

INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps. Each original is also photographed in one exposure and is included in reduced form at the back of the book.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

UMI

A Bell & Howell Information Company
300 North Zeeb Road, Ann Arbor MI 48106-1346 USA
313/761-4700 800/521-0600

ASSIMILATION OF DATA INTO LIMITED-AREA
COASTAL MODELS

By
Michael Dowd

SUBMITTED IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY
AT
DALHOUSIE UNIVERSITY
HALIFAX, NOVA SCOTIA
JANUARY 1997

© Copyright by Michael Dowd, 1997



National Library
of Canada

Acquisitions and
Bibliographic Services

395 Wellington Street
Ottawa ON K1A 0N4
Canada

Bibliothèque nationale
du Canada

Acquisitions et
services bibliographiques

395, rue Wellington
Ottawa ON K1A 0N4
Canada

Your file Votre référence

Our file Notre référence

The author has granted a non-exclusive licence allowing the National Library of Canada to reproduce, loan, distribute or sell copies of this thesis in microform, paper or electronic formats.

L'auteur a accordé une licence non exclusive permettant à la Bibliothèque nationale du Canada de reproduire, prêter, distribuer ou vendre des copies de cette thèse sous la forme de microfiche/film, de reproduction sur papier ou sur format électronique.

The author retains ownership of the copyright in this thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without the author's permission.

L'auteur conserve la propriété du droit d'auteur qui protège cette thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation.

0-612-24773-2

Canada

DALHOUSIE UNIVERSITY

FACULTY OF GRADUATE STUDIES

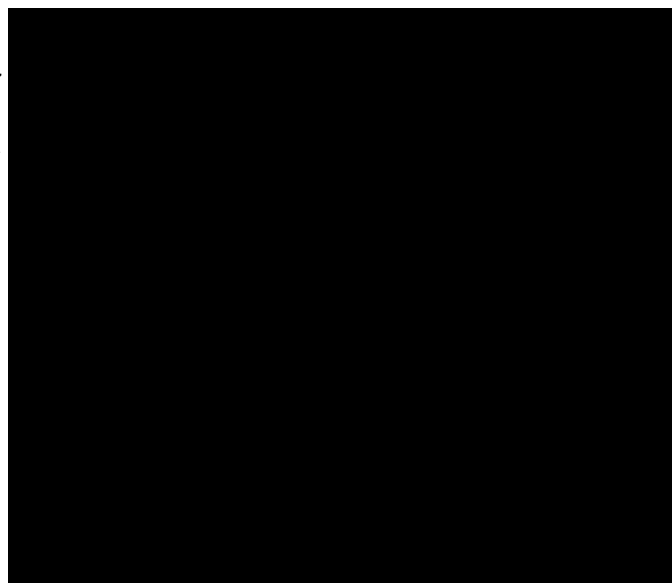
The undersigned hereby certify that they have read and recommend to the Faculty of Graduate Studies for acceptance a thesis entitled “Assimilation of Data into Limited-Area Coastal Models”

by Michael Dowd

in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

Dated: January 14, 1997

External Examiner
Research Supervisor
Examining Committee



DALHOUSIE UNIVERSITY

Date: **January 1997**

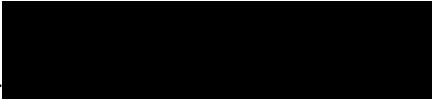
Author: **Michael Dowd**

Title: **Assimilation of Data into Limited-Area Coastal
Models**

Department: **Oceanography**

Degree: **Ph.D.** Convocation: **May** Year: **1997**

Permission is herewith granted to Dalhousie University to circulate and to have copied for non-commercial purposes, at its discretion, the above title upon the request of individuals or institutions.


Signature of Author

THE AUTHOR RESERVES OTHER PUBLICATION RIGHTS, AND NEITHER THE THESIS NOR EXTENSIVE EXTRACTS FROM IT MAY BE PRINTED OR OTHERWISE REPRODUCED WITHOUT THE AUTHOR'S WRITTEN PERMISSION.

THE AUTHOR ATTESTS THAT PERMISSION HAS BEEN OBTAINED FOR THE USE OF ANY COPYRIGHTED MATERIAL APPEARING IN THIS THESIS (OTHER THAN BRIEF EXCERPTS REQUIRING ONLY PROPER ACKNOWLEDGEMENT IN SCHOLARLY WRITING) AND THAT ALL SUCH USE IS CLEARLY ACKNOWLEDGED.

Contents

List of Tables	vii
List of Figures	viii
List of Symbols	xi
Acknowledgements	xiv
1 Introduction	1
1.1 Outline of Thesis	3
2 Background	7
2.1 General Inverse Methods and Regression	8
2.1.1 The Well-Posed Case	9
2.1.2 The Ill-Posed Case	11
2.1.3 Comments	14
2.2 Time Dependent Estimation Problems	15
2.3 Filtering and Forecasting	19
2.3.1 The Kalman Filter	19
2.3.2 Extensions and Applications	22
2.4 Smoothing	24
2.4.1 Regression Solution	24
2.4.2 The Kalman Smoother	25

2.4.3	The Adjoint Method	27
2.4.4	The Representer Solution	30
2.4.5	Application and Extensions	32
2.5	Summary	33
3	Extraction of Tidal Streams from a Ship-Borne ADCP	36
3.1	Observations	38
3.2	Tidal Model	46
3.2.1	Governing Equations	46
3.2.2	Discrete Form	49
3.3	Inverse Analysis	51
3.3.1	Comparing Model to Data	52
3.3.2	Regression Solution	54
3.3.3	Extensions	55
3.4	Application	56
3.4.1	Sensitivity Analysis	64
3.5	Summary and Discussion	66
4	Forecasting Coastal Circulation using an Approximate Kalman Filter	71
4.1	Background	74
4.2	Modal Representation	77
4.2.1	Dynamics	77
4.2.2	Interpretation of the Modes	79
4.2.3	Kalman Filter Equations	81
4.3	Selection of the Modes	82
4.3.1	Observability and Controllability	84
4.3.2	Energetics	85
4.4	Application	86
4.5	Discussion and Conclusions	101

5	Determining Dynamically Consistent Density and Velocity	108
5.1	Dynamics	111
5.2	Estimation	117
5.2.1	Summary	121
5.3	A Strong Constraint Approach	122
5.3.1	Model Formulation	123
5.3.2	Assimilation	124
5.3.3	Application	126
5.4	Discussion and Conclusions	142
6	Concluding Remarks	146
6.1	Suboptimal Data Assimilation	147
6.2	Future Directions	149
A	Least-Squares Regression	152
B	Probabilistic Approach	154
B.1	Filtering	154
B.1.1	The Prediction Process	155
B.1.2	The Measurement Process	155
B.1.3	The General Filtering Solution	156
B.2	Smoothing	157

