal{A}_X\mathcal{L}^{-1}\mathcal{L}\mathbf{p}_X = \lambda\mathbf{p}_X$ and, eventually $\mathcal{L}\mathcal{A}_X\mathcal{L}^{-1}(\mathcal{L}\mathbf{p}_X) = \lambda(\mathcal{L}\mathbf{p}_X)$. That is, if $\mathbf{p}_X$ is a POP of the time series $\mathbf{x}$, then $\mathcal{L}\mathbf{p}_X = \mathbf{p}_Y$ is a POP of $\mathbf{y}$ with the same eigenvalue λ.

The EOFs are *not* invariant against linear transformations $\mathcal{L}$, since in general the matrices $\mathcal{X}_0$ and $\mathcal{L}\mathcal{X}_0\mathcal{L}^T$ will have different eigenvalues and eigenvectors. Therefore, if the POP analysis is begun with a projection of the data on a truncated EOF expansion, the results of a POP analysis will change if the data are transformed into another coordinate system.

The POP coefficients. To get the POP coefficients, $z(t)$, two approaches are possible. One is to derive the adjoint patterns $\mathbf{p}^A$ and to use Eq. (8). An alternative is to not derive adjoint patterns but to derive the coefficients z by a least-square fit of the data $\mathbf{x}$ by minimizing

$$\left\|\mathbf{x} - z\cdot\mathbf{p} - [z\cdot\mathbf{p}]^*\right\| = \left\|\mathbf{x} - z^1\mathbf{p}^1 - z^2\mathbf{p}^2\right\| \tag{12}$$

if $\mathbf{p}$ is complex, or

$$\left\|\mathbf{x} - z\cdot\mathbf{p}\right\|. \tag{13}$$

2.2 POPs = Trivial Case of PIPs

State space models. Many complex dynamical systems, $\mathbf{x} \in R^n$, may conveniently be approximated as being driven by a simpler dynamical system, $\mathbf{z} \in R^m$, with a reduced number of degrees of freedom, $m \le n$. Mathematically, this may be described by a *state space model* which consists of a *system* equation

$$\mathbf{z}(t+1) = \mathcal{F}[\mathbf{z}(t), \alpha, t] + \text{noise}, \tag{14}$$

for the dynamical variables $\mathbf{z} = (z_1, \ldots, z_m)$ and an *observation* equation

$$\mathbf{x}(t) = \mathcal{P}\mathbf{z}(t) + \text{noise} = \sum_j z_j(t)\mathbf{p}_j + \text{noise} \tag{15}$$

for the observed variables $\mathbf{x}$. $\mathcal{P}$ is the matrix whose columns are the vectors, or *patterns*, $\mathbf{p}_j$. In general $\mathcal{P}$ is not a square-matrix. $\mathcal{F}[\mathbf{z}(t), \alpha, t]$ denotes a class of models which can be nonlinear in the dynamical variables $\mathbf{z}$ and which depends additionally on a set of free parameters $\alpha = (\alpha_1, \alpha_2, \ldots)$. Both equations, Eqs. (14,15), are disturbed by an additive noise.

Since $m \leq n$, the time coefficient $z_j(t)$ of a pattern $\mathbf{p}_j$ at a time t is not uniquely determined by the $\mathbf{x}(t)$. Instead, it may be obtained by a least-square fit, i.e.,

$$\mathbf{z}(t) = (\mathcal{P}^T\mathcal{P})^{-1}\mathcal{P}^T\mathbf{x}(t). \tag{16}$$

The intriguing aspect of state space models is that the dynamical behavior of complex systems often appears to be dominated by the interaction of only a few characteristic patterns $\mathbf{p}_j$. That is, even if the dynamics of the full system are restricted to the subspace spanned by the columns of $\mathcal{P}$, its principal dynamical properties are represented.

PIPs. When fitting the state space model Eqs. (14,15) to a time series, the following entities have to be specified: the class of models $\mathcal{F}$, the patterns $\mathcal{P}$, the free parameters α, and the dimension of the reduced system m. The class of models $\mathcal{F}$ has to be selected a priori on the basis of physical reasoning. Also, the number m might be specified a priori. The parameters α and the patterns $\mathcal{P}$ are fitted simultaneously to a time series by requesting them to minimize

$$\epsilon[\mathcal{P};\alpha] = \mathrm{E}\left\| \mathbf{x}(t+1) - \mathbf{x}(t) - \mathcal{P}(\mathcal{F}[\mathbf{z}(t),\alpha,t] - \mathbf{z}(t))\right\|^2 \tag{17}$$

where $\epsilon[\mathcal{P};\alpha]$ is the mean square error of the approximation of the (discretized) time derivative of the observations $\mathbf{x}$ by the state space model. The patterns $\mathcal{P}$, which minimize Eq. (17), are called *principal interaction patterns* (Hasselmann, 1988). If only a finite time series of observations $\mathbf{x}$ is available, the expectation E is replaced by a summation over time.

In general, the minimization of Eq. (17) is not unique. In particular, the set of patterns $\mathcal{P}' = \mathcal{P} \cdot \mathcal{L}$ with any nonsingular squared matrix $\mathcal{L}$ will minimize Eq. (17), if $\mathcal{P}$ does, as long as the corresponding model $\mathcal{F}' = \mathcal{L}^{-1}\mathcal{F}$ belongs to the a priori specified model class. This problem may be solved by requesting the solution to fulfill some constraints, e.g., that the linear term in the Taylor expansion of $\mathcal{F}$ is a diagonal matrix.

POPs as PIPs. The principal oscillation patterns can be understood as a kind of simplified principal interaction patterns. For that assume $m = n$. Then, the patterns $\mathcal{P}$ span the full $\mathbf{x}$-space, and their choice does not affect $\epsilon[\mathcal{P};\alpha]$. Also, let $\mathcal{F}$ be a linear model $\mathcal{F}[\mathbf{z}(t),\alpha] = \mathcal{A} \cdot \mathbf{z}(t)$, where the parameters α are the entries of $\mathcal{A}$. Then the dynamical equation Eq. (14) is identical to Eq. (11). The constraint mentioned above leads to the eigenvectors of $\mathcal{A}$ as being the PIPs of the particular, admittedly simplified, state space model.

2.3 Associated Correlation Patterns

Definition and representation. The *associated correlation pattern analysis* (von Storch et al., 1988) is a regression analysis to infer the spatial properties of a signal which is encoded in a two-dimensional index (a complex POP coefficient, for instance). If the parameter under consideration is $\vec{\mathbf{Y}}(t)$ and the bivariate index is $(z^1(t), z^2(t))$ the two associated correlation patterns $\vec{q}^1$ and $\vec{q}^2$ minimize

$$\sum_t \left\| \vec{\mathbf{Y}}(t) - \frac{z^1(t)}{\sqrt{2}}\vec{q}^1 - \frac{z^2(t)}{\sqrt{2}}\vec{q}^2 \right\|^2 = \min. \tag{18}$$

The normalization with $\sqrt{2}$ in Eq. (18) has been introduced so that $\vec{q}^1$ represents a typical state for $z^1(t) = 1$, $z^2(t) = 0$ and $\vec{q}^2$ a typical state for $z^1(t) = 0$, $z^2(t) = 1$. The solution of Eq. (18) is straightforward and requires the solution of a 2×2 linear equation at each location r of the input field $\vec{\mathbf{Y}} = (\mathbf{y}_r)$.

The associated correlation patterns can be displayed directly by the two patterns $\vec{q}^1$ and $\vec{q}^2$ or by amplitude distributions and phase distributions (Figure 6). The amplitude A and the phase ψ at the location r is given by

$$A = \sqrt{(\vec{q}^1)^2 + (\vec{q}^2)^2} \tag{19}$$

$$\tan\left(2\pi\frac{\psi}{T}\right) = \frac{-\vec{q}^2}{\vec{q}^1} \tag{20}$$

with T being the period of the mode. The phase ψ has been defined such that $\psi = 0$ coincides with $z^2 = 0$ and $z^1 > 0$, and $\psi = T/4$ with $z^1 = 0$ and $z^2 < 0$ (compare with Eq. (7)).

Measure of skill. A number measuring the relative importance of a POP for a parameter $\mathbf{y}_r$ at the location r is the rate of explained $\mathbf{y}_r$ variance by the index (z^1, z^2). This rate is given by

$$\varepsilon(\mathbf{y}_r, z^1, z^2) = \frac{\mathrm{Var}(\mathbf{y}_r) - \epsilon_r^2}{\mathrm{Var}(\mathbf{y}_r)} \tag{21}$$

with

$$\epsilon_r^2 = \sum_t \left[y_r(t) - \frac{z^1(t)}{\sqrt{2}\,\sigma_1} \vec{q}_r^1 - \frac{z^2(t)}{\sqrt{2}\,\sigma_2} \vec{q}_r^2 \right]$$

being the local error in Eq. (18); $\epsilon = 1$ indicates a perfect model and $\epsilon = 0$ a model without skill.

3. POP ANALYSIS OF THE INTRASEASONAL VARIABILITY IN THE EQUATORIAL PACIFIC

General. The general analysis strategy is first to derive an index of the equatorial modes through a principal oscillation pattern (POP) analysis of the equatorial current meter moorings at 165°E, 140°W, and 110°W. The time series at these stations are relatively long and sample the equator fairly well. Zonal currents and temperatures, which ought to reflect equatorial Kelvin waves well, as well as meridional currents are monitored by these buoys. After having established that the index makes sense, all available data from the current meter moorings and from the ATLAS buoys are examined in an "associated correlation pattern" analysis. The purpose of this exercise is to infer the 3-dimensional spatial structure of the modes.

3.1 Raw Data

For the analysis, daily observations were available from two series of moored buoys (Hayes et al., 1991):

- Current meter moorings at four locations, the exact positions of which are given in Table 1. These buoys recorded zonal and meridional currents and temperature at various levels and near surface air temperature and zonal and meridional wind.

- ATLAS buoys located at 20 positions in the near-equatorial Pacific (for the exact positions, see Table 1). From these buoys, subsurface temperatures at various levels, as well as near surface air temperature and wind, are available.

The shortest time series is from 147°E, 5°N (9 months). Maximum length is 7 years (at 0°, 110°W and 140°W).

Mean State. The buoy data represent a good data base to sketch the mean distribution of currents and temperature in the equatorial Pacific. In Figure 3 are plotted the mean zonal current and temperature distributions along the equator as well as latitude-depth cross-sections of temperature along 165°E and 110°W. The mean equatorial temperature distribution is dominated by the sharp thermocline that separates water of 10–15°C at deeper layers from warm surface waters of 24°C in the east and 28°C in the west. If we identify the thermocline with the 20°C isotherm, then the thermocline rises from 180 m at

165°E to 100 m at 140°W to 60 m at 110°W. The zonal current is weakly westward at the surface with maximum values below 25 cm/s. Maximum eastward flowing currents, the *Equatorial Undercurrent*, prevail along the thermocline, with maximum values at about the 17.5°C isotherm. At 165°E the maximum current is below 50 cm/s, at 140°W maximum speeds are 100 cm/s, and at 110°W above 75 cm/s.

Maximum temperatures prevail north of the Equator in the east and south of the Equator in the west. The thermal wind relationship is nicely reflected in the mean distributions (Fig. 3a, c and d).

Table 1. Position of buoys from which data have been used in the present study. Also given is the maximum time interval for which at least one variable is available.

Instrument	Longitude	Latitude	Data interval	Parameters
CMM	0°	165°E	5/86 - 4/91	current, temperature, wind
CMM	0°	140°W	5/84 - 4/91	
CMM	0°	110°W	5/84 - 4/91	
CMM	7°N	110°W	5/88 - 4/91	
ATLAS	5°N	147°E	5/90 - 2/91	temperature, wind
ATLAS	8°N	165°E	5/90 - 4/91	
ATLAS	5°N	165°E	7/88 - 4/91	
ATLAS	2°N	165°E	7/87 - 4/91	
ATLAS	2°S	165°E	5/86 - 4/91	
ATLAS	5°S	165°E	7/87 - 4/91	
ATLAS	0°	169°W	5/88 - 4/91	
ATLAS	7°N	147°W	11/88 - 11/90	
ATLAS	9°N	140°W	5/88 - 4/91	
ATLAS	5°N	140°W	5/88 - 4/91	
ATLAS	2°N	140°W	5/87 - 4/91	
ATLAS	2°S	140°W	5/87 - 4/91	
ATLAS	5°S	140°W	10/90 - 4/91	
ATLAS	7°N	132°W	5/89 - 10/90	
ATLAS	0°	124°W	5/87 - 4/91	
ATLAS	5°N	110°W	5/86 - 4/91	
ATLAS	2°N	110°W	6/85 - 4/91	
ATLAS	2°S	110°W	5/85 - 4/91	
ATLAS	5°S	110°W	5/86 - 4/91	
ATLAS	8°S	110°W	5/86 - 6/87	

Variability around the annual cycle. The annual cycles have been removed from all data. To also exclude part of the Southern Oscillation-related variability, this removal of

the annual cycle was done for each May-to-April segment separately. The May-to-April segments were chosen to represent one "El Niño year" (Wright ,1985). As an example, three variables at 0°, 140°W are shown before and after the removal of the low-frequency variability (Figure 4).

At the equatorial buoy all parameters undergo marked variations on the interannual time-scale, some of which stem mostly from the regular annual cycle (e.g., the zonal wind). In the subsurface variables the irregular ENSO-related variations contribute most to the low frequency variability. The high-frequency variations are normally distributed. In the zonal wind the intraseasonal variations are almost white in time, whereas the subsurface parameters exhibit an oscillatory behavior with typical periods of 50–100 days. The zonal current seems to lead the temperature by a few days.

3.2 The POP Analysis of the Equatorial Current Meter Mooring Data

Preprocessing of the data. In the data field to be analysed, we have parameters that differ with respect to units as well with respect to their standard deviations. To allow all parameters to play the same role in the analysis, all data are standardized to zero mean and standard deviation one.

For the POP analysis it is often helpful if the data are preprocessed prior to the analysis with the purpose of suppressing space-time noise (see section 2.1). The spatial noise is taken out by doing the analysis in a low-dimensional subspace spanned by the first few EOFs, and the temporal variations on time scales irrelevant for the process under investigation are taken out by a time filter.

The data are first subjected to an EOF analysis. In this EOF analysis the entries γ_{ij} of the correlation matrix have been estimated from all available pairs of observations, i.e.,

$$\gamma_i = \frac{1}{n_{ij}} \sum_{t \in \mathcal{T}_{ij}} p_i(t) p_j(t) \tag{22}$$

where $p_i(t)$ represents the i-parameter of the data field $\vec{\mathbf{X}}(t)$ at time t. $\mathcal{T}_{ij}$ is the set of all times when both p_i and p_j have been observed and n_{ij} is the number of elements in $\mathcal{T}_{ij}$ Definition Eq. (22) is adequate for the case of gappy data. Only those pairs of indices (i,j) were considered for which n_{ij} was at least 50% of all possible observations.

The EOF coefficients $\alpha^k(t)$ are then no longer given as the dot product of the field $\vec{\mathbf{X}}(t)$ at time t and the respective EOF $\vec{e}^k$ but are determined as a least-square-fit

$$\left\| \vec{\mathbf{X}}(t) - \alpha^k(t) \times \vec{e}^k \right\| = \min. \tag{23}$$

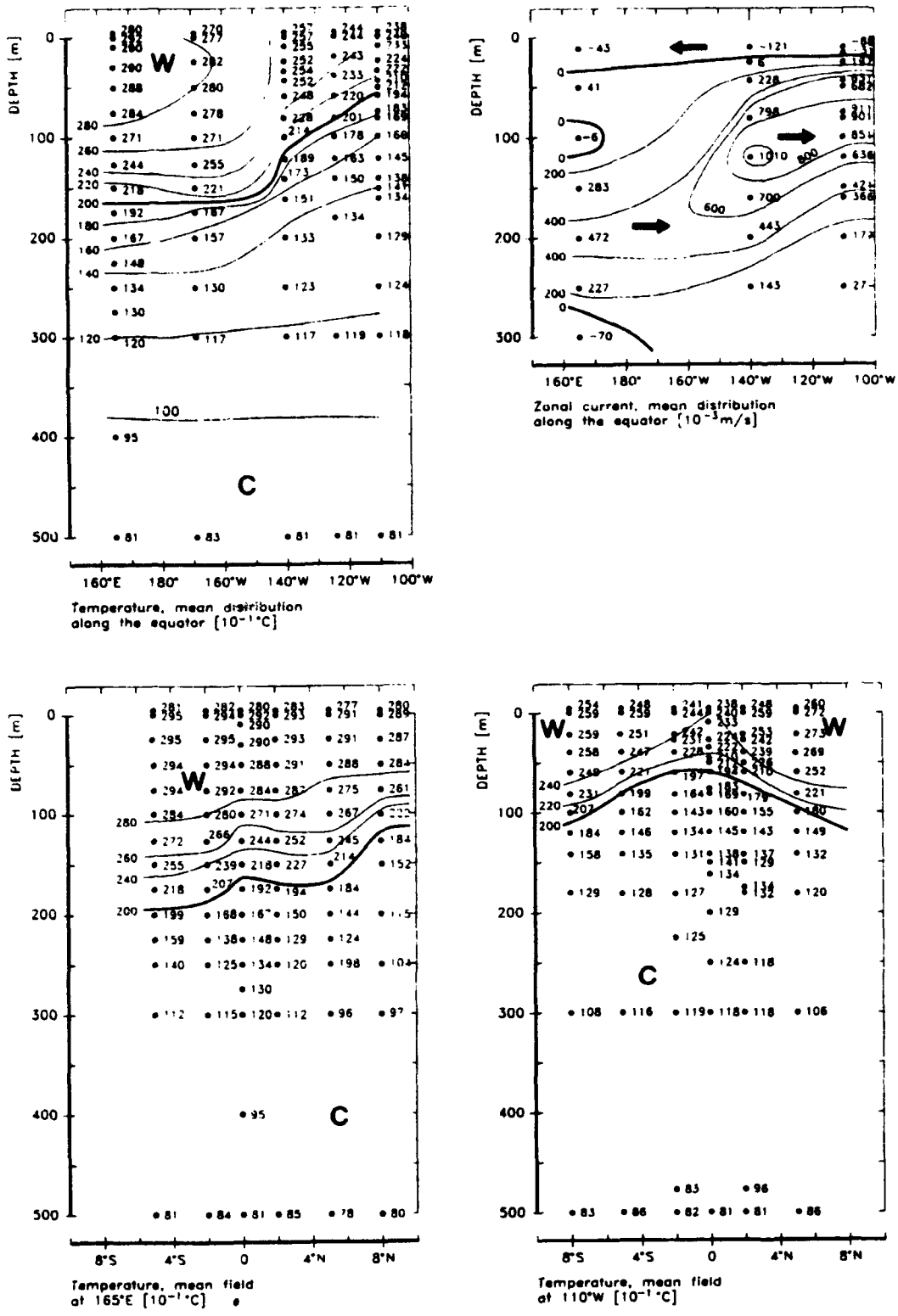


Figure 3. Mean distributions derived from the buoy data. The 20°C isotherm in the temperature distributions (in 10^{-1} °C) and the zero line in the current distribution (in 10^{-3} m/s) are given as heavy lines. Top: Longitude-depth cross sections of temperature and zonal current along the equator. Bottom: Latitude-depth cross-sections of temperature along 165°E and 110°W.

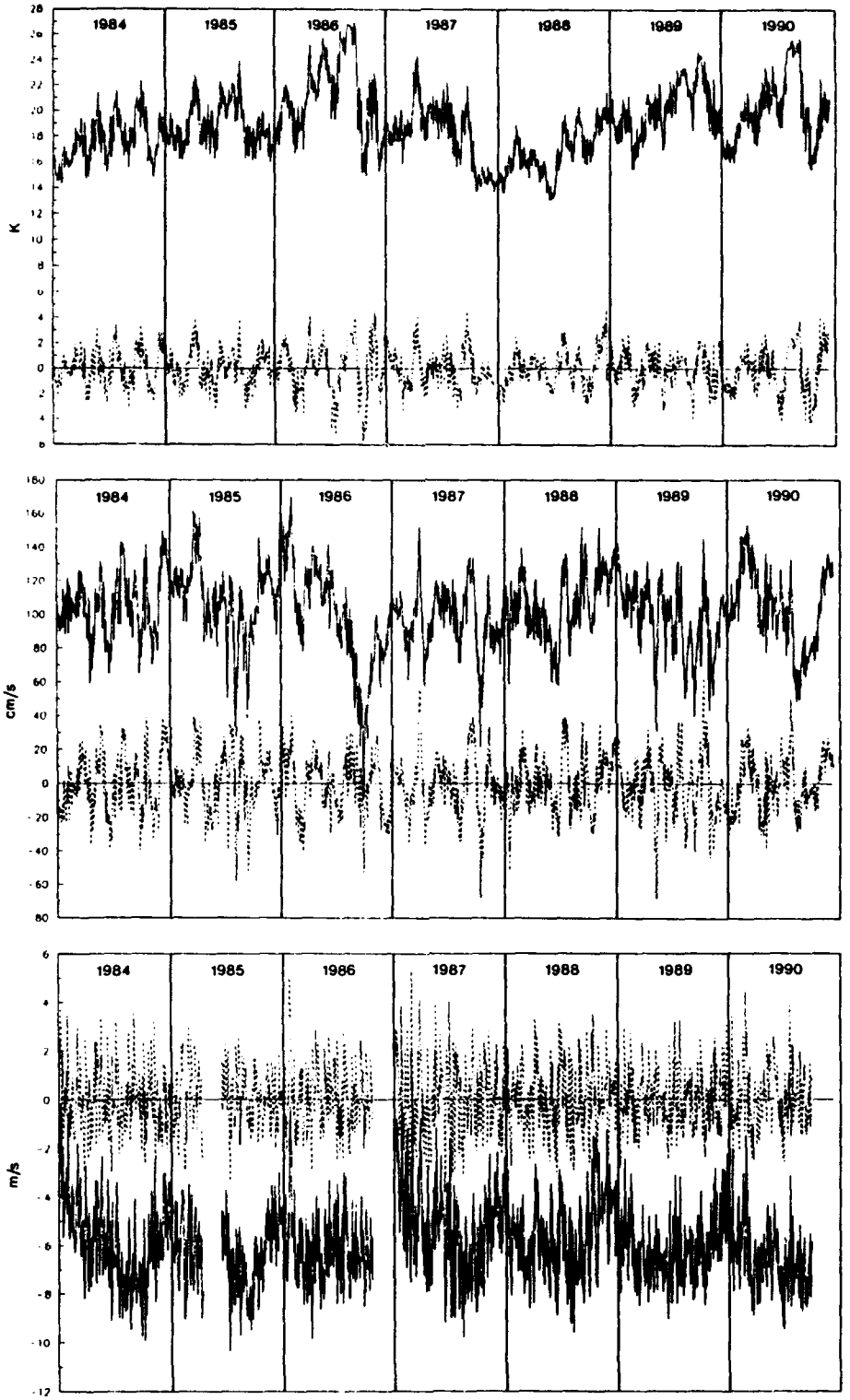


Figure 4. Time series at 0°, 140°W for temperature and currents at 120 m and for zonal wind, before and after subtraction of the annual cycle and of low-frequency variations. The years are given as May-to-April segments.

The EOF coefficients $\alpha^k(t)$ represent complete time series over the entire 7-year time interval from May 1984 through April 1991. These time series are time-filtered such that all variability below 10 days and above 180 days is completely eliminated and all variability on time scales between 40 and 150 days is not affected. In the windows between 10 and 40 days and 150 and 180 days the filter response function smoothly changes from 0 to 1.

Results of POP analysis. Two oscillatory modes are identified whose coefficient time series exhibit the desired high coherency and 90°-out-of-phase relationship. In Figure 5 the amplitude time series of the two complex POP coefficients are plotted. Note that the coefficient time series have been normalized so that $\mathrm{Var}(z^i(t)) = 1$. The coefficients were obtained by means of the adjoint patterns and Eq. (8).

One mode has a POP period $T = 65$ days, and an e-folding time $\tau = 73$ days. It represents about 16% of the variance of the band-pass filtered, EOF-truncated and normalized data (at all three locations, for temperature, zonal, and meridional currents as well as winds, and at all depths). In consistency with the POP period the maximum coherence is obtained for 60 days. The amplitude time series reveals a marked annual cycle, with a definite appearance of a semiannual component. The wave activity is strongest during solstice conditions and minimum activity during equinoctial conditions.

The second mode has an e-folding time of 106 days and a POP period $T = 120$ days. But the POP coefficients $z^1(t)$ and $z^2(t)$ have largest coherencies at 72 days, so that the POP period of 120 days likely is an overestimate of the true oscillation period. The POP coefficient represents 18% of the variance of the band-pass filtered, EOF-truncated, and normalized data. The amplitude time series in Figure 5 are hardly affected by the annual cycle. Instead the modification of the large-scale environment through the development of warm El Niño conditions leaves a clear mark on the time series. During the warm event in 1986/87 and the early phase of the warm event in 1990/91 the activity of the waves is enhanced.

The two modes are only weakly correlated. The correlations between the real and imaginary parts of the coefficient time series are very small, and the correlations between the real (imaginary) parts of the two modes are about –0.25.

3.3 The Spatial Signature of the Mode

General. In the present study, associated correlation patterns have been computed from various parameters for both modes separately. In all cases the annual cycle, as represented by the first two annual harmonics and the overall mean of each May-to-April segment, has been removed prior to the analysis. No more time-smoothing was done because of the wide gaps in the data. An implicit time-filtering has been introduced through the use of the

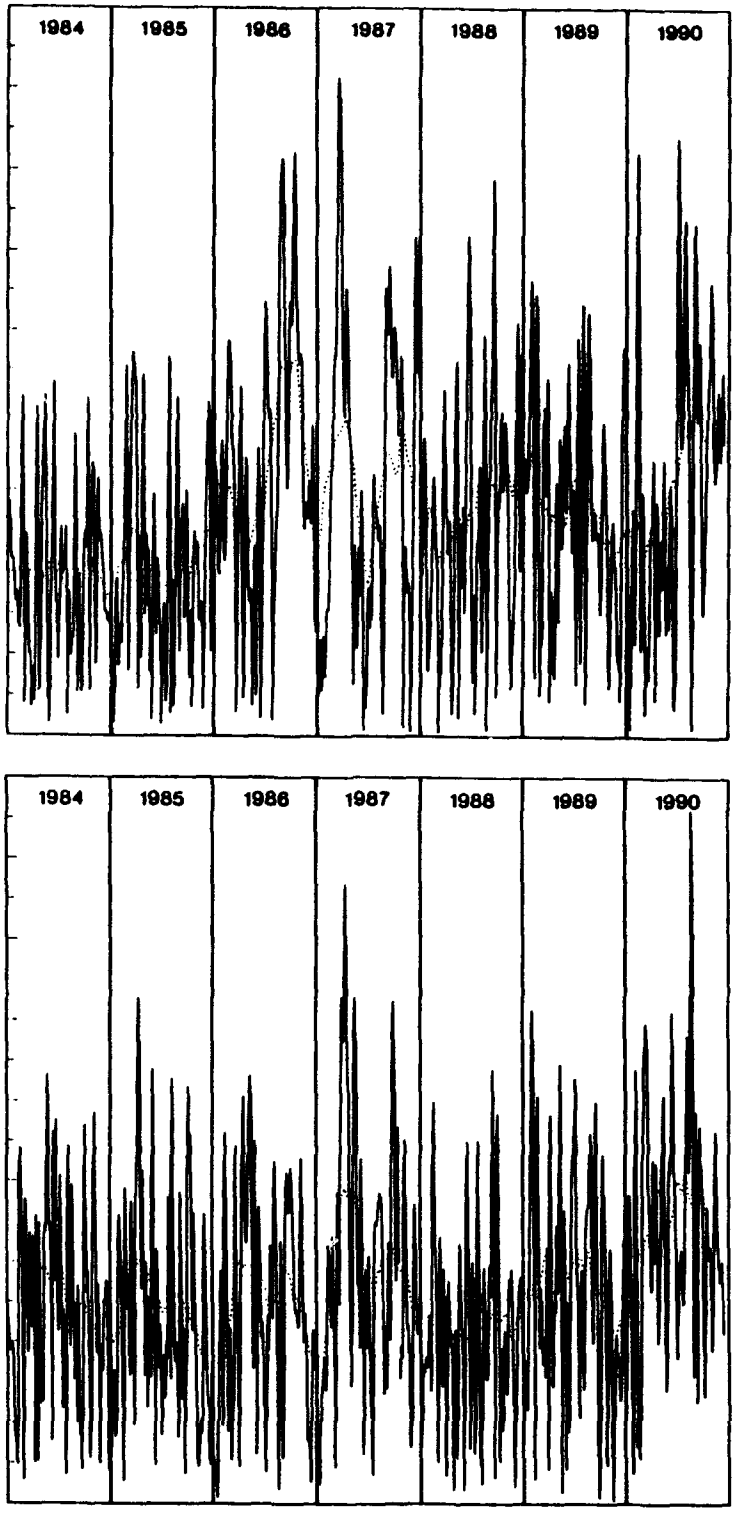


Figure 5. Time series of the amplitude of the two modes identified in the joint POP analysis of normalized data from equatorial current meter moorings. The years are drawn as May-to-April segments. (a) The 120 day mode. (b) The 65 day mode.

POP coefficient time series. Since these time series have been derived from time-filtered data (see above), they are themselves smooth. Unlike the POP analysis, the data are not normalized for the associated correlation pattern analysis.

Currents at the current meter moorings. The longitude-depth distributions of the amplitudes and phases of the two intraseasonal modes, with POP periods of 120 days and 65 days, are shown in Figure 6 for the zonal current. Both modes represent eastward propagating signals.

The 120-day mode has its largest amplitudes in the central part of the tropical Pacific, with maximum values of 16 cm/s, as typical anomalies, at 50 m depth at 165°E and 160 m depth at 140°W. In contrast, the 65-day mode has maximum zonal current anomalies at upper levels (50 m and above) in the eastern part of the basin, with a typical maximum of 12 cm/s at 140°W and 19 cm/s at 110°W.

In the 120-day mode, the zonal current signals need about 60 days to propagate from the 165°E buoy to the easternmost buoy at 110°W. If we accept the estimate of 120 days as a period, then the mean phase speed is 1.8 m/s. This number is increased to 2.4 m/s or 3.0 m/s if the period is set to 90 or even 72 days (see above). The phase lines are vertically tilted at 165°E and 140°W, with the upper levels lagging the lower levels by about 45° or 15 days (of a 120-day period).

The phase speed for the 65-day mode is estimated to be, on an average, 2.1 m/s. At the two eastern positions, the phase lines are again tilted, with the lower levels leading the upper levels by about 45° or 8 days (of a 65-day period). Maximum explained local variance of the zonal current field is 40% at 120 m at 140°W for the 120-day mode and 20% at 120 m at 110°W for the 65-day mode.

Current information is also available for one off-equatorial location from the 7°N, 140°W buoy. Here a maximum of 7% of explained variance is obtained for the 120-day mode at 40 m, where an amplitude of 5.4 cm/s is found (not shown). Thus the signal is weak at that location, but interestingly the sign at 7°N is opposite to that at the equator (not shown). A similar result is found for the 65-day mode.

In the meridional current the signal is negligible for the 120-day mode, but a well-defined signal is identified in the 65-day mode. Maximum percentages of explained local variance are 12% at 120 m and 160 m at 110°W. A maximum amplitude of 10 cm/s near the surface lags an amplitude of about 8 cm/s at lower levels by about 10 days (not shown). The phase relationship with the zonal current is that northward meridional current anomalies lead easterly zonal current anomalies by 10 days or so. An alternative interpretation is that easterly current anomalies lead southward current anomalies by 20 days or so.

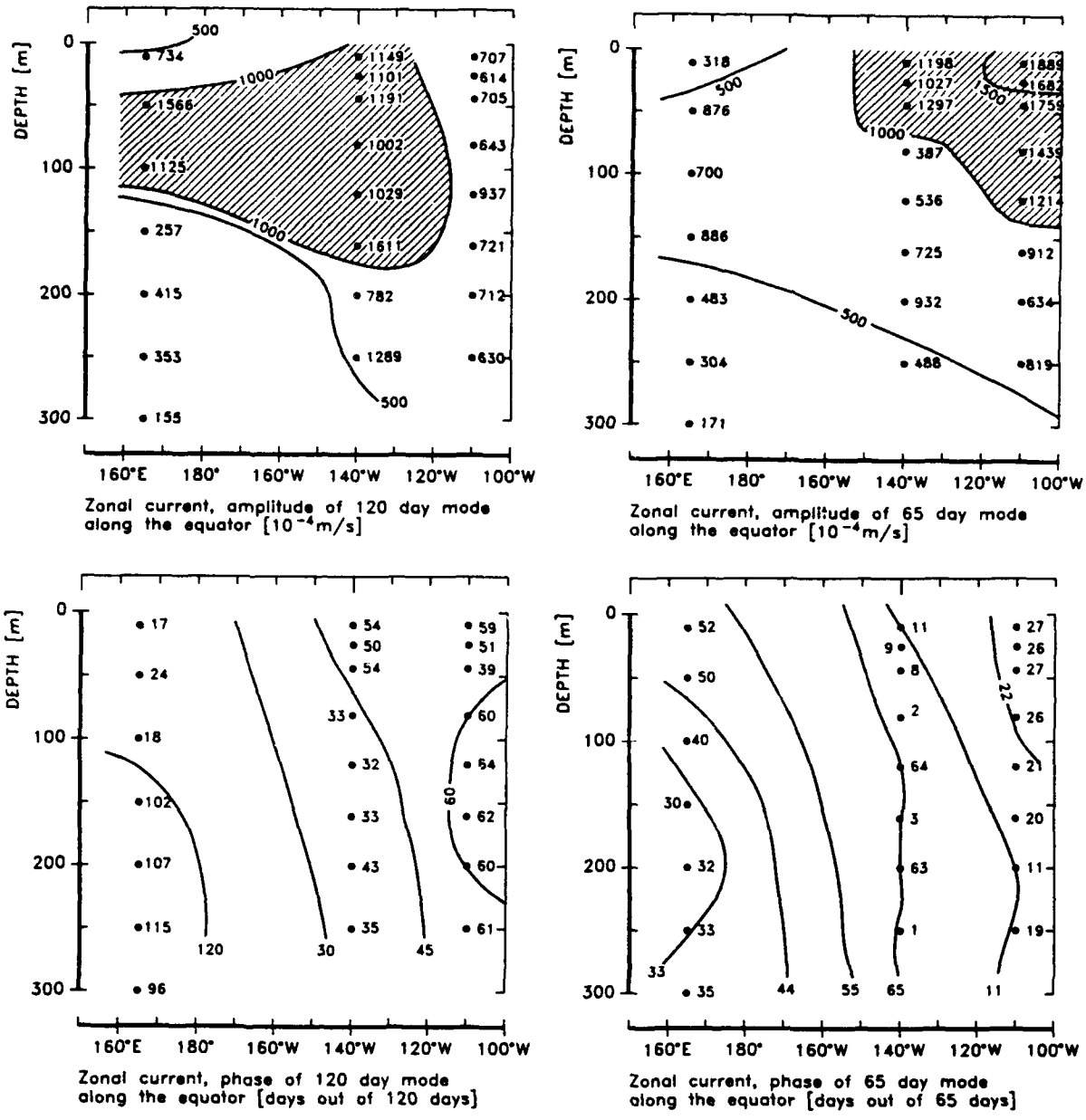


Figure 6. Longitude–depth cross-section of the zonal currents of the 120-day and 65-day mode along the equator. Top: The amplitudes A in 10^{-2} cm/s, and Bottom: The phases ψ in days (relative to base periods of 120 or 65 days).

Temperature at all buoys. For temperature, the amplitude distributions A and phase distributions ψ are shown as three cross sections through the tropical Pacific: a longitude-depth cross section along the equator (Fig. 7), a latitude-depth cross section at 110°W (Fig. 8), and a longitude-latitude cross section at 100 m (Fig. 9).

Maximum temperature amplitudes of both modes cluster along the thermocline (Fig. 3a) with maximum values of more than 1°C (Fig. 7). Overall, the temperature signal of the 120-day mode is stronger than that of the 65-day mode. The temperature signals propagate like the zonal current signals eastward along the equator. The 120-day temperature signal travels over the basin in about 90 days (relative to a base period of 120 days) so that the phase speed of temperature is 1.5 times that of the zonal current. At the 165°E buoy, the temperature and zonal current signals are almost in phase so that the later phase lags must stem from different travel times. The propagation of the temperature signal of the 65-day mode is mostly parallel to that of the zonal current signal but there is a uniform lag of about 10 days.

The latitude-depth cross sections of the associated correlation patterns at 110°W reveal maximum amplitudes of more than 1°C at about 100 m depth. In both modes are a marked amplitude minimum at 2°N and a maximum at 6°N. The activity of the 120-day mode is largest south of the equator, with a maximum amplitude of 1.4°C at 2°S, whereas the 65-day mode has its largest amplitude of 1.4°C at 6°N. Both modes exhibit complicated phase distributions. In the 120-day mode the phase varies mostly between 60 days at deeper levels and 90 days at upper levels. Only along the minimum at 2°N the phase is markedly lagging its neighborhood by 30 or more days. In the 65-day mode the maximum at 6°N is 180° out of phase with the temperature signal at the equator which, in turn, lags the secondary maximum at 2°S by another 10 to 15 days.

Figure 9 shows the latitude-longitude distributions of the amplitudes and phases of the two modes in 100 m depth. Maximum amplitudes of the order of 1°C at 140°W at the equator, where the thermocline is close to 100 m, tend to appear simultaneously with even larger (≈2°C) anomalies with opposite sign at 7°N. The eastward propagation is clearly visible in the 65-day mode, whereas in the 120-day mode the eastward propagation seems to be limited to the area west of 140°W. The isolated amplitude maximum at 5°N, 147°E should not be taken too seriously because of the shortness of the time series at that location (see Table 1).

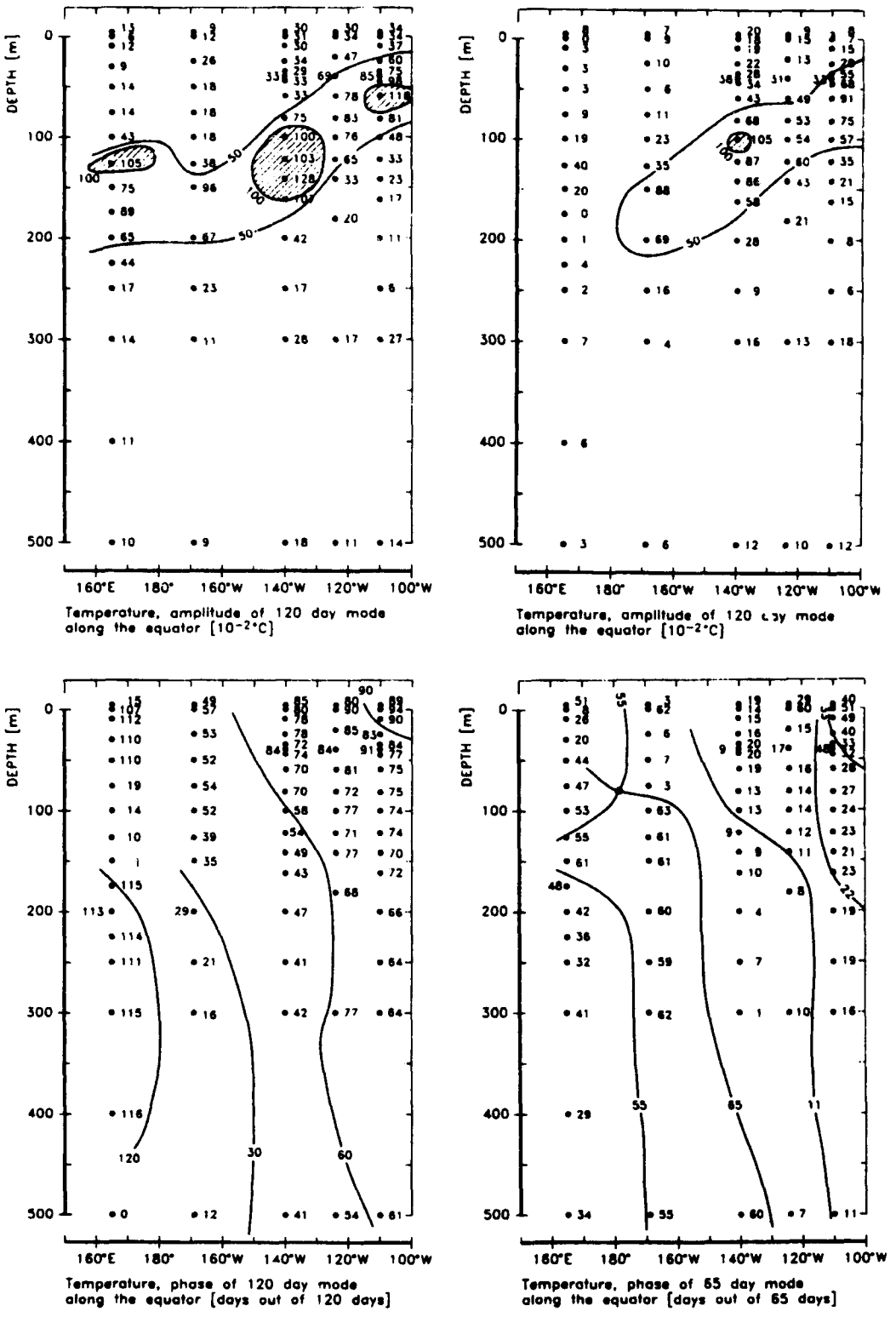


Figure 7. Longitude–depth cross-section of temperature of the 65 day mode and of 120 day mode along the equator. Top: The amplitude distributions A in 10^{-2} °C. Bottom: The phase distributions ψ (in days relative to the base periods of 120 and 65 days).

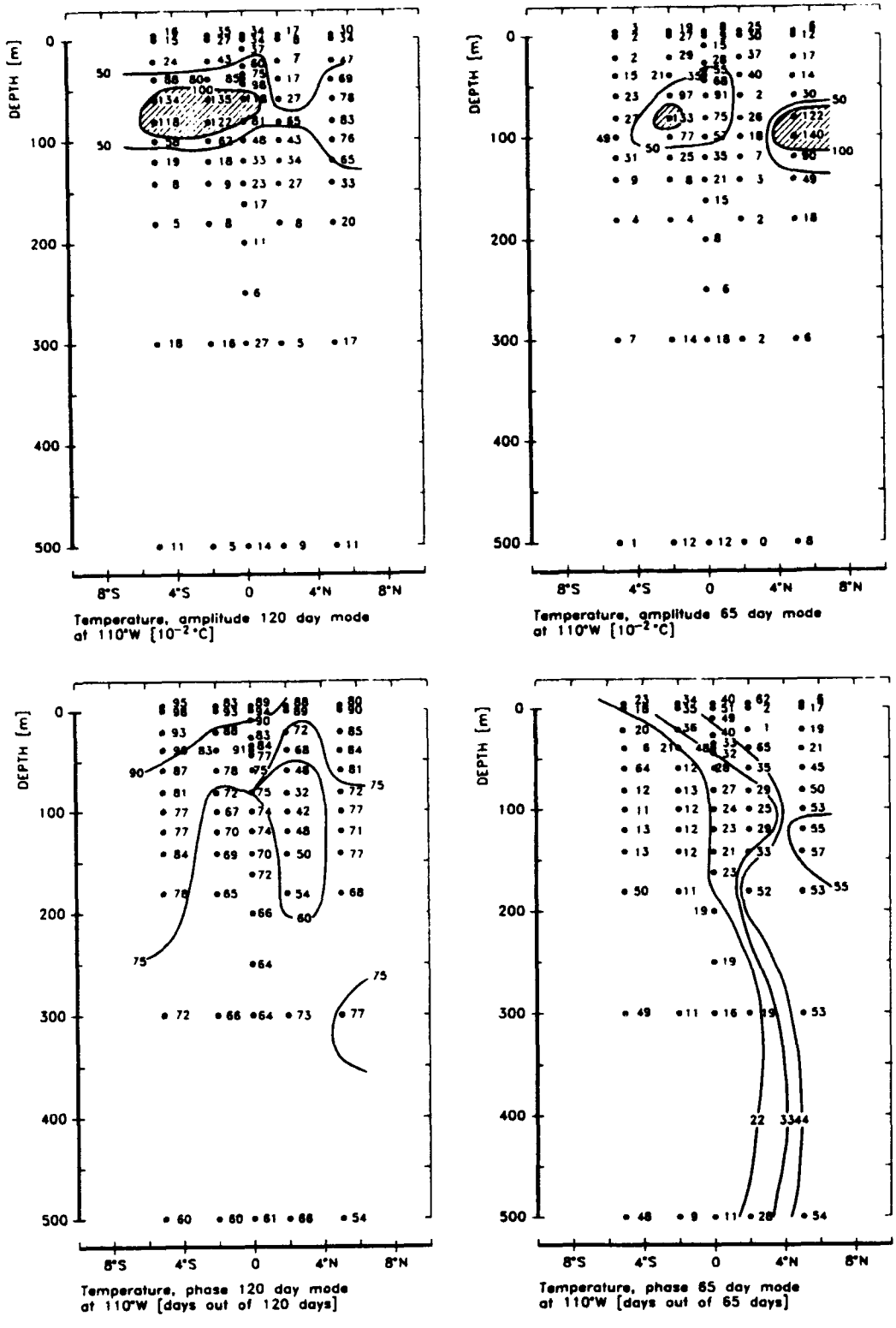


Figure 8. Latitude-depth distributions of temperature at 110°W and 140°W of the 120 day mode and of the 65 day mode. Top: Amplitude distributions A at 110°W (in 10^{-2} °C). Bottom: Phase distributions ψ at 110 °W (in days relative to the 120 day and 65 day base periods).

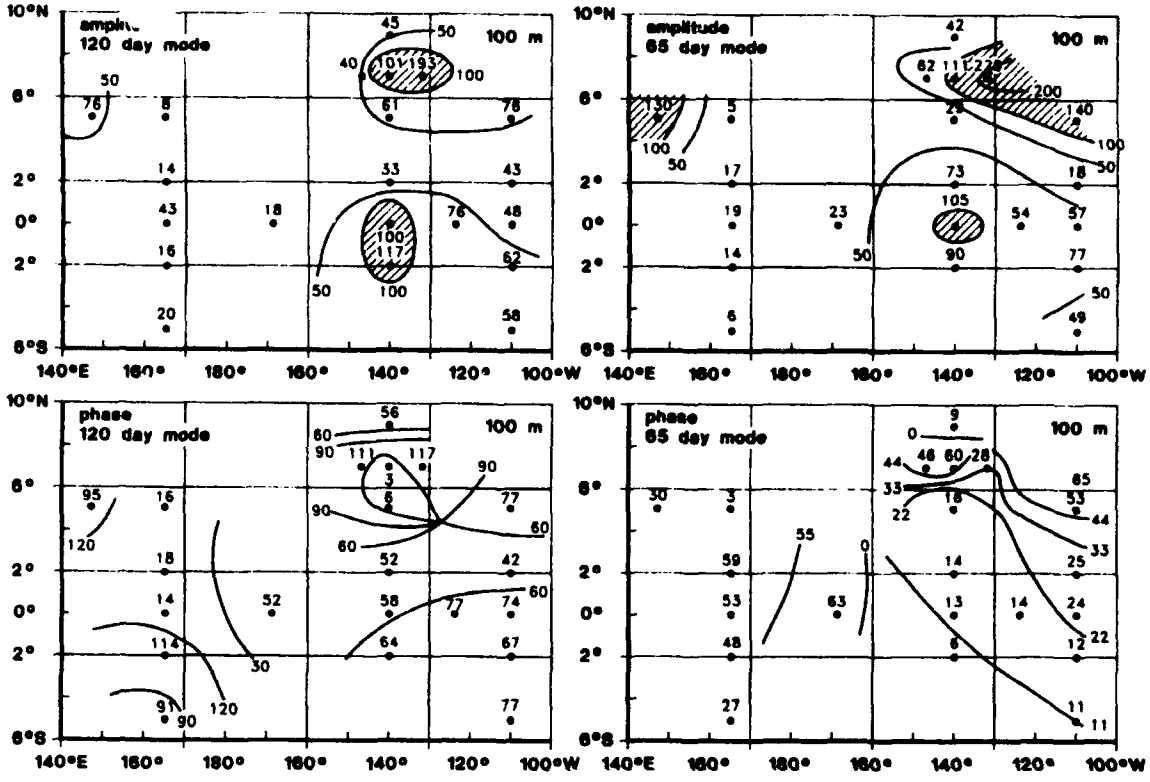


Figure 9. Horizontal distribution of temperature at a depth of 100 m of the 120 day mode and of the 65 day mode. Top: Amplitude distributions A in 10^{-2} °C. Bottom: Phase distributions ψ (in days relative to the 120 day and 65 day base periods).

Discussion: Equatorial temperature anomalies and advection. Because of the marked spatial gradients in the mean temperature field (Fig. 3) the temperature advection with the anomalous zonal currents might contribute significantly to the creation of temperature anomalies. Estimates of such temperature anomalies may be obtained for the equator since information on the currents is available there. If the anomalies are labelled by a * and the mean state by a $\bar{}$, then the effect of the anomalous currents on the temperature is approximated by

$$\mathbf{T}^{*} \approx \left[\mathbf{u}^{*} \frac{\partial \overline{\mathbf{T}}}{\partial x} + \mathbf{v}^{*} \frac{\partial \overline{\mathbf{T}}}{\partial y} \right] \times \frac{T}{2} \qquad (24)$$

with $\mathbf{T}$, $\mathbf{u}$, and $\mathbf{v}$ representing the temperature and zonal and meridional currents, and T the period; x refers to the zonal direction and y to the meridional direction. In the following we consider the situation at 140°W at 120 m depth.

The zonal gradient of the mean $\overline{\mathbf{T}}$ is approximately 2×10^{-8} K/cm (Fig. 3). For the 120 day mode the anomalous zonal current is 10 cm/s (Fig. 6) and the period is somewhere

between 80 to 120 days. Equation (24) yields with these numbers a temperature anomaly between 0.6 and 1.0°C, which compares well with the result of 1.0°C in Fig. 7. The 120 day mode is not connected with significant anomalies of the meridional current. Thus this back-of-the-envelope calculation Eq. (24) proposes that the equatorial temperature anomalies are due to anomalous zonal advection. This hypothesis is supported by the different travel times of the temperature and zonal current signal, which was found in a numerical experiment on the the response of the tropical Pacific to westerly wind bursts (Latif et al., 1988).

The typical zonal current anomalies of the 65 day mode are only 5.4 cm/s at 120 m (Fig. 6) and the characteristic time $T/2$ is only 32 days. Thus the effect of zonal advection is estimated as 0.3°C, which is significantly less than the predicted 0.9°C (Fig. 7). Thus zonal advection cannot fully explain the observed temperature anomalies—which is consistent with the coincidence of the temperature and zonal current travel times. The 65-day mode exhibits, however, a significant signal in the meridional current which could account for equatorial temperature anomalies of 0.3°C.

Kelvin waves? Are the modes identified and described so far what people call *Kelvin waves* (Moore and Philander, 1977)? The vertical structure of the modes along the equator, the horizontal scale, the eastward propagation and the time scale are broadly consistent with the concept of equatorial Kelvin waves. But several aspects are inconsistent with this concept. There are two modes, which have similar vertical structures, similar horizontal scales and time scales, that certainly cannot be accounted for as the first two Kelvin modes. The presence of a signal in the meridional signal in the 65 day mode does not fit the specification of a Kelvin wave nor has the rich structure found off the equator yet been described by the theory of equatorial Kelvin waves.

Johnson and McPhaden (1993) analyzed five years (1983-87) of current and temperature data from the 140°W and 110°W equatorial moorings and seven months of data from bouys at 2°S, 0° and at 2°N, 140°W. They used the complex empirical orthogonal functions (CEOFs, see also Section 4) and found one dominant mode that was broadly consistent with the idea of a first baroclinic Kelvin wave. The main differences from a conventionally defined Kelvin wave were these:

- A local maximum and a local minimum of the zonal velocity below and above the core of the equatorial undercurrent. This results holds for both modes identified in the POP analysis.
- An equatorial minimum of the temperature signal at the thermocline is straddled by two maxima at 2°S and 2°N. In the present POP analysis, on the other hand, the maximum at 2°S is reproduced, but north of the equator at 2°N a well-defined minimum is identified. Possibly Johnson and McPhaden's (1993) result is due to the short analysis period of only 210 days.

- A nonzero temperature signal at the surface lags the zonal current signal at the surface and the temperature signal at the thermocline by 90°. This result is confirmed by the POP analysis, in particular for the 120 day mode.

The biggest difference from Johnson and McPhaden (1993) is the presence of *two* modes which have uncorrelated coefficient time series but share substantial similarities in their spatial appearance. A reason for this difference might lie in the different analysis techniques. Johnson and McPhaden (1993) used CEOFs so that any two modes must be orthogonal in space whereas the POP analysis does not require orthogonality. If there are two orthogonal modes $(\mathbf{T}_i, \mathbf{u}_i)$ $(i = 1,2)$ with temperature signals $\mathbf{T}$ and zonal current signals $\mathbf{u}$ the orthogonality requires

$$\mathbf{T}_1^T \mathbf{T}_2 + \mathbf{u}_1^T \mathbf{u}_2 = 0. \tag{25}$$

Because of the sharp thermocline in the east equatorial Pacific the largest temperature anomalies will be centered around the thermocline so that $\mathbf{T}_1 \sim \mathbf{T}_2$. Thus to satisfy Eq. (25) a negative correlation of the current signals is needed, i.e., $\mathbf{u}_1 \sim -\mathbf{u}_2$. This latter condition represents a severe limitation without any physical justification. Therefore I speculate that the CEOF technique could not easily be used to identify two orthogonal modes in the equatorial $(\mathbf{T},\mathbf{u})$ data. This (admittedly handwaving) argument might help to resolve the apparent contradiction of only one mode in Johnson and McPhaden (1993) but two modes in the POP analysis. On the other hand, there is no support in the literature (as far as I know) for the idea of two non-orthogonal modes.

The 65 day mode is not envisaged by the theory of equatorial wave dynamics. This theory deals with the growth of small disturbances and not with the development or breakdown of finite amplitude disturbances. Schnur et al. (1993) have shown, for the case of synoptic-scale disturbances in the extratropical troposphere, that the POP analysis is an adequate tool to obtain not only the normal modes of a dynamical system but also modes that represent finite amplitude phases in the full spectrum of variability. I speculate that the 65 day mode might represent such a finite amplitude mode. It remains to be clarified if the results of this study will stand the test of more data, longer time series, and closer scrutiny. However, one has also to keep in mind that the present theory of equatorial Kelvin waves is based on a number of severe simplifications, one being the horizontal homogeneity of the background state.

4. CONCLUSIONS

The purpose of the present paper is two-fold. The main point is to introduce the POP technique to the oceanographic community. The minor point is to present first results from an analysis of data that are irregularly distributed in space and time.

The POP technique. The POP method is a powerful method to infer simultaneously the space-time characteristics of a vector time series. The basic idea is to isolate low-dimensional subsystems that are controlled by the linear dynamics of the full system. Even if the POP method represents the most consistent way of doing so, there are other techniques that can be used successfully for similar purposes. An alternative is the *complex empirical orthogonal functions* (CEOFs; Wallace and Dickinson 1972, Barnett and Preisendorfer 1981). CEOFs are obtained by applying the conventional EOF technique to a complex time series whose real part is the real time series that has to be analysed and whose imaginary part is the Hilbert transform of that real time series. (CEOFs are related to EOFs just like complex POPs to regular POPs). The main difference between CEOFs and POPs is that CEOFs are constructed under the constraint of a maximum of explained variance and mutual orthogonality. The characteristic times, the period and the damping time, are not an immediate result of the CEOF analysis but have to be derived a posteriori from the CEOF coefficient time series. The POPs, on the other hand, are constructed to satisfy a dynamical equation Eq. (11), and the characteristic times are an output of the analysis; also the complex POP coefficients $z(t)$ are not pairwise orthogonal. The non-orthogonality makes the mathematics less elegant, but it is not a physical drawback, because in most cases there is no reason to assume that different geophysical processes develop statistically independent from each other. The rate of variance explained by the POPs is not optimal and has to be calculated after the POP analysis from the POP coefficients.

The POP method is not a tool that is useful in all applications. If the analysed vector time series exhibit a strongly non-linear behaviour, as in turbulent flows, the POPs will fail to identify a useful sub-system, simply because a linear sub-system does not control a significant portion of the variability. The POP method will be useful if there are a priori indications that the processes under consideration are to a first approximation linear.

Equatorial Waves. We have found two modes of variability in the equatorial Pacific Ocean. The slower mode, with a nominal period of 120 days, resembles a first baroclinic Kelvin wave. The other, 65-day mode is different from theoretically derived modes and from previously empirically derived modes. More work is needed to ensure the reality and the signature of the two distinct modes.

Acknowledgments. The present work was mostly done during a two-month visit to the Joint Institute for Marine and Atmospheric Research in Honolulu. I want to thank Dennis Moore, Peter Müller and Gary Mitchum, among others, for a scientifically stimulating visit. The buoy data have been made available by the Director of the TOGA-TAO Project Office, Dr. Michael J. McPhaden. The buoy data were prepared by Artur Urbanowicz at the MPI. Marion Grunert prepared the professionally drawn diagrams.

References

Barnett, T.P. and R. Preisendorfer, 1981, Origins and levels of monthly and seasonal forecast skill for United States surface air temperature determined by canonical correlation analysis. *Mon. Wea. Rev.* 115, 1825-1850.

Blumenthal, B., 1991, Predictability of a coupled ocean-atmosphere model. *J. Climate* 4, 766-784.

Bürger, G., 1993, Complex Principal Oscillation Patterns. *J. Climate* (in press).

Gallagher, F., H. von Storch, R. Schnur and G. Hannoschöck, 1991, The POP Manual. Deutsches Klimarechenzentrum, Bundesstrasse 55, 2000 Hamburg 13, Germany.

Hasselmann, K.H., 1988, PIPs and POPs: The Reduction of Complex Dynamical Systems Using Principal Interaction and Oscillation Patterns. *J. Geophys. Res.* 93, 11 015-11 021.

Hayes, S.P., L.J. Mangnum, J. Picaut, A. Sumi and K. Takeuchi, 1991, TOGA-TAO: A moored array for real time measurements in the Tropical Pacific Ocean. *Bull. Am. Met. Soc.* 72, 339-347.

Johnson, E.S. and M.J. McPhaden, 1993, On the structure of intraseasonal Kelvin waves in the Equatorial Pacific Ocean. *J. Phys. Oceanogr.* 23, 608-625.

Latif, M., J. Biercamp and H. von Storch, 1988, The response of a coupled ocean-atmosphere general circulation model to wind bursts. *J. Atmos. Sci.* 45, 964-979.

Latif, M. and, M. Flügel, 1990, An investigation of short range climate predictability in the tropical Pacific. *J. Geophys. Res.* 96, 2661-2673.

Latif, M., A. Sterl, E. Maier-Reimer and M.M. June, 1993, Climate variability in a coupled GCM. Part I: The tropical Pacific. *J. Climate* 6, 5-21.

Latif, M., and A. Villwock, 1989, Interannual variability in the tropical Pacific as simulated in coupled ocean-atmosphere models. *J. Marine Sys.* 1, 51-60.

Mikolajewicz, U. and E. Maier-Reimer 1991, One example of a natural mode of the ocean circulation in a stochastically forced ocean general circulation model. In: *Strategies for Future Climate Research* (Ed. M. Latif), 287-318, Max-Planck-Institut für Meteorologie Hamburg.

Moore, D. and G. Philander, 1977, Modelling of the tropical oceanic circulation. In: *The Sea* 6, 319-361, John Wiley and Sons Inc.

Schnur, R., G. Schmitz, N. Grieger and H. von Storch, 1993, Normal Modes of the atmosphere as estimated by principal oscillation patterns and derived from quasi-geostrophic theory. *J. Atmos. Sci.* (in press).

von Storch, H., T. Bruns, I. Fischer-Bruns and K. Hasselmann, 1988, Principal Oscillation Pattern analysis of the 30- to 60-day oscillation in a General Circulation Model equatorial troposphere. *J. Geophys. Res.* 93, 11 022-11 036.

von Storch, H., and J. Xu, 1990, Principal Oscillation Pattern analysis of the tropical 30- to 60-day oscillation. Part I: Definition on an index and its prediction. *Clim. Dyn.* 4, 175-190.

von Storch, H., U. Weese and J. Xu, 1990, Simultaneous analysis of space-time variability: Principal Oscillation Patterns and Principal Interaction Patterns with applications to the Southern Oscillation. *Z. Meteor.* 40, 99-103.

von Storch, H., and D. Baumhefner, 1991, Principal Oscillation Pattern analysis of the tropical 30- to 60-days oscillation. Part II: The prediction of equatorial velocity potential and its skill. *Clim. Dyn.* 5, 1-12.

von Storch, H., and A. Smallegange, 1991, The phase of the 30- to 60-day oscillation and the genesis of tropical cyclones in the Western Pacific. Max Planck Institut für Meteorologie Report 64 (Max-Planck-Institut. für Meteorologie, Bundesstrasse 55, 2000 Hamburg 13, Germany).

Wallace, J.M. and R E. Dickinson, 1972, Empirical orthogonal representation of time series in the frequency domain. Part I: Theoretical considerations. *J. Appl. Meteor.* 11, 887- 892.

Weisse, R., U. Mikolajewicz and E. Maier-Reimer, Decadal variability of the Northern North Atlantic in an Ocean General Circulation Model. *J. Geophys. Res.* (in press).

Wright, P., 1985, The Southern Oscillation—An ocean-atmosphere feedback system? *Bull. Am. Met. Soc.* 66, 398-412.

Xu, J., 1990, Analysis and prediction of the El Niño Southern Oscillation phenomenon using Principal Oscillation Pattern Analysis. Max Planck Institut für Meteorologie Examensarbeiten 4 (Max-Planck-Institut für Meteorologie; Bundesstrasse 55; 2000 Hamburg 13, Germany).

Xu, J., 1992, On the relationship between the stratospheric QBO and the tropospheric SO. *J. Atmos. Sci.* 49, 725-734.

Xu, J., 1993, The observed global low frequency variability in the atmosphere-ocean system from 1967 to 1986. *J. Climate* 6, 816-838.

Xu, J. and H. von Storch, 1990, "Principal Oscillation Patterns"-prediction of the state of ENSO. *J. Climate* 3, 1316-1329.

ILLUSTRATING FREQUENTIST AND BAYESIAN STATISTICS IN OCEANOGRAPHY

George Casella, Biometrics Unit, Cornell University, Ithaca, NY 14853

ABSTRACT

Both frequentist and Bayesian methodologies provide means for a statistical solution to a problem. However, it is usually the case that, for a given situation, one methodology is more appropriate. Using a number of oceanographic examples we explore the components of a statistical solution and illustrate the most appropriate methodology. We argue that the statistical consideration of utmost importance is the type of inference and conclusion to be made. In some examples it is more appropriate to make this inference as a Bayesian, and in some it is more appropriate to make this inference as a frequentist.

"Still, it is an error to argue in front of your data. You find yourself insensibly twisting them round to fit your theories."

Sherlock Holmes
The Adventure of Wisteria Lodge

1. INTRODUCTION

An alternate title for this paper might well be "Conditional and Unconditional Inference in Oceanographic Studies," as a fundamental difference between frequentist and Bayesian statistics is their resulting inference. A frequentist inference is unconditional, applying to a series of repeated experiments (most always an imagined series). In contrast, a Bayesian inference is conditional, applying to the data at hand, and not directly addressing the concept of repeatability.

This paper is an introduction to these methods and illustrates their uses with some oceanographic data sets. The primary message is that each statistical view has a lot to offer, and, depending on the problem, one methodology is probably more appropriate. We illustrate this through the examples.

A second goal of this paper is to try to explain to the oceanographic community how a statistician approaches a problem. The purpose of this endeavor is to provide a structured approach to dealing with problems involving data, from their inception to ending. In doing so, perhaps the task of dealing with the ever-increasing data bases can be made a little easier.

The remainder of the paper is arranged as follows. In Section 2 we give a general outline of how to approach a problem statistically, illustrating this with an example in Section 3. Section 4 discusses the underlying differences between the frequentist and Bayesian approaches to statistics, and Sections 5 and 6 contain more examples illustrating these methodologies. Section 7 is a concluding discussion.

2. COMPONENTS OF A STATISTICAL SOLUTION

In the best of all possible worlds, a problem is planned, statistically, from beginning to end. Chronologically, the steps of a solution are these:

1. **Model** the Process
2. **Design** the Experiment
3. **Collect** the Data
4. **Estimate** and Verify the Model.
5. **Infer** and Conclude
6. **Implement** the Solution

Although the steps are performed in chronological order, they are best planned in reverse order. That is, when approaching any problem, the first consideration is "How will the knowledge we gain be implemented?" For example, if a study is proposed to examine wave magnitude and direction in the North Atlantic, the first consideration should be the use of the resulting knowledge. Will it be used to plan routes for oil tankers? Will it be used to increase our basic knowledge of ocean dynamics? By answering these questions first, the remainder of the steps of a statistical solution will fall into place, and the problem can be attacked in a very efficient fashion. Although this mechanism for solution is not usually taught in the classroom, it seems to be the one most preferred by statisticians. By concentrating on the final result, the entire study becomes focused.

With respect to frequentism or Bayesianism, the components of the statistical solution remain essentially unchanged. Of course, there are some differences in the approaches, with the major difference being in the modeling and inference stages. However, the overall attack is similar and is illustrated in the next section.

3. AN EXAMPLE CONCERNING ICEBERGS

Defant (1961, page 278) presented the following data on the frequency of icebergs off Newfoundland.

Table 1: Frequency of icebergs off Newfoundland south of 48°N (a) and south of the Grand Banks (b) for the period 1900-1926.

	Month												
	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	
(a)	3	10	36	83	130	68	25	13	9	4	3	2	386
(b)	0	1	4	9	18	13	3	2	1	0	0	0	51

For our example, we will look at the question of whether the yearly distribution of icebergs is the same in each location. A glance at Figure 1 will show that such a hypothesis is very likely, but for illustration we will step through both a Bayesian and a frequentist approach to the problem. We take as the goal of our study to be the description of the distribution of icebergs off Newfoundland.

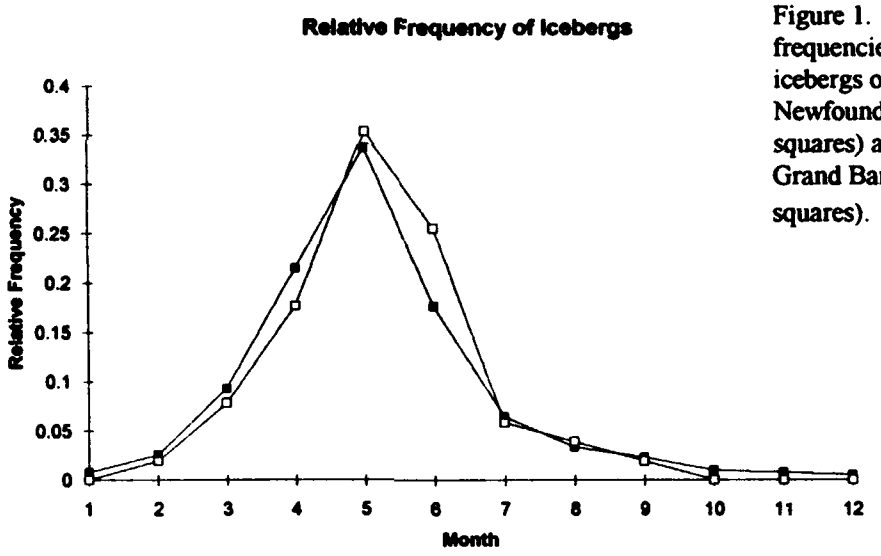


Figure 1. Relative frequencies of icebergs off (a) Newfoundland (solid squares) and (b) Grand Banks (open squares).

In both the Bayesian and frequentist approaches to this problem we assume that the data are distributed according to a multinomial distribution, and we wish to test the null hypothesis H_0: The distributions in locations (a) and (b) are the same. To test this as a frequentist we use a chi-squared test of association (see Snedecor and Cochran, 1989). The chi-squared test results in a p-value of 0.977, which is very strong evidence in favor of the null hypothesis.

To perform a Bayesian analysis a prior distribution must be specified, that is, a distribution that we subjectively believe describes the pattern of icebergs. We then use this distribution, in conjunction with the observed data, to assess the plausibility of the hypothesis. Since we really have no prior knowledge about the icebergs, we use a strategy

that attempts to model this ignorance and calculate the probability of every data table with the given marginal totals, using a hypergeometric distribution. This leads us to use Fisher's exact test (Fisher, 1970) and assess the probability of the null hypothesis as 0.994. Again, this is very strong evidence in favor of this hypothesis. (Strictly speaking, Fisher's exact test is not a Bayesian procedure but a conditional procedure, as it is calculated conditionally on the observed data. However, the important feature is that it yields a conditional inference.)

We now can clearly see the distinction between Bayesian and frequentist inferences. The frequentist bases inference on a frequency interpretation. A formal conclusion would be of the form, "the statistical procedure used (here the chi-squared test) would result in an erroneous inference less than 5% of the time in repeated experiments." In contrast, the Bayesian inference is conditional on the observed data, and we would formally conclude "based on the stated prior distribution and observed data, the probability is 0.994 that H_0 is true." We now look at these differences more closely.

4. WHERE DOES THE RANDOMNESS COME FROM?

The most important part of any statistical investigation is the resulting inference. In fact, it may even be said that the main reason for doing a statistical investigation is to produce a meaningful inference, because the inference applies to a wider population than is actually studied and measured. (For example, after measuring the activities of a number of waves in a certain area, we are then interested in making a statement (an inference) about all waves in that area.) To make this inference we need an underlying model of the phenomena, one that accounts for the randomness of the observations and allows an inference. Bayesians and frequentists have different approaches to this.

4.1 Frequency Randomness

The frequentist assumes repeatability of the experiment, that the experiment actually performed is one of an infinitely long sequence of identical experiments. If we denote this sequence of experiments $E_1, E_2, \ldots, E_k, E_{k+1}, \ldots$, then we make our inference to the entire sequence, even though only one experiment (say E_k) is actually performed. The rest of this imagined sequence builds the randomness into our model. We know that the results of each experiment (if performed) would be slightly different, and our inference will take these potential differences into account.

Thus, the frequentist inference is an unconditional one that applies to the entire sequence and does not single out the experiment actually performed. It is important to realize that the inference is about the performance of the *procedure* over the entire sequence of experiments, such as "The statistical procedure used will be correct in 95% of all

experiments performed." The actual outcome of the observed experiment will not change this inference.

4.2 Bayesian Randomness

In a Bayesian analysis the data are assumed to be fixed, and inference is made conditional on their observed values. Thus, no randomness comes from the data. The randomness in a Bayesian inference comes from the subjective prior distribution. This randomness, together with the information in the data, is combined into the posterior distribution. The posterior distribution is then used for inference. Of course, different subjective prior distributions may result in different inferences.

More precisely, suppose there are data, X, which vary according to a probability distribution $f(x|\theta)$, a distribution indexed by an unknown parameter θ. (For example, $f(\cdot|\theta)$ may be a Gaussian distribution with unknown mean θ.) We then assume that the parameter θ varies according to a prior distribution $\pi(\theta)$. This probability distribution reflects our knowledge about the parameter θ before observing the new data x. (In keeping with convention, an upper case X denotes an unseen random variable whereas a lower case x denotes an observed value. Thus the equation "$X = x$" means that we have observed the value x of the random variable X.) Using the laws of probability (or sometimes called Bayes rule) we calculate the posterior distribution of θ given $X = x$, $g(\theta|x)$ as

$$g(\theta|x) = \frac{f(x|\theta)\,\pi(\theta)}{\int f(x|\theta)\,\pi(\theta)\,d\theta},$$

where the integral is over all values of θ. (For more detail on such calculations, see Casella and Berger, 1990.) Our inference is then based on $g(\theta|x)$, which only considers the experiment actually performed, not any repeated sequence. For example, one might infer "On the basis of the specified $\pi(\theta)$ and observed x, we conclude that $\theta \geq 0$ with probability 95%." This inference would follow if it were the case that $\int_0^\infty g(\theta|x)\,d\theta = 0.95$.

4.3 The Appropriate Inference

As mentioned before, the purpose of this paper is not to make value judgments as to whether Bayesianism or frequentism is better. Rather, the purpose is to illustrate situations where one method is more appropriate. It then follows that the more appropriate methodology, and inference, is the one to use. From the previous two subsections, we see that the frequentist inference is more appropriate if repeatability is important, whereas the Bayesian inference is more appropriate if the inference is to be made conditional on the observed data. Returning to the iceberg data, it seems that the Bayesian inference is more

appropriate, as we are faced with a data set that is unrepeatable, and we are interested in an inference conditional on that data set. (Interestingly, it was argued during discussions at the workshop that one could consider the observed 26-year period as one of a sequence of 26-year periods, in which case the frequentist inference maybe more appropriate.) If it may be argued that either interpretation is valid, and hence either inference is appropriate, there is no problem. As long as the methodology is chosen to appropriately answer the question of interest, phrased in the manner of interest, the statistics have served their purpose.

5. AN EXAMPLE CONCERNING BREAKING WAVES

Hwang et al. (1990) report on an experiment concerning average height of breaking waves, H_B, measured as a function of RMS surface displacement, η. The data are presented in Figure 2. They conclude that $H_B<H_S$, the significant wave height, where $H_S = 4\eta$, and state, "In a random wave field, waves that break due to local instabilities are not necessarily the highest waves." Statistically, we can think of this as testing the hypotheses

$$H_0: \ H_B \le 4\eta \quad \text{vs} \quad H_1: \ H_B > 4\eta.$$

It seems here that frequency considerations are important, in that conclusions should apply to repetitions of the experiment. This concern seems implicit in the above quoted conclusion of Hwang et al. Thus, a frequentist analysis is more appropriate. Using a standard linear regression model with Gaussian errors, we obtain a p-value of 0.999 for the hypothesis H_0: $H_B \le 4\eta$, showing that there is overwhelming evidence to support

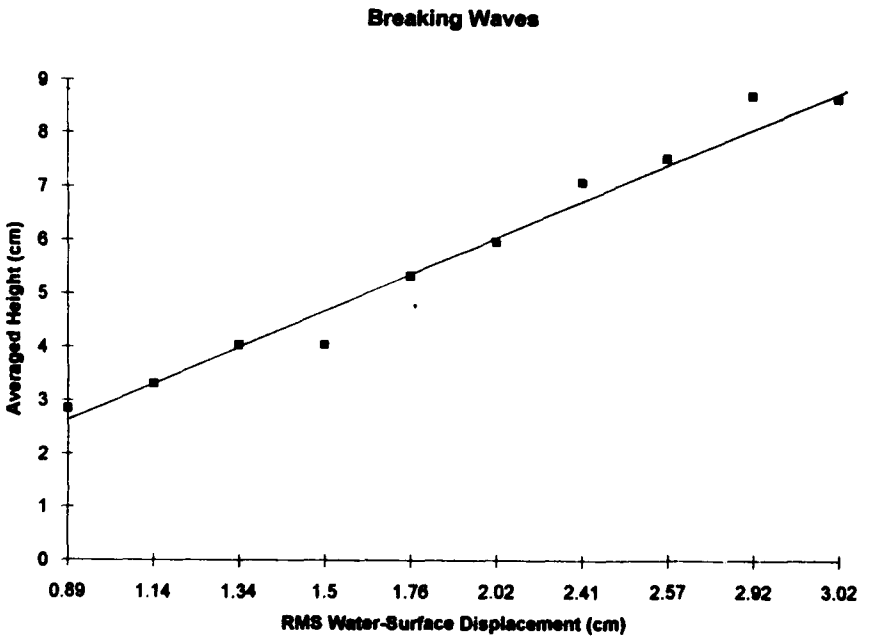


Figure 2. Averaged height of breaking waves, H_B, as a function of RMS water surface displacement, η. The line shown is the least squares line, with equation $H_B = 0.102 + 2.89\eta$ ($r^2 = 0.994$).

this hypothesis. (In fact, the hypothesis H_0: $H_B \leq 3\eta$ yields a p-value of 0.911, demonstrating extremely good support for this even stronger claim.)

Of course, a Bayesian analysis could also be performed, but the inference would not apply to a sequence of experiments. The conclusions would be conditional on the observed data. To do the Bayesian analysis we again use a standard linear regression model with Gaussian errors, but we also assume that $H_B = b\eta$, where b is a parameter with a specified prior distribution. We specify the prior to also be Gaussian, and we take the prior mean to be equal to the hypothesized value. (Thus, for testing H_0: $H_B \leq 4\eta$; we specify a Gaussian prior with mean 4. This strategy of centering the prior at the hypothesized value gives equal prior weight above and below the value, and may be considered an impartial prior specification.)

Combining our prior specification with the observed data, we calculate $\Pr(b \leq 4 | \text{ data}) = 0.999$ and $\Pr(b \leq 3 | \text{ data}) = 0.623$. That is, for the specified priors and conditional on the observed data, b is less than 4 with probability 0.999 and less than 3 with probability 0.623. Quantitatively, these conclusions are similar to those of the frequentist, and show overwhelming support for the null hypotheses. The only difference is in the scope of the inference.

Bayesian conclusions are, of course, dependent on the prior specification, and sometimes there might be concern about oversensitivity to this specification. Such a concern is easily addressed, however, by calculating posterior probabilities over a range of prior specifications. This is illustrated in Figure 3, where we display the posterior probabilities over a wide range of standard deviations. (The standard deviation of the data is 0.082, and the graph shows the prior standard deviation up to twice this value.) The figure shows that, for this range of prior standard deviations, the conclusions from the Bayesian analysis are relatively stable in their support of H_0.

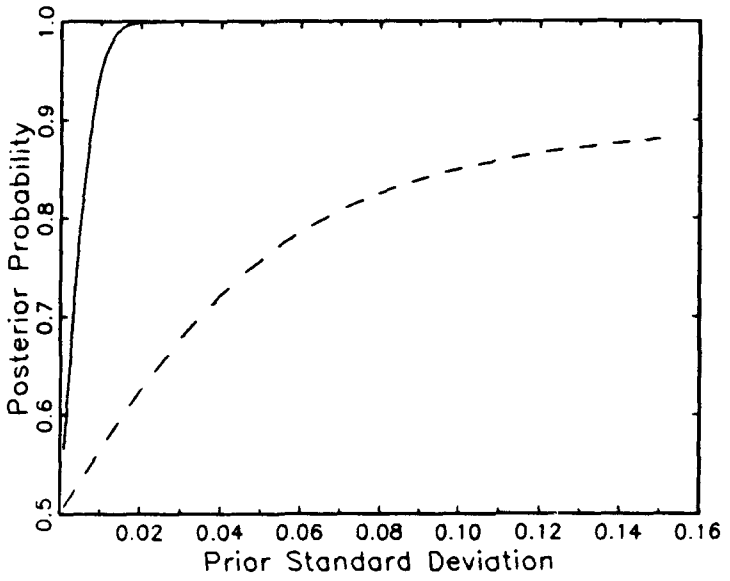


Figure 3. Posterior probabilities for the null hypotheses H_0: $b \leq 4$ (solid lines) and H_0: $b \leq 3$ (dashed lines), as a function of the prior standard deviation.

6. AN EXAMPLE CONCERNING BUBBLE POPULATIONS

The distribution of bubble populations is also investigated by Hwang et al. (1990). They collected data on bubble populations as a function of depth and wind velocity, as presented in Figure 4. For a given depth Z (cm) and wind velocity u (m/s), the logarithm of the bubble population $N(Z)$ (log cm^3) is modeled as

$$N(Z) = a_u + b_u Z + \epsilon \qquad u = 10,11, \ldots, 15$$

where ϵ represents random error and is assumed to have Gaussian distribution with mean 0 and variance σ^2.

A question of interest is whether the distribution of bubbles is the same at each depth. After some thought, it seems that the appropriate inference here is the frequentist inference. Concern about the repeatability of the inference leads to this conclusion, as we would like to be able to describe the bubble populations at a given depth and wind velocity when such conditions are again realized.

6.1 A Standard Frequentist Inference

A standard approach to this problem is to decide if the slopes are the same at each wind velocity, so we would test the null hypothesis H_0: $b_{10} = b_{11} = \cdots = b_{15}$. Doing so leads to a p-value of 0.063, which suggests rejection of H_0. Thus a standard frequentist analysis would lead us to fit separate regression lines for each wind velocity. So for each wind

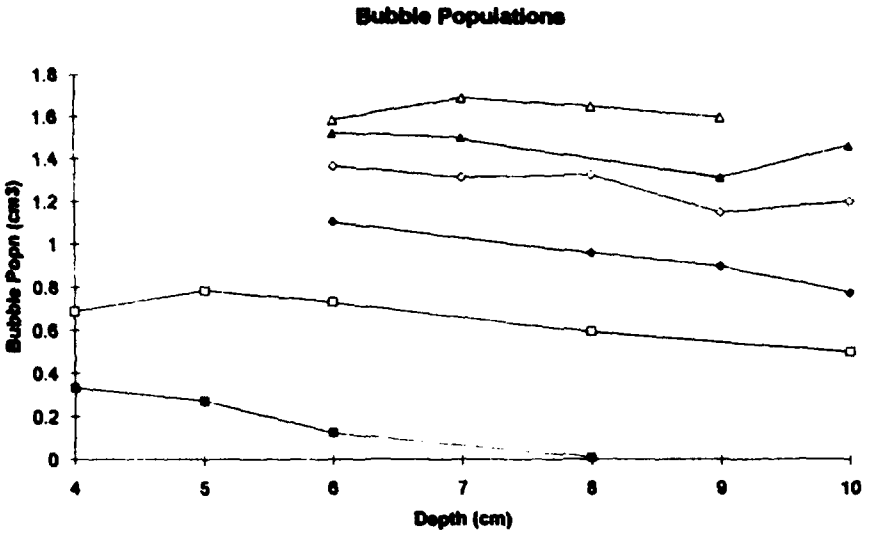


Figure 4. The groups are in order from lowest to highest wind velocity and are denoted by solid squares (10 m/s), open squares (11 m/s), solid diamonds (12 m/s), open diamonds (13 m/s), solid triangles (14 m/s) and open triangles (15 m/s). The data are connected merely to aid viewing. Data from Hwang et al. (1990).

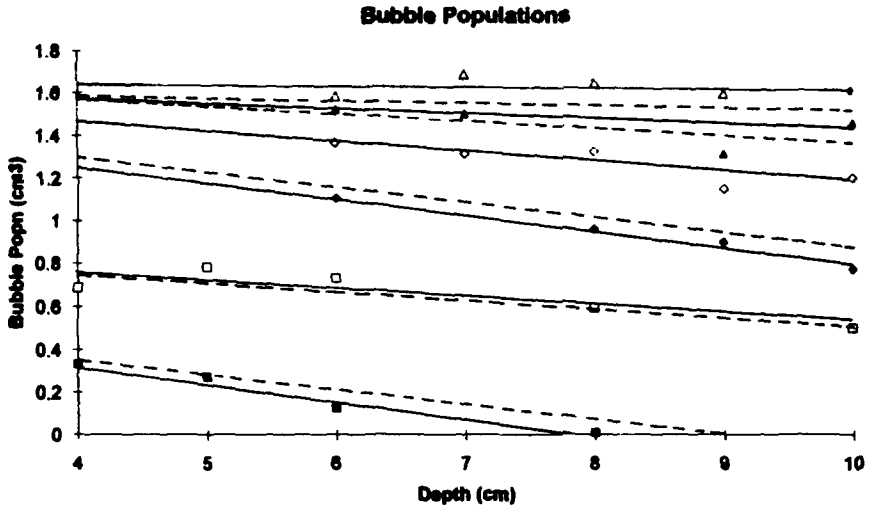


Figure 5. Standard frequentist (solid lines) and empirical Bayes (dashed lines) fits to the bubble data, coded as in Figure 4. The empirical Bayes lines (whose slopes are pulled toward -0.048) are under the least squares lines for 11, 14, and 15 m/s, and above the least squares lines for 10 and 12 m/s. The lines are virtually identical for 13 m/s.

velocity we would use a separate regression equation to predict the bubble population. See Table 2 and Figure 5.

6.2 An Empirical Bayes Analysis

The bubble population data are ideal for an empirical Bayes analysis—a mixture of frequency and Bayesian analyses that combines the best features of each. Here we will only briefly explain the methodology, for a more detailed introduction see the articles by Casella (1985, 1992).

To perform an empirical Bayes analysis we start with the frequentist model and inference structure. We append a Bayes model to the slopes

$$b_u \sim \text{Gaussian}\,(b, \tau^2), u = 10, 11, \cdots, 15,$$

that is, that the slopes come from a common Gaussian population with unknown mean b and variance τ^2.

The "empirical" part of empirical Bayesian is to now estimate these unknown parameters b and τ^2 from the data. (A standard Bayesian analysis would specify values for these parameters.) Using these estimated values allows the data to assess the tenability of the *submodel*, that the b_u's come from a common population. The empirical Bayes slope estimates are a convex combination of the common overall slope (-0.048) and the individual least squares slopes, given by

empirical Bayes slope = (0.221) (-0.048) + (0.0779) (least squares slope).

The weighting factors 0.221 (and 0.779 = 1 – 0.221) are data based estimates. The empirical Bayes slope estimates are valid under the model of frequentist repeatability. In fact, they are superior to the frequentist estimates using a criterion of expected mean squared error. Thus, on the average, the empirical Bayes estimates will be closer to the true values than the standard frequentist estimates. They combine the best features of Bayesian modeling and frequentist inference.

Figure 5 also shows the empirical Bayes regression lines. Although they are not very different from the standard frequentist lines, they do display a movement toward the common slope value. The empirical Bayes analysis has uncovered a small amount of common structure and has used this in improving each of the estimates.

Table 2: Coefficients for the standard regression analysis (frequentist) and empirical Bayes analyses of the bubble populations.

Wind Velocity	n	intercept	slope	std. dev.	empirical Bayes slope
10	4	0.666	-0.084	0.011	-0.076
11	5	0.924	-0.040	0.013	-0.042
12	4	1.594	-0.080	0.008	-0.073
13	5	1.669	-0.050	0.017	-0.050
14	4	1.698	-0.031	0.029	-0.035
15	4	1.635	-0.0009	0.027	-0.011

7. CONCLUSIONS

The statistical methodology to be used, whether Bayesian or frequentist, should be selected according to the type of inference that is desired (and is appropriate). The frequentist methodology is appropriate for inference over a series of repeated experiments, while the Bayesian methodology is appropriate for inference specific to the experiment that was done. This article has given examples and provided discussion of situations where each methodology is appropriate.

There is no brick wall between Bayesianism and frequentism. The methodologies are not at odds with one another; they are complementary to one another. When approaching a statistical problem "opportunism" is best. With that in mind, the appropriate analysis and inference can be chosen from all available statistical methodologies.

Both Bayesianism and frequentism are built on a set of assumptions, some more palatable than others. For a user of frequentist methods, perhaps the assumption most difficult to believe is that the process (including parameter values) remains constant over the imagined series of experiments. For a user of Bayesian methods, perhaps the assumption most difficult to believe is that the prior distribution is correct. These assumptions, however, can sometimes be checked and and maybe even relaxed. Moreover, their reasonableness in any particular situation may also form a basis for choosing an appropriate methodology. [See Berger's (1985) discussion of robust Bayesian analysis, which addresses these concerns]. Lastly, there is an enormous amount of research being done in statistics, and some of it is aimed at relaxing these assumptions. Such research has already given us techniques like empirical Bayes analysis, a synthesis of both Bayesian and frequentist methodologies which can often provide superior solutions.

This paper is technical report BU-1187-M, in the Biometrics Unit, Cornell University. This research was supported by National Science Foundation Grant No. DMS9100839 and National Security Agency Grant No. 90F-073.

REFERENCES

Berger, J.O. (1985): *Statistical Decision Theory and Bayesian Analysis*, Second Edition. New York: Springer-Verlag.

Casella, G. (1985): An Introduction to Empirical Bayes Data Analysis. *The American Statistician*, 39, 83-87.

Casella, G. (1992): Illustrating Empirical Bayes Methods. *Chem. and Intell. Lab. Sys.*, 16, 107-125.

Casella, G. and Berger, R.L. (1990): *Statistical Inference*, Pacific Grove: Wadsworth and Brooks/Cole.

Defant, A. (1961): *Physical Oceanography*, Volume I. New York: Pergamon Press.

Fisher, R.A. (1970): *Statistical Methods for Research Workers*, Fourteenth Edition. New York: Hafner (Reissued by Oxford University Press, 1990).

Hwang, P.A., Hsu, Y.-H.L., and Wu, Jin. (1990): Air Bubbles Produced by Breaking Wind Waves: A Laboratory Study, *J. Phys. Oceanogr.*, 20, 19-28.

Snedecor, G.W. and Cochran, W.G. (1989): *Statistical Methods*, Eighth Edition. Ames: Iowa State University Press.

BAYESIAN METHODS: AN INTRODUCTION FOR PHYSICAL OCEANOGRAPHERS

Joseph B. Kadane
Department of Statistics, Carnegie Mellon University, Pittsburgh, PA 15213

> "You could not step twice into the same river, for new waters are ever flowing on to you," Heraclitus, as quoted in Bartlett (1980).

ABSTRACT

The Bayesian approach to statistics is a conceptually simple method of treating uncertainty. It involves modeling uncertainty with probability, and conditioning on such data as become available. Because of its flexibility, there are many styles of application. Using the same examples as George Casella's paper in this volume, I discuss how this Bayesian method approaches such problems.

1. A GENERAL INTRODUCTION TO BAYESIAN IDEAS

Most statistical analyses begin with some data, denoted x, and a parameter, denoted θ. These may be discrete or continuous, and may have vector, matrix, or more complex structures. For the purposes of this section, the nature of x and θ do not matter, but in application they are very important.

The mechanism generating the data is called the likelihood function, and is written $f(x|\theta)$. Here f may be a probability mass function, if x is discrete, or a probability density, if x is continuous. In both cases it describes the probabilistic behavior of the data, x, for a fixed value of the parameter θ. The second part of a statistical model is a prior distribution $\pi(\theta)$, which again might be a probability mass function if θ is discrete, or a probability density if θ is continuous.

These two ingredients determine the joint distribution of x and θ as follows:

$$h(x,\theta) = f(x|\theta)\pi(\theta) \tag{1}$$

Once the data x are observed, the laws of probability prescribe how the conditional distribution of θ given x is to be calculated:

$$g(\theta|x)=\frac{h(x,\theta)}{p(x)}=\frac{h(x,\theta)}{\int_\Omega h(x,\theta)d\theta}=\frac{f(x|\theta)\pi(\theta)}{\int_\Omega f(x,\theta)\pi(\theta)d\theta} \quad (2)$$

The distribution is called the posterior distribution of θ, because it is the distribution of θ after having observed x. Thus the import of the data is to change the distribution of θ from the prior, $\pi(\theta)$, to the posterior, $g(\theta \mid x)$. Everything in this paper is a discussion or an application of these ideas.

The essential idea here is the use of probability to express uncertainty. Having decided to do that, formula (2) follows from formula (1) by very simple and non-controversial steps.

One important matter is the interpretation given to probability here; whose probabilities are these? Although there are some Bayesians who would give other answers, the dominant answer now, (and the one to which I subscribe) is that these probabilities are subjective, and reflect the opinion of the writer, or opinions the writer believes others hold. Bayesians do not come to this view gladly. We wish there were a way to guarantee that the equations written capture the objective truth, but such guarantees do not seem possible. We observe in science disagreements in which none of the sides has made a provable, mathematical error. The progress of a science might then be thought of as the development of informed opinion on a subject.

The name "Bayesian" incidentally, is in honor of Rev. Thomas Bayes, an eighteenth century English minister and "natural philosopher." He found the principle now embedded in (2), and hence this way of thinking about and doing statistics is named for him.

Finally, note that the quantities x and θ are simply random variables with some joint distribution, although one is written with a Roman letter and one with a Greek letter. If one began with the joint distribution $h(x,\theta)$ and learned θ, the posterior on x given θ would be $f(x|\theta)$, and would represent what was known about x after θ had been learned. Thus the model is symmetric in x and θ, although to encourage intuition it is customary to think of the former as data and the latter as a parameter.

In the remainder of this paper, I discuss Casella's iceberg example in section 2, breaking waves in section 3 and bubble data in section 4. In section 5, I give my views on frequentism and the possibility of compromises between Bayesian and frequentist ideas. Finally in section 6 I give my conclusions.

2. THE ICEBERG DATA

Before discussing the elements of a model, I think it is most useful to get the question straight, which corresponds most closely to Casella's steps 5 and 6.

Everyone with a modicum of liberal arts training knows about "compare and contrast" questions. The point is that there are always similarities and always differences. Either can be celebrated.

Looking at the graph of relative frequencies of icebergs, it is clear that most of the story here is in the similarity of the patterns. But one could also look for differences, for the "contrast." If you ask me to believe that the frequencies, month-by-month, of icebergs are exactly the same to an arbitrary number of decimal places, I must reply that I cannot. Thus I regard Casella's null hypothesis as foolishness. I put zero prior, and hence zero posterior, on its truth. So I need a better question.

Suppose instead I ask what I consider to be a better question: how far apart are $\theta^N = (\theta_1^N, \ldots, \theta_{12}^N)$, frequencies of icebergs south of 48°N, and $\theta^S = (\theta_1^S, \ldots, \theta_{12}^S)$ frequencies of icebergs south of Grand Banks? I could measure this in a variety of ways, such as

$$d_1 = \sum_{i=1}^{12} (\theta_i^S - \theta_i^N)^2$$

and

$$d_2 = \sum_{i=1}^{12} \left| \theta_i^S - \theta_i^N \right|.$$

Now a prior on (θ^S, θ^N) and a likelihood on counts given (θ^S, θ^N) will yield a posterior, and I can compare what I thought about a distance measure d before I saw the data with what I think after I see the data.

This is a measure of what I have learned from the data about how different θ^S and θ^N are. So this is how I think a modern Bayesian would structure the problem.

What are the data? If they are a complete census of all icebergs from 1900 to 1926, then we know that the hypothesis H_0 is false. So suppose that these are a random sample of a larger population of icebergs. How do these particular icebergs come to be in the data set? Because someone observed them, presumably. Is it reasonable to assume that icebergs have the same chance of being observed, regardless of month? I should think that the summer months are easier to observe than the winter months, because more observers will be around and weather conditions are better. The critical issue is whether the observation bias, I should believe, is the same for the two areas. Thus if θ_i^N is the probability of a random iceberg in region N being there in month i, and η_i^N is the probability of its being observed, then $\eta_i^N \theta_i^N$ is the probability of an iceberg being there and being observed, and the frequencies observed have probabilities

$$\psi_i^N = \frac{\eta_i^N \theta_i^N}{\sum_{i=1}^{12} \eta_i^N \theta_i^N} \tag{3}$$

Note that if I believe that iceberg generation is constant by month ($\theta_1^N = \ldots \theta_{12}^N = 1/12$), then ψ_i^N gives information about η_i's, the observation intensities. Which interpretation to give to the data depends on what you believe. The Bayesian method can't say which is right or wrong, but it does provide for (and insist on having) a full, probabilistic statement of what the investigator believes. Reasonable people need not agree on these matters. This allows readers to judge those beliefs, and possibly approximate the calculations the reader might do with his own beliefs. The argument affects the likelihood as well as the prior; both are subjective. Note that I now have more parameters than I have data points. Hence a frequentist treatment of such a model is impossible. Frequentist analysis thus encourages you not to delve too deeply, not to ask such questions.

Even the above formulation is too simplistic, since it assumes that the probabilities θ and η are constant over years. Since during the period of the data collection both the sinking of the Titanic and World War I occurred, it is hard to believe that η, the observation probabilities, were constant. A careful modeling of the data would have to take this into account and would treat skeptically claims of a vast increase in icebergs in the latter half of the period.

Priors on θ are important for the inference in question. The first tool a statistician would think of in this regard is a Dirichlet distribution (a multivariate Beta distribution) on the vector $(\theta_1,\ldots,\theta_{12})$. However the Dirichlet has some unattractive features for this purpose, principally that it treats all the months symmetrically, without making use of the adjacency of them. I would prefer to think of a continuous model in which the critical parameter is an angle, which could be given a Fisher/von Mises distribution, which is like a normal (or Gaussian) distribution for angles and has as hyperparameters a central tendency, ν, and a measure of spread, τ^2. Thus ν would indicate the direction of greatest iceberg intensity, thinking of time through the year as circular. Looking at the data, perhaps a good estimate would be $\hat{\nu}$ = May 10. The measure of spread, τ^2, would indicate how peaked the distribution is. To complete the model, a prior on both ν's (North and South), and both τ's would be necessary.

In these terms, I think that the quantity $d_3 = \nu^S - \nu^N$ would be useful as an alternative to d_1 and d_2. The main advantage of d_3 is that its units are days, which is natural and might have a physical interpretation.

It is now time to turn attention to inference, in Casella's sense. The frequentist statement, applied to this situation, is that if the null hypothesis of no difference were true, and the experiment were repeated an infinite number of times with the same parameter values, the

data would be as or more extreme in only $1 - 0.977 = 0.023$ proportion of the cases. Thus the conclusion is that either the null hypothesis is false or something unusual has happened. But frequentists cannot say which, or even give a probability on which. Note that 0.023 is **NOT** the probability that the null hypothesis is false. Not believing H_0, I don't find this frequentist probability calculation useful.

Casella's version of a Bayesian treatment of this problem is not recognizably Bayesian to me. All it does is condition on both margins in the table (total icebergs observed by month, and total icebergs observed N and S), and then calculates a frequentist p-value. The only warranted statement from his calculation is again that either something unusual happened (with probability less than 0.006), or the null hypothesis is false. But again he cannot say which, nor give a probability for it. I see no justification for Casella's statement that "the probability of the null hypothesis is 0.994."

One interesting way to think about these statistical procedures is to ask what happens as the sample size grows large. In frequentist statistics, no sharp null hypothesis (such as $\theta^N = \theta^S$) is significant if the sample size is small. However as the sample size grows large, every such hypothesis will turn out to be significant. Thus significance measures sample size more powerfully than it does the extent to which the "straw-man" null hypothesis is false. Since better measures of sample size are generally available, significance testing is, in my judgment, not very useful.

By contrast, in the Bayesian analyses I have been discussing, as the sample size grows, the posterior distribution of whichever d you like will converge to a point. You will then effectively know how far from true the hypothesis of equality is, by your chosen measure. What to make of it then depends on what you are doing scientifically, whether you want to emphasize the "compare" or the "contrast" side.

I have written at some length about the iceberg data because it gives me an opportunity to illustrate how Bayesian thinking helps me to model a process. The important points, in my view, are

- The frequentist hypothetical infinite sequence of identical circumstances is a figment of their imaginations.
- Priors and likelihoods are important because they correspond to something real: what you believe about the data.
- Frequentist ideas can get in the way of good modeling because you can easily get too many parameters.
- Testing sharp null hypotheses is generally a foolish undertaking, because they are each, to a greater or lesser degree, wrong.

3. BREAKING WAVES

The principal difference between this example and the previous one is that the null hypothesis is no longer sharp. That is, inferential attention focuses on a single parameter b, and whether $b \leq 4$ or $b \leq 3$.

Unlike Casella, I would not center the prior at the hypothesized value, but would instead have it represent my honest opinion, or my view of what some other scientific opinion might honestly be. My summary would be the posterior distribution on the parameter b, from which one could calculate $P(b \leq 4 | \text{ data})$, $P(b \leq 3 | \text{ data})$, and any other probabilities that might be of interest.

4. BUBBLE DATA

This is similar to the breaking wave data, except that there are several regressions instead of a single one. Such a model is called hierarchical. These have proven useful in a very wide variety of domains.

At the first level, the log bubble population is modeled as

$$N(Z) = A_u + b_u Z + \epsilon$$

where ϵ is Gaussian with mean 0 and variance σ^2, $N(Z)$ and Z are observed, and a_u, b_u, and σ^2 are parameters. At the second level, there might be a bivariate Gaussian distribution on (a_u, b_u) with some mean (a,b) and some covariance matrix Σ. Finally, a third level would specify a prior in (a,b,Σ,σ^2). Such a model is complete if each quantity mentioned has a distribution. A complete model permits a Bayesian analysis, conditioning on the observed data, as a Bayesian should. Interest may focus on the parameters at any level: (a_u, b_u) might be of interest, or (a,b), or any of the others.

5. ON COMPROMISES

As explained just above, a complete hierarchical model is fully Bayesian, and not a compromise. "Empirical Bayesian models" are incomplete; they forget the upper levels of a hierarchy and treat the remaining parameters frequentistically. There is no advantage to a Bayesian in doing so. If the posterior distribution is peaked in the parameters taken to be fixed, there may not be too much loss in this method as an approximation. However in great generality estimates of uncertainty derived from the "empirical Bayesian method" will be underestimates of the same measure derived from a fully Bayesian approach, because parameters are taken as known with certainty that are not known with certainty.

To be successful, a compromise must offer something to each party. Empirical Bayes methods do represent a compromise on the frequentist side, because some (but not all) parameters are treated as random variables with distributions. But to a Bayesian, this "compromise" offers no advantages over a straight Bayesian analysis.

6. PRAGMATIC CONCLUSIONS

In principle, I am convinced that Bayesian ideas are the right way to structure thinking about inference. We are still learning how to use this powerful tool in an effective way. If the problem you have can't be done now in a Bayesian way, then you have to work your problem as best you can, approximating a fully Bayesian analysis.

Even the pre-Socratic philosopher Heraclitus understood that frequentism does not apply to oceanographic problems.

Research supported by NSF SES-8900025 and DMS-9005858, and ONR N00014-89-J-1851.

REFERENCES

Bartlett, John (1980). Familiar Quotations: A Collection of Passages, Phrases and Proverbs, traced to their sources in Ancient and Modern Literature, 15th edition, Little-Brown & Co., Boston, p. 70.

Casella, G. (1993). Illustrating Frequentist and Bayesian Statistics in Oceanography, this volume.

A BAYESIAN APPROACH TO OBSERVATION QUALITY CONTROL IN VARIATIONAL AND STATISTICAL ASSIMILATION

Andrew C. Lorenc
Forecasting Research, Meteorological Office, Bracknell, England

1. INTRODUCTION

Bayesian methods are ideally suited to the ongoing operational data assimilation needed for numerical weather prediction (NWP). Observational errors can be treated as random variables, and we have a long experience of previous observations over which to build up an estimate of their distribution. This experience tells us that observation error distributions are typically non-Gaussian; there are more large errors than expected. It is the handling of these gross errors that we call quality control. As well as the observations, we also need, and have, much other information about the atmosphere. Indeed this prior information is more valuable than that from the observations at any one time. We have a forecast "background field," based on the accumulated knowledge from previous observations, which is usually rather accurate. A forecast based on the background, with no new observations, would probably be more accurate than one based on a batch of observations, with no background. So it is essential to give proper weight to this prior knowledge; the Bayesian approach allows us to do this.

In section 2 we review the Bayesian derivation of the posterior probability of atmospheric states, and hence the equation used to combine observations and background to produce an "analysis"[1] for NWP. With Gaussian distributions, the posterior distribution has mean and variance given by equations which are often derived by a statistical approach, referred to as optimal interpolation (OI). For NWP we need to find the "best" analysis, without necessarily evaluating the complete posterior probability density function (p.d.f.). This can be done by a variational approach, which for Gaussian errors is shown to be equivalent to OI. With non-Gaussian errors, we have to be more careful in defining "best." Appropriate definitions and their interpretation for multi-modal p.d.f.s are discussed.

In section 3 we introduce a simple model of observational errors as the sum of a no-information distribution of gross errors and a Gaussian distribution of good data. Despite its simplicity, this distribution has been found to be sufficient to derive an effective quality control scheme for the majority of observations. The gross errors leads to a posterior p.d.f. which may be multi-modal. Variational methods using a descent algorithm are not guaranteed to find the best analysis.

[1] This terminology is traditional in NWP. "Synthesis" would be better.

The traditional approach to dealing with gross errors is to apply a quality control procedure to reject "bad" observations, then to perform the analysis with the remaining observations, assuming they have Gaussian errors. In section 4 we provide a Bayesian justification of criteria for doing this. We derive an expression for the posterior probability of gross error and reject a datum based on this. (A similar, but not identical, probability is implicit in variational descent algorithms). The posterior probability can be evaluated for gross errors in each observation—individual quality control (IQC), or for each combination of gross errors—simultaneous quality control (SQC). The operational quality control procedure at the Met Office is based on IQC, while that at the European Centre for Medium-Range Weather Forcasting (ECMWF) is based on SQC. The approaches differ subtly in the assumptions made about the posterior p.d.f. when defining the "best" analysis. More significantly, they differ in the further approximations which have to be made in a practical implementation. In section 5, a simple example is studied illustrating the differences between the variational method, IQC, and SQC.

2. BAYESIAN DERIVATION OF ANALYSIS EQUATION

This derivation mainly follows Lorenc (1986).

2.1 Notation

$\mathbf{x}$	atmosphere as represented in model
$\mathbf{x}_t$	model representation of the true state of the atmosphere
$\mathbf{x}_b$	prior estimate of $\mathbf{x}_t$ (e.g., from forecast)
$\mathbf{y}$	observations
$\mathbf{y}_t$	observations that would be given by error-free instruments
$K(\mathbf{x})$	forward operator for calculating $\mathbf{y}$ from $\mathbf{x}$
$\mathbf{K}$	tangent linear operator of K, such that $K(\mathbf{x}+\delta\mathbf{x})=K(\mathbf{x})+\mathbf{K}\delta\mathbf{x}+O(\delta\mathbf{x}^2)$
P	probability
p	probability distribution function
$P(\mathbf{x})$	= probability that $\mathbf{x}\le\mathbf{x}_t<\mathbf{x}+\mathbf{dx}$ $= p(\mathbf{x})\mathbf{dx}$

N.B. We use $\mathbf{x}$ both for the vector of values and for the event $\mathbf{x}\le\mathbf{x}_t<\mathbf{x}+\mathbf{dx}$.

P(A|B) is the conditional probability of A, given B.

2.2 Probability equations

Probabilities are used in a Bayesian way to describe the state of information. We have some prior information about $\mathbf{x}$. We add to this information from observations $\mathbf{y}$. We need to know the posterior knowledge about $\mathbf{x}$. Operator K does not have a normal inverse.

From now on all probabilities are conditional on knowing $\mathbf{x}_b$. To simplify notation we write $P(\cdot)$ instead of $P(\cdot \mid \mathbf{x}_b)$.

The basis of the derivation is the identity

$$\begin{aligned} P(\mathbf{x} \cap \mathbf{y}) &= P(\mathbf{x} \mid \mathbf{y}) \quad P(\mathbf{y}) \;=\; P(\mathbf{y} \mid \mathbf{x}) \quad P(\mathbf{x}) \\ &= p(\mathbf{x} \mid \mathbf{y})\mathbf{dx}\; p(\mathbf{y})\mathbf{dy} = p(\mathbf{y} \mid \mathbf{x})\mathbf{dy}\; p(\mathbf{x})\mathbf{dx}. \end{aligned} \tag{1}$$

What we want an expression for is

$P(\mathbf{x}|\mathbf{y}) = p(\mathbf{x}|\mathbf{y})\mathbf{dx}$, the analysis probability, i.e., the probability that $\mathbf{x} \le \mathbf{x}_t < \mathbf{x}+\mathbf{dx}$, given the background $\mathbf{x}_b$ and the observations $\mathbf{y}$.

We assume we know certain distributions, based on our prior experience and our knowledge of the physics:

$P(\mathbf{x}) = p(\mathbf{x})\mathbf{dx}$, is the probability that $\mathbf{x} \le \mathbf{x}_t < \mathbf{x}+\mathbf{dx}$, given only the prior knowledge of $\mathbf{x}_b$.

$p(\mathbf{y} \mid \mathbf{y}_t \cap \mathbf{x})$ is the instrumental error distribution.

$p(\mathbf{y}_t \mid \mathbf{x})$ is the forward operator error distribution.

From the last two distributions, we can find

$P(\mathbf{y}|\mathbf{x}) = p(\mathbf{y}|\mathbf{x})\mathbf{dy}$, the probability of getting observations $\mathbf{y}$ given $\mathbf{x}=\mathbf{x}_t$.

$$p(\mathbf{y} \mid \mathbf{x}) = \int p(\mathbf{y} \mid \mathbf{y}_t \cap \mathbf{x})\, p(\mathbf{y}_t \mid \mathbf{x})\mathbf{dy}_t. \tag{2}$$

From this, and our prior knowledge of $\mathbf{x}$, we can find

$P(\mathbf{y}) = p(\mathbf{y})\mathbf{dy}$, the probability of getting observations $\mathbf{y}$.

$$\begin{aligned} p(\mathbf{y}) &= \int p(\mathbf{y} \mid \mathbf{x})\, p(\mathbf{x})\mathbf{dx} \\ &= \iint p(\mathbf{y} \mid \mathbf{y}_t \cap \mathbf{x})\, p(\mathbf{y}_t \mid \mathbf{x})\mathbf{dy}_t\; p(\mathbf{x})\mathbf{dx}. \end{aligned} \tag{3}$$

Bayes' Theorem, which follows from the basic identity (1), is

$$p(\mathbf{x} \mid \mathbf{y}) = p(\mathbf{y} \mid \mathbf{x})\, p(\mathbf{x}) / p(\mathbf{y}). \tag{4}$$

We can substitute the expressions derived above to give

$$p(\mathbf{x}|\mathbf{y}) = \frac{\int p(\mathbf{y}|\mathbf{y}_t \cap \mathbf{x})\, p(\mathbf{y}_t|\mathbf{x})\, d\mathbf{y}_t\; p(\mathbf{x})}{\iint p(\mathbf{y}|\mathbf{y}_t \cap \mathbf{x})\, p(\mathbf{y}_t|\mathbf{x})\, d\mathbf{y}_t\; p(\mathbf{x})\, d\mathbf{x}}. \tag{5}$$

This p.d.f. describes our total posterior information about $\mathbf{x}$, given $\mathbf{x}_b$ and $\mathbf{y}$.

2.3 Solution using Gaussian probability distributions

We assume K can be linearized in the region of $\mathbf{x}_b$ and $\mathbf{x}_a$ such that

$$K(\mathbf{x}_a) = K(\mathbf{x}_b) + K(\mathbf{x}_a - \mathbf{x}_b). \tag{6}$$

We assume all the p.d.f.s are Gaussian, and use the notation

$$\mathrm{N}(\mathbf{x}|\,\mathbf{m},\mathbf{B}) = ((2\pi)^N|\,\mathbf{B}\,|)^{-1/2} \exp(-\tfrac{1}{2}(\mathbf{x}-\mathbf{m})^T \mathbf{B}^{-1}(\mathbf{x}-\mathbf{m})) \tag{7}$$

where $\mathbf{B}$ is an N×N positive definite matrix, and $|\mathbf{B}|$ is its determinant.

We assume that we know

the background error distribution	$p(\mathbf{x})$	$= \mathrm{N}(\mathbf{x}\|\mathbf{x}_b, \mathbf{B})$,
the instrumental error distribution	$p(\mathbf{y}\|\mathbf{y}_t \cap \mathbf{x})$	$= \mathrm{N}(\mathbf{y}\|\,\mathbf{y}_t, \mathbf{O})$,
the forward operator error distribution	$p_f(\mathbf{y}_t\|\,\mathbf{x})$	$= \mathrm{N}(\mathbf{y}_t\|K(\mathbf{x}), \mathbf{F})$,

where $\mathbf{B}$, $\mathbf{O}$, and $\mathbf{F}$ are covariances.

Then, using the properties of Gaussians, the observational error distribution is given by the convolution

$$\begin{aligned} p(\mathbf{y}|\,\mathbf{x}) &= \int p(\mathbf{y}|\,\mathbf{y}_t \cap \mathbf{x})\, p(\mathbf{y}_t|\,\mathbf{x})\, d\mathbf{y}_t \\ &= \mathrm{N}(\mathbf{y}|\,K(\mathbf{x}_t), \mathbf{O}+\mathbf{F}) \end{aligned} \tag{8}$$

where $\mathbf{O}+\mathbf{F}$ ($=\mathbf{E}$) is the observational error covariance.

The observation distribution, only knowing $\mathbf{x}_b$, is given by

$$p(\mathbf{y}) = \mathrm{N}(\mathbf{y}|K(\mathbf{x}_b), \mathbf{O}+\mathbf{F}+\mathbf{K}\mathbf{B}\mathbf{K}^{\mathrm{T}}). \tag{9}$$

Substituting these into Bayes' Theorem (4) gives

$$\begin{aligned} p(\mathbf{x}|\mathbf{y}) &= \mathrm{N}(\mathbf{y}|K(\mathbf{x}_t), \mathbf{O}+\mathbf{F})\,\mathrm{N}(\mathbf{x}|\mathbf{x}_b, \mathbf{B}) / \mathrm{N}(\mathbf{y}|K(\mathbf{x}_b), \mathbf{O}+\mathbf{F}+\mathbf{K}\mathbf{B}\mathbf{K}^{\mathrm{T}}) \\ &= \mathrm{N}(\mathbf{x}|\mathbf{x}_a, \mathbf{A}). \end{aligned} \tag{10}$$

where $\mathbf{x}_a$ and $\mathbf{A}$ are defined by

$$\begin{aligned} \mathbf{A} &= \mathbf{B} - \mathbf{B}\mathbf{K}^{\mathrm{T}}(\mathbf{K}\mathbf{B}\mathbf{K}^{\mathrm{T}} + \mathbf{O} + \mathbf{F})^{-1}\mathbf{K}\mathbf{B} \\ \mathbf{x}_a &= \mathbf{x}_b + \mathbf{B}\mathbf{K}^{\mathrm{T}}(\mathbf{K}\mathbf{B}\mathbf{K}^{\mathrm{T}} + \mathbf{O} + \mathbf{F})^{-1}(\mathbf{y} - K(\mathbf{x}_b)). \end{aligned} \tag{11}$$

It is normal to assume that the "best" estimate of $\mathbf{x}_t$ is given by the mean $\mathbf{x}_a$ of the Gaussian posterior distribution. Thus using the above equation we can calculate $\mathbf{x}_a$ directly. Equation (11) is the "OI" equation, often derived as the minimum variance best estimate, without relating it to the p.d.f. (10).

2.4 Non-Gaussian Bayesian analysis

If K is more nonlinear, or the p.d.f.s are non-Gaussian, then (10) and (11), capable of direct solution, cannot be used. Although the Bayes' Theorem (4) for the analysis p.d.f. is still valid, the expression for p which results is usually too complicated to be very useful in describing our knowledge about $\mathbf{x}$; we want an estimate of the "best" $\mathbf{x}$, without evaluating the full p.d.f. First, to define "best," we define a *loss function* $L(\mathbf{x}_1, \mathbf{x})$ giving the cost to us of making an estimate $\mathbf{x}_1$ when the true value is $\mathbf{x}$. The expected loss R is

$$R(\mathbf{x}_1) = \int L(\mathbf{x}_1, \mathbf{x})\, p(\mathbf{x}|\mathbf{y})\, d\mathbf{x}. \tag{12}$$

The best estimate is the $\mathbf{x}_1$ which minimizes this. In general this requires evaluating all of $p(\mathbf{x}|\mathbf{y})$. This can be avoided by making L a negative delta function, so that there is a gain from getting exactly the correct $\mathbf{x}$, while all other values are equally worthless. With this spike loss

$$L(\mathbf{x}_1, \mathbf{x}) = -\delta(\mathbf{x}_1 - \mathbf{x}) \tag{13}$$

$$R(\mathbf{x}) = -p(\mathbf{x}|\mathbf{y}). \tag{14}$$

Substituting in the Bayesian expression for $p(\mathbf{x}|\mathbf{y})$, and since $p(\mathbf{y})$ is independent of $\mathbf{x}$, the $\mathbf{x}$ that minimizes $R(\mathbf{x})$ is the same as the $\mathbf{x}$ that minimizes a penalty functional $\mathfrak{I}$ given by

$$\mathfrak{I} = -\ln(p(\mathbf{y}|\mathbf{x})) - \ln(p(\mathbf{x})). \tag{15}$$

If we substitute the Gaussian p.d.f.s of the last section into this, we get

$$\mathfrak{J} = \frac{1}{2}(\mathbf{y} - K(\mathbf{x}))^{\mathrm{T}}(\mathbf{O}+\mathbf{F})^{-1}(\mathbf{y} - K(\mathbf{x})) + \frac{1}{2}(\mathbf{x}_b - \mathbf{x})^{\mathrm{T}}\mathbf{B}^{-1}(\mathbf{x}_b - \mathbf{x}) + \text{constant}. \tag{16}$$

If, furthermore, we make K linearizable, we see why the linear problem with Gaussians is easier to solve: $\mathfrak{J}$ becomes a quadratic in $\mathbf{x}$. Using the same algebraic manipulations as are needed to establish the properties of Gaussians used in the last section, and the same definitions (11) of $\mathbf{x}_a$ and $\mathbf{A}$, gives

$$\mathfrak{J} = \frac{1}{2}(\mathbf{x}_a - \mathbf{x})^{\mathrm{T}}\mathbf{A}^{-1}(\mathbf{x}_a - \mathbf{x}) + \text{constant}. \tag{17}$$

For large problems it is easier to find $\mathbf{x}_a$ iteratively, even if $\mathfrak{J}$ is quadratic. If K cannot be linearized over the whole range containing $\mathbf{x}_b$ and possible $\mathbf{x}_a$s, then an explicit solution is not possible. If K is still differentiable, so that

$$K(\mathbf{x}+\delta\mathbf{x}) = K(\mathbf{x}) + \mathbf{K}_{\mathbf{x}}\delta\mathbf{x}, \text{ as } \delta\mathbf{x} \to 0 \tag{18}$$

then we can look for the minimum of $\mathfrak{J}$ using a descent algorithm. At the minimum, the gradient of $\mathfrak{J}$ with respect to the components of $\mathbf{x}$ is zero:

$$\mathfrak{J}' = -\mathbf{K}_{\mathbf{x}}^{\mathrm{T}}(\mathbf{O}+\mathbf{F})^{-1}(\mathbf{y} - K(\mathbf{x})) - \mathbf{B}^{-1}(\mathbf{x}_b - \mathbf{x}) = 0. \tag{19}$$

This formula is exact; we can find the most probable $\mathbf{x}$. The next stage of generalization is to allow the p.d.f.s to be weakly non-Gaussian. That is, we use the Gaussian formulae with $\mathbf{O}_{\mathbf{x}}$, $\mathbf{F}_{\mathbf{x}}$, and $\mathbf{B}_{\mathbf{x}}$ being slowly varying functions of $\mathbf{x}$, whose derivatives we can neglect. We also neglect derivatives of $\mathbf{K}_{\mathbf{x}}$. Then if we define $\mathbf{x}_a$ as the $\mathbf{x}$ which minimizes $\mathfrak{J}$, i.e.,

$$\mathfrak{J}' = -\mathbf{K}_{\mathbf{x}_a}^{\mathrm{T}}(\mathbf{O}_{\mathbf{x}_a} + \mathbf{F}_{\mathbf{x}_a})^{-1}(\mathbf{y} - K(\mathbf{x}_a)) - \mathbf{B}_{\mathbf{x}_a}^{-1}(\mathbf{x}_b - \mathbf{x}_a) = 0. \tag{20}$$

Then

$$\mathfrak{J}'' \cong \mathbf{K}_{\mathbf{x}_a}^{\mathrm{T}}(\mathbf{O}_{\mathbf{x}_a} + \mathbf{F}_{\mathbf{x}_a})^{-1}\mathbf{K}_{\mathbf{x}_a} + \mathbf{B}_{\mathbf{x}_a}{}^{-1} = \mathbf{A}^{-1}. \tag{21}$$

Then, in the neighbourhood of $\mathbf{x}_a$,

$$p_a(\mathbf{x}|\mathbf{y}) \propto \mathrm{N}(\mathbf{x}|\mathbf{x}_a, \mathfrak{J}''^{-1}). \tag{22}$$

If K is sufficiently nonlinear, or the p.d.f.s are sufficiently non-Gaussian, $p_a(\mathbf{x}|\mathbf{y})$ may have multiple maxima. We have then to consider how to decide which is best. We can generalize on the spike loss, by allowing the loss function to be a Gaussian:

$$L(\mathbf{x}_1, \mathbf{x}) = -\mathrm{N}(\mathbf{x}_1 | \mathbf{x}, \mathbf{L}). \tag{23}$$

As **L** tends to zero this gives us the spike loss. For the Gaussian analysis problem we can evaluate the convolution explicitly:

$$R(\mathbf{x}_1) = -\mathrm{N}(\mathbf{x}_1 | \mathbf{x}_a, \mathbf{A} + \mathbf{L}). \tag{24}$$

Thus the loss is minimum when $\mathbf{x}_1 = \mathbf{x}_a$, as we would expect. We can use this expression to help us in deciding between peaks in a non-Gaussian posterior p.d.f., by assuming that the peaks can be approximated by a local Gaussian. We assume the spread of the entire posterior p.d.f. can be characterized by **S** (i.e., **S** describes the distance between peaks). If **L>>S** then the loss function is quadratic over the range of significant probabilities, and the best estimate is the mean of the full p.d.f. (which may fall between two peaks). But if **L<<S** then we may consider the peaks separately. Then if in the vicinity of the ith local maximum the p.d.f. is

$$p(\mathbf{x} | \mathbf{y}) \cong \mathrm{P}_i \ \mathrm{N}(\mathbf{x} | \mathbf{x}_i, \mathbf{A}_i) \tag{25}$$

Then the loss associated with choosing the analysis to be at this maximum of $p(\mathbf{x} | \mathbf{y})$ is given by

$$R(\mathbf{x}_i) = -\mathrm{P}_i \, \mathrm{N}(\mathbf{x}_1 | \mathbf{x}_i, \mathbf{A}_i + \mathbf{L}). \tag{26}$$

If $\mathbf{S} >> \mathbf{A}_i >> \mathbf{L}$ then $R(\mathbf{x}_i) \cong -p(\mathbf{x}_i | \mathbf{y})$, and the best peak is the highest. If $\mathbf{A}_i << \mathbf{L} << \mathbf{S}$ then $R(\mathbf{x}_i) \cong -\mathrm{P}_i \times$ constant (independent of i), and the best peak is that with the largest area.

3. NON-GAUSSIAN OBSERVATIONAL ERRORS

3.1 Gross error model

Lorenc and Hammon (1988) introduced a simple model of observational errors: They are uncorrelated, so that each observation can be considered separately. For each, either the observation is good, in which case its error comes from a Gaussian, or it has a gross error, in which all observed values over a range of plausible values are equally likely. Thus we have (for "plausible" y)

$$p(\mathrm{y} | \mathbf{x}) = (1 - \mathrm{P(G)}) \, \mathrm{N}(\mathrm{y} | K(\mathbf{x}), \mathrm{E}) + \mathrm{P(G)} \, \mathrm{k} \tag{27}$$

where E is the observational error variance (=O+F), P(G) is the probability of gross error, **x** is the true value, and k is given by

$$\int_{\text{plausible values}} \mathrm{k} \, \mathrm{dy} = 1. \tag{28}$$

3.2 Posterior p.d.f. with gross errors

It is instructive to look at some simple posterior p.d.f.s resulting from this model, before going on to the full multivariate analysis problem. The simplest case is for a single observation of one parameter, and a prior (background) estimate y_b ($=K(x_b)$) from a Gaussian distribution. Because $P(y \mid x)$ is non-Gaussian, the shape of the posterior p.d.f. depends on the difference between y and y_b, as illustrated in Figure 1. Even in this, the simplest case, there are multiple maxima, and there are configurations in which a variational search, starting from the prior estimate y_b, will not find the best value.

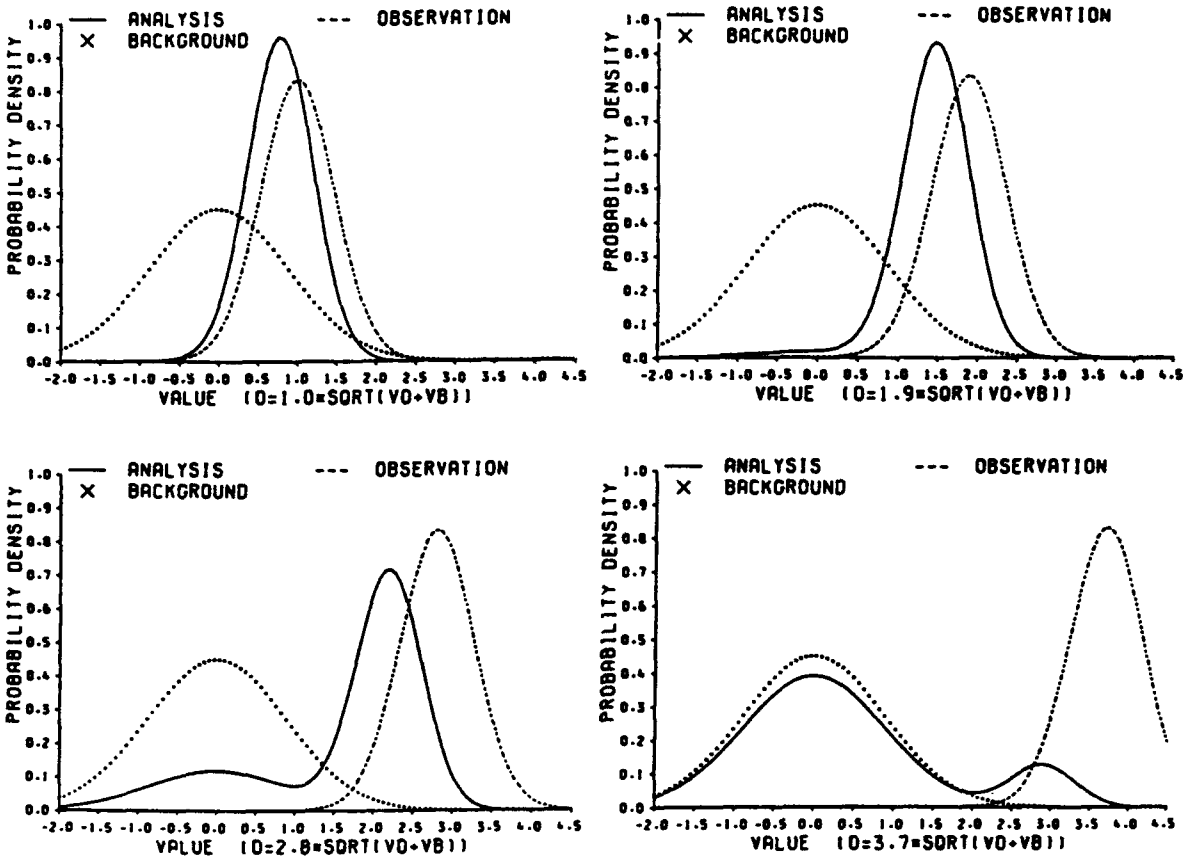


Figure 1. The p.d.f.s for an observation, background, and Bayesian analysis for a selection of observation-background differences o. The p.d.f.s are appropriate for ship observations of surface pressure, with P(G)=0.05. (Lorenc and Hammon 1988).

Figure 2 shows a similar error model applied to two realizations, each of ten observations, from an idealized Doppler observing system. With poor signal to noise ratio, P(G) may be large for such an instrument; we have used P(G)=0.5. In the lower example, it is not clear which is the "best" estimate; no method can consistently find it.

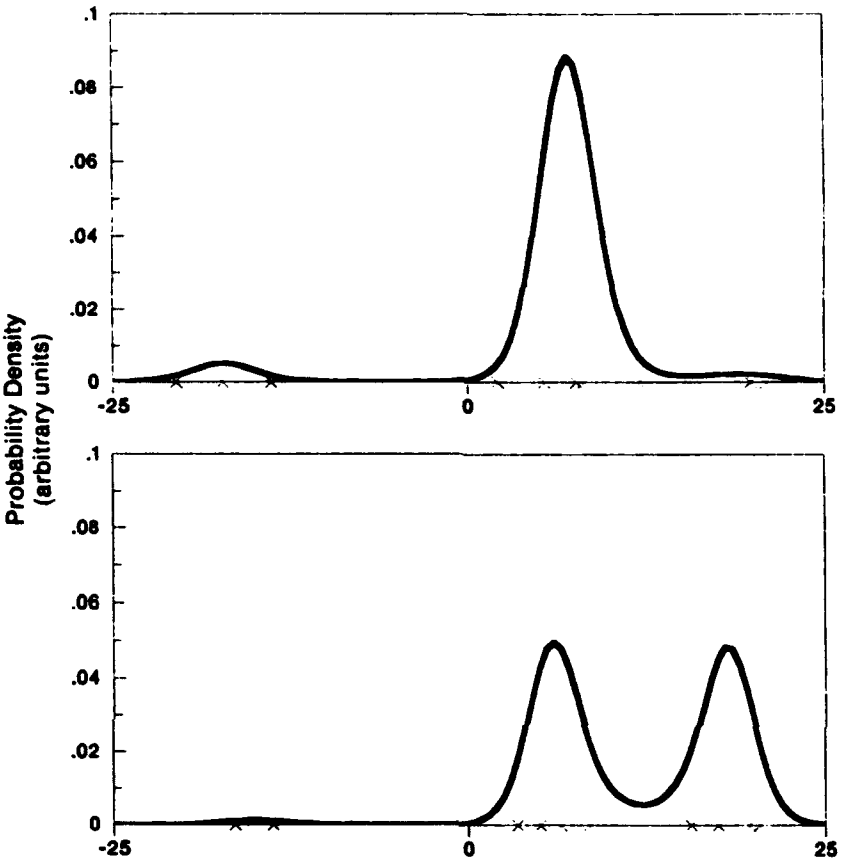


Figure 2. Two examples of p.d.f.s from simulated doppler wind observations, with x_t=7, good observations having E=9 and P(G)=0.5. (Dharssi et al. 1992).

3.3 Variational descent algorithms in the presence of gross errors

Even in the top example of Figure 2 there are multiple maxima, which become more obvious minima if we convert to a $\ln(p)$ penalty function $\mathfrak{I}$, so a descent algorithm must start near the correct value, if it is to find the absolute minimum.

Lorenc (1988) used an observational error distribution like (27) in a variational analysis based on minimizing (15). The possibility of gross errors converts the quadratic penalty function of (16) into one with plateaus (Figure 3). If the current estimate in an iterative algorithm is on one of these, the gradient does not well indicate which way to adjust towards the minimum. Note that the width of the minimum depends on E, while the spread of the deviations between initial estimate and observations depends on B+E. So if B is large, the iteration may not move towards the absolute minimum. This was the case in the

experiments of Lorenc (1988). He tried various methods to improve the first-guess of the iteration, for instance by first setting P(G)=0, but with limited success.

Dharssi et al. (1992) had more success in their examples. In simple single value problems like those shown in Figure 2, they found that increasing the observational error E in early iterations helped the iterative estimate move towards the best value. In a two-dimensional simulation of winds from a scanning lidar, they found that for relatively dense but unreliable (P(G)=0.5) observations, the iteration did converge. It is an open question whether descent algorithms, suitably modified in early iterations, will be sufficient for practical applications, or whether we will still need the decision algorithms described in the next section.

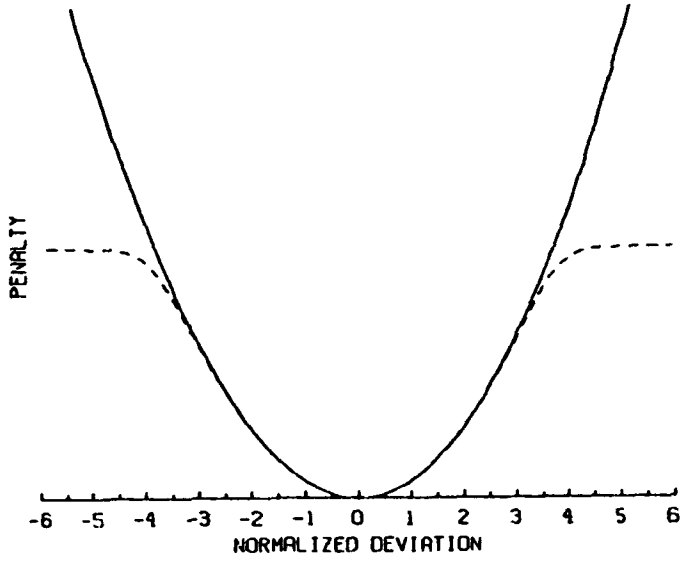


Figure 3. Solid line: quadratic penalty function for a single observation, dashed line: penalty function assuming a P(G)=0.05 (Lorenc 1988).

4. QUALITY CONTROL

4.1 Posterior probability of gross error

The posterior p.d.f.s shown in Figure 1 are each the sum of two Gaussians, one corresponding to there being a gross error (G), one corresponding to the observation being correct ($\overline{G}$). Lorenc and Hammon (1988) proposed applying Bayes' theorem directly to the gross error event G:

$$\begin{aligned} P(G|y) &= p(y|G)\,P(G)/p(y) \\ &= p(y|G)\,P(G)/(p(y|G)\,P(G)+p(y|\overline{G})\,P(\overline{G})). \end{aligned} \tag{29}$$

Using (2), (27), and (28), we have

$$p(y|G) = k. \tag{30}$$

Using (2), (27), and (9), we have

$$p(y|\overline{G}) = N(y|K(\mathbf{x}_b), \mathbf{E}+\mathbf{KBK}^T) \tag{31}$$

so (29) can be readily evaluated. The two Gaussians in Figure 1 are weighted by $P(G|y)$ and $P(\overline{G}|y)$ respectively. Thus accepting or rejecting an observation depending on whether $P(G|y)$ is greater than or less than 0.5 is consistent with the "best" analysis in terms of a Gaussian loss function, as discussed in relation to (26), as long as **S>>L>>A**. This is the basis of the decision taking algorithms used in Bayesian quality control schemes.

Dharssi et al. *(1992) pointed out* an interesting relationship between the variational method and the posterior probability of gross error. If we calculate $\Im'$ using the error model (27), then we get

$$\Im' = -\mathbf{K}_x^T(\mathbf{E}_x)^{-1}(y-K(x))-\mathbf{B}^{-1}(\mathbf{x}_b-\mathbf{x}) = 0 \tag{32}$$

where the diagonal element of $\mathbf{E}_x$, for observation i, is given by

$$(E_x)_{ii} = E_i / P(\overline{G}_i | \mathbf{x} \cap y_i). \tag{33}$$

The E_i is the observational error variance of observation i if it does not have a gross error, and $P(\overline{G}_i | \mathbf{x} \cap y_i)$ is the posterior probability that it does not have a gross error, given that $\mathbf{x}=\mathbf{x}_t$. We are effectively increasing the assumed error variance of observations that are unlikely to be correct. (This is not the same as the artificial increase discussed in section 3.3, where E_i is increased when calculating $E_i / P(\overline{G}_i | \mathbf{x})$ in early iterations, to aid convergence towards the global minimum). Equation (32) has the same form as (19), so by using (33) each iteration, a variational method for Gaussian errors is converted to one for non-Gaussian errors.

At convergence, there will exist a final estimate of $P(\overline{G}_i \mid \mathbf{x} \cap y_i)$ for each observation. It can be considered to be a variational quality control (VQC) decision about the observations' quality.

4.2 Individual quality control (IQC)

Equation (29) can be extended to consider more than one observation. Lorenc and Hammon (1986) give the derivation for two observations:

$$P(G_1 \mid \mathbf{y}) = P(G_1 \mid y_1) / (p(\mathbf{y}) / p(y_1) p(y_2)) \tag{34}$$

$$p(\mathbf{y}) / p(y_1) p(y_2) = 1 - P(\overline{G}_1 \mid y_1) P(\overline{G}_2 \mid y_2) \{1 - p(\mathbf{y} \mid \overline{G}_1 \cap \overline{G}_2) / (p(y_1 \mid \overline{G}_1) p(y_2 \mid \overline{G}_2))\}. \tag{35}$$

Ingleby and Lorenc (1992) give a more general derivation. The number of terms to be considered in the extended equation goes as 2^n, where n is the number of observations, so evaluation of the exact equation rapidly becomes impractical. Lorenc and Hammon (1988) suggest sequential application of the "buddy check" equation for two observations as an approximation. This is the method used operationally at the Met Office. The decision about whether to use each observation i is made individually, based on an approximation to its posterior probability of gross error $P(G_i \mid \mathbf{y})$. The analysis is then made using the accepted observations, assuming they have Gaussian errors.

4.3 Simultaneous quality control

The 2^n terms in the full expression for $P(G_1 \mid \mathbf{y})$ come from the various combinations of accepted and rejected observations. Each combination C_α is associated with a multivariate normal distribution, each individually calculated using (10), so that the total p.d.f. is given by Ingleby and Lorenc (1992):

$$p(\mathbf{x} \mid \mathbf{y}) = \sum_{\alpha=0}^{2^n-1} p(\mathbf{x} \mid \mathbf{y} \cap C_\alpha) P(C_\alpha \mid \mathbf{y}). \tag{36}$$

The posterior probability for each combination of gross errors can be found using Bayes' theorem:

$$P(C_\alpha \mid \mathbf{y}) = p(\mathbf{y} \mid C_\alpha) P(C_\alpha) / p(\mathbf{y}). \tag{37}$$

If we assume that each of the Gaussians which makes a significant contribution to (36) has a distinct peak, then we can apply (26) to decide which gives the best estimate of $\mathbf{x}$. If **S>>L>>A** it is the one with the maximum $P(C_\alpha \mid \mathbf{y})$.

Evaluating all 2^n probabilities is impossible for large n. Since $p(\mathbf{y})$ is the same for each C_α, we can instead search for the combination with the maximum $P(\mathbf{y}|C_\alpha)\,P(C_\alpha)$. The states C_α correspond to the vertices of an n-dimensional hypercube. One possible algorithm for searching only a small subset of possible combinations is related to the SIMPLEX algorithm in integer linear programming. We start with an estimate of the best, and then search to see if any of its neighbours is more likely. Moving from one C_α to a neighbour corresponds to changing the quality control decision on one observation, while keeping those on other observations the same. If one of the neighbouring combinations is more likely, we can then search its neighbours, and so on. This is the basis of the OI quality control algorithm of Lorenc (1981), which is used at ECMWF[2]. Rather like the variational descent algorithms, this search algorithm relies on having a good first guess of the best C_α, since there will in general be multiple local maxima.

5. COMPARISON OF QUALITY CONTROL CRITERIA

Figure 4 shows an example chosen to illustrate the differences between the approaches. The solid line shows the posterior p.d.f. given by (36), while the dotted lines are the constituent Gaussians. Variational analysis, using a spike loss function, will pick the highest peak (VAN). Note however that a simple descent algorithm would have to start quite close to $\mathbf{x}_{VAN}$ if it is to converge to the correct value; starting from $\mathbf{x}_b$ will not do.

Assuming this $\mathbf{x}_{VAN}$ is correct, all the observations have $P(\overline{G}_i|\mathbf{x}_{VAN} \cap y_i) > 0.5$, so if we were to use this as an acceptance criterion, and do a Gaussian analysis using the observations, we would get the value corresponding to the peak VQC.

Calculating the $P(G_i|\mathbf{y})$ for each observation (IQC), the two observations of -9 both have posterior probabilities less than 0.5 (i.e., they fail) while the observation of -6 just passes. This pass is in part due to contributions from the possibility that the other observations were actually correct; IQC can give inconsistent decisions.

Simultaneous quality control does look for a consistent decision; in this case the Gaussian with the largest area is that labelled SQC. It corresponds to rejection of all the observations, i.e., it is the background distribution. Note that the SIMPLEX algorithm will not work well in this case. There is one local maximum for the combinations accepting both observations of -9, and another for the combinations rejecting them both. The SIMPLEX algorithm will converge to one of these; it cannot get from one to the other because intermediate combinations (accepting one and rejecting the other) are less likely.

[2]The ECMWF scheme sets rejection tolerances directly, but an equivalent formation similar to (29) is possible.

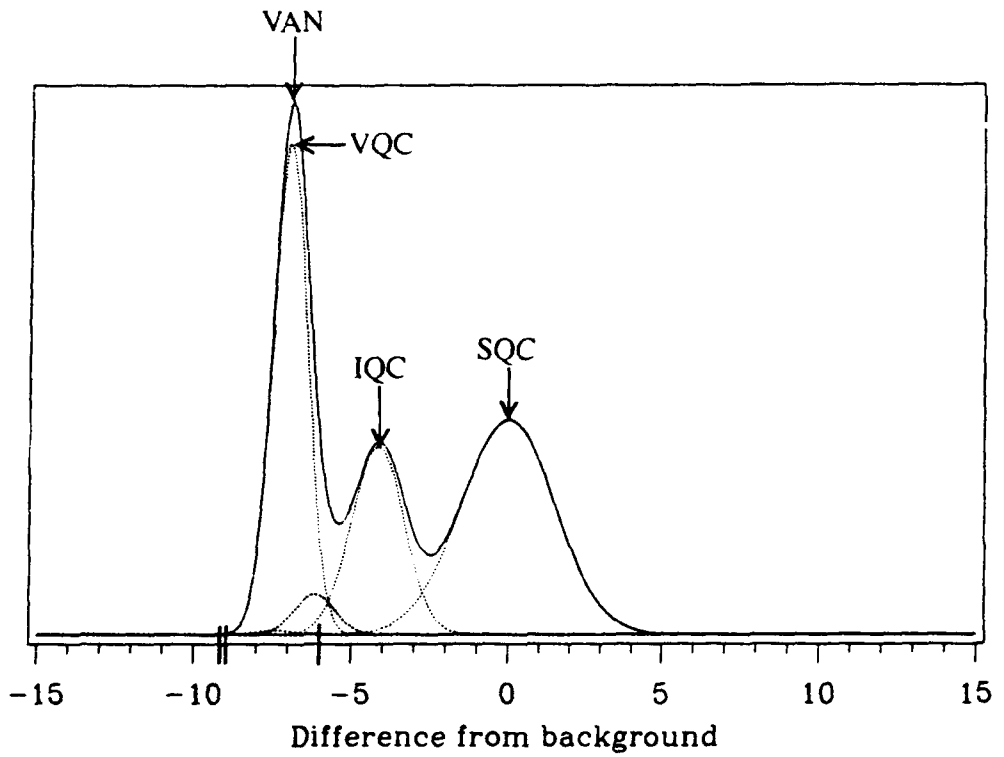


Figure 4. Solid curve: P(x | y), dotted curves: $P(\mathbf{x}|\mathbf{y} \cap C_\alpha)$ for y_i =-9, -9 and -6, x_b =0, and other values appropriate for sea-level-pressure observations (E=1, B=2.25, k=0.043, and P(G)=0.04), from Ingleby and Lorenc (1992). For meaning of annotations, see text.

6. CONCLUDING REMARKS

We have shown that the Bayesian approach provides a sound method for combining observations and background information. If distributions are Gaussian, it leads to the statistical interpolation (OI) equations and to a variational analysis with a quadratic penalty function. It also indicates how the method can be extended to observations with non-Gaussian distributions.

The proper "best" analysis depends on an appropriately defined loss function. Finding it requires convolutions over the posterior probability density function, which for non-Gaussian distributions is impractical. Variational analysis (VAN and VQC) and quality control algorithms (IQC and SQC) are making approximations to the ideal loss function. In NWP, we have a background $\mathbf{x}_b$ which usually would lead to a forecast that is not too bad. Large improvements on this accuracy are not required, so $\mathbf{L} \simeq \mathbf{B}$. Individual peaks in the p.d.f. have $\mathbf{A}_i < \mathbf{B}$. So the assumption that the region of useful analyses is larger than each peak, but smaller than the distance between peaks ($\mathbf{S} >> \mathbf{L} >> \mathbf{A}_i$) may not be too bad for NWP assimilation.

There have also to be approximations in implementation; none of the methods can be implemented perfectly in practical NWP problems. In the approximate forms discussed here,

VAN and VQC use a descent algorithm, with a modified penalty function in early iterations to try to get convergence to the best **x** from as wide a range as possible of first-guesses. This has been tried on simulated data by Dharssi et al. (1992) and is an attractive candidate for future variational NWP assimilation systems.

IQC, as used at the Met Office (Lorenc and Hammon 1988), uses a sequential pairwise buddy check to approximate the method for >2 close observations. Some tuning of this has been found to be necessary.

SQC, with a SIMPLEX search, does not necessarily correctly handle close observations which agree with each other, but might both be wrong. The method used at ECMWF (Lorenc 1981) is similar to this (although the rejection tolerances are set directly, rather than via P(G)).

The Bayesian approach has allowed us to understand the relationship between these different methods.

Acknowledgments. I am grateful to Imtiaz Dharssi and Bruce Ingleby for their contributions to our joint papers quoted here, and to Jim Purser for initiating me in Bayesian analysis and quality control theory.

REFERENCES

Dharssi,I., Lorenc,A.C. and Ingleby,N.B. 1992: Treatment of gross errors using maximum probability theory. *Quart. J. Roy. Met. Soc.*, 118, 1017-1036.

Ingleby, N.B., and Lorenc, A.C. 1992: Bayesian quality control using multivariate normal distributions. MetO(S) Sci.Paper No.10.

Lorenc, A.C. 1981: A global three-dimensional multivariate statistical analysis scheme. *Mon. Wea. Rev.* 109, 701-721.

Lorenc, A.C. 1986: Analysis methods for numerical weather prediction. *Quart. J. Roy. Met. Soc.* 112, 1177-1194.

Lorenc, A.C. and Hammon, O., 1988: Objective quality control of observations using Bayesian methods - Theory, and a practical implementation. *Quart. J. Roy. Met. Soc.* 114, 515-543.

Lorenc, A.C. 1988: Optimal nonlinear objective analysis. *Quart. J. Roy. Met. Soc.* 114, 205-240.

OPTIMAL SPACE-TIME INTERPOLATION OF GAPPY FRONTAL POSITION DATA

Toshio M. Chin and Arthur J. Mariano
Department of Meteorology and Physical Oceanography, RSMAS, University of Miami

INTRODUCTION

The spatial and temporal variability of Gulf Stream meanders has been studied by many including Watts and Johns (1982), Halliwell and Mooers (1979 and 1983), Olson et al. (1983), and Cornillon (1986). The majority of these studies use the northern edge or north wall, determined from the largest spatial gradient in advanced very high resolution radiometer (AVHRR) data, as the Gulf Stream path indicator. The advantages of using AVHRR data for locating the Gulf Stream are (i) the large contemporaneous spatial coverage, (ii) the measurements have been collected daily since 1978, and (iii) the frontal locations are the strongest signal in the data. The chief disadvantages are the amount of processing (geometric corrections, cloud-screening/compositing, and manual digitizing of frontal positions from images) required and that the satellite sensor cannot see through the clouds. Consequently, there are large spatial (2-6 degrees) and temporal (3-6 days) gaps in the Gulf Stream north wall position (GSNWP) data set. Mariano (1988 and 1990) devised a new approach, termed *contour analysis*, for melding of oceanic data and for space-time interpolation of gappy frontal data sets. The key elements of contour analysis are feature matching and averaging in a coordinate system determined from the contour positions. In applying his approach to the GSNWP, Mariano assumed a dominant one-dimensional east-west phase speed in his algorithm. This assumption restricted the application of this algorithm to other frontal data sets, such as the Brazil-Malvinus confluence (Garzoli et al., 1992) where the north-south phase speeds are also important, and led to poor estimates of the GSNWP when the north-south phase speed was significant.

The primary goal of this study is to develop an improved algorithm for space-time interpolation of gappy frontal data sets. The major improvements are the inclusions of (i) two-dimensional phase speed, (ii) a more autonomous algorithm, (iii) a better feature matching algorithm, and (iv) the inclusion of a temporal smoothness constraint. The space-time interpolator is formulated in the framework of probabilistic (Bayesian) estimation. This report first reviews such an estimation theoretic framework and, in particular, a Kalman filter-based interpolation algorithm. Then, feature detection and matching algorithms are discussed, followed by presentation and discussion of some preliminary results.

BACKGROUND

The approach described in this report is a two-step process: First, the locations of the sea surface temperature (SST) "edges" (gradient maximums) are detected and digitized by trained personnel at the University of Rhode Island (URI). Then, the longitude-latitude coordinates of the digitized points are interpolated by an autonomous computer program. This report describes this second step—a probabilistic approach to the development of a space-time interpolation algorithm.

The space-time interpolation problem is formulated as a quadratic optimization problem. Here, we review how the cost function can be optimized using a Bayesian estimation framework (with additive white Gaussian noise models) and how the solution can be obtained time-recursively using Kalman filters.

1. Space-only interpolation

We first discuss the problem of interpolating points digitized from a *single* frame of image, as this is the first step of our space-time interpolation algorithm. Let $(\tilde{x}_i, \tilde{y}_i), i = 1,2,\ldots,m$ be the longitude-latitude coordinates of the digitized points. We assume, for the time being, that the latitudes y of the GSNWP can be described by a function of the longitudes x only, i.e., there exists a *single-valued* function $y(x)$. This is a mathematically convenient description used in the previous studies of Gulf Stream variability, but it is not always appropriate for Gulf Stream meanders. The bi-variate formulation for *multi-valued* features, such as "S" and "Ω" shaped meanders, is discussed after analyzing the simpler single-valued case.

The function $y(x)$ is interpolated based on the measurements $(\tilde{x}_i, \tilde{y}_i)$ by finding the function that optimizes

$$\min_y \sum_{i=1}^{m} v_i |\tilde{y}_i - y(\tilde{x}_i)|^2 + \int_{\mathcal{D}} \left[\alpha_1 \left| \frac{\partial}{\partial x} y \right|^2 + \alpha_2 \left| \frac{\partial^2}{\partial x^2} y \right|^2 \right] dx \qquad (1)$$

where v_i are the weights representing our confidence in the corresponding measurements. The two integral terms, weighted by the parameters α_1 and α_2, control *continuity* ("tension") and *linearity* ("smoothness") of the interpolated curve, respectively. This optimization approach finds applications in general geophysical interpolation and variational problems (e.g., Inoue, 1986).

2. Maximum likelihood estimation

To obtain a numerical solution of Eq. (1), the longitude is discretized as $x = j\Delta x, j = 1,2\ldots,n.$. The interval Δx is chosen small enough for the discrete domain to include (within a reasonable quantization error) the measurements as $\{\tilde{x}_i\} \subset \{x(j\Delta x)\}$, which implies $m < n$—the number of points to be estimated is usually three to four times the number of data points. The corresponding latitudes are represented by an n-dimensional column vector **y** whose elements are $y(j\Delta x), j \in [1,n]$, while the measurements of the latitudes are organized as an m-dimensional vector **z** whose elements are $\tilde{y}_i, i \in [1,m]$. A discrete version of Eq. (1) is

$$\min_{\mathbf{y}} \alpha_1 \|\mathbf{S}_1\mathbf{y}\|^2 + \alpha_2 \|\mathbf{S}_2\mathbf{y}\|^2 + \|\mathbf{z}-\mathbf{H}\mathbf{y}\|_{\mathbf{M}}^2 \tag{2}$$

where the vector-norms are weighted 2-norms, e.g., $\|\mathbf{z}-\mathbf{H}\mathbf{y}\|_{\mathbf{M}}^2 \equiv (\mathbf{z}-\mathbf{H}\mathbf{y})^T \mathbf{M}(\mathbf{z}-\mathbf{H}\mathbf{y})$. (The superscript ^{T} denotes matrix transpose.) The matrixes $\mathbf{S}_1$ and $\mathbf{S}_2$ are the first and second order difference operators, respectively, while **M** is a diagonal matrix whose diagonal elements are the measurement weights $v_i, i \in [1,m]$. The $m \times n$ matrix **H** is the *data-estimate correspondence* operator whose $(i,j)^{th}$ element h_{ij} is defined as

$$h_{ij} = \begin{cases} 1 & \text{if } \tilde{x}_i = j\Delta x \\ 0 & \text{if otherwise.} \end{cases} \tag{3}$$

The process of determining the matrix **H**—the correspondence problem—is straightforward in this case where the latitudes are treated as a function of the longitudes. Some GSNWP features, such as an "S" shaped meander, can make the correspondence problem quite complex. Mariano (1990) showed that detecting and matching such features based on the sparse sets of data points are the key (and most difficult) components for a successful interpolation scheme. Our solution to the correspondence problem is presented in the next section.

The minimizing solution $\hat{y}$ of Eq. (2) is exactly the maximum likelihood estimate **y** based on the observation equation

$$\begin{bmatrix} \mathbf{z} \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} \mathbf{H} \\ \mathbf{S}_1 \\ \mathbf{S}_2 \end{bmatrix} \mathbf{y} + \begin{bmatrix} \mathbf{v}_H \\ \mathbf{v}_1 \\ \mathbf{v}_2 \end{bmatrix} \tag{4}$$

where the additive observation noise v_H, v_1, and v_2 are mutually independent zero-mean Gaussian random vectors with covariance $\mathbf{M}^{-1}, \alpha_1^{-1}\mathbf{I}$, and $\alpha_2^{-1}\mathbf{I}$, respectively. The solution of this probabilistic estimation problem requires minimization described in Eq. (2) (Lewis, 1986); thus, the maximum likelihood formulation based on Eq. (4) constitutes a probabilistic interpretation of Eq. (2). An advantage of this probabilistic version is that the estimation error covariance can be computed, along with the estimate itself, allowing us to quantify confidence/uncertainty in the solution. For Eq. (4), the optimal estimate $\hat{y}$ and estimation error covariance $\mathbf{P}$ are given by

$$\hat{y} = \mathbf{L}^{-1}\mathbf{H}^T\mathbf{M}\mathbf{z} \tag{5}$$

$$\mathbf{P} = \mathbf{L}^{-1} \tag{6}$$

where $\mathbf{L} \equiv \mathbf{H}^T\mathbf{M}\mathbf{H} + \alpha_1\mathbf{S}_1^T\mathbf{S}_1 + \alpha_2\mathbf{S}_2^T\mathbf{S}_2$ is a sparse penta-diagonal matrix. Alternatively, the minimization problem Eq. (2) can also be reformulated as a Bayesian estimation problem in which the first two terms in Eq. (2) are interpreted as the prior statistics for the unknown $\mathbf{y}$ (Szeliski, 1989). Both the Bayesian and maximum likelihood formulations are equivalent when Gaussian noise models are used, as they yield the same solution.

In terms of selecting the parameters for the interpolation problem, the probabilistic formulation must be specified slightly more precisely than its variational counterpart: In Eq. (2) the weights α_1, α_2, and $\mathbf{M}$ are only required to be specified up to a multiplicative constant—only the *ratios* among the weights need to be controlled. The same parameters in the probabilistic formulation Eq. (4) play the roles of noise covariances whose *values* (not just the ratios among them) must exactly be given. This extra bit of precision is necessary for the computed $\mathbf{P}$ to be interpreted meaningfully as the estimation error covariance.

3. Time-extension and Kalman filtering

Equation (1) can be extended temporally to perform space-time interpolation for $y(x,t)$ using an additional continuity constraint over time:

$$\min_y \sum_{k=1}^{K}\sum_{i=1}^{m(k)} v_i(k)\left|\tilde{y}_i - y(\tilde{x}_i(k), k\Delta t)\right|^2 + \int_0^T \int_{\mathcal{D}} \left[\alpha_1\left|\frac{\partial}{\partial x}y\right|^2 + \alpha_2\left|\frac{\partial^2}{\partial x^2}y\right|^2 + \beta_1\left|\frac{\partial}{\partial t}y\right|^2\right] dx\,dt \tag{7}$$

where the time variable is discretized as $t = k\Delta t, k = 1,2,\ldots,K$ and the variables associated with the measurements are indexed by k. In the GSNWP estimation problem, Δt is two days. The parameter β_1 controls the strength of the temporal constraint.

A discrete and probabilistic interpretation of Eq. (7) can be obtained by supplementing Eq. (2) with an evolution equation (8) representing the time-continuity constraint. The result is a stochastic dynamic system indexed by the time variable k:

$$\mathbf{y}(k) = \mathbf{y}(k-1) + \mathbf{w}(k) \tag{8}$$

$$\begin{bmatrix} \mathbf{z}(k) \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} \mathbf{H}(k) \\ \mathbf{S}_1 \\ \mathbf{S}_2 \end{bmatrix} y(k) + \begin{bmatrix} \mathbf{v}_H(k) \\ \mathbf{v}_1(k) \\ \mathbf{v}_2(k) \end{bmatrix} \tag{9}$$

where $\mathbf{w}(k)$ is a zero-mean Gaussian random vector with covariance $\beta_1^{-1}\mathbf{I}$. Representing the space-time interpolation problem as a dynamic system is attractive because the Kalman filtering algorithm (Gelb, 1974) allows computational efficiency (time-recursive computation) and flexibility (*filtered*, *predicted*, and *smoothed* estimates). Numerical solution of the space-time interpolation Eq. (7) is given by the smoothed estimate, which can be computed as a linear combination of forward and backward filtered estimates based on the system Eqs.(8,9): Let $(\hat{y}_f(k), \mathbf{P}_f(k))$ be the estimate-covariance pair (the forward estimate) produced by the Kalman filter based on the system equations. Then, Eq. (8) is replaced by a backward dynamic equation $\mathbf{y}(k) = \mathbf{y}(k+1) + \mathbf{w}(k+1)$ to compute the backward filtered estimates and covariances $(\hat{\mathbf{y}}_b(k), \mathbf{P}_b(k))$. The smoothed estimate-covariance pair $(\hat{\mathbf{y}}(k), \mathbf{P}(k))$ is given by

$$\hat{\mathbf{y}}(k) = \mathbf{P}(k)\left\{\mathbf{P}_f^{-1}(k)\mathbf{y}_f(k) + \mathbf{P}_b^{-1}(k)\mathbf{y}_b(k) - \mathbf{H}^T(k)\mathbf{M}\mathbf{z}(k)\right\}, \tag{10}$$

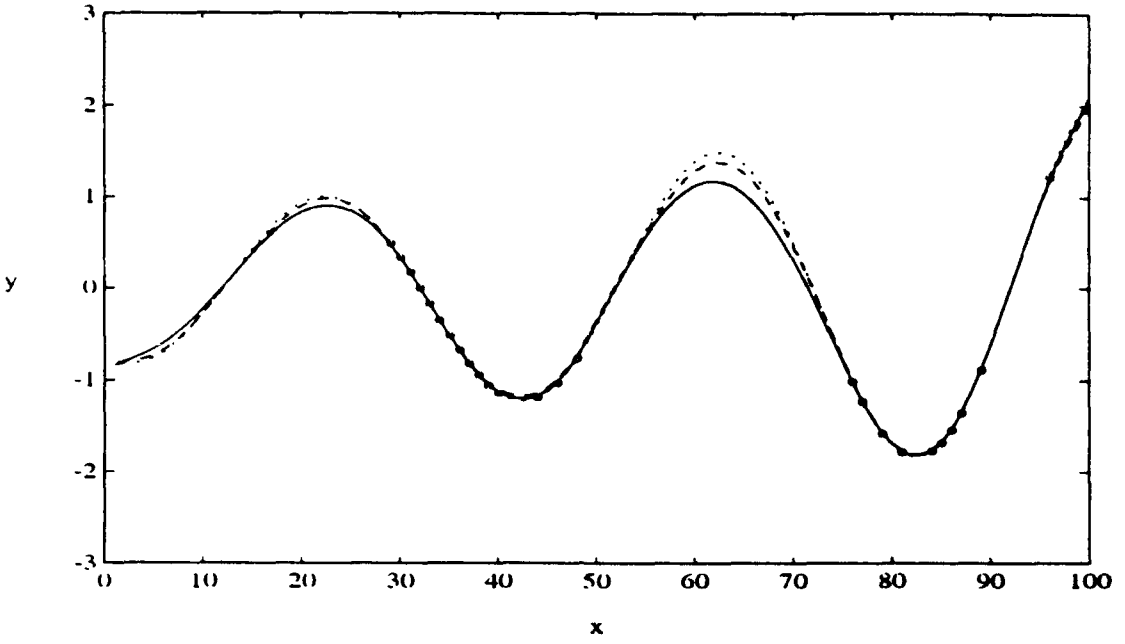
$$\mathbf{P}(k) = \left\{\mathbf{P}_f^{-1}(k) + \mathbf{P}_b^{-1}(k) - \mathbf{H}^T(k)\mathbf{M}\mathbf{H}(k) - \alpha_1\mathbf{S}_1^T\mathbf{S}_1 - \alpha_2\mathbf{S}_2^T\mathbf{S}_2\right\}^{-1}. \tag{11}$$

Detailed derivations can be found in textbooks such as Lewis (1986) and Anderson and Moore (1979).

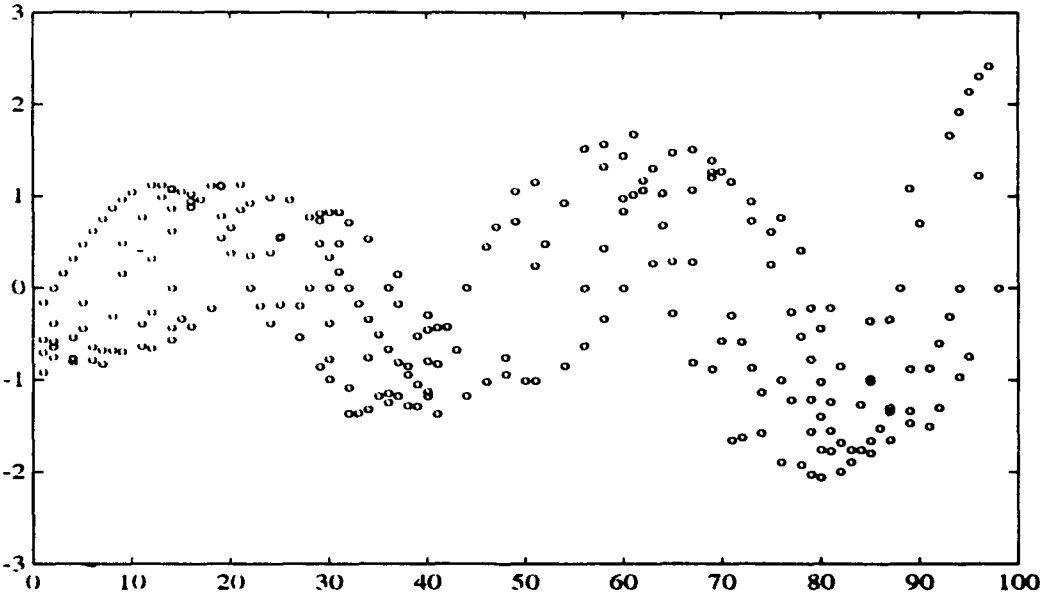
Figure 1a illustrates that the formulation Eq. (7) performs adequate interpolation for a simple ideal case in which y is in fact a function of x. Here, for each integer value of $x \in [1,100]$ and $t \in [1,10]$, y, is computed as

$$y = (1+u)\sin\left(\frac{x-2t}{20}\pi\right)\exp\left(\frac{x-5t}{100}\right)$$

where $u \in [0,0.2]$ is a uniformly distributed random number. The "measurements" are made by selecting 25 points along the curve for each t (Fig. 1a). All measurements over the 10 time-frames are shown in Figure 1b by superposition. The interpolated curve (the



(a)



(b)

Figure 1. (a) An example of space-time interpolation using the formulation Eq. (7) is shown as the solid curve. The dotted curve is the "truth" while the circles are the "measurements" made in this particular time-frame. The dashed curve is a result obtained by adding the "temporal linearity" term (cf. Eq. (15)) into the formulation. (b) All the "measurements" superimposed over time.

solid line in Fig. 1a) estimated the crests of the waveform reasonably well. The parameters used were $\mathbf{M} = \mathbf{I}$, $\alpha_1 = 0.01$, $\alpha_2 = 1$, and $\beta_1 = 0.1$.

FORMULATIONS

1. Bi-variate unknown

Problems with uni-variate formulation (i.e., assuming that y is a function of x) include inability of representing certain frequently occurring shapes of meanders (e.g., large "S" and "Ω" shapes) and inability to model uncertainty in the measurements of the longitudes x. The spatial domain of interpolation must be dynamic, rather than fixed, to correctly assimilate measurements in time under temporal movements of the GSNWP. A dynamic reference frame is crucial to GSNWP interpolation as smoothing over a fixed spatial grid will smear out meanders and other important shape features along the contours, as described by Mariano (1990) in a more general context of data melding. It is an adaptive ("object-oriented") reference frame similar in spirit to the Lagrangian frame. Unlike typical Lagrangian formulations, in which physical motion models are available, our problem must deal with phenomenologically characterized motions of the GSNWP contours, making the formulation challenging because of lack of accurate mathematical models.

We will convert Eq. (7) to a bi-variate formulation. Let $p(s,t) \equiv [x(s,t), y(s,t)]^T$ be the true contour location, where the spatial domain s is the arclength along the contour at a given t. We denote the points digitized from the k^{th} SST image as $\tilde{p}_i(k), i \in [1, m(k)]$. The bi-variate version of Eq. (7) is

$$\min_p \sum_{k=1}^{K} \sum_{i=1}^{m(k)} v_i(k) \left\| \tilde{p}_i(k) - p(s_i(k), k\Delta t) \right\|^2 + \int_0^T \int_C \left[\alpha_1 \left\| \frac{\partial}{\partial s} p \right\|^2 + \alpha_2 \left\| \frac{\partial^2}{\partial s^2} p \right\|^2 + \beta_1 \left\| \frac{\partial}{\partial t} p \right\|^2 \right] ds\, dt. \tag{12}$$

This minimization is more complex than Eq. (7) because $s_i(k)$, the spatial coordinates (in terms of arclength) of the digitized points, are unknown. Specifically, the origin of the spatial index s is difficult to define, since there is no guarantee (even though it is a reasonable assumption for the Gulf Stream) that all contours pass through a given point (i.e., the origin) on the x-y plane. Also, $s_i(k)$ must be determined concurrently as the contours are interpolated. The arclength, in fact, cannot be specified exactly without knowing the contour $p(k)$ itself! A Kalman filter-based solution for Eq. (12) becomes an adaptive filtering/smoothing problem:

$$\mathbf{p}(k) = \mathbf{p}(k-1) + \mathbf{w}(k) \tag{13}$$

$$\begin{bmatrix} \mathbf{q}(k) \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} \mathbf{H}(\mathbf{p}(k),k) \\ \mathbf{S}_1 \\ \mathbf{S}_2 \end{bmatrix} \mathbf{p}(k) + \begin{bmatrix} \mathbf{v}_H(k) \\ \mathbf{v}_1(k) \\ \mathbf{v}_2(k) \end{bmatrix} \tag{14}$$

where the components of the vector $\mathbf{q}(k)$ are $\tilde{p}_i(k), i \in [1, m(k)]$. Note that the data-estimate correspondence matrix $\mathbf{H}(p(k),k)$ is now dependent on the state $\mathbf{p}(k)$.

Clearly, Eq. (12) must be optimized adaptively: For each k, either of $s_i(k)$ and $p(k)$ is estimated alternately using the best guess for the other, and this process is iterated for a fixed number of times or until an agreement between the two estimates is obtained within an accuracy parameter. Because of the gaps in the measurements, the estimates at the previous frame (i.e., $\hat{p}(k-1)$) are often the best estimates of the general shape of the contour at the current time. Thus, the problem of establishing correspondence can be approached by incrementally matching the best available estimate of the current contour based on the previous contour and that based on the spatially sparse measurements. This important feature matching problem will be addressed in the next section.

2. Imposing linearity over time

Once the data-estimate correspondence is established, it is straightforward to expand the dynamic system formulation Eqs.(13,14) to incorporate various structural models for the GSNWP contours. For example, we can impose a linearity constraint over time by inserting an additional integrand term

$$\beta^2 \left\| \frac{\partial^2}{\partial t^2} p \right\|^2 \tag{15}$$

to Eq. (12). The corresponding change in the dynamic system is augmentation of the state vector; the dynamic equation is changed to

$$\begin{bmatrix} \mathbf{p}(k) \\ \mathbf{p}(k+1) \end{bmatrix} = \begin{bmatrix} \mathbf{I} & 0 \\ \mathbf{I} & -2\mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{p}(k-1) \\ \mathbf{p}(k) \end{bmatrix} + \begin{bmatrix} \mathbf{w}_1(k) \\ \mathbf{w}_2(k) \end{bmatrix} \tag{16}$$

where $\mathbf{w}_1(k)$ and $\mathbf{w}_2(k)$ are zero-mean Gaussian random vectors with covariance $\beta_1^{-1}\mathbf{I}$ and $\beta_2^{-1}\mathbf{I}$, respectively. Equation (16) can be written in a more attractive form which includes the local displacement $\mathbf{d}(k) \equiv \mathbf{p}(k+1) - \mathbf{p}(k)$ as the extra component of the state vector. The estimates of the local displacement field are of interest in their own right for statistical characterization of Gulf Stream dynamics. The resulting reformulation consists of a modified dynamic equation and an additional row in the observation equation:

$$\begin{bmatrix}\mathbf{p}(k)\\ \mathbf{d}(k)\end{bmatrix}=\begin{bmatrix}\mathbf{I} & \mathbf{I}\\ 0 & \mathbf{I}\end{bmatrix}\begin{bmatrix}\mathbf{p}(k-1)\\ \mathbf{d}(k-1)\end{bmatrix}+\begin{bmatrix}0\\ \mathbf{w}_2(k)\end{bmatrix} \tag{17}$$

$$0=\mathbf{d}(k)+\mathbf{v}_d(k) \tag{18}$$

where $\mathbf{v}_d(k)=-\mathbf{w}_1(k+1)$. By replacing Eq. (13) with Eq. (17) and adding Eq. (18) to Eq. (14), we can jointly estimate the GSNWP $\mathbf{p}(k)$ and the local displacement $\mathbf{d}(k)$. This formulation is the same in spirit as the approach used by Mariano (1990), except that the formulation presented here uses two-dimensional (bi-variate) displacement vectors, instead of one-dimensional in the previous approach, and that the presented formulation is optimal in the least square sense.

The formulation based on Eqs.(17,18) is applied to the uni-variate example in the previous section, i.e., a temporal linearity constraint (i.e., Eq. (15) imposed on y instead of p) is added to Eq. (7). The dashed line in Figure 1a shows one of the resulting interpolated curves. The figure shows that the curve has gained more "stiffness" and the crests of the waves are estimated more accurately with this extra constraint (dashed line) than without it (solid line). The parameter used for the constraint was β_2=0.1.

FEATURE MATCHING

This section describes an approach to establish the data-estimate correspondence. For conciseness in discussion we discuss the filtering problem based on the dynamic system Eqs.(13,14). As mentioned before, we adopt an adaptive filtering approach where best predictions of the GSNWP contour at a given time-frame k are used to estimate the positions, i.e., arc-length indexes $s_i(k)$, of the measurements along the contour. Specifically, two rudimentary contours, one predicted ahead in time based on the estimated contour at k-1 and the other interpolated only over space based on the measurements at k, are "matched" for correspondence, allowing incorporation of the measurements to update the predicted GSNWP estimate. In another words, the matrix $\mathbf{H}(p(k),k)$ in Eq. (14) is evaluated as $\mathbf{H}(\hat{\mathbf{p}}_f(k-1),k)$ in the forward filter and as $\mathbf{H}(\hat{\mathbf{p}}_b(k+1),k)$ in the backward filter, where $\hat{\mathbf{p}}_f(k)$ and $\hat{\mathbf{p}}_b(k)$ represent the forward and backward filtered estimates, respectively. The two contours are matched hierarchically—using larger-scale "features" first and then smaller, more local, inflections of the curves.

1. Feature detection

Large bends, especially those at the apexes of the meanders, are the major features along the GSNWP contours. Although these features are always associated with relatively large values of curvature (second-order derivative along the arc), such *local* attributes alone are not necessarily useful in isolating large meanders among a variety of contour inflections

with much smaller magnitudes. In fact, the magnitudes of the inflections themselves can be used to identify the meander features more directly. These magnitudes are computed as the deviations from a progressively fine-scaled, piece-wise linear approximation of the contour shape. Specifically, consider a segment of the curve between two arbitrary points $p(s_a)$ and $p(s_b)$. Let the deviation $\zeta(s,s_a,s_b)$ be the (perpendicular) distance from the point $p(s)$ along the segment ($s \in [s_a, s_b]$) to the line connecting points $p(s_a)$ and $p(s_b)$, as shown in Figure 2. The points along the curve where large deviations occur are used to segment the curve into a piece-wise linear "skeleton", exemplified in Figure 3. Those points associated with large deviations are the *nodes* of the skeleton of the curve. The following is an iterative algorithm to compute the set of nodes, or *node set*, given the tolerance parameter ϵ for the deviations:

1. Initialize the node set with the two end-points of the curve.
2. Let the number of nodes in the set be L. Let the indexes of the nodes be s_ℓ so that $s_\ell < s_{(\ell+1)}$ for $\ell = 1,2,\ldots,(L-1)$.
3. Find the maximum deviation d^* over the entire curve, i.e., for $\ell = 1,2,\ldots,(L-1)$,

$$d^* = \max_{\ell} \max_{s} \zeta(s, s_\ell, s_{(\ell+1)})$$

   Let s^* be the spatial index for the point where the maximum deviation occurs.
4. If $d^* > \epsilon$, include s^* into the node set; then, go back to step 2 and repeat. Otherwise ($d^* \leq \epsilon$), stop.

The internal node points, $p(s_2), p(s_3), \ldots, p(s_{L-1})$, after the final iteration are referred to as the *feature points*.

2. Feature matching

Let us consider matching feature points from two curves. Each feature point is at the apex of a corner on the skeleton of a curve. A cost is assigned to each of possible matching pairs of feature points as a sum of the costs associated with the distance, angle, and direction of the corner. Let $p(a)$ and $p(b)$ be feature points from each of the two curves. Each feature point, say $p(a)$, is a junction of two line segments of the skeleton; let the two unit vectors pointing along these line segments and originating in the feature point $p(a)$ be u_{a1} and u_{a2}. Let u_{b1} and u_{b2} be similarly defined unit vectors around the feature point $p(b)$. Also, we measure the direction of a vector v as the angle $\angle(v)$ in radians (in the longitude-latitude coordinate system). The costs are defined as

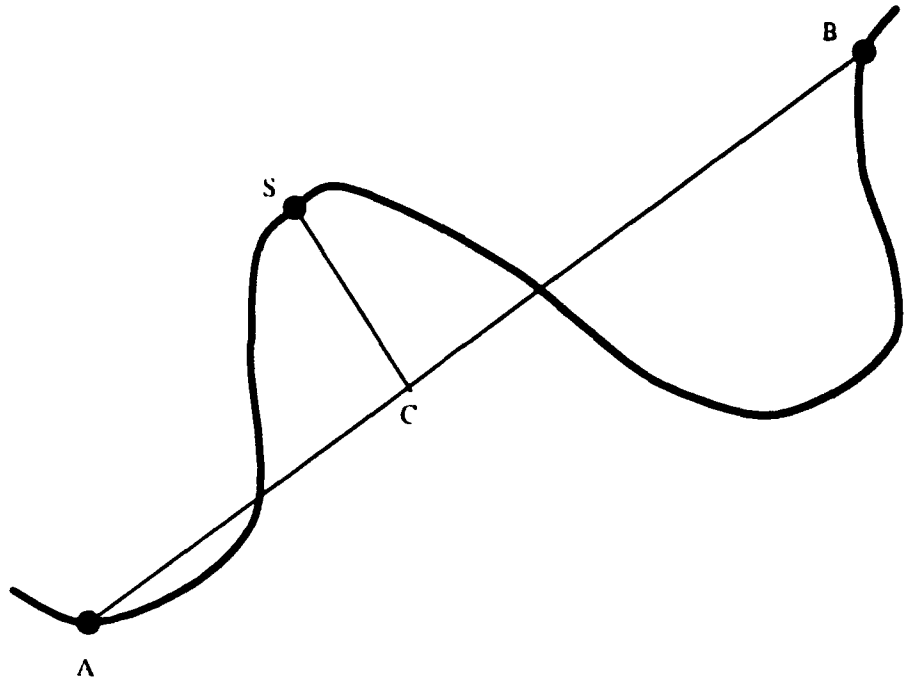


Figure 2. $\zeta(s,s_a,$ or $s_b)$ equals the length of the line segment $\overline{\mathbf{SC}}$, where points A, B, and S correspond to $p(s), p(s_a)$, and (s_b), respectively.

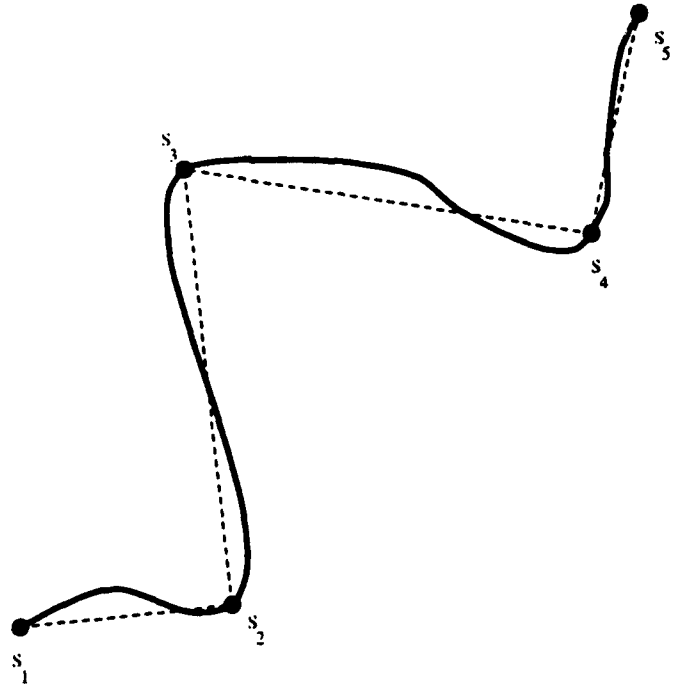


Figure 3. The contour is segmented by the set of nodes $\{s_1, s_2, \ldots\}$. The dashed line represents a skeleton for this contour.

1. Distance. $C_1 = \|p_a - p_b\|^2$
   The distance between the pair of points.
2. Angle. $C_2 = (|\angle(u_{a1}) - \angle(u_{a2})| - |\angle(u_{b1}) - \angle(u_{b2})|)^2$
   The absolute value of the difference between the angles of the corners associated with each of the two feature points.
3. Direction. $C_3 = |\angle(u_{a1} + u_{a2}) - \angle(u_{b1} + u_{b2})|^2$
   The difference between the directions of the openings of the two corners, i.e., the directions of the vectors bisecting the angles.

We penalize large values of these cost functions more heavily (i.e., more than by a linear proportion) than relatively small values. This is achieved by post-distorting the cost by a piece-wise linear mapping function, such as that shown in Figure 4, which discounts smaller cost values and inflates larger values by multipliers (slopes in the figure) smaller and larger, respectively, than 1.

The pairs of feature points with smaller total costs (the sum of three post-processed cost functions) are considered to be matching pairs, with the following constraints:

- The total cost for any matching pair must be smaller than a specified value, which we will refer to as C_{max}.
- A feature point cannot be matched to more than one other feature point.
- The line segments connecting matched feature points can never cross each other.

The last constraint reflects the structural integrity of the meanders (features): The GSNWP meanders can only appear and disappear; they cannot change their sequencing order along the contour.

To summarize, the number of the parameters to be specified for feature point matching is 10: C_{max}, and the two multipliers and a threshold value (the slopes and "th" in Fig. 4 for each of the three cost functions C_1, C_2, and C_3.

3. Local matching

Once correspondence of major features is established, non-feature points can be matched by a simple proportional mapping, leading to a correspondence match of the two contours in their entireties. In Figure 5, for example, the pairs of points (A,A') and (B,B') represent matched feature points, and arc-length indexes s and s' along the two contour segments between the respective feature points are considered to be a matching pair if

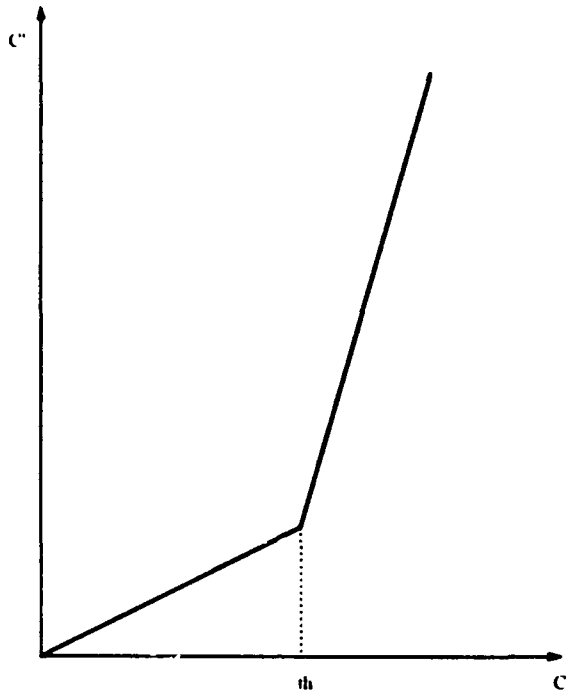


Figure 4. A typical mapping function for postprocessing of the cost "C" (representing C_1, C_2, or C_3). The values smaller than the threshold "th" are discounted while values larger are inflated.

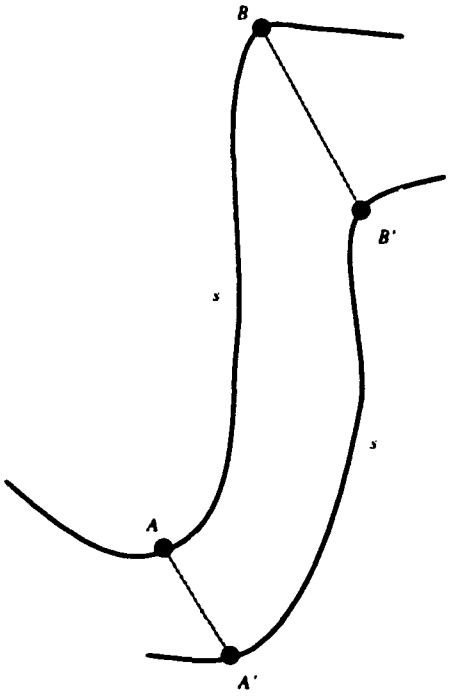


Figure 5. Mapping contour segment AB to segment $A'B'$.

$$\frac{s - s_A}{s_B - s_A} = \frac{s' - s'_A}{s'_B - s'_A} \tag{19}$$

where s_A, s'_A, s_B, and s'_B are indexes of the feature points.

Unfortunately, matched pairs of feature points are sometimes too sparse to be able to guide correspondence of the two contours reliably: Distance between adjacent feature points on a contour can be larger than the phenomenological length scale, a gap in measured points can occur between feature points, and some measurements do not contain any significant meander features.

To remedy this, we need a *secondary* method to register the indexes for two given contours without relying on feature identification and matching. One way of performing such a task is to deform one of the contours toward another using a variational formulation involving cost terms for structure of the deformed contour and for distances between points on two contours. Let $p_1(s)$ and $p_2(s)$ be the two contours to be matched and $\rho(s)$ be a deformation of $p_1(s)$. The deformed contour $\rho(s)$ inherits the indexes of $p_1(s)$; thus, by physically registering $\rho(s)$ onto $p_2(s)$, correspondence between the two index sets can be found. [Such a technique for contour registration is generically known as "snake" in computational vision (Kass et al., 1988)]. Specifically, we consider the optimization problem

$$\begin{aligned} \min_{\rho} \int_{C_1} & F(p_2,\rho) + \alpha_1 \left\| \frac{\partial}{\partial s}\rho \right\|^2 + \alpha_2 \left\| \frac{\partial^2}{\partial s^2}\rho \right\|^2 \\ & + \gamma_0 \| \rho - p_1 \|^2 + \gamma_1 \left\| \frac{\partial}{\partial s}(\rho - p_1) \right\|^2 + \gamma_2 \left\| \frac{\partial^2}{\partial s^2}(\rho - p_1) \right\|^2 ds \end{aligned} \tag{20}$$

where the "gravity" term $F(p_2, \rho)$ works to minimize the distances between points along $\rho(s)$ and $p_2(s)$ and is given by

$$F(p_2,\rho) \equiv -\int_{C_2} \exp\left(-\frac{1}{2} \| \rho(s) - p_2(s') \|^2 \right) ds'. \tag{21}$$

The domains C_1 and C_2 of the integrations are given by the contours p_1 and p_2, respectively. The three cost terms, with coefficients γ_0, γ_1, and γ_2 contain the shape of $\rho(s)$ from becoming radically different from that of $p_1(s)$. The minimizing $\rho(s)$ is given by the non-linear Euler-Lagrange equation

$$2\left[(\alpha_2+\gamma_2)\frac{\partial^4}{\partial s^4}-(\alpha_1+\gamma_1)\frac{\partial^2}{\partial s^2}+\gamma_0\right]\rho - 2\left[\gamma_2\frac{\partial^4}{\partial s^4}-\gamma_1\frac{\partial^2}{\partial s^2}+\gamma_0\right]p_1+\frac{\partial}{\partial\rho}F(p_2,\rho)=0 \tag{22}$$

which, since ρ is the only variable, can be written concisely as

$$2\left[(\alpha_2+\gamma_2)\frac{\partial^4}{\partial s^4}-(\alpha_1+\gamma_1)\frac{\partial^2}{\partial s^2}+\gamma_0\right]\rho-\bar{p}+\frac{\partial}{\partial\rho}F(\rho)=0 \tag{23}$$

where

$$\bar{p}\equiv 2\left[\gamma_2\frac{\partial^4}{\partial s^4}-\gamma_1\frac{\partial^2}{\partial s^2}+\gamma_0\right]p_1.$$

Given a parameter κ, Eq. (23) can be solved iteratively (Kass et al., 1988) as

$$2\left[(\alpha_2+\gamma_2)\frac{\partial^4}{\partial s^4}-(\alpha_1+\gamma_1)\frac{\partial^2}{\partial s^2}+\gamma_0+\kappa\right]\rho_\ell = \bar{p}+\kappa(\rho_\ell-\rho_{(\ell-1)})\frac{\partial}{\partial\rho}F(\rho_{(\ell-1)}) \tag{24}$$

which is equivalent to Eq. (23) if $\rho_\ell \to \rho$ as $\ell \to \infty$. Given $\rho_{(\ell-1)}$, Eq. (24) can be solved simply by inversion of a linear differential operator as

$$2\left[(\alpha_2+\gamma_2)\frac{\partial^4}{\partial s^4}-(\alpha_1+\gamma_1)\frac{\partial^2}{\partial s^2}+\gamma_0+\kappa\right]\rho_\ell = \bar{p}+\kappa(\rho_\ell-\rho_{(\ell-1)})\frac{\partial}{\partial\rho}F(\rho_{(\ell-1)}) \tag{25}$$

which we have implemented numerically. The iterations are initialized with $\rho_0 = p_1$. Graphically, as the iterations progress, the contour $\rho_\ell(s)$ approaches $p_2(s)$ in a structually constrained manner (from which the name "snake" is derived). When κ is large convergence is slow; when it is small the solution becomes unstable. We have chosen a relatively small value of κ for the first few iterations and then a larger value of κ for the rest of the iterations to ensure convergence. We used a total of about 20 such iterations per solution of Eq. (22).

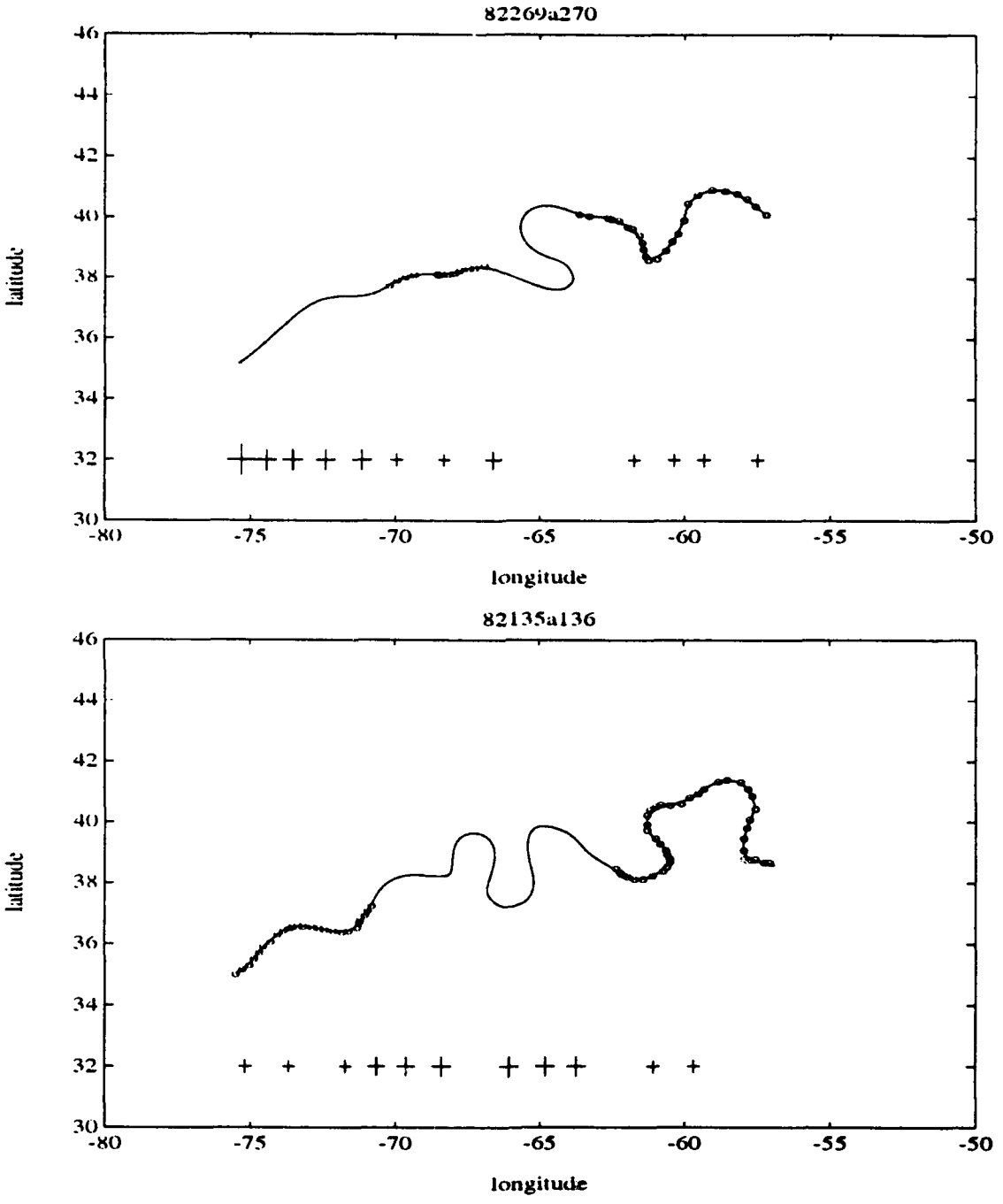


Figure 6. Examples of interpolated GSNWP contours (solid lines). The small circles are the digitized data points. The cross hairs along the 32°N lines are the standard deviations associated with the estimated contour points directly above them. The lengths of the two arms of each cross represent standard deviations in the estimates of the longitude and latitude associated with the estimated point.

RESULTS

Equation (12), along with the feature detection and matching scheme discussed in the previous section, has been used to interpolate 150 frames of data from the period April 1982 ~ February 1983. Figure 6 shows two of the interpolated GSNWP contours along with the digitized data points (small circles), indicating that the bi-variate formulation is able to reconstruct macroscopic features like the "S" and "Ω" shapes by interpolating data from nearby time-frames. The data points from nearby frames are shown in Figure 7. Also, the standard deviations (produced by the Kalman filter-based algorithm) in the longitude/latitude estimates of selected points are depicted in Figure 6 by the crosshairs (see the figure caption). As expected, the standard deviations are larger away from the data points and smaller near the data points.

The algorithm has been tested further by "hind-forecasting": a particular frame of data points is removed, and the contour in that frame is then predicted by interpolation based only on data in other frames. Ideally, the predicted contour matches well with the actual data points which did not participate in the interpolation. (Note, however, that the digitized points in a given frame can sometimes misrepresent the true frontal location because of imaging noise, inconsistency among the personnel who perform the digitization task, etc.) Figure 8 shows the hind-forecasted contours of the same two frames as those in Figure 6, while Figure 9 (cf. Fig. 10) shows the hind-forecasts for another pair of frames. In these figures, the data points match fairly well with the hind-forecasts, and, in fact, the agreement between the data and hind-forecasts is observed generally throughout our test. There are, however, several inconsistent hind-forecasts, two of which are shown in Figure 11 (cf. Fig. 12). As indicated in the figure, a major flaw in these hind-forecasts is inability to resolve some fast movements of the meanders and to detect transformations of the meanders into rings. Obviously, simple smoothness constraints like those in Eq. (12) by themselves are not able to handle events such as formation of rings and are heavily dependent on the data to resolve such events.

DISCUSSION

Although the present-day pattern recognition and matching algorithms have yet to realize flexibility and sensitivity of trained personnel, major advantages of a mechanized system in GSNWP estimation are speed, objectivity, and consistency, which are important in high volume production of the estimates. Also, a probabilistic formulation, such as that presented in this report, yields a measure of confidence in the estimates in the form of the second order statistics to facilitate interpretation of the results. We feel that such a statistical interpretation will be enhanced if the uncertainty (noise variance) in each digitized data point is quantified by using a probabilistic edge-detection algorithm (e.g., Canny, 1986) on the SST images. A new edge detection algorithm using both spatial and temporal constraints is being tested by Cayula and Cornillon (cf. 1990) at URI. A symbiotic merging of such an edge detection with our interpolation algorithm should

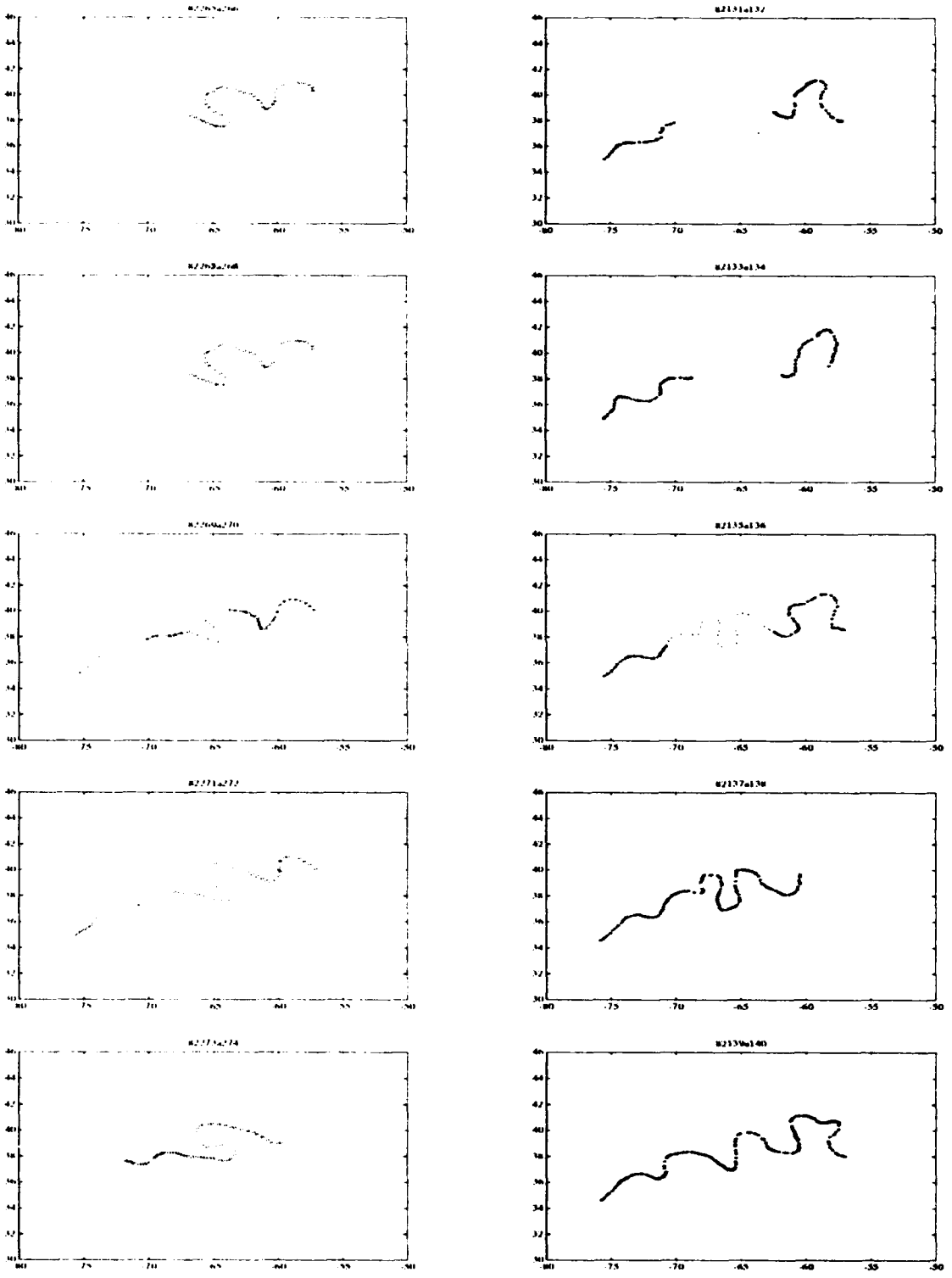


Figure 7. The digitized data points from five frames centered around the two frames depicted on Figure 6. Each of two columns of five frames shows a time-sequence of the digitized data points, with the third frame being the frame from Figure 6.

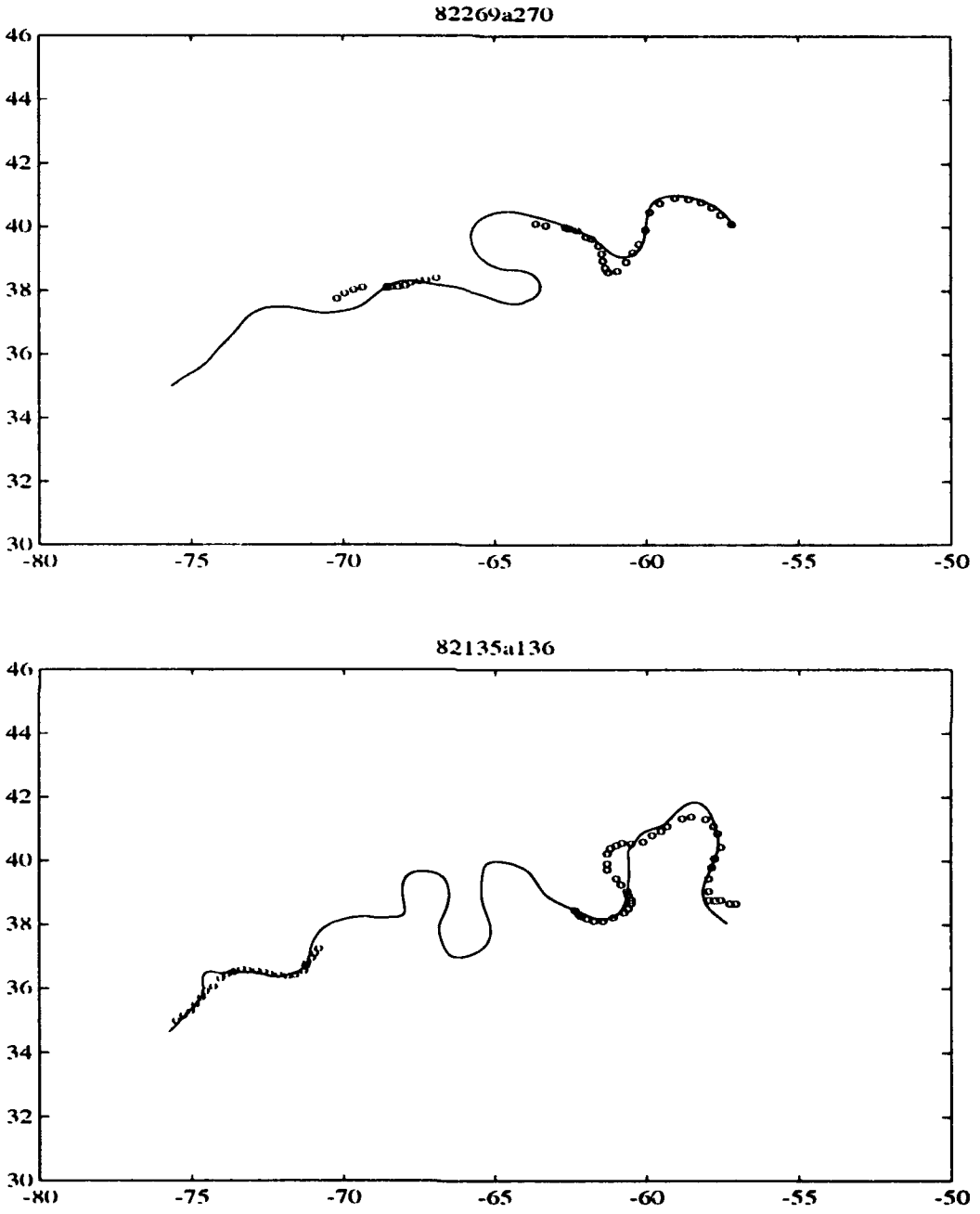


Figure 8. Hind-forecasts for the two frames in Figure 6.

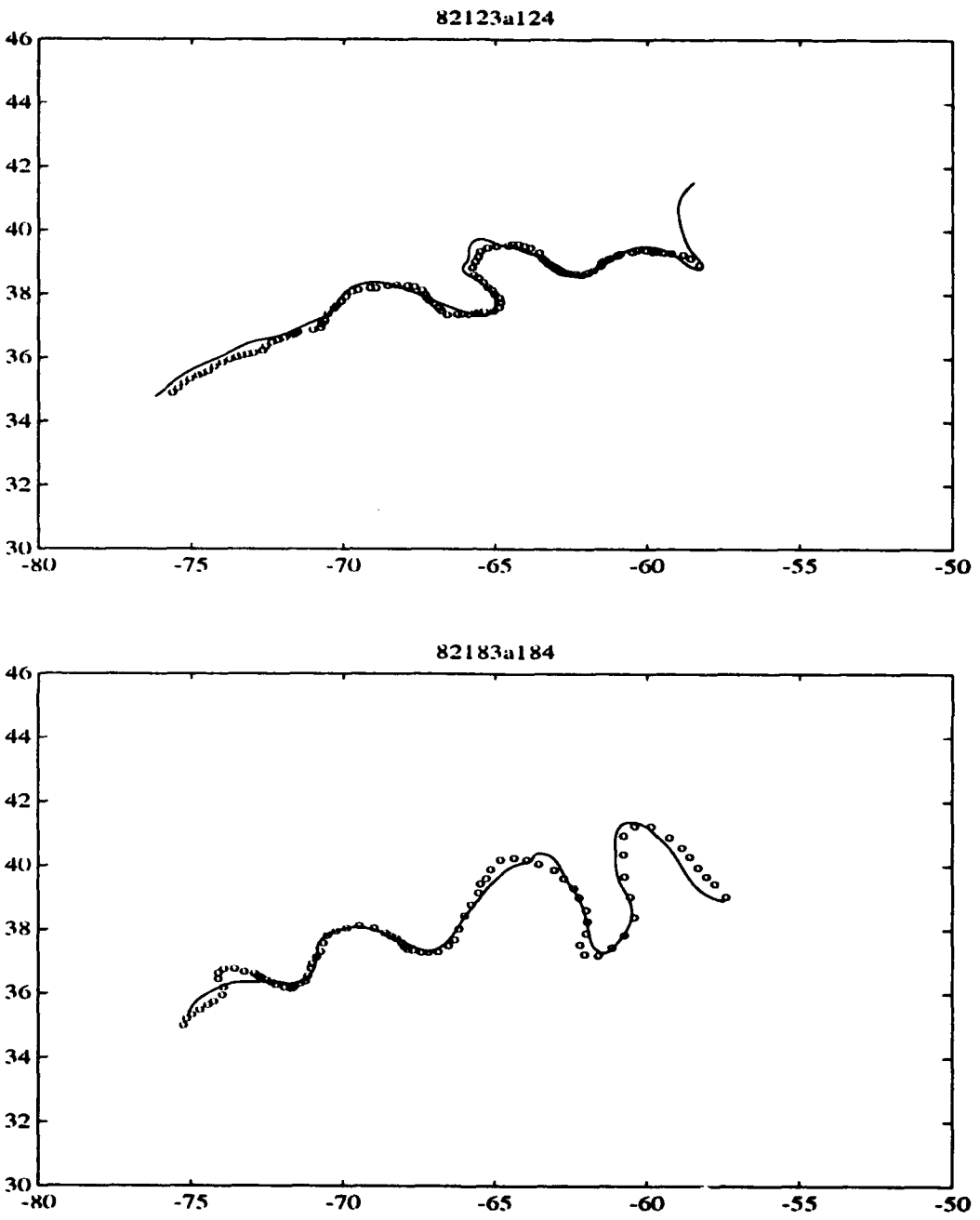


Figure 9. Two more examples from the hind-forecasting test.

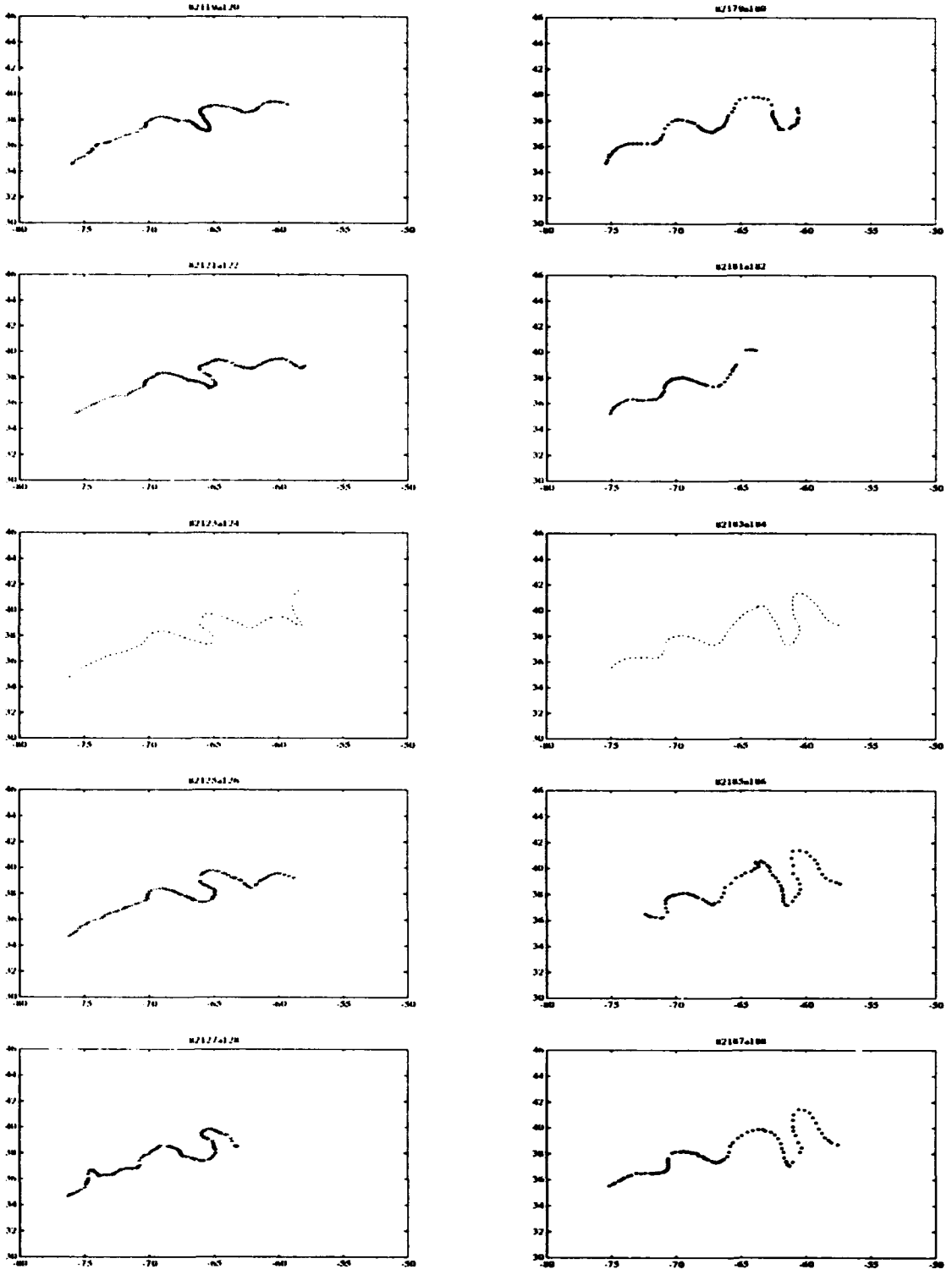


Figure 10. The digitized data points from five frames centered around the two frames depicted on Figure 9. Each of two columns of five frames shows a time-sequence of the digitized data points, with the third frame being the frame from Figure 9.

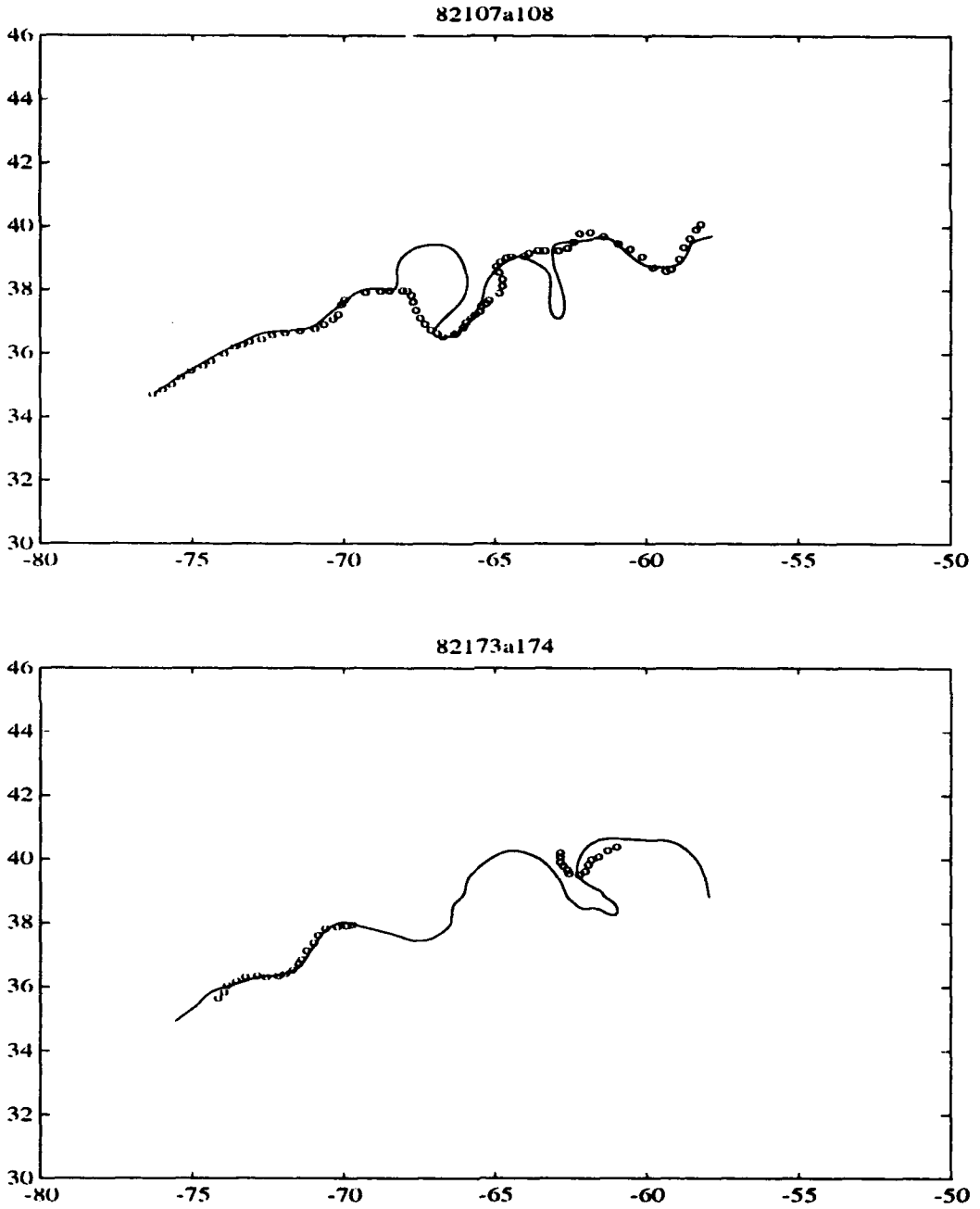


Figure 11. Two cases where hind-forecasts have failed, due to temporal Gulf Stream dynamics unresolvable from this particular data sequence.