List of Tables

- 3.1 Variance partitioning of current meter and ADCP velocity. 46
- 3.2 Definition of selected matrices and their dimensions. 52

List of Figures

2.1	Types of data assimilation problems.	18
3.1	Map of Scotian Shelf and model domain	40
3.2	Time series of current meter velocities.	41
3.3	Time series of depth averaged ADCP velocity.	43
3.4	Power spectra of depth averaged ADCP velocity.	44
3.5	Complex demodulation of ADCP velocity.	45
3.6	Vector plots of ADCP velocity.	47
3.7	Tidal ellipse maps estimated from ADCP.	59
3.8	Tidal ellipses from current meters.	60
3.9	Tidal time series from ADCP.	61
3.10	Vector plots of tidal residual velocity from ADCP.	63
3.11	Comparison of ADCP derived residual circulation with diagnostic calculation.	65
3.12	Sensitivity of ADCP tidal ellipses to grid resolution.	67
4.1	Model domain and location of observation array.	88
4.2	Example of propagating mode.	90
4.3	Scatter plot of decay time versus oscillation period.	91
4.4	Time series of Louisbourg sea level and Sable Island wind stress during CASP.	93
4.5	Controllability and observability of the modes.	95
4.6	Selection criteria for modes.	96

4.7	Approximate Kalman filter results at mooring 1.	98
4.8	Convergence of the Kalman gain matrix.	99
4.9	Approximate Kalman filter results for moving observation array. . . .	100
4.10	Comparison of actual versus estimated flow field.	102
4.11	Comparison of actual versus estimated forcing.	103
5.1	Geometry of the idealized coastal model and observing locations. . . .	128
5.2	Baseline results for the barotropically dominated case.	131
5.3	Baseline results for the baroclinically dominated case.	132
5.4	Sample diagnostic calculation.	134
5.5	Actual versus diagnostic JEBAR.	135
5.6	Point observations of density.	137
5.7	Performance diagnostics for the barotropic case.	138
5.8	Contour plots of the cost function.	140
5.9	Performance diagnostics for baroclinic case.	141

Abstract

This thesis investigates assimilation of oceanographic data into limited-area coastal circulation models. The approach taken analyses coastal data using simple, process-oriented ocean dynamics which isolate the essential physics of the problem. It is first demonstrated that oceanographic data assimilation can be treated in the framework of regression, and its extension to the time dependent case. Commonly used techniques are reviewed in this context. Three studies are then carried out covering a range of oceanographic data and dynamics: (1) A statistical-dynamical method is proposed to extract the barotropic tide from a ship-borne acoustic Doppler current profiler (ADCP). A limited-area tidal model, posed in the frequency domain, is fit to the time-space series of ADCP velocity using a boundary control approach. The procedure is applied to ship ADCP data from the Western Bank region of the Scotian Shelf. ADCP derived tides were in good agreement with those from fixed current meters, and the tidal residual was also found to be consistent with a diagnostic calculation of the flow. (2) An approximate Kalman filter is derived for forecasting coastal circulation. The original ocean model is reformulated in terms of its dynamical modes, and a reduced model is obtained using a subset of the modes preferentially excited by forcing. This retains the dynamics necessary for model forecasts and error propagation, yet allows the Kalman filter to be efficiently implemented. The approximate filter was demonstrated using a prototype shallow water model of the Scotian Shelf and focused on the variability associated with wind and boundary driven flows. (3) The estimation of circulation from density data is investigated. In particular, the consequences of including a prognostic density equation, together with the usual set of diagnostic (thermal wind) equations, are considered. The advantage of this approach is that dynamically consistent density and velocity estimates can be obtained from hydrographic data. A unique limit, wherein the dynamics are treated as a strong constraint in the assimilation, is explored using an idealized coastal model. Buoyancy fluxes across the open boundaries into the model domain are determined from interior point observations of the density field. Numerical experiments are performed to illustrate the issues arising in this joint estimation problem. Application of the method to realistic ocean models is discussed.

List of Symbols

English Letters

A	matrix of autoregressive coefficients
b	vector of unknown parameters
$\hat{\mathbf{b}}$	estimate for b
c	representers
C	matrix of representers
d	nonlinear dynamics operator
D	dynamics matrix
\mathbf{D}^+	generalized inverse of D
e	error vector
\mathbf{e}^f	forecast error
\mathbf{e}^o	observation error
\mathbf{e}^m	model error (system noise)
f	Coriolis parameter
g	acceleration of gravity
G	projection matrix for \mathbf{e}^m on to x
h	bottom depth
h	nonlinear observation operator
H	observation matrix
I	identity matrix
J	general cost function
J	Jacobian operator
$\bar{\mathbf{k}}$	unit vertical vector
K	horizontal diffusivity
K	Kalman gain matrix
L	Lagrange function
M	error covariance matrix associated with $\bar{\mathbf{x}}$
$\bar{\mathbf{n}}$	unit outward normal vector
P	Pressure
P	error covariance matrix associated with $\hat{\mathbf{x}}$
Q	covariance matrix \mathbf{e}^m

r	friction coefficient
\mathbf{R}	covariance matrix for \mathbf{e}^o
s	selection criterion
\mathbf{s}	spatial interpolator
\mathbf{S}	regularization matrix
\mathbf{x}	ocean state vector
$\bar{\mathbf{x}}$	model forecast of ocean state
$\hat{\mathbf{x}}$	assimilation estimate of ocean state
\bar{u}	horizontal velocity
\bar{u}_b	barotropic horizontal velocity
\bar{u}_s	baroclinic horizontal velocity
\bar{U}	horizontal transport
\bar{U}_b	barotropic horizontal transport
\bar{U}_s	baroclinic horizontal transport
w	vertical velocity
z	single observation
\mathbf{z}	observation vector
$\hat{\mathbf{z}}$	model estimates of observations

Greek Letters

α	modal coefficients
β	representer coefficients
γ	Lagrange multiplier
Γ	spectral density
ε	density anomaly
η	sea surface elevation
θ	resonant frequency of modes
κ	weighting term
λ	vector of eigenvalues
Λ	diagonal matrix of eigenvalues
ρ	density
σ	square root of variance
Σ	covariance matrix
$\bar{\tau}$	stress term
\mathbf{v}	eigenvector used in singular value decomposition
Υ	matrix with \mathbf{v} as columns
ϕ	eigenvector
Φ	matrix with ϕ as columns
Ψ	transport streamfunction
ω	frequency

Other

∇	horizontal gradient operator
∂	partial differential
\Re	real part of a complex number
\Im	imaginary part of a complex number
i	square root of -1

Acknowledgements

I would foremost like to thank my supervisor, Keith Thompson, for his exemplary guidance during my years at Dalhousie. Thanks is also due to my thesis committee, Barry Ruddick, Bruce Smith, Peter Smith, and Dan Wright, as well as my external examiner, Carlisle Thacker, for a thorough and detailed reading of my thesis, and their many insightful comments on my work. I am also grateful to Jinyu Sheng, who provided a great deal of help and useful discussion. My fellow students Karin Bryan, Mark Buehner, Josko Bobanovic, Jim Burke, Phil MacAulay, Doug Mercer, Conrad Pilditch, Jacquie Witte, and Zhigang Xu, provided much support and assistance. The technical support provided by Jackie Hurst and Steve Matheson throughout was also much appreciated.