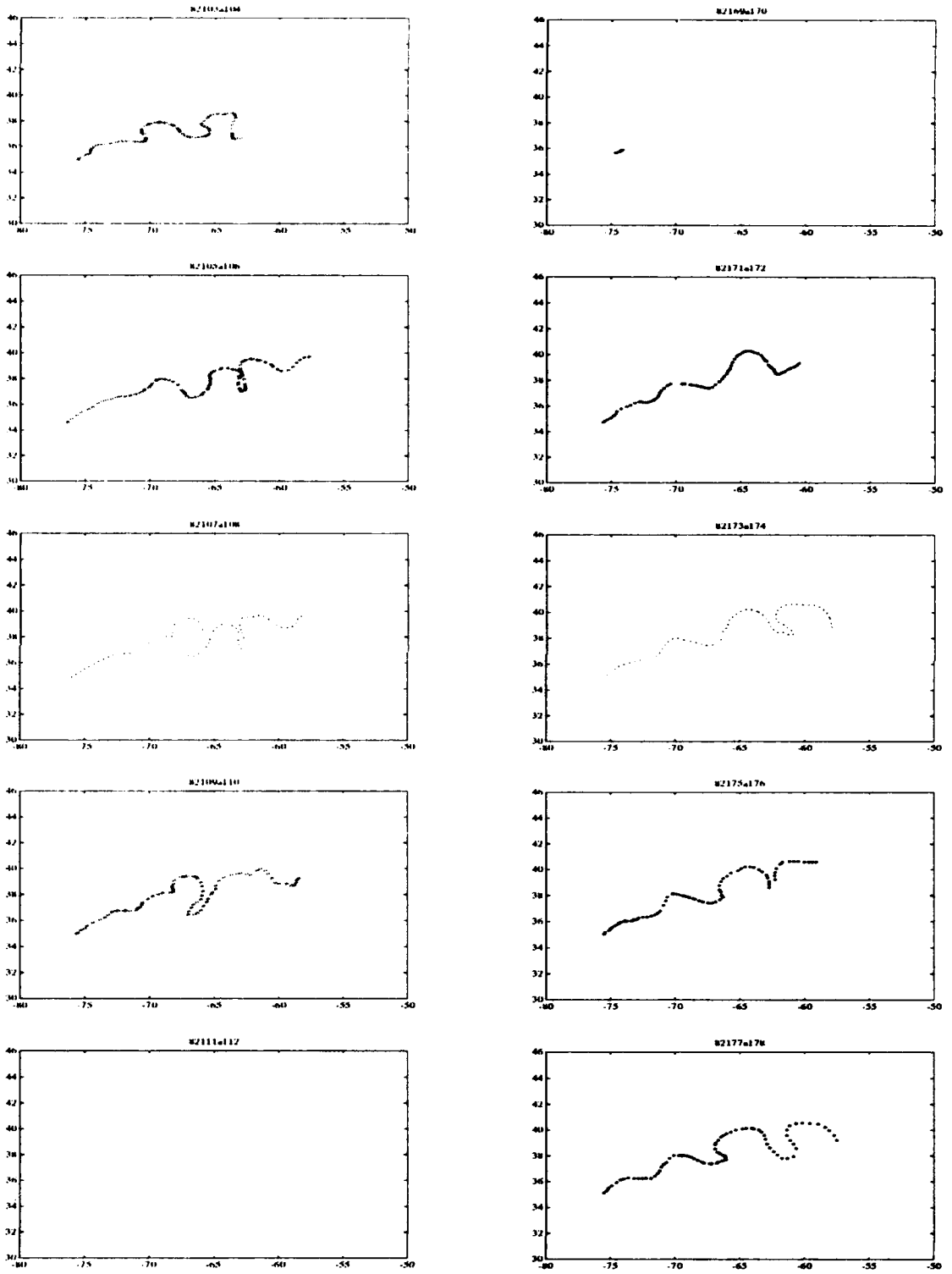


Figure 12. The digitized data points from five frames centered around the two frames depicted on Figure 11. Each of two columns of five frames shows a time-sequence of the digitized data points, with the third frame being the frame from Figure 11.

reduce the effect of inconsistencies in the initial frontal locations. We are also considering a higher order model for contour dynamics (Pratt and Stern, 1986) as an extension of the work presented in this report.

In the near future, all available digitized frontal locations in the Gulf Stream, Brazil-Malvinus confluence, and Kuroshio current systems will be interpolated. The spatial/temporal variability and phase speed distribution of the resulting complete frontal locations will be documented.

Acknowledgments. This work is supported by Office of Naval Research Random Field in Oceanography ARI under Grant N00014-91-J-1120. We would also like to thank Tong Lee and Peter Cornillon for sharing the GSNWP data set with us.

REFERENCES

Anderson, B. D. O., and J. B. Moore, 1979: *Optimal Filtering*. Prentice-Hall, Englewood Cliffs, N.J.

Canny, J. F., 1986: A computational approach to edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-8, 679–698.

Cayula, J. F., and P. Cornillon, 1990: Edge detection applied to SST fields. *Proc. SPIE (Digital image processing and visual communications technologies in the earth and atmospheric sciences)*, 1301, 13–24.

Cornillon, P., 1986: The effect of the New England seamounts on Gulf Stream meandering as observed from satellite IR imagery. *J. Phys. Oceanog.*, 16, 386–389.

Garzoli, S. L., Z. Garraffo, G. Podesta, and O. Brown, 1992: Analysis of a general circulation model product 1. frontal systems in the Brazil/Malvinas and Kuroshio/Oyashio regions. *J. Geophys. Res.*, 97, 20117–20138.

Gelb, A., editor, 1974: *Applied Optimal Estimation*. MIT Press, Cambridge, MA.

Halliwell, G. R., and C. N. K. Mooers, 1983: Meanders of the Gulf Stream downstream from Cape Hatteras 1975-1978. *J. Phys. Oceanog.*, 13, 1275–1292.

Halliwell, G. R., and C. N. K. Mooers, 1979: The space-time structure and variability of the shelf water–slope water and Gulf Stream surface temperature fronts, and associated warm-core eddies. *J. Geophys. Res.*, 84, 7707–7726.

Inoue, H., 1986: A least-squares smooth fitting for irregularly space data: finite-element approach using the cubic *B*-spline basis. *Geophysics*, 51, 2051–2066.

Kass, M., A. Witkin, and D. Terzopoulos, 1988: Snakes: active contour models. *Intl. J. Computer Vision*, 1, 321–331.

Lewis, F. L., 1986: *Optimal Estimation*. John Wiley & Sons, New York.

Mariano, A. J., 1990: Contour analysis: a new approach for melding geophysical fields. *J. Atmos. Oceanic Technol.*, 7, 285–295.

Mariano, A. J., 1988: *Space-time interpolation of Gulf Stream north wall positions*. Harvard Open Ocean Model Reports 29, Harvard University.

Olson, D. B., O. B. Brown, and S. R. Emmerson, 1983: Gulf Stream frontal statistics from Florida Straits to Cape Hatteras derived from satellite and historical data. *J. Geophys. Res.*, 88, 4569–4577.

Pratt, L. J., and M. E. Stern, 1986: Dynamics of potential vorticity fronts and eddy detachment. *J. Phys. Oceanogr.*, 16, 1101–1120.

Szeliski, R., 1989: *Baysian Modeling of Uncertainty in Low-level Vision*. Kluwer Academic Publishers, Norwell, Massachuesetts.

Watts, D. R., and W. E. Johns, 1982: Gulf Stream meanders: observations on propagation and growth. *J. Geophys. Res.*, 87, 9467–9476.

SOME NOTES ON DATA ASSIMILATION IN PHYSICAL OCEANOGRAPHY

James J. O'Brien
Mesoscale Air-Sea Interaction Group, The Florida State University, Tallahassee, FL 32306-3041

INTRODUCTION

This paper is a discussion of the author's emphasis on data assimilation in physical oceanography. The work draws on recent work by members of the MASIG Team. Our approach has focused on time-dependent models in which parameters are estimated through data assimilation using the variational adjoint method.

It is useful to adapt a paradigm for classifying all data assimilation methods. I chose to define three groups of assimilation schemes: (A) local polynomial interpolation methods, (B) statistical (including optimal) interpolation methods, and (C) variational numerical analysis methods.

In (A), the idea is to expand the data misfit in terms of some interpolating polynomial in the spatial vicinity of the data location; direct insertion or substitution or "bogusing" are some simple examples; Cressman filters are a commonly used meteorological assimilation technique. No knowledge of the statistical property of the data or the model is used.

In (B), we use statistical information of the data error field or the model variability to determine the adjustment in space and time. In principle, one could estimate the cross-correlation function of the data misfit and adopt some rules to adjust the model solution. The simplest idea is the so-called nudging method where an inverse-time parameter is used to estimate the variability of the data misfit. The most sophisticated example is the Kalman-Bucy filtering method. In all the implementation schemes one should imagine that the physical model is evolving in time, and a moment arrives when a data value is encountered. The data misfit is then added to the model field in time and space. If the covariance matrix structure is primarily spatial, then the simplest time structure for the variability is nudging where a linear time decay processes is added to the prognostic model.

In (C), the assimilation scheme defines a statistically weighted data misfit field, which is minimized in a construct such that the complete physics of the prognostic model is included as dynamical constraints. I will concentrate on examples of this latter method.

THE VARIATIONAL ADJOINT METHOD

The essential ingredients in this data assimilation are a "nice" model, availability of some "useful" data, and a willingness to adjust the model in some manner. Each of these elements must each be appreciated. The model should produce validated solutions that are reasonable and "liked." The data may be estimates of model-dependent state variables or the data may be any function of a dependent state variable as long as an estimate of the function can be calculated from the model output. A simple example would be ocean altimeter cross-over data. The difference in time between two altimeter readings at a point can be estimated from the solution to any ocean model that simulates sea level, and therefore altimeter cross-overs can be assimilated.

For a contrived example, let us consider the following model. Suppose a scalar field, $c(x,t)$, is advected by an unknown advection field, $u(x)$, and other processes are represented by $g(c,\beta)$ where β is a poorly defined parameter. We "like" our model after we guess u,β and the initial conditions, $c'(0,x)$. We acquire some data, $F'(c)$, where $F(c)$ is any function of c which we can estimate from the output of $c(x,t)$. The model is

$$c_t + uc_x = g(c,\beta). \tag{1}$$

There are many avenues to arrive at the variational problem. I choose simply to write down the functional

$$\begin{aligned} H(c,\lambda,u,\beta) = &\int_{x,t}^{T} \lambda(c_t + uc_x - g)\,dx\,dt \\ &+ \int_{x,t}^{T} \frac{K_c}{2}(F(c) - F'(c))^2\,dx\,dt \\ &+ \int_{x,t}^{T} \frac{K_u}{2}(u - u')^2\,dx\,dt \\ &+ \int_{x,t}^{T} \frac{K_\beta}{2}(\beta - \beta')^2\,dx\,dt \end{aligned} \tag{2}$$

where $\lambda(x,t)$ is a Lagrange multiplier, K_c, K_u, K_β are called Gauss precision modulae. The range of space is over all x, say, $x \in [0,L]$ and periodic, e.g., $c(t,x) = c(t,x+L)$. The last three terms are called the cost function, which is to be minimized subject to the contraint that the data, F', and the advection function, u, and the parameter, β, must satisfy the model. The range of time is $[0,T]$; T is a time later than the last observed datum.

The minimum is determined by the usual approach,

$$\frac{\partial H}{\partial \lambda} \text{ yields } c_t + uc_x = g(c,\beta) \tag{3}$$

where u and β are now not known.

$$\frac{\partial H}{\partial u} \text{ yields } u(x) = u'(x) - \frac{1}{TK_u}\int_t^T \lambda c_x \, dt \tag{4}$$

where we observe that the correction of u from its guess field depends on the average of the product of the Lagrange multiplier and the spatial gradient of the dependent state variable.

$$\frac{\partial H}{\partial \beta} \text{ yields } \beta = \beta' - \frac{1}{TLK_\beta}\int_{x,t}^{T} \lambda \frac{\partial g}{\partial \beta} dx\, dt \tag{5}$$

and

$$\begin{aligned}\frac{\partial H}{\partial c} \text{ yields } \frac{\partial \lambda}{\partial t} + (u\lambda)_x = -\lambda\frac{\partial g}{\partial c} + K_c\left[F(c) - F'(c)\right]\frac{\partial F}{\partial c} \\ + \int_x \lambda(0,x)\, c(0,x)\, dx + \int_x c(0,x)\, dx.\end{aligned} \tag{6}$$

The next to last integral vanishes using the lemma that a product of periodic functions is periodic. We have used the natural spatial boundary conditions for λ and chosen $\lambda(T,x) = 0$. Note that the last integral is zero except at $t = 0$.

The solution procedure is as follows:

1. Guess u', β', $c'(0,x)$ and calculate the solution forward over the time $[0,T]$ from Eq. (1).
2. Calculate the data misfit, F-F', and the data misfit transfer function, F_c, and integrate Eq. (6) backwards in time from T to zero.
3. Next adjust the initial conditions and $u(x)$ and β using Eqs. (4, 5, and 6) (for $c(0,x)$).
4. Repeat 1,2, and 3 as often as desired in order to assimilate the data, $F(c)$.

There are many advantages to this algorithm. It will almost always converge; thus all the data are used and it is eloquent. I am told that one can contrive a case where it will not converge. There are some disadvantages. It is very expensive because we have to integrate two models and save the solution from both models, particularly when the physical model is nonlinear. It may take many forward and backward integrations to find the miminum. An emphasis of current research is to identify algorithms which find the minimum in as few integrations as possible. The present view is to implement an efficient conjugate gradient algorithm. A further disadvantage is that this method is difficult to teach to scientists. We are beginning to have several simple examples that will demonstrate the method to scientists.

A SIMPLE REAL OCEAN EXAMPLE

There have only been a few modern superb upper-ocean data expeditions that have measured meteorological and upper ocean currents. One such experiment is LOTUS from which Briscoe, Price and Weller have provided us data to develop data assimilation methods. Suppose we wish to assimilate wind and ocean current data and determine the momentum drag coefficient and the mixing function for momentum, $A(z)$.
The model equation is

$$\frac{\partial w}{\partial t} + ifw = \frac{\partial}{\partial z}(A\frac{\partial w}{\partial z}) \tag{7}$$

where $w = u + iv$. The boundary conditions are at

$$z = 0, \text{ and } \rho_w A\frac{\partial w}{\partial z} = \tau \tag{8}$$

where the wind stress is calculated from

$$\tau = \rho_a c_D |w_a| w_a.$$

At the bottom

$$z = -H, \text{ and } A\frac{\partial w}{\partial z} = 0. \tag{9}$$

The initial condition for this dynamic system is at $t = 0$ and $w = w_0$. We chose to nondimensionalize the system as follows:

$$t' = \frac{t}{T_f},\ w' = \frac{w}{U},\ z' = \frac{z}{D},\ A' = \frac{A}{s_a},\ c_D' = \frac{c_D}{s_c},\ w_a' = \frac{w_a}{U_a}$$

where

$$T_f = f^{-1},\ D = \sqrt{\frac{s_a}{f}},\ \text{and } U = \left(\frac{\rho_a}{\rho_w} s_c\right)\frac{U_a^2}{\sqrt{s_a f}},$$

which yields the model

$$\frac{\partial w}{\partial t} + iw = \frac{\partial}{\partial z}\left(A\frac{\partial w}{\partial z}\right) \tag{10}$$

with

$$A\frac{\partial w}{\partial z} = \begin{cases} c_D |w_a| w_a & \text{for } z = 0 \\ 0 & \text{for } z = -\dfrac{H}{D} \end{cases} \tag{11}$$

$$\text{and } w = w_0 \text{ for } t = 0.$$

Following the formalism developed in the previous section, we define the cost function, J:

$$J(w,A,c_D)=\frac{1}{2}K_m\int_t\int_z(w-\hat{w})^2\,d\zeta\,dt \\ +\frac{1}{2}K_aT\int_z(A-\hat{A})^2\,d\zeta+\frac{1}{2}K_cTH(c_D-\hat{c}_D)^2. \tag{12}$$

The functional, L, is the sum of the cost function and the constraint

$$L(w,A,C_D,\lambda)=J+\int_t\int_z\left\{\lambda\left(\frac{\partial w}{\partial t}+ifw-\frac{\partial}{\partial z}\left(A\frac{\partial w}{\partial z}\right)\right)\right\}d\zeta\,d\tau. \tag{13}$$

The solution is found as usual by solving

$$\frac{\partial L(w,A,c_D,\lambda)}{\partial\lambda}=0$$
$$\frac{\partial L(w,A,c_D,\lambda)}{\partial w}=0$$
$$\frac{\partial L(w,A,c_D,\lambda)}{\partial A}=0$$
$$\frac{\partial L(w,A,c_D,\lambda)}{\partial c_D}=0.$$

This yields the model plus

$$\frac{\partial\lambda}{\partial t}+i\lambda+\frac{\partial}{\partial z}\left(A\frac{\partial\lambda}{\partial z}\right)=K_m(w-\hat{w}) \tag{14}$$

$$c_D=\hat{c}_D+\frac{1}{K_cTH}\int_t\left(|w_a|u_a\lambda_{uz=0}+|w_a|v_a\lambda_{vz=0}\right)d\tau \tag{15}$$

$$A=\hat{A}+\frac{1}{K_aT}\int_t\left(\frac{\partial u}{\partial z}\frac{\partial\lambda u}{\partial z}+\frac{\partial v}{\partial z}\frac{\partial\lambda v}{\partial z}\right)d\tau. \tag{16}$$

Note that we have assumed that c_D is a constant and A is only a function of depth. We can rescale the parameters, K, by using

$$\frac{\lambda}{K_M}=\lambda',\ \frac{K_c}{K_m}=K_c',\ \text{and}\ \frac{K_a}{K_m}=K_a'.$$

This yields

$$\frac{\partial \lambda}{\partial t} + i\lambda + \frac{\partial}{\partial z}\left(A\frac{\partial \lambda}{\partial z}\right) = (w - \hat{w}) \tag{17}$$

$$c_D = \hat{c}_D + \frac{1}{K_c TH}\int_t \left(|w_a| u_a \lambda_{uz=0} + |w_a| v_a \lambda_{vz=0}\right) d\tau \tag{18}$$

$$A = \hat{A} + \frac{1}{K_a T}\int_t \left(\frac{\partial u}{\partial z}\frac{\partial \lambda u}{\partial z} + \frac{\partial v}{\partial z}\frac{\partial \lambda v}{\partial z}\right) d\tau. \tag{19}$$

In order to solve these equations we need to define a solution space. This is shown in Figure 1.

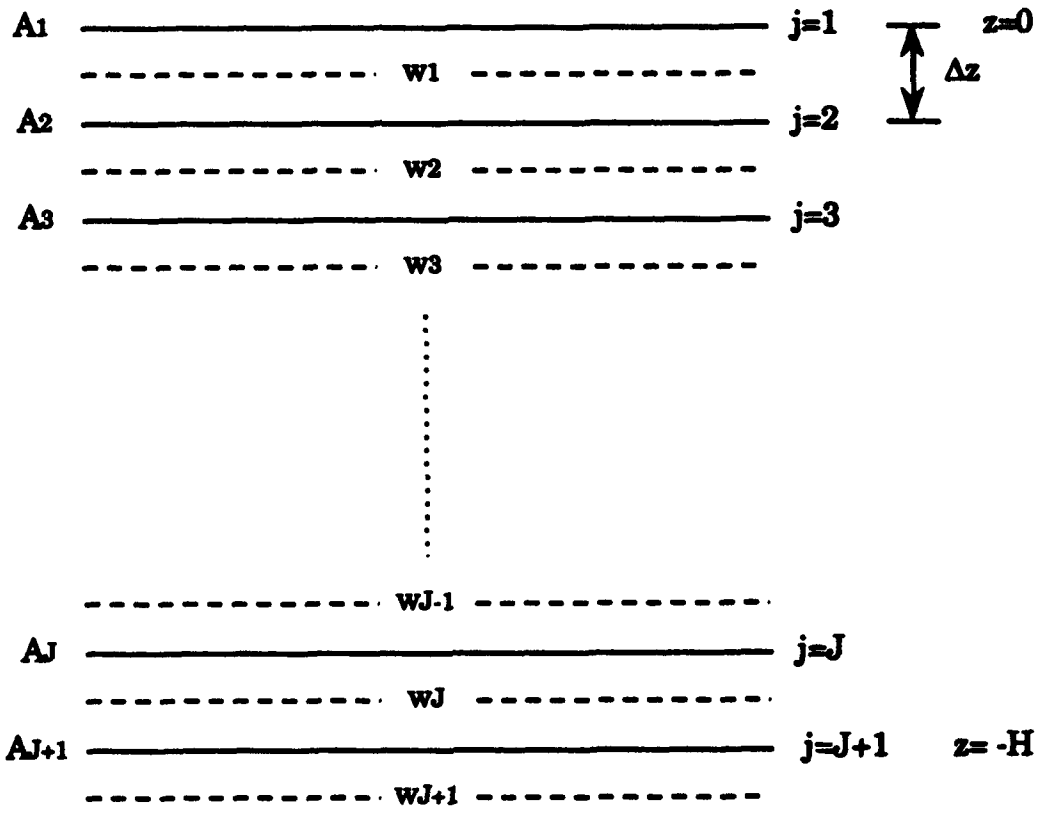


Figure 1. Diagram of the vertical structure of the numerical model.

Our procedure for using the variational method to solve this system can be fully described:

(1) Begin with a best initial estimate for the control parameters A and c_D.
(2) Integrate the model equation (7) forward in time and calculate the value of the cost function.
(3) Compute the data misfits $(w-\hat{w})$.
(4) Integrate the adjoint equation (17) backward in time.
(5) Use equations (18) and (19) to calculate the gradients of the cost functions ∇J corresponding to A and c_D with solutions for λ and w from steps (2) and (4).
(6) With the gradient information, apply the descent algorithm to obtain the new values of A and c_D simultaneously.
(7) Check if the minimization process is done. The convergence criterion is satisfied if $|\nabla J|/|\nabla J_0| < 10^{-2}$, where ∇J_0 is the value at the initial iteration.
(8) Return to step (2) if the optimal solution is not found.

We will demonstrate a solution using currents over 10 days in the summer in the North Atlantic during the LOTUS experiment. Figure 2 shows the observed currents at 5 and 15 meters. Only a low frequency trend has been omitted from the original data. The cost function is shown and the gradient are shown in Figure 3 as a function of interation. Note that the cost function reaches a "*practical*" *minimum in four iterations.* The profile of the eddy viscosity coefficient and the drag coefficient are shown in Figure 4. The surface value of 0.003 implies an "Ekman Layer Depth" of about 6-8 m. The comparison of the assimilated data with the data is shown in Figures 5 and 6. It is seen that the model reproduces the current meter data above 65 m quite well and very poorly below. This is a simple example of ocean data assimilation. This research is available in detail in Yu and O'Brien (1991). In Yu and O'Brien (1992), we also change the initial condition with improvement (Table 1). There are additional, completed examples of this work showing how to assimilate sea level. In this report I have not tried to reference all the important works by other research teams.

Table 1.
Change of Correlation Coefficient with Depth

Depth Z	New r*	Old r
5	0.92	0.87
25	0.88	0.81
35	0.71	0.67
75	0.34	0.28
95	0.53	0.44
*Max A	1.4×10^{-3}	2.9×10^{-3}
c_D	1.2×10^{-3}	1.3×10^{-3}

* Initial condition adjusted.

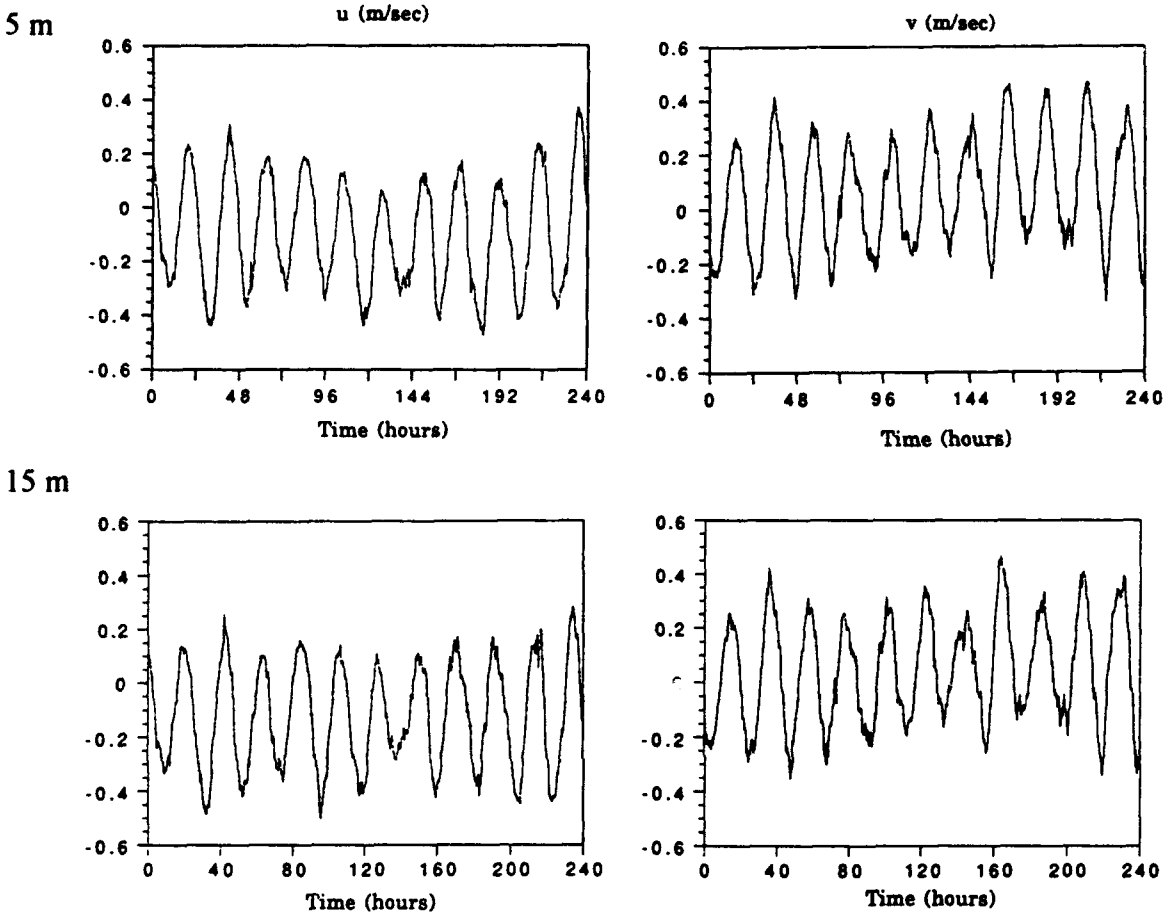


Figure 2. Current observations at 5 m (top) and 15 m (bottom).

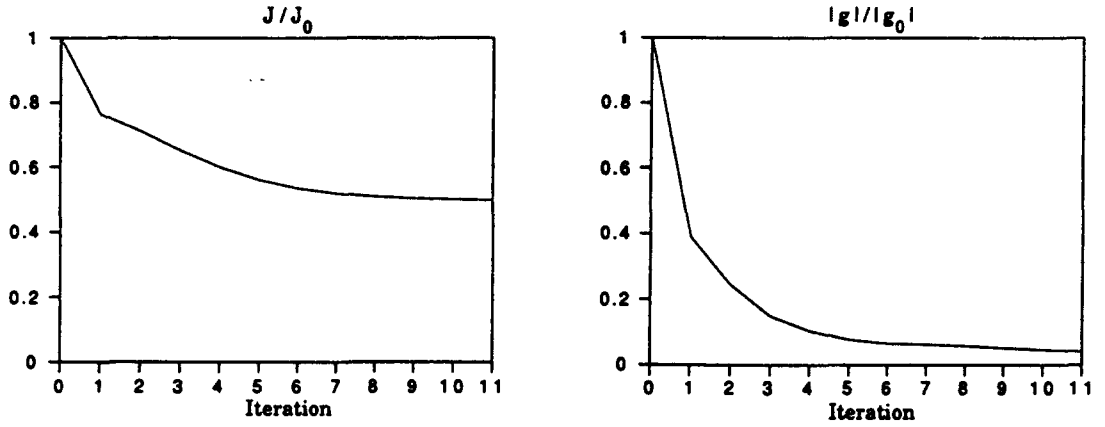


Figure 3. The variation of (left) the cost function and (right) the gradient with the number of iterations.

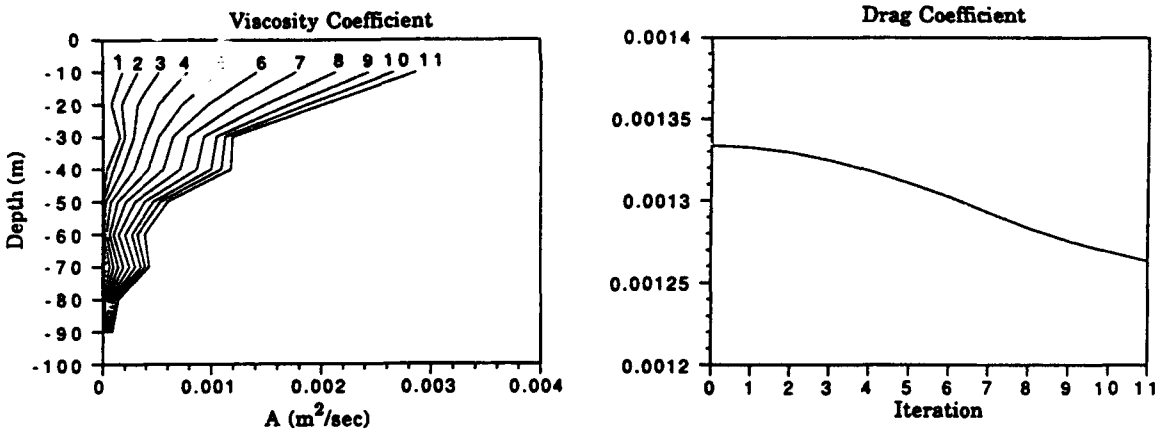


Figure 4. (left) The variation of the eddy viscosity coefficient during the iterative process, and (right) the variation of the drag coefficient with the number of iterations.

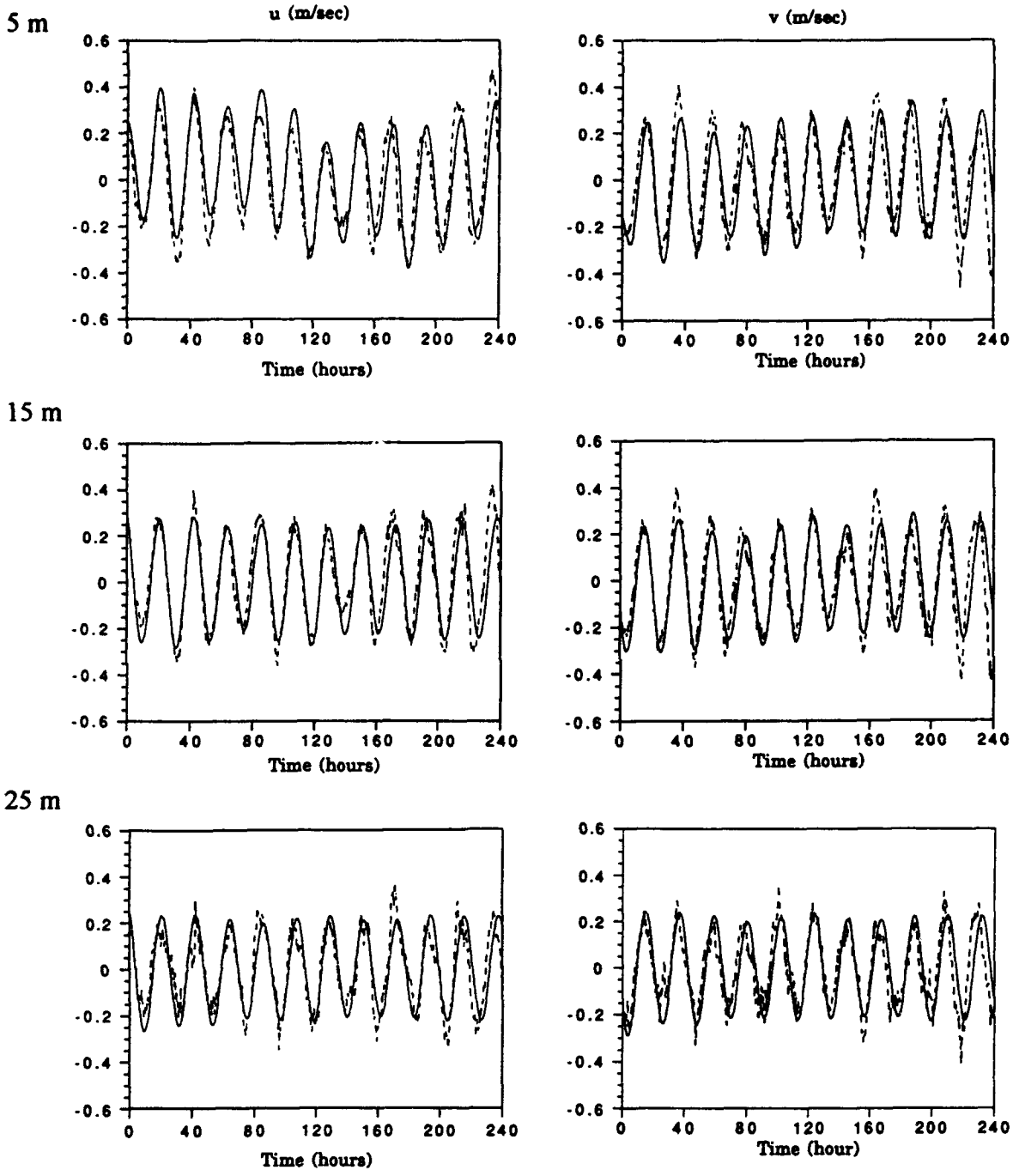


Figure 5. Comparison of modelled (solid lines) and observed (dashed lines) current speeds u (left) and v (right) for 5, 15, and 25 m.

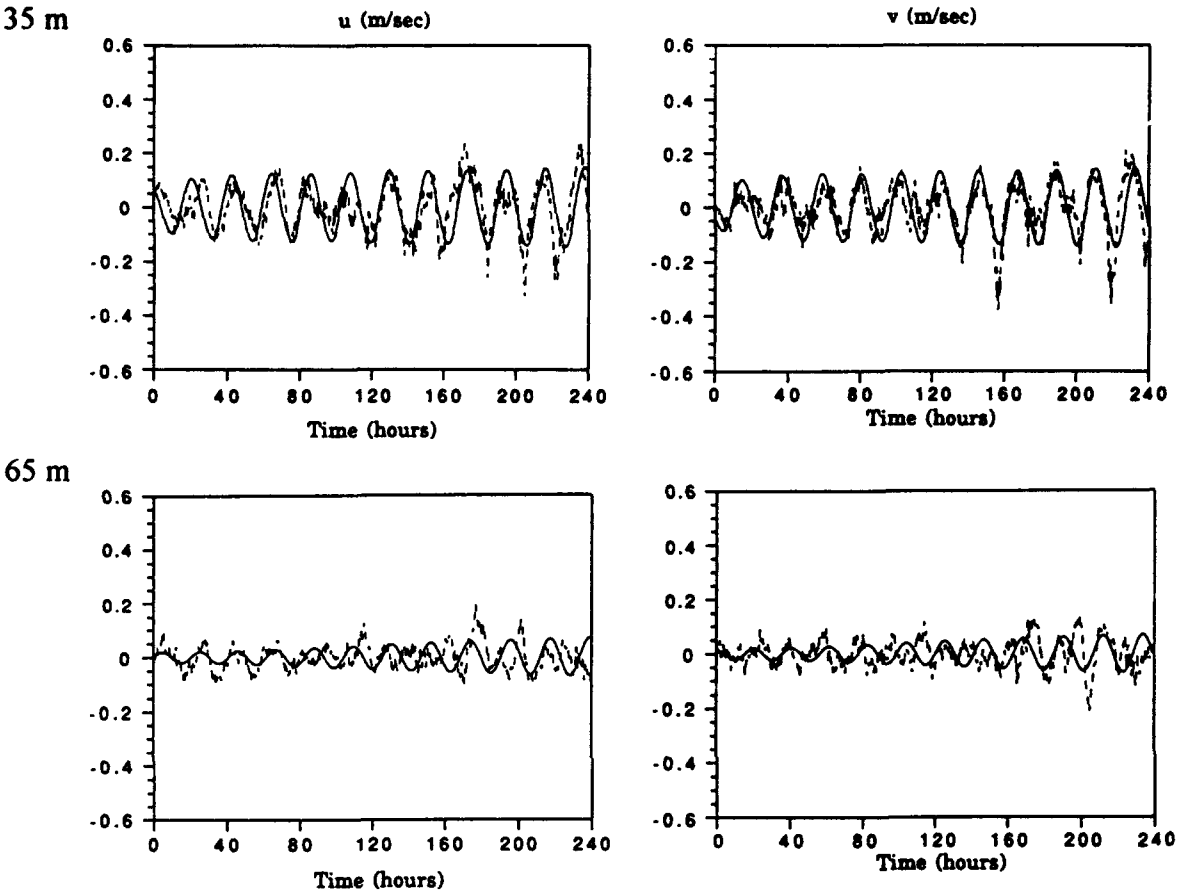


Figure 6. Comparison of modelled (solid lines) and observed (dashed lines) current speeds for *u* (left) and *v* (right) at, 35 and 65 m.

Acknowledgments

This work has been supported by NASA Oceanic Processes. Additional support from ONR is important. The data were supplied by Bob Weller, WHOI. In addition, Mel Briscoe and Jim Price were very helpful. The calculations were done by Lisan Yu. Rita Kuyper did a great job on the typing and everything else.

REFERENCES

Yu, L. and J. J. O'Brien, 1991: Variational Estimation of the Wind Stress Drag Coefficient and the Oceanic Eddy Viscosity Profile, *J. Phys. Oceanog.*, 21, No. 5, 709-719.

Yu, L. and J. J. O'Brien, 1992: On the Initial Condition in Parameter Estimation, *J. Phys. Oceanog.*, 22, 1361-1364.

ESTIMATION OF FREE PARAMETERS BY INVERSE MODELING

Jens Schröter
Alfred-Wegener-Institute for Polar and Marine Research
Am Handelshafen 12, D 2850 Bremerhaven, F.R.G.

ABSTRACT

An objective procedure is presented which allows the systematic determination of free model parameters in numerical models. A nonlinear inverse technique is applied to fit the model to observations. Optimal values for the free parameters are found in a systematic way by minimizing the least squares distance between modeled and observed data.

The method is applied to a general circulation model (GCM) of the Atlantic Ocean. The GCM includes an embedded mixed layer model based on the equation of turbulent kinetic energy. A number of free parameters describe, e.g., the efficiency of wind stirring, decay scales of turbulence, and so forth. They are determined by fitting the annual cycle of the modeled mixed layer depth to climatological data. The parameter values and their error covariance matrix are computed.

INTRODUCTION

Adjustable model parameters are involved in almost all numerical ocean models. As an example, horizontal exchange of momentum or of tracers that is due to small scale processes is commonly described by a diffusion term. Another example is the drag coefficient that is used to convert surface wind speed in the atmosphere to surface stress of the ocean.

These parameters may serve different purposes. The diffusion coefficient is primarily intended to describe directly the effect of mixing and stirring. On the other hand a much larger coefficient may be necessary to insure numerical stability or damp out computational modes. It is therefore necessary to define exactly the purpose of the parameter involved before values are assigned to it.

Values are often chosen according to the intuition of the modeler. If the intuition fails, a small parametric study may help. One example for this type of study is the treatment of bottom friction. Here models are frequently "tuned" by trying out a few bottom friction coefficients that span two or three orders of magnitude. The coefficient that leads to the "best results" is then chosen.

Parameters of mixed layer models have been tuned to fit the data of a certain location, such as Ocean Station Papa (Martin, 1985), or of the equatorial Pacific (Garwood et al.,

1985a,b). Some of the parameters, such as the efficiency of wind stirring m_o, can be measured in laboratories. It is now our task to find out if the same values are applicable in the context of a global model.

An objective way to fit models to data is the application of inverse techniques. Distributions of active or passive tracers may be inverted to derive flow velocities and diffusion coefficients (Wunsch, 1985, Fiadero and Veronis 1984, Olbers et al., 1985). In these inversions more or less complicated models are applied. In the following we will describe a method to determine free parameters of highly nonlinear models. The technique is iterative and very general. In the example given below it is applied to a mixed layer model that is coupled to a general circulation model. Following ideas suggested by Tarantola and Valette (1982), a sequence of linear subproblems is solved wherein each solution is a compromise between observation and prior information.

A brief summary of the general circulation model and the mixed layer model is given in the next sections, followed by the presentation of the data, the inverse method and finally the results.

ISOPYCNIC OCEAN CIRCULATION MODEL

A general circulation model that uses isopycnical coordinates in the vertical was used in this study. The model was developed by Oberhuber (1993a,b) and is known under the name "OPYC." It includes an ice model with viscous-plastic rheology. The surface layer is modeled as a fully active mixed layer of variable depth in which temperature and salinity may change arbitrarily. The mixed layer is coupled interactively to the ice model as well as to the deeper, isopycnic layers. One of the intentions in deriving the isopycnic model is its use in climate studies. For this purpose the model formulation was made rather complete. It combines primitive equations and the full thermohaline dynamics, a realistic equation of state, convection and detailed mixed layer dynamics with an isopycnical description of the deep ocean. Topography is arbitrary.

An early version of the model has been applied to the tropical Pacific (Miller et al. 1991). The most intensive studies were, however, performed in the Atlantic Ocean. The model has been described in detail in Oberhuber (1993a,b). The present study was undertaken in support of the model development and an earlier version of OPYC was applied. The results presented here are, accordingly, only preliminary and the successes of the isopynic model should be judged by the more recent work of Oberhuber. The model version used here has a horizontal resolution of 2° by 2° and seven vertical layers. It covers the Atlantic Ocean from 30°S to 80°N where it is closed by artificial boundaries.

The model is driven by surface fluxes of momentum, heat, fresh water, turbulent kinetic energy, and buoyancy. The fluxes are calculated from monthly mean climatological values. Windstress is taken from Hellermann and Rosenstein (1983). The other fluxes are based

on the COADS data set (Woodruff et al., 1987). The climatology was calculated by Wright (1988) for the years 1950 to 1979 on a 2° by 2° grid. From these Oberhuber (1988) derived all other quantities necessary to drive the model.

MIXED LAYER MODEL

Mixing in the surface layer is caused by turbulence generated by wind stirring and buoyancy fluxes. The turbulence produces a uniform vertical distribution of temperature and salinity. However the turbulent kinetic energy (TKE) may vary with depth within the mixed layer. Models of the mixed layer are generally based on a budget equation of the TKE: The input of TKE at the surface is balanced partly by dissipation and partly used for the production of mean potential energy by the entrainment of underlying denser water. While wind stirring always acts as a source for TKE the buoyancy flux may change sign. Cooling and evaporation act as production terms while precipitation and heating of the surface layer increase the stability and limit vertical mixing. When the warming is sufficiently strong detrainment occurs. A new shallow mixed layer is established in which the input of TKE by the wind is used to distribute the heat vertically and produce potential energy. In OPYC the underlying old mixed layer is redistributed into the isopycnic layers below. While entrainment is modeled prognostically the detrainment is treated separately.

At an early stage in the development of OPYC the mixed layer models of Kraus and Turner (1967), Niiler (1975), Niiler and Kraus (1977) and Garwood et al. (1985a,b) were applied. The experience gained from these models led to a new formulation for the mixed layer equations. The major process that governs the mixed layer depth (MLD) is the entrainment/detrainment cycle. Additionally, the model includes changes in MLD due to convergence of mass or heat.

The entrainment rate w is modeled by

$$\begin{aligned} whg' = {} & wRi_{crit}(\Delta u^2 + \Delta v^2) + 2m_o a u_*^3 + hbe(B - \gamma B_S) \\ & + be\gamma B_S\left[h\left(1+\exp\left(\frac{-h}{h_B}\right)\right) - 2h_B\left(1-\exp\left(\frac{-h}{h_B}\right)\right)\right] \\ & - c\frac{12}{7}m_o\tau_x\Omega_y - dh - d' \end{aligned} \tag{1}$$

where

$$g' = \frac{\Delta\sigma}{\sigma}g \tag{2}$$

$$B = \frac{g}{c_p\rho^2}(\alpha Q + \beta R) \tag{3}$$

$$R = \frac{c_p \rho}{S}(P - E) \tag{4}$$

$$B_S = \frac{\alpha g}{c_p \rho^2} Q_S \tag{5}$$

$$a = \exp\left(-\frac{h}{\kappa}\frac{f}{u_*}\right) \tag{6}$$

$$b = \begin{cases} \exp\left(-\frac{h}{\kappa}\frac{f}{u_*}\right) & B < 0 \\ \exp\left(-\frac{h}{\mu}\frac{f}{u_*}\right) & B > 0 \end{cases} \tag{7}$$

the h is the MLD, g' the reduced gravity between the mixed and the underlying layer. The critical Richardson number (0.25) is denoted Ri_{crit}, Δu and Δv are the differences in the horizontal velocities between the mixed and the underlying layer. The friction velocity is denoted u_*, B is the total buoyancy flux through the surface comprising the total heat flux Q and the equivalent heat flux R due to the fresh water flux (P-E). The buoyancy flux B_s is produced by the solar radiative heat flux Q_s; γ describes the fraction of solar radiation that is not immediately absorbed and that enters the ocean. The scaling depth for the penetration is h_B. If the MLD is sufficiently shallow, lower layers may gain heat by solar radiance. In the term involving $\tau_x \Omega_y$ the northward component of the planetary rotation Ω_y allows the exchange between horizontal and vertical turbulence according to Garwood et al. (1985a,b). The two dissipation terms d and d' will act proportional to and independent of the MLD, respectively.

The finding that less turbulent kinetic energy is needed for mixing at high latitudes than at low ones is modeled as an efficiency term which depends on the Ekman scale. Two functions, denoted a and b, describe which part of the kinetic energy input is available for conversion into potential energy at the mixed layer depth h. They describe an exponential decay that depends on $-hf / \kappa u_*$ where f is the coriolis parameter. Functions a and b differ in their length scales κ and μ. Here negative buoyancy fluxes are treated like wind stirring (function a) while a positive buoyancy flux such as cooling is considered to be more efficient (function b). Buoyancy fluxes can be scaled independently from wind stirring with the coefficient e.

When sea ice is present additional terms appear in eq. (1) (Oberhuber, 1993a). In this study, however, these terms were not treated as variable and are not shown here for simplicity.

In the detrainment phase, the entrainment velocity w is set to zero and eq. (1) is solved diagnostically. Additionally the resulting Monin-Obukhov depth is bounded for small values by the MLD due to vertical velocity shear h_{Ri}:

$$h_{Ri} = Ri_{crit}(\Delta u^2 + \Delta v^2) / g' . \qquad (8)$$

The set of adjustable parameters under consideration now consists of the efficiency of the wind stirring at the surface m_o, the decay scales κ and μ which determine the decay functions a and b, the fraction γ of the solar heating B_s and its penetration scale h_B in the ocean, the efficiency e of buoyancy forcing, the coefficient c that governs the $\tau_x\Omega_y$ term, and the two dissipation coefficients d and d'.

DATA

Climatological hydrographic data compiled by Levitus (1982) are used to determine the annual cycle of the MLD. The problem is that measurements based on turbulence are not available for the whole area of the Atlantic Ocean. Instead our definition must be based on the effect of turbulence on the vertical structure of mean quantities. Several ways are possible to define how deeply the surface layer is mixed. A common approach is to define the depth of the mixed layer as the depth at which either density or temperature deviate from their surface value by a certain margin. When the bottom of the mixed layer is characterized by large steps in mean values of temperature and salinity the choice of the criterion is not critical. However, a definition of the measured MLD based on temperature differences will work more reliably in low latitudes, whereas for high latitudes with their low vertical temperature gradients a criterion using density differences is preferable. As our model includes latitudes up to 80 degrees north, we have chosen a difference in σ_t to define the mixed layer depth.

Monthly mean values of temperature and salinity are used to calculate the mean density profile at standard levels. Linear interpolation is applied to determine the depth at which the density differs from the surface density by 0.125 kg/m^3. Values of the MLD of less than 10 m were set to 10 m while values higher than 400 m were excluded from the comparison with modeled MLD.

Figure 1 shows the monthly distribution of the measured MLD. January is depicted in the upper left panel, April in the upper right, etc., until December in the lower right corner. Contour lines are at every 100 m with additional contours at 25 m and 50 m. The MLD in the Gulf of Guinea is always shallower than 25 m. Values increase poleward to over 400 m at the northern wall. Main features are a shallow mixed layer in the equatorial band and a strong seasonal cycle in mid latitudes.

Figure 1. Monthly mean depth of mixed layer derived from climatological data of Levitus. The depth is defined by a density difference of $\Delta\sigma_t = 0.125$ kg m^{-3}. Upper left: January to lower right: December. Contours at 25, 50, 100, 200, 300, 400 m depth.

OPTIMIZATION METHOD

Parameters are found objectively by minimizing the rms misfit between modeled and measured MLD. The method involves an iterative technique wherein first the model sensitivity is calculated and second the optimal set of parameters is estimated, followed again by a sensitivity analysis, and so forth. Prior information on the set of parameters is taken into account by using the error covariance matrix during the optimization (Tarantola and Valette, 1982).

After choosing a first guess for the parameters the full coupled sea ice-mixed layer-isopycnic model OPYC is integrated. After five years of model time the mixed layer has reached an almost cyclo-stationary state. Monthly averages of the MLD of the last year, denoted as h_o, are stored for future computations.

To find the minimum of the data misfit the model is linearized around its current set of parameters. For this purpose all parameters under consideration are perturbed individually. For each perturbation OPYC is integrated for five years. This integration period seemed to be necessary for the MLD to come to a cyclo-stationary state after major parameter changes such as the introduction of the $\tau_x\Omega_y$ term. To ensure that differences in the MLD result from parameter changes and are not due to an undetected trend in the model, the same initial conditions and integration time as in the control experiment are used in the perturbation runs. The differences h_i between the modeled MLD of the last year and h_o are stored again.

Our linear model for the MLD then consists of the reference solution plus a linear combination of the perturbations

$$h_{mo}(X) = h_0 + \sum_i^n x_i h_i. \tag{9}$$

The vector X consists of n components x_i. They describe which fractions of the parameter changes applied to calculate h_i are used to compute the linear approximation h_{mo} of the MLD. For the linear model the data misfit J_{dat} can be written as

$$\begin{aligned} J_{dat} &= (h_{meas} - h_{mo})^T W (h_{meas} - h_{mo}) \\ &= \left(h_{meas} - h_0 - \sum_i^n x_i h_i \right)^T W \left(h_{meas} - h_0 - \sum_i^n x_i h_i \right) \end{aligned} \tag{10}$$

with a diagonal weighting matrix W defined as

$$W = \left(\frac{\cos(\varphi)}{10m} \right)^2 I. \tag{11}$$

The weights are proportional to the area represented by the measurement. They are normalized with a uniform rms of 10 m. The errors are assumed to be uncorrelated. Of course the weighting can be changed to represent the error of the individual estimates of the MLD. For instance a weighting proportional to the MLD itself was tried out as an alternative to (11). The change in the optimal parameters was, however, small. The sensitivity of the results to the choice of W seems to be low. Of course the absolute values in W are important only in comparison to the standard deviations s_i of the parameters.

The s_i are used to describe our *a priori* knowledge about the different parameters. This information is built up during many previous iterations and model reconfigurations. As the result of early iterations some of the parameters were discarded (e.g., by putting them to

zero) or fixed to specific values. Values for the remaining parameters are known better and better during the iteration process. Furthermore the sensitivity of the MLD to changes in the parameters is known from previous experiments. This information is used for the first guess and the variations of the parameters.

The total function to be minimized consists of the data misfit and an additional regularization term, which penalizes the deviation of the solution from its first guess,

$$J_{tot} = J_{dat} + X^T S X \tag{12}$$

where $\mathbf{S} = \text{diag}\ (\sigma_i^{-2})$ is the inverse of the *a priori* covariance matrix of X. The minimum of J_{tot} can easily be found by setting the partial derivatives to zero.

$$\frac{\partial J_{tot}}{\partial X} = -2\left(h_{meas} - h_0 - HX\right)^T WH + 2X^T S = 0 \tag{13}$$

where the matrix $\mathbf{H}$ consists of the MLD differences h_i. Solving (13) for X yields

$$AX = Y \tag{14}$$

with

$$A = H^T WH + S \tag{15}$$

and

$$Y = \left(h_{meas} - h_0\right)^T WH. \tag{16}$$

From the retrieved X we can directly calculate the optimal set of parameters.
For the estimation of the *a posteriori* error covariance matrix $\mathbf{E}$ of X we apply the singular value decomposition of $\mathbf{A}$. The advantage of this approach is that we can easily use alternative truncated solutions with their resolution and error covariance analysis (Wunsch, 1989).

$$A = U \Lambda V^T \tag{17}$$

where $\mathbf{U}$ and $\mathbf{V}$ consist of the eigenvectors of $\mathbf{A}$. Eigenvalues λ_i are stored in descending order in the diagonal matrix $\mathbf{L}$. $\mathbf{E}$ can now be calculated as

$$\begin{aligned}\mathbf{E} &= \left\langle (\hat{x}-x),(\hat{x}-x)^T \right\rangle \\ &= \sum_i^n \sum_j^n \frac{u_i^T S^{-1} u_j}{\lambda_i \lambda_j} v_i v_j^T \\ &= \sum_i^n \sum_j^n \frac{\sigma^2 \delta_{ij}}{\lambda_i \lambda_j} v_i v_j^T \\ &= \sum_i^n \frac{\sigma^2}{\lambda_i^2} v_i v_j^T\end{aligned} \tag{18}$$

if

$$S^{-1} = \sigma^2 I \tag{19}$$

The only problem that remains is to find suitable perturbations of the parameters, which turns out to be quite an art. A lot of intuition and experience from previous iterations is involved. The difficulties in deriving a "reasonable" set will become clear in the following.

First, it is better to interpolate than to extrapolate: When we calculate the local gradient of the MLD only small perturbations are used and the computed h_i will be small, in general. In order to produce MLD differences of appreciable size the corresponding x_i must be large, i.e., >>1. With these large coefficients the linear model h_{mo} extrapolates and the MLD will be quite different from the nonlinear OPYC using the optimized parameters. In some areas the extrapolated h_{mo} is considerably deeper than in its neighborhood or, on the contrary, may even become negative. The reason for this unrealistic behavior lies in the strong nonlinearities of the mixed layer dynamics. For every gridpoint there is a time in the seasonal cycle when the entrainment period terminates and detrainment occurs with a corresponding rapid change in the MLD. This decrease is often on the order of 100 m. A small perturbation in the model parameters will change the MLD both in the entrainment and the detrainment phase only slightly. However, it will shift the onset of the detrainment by a few days. For these few days we compute large differences in the MLD which multiplied by $x_i >> 1$ produces unrealistic results. Reducing the perturbations only intensifies the spiky appearance of the h_i and makes the response more local in space and time. As we will see below the error between modeled and measured MLD consists partly in a bias and partly in a phase shift. Such a phase shift cannot be modeled successfully with spiky h_i. As a consequence we impose a constraint to ensure that the modeled MLD will be an interpolation between meaningful solutions and no extrapolation:

$$0 \le x_i \le 1, \quad i = 1..n \tag{20}$$

The changes in the parameters applied to compute h_i must be chosen accordingly. Ideally the x_i should be approximately 0.5 at the solution to ensure a good compromise between gradient calculation and interpolation. Constraint (20) implies that for positive and negative parameter disturbances separate model integrations have to be performed.

Only a limited number of perturbation experiments were done because every run requires several hours of CPU time on a Cray computer for the integration of OPYC. Therefore we restricted the number of perturbations to the minimum. Only when it turned out (which it frequently did) that a perturbation was of the wrong sign or was made too small was a new model integration performed and the set of h_i augmented.

Because of the constraint (20) the solution of (13) is slightly more complicated than described previously. If the optimal x_i turns out to be zero we need another perturbation run for the corresponding parameter with a changed sign; $x_i = 1$ on the other hand makes a larger perturbation necessary. Rather than overwriting the corresponding h_i we set the corresponding x_i to zero and augment the set of variables. The frequent changes in the set of h_i make it necessary to retain all information until the final solution is found. Of course optimal parameter changes cannot be positive and negative at the same time. In this case the smaller change is discarded. The optimal x_i then consist of a number of zeros and values smaller than one where only nonzero values are used for the solution. Once the final set of variations x_i and h_i have been found we can again apply equations (13) and following to compute the solution and its error covariance matrix.

It is still possible to find gridpoints where $h(X)$ behaves unreasonably. For instance, in the control experiment an area might be marginally unstable and convection produces a deep MLD. In most perturbation runs convection does not occur and we have a situation where locally many h_i are large (and negative). Their weighted sum may produce an $h < 0$, that is, a negative thickness. A similar argument can be given for extremely deep values for the linear model. To safeguard against such a behavior we could add another constraint

$$\sum_{i=1}^{n} x_i \leq 1 \tag{21}$$

The disadvantage of (21) is that the solution now may depend on the number of variable n. Also we expect reasonable values of the x_i to be around 0.5. These disadvantages are avoided by requiring that h_{mo} lies in the same depth interval as the measurements, i.e.

$$10\ m \leq h_{mo} \leq 400\ m \tag{22}$$

Values which violate (22) are excluded from the calculation. Thus the number of 2° by 2° boxes involved in the optimization may vary during the iteration.

RESULTS

a) Early results

It was soon found that some of the parameters retrieved were close to their theoretical or their values measured in laboratories. Therefore γ was set to 0.42 and Ri_{crit} to 0.25. Both parameters were considered fixed subsequently. Another early result was the latitude dependence of the damping terms a and b. Attempts to model them independent of φ failed. A scaling depth depending on u_*/f, i.e., a scaling proportional to the Ekman depth, was clearly superior. Accordingly, damping independent of φ was no longer pursued. Value for the damping parameters d and d' were determined to be very small and we set $d=d'=0$. In the same way an independent efficiency parameter e for buoyancy was found to be unnecessary and e was fixed at unity.

b) Reference solution

The modeled MLD is depicted in Figure 2. It shows the same characteristics as the measured MLD (Fig. 1). The shallow equatorial MLD with little seasonal variation can be clearly seen. Farther north the annual cycle is the dominant signal with deepest values in March and values below 25 m during summer. In the South Atlantic, values below 25 m occur during Austral summer, i.e. December to February. High values for the MLD are found north of 50° N during winter and spring. These are also the areas of highest error in the MLD where the model is much too shallow compared to observations (Figure 1). Farther south the errors are smaller with the exception of a phase error in the retreat of the MLD at 20° N during spring warming.

Modeling of the mixed layer temperature (Figure 3) is relatively successful. In comparison with measured sea surface temperature, we find errors below 1 K for high temperatures. Errors increase to the north where they reach -4 K (model too cold) at 60°N. Both model deficiencies have been reduced noticeably in the meantime. The major improvements were due to the removal of the northern wall in favor of modeling the Arctic Ocean together with the Atlantic (Oberhuber, 1993b).

Details of the annual cycle of the MLD are difficult to perceive in Figures 1 and 2. Isolines are gappy because of undefined values. To give a clearer picture, a number of Hovmöller diagrams are shown below that depict conditions along 30°W as a function of month. Diagrams of measured and modeled MLD are given in Figures 4 and 5, respectively. We notice measurements of a deep mixed layer in winter increasing to the north. Maximum values are found in April when the depth of 200 m extends south to 40°N and the 100 m isoline almost reaches 20°N. During warming in spring and summer the depth is reduced gradually until minimal values are found in July and August. The seasonal cycle of the 100 m isobath is modeled fairly well. A deeper mixed layer is underestimated and a shallower

Figure 2. Monthly mean of modeled MLD. Upper left: January to lower right: December. Contours as in Figure 1. The MLD is shallow at the equator and becomes deeper in the region of the trade winds. Farther north the annual cycle is prominent with very deep MLD during winter.

MLD is overestimated. In most of the northern hemisphere the seasonal cycle is underestimated (Figure 6) while in the southern hemisphere little variation is observed.

For completeness the Hovmöller diagram of the mixed layer temperature along 30°W is given in Figure 7. In the northern hemisphere temperatures are lowest in March and warmest in August. The seasonal cycle is most pronounced around 30°N. Differences between modeled mixed layer temperatures and sea surface temperature measurements are given in Figure 8. Throughout the year the error is always below 1 K in the region south of 40°N. The annual cycle is mostly visible in the error around 60°N, i.e., close to Greenland. During winter and spring the model is too cold by up to 4 K and the error is smallest (1 K) during November.

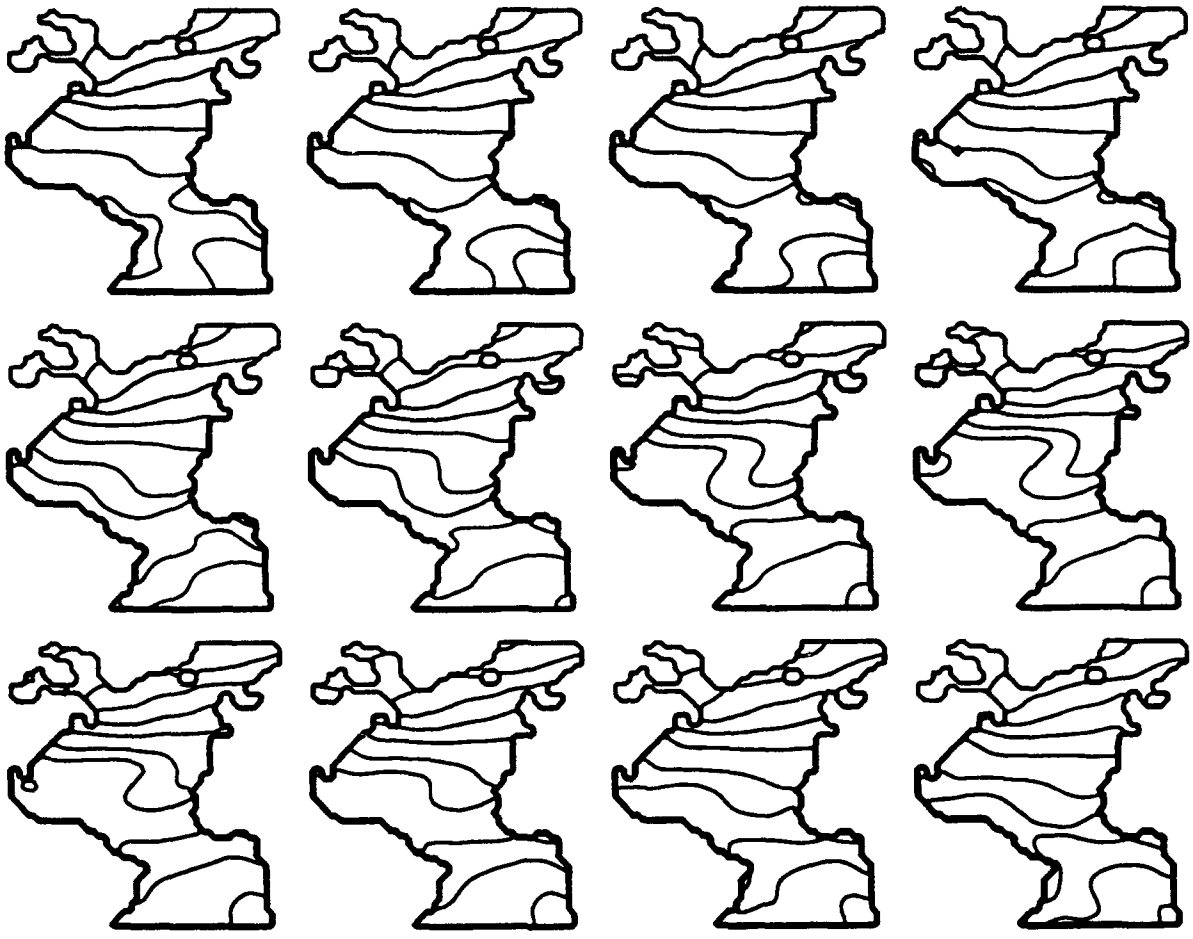


Figure 3. Monthly mean of modeled mixed layer temperature. Upper left: January to lower right: December. Contour interval 5° from 0° to 25°, additional contour at 27.5°. The general temperature distribution is close to observations except in the Gulf Stream region and in the vicinity of the northern boundary.

c) Perturbation experiments

After many iterations, five parameters remained to be determined by optimization. They were m_o, h_B, c, κ, and μ. Their *a priori* values and variances were chosen according to our knowledge gained so far (Table 1). Parameter disturbances were taken as twice the respective rms, which implies expected rms values of all non dimensional x_i of 0.5 to fit our requirements for interpolation. The corresponding σ_i^{-2} of matrix S are then 4.0. We will now discuss the results of the sensitivity experiments, i.e., the fields h_i. Again Hovmöller diagrams along 30°W are chosen to show both the annual cycle and the latitudinal dependence of the MLD differences.

Table 1. Values and standard deviations of estimated parameters

Parameter	*a priori*		*a posteriori*	
	mean	rms	mean	rms
κ	0.4	0.083	0.396	0.027
m_0	1.2	0.2	1.060	0.129
$h_{B\,[m]}$	10	2.5	8.077	1.640
μ	5	2.2	3.541	1.395

The result of a change in m_o from 1.2 to 0.8 is depicted in Figure 9. The decrease in the input of wind-induced TKE at the sea surface leads to a corresponding decrease in the MLD over the whole area. Values range from 5 m at the equator to 50 m at 50° N. The sensitivity is highest during the detrainment period in both hemispheres.

Changing the penetration depth h_B for the solar radiation from 10 m to 5 m results in a general reduction of the MLD too (Figure 10). Solar heating is concentrated more toward the surface, the buoyancy input is more negative, and the MLD reduced (see equation (1)). As with m_o, the highest sensitivity is during the retreat phase. But here we find maxima at 25°N and 25°S. The large positive change during summer occurs at the coast of Greenland. It must be attributed to a combined effect of advection and convection. Note that the MLD is undefined prior to June.

Considering the $\tau_x\Omega_y$ term (Figure 11) we also find noticeable changes concentrated in May. According to the latitudinal distribution of the windstress τ_x we find a decrease in MLD in the area of the westerlies north of about 30°N. Closer to the equator, easterlies prevail and the MLD becomes deeper.

The sensitivity of the decay functions a and b appears to be quite different. The effect of changing κ from 0.4 to 0.33 is small and concentrated mainly in the northern area (Figure 12). On the other hand changing μ from 5 to 2 results in a large decrease of the MLD outside the equatorial band (Figure 13). Differences are on the order of 20 m during the time when the MLD is deepest. Of course a reduction in μ will diminish the MLD only in regions with a positive buoyancy flux B, i.e. cooling or evaporation. Outside these areas there will be no change. The local deepening found at the coast of Greenland during summer is similar to Figure 10.

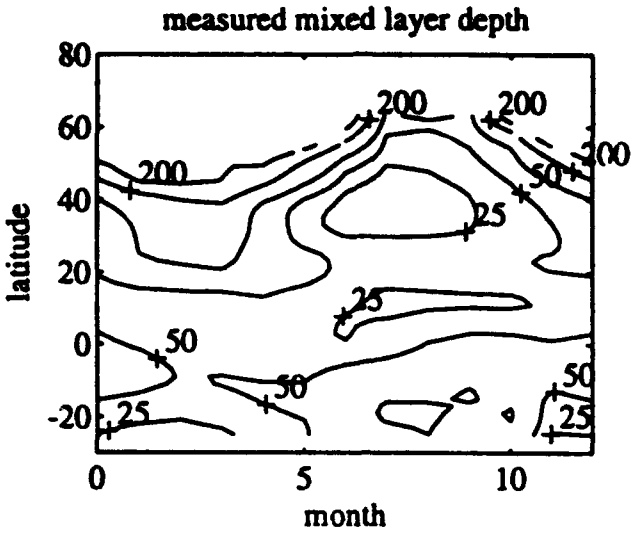


Figure 4. Hovmöller diagram of measured mixed layer depth at 30°W as a function of latitude and time. Contour intervals as in Fig. 1. There is a strong seasonal cycle north of 40°N.

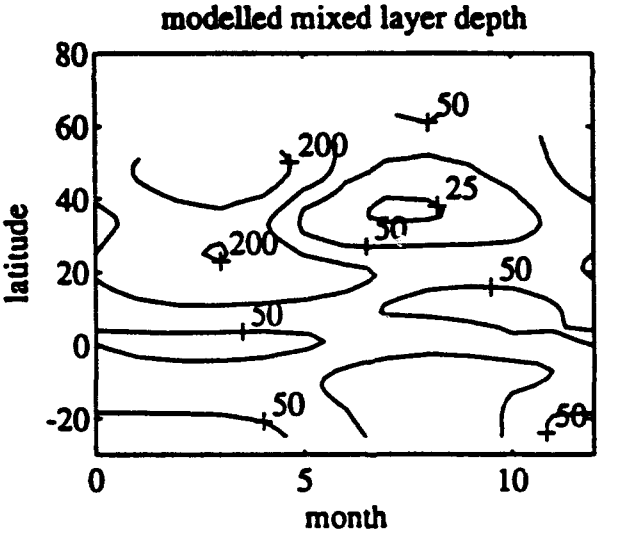


Figure 5. Hovmöller diagram of modeled mixed layer depth at 30°W. Contour intervals as in Fig. 1. The seasonal cycle is less pronounced compared to measurements (Fig. 4).

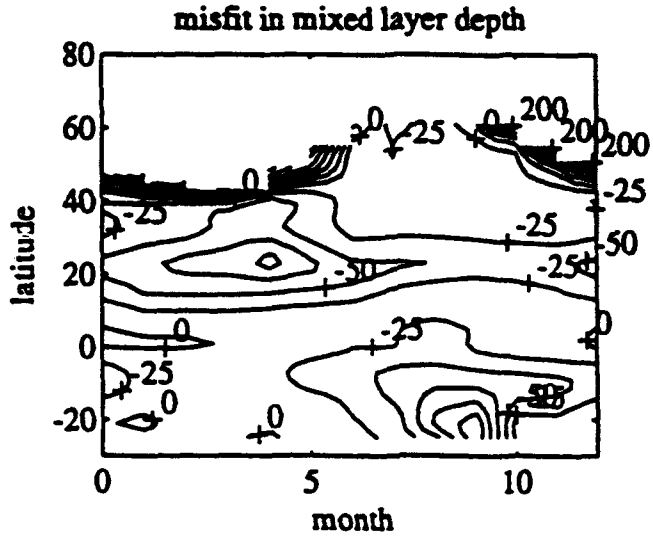


Figure 6. Hovmöller diagram of error in mixed layer depth at 30°W. C.i. = 25 m. During winter the model is too shallow in the north and too deep in the south. Detrainment in the spring is delayed.

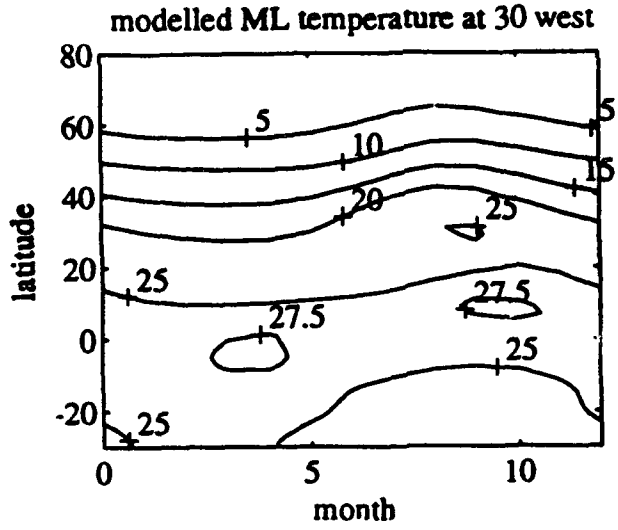


Figure 7. Hovmöller diagram of the modeled mixed layer temperature at 30°W. Temperature rises to 28° C at the equator.

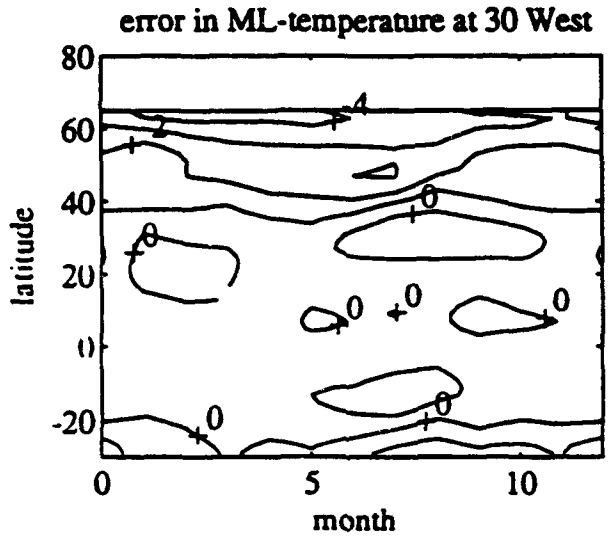


Figure 8. Hovmöller diagram of error in mixed layer temperature at 30°W. C.i. = 1 K. Temperature differences are small except north of 40°N where they increase. The model is too cold by as much as 4 K near Greenland.

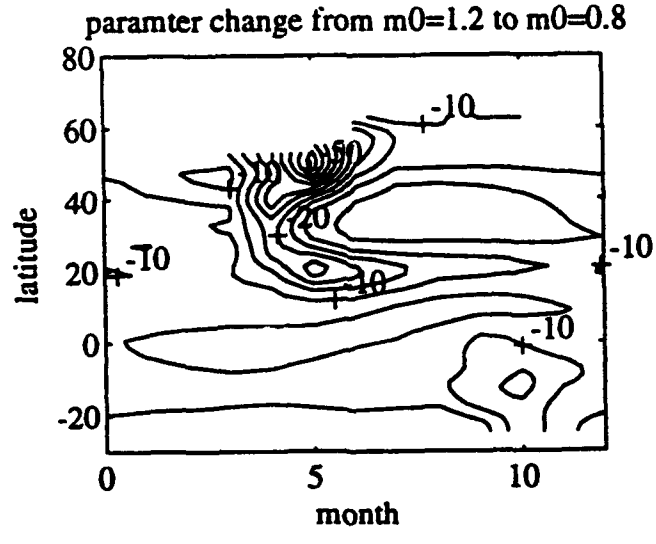


Figure 9. Hovmöller diagram of difference in mixed layer depth at 30°W. Parameter m_O was changed from 1.2 to 0.8. C.i. = 5 m. Because of the decrease in the wind input of TKE the MLD becomes shallower by up to 50 m.

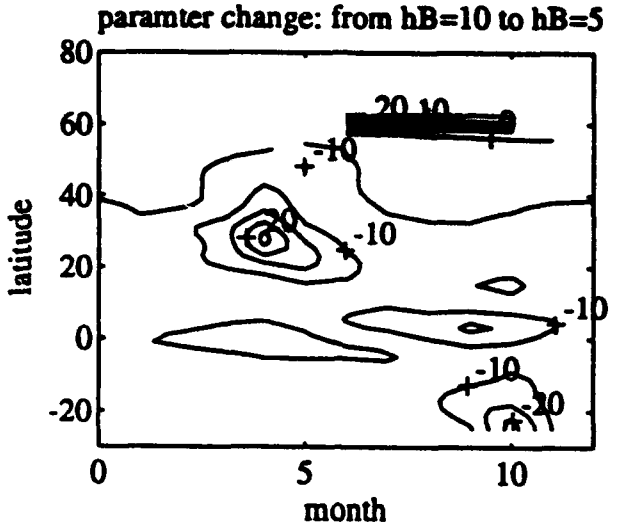


Figure 10. Hovmöller diagram of difference in mixed layer depth at 30°W. Parameter h_B was changed from 10 to 5 m. C.i. = 5 m. Less penetration of solar heating concentrates the input of negative buoyancy more toward the surface. This stabilizes the mixed layer and decreases its thickness by 10 m on the average.

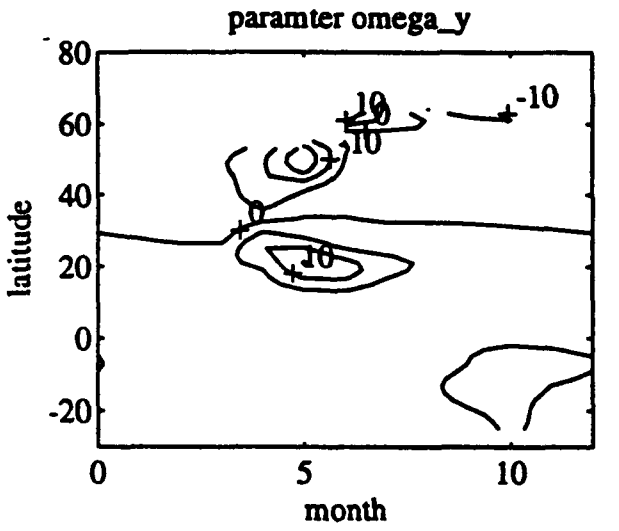


Figure 11. Hovmöller diagram of difference in mixed layer depth at 30°W. The term involving $\tau_x \Omega_y$ is included in the calculation. C.i. = 5 m. In the region of westerly winds we observe a retreat in the MLD, whereas easterlies lead to a deepening.

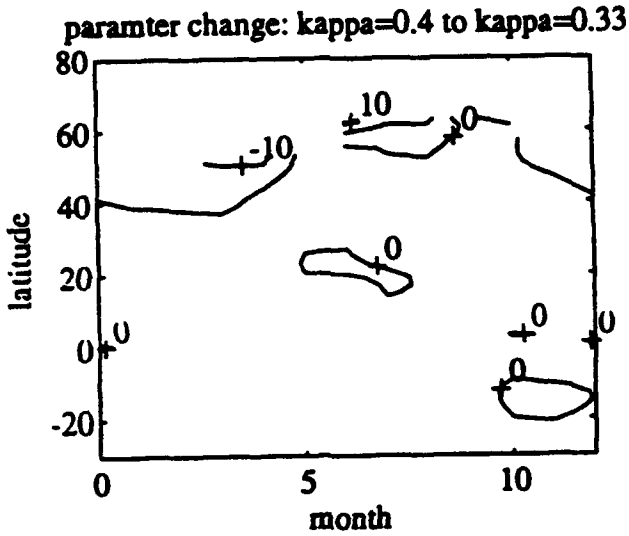


Figure 12. Hovmöller diagram of difference in mixed layer depth at 30°W. The depth scale κ was changed from 0.4 to 0.33. C.i. = 5 m. Less TKE is available at the bottom of the MLD resulting in a shallower mixed layer.

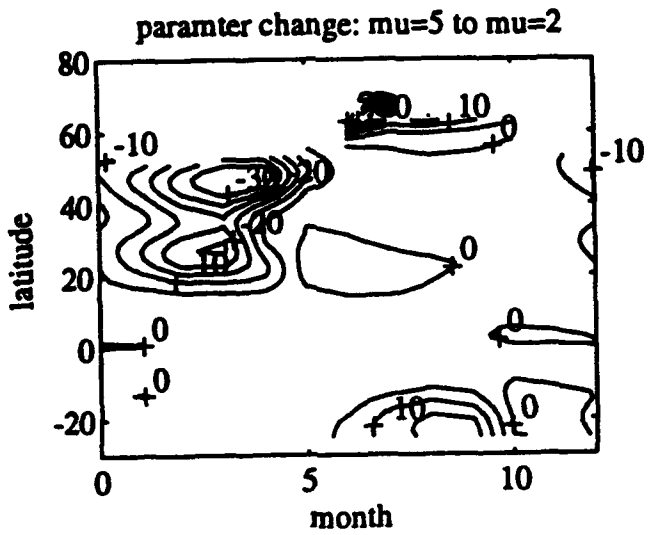


Figure 13. Hovmöller diagram of difference in mixed layer depth at 30°W. The depth scale μ was changed from 5 to 2. C.i. = 5 m. Much stronger damping of the turbulence produced by positive buoyancy results in a retreat of the MLD. The decrease is restricted to areas of positive buoyancy flux *B*.

d) Inverse solution

The h_i calculated above are now used to solve (13) for the optimal vector X. It should be mentioned first that most of the improvement in the data misfit was made in previous iterations. The remaining error was predominantly systematic and could only slightly be reduced. Differences between the optimal solution and the reference solution are small. They amount to a reduction of the MLD of less than 20 m for most cases (Figure 14). Only in areas with convection changes are high and local.

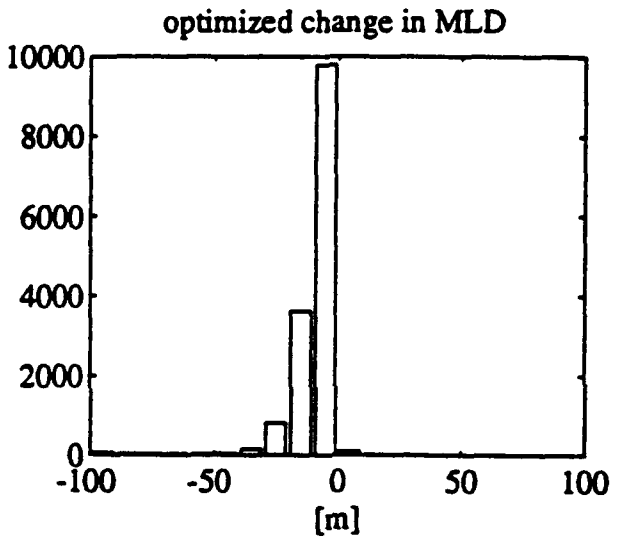


Figure 14. Histogram of the change of the modeled MLD resulting from the optimization of the parameters. Most changes reduce the MLD by up to 10 m. Positive changes (deepening) are rare.

One of the most important findings is that parameter c, which is the coefficient of the $\tau_x \Omega_y$ term, finally turns out to be zero. The corresponding h_i (Figure 11) leads to a deepening in the tropical regions where the model is already too deep. Farther north the reduction in MLD is of benefit. However, in this area other parameters, such as μ or m_o are more effective and are preferred. The modeled MLD would only be improved with a negative c, which violates the theory of Garwood et al. (1985a,b). Consequently this theory must be rejected in the context of our modeling effort.

It is worth mentioning that the data misfit (10) is larger for each individual perturbation than for the control experiment. One could be tempted to believe that no improvement is possible. However, a combination of small x_i leads to a reduction in J_{dat}. This is evident from the gradient $\partial J_{tot} / \partial x_i$ calculated at the reference solution. It is negative for c and positive for the other four parameters. Of course the gradient must be zero at the optimized solution.

The remaining parameters are found by solving

$$\begin{pmatrix} 4.273 & 0.264 & 0.261 & 0.287 \\ 0.264 & 4.968 & 0.627 & 0.419 \\ 0.261 & 0.627 & 4.772 & 0.382 \\ 0.287 & 0.419 & 0.382 & 4.728 \end{pmatrix} X = \begin{pmatrix} 0.467 \\ 2.126 \\ 2.189 \\ 1.604 \end{pmatrix} \tag{23}$$

The x_i are positive and less than one as required by (20). The regularization term in (12) has a strong influence on the solution. Although J_{dat} accounts for 97 % of J_{tot} most of this error is systematic and cannot be improved much in the final iteration. Most of the progress has already been made previously. The diagonal elements of matrix **A** are dominated by σ_i^{-2}. They imply that after many iterations we have reached a state where the solution now depends more on our *a priori* knowledge and less on data.

Non-dimensional x_i are converted to the corresponding parameter difference and the optimal set of parameters is calculated (Table 1). Values for m_o and κ appear to be reasonable. The closeness of κ to the Kármán constant of 0.4 is striking. However, we would never propose to determine the Kármán constant via assimilation of measured mixed layer thickness. A μ of 3.5 is reasonable too as it allows less damping of buoyancy induced turbulence compared to mechanically generated turbulence. Finally a penetration scale of solar heating h_B of 8 m seems to be too small. In clear sea water h_B is on the order of 20 m.327

Error analysis

The optimal MLD h_{mo} is similar to the first guess h_0. A histogram of the remaining error is shown in Figure 15. The result is still biased with a mean of the error of -7.8 m. Most of the MLD is overestimated by some tens of meters. A noticeable number of underestimations by 100 m and more are also found. The rms error after optimization amounts to 48.7 m. If we now construct a linear model T_{mo} for the mixed layer temperature in analogy to h_{mo}, where the T_i are calculated from the temperature differences in the perturbation experiments and the x_i are taken from the optimization of the MLD, we find only small temperature changes in comparison with the reference solution. Figure 16 shows the histogram of the final temperature errors. The differences to the measured sea surface temperature are only slightly biased. The mean temperature error amounts to -0.49 K (model too cold) and the rms error is 1.35 K.

The seasonal cycle of the error is shown in Figure 17. The rms error (upper curve) and the bias (lower curve) are plotted as a function of time. Straight lines depict the annual mean of the rms (48.7 m) and of the bias (7.8 m). We notice a moderate seasonal cycle in the data misfit with the smallest values in May. The bias, on the other, hand shows no systematic time dependence. Errors of the mean are low in March when the MLD is deep

and in August to November when the MLD is shallow. In between the MLD is overestimated in the mean by as much as 20 m.

Next we study the latitudinal dependence. We find that the error (full line in Figure 18) is very small at the equator and increases poleward. The maximum of about 120 m rms is found at 60°N. However, here the number of points that enter the optimization (thin line in Figure 18) has already dropped considerably down from its maximum at 30°N. The total misfit J_{dat} is therefore only moderately influenced by the errors at 60°N and farther north.

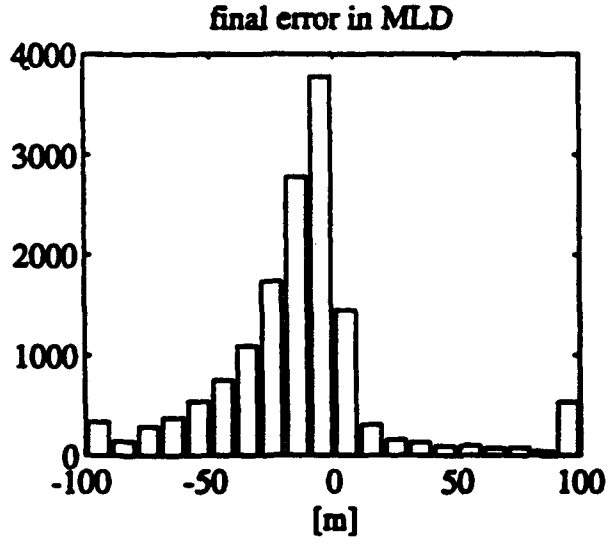


Figure 15: Histogram of the error of the modelled MLD. The distribution of the model error is strongly skewed. Most model values are too deep. However, there is a strong contribution by the values which are too shallow by 100 m and more.

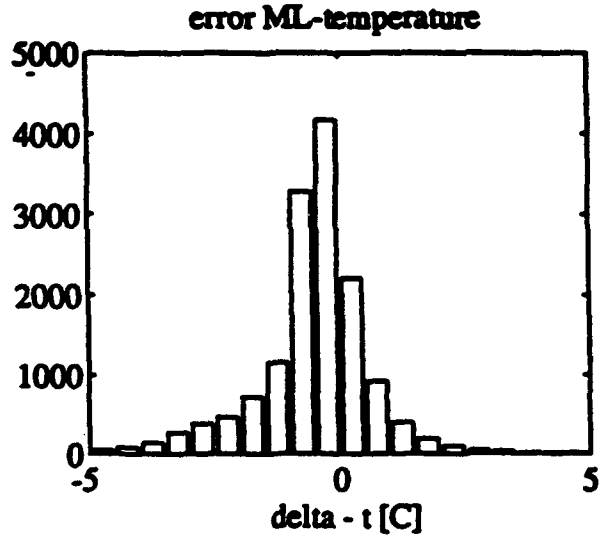


Figure 16. Histogram of the error of the modeled mixed layer temperature. The model is slightly too cold (0.14 K on average). Most errors are smaller than 1 K.

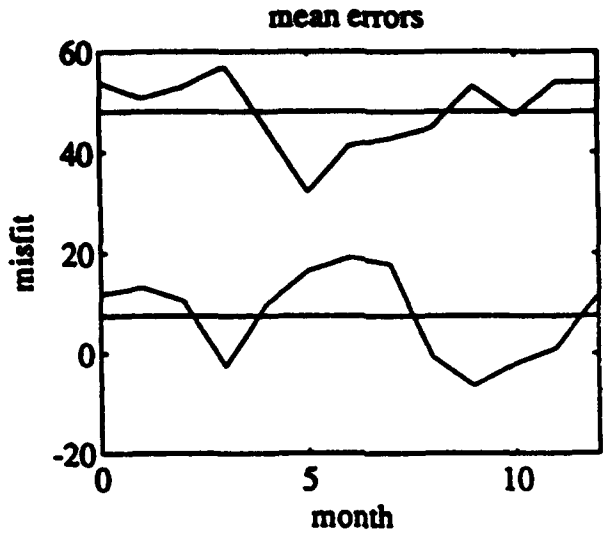


Figure 17. Seasonal cycle of the rms error in MLD and the mean of the misfit. The rms error is smallest in May, the error in the mean is highest in June. Straight lines depict the annual mean. On average the modeled MLD is too deep by 6.9 m.

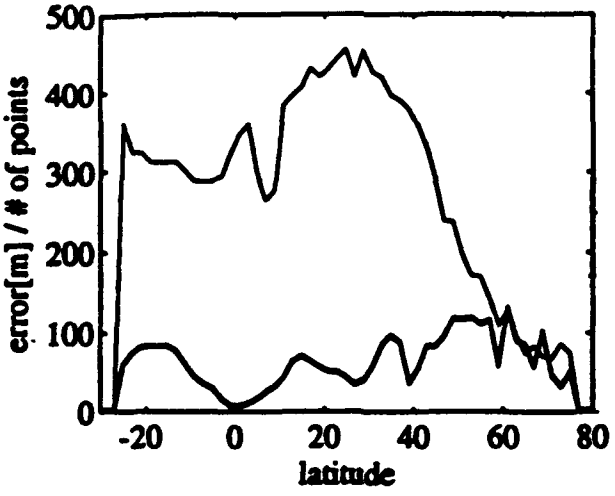


Figure 18. Rms error of the MLD in m (solid line) and number of 2° by 2° grid-cells used in the optimization (thin line) as a function of latitude. The error is very small at the equator and increases poleward with maximum values of 180 m at 60°N. The number of points drops sharply north of 30°N mainly as a result of the increase in undefined MLD.

Large errors in temperature occur in the same areas as large MLD errors, i.e. mainly north of 60°N. There is, however, no distinct connection between the large errors. We have recalculated the whole data assimilation retaining only MLD where the temperature error was below 1 K. Nevertheless, the optimized parameters were practically the same as before. The x_i changed by less than 10%, the total number of points was reduced from 14782 to 10043 and the rms error from 48.7 to 41.3 m.

The improvement in modeling of the MLD can be seen in the differences in the misfit before and after the data assimilation. Figure 19 shows the probability density function (pdf) of the remaining error versus the initial error. For practically all negative errors (model too deep) values are above the 45° line. The improvement amounts to changes between 5 and 15 m and is higher for large errors.

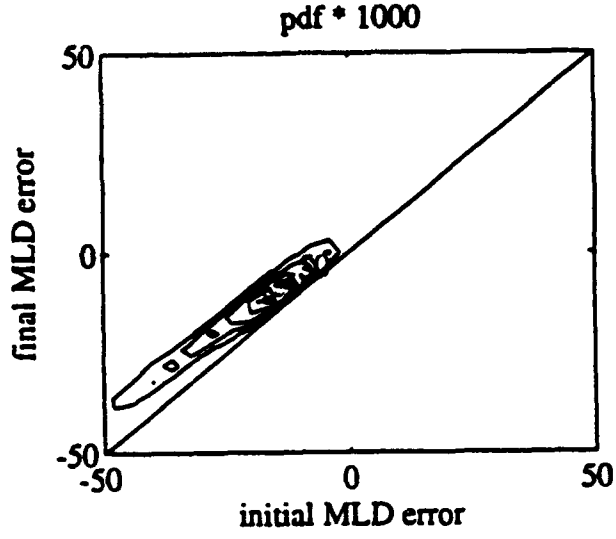


Figure 19. Frequency distribution of final error in MLD versus initial error. Most errors are negative (model too deep) with a maximum of probability at about -20 m. The final optimization improves the errors by 10 m and more.

Finally we calculate the error covariance matrix **E** according to (18). As the dimensions of the parameters are different and their variances (Table 1) have a different order of magnitude we have chosen to show the error correlation matrix (Table 2) instead of **E**. As can be deducted from the similarities in the h_i all cross-correlations are negative. That is, overestimation of one parameter is correlated with underestimation of the others. Most cross-correlations are, however, very small with the exception of the anti-correlation of the errors in h_B and m_o.

Table 2. Correlation matrix

Parameter	κ	m_0	h_B [m]	μ
κ	1.000	-0.046	-0.046	-0.056
m_o	-0.046	1.000	-0.131	-0.073
h_B [m]	-0.046	-0.131	1.000	-0.073
μ	-0.056	-0.073	-0.073	1.000

Conclusions

An inverse method has been applied to determine the free parameters of a non linear mixed layer model. The method is successful in reducing the misfit between modeled and measured mixed layer depth. Parameters and their error covariance matrix are determined by data assimilation. The major advantage is an objective test of different competing model formulations. A number of different damping mechanisms were examined and finally an exponential decay of turbulent kinetic energy that scales with the Ekman depth is selected. In the same way we tested the hypothesis of Garwood et al. (1985a,b), which introduces a source of turbulent kinetic energy for easterly and a sink for westerly winds. This theory was rejected because it did not fit the data.

An error analysis is a necessary step to determine systematic errors which cannot be removed by parameter optimization. On the contrary, one has to be careful that some of the parameters are not tuned to alleviate these errors. For instance, in our case the penetration depth of the solar heating was diminished to reduce a systematic overestimation of the mixed layer depth. For the same reason the efficiency of wind stirring was underestimated in comparison to more advanced model versions. We would like to point out that currently the following set of parameters is advised for OPYC:

$$
\begin{aligned}
&m_o = 1.2,\ \kappa = 0.4,\\
&\mu = 2,\ Ri_{crit} = 0.25,\\
&c = d = d' = 0,\\
&e = 1,\ \gamma = 0.42 \text{ and}\\
&h_B = 23 \text{ m}.
\end{aligned}
$$

Acknowledgments

Josef Oberhuber provided the numerical code of OPYC, assisted in the setting up of the model, and took part in many fruitful discussions about the interpretation of the results. Ralf Giering performed all of the numerous model integrations. Their help and their contributions to this work are gratefully appreciated.

References

Fiadero, M. E. and G. Veronis, 1984: Obtaining velocities from tracer distributions. *J. Phys. Oceanogr.*, **14**, 1734-1746.

Garwood, R. W., Jr., P. C. Gallacher and P. Müller, 1985a: Wind direction and equilibrium mixed layer depth: General theory. *J. Phys. Oceanogr.*, **15**, 1325-1331.

Garwood, R. W., Jr., P. Müller and P. C. Gallacher, 1985b: Wind direction and equilibrium mixed layer depth in the tropical Pacific Ocean. *J. Phys. Oceanogr.*, **15**, 1332-1338.

Hellermann, S. and M. Rosenstein,1983: Normal monthly wind stress over the world ocean with error estimates. *J. Phys. Oceanogr.*, **13**, 1093-1104.

Kraus, E. B. and J. S. Turner, 1967: A one-dimensional model of the seasonal thermocline. *Tellus*, **1**, 88-97.

Levitus, S., 1982: *Climatological Atlas of the World Ocean.* NOAA Prof. Paper No. 13, U. S. Govt. Printing Office, 173 pp., 17 fiche.

Martin, P. J., 1985: Simulations of the mixed layer at OWS November and Papa with several models. *J. Geophys. Res.*, **90**, 903-916.

Miller, A. J., J. M. Oberhuber, N. E. Graham and T. P. Barnett, 1992: Tropical Pacific response to observed winds in a layered general circulation model. *J. Geophys. Res.*, **97**, 7317-7340.

Niiler, P. P., 1975: Deepening of the wind mixed layer. *J. Mar. Res.*, **33**, 405-422.

Niiler, P. P. and E. B. Kraus, 1977: One-dimensional model of the seasonal thermocline. *The Sea*, Vol. **VI**, Wiley-Interscience, 97-115.

Oberhuber, J. M., 1988: An Atlas based on the 'COADS' data set: The budgets of heat, buoyancy and turbulent kinetic energy at the surface of the Global Ocean. Max-Planck-Institut für Meteorologie / Hamburg, Rep. 15, 199 pp.

Oberhuber, J. M., 1993a: Simulation of the Atlantic circulation with a coupled sea ice-mixed layer-isopycnal general circulation model. Part I: Model description. *J. Phys. Oceanogr.*, **23**, 808-829.

Oberhuber, J. M., 1993b: Simulation of the Atlantic circulation with a coupled sea ice-mixed layer-isopycnal general circulation model. Part II: Model experiment. *J. Phys. Oceanogr.*, **23**, 830-845.

Olbers, D. J., M. Wenzel and J. Willebrand, 1985: The inference of North Atlantic circulation patterns from climatological hydrographic data. *Geophys. Rev.*, **23**, 313-356.

Tarantola, A. and B. Valette, 1982: Generalized nonlinear problems solved using the least squares criterion, *Rev. Geophys.*, **20**, 219-232.

Woodruff, S. D., R. J. Slutz, R. L. Jenne and P. M. Steurer, 1987: A Comprehensive Ocean-Atmosphere Data Set. *Bull. Amer. Meteor. Soc.*, **68**, 1239-1250.

Wright, P., 1988: An Atlas based on the 'COADS' data set: Fields of mean wind, cloudiness and humidity at the surface of the Global Ocean. Max-Planck-Institut für Meteorologie / Hamburg, Rep. 14, 70pp.

Wunsch, C., 1985: Can a tracer field be inverted for velocity?, *J. Phys. Oceanogr.*, **11**, 1521-1531.

Wunsch , C., 1989: Tracer inverse problems. In: D. L. T. Anderson and J. Willebrand (eds.), *Oceanic Circulation Models: Combining Data and Dynamics.* Kluwer Academic Publishers, 1-77.

AN ADAPTIVE INVERSE METHOD FOR MODEL TUNING AND TESTING

Claude Frankignoul and Nathalie Scoffier
L.O.D.Y.C., Université Pierre et Marie Curie, Paris 75005, France

Mark A. Cane
Lamont-Doherty Geological Observatory of Columbia University, Palisades, NY 10964

ABSTRACT

To determine the value of the adjustable parameters of an ocean model that are required to optimally fit the observations, an adaptive inverse method is developed and applied to a sea surface temperature (SST) model of the tropical Atlantic. The best-fit calculation is performed by minimizing the misfit between observed and simulated data, which depends on the observational and the modelization errors. An adaptive procedure is designed where the model that is being tuned is also used to construct a sample estimate of the observational error covariance matrix. Assuming idealized modelization errors, the procedure is applied to the SST model of Blumenthal and Cane (1989), yielding improved estimates for several model and heat flux parameters. The tuned model provides a better simulation of the mean annual SST, but the model's ability to represent the seasonal and the interannual variability is not improved, and the model-observation discrepancies remain too large. The existence of larger model deficiencies than was originally assumed in the model errors is confirmed by a statistical test of the correctness of the assumptions in the inverse calculation.

1. INTRODUCTION

All oceanic models contain parameterizations of such physical processes as convection and mixing. Surface forcing also depends on poorly known parameters. Parameterizations are based on physical ideas, but typically yield forms that contain parameters whose values are not known precisely. A parameter is often model dependent (e.g., mixing is a function of grid spacing), hence parameter tuning may be in part model dependent. In view of their inherent imprecision, the uncertain parameters should be tuned against observed data. At the same time, models should be consistent with known physics to within the tolerances allowed by the approximations made.

Particularly in the tropics where observations are sparse, both forcing and verification data are imprecisely known. Hence, the accuracy to be expected in model simulations is limited, even if the physics are perfectly represented, and data uncertainties should be taken into account in parameter tuning. Frankignoul et al. (1989) have developed a multivariate model testing procedure that provides an objective measure of the fit between ocean

model simulations and observations, taking into account the data uncertainties. By using a trial and error approach, the method can be used for model tuning (Duchêne and Frankignoul, 1991; Braconnot and Frankignoul, 1993). However, this requires that the number of adjustable parameters is small.

A more efficient tuning approach is that of Blumenthal and Cane (1989), who used inverse modeling procedures to determine the parameter values required to optimally fit sea surface temperature (SST) in a simplified tropical SST model. A priori knowledge constraining the parameter range was included in the calculation, but only a highly idealized model was used for the data errors. The error model enters the measure of the misfit between observed and predicted data which is minimized in the best-fit calculation. Thus, the atmospheric forcing uncertainties need to be properly represented, as they introduce large uncertainties in the model response.

As the forcing uncertainties have large and poorly known correlation scales, the error estimates are best derived from direct simulations. We have thus developed an adaptive tuning procedure, where the model that is being tuned is also used to construct the observational error model for the best-fit calculation. The tuned model is then tested against observations, and if it agrees with the data to within expected errors, it will be judged adequate. Such an adaptive technique combines the model tuning of Blumenthal and Cane (1989) and the model testing of Frankignoul et al. (1989). Although the procedure is developed in the context of a simplified tropical sea surface temperature model, it is general as long as the parameter dependence is linear. The adaptive procedure requires little computation and programming, and is much simpler to implement than the adjoint method. However, since the effective degrees of freedom of the error estimates is limited by the length of the sample, the number of parameters that can be tuned is limited.

The emphasis here is on the adaptive inverse procedure, although it is introduced in the context of a tropical SST model. An in-depth discussion of the results is given in Scoffier et al. (1993).

2. MODELING SEA SURFACE TEMPERATURE VARIATIONS

a. Ocean model and surface heat flux

The ocean model is that of Blumenthal and Cane (1989, hereafter BC). The velocity is predicted with a linear, multimode equatorial beta-plane model with a surface mixed layer of constant depth h= 35 m, which adds a direct Ekman flow to the modal currents. The model has five vertical modes, which are characteristic of mean tropical Atlantic conditions and have gravity wave speed of 2.36, 1.38, 0.89, 0.69 and 0.53 m/s, respectively. The model basin extends from 30°N to 20°S and has a simplified geometry; its resolution is 1° in longitude and 0.5° in latitude and the time step is one week. The equations are solved in the longwave approximation, so that the model is only appropriate

away from the western boundary. In the following, we only consider the domain in Figure 1, which should not be affected by the model artificial boundaries.

The SST is uniform in the mixed layer and determined from the net balance of horizontal advection, upwelling, horizontal diffusion, and surface heat exchanges:

$$\partial_t T + u\partial_x T + v\partial_y T + \frac{\gamma w(T - T_d)}{h} = \kappa(\partial_{xx} + \partial_{yy})T + \frac{Q}{\rho C_p h} \tag{1}$$

where w is the vertical velocity at the mixed layer base in case of entrainment and zero otherwise, κ a horizontal diffusion coefficient and Q the surface heat flux into the mixed layer, and T_d the temperature below the mixed layer which is parametrized as a function of the thermocline depth. As in BC, the parameterization of T_d is done in two parts: First the temperature at the mixed layer base is fit to the depth of the 20°C isotherm in the equatorial zone using observations, then the 20°C isotherm depth is fit to the model prediction of the thermocline depth. The upwelling term is usually written as $w(T - T_e)$, where T_e is the temperature of the water entrained into the mixed layer, but the two forms are equivalent if

$$T_e = (1 - \gamma)T + \gamma T_d \tag{2}$$

where the "entrainment efficiency" γ is an adjustable parameter that should be less than one, because T_e is between T and T_d.

The surface heat flux parameterization is that of Seager et al. (1988, henceforth SZC), which was designed to avoid using either the (poorly measured) air-sea temperature differences found in the bulk formulae or the artificial feedback to a prescribed climatological air temperature often imposed in ocean simulations. This parameterization only makes use of wind speed v^a and fractional cloud cover C as measured variables:

$$Q = 0.94 Q_0 (1 - a_c c + a_\alpha \alpha) - \rho C_E L v^a a_{rh} q_s(T) - a_T (T - T_r). \tag{3}$$

The first term is the (usual) short wave radiation, where Q_0 is the clear sky solar flux reduced by the effects of a constant surface albedo and by the absorption and reflection of the atmosphere, which depends on C and solar angle α. The second term represents the latent heat flux, which is computed from the standard bulk formula using a fixed percentage a_{rh} of the saturation humidity $q_s(T)$ as evaporation potential; this assumes that the moisture content of the air has equilibrated with the ocean temperature, which is a reasonable assumption sufficiently far from the coasts. To compensate for the loss of variability in using monthly winds, v^a is not allowed to fall below 4 m/s. The smaller sensible heat flux and back radiation are simply modeled together in the last term as being proportional to T minus a constant reference temperature T_r.

In the SST equation and the heat flux formulation, there are a number of parameters that are not precisely known, but were assigned a "reasonable" value by SZC. Here we assume that seven parameters are adjustable within reasonable ranges: the entrainment efficiency γ, the horizontal diffusion κ, and the heat flux parameters $a_c, a_\alpha, a_{rh}, a_T$, and $a_T T_r$ in (3), which we represent below by the seven-dimensional vector $\mathbf{a}$. The a priori values of the tunable parameters, denoted by $\mathbf{a}_p$, are those of SZC, namely $\gamma = 0.5$, $\kappa = 2 \times 10^8$ $m^2 s^{-1}$, $a_c = 0.62$, $a_\alpha = 0.0019$, $a_{rh} = 0.3$, $a_T = 1.5$ W m^{-2} K^{-1} and $T_r = 273.15$ K. The drag coefficient for the wind stress is not allowed to vary because its uncertainty is simulated explicitly.

b. Simulation of the tropical Atlantic SST climatology

After spin-up, the model is forced by a monthly wind stress derived from ship reports for the period 1964-1986 and described in Frankignoul et al. (1989, henceforth FDC). To simulate the drag coefficient uncertainty, we follow the Monte Carlo approach of Braconnot and Frankignoul (1993) and use five different, equally plausible drag coefficients in the bulk formula. They are calculated by prescribing a relative humidity of 80% and using either a constant air-sea temperature difference of −1°C (for the parameterization of Cardone (unpublished manuscript)), or a climatological monthly air-sea temperature difference derived from the COADS data (for the parameterizations of Liu et al. (1979), Large and Pond (1981), Isemer and Hasse (1987), and Smith (1988)). To avoid smoothing, the monthly mean wind stresses were corrected to insure that linear interpolation on the model time step would not alter the original means. Cloudiness data are of poorer quality, so that cloud cover is prescribed from the monthly climatology of Esbensen and Kushnir (1981), with an added normal noise of 0.1 standard deviation to crudely simulate its short space-time scale variability.

Ignoring the first year to eliminate the effects of the unknown initial conditions, we have five 22-year simulations of the SST whose dispersion is representative of both the interannual variability and the drag coefficient uncertainty. The mean cycle of simulated SST is warmer than the observations, as illustrated in Figure 1 for January, April, July, and October by a comparison with the mean SST over the same period calculated from the data of Servain et al. (1985).

The differences between the SST predictions and the observations are due to (a) errors in the atmospheric forcing (wind stress, cloud) and the SST observations, (b) model shortcomings due to over-simplification of the physics, or (c) poor choice of the model parameters. To assess the validity of the SST model, we must take (a) into account and minimize (c) by an optimal tuning. Remaining discrepancies should then point to the model deficiencies (b).

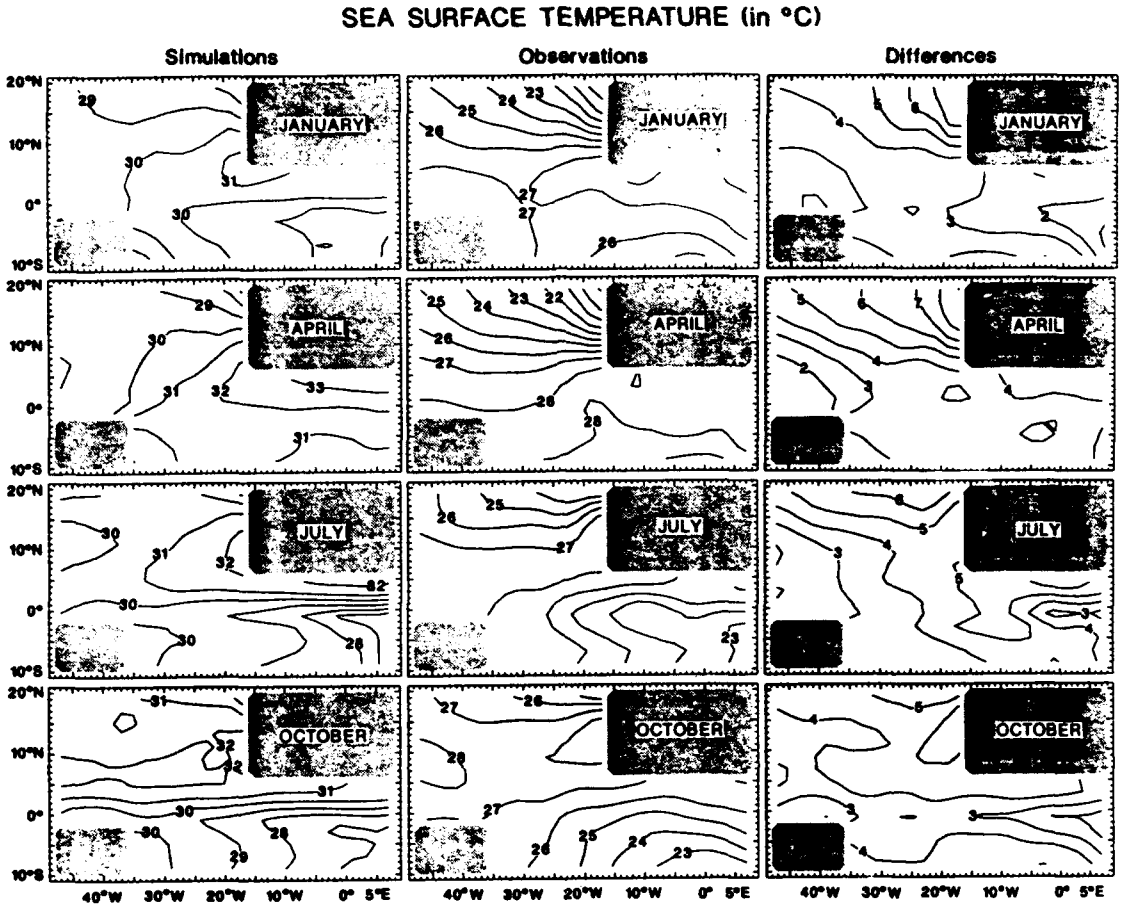


Figure 1. (left) Mean SST in °C during January, April, July and October for the period 1965-1986 as predicted using the a priori values of the model parameters. (center) Corresponding SST as derived from the observations by Servain et al. (1985). (right) Differences between simulations and observations.

Root-mean-square (rms) SST differences between model and observations on the 2° × 2° grid of the latter are given in Table 1 (left column), where we distinguish between annual mean, mean seasonal variations around the annual mean (hereafter the mean seasonal variability), and SST anomalies. The model-observation differences are large, particularly for the long term mean which is strongly affected by a 3.9°C mean bias.

A more quantitative estimation of the model performances taking into account some of the uncertainties in the oceanic observations and the atmospheric forcing, as well as their space-time correlations, has been made for the mean seasonal cycle obtained with Cardone's drag coefficient. Following the multivariate approach of FDC, we calculate the misfit

$$T^2 = (\overline{\mathbf{T}} - \overline{\mathbf{T}}_0)' \mathbf{D}^{-1} (\overline{\mathbf{T}} - \overline{\mathbf{T}}_0), \tag{4}$$

where $\overline{\mathbf{T}}$ and $\overline{\mathbf{T}}_0$ describe the mean seasonal cycle of modeled and observed SST, respectively, the vector space including all grid points (on the observational grid) and the twelve months. The overbar denotes the 22-year mean, the prime denotes the vector transpose, and **D** is the error covariance matrix of $(\overline{\mathbf{T}} - \overline{\mathbf{T}}_0)$. In the calculation reported here, **D** is estimated from the five 22-year samples, assuming for simplicity that each year is statistically independent. It takes into account the uncertainties in the mean seasonal variations that are due to interannual variability, non-systematic observational errors of SST, wind, and cloud cover. Not represented in **D** are systematic observational errors (e.g., incorrect Beaufort scale), drag coefficient uncertainty, lack of high frequency variability, and limited resolution of the wind stress curl. As the dimension of the SST field is much larger than the degrees of freedom of **D**, the misfit (4) is calculated in a truncated space which is sufficiently small to calculate **D** reliably while representing the main space-time patterns of $(\overline{\mathbf{T}} - \overline{\mathbf{T}}_0)$.

Table 1: Rms difference in °C between observed and modeled SST before and after tuning in the 20°N-10°S region. The correlation between observed and simulated monthly anomalies during 1965-86 is given in italic.

$(SST_{mod}\text{-}SST_{obs})$	before tuning	after tuning
annual mean	4.0	1.9
seasonal variability	0.7	0.8
anomaly correlation	*0.13*	*0.10*

If the SST fields are multinormal, the null hypothesis that the model response to the true forcing is equal to the true SST can be tested because the test statistic (4) is then Hotelling's T^2 statistic. As shown in Table 2, T^2 is much larger than the critical value at the 5% level, especially for the yearly mean difference. Although only part of the observational errors have been considered in the test, the data uncertainties are clearly insufficient to explain all the model-observation discrepancies, which must be mainly attributed to model shortcomings and poor parameter tuning.

Table 2: Misfit between model and observations in the 20°N-10°S region, before and after tuning. The critical values for rejecting the null hypothesis of no modelization error are given for the 5% level (right).

Misfit	before tuning	after tuning	critical value
annual mean	906	277	4
seasonal variability	2012	1694	73

3. AN ADAPTIVE PROCEDURE FOR MODEL TUNING

a. Linear model corrections

To see how the tunable parameters enter the SST calculation, it is convenient to write equation (1) in matrix form

$$\mathbf{L(T)} + \mathbf{M(T)}\mathbf{a}_p = 0 \tag{5}$$

where the vector **T** represents temperature at all the points in space and time where a model solution has been obtained, $\mathbf{a}_p = (\gamma, \kappa, a_c, a_\alpha, a_{rh}, a_T, a_T T_r)$ is the vector of a priori parameter values, **M(T)** and **L(T)** are linear operators determined at all space/time points by retaining the terms of the model equations (1) and (3) that are and are not affected by parameter changes, respectively. Specifically, the i^{th} row of **L(T)** includes the contribution at space/time point i from

$$\partial_t T + u\partial_x T + v\partial_y T - 0.94Q_0,$$

while the i^{th} row of **M(T)** correspondingly represents the transpose of the terms

$$\begin{bmatrix} w(T-T_d)/h \\ -(\partial_{xx}+\partial_{yy})T \\ 0.94Q_0 C \\ -0.94Q_0\alpha \\ -\rho C_E L v^a q_s(T) \\ T \\ -1 \end{bmatrix}.$$

Both **L** and **M** depend on the atmospheric forcing, which is imperfectly known, so that even if the model was perfect and the uncertain parameters optimally chosen, the model predictions would differ from the observations.

Because SST is a relatively well-measured variable, we follow BC and estimate the "corrective heat flux" $\delta\mathbf{q}$ that, for the a priori values of the uncertain model parameters, would be necessary to make the model SST match the observed SST exactly. To do so, we run the model using the observed SST, denoted by $\mathbf{T_0}$, instead of the calculated one, after interpolation on the model grid. Equation (5) is then only satisfied by adding a "heat flux correction" $\delta\mathbf{q}$:

$$\mathbf{L}(\mathbf{T_0}) + \mathbf{M}(\mathbf{T_0})\,\mathbf{a_p} + \delta\mathbf{q} = 0 \tag{6}$$

As expected from the limited SST agreement, the heat flux correction $\delta\mathbf{q}$ is rather large, and additional cooling would be needed for realistic simulations (Fig.2a).

Because $\delta\mathbf{q}$ depends linearly on the tunable model parameters, the estimation of their optimal value can be formulated as the linear inverse problem

$$\delta\mathbf{q} = \mathbf{M}(\mathbf{T_0})\,\delta\mathbf{a}, \tag{7}$$

where $\delta\mathbf{a}=(\delta\gamma, \delta\kappa,..., \delta\alpha_T T_r)$ represent the parameters changes that minimize the heat flux correction $\delta\mathbf{q}$, yielding

$$(\delta\mathbf{q})_{\min} = \delta\mathbf{q} - \mathbf{M}(\mathbf{T_0})\,\delta\mathbf{a}, \tag{8}$$

A good estimator of $\delta\mathbf{a}$ must take errors into account, as well as our knowledge of the expected parameter range.

There are many sources of errors in the estimates appearing in (7). The wind stress and the cloud data used to force the model have significant errors, resulting in model response uncertainties with large correlation scales, particularly in the equatorial waveguide. The observed SST is noisy as well, although to a lesser extent. When the best-fit calculation is based on a mean seasonal cycle as in this paper, there are also sampling errors which reflect the interannual variability and have large correlation scales. Finally, there are "irreducible" modelization errors inherent in the ocean model formulation, e.g., errors due to subgridscale phenomena or to the oversimplification of the ocean dynamics and the air-sea fluxes, which cannot be expected to be reduced by model tuning,. The modelization errors (called system errors in the Kalman filter literature) thus represent the errors that would exist if there were no observational errors and the uncertain parameters were at their true value.

Using a Bayesian viewpoint, Tarantola (1987) discusses the general inverse problem in the case of an inaccurate theory. When the forward problem is linear as in (7) and there are Gaussian modelization errors in **M**, described by the covariance $\mathbf{C_T}$, the solution of the

inverse problem takes a simple form if the observational errors in $\delta\mathbf{q}$ are Gaussian and statistically independent from the modelization errors. If the a priori value of the parameter correction $\delta\mathbf{a}$ is zero, as in the present case, the optimal solution is given by the minimum of the misfit function

$$S(\delta\mathbf{a}) = [(\mathbf{M}\delta\mathbf{a} - \delta\mathbf{q})' \mathbf{C}^{-1} (\mathbf{M}\delta\mathbf{a} - \delta\mathbf{q}) + \delta\mathbf{a}' \mathbf{C}_\mathbf{a}^{-1} \delta\mathbf{a}] / 2 \quad (9)$$

with $\mathbf{C} = \mathbf{C}_\mathbf{T} + \mathbf{C}_\mathbf{d}$, where $\mathbf{C}_\mathbf{d}$ is the error covariance matrix of the observations $\delta\mathbf{q}$, and the covariance matrix $\mathbf{C}_\mathbf{a}$ describes the a priori uncertainty of $\delta\mathbf{a}$. The solution is

$$\delta\mathbf{a} = (\mathbf{M}' \mathbf{C}^{-1} \mathbf{M} + \mathbf{C}_\mathbf{a}^{-1})^{-1} \mathbf{M}' \mathbf{C}^{-1} \delta\mathbf{q}. \quad (10)$$

BC followed this formalism, assuming for simplicity that the observational noise only affected the model matrix $\mathbf{M}$, and the modelization error only the heat flux correction $\delta\mathbf{q}$. On the basis of order of magnitude estimates, they used a constant rms error of 35 W/m^2 (10 W/m^2) with a simple exponential decay for the total (modelization) errors. There are a number of simplifications in this approach. As shown by (6), both $\delta\mathbf{q}$ and $\mathbf{M}$ depend on the input data (e.g., the surface wind stress affects both the heat exchanges and the ocean dynamics), hence they are both affected by data uncertainties and modelization errors. The errors in $\delta\mathbf{q}$ and $\mathbf{M}$ are thus *not statistically independent*, and the model matrix really is a stochastic regression matrix. Unfortunately, ordinary and generalized least squares estimators are in general not consistent in this case of nonlinear coupling between model and data errors. The error models used by BC are also highly idealized. Since the results of the tuning are sensitive to the assumed error models, we adopt a more elaborate strategy to achieve a refined estimate.

b. The adaptive procedure

The correlation scales of the model response errors due to forcing and SST uncertainties are large and complex, hence difficult to represent a priori. However, they can be estimated by using the different wind stress products and the long SST time series of section 2b, since many plausible realizations of the model seasonal response are available. We thus perform the optimization on the mean seasonal cycle, which is least noisy, and use the dispersion of the model seasonal responses as independent information to construct a more realistic model for the observational errors.

Assuming that the parameters do not vary in time, we can write for each year t (here $t = 1, 22$) and for each forcing i (here $i = 1, 5$), denoted by the upper index, that the linear model (7) holds:

$$\mathbf{L}^{t,i}(\mathbf{T_0}^t) + \mathbf{M}^{t,i}(\mathbf{T_0}^t)\ \mathbf{a_p} + \delta\mathbf{q}^{t,i} = 0. \quad (11)$$

Denoting long-term sample means by an overbar and the mean over the different forcing by an angle brace, we write relation (6) under the form

$$< \overline{\mathbf{L}(\mathbf{T}_0)} > + < \overline{\mathbf{M}(\mathbf{T}_0)} > \mathbf{a}_p + < \overline{\delta \mathbf{q}} > = 0. \tag{12}$$

The errors in (11) and (12) are due to forcing and SST uncertainties, and to model inadequacies.

Let us write the parameter estimation as the linear statistical model

$$< \overline{\delta \mathbf{q}} > = < \overline{\mathbf{M}} > \delta \mathbf{a} + < \bar{\mathbf{e}} > \tag{13}$$

where $< \bar{\mathbf{e}} >$ represents the errors, which are assumed to be Gaussian, with zero mean and unknown true covariance matrix **C**. Because of the statistical dependence between $< \overline{\delta \mathbf{q}} >$ and $< \overline{\mathbf{M}} >$, an estimate of $\delta \mathbf{a}$ is required before one may estimate the random errors from the sample. Thus, an adaptive approach is used, where the estimates of the observational error covariance and the model parameters are updated as part of an iterative procedure. If we have a first estimate of $\delta \mathbf{a}$, say $\delta \mathbf{a}_0$, which we will take equal to zero, then we can estimate for each year t the mean error over the different forcing, $<\mathbf{e}_1^t>$, by

$$<\mathbf{e}_1^t> = < \delta \mathbf{q}^t > - < \mathbf{M}^t > \delta \mathbf{a}_0. \tag{14}$$

A first sample estimate of the error covariance matrix associated with the random wind, cloud and SST errors is

$$\mathbf{S}_{r1} = \frac{1}{21 \times 22} \sum_{t=1}^{22} (< \mathbf{e}_1^t > - < \bar{\mathbf{e}}_1 >)(< \mathbf{e}_1^t > - < \bar{\mathbf{e}}_1 >)' \tag{15}$$

where we have assumed for simplicity that observations are independent at yearly intervals. We can also estimate for each forcing i the long-term mean error, $\bar{\mathbf{e}}_1^i$, by

$$\bar{\mathbf{e}}_1^i = \overline{\delta \mathbf{q}}^i - \overline{\mathbf{M}}^i \delta \mathbf{a} \tag{16}$$

and a first sample estimate of the error covariance matrix associated with the drag coefficient uncertainties is

$$\mathbf{S}_{f1} = \frac{1}{4 \times 5} \sum_{i=1}^{5} (\bar{\mathbf{e}}_1^i - < \bar{\mathbf{e}}_1 >)(< \bar{\mathbf{e}}_1^i > - < \bar{\mathbf{e}}_1 >)' \tag{17}$$

A first sample estimate of the error covariance associated with the observational uncertainties, say $\mathbf{S}_{d1}$, can then be obtained by

$$\mathbf{S}_{d1} = \mathbf{S}_{r1} + \mathbf{S}_n$$

and it can be used to compute an estimated generalized least squares estimate of $\delta \mathbf{a}$, say $\delta \mathbf{a}_1$. As in (10), we incorporate the modelization errors and our a priori knowledge on the model parameters,

$$\delta \mathbf{a}_1 = (\overline{\mathbf{M}}' \mathbf{S}_1^{-1} \overline{\mathbf{M}} + \mathbf{C}_a^{-1}) \overline{\mathbf{M}}' \mathbf{S}_1^{-1} \overline{\delta \mathbf{q}}, \tag{18}$$

with

$$\mathbf{S}_1 = \mathbf{S}_{d1} + \mathbf{C}_T \tag{19}$$

The procedure is repeated by using $\delta \mathbf{a}_1$ in (14) to get an improved estimate $\mathbf{S}_2$, leading to the parameter correction $\delta \mathbf{a}_2$, and so on. If $\delta \mathbf{a}_0$ represents a reasonable first guess and if the inverses in (18) are well-conditioned, the procedure should converge rapidly. The end result is a data error structure consistent with the results of the multi-year model run, and thus presumably a better parameter estimation.

The error model $\mathbf{S}_n$ represents most of the nonsystematic data and model errors; it also includes such data errors as artificial trends in wind and SST data. The true interannual variability is not treated as an error since it appears in both $\delta \mathbf{q}^t$ and $\mathbf{M}^t$ in (14). The weighting in the least squares fit is therefore based on data noise and uncertainties and it takes into account, at least approximately, the lack of independence between $\overline{\mathbf{M}}$ and $\overline{\delta \mathbf{q}}$. On the other hand, the weighting is not affected by the systematic errors that recur every year; model deficiencies, or systematic data biases, must be dealt with explicitly.

Because of the limited sample, the error covariance matrix $\mathbf{S}_{dn}$ is of strongly reduced rank and the inverse of $\mathbf{S}_n$ dominated by unreliable information. Hence, the problem is ill-conditioned. To circumvent the difficulty, we strongly reduce the dimension of the fields and tune the model in the highly truncated space. The iterative method is implemented in reduced space: for each forcing, each individual year is projected onto the reduced base, thereby defining a reduced heat flux correction and a reduced model matrix . By projection, a reduced modelization error matrix is also constructed. The sample error covariance matrix associated with the observational uncertainties and the optimal parameter corrections are then directly calculated in reduced space, so that the computational costs are very limited.

c. Model testing

The correctness of the SST model and the main assumptions in the inverse calculation (e.g., modelization and data errors) can be checked by looking at the residuals after optimization, but this ignores useful information on correlation scales. To take the multidimensional aspects of the fields into account, we generalize a multivariate test derived by Tarantola (1987) and consider the minimum of the misfit function (9), given by

$$2S(\delta \mathbf{a}_n) = \overline{\delta \mathbf{q}'} (\overline{\mathbf{M}} \mathbf{C}_a \overline{\mathbf{M}}' + \mathbf{S}_n)^{-1} \overline{\delta \mathbf{q}} \tag{20}$$

with $\mathbf{S}_n = \mathbf{S}_{dn} + \mathbf{C}_T$. The null hypothesis that the only errors besides the observational ones are the modelization errors can be tested since the test statistic (20) is distributed as Hotelling's T^2 with degrees of freedom η (the reduced dimension) and τ (the equivalent degrees of freedom of $\mathbf{S}_n$). If (20) exceeds the critical value at a given level of confidence, then some of the assumptions are unlikely to be acceptable. Since, except for possible biases, the observational uncertainties are represented by an error model which is, by construction, consistent with the available observations, the most likely interpretation is that the model is not as accurate as it has been assumed, i.e., the modelization errors have been underestimated.

4. TUNING THE TROPICAL ATLANTIC SST MODEL

The monthly values of $\delta \mathbf{q}^{t,i}$ and $\mathbf{M}^{t,i}$ are first spatially smoothed with a 5° × 5° running average. The fit is then done in the region between 10°S and 20°N, by considering

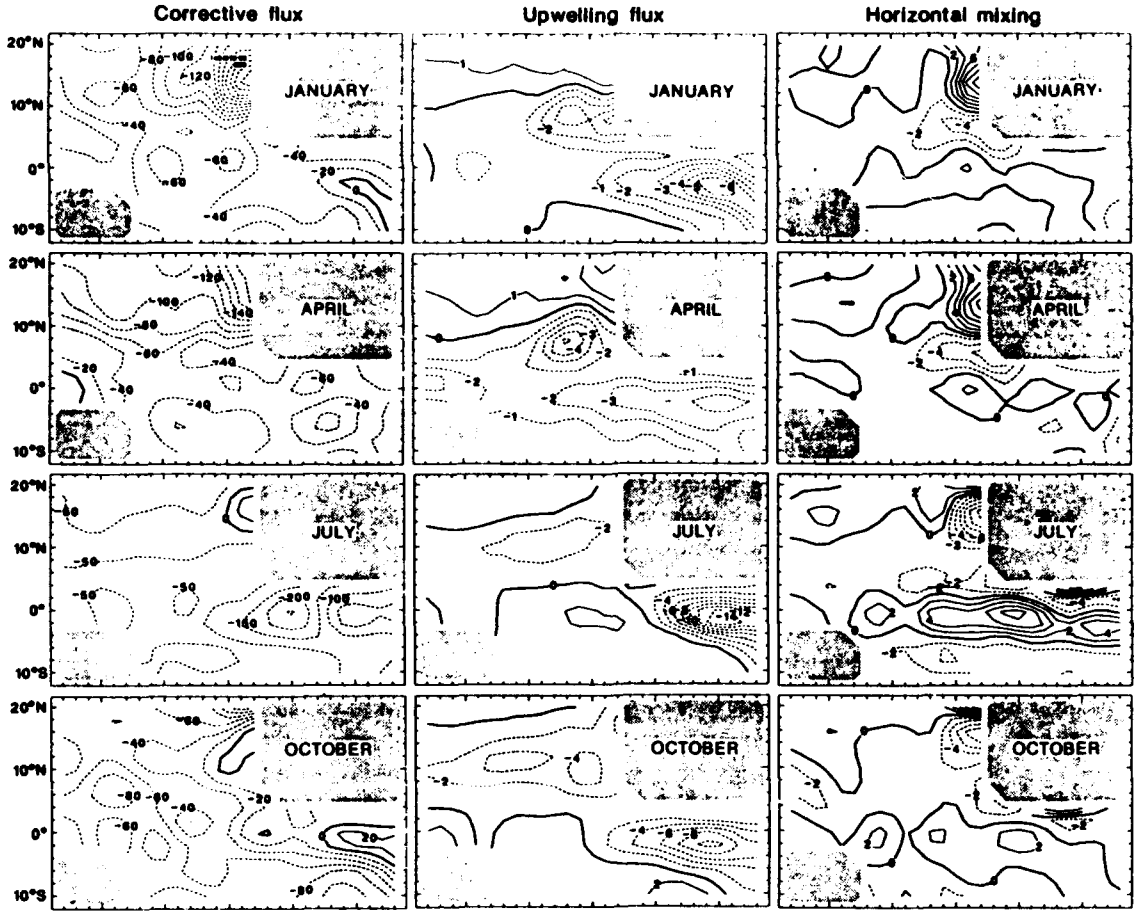


Figure 2a. (left) Mean heat flux correction in Wm^{-2} during January, April, July and October for the period 1965-1986, when using the a priori values of the model parameters. Corresponding values of (center) the upwelling flux and (right) horizontal diffusion.

January, April, July, and October, which are representative of the various SST regimes. The data dimension p is $322 \times 4 = 1288$.

The mean heat flux correction $< \overline{\delta q} >$ is represented in Figure 2a. The rms value is large (69 Wm^{-2}), and negative values in excess of -100 Wm^{-2} are found off Africa and in the Gulf of Guinea, mostly where the largest SST differences are observed. The tuning can be viewed as determining the best fit of the heat flux correction vector in Figure 2a by the seven column vectors of $< \overline{\mathbf{M}(\mathbf{T}_0)} >$, which are represented in Figures 2a,b (units are arbitrary). The upwelling pattern (Fig. 2a, center) has a large signal in the Gulf of Guinea with maximum amplitude during the upwelling season in July; a smaller signal is seen in the ITCZ with maximum amplitude off Africa, except in April. The meridional scaled of the diffusion pattern (Fig. 2a, right) is slightly smaller than that of upwelling. The cloud pattern cloud pattern (Fig. 2b, left) has broader scales and its seasonal changes reflect

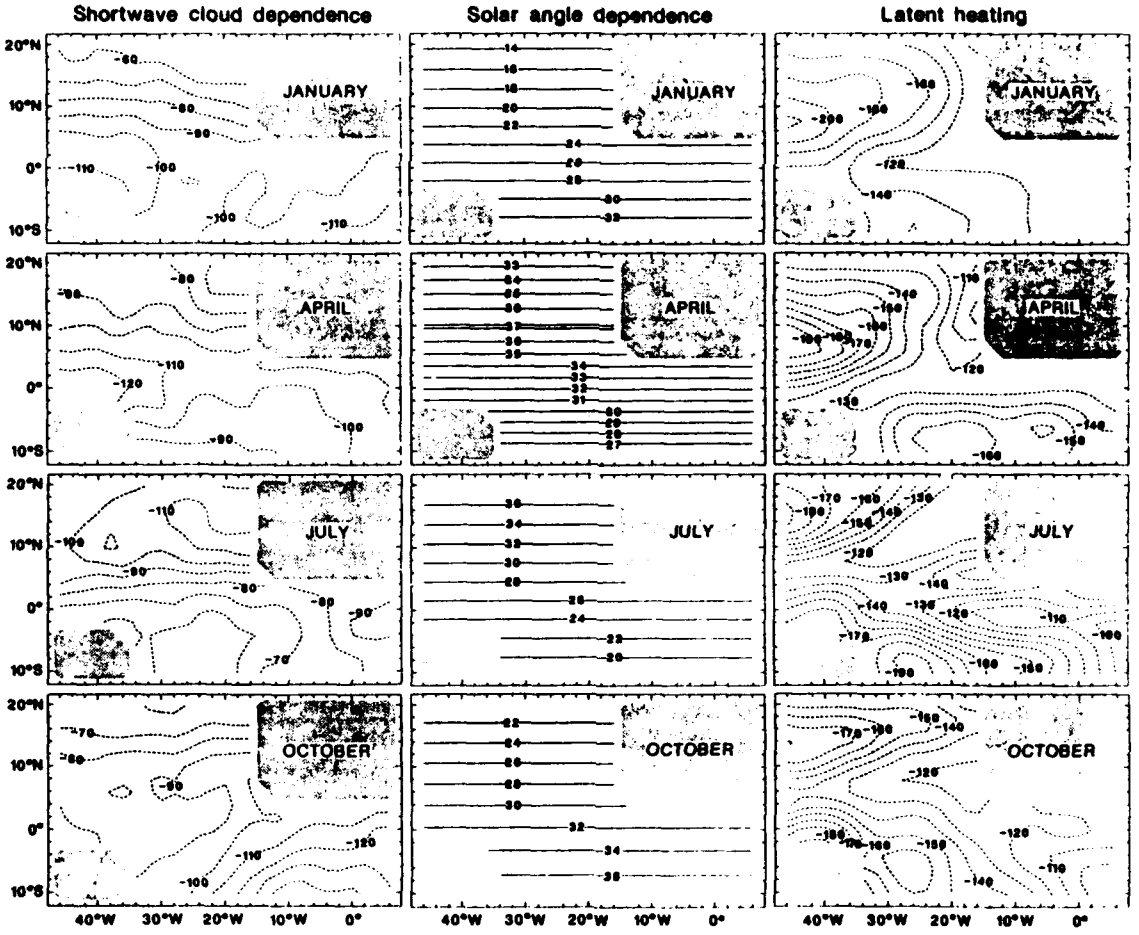


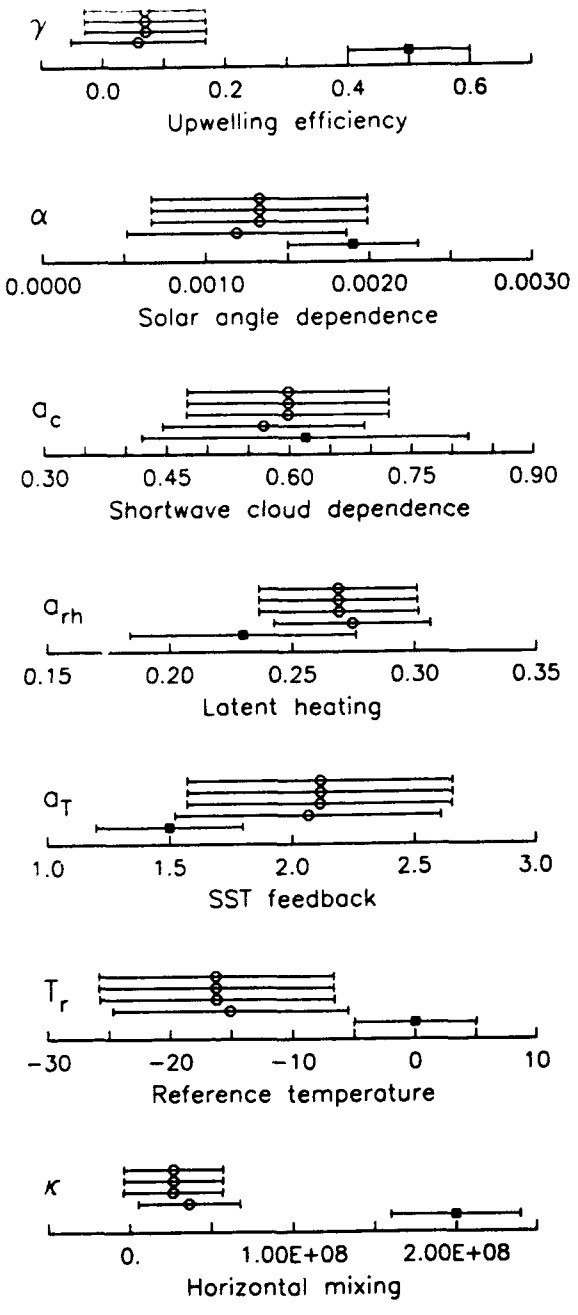
Figure 2b. (left) Cloud factor values during January, April, July and October for the period 1965-1986, when using the a priori values of the model parameters, (center) solar angle, and (right) latent heat flux.

those of Q_0 and C. The evaporation pattern (Fig. 2b, right) has a large meridional scale and strong zonal gradients. Additional patterns are the insolation pattern in Figure 2b (right), a constant, and the observed SST pattern shown in Figure 1.

The data compression is done by working in the space defined by orthonormalizing the eight vectors consisting of $< \overline{\delta \mathbf{q}} >$ and the seven column vectors of $< \overline{\mathbf{M}} >$. As the dimension η of the subspace is the number of adjustable parameters plus one, the inverse problem remains formally overdetermined. As described in section 3, $\delta \mathbf{q}^{t,i}$ and $\mathbf{M}^{t,i}$ are projected onto the reduced base for each year t, and the sample error covariance matrix directly estimated in reduced space at each iteration n. Because $\mathbf{S}_{\mathbf{dn}}$ has limited degrees of freedom, its elements are inaccurately known (large sampling errors) and the condition number of the matrix $\mathbf{S}_{\mathbf{n}}$ is very large. Lacking precise information on the modelization errors, we use BC's model, but double the rms error to 20 Wm^{-2}. This modelization error matrix is not sufficient to insure good conditioning, so a singular value decomposition is used to invert $\mathbf{S}_{\mathbf{n}}$ in (18). In practice, we apply a taper which is an estimate of the accuracy of the elements of $\mathbf{S}_{\mathbf{dn}}$.

For simplicity, we use zero for the parameter correction $\delta \mathbf{a}_0$, but the results are similar when using a different initial value. Convergence is reached in two or three iterations, with the largest changes occuring after the first iteration. Figure 3 shows the a priori and a posteriori values of the adjustable parameters with twice their standard deviation (an approximation to the 95% confidence interval). Of the seven adjustable parameters, two strongly decrease to values that are positive, but not significantly different from zero at the 5% level: the upwelling efficiency γ and the horizontal diffusion κ. Both parameters are well-resolved by the data set and independently resolved. However, such a small value for the upwelling efficiency is unlikely from a physical point of view. Although the changes in the cloud factor a_c and the latent heat flux a_{rh} are also well-resolved, they are not statistically significant at the 5% level, which suggests that the a priori choices were good, needing only little adjustment. However, the two parameters are not independently resolved and are anticorrelated, and correlated with the three remaining parameters, a_α, a_T and $a_T T_r$, which are poorly resolved by the data set.

Figure 4 shows the heat flux correction (8) after tuning. The amplitudes are smaller than in Figure 2: the rms value has dropped to 32 Wm^{-2} and the space-time average to -7 Wm^{-2}, suggesting that the warm SST bias in Figure 1 should be mostly corrected. However, heat flux corrections larger than 100 Wm^{-2} can still be seen off the North African coast during winter and in the equatorial upwelling region during summer. These are too large to be explainable by the data uncertainties and are associated with model deficiencies, as discussed by BC and Scoffier et al. (1993).



To verify the consistency of the inverse calculation, we apply the test of section 3c. Although the critical value of the test statistic (20) is difficult to establish as the total error covariance is the sum of a sample one and an (assumed to be) true one, upper and lower bounds can easily be found. For true covariances, the critical value, given by the χ^2 distribution with 8 degrees of freedom (the dimension of the space), would be 16 at the 5% level (lower bound). For sample covariance matrices, it would be given by Hotelling's T^2 and equal to 32 (upper bound). The test is 385, which largely exceeds the latter value. This confirms that the modelization errors have been strongly underestimated. In particular, there are large modelization biases, not only random modelization errors as assumed.

Figure 3. Evolution of the parameter corrections as a function of the number of iterations. The error bars represent the 95 % confidence intervals.

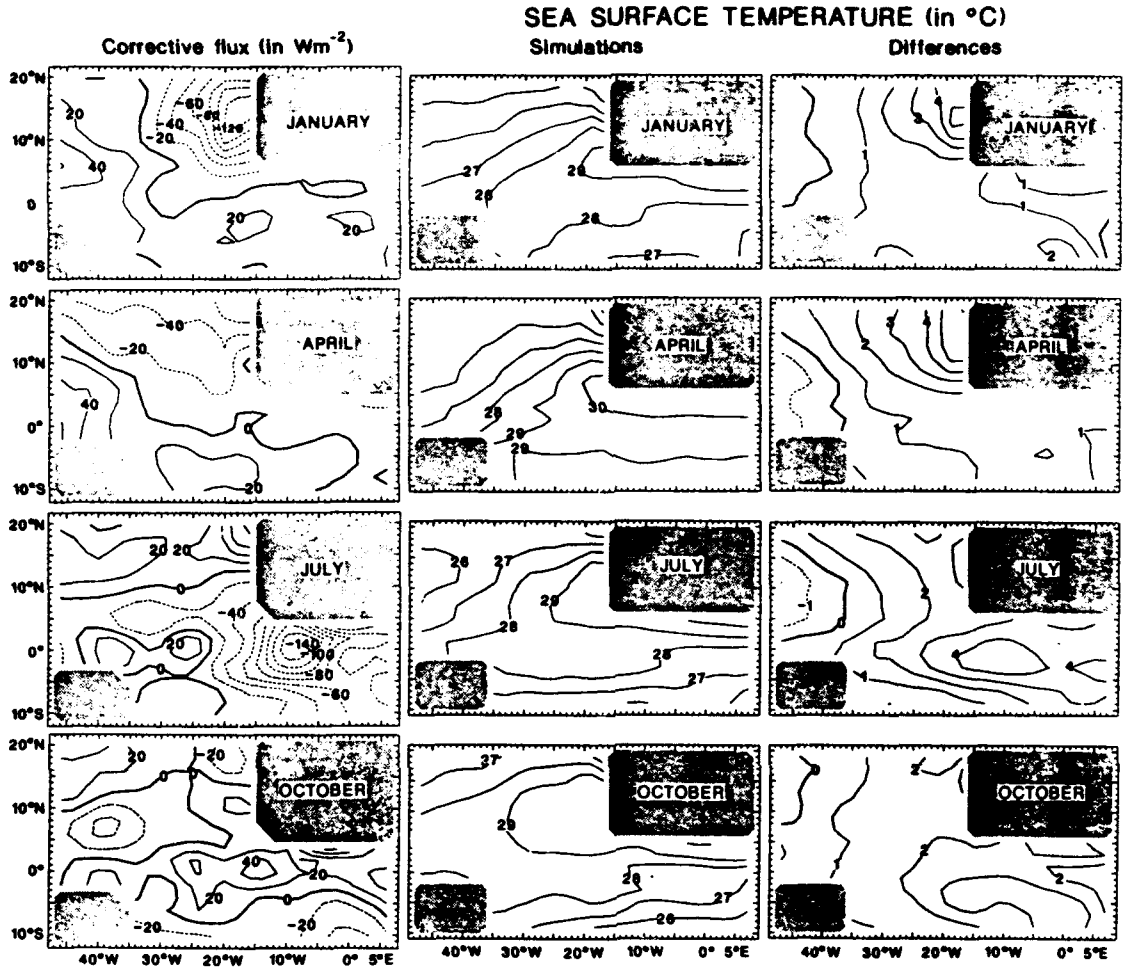


Figure 4.(left) Mean heat flux correction in Wm^{-2} during January, April, July and October for the period 1965-1986, after optimization; (center) Corresponding SST predictions; (right) Differences between simulated and observed SST.

Because the tuning minimizes the heat flux correction (more precisely a weighted form of it), it is of interest to verify whether the SST predictions have been improved by the parameter changes. The tuned model was thus run with the same forcing fields as before. As expected, a more realistic SST field is obtained (Fig. 4, center), although model-observation differences of a few degrees can still be seen in the upwelling region off Africa during the first part of the year and in the Gulf of Guinea during the second part (Fig. 4, right). Tables 1 and 2 suggest that the model improvements are limited to a decrease of the warm SST bias, although it still averages to 1.5°C. The mean seasonal variability and the observed SST anomalies are not significantly improved, so that the SST model remains largely inconsistent with the observations: the tuning is unable to compensate the model shortcomings.

The method is not very sensitive to the details in the calculation. The largest parameter corrections were obtained when working with low-passed seasonal data, because filtering

decreases the magnitude of the observational and modelization errors, thereby giving more weight to the observations in the best-fit calculation. Unfortunately, the increased resolution by the data set leads to vanishing upwelling efficiency, which is not acceptable. Although a larger upwelling efficiency could be obtained by constraining more γ, this stresses the inadequacy of the upwelling representation for the tropical Atlantic.

5. CONCLUSIONS

We have developed an adaptive inverse method to tune the adjustable parameters of a tropical SST model in a way that optimally takes into account the large uncertainties of the atmospheric forcing and the oceanic data, the expected modelization errors and our a priori knowledge of the parameter values. This is achieved by performing the model optimization for the mean seasonal SST cycle and using the dispersion of the model responses for each year and (equally plausible) forcing field as independent information to construct a sample estimate of the observational error covariance matrix. The procedure is more refined than that of BC in that the nonlinear nature of the inverse problem is taken into account and the large correlation scales of the forcing uncertainties are represented realistically. The method is general as long as the parameters enter the SST equation linearly, and it can be extended to the nonlinear case by using an iterative approach. Since the optimization is performed in a strongly reduced space, the computational cost is limited. However, the estimation of the observational errors requires that several multi-year model runs be available.

The method has been applied to tuning BC's SST model of the tropical Atlantic. The optimization reduces the warm SST bias of the model, but brings no significant improvement in its ability at representing the seasonal or interannual SST fluctuations. A statistical test of the correctness of the assumptions in the inverse calculation shows that the modelization errors are much larger than assumed. The model flaws are discussed in Scoffier et al. (1993), who show that the model's inability to properly represent SST cooling by upwelling is linked to the parameterization of T_d in (1) and, as seen in Figure 2a (center), may result in SST heating by upwelling when the SST is low and the thermocline deep, which is not realistic.

Finally, it should be noted that the adaptive tuning procedure provides an alternative to imposing the "correction flux" that is often needed to avoid climate drift when coupling an SST model to an atmospheric model. Indeed, the decrease in mean SST bias should decrease climate drift in the coupled mode without introducing the drawbacks of the correction flux method, because the correction more properly takes place via model parameters, without altering the SST dynamics.

Acknowledgements

We would like to thank B. Blumenthal for his help, C. Wunsch and F. Martel for useful discussions. This research was supported by grants from the DRET and PNEDC, and done in part at the Lamont Doherty Geological Observatory, under ONR contract N00014-J-15951 with Columbia University, and at the Massachusetts Institute of Technology, whose hospitality to one of us (CF) is gratefully acknowledged.

REFERENCES

Blumenthal, M.B. and M.A. Cane, 1989: Accounting for parameter uncertainties in model verification: an illustration with tropical sea surface temperature. *J. Phys. Oceanogr., 19,* 815-830.

Braconnot, P., and C. Frankignoul, 1993: Testing model simulations of the thermocline depth variability in the tropical Atlantic from 1982 through 1984. *J. Phys. Oceanogr.*, in press.

Duchêne, C., and C. Frankignoul, 1991: Seasonal variations of surface dynamic topography in the tropical Atlantic: Observational uncertainties and model testing, *J. Mar. Res., 49,* 223-247.

Esbensen, S. K. and Y. Kushnir, 1981: The heat budget of the global ocean: an atlas based on estimates from marine surface observations. *Climatic Research Institution, Report 29,* Oregon State University, Corvallis, 27 pp.

Frankignoul, C., C. Duchêne and M. Cane, 1989: A statistical approach to testing equatorial ocean models with observed data. *J. Phys. Oceanogr., 19,* 1191-1208.

Isemer, H.-J. and L. Hasse, 1987: *The Bunker Climate Atlas of the North Atlantic Ocean. Vol. 2: Air-sea Interactions.* Springer Verlag, 252 pp.

Large, W. G., and S. Pond, 1981: Open ocean momentum flux measurements in moderate to strong winds. *J. Phys. Oceanogr.*, **11,** 324-336.

Liu, W. T., K. B. Katsaros, and J. A. Businger, 1979: Bulk parameterization of air-sea exchanges of heat and water vapor including the molecular constraints at the interface. *Atmos. Sci.,36,* 1722-1735.

Scoffier, N., C. Frankignoul and M.A. Cane, 1993: An adaptive procedure for tuning a sea surface temperature model. In preparation.

Seager, R., S.E. Zebiak and M.A. Cane, 1988: A model of the tropical Pacific sea surface temperature climatology. *J. Geophys. Res. 93,* 1265-1280.

Servain, J., J. Picaut and A.J. Busalacchi, 1985: Interannual and seasonal variability of the tropical Atlantic ocean depicted by 16 years of sea surface temperature and wind stress. *In Coupled Ocean-atmosphere Models*, 211-237, ed. J.C.J. Nihoul, Elsevier.

Smith S. D., 1988: coefficicent for sea surface wind stress, heat fluxes and wind profiles as a function of wind speed and temperature. *J. Geophys. Res., 93,* 15467-15472.

Tarantola, A. 1987: *Inverse problem theory*. Elsevier, 613 pp.

NEW DEVELOPMENTS IN STIRRING AND CHAOS: POSSIBLE ROLE IN OCEAN SCIENCES

Julio M. Ottino
R. R. McCormick School of Engineering and Applied Science, Northwestern University
Evanston, Illinois 60208-3120

1. Introduction and Setting

Getting asked to comment outside one's area is both flattering and healthy. However, the intersection between what one might know and what people might like to hear—especially when one cannot accurately gauge the needs of an audience technically far-removed from one's own—might, in fact, be remarkably small. In spite of having heard much about oceanography during the 'Aha Huliko'a Hawaii workshop held in January 1993, such still might be my predicament in this particular case. My role here is to present a view of mixing and chaos theory and indicate what relevance it might have in problems of interest in oceanography. My assumption is that the reader is at least vaguely familiar with some aspects of dynamical chaos.

It probably has not escaped anybody's attention that during the past few years there has been considerable interest in chaos. The theoretical foundations of the subject are on firm footing and demonstrations of chaos have been firmly established by analytical, computational, and experimental means. So much has been the bulk of the work generated that hardly a month goes by without a book being published and at the last count there were at least half a dozen journals largely devoted to the topic. The collective impact of the body of work so generated, with no apparent signs of slowing down, can be compared to the emergence of a new paradigm. Regrettably, as in any emerging area, sometimes to the chagrin of its creators, there is some degree of overshoot and less than guaranteed unbounded enthusiasm. Not everything that claims to be useful is likely going to pass the test of time, but it is also doubtful that no permanent mark will be left. Undoubtedly, the way that people will be educated will change (in fact, it is already changing; college physics textbooks now have sections devoted to chaos). A non-trivial consequence of this trend is that data that could have been discarded a decade or so ago as being unanalyzable will be scrutinized in the future in more detail for trends and patterns.

The most intuitively understandable definition of chaos is magnification of small errors and the impossibility of making predictions for long times. This statement—so often repeated—has produced the impression that chaotic systems cannot be predicted at all. Strictly speaking this is far from being true. What cannot be predicted is the detailed evolution of a *specific* initial condition. The behavior of the system at large—that of a multitude of initial conditions—may be quite robust, and this is, in fact, what matters in many situations of practical interest. As we shall see, a particularly important example is provided by mixing of fluids.

Claims for applications of chaos theory abound. In fact, in the context of dynamical systems, the absence of chaos seems to be the exception rather than the rule. The applications, however, have been largely *a posteriori*; that is, explaining existing (complex) behavior and demonstrating that the complexity stems from an underlying deterministic cause. Much less has been done on the predictive side; using theory to predict the state of systems for long times. It is apparent that there might be a need for both types of works in the context of oceanography: interpretation of seagoing data being in the first class, prediction based on available information being the second. An alternative breakdown might divide the tasks between analysis of observational data on one side and analysis of output from numerical models, such as general circulation models, on the other.

The objective of this article is to provide a brief overview of some of our past work on mixing and chaotic advection including a few remarks not made before. However, in order to accomplish this objective and setting things in perspective, a number of remarks pertaining to general aspects of chaos theory will be made. As there are a large number of references for this material, no review is attempted. The second part of the presentation involves issues in chaotic advection. One general reference on this topic is available (Ottino, 1989a) and an introductory review to chaotic mixing is given in Ottino (1989b). Several other reviews are available (Aref, 1991; the entire issue of *Physics of Fluids A*, *3*, May 1991, is entirely devoted to stirring and mixing).

2. Dynamical Chaos: Brief Review of Essential Concepts

During the past few years there has been a realization that nonlinear dynamical systems are able to display a variety of what superficially might be regarded as two contradictory—but, in fact, perfectly coexisting—behaviors. On one hand the output can be order (e.g. solitons), on the other it can be chaos (See Fig. 1). Often a system exhibits both behaviors simultaneously.

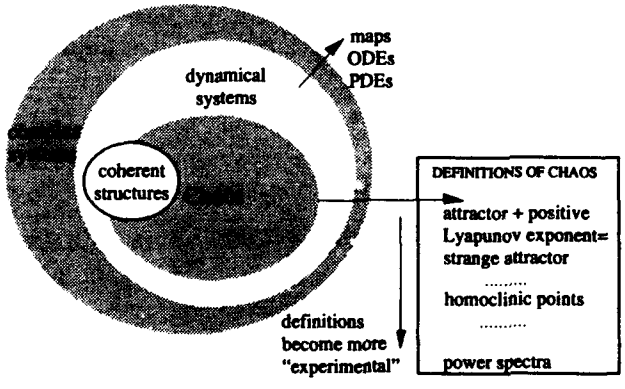


Figure 1. Overview of dynamical systems and definitions of chaos.

In the context of our discussion dynamical systems are given by sets of ordinary differential equations (ODEs) or maps

$$\frac{d\mathbf{x}}{dt} = \mathbf{f}(\mathbf{x}, p), \mathbf{x}_{n+1} = \mathbf{g}(\mathbf{x}_n, p), \tag{1}$$

where $\mathbf{x}=(x_1,...,x_k)$ with $k\geq 1$. In order for the system to be chaotic $k\geq 3$ for ODEs, $k\geq 1$ for maps. The components of the vector $\mathbf{x}$ might have either a transparent physical meaning or not according to the problem in question. For example, in chaotic advection $\mathbf{x}$ denotes the actual physical space but in problems where the set of ODEs is arrived at through truncation, as in the case of Lorenz's equations, the variables have a less transparent one. The space spanned by $\mathbf{x}$ is called the phase space of the system and p, or a set of p's, are parameters such as the Rayleigh number or Reynolds number. In the case of systems described by partial differential equations (PDEs) the number of degrees of freedom is infinite and therefore the phase space is infinite as well. According to the form of $\mathbf{f}(\mathbf{x})$, more precisely the sign of $\nabla \cdot \mathbf{f}(\mathbf{x})$, we can speak of two kinds of systems. In one class there is volume contraction in phase space ($\nabla \cdot \mathbf{f}(\mathbf{x})<0$); these are dissipative *systems*. The other kind of systems are those that conserve volume in phase space ($\nabla \cdot \mathbf{f}(\mathbf{x})=0$), and of those the most important sub-class is given by the so-called *Hamiltonian systems* [a system can be volume preserving and not be Hamiltonian; however, if it is Hamiltonian it is volume preserving]. The prototypical example of a dissipative system is the forced pendulum with friction; the prototypical Hamiltonian system is a forced pendulum without friction. The bulk of the presentation here will be restricted to volume preserving systems. However, in order to place the topic in perspective a few remarks pertaining to dissipative systems might be in order (for mathematical presentations of dynamical systems see Guckenheimer and Holmes, 1983 and Wiggins, 1991; for a collection of classical papers the reader can consult Hao, 1984; an accessible introduction to chaos in both dissipative and non-dissipative systems is given by Doherty and Ottino, 1988).

Dissipative systems are typically associated with one dimensional maps (such as the logistic equation, volume contracting systems of ordinary differential equations—such as in the Lorenz equations—and strange attractors characterized by fractal dimensions. If the model is continuous, a dissipative system must consist of at least three (autonomous) ordinary differential equations in order to exhibit chaos (as in the Lorenz model). On the other hand, if the model is represented by a mapping $\mathbf{x}_{n+1}= \mathbf{g}(\mathbf{x}_n, p)$, it can display chaos in one dimension, i.e. with $\mathbf{x}_n$ being real (as in the logistic equation). By contrast, a volume preserving mapping must be at least two-dimensional to be chaotic. As opposed to dissipative systems, Hamiltonian systems have no stable steady states, the phase space does not contract, and there are no attractors, strange or otherwise. Dissipative and Hamiltonian systems have their own ways of "going chaotic." However, both types of systems have a few things in common. One of the connections is a stretching-and-folding mechanism in phase space; this is what might be regarded as the basic mechanism leading to amplification of errors in chaotic systems.

The closest applications related to oceanography are probably those in meteorology (for reviews see Tsonis and Elsmer, 1989; Yang, 1991; Zeng at al., 1993), a well-known application being the model of El Niño-Southern Oscillation system (Vallis, 1986). Literature in this area appears voluminous when compared with oceanographic applications. When averaged across all fields probably over 95% of the current applications of chaos involve dissipative systems. In meteorology the ratio is close to 100%.

The first question that should be asked when facing a complex system or signal is to determine if it is stochastic or chaotic (Sigeti and Horsthemke, 1987). If the process is indeed chaotic, the next task is to determine whether or not it possesses a strange attractor, the hope being that no matter how large the original system might be, the dynamics might be captured by the motion in a subspace of much smaller dimension (Fig. 2). These reconstruction techniques can be based on the measurement of one or more components of the vector **x** (Packard et al., 1980; Wolf et al., 1985). Subsequently, the "amount of chaos" in the projection of the attractor can be characterized by determining its dimensions, by measuring one or more Lyapunov exponents, and so forth (for a practical application of these ideas, see Parker and Chua, 1989). Naturally, there are instances when the analysis starts with the equations themselves (for example the Navier-Stokes equations in a problem in fluid mechanics). However, in many cases the equations are unmanageable and they have to be transformed in a way that is suitable for analysis. This is where the issue of representing a PDE in terms of finite degrees of freedom appears. The most famous example belonging to this class is the reduction of the Rayleigh-Bénard flow problem to the Lorenz equations (Lorenz, 1962). A question in this case is whether the chaos that is seen in the 3x3 truncated system would actually appear in the full problem or not. This issue was studied by, among others, Wiin-Nielson (1992). The answer, not surprisingly, is that, yes, the details of the process might depend heavily on the number of equations considered and that extreme care should be exercised in extrapolating conclusions outside the range of applicability of the equations.

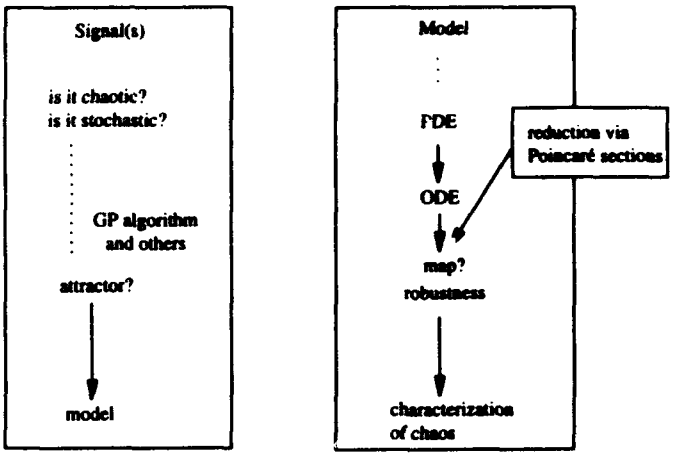


Figure 2. Typical modes of analyses of chaotic systems.

In the viewpoint advocated here truncation is not an issue. The viewpoint adopted is a purely kinematical one which is only suited to the analysis of fluid mechanical issues. However, to the extent that oceanography is routinely faced with such issues, this does not seem to be a terribly important drawback. The dynamical system is the velocity field itself. An important fringe benefit of this approach is the rather transparent connection between the underlying mathematics and their associated physical meaning.

3. Chaotic Advection: Kinematics

The study of mixing begins with the analysis of the motion due to an imposed velocity field; i.e., the study of the dynamical system

$$\frac{d\mathbf{x}}{dt} = \mathbf{v}(\mathbf{x},t) \tag{2}$$

where $\mathbf{v}(\mathbf{x},t)$ is usually obtained by solution of the Navier-Stokes equations and is volume preserving (i.e., $\nabla \cdot \mathbf{v}=0$). The solution of (2) with the initial condition that $\mathbf{x}=\mathbf{X}$ at $t=0$:

$$\mathbf{x}(t) = \phi(\mathbf{X},t) \text{ such that } \mathbf{X} = \phi(\mathbf{X},0) \tag{3}$$

This solution is called the *flow* or *motion*. Although traditional, and probably by now unchangeable, it should be noted that Eqns. (2-3) represent an abuse of notation. The variable $\mathbf{x}$ has two meanings that can be inferred according to context. In the first one, as in the right hand side of (2), $\mathbf{x}$ represents a fixed position in space; in the second one, as in the left hand side of (3), $\mathbf{x}$ represents the position of particle $\mathbf{X}$ at time t. Note also that it is common to refer to a specific fluid particle as "particle $\mathbf{X}$," when in fact we mean *the fluid particle that was initially located at position* $\mathbf{X}$. Equations written in terms of $\mathbf{X}$ are

referred to as Lagrangian; equations written in terms of $\mathbf{x}$ are referred to as Eulerian. These two viewpoints are classical in fluid mechanics. The key idea in chaotic advection is that whereas $\mathbf{v}(\mathbf{x},t)$ might be simple, $\mathbf{v}(\mathbf{X},t)$ can be extremely complicated.

The traditional characterization of velocity fields, usually $\mathbf{v}(\mathbf{x},t)$, is in terms of streamlines, streaklines, and pathlines. A graph of equation (3) for a single $\mathbf{X}$, with t as a parameter, gives the *pathline* of particle $\mathbf{X}$. The *streamlines* corresponding to the velocity field $\mathbf{v}(\mathbf{x},t)$ at time t is the solution of $d\mathbf{x}/ds = \mathbf{v}(\mathbf{x},t)$, where s is a parameter and t is fixed. The streakline passing through $\mathbf{x}'$ at time t is the locus of all particles which passed through $\mathbf{x}'$ during the interval 0 to t. Physically, this corresponds to the curve traced out by a non-diffusive dye which is injected at $\mathbf{x}'$.

A description of a velocity field in terms of streaklines and pathlines represents a nearly complete characterization of the flow. However, analytical examples of streamlines, streaklines, and pathlines are rare unless the flows happen to be trivial. The reason has to do with the fact that many solutions are chaotic and therefore cannot possibly be written down. In all the examples considered here the velocity fields are two dimensional and time periodic. It should be pointed out that steadiness does not preclude chaos. The velocity field, however, has to be three dimensional for this to occur.

The most studied case of chaotic advection corresponds to time-periodic velocity fields. A time-periodic flow can be regarded as a composition of motions or, equivalently, the iteration of a map. A few remarks regarding the composition of motions seem in order, because there are subtle points which are often misunderstood. When two different motions, $\phi^{(1)}$ and $\phi^{(2)}$, follow each other, they can be composed as

$$\phi^{(2)}\left[\phi^{(1)}(\mathrm{X},t_a),t_b\right] \tag{4}$$

Here, the first motion acts for time t_a, and the second motion acts for time t_b. It is understood that the final position of the particle after completion of the first motion constitutes the initial position for the particle for the second motion. In general this is not equivalent to

$$\phi^{(1)}\left[\phi^{(2)}(\mathrm{X},t_b),t_a\right] \tag{5}$$

In this case, the first motion acts for time t_b, and the second motion acts for time t_a.

Even composing a single flow with itself can be a bit subtle. A straightforward composition of flows, i.e., transforming $\mathbf{X}$ with ϕ for t, and then transforming again for τ yields, in general, incorrect results. This occurs because the velocity field is time dependent. When the Eulerian velocity field is unsteady, it matters not only *where* a particle is located, but when it is found there. By contrast, when the velocity field is time

independent, a flow may be composed with itself, and the composition is also commutative, i.e.,

$$\mathbf{x}(t+\tau) = \phi\big[\phi(\mathbf{X},\tau),t\big] = \phi\big[\phi(\mathbf{X},t),\tau\big] = \phi(\mathbf{X},t+\tau) \tag{6}$$

If the velocity field is time periodic, i.e., $\mathbf{v}(\mathbf{x},t) = \mathbf{v}(\mathbf{x},t+T)$, then a flow may be composed with itself, but only for an amount of time which is an integer multiple of the period of the velocity field:

$$\mathbf{x}(t+T) = \phi\big(\phi(\mathbf{X},T),t\big) \tag{7}$$

$$\mathbf{x}(t+nT) = \phi\big(\phi\big(\phi\ldots\phi(\mathbf{X},T)\ldots,T\big),T\big),t\big) = \phi\big(\phi(\mathbf{X},nT),t\big) \tag{8}$$

Flows due to a time periodic velocity field are frequently written as a *mapping*:

$$\mathbf{x}_{n+1} = \mathbf{M}\mathbf{x}_n \tag{9}$$

Customarily, the parenthesis around $\mathbf{x}_n$ are omitted. In mapping notation, usually the initial particle position is denoted as $\mathbf{x}_0$, rather than $\mathbf{X}$. Equation (8) gives the position of a particle at the end of the $(n+1)$ period, given its position at the end of the nth period. Since they are derived from periodic velocity fields, a mapping may be composed with itself:

$$\mathbf{x}_{n+2} = \mathbf{M}\mathbf{M}\mathbf{x}_n = \mathbf{M}^2\mathbf{x}_n \tag{10}$$

$$\mathbf{x}_{n+k} = \mathbf{M}^2\mathbf{x}_n^k \tag{11}$$

Of course, two different mappings may be composed together; i.e., if $\mathbf{x}_1 = \mathbf{M}\mathbf{x}_0$ and $\mathbf{x}_2 = \mathbf{N}\mathbf{x}_1$, then $\mathbf{x}_2 = \mathbf{N}\mathbf{M}\mathbf{x}_0$. In some respects, a mapping does not contain quite as much information as the corresponding motion. However, it does possess most of the important qualitative characteristics. It might be argued that these considerations apply to too simple cases. However, a complete understanding of time-periodic flows seems necessary before venturing into general unsteady flows.

4. Stretching and Regular Flows

Stretching lies at the heart of mixing. Stretching governs the fine scale of passive scalars dispersed in the flow and acts as a fabric for the evolution of diffusing scalars in the flow. To quantify the amount of stretching which occurs around a particle we follow a small material vector $\delta\mathbf{X}$ attached to the particle. The length stretch, λ, is simply the ratio of the length a time t, $\delta\mathbf{x}$, to the initial length:

$$\lambda = \frac{|\delta \mathbf{x}|}{|\delta \mathbf{X}|} \tag{12}$$

The orientation vector, denoted **m**, is simply $\delta\mathbf{X}$ normalized to unit length:

$$\mathbf{m} = \frac{\delta \mathbf{x}}{|\delta \mathbf{x}|} \tag{13}$$

The time evolution of the length stretch can be written as

$$\frac{\dot{\lambda}}{\lambda} = (\nabla \mathbf{v})^T : \mathbf{mm} = \mathbf{D} : \mathbf{mm} \tag{14}$$

where **D** is the symmetric part of the velocity gradient tensor, $\nabla\mathbf{v}$. By the Cauchy-Schwarz inequality, $\dot{\lambda}/\lambda$ is bounded by $(\mathbf{D}:\mathbf{D})^{1/2}$ (since the magnitude of the dyad **mm** is equal to one). The normalized stretching rate is called the stretching efficiency:

$$e = \frac{\mathbf{D} : \mathbf{mm}}{(\mathbf{D} : \mathbf{D})^{1/2}} \tag{15}$$

In an n-dimensional flow, the efficiency can attain a maximum value of $(1\text{-}1/n)^{1/2}$.

In many flow systems, the instantaneous values for both specific stretch rate and efficiency vary erratically in time. More useful quantities are the time averaged values, α_{avg} and e_{avg}:

$$\alpha_{avg} = \frac{1}{t}\int_0^t \mathbf{D} : \mathbf{mm}\, dt' = \frac{1}{t}\int_0^t \frac{d(\ln \lambda)}{dt'}\, dt' = \frac{\ln \lambda}{t} \tag{16}$$

$$e_{avg} = \frac{1}{t}\int_0^t \frac{\mathbf{D} : \mathbf{mm}}{(\mathbf{D} : \mathbf{D})^{1/2}}\, dt' \tag{17}$$

A system is considered efficient for mixing if the long time value (i.e., as $t \to \infty$) of α_{avg} (or equivalently e_{avg}) tends to a positive value, regardless of the initial orientation of the material filament $\delta\mathbf{X}$.

A complicated stretching function, with a nearly constant time average, is a symptom of "chaotic advection." Steady two-dimensional flows with $\nabla \cdot \mathbf{v}=0$ cannot produce chaotic advection; stretching is linear in time, the stretching function decays as $1/t$, and the efficiency decays to zero. This can be seen in various ways. A steady area preserving two-

dimensional flow is characterized by the streamfunction $\psi(x,y)$. Level curves $\psi(x,y,t=\text{fixed})$ give the instantaneous picture of the streamlines which in this case coincides with the pathlines and streaklines. If the flow is *bounded*, the flow can be divided into regions of closed streamlines and the stretching within each region is poor. In fact, if we let $T(\psi)$ denote the period in the streamline ψ, it is then possible to show that $\mathrm{dx}(t)$ is mapped into $\mathrm{dx}(t+T)$ at time $t+T$:

$$\mathrm{dx}(t+T) = \mathrm{dx}(t)\cdot\left[1-(\mathrm{d}T/\mathrm{d}\psi)(\nabla\psi)\mathbf{v}\right] + \text{ higher order terms in dx} \tag{18}$$

and that the orientation of the filament after n cycles of the flow is given by

$$\mathbf{m}_{t+T} = \mathbf{m}_0\left[1-(dT/d\psi)(\nabla\psi)\mathbf{v}\right]^n / \lambda, \tag{19}$$

where $\mathbf{m}_0$ is the initial orientation. As the number of cycles goes to infinity, the filament becomes aligned with the streamlines and the stretching λ becomes linear with time (Franjione and Ottino, 1991).

5. Chaos in Area-Preserving Flows

The most understood case of chaotic advection corresponds to area-preserving flows. The understanding of this case resides in knowing something about the periodic points of the flow and their associated manifolds. Let us review briefly some of the main concepts. Given a flow $\mathbf{x} = \phi(\mathbf{X},t)$, $\mathbf{P}$ is a *fixed point* of the flow if

$$\mathbf{P} = \phi(\mathbf{P},T) \tag{20}$$

for all time t (i.e., the particle located at the position $\mathbf{P}$ stays at $\mathbf{P}$). On the other hand, the point $\mathbf{P}$ is periodic, of period T, if

$$\mathbf{P} = \phi(\mathbf{P},nT) \tag{21}$$

for $n = 1, 2, 3,...$ but not for any $t<T$. That is, the material particle that happened to be at the position $\mathbf{P}$ at time $t=0$ will be located in exactly the same spatial position after a time nT [it could be anywhere for $nT<t<(n+1)T$]. Similar definitions apply to a period-p points (for example, a period-2 point returns to $\mathbf{P}$ for $n = 2, 4, 6,...$). It is important to stress that the concept of periodicity depends on the frame of reference. Thus, for example, there are periodic points in a moving frame in the cat-eyes portrait in a shear flow, but there are none in a fixed frame (see Ottino, 1989a; Shariff et al., 1991). Periodic points can be classified as hyperbolic, elliptic, or parabolic, according to the deformation of the fluid in the neighborhood of the periodic point (the parabolic case being degenerate). The character of

the flow in the neighborhood of the periodic point is given by the eigenvalues of the linearized mapping:

$$D\phi(\mathbf{P},T)\cdot\xi_k = \lambda_k\xi_k \tag{22}$$

where D denotes the operation $\partial(\cdot)_i/\partial X_j$. According to the value of the eigenvalues λ_k, the point **P** is called hyperbolic, elliptic, or parabolic:

$$\text{Hyperbolic } |\lambda_1| > 1 > |\lambda_2|,\ \lambda_1\lambda_2 = 1, \tag{23a}$$

$$\text{Elliptic } |\lambda_k| = 1\ (k = 1,\ 2) \text{ but } \lambda_k \neq 1, \tag{23b}$$

$$\text{Parabolic } \lambda_k = \pm 1\ (k = 1,\ 2). \tag{23c}$$

The net motion in the neighborhood of an elliptic periodic point is rotation; the motion in the neighborhood of hyperbolic point is contraction in one direction and stretching in another.

Hyperbolic points have associated invariant regions of inflow and outflow called the stable $[W^s(\mathbf{P})]$ and unstable $[W^u(\mathbf{P})]$ manifolds:

$$W^S(\mathbf{P}) = \{\text{all } \mathbf{X} \in \mathrm{R}^2 \text{ s.t. } \phi_t(\mathbf{X}) \to \mathbf{P} \text{ as } t \to \infty\} \tag{24a}$$

$$W^U(\mathbf{P}) = \{\text{all } \mathbf{X} \in \mathrm{R}^2 \text{ s.t. } \phi_t(\mathbf{X}) \to \mathbf{P} \text{ as } t \to -\infty\} \tag{24b}$$

Fluid particles leave the neighborhood of **P** through $W^u(\mathbf{P})$ and get back to **P** via $W^s(\mathbf{P})$. Physically, the unstable manifold corresponds to a streakline injected at the periodic point. By definition, the sets $W^s(\mathbf{P})$ and $W^u(\mathbf{P})$ are invariant; a particle belonging to one of the sets does so permanently and cannot escape from it. In bounded steady flows, the outflow $W^u(\mathbf{P})$ joins smoothly into the inflow $W^s(\mathbf{P})$; in this case nothing interesting happens.

In time-periodic flows the manifolds might intersect non-tangentially. A point belonging simultaneously to both the stable and unstable manifolds of two different fixed (or periodic) points **P** and **Q** is called a *transverse heteroclinic point.* If **P=Q** the point is called homoclinic; if $\mathbf{P} \neq \mathbf{Q}$ the point is called heteroclinic. One intersection implies infinitely many and sensitivity to initial conditions. The sensitivity to initial conditions, or exponential divergence of initial conditions, is measured by means of Lyapunov exponents. The Lyapunov exponent is the long-time average of the specific rate of stretching, $D\ln\lambda / Dt$.

$$\sigma_i(\mathbf{X},\mathbf{M}_i) \equiv \lim_{t\to\infty}\left\{\frac{1}{t}\int_0^t \left(\frac{D\ln\lambda}{Dt}\right)dt'\right\} = \lim_{t\to\infty}\left\{\frac{1}{t}\ln\lambda(\mathbf{X},\mathbf{M}_i,t)\right\} \tag{25}$$

Thus, the average stretching efficiency can be interpreted as a normalized Lyapunov exponent [with respect to $(\mathbf{D}:\mathbf{D})^{1/2}$].

An important, and often misunderstood, distinction should be made between fixed points of velocity fields and maps. Given a flow $\mathbf{x} = \phi(\mathbf{X},t)$, $\mathbf{P}$ is a *fixed point* of the flow if

$$\mathbf{P} = \phi(\mathbf{P},t) \tag{26}$$

for all time t (i.e., the particle located at the position $\mathbf{P}$ stays at $\mathbf{P}$); equivalently $\mathbf{v}(\mathbf{P},t)=\mathbf{0}$ for all t. A critical point, on the other hand, corresponds to locations such that $\mathbf{v}(\mathbf{P},t)=\mathbf{0}$ at some time t. Fixed and critical points corresponding to isochoric two-dimensional flows can be hyperbolic or saddle type, elliptic, or parabolic; the character of the fixed point can be obtained by linearizing the *velocity field* (as opposed to the motion) near $\mathbf{P}$. There is a key difference between periodic points and critical points. A periodic point *is* a material point; a critical point is not. Thus, if one were able to place a labeled fluid particle at any arbitrary time on a periodic point the particle will faithfully record the motion of the periodic point for all times. Such a thought experiment is not possible with a critical point. A critical point might appear or disappear according to when the flow is looked at; a periodic point cannot possibly disappear. This is a point that often escapes people interested in visualizing flows. An estimation of the mixing abilities of flows based on streamline portraits can be misleading. This has been pointed out in the past (Hama, 1962), but is worth repeating, primarily when viewed in the context of what happens in two-dimensional time periodic chaotic flows.

A final comment should be made about periodic points. It often happens that the simple prototypical chaotic systems studied in the context of chaotic advection present symmetry properties. Mathematically, two maps $\mathbf{A}$ and $\mathbf{B}$ are said to be symmetric to each other if there exists a transformation $\mathbf{S}$ such that

$$\mathbf{B} = \mathbf{SAS}^{-1} \tag{27}$$

If $\mathbf{A} = \mathbf{B}$, the symmetry is termed ordinary; if $\mathbf{A}^{-1} = \mathbf{B}$, the symmetry is termed time-reversal. In general, $\mathbf{S}$ can be a rotational symmetry or reflectional symmetry. An important consequence of this is that if a map possesses symmetry, the periodic points are found in symmetric arrangements.

6. Statistical Tools

The bulk of the systems studied, to date, in chaotic advection are deterministic, exceptions being attempts to introduce molecular diffusion into the description given by equation (2) or random forcing instead of periodic forcing. However, to the extent that outcomes are chaotic, statistical tools provide useful guidance in the analysis of various systems. A particularly useful tool is single-parameter scaling, which sits somehow in the broader context of multiplicative processes (Redner, 1990). A simple explanation of the main facts can be put forth in terms of stretching.

Consider a large number of points—each with an associated vector $\delta\mathbf{X}$—advected by a time periodic flow. Let $dN(\lambda)$ be the number of points with stretching between λ and $\lambda + d\lambda$. The probability of a point having a stretching λ after n periods is $F_n(\lambda) = dN(\lambda)/d\lambda$; similarly $H_n(\log\ \lambda) = dN(\log\ \lambda)/d(\log\ \lambda)$; the distributions $F_n(\lambda)$ and $H_n(\log\ \lambda)$ are related by $H_n(\log\ \lambda) = \lambda F_n(\lambda)$. Such distributions may be analyzed by single parameter scaling.

The main idea is the following. A distribution, $G(\cdot)$, is said to have single-parameter self-similarity if under a transformation of variables

$$x \rightarrow y = x / X(n), \tag{28}$$

$$G_n(x) \rightarrow \mathcal{G}(y) = K(n) \cdot G_n(x), \tag{29}$$

the function $\mathcal{G}(y)$ becomes (asymptotically) independent of n; $X(n)$ can be obtained as the ratio of two successive convergent moments, $X(n) = m_i/m_{i-1}$ where m_i is given by

$$m_i(n) = \int_0^\infty x^i G_n(x) dx, \tag{30}$$

whereas $K(n)$ is given by

$$K(n) = C_1 X(n)^2 / m_1(n) \tag{31}$$

where C_1 is a constant. It is apparent that this technique allows for the computation of the evolution of the moments of the distribution (Muzzio et al., 1991).

Another potentially useful technique is multifractal scaling. The most fruitful application of this concept in fluid mechanics, so far, has been in the context of turbulence (Sreenivasan, 1991). The explanation, again, is in terms of stretching. Consider the field of $\lambda(\mathbf{x},t)$, corresponding to a very large number of initial conditions $\mathbf{X}$ distributed in a domain V. Divide V into boxes of equal size r and label each box by an index i. The measure $\mu_r(i)$ is

the amount of λ in the i^{th} box, of volume V_r, normalized by the total amount of λ in all the boxes:

$$\mu_r(i) = \left[\int_{V_{r(i)}} \lambda(\mathbf{x})dV\right] / \left[\sum_i \int_{V_{r(i)}} \lambda(\mathbf{x})dV\right]. \tag{32}$$

In turn the measure $\mu_r(i)$ can be used to define the strength $\alpha(i)$ as

$$\mu_r(i) \sim r^{\alpha(i)},\ \alpha(i) = \log(\mu_r(i)) / \log(r). \tag{33}$$

Multifractal behavior corresponds to the case where the probability density function of α exhibits self-similar behavior over a range of length scales r. This implies that the number of boxes $N_r(\alpha)$ where α has values in a range between α and $\alpha + d\alpha$ can be expressed in terms of an invariant function $f(\alpha)$, according to

$$N_r(\alpha)d\alpha \sim r^{f(\alpha)}d\alpha. \tag{34}$$

$f(\alpha)$ is called the multifractal spectrum. The use of multifractal concepts in chaotic advection is discussed in Muzzio et al. (1992).

7. Systems Studied

It might be argued that the typical systems studied, to date, in the context of chaotic advection are unrealistic—and hence irrelevant—for an oceanographic viewpoint. That would be a mistake. The proper way to understand these examples is not as faithful representations of real systems but rather as analyzable prototypes yielding physical insight and increased basic knowledge. They act, in short, as a sort of yardstick with respect to which we can measure the understanding of realistic advection problems. Undoubtedly there are situations, such as tidal systems, that are well suited for immediate applications (Ridderinkhof and Zimmerman, 1992). Applications to more complex systems still lie in the future.

Possibly the simplest systems are the tendril-whorl flow (Khakhar et al., 1986) and the egg-beater flow (Franjione and Ottino, 1992). The tendril-whorl flow is a discontinuous succession of extensional flows and twist maps. The physical motivation for this flow is that, locally, any velocity field can be decomposed into extension and rotation. The egg-beater flow on the other hand can be seen as a flow occurring in a square region of observation periodically invaded by shear flows entering at right angles from each other. The first shear flow acts in a "horizontal" direction:

$$x_{n+1} = x_n + Tv(y_n) \tag{35a}$$

$$y_{n+1} = y_n, \tag{35b}$$

where T is the duration of the flow. This flow is written as $\mathbf{x}_{n+1} = \mathbf{H}\mathbf{x}_n$, where $\mathbf{x}=(x,y)$. The second flow acts in a "vertical" direction:

$$x_{n+1} = x_n \tag{36a}$$

$$y_{n+1} = y_n + Tv(x_n) \tag{36b}$$

and is written as $\mathbf{x}_{n+1} = \mathbf{V}\mathbf{x}_n$. The flow occurs in a domain which is periodic in both the x and y directions. The overall mapping may be written as the composition of both maps, i.e.,

$$\mathbf{x}_{n+1} = \mathbf{V}\mathbf{H}\mathbf{x}_n = \mathbf{E}\mathbf{x}_n. \tag{37}$$

A sequence of actions of the horizontal and vertical components, **H** and **V**, is denoted as **VHVHVH...** and is an example of a *mixing protocol.*

The next simplest, but historically, the first flow analyzed in the context of chaotic advection, is the blinking-vortex flow (Aref, 1984; Khakhar et al., 1986) which consists of two corotating fixed point vortices that blink on and off periodically with a constant period T. At any given time, only one of the vortices is on, so that the motion is made up of consecutive twist maps about different centers.

All these flows are computational. There are several experimentally realizable flows though, mostly two-dimensional, although a couple of experiments have been carried out in three dimensional flows as well. The first example of a two-dimensional flow is the cavity flow (Chien et al., 1986; Leong and Ottino, 1989). The cavity flow consists of a rectangular region capable of producing a two-dimensional velocity field in the x-y plane. Two opposing walls can be moved in a steady- or time-dependent manner inducing circulation within the cavity with one of multiple cells according to the aspect ratio of the cavity and the mode of operation of the walls. Several new studies are focusing on transport away for open cavities (Jana and Ottino, 1992) as well as systems involving one or two cylinders rotating in a circular containers. The two cylinder case is the so-called journal bearing flow (Chaiken et al., 1987; Swanson and Ottino, 1990). Only a few studies have been reported for three-dimensional flows (Kusch and Ottino, 1992).

There are several insights that have been gained in terms of these flows. The first insight is that passive structures in time-periodic flows evolve in an iterative fashion; an entire structure is mapped into a new structure with persistent large-scale features, but finer and finer scale features are revealed at each period of the flow. Thin striations are produced at

the expense of thicker ones, and length scales (characterized by the first moment of a striation thicknesses distribution) decrease exponentially in time. The length stretch and striations thicknesses are inversely related. It has also been found that islands form coherent regions that translate, stretch, and contract periodically and undergo a net rotation, preserving their identity. Islands display symmetry at regular intervals of time. Island symmetry is caused by symmetric placement of elliptic points. The flow within islands is weakly rotational, the stretching is linear, and the rates of rotation are usually much slower than in the rest of the flow. Rotation notwithstanding, it would be a gross mistake to identify these coherent regions as regions of vorticity.

Another insight has to do with island destruction. For example, using the case of the egg-beater flow, it is known what sequences of H's and V's lead to best mixing in a minimum number of periods. Another insight has to do with resonance and conditions leading to coupling between a base flow and a perturbation. Some simple cases admit analytical treatment. A recent example in the context of oceanography is the paper by Samelson (1992) addressing the issue of fluid exchange across a meandering jet in terms of the Melnikov method.

Another general statement that can be made regarding chaotic advection and transport is that the rate of spreading is controlled by the unstable manifolds of the hyperbolic points belonging to the lowest order periodic points. The stretching is roughly proportional to the value of the eigenvalues and is inversely proportional to the period of the point. An analysis in terms of manifolds can yield valuable information regarding the transport of material in the flow. All applications so far have been in terms of rather idealized flows. Consider, now, the application of these concepts to one of the most studied flows in fluid mechanics, but from the viewpoint of transport still a rather poorly understood flow. The flow considered is the time-periodic vortex shedding past a two-dimensional circular cylinder with diameter D placed in a stream of fluid moving with uniform speed U in x-direction. As is well known if $\mathrm{Re} = \rho UD/\mu << 1$ the flow is symmetric with respect to both the x-axis and the y-axis; as the Re increases the flow loses y-symmetry and two attached eddies form behind the cylinder which grow in size with increasing Re until, at $\mathrm{Re} \approx 40$, the flow ceases to be steady and becomes time-periodic. Experiments show that when $\mathrm{Re} \approx 100$ eddies are shed periodically from the top and bottom part of the cylinder: all the vortices originating from the top rotate in one direction; all the vortices originating from the bottom rotate in the opposite direction while the whole pattern of vortices travels downstream but with a speed smaller than U. As Re is increased above 200 or so, the flow develops three-dimensionality, time-periodicity is lost and the flow ultimately produces a turbulent wake. The case of interest here is in the range $100<\mathrm{Re}<200$. Streakline experiments produce the so-called von Kármán wake, something that has been known for eight decades or so. However, an understanding on how transport proceeds in this flow, i.e., how parcels of fluids move from one place to another and entrain material, is still far from being clear. Instantaneous streamlines offer only partial help, even though most of the recent attempts at explaining this topic address the problem from this

viewpoint (Perry et al., 1982). An analysis in terms of manifolds improves the picture considerably (Shariff et al., 1991).

Such an analysis relies on the identification of two classes of periodic points: (i) parabolic periodic points associated with separating and attaching streamlines, which produce unstable and stable manifolds that are associated with zero wall shear, parabolic points, and (ii) a period-one hyperbolic point located in the wake itself. Such points move as function of time and return to their original location after one full period. The typical picture is as follows: the streamline corresponding to one of the separation points joins smoothly with an attachment point to form a separation bubble whereas another separation streamline goes into the wake of the flow. Stable and unstable manifolds produce heteroclinic and homoclinic intersections. In this particular flow, four types of transversal intersections are possible: heteroclinic intersections are produced by intersections of stable manifolds of the period-one hyperbolic with unstable manifolds of the periodic points attached to the cylinder as well as by unstable manifolds of the hyperbolic point intersecting the stable manifolds attached to the surface of the cylinder, homoclinic intersections are produced by crossings of stable and unstable manifolds belonging to the hyperbolic point as well as those of parabolic points attached to the cylinder. The complete manifold picture of the system is, however, more complex since there are additional period-one hyperbolic points close to the surface of the cylinder as well as higher order periodic points; they however, seem to contribute much less to the gross aspects of the transport in the flow. The manifold structure results in the qualitative picture shown in Figure 3. Figure 4 shows computed pictures corresponding to Re=180. For the sake of clarity Figure 4a shows only the manifold structure corresponding to the upper wake; whereas Figure 4b shows the manifold structure corresponding to the lower wake.

The manifold structure provides a template for stretching and transport and provides a qualitative picture for the stretching and folding of a streakline in the wake. More and more details are revealed as the system evolves. This sort of iterative process has implications for the distribution of stretching within the flow. This is particularly clear in the case of time periodic flows. In this case the stretching between period 0 and period n, $\lambda_{0,n}$, can be expressed as a multiplication of stretching corresponding to individual periods, i.e.,

$$\lambda_{0,n} = \lambda_{0,1}\lambda_{1,2}\ldots\lambda_{n-1,n}, \tag{38}$$

where $\lambda_{i-1,i}$ is the stretching experienced in the interval i-1 to i. Moreover, due to the chaotic character of the flows, the $\lambda_{i-1,i}$'s quickly become uncorrelated. These two observations suggest that stretching can be considered as a multiplicative process with loosely correlated steps and are, therefore, an ideal situation for the application of scaling

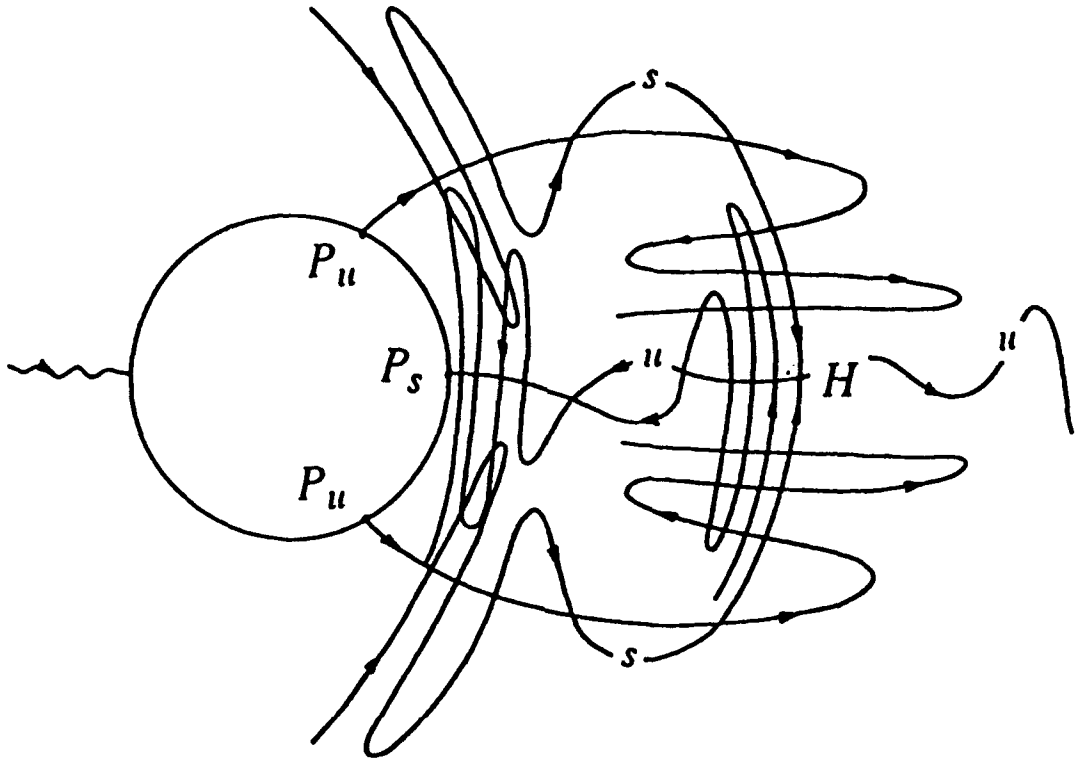


Figure 3. Qualitative picture of manifold structure in the vortex shedding regime of flow behind a circular cylinder.

concepts. The application of single parameter scaling concepts shows that as the number of periods increases beyond 5 or so a wide portion of the probability density functions of stretching overlap when re-plotted in scaled form. Closer examination of the scaled results reveals additional insight; in general, flows with islands exhibit spatial segregation with respect to stretching even within chaotic regions; one set of points wanders throughout the 'bulk of the chaotic region' and undergoes exponential stretching; the other stays close to regular islands for many periods and stretches very slowly.

Another useful tool is multifractals. The simplest application of multifractal concepts to mixing arises in the case of flows with no islands. In this case, the spatial distribution of stretching is well described by multifractal scaling if the very high tail of the distribution of stretchings is neglected. Moreover, different methods for obtaining the multifractal spectrum $f(\alpha)$ agree reasonably well, producing a time-independent self-similar distribution. For flows with islands (e.g., the flow between eccentric cylinders), the spectrum $f(\alpha)$ is time-dependent and therefore, it is not self-similar). However,

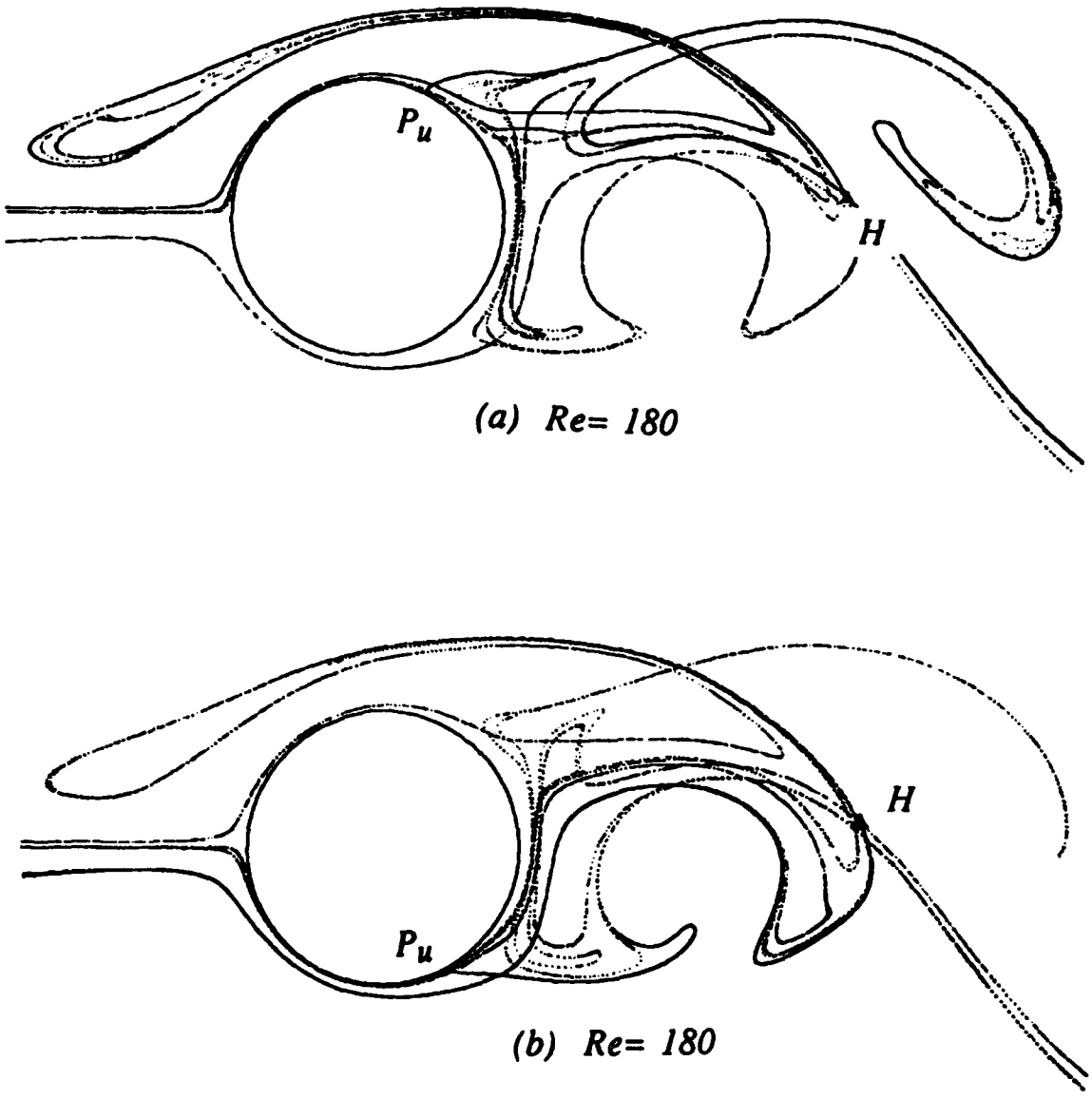


Figure 4. Intersection between unstable manifolds associated with parabolic points attached to the cylinder, P_u, and stable manifolds associated with a periodic hyperbolic point, H, at Reynolds number 180: (a) represents the manifold structure corresponding to the upper wake, (b) the represents the manifold structure corresponding to the lower wake.

multifractal concepts suggest a single-parameter scaling for the distribution of Lyapunov exponents that works well for flows without islands (Muzzio et al., 1992). A possible point of confluence of scaling concepts, multifractal descriptions and transport is in the interpretation and prediction of dispersion of passive scalars. Some work has been done (Pasmanter, 1988), but it is obvious that much more remains to be done.

8. Conclusions

Some familiarity with chaotic advection appears to be a necessary ingredient in developing an understanding of mixing and dispersion in complex flows. It is apparent that the current kinematical vocabulary necessary to deal with stirring and mixing needs to be amplified. Chaotic advection clearly demonstrates the pitfalls of flow visualization in terms of velocity field information such as instantaneous streamlines and particle paths; both can be relatively simple and streaklines extremely complex. Concepts such as periodic points and manifolds seem both useful and necessary in interpreting issues involving coherence and transport. The flow within coherent islands in two-dimensional chaotic flows is weakly rotational (in the sense that there is a net twist) but that rotation notwithstanding, it would be a gross mistake to identify these coherent regions as regions of vorticity.

Advances, to date, are mostly in the form of physical insight and basic knowledge obtained in terms of computational and experimental studies in simple flows. Currently available results can be used in two different ways: (i) to make qualitative predictions regarding the behavior or more complex systems, (ii) as a yardstick with respect to which we can measure the understanding (or lack thereof) of such problems. Most studies are for two-dimensional flows but attempts at extending analyses to three-dimensional cases are currently underway. However, many problems of interest in ocean sciences are inherently two-dimensional. The most obvious example might involve lateral mixing descriptions in terms of large circulation models. Other problems can be encountered at smaller scales. Examples might include stirring in tidal systems, an inherently time periodic case, transport and entrainment in meandering jets, and penetrative convection under ice shelves.

Acknowledgments
I would like to thank Peter Müller and the organizers of the 'Aha Huliko'a Hawaiian Workshop for an enjoyable conference and for the unique opportunity to interact with such a diverse audience.

REFERENCES

Aref, H. (1984). Stirring by chaotic advection. *J. Fluid Mech., 143*, 1-21.

Aref, H. (1991). Stochastic particle motion in laminar flows. *Physics Fluids A, 3*, 1009-1016.

Chaiken, J., Chevray, R., Tabor, M., and Tan, Q.M. (1986). Experimental study of Lagrangian turbulence in Stokes flow. *Proc. Roy. Soc. London*, A408: 165-74.

Chien, W.-L., Rising, H., and Ottino, J.M. (1986). Laminar mixing and chaotic mixing in several cavity flows. *J. Fluid Mech., 170*, 355-77.

Doherty, M. F., and Ottino, J. M. (1988). Chaos in deterministic systems: strange attractors, turbulence, and applications in chemical engineering. *Chem. Eng. Sci.*, *43*, 139–183.

Franjione, J.G. and Ottino, J.M. (1991). Stretching in Duct Flows. *Phys. Fluids A.*, 3, 2819-2821.

Franjione, J.G. and Ottino, J.M. (1992). Symmetry Concepts for the Geometric Analysis of Mixing Flows. *Phil. Trans. Roy. Soc. Lond.*, *338*, 301-323

Franjione, J.G., Leong, C.W. and Ottino, J.M. (1989). Symmetries within Chaos: a Route to Effective Mixing. *Phys. Fluids A.*, *1*, 1772-1783

Guckenheimer, J., and Holmes, P. (1983). Nonlinear oscillations, dynamical systems, and bifurcations of vector fields. New York: Springer–Verlag, 1983.

Hama, F.R. (1962). Streaklines in a perturbed shear flow. *Phys. Fluids*, *5*: 644-50.

Hao, B.-L. (1984). *Chaos*. World Scientific.

Jana, S.C., and Ottino, J.M. (1992). Chaos-Enhanced Transport in Cellular Flows. *Proc. Roy. Soc. London A.*, *338*, 519-532.

Khakhar, D.V., Rising, H., and Ottino, J.M. (1986). An analysis of chaotic mixing in two chaotic flows. *J. Fluid Mech.*, *172*, 419-51.

Kusch, H.A., and Ottino, J.M. (1992). Experiments on Mixing in Continuous Chaotic Flows, *J. Fluid Mech.*, *236*, 319-348.

Leong, C.-W. and Ottino, J.M. (1989). Experiments on mixing due to chaotic advection in a cavity. *J. Fluid Mech.*, *209*, 463-499.

Lorenz, E.N. (1962). Deterministic nonperiodic flow. *J. Atmos. Sci.*, *20*, 130-141.

Muzzio, F.J., Meneveau, C., Swanson, P.D., and Ottino, J.M. (1992). Scaling and Multifractal Properties of Mixing in Chaotic Flows. *Phys. Fluids A*, *4*, 1439-1456.

Muzzio, F.J., Swanson, P.D., and Ottino, J.M. (1991). The Statistics of Stretching and Stirring in Chaotic Flows. *Phys. Fluids A, 5*, 822-834.

Ottino, J.M. (1989a). The kinematics of mixing: stretching, chaos, and transport, Cambridge: Cambridge University Press (reprinted 1990).

Ottino, J.M. (1989b). The mixing of fluids. *Scientific American*, *260*, 56-67.

Packard, N.H., Crutchfield, J.D., Farmer, J.D., and Shaw, R.S. (1980). Geometry from a time series. *Phys. Rev. Lett.*, *45*, 712-716.

Parker, T.S., and Chua, L.O. (1989). Practical Numerical Algorithms for Chaotic Systems. (Springer, Berlin, Heidelberg, New York).

Pasmanter, R. (1988). Anomalous diffusion and anomalous stretching in vortical flows. *Fluid Mech. Res.*, 320-26.

Perry, A.E., Chong, M.S., and Lim, T.T. (1982). The vortex-shedding process behind two-dimensional bluff bodies. *J. Fluid Mech.*, *116*, 77-90.

Redner, S. (1990). Random multiplicative processes: an elementary tutorial. *Amer. J. Phys.*, *58*, 267-273.

Ridderinkhof, H., and Zimmerman, J.T.F. (1992). Chaotic stirring in a tidal system. *Science*, *258*, 1107-1111.

Samelson, R.M. (1992). Fluid exchange across a meandering jet. *J. Phys. Oceanography*, *22*, 431-440.

Shariff, K., Pulliam, T.H., and Ottino, J.M. (1991). Dynamical Systems Analysis of Kinematics in the Time-Periodic wake of a Circular Cylinder, in Vortex Dynamics and Vortex Methods, Editors C. Anderson and C. Greengard, Lectures in Applied Mathematics. *American Mathematical Society*, *28*, 613-646.

Sigeti, D. and Horsthemke, W. (1991). High-frequency power spectra for systems subject to noise. *Phys. Rev. A*, *35*, 2276-2282.

Sreenivasan, K.R. (1991). Fractals and multifractals in turbulence. *Ann. Revs. Fluid Mech.*, *23*, 539-600.

Swanson, P.D. and Ottino, J.M. (1990). A Comparative Computational and Experimental Study of Chaotic Mixing of Viscous Fluids. *J. Fluid Mech.*, *213*, 227-249.

Tsonis, A.A. and Elsner, J.B. (1989). Chaos, strange attractors, and weather. *Bull. Amer. Meteor. Soc.*, *70*, 14-23.

Vallis, G.K. (1986). El Niño: a chaotic dynamical system? *Science*, *232*, 243-245.

Wiggins, S. (1991). Global bifurcations and chaos: analytical tools, New York: Springer.

Wiin-Nielson, A. (1992). Comparisons of low-order atmospheric dynamic systems. *Atmosphera*, *5*, 135-155.

Wolf, A., Swift, J.B., Swinney, H.L., and Vastano J.A. (1985). Determining Lyapunov exponents from a time series. *Physics D*, *16*, 285-317.

Yang, P. (1991). On the chaotic behavior and predictability of the real atmosphere. *Adv. in Atmos. Sci.*, *8*, 407-420.

Zeng, X., Pielke, R.A., and Eykholt, R. (1993). Chaos theory and applications to the atmosphere. *Bull. Amer. Meteor. Soc.*, *74*, 631-644.

CHAOS IN OCEAN PHYSICS

Michael G. Brown
RSMAS, University of Miami, Miami FL 33149

ABSTRACT

Three topics relating to chaotic ocean physics are discussed. These are (1) low order El Niño dynamics, (2) lateral stirring processes, and (3) linear ocean waves in the geometric limit. Each topic is discussed separately; emphasis, in each case, is given to the manner in which ideas associated with chaos and low-order dynamical systems complement more traditional approaches to the same problem.

INTRODUCTION

In this paper three topics relating to chaotic ocean physics are discussed. This list is not intended to be an exhaustive list of topics in ocean physics to which ideas relating to chaos can be applied. Our discussion of these three topics—which were chosen because the author has some familiarity with them—serves to illustrate several important concepts likely to be useful in other oceanographic applications as well. It is our feeling that the ideas relating to chaotic dynamical systems discussed in this paper are useful but must be applied in a sober fashion which complements more traditional approaches. When properly applied, these ideas provide a vehicle to increase our understanding of various physical processes in the ocean in an evolutionary fashion. Expectations of gaining new insight of a revolutionary nature are not likely to be realized.

In each of the three sections that follow, we discuss a topic in ocean physics (low-order El Niño dynamics, lateral stirring processes, linear ocean waves in the geometric limit) to which ideas associated with chaos can be applied. Background material relating to dynamical systems and chaotic dynamics is introduced as necessary in the context of the problems treated. This approach is natural inasmuch as our intention is not to provide a tutorial on chaos; instead, we seek to demonstrate that these ideas are useful in the context of specific problems in ocean physics. All three topics discussed in this paper are treated in more detail elsewhere; references are provided below. So as not to duplicate this material, we focus here on the rationale for applying ideas relating to chaos and low-order dynamical systems. Stated somewhat differently, in this paper we focus more on the questions being addressed than on details of the subsequent analysis. Some unifying comments and observations concerning chaotic ocean physics are included in the final section.

LOW-ORDER EL NIÑO DYNAMICS

The El Niño/Southern Oscillation (ENSO) system is a quasi-periodic oscillation of the tropical Pacific Ocean and overlying atmosphere (see, e.g., Enfield, 1989). The ENSO system involves interactions among eastern basin sea surface temperature (SST), zonal trade winds and the thermocline depth (Bjerknes, 1969; Wyrtki, 1975). El Niño events—characterized by anomalously high eastern basin SST, weak trade winds, and a shallow western basin thermocline—are separated by three to five years, typically.

Models of the ENSO system vary considerably in complexity. At one extreme are coupled ocean-atmosphere general circulation models (see, e.g., Neelin, 1990). That such models produce ENSO-like behavior should come as no surprise; ENSO behavior is surely contained in the complicated coupled equations of motion/state which were numerically solved. It is our feeling that simpler models—provided they adequately reproduce essential features of the system being modeled—are more insightful inasmuch as they better elucidate the essential physical processes involved. This leads naturally to the question of whether the essential physics of the ENSO system can be captured in simpler models.