Chapter 1

Introduction

Studying ocean circulation proceeds through the complementary techniques of theory, observation, and numerical modelling. In recent years, insights from statistical estimation theory have been applied to the oceanographic problem of combining models and data. The blending of these two components has come to be known as data assimilation. Its goal in oceanography is to extract the maximum amount of information from available data sets using mathematical models based on an acquired knowledge of the underlying physical processes.

A distinguishing feature of oceanographic data are their irregular and asynchronous distribution, as well as the diversity of data types available. Optimal interpolation has been perhaps the most widely used analysis method for statistically mapping irregularly spaced data (Daley 1991). The techniques of principal oscillation patterns (Hasselmann 1988) and extended empirical orthogonal function analysis (Brillinger 1981) further offer the ability to account for temporal as well as spatial coherence but, like optimal interpolation, are essentially empirical methods. To interpolate and extrapolate data in a manner consistent with our knowledge of geophysical fluid dynamics requires combining ocean dynamics and data.

Interest in oceanographic data assimilation has increased dramatically over the last decade. New observation technologies, such as satellite altimetry and acoustic tomography, began to offer the promise of large volumes of high quality oceanographic

data, adding to the already extensive databases of ocean hydrography. It was soon recognized that traditional data analysis and mapping techniques were not well suited for making circulation estimates from these data sets. At the same time, advances in ocean modelling and increases in computer power made possible realistic models of ocean circulation. Thus was raised the possibility of data analysis techniques based on numerical ocean models.

Optimal control methods were developed in the 1950s and 1960s as a practical means to combine data with models and thereby control the evolution of dynamic systems (see Bryson and Ho 1969). These techniques provide the basis for both oceanographic and atmospheric data assimilation. Sasaki (1970) was the first to suggest applying these variational techniques to numerical weather prediction, but little interest was shown at the time. LeDimet and Talagrand (1986) revisited the idea and proposed its use for the initialization of operational weather forecasting models. In oceanography, the pioneering study of Wunsch (1978) introduced inverse techniques in the context of determining deep reference velocities from hydrographic section data. Later, Thacker and Long (1988) demonstrated the practicality of these control methods in fitting time dependent ocean models to data. Since this time, a great deal of work has been carried out in atmospheric and oceanographic data assimilation (for example, see the textbooks of Daley 1991 and Bennett 1992).

Although data assimilation in oceanography and meteorology bear many similarities, there are also some key distinctions. The baroclinic Rossby radius in the ocean is 10-30 km, whereas in the atmosphere it is typically 1000 km (Gill 1982, section 7.5). Clearly, the important scales of motion which must be resolved in the ocean are much smaller than those found in the atmosphere and in this regard the ocean is grossly under-sampled (e.g. Ghil and Malanotte-Rizzoli 1992). Complex geometry also strongly influences ocean dynamics. Coastlines and bottom topography have a wide range of dynamical effects on circulation. As a result, regional or basin scale models of limited horizontal extent are frequently used with the influence of the remainder of the ocean represented by flows across their open lateral boundaries.

Another important difference is that the development of atmospheric data assimilation has been driven in large part by the operational need for weather forecasts. In oceanography, the main emphasis to date has been on the use of data assimilation as a research tool.

One area in oceanography currently receiving increased attention is the operational prediction of coastal circulation. Accurate estimates of circulation are required to address such coastal zone problems as storm surge forecasting, predicting oil spill and iceberg trajectories, and assessing the impact of point source pollutants. A number of efforts are presently underway to develop operational nowcasting and forecasting systems for the coastal ocean (e.g. Heemink and Van-Stijn 1993, Griffin and Thompson 1995, Aikman *et al.* 1996). These must synthesize a wide range of data from sources including coastal tide gauges, current meters, CTDs, drifters, ship-borne acoustic Doppler current profilers, and remotely sensed images. As a result, data assimilation forms an integral part of these schemes. It is the operational prediction of coastal flow fields which largely motivates the work carried out in this thesis.

1.1 Outline of Thesis

The objective of this thesis is to investigate the assimilation of data into limited-area coastal models. A process-oriented approach is taken whereby oceanographic data are analyzed using relatively simple models. This allows the important features of the dynamics to be isolated and their role in the estimation problem studied in more detail. This approach is intended to lead to practical techniques for assimilating oceanographic data in more complex coastal circulation models.

In this regard, a notable characteristic of the coastal region that must be considered is the range of timescales over which important processes act, as well as the variety of data sources. A complete coastal circulation model that simultaneously considers these timescales and data types could be used as the basis for an assimilation exercise. Instead, we have chosen to decompose the circulation so that

the variability associated with tides (days), wind forcing (days-weeks) and buoyancy fluxes (weeks-years) are each treated separately. Similarly, data sources have been treated as distinct and careful consideration given to the nature and properties of these data, as well as the appropriateness of the analysis scheme used. It is felt that such a treatment provides a necessary first step in the synthesis of such data with more complex coastal circulation models.

Another issue which inevitably arises when using limited-area models is the treatment of the open boundaries. Important forcing occurs through these boundaries since they must account for the effect of remotely generated disturbances propagating into the model domain. It is rare that direct observations of the boundary state are available, and we must therefore rely on information contained in the interior observations about these remote effects. The inference of boundary conditions from interior data is an issue addressed throughout this thesis.

The studies which are carried out in this thesis include: extracting the tidal signal from irregularly distributed velocity measurements; capturing the important components of the synoptic variability due to wind and remote forcing for the purpose of forecasting coastal circulation; and determining the quasi-steady circulation from density data. These topics cover a range of timescales, use both linear and nonlinear models, and consider data assimilation using discrete inverse methods, the Kalman filter, as well as nonlinear optimization. A detailed outline of the thesis is as follows:

- Chapter 2 provides a general overview of the techniques commonly used in oceanographic data assimilation. These are unified under the common framework of regression analysis, which is extended to cover the case of time dependent ocean models and observations. The chapter includes general inverse methods, the Kalman filter and smoother as well as optimal control and representer methods. Linear theory is used to derive and demonstrate the optimality properties after which nonlinear extensions are discussed. Two appendices are provided as complements to this chapter. Appendix A briefly covers least squares regression, and Appendix B provides a discussion of the filtering and

smoothing problems based on a probabilistic approach.