The simplest type of model of the ENSO system which has been proposed consists of a small number (n, say) of autonomous ordinary differential equations,

$$\frac{d\underline{x}}{dt} = \underline{f}(\underline{x}). \tag{1}$$

The solution $\underline{x}(t)$ of these equations describes the temporal evolution of the system. The x_i's ($i = 1,2,...n$) in such a model would include variables such as anomalies of eastern basin SST, zonal winds and western basin thermocline depth. Vallis (1986) was the first to propose a model of the ENSO system of this type. That this model produces unphysical behavior for some choices of parameters (see, e.g., Vallis, 1988) is, in our opinion, not terribly important: the significance of the Vallis (1986) paper is the suggestion that the essential physics of the ENSO system can be captured in severely truncated physical model consisting of a low-order, autonomous dynamical system. The word autonomous means that the function f in (1) does not depend explicitly on time; physically, this restriction means that any quasi-oscillatory behavior in $x_i(t)$—which might be associated with the occurrence of El Niño events—is the result of internal, self-sustained dynamical processes rather than being the response to external stochastic forcing. More recently, improved low-order models (autonomous dynamical systems) of the ENSO system have been proposed by Schopf and Suarez (1988) and Münnich et al. (1991). Before proceeding, it is worth noting that the notion of simple ENSO dynamics—during the growth phase of El Niño events, at least—is generally accepted and dates back to the seminal work of Bjerknes (1969) and Wyrtki (1975); the notion that the complete ENSO cycle—and, in particular, the triggering of El Niño events—results from

internal dynamics (i.e., that these oscillations are self-sustained) is not universally accepted.

These considerations led Bauer and Brown (1992) to address the question of whether *observations* of the ENSO system are consistent with underlying low-order dynamics. This question was addressed via the process of phase space reconstruction whereby discrete samples of a single variable, $y(t_j)$, $j = 1,2, \ldots$ (monthly samples of eastern basin SST were used in the Bauer and Brown analysis), are used to construct a discretely sampled multidimensional phase space portrait, $\underline{y}(t_k)$, $k = 1,2, \ldots$ A simple way to carry out this process is to use delay time coordinates: $y_1(t_k) = y(t_k)$, $y_2(t_k) = y(t_{k+1})$, $y_3(t_k) = y(t_{k+2})$, etc. Surprisingly, perhaps, the reconstructed, discretely sampled phase space trajectory $\underline{y}(t_k)$ constructed in this fashion can be shown (Broomhead and King, 1986), under appropriate conditions, to reproduce with only minor distortion (a diffeomorphism) the true multidimensional phase space portrait $\underline{x}(t)$ of the underlying dynamical system. Unfortunately, this procedure is sensitive to noise and therefore generally works poorly on geophysical data. The shortcoming was overcome by Bauer and Brown by using a technique developed by Broomhead and King (1986)—see also Vautard and Ghil (1989)—wherein temporal empirical orthogonal functions are used as basis functions for the reconstructed phase space trajectory. Details of this analysis will not be repeated here. The results of this analysis suggest that the underlying ENSO dynamics are approximately those of a low-order system; we urge the reader to carefully assess the evidence presented and come to his/her own conclusion.

It is worth emphasizing in this context that the question of chaotic ENSO dynamics is secondary to the question of whether ENSO dynamics are approximately those of a low-order system. If the later question is answered affirmatively, then questions concerning chaotic behavior become relevant. Among these are (1) Does the system evolve chaotically, and if so, what is the predictability timescale (reciprocal of the largest positive Lyapunov exponent)? (2) What is the dimension of the corresponding attractor? At the present time these questions are, in our opinion, premature. It is worth pointing out, however, that if the underlying dynamics are approximately those of a low-order system—even a chaotic one—this would lead to some long-term predictability in the sense that it would be known that the system's state vector $\underline{x}$ must, at all times, lie on some attractor—although its precise position may not be predictable.

LATERAL OCEAN STIRRING PROCESSES

In the ocean, many water properties such as temperature, salinity, oxygen content or pollutant concentration can be treated approximately as passive fluid parcel markers. Passive means that the flow field evolves independently of the initial distribution of the tracer. In order to understand the distribution of these oceanic tracers and how they evolve in time, one needs to understand the process by which passive tracers get redistributed. Our discussion of this process focuses on the lateral stirring (advective

tracer transport) process; we ignore the quasi-diffusive 3-d behavior that takes place at the smallest scales (internal wave and smaller).

The advective transport of a passive tracer in a two dimensional incompressible flow is described by the equation,

$$\frac{\partial \theta}{\partial t} - \frac{\partial \psi}{\partial y}\frac{\partial \theta}{\partial x} + \frac{\partial \psi}{\partial x}\frac{\partial \theta}{\partial y} = 0 \tag{2}$$

subject to the initial condition $\theta(x,y,0) = \theta_0(x,y,t)$. Here $\theta(x,y,t)$ is the tracer concentration and $\psi(x,y,t)$ is the streamfunction. It follows from (2) that θ is constant following particle trajectories, $x(t)$, $y(t)$, which satisfy

$$\frac{\partial x}{\partial t} = -\frac{\partial \psi}{\partial y}, \quad \frac{\partial y}{\partial t} = \frac{\partial \psi}{\partial x} \tag{3}$$

Thus, in order to understand the temporal evolution of $\theta(x,y,t)$—even in a statistical sense—one needs to understand the behavior of particle trajectories and understand the implications of the form of the Lagrangian equations of motion (3).

The Lagrangian equations of motion constitute a generally nonautonomous Hamiltonian system with one degree of freedom; $\psi(x,y,t)$ plays the role of the Hamiltonian $H(p,q,t)$. It is extremely important to distinguish integrable Hamiltonian systems from nonintegrable ones. For the system (3) integrability implies that there exists a single-valued function $\chi(x,y,t)$ which is constant following particle trajectories, $d\chi/dt = 0$. If the flow is steady, $\partial\psi/\partial t = 0$, then the system of equations is said to be autonomous and the streamfunction is the required constant of the motion, $d\psi/dt = 0$. This follows from equations (3). In nonsteady flows, however, the equations of motion (3) are nonautonomous and are generally nonintegrable. This observation is important inasmuch as nonintegrability is a necessary—but not sufficient—condition for chaotic motion (see, e.g., Tabor, 1989).

The distinction between chaotic and nonchaotic particle trajectories is extremely important in the context of passive tracer transport. The reason is that chaotic particle trajectories exhibit extreme sensitivity to their initial conditions. This means that neighboring particle trajectories diverge from one another at an exponential rate, on average. It follows that material lines of fluid will also grow exponentially, on average. This type of behavior leads to very efficient stirring (advective transport) of a tracer, and, in turn, enhances the mixing (diffusive transport) of the tracer at smaller scales. These ideas are discussed in more detail by Ottino (1990) (see also the contribution by Ottino in this volume) and Brown and Smith (1990, 1991). The latter publications also address the question of whether proxy ocean particle trajectories (acoustically tracked submerged SOFAR floats) exhibit extreme sensitivity. Previously, Osborne et al. (1986) had addressed this question

using satellite-tracked surface drifters. This work suggests that float/drifter trajectories do exhibit the important property of extreme sensitivity which is associated with chaotic systems.

It is important to note, however, that typical oceanographic realizations of $\psi(x,y,t)$ are significantly more complicated than the idealized systems to which notions relating to chaos are normally associated. Specifically, almost integrable systems with periodic time-dependence are fairly well understood (see, e.g., Tabor, 1989). In such systems, the onset of chaos is associated with resonances between periodic motion in the nearby integrable system and the period of the temporal variations of the streamfunction. It is not clear whether results which apply to time-periodic streamfunctions carry over to the problem where the streamfunction has more general time dependence; there remains a significant gap between the complexity of the ocean and that of the idealized systems treated in textbooks on nonlinear dynamics.

This gap in complexity offers challenges to both oceanographers and nonlinear dynamicists and provides the opportunity for the two groups work together in a mutually beneficial fashion. In fact, this has already happened. In the aforementioned work of Osborne et al. (1986), the authors argued that the fractal characteristics of drifter trajectories was attributable to underlying stochasticity (power law energy spectrum of the velocity field) rather than being associated with a strange attractor. This work led to several studies on the relationship between stochasticity and fractal behavior.

LINEAR OCEAN WAVES IN THE GEOMETRIC LIMIT

In the geometric (ray theoretical) limit, any type of linear wave motion can be described using a ray approximation (see, e.g., Lighthill, 1978). Such a description is valid when the properties of the ocean, including its boundaries, vary slowly on a scale of wavelengths. The ray equations are

$$\frac{dx_i}{dt} = \frac{\partial w}{\partial k_i}, \quad \frac{dk_i}{dt} = -\frac{\partial w}{\partial x_i} \tag{4}$$

where

$$\omega = \omega(\underline{k},\underline{x}). \tag{5}$$

Here the x_i's are position coordinates and the k_i's are the corresponding components of the wavenumber vector. The form of the function $\omega(\underline{k}, \underline{x})$—the dispersion relation—depends on the type of wave being considered. For example, for surface gravity waves propagating in water of variable depth $h(\underline{x}) = h(x,y)$,

$$\omega(\underline{k},\underline{x}) = \left[g|\underline{k}| \tanh(|\underline{k}|h(x))\right]^{1/2} \tag{6}$$

In the following, some important ideas are illustrated using this form of the dispersion relation. We emphasize, however, that equations (4) and (5) are very general and that the following considerations are applicable to any type of linear wave motion.

Equations (4) and (6) constitute an autonomous Hamiltonian system with two degrees of freedom; $\omega(\underline{k}, \underline{x})$ serves as the Hamiltonian $H(\underline{p}, \underline{q})$. Autonomous means that the Hamiltonian function does not depend explicitly on the dependent variable, time. Integrability of such a system requires that two independent constants of the motion exist. One of these is $\omega(\underline{k}, \underline{x})$; it follows from equations (4) that $d\omega/dt = 0$. Only for very special bathymetric variations $h(x,y)$ does the second required constant of the motion exist. For example, if $h = h(x)$, then it follows from the second of equations (4) that $dk_y/dt = 0$; under such conditions k_y is a second constant of the motion. Such behavior is not typical, however.

In the absence of a second constant of the motion, the possibility of chaotic ray motion exists. Numerical experiments strongly support the expectations that, under such conditions, ray trajectories exhibit chaotic behavior (see, e.g., Brown et al., 1991; Smith et al., 1992; Abdullaev and Zaslavskii, 1989). These studies, however, assume spatially periodic ocean properties. This assumption allows readily available mathematical tools to be exploited. Unfortunately, a similar set of tools is not available to treat problems involving more realistic (nonperiodic) ocean structure. Chaotic behavior, which presumably persists in some form in realistic ocean environments, is characterized by exponential growth of small errors and leads unavoidably to the conclusion that, under such conditions, predictability of ray trajectories is limited to small times.

Does this imply a lack of predictability of the corresponding wavefield? Probably not. The reason is that the ray description of the wave motion is a nonlinear approximation to a linear wave equation. (For the system described by (4) and (6) the corresponding linear wave equation is the mild slope equation—see, e.g., Mei, 1983). Because nonlinearity is a necessary condition for chaos, the linear wave equation does not admit chaotic solutions. These solutions may have different properties, however, depending on whether the corresponding ray trajectories are chaotic or not. (There is a vast literature on the corresponding quantum mechanical problem—see Reichl, 1992, for an excellent recent review.) Wavefield statistics, for example, may be very different depending on whether the corresponding ray trajectories are chaotic or not.

It should also be noted that for ocean waves the linear wave equation is itself an approximation to a nonlinear wave equation. This leads to more questions. Does the nonlinear wave equation admit chaotic solutions, and, if so, is there any connection

between this chaos and chaotic behavior in the corresponding ray trajectories? The answers are currently not known.

FINAL REMARKS

Our discussion of the topics in the three preceding sections led to a number questions relating to chaotic ocean dynamics. For the most part, the questions corresponding to the different topics were not the same. This is consistent with our view that chaotic ocean physics should not be treated as a unified branch of ocean physics. Rather, results from studies of low-order dynamical systems should be thoughtfully applied to selected problems in ocean physics in a manner which complements more traditional approaches to the same problem.

Not surprisingly, we have seen that the ocean is more complicated than the systems normally studied in the context of nonlinear dynamics. This discrepancy should be viewed as a challenge to both physical oceanographers and nonlinear dynamicists; both groups stand to benefit from collaborating. The example given earlier of Osborne et al.'s (1986) work motivating studies on the relationship between stochasticity and fractal behavior is an excellent example of precisely this type of symbiotic relationship.

Acknowledgments

Thanks to Peter Müller and Phyllis Haines for organizing a physically relaxing and intellectually stimulating workshop. The author's original work on the three topics discussed in this paper was done in collaboration with F. Tappert, K. Smith and S. Bauer. This work was supported by the Office of Naval Research and the National Science Foundation.

REFERENCES

Abdullaev, S.S., and Zaslavskii, G.M., 1989: Fractals and ray dynamics in a longitudinally inhomogeneous medium, *Sov. Phys. Acoust., 34*, 334-336.

Bauer, S.T., and M.G. Brown, 1992: Empirical low-order ENSO dynamics, *Geophys. Res. Lett., 19*, 2055-2058.

Bjerknes, J., 1969: Atmosphere teleconnections from the equatorial Pacific, *Mon. Wea. Rev., 97*, 163-172.

Broomhead, D.S., and G.P. King, 1986: Extracting qualitative dynamics from experimental data, *Physica., 20 D*, 217-235.

Brown, M.G., F.D. Tappert, and S.E.R.B. Sundaram, 1991: Chaos in small amplitude surface gravity waves over slowly varying bathymetry, *J. Fluid Mech., 227*, 35-46.

Brown, M.G., and K.B. Smith, 1990: Are SOFAR float trajectories chaotic? *J. Phys. Oceanog., 20*, 139-149.

Brown, M.G., and K. B. Smith, 1991: Ocean stirring and chaotic low-order dynamics, *Phys. Fluids A, 3*, 1186-1192.

Enfield, D.B., 1989: El Niño, past and present, *Rev. Geophys., 27*, 159-187.

Lighthill, J., 1978: *Waves in Fluids*, Cambridge University Press, Cambridge, 504 pp.

Mei, C.C., 1983: *The Applied Dynamics of Ocean Surface Waves*, Wiley-Interscience, New York, 364 pp.

Münnich, M., M.A. Cane, and S.E. Zebiak, 1991: A study of self-excited oscillations of the tropical ocean-atmosphere system. Part II: Nonlinear cases, *J. Atmos. Sci., 48*, 1238-1248.

Neelin, J.D., 1990: A hybrid coupled general circulation model for El Niño studies, *J. Atmos. Sci., 47*, 674-693.

Osborne, A.R., A.D. Kirwan, A. Provenzale and L. Bergamasco, 1986: A search for chaotic behavior in large and mesoscale motions in the Pacific Ocean. *Physica, 23 D*, 75-83.

Ottino, J.M., 1990: *The Kinematics of Mixing: Stretching, Chaos and Transport*, Cambridge University Press, Cambridge.

Reichl, L.E., 1992: *The Transition to Chaos in Conservative Classical Systems: Quantum Manifestations*, Springer, New York, 551 pp.

Schopf, P.S. and M.J. Suarez, 1988: Vacillations in a coupled ocean-atmosphere model, *J. Atmos. Sci., 45*, 549-566.

Smith, K.B., M.G. Brown and F.D. Tappert, 1992: Ray chaos in underwater acoustics, *J. Acoust. Soc. Am., 91*, 1939-1949.

Tabor, M., 1989: *Chaos and Integrability in Nonlinear Dynamics*, Wiley-Interscience, New York, 364 pp.

Vallis, G.K., 1986: El Niño: A chaotic dynamical system? *Science, 232*, 241-245.

Vallis, G.K., 1988: Conceptual models of El Niño and the southern oscillation, *J. Geophys. Res., 93*, 13,979-13,991.

Vautard, R. and M. Ghil, 1989: Singular spectrum analysis in nonlinear dynamics with applications to paleoclimatic time series, *Physica. 35 D*, 395-424.

Wyrtki, K, 1975: El Nino--the dynamic response of the equatorial Pacific Ocean to atmospheric forcing, *J. Phys. Oceanog.*, v, 572-584.

MEASUREMENTS OF CHAOS IN THE OCEAN

Everett F. Carter Jr.
Department of Oceanography, Code OC/CR
Naval Postgraduate School, Monterey CA 93943

Abstract

The theory of dissipative chaos appears to promise great insights into the behavior of natural systems like the ocean. Results based upon model simulations show the possibility that phenomena such as El Niño are chaotic. Chaotic phenomena also demonstrate that certain traditional methods are not appropriate for chaotic systems. For example a perturbation from the linear solution provides *no* insight into the behavior of the nonlinear system if that system is chaotic, even if the nonlinear terms are small. The existence of chaos implies an inherent limit to the predictability of a system, this is one reason why it is important to determine if a system is chaotic.

However, when one attempts to make estimates of measures of chaos (dimensions, Lyapunov exponents, etc.) from oceanographic *data* one is faced with the fact that the methods that quantify chaotic properties of systems from data require an enormous number of degrees of freedom for any reasonable degree of confidence. Again traditional analysis techniques can make matters worse and not better. An example of this is the use of a smoothing filter: the filter can *increase* the dimension of the resulting data set by as much as 1.

1 What chaos might contribute

There are several ways that ideas from chaotic dynamics may contribute to an understanding of the ocean. The primary question is whether or not any oceanic phenomena are chaotic.

If an oceanic phenomenon is chaotic, that will automatically impose inherent limits to the predictability of the system. If this is so, it is important to be able to quantify what the predictability limit is.

1.1 Are phenomena such as El Niño chaotic?

A first question to ask is whether any oceanic phenomena are actually driven by chaotic dynamics. The identification of chaos in the ocean would mean that the relatively complicated behavior that is observed could be described in terms of a system with a small number of degrees of freedom. This possibility that El Niño is chaotic has been investigated by looking at the available data (Fraedrich, 1988), and by model studies (Vallis, 1986).

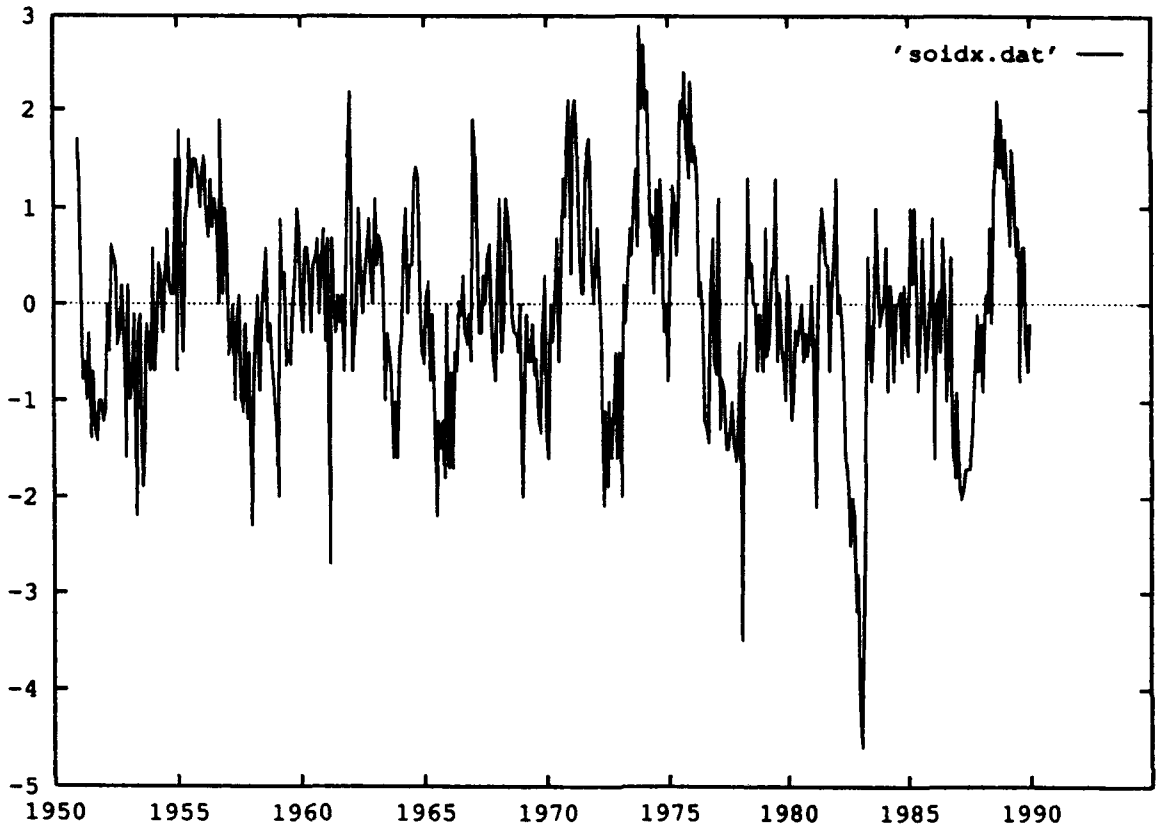


Figure 1. The Southern Oscillation Index, a monthly time series of sea level pressure differences between Tahiti and Darwin, Australia (These data are scaled to standardized dimensionless units so that the series has a zero mean and a unit standard deviation.)

Figure 1 shows the Southern Oscillation Index; its irregularity is visually reminiscent of chaotic time series. This time series has fewer than 500 data points in it, which is unfortunately too few to make reliable calculations of the dimension of the underlying system. Model studies of El Niño indicate that it is possible to mimic time series such as the Southern Oscillation Index with models that are chaotic. Figure 2 shows the Vallis (1986) model. This very simple model produces an El Niño event with about the right periodicity. The system is chaotic and has a Lyapunov dimension of 2.088 (see Fig. 3).

1.2 **Chaotic Lagrangian trajectories?**

The irregular nature of drifter trajectories is suggestive of either **turbulence** or **chaos**, (see Fig. 4). The possibility that these trajectories are fractal has been investigated by several people (Osborne, Brown and others). The major problem with these analyses is that the data records are short (typically about 1000 points), while the methods used in chaotic analysis require one or two orders of magnitude more data for confident estimates.

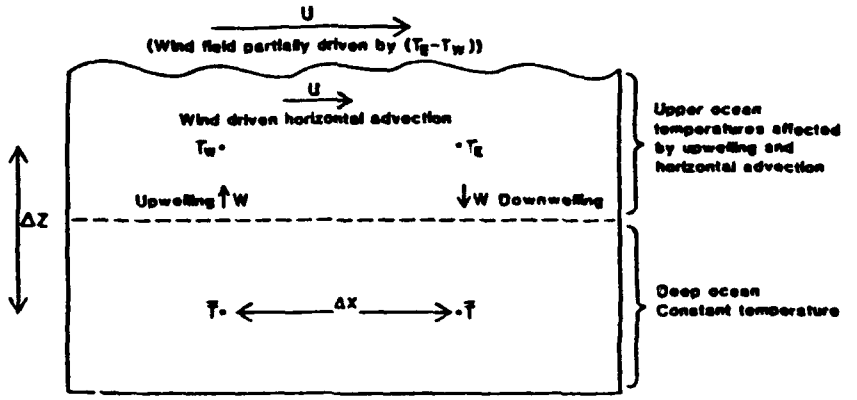
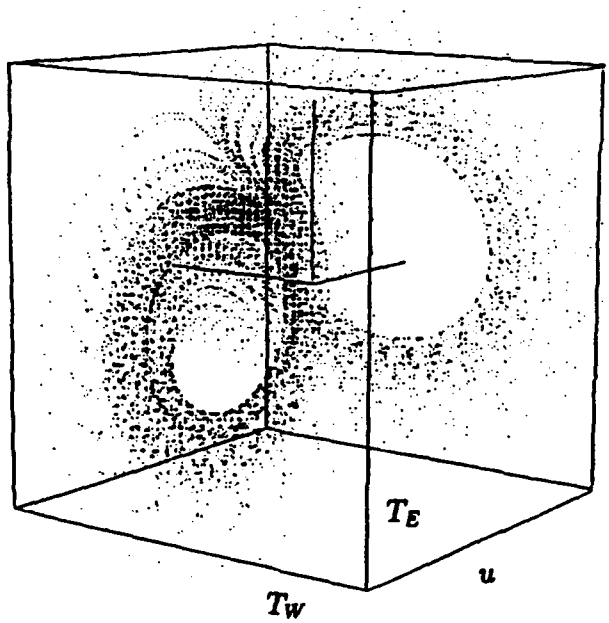


Figure 2. The Vallis (1986) ENSO model. Top: west-east section of the equatorial Pacific Ocean, defining symbols used in the model. Center: model equations. Bottom: the chaotic attractor resulting from the model equations with parameters A= 1 year^{-1} and B=2m^{-2} s^{-2} °C^{-1}.

$$\frac{du}{dt} = B(T_E - T_W)/2\Delta x - C(u - u^*)$$

$$\frac{dT_W}{dt} = \frac{u}{2\Delta x}(\bar{T} - T_E) - A(T_W - T^*)$$

$$\frac{dT_E}{dt} = \frac{u}{2\Delta x}(T_W - \bar{T}) - A(T_E - T^*)$$



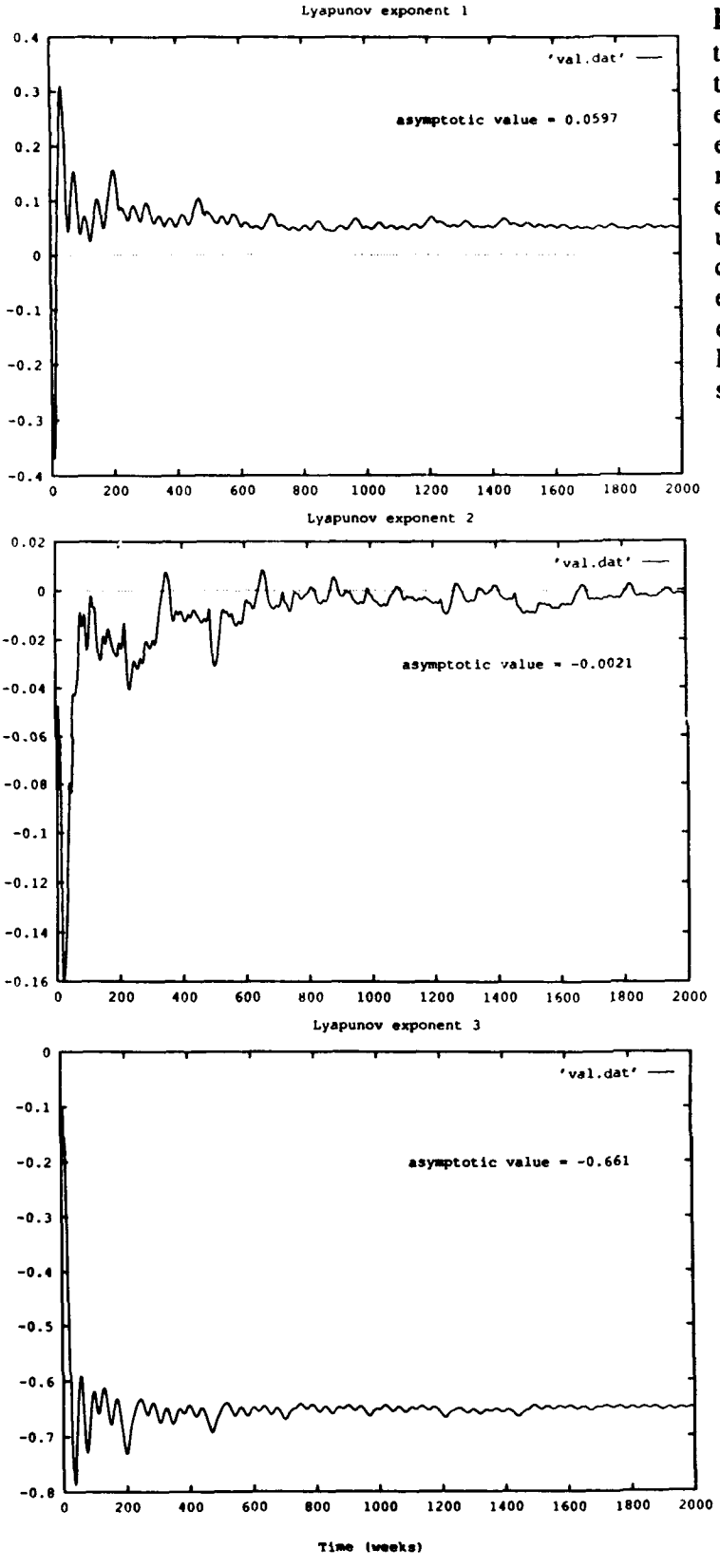


Figure 3. The Lyapunof spectrum of the Vallis attractor. The panels show the convergence of a numerical estimate of the respective Lyapunov exponents as a function of time. The noted asymptotic value is the final estimate of the exponent. The time units are nondimensional and correspond to one unit being equivalent to one week. The Lyapunov dimension (calculated using the Kaplan-Yorke equation (28)) of this system is D_Λ=2.087.

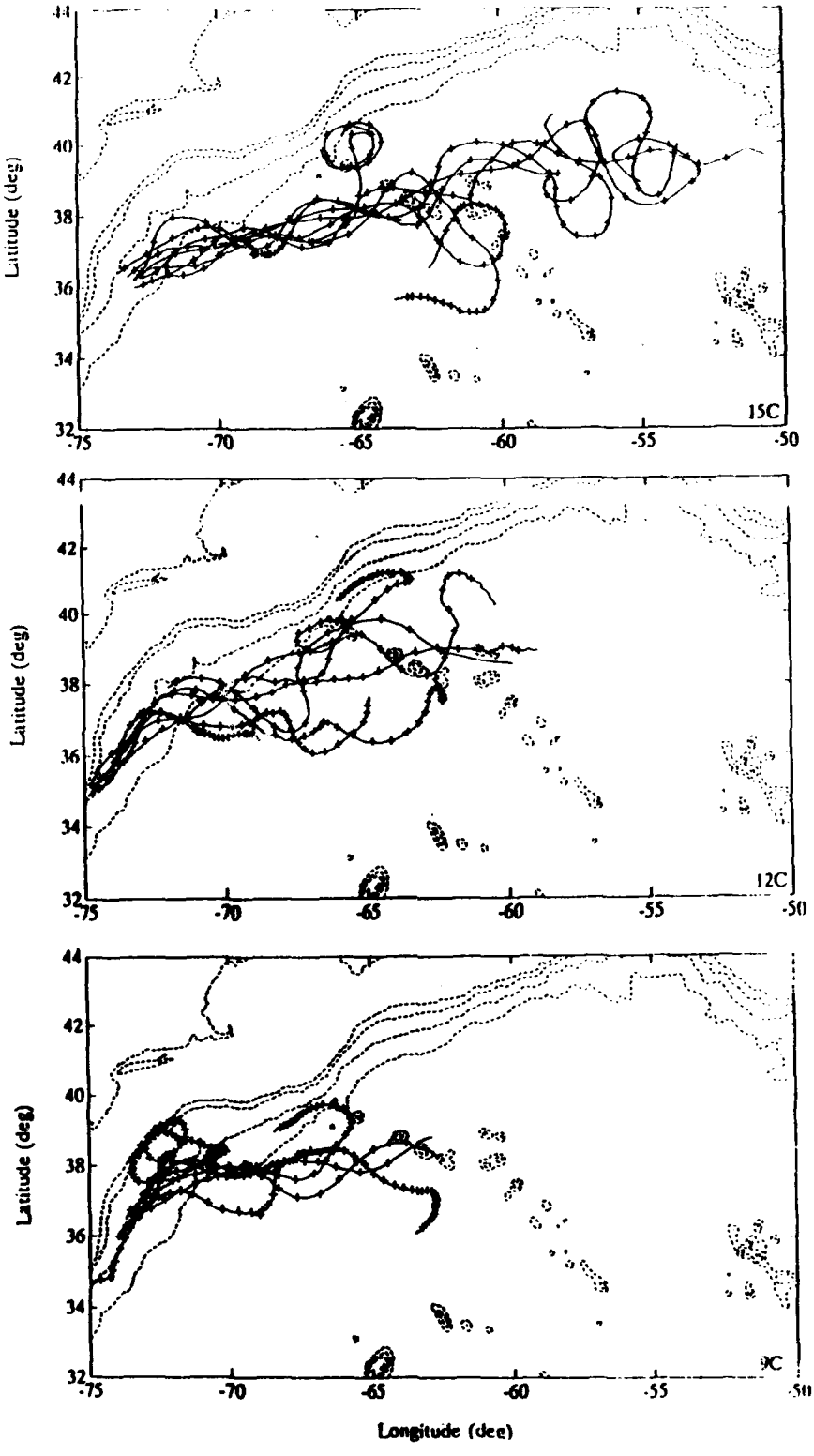


Figure 4. Complete trajectories of RAFOS floats in the Gulf Stream. The tick marks are at daily intervals; the typical float track is 45 days long. Floats in the upper panel were deployed on the 15°C surface, the middle panel at 12°C, and the lower at 9°C.

Early calculations by Osborne et al. (1986) for a year of measurements of three surface drifters indicated a correlation dimension of about 1.4. More recent calculations on SOFAR float trajectories (Brown and Smith, 1990) are more ambiguous. Based on available observations, the current conclusion is that float trajectories are probably not chaotic. They are more likely to be controlled by turbulent processes.

1.3 **Limits to predictability**

If a system is chaotic, then trajectories that are nearby in phase space will diverge **exponentially**. Increasing the accuracy of the observations does not help, since predictability only increases **linearly** with the number of digits.

Another possible situation that can impose limits on predictability is the possibility that the boundary between the states of the ocean/atmosphere is fractal. As an illustration of this possibility, consider the determination of the basins of attraction (i.e., the root that is reached for a given starting point) for the problem of finding the roots of

$$z^3 - 1 = 0$$

for complex z, by using Newton's method. Here Newton's method for this complex polynomial is the "physics" for a system which ultimately reaches one of three states. It turns out that the *boundaries* of the regions that reach a given root are fractal and have the remarkable property that any boundary point is a boundary between *all three domains*, these boundary points define a set known as a Julia set (see Fig. 5). The implication for predictability is that for measurements with a given finite error, there are some regions that are perfectly predictable and other regions where there is no predictability at all.

1.4 **Perturbation expansion of chaotic models**

One common technique in solving nonlinear systems is to do a perturbation expansion about some small parameter. We demonstrate here that a conventional perturbation expansion may *not* be helpful when the system is chaotic because the perturbation solution has no chaotic behavior.

Look at the Lorenz system of equations,

$$\dot{x} = \sigma(y - x) \tag{1}$$
$$\dot{y} = -y - xz + rx \tag{2}$$
$$\dot{z} = xy - bz. \tag{3}$$

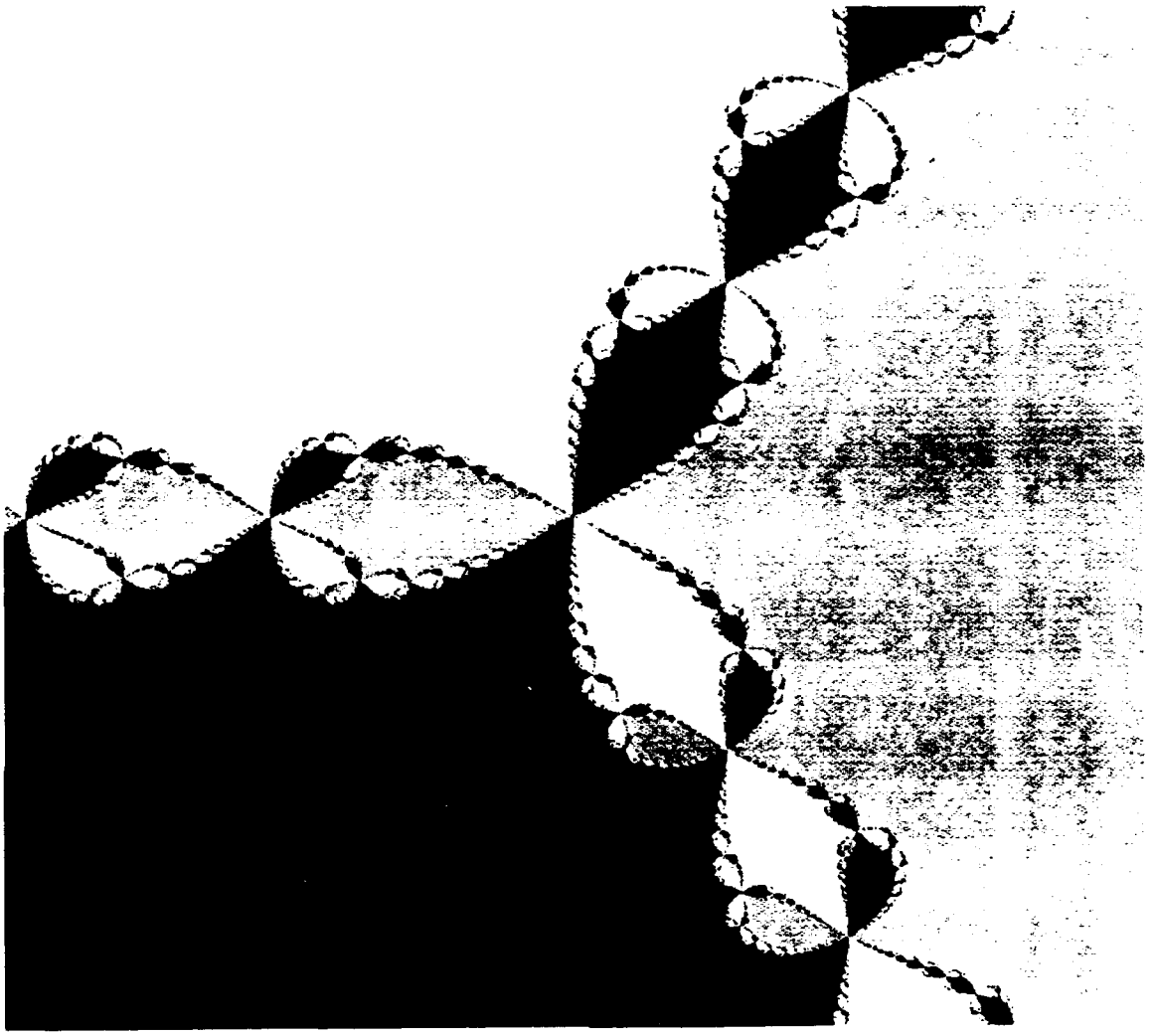


Figure 5. The basins of attraction for the roots of $z = (-1 - i\sqrt{3})/2)$, for complex z, using Newton's method. The starting points that converge to the root $z = 1$ are colored grey, points that converge to the root $z = (-1 + i\sqrt{3})/2)$, and points that converge to the root $z = (-1 - i\sqrt{3})/2)$ are black. (The center of the figure is at the origin.)

The parameter r is the ratio of the Rayleigh number divided by the critical Rayleigh number. The parameter σ is the Prandtl number. The third parameter b is related to the horizontal wave number of the system. Typical values, $r = 28$, $o = 10$, $b = 8/3$, dimension = 2.05. A common second set of values, $r = 45.92$, $\sigma = 16$, $b = 4$, dimension = 2.067.

The interesting cases are where the Rayleigh number ratio r is large, which suggests that we could expand the system of equations around a parameter proportional to the reciprocal of r (which would be small).

If we define

$$\varepsilon = r^{-\frac{1}{2}} \tag{4}$$

and let

$$\begin{aligned} x' &= \varepsilon x \\ y' &= \varepsilon^2 \sigma y \\ z' &= \sigma(\varepsilon^2 z - 1) \\ t' &= t/\varepsilon, \end{aligned} \tag{5}$$

Then equations (1) - (3) become (after dropping the primes)

$$\dot{x} = y - \varepsilon\sigma x \tag{6}$$
$$\dot{y} = -xz - \varepsilon y \tag{7}$$
$$\dot{z} = xy - \varepsilon b(z + \sigma). \tag{8}$$

Now consider the expansion of $x, y,$ and z in terms of the parameter ε

$$\begin{aligned} x &= x_0 + \varepsilon x_1 + \varepsilon^2 x_2 + \cdots \\ y &= y_0 + \varepsilon y_1 + \varepsilon^2 y_2 + \cdots \\ z &= z_0 + \varepsilon z_1 + \varepsilon^2 z_2 + \cdots. \end{aligned} \tag{9}$$

Introducing (9) in (6) - (8), gives the order 0 equations,

$$\dot{x}_0 = y_0 \tag{10}$$
$$\dot{y}_0 = -x_0 z_0 \tag{11}$$
$$\dot{z}_0 = x_0 y_0 \tag{12}$$

and at order ε, the system,

$$\dot{x}_1 = y_1 - \sigma x_0 \tag{13}$$
$$\dot{y}_1 = -x_0 z_1 - x_1 z_0 - y_0 \tag{14}$$
$$\dot{z}_1 = x_0 y_1 + x_1 y_0 - b(z_0 + \sigma). \tag{15}$$

The interdependence of the order 0 equations can removed with some algebraic manipulation. Use (11) and (12) to eliminate x_0

$$\begin{aligned} y_0 \dot{y}_0 + z_0 \dot{z}_0 &= 0 \\ \frac{d}{dt}\left[y_0^2 + z_0^2\right] &= 0. \end{aligned} \tag{16}$$

Integrating this,

$$y_0^2 = -z_0^2 + [y_0^2(0) + z_0^2(0)] \tag{17}$$

where the terms in the brackets of equation (17) are the initial values of y_0 and z_0. We will define this (constant) term as

$$C_{23} = y_0^2(0) + z_0^2(0) \tag{18}$$

If we go back to (10) and (12) to eliminate y_0,

$$\begin{aligned} \dot{z}_0 &= x_0 \dot{x}_0 = \frac{1}{2}\dot{x}_0^2 \\ \frac{d}{dt}\left[z_0 - \frac{1}{2}x_0^2 \right] &= 0. \end{aligned} \tag{19}$$

Integrating this gives

$$z_0 = \frac{1}{2}x_0^2 - \left[\frac{1}{2}x_0(0) - z_0 \right]; \tag{20}$$

here the terms in the brackets of equation (20) are the initial values of x_0 and z_0. We will define this term as

$$C_{13} = \frac{1}{2}x_0(0) - z_0(0). \tag{21}$$

Using (17) in (10) gives

$$(\dot{x}_0)^2 = -z_0^2 + C_{23}. \tag{22}$$

Now using (19)

$$(\dot{x}_0)^2 = -\frac{1}{4}x_0^4 + C_{13}x_0^2 + [C_{23} - C_{13}^2]. \tag{23}$$

Given the solution to this equation, y_0 can be solved for by using (10). Then given x_0 and y_0, z_0 can be solved for by using (12). An equation for z_0 can also be derived by using manipulations similar to that used in deriving (23) (using equations (17) and (20) in (12) to eliminate x_0 and y_0) , to give

$$(\dot{z}_0)^2 = -2z_0^3 - 2C_{13}z_0^2 + 2C_{23}z_0 + 2C_{13}C_{23}. \tag{24}$$

Equation (22) can be solved analytically, its solution is a **Jacobi elliptic function**

$$x_0 = A\,\mathrm{sn}(t|m).$$

The other components can also be determined,

$$\begin{aligned} y_0 &= A\,\mathrm{cn}(t|m)\,\mathrm{dn}(t|m) \\ z_0 &= \mathrm{dn}^2(t|m)\; m\,\mathrm{cn}^2(t|m) \end{aligned} \tag{26}$$

(where $m = -A^2 / 4$) so the system is well behaved (not chaotic). The first order equations (13) - (15) are linear so they cannot possibly lead to chaotic solutions. Thus we have shown that while the actual system can be chaotic, the perturbation solutions may not be

Figure 6 shows a phase portrait of the solution of the full system (6) - (8), the zero order system (10) - (12), and the perturbation solution to first order (i.e., with ε times the solution of (13) - (15) added to the zero order solution). The perturbation solution tracks the nonlinear solution for a short while then it moves off in a different direction. The perturbation solution also rapidly grows to order one, so that the expansion (9) is valid for only a limited time.

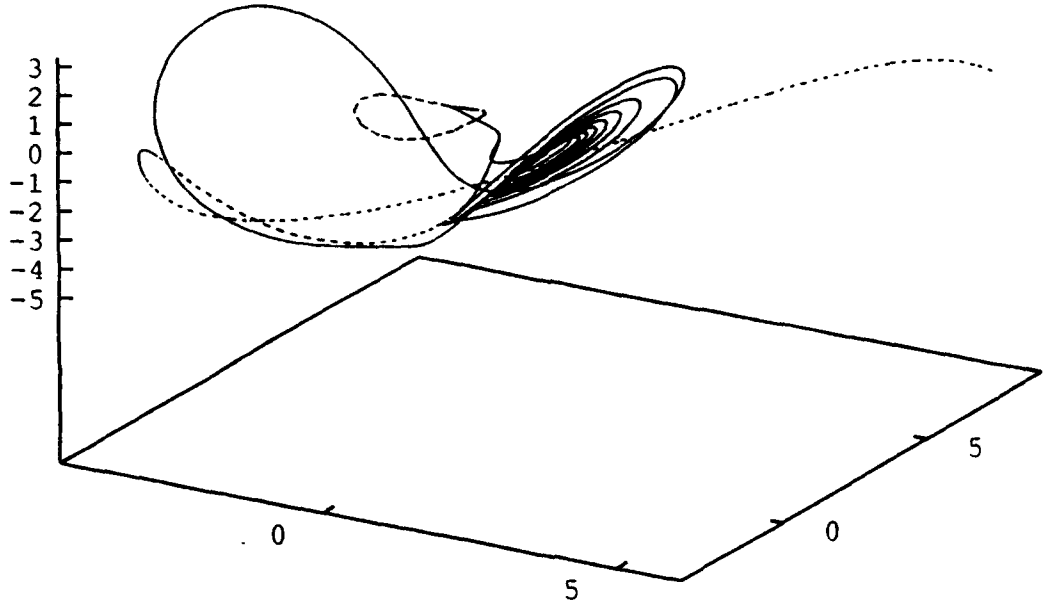


Figure 6. Solutions to the Lorenz equations for large Rayleigh number ratio (equations (6)-(8)). The solid line is the solution to the nonlinear (chaotic) system. The long dashed line is the solution to the zero order perturbation experiment. The short dashed line is the perturbation solution to first order.

2. **Practical problems in estimating chaotic parameters from actual data**

Most methods developed for quantifying chaos (e.g., the Grassberger-Procaccia (1983) method) require very long data sets in order to converge with a reasonable uncertainty. Such lengthy data sets do not exist in oceanography, so methods that work with short data sets (see for example, Ellner (1988), Havstad and Ehlers (1989) or Abraham et al., 1986) must be used. Also the presence of noise (either due to measurement errors or to small scale oceanic process) complicates the calculations. In addition, the ill-considered use of filters applied to the data can make things worse, not better.

2.1 **The effect of noise**

Random errors in the observations of a system can complicate the estimation of the fractal dimension of a system. It has the effect of **increasing** the apparent dimension of the system. This is unfortunate since estimation methods have data requirements that grow exponentially with the dimension of the system.

In addition, while truly random processes ought to be infinitely dimensional, biases in commonly used dimension algorithms indicate finite dimension when presented with random data.

For **colored noise**, the correlations between nearby points can produce effects that mimic a finite correlation dimension (Theiler, 1991). Osborne and Provenzale (1989) provide an example of this effect. Kennel and Isabelle (1992) have investigated the possibility of distinguishing colored noise effects from chaos.

2.2 **The effect of filtering the observations**

One traditional way to deal with noise in the observations is to apply a filter in an attempt to remove the frequencies that are attributed to the noise. With chaotic systems, the effect of the filter is to potentially **increase** the apparent dimension of the system (Badii et al., 1988).

Consider a physical system $\dot{u}(t) = -F(u)$ and an ideal lowpass filter, which can be described as a differential equation that adds to the original system:

$$\dot{z}(t) = -\eta z(t) + X(t) \tag{27}$$

where $z(t)$ is the filter output, and η is the filter cutoff frequency.

With this filter present, the Lyapunov exponents of the system consist of the original Lyapunov exponents plus a new one $\lambda_f = -\eta$ resulting from the filter.
From the Kaplan-Yorke equation for the Lyapunov dimension

$$D_L = j + \sum_{k=1}^{j} \frac{\lambda_k}{|\lambda_{j+1}|} \tag{28}$$

The dimension, D_L of the system will remain unchanged as long as

$$\eta \geq |\lambda_{j+1}|.$$

Otherwise the dimension of the filtered system will **increase**. In fact, depending upon the size of η compared to the other Lyapunov exponents, D_L can increase as much as 1. There has been some work (e.g., Chennaoui et al., 1990) to remove this effect of filtering on chaotic time series by (at least in a topological sense) unfiltering the time series.

3 Methods from systems dynamics

Even if it turns out that the ocean is not chaotic, certain techniques developed for analyzing chaotic systems may prove useful. For many of these methods the fact that a nonlinear system is a chaotic one is not essential for the analysis method to be usable.

3.1 Mutual information and dynamical connections

The mutual information of two (discrete scalar) messages S and Q is (Fraser and Swinney, 1986)

$$I(Q,S) = H(Q) + H(S) - H(Q,S) \tag{30}$$

where

$$H(Q) = -\sum_i P_q(q_i) \log(P_q(q_i)) \tag{31}$$

(and similarly for S)

$$H(Q,S) = -\sum_i P_{qs}(q_i,s_j) \log(P_{qs}(q_i,s_j)). \tag{32}$$

When Q is a set of time delayed measurements ($q(t+\tau)$) then the **first** minimum of I as a function of τ is a good choice of the lag time in the higher dimensional reconstruction (Fraser, 1986).

By taking the appropriate limits, we can calculate an **information dimension** from the mutual information

$$D_I = D_q + D_s - D_{qs}. \tag{33}$$

D_I is nonnegative and has the following properties:

$D_I = D_q$ when $q = s$

$D_I < D_q$ when q and s are time shifted versions of each other or when they are dynamically related (and have the same dimension)

$D_I = 0$ when q and s are dynamically independent.

Hence we have a test for **synchronizability** and for **dynamical relatedness**. This could be exploited to determine if two different time series (say one from a model and another from actual observations) are controlled by the same dynamics or not.

3.2 A theorem on dynamic dependence

The dimensions and entropies of series can also be used to determine whether two systems are dynamically independent or not. The following theorem is due to Hartt and Kahn, 1990.

Consider a composite system

$$\vec{x}_{ab,i} = \begin{bmatrix} \vec{x}_{a,i} \\ \vec{x}_{b,i} \end{bmatrix} \tag{34}$$

where

$$\vec{x}_{a,i} = [Y(t_i), Y(t_i+\tau), \cdots Y(t_i+(d-f-1)\tau)]^T \tag{35}$$

$$\vec{x}_{b,i} = [Z(t_i), Z(t_i+\tau), \cdots Z(t_i+(f-1)\tau)]^T \tag{36}$$

with combined dimension of $d, ((d-f)+(f))$. We investigate the effects of the dependence and independence of these subsystems. The supremum norm gives

$$\rho_{ab}(i,j) = \text{dist}(\vec{x}_{ab,i}, \vec{x}_{ab,j}) = \max_{k=0,d-1} \left| X_{ab} \; - X_{ab,j,k} \right| \tag{37}$$

where k represents a component. It follows that

$$\rho_{ab}(i,j) = \max(\rho_a(i,j), \rho_b(i,j)) \rho_{ab}(i,j) = \max(\rho_a(i,j), \rho_b(i,j)). \tag{38}$$

The simplest way to obtain dimensions and entropies is to evaluate the generalized correlation integrals

$$C_d^q(\ell) = \left[\frac{1}{N_r} \sum_i \{ \frac{1}{Ns} \sum_{j \neq i} \theta(\ell - \rho_d(i,j)) \}^{q-1} \right]^{\frac{1}{q-1}} \tag{39}$$

where N_r = number of reference points and N_s = number of sample points in the vector time series. Then

$$\theta(\ell - p_{ab}(i,j)) = \theta(\ell - p_a(i,j)) \theta(\ell - p_b(i,j)). \tag{40}$$

There are two important special cases:

- Identical subsystems
- Independent subsystems

3.2.1 **Identical subsystem**

In this case $p_a = p_b$, and $(f = \frac{d}{2} = d - f)$. Here, $\theta(\ell - p_a(i,j))\theta(\ell - p_b(i,j)) = \theta(\ell - p_a(i,j))$ and $C^q_{ab,d}(\ell) = C^q_{a,\frac{d}{2}}(\ell) C^q_{ab,d}(\ell) = C^q_{a,\frac{d}{2}}(\ell)$ for all q and ℓ. Asymptotically for $\ell \to 0$,

$$C^q_{ab,\frac{d}{2}}(\ell) \sim \ln \ell^{\nu_{ab}} \exp(-d\tau K^q_{ab}(d,\tau)) \tag{41}$$

and similarly for $C^q_{ab,d}(\ell)$. Then

$$\begin{aligned} \ln C^q_{ab,d}(\ell) &= \nu_{ab} \ln \ell - d\tau K^q_{ab}(d,\tau) \\ &= \nu_a \ln \ell - \frac{d}{2} \tau K^q_a(\frac{d}{2}) \end{aligned} \tag{42}$$

from which we arrive at

$$\nu_{ab} = \nu_a \tag{43}$$

and

$$K^q_{ab}(d) = K^q_a(\tfrac{d}{2}). \tag{44}$$

3.2.2 **Dynamically independent subsystems**

Here, $\rho_b(i,j)$ takes values that are independent of $\rho_a(i,j)$. Then $\theta(\ell - p_b(i,j))$ can be replaced by its average value over the entire series. The cases $q = 1$ and $q = 2$ are especially important. In both of these cases it follows

$$C^{1,2}_{ab,d}(\ell) = C^{1,2}_{a,d-f}(\ell) C^{1,2}_{b,f}(\ell) \tag{45}$$

from which asymptotically,

$$\nu^{1,2}_{ab} = \nu^{1,2}_a + \nu^{1,2}_b \tag{46}$$

and in the case $\frac{d}{2} = f = d - f$,

$$K^{1,2}_{ab} = \frac{1}{2}\left[K^{1,2}_a + K^{1,2}_b\right]. \tag{47}$$

Clearly, $C(i,j) < C(i)$.

4 Summary

- Several oceanic phenomena, El Niño and drifter trajectories in particular, are suggestive of chaos. For El Niño, the presence of chaos is inconclusive. Drifter trajectories, on the other hand, are probably not chaotic.
- Limitations on the quantities of data have prevented a definitive conclusion on the existence of chaos in the ocean.
- The existence of chaos means that special care must be used when dealing with both the equations and the data.
- The properties of dynamically connected chaotic systems may be useful in identifying the dynamical system.

References

Abraham, N.B., A.M. Albano, B. Das, G. De Guzman, S. Young, R.S. Gioggia, G.P. Puccioni, and J.R. Tredicce, 1986; Calculating the Dimension of Attractors from small data sets, *Phys. Lett. A*, 114, 217–221.

Badii, R., G. Broggi, B. Derighetti, M. Ravani, S. Ciliberto, A. Politi, and.M.A. Rubio, 1988, Dimension Increase in Filtered Chaotic Signals, *Phys. Rev. Lett.*, 60, 979–982.

Brown, M.G, and K.B. Smith, 1990, Are SOFAR Float Trajectories Chaotic?, *J. Phys. Oceanog.*, 20, 139–149.

Chennaoui, A., K. Pawelzik, W. Liebert, H.G. Schuster, and G. Pfister, 1990; Attractor reconstruction from filtered chaotic time series, *Phys. Rev. A*, 41, 4151–4159.

Ellner, S., 1988; Estimating attractor dimensions from limited data: A new method, with error estimates, *Phys. Lett. A*, 133, No. 3, pp. 128–133.

Fraedrich, K., 1988; El Niño/Southern Oscillation Predictability, *Mon. Weather Rev.*, 116, 1001–1012.

Fraser, A., 1986; Using mutual information to estimate entropy, in G. Mayer-Kress, ed, *Dimensions and Entropies in Chaotic Systems*, Springer-Verlag, Berlin.

Fraser, A., and H.L. Swinney, 1986; Independent coordinates for strange attractors from mutual information, *Phys. Rev. A*, 33, 1134–1140.

Grassberber, P. and I. Procaccia, 1983; Measuring the strangeness of strange attractors, *Physica* 9D, 189–208.

Hartt, K., and L.M. Kahn, 1990; Seeking dynamically connected chaotic variables, in N.B. Abraham, A.M. Albano, A. Passamante, and P.E. Rapp, editors, Proceedings of Workshop: *Quantitative Characterization of Dynamical Complexity in Nonlinear Systems*, June 22-24, 1989, Bryn Mawr College, Plenum Press, New York, 475 pp.

Havstad, J.W. and C.L. Ehlers, 1989; Attractor dimension of nonstationary dynamical systems from small data sets, *Phys. Rev. A*, 39, 845–853.

Kennel, M.B., and S. Isabelle, 1992; A Method to Distinguish Possible Chaos from Colored Noise and Determine Embedding Parameters, *Phys. Rev. A. submitted preprint.*

Lorenz, E.N., 1963, Deterministic Nonpriodic Flow, *J. Atmos. Sci.*, 20, 130–141.

Lorenz, E.N., 1964; The problem of deducing the climate from the governing equations, *Tellus*, XVI, 1–11.

Osborne, A.R., A.D. Kirwan Jr, A. Provenzale, and L. Bergamasco, 1986; A Search for Chaotic Behavior in Large and Mesoscale motions in the Pacific Ocean, *Physica* 23D, 75–83.

Osborne, A.R., A. Provenzale, 1989; Finite Correlation Dimension for Stochastic Systems with Power-Law Spectra, *Physica* 35D, 357–381.

Theiler, J., 1991; Some Comments on the Correlation Dimension of $f^{-\alpha}$ Noise, *Phys. Lett. A*, 155, 480–489.

Vallis, G.K., 1986; El Niño: A Chaotic Dynamical System ?, *Science*, 232, 243–245.

GEOMETRIC THERMODYNAMICS AS A TOOL FOR ANALYSIS AND PREDICTION IN OCEANOGRAPHY

Nessan Fitzmaurice, Wojbor Woyczynski
Department of Mathematics, Case Western Reserve University, Cleveland, Ohio, 44106

Anatoly Odulo
Applied Science Associates Inc., Narragansett, Rhode Island

ABSTRACT

The physical parameters that are important to oceanographers often have a stochastic nature and can be represented as the sum of a deterministic average and a random component of zero mean. Coastline shapes, water depth and fluid density are examples of such quantities. When the random components are small, perturbation methods can be used to calculate their effects on the mean flow. However, in certain cases it is the derivative of the random component which is of importance and that can have a very large magnitude. Consequently, the ostensibly small stochastic part may well be more influential than the smooth average component. In this paper we present a technique for quantifying roughness that can be easily implemented for experimental data sets and apply the method to some bathymetric examples. Moreover, to examine how such randomness will influence ocean flows we consider the problem of predicting the dispersion relations for topographic Rossby waves propagating in the presence of a rough ocean floor. The random depth and its *derivative* act as coefficients in the equations governing topographic Rossby waves. In this paper we analytically and numerically examine the solutions to those equations and consider how they change as the roughness of the bottom increases.

1. INTRODUCTION

Many physical characteristics of importance to the oceanographer have a stochastic nature and can be represented as the sum of a deterministic average and a random component of zero mean. Quantities that come to mind include coastline shapes, water depth and fluid density.

When the random components are small, perturbation methods can be used to calculate their effects on the mean flow (see for example Mysak (1978)). However, in certain cases it is the derivative of the random component which is of importance and that can have a very large magnitude. Consequently, the influence of the ostensibly small stochastic part may well be of the same, or even larger order, than that of the smooth average component.

For example, it has long been recognized that variations in the sea floor topography allows for the propagation of a class of disturbances known as topographic Rossby waves (see for example Pedlosky (1989)). These flows are spatially extensive and have

large temporal periods. The critical coefficient in the governing equations for such waves depends on the derivative of the undisturbed water depth. This depth might well be considered random and rather small ($O(1)$ km.) when compared to the spatial extent of the waves in question ($O(100)$ km.). However, the *derivative* of the depth which appears in the equation can not be treated as a small term.

In past studies, the ocean floor was often treated as a plane with a slight slope and indeed mathematical analysis then predicts the existence of topographic waves. While it true that many regions of the ocean can be characterized by having a small mean slope, it is not apparent that one can ignore other variations in topography in favor of this slope.

A natural question arises then as to whether one can quantitatively characterize the roughness inherent in bathymetric data. Can one give some reasonable estimate of the "average slope" of such a data set? In recent years, fractal techniques have proved popular in similar quests by researchers in many fields. However, dimension estimates depend on infinite scaling properties that are often not physically justifiable as real surfaces are self similar only over a limited range of scales. Moreover, as a practical matter, fractal analyses are simply not formulated with discrete data sets in mind.

In this paper we take an alternative approach based on the *geometric thermodynamic* theory for curves and surfaces. The essential ideas are developed in the next section and are illustrated there by a number of simple examples. We merely note now that the theory allows one to compute a temperature for a curve or surface. The temperature of a straight line is zero and the value increases as the curve roughness increases. The quantity can be successfully measured even for data sets of finite resolution and efficient algorithms to accomplish this are presented in section 3.

Tools to analyze rough data sets are very useful but it is even more intriguing to apply those tools to predict how roughness influences oceanographic flows. To that end we take up the problem of computing the dispersion relations that govern linear topographic Rossby wave disturbances for an ocean of random depth. The governing equations, some properties of their solutions, and the numerical methods used to solve them are described in section 4.

To render the problem computationally tractable, we restrict attention to topographies that vary in one direction only, i.e. we consider oceans with corrugated floors. Past investigations by Thomson (1975), Odulo and Pelinovsky (1978) and others have shown that if the waves are constrained to propagate in the same direction as the bottom relief then even for simple floors with periodic ripples wave dissipation and reflection are observed. In particular, the second study showed that the characteristic damping time for Rossby waves is $T_d \sim (\Delta h/h_0)^{-2}$. Typical values for the ocean are fairly large—in the 4 months to 3 years range. In this paper we will consider waves propagating in a direction that is **not** parallel to the bottom topography and in this case the waves do not dissipate.

It should be pointed out from the start that while the ocean floor can be modelled stochastically, it is far removed from a white noise state. In this paper most of the

simulations were done for synthetic bottom profiles though some preliminary analysis has been carried for data sets collected in the North Atlantic and Pacific. Some of the methods we used to synthesize rough bottom profile are presented in section 5 and their geometric thermodynamic characteristics are computed.

In section 6 we present results for bathymetries of various temperatures.

2. GEOMETRIC THERMODYNAMICS

In this section we explain how one can formulate a thermodynamic theory for geometric objects and how one can use that theory to construct quantitative estimates of the "roughness" of curves. We illustrate the concepts with a number of simple examples.

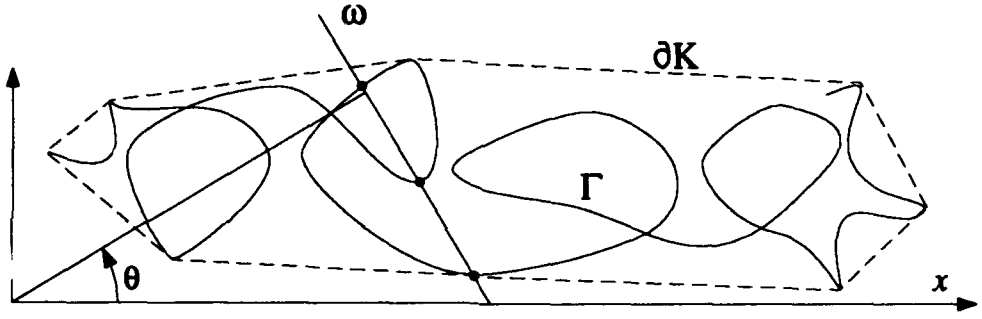


Figure 1 Random straight line ω intersecting curve Γ at three points. The convex hull of Γ is the set of points that lie inside the dashed boundary line.

The fundamental quantity we measure for a curve is the average number of intersections it has with randomly chosen straight line segments. In general, the rougher the curve, the larger this number will be. To formalize the idea let Γ be a rectifiable curve in the plane and let $\Omega(\Gamma)$ be the set of all straight lines intersecting Γ. Directly measuring the number of intersections between a random element $\omega \in \Omega(\Gamma)$ and Γ is a computationally intensive task but Blaschke (1936) has shown that if one picks ω at random, with the natural (and as it turns out unique) distribution m that is invariant under rigid motions of the plane, then the average number of intersections between the line ω and the curve Γ is given by

$$\frac{2|\Gamma|}{|\partial K|}, \tag{1}$$

where the convex hull of Γ, K has boundary ∂K and we use $|\cdot|$ to denote the length of a curve. A detailed definition of the term convex hull is given in the next section but its intuitive meaning should be clear from figure (1).

An easy derivation of Blaschke's formula for smooth (piecewise differentiable) curves can be found in Santalo (1976) and depends on the formula

$$\int_{\Omega(\Gamma)} n_\Gamma(w)\, d\omega = \int_0^{|\Gamma|} ds \int_0^{\pi} |\sin\theta|\, d\theta = 2|\Gamma|. \tag{2}$$

Here the line ω is parameterized by its perpendicular length, s, from the origin and by the angle θ subtended by the normal to ω with the x axis. $n_\Gamma(\omega)$ is defined as the number of points at which the line ω intersects Γ ($n_\Gamma(\omega) = 3$ in figure (1)). Some manipulation of this formula quickly gives

$$\frac{1}{m(\Omega(\Gamma))} \int_{\Omega(\Gamma)} n_\Gamma(\omega)\, d\omega = \frac{2|\Gamma|}{|\partial K|}, \tag{3}$$

as claimed at the beginning of the section.

Steinhaus (1954) observed that while the quantity on the right hand side of (3) only makes sense for rectifiable curves, the left hand side, representing the average number of intersections with lines, makes sense for any planar set, whatever its complexity. The set need not be a curve representing a single valued function or even a curve at all. He then suggested that the average be considered as a measure for the "length" of such a set. This is the starting point of our paper and suggests a way of measuring the roughness of interfaces that can be much more general than those described by functions of one or two variables.

DuPain, Kamae and Mendes-France (1986) extended Steinhaus's approach by applying ideas from the field of statistical mechanics. They considered the family $M^*(\Gamma)$ of all probability measures on $\Omega(\Gamma)$ which gave the same average number of intersections of lines ω with Γ as is given by the isotropic homogeneous measure m. For a curve Γ which has the property that for any positive integer k there exists a line ω which intersects Γ exactly k times, one can associate a geometric entropy function $\sigma : M^*(\Gamma) \to \mathbf{R}$ by defining

$$\sigma(m) = -\sum_{k=1}^{\infty} m(\omega :\ |\omega \cap \Gamma| = k)\ \log m(\omega :\ |\omega \cap \Gamma| = k) \tag{4}$$

where $|\omega \cap \Gamma|$ stands for the number of intersections between ω and Γ.

By a straightforward application of the method of Lagrange multipliers one can then find a "Gibbs" measure $g \in M^*(\Gamma)$ which maximizes the geometric entropy over $M^*(\Gamma)$. It turns out that

$$g(\omega :\ |\omega \cap \Gamma| = k) = Ce^{-\beta k}, \tag{5}$$

where

$$e^{\beta} = \frac{2|\Gamma|}{2|\Gamma| - |\partial K|}. \tag{6}$$

The maximum geometric entropy is then

$$\sigma(g) = \log\left(\frac{2|\Gamma|}{|\partial \mathrm{K}|}\right) + \frac{\beta}{e^{\beta}-1}, \tag{7}$$

and C^{-1} is the partition function with $C = e^{\beta} - 1$.

Other geometric "thermodynamic" quantities can easily be defined including the geometric temperature $\tau = \beta^{-1}$, the geometric pressure $\Pi = |\partial \mathrm{K}|^{-1}$, the geometric volume $V = |\Gamma|$, the geometric heat $Q = \left(e^{\beta} - 1\right)^{-1}$ and the geometric free energy $F = \beta^{-1}\log\left(e^{\beta} - 1\right)$. A particular quantity that we shall make further use of is the geometric internal energy

$$U = \frac{2|\Gamma|}{|\partial \mathrm{K}|} = \frac{e^{\beta}}{e^{\beta}-1}. \tag{8}$$

Although the construction used above is only valid for a limited class of curves, the quantities β and σ can easily be extended to all rectifiable curves. Mann, Rains and Woyczynski (1991) contains further details of the application of these ideas to the roughness of surfaces but in this paper we will only consider one dimensional objects.

Note that if Γ is itself a straight line segment then $U = 1$ and $\tau = 0$. Thus the least interesting curves, straight lines, all have zero temperatures!

To gain some familiarity with the concepts outlined above let us consider some other simple examples where Γ is a portion of an infinite curve with small scale roughness. It is then reasonable to replace $|\partial \mathrm{K}|$ by $2L$ where L is the distance between the end points. This is because for a periodic extension of Γ the convex hull is an infinite strip and the section of $\partial \mathrm{K}$ corresponding to Γ has perimeter approximately equal to $2L$. In a later section we will explain algorithms that can be used to precisely measure the convex hull for more complicated situations. With that approximation $U = |\Gamma|/L$.

Example 1: Sawtooth Curves

Let Γ_N be the periodic sawtooth curve with period a and amplitude ϵh_0 shown on the left of figure (2).

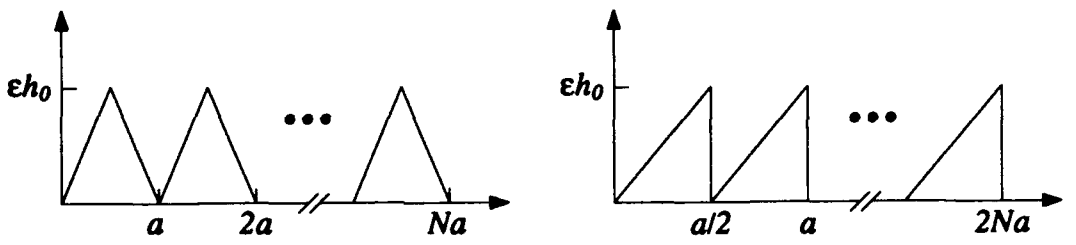


Figure 2 Symmetric and asymmetric sawtooth curves

By inspection

$$|\Gamma_N| = 2N\sqrt{\left(\frac{a}{2}\right)^2 + (\epsilon h_0)^2}, \tag{9}$$

while $|\partial \mathrm{K}| \approx 2Na$. Hence

$$U \approx \sqrt{1 + (2\epsilon h_0 a^{-1})^2}. \tag{10}$$

Each line making up the sawtooth has the same slope in absolute magnitude so the average value for the absolute slope is

$$\delta = 2\epsilon h_0 a^{-1}. \tag{11}$$

Note that

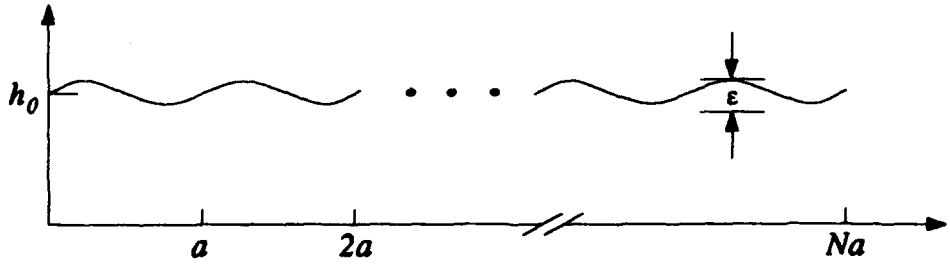
$$\delta = \sqrt{U^2 - 1}. \tag{12}$$

Another observation will be of some consequence later is that the profile on the right of figure (2), which is not invariant with respect to a change of direction $x \to -x$, still gives rise to the same values for U and δ.

Example 2: Sinusoidal Curves

For our next example we consider the sinusoidal profile

$$\Gamma_N(x) = h_0\left(1 + \epsilon \sin\frac{2\pi}{a}x\right), \quad 0 \le x \le Na. \tag{13}$$



For this curve the average value of the absolute value of the slope is easily calculated as

$$\delta = \frac{1}{Na}\int_0^{Na}\left|\frac{2\pi\epsilon h_0}{a}\cos\frac{2\pi}{a}x\right| dx = \frac{4\epsilon h_0}{a}. \tag{14}$$

The length of the curve is

$$|\Gamma_N| = \int_0^{Na} \sqrt{1 + \Gamma_N'(x)^2}\, dx = \frac{2Na}{\pi} \int_0^{\pi/2} \sqrt{1 + \left(\frac{2\pi\epsilon h_0}{a} \cos y\right)^2}\, dy, \tag{15}$$

while $|\partial \mathrm{K}| \approx 2Na$. Hence

$$U \approx \frac{2}{\pi} \int_0^{\pi/2} \sqrt{1 + \left(\frac{2\pi\epsilon h_0}{a} \cos y\right)^2}\, dy = \frac{2}{\pi} \int_0^{\pi/2} \sqrt{1 + \left(\frac{\pi\delta}{2} \cos y\right)^2}\, dy. \tag{16}$$

Setting $p = \pi\delta/2$ this integral can be expressed in terms of the elliptic function of second kind E in the form

$$\begin{aligned} U &= \sqrt{1+p^2}\, E\left(\frac{\pi}{2}, \frac{p}{\sqrt{1-p^2}}\right), \\ &= \sqrt{1+p^2}\left[1 - \frac{1}{2}\left(\frac{p^2}{1-p^2}\right) - \cdots - \frac{(2n-1)!!}{2^n n!}\frac{1}{2n-1}\left(\frac{p^2}{1-p^2}\right)\cdots\right], \end{aligned} \tag{17}$$

so that in the first approximation

$$U \approx \sqrt{1 + \left(\frac{\pi\delta}{2}\right)^2}, \tag{18}$$

and once again the average slope is proportional to $\sqrt{U^2 - 1}$

$$\delta \approx \frac{2}{\pi}\sqrt{U^2 - 1}. \tag{19}$$

3. MEASURING THE CONVEX HULL

A domain $D \subset \mathbf{R}^n$ is said to be convex if for every pair of points $p_1, p_2 \in D$, the line segment $\overline{p_1 p_2}$ is entirely contained in D. Given an arbitrary set of points $S \subset \mathbf{R}^n$, the convex hull $\mathrm{conv}(S)$ of S is defined to be the smallest convex domain containing S. The hull of a bounded set will always be a convex polytope.

For any finite set of points on the plane it is easy to visualize the convex hull by imagining that the points are marked on a board with protruding nails. To find the hull, stretch a rubber band so that it encloses all of the points and release it. The band will be caught on the nails located at the extreme points of the set and form the polygonal boundary of the convex hull.

In order to apply Blaschke's formula (1) to general sets of points we must implement an algorithm for computing the convex hull. Several algorithms for this purpose exist for

planar sets of points and we will merely mention a couple of techniques here. The reader is referred to the text by Preparata and Shamos (1985) for further details of the theory.

The *package wrapping* technique is the simplest algorithm for extracting the subset of the points that form the convex hull. While it is not the fastest method for sets of points on the plane, it deserves attention because it is one of the few that can be generalized to deal with higher dimensional data. This is an important consideration because we will eventually want to handle large three dimensional sets of topographic data.

The method parallels how a human might draw the boundary of the convex hull. Start with some point that is guaranteed to be on that boundary, say the one with smallest y coordinate. Fix one end of a horizontal line to this point and rotate it upwards until it encounters another point in the set. That point must also belong to the convex hull. Use it as a new anchor for the horizontal line and repeat the procedure. Continue in this fashion until you form a package that completely wraps around the original set of points. The package is precisely the boundary of the convex hull.

Of course, instead of sweeping horizontal lines around to see which point in the set they hit first, one actually looks at all the segments between the current anchor and the other points not yet accounted for by the convex hull boundary. The end point of the segment that subtends the smallest angle with the x axis will be the next point in the hull and it will also be the new anchor. The major computational costs associated with the algorithm are the calculation of lots of angles followed by some form of sorting procedure on those angles. It can be shown that the technique takes $O(N^2)$ operations for sets with N points. The constant in front of the N^2 is large however.

Several improvements on the basic algorithm can be made. For one thing, it is possible to cheaply eliminate many of the points *before* we call the convex hull routine. One way to do this is to construct an extreme quadrilateral by searching for those points in the set that have the largest and smallest x and y coordinates. This search can be done in $3N$ operations and will typically yield four different vertices. Points that lie inside the region defined by those vertices cannot be on the convex hull boundary. By eliminating them (another linear time process) one effectively reduces N, the number of points submitted to the more expensive package wrapping technique. If one happens to know something about the distribution of the set of points even better quadrilaterals can be chosen to maximize this effect.

Additional savings come from the realization that a lot of time is spent computing angles. A naive implementation might calculate $\theta = \arctan(\Delta y/\Delta x)$ but evaluating the arctangent is relatively expensive, particularly on RISC machines that do not perform the computation in hardware. In any case, the precise value of the angle is of little interest here as we are only using it as a key in the sorting process. What is required is a cheap alternative to the arctangent that preserves the ordering properties of that function. A good candidate for this purpose is $\Delta y/(|\Delta y| + |\Delta x|)$ with appropriate modifications for positive and negative values of Δx and Δy.

The two techniques just mentioned, eliminating "obvious" points and replacing an expensive calculation with a cheaper one, can both be used to good effect for data in any dimension. Further savings are possible for planar points. Graham (1972) suggested that one first form any simple closed polygon that contains all the points. Having found this polygon one then proceeds to eliminate from it those points that do not belong to the convex hull. The major cost of this effort is the initial construction of the closed polygon and this is done by a sorting procedure based on angles from say the lowest point in the set. The number of operations for the Graham scan is thus dominated by the sorting process which can be done in $N \log N$ operations for N input points. Even more sophisticated divide and conquer algorithms are available in the literature but we have found that even for large sets of data the Graham scan technique coupled with interior elimination provides adequate efficiency.

Figure (3) shows the calculation of the convex hull boundary for 100 points which were chosen to be uniformly distributed in a square.

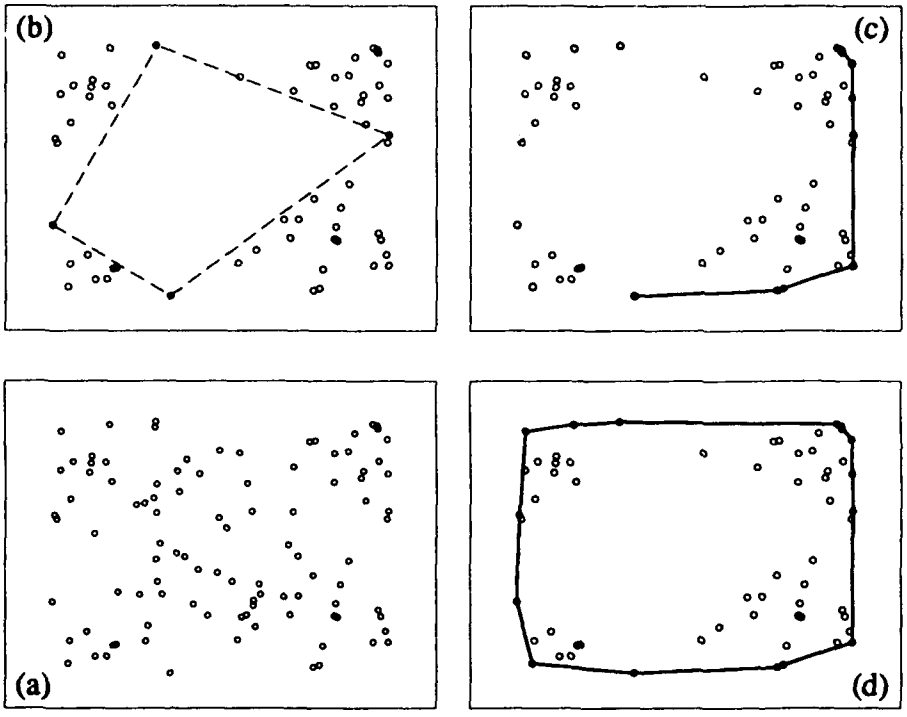


Figure 3 These four plots show: (a) the original set of points, (b) the set with the "interior" points removed, (c) the package wrapping algorithm in progress, (d) the completed convex hull boundary.

Roughness Calculations for Bathymetric Data

The topographic wave dispersion relation computations carried out in the later sections of this paper are for synthesized bottom profiles only. Indeed, at this early stage of our investigations we are primarily interested in profiles having a *controllable* degree

of roughness. However, it is naturally interesting to examine the degree of roughness that is present in real bathymetric data. Therefore we have analyzed some of the high quality tracklines that are present in the large database assembled by the National, Geophysical and Solar-Terrestrial Data Center/NOAA (NGSDC/NOAA 1977). The reader is referred to Dworski and Holloway (1983) for a statistical study of this data.

We present a sample calculation here. The data came from a cruise by the R/V Melville II from Adak to Tokyo in October 1973. Bathymetry data at the start and the end of cruise were ignored until a reading of 5000 meters was encountered. Some 2408 depths were recorded corresponding to about one reading every 1.75 kilometers along the track. On the Mercator map the trackline is approximately a straight line starting at (178°W, 52°N) near Adak in the North Pacific and proceeding south west to(143°E, 35°N) west of Tokyo. Figure (4) shows the total depth profile.

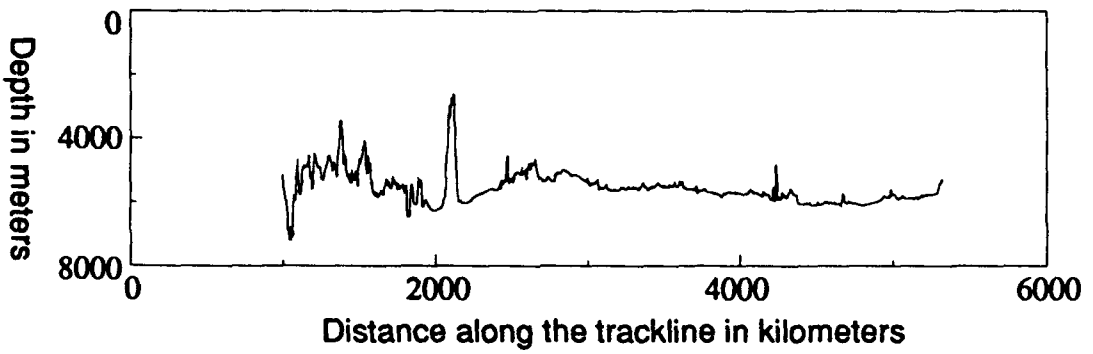


Figure 4 Bathymetric data from the cruise of the R/V Melville II from Adak to Tokyo in October 1973.

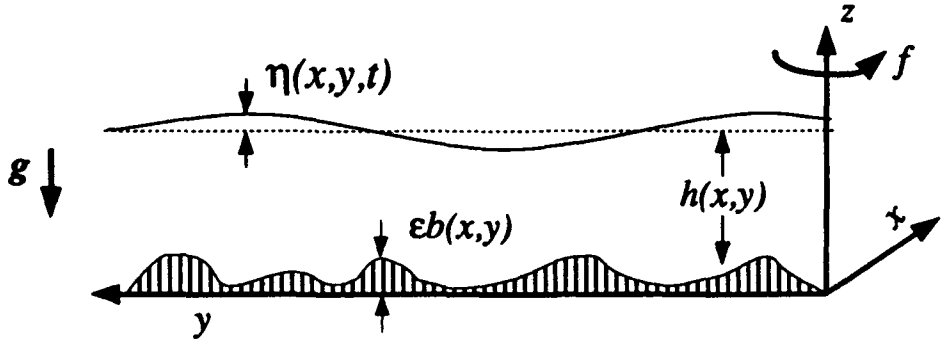
Looking at figure (4) it is clear that the data are rougher in some sections than in others. In the following table we present some thermodynamic characteristics for the curve as a whole and then separately for four 1000 kilometer sections along its length. We note that the thermodynamic statistics are all perfectly well defined for sections of the curve—in the future we expect to make use of this trait to focus our computational energies on those parts of the boundaries that are likely to provide the greatest challenge for flow simulations. The fact that the theory is well posed for even the crudest of data sets makes it a useful diagnostic for adaptive computations.

Section	Number of data points	Number eliminated by interior check	$U = 2\|\Gamma\|/\|K\|$	Temperature τ
Full track	2408	1726	1.0007	0.1378
1000-2000 km.	520	457	1.0017	0.1566
2000-3000 km.	544	495	1.0009	0.1416
3000-4000 km.	575	295	1.0001	0.1033
4000-5000 km.	564	260	1.0004	0.1271

In the table we report on the number of points that were present in the experimental data and also the number of those that were eliminated by the interior check procedure before the convex hull routine was called. On the average, some 70% of the data points were eliminated by this check and in fact the calculations could easily be carried out in near real time on a moderate workstation or personal computer. We note that the temperature of the first 1000 kilometer stretch is the largest which corresponds well with our intuitive sense that data sections that are visually "roughest" should give rise to larger temperatures.

4. LINEAR TOPOGRAPHIC ROSSBY WAVES

In this section we consider perturbations of the rest state for a rotating inviscid ocean of variable depth. The motions to be considered will be characterized as having a large horizontal extent when compared to the maximum water depth and therefore use is made of the hydrostatic approximation. The curvature of the earth is ignored and we will denote by f the local vertical component of the earth's rotation vector—the Coriolis parameter which we take to be a constant.



The equations linearized about the rest state are (cf. LeBlond and Mysak(1978))

$$\begin{aligned} u_t - fv &= -g\eta_x, \\ v_t + fu &= -g\eta_y, \\ \eta_t + (hu)_x + (hv)_y &= 0, \end{aligned} \tag{20}$$

where u, v are the perturbation velocity components in the x, y directions, $h(x, y)$ measures the undisturbed water depth and $\eta(x, y, t)$ measures the departure of the free surface from the rest state.

These are easily reduced to the following set

$$\begin{aligned} \partial_t\left[(\partial_{tt} + f^2)\eta - g\nabla\cdot(h\nabla\eta)\right] - gf\mathcal{J}(h,\eta) &= 0, \\ (\partial_{tt} + f^2)u &= -g(\partial_{xt} + f\partial_y)\eta, \\ (\partial_{tt} + f^2)v &= -g(\partial_{yt} - f\partial_x)\eta, \end{aligned} \tag{21}$$

where $\mathcal{J}(h, \eta) = h_x\eta_y - h_y\eta_x$.

Even in the case of a flat ocean floor when $h(x, y) =$ constant, the equations admit wave solutions. These *gravity waves* have relatively short periods which are $\leq 1/f$ and are not of interest in the current study. It is convenient to eliminate them from the start and to concentrate on the longer period waves that are only seen in the presence of a non-trivial topography. A scaling analysis shows that for motions with long temporal periods (typically 50 or more days) the ∂_{tt} terms in (21) are negligible. Moreover, for the periodic boundary data under consideration, the equations for u and v decouple entirely from the η equation allowing us to concentrate on

$$\partial_t\left[f^2\eta - g\nabla\cdot(h\nabla\eta)\right] - gf\mathcal{J}(h,\eta) = 0. \tag{22}$$

If L is a characteristic horizontal length and D is say the maximum undisturbed ocean depth we can introduce non-dimensional (starred) variables as follows

$$\begin{aligned} x = \frac{L}{2\pi}x^*,\ y = \frac{L}{2\pi}y^*,\ h = Dh^*,\ \eta = D\eta^*, \\ t = f^{-1}t^*,\ u = \frac{Lf}{2\pi}u^*,\ v = \frac{Lf}{2\pi}v^*, \end{aligned} \tag{23}$$

and arrive at the following equation for η^*

$$\partial_{t^*}\left[\rho_T^2\eta^* - \rho_L\nabla^*\cdot(h^*\nabla^*\eta^*)\right] - \mathcal{J}^*(h^*,\eta^*) = 0. \tag{24}$$

All derivatives are now taken with respect to the non-dimensional variables while

$$\rho_T = \frac{\sqrt{L/2\pi g}}{f^{-1}},\quad \rho_L = \frac{D}{L/2\pi} \tag{25}$$

are small non-dimensional time and length ratio parameters respectively. From now on we shall drop the stars and all references will be to the non-dimensional equations and variables.

It is our intent to solve (24) for random, periodic $h(x,y)$. Note that it is the derivatives of this random function that are important in the current context. To render the problem computationally tractable we will only consider the case where $h = h(y)$. A normal mode decomposition of the following form is then employed

$$\eta(x,y,t) = \hat{\eta}(y)e^{i\alpha x}e^{i\sigma t} \tag{26}$$

yielding the equation

$$\sigma\left[(\rho_T^2 + \alpha^2\rho_L h)\hat{\eta} - \rho_L(h\hat{\eta}')'\right] + \alpha h'\hat{\eta} = 0. \tag{27}$$

where the prime denotes the derivative with respect to y.

Properties of the Governing Equation

Introducing new parameters

$$\lambda = -\alpha/\rho_L\sigma, \qquad \rho = \rho_T^2/\rho_L. \tag{28}$$

equation (27) becomes

$$\frac{d}{dy}\left[h(y)\frac{d\hat{\eta}}{dy}\right] + \left[\lambda h'(y) - (\alpha^2 h(y) + \rho)\right]\hat{\eta} = 0. \tag{29}$$

which is to be solved with periodic boundary data. In this format the equation is similar to a periodic Sturm Liouville system (see for example Birkhoff and Rota (1978)) except for the important fact that $h'(y)$, the coefficient multiplying the eigenvalue, is not necessarily positive. Nevertheless, many of the results for Sturm-Liouville systems still apply. For example we can easily prove the following orthogonality theorem.

Theorem:

Eigenfunctions corresponding to different eigenvalues are orthogonal with respect to dh i.e. if $\hat{\eta}^{(1)}$ and $\hat{\eta}^{(2)}$ are eigenfunctions belonging to distinct eigenvalues $\lambda^{(1)}$ and $\lambda^{(2)}$ then

$$\int_0^{2\pi} \hat{\eta}^{(1)}(y)\hat{\eta}^{(2)}(y)\, h'(y)dy = 0 \tag{30}$$

Proof: Define the operator L by

$$L[\hat{\eta}] = \frac{d}{dy}\left[h(y)\frac{d\hat{\eta}}{dy}\right] - (\alpha^2 h(y) + \rho)\hat{\eta}. \tag{31}$$

Then

$$L\left[\hat{\eta}^{(i)}\right] = -\lambda^{(i)} h'(y)\hat{\eta}^{(i)} \text{ for } i = 1, 2. \tag{32}$$

It is easily verified directly that

$$\hat{\eta}^{(1)} L\left[\hat{\eta}^{(2)}\right] - \hat{\eta}^{(2)} L\left[\hat{\eta}^{(1)}\right] = \frac{d}{dy}\left\{h(y)\left[\hat{\eta}^{(1)}(y)\frac{d\hat{\eta}^{(2)}}{dy} - \hat{\eta}^{(2)}(y)\frac{d\hat{\eta}^{(1)}}{dy}\right]\right\}. \tag{33}$$

Integrating from 0 to 2π substituting for the operators on the left hand side yields

$$\left(\lambda^{(1)} - \lambda^{(2)}\right)\int_a^b h'(y)\,\hat{\eta}^{(1)}(y)\hat{\eta}^{(2)}(y)\,dy = h(y)\left[\hat{\eta}^{(1)}(y)\frac{d\hat{\eta}^{(2)}}{dy} - \hat{\eta}^{(2)}(y)\frac{d\hat{\eta}^{(1)}}{dy}\right]_{y=0}^{2\pi} \tag{34}$$

which is zero due to the periodic nature of the coefficients and the eigenfunctions. Hence if the eigenfunctions are distinct we have the orthogonality result.

Other properties of interest include

- Eigenpairs come in conjugates i.e. if $(\lambda, \hat{\eta}(y))$ is an eigenpair so also is $(\lambda^*, \hat{\eta}^*(y))$.
- The eigenvalues λ and thus the wave speeds σ are purely imaginary. Thus the waves are not dissipative.
- The number of zeros in the eigenfunctions increases as σ decreases.

The latter result has both physical and computational significance. As was mentioned earlier, the topographic waves of greatest importance are those with the largest wavelengths. In the x-direction this concern with wavelength causes us to pay particular attention to small values of the wavenumber parameter α. By the same token we wish to characterize the eigenfunctions $\hat{\eta}(y)$ according to the number of oscillations they make in the non-dimensional interval $y \in [0, 2\pi[$ and concentrate on those that have the fewest oscillations, and thus the fewest zeros in that domain. This idea is depicted in figure (5).

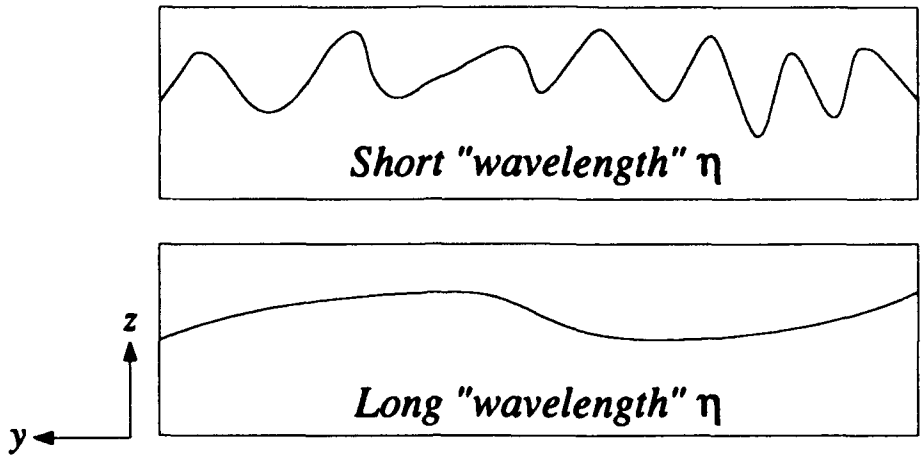


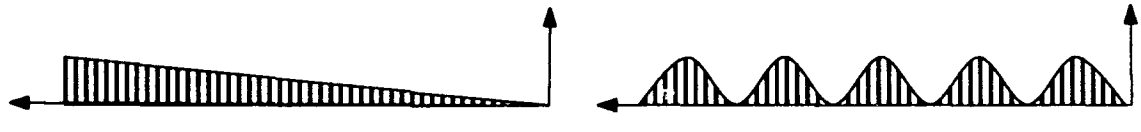
Figure 5 "Long" and "short" wavelength solutions.

This ability to label the eigenfunctions is crucial in the computational setting where the bottom profile is modeled by a random process. Each different realization of the bottom topography yields a different spectrum of σ's. It is only by labelling the σ's by the number of zeros in matching $\hat{\eta}$'s that we can do any sort of reasonable statistical analysis on the dispersion relations. Essentially it allows us to compare like with like from run to run.

The details of the proofs of these and other properties of a mathematical nature will be published later. We note that in particular we can deduce some asymptotic results for the Lyapunov exponent and rotation number associated with this equation when the bottom profile $h(y)$ is a piecewise linear curve such that the slope $h'(y)$ is a "telegraphic" random process—formally this is a stationary ergodic Markov process where the slope switches between two states $\{+H, -H\}$ at nodes that are exponentially distributed along the y axis.

Prior Results

Most of the previous work on this problem was done for deterministic bottom profiles. In particular the two profiles shown in the figure below were investigated by a number of researchers and we mention a couple of relevant results from those studies now.



In the case of a small constant slope profile

$$h(y) = [1 - \epsilon y] \tag{35}$$

the following quantized set of dispersion relations are easily found (see for example Pedlosky)

$$\sigma_n(\alpha) = i\frac{\epsilon}{\rho_L}\left[\frac{\alpha}{n^2+\alpha^2+\rho_T^2}\right] \tag{36}$$

The corresponding topographic waves propagate along the positive x-direction

For the second case in the figure, that of a small purely sinusoidal bottom profile with period $2\pi/\mu$

$$h(y) = 1 - \epsilon \sin \mu y \tag{37}$$

the governing equation is of the Hills type. For small α we get periodic solutions (periods $2\pi/n$). Rhines and Bretherton (1973) found that asymptotically ($\rho_T = 0$)

$$\sigma_n(\alpha) = \pm i\frac{\epsilon}{\rho_L}\frac{1}{\sqrt{(n/\mu)^2+\alpha^2}} \tag{38}$$

Topographic waves propagating in both directions along the corrugations of the bottom profile. This is not surprising as the bottom slope varies periodically from positive to negative. This result can serve as a useful test of the numerics.

Numerical Simulations

Next we turn to numerical simulations carried out for other bottom profiles. Having expressed everything is in terms of nondimensional coordinates, we assume that the bottom topography is a periodic extension of the fundamental interval $y \in [0, 2\pi)$ and use the Fourier series expansions

$$h(y) = \sum_{-\infty}^{\infty} \hat{h}_k e^{iky}, \quad \hat{\eta} = \sum_{-\infty}^{\infty} \hat{\hat{\eta}}_k e^{iky} \tag{39}$$

to reduce the differential eigen-problem (27) to the generalized algebraic eigen-problem

$$A\hat{\hat{\eta}} = i\lambda B\hat{\hat{\eta}} \tag{40}$$

where

$$\begin{aligned} A_{jk} &= (\alpha^2 + jk)\hat{h}_{j-k} + \rho\delta_{jk} \\ B_{jk} &= (j-k)\hat{h}_{j-k} \end{aligned} \tag{41}$$

It is convenient to introduce the bottom profile by the relation

$$h(y) = 1 - \epsilon b(y). \tag{42}$$

In terms of the Fourier coefficients of this profile we have

$$\begin{aligned} A_{jk} &= (\alpha^2 + jk)\hat{b}_{j-k} - \epsilon^{-1}(\alpha^2 + jk + \rho)\delta_{jk} \\ B_{jk} &= (j-k)\hat{b}_{j-k} \end{aligned} \tag{43}$$

The eigenvalues produced from equation (40) depend on all the parameters in the problem

$$\lambda = \lambda\left(\alpha, \epsilon, \rho, \left\{\hat{b}_j\right\}\right) \tag{44}$$

and also on the resolution chosen for the eigenfunction and the bottom profile. The actual values were produced using the standard QZ algorithm on the generated A, B matrices.

5. SIMULATING RANDOM BOUNDARIES

Most of the results presented in this paper are for simulated models of a rough ocean floors. There is of course an element choice in the way one simulates rough surfaces. Various methods are discussed by Oglivy (1991). Our own choice was motivated by both practical and theoretical considerations:

- It is natural to refer vertical distances to the maximum depth of the undisturbed ocean, as was done above in the non-dimensionalization process. Consequently we want $b_k \geq 0$ for all k. This condition is ensured by using an exponential distribution.
- There is strong evidence from experimental bathymetry data that the floor of the ocean is non-Gaussian (see for example, Dworski and Holloway (1983)). Indeed our simulated profiles are somewhat reminiscent of experimental measurements.
- From a theoretical point of view, it is desirable to work with a process for which all of the moments are finite as is the case for the exponential distribution. Although this does not play a major role here, several theoretical statistical results for moving average processes of the type described below only hold under the assumption that the moments of higher order exist (see for example, Grenauder and Rosenblatt (1956)).

The principal tool we have used to produce synthetic bottom profiles are wide-sense, discrete-"time", stationary stochastic processes where at any point $y_k = k\Delta y$ in physical space the bottom boundary is represented by the moving average

$$b(y_k) \equiv b_k = \sum_{j=-\infty}^{\infty} a_j V_{k-j}, \qquad k = 0, 1, 2, \ldots \tag{45}$$

The V_j were chosen to be independent random variables having a common exponential distribution function so that the probability that V_j is less than v is given by

$$P(V_j \leq v) = 1 - e^{-v}. \tag{46}$$

The infinite sums must be truncated for computations and in this paper we take as weights

$$a_j = \begin{cases} 1/W & \text{for } j = 0, 1, \ldots, W-1, \\ 0 & \text{otherwise} \end{cases} \tag{47}$$

where W is the *averaging width*. Thus a set of N_b points were generated according to the prescription

$$b_k = \sum_{j=k}^{k+W-1} \frac{1}{W} V_j \text{ for } k = 0, 1, \ldots, N_b - 1. \tag{48}$$

The Fourier transform of these values then gives the $\hat{b}_k$ coefficients that are used to produce the A and B matrices above.

Note that in practice the quantities of interest are ϵV_j which are also exponentially distributed, but with parameter $1/\epsilon$ so that

$$P(\epsilon V_j \leq v) = P\left(V_j \leq \frac{v}{\epsilon}\right) = 1 - e^{-\frac{1}{\epsilon}v}. \tag{49}$$

Then the mean and variance for the corresponding ϵb_k's are

$$E\{\epsilon b_k\} = \epsilon, \quad \text{Var}\{\epsilon b_k\} = \epsilon^2 \tag{50}$$

while the correlations are given by

$$\text{Cor}(b_j, b_k) = \begin{cases} 1 - (j-k)/W & \text{if } |j-k| \leq W, \\ 0 & \text{otherwise.} \end{cases} \tag{51}$$

Larger values of the parameter ϵ increase the mean value of the bottom profile and also the deviations from that mean while increasing W makes points on the boundary more correlated and tends to smooth it out. This is observed in the figure (6) which shows profiles $\epsilon b(y)$ for some different values of ϵ and W. In each case $N_b = 256$. The number τ reported on each graph is the geometric temperature which was discussed earlier. We point out that larger values of τ are clearly associated with "wilder" boundaries.

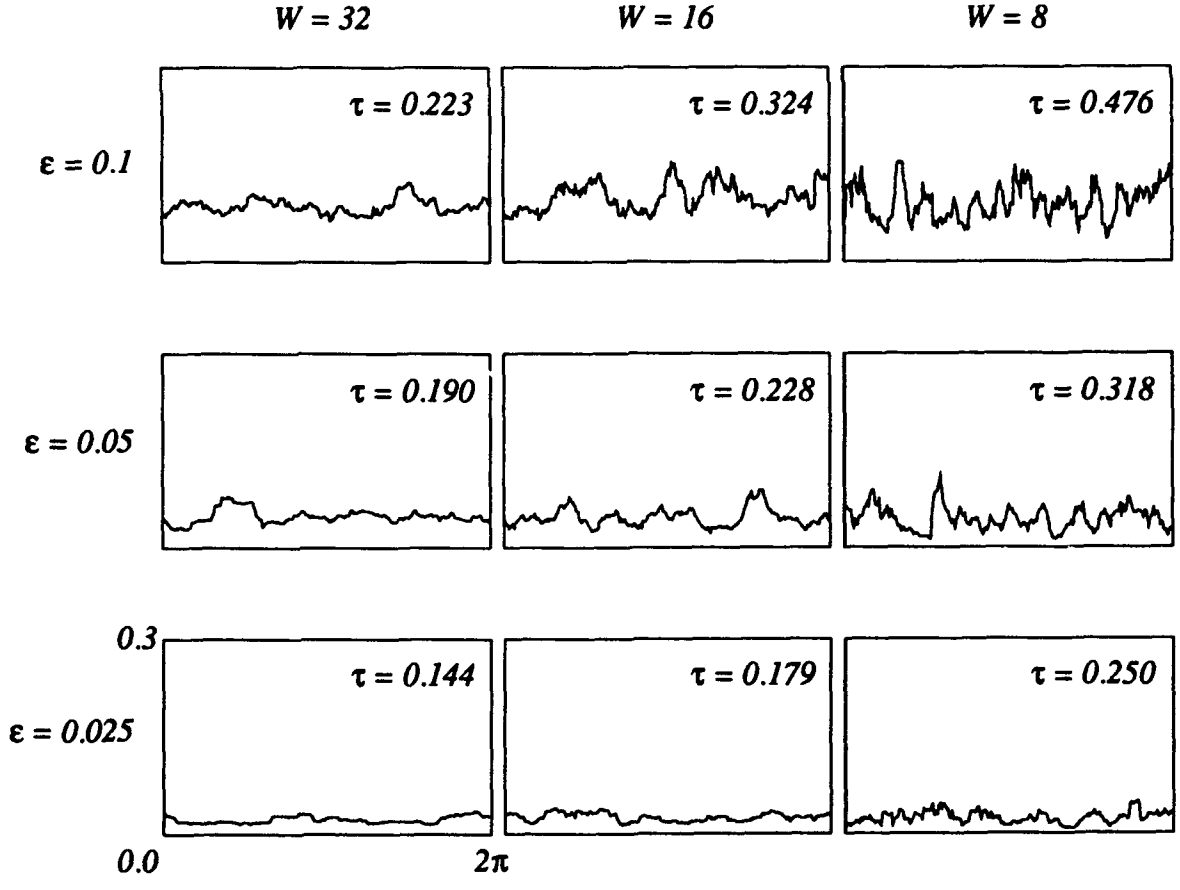


Figure 6 Some realizations of bottom profiles for different values of ϵ and W with $N_b = 256$. The scale on each is identical to that shown on the lowest left graph.

6. DISPERSION RELATION RESULTS

In this section we take up the problem of numerically solving the algebraic eigenvalue problem (40). Note that the input function $b(y)$ is real valued and thus $\hat{b}_j = \hat{b}^*_{-j}$ where the superscript star denotes the complex conjugate. Using this, it is easy to show that the matrix A is Hermitian while the matrix B is skew Hermitian. Actually, for the results presented in this section, we also assumed that the bottom profile is symmetric, $b(y) = b(-y)$. The matrices are then real which simplifies the numerical calculation of the eigenvalues somewhat.

The eigenvalues can be considered as functions of all the parameters in the problem, $\lambda = \lambda\left(\alpha, \epsilon, \rho, \left\{\hat{b}_j\right\}\right)$. The $\hat{b}_j$ depends the floor data $b_k \equiv b(y_k)$ which in turn are determined by the averaging width W described in the previous section. In real long wave flows the parameter ρ is tiny and we have taken it to be zero in all our simulations. Therefore $\lambda = \lambda(\alpha, \epsilon, W)$.