- In Chapter 3, extraction of the barotropic tide from the time-space series of horizontal velocity obtained by a ship mounted acoustic Doppler current profiler (ADCP) is investigated. It is based on fitting a limited area tidal model, posed in the frequency domain, to the ADCP record in order to obtain optimal tidal boundary conditions. An application of the method using ship ADCP data from the Western Bank region of the Scotian Shelf is also included.
- Chapter 4 considers the role of wind and boundary forced circulation on the development of an operational forecasting system based on the Kalman filter. In particular, an approximate, reduced dimension Kalman filter is proposed. It is based on reformulating an ocean model in terms of its dynamical modes. A subset of the modes preferentially excited by the forcing is then chosen as the basis for a reduced ocean model. Solving the Kalman filter equations using this reduced model retains the important components of the dynamics necessary for model forecasts and error propagation, yet allows for a computationally efficient means to implement this data assimilation scheme. A variety of tests are carried out based on a prototype model of the Scotian Shelf using wind and boundary forcing as well as fixed and moving observation arrays.
- Chapter 5 investigates the steady circulation associated with density variations, with a focus on coastal regions. The general problem of determining circulation from density data is reviewed and the consequences of including a prognostic density equation together with the usual set of diagnostic equations is investigated. An inverse formulation of the problem is presented and a particular example undertaken based on a data assimilation scheme which treats the dynamics, including the density equation, as a strong constraint. Experiments are carried out for this case using an idealized, limited-area shelf model in order to investigate the joint estimation of dynamically consistent density and velocity.
- Chapter 6 remarks on some important issues arising from the work of this thesis.

In particular, the need for suboptimal data assimilation methods suitable for operational forecasting of circulation is emphasized and offered as an area for future research.

Chapter 2

Background

Practical applications of oceanographic data assimilation are concerned with combining data and dynamics in order to produce optimal estimates of ocean circulation and hydrography. Another purpose is to test hypotheses about the dynamical origins of these fields. An apparently large number of approaches to this problem are available. However, this variety arises mainly as a result of whether the estimation problem is treated as: (i) continuous or discrete in its time-space coordinates, (ii) deterministic or stochastic, (iii) linear or nonlinear, or (iv) with dynamics acting as a weak or strong constraint. Each treatment has advantages and disadvantages in discussing the theory, issues, and techniques of oceanographic data assimilation.

In this chapter, the basic concepts of oceanographic data assimilation are discussed in the context of regression theory and its generalization to time dependent models. Regression is a widely used and well developed technique concerned with fitting models to data. It provides a unifying framework for understanding the methods commonly used in oceanographic data assimilation. Being discrete in nature, regression is both consistent with, and complementary to, the use of numerical ocean models. It can be approached from a stochastic or deterministic standpoint and has nonlinear extensions. The major drawback is that it does not deal explicitly with time-stepping models but does provide a basis on which to develop the appropriate techniques for these cases.

The major sources for the material in this chapter include the textbooks on regression (Sen and Srivastava, 1981), control theory (Bryson and Ho 1969), filtering theory (Jazwinski 1970), and optimal estimation (Gelb 1974). Specific to atmospheric and oceanographic data assimilation are the books of Bennett (1992) and Daley (1991) and the review article of Ghil and Malanotte-Rizzoli (1991).

The chapter is structured as follows. Section 2.1 discusses general inverse methods and regression in a linear framework with an emphasis on the important aspects common to all estimation problems. Section 2.2 then turns to oceanographic data assimilation. A general representation of time dependent ocean models and observations is introduced and the problem of blending these two sources of information is discussed. In Section 2.3, the Kalman filter is derived in terms of regression. Optimal (Kalman) smoothing is then developed in Section 2.4. Adjoint based smoothing and representer techniques are also derived and discussed. Finally, Section 2.5 provides a brief summary. Appendices A and B are also included in the thesis to complement this chapter.

2.1 General Inverse Methods and Regression

Suppose we wish to estimate a set of unknown, or uncertain, oceanographic quantities from data. These might, for example, take the form of model parameters, initial or boundary conditions, or the ocean state itself. Consider the following regression equation

$$\mathbf{z} = \mathbf{D}\mathbf{b} + \mathbf{e}. \quad (2.1)$$

Here, the $n \times 1$ vector \mathbf{z} represents data. The unknown quantities to be estimated are given by the $p \times 1$ vector \mathbf{b} . The $n \times p$ matrix \mathbf{D} represents the model which (linearly) relates the unknown \mathbf{b} to the data \mathbf{z} . In oceanographic data assimilation, the matrix \mathbf{D} includes ocean dynamics (determined from a finite difference form of the governing equations) as well as a mapping to the observing locations. An error term \mathbf{e} is included to reflect the fact that the model, for a variety of reasons, will not

reproduce the observations exactly. Note that no time index has been introduced for either the model or the data, although it could be included implicitly in the general development of regression given here (see Section 2.4).

The goal of the assimilation procedure is to choose an estimate of the unknown quantity \mathbf{b} such that the model estimates $\hat{\mathbf{z}}$ are a 'best-fit' to the actual observations \mathbf{z} . This problem is naturally approached in the framework of regression analysis. The general solution to the regression problem takes the form (see Appendix A)

$$\hat{\mathbf{b}} = \mathbf{D}^+ \mathbf{z}, \quad (2.2)$$

where $\hat{\mathbf{b}}$ is an estimate for \mathbf{b} and \mathbf{D}^+ is the generalized inverse of \mathbf{D} .

Two desirable properties (Jackson 1972) of the generalized inverse are

1. $\mathbf{D}\mathbf{D}^+$ should act as $\mathbf{I}_{n \times n}$. By pre-multiplying both sides of (2.2) by \mathbf{D} , it is seen that this condition may be interpreted as an overall measure of how well the model predictions $\hat{\mathbf{z}}$ fits to the data \mathbf{z} .
2. $\mathbf{D}^+\mathbf{D}$ should act as $\mathbf{I}_{p \times p}$. Pre-multiplying (2.1) by \mathbf{D}^+ and using (2.2), shows that this condition provides a comparison between the estimate $\hat{\mathbf{b}}$ and its true value \mathbf{b} .

These properties provide a general basis on which to construct a specific inverse and to assess the problem of fitting models to data.