There are numerical resolution parameters to be considered also—how many Fourier modes, or equivalently how many points in physical space, are used to resolve the bottom boundary and the disturbance $\hat{\eta}(y)$? The point of view we have taken is that if N_b modes are used for $b(y)$ then one should increase the number of modes used for $\hat{\eta}(y)$ until convergence is seen. In our study several such resolution studies were performed. To capture the "longest" mode (the $\hat{\eta}(y)$ with the fewest zeros) it was found that using an expansion with for $\hat{\eta}(y)$ with N_b modes was always more than adequate. Typical values of N_b were $128, 256,$ and 512. Note that extracting the eigenvalues is $O(N^3)$ process so going to higher resolutions is prohibitively expensive.

For each choice of the parameters ϵ and W one can generate many realizations of a bottom topography, each of which has approximately the same thermodynamic properties. Figure (7) shows how the temperature of the bottom profile changes for twenty different realizations each for $\epsilon = 0.025, 0.050, 0.075$ and $W = 4$. In practice each trial corresponds to choosing a fresh seed for a random number generator. It is clear from the plot that increasing ϵ guarantees an increase in τ although there is also some variability from realization to realization.

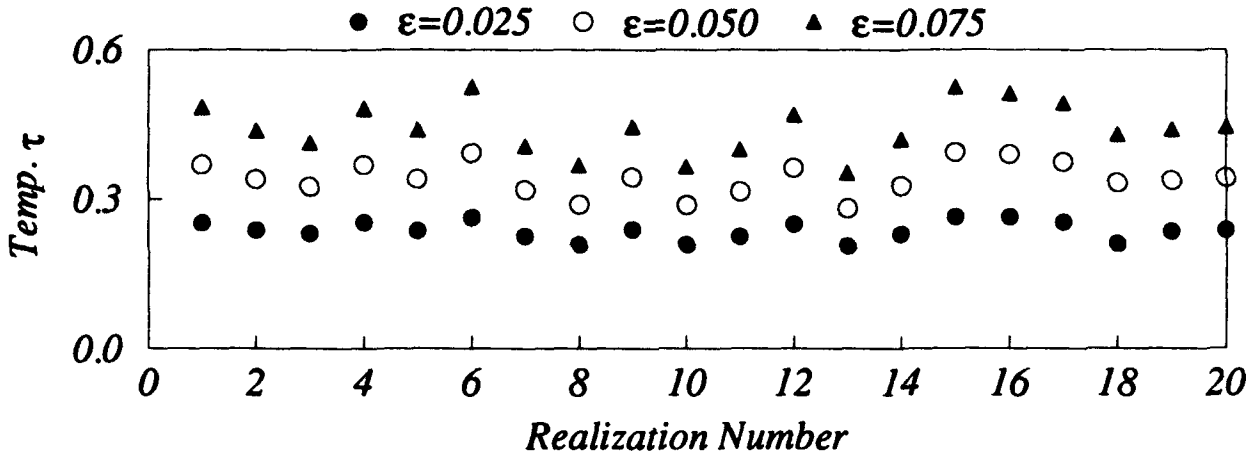


Figure 7 The temperature of twenty different realizations of bottom profiles for three different values of ϵ. In each case W=4.

Although for fixed values of ϵ and W the curve temperature remains fairly close to some constant value there can be quite a range for the curve ordinates. This is depicted in (8) which shows the mean, and the upper and lower bounds found for $\epsilon b(y_k)$ over 20 realizations, each with $\epsilon = 0.05$, $W = 4$. Also clearly visible in this plot is the symmetry assumption mentioned earlier. That assumption will be removed in a later paper. Also note the mean profile has $\epsilon b(y) \approx \epsilon$ as we would expect.

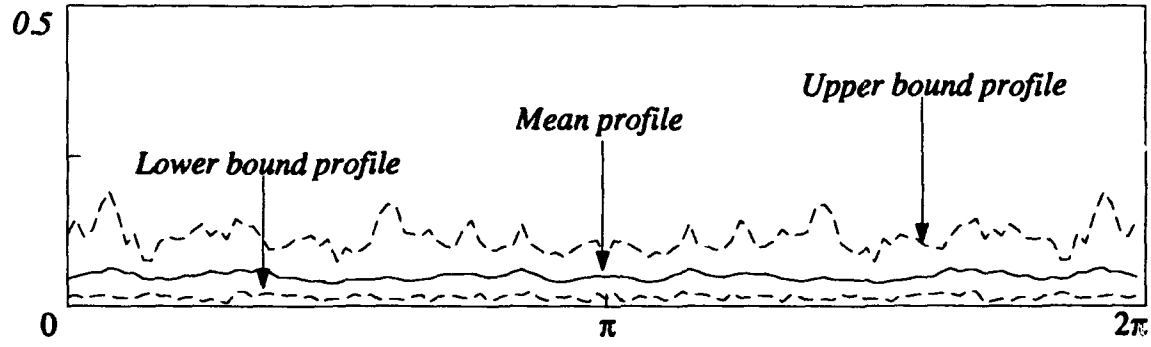


Figure 8 Average and extreme values of the bottom profile for ϵ=0.05, W=4.

For each individual bottom profile we look at a range of wavenumbers α, fill the matrices A, B and solve the eigenvalue problem. The eigenvalues are sorted according to their size and the largest ones are output—we already know that the corresponding eigenfunction will have the fewest zeros and thus correspond to the largest wave. We then can make a plot of the *dispersion relation* which shows the wave speed σ as a function of the x-wavenumber α.

Large numbers of eigenvalue problems are tackled in this process. The total number can be expressed as $N_W N_\epsilon N_\tau N_\alpha$ where the four N's respectively represent the number

of averaging widths tried, the number of ϵ's used, the number of realizations generated, and the number of axial wavenumbers investigated per realization. Many of these runs are independent so if a distributed network of workstations is available they can be used with good effect to reduce the computational burden.

In figure (9) we show the mean value of the dispersion curve (for the longest wave) found for three different values of ϵ. In each case $W = 4$ and runs were done for 20 realizations of the bottom topography. Also reported on the graph is the mean value of the temperature of the bottom in each of the cases. Clearly as the temperature rises so also does the mean value of the wave speed σ.

A more detailed statistical study of the variation of σ with τ will the subject of another paper. However, it is clear that there is a correlation between the roughness parameter and the predicted wave speed of long waves.

Of course there is also some variability in the computed dispersion relations for different realizations at a fixed value of ϵ. In figure (10) plots are shown of the minimum and maximum eigenvalues found across all the bottom profiles run. This is done for $\epsilon = 0.025$ and 0.050 corresponding to profiles with temperatures close to $\tau = 0.24$ and $\tau = 0.34$ respectively. The region enclosed by minimum and maximum plots on the left is clearly different from that enclosed on the right.

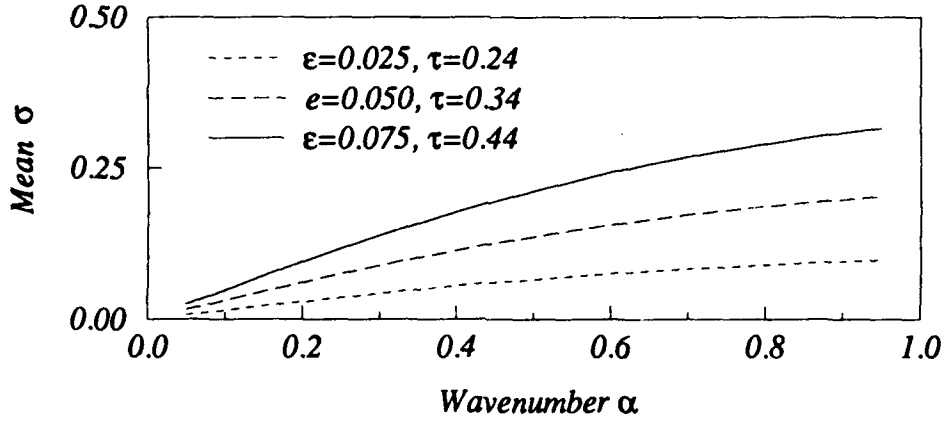


Figure 9 The dispersion relation for the longest wave averaged over 20 different realizations with the same ϵ. W=4 in each case.

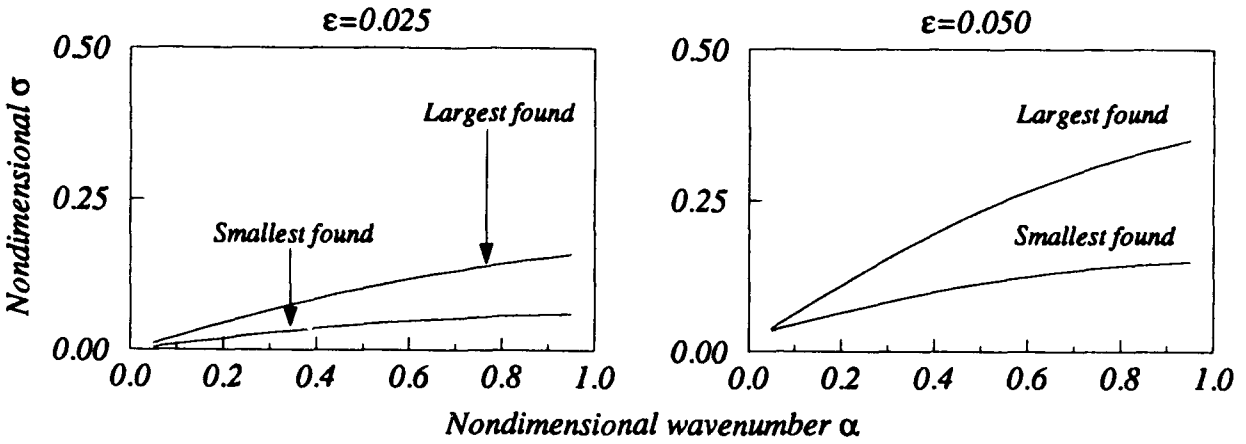


Figure 10 The range of values found for the eigenvalues over all the runs for two different values of ϵ.

7. SUMMARY

We have presented a method for quantifying roughness of curves and surfaces that is easily implemented for experimental data sets. In contrast to fractal analyses, the technique is based on probabilistic concepts that validly apply to data sets of finite resolution. Intuition as to the meaning of the temperature of a curve was developed by means of simple examples, and efficient algorithms and implementations were discussed for more realistic data.

While there is no doubt that geometric thermodynamics is a useful identification tool it is an open question as to whether it can be used in a predictive fashion in oceanography. To that end we are currently studying the testbed problem of topographic waves in an ocean of random depth and have presented some early results in this paper for synthesized bottom profiles. It will be interesting to make use of real bathymetric data in these simulations also.

The coefficients in the governing equations are the derivatives of random functions and are therefore not necessarily small. One question that we are now investigating is whether it possible to replace a complex boundary with a much simpler one having the same "average" slope where that quantity is proportional to $U = 2|\Gamma|/|\partial \mathrm{K}|$. Unfortunately, as was mentioned at the end of section 2, U by itself is insensitive to some features of a boundary that we would not expect the flow to be insensitive to. Work on this and other points is ongoing and we will present a more detailed analysis in a future paper.

Acknowledgment

The authors thank Skip Carter of the Naval Postgraduate School in California, and Greg Holloway of the Institute for Ocean Sciences in Sydney, British Columbia, for

providing access to bathymetric data sets. It was a pleasure to plot data tracks on a snowy February morning in Cleveland, Ohio and dream of cruises from Pago-Pago to Honolulu.

REFERENCES

Birkhoff, G. and Rota, G. C., 1978: *Ordinary Differential Equations* John Wiley.

Blaschke, W., 1936: *Vorlesungen Uber Integral Geometrie* Tuebner.

DuPain, Y., Kamae, T., and Mendes-France, M., 1986: Can one measure the temperature of a curve, *Archive for Rational Mechanics and Analysis*, 94, 155–163.

Dworski, J. and Holloway, G., 1983: Statistical representation of seafloor topography, Special Report 94, School of Oceanography, University of Washington.

Graham, R., 1972: An efficient algorithm for determining the convex hull of a finite planar set, *Info. Proc. Lett.*, 1, 132–133.

Grenauder, V. and Rosenblatt, M., 1956: *Statistical Analysis of Stationary Time Series* Almqvist and Wikjell.

LeBlond, P. H. and Mysak, L. A., 1978: *Waves in the Ocean* Elsevier.

Mann, J., Rains, E., and Woyczynski, W., 1991: Measuring the roughness of interfaces, *Chemometrics and Intelligent Laboratory Systems*, 12, 169–180.

Mysak, L. A., 1978: Wave progagation in random media with oceanic applications, *Reviews of Geophysics and Space Physics*, 16, 233–261.

Odulo, A. B. and Pelinovsky, Y. N., 1978: Effect of random inhomogeneities of ocean bottom relief on the propagation of Rossby waves, *Oceanology*, 18.

Oglivy, J., 1991: *Theory of Wave Scattering from Random Rough Surfaces* Adam Hilger.

Pedlosky, J., 1989: *Geophysical Fluid Dynamics* Springer-Verlag.

Preparata, F. and Shamos, M., 1985: *Computational Geometry* Springer–Verlag.

Rhines, P. and Bretherton, F., 1973: Topographic Rossby waves in a rough bottomed ocean, *Journal of Fluid Mechanics*, 61, 3.

Santalo, L. A., 1976: *Integral Geometry and Geometric Probability* Addison Wesley.

Steinhaus, H., 1954: Length, shape and area., *Colloquium Mathematicum*, 3, 1.

Thomson, R. E., 1975: The propagation of planetary waves over random topography, *Journal of Fluid Mechanics*, 70.

NON-EDDY RESOLVING MODEL OF β-PLANE TURBULENCE

Boris Galperin
Department of Marine Science, University of South Florida, St. Petersburg, FL 33701

Semion Sukoriansky
Department of Mechanical Engineering, Ben Gurion University of the Negev,
Beer Sheva 84105, Israel

Steven A Orszag
Applied and Computational Mathematics, Princeton University, Princeton, NJ 08544

Ilya Staroselsky
Cambridge Hydrodynamics, Inc., P.O. Box 1403, Princeton, NJ 08542

ABSTRACT

In this paper, we develop grid-scale dependent parameters, including eddy viscosity and eddy β, for non-eddy-resolving simulations of β-plane turbulence. These eddy parameters account for the effect of subgrid scale (SGS) turbulence and Rossby waves on the resolved scales of motion and are derived in a self-consistent framework provided by the renormalization group (RG) theory of turbulence. The RG formalism allows a coarsened description of a strongly non-linear system with a very large number of degrees of freedom, or wave numbers in Fourier space, by successive elimination of small shells of wave numbers corresponding to unresolved scales. The resulting equation of motion for large, resolved scales is structurally similar to the initial equation but its dimensional parameters, viscosity and β, are rescaled, or renormalized, and depend on the wave number. In the resulting description of β-plane turbulence, the flow field at relatively small scales, or large wave numbers k, has behavior typical of 2-D turbulence; however, as $k \to 0$, the β-effect becomes significant and the flow characteristics develop strong anisotropy. The energy transfer, energy spectra and two-parametric viscosity and β are calculated in the energy sub-range. At relatively large k the energy spectrum is isotropic and follows the Kolmogorov ($-\frac{5}{3}$) law; with decreasing k the spectrum becomes substantially anisotropic as the energy is preferentially transported into zonal flows and zonally propagating Rossby waves that develop a power law with exponent approximately $-\frac{7}{2}$ in the zonal direction. We conclude that the anisotropization of the energy transfer is associated with the mechanism of generation and maintenance of mean zonal flows and radiation of zonally propagating Rossby waves by non-linear interactions. It is argued that the two-parametric viscosity and β should be used as eddy viscosity and eddy

β in non-eddy-resolving simulations of β-plane turbulence. In physical space, the large-scale dynamics is described by a Kuramoto–Sivashinsky-type equation, which includes a negative (destabilizing) Laplacian, positive (stabilizing) biharmonic friction term, and, possibly, higher order hyperviscosities. Being numerically stable, this equation naturally incorporates negative viscosity phenomena. The results and their implications in the context of non-eddy-resolving modeling in geophysical fluid dynamics are discussed.

1. INTRODUCTION

The rapid development of computer technology has enabled ever increasing resolution in ocean circulation models. At the present time, there exist oceanic general circulation models (OGCM) with grids as small as $\frac{1}{4}^\circ$ that are capable of resolving mesoscale eddies, i.e., processes at the scales of the local deformation radius. Such eddies play a key role in transport of vorticity, mass, salt, and heat in the world ocean. The feasibility of eddy-resolving modeling of the global ocean circulation has been demonstrated by Semtner and Chervin (1988). Recently, they presented results of extensive eddy-resolving simulations of the world ocean executed on the largest supercomputer available at that time (Semtner and Chervin, 1992). They found that even with marginal resolution at high latitudes, the simulated three-dimensional fields and major features of the global circulation, such as western boundary currents and the Antarctic Circumpolar Current were quite realistic. On the other hand, other key features of the global circulation, such as separation of the Gulf Stream are still not faithfully captured by present eddy-resolving models (Haidvogel et al., 1992). Semtner and Chervin (1992) suggest that further improvement in the modeling of ocean circulation, particularly in the areas of high variability, will be achieved with increased resolution; as the next step, simulations with resolution of $\frac{1}{8}^\circ$ are envisioned.

Eddy-resolving simulations, as well as the observational data summarized in Stammer and Böning (1992) indicate that processes on the scales of the local deformation radius are crucial for ocean dynamics and should be adequately represented. Semtner and Chervin (1992), however, propose an even more conservative resolution criterion that includes a part of the inertial subrange.

Consideration of the inertial subrange opens the possibility of a turbulence-based subgrid scale (SGS) parameterization for models of large-scale circulation (see the review by Holloway, 1989). Sensitivity of OGCM predictions to the SGS parameterization has been well documented (see, for instance, Bryan, 1987). Semtner and Chervin (1992) note significant changes in their results upon replacing the classical Laplacian friction by a biharmonic one. They also mention the importance of effects related to the phenomenon of "negative viscosity." Among the reasons why SGS parameterization is so important are, first, that a considerable amount of energy is concentrated in subgrid scales (see, for instance, Holloway, 1992, who estimated that the energy flux due to subgrid topographic effects on scales of the order

300 m is comparable in magnitude with those due to other sources), and, second, that the inverse energy transfer developing in large-scale, quasi-two-dimensional, turbulent flows facilitates an efficient energy exchange between SGS and resolved eddies. The inverse energy transfer arising from nonlinear interactions is intimately related to negative viscosity phenomena (Kraichnan, 1976). On the other hand, existing models of the large-scale circulation parameterize the SGS effects by highly dissipative operators that are designed to dissipate the large-scale energy and ensure numerical stability, but cannot properly account for the complex interaction between resolved and unresolved modes.

The problem of SGS parameterization is one of the hardest in geophysical modeling yet it cannot be resolved by a mere increase in resolution which quickly becomes computationally prohibitive even for the fastest supercomputers. This problem becomes even more acute in climate models that operate on very long time scales; in these models, any increase in resolution must come at the expense of other, possibly more important information (see the discussions in Holloway, 1992, and elsewhere in this volume). Therefore, along with further development of eddy resolving models, one needs to invest more effort in the better representation of SGS processes. The latter line of research should not only improve the performance of eddy resolving models but should also allow one to develop a generation of non-eddy-resolving models in which SGS parameterization extends up to, and possibly beyond, the scales of the local deformation radius thus relaxing resolution requirements. If such models incorporate a computationally efficient algorithm for the calculation of SGS parameters, they should become a valuable resource for modeling large-scale ocean and atmosphere circulations, coupled atmosphere-ocean systems, and climate. A non-eddy-resolving model of this kind, based upon the renormalization group theory of turbulence, is described in the present paper.

To account for non-local interactions typical of 2D turbulence, it is convenient to operate in Fourier space. However, spectral closure methods, already complicated in the case of purely 2D turbulence, quickly become intractable when spectral anisotropy and/or waves are added to the picture (for a review of spectral closures, see Orszag, 1977; Lesieur, 1990; Herring and Kerr, 1993). To circumvent some of these problems, the ideas of thermal equilibrium statistical mechanics have been utilized by some authors (Salmon et al., 1976; Holloway, 1992, 1993; Griffa and Salmon, 1989; Griffa and Castellari, 1991). According to these ideas, a non-linear, dissipative and forced real system is replaced by a non-dissipative and unforced system that is allowed to reach thermal equilibrium, i.e., a state of maximum entropy. In this state, the system characteristics can be calculated using variational principles (Salmon et al., 1976; Robert and Sommeria, 1991a,b, 1992). Although the state of thermal equilibrium is never achievable in reality, the tendency toward this state governs the evolution of real systems (Rose and Sulem, 1978; Kraichnan and Montgomery, 1980). The practical use of the thermal equilibrium approach for purposes of large-scale oceanographic modeling was outlined and implemented by Holloway

(1992, 1993). He calculated the barotropic thermal equilibrium stream function that reflects topographic effects based upon the variational principle of maximum entropy, and then required that the barotropic stream function calculated by a fully forced OGCM relax towards the thermal equilibrium value. Such an approach improved the representation of topographic effects in the GFDL OGCM (see articles by Holloway and Eby and Holloway in this volume).

In this paper, an application of a different statistical mechanical approach to oceanographic modeling is described. This approach is formulated for a fully non-linear, forced and dissipative system and is based upon the (renormalization group) RG theory. The RG theory has been particularly successful in the description of large-scale, long-time behavior of systems associated with phase transitions and critical phenomena (see Wilson and Kogut, 1974; Ma, 1976; Amit, 1978). Applied to a strongly non-linear system with a very large number of degrees of freedom, the RG theory allows one to "coarsen" the description of the system by "mapping" it onto a system with a significantly reduced number of modes. This reduced system is described by an equation structurally identical to the one describing the initial system, but its dimensional parameters (such as viscosity) are renormalized, or multiplicatively rescaled in terms of the scales being removed. The final product of this approach resembles the traditional eddy viscosity parameterization, but the eddy parameters emerging from it are calculated by a self-consistent algorithm based upon the physics of the problem.

The RG techniques were first applied to artificially forced fluid turbulence (see Forster et al., 1977), and then extended to realistic 3D turbulence (Yakhot and Orszag, 1986). Since then, the RG methods have been widely used in a variety of applications, including simulations of transitional, incompressible and compressible flows, turbulent combustion and derivation of $k - \epsilon$ turbulence transport models. A review of various applications of the RG methods can be found in Galperin and Orszag (1993). In analogy to its applications in theoretical physics, the RG technique for turbulence allows the study of the large-scale, long-time behavior of turbulent flow fields and their correlation functions. In addition, the RG formalism can be used for the derivation of SGS models in both 3D and 2D flows. The RG-based spectral closures are generally simpler than those obtained in other theories, which is a promising feature for efficient SGS parameterization. The present paper describes one of the first applications of the RG technique to geophysical flows and considers the large-scale, long-time behavior of β-plane turbulence, as well as its SGS parameterization. The flow chosen is one of the simplest "building block" geophysical flows and has been quite well studied theoretically and numerically. Being relatively simple, it combines features which have hindered the application of spectral closures to such kinds of flows in the past: spectral anisotropy and Rossby waves. The methodology of RG not only allows one to advance the analytical understanding of this flow, but provides an efficient SGS parameterization that can be used in large eddy simulations and non-eddy-resolving modeling of β-plane

turbulence.

In the next section, we provide the mathematical formulation of the problem based upon the barotropic vorticity equation on the β-plane in physical and Fourier space. In Section 3, we describe the application of the RG procedure and show how the process of small-scale elimination produces rescaled SGS parameters. In Section 4, an analysis of characteristic time scales is given and the regions dominated by 2D turbulence or Rossby waves are identified. Also, anisotropic energy spectra are calculated and analyzed. In Section 5, the SGS (or eddy) parameters are introduced based upon two-point turbulence characteristics. It is shown how the eddy parameters can be derived using the RG theory. Then, in Section 6, we describe the RG-calculated anisotropic spectral energy transfer and relate it to existing information on β-plane turbulence. In Section 7, we show how the RG-based eddy parameters can be used in practical oceanographic simulations. Finally, in Section 8 we summarize the results obtained here.

2. MATHEMATICAL FORMULATION OF THE β-PLANE PROBLEM AND RESULTS OF PREVIOUS RESEARCH

The subject of the present paper is the barotropic vorticity equation on the β-plane:

$$\frac{\partial \zeta}{\partial t} + \frac{\partial \left(\nabla^{-2}\zeta, \zeta\right)}{\partial (x,y)} + \beta_o \frac{\partial}{\partial x}\left(\nabla^{-2}\zeta\right) = \nu_o \nabla^2 \zeta, \tag{1}$$

where ζ is the barotropic vorticity and ν_o is the molecular viscosity; x and y are directed eastward and northward, respectively. The constant β_o is the background vorticity gradient; it describes the latitudinal variation of the vertical component of the Coriolis parameter, f, in the β-plane approximation, $f = f_o + \beta_o y$.

This equation describes one of the simplest "building block" systems relevant to geophysical flows (see, e.g., Pedlosky, 1987). For β_o=0, this equation describes isotropic 2D turbulence (see the reviews by Kraichnan and Montgomery, 1980; Vallis, 1992). In its linearized form, it describes Rossby waves with the dispersion relation

$$\omega = -\beta_o k_x / k^2 + i\nu_o k^2. \tag{2}$$

With the full non-linear terms and $\beta_o \neq 0$, (1) describes the interaction between 2D turbulence and Rossby waves.

Studies of (1) from the point of view of a non-linear system combining turbulence and waves were first reported by Rhines (1975). He found that, at large k, the β-effect is small and the flow behaves largely like 2D turbulence. With decreasing k, the inverse energy cascade terminates and the flow evolves towards the regime

of linear Rossby waves. The transition from turbulence to the Rossby waves dominated regime takes place at wave numbers of the order of $k_\beta = (\beta/2U)^{1/2}$, where U is a measure of velocity fluctuations in the system. Maltrud and Vallis (1991) and Vallis and Maltrud (1993) expressed k_β in terms of the inverse energy cascade rate $\overline{\epsilon}$: $k_\beta = (\beta^3/\overline{\epsilon})^{1/5}$. Rhines (1975) noted that as k_β is approached, the energy transfer not only slows down, but also becomes progressively anisotropic preferring the zonal direction. Later, Holloway and Hendershott (1977) extended studies of β-plane turbulence by the use of the test field model (TFM) of Kraichnan (1971), a spectral closure theory. They reconfirmed the observations of Rhines (1975) about the slowing down of the inverse energy transfer and its anisotropization with a preferred zonal direction. They redefined the transitional wave number k_β in terms of the root mean square vorticity, $\overline{\zeta}$, such that Galilean invariance could be satisfied automatically; in their notation, $k_\beta = \beta/\overline{\zeta}$. Further theoretical accounts can be found in Carnevale and Martin (1982), Salmon (1982), Maltrud and Vallis (1991), and Vallis and Maltrud (1993). In particular, Vallis and Maltrud (1993) noted that k_β is an anisotropic parameter, and derived an analytical expression describing its angular variation. This issue will be revisited in Section 4. Bartello and Holloway (1991) used the TFM framework for analytical and numerical studies of diffusion on the β-plane. Holloway (1986) provided a comprehensive review of the β-plane research.

Turbulence on the β-plane has been extensively studied by numerical experimentation. In all simulations, robust generation of zonal flows has been observed, either in Cartesian (Rhines, 1975; Maltrud and Vallis, 1991; Vallis and Maltrud 1993; Bartello and Holloway, 1991; Panetta, 1993) or spherical (Williams, 1978; Yoden and Yamada, 1993) coordinate systems. Maltrud and Vallis (1991) observed that on the β-plane, large-scale vortices tend to radiate their energy as Rossby waves, a phenomenon that improves the applicability of statistical theories to β-plane turbulence.

It should be emphasized that in these previous studies, the process of spectral transfer anisotropization has not been fully quantified or related to SGS parameterization. Moreover, large eddy simulations (LES) of β-plane turbulence, in which large-scale modes of the flow are resolved, but small scales are parameterized, have never been attempted. In the present paper, we quantify the inverse energy transfer, its anisotropization due to the β-effect, and the interaction between turbulence and Rossby waves, and develop SGS models for LES of β-plane turbulence. This effort can be considered as a case study of non-eddy-resolving modeling, in which SGS eddies are not resolved but their effects are properly incorporated. The results of this "building block" study can be extended for non-eddy-resolving modeling of more realistic oceanic and atmospheric systems.

The RG analysis is performed in wave number-frequency space for an infinite domain. Using the space-time Fourier transform of vorticity, one can derive the

Fourier-space representation of (1):

$$(i\omega + i\beta_o k_x k^{-2} + \nu_o k^2)\zeta(\hat{k}) = \int \frac{\mathbf{k}\times\mathbf{q}}{q^2}\frac{\zeta(\hat{q})\zeta(\hat{k}-\hat{q})}{(2\pi)^{d+1}}d\hat{q}, \tag{3}$$

where d is the dimension of space (d=2 in this study), and $\hat{k}$ and $\hat{q}$ are three-dimensional vectors $(\mathbf{k},\omega)$ and $(\mathbf{q},\Omega)$, respectively.

Here we study a forced system with the forcing concentrated at a high wave number k_o. According to the classical results of Batchelor (1969) and Kraichnan (1967), conservation of energy and enstrophy in the inviscid limit may lead to the development of two inertial sub-ranges: the energy sub-range for $k < k_o$, where energy is transported up-scales and the energy spectrum is the Kolmogorov $E(k) \propto k^{-5/3}$ law, and the enstrophy sub-range for $k > k_o$, where enstrophy is transported down-scales and the energy spectrum is $E(k) \propto k^{-3}$. The subject of the present study is the energy sub-range. Owing to the inverse transfer mechanism, energy is being pumped into ever increasing scales of motion, so that a global steady-state solution is unreachable. However, if the energy injection rate at k_o is constant in time and equals $\bar{\epsilon}$, then after the energy "front" sweeps over a wave number k, a local steady state will develop at k, in which energy will be passing through k with the constant rate $\bar{\epsilon}$. This local statistically steady state will be analyzed here. It can be shown (Orszag et al., 1993a) that the effect of the forcing localized at k_o on the initial equation of motion that resolves all scales is equivalent to the effect of spatially-homogeneous forcing $\xi(\mathbf{k},\omega)$ on the "coarsened" equation that results from the application of the RG procedure. The forcing $\xi(\mathbf{k},\omega)$ is zero-mean, Gaussian, white in time, and its correlation function is

$$\langle \xi(\mathbf{k},\omega)\,\xi(\mathbf{k}',\omega')\rangle = 2D_o k^{-s}(2\pi)^{d+1}\delta(\mathbf{k}+\mathbf{k}')\delta(\omega+\omega'), \tag{4}$$

where s and D_o are parameters to be specified later. If this forcing is inserted into (1) explicitly, after Fourier-transform it results in the following modification of (3):

$$\zeta(\hat{k}) = G^o(\hat{k})\int \frac{\mathbf{k}\times\mathbf{q}}{q^2}\frac{\zeta(\hat{q})\zeta(\hat{k}-\hat{q})}{(2\pi)^{d+1}}d\hat{q} + G^o(\hat{k})\xi(\hat{k}), \tag{5}$$

where $G^o(\hat{k}) = (i\omega + i\beta_o k_x k^{-2} + \nu_o k^2)^{-1}$ is the bare Green function.

Equation (5) is the subject of the RG analysis given below in which the object is to derive an effective equation for large-scale components of ζ. This derivation will be based upon gradually eliminating small-scale components of the flow field with characteristic wave numbers of the order of the dissipation cutoff and moving the dissipation cutoff toward larger scales. In this process, the initially constant molecular viscosity ν_o and β_o get modified, or renormalized, and become functions of the dissipation cutoff.

3. SMALL SCALE ELIMINATION BY THE RG PROCEDURE

As formulated by Yakhot and Orszag (1986) for 3D turbulence and adapted here for 2D turbulence, the RG theory seeks to answer the following question: "How are the long wave length modes $\zeta^<(\hat{k})$ belonging to the interval $0 < k < \Lambda$ affected by the short wave length modes $\zeta^>(\hat{k})$ from a narrow wave vector band $\Lambda - \delta\Lambda < k < \Lambda$?" The answer to this question is obtained by a formal RG procedure. Repeated many times, it allows one to consider the limit $\Lambda \to 0$, corresponding to the large-scale asymptotics; if Λ remains finite, only a finite shell of short wave length modes is eliminated leading to SGS parameterization.

Let us assume that initially (5) is defined on the interval $0 < k < \Lambda_o$, where Λ_o is the dissipation cutoff. Following the Yakhot and Orszag (1986) RG procedure for 3D turbulence, the formal RG procedure for 2D turbulence consists of two steps. First, one introduces a narrow band of wave vectors near the dissipation cutoff, $\Lambda_o - \delta\Lambda_o < k < \Lambda_o$, and the vorticity field and stirring force are decomposed into two parts: "fast" modes $\zeta^>(\hat{k}), \xi^>(\hat{k})$ with wave vectors satisfying $\Lambda_o - \delta\Lambda_o < k < \Lambda_o$, and "slow" modes $\zeta^<(\hat{k}), \xi^<(\hat{k})$ for which $0 < k < \Lambda_o - \delta\Lambda_o$. With this decomposition, (5) becomes:

$$\zeta(\hat{k}) = \lambda_0 G^o(\hat{k}) \int \frac{\mathbf{k} \times \mathbf{q}}{q^2} \left[\zeta^<(\hat{q})\zeta^<(\hat{k}-\hat{q}) + 2\zeta^<(\hat{q})\zeta^>(\hat{k}-\hat{q}) + \zeta^>(\hat{q})\zeta^>(\hat{k}-\hat{q})\right] \frac{d\hat{q}}{(2\pi)^{d+1}} + G^o(\hat{k})\xi(\hat{k}), \tag{6}$$

where λ_0 is the formal expansion parameter introduced for the purpose of developing a perturbative solution to (6); eventually, λ_0 is set to 1. The perturbative solution of the non-dimensionalized (6) involves expansion in a non-dimensional coupling parameter $\overline{\lambda}_0 = \lambda_0 D_o^{1/2}/\nu_o^{3/2}\Lambda_o^{\epsilon/2}$, where $\epsilon = 6 + s - d$. Note that $\overline{\lambda}_0$ is in fact a "local" Reynolds number determined by the "local" viscosity $\nu(\Lambda_o)$ at the "local" wave number Λ_o. After repetitive elimination of small wave number bands $\delta\Lambda_o$ described below, the local Reynolds number remains O(1) because the "local" viscosity increases and Kolmogorov-like viscous scaling holds. Thus, while the RG procedure is based upon expansion in terms of $\overline{\lambda} = O(1)$, the procedure is likely to yield much better results than other methods which employ expansions in terms of the conventional bare Reynolds number which is very large in real flows.

The "fast" modes $\zeta^>(\hat{k})$ are eliminated from (6) through recursive substitution of the formal solution (5) written for $\zeta^>(\hat{k})$. This yields a solution for $\zeta^<(\hat{k})$ in terms of an infinite series in powers of $\overline{\lambda}_0$.

The second step of the RG procedure consists of taking averages over the short wave length modes of the stirring force $\xi^>$. The derivations are given by Yakhot and Orszag (1986), Forster et al. (1977), and in further detail by Smith and Reynolds

(1992). The result is a set of effective dynamical equations for the slow modes with $0 < k < \Lambda_o - \delta\Lambda_o$. This process is then iterated to remove further infinitesimal bands of modes, resulting in a set of ordinary differential equations for ν, β and $\overline{\lambda}$ as functions of k. Particularly important in the RG theory are those solutions for which $d\overline{\lambda}(k)/dk = 0$; a solution of this kind is called a *fixed point* (Amit, 1978; Creswick et al., 1992). In the case of 3D isotropic turbulence, Yakhot and Orszag (1986) showed that at the fixed point $\overline{\lambda} \propto \epsilon^{1/2}$. For the case of randomly stirred flows near thermal equilibrium including 2D flows, similar analysis was performed by Forster et al. (1977). If $\epsilon \to 0$, then also $\overline{\lambda} \to 0$, and the results are exact.

For finite ϵ, the so-called ϵ-expansion (Wilson and Kogut, 1974; Creswick et al., 1992) is applied in which quantities are evaluated asymptotically only to the lowest nontrivial order in powers of ϵ. If ϵ were equal to zero or at least small in real turbulence, the problem of turbulence would then be solvable using expansions in powers of $\overline{\lambda}$ or ϵ. Unfortunately, "real" values of ϵ are not small; indeed, $\epsilon = 4$ for 3D turbulence. Does this fact invalidate the ϵ-expansion? Fortunately, the situation is not so gloomy, although the mathematical justification for the ϵ-expansion procedure for fluid turbulence does not yet exist. Using the ϵ-expansion for 3D turbulence, Yakhot and Orszag (1986) succeeded in calculating many basic constants of turbulence, such as the Kolmogorov and Batchelor constants from first principles of the RG theory. Also, RG-derived turbulence transport models have achieved considerable success in calculations of complex flows in complex geometries (Orszag et al., 1993b). Applying the ϵ-expansion to isotropic 2D turbulence, Staroselsky and Sukoriansky (1993) calculated the Kolmogorov constant which was in good agreement with available data. They also calculated a two-parametric viscosity, which will be discussed later, and found it to be in good agreement with that calculated by Kraichnan (1976) using an entirely different approach, the TFM closure theory. Based upon the previous success in application of the ϵ-expansion technique to both 3D and 2D turbulence, in the present paper we also apply the ϵ-expansion technique to achieve nontrivial results for anisotropic turbulence on the β-plane; in this procedure, only terms up to second order in $\overline{\lambda}$ are kept.

At the first iteration of the scale elimination procedure, the equation for $\zeta^<$ becomes

$$
\begin{aligned}
[G^0(\hat{k})]^{-1}\zeta(\hat{k}) &= \lambda_0 \int \frac{\mathbf{k}\times\mathbf{q}}{q^2}\zeta^<(\hat{q})\zeta^<(\hat{k}-\hat{q})\frac{d\hat{q}}{(2\pi)^{d+1}} \\
&- \lambda_0^2 \int^> \frac{k^2q^2-(\mathbf{kq})^2}{k^2}(|\mathbf{q}-\mathbf{k}|^{-2}-q^{-2})G^o(\hat{k}-\hat{q})|G^o(\hat{q})|^2 D_o q^{-s}\zeta^<(\hat{k})\frac{d\hat{q}}{(2\pi)^{d+1}} \\
&+ \xi(\hat{k}) + \lambda_0 \int \frac{\mathbf{k}\times\mathbf{q}}{q^2}G^o(\hat{q})G^o(\hat{k}-\hat{q})\langle\xi^>(\hat{q})\xi^>(\hat{k}-\hat{q})\rangle\frac{d\hat{q}}{(2\pi)^{d+1}} + O(\lambda_0^3), \qquad (7)
\end{aligned}
$$

where $\int^>$ denotes integration over the band being removed. Equation (7) will be analyzed in a small-k approximation, i.e., in the asymptotic limit when $k/\Lambda_o \to 0$.

The analysis of neglected terms is quite similar to the case of the 3D Navier-Stokes equations considered by Yakhot and Orszag (1986) and Forster et al. (1977).

The first term on the right of (7) is the usual non-linear term for the slow modes $\zeta^<$ while the second term is generated by non-linear fast mode interactions. In the limit $k/\Lambda_o \to 0$, after performing a frequency integration over Ω and setting $\omega \to 0$, this term can be represented as $\delta[G^o(\mathbf{k})]^{-1}\zeta^<(\mathbf{k})$, where

$$\delta[G^o(\mathbf{k})]^{-1} = k^{-2}\int^> \frac{D(q)\left(q^{-2}-|\mathbf{q}-\mathbf{k}|^{-2}\right)}{2\mu(\mathbf{q})\left[\mu(\mathbf{q})+\mu(\mathbf{k}-\mathbf{q})\right]}\left[k^2q^2-(\mathbf{kq})^2\right]\frac{d^dq}{(2\pi)^d}, \qquad (8)$$

with $\mu(\mathbf{q}) = \nu_o q^2 + i\beta_o q_x q^{-2}$. Thus $\delta[G^o(\mathbf{k})]^{-1}$ is the correction to the inverse Green (or response) function, such that its $O(k^2)$ and $O(k^{-1})$ terms describe corrections $\delta\nu_o$ and $\delta\beta_o$ to the bare viscosity ν_o and bare β_o, respectively. It is clear however that since the integral (8) is calculated as a power series expansion in powers of k/Λ_o in the limit $k/\Lambda_o \to 0$, terms of the form $O(k^{-1})$ do not appear in the result, i.e., the β-term does not renormalize, $\delta\beta_o = 0$, and β_o remains constant and equal to its bare value in the process of small-scale elimination. Keeping this in mind, the subscript o will be removed from β_o in the following.

The last term in (7) is a correction to the stirring force. In the formal limit $\omega \to 0$ at the fixed point corresponding to $\epsilon \to 0$ this term develops a singularity of the kind $1/(-2\epsilon - d + 2)$, which is specific to 2D turbulence ($d = 2$). The physics of this singularity is clear: the inverse cascade cannot possibly exist as a truly stationary state of a closed system with no energy sink at the largest scales. This singularity can be removed by retaining finite but small ω in calculation of the force renormalization integral. This is not a rigorous way of singularity removal but rather a use of intermediate asymptotics corresponding to finite (non-infinite) times or finite (non-zero) frequencies. Upon removing the singularity, it can be shown that this term introduces an $O(\overline{\lambda}^2)$ correction to the correlation function (4) which in turn generates $O(\overline{\lambda}^4)$ terms at the next iteration steps, such that this term can be neglected altogether.

After the integration over the shell $\delta\Lambda_o$ is completed, the *renormalized* form of equation (5) is obtained. Structurally, it is the same as the original (5), but its parameters are renormalized, and it is defined in the reduced interval of wave numbers $0 < k < \Lambda_o - \delta\Lambda_o$. Repeating this procedure many times, one can remove a finite band of modes. The largest remaining wave number will be called the moving dissipation cutoff and denoted by k_c.

In the fixed point, a fully renormalized equation of motion is obtained; it reads

$$\zeta(\hat{k}) = G_r(\hat{k})\xi(\hat{k}), \qquad (9)$$

where $G_r(\hat{k}) = [i\omega + i\beta k_x k^{-2} + \nu(\mathbf{k})k^2]^{-1}$ is the renormalized Green function. This equation will be used in the next section for calculation of characteristic time scales, correlation functions and spectra.

Let us apply now the RG-scheme of small scale elimination to the case of isotropic 2D turbulence ($\beta = 0$) (Staroselsky and Sukoriansky, 1993). It can be shown that for the energy sub-range, $s = 0$ and $\epsilon = 4$. Equating the total energy transport through the system to the constant rate of energy injection, $\bar{\epsilon}$, Staroselsky and Sukoriansky (1993) established that $D_o = 64\bar{\epsilon}$. These parameters lead to the classical Kolmogorov energy spectrum $E(k) = C_K \bar{\epsilon}^{2/3} k^{-5/3}$. In the enstrophy sub-range, $s = 2$, $\epsilon = 6$, $D_o \propto \bar{\eta}$, where $\bar{\eta}$ is the constant rate of enstrophy dissipation, and the energy spectrum is $E(k) = C_\omega \bar{\eta}^{2/3} k^{-3}$. In the limit $k \to 0$, $\omega \to 0$ and in the lowest order of the ϵ-expansion the recursive differential relation for $\nu(k)$ is

$$\frac{d\nu(k)}{dk} = -\frac{1}{16\pi} \frac{D_o}{\nu^2 k^{\epsilon+1}}. \tag{10}$$

The solution to (10) is

$$\nu(k_c) = \nu_o \left\{ 1 + \frac{3}{16\pi} \frac{D_o}{\epsilon \nu_o^3 \Lambda_o^\epsilon} \left[\left(\frac{k_c}{\Lambda_o} \right)^{-\epsilon} - 1 \right] \right\}^{1/3}, \tag{11}$$

which indicates that in isotropic 2D turbulence $\nu(k)$ grows like $k^{-\epsilon/3}$.

Let us now consider the general case $\beta \neq 0$. Then, the integrand in (8) contains angle-dependent functions $\nu(\mathbf{q})$ and $\mu(\mathbf{q}) = \nu(\mathbf{q})q^2 + i\beta q_x q^{-2}$. In principle, the renormalized force could also become anisotropic, leading to $D = D(\mathbf{q})$. This would add further difficulties to the already difficult problem of force renormalization discussed earlier. The possible anisotropization of the renormalized forcing will be neglected in the present study. On the one hand, such an assumption seems to be justified by the fact that the renormalized Green function $G_r(\hat{k})$ absorbs considerable anisotropy; on the other, some sensitivity studies discussed later indicate that the results are not very sensitive to the forcing anisotropy.

The anisotropy due to $\beta \neq 0$ makes the angular integration in (8) nontrivial, so that a self-contained equation similar to (10) cannot be obtained for anisotropic $\nu(\mathbf{k})$. Instead, recast in terms of the non-dimensional parameters $M \equiv \nu k^3/\beta$ and $z \equiv D_o k^{9-\epsilon}/\beta^3$, $\nu(\mathbf{k})$ is described by the integro-differential equation

$$\frac{dM(z,\phi)}{dz} = \frac{3}{9-\epsilon} \frac{M(z,\phi)}{z} + \frac{1}{2} \frac{1}{9-\epsilon} \frac{1}{2\pi} \int_0^{2\pi} \frac{d\theta}{2\pi} \frac{F(M,\theta,\phi)}{M^2(z,\theta)}, \tag{12}$$

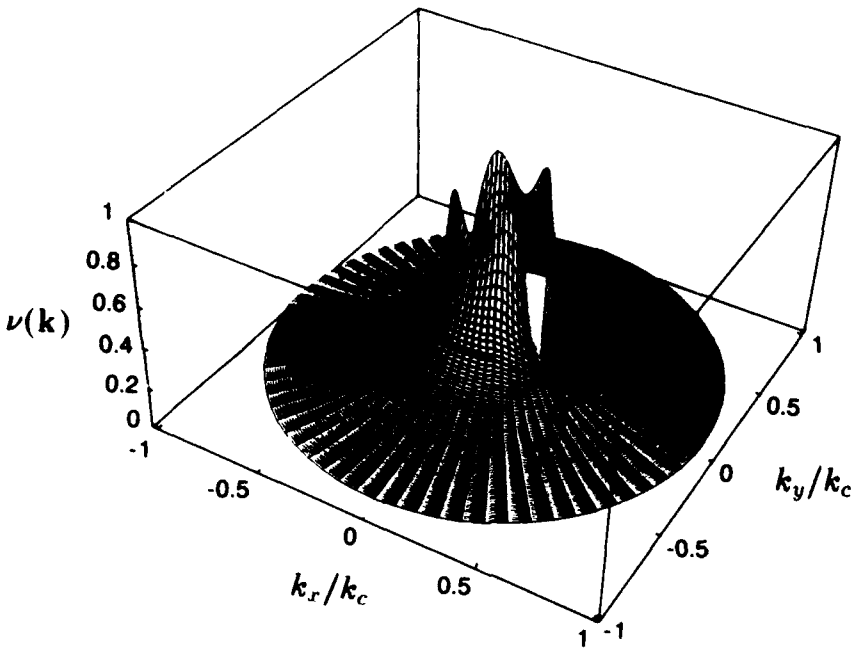


Figure 1. Eddy damping parameter $\nu(\mathbf{k})$ normalized by its maximum value.

where $\phi = \arctan(k_y/k_x)$, and

$$F(M,\theta,\phi) = \frac{1-\cos^2\theta}{M^2(z,\theta)+\cos^2(\theta+\phi)}\{\cos(\theta+\phi)\cos\theta[\cos\phi - 2\cos(\theta+\phi)\cos\theta] + M^2(z,\theta)(1-6\cos^2\theta)\}. \quad (13)$$

The analysis of (12), (13) reveals that for $z >> 1$, turbulence is essentially isotropic, and $\nu(\mathbf{k}) \propto k^{-\epsilon/3}$, as in pure 2-D turbulence. The anisotropy induced by the β-effect develops at $z = O(1)$. The results below pertain to the case of the energy sub-range, where $s = 2$, $\epsilon = 4, z = D_o k^5/\beta^3$, $D_o = 64\bar{\epsilon}$, giving $z = 64\bar{\epsilon}k^5/\beta^3 = 64(k/k_\beta)^5$, where $k_\beta = (\beta^3/\bar{\epsilon})^{1/5}$. As will be discussed in the next section, k_β is a transitional wave number that separates regions of 2D turbulence and Rossby wave domination.

Solving (12), (13) gives $\nu(\mathbf{k})$ as depicted in Fig. 1 in cylindrical surface coordinates as a function of $\mathbf{k}/k_c$; k_c here and in Figs. 2, 4 below is just a dimensional scale that corresponds to $z = 10^3$, or $k_c = (10^3/64)^{1/5}k_\beta = 1.73k_\beta$. At small k, $\nu(\mathbf{k})$ grows sharply for $\phi \in [\pi/4, 3\pi/4]$ and $[5\pi/4, 7\pi/4]$; it abruptly decreases to zero along $\phi = 0, \pi$ causing a singularity in the numerical solution of (12) at $z \approx 0.04$, or $k \approx 0.23k_\beta$.

4. CHARACTERISTIC TIME SCALES AND SPECTRA IN β-PLANE TURBULENCE

The relative magnitudes of the turbulence time scale, the eddy turnover time $\tau_{tu} = [\nu(\mathbf{k})k^2]^{-1}$, and the Rossby wave period, $\tau_R = (\beta \cos\phi/k)^{-1}$, determine which process dominates. In Fig. 2a we plot $\tau_{tu}/\tau_R = \cos\phi/M(k,\phi)$. At large k, this ratio is smaller than 1 and the flow is turbulence dominated. With decreasing k, the ratio becomes progressively anisotropic; it remains much smaller than 1 for the directions close to $\phi = \pm\pi/2$ but rapidly grows in the vicinity of $\phi = 0$ and π indicating progressive domination of Rossby waves. In Fig. 2b, we plot only the Rossby wave dominated region of Fig. 2a, i.e., only the part where $\tau_{tu}/\tau_R \geq 1$. The base of this surface, i.e., the curve at which $\tau_{tu}/\tau_R = 1$, can be identified as the threshold between turbulence and Rossby wave domination. Vallis and Maltrud (1993) derived an analytical expression for such a curve using scaling relations based upon isotropic 2D turbulence. For $k_\beta = (\beta^3/\bar{\epsilon})^{1/5}$, they found

$$k_{x\beta} = k_\beta \cos^{8/5}\phi, \tag{14a}$$

$$k_{y\beta} = k_\beta \sin\phi \cos^{3/5}\phi, \tag{14b}$$

such that the anisotropic transitional wave number, $k_t(\phi)$, becomes

$$k_t(\phi) = \left(k_{x\beta}^2 + k_{y\beta}^2\right)^{1/2} = k_\beta \cos^{3/5}\phi. \tag{15}$$

The curve given by (15), described by Vallis and Maltrud (1993) as a "dumb-bell shape," is shown in Fig. 3; it seems to agree well with the contour $\tau_{tu}/\tau_R = 1$ in Fig. 2b.

It is important to note that the singularity in $\nu(\mathbf{k})$ at $z = 0.04$, or $k = 0.23k_\beta$ for $\phi = 0, \pi$ mentioned in the previous section, resides well inside the dumb-bell shape, such that it should not significantly affect turbulence and wave-turbulence transitional processes that take place at much larger k. It is thus expected that 2D turbulence-Rossby waves interactions are captured faithfully by the RG model.

The preceding results indicate that as $k \to 0$, the β-term significantly affects the nature of the flow field making it anisotropic and either turbulence- or waves-dominated. One should expect that the energy spectrum would also develop dependence upon ϕ and that an anisotropic spectrum $E(\mathbf{k})$ must be considered.

The energy spectrum, $E(\mathbf{k})$, is related to the vorticity correlation function $U(\mathbf{k},\omega)$ which in turn is expressed in terms of the correlation function of the stirring force (4) using (9):

$$U(\mathbf{k},\omega) = \frac{\langle\zeta(\mathbf{k},\omega)\zeta(\mathbf{k}',\omega')\rangle}{(2\pi)^{d+1}\delta(\mathbf{k}+\mathbf{k}')\delta(\omega+\omega')} = 2D_o k^{-s} G_r(\mathbf{k},\omega)G_r(-\mathbf{k},-\omega), \tag{16}$$

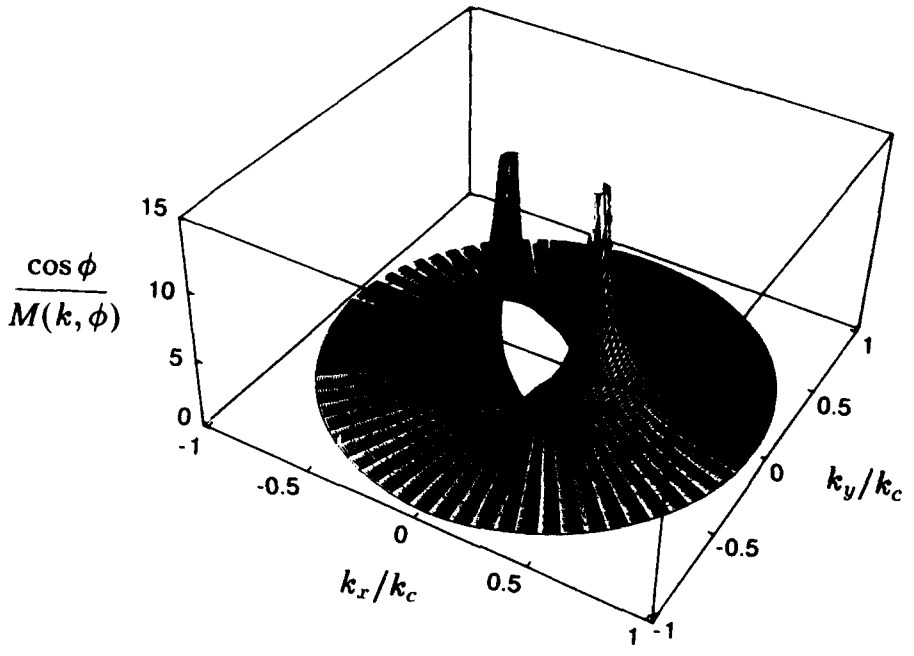


Figure 2a. The ratio of the eddy turnover to Rossby wave time scales, τ_{tu}/τ_R.

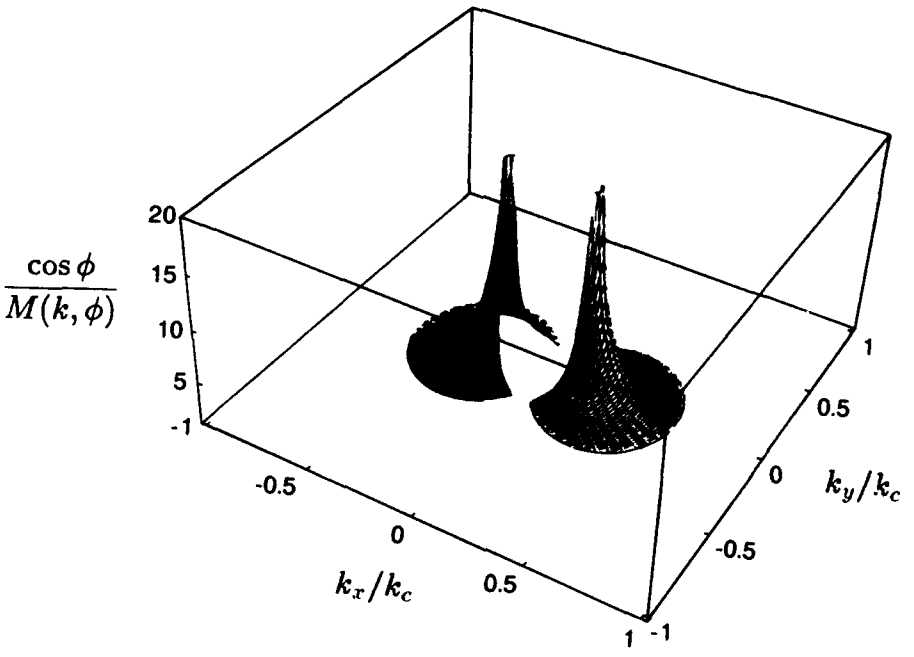


Figure 2b. The region of Rossby wave domination, $\tau_{tu}/\tau_R \geq 1$.

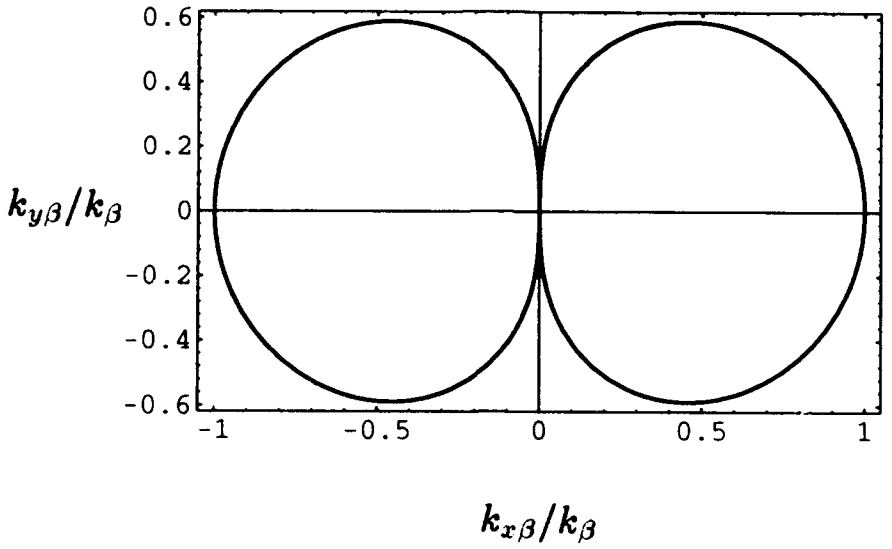


Figure 3. The "dumb-bell shape" of Vallis and Maltrud (1993) for the anisotropic transitional wavenumber $k_t(\phi)$ given by (15).

$$E(\mathbf{k}) = \frac{1}{4\pi k}\int_{-\infty}^{\infty}\frac{d\omega}{2\pi}U(\mathbf{k},\omega). \tag{17}$$

In Figs. 4a,b, we plot the compensated energy spectra $E(\mathbf{k})k^{5/3}$ and $E(\mathbf{k})k^{7/2}$, respectively. One can see that for large k, $E(\mathbf{k})$ is isotropic and proportional to $k^{-5/3}$, which is the classical Kolmogorov spectrum found in the energy sub-range of isotropic 2D turbulence. As $k \to 0$, the spectral anisotropy develops; the results plotted in Fig. 4b indicate that $E(\mathbf{k}) \propto k^{-7/2}$ is a good approximation for $\phi = 0, \pi$. One could speculate that this spectrum is generated by interacting non-linear Rossby waves, which is indeed the mechanism considered, for instance, by Monin and Piterbarg (1987) and Reznik (1986). One should note, however, that the $k^{-7/2}$ spectrum is rather qualitative since it occupies a relatively small range of k and therefore should be considered cautiously. On the other hand, the anisotropic energy spectrum and particularly the power law $k^{-7/2}$ along $\phi = 0, \pi$ are qualitatively new predictions of the RG theory which are yet to be compared with data.

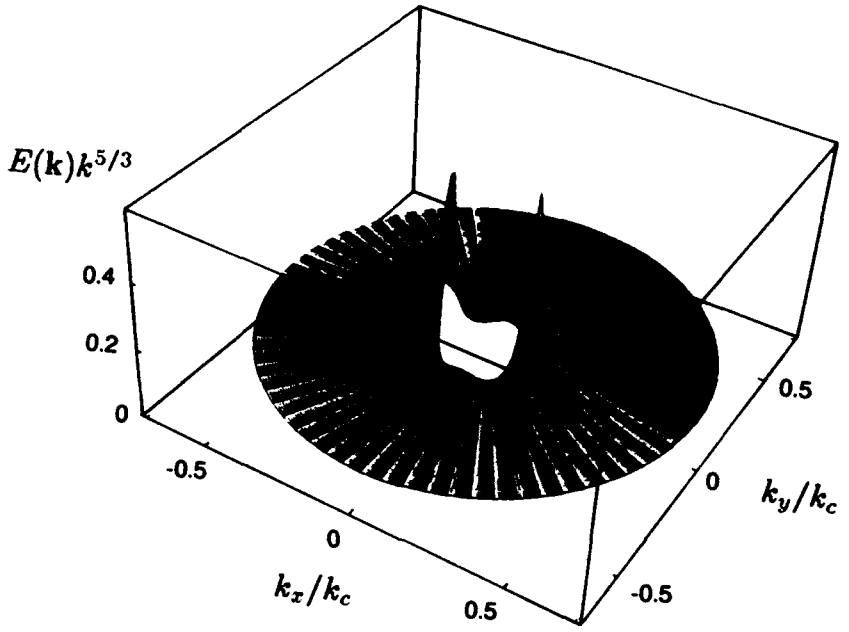


Figure 4a. Compensated energy spectrum $E(\mathbf{k})k^{5/3}$.

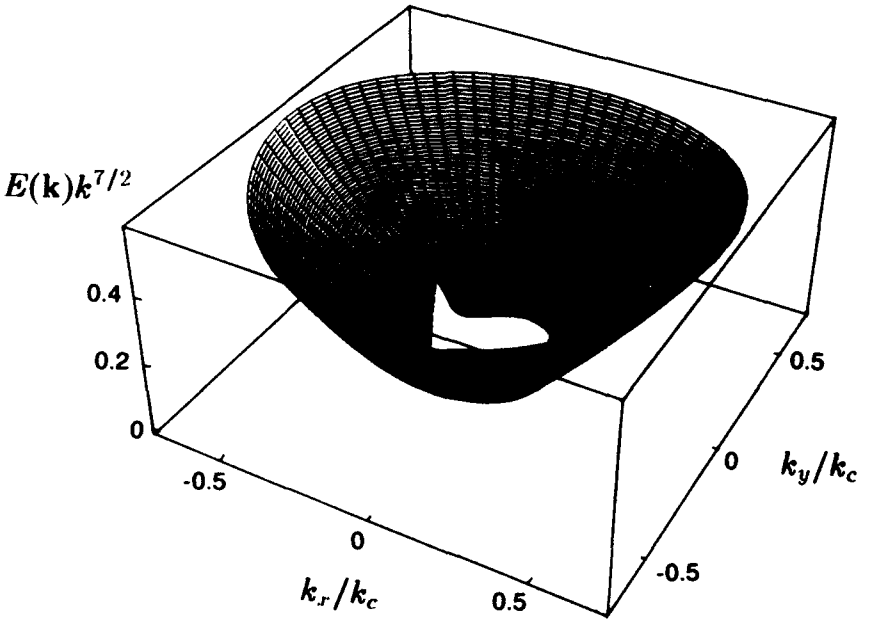


Figure 4b. Compensated energy spectrum $E(\mathbf{k})k^{7/2}$.

5. TWO-PARAMETRIC TURBULENCE CHARACTERISTICS; EDDY VISCOSITY AND EDDY β

Although $\nu(\mathbf{k})$ is derived from the bare, or molecular viscosity ν_o, it is not what is often comprehended as an "eddy" viscosity. According to (9), $\nu(\mathbf{k})$ and β are parts of the renormalized Green function $G_r(\hat{k})$ that describes the response of the mode $\zeta(\hat{k})$ to stochastic forcing at the same wave number, $\xi(\hat{k})$. Mathematically, $G_r(\hat{k})$ is a *response function* which can be formally calculated from (9) by taking a functional derivative of $\zeta(\hat{k})$ with respect to the forcing $\xi(\hat{k})$, and β and $\nu(\mathbf{k})$ are thus *response parameters*. They allow one to calculate the vorticity correlation function, $U(\mathbf{k},\omega)$ in (16), and the energy spectrum (17), but they do not directly relate to enstrophy and energy transfers and dissipation. Furthermore, $[\nu(\mathbf{k})k^2]^{-1}$ can be viewed as a characteristic time scale of information loss at given $\mathbf{k}$ caused by non-linear scrambling of all other modes (see Dannevik et al., 1987). Therefore, $\nu(\mathbf{k})$ may be interpreted as an eddy damping parameter and is substantially a one-point turbulence characteristic. In this Section, proper "eddy" parameters will be introduced and it will be shown how they can be calculated using RG-based response parameters.

To analyze energy and enstrophy transfer, one needs to consider two-point characteristics that account for interactions between a given large-scale mode $k < k_c$ and all SGS modes $k > k_c$, where k_c, as before, is the moving dissipation cutoff. A fundamental characteristic of this kind, the two-parametric viscosity $\nu(k|k_c)$, was introduced by Kraichnan (1976) for isotropic 2D and 3D turbulence based upon one-time, two-point correlation functions; it characterizes the energy transfer from all SGS modes to a resolved mode with the wave number k. This approach is sufficient if a system does not support waves, i.e., if the system does not have dispersion, or its Green function is real at $\omega = 0$. If a system is anisotropic and, in addition, waves are present, as, for instance, the Rossby wave term in (3), then, as suggested by Kaneda and Holloway (1992; also this volume), a two-point, two-time vorticity correlation function, $U(\mathbf{k},t,t') \equiv \langle\zeta(\mathbf{k},t)\zeta(-\mathbf{k},t')\rangle$, should be considered. Here $U(\mathbf{k},t,t')$ is described by a von Kármán–Howarth-type equation

$$\left(\partial_t + i\beta k_x/k^2 + \nu k^2\right) U(\mathbf{k},t,t') = \mathcal{T}(\mathbf{k},t,t'), \tag{18}$$

where $\mathcal{T}(\mathbf{k},t,t')$ is the two-time, anisotropic, spectral enstrophy transfer function. Assuming quasi-stationarity, the dependence on t is negligible compared to that on $t - t'$, and will be ignored. When the limit $t \to t'$ is considered, then $\mathcal{T}(\mathbf{k},t,t)$ is complex and, as will be seen later, describes the effect of unresolved on resolved, or explicit, modes of motion, which results in a modification of both ν and β in (18). Generalizing Kraichnan's (1976) definition for a spectrally anisotropic β-plane turbulence, one can introduce two two-parametric characteristics, $\nu(\mathbf{k}|k_c)$ and $\beta(\mathbf{k}|k_c)$:

$$\nu(\mathbf{k}|k_c) = -\frac{\Re[\mathcal{T}(\mathbf{k}|k_c)]}{k^2 U(\mathbf{k})}, \tag{19}$$

$$\beta(\mathbf{k}|k_c) = -\frac{\Im[\mathcal{T}(\mathbf{k}|k_c)]\, k^2}{k_x U(\mathbf{k})}, \tag{20}$$

where

$$\mathcal{T}(\mathbf{k}|k_c) = \iint_{\Delta} \Theta_{-\mathbf{k},\mathbf{p},\mathbf{q}}(p^2 - q^2)\sin\alpha \left[\frac{p^2 - q^2}{p^2 q^2} U(\mathbf{p})U(\mathbf{q}) - \frac{k^2 - q^2}{k^2 q^2} U(\mathbf{q})U(\mathbf{k}) + \frac{k^2 - p^2}{k^2 p^2} U(\mathbf{p})U(\mathbf{k}) + 3 \text{ similar terms}\right] d\mathbf{p}d\mathbf{q}. \tag{21}$$

Here, $\Theta_{-\mathbf{k},\mathbf{p},\mathbf{q}}$ is a complex triad interaction characteristic; its real part is the familiar triad relaxation time while the imaginary part describes the SGS effect on phase properties of the resolved modes. Also, α is the angle between the vectors $\mathbf{p}$ and $\mathbf{q}$, and $\iint_{\Delta}$ denotes integration over all triangles $(\mathbf{k}, \mathbf{p}, \mathbf{q})$ such that p and/or q are greater than k_c. Not shown are the terms that correspond to the mirror image of the triangle with respect to $\mathbf{k}$. The two-parametric viscosity $\nu(\mathbf{k}|k_c)$ in (19) is a measure of the energy transfer from the unresolved flow scales (turbulence and Rossby waves) to the resolved wave number $\mathbf{k}$. Similarly, the two-parametric $\beta(\mathbf{k}|k_c)$ in (20) accounts for the total effect of the SGS processes on the resolved Rossby wave with the natural frequency $\beta k_x/k^2$ and dispersion relation (2). Such a Rossby wave frequency shift has been discussed by Holloway (1986) and calculated by Kaneda and Holloway (1992; this volume) using the Lagrangian renormalized approximation in the assumption of small β.

The two-parameter quantities $\nu(\mathbf{k}|k_c)$ and $\beta(\mathbf{k}|k_c)$ thus defined generalize the notion of eddy viscosity for flows that involve both turbulence and waves. Here we suggest that in LES of β-plane turbulence, $\nu(\mathbf{k}|k_c)$ and $\beta(\mathbf{k}|k_c)$ should be used as eddy viscosity and eddy β, respectively.

Another useful interpretation of the response and eddy parameters is in associating the former with the effective dispersion relation (similar to the bare dispersion relation (2)) for SGS modes and the latter with the effective dispersion relation for resolved modes.

Different spectral closures provide different expressions for $\Theta_{-\mathbf{k},\mathbf{p},\mathbf{q}}$ but they all involve eddy damping characteristics that should be found in conjunction with (3). This leads to the necessity either to introduce phenomenological considerations to parameterize the eddy damping or to solve a coupled field problem which is a difficult task, particularly when spectral anisotropy and waves are present. Kraichnan

(1976) has solved this problem numerically using the TFM for isotropic 2D turbulence, while Holloway and Hendershott (1977) applied TFM to β-plane turbulence assuming weak anisotropy.

The function $\Theta_{-\mathbf{k},\mathbf{p},\mathbf{q}}$ can be calculated using the RG-derived response parameter $\nu(\mathbf{k})$ or renormalized Green function $G_r(\mathbf{k},\omega)$; it can be shown that in the lowest nontrivial order of the ϵ-expansion

$$\begin{aligned}\Theta_{-\mathbf{k},\mathbf{p},\mathbf{q}} &= \frac{1}{2}\left[G_r^{-1}(-\mathbf{k},0) + G_r^{-1}(\mathbf{p},0) + G_r^{-1}(\mathbf{q},0)\right]^{-1} \\ &= \frac{1}{2}\frac{(\nu_\mathbf{k}+\nu_\mathbf{p}+\nu_\mathbf{q}) - i(\omega_{-\mathbf{k}}+\omega_\mathbf{p}+\omega_\mathbf{q})}{(\nu_\mathbf{k}+\nu_\mathbf{p}+\nu_\mathbf{q})^2 + (\omega_{-\mathbf{k}}+\omega_\mathbf{p}+\omega_\mathbf{q})^2},\end{aligned} \tag{22}$$

where $\nu_\mathbf{k} \equiv \nu(\mathbf{k})k^2$, and $\omega_\mathbf{k} \equiv \beta k_x/k^2$.

The idea of using the RG-derived response parameters in second order spectral closures was first suggested by Dannevik et al. (1987); in particular, they showed that the Eddy-Damped, Quasi-Normal, Markovian (EDQNM) approximation (Orszag, 1977) is obtained from the RG theory in the lowest nontrivial order of the ϵ-expansion.

The use of RG-derived response parameters in second order spectral closures appears to be a powerful idea because renormalized Green functions in RG are decoupled from correlation functions and thus can be calculated independently.

From the point of view of practical applications, the RG-based spectral closures and thus eddy parameters, can be calculated in a self-consistent way free of phenomenological approximations with full account for spectral anisotropy and turbulence-wave interactions. The models implementing such eddy parameters should have the correct SGS physics; they are expected to perform rather well in a variety of complicated flows typical of physical oceanography, in both eddy-resolving and non-eddy-resolving configurations.

In the limit $(\nu_\mathbf{k}+\nu_\mathbf{p}+\nu_\mathbf{q})/(\omega_{-\mathbf{k}}+\omega_\mathbf{p}+\omega_\mathbf{q}) \to 0$, $\Theta_{-\mathbf{k},\mathbf{p},\mathbf{q}}$ reduces to a δ-function, $\Theta_{-\mathbf{k},\mathbf{p},\mathbf{q}} \to \pi\delta(\omega_{-\mathbf{k}}+\omega_\mathbf{p}+\omega_\mathbf{q})$, thus revealing the resonance condition for wave triads broadened by turbulence. In this limit, one recovers the approximation of weakly non-linear waves (Holloway, 1986; Carnevale and Martin, 1982; Salmon, 1982; Reznik, 1986) that leads to the kinetic, or Boltzmann, equation for waves.

In Fig. 5, we plot $\nu(k|k_c)$ for isotropic 2-D turbulence ($\beta = 0$) calculated using the RG-based response function $\nu(k)$ in (10). As in Kraichnan (1976), $\nu(k|k_c)$ has a positive cusp near k_c and becomes negative as $k \to 0$. Numerical values of the RG-based $\nu(k|k_c)$ are close to those derived by Kraichnan (1976).

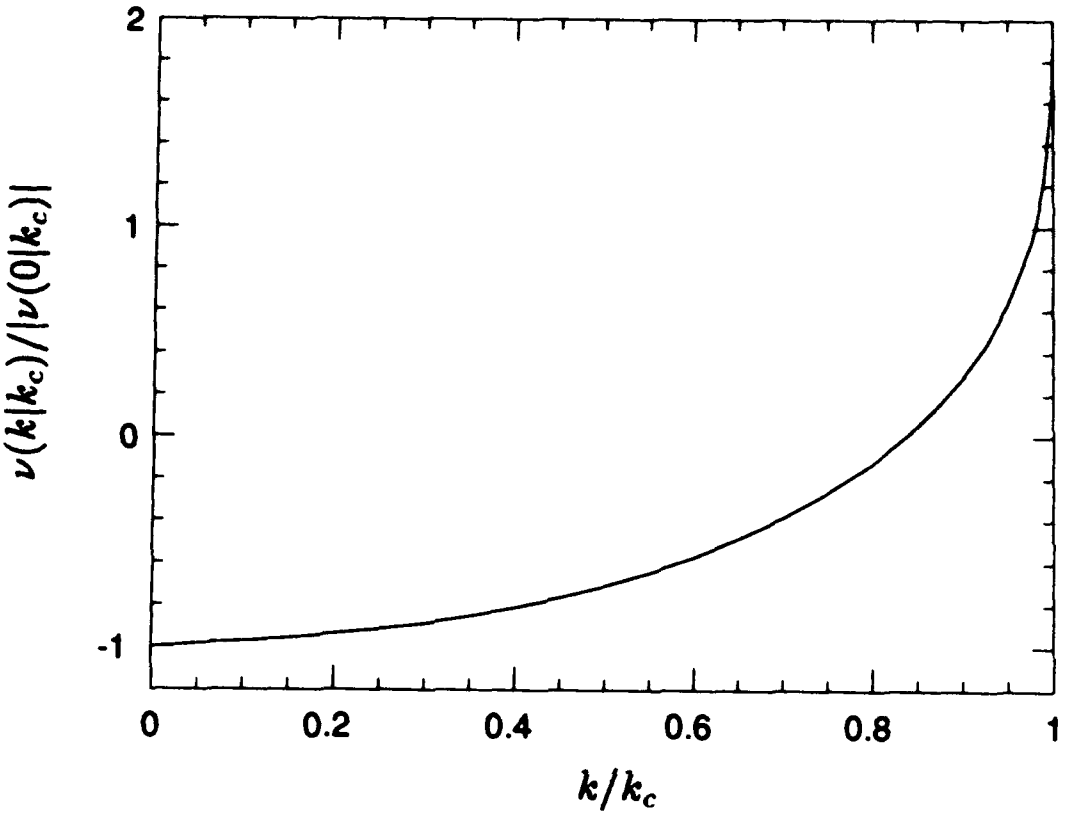


Figure 5. Normalized two-parametric viscosity $\nu(k|k_c)/|\nu(0|k_c)|$ for isotropic ($\beta = 0$) 2D turbulence.

In Fig. 6a, we plot the angle-dependent, RG-based $\nu(\mathbf{k}|k_c)$ for isotropic 2D turbulence ($\beta = 0$). Obviously, the result plotted here is the body of revolution formed by the curve shown in Fig. 5. The isotropy of Fig. 6a is broken in Fig. 6b, where $\nu(\mathbf{k}|k_c)$ for $\beta \neq 0$ is shown. For $k/k_c \geq 0.6$, or $k/k_\beta \geq 1.0$, the β-effect is not pronounced and $\nu(\mathbf{k}|k_c)$ behaves quite similarly to isotropic 2D turbulence; there is a positive cusp and then $\nu(\mathbf{k}|k_c)$ becomes negative. As $k \to 0$, the effect of the β-term becomes stronger; $\nu(\mathbf{k}|k_c)$ remains negative in the vicinity of $\phi = \pm\pi/2$ but increases in other directions. The negativity of $\nu(\mathbf{k}|k_c)$ along $\phi = \pm\pi/2$ indicates a strong inverse energy transfer to these directions which, in physical space, correspond to energy funneling into zonal flows $\mathbf{v} = (v_x(y), 0)$. This demonstrates that zonal flows, typical of both Earth and planetary circulations (Ingersoll, 1990) result from and are sustained by the self-organization of the quasi-2-D turbulence on the β-plane.

To single out the mechanism causing $\nu(\mathbf{k}|k_c)$ to remain negative along $\phi = \pm\pi/2$ for small k, or the mechanism of zonalization in the physical space, $\nu(\mathbf{k}|k_c)$ was calculated with the RG-based vorticity correlation function (16) for isotropic turbulence, in which the non-zero β-term was retained only in the triad relaxation

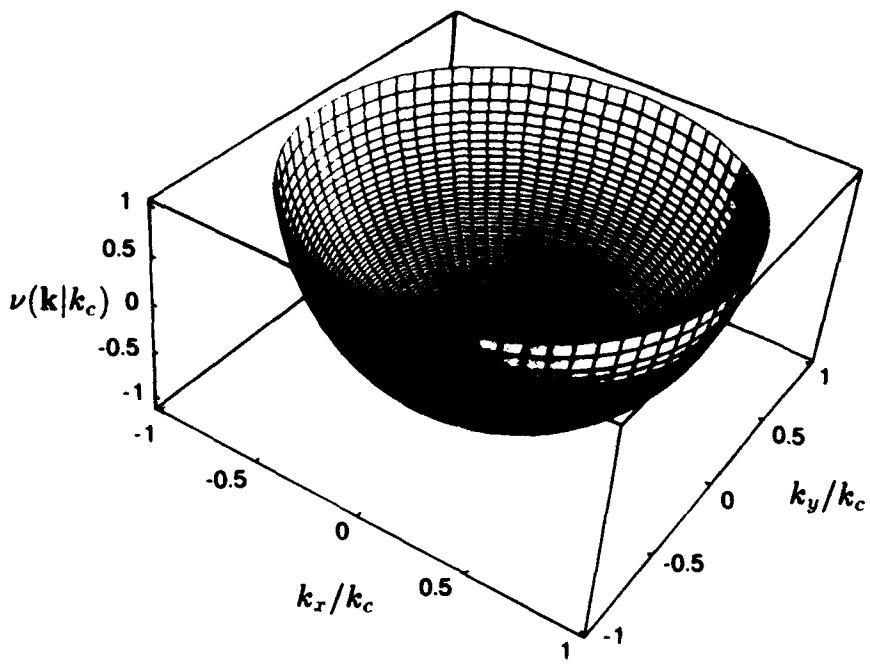


Figure 6a. Two-parametric viscosity $\nu(\mathbf{k}|k_c)$ normalized by its maximum value for isotropic (β=0) 2D turbulence.

characteristic $\Theta_{-\mathbf{k},\mathbf{p},\mathbf{q}}$ in (21). In Fig. 6c, we plot $\nu(\mathbf{k}|k_c)$ in this case and observe the same general features as the two-parametric viscosity calculated with the full model (cf. Fig. 6b). Particularly strong negative values along $\phi = \pm\pi/2$ are also present in Fig. 6c. This result indicates that the energy funneling into zonal flows is more the result of the β-effect on $\Theta_{-\mathbf{k},\mathbf{p},\mathbf{q}}$ than on the correlation function $U(\mathbf{k})$.

In Fig. 7, we plot $\beta(\mathbf{k}|k_c)/\beta$. Similarly to the two-parametric viscosity, this characteristic also reveals a positive cusp for $k/k_c \approx 1$, but unlike $\nu(\mathbf{k}|k_c)$, $\beta(\mathbf{k}|k_c)$ varies significantly at large k; it is larger inside the sectors $(\pi/4, 3\pi/4)$ and $(5\pi/4, 7\pi/4)$ than outside. Since $\beta(\mathbf{k}|k_c)$ is in fact a correction to β, the result plotted in Fig. 7 indicates that at $k/k_c \approx 1$ the SGS contribution to β is comparable to β itself. In the limit $k \to 0$, the SGS contribution to β decreases; the eddy β remains small for all directions.

6. ANISOTROPIC ENERGY TRANSFER

The spectral energy transfer, $-\mathcal{T}_e(\mathbf{k}|k_c) = 2k^2\nu(\mathbf{k}|k_c)E(\mathbf{k})$, is plotted in Fig. 8. By definition, $\mathcal{T}_e(\mathbf{k}|k_c)$ accounts for the *total* energy transfer, that is due to turbulence and non-linear waves. As $k \to 0$, $-\mathcal{T}_e(\mathbf{k}|k_c)$ approaches 0 for all directions except in the vicinity of $\phi = \pm\pi/2$ and $\phi = 0, \pi$, where it develops four

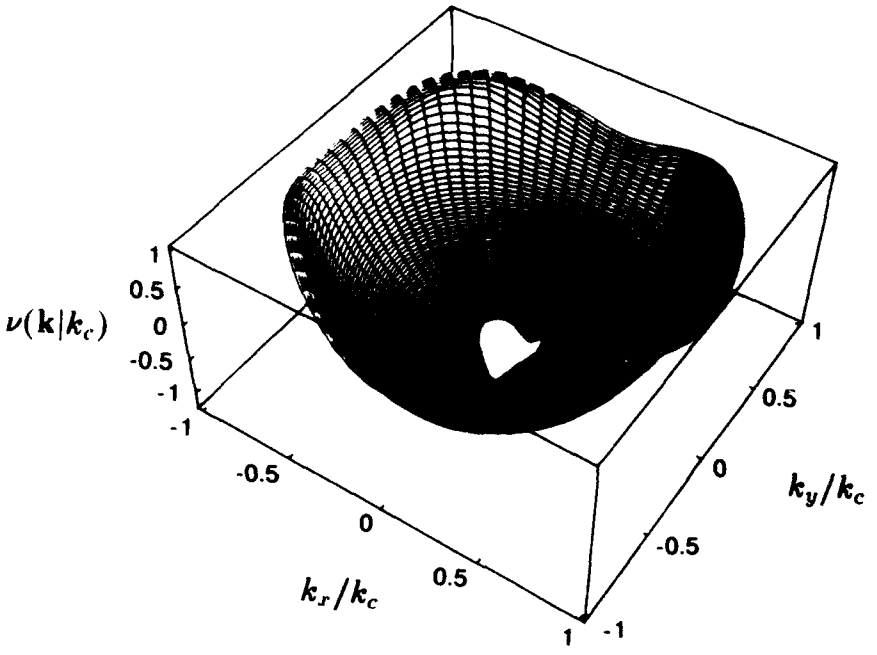


Figure 6b. Two-parametric viscosity $\nu(\mathbf{k}|k_c)$ normalized by its maximum value for β-plane turbulence. Here $k_c = 1.73k_\beta$.

negative dips. The dips along $\phi = \pm\pi/2$ have been identified earlier with the flow zonalization. The interpretation of the other two dips is more subtle. As was shown in Fig. 2b, the region in the vicinity of $\phi = 0, \pi$ is strongly dominated by Rossby waves. It appears now that the energy of these zonally propagating waves is sustained by the anisotropic transfer. As $k \to 0$, the inverse energy cascade becomes less efficient because wave triads should satisfy an additional resonance condition. However, since total energy must be conserved, the increasing amount of energy funnels into $\phi = 0, \pi$ and $\phi = \pm\pi/2$. Such a large-scale flow organized into zonal jets and zonally propagating Rossby waves agrees with results of numerical simulations (Rhines, 1975; Williams, 1978; Yoden and Yamada, 1993). The energy flux into $\phi = 0, \pi$ is manifested in the annihilation of large-scale eddies due to radiation of their energy by Rossby waves. The tendency of the β-effect to destroy coherent vortex structures has been demonstrated in numerical simulations (Maltrud and Vallis, 1991). The Rossby waves radiation does not occur only for structures with $k_x \to 0$, i.e., zonal jets, which thus become an additional attracting large-scale flow configuration.

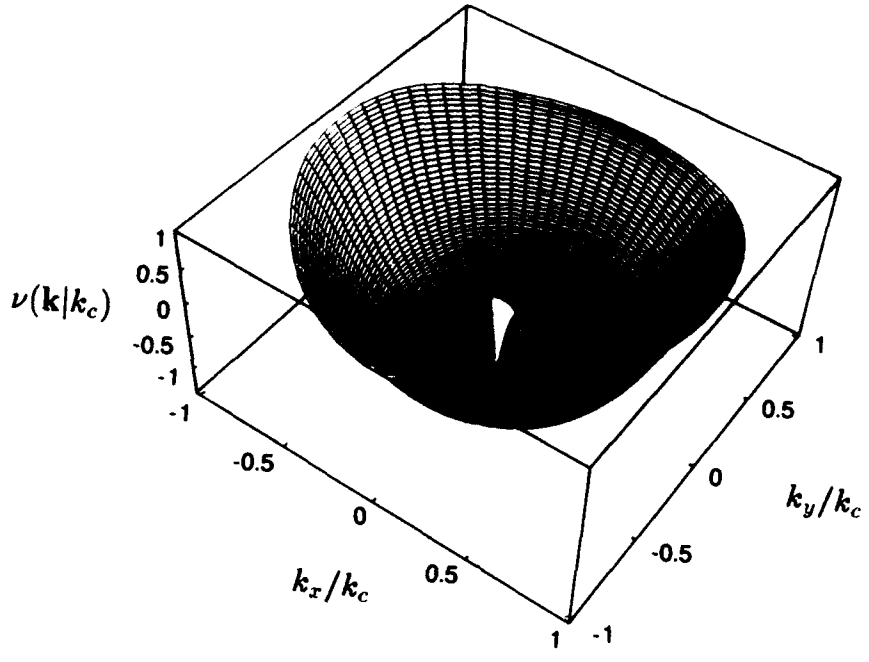


Figure 6c. Two-parametric viscosity $\nu(\mathbf{k}|k_c)$ normalized by its maximum value. Here $\beta \neq 0$ but $U(\mathbf{k})$ is isotropic. To better resolve the region of small k, k_c is reduced to $k_c = 0.66k_\beta$.

One may apply these results to tackle the problem of the Gulf Stream separation and maintenance from the point of view of non-linear dynamics. An energetic jet stream is formed due to the western intensification and flows northward along the east coast of the US. Topographic deflection and, say, adverse pressure gradient (Haidvogel et al., 1992) at Cape Hatteras facilitate the funneling of the jet's energy in the zonal direction such that it leaves the coast. The energy funneling mechanism sustains this organized jet flow in the open ocean and is in fact responsible for the Gulf Stream's existence. Such a barotropic picture of the Gulf Stream separation is, of course, an oversimplification because baroclinicity plays an important role in the stream's dynamics. However, the present results strongly indicate that Gulf Stream separation and maintenance may be facilitated by essentially non-linear processes, a direction almost unexplored in current Gulf Stream research (see the survey by Haidvogel et al., 1992).

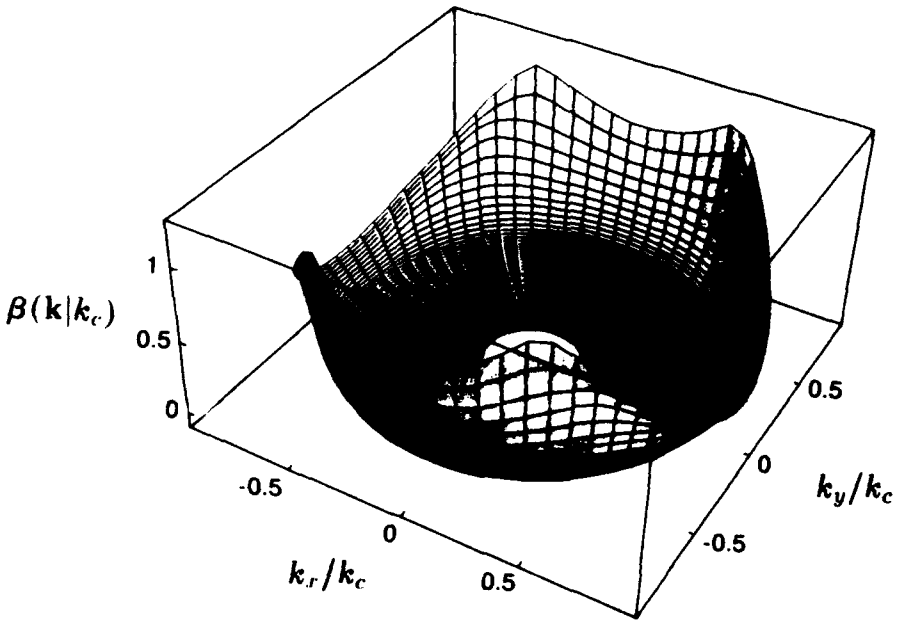


Figure 7. Two-parametric $\beta(\mathbf{k}|k_c)/\beta$ for β-plane turbulence. Here $k_c = 1.73k_\beta$.

7. APPLICATION OF THE RG THEORY FOR PRACTICAL OCEANOGRAPHIC SIMULATIONS

The present results can be used for large eddy simulation of β-plane turbulence. For this purpose, k_c should be identified with the boundary between explicit and subgrid scales, and $\nu(\mathbf{k}|k_c)$ and $\beta(\mathbf{k}|k_c)$ should be used as spectral eddy viscosity and eddy β, respectively. These and similar eddy parameters can also be adopted as SGS characteristics in models of large-scale ocean circulation. There are two obvious difficulties, however, that seem to be able to hamper the application of these eddy parameters to practical oceanographic problems:

- All the derivations here are performed in Fourier space while the majority of OGCMs are developed in physical space;
- The presence of negative viscosity may lead to inherent numerical instability of OGCMs.

Here, it will be shown how these both problems can be resolved for the example of isotropic 2D turbulence. In that case, the two-parametric viscosity, shown in Fig. 5, can be interpolated by a polynomial expression in powers of $(k/k_c)^2$:

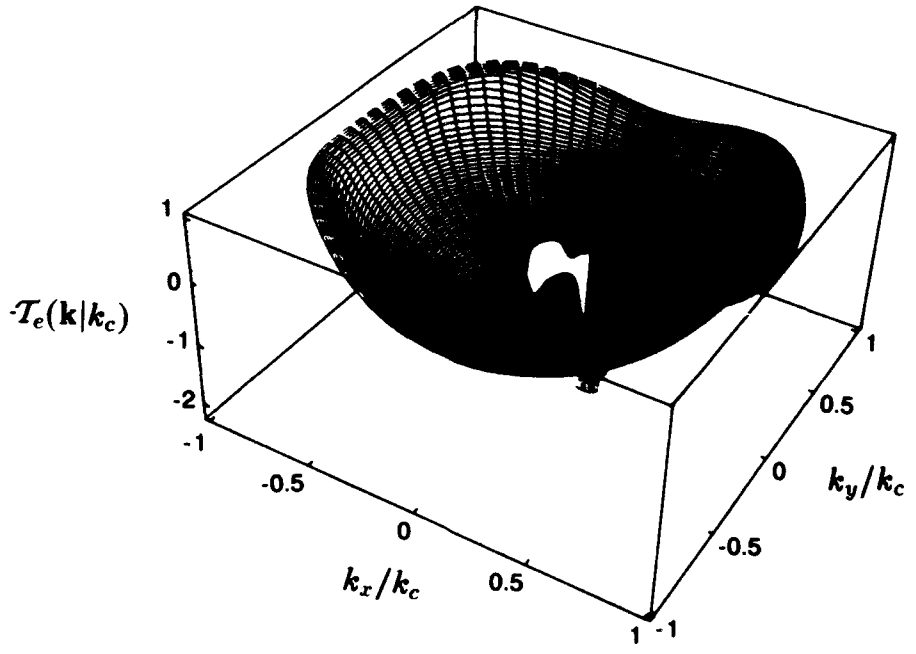


Figure 8. Spectral energy transfer for β-plane turbulence, $-\mathcal{T}_e(\mathbf{k}|k_c)$, normalized by its maximum value. Here $k_c = 1.73k_\beta$.

$$\nu(k|k_c)\left(\frac{k}{k_c}\right)^2 \approx -|\nu(0|k_c)|\left(\frac{k}{k_c}\right)^2 + |\nu(0|k_c)|\left(\frac{k}{k_c}\right)^4 + 0.5|\nu(0|k_c)|\left(\frac{k}{k_c}\right)^8 + 0.125|\nu(0|k_c)|\left(\frac{k}{k_c}\right)^{12} + \ldots\ , \quad (23)$$

where

$$\nu(0|k_c) = \left(1 - \frac{\epsilon}{3}\right)\nu(k_c), \quad (24)$$

which gives

$$\nu(0|k_c) = -\frac{1}{3}\nu(k_c) < 0 \quad (25)$$

for the energy sub-range and

$$\nu(0|k_c) = -\nu(k_c) < 0 \quad (26)$$

for the enstrophy sub-range. These results are consistent with Kraichnan's (1976) assertion that at low k, $\nu(k|k_c)$ reaches saturation at negative values. Equations

(24-26) not only reconfirm Kraichnan's (1976) result, but quantify it in terms of $\nu(k_c)$.

Approximated by the first two terms, the inverse Fourier transform of (23) produces a dissipation term in (1) of the form

$$-\nu_1(\Delta)\nabla^2\zeta - \nu_2(\Delta)\nabla^4\zeta, \tag{27}$$

where Δ is the grid resolution in physical space. Equation (27) is a linear superposition of a **negative** Laplacian and a **positive** biharmonic viscosity. The negative Laplacian viscosity is the destabilizing term that may be responsible for initiating and maintaining the eddy activity, while the biharmonic (and higher order) friction term provides an efficient dissipation mechanism that also insures numerical stability. Equations of similar structure have appeared in different branches of physics (Kuramoto and Tsuzuki, 1976; Kuramoto, 1984; Sivashinsky, 1979, and references therein) and are known as Kuramoto–Sivashinsky-type equations. A very important feature of these equations is that they have regular solutions due to the stabilizing biharmonic term and thus produce well-posed problems despite the presence of the negative Laplacian.

Models of horizontal mixing used in existing OGCMs mostly employ either positive Laplacian or negative biharmonic operators with constant or Smagorinsky-type (Smagorinsky, 1963, 1993) eddy viscosities. The closest model to that of (23) with a scale-selective representation of the SGS processes used in today's geophysical simulations is that given by the anticipated potential vorticity (APV) method (Sadourny and Basdevant, 1985), in which potential enstrophy dissipation is parameterized by a diffusion operator in the form of an iterated Laplacian, $\tau^{-1}k_c^{-2n}(-\nabla^2)^n$, τ being a characteristic eddy turnover time at the grid scale, and $n \sim 8$. A detailed analysis of the APV method is given by Vallis and Hua (1988) who pointed out that its major advantage is in that it conserves energy and dissipates enstrophy. By virtue of energy conservation, the APV method produces effective eddy viscosities that are cusp-like near k_c and negative at small k. However, its implementation involves a certain degree of phenomenology, for instance, with respect to determination of τ. Another disadvantage of the APV method is its lack of Galilean invariance.

In practical situations where spectral anisotropy and/or waves may be present, the RG-based parameterization of SGS processes will not be as simple as that given by (23), (24), (27). As an example, in the case of β-plane turbulence, the two-parametric viscosity shown in Fig. 6b is a complicated cylindrical surface with pronounced anisotropy. The corresponding viscosity operators in physical space will also be complicated and will acquire tensorial properties. However, these operators can be calculated from the RG theory in a self-consistent way with no appeal to phenomenological or empirical considerations; they incorporate correct physics that include Galilean invariance, conservation of energy, dissipation of enstrophy, Rossby wave-turbulence interaction and the negative viscosity phenomena that long have

been known to play an important role in geophysical and planetary circulations (Starr, 1968).

8. CONCLUSIONS AND DISCUSSION

We have developed a self-consistent theory of β-plane turbulence using the RG theory. This theory is rooted in the basic physics of 2D turbulence and Rossby waves and recovers most of the known theoretical and numerical results on β-plane turbulence.

At large wave numbers, the flow is turbulence dominated and reveals isotropic 2D turbulence-like behavior. With decreasing k, the β-effect progressively becomes more pronounced. The transition from 2D turbulence to the Rossby wave dominated regime takes place near the dumb-bell curve (15) that is the anisotropic generalization of the transitional wave number introduced by Rhines (1975). Inside the dumb-bell curve, the flow dynamics approach the limit of weakly interacting non-linear Rossby waves described by the kinetic, or Boltzmann, equation.

With decreasing k, the energy spectrum undergoes a smooth transition from the isotropic $k^{-5/3}$ Kolmogorov law to a strongly anisotropic $k^{-7/2}$ law along $\phi = 0, \pi$. The energy transfer has a cusp-like behavior near the wave number k_c [defined above (9)] and becomes negative for smaller k revealing inverse energy transfer typical of 2D turbulence. At yet smaller k, the inverse transfer diminishes for all directions but $\phi = 0, \pi$ and $\phi = \pm\pi/2$, where it remains negative for all k indicating that the flow undergoes self-organization into zonal flows and zonally propagating Rossby waves.

The eddy viscosity and eddy β parameters for LES of β-plane turbulence have been obtained using the two-time von Kármán–Howarth-type equation for the vorticity correlation function and fully account for the turbulence-Rossby wave interaction and spectral anisotropy. When converted back to physical space, the spectral SGS representation produces a Kuramoto–Sivashinsky-type equation, with negative Laplacian term, a positive biharmonic friction, and, possibly, higher order dissipative terms. The negative Laplacian is a destabilizing term which is directly responsible for initiating and sustaining the eddy activity in the world ocean, atmosphere and other quasi-2D, high Reynolds number systems. The higher order biharmonic and hyperviscosities provide efficient scale-selective dissipation mechanisms that ensure the well-posedness and numerical stability of the problem. These results have direct implications for numerical modeling of the oceanic circulation. The action of negative viscosity is not only a persistent source of the SGS energy, but also a vehicle for transporting the SGS energy to ever larger scales. Compounded with spectral anisotropy and tendencies to self-organization, the negative viscosity phenomena may play a very important role in the ocean circulation physics. It is clear therefore that SGS representation is an important element in both eddy resolving and non-eddy-resolving models. The RG theory of turbulence provides a self-consistent

framework capable of addressing some of the key issues of SGS parameterization for oceanographic modeling; the RG-based SGS models are Galilean invariant, conserve energy, dissipate enstrophy, include Rossby wave-turbulence interaction and incorporate the negative viscosity phenomena.

The RG-based SGS parameterization of β-plane turbulence indicates that the moving dissipation cutoff k_c can be chosen quite close to k_β, the scales at which the Rossby wave dynamics become dominant. This fact gives rise to the hope that if the RG-based SGS parameterization is developed for processes at scales of the deformation radius, the grid resolution could be of the order of the deformation radius itself. This would lead to creation of a non-eddy-resolving model in which eddy effects are properly incorporated.

The present study is concerned with only a "building block" geophysical flow, viz. β-plane turbulence. Real oceanic flows are much more complicated. However, the general approach developed in this paper may be generalized for more complicated situations. In particular, the effects of topography and stratification can be incorporated in the vorticity equation (Charney and Stern, 1962; Rhines, 1979; McWilliams, 1989) which enables the extension of the RG analysis to these flows. For example, barotropic quasi-2D turbulence over topography can in some cases be described by the barotropic vorticity equation with a β-like term due to the topographic gradient. As for β-plane turbulence, the topographic β-term generates mean flows and topographic Rossby waves in the direction normal to the topographic gradient, with direct implications for coastal oceanography. These conclusions are supported by the results of direct numerical simulations (Vallis and Maltrud, 1993).

9. ACKNOWLEDGMENTS

This research has been partially supported by ONR Grants N00014-92-J-1363, N00014-92-C-0118, and N00014-92-C-0089, NSF Grant OCE 9010851, and the Perlstone Center for Aeronautical Engineering Studies.

REFERENCES

AMIT, D. J., 1978: *Field Theory, the Renormalization Group, and Critical Phenomena.* McGraw-Hill.

BARTELLO, P., AND HOLLOWAY, G., 1991: Passive scalar transport in β-plane turbulence. *J. Fluid Mech.*, **223**, 521–536.

BATCHELOR, G.K., 1969: Computation of the energy spectrum in homogeneous two-dimensional turbulence. *Phys. Fluids Suppl. II*, **12**, 233–238.

BRYAN, F., 1987: Parameter sensitivity of primitive equation general circulation models. *J. Phys. Oceanogr.*, **17**, 970–985.

CARNEVALE, K.F., AND MARTIN, P.C., 1982: Field theoretical techniques in statistical fluid dynamics: with application to nonlinear wave dynamics. *Geophys. Astrophys. Fluid Dyn.*, **20**, 131–164.

CHARNEY, J.G., AND STERN, M., 1962: On the stability of internal baroclinic jets in a rotating atmosphere. *J. Atmos. Sci.*, **19**, 159–172.

CRESWICK, R.J., FARACH, H.A., AND POOLE, C.P., JR., 1992: *Introduction to Renormalization Group Methods in Physics.* John Wiley & Sons.

DANNEVIK, W.P., YAKHOT, V., AND ORSZAG, S.A., 1987: Analytical theories of turbulence and the ϵ expansion. *Phys. Fluids*, **30**, 2021–2029.

FORSTER, D., NELSON, D.R., AND STEPHEN, M.J., 1977: Large distance and long-time properties of a randomly stirred fluid. *Phys. Rev. A*, **16**, 732–749.

GALPERIN, B., AND ORSZAG, S.A., EDS., 1993: Large Eddy Simulation of Complex Engineering and Geophysical Flows. Cambridge University Press, in press.

GRIFFA, A., AND CASTELLARI, S., 1991: Nonlinear general circulation of an ocean model driven by wind with a stochastic component. *J. Mar. Res.*, **49**, 53–73.

GRIFFA, A., AND SALMON, R., 1989: Wind-driven ocean circulation and equilibrium statistical mechanics. *J. Mar. Res.*, **47**, 457–492.

HAIDVOGEL, D.B., MCWILLIAMS, J.C., AND GENT, P.R., 1992: Boundary current separation in a quasigeostrophic, eddy-resolving ocean circulation model. *J. Phys. Oceanogr.*, **22**, 882–902.

HERRING, J.R., AND KERR, R.M., 1993: Some contributions of two-point closure to large eddy simulations. In: *Large Eddy Simulation of Complex Engineering and Geophysical Flows.* Eds. B. Galperin and S.A. Orszag. Cambridge University Press, in press.

HOLLOWAY, G., 1986: Eddies, waves, circulation, and mixing: Statistical geofluid mechanics. *Ann. Rev. Fluid Mech.*, **18**, 91–147.

HOLLOWAY, G., 1989: Subgridscale representation. In: *Oceanic Circulation Models: Combining Data and Dynamics.* Eds. D.L.T. Anderson and J. Willebrand, Kluwer, pp. 513–593.

HOLLOWAY, G., 1992: Representing topographic stress for large-scale ocean models. *J. Phys. Oceanogr.*, **22**, 1033–1046.

HOLLOWAY, G., 1993: The role of oceans in climate change: a challenge to large eddy simulation. In: *Large Eddy Simulation of Complex Engineering and Geophysical Flows.* Eds. B. Galperin and S.A. Orszag. Cambridge University Press, in press.

HOLLOWAY, G., AND HENDERSHOTT, M.C., 1977: Stochastic closure for nonlinear Rossby waves. *J. Fluid Mech.*, **82**, 747–765.

INGERSOLL, A.P., 1990: Atmospheric dynamics of the outer planets. *Science*, **248**, 308–315.

KANEDA, Y., AND HOLLOWAY, G., 1992: Rossby wave speeds in β-plane turbulence. In: *Analysis of Nonlinear Phenomena and Its Applications*. Ed. T. Nishida, Showado Publishing Corp., pp. 42–45,

KRAICHNAN, R.H., 1967: Inertial ranges in two-dimensional turbulence. *Phys. Fluids*, **10**, 1417–1423.

KRAICHNAN, R.H., 1971: An almost-Markovian Galilean-invariant turbulence model. *J. Fluid Mech.*, **47**, 513–535.

KRAICHNAN, R.H., 1976: Eddy viscosity in two and three dimensions. *J. Atmos. Sci.*, **33**, 1521–1536.

KRAICHNAN, R.H., AND MONTGOMERY, D., 1980: Two-dimensional turbulence. *Rep. Prog. Phys.*, **43**, 547–619.

KURAMOTO, Y., 1984: *Chemical Oscillations, Waves, and Turbulence*. Springer-Verlag.

KURAMOTO, Y., AND TSUZUKI, T., 1976: Persistent propagation of concentration waves in dissipative media far from thermal equilibrium. *Progr. Theor. Phys.*, **55**, 356.

LESIEUR, M., 1990: *Turbulence in Fluids*, 2nd Edition. Kluwer.

MA, S.-K., 1976: *Modern theory of critical phenomena*. Benjamin/Cummings.

MALTRUD, M.E., AND VALLIS, G.K., 1991: Energy spectra and coherent structures in two-dimensional and beta-plane turbulence. *J. Fluid Mech.*, **228**, 321–342.

MCWILLIAMS, J.C., 1989: Statistical properties of decaying geostrophic turbulence. *J. Fluid Mech.*, **198**, 199–230.

MONIN, A.S., AND PITERBARG, L.I., 1987: On the kinetic equation for Rossby–Blinova waves (in Russian). *Dokl. Akad. Nauk SSSR*, **295**, 816–820.

ORSZAG, S.A., 1977: Statistical theory of turbulence. In: *Fluid Dynamics 1973, Les Houches Summer School in Physics*. Eds. R. Balian and J.-L. Peabe, Gordon and Breach, pp. 237–374.

ORSZAG, S.A., STAROSELSKY, I., AND YAKHOT, V., 1993a: Some basic challenges for large eddy simulation research. In: *Large Eddy Simulation of Complex Engineering and Geophysical Flows*. Eds. B. Galperin and S.A. Orszag. Cambridge University Press, in press.

ORSZAG, S.A., YAKHOT, V., FLANNERY, W.S., BOYSAN, F., CHOUDHURY, D., MARUZEWSKI, J., AND PATEL, B., 1993b: Renormalization group modeling

and turbulence simulations. In: *Near-Wall Turbulent Flows*. Eds. R.M.C. So, C.G. Speziale, and B.E. Launder, Elsevier Science Publishers, pp. 1031–1046.

PANETTA, R.L., 1993: Zonal jets in wide baroclinically unstable regions: persistence and scale selection. *J. Atmos. Sci.*, **50**, 2073–2106.

PEDLOSKY, J., 1987: *Geophysical Fluid Dynamics*, 2nd Edition. Springer-Verlag.

REZNIK, G.M., 1986: Weak turbulence on the β-plane. In: *Synoptic Eddies in the Ocean*. Eds. V.M. Kamenkovich, M.N. Koshlyakov, and A.S. Monin, D. Riedel Publishing Company, pp. 73–107.

RHINES, P.B., 1975: Waves and turbulence on a β-plane. *J. Fluid Mech.*, **69**, 417–443.

RHINES, P.B., 1979: Geostrophic turbulence. *Ann. Rev. Fluid Mech.*, **11**, 401–441.

ROBERT, R., AND SOMMERIA, J., 1991a: Statistical equilibrium states for two-dimensional turbulence. *J. Fluid Mech.*, **229**, 291–310.

ROBERT, R., AND SOMMERIA, J., 1991b: Final equilibrium state of a two-dimensional shear layer. *J. Fluid Mech.*, **233**, 661–689.

ROBERT, R., AND SOMMERIA, J., 1992: Relaxation towards a statistical equilibrium state in two-dimensional perfect fluid dynamics. *Phys. Rev. Lett.*, **69**, 2776–2779.

ROSE, H.A., AND SULEM, P.L., 1978: Fully developed turbulence and statistical mechanics. *J. Phys.*, **39**, 441–484.

SADOURNY, R., AND B[illegible]DEVANT, C., 1985: Parameterization of subgrid scale barotropic and baroclinic eddies in quasi-geostrophic models: Anticipated potential vorticity method. *J. Atmos. Sci.*, **42**, 1353–1363.

SALMON, R., 1982: Geostrophic turbulence. In: *Topics in Ocean Physics*. Proc. Intnl. School Phys. Enrico Fermi, Varenna, Italy, pp. 30–78.

SALMON, R., HOLLOWAY, G., AND HENDERSHOTT, M.C., 1976: The equilibrium statistical mechanics of simple quasi-geostrophic models. *J. Fluid Mech.*, **75**, 691–703.

SEMTNER, A.J., JR., AND CHERVIN, R.M., 1988: A simulation of the global ocean circulation with resolved eddies. *J. Geophys. Res.*, **93**, 15502–15,522; 15,767–15,775.

SEMTNER, A.J., AND CHERVIN, R.M., 1992: Ocean general circulation from a global eddy-resolving model. *J. Geophys. Res.*, **97**, 5493–5550.

SIVASHINSKY, G.I., 1979: On self-turbulization of a laminar flame. *Acta Astronautica*, **6**, 569.

SMAGORINSKY, J., 1963: General circulation experiments with the primitive equations, Part I: The basic experiment. *Mon. Wea. Rev.* **91**, 99–152.

SMAGORINSKY, J., 1993: Some historical remarks on the use of nonlinear viscosities. In: *Large Eddy Simulation of Complex Engineering and Geophysical Flows.* Eds. B. Galperin and S.A. Orszag. Cambridge University Press, in press.

SMITH, L.M., AND REYNOLDS, W.C., 1992: On the Yakhot-Orszag renormalization group method for deriving turbulence statistics and models. *Phys. Fluids A*, **4**, 364–390.

STAMMER, D., AND BÖNING, C.W., 1992: Mesoscale variability in the Atlantic ocean from Geosat altimetry and WOCE high-resolution numerical modeling. *J. Phys. Oceanogr.*, **22**, 732–752.

STAROSELSKY, I., AND SUKORIANSKY, S., 1993: Renormalization group approach to two-dimensional turbulence and the ϵ-expansion for the vorticity equation. In: *Advances in Turbulence Studies.* Eds. H. Branover and Y. Unger, Vol. **149**, *Progress in Astron. and Aeron.*, AIAA, pp. 159–164.

STARR, V.P., 1968: *Physics of Negative Viscosity Phenomena.* Mc-Graw Hill.

VALLIS, G.K., 1992: Problems and phenomenology in two-dimensional turbulence. In: *Nonlinear Phenomena in Atmospheric and Oceanic Sciences.* Eds. G.K. Carnevale and R.T. Pierrehumbert, Springer-Verlag, pp. 1–25.

VALLIS, G.K., AND HUA, B.L., 1988: Eddy viscosity of the anticipated potential vorticity method. *J. Atmos. Sci.*, **45**, 617–627.

VALLIS, G.K., AND MALTRUD, M.E., 1993: Generation of mean flows and jets on a beta-plane and over topography. *J. Phys. Oceanogr.*, **23**, 1346–1362.

WILLIAMS, G.P., 1978: Planetary circulations: 1. Barotropic representation of Jovian and terrestrial turbulence. *J. Atmos. Sci.*, **35**, 1399–1426.

WILSON, K.G., AND KOGUT, J., 1974: The renormalization group and the ϵ-expansion. *Phys. Rep.*, **12C**, 75–199.

YAKHOT, V., AND ORSZAG, S.A., 1986: Renormalization group analysis of turbulence. I. Basic theory. *J. Sci. Comp.*, **1**, 3–51.

YODEN, S., AND YAMADA, M., 1993: A numerical experiment on two-dimensional decaying turbulence on a rotating sphere. *J. Atmos. Sci.*, **50**, 631–643.

FREQUENCY SHIFTS OF ROSSBY WAVES IN β-PLANE TURBULENCE

Yukio Kaneda
Department of Applied Physics, Nagoya University, Nagoya 464-01, Japan

Greg Holloway
Institute of Ocean Sciences, Sidney, BC V8L 4B2, Canada

ABSTRACT

Frequency shifts in Rossby wave propagation due to nonlinear interactions in geostrophic (beta-plane) turbulence are studied by direct numerical simulations and a statistical closure theory. The shifts are of systematic sign and can be quite large as compared with the linear Rossby frequency. Under certain conditions, the closure equations yield a simple approximation for the shifts. An explanation of the shifts is given by a model that includes oscillating random sweeping and strain of large eddies.

INTRODUCTION

The beta-plane model is one of the simplest turbulence models that takes into account planetary gradient of Coriolis effect, obeying

$$\frac{\partial \zeta}{\partial t} + \frac{\partial(\zeta, \psi)}{\partial(x, y)} - \beta \frac{\partial \psi}{\partial x} = \nu \nabla^2 \zeta, \tag{1}$$

where ψ is the stream function related to the fluid velocity as $\mathbf{u} = (\partial\psi/\partial y, -\partial\psi/\partial x)$, $\zeta = -\nabla^2\psi$ is the vorticity, and ν the viscosity. The β term represents Coriolis effect. In the absence of the nonlinear Jacobian term, Eq.(1) exhibits just the Rossby wave propagation, while in the absence of the β term, Eq.(1) is the two-dimensional Navier-Stokes equation. Thus the model provides a simple prototype of wave/turbulence system.

We consider in this paper the frequency shifts of the the Eulerian two-time correlation function in homogeneous and quasi-stationary beta-plane turbulence. Under periodic boundary conditions, in the Fourier space given by

$$\mathbf{u}(\mathbf{x}, t) = \sum_{\mathbf{k}} \mathbf{u}(\mathbf{k}, t) \exp(\, \mathrm{i}\mathbf{k} \cdot \mathbf{x}),$$

the Eulerian correlation spectrum U defined by

$$U(\mathbf{k},\tau,t) =< \mathbf{u}(\mathbf{k},\tau)\cdot\mathbf{u}(-\mathbf{k},t) >, \tag{2}$$

obeys

$$[\frac{\partial}{\partial\tau} + \nu k^2 + \mathrm{i}\omega_0(\mathbf{k})]U(\mathbf{k},\tau,t) = T(\mathbf{k},\tau,t) \equiv - < J(\mathbf{k},\tau)\psi(-\mathbf{k},t) >, \tag{3}$$

where $\omega_0(\mathbf{k}) = -\beta k_x/k^2$ is the linear frequency, and J is the Fourier transform of the Jacobian term in Eq.(1).

The real part of T represents the energy transfer due to nonlinear interactions;

$$[\frac{\partial}{\partial t} + 2\nu k^2]U(\mathbf{k},t) = 2\mathrm{Re}T(\mathbf{k},t),$$

where $U(\mathbf{k},t) \equiv U(\mathbf{k},t,t)$ and $T(\mathbf{k},t) \equiv T(\mathbf{k},t,t)$, while the imaginary part normalized by the energy spectrum U gives

$$\frac{\mathrm{Im}\,T(\mathbf{k},t)}{U(\mathbf{k},t)} = \Delta\omega(\mathbf{k}) \equiv \mathrm{Re}[\bar{\omega}(\mathbf{k})] + \omega_0(\mathbf{k}), \tag{4a}$$

where

$$\bar{\omega}(\mathbf{k}) \equiv \int \omega U(\mathbf{k},\omega)d\omega / \int U(\mathbf{k},\omega)d\omega, \tag{4b}$$

$$U(\mathbf{k},t+\tau,t) = \int U(\mathbf{k},\omega)\exp(i\omega\tau)d\omega.$$

Here and hereafter, we assume quasi-stationarity of turbulence such that the $t-$dependence of $U(\mathbf{k},t+\tau,t)$ is negligible compared with its τ-dependence, and omit the argument t at will. In the absence of the nonlinear interactions, $\Delta\omega$ is zero, and $\Delta\omega$ is therefore a measure of the frequency shifts due to nonlinear interactions.

Direct numerical simulations (DNS) of beta-plane turbulence in planar and spherical geometries so far have suggested that the shifts are westward and nearly proportional to k_x (see the review by Holloway, 1986). There have been theoretical studies on renormalized frequencies that take into account the nonlinear interactions (Legras, 1980; Carnevale and Martin, 1982). However, the reason for the observed shifts remained unknown.

The primary purpose of this paper is to study the frequency shifts by DNS and a two-point closure theory, and to understand the reason. As for the closure theory, we use the Lagrangian renormalized approximation (LRA; Kaneda, 1981), which is derived by systematic Lagrangian renormalized expansions.

STATISTICAL APPROXIMATION (LRA)

In a wide class of two-point closure theories of turbulence, the transfer function T for homogeneous turbulence in a quasi-stationary state is given by an equation of the form

$$T(\mathbf{k}) = \frac{1}{2}\sum_{\mathbf{p},\mathbf{q}}^{\Delta} \frac{|\mathbf{p}\times\mathbf{q}|^2}{k^2p^2q^2}\theta(-\mathbf{k},\mathbf{p},\mathbf{q})$$

$$\times[(p^2-q^2)^2U(\mathbf{p})U(\mathbf{q}) - 2(p^2-q^2)(k^2-q^2)U(\mathbf{k})U(\mathbf{q})], \qquad (5)$$

where $\sum_{\mathbf{p},\mathbf{q}}^{\Delta}$ denotes the sum over $\mathbf{p},\mathbf{q}$ satisfying $\mathbf{p}+\mathbf{q}=\mathbf{k}$. The principal difference between various closure theories comes from the difference of the triple relaxation factors θ.

The application of the LRA to the β-plane model equation (1) yields Eq.(5) with

$$\theta(-\mathbf{k},\mathbf{p},\mathbf{q}) \equiv \int_0^\infty G(-\mathbf{k},\tau)G(\mathbf{p},\tau)G(\mathbf{q},\tau)d\tau, \qquad (6)$$

where G is an appropriately defined Lagrangian response function obeying

$$[\frac{\partial}{\partial\tau} + \nu k^2 + i\omega_0(\mathbf{k})]G(\mathbf{k},\tau) = -2\sum_{\mathbf{p},\mathbf{q}}^{\Delta}\frac{|\mathbf{p}\times\mathbf{q}|^4}{k^2p^2q^2}\int_0^\tau G(-\mathbf{q},s)dsU(-\mathbf{q})G(\mathbf{k},\tau), \quad (7)$$

$$G(\mathbf{k},0) = 1,$$

(Kaneda, 1981; Kaneda and Gotoh, 1988).

In terms of ϕ defined by

$$G(\mathbf{k},\tau) = \exp[-\phi(\mathbf{k},\tau)],$$

Eq.(7) may be written as

$$\frac{\partial^2}{\partial\tau^2}\phi(\mathbf{k},\tau) = 2\sum_{\mathbf{p},\mathbf{q}}^{\Delta}\frac{|\mathbf{p}\times\mathbf{q}|^4}{k^2p^2q^2}\exp[-\phi(-\mathbf{q},\tau)]U(\mathbf{q}), \qquad (8)$$

where

$$\phi(\mathbf{k},0) = 0, \qquad \phi_\tau(\mathbf{k},0) = \nu k^2 + i\omega_0(\mathbf{k}),$$

and we have used $U(\mathbf{q}) = U(-\mathbf{q})$. For small τ, ϕ may therefore be expanded as

$$\phi(\mathbf{k},\tau) = [\nu k^2\tau + \frac{A(\mathbf{k})}{2}\tau^2 + ...] + i[\omega_0(\mathbf{k})\tau + \frac{B(\mathbf{k})}{6}\tau^3 + ...], \qquad (9a)$$

where

$$A(\mathbf{k}) = 2\sum_{\mathbf{p},\mathbf{q}}^{\Delta} |\hat{\mathbf{k}} \times \hat{\mathbf{q}}|^4 \frac{k^2 q^2}{p^2} U(\mathbf{q}), \tag{9b}$$

$$B(\mathbf{k}) = -2\beta \sum_{\mathbf{p},\mathbf{q}}^{\Delta} |\hat{\mathbf{k}} \times \hat{\mathbf{q}}|^4 \frac{k^2 q_z}{p^2} U(\mathbf{q}), \tag{9c}$$

$\hat{\mathbf{k}} = \mathbf{k}/k$, and we have used $\mathbf{p} \times \mathbf{q} = \mathbf{k} \times \mathbf{q}$ for $\mathbf{k} = \mathbf{p} + \mathbf{q}$.

In the following section, we need an estimation for the imaginary part of the triple relaxation factor $\theta(-\mathbf{k}, \mathbf{p}, \mathbf{q})$ with $k \gg q$. If $\phi(\mathbf{p}) \sim \phi(\mathbf{k})$ for $\mathbf{p} = \mathbf{k} - \mathbf{q}$ and $k \sim p \gg q$, then Eq.(6) gives

$$\theta(-\mathbf{k}, \mathbf{p}, \mathbf{q}) \sim \theta(-\mathbf{k}, \mathbf{k}, \mathbf{q}) = \int_0^\infty \exp[-2\phi_R(\mathbf{k}, \tau) - \phi_R(\mathbf{q}, \tau) - i\phi_I(\mathbf{q}, \tau)] d\tau, \tag{10}$$

for $k \gg q$, where ϕ_R and ϕ_I are the real and imaginary parts of ϕ, respectively, and we have used $\phi_R(\mathbf{k}) = \phi_R(-\mathbf{k})$, and the term $\phi_I(\mathbf{k})$ has disappeared because $\phi_I(\mathbf{k}) + \phi_I(-\mathbf{k}) = 0$.

In order to get a rough estimate of the imaginary part, we assume that the terms of higher order in τ may be neglected when Eq.(9a) is substituted into Eq.(10), i.e., we may substitute

$$\phi_R(\mathbf{k}) \sim \nu k^2 \tau + A(\mathbf{k})\tau^2, \qquad \phi_I(\mathbf{k}) \sim \omega_0(\mathbf{k})\tau + B(\mathbf{k})\tau^3, \tag{11}$$

into Eq.(10). This substitution does not imply that the value of ϕ itself is assumed to be well approximated by Eq.(11) in the entire range of τ. It is clear that Eq.(11) may be a poor approximation for large τ, although it may be good for small τ. The substitution implies that we assume the magnitude of the integrand in Eq.(10) to be sufficiently small for large τ (where Eq.(11) may be wrong), and the error caused by the substitution to be not serious.

For the sake of simplicity, we further assume that the anisotropy of the energy spectrum may be negligible, or we may discard the anisotropic part of U, for example by the smallness of β. Then $A(\mathbf{k})$ is a function of only the magnitude k and, Eq.(9c) may be reduced to

$$B(\mathbf{k}) = \omega_0(\mathbf{k}) C(k),$$

where C is a function of only k.

The substitution of Eq.(11) into Eq.(10) then gives

$$\theta(-\mathbf{k}, \mathbf{p}, \mathbf{q}) \sim \int_0^\infty \exp\{-\nu[2k^2 + q^2]\tau - [2A(k) + A(q)]\tau^2 - i\omega_0(\mathbf{q})\tau[1 + C(q)\tau^2]\} d\tau,$$

where we have put $A(\mathbf{k}) = A(k)$. If the viscous term is negligible, and

$$2A(k) \gg A(q), \qquad 2A(k) \gg |C(q)|, \qquad [2A(k)]^{1/2} \gg |\omega_0(\mathbf{q})|, \qquad (12a,b,c)$$

this gives

$$\mathrm{Im}\,\theta(-\mathbf{k},\mathbf{p},\mathbf{q}) \sim -\gamma(k)\omega_0(\mathbf{q}), \qquad (13a)$$

where

$$\gamma(k) \equiv \int_0^\infty \tau \exp[-2A(k)\tau^2]d\tau. \qquad (13b)$$

Although it is not easy to estimate A and C under general conditions, simple estimations are possible under certain idealized conditions and assumptions as follows. Let E and Z be the total energy and enstrophy defined by

$$E \equiv \sum_{\mathbf{q}} U(\mathbf{q}), \qquad Z \equiv \sum_{\mathbf{q}} q^2 U(\mathbf{q}),$$

respectively, and let us assume that most energy and enstrophy are from wavenumbers near k_E and k_Z, respectively, so that E and Z may be approximated as

$$E \sim \sum_{q<K_E} U(\mathbf{q}), \qquad Z \sim \sum_{q<K_Z} q^2 U(\mathbf{q}), \qquad (14a,b)$$

where K_E and K_Z are of the same order with k_E and k_Z, respectively. If $k \gg k_Z$, and the dominant contributions to $A(\mathbf{k})$ and $B(\mathbf{k})$ in Eqs.(9b) and (9c) come from the domain $q < K_Z$, then it is shown after some algebra that Eqs.(9b,c) and (14b) give

$$A(k) \sim \frac{3}{4}Z, \qquad C(k) \sim \frac{1}{4}Z, \qquad (15a,b)$$

while the dimensional consideration based on Eqs.(9b,c) and (14a) gives

$$A(q) = O(k_E^2 E), \qquad C(q) = O(k_E^2 E),$$

for $q \sim k_E \ll k$, provided that the dominant contributions to $A(q)$ and $C(q)$ are from the energy containing range.

The conditions (12a) and (12b) are then well satisfied if

$$Z \gg k_E^2 E. \qquad (16)$$

The numerical factor 2 in front of $A(k)$ in Eq.(12) is insignificant in the estimation of the order of magnitude of the terms, but may be significant numerically in real DNS of limited resolution, where the strong inequalities in Eqs.(12) and (16) may hold only in a weak sense, i.e., the strong inequality "$\gg$" is to be changed to the weaker "$>$".

SIMPLIFIED APPROXIMATION

Let $T^<(\mathbf{k}|K)$ be the contribution from the interactions among the modes $(\mathbf{k},\mathbf{p},\mathbf{q})$ in Eq.(5) with p or $q < K$. The contribution can be estimated in the same way as Kraichnan (1966). Since

$$(p^2-q^2)[(p^2-q^2)U(\mathbf{p})-(k^2-q^2)U(\mathbf{k})] \sim -k^2(\mathbf{q}\cdot\nabla_{\mathbf{k}})[k^2U(\mathbf{k})],$$

for $\mathbf{k}=\mathbf{p}+\mathbf{q}$ and $k \sim p \gg q$, we have for $k \gg K$,

$$T^<(\mathbf{k}|K) \sim -\sum_{q<K}^{\Delta} \theta(-\mathbf{k},\mathbf{p},\mathbf{q})|\hat{\mathbf{k}}\times\hat{\mathbf{q}}|^2U(\mathbf{q})(\mathbf{q}\cdot\nabla_{\mathbf{k}})[k^2U(\mathbf{k})], \tag{17}$$

where $\sum_{q<K}^{\Delta}$ denotes the sum over $\mathbf{q}$ satisfying $\mathbf{k}=\mathbf{p}+\mathbf{q}$ and $q<K$.

In order to get a simplified approximation for the frequency shift $\Delta\omega(\mathbf{k})$ for large wavenumber k, we introduce here the following three assumptions.
(I): The dominant contributions to $\mathrm{Im}\,T(\mathbf{k})$ for $k \gg K_Z$ come from nonlocal interactions with low wavenumbers, so that

$$\mathrm{Im}\,T(\mathbf{k}) \sim \mathrm{Im}\,T^<(\mathbf{k}|K_Z).$$

(II): β is not very large so that the anisotropy of the energy spectrum U is weak and we may therefore neglect its anisotropic part, i.e., we may put

$$U(\mathbf{q}) \sim U(q).$$

(III): The imaginary part of the triple relaxation for $k \gg K_Z$ may be approximated by Eq.(13) in the estimation $\mathrm{Im}\,T^<(\mathbf{k}|K_Z)$.

Under the assumption (III), Eq.(17) yields

$$\mathrm{Im}\,T^<(\mathbf{k}|K_Z) \sim \gamma(k)\sum_{q<K_Z}\omega_0(\mathbf{q})|\hat{\mathbf{k}}\times\hat{\mathbf{q}}|^2U(\mathbf{q})(\mathbf{q}\cdot\nabla_{\mathbf{k}})[k^2U(\mathbf{k})], \tag{18}$$

and under the isotropic assumption (II) this may be further reduced to

$$\mathrm{Im}\,T(\mathbf{k}|K_Z) \sim -\frac{\beta\hat{k}_z\gamma(k)}{8}\sum_{q<K_Z}U(q)\frac{\partial[k^2U(k)]}{\partial k}. \tag{19}$$

A rough estimate of $\gamma(k)$ may be obtained by substituting Eq.(15a) into Eq.(13b), which results in

$$\gamma(k)=\frac{2}{3Z}. \tag{20}$$

If we further assume Eq.(14a), then the assumption (I) and Eq.(19) give

$$\mathrm{Im}\, T(\mathbf{k}) \sim -\frac{\beta \hat{k}_x E}{12Z} \frac{\partial [k^2 U(k)]}{\partial k},$$

and therefore

$$\Delta\omega(\mathbf{k}) = \frac{\mathrm{Im}\, T(\mathbf{k})}{U(\mathbf{k})} = -\frac{\beta \hat{k}_x E}{12ZU(k)} \frac{\partial [k^2 U(k)]}{\partial k}. \tag{21}$$

If $U(k) \sim k^{-m}$, then Eq.(21) yields for $k \gg k_Z$,

$$\Delta\omega(\mathbf{k}) = \frac{(m-2)\beta E}{12Z} k_x. \tag{22}$$

The comparison of Eq.(22) with the linear frequency yields

$$\frac{\Delta\omega(\mathbf{k})}{\omega_0(\mathbf{k})} = -\frac{(m-2)}{12} \frac{k^2}{k_Z^2},$$

while the comparison with the random sweeping frequency, yields

$$\frac{\Delta\omega(\mathbf{k})}{u'k} = -\frac{(m-2)}{6} \frac{k_R^2}{k_Z^2} \frac{k_x}{k},$$

where $k_Z = \zeta'/u'$ is a representative wavelength for a flow with an rms velocity $u' = E^{1/2}$ and an rms vorticity $\zeta' = Z^{1/2}$ and $k_R = (\beta/2u')^{1/2}$ is a representative wavelength obtained by comparing representative wave speed with u' (Rhines, 1975).

Equations (21) and (22) suggest that the frequency shifts are independent of the amplitude of the turbulent flow. However, it is to be remembered that Eq.(21) is based on the assumption (III) or Eq.(13), and in the derivation of Eq.(13) we have assumed Eq.(12) and that the viscosity is negligible. In the limit of weak nonlinearity, Eq.(13) does not hold, but

$$\mathrm{Im}\, \theta(-\mathbf{k}, \mathbf{p}, \mathbf{q}) \sim -\frac{\omega_0(\mathbf{q})}{(2\nu k^2 + \nu q^2)^2 + \omega_0^2(\mathbf{q})}.$$

This can be justified by noting that neglecting the right-hand side of Eq.(8) yields

$$\phi(\mathbf{k}, \tau) = \nu k^2 \tau + i\omega_0(\mathbf{k})\tau.$$

Retracing the derivation of Eq.(21) then gives for $k \gg k_Z$,

$$\Delta\omega(\mathbf{k}) = -\frac{\beta \hat{k}_x E}{32(\nu k^2)^2 U(k)} \frac{\partial [k^2 U(k)]}{\partial k},$$

instead of Eq.(21), provided that the assumptions (I) and (II) are still valid in this limit, and $\nu k^2 \gg |\omega_o(\mathbf{q})|$ for $k \gg K_Z > q$.

It is also to be noted here that if $m < 2$ then the integration of Eq.(18) or Eq.(19) over $\mathbf{q}$ does converge at low q. This implies that the dominant contributions come from local or high wavenumbers, and this is incompatible with the assumption (I). Hence m must satisfy $m > 2$ unless the form $U(k) \sim k^{-m}$ is assumed to be valid only in a local sense.

The approximation (22) has the advantage of simplicity, as compared with Eq.(5), but is based on several assumptions. It is therefore interesting to compare the simplified approximation (22) with the estimate obtained from Eq.(5) without using the assumptions. In the next section, we try such a comparison as well as the comparison of the theory with DNS.

DNS AND NUMERICAL SOLUTION OF THE LRA

Fields satisfying Eq.(1) under periodic boundary conditions were generated by alias-free spectral method with wavenumber increment $\Delta k = 1$ in each of k_x and k_y directions, and retained wavevector domain $k < K_{max}$, where K_{max} is about 85. The initial values of the Fourier components $\mathbf{u}(\mathbf{k})$ were chosen to be normally distributed with given initial isotropic spectrum $U(k, t = 0)$. In the runs reported here, we used $\nu = 0.004$ and $E(k) \equiv \pi k U(k, t = 0) = Ck \exp(-2k/k_0)$, where k_0 is a constant, and the constant C is so normalized that $E = 1$ in each realization. In a series of run (Series B), k_0 was fixed at $k_0 = 5.0$ and β was changed as $\beta = 2.5$, 5.0 and 10.0. These runs are called here as B25, B5 and B10, respectively. In another series (Series K), β was fixed at $\beta = 5.0$ and k_0 was changed as $k_0 = 2.5, 5.0$ and 10.0. These runs are called as K25, K5 and K10, respectively.

In order to avoid the initial rapidly changing phase, we started to take time averages after $t = 0.8$. The averages here are time averages from $t = 0.8$ to $t = 1.0$. In all the runs, the time averaged spectrum $k^4 U(\mathbf{k})$ was observed to be nearly isotropic, and the slope of U was steeper than k^{-4} at high k. The representative wavenumbers $k_Z = \zeta'/u' = \sqrt{Z/E}$ and the total enstrophy in the runs were as follows:

	B5/K5	B25	B10	K25	K10
k_Z	4.64	4.64	4.64	2.95	6.55
Z	17.5	17.5	17.6	8.13	25.4

Thus k_Z and Z are larger for larger k_0 in Series K, as would be expected.

If we take the characteristic eddy-damping time scale as $\tau_D \sim \sqrt{4/3Z}$, which

is suggested from Eqs.(11) and (15a), and the eddy turn over time as $\tau_T \sim 2\pi/\zeta'$, then, for example, for K5/B25 they are given by $\tau_D \sim 0.28$, and $\tau_T \sim 1.5$. Thus the time interval of the averages is comparable to or shorter than the damping and eddy turn over times. (The time interval is limited in our DNS, because in order to avoid extra complexity caused by the introduction of external driving force, we are considering here only freely decaying turbulence in which the statistics cannot be stationary in a strict sense. Better statistics could be obtained by increasing the number of realizations. However, a preliminary test of taking averages over 6 realizations suggested that the results are qualitatively not significantly different from those by one realization.)

The LRA approximation Eq.(5) for $T(\mathbf{k})$ with Eq.(6) was also estimated by numerical computation. The sums over $\mathbf{p}, \mathbf{q}$ in Eqs.(5) and (8) were computed by an alias free spectral method based on the use of Fast Fourier Transform (FFT) as in a previous study, (Gotoh and Kaneda, 1991). In order to avoid large fluctuations in the simulated energy spectrum, we substituted to $U(\mathbf{k})$ the isotropic band-averaged as well as time- averaged spectrum. The wavenumber increment in the computation is $\Delta k = 1$ as in DNS, and the retained wavevector domain was $k < K_{max} \sim 85$. As a preliminary check, we computed $\Delta\omega$ by two ways; one is by using the single-precision FFT and the other by double-precision FFT. Although the value of $\Delta\omega(\mathbf{k})$ at high wavenumbers was found to be very sensitive to the precision, no significant difference was observed at k less than about 40. We therefore present results only for $k_x < 40$, in the followings. The sensitivity at high k is presumably because $U(\mathbf{k})$ is there very small and $\Delta\omega$ has the denominator $U(\mathbf{k})$ as in Eq.(4a).

Figures 1,2 and 3 show the frequency shifts by DNS and the LRA in Series B, while Figs. 1,4 and 5 show the shifts in Series K. In the figures, the values by the simplified approximation (22) are also plotted, where the value $m = 7$, which was guessed from the energy spectrum at $k \sim 20$ or so, is used. The energy spectrum is not rigorously of power low form in the DNS, and this exponent should not be taken too seriously.

Although it is difficult to make detailed quantitative comparisons due to relatively large fluctuations in the simulated values of $\Delta\omega$ taken from short time interval as noted above, the figures show that the slope $\Delta\omega/k_x$ increases with β in Series B, and decreases with k_0, i.e., with the total enstrophy in Series K. The DNS results suggest that the shifts are nearly proportional to k_x, the slopes in the figures are positive (i.e.,$\Delta\omega/k_x > 0$) and $\Delta\omega$ exhibits only weak dependence on k_y. The positivity of the slopes means that the shifts are in the direction of westward phase propagation. These results are in agreement with previous studies, (cf. Holloway, 1986). The results of the LRA as well as the simplified approximation (22) are seen to agree qualitatively with DNS.

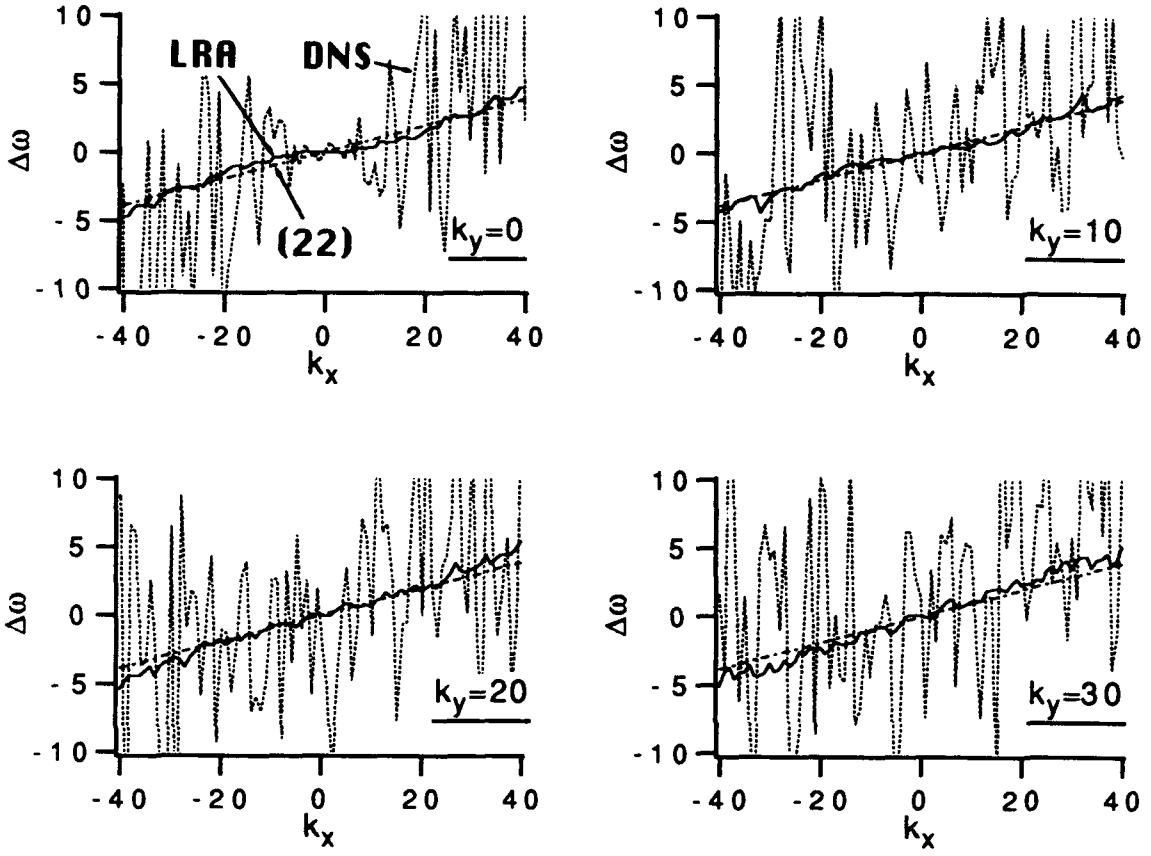


Figure 1. Frequency shift $\Delta\omega$ by LRA, DNS and simplified approximation (22) with m=7 for Case B5/K5 (k_0=5, β=5) at k_y=0,10,20,and 30.

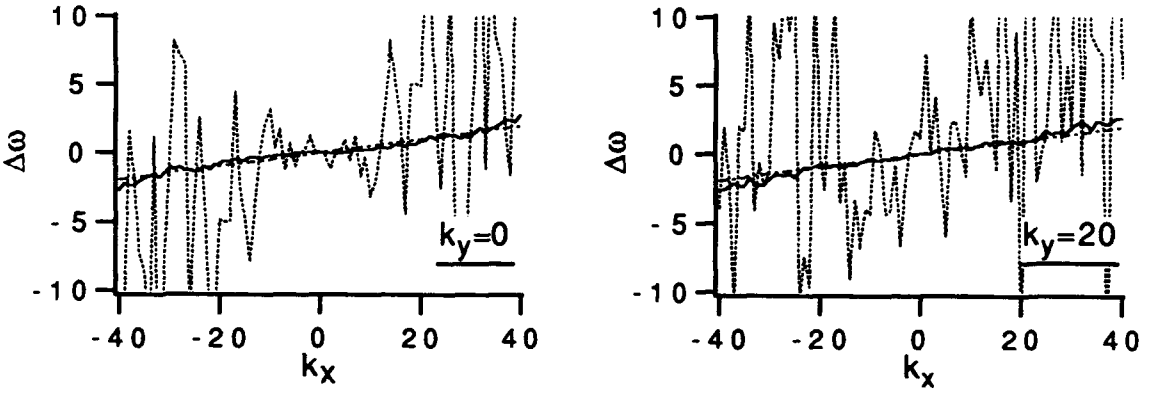


Figure 2. Same as Figure 1, but for Case B25 (k_0=5, β=2.5) at k_y=0 and 20.

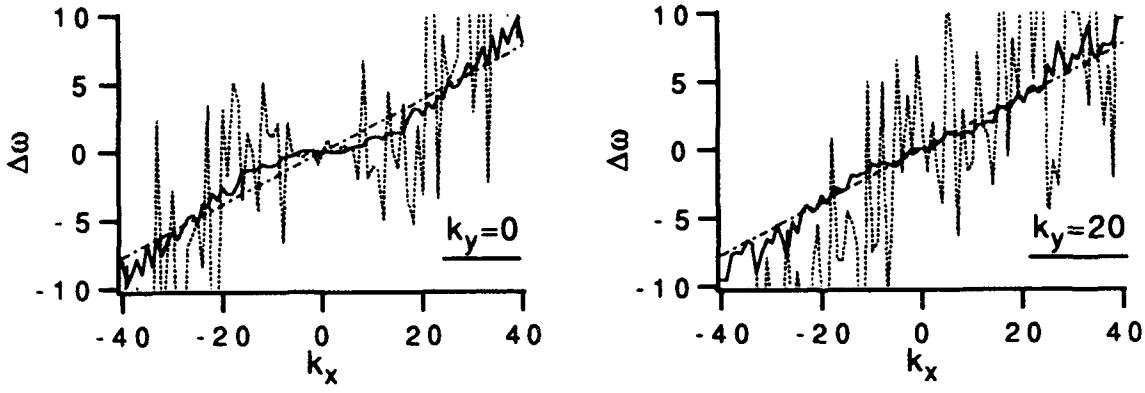


Figure 3. Same as Figure 2, but for Case B10 (k_0=5, β=10).

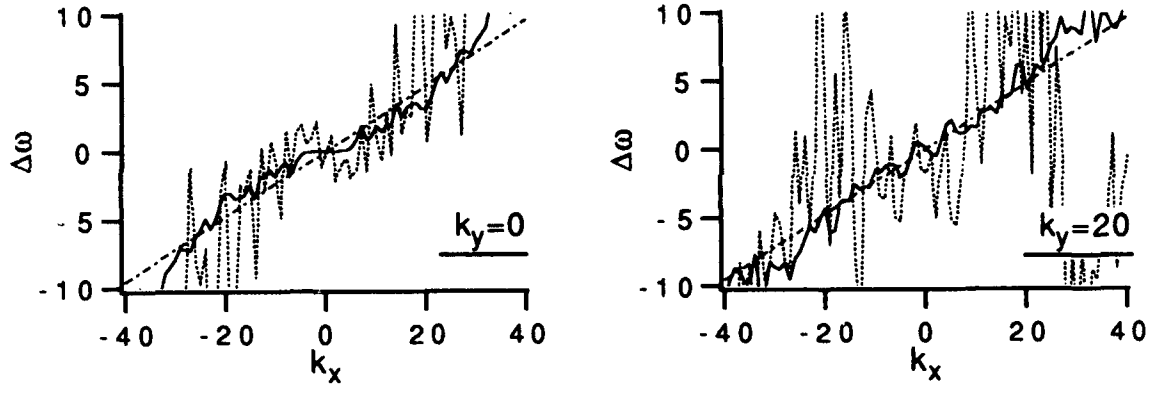


Figure 4. Same as Figure 2, but for Case K25 (k_0=2.5, β=5).

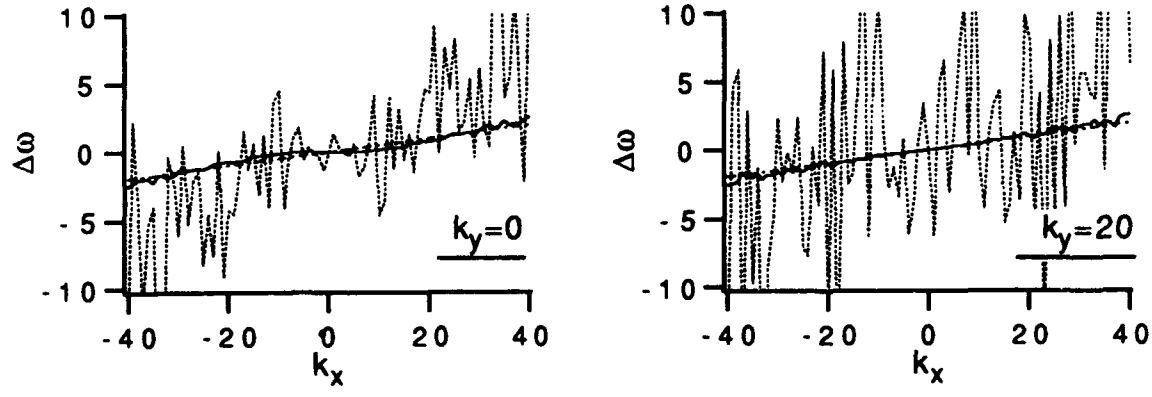


Figure 5. Same as Figure 2, but for Case K10 (k_0=10, β=5).

OSCILLATING RANDOM VELOCITY GRADIENT MODEL

In order to get some idea on the physics underlying the frequency shifts discussed in the previous sections, let us consider the following model equation for the vorticity $\zeta(\mathbf{k})$ of small eddies;

$$[\frac{\partial}{\partial t} + \lambda V]\zeta(\mathbf{k},t) = -\mu(\mathbf{k})\zeta(\mathbf{k},t) + f(\mathbf{k},t), \tag{23}$$

where the parameter λ is introduced for the later convenience, μ a real time-independent deterministic damping factor satisfying $\mu(\mathbf{k}) = \mu(-\mathbf{k})$, f a statistically homogeneous and stationary white-noise random process with zero mean and $f(-\mathbf{k}) = f^*(\mathbf{k})$, and V an operator defined by

$$V = V(\mathbf{k},t) = ik_a U_a(t) + ik_a S_{ab}(t)\frac{\partial}{\partial k_b},$$

in which U_a and S_{ab} are wavevector-independent random variables with zero mean.

The U_a- and S_{ab}-terms are supposed to be models for the effects of random sweeping and random straining of the vorticity field by large eddies, respectively. Such a representation of the effects of the large eddies has been used in studies of the role of large eddies on small eddies, (see, for example, Townsend, 1976). The $\mu-$ term is supposed to represent the viscous damping as well as the eddy-damping due to the nonlinear interactions that are not taken into account by the $V-$ term.

Under the existence of the uniform strain term (S_{ab} term), the two-time Eulerian correlation function, unlike single-time correlation, of ζ obeying Eq.(23) is not homogeneous. This implies that $< \zeta(\mathbf{x},t)\zeta(\mathbf{x}',t') >$ depends on the space variables $\mathbf{x}$ and $\mathbf{x}'$ not only through $\mathbf{x} - \mathbf{x}'$ unless $t = t'$, (cf. Gotoh and Kaneda, 1991). We consider here the Fourier transform of $< \zeta(\mathbf{x},t)\zeta(\mathbf{x}',s) >$ with respect to $\mathbf{x}$ for $\mathbf{x}' = 0$, and define the spectrum $U(\mathbf{k},\tau,t)$ as

$$U(\mathbf{k},\tau,t) \equiv \int < \zeta(\mathbf{k},\tau)\zeta(\mathbf{p},t) > d^2\mathbf{p}/k^2.$$

This definition of U is equivalent to Eq.(2) for homogeneous turbulence.

Multiplying Eq.(23) with $\zeta(\mathbf{p},s)$ and taking the average yield

$$\frac{\partial}{\partial t} < \zeta(\mathbf{k},t)\zeta(\mathbf{p},s) > |_{t=s} = T_\zeta - \mu(\mathbf{k}) < \zeta(\mathbf{k},t)\zeta(\mathbf{p},t) > + < f(\mathbf{k},t)\zeta(\mathbf{p},t) >, \tag{24}$$

where

$$T_\zeta \equiv T_\zeta(\mathbf{k},\mathbf{p},t) = -\lambda < V(\mathbf{k},t)\zeta(\mathbf{k},t)\zeta(\mathbf{p},t) > .$$

Since the imaginary parts of the second and third terms on the right-hand-side of Eq.(24) are zero, Eq.(24) gives

$$\mathrm{Re}[\bar{\omega}(\mathbf{k})] = \frac{\mathrm{Im}\, T_\zeta(\mathbf{k})}{k^2 U(\mathbf{k})}, \tag{25}$$

where

$$T_\zeta(\mathbf{k}) \equiv \int d\mathbf{p} < T_\zeta(\mathbf{k},\mathbf{p},t) >,$$

and $\bar{\omega}(\mathbf{k})$ is defined in the same way as Eq.(4b) through the frequency spectrum $U(\mathbf{k},\omega)$. Because the linear frequency $\omega_0(\mathbf{k})$, i.e. the frequency in the absence of $V-$term, is zero in Eq.(24), Eq.(25) also represents the frequency shift $\Delta\omega(\mathbf{k})$ due to nonlinear interactions, i.e.,

$$\Delta\omega(\mathbf{k}) = \frac{\mathrm{Im}\, T_\zeta(\mathbf{k})}{k^2 U(\mathbf{k})},$$

in the present model (cf. Eq.(4a)).

Since Eq.(23) is linear in ζ, it is possible to solve ζ analytically, but the expression for $\Delta\omega$ would be then quite complicated. Hence we try here a perturbative expansion of $\Delta\omega$ in powers of λ. When $\lambda = 0$, Eq.(23) is just the wellknown Langevin equation. Let ζ_0 be the zeroth order solution of Eq.(23) with $\lambda = 0$, and

$$< \zeta_0(\mathbf{k},t)\zeta_0(\mathbf{q},t) >= \delta(\mathbf{k}+\mathbf{q})k^2 U_0(\mathbf{k}).$$

By discarding terms of $0(\lambda^3)$ and putting $\lambda = 1$, we obtain after some straightforward algebra,

$$\frac{T_\zeta(\mathbf{k})}{k^2} = -\int_0^\infty d\tau < U_\alpha(0) S_{ab}(-\tau) > \exp[-2\mu(\mathbf{k})\tau]\hat{k}_\alpha \hat{k}_a \frac{\partial}{\partial k_b}[k^2 U_0(\mathbf{k})], \tag{26}$$

provided that the term second order in $S_{ab}k_a(\partial/\partial k_b)$ is negligible.

A specific model of the correlation between U and S in Eq.(26) may be obtained by assuming

$$U_a(t) \sim \sum_{q<K} u_a(\mathbf{q},t), \qquad S_{ab}(t) \sim -\sum_{q<K} q_b u_a(\mathbf{q},t),$$

with

$$< U_\alpha(t) S_{ab}(s) >= \sum_{q<K} < u_\alpha(\mathbf{q},t) q_b u_a(-\mathbf{q},s) >, \tag{27}$$

where $K \ll k$. Without loss of generality, we may put

$$< u_\alpha(\mathbf{q},0)u_a(-\mathbf{q},-\tau) >= D_{\alpha a}(\mathbf{q})U(\mathbf{q})\exp[-\phi(\mathbf{q},\tau)], \tag{28}$$

in which $\phi = \phi_R + i\phi_I$ is a function of $\mathbf{q}$ and τ, and the factor $D_{\alpha a}(\mathbf{q}) \equiv \delta_{\alpha a} - \hat{q}_\alpha \hat{q}_a$ ensures the incompressibility condition of the velocity field.

Substituting Eq.(27) with Eq.(28) into Eq.(26) gives

$$\Delta\omega(\mathbf{k}) = \frac{\mathrm{Im}T_\zeta(\mathbf{k})}{k^2U(\mathbf{k})} = -\sum_{q<K} \mathrm{Im}\theta(\mathbf{k},\mathbf{q})|\hat{\mathbf{k}}\times\hat{\mathbf{q}}|^2U(\mathbf{q})(\mathbf{q}\cdot\nabla_\mathbf{k})[k^2U(\mathbf{k})]/U(\mathbf{k}). \tag{29}$$

to the lowest order in λ, where

$$\theta(\mathbf{k},\mathbf{q}) = \int_0^\infty \exp[-2\mu(\mathbf{k})\tau - \phi(\mathbf{q},\tau)]d\tau, \tag{30}$$

and we have used $D_{\alpha a}(\mathbf{q})\hat{k}_\alpha\hat{k}_a = |\hat{\mathbf{k}}\times\hat{\mathbf{q}}|^2$. The right-hand side of Eq.(29) multiplied by $U(\mathbf{k})$ is of the same form as the imaginary part of Eq.(17) except that $\mathrm{Im}\theta(\mathbf{k},\mathbf{p},\mathbf{q})$ is replaced by $\mathrm{Im}\theta(\mathbf{k},\mathbf{q})$ in Eq.(29).

If we choose $\phi(\mathbf{q},\tau) = [\mu(\mathbf{q}) + i\omega(\mathbf{q})]\tau$, then Eq.(30) may be written as

$$\theta(\mathbf{k},\mathbf{q}) = \frac{1}{2\mu(\mathbf{k}) + \mu(\mathbf{q}) + i\omega(\mathbf{q})}. \tag{31}$$

If $\omega(\mathbf{q}) \sim \omega_0(\mathbf{q})$ and $\mu(\mathbf{k}) \gg \mu(\mathbf{q}), |\omega_0(\mathbf{q})|$ for $k \gg K > q$, then

$$\mathrm{Im}\theta(\mathbf{k},\mathbf{q}) \sim -\gamma(\mathbf{k})\omega_0(\mathbf{q}), \qquad \gamma(\mathbf{k}) = \frac{1}{4\mu^2(\mathbf{k})}.$$

By choosing μ as

$$\frac{1}{\mu^2(\mathbf{k})} \sim \frac{8}{3Z}, \tag{32}$$

(i.e., $\gamma(\mathbf{k}) = 2/(3Z)$ as in (20)), and retracing the derivation of (21) from (18), we can recover Eq.(21) from Eq.(29) under the isotropic assumption (II). When $U(k) \sim k^{-m}$, Eq.(29) becomes identical to Eq.(21).

The above model suggests that the correlation $< U_a(0)S_{ab}(\tau) >$ between the random sweeping velocity and strain of large eddies may yield the systematic westward frequency shifts of small eddies. Equation (26) shows that the frequency shifts are smaller for larger damping factor $\mu(\mathbf{k})$ of small eddies. The small eddies have a characteristic life time of order $1/\mu(\mathbf{k})$ associated with the damping factor in Eq.(23). Equation (32) or (11) with (15a) suggests that the life time is shorter for larger total vorticity Z under certain conditions. This results in smaller frequency shifts for larger Z. The result Eq.(29) with Eq.(31) shows that the increase of frequency $\omega(\mathbf{q})$ of large eddies yields larger frequency shifts of small eddies when $|\omega(\mathbf{q})| \ll 2\mu(\mathbf{k})$, but the frequency shifts decrease with the increase of $\omega(\mathbf{q})$ in the opposite limit $|\omega(\mathbf{q})| \gg 2\mu(\mathbf{k})$.

OTHER QUANTITIES

A) Frequency Shifts of Eulerian Response Function

In this paper we have considered the frequency shift $\Delta\omega$ of the Eulerian two-time correlation function U. It might be tempting to relate the shift with that of the Eulerian response function G (or the so-called Eulerian renormalized propagator), which may be defined, corresponding to our use of Eq.(1), as

$$G(\mathbf{k},t,s)\delta(\mathbf{k}+\mathbf{q}) \equiv < \hat{G}(\mathbf{k},\mathbf{q},t,s) >,$$

where $\hat{G}$ is defined as

$$\delta\zeta(\mathbf{k},t) = \int d^2\mathbf{q} \int_{-\infty}^{t} ds \hat{G}(\mathbf{k},\mathbf{q},t,s)\delta f(\mathbf{q},s),$$

in which δf is an infinitesimal disturbance added to the right-hand-side of Eq.(1) and $\delta\zeta$ is the response to the disturbance, and $\hat{G}$ obeys

$$[\frac{\partial}{\partial t} + \nu k^2 + i\omega_0(\mathbf{k})]\hat{G}(\mathbf{k},\mathbf{k}',t,s) = \sum_{\mathbf{p},\mathbf{q}}^{\Delta}(q_x p_y - q_y p_x)(\frac{1}{p^2} - \frac{1}{q^2})\zeta(\mathbf{p},t)\hat{G}(\mathbf{q},\mathbf{k}',t,s), \tag{33a}$$

$$\hat{G}(\mathbf{k},\mathbf{k}',t,t) = 1. \tag{33b}$$

Because $\hat{G}$ is deterministic at $t = s$ and satisfies Eq.(33b) and $< \zeta >= 0$, Eq.(33a) gives

$$[\frac{\partial}{\partial t} + \nu k^2 + i\omega_0(\mathbf{k})]G(\mathbf{k},\mathbf{k}',t,s) = 0, \qquad \text{at} \quad t = s.$$

Unlike to the frequency shift $\Delta\omega$, there is therefore no contribution from the nonlinear interactions to $\Delta\omega_G$, where $\Delta\omega_G$ is defined similarly to Eq.(4a) with T replaced by the average of the right-hand side of Eq.(33a). Thus it is wrong to assume the so-called fluctuation-dissipation approximation

$$U(\mathbf{k},t,s) = U(\mathbf{k})G(\mathbf{k},t,s),$$

as far as the shift $\Delta\omega_G$ is concerned, and the shift $\Delta\omega$ of Eulerian two-time correlation should not be confused with the shift $\Delta\omega_G$ of the response function.

B) Frequency Shifts of Lagrangian Correlation Function

Another quantity which might be related to the shift $\Delta\omega$ is the shift of Lagrangian correlation. Let $U_L(\mathbf{k},\tau,t)$ be the Fourier transform with respect $\mathbf{r}$ of the Lagrangian two-time velocity correlation $< \mathbf{v}(\mathbf{x}+\mathbf{r},t;\tau)\cdot\mathbf{v}(\mathbf{x},t;t) >$, where $\mathbf{v}(\mathbf{x},t;\tau)$ is the velocity at time τ of the fluid particle that was at $\mathbf{x}$ at time t.

Because

$$\frac{\partial}{\partial\tau} < \mathbf{v}(\mathbf{x}+\mathbf{r},t;\tau)\cdot\mathbf{v}(\mathbf{x},t;t) >=< [\frac{\partial}{\partial\tau}\mathbf{v}(\mathbf{x}+\mathbf{r},t;\tau)]\cdot\mathbf{u}(\mathbf{x},t) >,$$

and

$$\frac{\partial}{\partial\tau}\mathbf{v}(\mathbf{x},t:\tau)|_{\tau=t} = [\frac{\partial}{\partial t}+(\mathbf{u}(\mathbf{x},t)\cdot\nabla)]\mathbf{u}(\mathbf{x},t) = -\nabla p - (\text{ terms linear in } \mathbf{u}),$$

it is shown that

$$[\frac{\partial}{\partial\tau}+\nu k^2+ \mathrm{i}\omega_0(\mathbf{k})]U_L(\mathbf{k},\tau,t) = 0, \quad \text{at } \tau = t, \tag{34}$$

where we have used $< \nabla p\cdot\mathbf{u} >= 0$ in homogeneous turbulence, in which p is the pressure. Unlike the Eulerian spectrum U in Eq.(3), there is therefore no contribution from the nonlinear interactions to the $\tau-$ derivative at $\tau = t$ of the Lagrangian spectrum U_L. Thus the frequency shift $\Delta\omega$ should not be confused with that of Lagrangian correlation U_L.

In the LRA, U_L is given by $U_L(\mathbf{k},\tau,t) = G(\mathbf{k},\tau,t)U(\mathbf{k},t)$ and the LRA with Eq.(7) is consistent with Eq.(34). Because

$$(\partial/\partial t)U(\mathbf{k},t) = (\partial/\partial\tau)U_L(\mathbf{k},\tau,t)+(\partial/\partial\tau)U_A(-\mathbf{k},\tau,t), \quad \text{at } \tau = t,$$

and $U(\mathbf{k},t)$ is real, Eq.(34) also implies that there is neither contribution from the nonlinear interactions to $\mathrm{Im}(\partial/\partial\tau)U_A(\mathbf{k},\tau,t)$ at $\tau = t$, where $U_A(\mathbf{k},\tau,t)$ is the Fourier transform of $< \mathbf{v}(\mathbf{x}+\mathbf{r},\tau;\tau)\cdot\mathbf{v}(\mathbf{x},\tau;t) >$ and $(\partial/\partial\tau)U_A(-\mathbf{k},\tau,t)$ is the key quantity in the Abridged Lagrangian History Direct Interaction Approximation by Kraichnan (1965).

C) Frequency Shifts in Inviscid Truncated System

The inviscid truncated model of Eq.(1) with a retained wavevector domain D has an equilibrium state characterized by the equilibrium energy spectrum

$$U(\mathbf{k}) = \frac{1}{a+bk^2} \quad \text{in D},$$

where a and b are constants (Salmon et al., 1976). Since Eq.(5) gives

$$\frac{T(\mathbf{k})}{U^2(\mathbf{k})} = \frac{1}{2}\sum_{\mathbf{p},\mathbf{q}\in D}^{\Delta} \theta(-\mathbf{k},\mathbf{p},\mathbf{q})\frac{|\mathbf{p}\times\mathbf{q}|^2}{k^2p^2q^2}U(\mathbf{p})U(\mathbf{q})\{[\frac{p^2-q^2}{U(\mathbf{k})}-\frac{k^2-q^2}{U(\mathbf{p})}]^2-[\frac{k^2-q^2}{U(\mathbf{p})}]^2\},$$

(a similar expression has been derived by Carnevale et al., 1981), it is shown that if U is given by the equilibrium spectrum and if the triple relaxation factor satisfies the symmetry $\theta(-\mathbf{k},\mathbf{p},\mathbf{q}) = \theta(-\mathbf{k},\mathbf{q},\mathbf{p})$ between $\mathbf{p}$ and $\mathbf{q}$, then $T(\mathbf{k})$ is identically zero, i.e., not only the real but also the imaginary part of $T(\mathbf{k})$ are zero. The triple relaxation factor of the LRA given by Eq.(6) in fact satisfies the symmetry, and the LRA therefore yields $\Delta\omega(\mathbf{k}) = 0$ at the inviscid equilibrium state.

D) Complex Eddy Viscosity

There are various ways to define eddy viscosity. Following Kraichnan (1976), we consider here the following definition of the eddy viscosity ν_T;

$$\nu_T(\mathbf{k}|K) \equiv -T^{>}(\mathbf{k}|K)/[k^2U(\mathbf{k})],$$

where $T^{>}(\mathbf{k}|K)$ is the contribution to $T(\mathbf{k})$ from the interactions among the modes $(\mathbf{k},\mathbf{p},\mathbf{q})$ with p or $q > K$. By assuming $U(\mathbf{q}) \ll U(\mathbf{k})$ for $q \gg k$, and noting that Eq.(5) gives

$$\frac{T^{>}(\mathbf{k}|K)}{U(\mathbf{k})} \sim \frac{1}{2}\sum_{q>K}^{\Delta} \theta(-\mathbf{k},\mathbf{p},\mathbf{q})\frac{|\hat{\mathbf{p}}\times\hat{\mathbf{q}}|^2}{k^2}(q^2-p^2)\{[p^2U(\mathbf{p})-q^2U(\mathbf{q})]+k^2[U(\mathbf{q})-U(\mathbf{p})]\},$$

for $k \ll K$, we obtain

$$\nu_T(\mathbf{k}|K) = \sum_{q>K}^{\Delta} \theta(-\mathbf{k},\mathbf{p},\mathbf{q})(\hat{\mathbf{k}}\times\hat{\mathbf{q}})^2\frac{(\hat{\mathbf{k}}\cdot\mathbf{q})}{q^2}(\hat{\mathbf{k}}\cdot\nabla_{\mathbf{q}})[q^2U(\mathbf{q})], \tag{35}$$

for $k \ll K$, where the triple relaxation factor may be approximated as

$$\theta(-\mathbf{k},\mathbf{p},\mathbf{q}) \sim \theta(-\mathbf{k},\mathbf{q},\mathbf{q}) = \int_0^{\infty} \exp[-2\phi_R(\mathbf{q},\tau) - \phi_R(-\mathbf{k},\tau) - i\phi_I(-\mathbf{k},\tau)]d\tau,$$

provided that $\phi(\mathbf{p}) \sim \phi(-\mathbf{q})$ for $\mathbf{p} = \mathbf{k} - \mathbf{q}$ and $k \ll q \sim p$.

If we suppose $U(\mathbf{q}) \sim U(q)$ and

$$\phi_I(\mathbf{k},\tau) \sim \alpha(k)\omega_0(\mathbf{k})\tau, \qquad |\phi_I(\mathbf{k},\tau)| \ll \phi_R(\mathbf{q},\tau),$$

for $\tau = O(\tau_R(\mathbf{k}))$ in Eq.(35), then

$$\mathrm{Im}\nu(\mathbf{k}|k_L) = \frac{\alpha(k)\omega_o(\mathbf{k})}{8} \sum_{q>k_L} \frac{\gamma(\mathbf{q})}{q} \frac{\partial[q^2 U(q)]}{\partial q},$$

where $\tau_R(\mathbf{k})$ is the characteristic time scale of $\phi_R(\mathbf{k}, \tau)$, and

$$\gamma(\mathbf{q}) = \int_0^\infty \tau \exp[-\phi(\mathbf{q}, \tau)] d\tau.$$

Thus the imaginary part of the viscosity ν_T may be nonzero.

CONCLUSION

The results obtained in the present paper may be summarized as follows.

I]. The DNS and the LRA agree in the following points:
(1) the shifts are westward, i.e., $\Delta\omega > 0$ for $k_x > 0$,
(2) the shifts are nearly proportional to k_x,
(3) the shifts increase with β,
and (4) the shifts increase with E/Z, but are independent of either amplitude under certain conditions.

II]. The above properties may be explained by a model that includes
(1) oscillating random sweeping and strain of large eddies,
and (2) eddy-damping of small eddies.
These are represented by the $V-$ and $\mu-$ terms in the model (23). The LRA as well as the model suggests that the shifts may occur even if the energy spectrum is nearly isotropic.

III]. The time dependence of Eulerian correlation should not be confused with those of Eulerian response function and/or Lagrangian correlation. It is wrong to assume the fluctuation-dissipation relation for Eulerian correlation. An analysis of the nonlocal interactions suggests that eddy viscosity may be complex.

The present paper treats only cases of small β, and the effects of high β and strong anisotropy are remained to be studied. The role of coherent structure, which was not taken into account in the theory, remains an open question.

REFERENCES

Carnevale G.F., U.Frisch and R.Salmon, 1981: H theorem in statistical fluid dynamics, *J. Phys. A*,**14**,1701-1718.

Carnevale G.F. and P.C. Martin, 1982: Field theoretical techniques in statistical fluid dynamics, *Geophys. Astrophys. Fluid Dyn.*,**20**,131-164.

Gotoh T., and Y.Kaneda, 1991: Lagrangian velocity autocorrelation and eddy viscosity in two-dimensional anisotropic turbulence, *Phys.Fluids*,**3**,2426-2437.

Holloway, G., 1986: Eddies, waves, circulation, and mixing: Statistical geofluid mechanics, *Ann. Rev. Fluid Mech.*,**18**,91-147.

Kaneda Y., 1981: Renormalized expansions in the theory of turbulence with the use of the Lagrangian position function, *J.Fluid Mech.*,**107**,131-145.

Kaneda Y., and T. Gotoh, 1988: A generalized expression and applications of a Lagrangian renormalized closure, *Research Report of Res. Inst. Math. Sci. (Kyoto Univ.)*, **652**,159-171.

Kraichnan R.H., 1965: Lagrangian-history closure approximation for turbulence, *Phys.Fluids*,**8**,575-598.

Kraichnan R.H., 1966: Isotropic turbulence and inertial-range structure, *Phys.Fluids*,**9**,1728-1752.

Kraichnan, R.H., 1976: Eddy viscosity in two and three dimensions, *J.Atmos.Sci.*, **33**,1521-1536.

Legras B., 1980: Turbulent phase shifts of Rossby waves, *Geophys. Astrophys. Fluid Dyn.*,**15**,253-281.

Rhines, P.B., 1975: Waves and turbulence on a beta-plane, *J.Fluid Mech.*,**69**,417-443.

Salmon R., G.Holloway and M.C.Hendershott, 1976: The equilibrium statistical mechanics of simple quasi-geostrophic models, *J.Fluid Mech.*,**75**,691-703.

Townsend A.A., 1976: *The structure of turbulent shear flow*, Cambridge University Press.

STATISTICAL MECHANICS, TURBULENCE, AND OCEAN CURRENTS

Geoffrey K. Vallis
Institutes of Nonlinear Science and Marine Science, University of California, Santa Cruz

ABSTRACT

This paper examines the formulation, application and utility of certain ideas from equilibrium statistical mechanics to physical oceanography. In particular we discuss the connection of these ideas to selective decay (minimum enstrophy) theories and to the production of nonlinearly stable states, as well as its limitations when dealing with forced-dissipative flows. A robust prediction of the theories discussed is the generation of mean flows around topography, which should be amenable to observational verification or falsification.

1. INTRODUCTION

The goal of this paper is to try to put into an oceanic framework various concepts from the fields of equilibrium statistical mechanics and from turbulence, and to attempt to understand their importance and relevance, if any, to the circulation of the world's oceans. To these ends we first briefly summarize the theory of equilibrium statistical mechanics as applied to geophysical fluids, determining the conditions under which it applies, and, in those conditions, what the predictions of the theory are. Then, we discuss whether numerical simulations of the equations of motion do in fact give rise to the predicted (maximum entropy) solutions under the appropriate conditions.

The statistical equlibrium has often been compared to an alternative theory, the 'selective decay' or 'minimum enstrophy' hypothesis, which predicts evolution toward the nonlinearly stable 'minimum enstrophy' state, and we shall discuss this connection. Both of these theories are in a sense equilibrium theories: the statistical mechanics applies to time or ensemble averages, and the selective decay theory predicts the end-state of a weakly decaying sysem. Neither can describe certain disequilibrium phenomena arising in forced-dissipative situations. We shall show that certain important phenomena, such as the formation of jets on a beta-plane, cannot in fact be described by such theories. Finally, we discuss the application of both equilibrium and non-equilibrium theories to real oceanic flows, and briefly discuss where their predictions could be observationally tested.

2. THEORETICAL FORMULATION

The simplest model system with which to fix ideas is the barotropic vorticity equation, to wit:

$$\frac{\partial \zeta}{\partial t} + (\mathbf{u} \cdot \nabla)\zeta = 0 \tag{2.0}$$

where $\mathbf{u}$ is the two-dimensional velocity and ζ is the vorticity. In terms of a stream function, $\mathbf{u} = \nabla \times \mathbf{k}\psi$ and $\zeta = \nabla^2 \psi$. If the domain is homogeneous, there is no mean flow, and energy and enstrophy are both conserved. In fact, any integral function of the vorticity is conserved. To see this, note that the equation states that the evolution consists merely of a continuous re-arrangement of the vorticity, which is conserved on each parcel. Similarly, any function of vorticity is conserved on parcels. Thus, an integral over the domain of any function of the vorticity is preserved, since the integration is indifferent to the location of the parcels themselves. The quadratic invariants, energy E, and enstrophy Z_2 are given by

$$E = \frac{1}{2}\int_S \mathbf{u} \cdot \mathbf{u}\, dx \tag{2.1}$$

$$Z_2 = \frac{1}{2}\int_S \zeta^2\, dx. \tag{2.2}$$

Circulation,

$$Z_1 = \int_S \zeta\, dx, \tag{2.3}$$

is also conserved. Of all the integral invariants, these three have assumed a special importance in the equilibrium theory, as discussed further in section 4.

The *inviscid* equation of motion may be written in the form

$$\frac{d\zeta_k}{dt} = \sum_{kpq} A_{kpq}\zeta_p\zeta_q \tag{2.4}$$

where ζ_k is the spectral coefficient of the kth wavevector, and the geometric interaction coefficients A_{kpq} are zero unless the wavevectors form a triad $\mathbf{k} + \mathbf{p} + \mathbf{q} = 0$. Also, if two of the three members are equal $A_{kpq} = 0$. These conditions lead to

$$\frac{d\dot{\zeta}_k}{d\zeta_k} = 0. \tag{2.5}$$

This important property means the system is Louivillean (in fact the system satisfies the detailed Louivillean property). In the phase space of the spectral coefficients the motion is therefore incompressible. If an ensemble of system states is represented by a cloud in the phase space, the cloud preserves its volume. Statistical mechanics, in particular the notion that the properties of a system will be given by the maximum entropy state, may then be applied. There are now two ways to proceed. In the first—perhaps the more conventional—one assumes that a single system will explore all accessible phase space with equal likelihood—this is the *ergodic hypothesis*. Alternatively, one may assume, without reference to the ergodic hypothesis, that the least biased assumption one can make about the averaged properties of a system is that its time average state is given by the maximum likelihood state. This is the information theory approach (Jaynes 1979). Although the underlying philosophy is different, either way one must compute the maximum entropy state. The difference in these two attitudes far transcends applications in geophysical fluid dynamics; it goes to the heart of statistical mechanics, and we will not discuss in any detail the differences here. The information theory approach requires no assumptions about the behaviour of a system; it merely says "this is what we know about a system, and this is what we can predict without implicitly making additional assumptions." No 'mixing' hypotheses, for example, are required. It is essentially a Bayesian approach (see e.g., Gull 1991, and other articles in Buck and Macaulay 1991). The prior constraints are the known invariants: in principle any invariant or other constraint could be built in. Aside from these constraints equal probability is assigned to each micro-state—a principle of least bias, which says nothing about how a system may actually behave. In spite of this seemingly rational basis, many physicists are uncomfortable with the information theory approach, since for any given system it offers little assurance that its predictions will be of any use whatsoever, and it seems to divorce the predictions one makes of a system from its physics. The ergodic hypothesis, on the other hand, makes the explicit physical assumption that a system will explore all regions of phase space available to it, constrained by the global integral invariants. However, it is generally extremely difficult to rigorously prove that a particular system is ergodic, and for most systems it remains an assumption.

In either case, the time or ensemble averaged state is given by the maximum entropy state. The problem is thus to maximize the entropy,

$$S = \sum_i p_i \log p_i \tag{2.6}$$

where p_i is the probability of the system being in the i^{th} microstate, subject to the inviscid constraints. Assuming for the moment that only the quadratic constraints (2.1) and (2.2) and the circulation (2.3), are relevant, the system will satisfy a Gibbs distribution

$$p_i = \exp -(\alpha E + \gamma Z_1 + \delta Z_1) \tag{2.7}$$

where the parameters α, γ and δ are Lagrange multipliers. (See e.g. Tolman 1938. See Holloway 1986 for a review of many applications of statistical methods to geophysical fluid mechanics.) Fairly standard methods can then be used to make predictions of the mean flow and its variance (e.g. Kraichnan 1975, Salmon, Holloway and Hendershott 1976). The spectrum of eddy kinetic energy of the maximum entropy state is given by (Kraichnan 1975):

$$E(k) = \frac{\pi k}{\alpha(\mu + k^2)} \tag{2.8}$$

In a homogeneous environment (for example a doubly periodic flow with no topography) there is no mean flow. However, the presence of topography, or of boundaries, will in general produce a mean flow. The equation of motion is then

$$\frac{\partial q}{\partial t} + J(\psi, q) = 0 \tag{2.9}$$

where $q = \nabla^2\psi + h(x,y) + \beta y$. The beta effect appears in formally the same way as topography. In a closed domain with boundary conditions of no normal flow (or in a channel geometry) the enstrophy constraint (2.2) is replaced by the condition that potential enstrophy Q_2 is conserved, where

$$Q_2 = \int_S q^2 \, dx. \tag{2.10}$$

The steady component of the maximum entropy flow is then given by the linear relationship,

$$< q >= \mu < \psi > + \lambda, \tag{2.11a}$$

which gives the Helmholtz equation

$$(\mu - \nabla^2) < \psi >= \beta y + h(x,y) - \lambda. \tag{2.11b}$$

The values of λ, μ, and α are determined implicitly by the values of the energy, enstrophy and circulation. In a doubly periodic domain the β-effect plays no direct role. This is because in such a domain the inviscid invariants do not depend on beta: Z_2 remains invariant, and there is no mean flow, as demanded also by homogeneity. This is however, rather a special case because Z_2 is not a Casimir: its conservation depends on the special relationship between ζ and ψ. In a closed domain, or in a zonal channel, the Casimir Q_2 is conserved, and β will in general affect the mean flow.

3. EXPERIMENTAL VERIFICATION

We now examine for a few cases whether a single system does indeed evolve into a maximum entropy state. A strict information theorist might argue that it is irrelevant to the theory whether or not a system is ergodic; however, ergodicity is an interesting property in its own right, regardless of its role in the foundations of statistical mechanics. Most numerical models conserve only the quadratic invariants, plus circulation, and are not guaranteed to respect the higher order invariants. Thus, one aspect that will be of interest is whether a numerical simulation will evolve into a state governed by only by the quadratic invariants, or whether higher order invariants may nevertheless somehow play a role.

The equilibrium spectrum of inviscid two-dimensional fluids in doubly periodic domains has been demonstrated by Carnevale (1982) and Carnevale and Vallis (1984). Using a de-aliased spectral code which exactly conserves energy and enstrophy—such a model may be termed 'quadratically inviscid'—then for a long enough time average the energy spectrum is found to be that of (2.8), and illustrated in Figure 1.

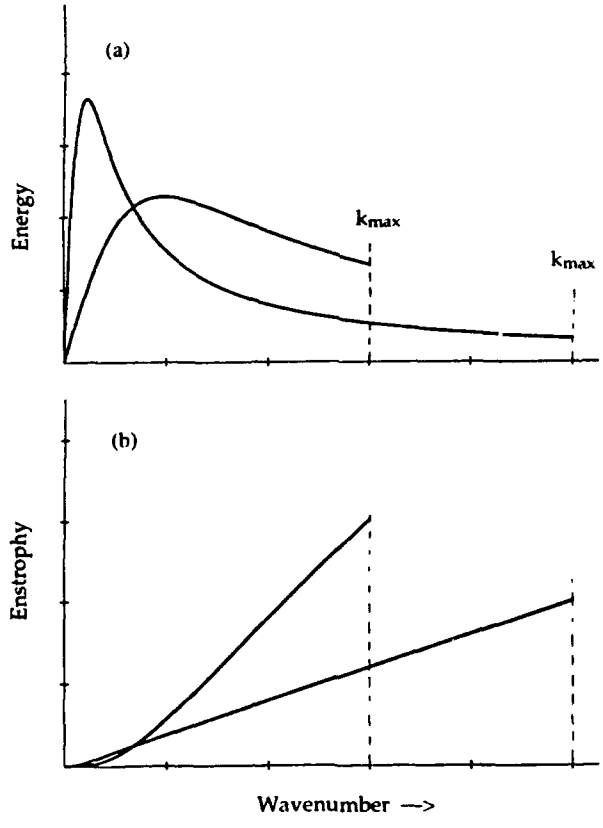


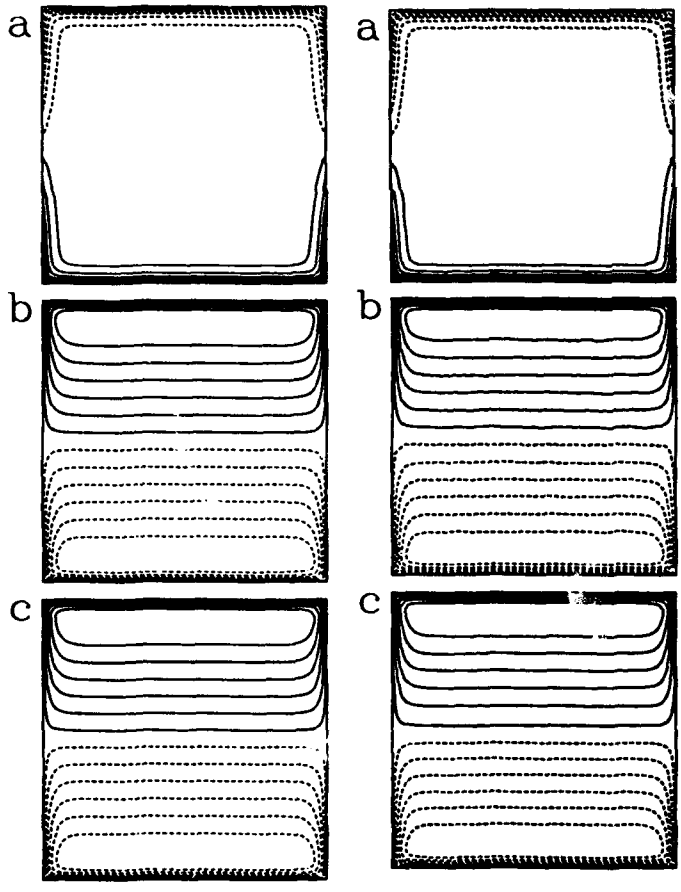
Figure 1. Predicted equilibrium energy spectrum (a), and enstrophy spectrum (b), in a spectrally truncated inviscid model, for two values of the truncation wavenumber k_{max}.

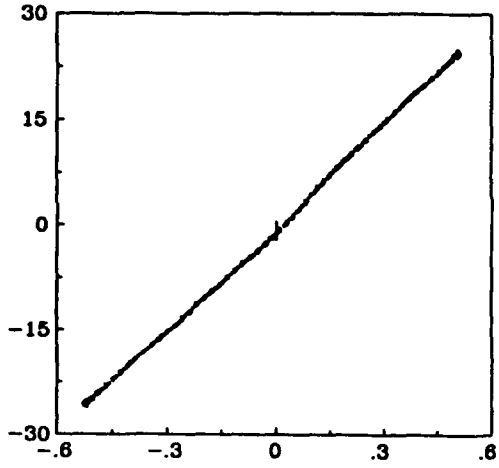
In a closed domain, the mean flows of the maximum entropy states are known as Fofonoff flows. Fofonoff (1954 and 1962) studied the analytical solution of equation (2.11) in a square basin with no-normal flow boundary condition. Wang and Vallis (1993) integrated (2.9) in a closed domain with a quadratically inviscid numerical model. A typical (analytic) Fofonoff solution, for positive μ, is shown in Figure 2. The absolute vorticity field is parallel to the streamfunction field, as required by the linear relationship. The relative vorticity is confined to the boundary layer, the thickness of which, l, is given by

$$l \sim \frac{1}{\sqrt{u}} \sim R_o^{2/3} L, \tag{3.0}$$

where R_o is the β-plane Rossby number, defined as $R_o \equiv U_{rms} / \beta L^2$, where U_{rms} is the root mean square velocity, and L is the basin size. The boundary layer gets thinner as β increases, or total energy decreases. The absolute vorticity field is dominated by the the planetary vorticity βy inside the basin, where the flow is westward, the flow returns to the eastern boundary through northern and southern boundary layers, forming two gyres, anticyclonic in the northern basin, cyclonic in southern basin. The parameter λ affects the symmetry of the fields: in a domain stretching from $y = -L$ to $+L$ then, for zero λ, the fields are symmetrical about $y = 0$; for general non-zero λ, one gyre will be enlarged, while the other will be squeezed, and in the extreme case one gyre can fill out the whole basin.

Just as for the spectral doubly periodic case, the inviscid simulations do show an approach to a maximum entropy state.† Figure 3 shows the resulting time averaged streamfunction and Figure 4 shows a scatter plot of streamfunction versus potential vorticity. Many other





(left) Figure 2. Analytic Fofonoff solution in a closed domain ($\mu = 40$, $\beta = 10$ and $\lambda = 0$. (a) Relative vorticity. (b) Potential vorticity. (c) Streamfunction.

(center) Figure 3. As for Figure 2, but the results are now obtained as the time average of an energy and enstrophy conserving numerical simulation.

(above) Figure 4. Scatter plot of $q - \psi$, corresponding to the solution in Figure 3.

† Strictly, this and other numerical demonstrations in this article are not rigorous demonstrations of ergodicity. They merely demonstrate that the system reaches a macro-state similar to that of the maximum entropy state.

simulations and details are presented in the paper by Wang and Vallis (1993). They consider different shaped domains, various parameter values and cases with topography. It appears that, in general, inviscid integrations do indeed evolve in such a way that the time averaged flow is very similar to the maximum entropy state. In other words, the simulations suggest the flow is ergodic.

4. SELECTIVE DECAY, STABILITY THEORY AND HIGH ORDER INVARIANTS

Minimum enstrophy states

Related, certainly in the minds of many oceanographers, to maximum entropy theories are so-called selective decay theories (Bretherton and Haidvogel 1976). Although apparently not based on such a fundamental tenet as maximizing entropy, these can and have been applied in a number areas of physics, such as magneto-hydrodynamics, as well as in geophysics. (Other variational principles, such as 'minimum energy dissipation,' exist. Montgomery and Phillips, 1990, argue that minimum energy dissipation is a consequence of maximum entropy, and we shall see that is also true for the minimum enstrophy principle.) Consider two-dimensional or quasi-geostrophic flows. Then, in any turbulent situation enstrophy may be expected to be dissipated by viscosity at a much faster rate then energy. This is because (in the classic theory of two-dimensional turbulence) energy is trapped at the relatively inviscid large scale whereas enstrophy is transferred to small scales where it may be dissipated by viscosity. Indeed, in the limit of zero viscosity, enstrophy is dissipated whereas energy is conserved. (This is an equilibrium prediction, which does not violate the regularity results that enstrophy dissipation is zero if viscosity is zero. See Vallis 1985 and 1992.) Thus, the end state of decaying system may be expected to be close to a minimum enstrophy state for a given energy. Consider arbitrary variations satisfying $\psi = 0$ on the boundary Γ, minimizing potential enstrophy Q_2 for given circulation Q_1 and energy E. We require

$$\delta\int_S \frac{1}{2}(\nabla^2\psi+\beta y)^2\,dx\,dy+\mu\delta\int_S \frac{1}{2}(\nabla)^2\,dx\,dy-\lambda\delta\int_S \nabla^2\psi\,dx\,dy=0. \tag{4.0}$$

After integrating by parts this yields

$$\int_S \nabla^2(\nabla^2\psi+\beta y-\mu\psi)\delta\,dx\,dy+\int_\Gamma(\nabla^2\psi+\beta y-\lambda)\frac{\partial\delta\psi}{\partial n}ds=0 \tag{4.1}$$

which gives, since both $\delta\psi$ and boundary value of $\partial\delta\psi/\partial n$ are arbitrary,

$$\nabla^2(\nabla^2\psi+\beta y-\mu\psi)=0 \text{ within } S, \tag{4.2}$$

and

$$\nabla^2\psi + \beta y - \lambda = 0 \text{ on } \Gamma. \tag{4.3}$$

Thus, using $\psi = 0$ on Γ, we obtain

$$(\mu - \nabla^2)\psi = \beta y - \lambda \text{ everywhere.} \tag{4.4}$$

Hence, minimization of potential enstrophy gives the same linear relationship between absolute vorticity and streamfunction as the maximum entropy prediction, although we have not yet shown that the parameters are the same. But in fact, in the limit of infinite resolution, the maximum entropy state is identical to the minimum potential enstrophy state

For the sake of discussion, consider flow in a periodic domain, with $\beta = 0$. Carnevale and Frederikson (1987) show that a steady flow defined by the form (2.11a) or (4.4) is stable in the sense of Lyapunov, in that the maximum amplitude to which a perturbation may grow is bounded by its initial amplitude, if $\mu > k_o^2$ where k_o is a wavenumber smaller than the smallest wavenumber of the topography. (If $q'(\psi)$ is positive everywhere, stability follows immediately by Arnol'd's first theorem; Arnol'd 1966.) Stability occurs physically because the branch of solutions with $\mu > k_o^2$ corresponds to a mimimum enstrophy state. (A state of maximum enstrophy for a given energy is also stable, although it does not correspond to a physically realizable state at infinite resolution.) In general, any physical state which corresponds to an extremum of conserved quantities must be stable, for if the system is perturbed from that state, it must remain close to the extremum state. Thus, a minimum enstrophy state for a given energy is stable, since this is an extremum of the conserved quantitiy $Q_2 + \mu E$. It is equivalent to maximum energy state for a given enstrophy. Now, the maximum entropy state is not a steady state, and it is not appropriate to call it a 'stable' state. However, in the limit of infinite resolution, it can be shown that the statistical mechanical equilibrium becomes a steady Arnol'd stable state, identical to the minimum enstrophy state. That is, the eddy energy vanishes at all finite wavenumbers. The proof is to be found in Carnevale and Frederikson (1987).

Thus, there is a close connection between the maximum entropy and minimum enstrophy theories. To see the underlying physical connection, ask the question 'why is enstrophy dissipated faster than energy?' An answer may be found in statistical mechanics. Consider an inviscid spectrally truncated flow with no topography, with energy and enstrophy localised around some wavenumber, and suppose that the turbulence begins. As the system increases its entropy, it evolves toward a distribution of Figure 1; energy will be confined to the large wavenumbers, whereas enstrophy is moved to higher wavenumber. Now imagine increasing the cut-off wavenumber. The energy remains trapped, whereas the enstrophy moves to higher and higher wavenumber, and is essentially flushed from the large scales. Indeed, at infinite resolution, all the enstrophy is at large wavenumber, and all the energy is confined to the small wavenumber (Kraichnan 1975). Thus, at finite

wavenumbers a 'minimum enstrophy state' emerges. The generalization to the topographic state is straightforward, and at infinitely high resolution maximum entropy and minimum enstrophy are identical. There is no eddy (time-varying) flow, just steady flow locked to the topography by (2.11b). The flow at finite wavenumber is steady and nonlinearly stable

Higher Order Invariants

As mentioned, the continuous equations conserve an infinity (an uncountable one!) of integral invariants. Canonical equilibrium theory based on the conservation of energy and enstrophy has but one mean state, $\mu < \psi >=< q >$. A general stationary state satisfies

$$\psi = F'(q) \tag{4.5}$$

where F is an arbitrary differentiable function. If $F''(q)$ is strictly positive, i.e. $0 \leq c \leq F''(q) \leq C < \infty$ then nonlinear stability ensues (Arnol'd 1966). Suppose that F is chosen to satisfy the stability criterion, and consider a system close to (4.5). Then, the system cannot deviate too far from its initial state. If F is chosen to be a nonlinear function, the system certainly cannot be expected to produce time average statistics which satisfy a linear $q-\psi$ relationship. By the time-reversibility of the dynamics, a system which begins its evolution in some other state far from (4.5) can never approach that state too closely. Some regions of phase space are forbidden to it, and ergodicity on the energy-enstrophy surface will again not arise. Shepherd (1987) has explicitly demonstrated that beta plane dynamics are not ergodic on the energy-enstrophy surface: if beta is sufficiently strong and the system is initially in a sufficiently anisotropic state, then the system will remain anisotropic, because the higher order invariants prevent the system from ever becoming isotropic. These results, however, are not criticisms of the statistical mechanical method *per se*; it is simply that we have not incoporated all the known constraints. Since potential vorticity is conserved on parcels, arbitrary integral functions of potential vorticty are invariant. Thus, with the invariant

$$H = E + G(q) \tag{4.6}$$

where E is the energy and the the Lagrange multiplier is aborbed into the definition of the arbitrary function G, the appropriate Gibbs distribution is

$$P \propto e^{-\alpha H} \tag{4.7}$$

where α is positive for normalizability. Then, in a manner quite analogous to that which produced (2.11) we obtain the flow,

$$< \psi >=< G'(q) > \tag{4.8}$$

Thus, we can in fact regain arbitrary stationary flows from the statistical mechanics, although since G can be almost any function, it might appear that the statistical mechanics is really a quite unhelpful predictive tool. Furthermore, note that (4.8) differs from $< \psi >=< G(q) >$ unless G' is a linear function, so that the equilibrium state cannot be directly calculated unless it is steady.

The flow produced by (4.5) will not, however, be necessarily stable in a truncated finite difference or spectral numerical model, unless F' is a linear function. This is because the stability for such a flow requires the the integral of F to be conserved, which does not in general hold for a truncated model. The role of such higher order constraints is rather unclear at the moment, especially in the light of interesting results by Robert and Sommeria (1991) which purport to explain the prevalence of coherent structures in two-dimensional turbulence via the use of a statistical mechanical theory that formally maintains all the invariants, and the work by Miller (1990). It does appear, though, that the stability of coherent structures (e.g. modons) may rely on higher order invariants, not captured by a theory which only preserves the quadratic invariants.

5. NON-EQUILIBRIUM FLOWS AND JETS

Non-equilibrium simulations

Although strictly inviscid flows have been observed to evolve into statistical equilibrium, the presence of viscosity can nevertheless have large effects in preventing the realisation of statistical equilibrium. Wang and Vallis (1993) considered the effects of viscosity in modifying Fofonoff flows (see also Griffa and Salmon 1989; Cummins 1992). They found that the additional boundary conditions that a viscous solution must satisfy are responsible for producing time-averaged states different from Fofonoff flows, with $q-\psi$ relationships which showed strong deviations from linearity. For example, with free-slip boundary conditions, the potential vorticity is constrained to the boundary value βy, and the $q-\psi$ scatter plots show considerable deviations from linearity in the neighborhood of the boundary. The interior flow is more free to evolve into a free state (really a minimum enstrophy rather than maximum entropy state), although this too is prevented from complete realization by potential vorticity homogenization in closed gyres. With so-called 'super-slip' boundary conditions, in which the normal derivative of vorticity is set to zero at a boundary, the boundary layer effects are reduced, although homogenization still occurs.

In the rest of this section, we would like to discuss another disequilibrium phenomena, the production of jets in the presence of a large-scale potential vorticity gradient, a phenomenon not captured by the equlibrium theories. The presence of a beta-effect (apart from the rather special homogeneous geometry) does in fact produce an anisotropic mean

flow which the equilibrium theories can capture. In a channel geometry the equilibrium solution will be a zonal flow, given by the solution of (2.11b) with periodic boundary conditions in the x-direction. Taking $\psi = 0$ at $y = -1$ and $y = 1$ the solution is

$$\psi(x,y) = \frac{\beta}{\mu}\left\{ y - \frac{e^{\sqrt{\mu}y} - e^{-\sqrt{\mu}y}}{e^{\sqrt{\mu}} - e^{-\sqrt{\mu}}} \right\} - \frac{\lambda}{\mu}\left\{ 1 - \frac{e^{\sqrt{\mu}y} + e^{-\sqrt{\mu}y}}{e^{\sqrt{\mu}} + e^{-\sqrt{\mu}}} \right\}. \tag{5.0}$$

For positive μ the scale of this purely zonal flow is the scale of the channel (Fig. 5). It is the 'Fofonoff channel flow' It *is* anisotropic. (The fact that the equilibrium doubly-periodic beta-plane flow is isotropic is a slightly artificial result, consequent on the homogeneous geometry.) If the flow is required to be symmetric across the channel then $\lambda = 0$. Then, the mean flow is only non-zero if $\beta \neq 0$. If $\mu < 0$ then the flow may produce jet-like features. However, these correspond to a maximum enstrophy state and are not necessarily stable.

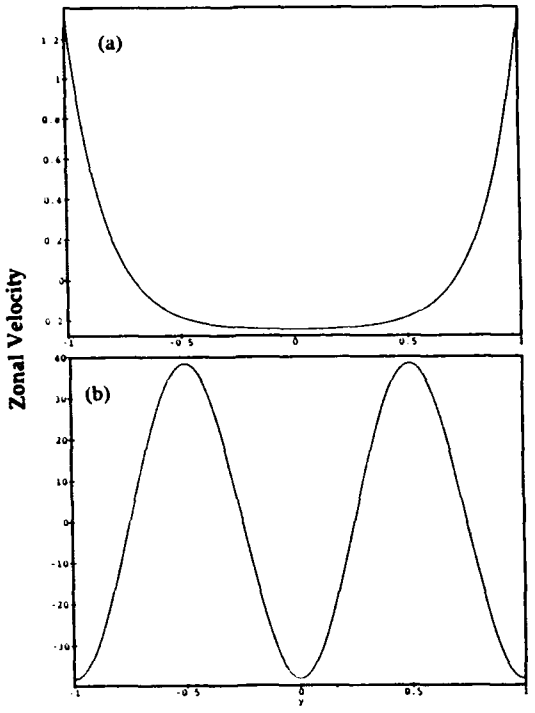


Figure 5. Fofonoff flow in a channel. Shown is the zonal flow, namely $u = -\partial\psi / \partial y$, where ψ is given by (5.0) with λ and: (a) μ =40, β =10, and (b) μ =-40, β =10.

In a forced-dissipative turbulent flow, another mechanism comes into play which the equilibrium theory does not capture, and jets may be produced (Vallis and Maltrud 1993). Briefly, the mechanism is as follows. The frequency associated with a Rossby wave is $\beta k_x / k^2$, whereas the 'frequency' associated with turbulent motion is more like Uk, where U is the rms velocity of the flow. (Vallis and Maltrud discuss other possibilities for the 'turbulent frequency,' and show that other choices, for example ζ^{-1} where ζ is the mean vorticity, make little difference to the following argument.) If the Rossby wave frequency is much higher than the 'turbulent frequency,' then wave-like motion dominates over turbulent motion. However, it will be very difficult to excite such Rossby waves for that same reason—their natural frequency is much higher than that of the forcing turbulent motion. Now, the 'turbulent frequency' is, to lowest order, isotropic. However, the Rossby wave freqency is most decidedly not. Figure 6

shows the wave-turbulence boundary: within the the dumb-bell shape the Rossby wave frequency is higher than the turbulent frequency, and energy transfer into this region is inhibited. Energy cascading to larger scales 'avoids' the modes within the wave region. The cascade to large scales is then most efficiently achieved by the excitation of *zonal* flow. The isotropic cross-over scale between waves and turbulence is given by

$$k_\beta = \sqrt{\frac{\beta}{U}} \tag{5.1}$$

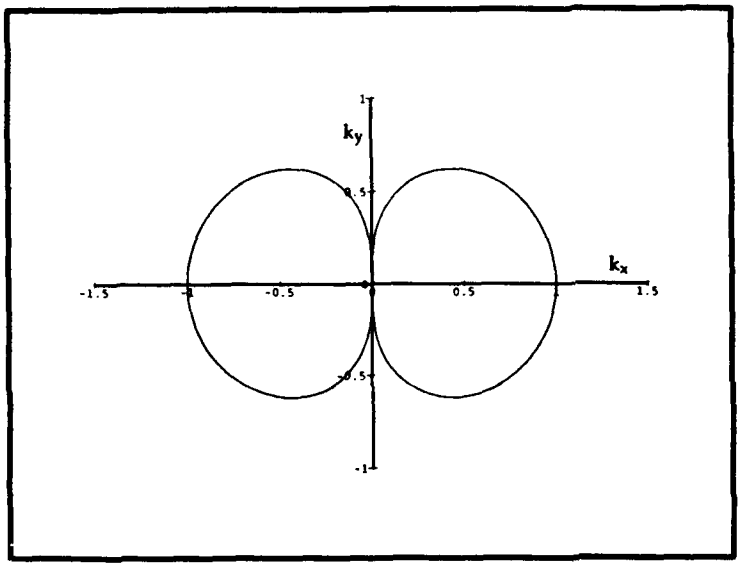


Figure 6. Wave-turbulence boundary in k-space. Plotted is the locus of points whose Rossby wave frequency, $\beta k_x / k^2$ equals a 'turbulent' frequency Uk. Within the 'dumb-bell' the frequency of Rossby waves exceeds that of the inverse turbulence timescale.

if the simple expression Uk is used for the turbulent frequency. (Other expressions are found if the turbulent frequency is parameterized differently.) The scale of the zonal flow will not quantitatively be found at this scale, because as seen in Figure 6 Rossby waves give no restriction on the scale of the zonal motion, because for the zonal flow the Rossby wave frequency vanishes. However, we should expect the scale of the zonal jets to qualitatively have the scale k_β, since the cascade to larger scales will be very inefficient once the energy has become largely zonal.

This robust mechanism seems responsible for the production of zonal flow in forced-dissipative beta-plane simulations. (A related but slightly different mechanism was first proposed by Rhines 1975.) However, although it does not rely on dissipation to work, it is not a feature of the inviscid statistical mechanical simulations, because it is a *transient* effect. Although it is more 'difficult' to initially excite modes in the wave regime, in time energy will nevertheless creep into the wave region and remain. However, if energy is being removed at low wavenumbers, by viscosity or Ekman friction, then a constant state of anisotropy can be maintained. Thus, in forced dissipative flows, jet-like zonal structures

timescale on which a statistical mechanical equilibrium can be maintained. The equilibrium state of forced-dissipative beta-plane turbulence is zonal flow, whereas the equilibrium state of inviscid beta-plane turbulence, in a homogeneous domain, is isotropic flow.

Flow over topography provides a very similar example of where dissipative flow can be different from the inviscid equilibrium, or from the minimum enstrophy state. Again two mechanisms are involved, only the first of which the equilibrium theory is able to capture. This is the mechanism which first generates a mean flow over the topography; it can be interpreted as one of vortex segregation. Consider, say, a single hump (a 'sea-mount') in an eddy field. Fluid moved up the hump conserves its potential vorticity, so its relative vorticity falls. Similarly, fluid moving off the hill into a valley increases its relative vorticity. The upshot is a negative correlation (for positive f) between topography and vorticity (or in general anti-cyclones over humps), and the generation of a mean flow. (This same mechanism is responsible for producing the zonal flow on a beta-channel.) Both maximum entropy and minimum enstrophy quantify this phenomena: although neither theory is aware of the conservation of potential vorticity on parcels, utilization of the quadratic invariants gives rise to a linear relationship between streamfunction and potential vorticity, as in (2.11b). Inviscid simulations (Wang and Vallis 1993) indeed show that a maximum entropy state is realized. However, the addition of viscosity can have an important effect. For topography comprising a single ridge, the simple prediction is of flow parallel to the ridge in a pseudo-westward fashion (that is, facing downstream; higher values of potential vorticity are to the right). In a forced-dissipative situation this prediction is not, however, always realized. For small values of the topography, flow similar to that

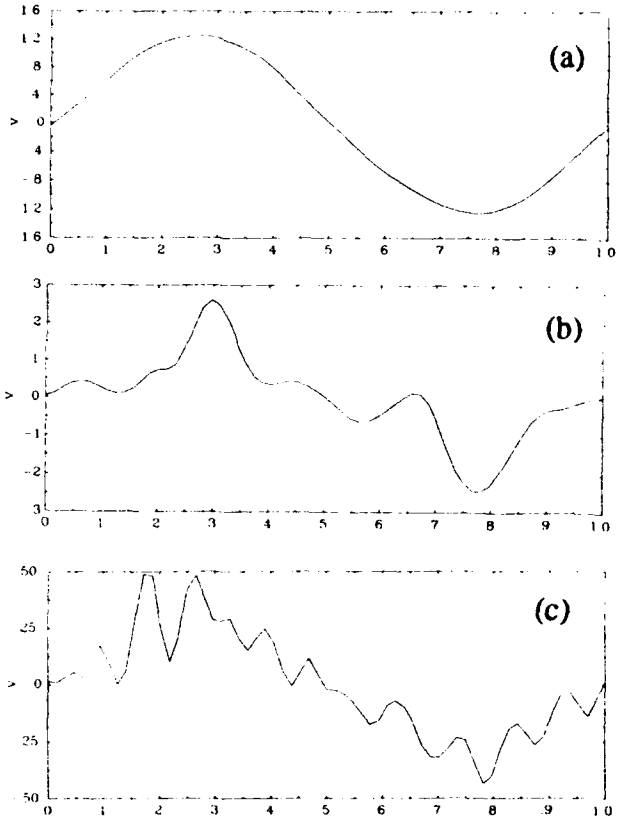


Figure 7. The time averaged velocity along a simple meridional ridge on the f-plane for various topographic heights in a barotropic forced-dissipative simulation, forced near wavenumber 12. The velocity is plotted as a function of cross-slope co-ordinate. The topography is peaked at the center line, with no long-slope variation. (a) The amplitude of the topography is $h = 20$; the flow has approximately the same scale as the topography. (b) $h = 200$; jets begin to appear, superimposed on the broad background flow. (c) $h = 1000$; stronger jets appear.

prediction is realised. However, for larger topographic heights, jets form parallel to the topographic contours (Fig. 7). Essentially, as far as the flow is concerned, the topography seems like a beta-effect, and for the same reason that jets form on a beta-plane, topographic jets form parallel to iso-lines of topography. The condition for the appearance of jets is that the jet scale be smaller than the cross topography scale, where the jet scale is given by (5.1), but with variation in topographic height replacing the beta effect as the cause of the large scale potential vorticity gradient.

7. DISCUSSION AND OCEANIC RELEVANCE

Statistical mechanical ideas have been applied in two general areas in ocean dynamics—to the gyre scale quasi-horizontal circulation, and to flow over topography. The question is whether the real ocean circulation should pay attention to any of the ideas we have discussed herein. Clearly, the ocean is neither unforced nor inviscid, and we cannot expect it to quantitatively reproduce a statistical mechanical equilibrium. For example, the spectrum of eddy kinetic energy is more likely to be a consequence of forced-dissipative geostrophic turbulence than an approach to inviscid equilibrium. If such ideas have any meaning, then, they will be found in the *tendency* of flow toward such equilibrium, constrained by the consequences of forcing and dissipation. This point of view has also been taken by Holloway (1992). Thus, the minumum enstrophy state, which is perhaps a step closer to a forced-dissipative reality, can be seen as a consequence of the nonlinear dynamics trying to evolve into a maximum entropy state, plus the effect of small-scale preferential dissipation of enstrophy.

The minimum enstrophy and maximum entropy states are 'free' solutions of the equations of motion, and so a tendency toward these states is a tendency toward free solutions. The notion of large scale quasi-horizontal circulation as a free solution (a Fofonoff state) is rather opposed to the forced-dissipative Stommel-like models. The reconciliation of these viewpoints is not obvious—or indeed if 'reconciliation' is the correct attitude—for the forced-dissipative viewpoint alone is simple and appealing. Yet to the extent that free nonlinear evolution is possible the system will attempt to evolve toward a free solution. Minimum enstrophy (or maximum entropy) states are *stable solutions* of the equations of motion. Are they attractors? The inviscid equations, being Hamiltonian, have no attractors and it is not correct to call the maximum entropy state an attractor. However, the related minimum enstrophy state is an attracting state, in the absence of forcing. The competitive roles of forcing and free evolution then will together determine the (statistically) steady state ultimately achieved.

The situation where free evolution is likely to be most apparent, and equilibrium solutions actually manifest themselves, is probably in mesoscale phenomena. Here the free evolution of turbulent motion has more rein to determine the mean flow. The production of mean flows around topographic features would be a direct consequence of such free evolution. The quasi-passive free advection of vorticity over topography will lead to a negative

correlation between vorticity and topography, that is anti-cyclonic motion over hills and cyclones over valleys. The mean flow is pseudo-westward; that is, facing downstream, higher values of potential vorticity are on the right. Over ridges or continental slopes this results in polewards (equatorwards) mean flows on the western (eastern) sides of meridional ridges. One may conjecture that this is the cause of the almost ubiquitous polewards undercurrents in eastern boundary currents. Since such mean flows are generated by the interaction of topography and mesoscale eddies, the strength of the mean flow will directly depend on the strength of the eddy field. The eddy field must exist independently of the topography. It may be a result of baroclinic instability of a large scale flow giving rise to a sea of mesoscale eddies. However, if the eddies are themselves produced by an instability involving a large-scale mean flow and the topography, the phase relationships between the eddies and the topography may be quite different, and the eddies will not be passively advected over the topography.

If the topography is sufficiently steep then a second effect becomes noticeable, namely the concentration of the mean flow into narrow currents, via the topographic beta-effect—just as the more familiar beta-effect due to differential rotation produces zonal jets. The criterion to see such an effect is that the topography is sufficiently steep and sufficiently broad that the width of the topographic jets is narrower than the cross-slope scale of the topography. Possible locations for such phenomena are on continental slopes and mid-ocean ridges, although the criterion for multiple jets may never be actually satisfied in the ocean. Multiple jets do of course exist in Jupiter's atmosphere, and it has sometimes been suggested that the earth's atmosphere verges on having two jet-streams, rather than one.

Observational testing of these ideas is possible. One such test would be to demonstrate the unambiguous existence of mean currents flowing more-or-less parallel to the topography on mid-ocean ridges or around mid-ocean sea-mounts. The production of mean flows along continental borderlands is also predicted by the theory, and here the sense of the mean flow is to produce polewards flowing undercurrents along eastern edge of ocean basins and equatorward flowing currents along the western edge. These are *counter* to the mean surface flow of the large scale gyre structure and may be the cause of the ubiquitous counter currents, especially the polewards counter currents seen on the eastern edge of a number of ocean basins (Neshyba *et al.* 1989). However, there are other theories for that phenomena which do not rely on the topography but on the wind-stress and ageostrophic phenomena (MacCreary 1981), and the situation is not definitively resolved. If currents can be observed around seamounts where there is little systematic wind-forcing then it would be hard to avoid a theory involving eddy-topographic interactions, such as those described here. A prediction of the theory is that the strength of the mean flow is correlated with the strength of the eddy field, and this may be amenable to direct verification. Finally, a practically useful aspect of statistical mechanical concepts may lie in their use in subgrid-scale representation, and the interested reader is referred to the chapter by Holloway in this volume.

Acknowledgments. This work was supported by the ONR (N00014-90-J-1618), and part of it was presented at the 'Aha Huliko'a workshop at the University of Hawaii in 1993. Glenn Ierley suggested the term 'quadratically inviscid.' I would like to thank Jian Wang for many helpful discussions, and for some of the simulations.

REFERENCES

Arnol'd, V.I. 1966. On an a priori estimate in the theory of hydrodynamic stability. *Izv. Vyss. Uchebn. Zaved. Mat.* 54(5), 3–5; Eng. translation: *Am. Math. Soc. Transl. Ser.* 2 79, 267–269 (1969).

Buck, B. and Macaulay, V. 1991. *Maximum Entropy in Action.* Clarendon Press, Oxford.

Bretherton, F.P. and D.B. Haidvogel. 1976. Two dimensional turbulence above topography. *J. Fluid Mech.*, 78, 129–154.