2.1.1 The Well-Posed Case

Consider the case in which the number of observations equals the number of unknowns ($n = p$) and assume that the matrix \mathbf{D} is of full rank (i.e. has no zero eigenvalues). The unknown \mathbf{b} are estimated such that the model predictions $\hat{\mathbf{z}}$ match exactly the data \mathbf{z} . The concept of an error given in (2.1) is of little utility in this case, although it is undoubtedly present in both the observations and the model. The solution is in the sense of the usual definition of an inverse,

$$\mathbf{D}^+ = \mathbf{D}^{-1},$$

and the properties 1) and 2) are satisfied by definition. The important issue here is that while a unique solution has been obtained, the variance of $\hat{\mathbf{b}}$ may prove to be unacceptably large. This may occur if one or more of the eigenvalues of \mathbf{D} are near zero and/or if the error \mathbf{e} is large. In this case, some of the techniques illustrated in the next section may prove useful for reducing the variance of $\hat{\mathbf{b}}$ at the expense of introducing bias into the estimate.

The overdetermined case is encountered when the number of observations n is greater than the number of unknowns p (we assume $\mathbf{D}^T\mathbf{D}$ is of full rank, where superscript T denotes transpose). This implies that a unique solution could be obtained with fewer observations than are available. Since the observations (and model) are uncertain, the extra information is absorbed into the error term \mathbf{e} . The squared error

$$\begin{aligned} J &= \mathbf{e}^T\mathbf{e} \\ &= (\mathbf{z} - \mathbf{D}\mathbf{b})^T(\mathbf{z} - \mathbf{D}\mathbf{b}) \end{aligned} \tag{2.3}$$

is usually chosen as the appropriate measure of model fit, and the unknown \mathbf{b} chosen such that J is minimized. Differentiating J with respect to \mathbf{b} and setting the result to zero yields the inverse

$$\mathbf{D}^+ = (\mathbf{D}^T\mathbf{D})^{-1}\mathbf{D}^T.$$

If $\mathbf{e} \sim N(0, \sigma^2\mathbf{I})$, where \mathbf{I} is the identity matrix and σ^2 the variance. then $\hat{\mathbf{b}}$ is the maximum likelihood estimate of its true value \mathbf{b} . In the more general case of a positive definite error covariance matrix, i.e. $\mathbf{e} \sim N(0, \Sigma)$, the inverse is

$$\mathbf{D}^+ = (\mathbf{D}^T\mathbf{\Sigma}^{-1}\mathbf{D})^{-1}\mathbf{D}^T\mathbf{\Sigma}^{-1}.$$

which leads to the generalized least squares estimates for $\hat{\mathbf{b}}$. Again, these satisfy property 1) and match property 2) in the sense that they are minimum variance unbiased estimates.

2.1.2 The Ill-Posed Case

We now turn to the situation where the number of observations are less than the number of unknowns and an infinity of solutions for $\hat{\mathbf{b}}$ exists. This is the under-determined problem of rank-deficient regression and a situation often encountered in oceanographic data assimilation. Choosing a particular inverse \mathbf{D}^+ requires introducing additional (prior) information into the problem. Two approaches for determining a generalized inverse are given below.

The first procedure explicitly identifies the parts of $\hat{\mathbf{b}}$ which are resolvable for a given \mathbf{D} and obtains a unique solution on this basis. Singular value decomposition allows any $n \times p$ matrix \mathbf{D} to be expressed as

$$\mathbf{D} = \mathbf{\Upsilon} \mathbf{\Lambda} \mathbf{\Phi}^T. \quad (2.4)$$

The elements of $\mathbf{\Upsilon}$, $\mathbf{\Phi}$ and $\mathbf{\Lambda}$ are determined through generalized eigenvector analysis of an $n \times p$ matrix (Lanczos 1961) as follows. Define

$$\mathbf{D}\phi_j = \lambda_j \mathbf{v}_j \quad (2.5)$$

$$\mathbf{D}^T \mathbf{v}_i = \lambda_i \phi_i \quad (2.6)$$

or

$$\mathbf{D}^T \mathbf{D} \phi_j = \lambda_j^2 \phi_j \quad j = 1, \dots, p \quad (2.7)$$

$$\mathbf{D} \mathbf{D}^T \mathbf{v}_i = \lambda_i^2 \mathbf{v}_i \quad i = 1, \dots, n. \quad (2.8)$$

The $l \times l$ matrix $\mathbf{\Lambda}$ is a diagonal matrix of the non-zero λ which are common to the two sets of eigenvalues, where l is less than or equal to the minimum of p or n . The columns of $\mathbf{\Upsilon}$ ($n \times l$) and $\mathbf{\Phi}$ ($p \times l$) are the eigenvectors \mathbf{v} and ϕ associated with non-zero singular values in $\mathbf{\Lambda}$.

The inverse associated with the singular value decomposition is

$$\mathbf{D}^+ = \mathbf{\Phi} \mathbf{\Lambda}^{-1} \mathbf{\Upsilon}^T, \quad (2.9)$$

which follows from the orthonormal property of Υ and Φ . The eigenvectors \mathbf{v}_i and ϕ_j provide bases for the vector spaces spanned by \mathbf{z} and \mathbf{b} , respectively. The above inverse has been constructed such that their associated null spaces (represented by zero eigenvalues) are explicitly identified and then neglected from any further involvement in the estimation procedure. As a result, the estimate for $\hat{\mathbf{b}}$ has minimum norm $(\hat{\mathbf{b}}^T \hat{\mathbf{b}})^{1/2}$ (this is made more clear below). The inverse also provides the closest fit, in a least squares sense, to the identity matrices in both properties 1) and 2). The singular value decomposition may then be said to choose from the infinity of possible solutions, a least squares inverse which yields the estimate for $\hat{\mathbf{b}}$ which has minimum length.

Next, we consider an alternate, but related, procedure for treating undetermined systems based on introducing prior information, or regularization, into the problem. A justification of this procedure is given in this quotation from Jackson (1972):

Let us examine in more detail the case of the strongly underdetermined system. This case will include those problems (which are) the result of discretizing a continuous relationship between a known function and an unknown function, because we may handle only finite amounts of data, yet we would in principle like to know an infinitude of details about the unknown function. *A wise procedure is to use more parameters to describe the unknowns than are likely to be uniquely determined by the data.* One may then form a family of inverses, compare the tradeoff between resolution and variance for this family, and select that inverse which is most appropriate for interpreting the solution.

In oceanography, the model \mathbf{D} being fit to data results from the discretization of a set of governing partial differential equations. These relations, apart from initial and boundary conditions and certain internal parameterizations, are often considered as essentially known. To estimate the unknown (and underdetermined) parts of the function, we can introduce prior information based on our physical understanding of the problem in order to select a unique solution.