Carnevale, G.F., 1982. Statistical features of the evolution of two-dimensional turbulence. *J. Fluid. Mech.*, 122, 143–153.

Carnevale, G.F. and J.D. Frederiksen. 1987. Nonlinear stability and statistical mechanics of flow over topography. *J. Fluid Mech.*, 175, 157–181.

Carnevale, G.F. and Vallis, G.K. 1984. Applications of entropy to predictability theory. In *Predictability of Fluid Motion,* AIP Conference Proceedings 106. G. Holloway and B. West (eds.), 577–592.

Cummins, P. F. 1992. Inertial gyres in decaying and forced geostrophic turbulence. *J. Mar. Res.*, 50, 545–566

Fofonoff, N.P. 1954. Steady flow in a frictionless homogeneous ocean. *J. Mar. Res.*, 13, 254–262.

Fofonoff, N.P. 1962. Dynamics of ocean currents. In *The Sea, 1: Physical Oceanography* (ed. M.N. Hill). Interscience, New York, 323–395.

Griffa, A. and Salmon, R. 1989. Wind-driven ocean circulation and equilibrium statistical mechanics. *J. Mar. Res.*, 49, 53–73.

Gull, S.F. 1991. Some misconceptions about entropy. In *Maximum Entropy in Action,* B. Buck and V. Macaulay (eds.). Clarendon Press, 171–186.

Holloway, G. 1986. Eddies, waves, circulation and mixing: statistical geofluid mechanics. *Ann. Rev. Fluid Mech.*, 18, 91–147.

Holloway, G. 1992. Representing topographic stress for large scale ocean models. *J Phys Oceanog.*, 22, 1033–1046.

Jaynes, E. 1979. Where do we stand on maximum entropy? In *The Maximum Entropy Formalism*, R. D. Levine and M. Tribus (eds.) MIT Press, Cambridge, MA. pp. 15–118.

Kraichnan, R.H. 1975. Statistical dynamics of two-dimensional flow. *J. Fluid Mech.*, 67, 155–175.

McCreary, J.P. 1981. A linear stratified ocean model of the coastal undercurrent. *Phil. Trans. Roy. Soc. Lond.*, 302, 385–413.

Miller, J. 1990. Statistical mechanics of Euler equation in two dimensions. *Phys. Rev. Lett.*, 22, 2137–2140.

Montgomery, D. and Phillips, L. 1990. Minimum dissipation and maximum entropy. In *Maximum Entropy and Bayesian Methods*. P. F. Fougere (ed.), Kluwer Academic Publishers, 281–296.

Neshyba, S., Mooers, C.N.K., Smith, R.L. and Barber, R. 1989. (eds). *Polewards Flows Along Eastern Ocean Boundaries*. Springer-Verlag. 374 pp.

Rhines, P.B. 1975. Waves and turbulence on a beta-plane. *J. Fluid Mech.*, 69, 417–443.

Robert, R. and J. Sommeria. 1991. Statistical equilibrium states for two-dimensional flows. *J. Fluid Mech.*, 229, 291–310.

Salmon, R., G. Holloway and M. C. Hendershott. 1976. The equilibrium statistical mechanics of simple quasigeostrophic models. *J. Fluid Mech.*, 75, 691–703.

Shepherd, T.G. 1987: Non-ergodicity of inviscid two-dimensional flow on a beta-plane and on the surface of a rotating sphere. *J. Fluid Mech.*, 184, 289–02.

Vallis, G.K. 1985: Remarks on the predictability properties of two- and three-dimensional flow. *Q. J. Roy. Meteor. Soc.* 111, 1039–1049.

Vallis, G.K. 1992: Problems and phenomenology in two-dimensional turbulence. In *Nonlinear Phenomena in Ocenic and Atmospheric Science*. Eds., G. Carnevale and R. Pierrehumbert. Springer-Verlag. pp. 1–125.

Vallis, G.K. and Maltrud, M. 1993: Generation of mean flows and jets on a beta-plane and over topography. *J. Phys. Oceanog.* 23, 1346–1362.

Wang, J. and Vallis, G.K. 1993: Emergence of Fofofonoff flows in inviscid and viscous ocean circulation models. *J. Mar. Res.* (In press.)

OVERVIEW OF STATISTICAL MECHANICS, WITH PRACTICAL APPLICATION FOR OCEAN MODELING

Greg Holloway
Joint Program for Ocean Dynamics, Institute of Ocean Sciences,
Sidney, BC, Canada V8L4B2, and
Centre for Earth and Ocean Research, University of Victoria, Canada V8W2Y2

ABSTRACT

Because oceans are bigger than the computers that model them, most of what goes on in oceans cannot be represented adequately. Ability to observe the ocean is limited. These considerations compel a probabilistic view, both of the "observed" ocean and especially taking account that models really solve for moments of probability of possible states of the ocean. We rethink how models should work, here taking into account statistical mechanical ideas about ocean circulations. There is a difficulty. The problems that can be treated by methods of statistical mechanics are far from the practical problems of ocean modelling. We entertain a hybrid approach—employing conventional ocean modeling to deal with application of large scale forcing while extending model physics to recognize the oceans' internal dynamical tendency toward higher system entropy.

INTRODUCTION

There are many reasons to apply statistical methods in physical oceanography, as seen in the many contributions in this volume. In the present article we focus on a particular aspect. We ask to what extent we may treat *statistics* of flows as *dynamical objects.* The challenge is to determine what are the *equations of motion of statistics* of flows.

This invites us to reconsider what the "fluid dynamic enterprise" is about. In its usual context, fluid dynamics deals with partial differential equations describing fields of momenta, density, and so forth. Given boundary and initial conditions, the goal is this: solve. Often "solve" is too tough, so the strategy may be to obtain simplifying approximations or idealizations, and then solve. For most practical applications, "solve" includes also a finite discretization to some numerical representation of the intended equations of motion.

In ocean modeling is this *really* what we do? I think not. Even in a domain as small as a bay or a harbour, we aren't given the initial conditions and boundary conditions at the fine scales for which actual equations of fluid flow apply. Moreover we likely couldn't "solve" the equations of motion if we did have this information! And the global ocean is so much bigger.

What to do? To a large extent, ocean modeling succeeds by luck and by cheating. We take solutions to idealized problems and compare with our partial information about the real ocean. [Although computer models are often characterized as "realistic," that should be read only as "less idealized" than some other model.] When we compare solutions with reality, we don't truly ask if reality coincides with the solution. In reality, when we measure velocity or temperature or elevation somewhere at some time, we see wiggles, whirls, blibs and so on. In part there is always measurement error. In a greater part though, we appreciate that oceanic flows are nearly always characterised by a lot of wiggles and whirls, over many scales of motion. Thus, when we compare idealized solutions or model output with "reality," we should be obliged to append a phrase "in the mean," desperately hoping no astute reader asks what we mean by "the mean."

Something statistical has got into this "fluid dynamic enterprise." We have compared 'apples with oranges,' testing explicit, fully determined solutions to idealized problems against some sort of statistical measures of reality.

THE PHASE SPACE OF THE OCEAN

To pose the question consistently, we might deal with probability throughout. Let $\mathbf{Y}$ represent the state of the fluid at any instant. In practice, $\mathbf{Y}$ will consist of some finite representation, perhaps the velocities, densities and whatever at many grid points, or perhaps the coefficients from expansion on some set of basis functions. $\mathbf{Y}$ may have a huge number of components, perhaps a million or more if we think of large scale supercomputer representations or we may speak of zillions (any number) of components of $\mathbf{Y}$. Vector $\mathbf{Y}$ is a "point" in the multi-dimensional "phase space" of all possible $\mathbf{Y}$. Determinisitic equations of motion yield a trajectory $d\mathbf{Y}/dt = \mathbf{G}(\mathbf{Y})$. In reality, $\mathbf{Y}(t)$ is fantastically complicated, representing every tiny whirl and wiggle in the ocean. It seems doubtful we could ever have so much information or would ever want it if we could have it.

> Aside: How big is the phase space of the ocean? If we think of continuous fields, then size is power of continuum. However, we recognize that there is some scale below which viscous-diffusive effects smooth the fields. That scale depends upon turbulent intensity, which varies greatly. Moreover, velocity, temperature and salinity will be smoothed at different scales because of their different molecular diffusivities. The result overall is that in more intense regions in the upper ocean, the smooth scale will be significantly less than 1 cm. In a weakly turbulent deeper ocean, the scale may be several cm. To make a kind of "average" for back-of-envelope estimation, say the number is "around" 2 cm. In the ocean there are roughly 1.3×10^{24} cc of water. If we take the standard incompressibility assumption, we will have four dependent variables (two components of velocity, temperature and salinity, say) in each $2 \times 2 \times 2$ cc volume element. Thus the size of the phase space is $1.3 \times 10^{24} \times 4 / 2^3$, or something over 6×10^{23}. What would Avogadro have thought of that?!

Of necessity, as well as by thoughtful intent, we wear "smoky glasses" when looking at the ocean. We do not see a "point" **Y**, we see a "blur," a cloud of possible **Y**. Thus it only makes sense to speak of the ocean in terms of probability $p(\mathbf{Y})dY$ that the actual state of the ocean lies within phase volume dY of some **Y**. We then pose the ocean problem by saying that initially we have some $p(\mathbf{Y};t=0)=p_0(\mathbf{Y})$, some probabilistic statement of boundary and forcing conditions, and we wish to solve for $p(\mathbf{Y};t)$ at future t.

It seems we've made the problem worse. Before we had too many **Y**. Now, for every **Y**, we also want a continuous function $p(\mathbf{Y})$. Moreover, we had at least an equation of motion for **Y**; what is the equation of motion for $p(\mathbf{Y})$? Happily, things start to get better. Some of what we really want are only moments of $p(\mathbf{Y})$, starting with first and second moments: $<\mathbf{Y}>=\int \mathbf{Y}\, p(\mathbf{Y})\, dY$ and $<\mathbf{YY}>=\int \mathbf{YY}\, p(\mathbf{Y})\, dY$. These include things like the "average" ("expected") current, temperature or salinity, or average heat transport or eddy energy, for example. As well, when we appreciate that we are only interested in moments of **Y**, we *choose* not to examine **Y** in all its 10^{23} phase space detail; 10^6 or 10^3 or fewer numbers might be all we care about. Although this discussion may provide a viewpoint, actual value rests on displaying an explicit means of calculation. How do we obtain useful <**Y**>, say? What are the equations of motion of <**Y**>?

> Aside: A topic often mentioned at this 'Aha Huliko'a, and elsewhere, is chaos. It may be argued that chaotic behavior in low order deterministic systems reveals a kind of dynamics for which we thought ideas of probability were needed. Is the ocean chaotic? If the question asks if nearby trajectories Y(t) diverge exponentially, the answer is surely yes. If the question asks whether there exists a lower-dimension attracting object in the phase space, again the answer is surely yes. If the *practical* question is reducing dimension from 10^{23} to a mere 10^{17}, say, then there is little utility in finding such an attractor (*if* we can find such an attractor). A point is that even if we could deal with deterministic dynamics, we might *wish* to introduce $p(\mathbf{Y})$ as the *object* of investigation, with practical goals to obtain expectations <**Y**> and <**YY**>, say.

OCEAN MODELING

First consider the ocean modelers' cheat. We guess and hope that equations for <**Y**> are a lot like the textbook equations for **Y**. We observe that if the equation for **Y** were linear in **Y**, we could pass <•> over this equation and be done. Easy. Unhappily, the equation is not linear and we are faced with unknown <**YY**> in the equation for <**Y**>. So we replace <**YY**> by <**Y**><**Y**> + <**Y'Y'**> where **Y'**= **Y** – <**Y**>. This is Reynolds averaging, here under <•>. It helps some, but leaves unknown <**Y'Y'**>. Now we complete the cheat by copying someone else's cheat. (When cheating it is ever-so-helpful to copy others' cheats. If called out, you can appeal to the list of all the cheaters who have gone before.) The standard

cheat is to characterize flux components of $\langle \mathbf{Y}'\mathbf{Y}' \rangle$ by a Fickian relation to spatial gradients of $\langle \mathbf{Y} \rangle$. It's "eddy viscosity." The lovely thing about this cheat is that the equation of motion for $\mathbf{Y}$ already has a term like that, ascribed to diffusion by molecular chaos. Thus the equation for $\langle \mathbf{Y} \rangle$ really is just the equation for $\mathbf{Y}$ if only we fudge certain coefficients.

Does the cheat work? It is believed to work (sort of) in a variety of turbulent flows, including many industrial applications. That's encouraging (maybe). And ocean models work (sort of) ... don't they? The answer depends on what we mean by "work." Gross features of upper ocean circulation may be more-or-less directly forced by wind or buoyancy. Integral measures such as Ekman transport, Sverdrup relation and volume conservation already constrain gross $\langle \mathbf{Y} \rangle$. So long as models respect these relations, they will work (sort of). Examined more closely, problems appear. Even the surface circulation, which most feels direct forcing, can be problematic in, e.g., western boundary current separation regions. Moreover, problems seen near surface get worse as one looks deeper in the water column. Flows that run poleward along eastern boundaries may run the wrong way in models; undercurrents along western boundaries may be absent or weak. Perhaps the standard cheat doesn't really work so well.

What to do? There is a standard fix for the standard cheat: Get a bigger computer. At finer resolution, less of $\mathbf{Y}$ is left in $\mathbf{Y}'$, so $\langle \mathbf{Y}'\mathbf{Y}' \rangle$ is smaller and the cheat can be made smaller. Modern supercomputers may advance a state vector of length 10^7. We need only await 10^{16}-fold increase in computing power (speed+memory) and the problem is solved. (Feasibly one hopes that a "mere" 100-fold increase might substantially improve the mesoscale eddy problem, as one part of the bigger problem.)

STATISTICAL MECHANICS: Equilibrium and Disequilibrium

If the only available method is increased computing power, and if answers about ocean circulation are needed sorely enough, then the necessary computing resource will have to be created and dedicated to this purpose. When (if) that could happen, at what cost, I can't guess. The question we ask here is to what extent theory of statistical mechanics can provide a practical complement to 'brute force' computing.

Statistical mechanics comes in two flavours: equilibrium and disequilibrium. Equilibrium statistical mechanics addresses isolated dynamical systems, seeking the $p(\mathbf{Y})$ "in equilibrium" (i.e., if the system has been isolated forever). Although this is clearly an idealization, it is the basis for understanding much of classical thermodynamics. The more difficult problems arise in disequilibrium statistical mechanics, including circumstances of open systems where energy or information passes through a system, or where conditions change rapidly in time. Applied to macroscopic fluid flows, disequilibrium statistical mechanics is better known as turbulence closure theory. One might, for example, imagine

turbulence theory helping with the specification of eddy viscosities. In fact we'll see (shortly) a much bolder result, one that suggests the *equations of motion* used by *ocean models* are *wrong*—and not only by uncertain coefficients.

Both the equilibrium and disequilibrium flavours have been exercised with respect to geophysical flows. It is beyond the scope of this article to review, or even make mention of, those exercises. Reviews can be found in Holloway (1986) or Lesieur (1990). A recent review of turbulence theories is in McComb (1990).

Here let us recall two of the simplest examples, one from equilibrium and one from disequilibrium. The examples are chosen because they feed directly into the practical application which follows.

Equilibrium

First consider the idealized case of barotropic vorticity advection on an f-plane, with rigid lid and a bottom of variable depth $H(x)$. Without forcing or dissipation, the equation of motion for $\mathbf{Y}$ is

$$D/Dt(\zeta + h) = 0 \tag{1}$$

where D/Dt is the material derivative $(\partial/\partial t + \mathbf{u}\cdot\nabla)$, ζ is the vertical component of vorticity $\nabla\times\mathbf{u}$, and $h = f(H_0 - H)H_0$ is a potential vorticity due to variation of H about reference depth H_0. Variation of H is presumed small so $|h| << f$. Suppose the initial conditions are random eddies without mean flow, hence $<\zeta> = 0$. We seek $<\zeta>$ at later $t > 0$. For a problem this simple, we can directly solve (1) for a number of realizations of ζ, then average to get $<\zeta>$. What do we guess may be the outcome? If $<\zeta> = 0$ at $t = 0$ then does $<\zeta> = 0$ for all time?

The numerical experimental result for mean velocity $<\mathbf{u}>$ is shown in Figure 1a. Here $h(x)$ has been chosen to resemble the Arctic Ocean. That's just for fun, appreciating the extraordinary idealization in (1). Clearly the answer is not $<\zeta> = 0 = <\mathbf{u}>$. In fact the answer is only a subset of the more general answer given by Salmon et al. (1976). Expressed in terms of streamfunction ψ(from $\nabla^2\psi = \zeta$),

$$(\alpha_1/\alpha_2 - \nabla^2) < \psi >= h. \tag{2}$$

The derivation of (2) is based on the observation that (1) conserves two integrals over the flow field: energy $E = \frac{1}{2}\int dA|\nabla\psi|^2$ and total enstrophy $Q = \int dA\left(\nabla^2\psi + h\right)^2$. Phase vector $\mathbf{Y}$ may be the collection of values of ψ at grid points or the coefficients of ψ expanded on

some basis set. A great many **Y** are consistent with most possible E and Q. Salmon et al. make the ergodic hypothesis that **Y** may equally likely visit all possible configurations consistent with E and Q. Salmon et al. show that the result asymptotes to (2) where α_1 and α_2 are Lagrange multipliers to enforce the constraints to E and Q. A simple route to the result, maximizing entropy $S = \int dY\, p(\mathbf{Y}) \log p(\mathbf{Y})$ subject to E and Q is in an appendix to Holloway (1992, hereafter H92).

An immediate consequence of (2) is that the answer is *not* $<\mathbf{u}>=0$. Observe that this is *qualitatively contrary* to any manner of *eddy viscosity* that ultimately seeks to drag mean flow toward a state of rest. We observe also that in (2), only the ratio α_1/α_2 appears, which may be expressed in terms of a length scale $L^2 = \alpha_2/\alpha_1$. If it happens that we are only interested in $<\psi>$ on scales larger than L, we can omit ∇^2 in (2) and write approximately the simplest ever "theory" of ocean circulation: $<\psi> = L^2 h$. No wind, no sun, no rain, no moon, no ice, no whales.

Can it be so? Figure 1b shows $<\mathbf{u}>$ obtained from our simplest ever $<\psi> = L^2 h$. Although the two panels look similar, they are not identical if overlaid. Figure 1a was produced by averaging eleven realizations after a few thousand timesteps each, at cost of about 80 hours CPU on an Alliant FX40. (I meant to collect an even dozen realizations but ran out of time.) My hunch is that after another 80 hours the average of direct realizations would have got closer to Figure 1b. The simplest ever calculation of $<\mathbf{u}>$ used about 1 second on a Macintosh (wall clock). [A note: Overall velocity amplitude is not shown in Figure 1.

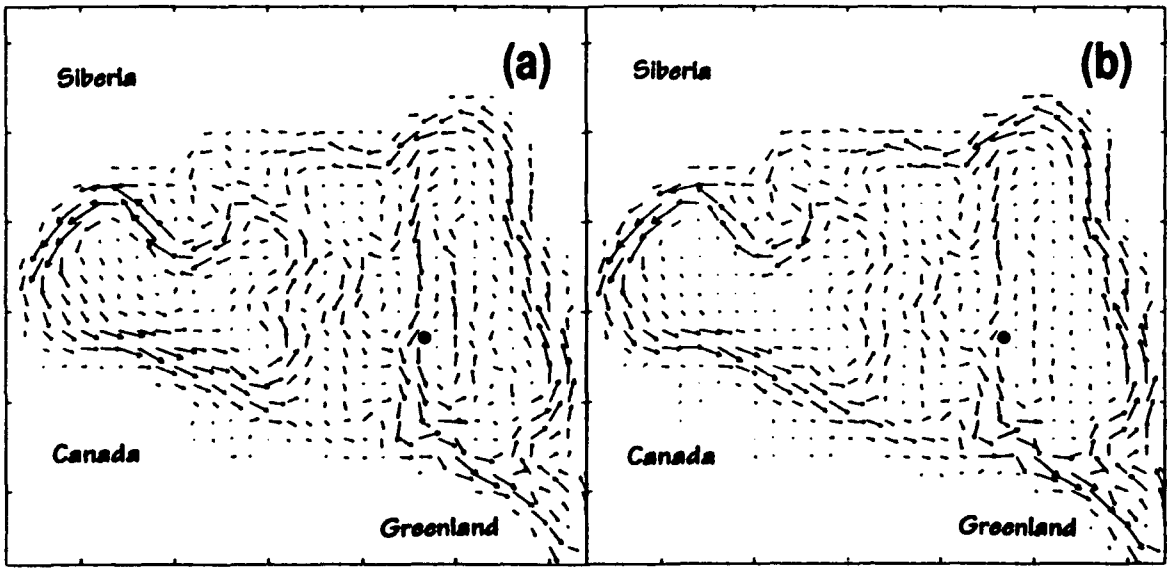


Figure 1. (a) An ensemble average at statistical stationarity over eleven realizations at 256 × 256 resolution of solutions of (1) from initial conditions consisting of random eddies. The ensemble average flow is presented at 32 × 32 resolution. (b) The approximate theoretical flow, given by $<\psi> = L^2 h$.

This can be scaled, so Figures 1a and 1b can be made to fit in terms of overall amplitude.] It may seem a bit uncanny that Figure 1 has a lot in common with the synthesis of Arctic circulation developed by Aagaard (1989). Happenchance, doubtless.

> Aside: Even without addressing really difficult matters such as forcing and dissipation, the equilibrium statistical mechanics of (1) is not as straightforward as I've made it seem. The concern is that we've sought the maximum entropy solution subject only to constraints to E and Q. In fact, within an enclosed basin or for other special boundary conditions such as channel flow, the continuum solution to (1) preserves $\int dA\, g(\nabla^2\psi + h))$ where g is any function. The solution is enormously more constrained than we've taken into account. So why does simple constraint by E and Q seem to "work"? Discussions on this point have gone on for years, and can't be dealt with in this limited space. There are just a few comments: (a) The discrete numerical representation of (1) is not faithful to the continuum invariants of (1), and we really test statistical mechanics against the numerical (1). (b) It may be that even if a hierarchy of invariants does exist, the invariants don't constrain the available phase space in ways that very much affect low order moments such as <Y>. (c) One invariant which numerical schemes may respect is circulation $C = \int dA\, \nabla^2\psi$. In fact this can be taken into account, adding another Lagrange multiplier α_3 on the right side of (2), while stipulating also a value $\psi = \psi_b$ on a closed domain boundary. The result is to modify a boundary current of width L.

In truth, simplest ever $<\psi> = L^2 h$ can't really serve as a theory of ocean circulation—in part because of what model (1) leaves out. Like sun and wind and seagulls. The difficulty is that equilibrium statistical mechanics applies to isolated (closed) systems, whereas the ocean is subject to external forcing and internal dissipation. The more difficult task to include forcing and dissipation falls under the category of disequilibrium statistical mechanics.

Disequilibrium

Consider the simplest extension of our simple model (1), including some dissipation operator such as $-a\zeta + b\nabla^2\zeta$. To achieve statistical stationarity, the flow might be excited Wunder some probability distribution of random torques. These are the open connections which prevent application of simpler equilibrium methods. For disequilibrium calculation one often permits "slow" time dependence of low order statistics, "slow" in the sense that higher moments remain in quasi-steady adjustment to low order moments. In fluid dynamics context, this leads to the turbulence "closure" problem.

Let us maintain simplicity by assuming in (1) that topography h is entirely random. Among realizations, the topography is chosen independently without any mean $\langle h \rangle$. This permits us to consider the case of no mean $\langle \zeta \rangle$, simplifying ensuing moment equations. Assuming that second moments of h, say a wavenumber variance spectrum $H(k)$, is given, the first nontrivial moments we come to are $\langle \zeta\zeta \rangle$ and $\langle \zeta h \rangle$, perhaps characterized by wavenumber spectra of vorticity variance and of vorticity-topography correlation. The straightforward approach is to multiply (1) by ζ or by h, and average $\langle \cdot \rangle$. The problem is that this generates new terms $\langle \zeta\zeta\zeta \rangle$. $\langle \zeta\zeta h \rangle$ and $\langle \zeta hh \rangle$. Continuing by building equations for these $\langle \mathbf{YYY} \rangle$ only generates new unknown $\langle \mathbf{YYYY} \rangle$, and so on indefinitely in explosive proliferation. Many efforts have been made to "close" the hierarchy of moment equations, which I'll not begin to recount here. An excellent recent reference is McComb (1990).

The first disequilibrium theoretical results for the case of barotropic vorticity-topography interaction were those of Herring (1977) and Holloway (1978), the latter comparing with direct numerical experiments. Comparisons were encouragingly (surprisingly?) good. So this is the way to go?

Unfortunately it's not so good for several reasons. First, the "theories" are all subject to a certain amount of "tuning". [My personal view is that even recent theories which claim to be "free of phenomenological constants" have actually only found more clever ways to hide the "adjustments."] One may lack confidence in the generalizing power of these theories. A second, and more damaging, shortcoming is that actual calculation from these theories appears only to be feasible in highly idealized geometries corresponding to statistically homogeneous (or nearly so) fields. Often one appeals as well to statistical isotropy (or near thereto). Third, even with these idealizations, calculation of the theoretical results demands nearly as much computing effort as direct simulations. [This is especially disappointing when one has to do the direct simulations anyway—to see if the theory is right.] Fourth, upon attempting to simplify the theoretical results (Holloway, 1987), they remain too unwieldy for practical exercise in ocean models. Finally, the equations of motion for which these results are available may be only idealizations (such as quasigeostrophy) from the equations intended for actual ocean models.

The disequilibrium studies, like the equilibrium studies, may be regarded as esoteric playthings for theoreticians. Yet there are lessons to be learned. First, there is an important connection between the two approaches. It is entropy. In the disequilibrium studies, particularly as seen in turbulence closures, often there is no explicit discussion on entropy. However, when a theory yields any set of second moments $\langle \mathbf{YY} \rangle$, say, one may evaluate the enstropy S subject to those $\langle \mathbf{YY} \rangle$. George Carnevale in his thesis (1979) has carefully considered this, including the case of vorticity-topography interaction. The entropy can be expressed $S = 1/2 \log \det \langle \mathbf{YY} \rangle$, and George shows that a broad class of turbulence closures demonstrate the Second Law: $dS/dt \geq 0$ in the absence of external forcing or

dissipation (Carnevale et al., 1981). Turbulence closure extends the calculation to show "outside" influence (forcing and dissipation, the latter "outside" the **Y** treated by models). As well, authors have considered the insights gleaned from consideration of equilibrium solutions even when one ultimately has forced-dissipative reality in mind. See Frederiksen (1982, 1985, 1986), Carnevale and Frederiksen (1987) or Holloway (1986). Nonetheless, "insight" is a matter of point-of-view, and the present point-of-view is far outside mainstream ocean dynamics.

ALTERNATIVELY, A HYBRID?

If statistical mechanics only provides a point-of-view, whose insights are a matter of taste, what can we do practically about ocean modelling? Get a bigger computer. Sure. Use eddy viscosity, explictly or via numerical diffusion (because our forefathers have always done so). Business as usual.

We may try to do better than that for two reasons. First, it is not clear that business-as-usual is on the path to success. Doubtless bigger computers will help, but we also *know* that eddy viscosities are wrong because we that *know* eddy-topography interactions (for example) *drive* rather than *damp* mean flows. Second, no matter how big the computer, there will be a host of pressing questions we seek to answer, all of which will be compromised if we must expend computing resource to achieve super-high resolution.

Is there an alternative? It's not clear. To proceed beyond the highly idealized statistical mechanical calculations requires bold leaps, perhaps along lines suggested in H92. While bold leaps may be exhilarating, are they scientific and do they practically contribute? [We should remain aware that eddy viscosity is a leap, one that is plain wrong but only sanctioned by past use.] For the present we consider the leaps—and their consequences. Let me recall only briefly discussion from H92 by way of introducing the following (this volume) paper by Eby and Holloway (hereafter EH).

We begin by recognizing that oceans are subject to external forcing and internal dissipation. To the extent that forcing is on relatively large scales (atmospheric synoptic scale up to planetary scale), this does not present a severe problem. Of course data uncertainty is always an issue. Coastal zone forcing may be a largely under-appreciated problem. Internal dissipation is parameterized in some haphazard fashion; but it is beyond the scope of this paper to address that. Thus we arrive at rudimentary ocean modelling: calculating the (parameterized) viscous response to imposed forcing.

What this picture leaves out are tendencies due to the internal, largely "free" dynamics of the richly nonlinear, 10^{23}-mode ocean. On account of forcing and dissipation, ocean models are dragged away from the higher entropy state to which the internal dynamics would otherwise tend. As the ocean model is drawn away, it ought to feel a force tending to restore toward higher entropy. Neither is this just a loose way of talking; the force the

ocean model should be feeling is very much like the tension in a stretched elastic (see the entropy calculus in Kubo (1965, ch. 1, ex. 13)). By omitting this "entropy force," ocean models get the equations of motion wrong.

To include the entropy tendency in ocean models, we need three things:

1. ability to characterize the maximum entropy configuration of the ocean,
2. a measure of difference between the model solution and maximum entropy, and
3. a "spring constant" for the strength of tendency toward higher entropy.

Although we have none of these three rigorously, plausible guesses can be made. Such guessing may annoy more serious-minded colleagues. However, to avoid guessing means falling back onto eddy viscosity—a far *worse guess*. I make the following guesses in part hoping to draw the attention of bright talent that one day will straighten this stuff out.

Unprejudiced circulation

We begin guessing from the barotropic, quasigeostrophic $\langle\psi\rangle = L^2 h$. There are two immediate objections: it's barotropic and quasigeostrophic. The barotropic aspect isn't so bad. If we have in mind large scale ocean modelling in which we do not choose to resolve the first internal radius of deformation, then the theory of Salmon et al. (1975) shows that statistical equilibrium is nearly barotropic on scales larger than the first radius. Of course the actual ocean is not barotropic; but we readily understand that in terms of the baroclinic projection of applied forcing.

Quasigeostrophy is more difficult, in part because the potential vorticity fluctuation $h = f(H_0 - H)H_0$ should involve only small departures of total depth H from a reference depth H_0. We mean to apply these ideas to primitive equation, full depth ocean models. Quasigeostrophy is ambiguous whether ψ refers to velocity streamfunction or to depth-integrated transport streamfunction. H92 considers both, the former leading to $\Phi^* = -fL^2H^2 / H_0$ and the latter to $\Phi^* = -fL^2H$, where Φ^* is introduced to denote transport streamfunction at maximum entropy. These formulae were suggested with constant f in mind, appreciating that spatial scales of variation of H will be small compared with planetary radius. Things will break down if we apply such formulae carelessly, say to a flat-bottomed beta-plane "ocean." For the present, the aim is to proceed most simply with realistic-geometry practical modelling in mind.

Serious-minded colleagues may be apalled by the uncertainty over which formula to use for Φ^*. So be it. In practice, EH find little difference between the two as compared with the larger differences from conventional modelling (the *guess* that $\Phi^* = 0$). Of the two formulae above, the former requires assigning H_0 while the latter present the possibility of velocity singularities as H→0. The latter difficulty seems harmless because, first, models approach H→0 discretely and, second, in shallower water direct forcing and dissipation

tend to over-ride the statistical mechanical tendency. Therefore, with short term expedience in mind, we adopt an "unprejudiced circulation"

$$\Phi^* = -fL^2H \tag{3}$$

where "unprejudiced" refers to the "least biased" (minimum information) aspect of maximum entropy. (A guess $\Phi^* = 0$ is, by comparison, one of extreme prejudice.) We are left to assign L^2. In the ideal case of inviscid equilibrium, $L^2 = \alpha_2/\alpha_1$ is a ratio of Lagrange multipliers determined from E and Q and the number of retained degrees of freedom. In practice, we have left L a disposable fudge factor, presumably reflecting a length scale somewhat shorter than actual eddy length scales. At absolute (ideal) equilibrium, one expects L^2 to take a single value characterizing all of the domain, in just the sense that temperature comes to be uniform for an isolated system. In practice, we understand that the disparate regions of the forced, dissipative ocean are only weakly in "thermal" (statistical mechanical) contact. Hence we expect that L^2 may have weak geographic dependence, tending to follow eddy length scales. A natural suggestion is to tie L to regionally smoothed first deformation radius; we (EH) have not yet explored this. The shorter term expedient is simply to allow that L should be somewhat larger at lower latitudes. In EH and subsequent experiments, we've let L range from a few km in the Arctic to a few tens of km near the equator. It is an uncertainty that ought to be narrowed in future work.

When an actual ocean model is executed, its output will differ from (3). How to measure that difference will depend upon specifics of the model under consideration. As a representative model (without implication concerning its putative strengths or weaknesses), EH have considered the GFDL "Modular Ocean Model," a successor to the model described by Cox (1984). Prognostic variables include velocity, temperature and salinity fields. Because the velocity field is split to external (depth integrated) and internal (baroclinic) parts, with the external part described by a transport streamfunction, it is natural to measure the difference between model streamfunction and Φ^*. Continuing in this simplest way, EH append a term in which the model relaxes toward Φ^* with a given time constant. This is not just a "quick fix." The appended term corresponds to the tension in a stretched elastic as mentioned earlier. It reflects missing physics in conventional model formulation.

Temperature and salinity equations are not affected by the entropy tendency here considered. This is because we only treat scales larger than the first radius of deformation. If one sought to apply these ideas at smaller scales, then "entropy forcing" terms should appear also in the temperature and salinity equations.

It remains to specify a time constant for the streamfunction restoring. With simplicity in mind, EH choose 25 days (while also exploring sensitivity to other choices). These first experiments with a hybrid model are discussed in a following article (EH).

> **Warning: Reader discretion is advised.**
> The following article (Eby and Holloway) is rated **R**. This material has been Rejected for responsible* publication. Not merely Rejected but Rejected with vehemence as "... silly ... sketchy ... at best trivial ... cavalier ... tenuous ... just another formula ..."
> * The Rejection committee notes with concern that 'Aha Huliko'a follows an alarming practice of irresponsibly publishing dangerous ideas.

The work reported by EH is only a very first exploration, trying to see if it is worth the effort to further pursue this line. We do feel encouraged that, first, inclusion of entropy tendency makes a difference and, second, the sense of the difference appears to be toward improving model fidelity. Can we do better?

There will need to be renewed theoretical effort with respect to matters like L^2 and how the "spring constant" varies at different scales of motion. Some aspects for improvement are clear, even if not precisely so. Relaxing streamfunction, as EH, means that the largest scales of motion tend to higher entropy as quickly as smaller scales, whereas both theory and idealized experiments show that smaller scales should adjust more quickly. This suggests filtering the relaxation process by some practically convenient operator such as ∇^2, which is already present in the momentum equation. *Eddy viscosity*! Our earlier complaint about eddy viscosity is not so much with the differential operator as with the aspect that conventional eddy viscosity drags models toward a state of rest (extreme prejudice!) If instead we define from (3) a maximum entropy, barotropic (at scales larger than first radius) flow

$$H\mathbf{u}^* = \mathbf{z} \times \nabla \Phi^* \tag{4}$$

then we can center the eddy viscosity not about the state of rest but about $\mathbf{u}^*$. The eddy viscosity term is given by $A \nabla^2 (\mathbf{u} - \mathbf{u}^*)$, where A might be a simple constant coefficient. Of course one could also try to be more sophisticated about how A may vary. For coarse resolution modelling, temperature and salinity equations remain unaffected. (In fact, horizontal eddy diffusion of density is consistent with the tendency toward barotropic $\mathbf{u}^*$.)

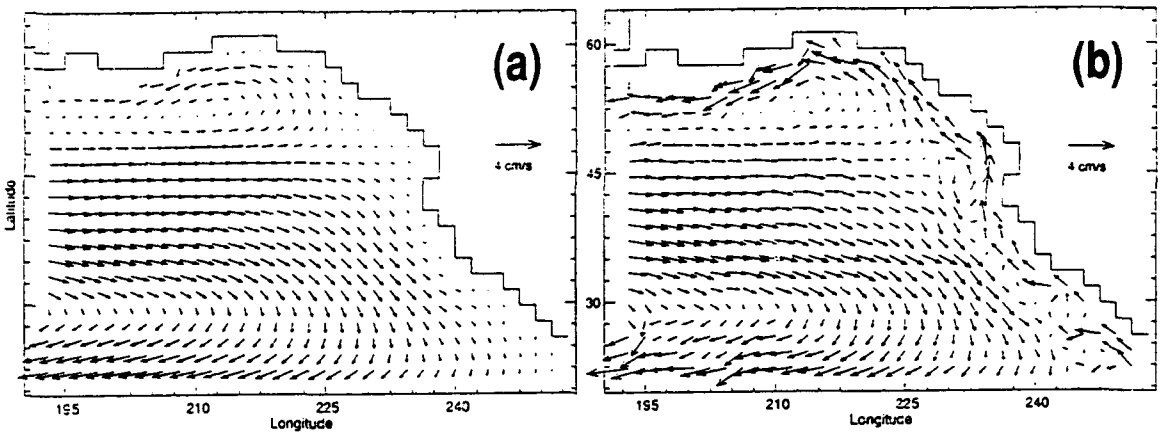


Figure 2. (a) Flow at 245 m after 150 years integration ("control case") under annual mean windstress and surface layer restoring to annual temperature and salinity. (b) Flow at 245 m. when eddy viscosity is modified to $A\nabla^2(\mathbf{u}-\mathbf{u}^*)$, with $A = 2\times10^5$ m^2/s and L (a factor in $\mathbf{u}^*$) increasing from 3 km at the pole to 15 km at the equator.

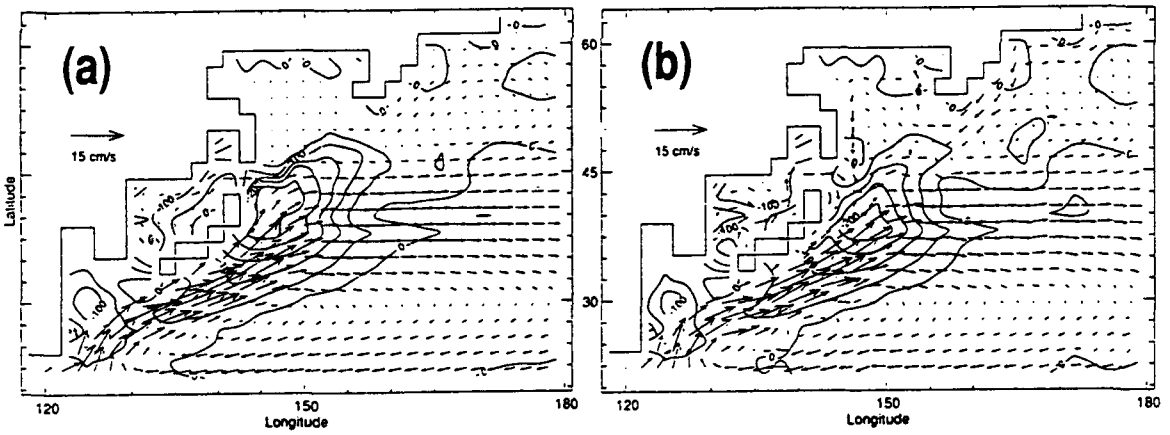


Figure 3. (a) Flow at 245 m. for the "control case" (Fig. 2). Contours show the implied air-sea heat exchange due to surface restoring to annual mean temperature. (b) Flow at 245 m and implied air-sea heat exchange when eddy viscosity is modified (as Fig. 2).

Michael Eby has performed newer experiments replacing conventional eddy viscosity by $A\nabla^2(\mathbf{u}-\mathbf{u}^*)$. While the experiments will be described fully in a subsequent paper (in prep.), Michael has kindly provided me a few figures by way of preview. Results are shown after 150 years under steady forcing by Hellerman-Rosenstein (1983) winds and surface layer relaxation of temperature and salinity toward mean Levitus (1982). Figure 2 shows flow at 245 m in the northeast Pacific, comparing the control case without $\mathbf{u}^*$ (panel a) and the test case (panel b). Differences between these two panels are evident. As well, panel b differs from EH insofar as the eastern boundary poleward flow is more

narrowly confined over the upper continental slope (appreciating the coarseness of resolution, here with roughly 1.9° grid spacing). Another difference between recent work and the earlier EH is a clearer upper ocean expression of the entropy tendency which is not being entirely over-ridden by direct forcing. Deeper in the water column (not shown), the entropy tendency dominates.

Figure 3 shows flow at 245 m in the northwest Pacific. Differences are apparent between the control (panel a) and test (panel b). Inclusion of the entropy tendency in panel b results in stronger Oyashio and Sakhalin Currents, with Kuroshio confluence at lower latitude. Note especially within the Sea of Japan that the circulations in panels a and b are of *opposite sign*, with panel b supporting a warm Tsushima Current along the west coast of Japan (and a colder southward Korea Current). These differences of circulation (here shown at 245 m) are so large that they affect implied air-sea heat and freshwater exchanges. Contours on Figure 3 are implied heat exchange (W/m^2) given by the model relaxation to Levitus atlas. Negative values denote oceanic heat loss. Differences between panels a and b are significant, with the control case (panel a) implying annual mean heat loss in excess of 300 W/m^2 with peak loss occurring north of 40N while the test case (panel b) shows peak heat loss reduced by about 100 W/m^2 and shifted to lower latitude.

ACONCLUSION

We are at a beginning, not a conclusion. I think we can say that the equations of motion assumed by conventional ocean modelling are wrong. They are wrong because they do not take account of forces which should arise due to the dependence of system entropy upon the model state. What we are just beginning is the attempt to improve (complete) the equations of motion. Certain calculations can be made with care for both equilibrium and disequilibrium statistical mechanics. Unhappily, the idealizations required to effect these calculations prevent direct practical application. To pass beyond this point with only present knowledge requires bold leaps. The unsupported, vague nature of those leaps may annoy more serious-minded colleagues. Yet I think the leaps can accomplish two things (apart from sheer fun): first, to get a glimpse of what may lie ahead, and, second, to attract interested talent that may strengthen the basis for surer leaps in the future.

First exercises have been chosen with simplicity in mind. We've considered large scale, primitive equation, prognostic ocean modelling. By considering only resolution coarse compared with first internal radius of deformation, it is possible to define a maximum entropy ("unprejudiced") circulation $\mathbf{u}^*$, which is barotropic at the coarse scale. [See Fig. 1 of EH.] Ambiguities arise when we seek to apply results from inviscid quasigeostrophy to a model based upon forced-dissipative, full-depth, primitive equations. Certain choices (leaps) were made that will surely be refined in future work.

Further exercises will look in two directions. We must look back, seeking a better footing for the uncertain leaps. And we continue to look forward. To date we have explored

modification of a particular ocean model, chosen for its history and wide usage. There are other prognostic ocean models, differently formulated in terms of variables or grids (isopycnal layers, terrain-following coordinates, ...) and in terms of equation sets (thermocline equations, semi-geostrophy, balance, ...). At finer resolution (coastal zone studies, individual seamounts, ...) effects of stratification will need be taken into account in the maximum entropy solution. The combination of stratification and large amplitude topography may be especially vexing. In coarse resolution exercises thus far, we've avoided eddy-active models. As higher resolution models "admit" eddies (whether or not adequately "resolving" them), some of the entropy tendency will be expected to appear within the explicit model. How to complement such eddy-active models with entropy tendency? (If one follows the $A\nabla^2(\mathbf{u}-\mathbf{u}^*)$ implementation, it is natural to reduce A at higher resolution. One might also substitute other operators such as ∇^4.)

What we do with prognostic models suggests modification also to diagnostic or inverse models. To the extent that an inverse model may be constrained by equations of motion, "improved" equations for prognostic modelling should improve the quality of inverse solutions. As well, inverse models often minimize a cost function. There it is natural to append a penalty for distance from maximum entropy, thereby seeking a "least prejudiced" inverse solution. An important consideration is to include uncertain parameters (L^2, ...) as parts of the inverse solution, evaluating them by best fit to data.

Finally, we can choose to ignore all this stuff. When computers someday get to be powerful enough (and if one hasn't anything else interesting to compute), then we can just clobber everything by brute force.

ACKNOWLEDGMENT. This work has been supported in part by the Office of Naval Research (N00014-92-J-1775).

REFERENCES

Aagaard, K., 1989, A synthesis of the Arctic Ocean circulation, *Rapp. P.-v. Reun. Cons. int. Explor. Mer*, 188, 11-22.

Carnevale, G. F., 1979, Statistical dynamics of nonequilibrium fluid systems, PhD thesis, Harvard Univ.

Carnevale, G. F., U. Frisch and R. Salmon, 1981, H-theorems in statistical fluid dynamics, *J. Phys. A.*, 14, 1701-1718.

Carnevale, G. F. and J. S. Frederiksen, 1987, Nonlinear stability and statistical mechanics of flow over topography, *J. Fluid Mech.*, 175, 157-181.

Cox, M. D., 1984, *A primitive equation, three-dimensional model of the ocean*, GFDL Ocean Group, Tech. Rept. No. 1, Princeton Univ.

Eby, M. and G. Holloway, Experiments with a hybrid statistical mechanics /ocean circulation model, 'Aha Huliko'a 1993 (this volume).

Frederiksen, J. S., 1982, Eastward and westward flows over topography in nonlinear and linear barotropic models, *J. Atmos. Sci.*, 39, 2477-2489.

Frederiksen, J. S., 1985, Strongly nonlinear topographic instability and phase transitions, *Geophys. Astrophys. Fluid Dyn.*, 32, 103-122.

Frederiksen, J. S., 1986, Stability properties of exact nonzonal solutions for flow over topography, *Geophys. Astrophys. Fluid Dyn.*, 35, 173-207.

Griffa, A. and R. Salmon, 1989, Wind-driven ocean circulation and equilibrium statistical mechanics, *J. Marine Res.*, 47, 457-492.

Hellerman, S. and M. Rosenstein, 1983, Normal monthly wind stress over the world ocean with error estimates, *J. Phys. Oceanogr.*, 13, 1093-1104.

Herring, J. R., 1977, Two-dimensional topographic turbulence, *J. Atmos. Sci.*, 34, 1731-1750.

Holloway, G., 1978, A spectral theory of nonlinear barotropic motion above irregular topography, *J. Phys. Oceanogr.*, 8, 414-427.

Holloway, G., 1986, Eddies, waves, circulation and mixing: statistical geofluid mechanics, *Ann. Rev. Fluid Mech.*,18, 91-147.

Holloway, G., 1987, Systematic forcing of large-scale geophysical flows by eddy-topography interaction, *J. Fluid Mech.*, 184 , 463-476.

Holloway, G., 1992, Representing topographic stress for large scale ocean models, *J. Phys. Oceanogr.*, 22, 1033-1046.

Kubo, R., 1965, *Statistical Mechanics: an Advanced Course with Problems and Solutions*, North-Holland Publ., 425 pp.

Lesieur, M., 1990, *Turbulence in Fluids*, 2nd Ed., Kluwer Academic, Dordrecht, Netherlands, 412 pp.

Levitus, S., 1982, *Climatological Atlas of the World Ocean*, NOAA Prof. Paper 13, Washington, D. C.

McComb, W. D., 1990, *The Physics of Fluid Turbulence*, Oxford Sci. Publ., 572 pp.

Salmon, R., G. Holloway and M. C. Hendershott, 1976, The equilibrium statistical mechanics of simple quasi-geostrophic models, *J. Fluid Mech.*, 75, 691-703.

EXPERIMENTS WITH A HYBRID STATISTICAL MECHANICS/ OCEAN CIRCULATION MODEL

Michael Eby and Greg Holloway
Joint Program for Ocean Dynamics
Institute of Ocean Sciences, Sidney, BC, V8L 4B2, Canada, and
Centre for Earth and Ocean Research, University of Victoria, V8W 2Y2, Canada

ABSTRACT

A hybrid statistical mechanics / ocean circulation model is tested. A conventional ocean model was revised to include a tendency for model streamfunction to relax toward a maximum entropy configuration which depends on the shape of topography. The tendency is called "topographic stress". Comparisons are made between three cases; a control case with streamfunction relaxation toward rest and two implementations of topographic stress (differing by their functional dependence on total depth). The two topographic stress cases perform similarly, but they differ from the control case in several regards. Topographic stress strengthens equatorward tendencies in deep western boundary currents, sustains a deep Alaska Stream, and leads to poleward eastern boundary undercurrents which are absent in the control. In the upper water column, where direct wind and buoyancy forcing dominate, the influence of topographic stress is slight.

INTRODUCTION

Oceanic general circulation models (OGCM) are used to advance our understanding of physical processes in the ocean. Increasingly, OGCMs are being coupled to atmospheric models and used to predict climate change. Most models, however, are not capable of simulating present day ocean climatology accurately enough to provide a confident basis for predictions. We are motivated to search for systematic defects which afflict these models. In particular, OGCMs which are global in their domain and used for prediction over decades or longer are necessarily of relatively coarse resolution. Oceanic eddies on length scales of tens of kilometers are either not resolved, or are only marginally resolved in ways that may corrupt their dynamics. It is therefore important to find a representation of unresolved eddies which is of sufficient skill to better recover present day ocean climatology, providing an improved basis for climate change studies.

It has been suggested by Holloway (1992) that eddies interact with bottom topography to generate pressure-slope correlations, possibly exerting large systematic forces (topographic stress) upon mean circulation. The usual eddy parameterizations in terms of bottom drag or eddy viscosity move a model towards a state of rest, whereas topographic stress may be a driving force behind mean flows. It is suggested that a more skillful representation of unresolved eddies may be given by the tendency toward higher system entropy.

Statistical dynamical tendencies were examined by Salmon et al. (1976) in the context of ideal quasi-geostrophic dynamics. Among their simplest results is the expectation that, on scales larger than the first deformation radius, motion should tend to be barotropic and

given by a streamfunction satisfying

$$(\alpha/\beta - \nabla^2) < \psi >= h \tag{1}$$

where ∇^2 is the 2-dimensional Laplacian, α/β is a ratio of Lagrange multipliers (due to dynamics which conserve energy and enstrophy), $h = f\delta H/H$ is the potential vorticity due to variation δH about mean depth H, and f is the Coriolis parameter. This equation implies that an ocean with no external forcing, filled with random eddies (without mean motion), would tend to set up a mean flow (ψ) that depends on the topography (h).

In reality, the ocean has external forcing and internal dissipation, and thus is not a closed system to which maximal entropy solutions apply. The state of actual ocean circulation is achieved as a balance between entropy-increasing tendencies on account of eddy interactions and entropy-limiting tendencies due to forcing and dissipation. OGCMs already have modest skill to include large scale forcing, while internal dissipation is parameterized more haphazardly (in part due to poorly understood eddies). What OGCMs omit is the eddy tendency toward increasing entropy. We investigate the effects of modifying an OGCM such that the models would relax not toward rest, but rather toward a solution such as (1). There is a theoretical leap in applying a parameterization based on quasi-geostrophy to a primitive equation model. For this reason, as well as uncertainty in how to characterize the competition between forcing-dissipation and topographic stress, we do not know precisely how to proceed. What we do hope is that this study will help motivate further theoretical work and demonstrate that the inclusion of a relatively crude parameterization of topographic stress already improves the quality of model simulations.

IMPLEMENTATION

The model chosen to study the effects of topographic stress was the GFDL Modular Ocean Model (MOM) (Pacanowski et al. 1991) which is based on code originally formulated by Bryan (1969) and further developed by Semtner (1974) and Cox (1984). Versions of this three dimensional, primitive equation model are widely used (Killworth et al. 1991).

MOM calculates velocity as internal (baroclinic) and external (barotropic) modes. From a vorticity tendency, the model solves an elliptic equation for transport streamfunction from which it obtains the external mode of velocity. We will be using MOM at a grid resolution more coarse than the first deformation radius, hence at scales for which the maximum entropy solution is barotropic. Thus, we can introduce a simple relaxation of the model streamfunction toward that given by (1).

There is a further simplification as well as certain ambiguities which arise in application based upon (1). The ratio α/β is not well defined in reality, since its theoretical motivation depends upon artifacts such as finite spectral truncation. However, $\alpha/\beta = 1/L^2$ defines a length scale which is plausibly related to eddy length scales. In what follows, we treat L as an adjustable parameter on the order of 10 km. The model resolution we will use is much coarser than L, so we may omit ∇^2 in (1), taking only $\psi = L^2 h$

Ambiguities arise also because (1) is based upon quasi-geostrophy whereas application will be made in primitive equation MOM. The range of variation of depth, expressed by h, should be small under quasi-geostrophy. In fact we will use the full range of oceanic depth, making $\psi = -fL^2 H/H_o$ where H_o is a reference depth. Under quasi-geostrophy, interpretation of ψ is

arbitrary; it may describe either a transport or velocity streamfunction. If we adopt the velocity streamfunction view, then ψ will be converted to a transport streamfunction for incorporation into MOM. Because variation in ψ is dominated by variation in H, an approximation for the maximum entropy transport streamfunction is given by

$$\Phi^* = -fL_v H^2 \quad \text{where} \quad L_v = \frac{\beta}{2\alpha H_o} \tag{2}$$

If we adopt the transport streamfunction view of (1), we multiply through by H_o, and the maximum entropy transport streamfunction becomes

$$\Phi^* = -fL_t^2 H \quad \text{where} \quad L_t^2 = \frac{\beta}{\alpha} \tag{3}$$

Ambiguities such as the different functional dependences in (2) or (3) may seem unnerving. There should be no pretense to sophistication here. Simply, our aim is to use MOM to explore sensitivity, comparing differences under (2) or (3) with the results from traditional subgridscale relaxation to rest (Φ^* = constant). Length scales L_v and L_t are treated as adjustable, with H_o absorbed into L_v. Moreover, one may consider that these length scales exhibit some weak spatial dependence. In particular, we will consider that L_v or L_t vary with latitude. Clearly this invites parameter tuning. At present, our aim is only to observe sensitivity to such issues.

Finally, the transport stream function (Φ) calculated by MOM is replaced at each time step with

$$\Phi + \frac{\delta t}{\tau}(\Phi^* - \Phi) \tag{4}$$

using either (2) or (3) for Φ^*. The model velocity time step δt is of order 1 hour and τ is an adjustable relaxation time of order 25 days.

MODEL SETUP

A coarse-resolution global model was created with grid spacing 3.75° in longitude and 3.711° in latitude. This resolution closely approximates the spectral T32 grid used by the global atmospheric general circulation model of the Canadian Climate Centre. Fifteen levels were used with layer thicknesses ranging from 20 to 870 m. Topography was extracted from ETOPO5 (1986) using a raised cosine weighted average of the data within a grid cell. Four islands were included: Madagascar, Australia, New Zealand and Antarctica.

The model was forced with annual mean Hellerman and Rosenstein (1983) windstress and a 50 day relaxation of surface salinity and temperature to annual mean Levitus (1982). The domain was limited at 69° North to avoid high grid latitudes, and salinity and temperature were relaxed to Levitus values on the artificial northern boundary with a time scale of 3 years. Horizontal viscosity, horizontal diffusion, vertical viscosity and vertical diffusion were set to 2×10^5 m^2 s^{-1}, 4×10^3 m^2 s^{-1}, 2×10^{-3} m^2 s^{-1} and 1×10^{-4} m^2 s^{-1} respectively. Velocity time steps were 1 hour and the tracer time step was 2 days.

In exploratory integrations, we have allowed relaxation time τ to vary from 10 to 200 days, and length scale L_t to vary from 10 to 30 km. Results overall were as expected — the

topographic stress parameterization caused larger model responses in cases with larger length scale or shorter relaxation times. These integrations also demonstrated that without a latitude dependence, topographic stress tendencies were relatively too strong at high latitudes. For longer integrations, we have assigned Φ^* a latitude dependance given by $1 - 0.9 \sin(\text{latitude})$. We have chosen relaxation time τ to be 25 days and the equatorial value of length scale L_t to be 22 km.

To compare the two implementations of topographic stress, relative values of the length parameters L_t and L_v must be assigned. Equating the topographic stress solutions given by (2) and (3), length parameters are related by $L_v H^2 = L_t^2 H$. Within the model, H varies discretely from 0 to 5.5 km; we equate both solutions at 5.5 km. Thus given a choice of 22 km for L_t, an equivalent L_v is 88 km. Although the extreme values of the two implementations are equivalent, the barotropic velocities will be different. A topographic stress which is proportional to L^2H, as in (2), will tend to produce relatively stronger barotropic velocities in shallow water compared to a topographic stress which is proportional to LH^2, as in (3). Maximum entropy (equilibrium) velocities corresponding to (2) and (3) are shown in Figure 1.

Integrations were carried out in parallel — two with relaxation to the equilibrium solutions described by (2) and (3), and a third with relaxation to zero (control). To keep the runs as similar as possible, the control case includes relaxation to zero streamfunction. This takes into account that topographic stress velocities are small compared with direct forced, upper ocean flows. Thus topographic stress is closer to relaxing streamfunction to zero than to no relaxation. As well, when including streamfunction relaxation, less explicit viscosity is required, so the model's total transports are largely unchanged. With respect to the damping processes, we try to keep the control case as similar as possible to the topographic stress cases.

Integrations were started from horizontally averaged Levitus data. Interior relaxation to Levitus temperature and salinity was continued for 10 years with a relaxation time scale of 1 year. This method of start-up avoids shocking the nodel with observed data that is incompatible with the model physics (Semtner and Chervin 1988). The model was then released from any interior relaxation and integrated for another 190 years. Although the model will not have reached equilibrium after 200 years, comparisons of trends can be made between parallel runs.

RESULTS

From Figure 1 we can anticipate some of the effects of topographic stress. Equilibrium velocities are poleward on the eastern, and equatorward on the western, slopes of basins. These velocities suggest currents which are opposite to many of the well known surface currents such as the Gulf Stream, Canary, Brazil and Benguela Currents or the Kuroshio, California, East Australia and Peru Currents. Magnitudes of statistical dynamical velocities are small, however, compared to the magnitude of the wind-driven surface currents, so we expect that topographic stress will have little effect on the surface circulation. At greater depths, where velocities from directly forced flows have smaller magnitude, tendency toward higher entropy has relatively greater effect. One anticipates the development or strengthening of poleward undercurrents along eastern boundaries and equatorward undercurrents along western boundaries.

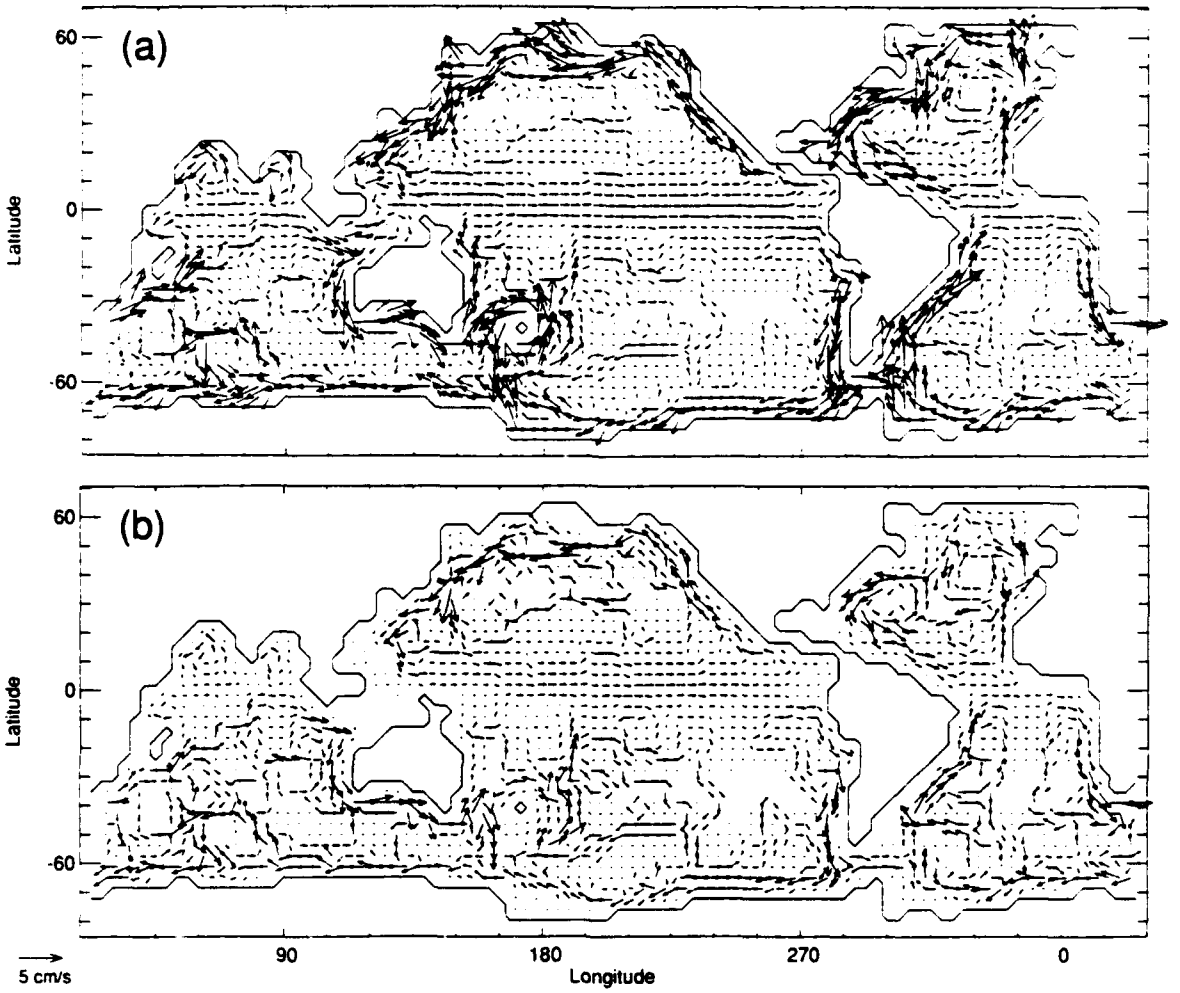


Figure 1. (a) Maximum entropy (equilibrium) velocities: (a) for the L^2H case and (b) for the LH^2 case.

Velocities for the first few levels are dominated by the wind and buoyancy driven surface circulation. Since plots from the three integrations are visually indistinguishable, only one plot at 35 m (level 2) is shown in Figure 2. Total transports are also very similar for all three runs since much of the transport occurs in the upper ocean. Small differences in transport are noticeable along continental margins, but these differences are more clearly seen when comparing velocity fields. Results from the model integrations at greater depths will be discussed for three geographic areas: the North Pacific, the Mid-Atlantic and the Indian Oceans.

We will show results at two depth levels: 850 m (level 8) and 2750 m (level 12). Shallower levels are dominated by direct forcing. With the choice of parameters used here, the competition between direct forcing and statistical dynamics is such that the statistical dynamical tendencies become apparent in the lower main thermocline, roughly 850 m. At greater depths, statistical dynamical tendencies become more dominant.

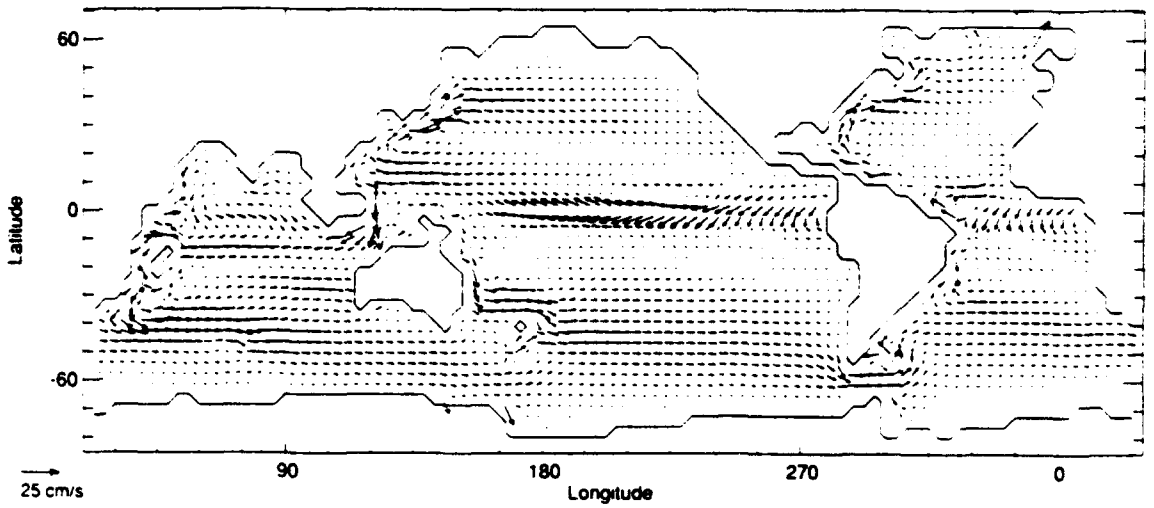


Figure 2. Velocities at 35 m for the control case (results including topographic stress are visually identical).

North Pacific

Velocities at 850 m for the three implementations are shown in Figure 3. Differences between the control run (Figure 3a) and the topographic stress runs (Figures 3b and 3c) can be seen along the continental margins. The control run exhibits no California undercurrent and only a weak Oyashio. The control also has a strong Kuroshio extension cutting off a weak Alaska Stream.

Observational evidence for an undercurrent along the West coast of North America is extensive, including work by Hickey (1979) off the coast of Southern Washington, Freeland et al. (1984) off Vancouver Island and Chelton (1984) off California. Measurements indicate poleward flow up to 15 cm s^{-1}, often with a width greater than 100 km, usually with a maximum at depths less than 700 m, but extending to more than 1000 m. Most of these studies were coastal in nature, thus the width and depth of the underflow has not been well established. The temporal and spatial persistence of the undercurrent is also not well known.

Observations by Warren and Owens (1988) indicate a deep Alaska Stream flowing westward, with mean velocities between 1 and 3 cm s^{-1}, along the northern side of the Aleutian Trench. They also report evidence for a deep, eastward jet which flows parallel to the trench, south of the Alaska Stream.

Figure 4 shows velocities at 2750 m for the three integrations. Coastal currents induced or strengthened by topographic stress at 850 m are seen more clearly at 2750 m, with a western boundary undercurrent now extending to the equator. Smaller differences between the control run and the topographic stress runs can be seen in the central Pacific.

Direct observations of deep western boundary currents in the North Pacific are few. Indirect inference from tracers such as silica (Tally and Joyce 1992) suggest northward deep flow along the western boundary. Current meter observations by Fukasawa et al. (1986) in the Shikoku Basin south of Japan (west of the region considered by Talley and Joyce 1992) show deep

mean currents of 5 to 10 cm s^{-1} toward the south-west (parallel to local isobaths). Northward flow of low silica water is contrary to that indicated by Figure 4, whereas the current meter observations are consistent with topographic stress.

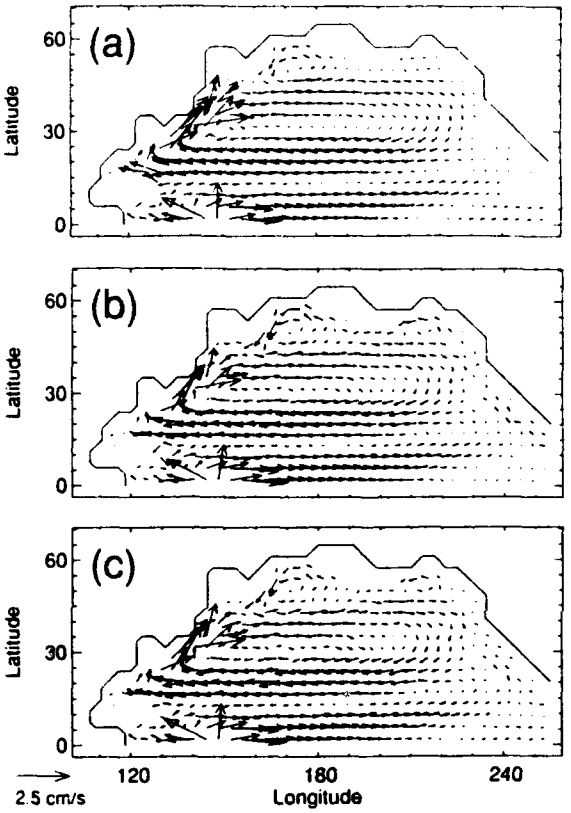


Figure 3. Velocities at 850 m in the North Pacific Ocean: (a) for the control, (b) for the L^2H and (c) for the LH^2 cases.

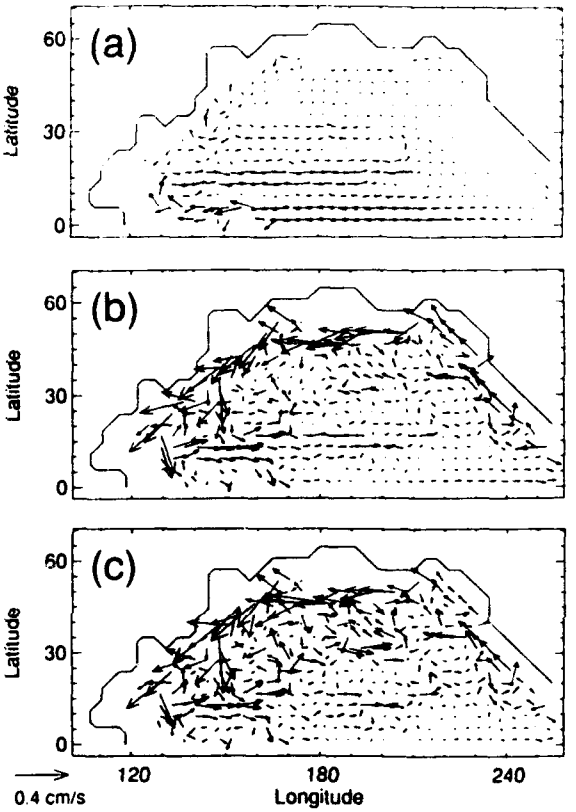


Figure 4. Velocities at 2750 m in the North Pacific Ocean: (a) for the control, (b) for the L^2H and (c) for the LH^2 cases.

The deep, narrow trenches found in the Pacific are not resolved by this model, but could be important in setting up counterflows such as the one observed by Warren and Owens (1988) south of the Alaska Stream. At small scale the baroclinic influence should be taken into account, however, the barotropic formulations for topographic stress (formulae 2 or 3) suggest a tendency for opposing currents on opposite sides of a trench. One could imagine cyclonic shear above the trenches in mid-depth and abyssal waters, supporting northward and eastward transport of tracers in the western Pacific while current meters on the inshore side of trenches show southward or westward flow.

Differences between the two topographic stress runs are subtle. The effects (when compared to the control run) produced by the L^2H implementation (Figures 3b and 4b) tend to be stronger along the coast and slightly weaker in the central Pacific than with the LH^2 implementation

(Figures 3c and 4c). Since plots of the two implementations of topographic stress are so similar, only plots from the control run and the L^2H run will be shown for the other regions.

Mid-Atlantic

Model velocities at 850 m are shown in Figure 5. Small differences between the control run (Figure 5a) and the topographic stress run (Figure 5b) can be seen along the continental margins. Topographic stress has weakened or reversed the control runs equatorward eastern boundary currents. The northward flow along the western margin is also weaker in the North Atlantic and stronger in the South Atlantic for the topographic stress run than for the control.

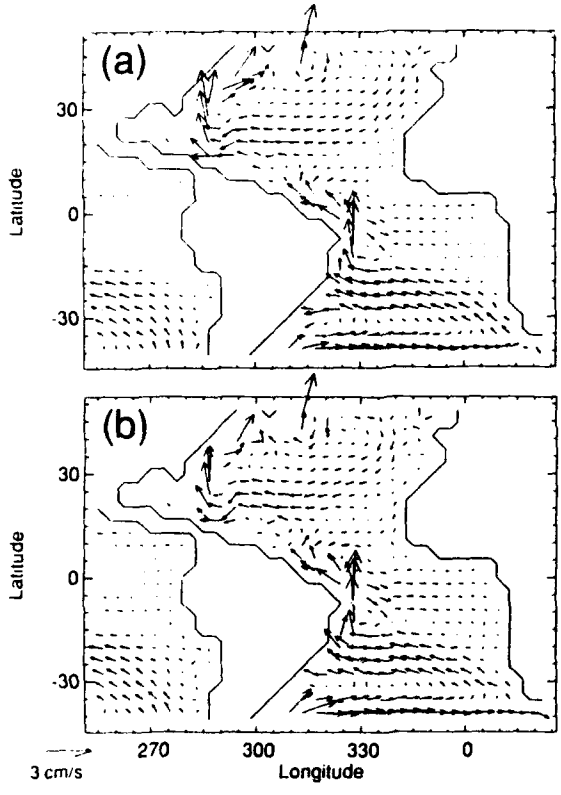


Figure 5. Velocities at 850 m in the Mid-Atlantic Ocean: (a) for the control and (b) for the L^2H cases.

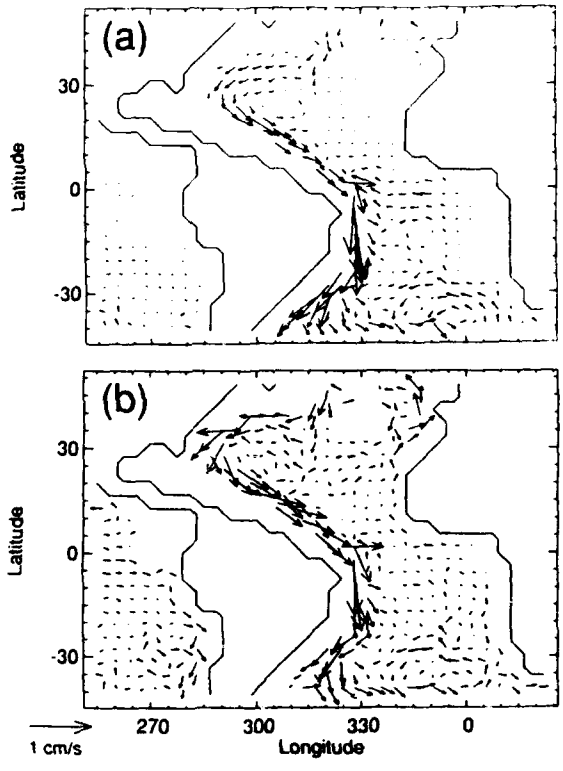


Figure 6. Velocities at 2750 m in the Mid-Atlantic Ocean: (a) for the control and (b) for the L^2H cases.

Differences between the control run and the topographic stress run are more obvious at 2750 m (Figure 6) than at 850 m. The control run does not develop any poleward eastern boundary undercurrents. The deep, southward flowing western boundary currents are also stronger in the North Atlantic, and weaker in the South Atlantic for the topographic stress run compared to the control run.

Poleward undercurrents have been observed off the west coast of South Africa (Nelson 1989) and off the coast of North Africa (Mittelstaedt 1989). Poleward flow has also been described off the Iberian Peninsula (Barton 1989). Although spatial and temporal knowledge of poleward

undercurrents along the Eastern Atlantic is limited, direct measurements indicate a flow of up to 10 cm s^{-1} with a width of 30 to 100 km, often with a maximum at about 300 m, but extending to great depths.

The Peru-Chilean undercurrent is also present in Figure 6b. Observations of the Peru-Chilean undercurrent are summarized by Fonseca (1989). This current has been seen from the surface to below 300 m.

While topographic stress may be a dominant force in the deep western boundary currents of the North Pacific, thermohaline forcing is clearly important in the Atlantic. Because the model domain is truncated at 69° North, the thermohaline forcing has in part been provided by relaxation toward mean Levitus at all depths along the artificial northern boundary. To test the sensitivity of the model to this northern boundary condition, two further integrations were performed, for the control case and for the L^2H case, without interior relaxation on the boundary. Velocities at 2750 m are shown in Figure 7. Without relaxation, the western boundary current is largely absent in the control case (Figure 7), but remains present in the topographic stress case (Figure 7b).

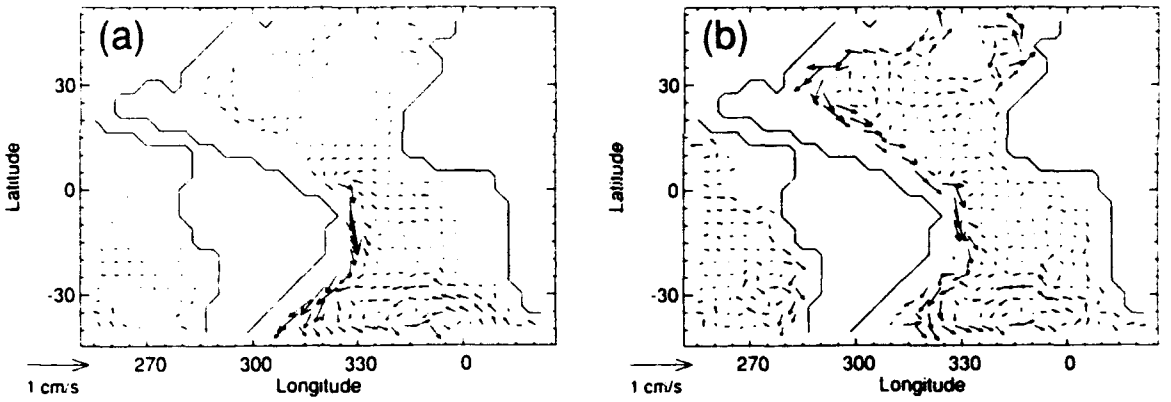


Figure 7. Velocities at 2750 m in the Mid-Atlantic Ocean without relaxation to Levitus along the northern boundary: (a) for the control and (b) for the L^2H cases.

It is clear that when the nature of imposed forcing happens to agree with a maximum entropy configuration, topographic stress may not play a very significant role. The stress depends upon how far forced-dissipative flows are from ideal maximum entropy. A second observation concerns the climatic implications of these results. One may speculate that variation in thermohaline forcing of the North Atlantic could lead to abrupt alteration of the pattern of deep circulation. Comparison of Figures 6a and 7a indeed appears to support this speculation. However, when topographic stress is included, Figures 6b and 7b show a deep circulation which is rather insensitive to changes in thermohaline forcing. The indication is that inclusion or omission of topographic stress in coupled ocean-atmosphere climate models could have significant effect on the overall sensitivity of the coupled system.

Indian

Model velocities at 850 m are shown in Figure 8. Effects of topographic stress include a slight weakening of the West Australian and the Agulhas Currents, and a slight strengthening of both the undercurrent off the east coast of Australia and the cyclonic circulations in the Northern Indian Ocean.

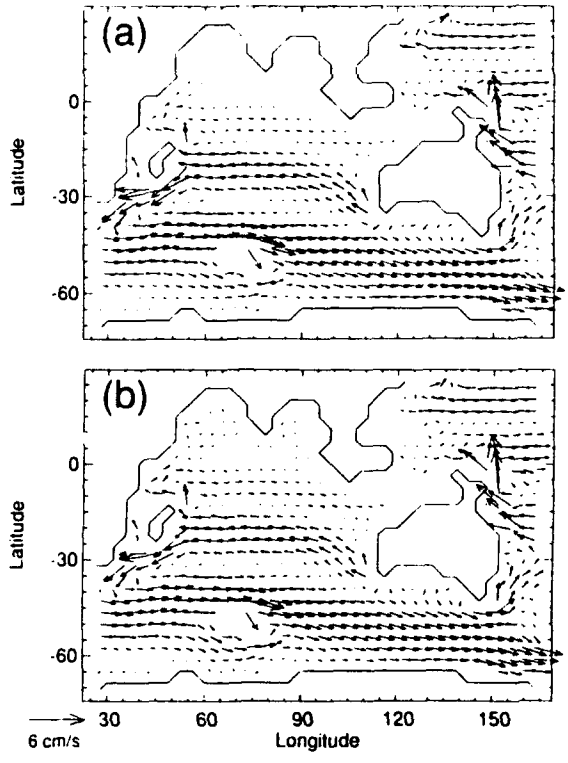


Figure 8. Velocities at 850 m in the Indian Ocean: (a) for the control and (b) for the L^2H cases.

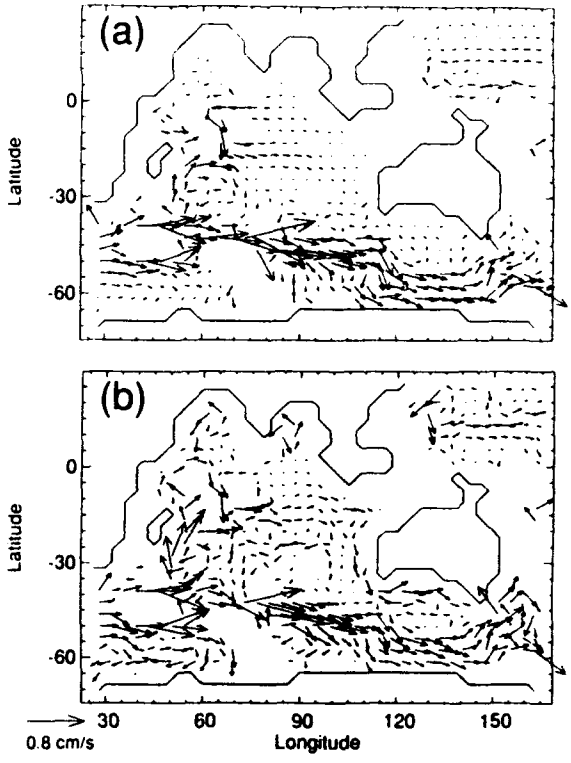


Figure 9. Velocities at 2750 m in the Indian Ocean: (a) for the control and (b) for the L^2H cases.

Figure 9 shows velocities for the two integrations at 2750 m. Differences between the two runs are more pronounced than at 850 m. Topographic stress induced effects include: an increase in the transport of deep Antarctic water northward into the equatorial Indian; a strengthened undercurrent along the south, west and east coasts of Australia; a strengthening of the circulation in the North Indian; and a slight weakening of the Circumpolar Current near the coast of Antarctica leading to the development of a countercurrent, west of the Kerguelen Plateau.

An undercurrent beneath the Leeuwin current has been observed to be equatorward (Church et al. 1989), contrary to the topographic stress tendency. Model results off the west coast of Australia demonstrate the competition between direct forcing and topographic stress. Figure 2 suggests a weak, poleward current which continues down to about 200 m, after which an equatorward undercurrent is present until about 1000 m (Figure 8). Although topographic stress opposes an equatorward undercurrent, other forces are overriding the

maximum entropy tendency. Between 1000 and 2000 m, both integrations generally show poleward flow. Differences between the two runs are most apparent between 2000 and 3500 m where slow, mixed flow is seen for the control run while a steady poleward flow is seen for the topographic stress run (Figure 9). Below 3500 m the flow is again equatorward for both integrations, although stronger for the control run.

SUMMARY

The GFDL Modular Ocean Model was used to explore proposed representations of topographic stress for large scale ocean modelling. We have sought to characterize the effect of unresolved eddy-topography interaction (topographic stress) in terms of a tendency for large scale flow to evolve toward a state of higher system entropy.

Two implementations of topographic stress were tested, differing in the assumed functional dependence of streamfunction upon depth. The two topographic stress cases perform similarly, but they differ from the control case in several respects. The most apparent effects of topographic stress are seen along continental boundaries, particularly in the development or strengthening of undercurrents. Integrations which include topographic stress produce many of the observed poleward eastern boundary undercurrents which the control run does not. Along western boundaries, the topographic stress tendency is equatorward. Differences from the control run are seen more clearly in the western Pacific than in the western Atlantic, since the Atlantic is also responding to stronger thermohaline forcing.

Acknowledgments

This work was carried out with the support of the Office of Naval Research (grants N00014-87-J-1262 and N00014–92–J-1775). The authors would like to thank Kelly Choo for his work on graphics.

REFERENCES

Barton, E. D., 1989, "The poleward undercurrent of the eastern boundary of the subtropical North Atlantic", in Neshyba et al., 1989.

Bryan, K., 1969, "A numerical method for the study of the circulation of the World Ocean", J. Comput. Phys., 4, 347-376.

Cox, M. D., 1984, "A primitive equation, three-dimensional model of the ocean", GFDL Ocean Group Tech. Rep. No. 1. [Available from Geophysical Fluid Dynamics Laboratory, NOAA, Princeton University, Princeton, NJ 08542]

Chelton, D. B., 1984, "Seasonal variability of alongshore geostrophic velocity off central California", J. Phys. Oceanogr., 12, 757–784

Church, J. A., G. R. Cresswell and J. S. Godfrey, 1989, "The Leeuwin Current", in Neshyba et al., 1989.

Freeland, H. J., W. R. Crawford and R. E. Thompson, 1984, "Currents along the Pacific coast of Canada", Atmos-Ocean, 22, 151-172.

Fukasawa, M., T. Teramoto and K. Taira, 1986, "Abyssal current along the northern periphery of Shikoku Basin", J. Oceanogr. Soc. Japan, 42. 459-472.

Hellerman, S., and M. Rosenstein, 1983, "Normal monthly wind stress over the World Ocean with error estimates", J. Phys. Oceanogr., 13, 1093-1104.

Hickey, B. M., 1979, "The California Current system - hypotheses and facts", Prog. Oceanogr., 8, 191-279.

Holloway, G., 1992, "Representing topographic stress for large-scale ocean models", J. Phy. Oceanogr., 22, 1033-46.

Killworth, P. D., D. Staniforth, D. J. Webb and S. M. Paterson, 1991, "The development of a free-surface Bryan-Cox-Semtner ocean model", J. Phys. Oceanogr., 21, 1333-48.

Levitus, S., 1982, "Climatological atlas of the World Ocean", NOAA Prof. Paper 13, Washington, D. C.

Mittelstaedt, E., 1989, "The subsurface circulation along the Moroccan slope", in Neshyba et al., 1989.

Nelson, G., 1989, "Poleward motion in the Benguela area", in Neshyba et al., 1989.

Neshyba, S. J., C. N. K. Mooers, R. L. Smith and R. T. Barber, eds., 1989, Poleward Flows Along Eastern Ocean Boundaries, Springer-Verlag, 374 pp.

Pacanowski, R., K. Dixion and A. Rosati, 1991, "The G.F.D.L. Modular Ocean Model Users Guide version 1.0", GFDL Ocean Group Tech. Rep. No. 2., [Available from Geophysical Fluid Dynamics Laboratory, NOAA, Princeton University, Princeton, NJ 08542]

Salmon, R., G. Holloway and M. C. Hendershott, 1976, "The equilibrium statistical mechanics of simple quasi-geostrophic models", J. Fluid Mech., 75, 375–386

ETOPO5, 1986, "Global 5' x 5' depth and elevation", [Available from National Geophysical Data Centre, NOAA, U.S. Dept. of Commerce, Code E/GC3, Boulder, CO 80303]

Semtner, A. J., 1974, "A general circulation model for the World Ocean", UCLA Dept. of Meteor. Tech. Rep. No. 8, 99 pp.

Semtner, A. J. and R. M. Chervin, 1988, "A simulation of the global ocean circulation with resolved eddies", J. Geophys. Res., 93, 15502-22.

Tally, L. D., and T. M. Joyce, 1992, "The double silica maximum in the North Pacific", J. Geophys. Res., 97, 5465–5480.

Warren, B. A. and W. B. Owens, 1988, "Deep currents of the central subarctic Pacific Ocean", J. Phys. Oceanogr., 18, 529-551.

STATISTICAL METHODS IN PHYSICAL OCEANOGRAPHY: MEETING REPORT

Peter Müller
Department of Oceanography, School of Ocean and Earth Science and Technology, University of Hawaii, Honolulu, Hawaii

and

Greg Holloway
Institute of Ocean Sciences, Sidney, BC, Canada

Physical oceanographers deal with randomness and uncertainties when analyzing ocean data or formulating ocean models. Concepts and results from probability theory, statistical inference, and stochastic processes are applied: space-time averages are interpreted as ensemble averages, variances and spectra estimated, and stochastic terms added to dynamical equations. Special aspects arise when the huge amount of real or model ocean data and the complexities of ocean physics are considered. Efficient data representation and analysis algorithms are sought, and idealized dynamics that incorporate statistical and chaotic tendencies are explored. Progress on such statistical methods was discussed at the seventh 'Aha Huliko'a Hawaiian Winter Workshop, held January 12-15, 1993 in Honolulu. Specifically, the participants considered the variety of oceanographic observations, methods for efficient flow and data representation, frequentist versus Bayesian inference, data assimilation, and idealized dynamics. The size and complexity of oceanographic problems often prevent the application of standard methods and physical oceanographers are faced with the task of inventing methods that deal with the peculiarities of their problems in a sensible manner. These special methods are discussed in more detail in this article. Names in parentheses refer to the authors of lectures given at the meeting and chapters published in the proceedings.

OCEANIC OBSERVATIONS

Oceanographic phenomena cover many space and time scales and the questions asked, probability assumptions made, and inferences drawn differ widely. A few examples were considered.

The smallest scales of variability are important for mixing. One of the oldest unsolved problems is how the strength of ocean mixing varies with location and over time. Traditional measurement methods have depended upon making estimates of dissipation rates from direct observations at scales of centimeters. Fluxes are inferred from

dissipation, after further uncertain assumptions. As instruments for direct observation of the fluxes become available, the relation of flux to dissipation should become clearer. However, the difficulty remains that direct observation at centimeter scales is a time-consuming (hence expensive) operation, limiting its applicability for larger scale ocean surveying. A possibility is that larger scale observation of vertical velocity by a modified acoustic Doppler current profiler may provide a basis for estimation of dissipation and mixing while enabling rapid surveying (A. Gargett). The value of this proposed method will depend upon extensive intercomparison with more traditional measurements, using statistics to calibrate, and to estimate the confidence of, inferences from acoustic Doppler surveys.

Over vertical scales from meters to tens of meters, it is natural to describe the variability in the ocean interior in terms of vertical displacements of isopycnals. The frequency of chosen isopycnals in a given vertical bin size may be fitted against a Poisson distribution while the separation between isopycnals may approximate a gamma distribution, constrained to have unit mean (R. Pinkel). The mean, variance, and skewness at many scales is thus described by a single (dimensional) parameter, whose physical significance remains elusive.

On larger scales, underway acoustic Doppler profiling together with accurate global positioning has made possible the surveying of upper ocean currents along ship tracks (E. Firing). Length scales of velocity variations, both in the horizontal and vertical, are observed to change with latitude and depth, including strong signatures when crossing the equator. No comprehensive statistical descriptions of the variability and its changes have been established.

Satellite altimeter data have allowed the mapping of sea level variance (D. Chelton). A close relation between the variance of transient eddies and the intensity of mean flows and the bathymetry is found. Wavenumber spectra show a break of slope near the first internal Rossby radius, then are descending steeply towards higher wavenumbers. At the crossovers of ascending and descending satellite ground tracks, the two components of surface geostrophic velocity can be determined. This has been used to investigate the anisotropy of velocity variance and lateral transfers of momentum by Reynolds stresses on a dense global grid. Further inferences are complicated by the unique space-time sampling characteristics of satellite observations. The choice of orbit parameters for satellite missions can be discussed in terms of these sampling characteristics and their implied filters, transfer functions, and biases.

Interest in oceanography is often focussed on the larger scales, and considerable ingenuity goes into averaging local observations to remove the "noise" and obtain the "signal." Noise can be suppressed by making observations that are inherently of an integrating type, such as acoustic thermometry, reciprocal tomography, electric field and bottom pressure, inverted echo soundings, cable voltages, polar motion, and length of day. Such integrating

measurements often provide cleaner signatures of underlying dynamical processes. For example, bottom pressure and electric voltage are excellent proxies for the barotropic flow component and are found to be well correlated with the atmospheric windstress curl, suggesting that the subinertial barotropic variability in the ocean is atmospherically forced (D. Luther).

FLOW REPRESENTATION

The proper choice of basis functions on which to decompose a flow field becomes crucial when one seeks to reduce the system space. Reduced bases should optimally resolve the underlying dynamics, representing the more significant flow patterns. Wavelet transforms might be such an optimal choice. Traditional analyses of turbulence have resorted either to Fourier spectra or to grid point (Dirac) treatment, suggesting superposition of waves or interaction of isolated vortices. These are extreme views, emphasizing either perfect wavenumber resolution with no spatial resolution or the converse. Wavelets make a compromise, offering limited wavenumber resolution with limited spatial resolution, suggesting 'coherent structures.' Applied to analyses from numerical two-dimensional turbulence (M. Farge), wavelets were shown to be efficient at retaining the full range of spectral information while permitting spatially local examination of regions dominated by rotation compared with regions dominated by straining.

Another generalization of the traditional Fourier transform is based on the solutions of exactly integrable nonlinear wave equations. The method, known as inverse scattering transform, decomposes a time series into a superposition of nonlinear oscillation modes, which include ordinary sine waves, cnoidal waves, solitary waves, and other special wave forms (A. Osborne). Applications of the method, based on the solutions of the Korteweg-de-Vries equation, have been efficient and insightful in the analysis of field and lab data.

Bases for representing actual data are often chosen in an empirical way by methods known as 'empirical orthogonal functions' (EOFs), 'principal component analyses' (PCAs), or 'factor analyses.' Eigenvalues from the data covariance matrix permit ranking the empirical eigenvectors according to what fraction of total data variance is represented by each eigenvector. This allows retention of a relatively few eigenvectors in lieu of the much larger dataset. Where data are seen to be "clumped" EOFs are rotated to more nearly recognize the "clumps." A major open question is the identification of the significant EOFs. Selection rules that claim to perform a significance test may lack a thorough statistical basis and the often applied rules-of-thumb are just what they claim to be—rules-of-thumb. A promising approach is the testing against artificial data (G. Mitchum). Data generated by a red noise process are particularly relevant to oceanographic fields such as sea level.

Time series from coefficients of EOFs may be subject to further analyses. By best fitting such time series to a first order vector Markov process, one identifies linear combinations

of EOFs that exhibit oscillation or propagation, termed 'principal oscillation patterns' (POPs) (H. von Storch). A POP analysis can be performed both on actual data and on the output from large numerical models. It can be a tool in identifying linear subsystems when these linear subsystems control a significant portion of the variability.

FREQUENTIST VERSUS BAYESIAN INFERENCE

Statistical inferences depend on the probability assumptions made. One interpretation of probability is from a 'frequentist' viewpoint. If an experiment can be regarded as being repeated many times, previous outcomes may be collected to estimate the probability for subsequent outcomes. In contrast, a Bayesian approach may be taken in a case where only one experiment is possible. Then one expresses one's prior beliefs concerning uncertain model parameters as a probability distribution, observes data, then computes a posterior distribution about those parameters. The frequentist's method assumes that the process and all its parameter values remain constant over the (often imagined) series of experiments. The Bayesian approach has to assume a prior distribution. Both methodologies have their advantages, and the appropriate statistical approach depends on the type of problem considered and the type of inference desired (G. Casella).

The Bayesian approach emphasizes a clear statement of prior beliefs. This may be valuable in bringing out 'hidden' assumptions (J. Kadane). How one formulates a null hypothesis can be crucial. If one formulates the null hypothesis as a 'sharp' statement, then collecting sufficient data will lead to rejection, unless the hypothesis is exactly true. Another danger is that formulating a null hypothesis after observing some data, for example in the case of Earth's climate record, may confuse the testing of such a hypothesis (H. von Storch). Bayesian methods have successfully been applied to the quality control of data used for assimilation in numerical weather prediction models where the prior knowledge about error distribution and background fields need to be properly weighted (A. Lorenc).

DATA ASSIMILATION

Combining data with models serves various purposes. The model may help to complete a data set, 'dynamically interpolating' to fill data gaps. An example is the reconstruction of Gulf Stream paths (M. Chin).

One of the ways that oceanography may differ from weather forecasting is that there is less emphasis upon ocean 'forecasting.' To a considerable extent, the role of data assimilation in the ocean may be more as a means to obtain information about uncertain parameters in the ocean physics. Among the more rigorous methods, oceanographers employ adjoint methods, integrating backward in time to obtain the gradient of the cost function. An example is the estimation of vertical eddy viscosity and surface drag coefficient from data in a wind forced Ekman layer (J. O'Brien). For more complex systems, simpler algorithms are being developed that seek to reduce a chosen cost

function toward a smaller value which nonetheless might not be a minimum. This can be seen in a time-dependent, three-dimensional circulation model of the North Atlantic, including an embedded mixed layer with uncertain parameters (J. Schröter). Repeated forward runs of the model are made using different choices of parameters, with outcome evaluated by its cost function.

Model errors, such as caused by imprecisely known physics or forcing fields, often represent a major uncertainty in the problem and need to be accounted for in the cost function. This is the case for the heat flux uncertainties in a model of the tropical sea surface temperature (C. Frankignoul). Since the heat flux uncertainties have poorly known correlation scales, an adaptive method is applied where the model being tuned is also used to determine the uncertainties in the heat flux field. Parameters are found that reduce the warm sea surface temperature bias and that might eliminate the need for a flux-correction term in climate studies.

There are various methods to seek extrema of a function. The adjoint equation technique yields estimates of the gradient of the cost function about a state of model variables, permitting use of some descent algorithm to seek a minimum cost. However, when the cost function has a complicated dependence on model variables, there is danger that descent algorithms may not converge or may locate only a local, rather than a global, minimum. Two alternatives were described (N. Frazer). Simulated annealing (by analogy to cooling from a melt), employs random perturbations of parameters while seeking to reduce a cost function (characterized as a Gibb's free energy, U). When a particular random perturbation reduces U, that perturbation is accepted and a subsequent perturbation is applied. When a perturbation increases U, the perturbation may be accepted with probability given by a Gibb's distribution $\exp(-U/T)$, where parameter T has the role of temperature. The 'art' is to reduce T according to a 'cooling schedule' which is efficient yet avoids 'freezing' into a local minimum of U. The second alternative, termed genetic algorithms, bears similarity to simulated annealing. A population of possible choices of sets of parameters is generated. Members are paired, and randomly chosen portions of the parameter set are exchanged, including some 'mutations.' Members are evaluated by a 'fitness' function (which might be the previous Gibb's distribution) to determine probable representation in a next generation.

CHAOS

Concepts from chaotic dynamics have been applied to account for fluid behavior that appears to be random. Laboratory experiments show aspects of how mixing comes about. A theoretical precept about turbulent mixing is that line elements advected with a fluid should tend to grow exponentially in length. It is argued that this is impossible in a two-dimensional steady flow. Beginning from this simple case, what further dynamics are needed before exponential line stretching occurs? Experiments (J. Ottino) show that a simple (periodic) time dependence is already sufficient to introduce chaotic regions,

yielding exponential line growth. Low order fixed points of the time-periodic mapping dominate the regions of chaotic mixing, where it is seen that hyperbolic fixed points are surrounded by elliptic islands. Unmixed regions stretch and contract, but form coherent islands in the midst of chaos. In three dimensional systems these regions might appear as tubes. These results have implications regarding the character of velocity fields inferred via flow visualization.

The study of chaotic dynamics may contribute to an understanding of several problems in ocean dynamics. If a system is chaotic, perturbation expansions are misleading and predictability is limited. It has been suggested that the El Niño/Southern Oscillation system can be modeled as a low order chaotic system. A phase-space reconstruction analysis (M. Brown) performed using a measured time series of eastern tropical Pacific sea surface temperature provides some support for this hypothesis. "Spaghetti plots" of float trajectories often suggest chaotic behavior. However, floats that have been seeded in an isolated eddy may undergo many rotations of the eddy as it translates without dispersing. It is suggested that such coherent eddies are essential to the observation of anomalous diffusion (rates of particle separation greater than would result from random walks). On the theoretical side, WKB ray paths of surface waves propagating through an ocean of periodically varying depth exhibit chaotic behavior when the amplitude of the topographic variation exceeds a critical value.

A major problem in the application of ideas relating to chaos is the determination of whether an ocean phenomenon is chaotic. Estimates of measures of chaos (such as dimension or the Lyapunov exponent) from oceanographic data require an enormous number of degrees of freedom for any reasonable degree of confidence and are often counter-intuitive. For example, filtering for the purpose of suppressing noise can sometimes be seen to increase the Lyapunov exponent (E. Carter). A related issue is the characterization of the roughness of seafloor topography. Spectra and fractal dimensions have been used. A novel approach considers the 'geometric temperature' of any curve, calculated from the number of intersections of the test curve with randomly selected straight lines (W. Woyczynzki). From geometric temperature, one may proceed to an analogous thermodynamics of curves and surfaces.

STATISTICAL DYNAMICS

The evolving statistics of flows have long been an object of turbulence theory. Renormalization group (RG) techniques have been applied to beta-plane turbulence, seeking an efficient non-eddy-resolving parameterization of small-scale processes to enable more cost-effective large scale modeling. Both viscosity and beta (Coriolis gradient) have to be renormalized or rescaled. Such renormalization re-interprets and quantifies previously obtained results and yields new insights. In particular, the RG-derived spectral energy transfer shows that beta-plane turbulence naturally tends to self-organize into zonal-jetlike flows and westward propagating Rossby waves. Two-parametric eddy viscosity and beta

coefficients account for the effect of unresolved turbulence and waves on resolved scales and are suggested for use in non-eddy-resolving simulations of beta-plane turbulence (B. Galperin). A further statistical study of beta-plane turbulence, utilizing Lagrangian-based second order closure, shows that the westward phase speed of Rossby waves is significantly enhanced by random nonlinear interaction (Y. Kaneda).

Outcome of many random interactions within a flow can sometimes be expressed in terms of a large scale statistical dynamics. Numerical experiments with inviscid geostrophic dynamics show random initial conditions forming basin-scale mean flows. When viscosity is included, regions of homogenized vorticity form in the basin interior. Inclusion of topography modifies the resulting mean flows. Sufficiently steep topographic slopes will cause jet-like flows along isobaths, corresponding to the zonal jets in Rossby wave turbulence (G. Vallis). A tendency for random interactions to yield large scale mean flows suggests a method to parameterize eddies by directly introducing the statistical dynamical tendencies at large scale. Experiments with a coarsely resolved global ocean model show that various observed large scale flows, thought to be due to eddy interactions, can be obtained by such parameterization (G. Holloway).

CONCLUSIONS

Oceanographic phenomena differ widely in the characteristics of data, the governing physics, the underlying probability space, and the inferences sought. The statistical methods reflect this diversity. They range from signal detection, via parameter estimation and data assimilation, to stochastic or chaotic models. A common challenge in most applications is the huge amount of data and the complexities of the physics. The system space or the number of degrees of freedom must be reduced to a manageable size. EOF and POP analyses are used to reduce large and complex data sets; wavelet and inverse scattering transforms are applied to represent efficiently the underlying physics; idealized dynamics are employed to understand chaotic and statistical tendencies. In applying these and other statistical methods, physical oceanographers must modify existing methods and must invent new ones to deal with the peculiarities of oceanographic problems in a sensible way. The workshop witnessed the considerable progress that is being made at this task.

ACKNOWLEDGMENTS

We thank the participants of the workshop for their input to this report and for their permission to quote unpublished material. Phyllis Haines is thanked for expert editorial advice. Copies of the proceedings are available from Peter Müller, Department of Oceanography, University of Hawaii, 1000 Pope Road, MSB 307, Honolulu, HI 96822. The seventh 'Aha Huliko'a Hawaiian Winter Workshop was supported by Department of the Navy grant N00014-93-1-0156 issued by the Office of Naval Research. The U.S. government has a royalty-free license throughout the world in all copyrightable material contained herein.

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)	5 MONITORING ORGANIZATION REPORT NUMBER(S)
6a NAME OF PERFORMING ORGANIZATION School of Ocean and Earth Science and Technology 6b OFFICE SYMBOL (If applicable)	7a NAME OF MONITORING ORGANIZATION Office of Naval Research
6c ADDRESS (City, State, and ZIP Code) Department of Oceanography, Univ. of Hawaii 1000 Pope Road Honolulu, HI 96822	7b ADDRESS (City, State, and ZIP Code) Department of the Navy 800 No. Quincy Street Arlington, VIrginia 22217
8a NAME OF FUNDING / SPONSORING ORGANIZATION 8b OFFICE SYMBOL (If applicable)	9 PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER N00014-93-1-0156
8c ADDRESS (City, State, and ZIP Code)	10 SOURCE OF FUNDING NUMBERS PROGRAM ELEMENT NO: / PROJECT NO: 4222055-02 / TASK NO: / WORK UNIT ACCESSION NO:

11 TITLE (Include Security Classification)
Statistical Methods in Physical Oceanography

12 PERSONAL AUTHOR(S)
Muller, Peter and Henderson, Diane (eds.)

13a TYPE OF REPORT	13b TIME COVERED	14 DATE OF REPORT (Year, Month, Day)	15 PAGE COUNT
Workshop proceedings	FROM 12/1/92 TO 11/30/93	November 1993	528

16 SUPPLEMENTARY NOTATION
Proceedings, 'AHa Huliko'a, Hawaiian Winter Workshop, January 1993, Honolulu, Hawaii

17 COSATI CODES		
FIELD	GROUP	SUB-GROUP

18 SUBJECT TERMS (Continue on reverse if necessary and identify by block number)
oceanic observations, flow representation, statistical inference, data assimilation, chaos, statistical dynamics

19 ABSTRACT (Continue on reverse if necessary and identify by block number)

These proceedings contain the lectures given at the seventh 'Aha Huliko'a Hawaiian Winter Workshop on "Statical Methods in Physical Oceanography" and a meeting report. The lectures and the meeting report cover special probabilistic, statistical, and stochastic methods employed by physical oceanographers to analyze ocean data and formuate ocean models.

20 DISTRIBUTION / AVAILABILITY OF ABSTRACT	21 ABSTRACT SECURITY CLASSIFICATION
☒ UNCLASSIFIED/UNLIMITED ☐ SAME AS RPT ☐ DTIC USERS	Unclassified
22a NAME OF RESPONSIBLE INDIVIDUAL	22b TELEPHONE (Include Area Code) / 22c OFFICE SYMBOL

DD FORM 1473, 84 MAR — 83 APR edition may be used until exhausted. All other editions are obsolete.

* U.S. Government Printing Office: 1986—607-044