Consider augmenting the cost function (2.3), which measures the squared error of the observation/model misfit, with a regularization term accounting for prior information, for example the biasing of the solution towards smoothness. This cost function could take the form

$$J = (\mathbf{z} - \mathbf{D}\mathbf{b})^T(\mathbf{z} - \mathbf{D}\mathbf{b}) + \kappa^2 \mathbf{b}^T \mathbf{S} \mathbf{b}, \quad (2.10)$$

where the last term on the right-hand-side constitutes the regularization term and κ^2 designates its relative strength. Minimizing J by differentiating with respect to \mathbf{b} and setting the result to zero yields

$$(\mathbf{D}^T \mathbf{D} + \kappa^2 \mathbf{S}) \mathbf{b} = \mathbf{D}^T \mathbf{z}. \quad (2.11)$$

Solving for \mathbf{b} leads to the inverse

$$\mathbf{D}^+ = (\mathbf{D}^T \mathbf{D} + \kappa^2 \mathbf{S})^{-1} \mathbf{D}^T, \quad (2.12)$$

whose existence depends on $\mathbf{D}^T \mathbf{D} + \kappa^2 \mathbf{S}$ being full rank. The philosophy behind this approach is that by including the prior information on \mathbf{b} , the null space (zero eigenvalues) associated with $\mathbf{D}^T \mathbf{D}$ can be resolved. If \mathbf{S} is the identity matrix, this method corresponds to ridge regression and the problem of choosing κ can be dealt with through generalized cross validation (Craven and Wahba 1979, McIntosh and Veronis 1993).

In the limit where κ^2 tends to zero, the inverse obtained by the regularization approach reduces to that found by the singular value decomposition. To demonstrate this, let $\kappa^2 = 0$ in (2.10). Minimizing J leads to an equation identical to (2.11) except that the term involving κ^2 is now removed. The matrix $\mathbf{D}^T \mathbf{D}$ cannot be inverted directly and thus \mathbf{D}^+ in (2.12) is not defined. However, using the singular value decomposition for \mathbf{D} given in (2.4) and premultiplying (2.11) by $\Phi \Lambda^{-2} \Phi^T$ yields the inverse \mathbf{D}^+ found in (2.9). That is, the inverse associated with the singular value decomposition is identical to that found by minimizing J in (2.10) when the weight $\kappa^2 = 0$. This demonstrates that the singular value decomposition solution

yields a solution vector \mathbf{b} of minimum length. Introducing prior information into the problem biases the solution to an extent dependent on the weight κ^2 and form \mathbf{S} of the regularization.

2.1.3 Comments

General inverse theory provides a framework in which to consider issues important to oceanographic data assimilation. Inverse methods are well established in geophysics (Backus and Gilbert 1967, 1968, Wiggins 1972) and reviewed in Tarantola (1987). Few direct applications of these techniques are found in the oceanographic literature (e.g. Wunsch 1978), mostly because straightforward usage of the matrix algebra is feasible only if the dimension of the matrices are kept manageable. In practice, this has limited oceanographic application to linear problems based on steady-state models. For the case of nonlinear models, generalized inverses can be determined via the minimization procedures used in nonlinear regression (Sen and Srivastava 1990).

The generalized inverse methods outlined in this section are not ideally suited for situations involving time dependent models and observations. The time dependent case could be absorbed into the framework given here (see Section 2.4). However this makes either the dimensionality of the matrix \mathbf{D} very large, or its form quite complicated. As a consequence, most efforts in oceanography have focused on data assimilation procedures which take advantage of the special structure imposed on the estimation problem by time-stepping models and these will be discussed for the remainder of this chapter. Nonetheless, it is important to emphasize that the data assimilation problem is still the problem of determining a generalized inverse and issues raised in the previous section are fundamental to the problem of fitting time dependent models to data.

2.2 Time Dependent Estimation Problems

In this section, ocean dynamics are introduced in the context of a generic, time-stepping numerical ocean model represented as a stochastic difference equation. Observations are related to these model variables through a measurement equation. These two components form the basis for fitting time dependent models to data.

Geophysical fluid dynamics provides a continuous form of the governing equations for ocean circulation (e.g. Pedlosky 1979). These have approximate counterparts in discrete time-space. Assume that a numerically stable discrete form of these governing equations exists and is given by

$$\mathbf{x}_t = \mathbf{d}(\mathbf{x}_{t-1}) + \mathbf{G}_t \mathbf{e}_t^m, \quad (2.13)$$

where \mathbf{x}_t is a column vector which contains the prognostic variables of the model defined on a spatial lattice at a given time t . The \mathbf{d} operator represents a discretized form of the dynamic equations. The system noise, or model error, is represented by \mathbf{e}_t^m and is assumed additive. It includes such processes as forcing, model uncertainties resulting from neglected dynamics, and any errors in the numerical representation of the governing partial differential equations (since it includes both forcing and model errors it may be viewed as a mixture of deterministic and stochastic parts). This system noise is projected on the ocean state \mathbf{x}_t using the matrix \mathbf{G}_t . The form of (2.13) may be viewed as a general representation of a nonlinear, time-stepping, numerical ocean model which requires initial and boundary conditions to be integrated.

Available oceanographic observations \mathbf{z}_t are related to the ocean state \mathbf{x}_t through an observation equation given by

$$\mathbf{z}_t = \mathbf{h}(\mathbf{x}_t, t) + \mathbf{e}_t^o, \quad (2.14)$$

where \mathbf{h} represents the operator which maps from model state to the observations. It includes a conversion from the model variables to the measured variables (for example, velocity measurements used in a vorticity based model), and an interpolation of the observation locations to the appropriate points on the model grid. The observation

operator \mathbf{h} is considered a function of time, mainly to allow for the possibility of an observation array which changes over time. The observation error term \mathbf{e}_t^o accounts for the fact that the model cannot reproduce the observations exactly. It includes any errors in the conversion and interpolation operations contained in \mathbf{h} , as well as instrument noise.

In developing the theory of data assimilation in this chapter, the linear versions of (2.13) and (2.14) are used. Suppose that a reference trajectory for the state $\mathbf{x}_t = \mathbf{x}_t^*$, $t = 1, \dots, N$ has been determined for the system (2.13) by specifying best guesses of the initial state and the forcing. The linearized versions of (2.13) and (2.14) are then

$$\mathbf{x}_t = \mathbf{D}_t \mathbf{x}_{t-1} + \mathbf{G}_t \mathbf{e}_t^m \quad (2.15)$$

$$\mathbf{z}_t = \mathbf{H}_t \mathbf{x}_t + \mathbf{e}_t^o, \quad (2.16)$$

where \mathbf{x}_t , \mathbf{z}_t , \mathbf{e}_t^m , and \mathbf{e}_t^o now represent deviations from their reference trajectory values. The matrices \mathbf{D}_t and \mathbf{H}_t are determined through the linearization of vector functions \mathbf{d} and \mathbf{h} about \mathbf{x}_t^* , i.e.,

$$D_{ij t} = \left. \frac{\partial d_i}{\partial x_j} \right|_{\mathbf{x}_t = \mathbf{x}_t^*}, \quad H_{ij t} = \left. \frac{\partial h_i}{\partial x_j} \right|_{\mathbf{x}_t = \mathbf{x}_t^*}, \quad (2.17)$$

where the subscripts i, j represent the elements of \mathbf{D} and \mathbf{H} . This approximation is valid for small deviations about the reference state. Time subscripts have been included in \mathbf{D}_t (and \mathbf{H}_t) to allow for the fact that the reference trajectory varies over time.

The overall goal of the data assimilation problem is to combine the model estimates (2.13) with the observations (2.14) to produce a better estimate of the overall ocean state than could be obtained with either models or data alone. Data assimilation methods are commonly divided into two categories: filtering and smoothing. Filtering methods use available information (i.e. observations and model predictions) over an interval $(0, N)$ to produce an estimate of the ocean state at the end of that interval (at time $t = N$). These methods are sequential in the sense that recursion relations are available for updating the state based on the previous estimates obtained.

Smoothing methods, on the other hand, use information over an interval $(0, N)$ to produce estimates at any, or possibly all, intermediate times $t = 0, \dots, N$. Figure 2.1 presents a schematic diagram contrasting the filtering and smoothing problems.

A probabilistic framework provides, perhaps, the most complete solution to the filtering and smoothing problems and it is worthwhile to note a few important points. Consider the conditional probability density function $p(\mathbf{x}_t | \mathbf{z}_0, \dots, \mathbf{z}_N)$. For $t = N$, a complete solution to the filtering problem is provided by knowledge of this density function. A similar statement applies for the smoothing problem for $t = 0, \dots, N$. Clearly then, the filtering problem is a special case of the smoothing problem. Elementary probability theory allows the conditional probability density function governing the ocean state to be determined from the dynamics (2.13) and the observations (2.14). The general solutions to the filtering and smoothing problems are outlined in detail in Appendix B.

While the probabilistic approach offers a general, and conceptually straightforward, description of the data assimilation problem, it is difficult to implement in practice. Application to realistic oceanographic problems is hampered by incomplete knowledge of the input probability density functions as well as the computational difficulty of transforming the full probability density functions using the ocean dynamics. Monte Carlo methods may offer some potential for dealing with these problems (Evensen 1994); however, it is usually argued (e.g. Jazwinski 1970) that a more appropriate way to treat the data assimilation problem is to use statistical estimation theory. In this case, we are primarily concerned with the first and second central moments (mean and covariance) of the probability density functions. Instead of maximum likelihood estimates, we consider estimates which minimize model and observational errors while taking into account appropriate statistical information. Under certain assumptions (notably, linear models and Gaussian noise) these are equivalent to maximum likelihood estimates derived from the complete probability structure of the problem.

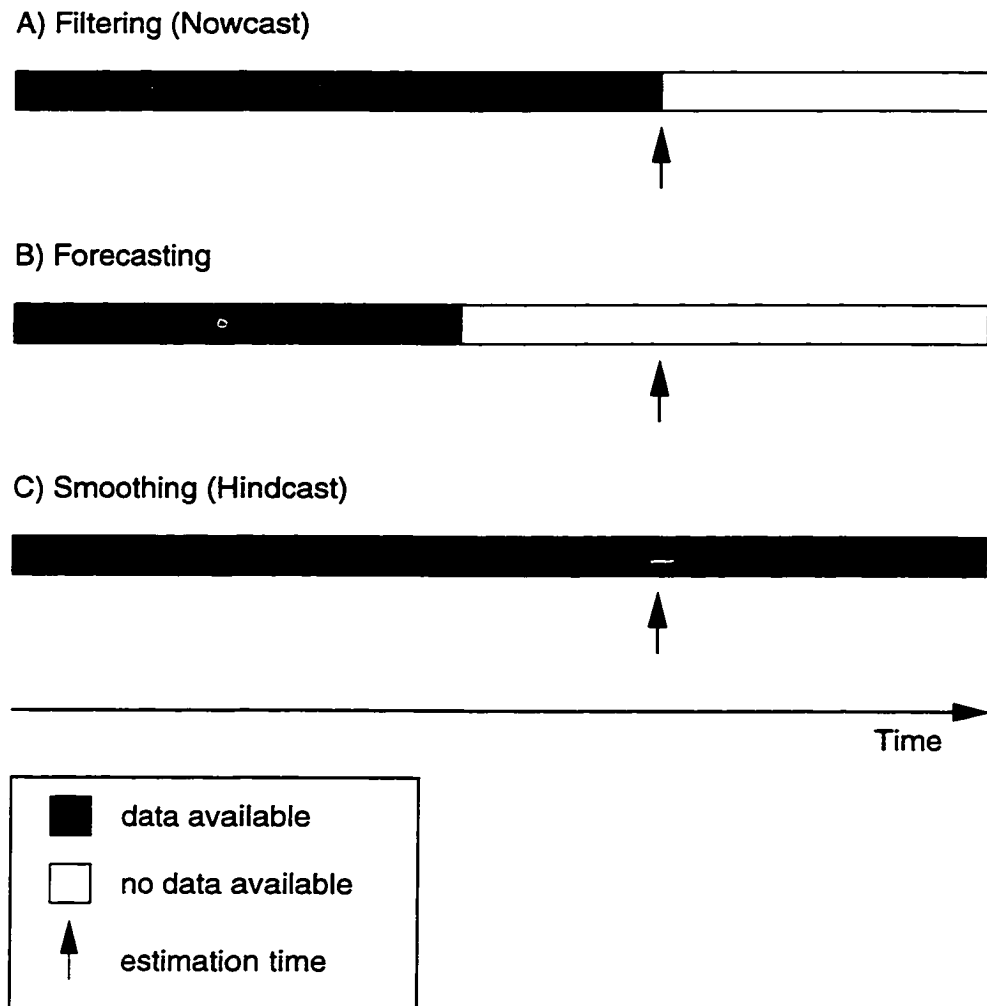


Figure 2.1: Schematic diagram contrasting the nowcasting (filtering), forecasting, and hindcasting (smoothing) problems.

2.3 Filtering and Forecasting

Filtering methods in oceanographic data assimilation are normally viewed in the context of the Kalman filter (Kalman 1960). It is a sequential scheme which is statistically optimal for linear systems with Gaussian noise statistics. It combines a model forecast at time t with observations at that time to produce a better overall estimate of the state. It is recursive and well suited to nowcasting and forecasting. In this section, we develop the Kalman filter using the linear model (2.15) and observation equation (2.16). Nonlinear extensions and application in oceanography are then reviewed.

2.3.1 The Kalman Filter

Suppose that an unbiased estimate of the ocean state \mathbf{x}_{t-1} is available from previous analysis. This prior estimate is denoted by $\hat{\mathbf{x}}_{t-1}$ and has an error covariance given by \mathbf{P}_{t-1} . It is also assumed that the system noise \mathbf{e}_t^m is a zero-mean, Gaussian process uncorrelated in time (but correlated in space) with covariance \mathbf{Q}_t . Using this information, the forecast of the state at the next time step $\bar{\mathbf{x}}_t$ is determined using (2.15) as

$$\bar{\mathbf{x}}_t = \mathbf{D}_t \hat{\mathbf{x}}_{t-1}. \quad (2.18)$$

The stochastic nature of this estimate is made more clear by writing

$$\bar{\mathbf{x}}_t = \mathbf{x}_t + \mathbf{e}_t^f, \quad (2.19)$$

where \mathbf{x}_t denotes the true value of the state, and \mathbf{e}_t^f the forecast error. It follows that the forecast error has zero-mean and a covariance given by

$$\text{var}(\bar{\mathbf{x}}_t - \mathbf{x}_t) \equiv \mathbf{M}_t = \mathbf{D}_t \mathbf{P}_{t-1} \mathbf{D}_t^T + \mathbf{G}_t \mathbf{Q}_t \mathbf{G}_t^T.$$

Suppose that observations \mathbf{z}_t are available at time t according to the linear measurement relation (2.16). *The filtering problem is then to find the best way to combine the forecast (2.18) with the observations (2.16).*

Equations (2.16) and (2.19) can be combined in standard regression form (2.1) as

$$\begin{pmatrix} \mathbf{z} \\ \bar{\mathbf{x}} \end{pmatrix} = \begin{pmatrix} \mathbf{H} \\ \mathbf{I} \end{pmatrix} \mathbf{x} + \begin{pmatrix} \mathbf{e}^o \\ \mathbf{e}^f \end{pmatrix}$$

where the the time index has been dropped for convenience. It is assumed that the forecast and observation errors are uncorrelated, i.e.

$$\text{var} \begin{pmatrix} \mathbf{e}^o \\ \mathbf{e}^f \end{pmatrix} = \begin{pmatrix} \mathbf{R} & 0 \\ 0 & \mathbf{M} \end{pmatrix},$$

where \mathbf{R} represents the error covariance matrix of the observations. The fact that the problem is now in standard regression form allows an optimal, in a least squares sense, $\hat{\mathbf{x}}$ to be obtained immediately as

$$\begin{aligned} \hat{\mathbf{x}} &= \mathbf{P} \left[(\mathbf{H}^T \mathbf{I}) \begin{pmatrix} \mathbf{R}^{-1} & 0 \\ 0 & \mathbf{M}^{-1} \end{pmatrix} \right] \begin{pmatrix} \mathbf{z} \\ \bar{\mathbf{x}} \end{pmatrix} \\ &= \mathbf{P} \mathbf{H}^T \mathbf{R}^{-1} \mathbf{z} + \mathbf{P} \mathbf{M}^{-1} \bar{\mathbf{x}}, \end{aligned} \quad (2.20)$$

where \mathbf{P} is the error covariance matrix of $\hat{\mathbf{x}}$ and given by

$$\begin{aligned} \mathbf{P} &= \left[(\mathbf{H}^T \mathbf{I}) \begin{pmatrix} \mathbf{R}^{-1} & 0 \\ 0 & \mathbf{M}^{-1} \end{pmatrix} \begin{pmatrix} \mathbf{H} \\ \mathbf{I} \end{pmatrix} \right]^{-1} \\ &= [\mathbf{H}^T \mathbf{R}^{-1} \mathbf{H} + \mathbf{M}^{-1}]^{-1}. \end{aligned} \quad (2.21)$$

The estimate $\hat{\mathbf{x}}_t$ is the Kalman filter estimate of the state. Its form is more commonly expressed as follows. Substituting (2.21) into (2.20) yields

$$\begin{aligned} (\mathbf{M}^{-1} + \mathbf{H}^T \mathbf{R}^{-1} \mathbf{H}) \hat{\mathbf{x}} &= \mathbf{H}^T \mathbf{R}^{-1} \mathbf{z} + \mathbf{M}^{-1} \bar{\mathbf{x}} \\ &= (\mathbf{M}^{-1} + \mathbf{H}^T \mathbf{R}^{-1} \mathbf{H}) \bar{\mathbf{x}} + \mathbf{H}^T \mathbf{R}^{-1} (\mathbf{z} - \mathbf{H} \bar{\mathbf{x}}). \end{aligned}$$

Pre-multiplying by $(\mathbf{M}^{-1} + \mathbf{H}^T \mathbf{R}^{-1} \mathbf{H})^{-1}$ gives

$$\begin{aligned} \hat{\mathbf{x}} &= \bar{\mathbf{x}} + (\mathbf{M}^{-1} + \mathbf{H}^T \mathbf{R}^{-1} \mathbf{H})^{-1} \mathbf{H}^T \mathbf{R}^{-1} (\mathbf{z} - \mathbf{H} \bar{\mathbf{x}}) \\ &= \bar{\mathbf{x}} + \mathbf{P} \mathbf{H}^T \mathbf{R}^{-1} (\mathbf{z} - \mathbf{H} \bar{\mathbf{x}}) \\ &= \bar{\mathbf{x}} + \mathbf{K} (\mathbf{z} - \mathbf{H} \bar{\mathbf{x}}), \end{aligned}$$

where we have defined the gain matrix $\mathbf{K} = \mathbf{P}\mathbf{H}^T\mathbf{R}^{-1}$.

This shows that the estimate for the optimal state is calculated using the forecast field plus a correction term based on the discrepancy between the observations and the forecast. The weighting factor \mathbf{K} is determined from the covariance structure of the problem (outlined below). The associated sequential updating consisting of the forecast, observation, and analysis procedure is known as the Kalman filter. For interpretation purposes, consider the case where z and \bar{x} are scalars with error covariance σ_o^2 and σ_f^2 respectively. If $\mathbf{H} = 1$, then the solution is

$$\hat{x} = \left(\frac{1}{1 + \sigma_o^2/\sigma_f^2} \right) z + \left(\frac{1}{1 + \sigma_f^2/\sigma_o^2} \right) \bar{x}.$$

This indicates that the filter estimate is just a weighted sum of observations and model forecast. In the limit where the observations have much greater uncertainty than the model forecast, i.e. $\sigma_o^2/\sigma_f^2 \gg 1$, then the estimate \hat{x} reverts to the forecast. Similarly, when $\sigma_o^2/\sigma_f^2 \ll 1$ then \hat{x} tends to the observations.

To summarize, the Kalman filter estimate of the state at time t is

$$\hat{\mathbf{x}}_t = \bar{\mathbf{x}}_t + \mathbf{K}_t(\mathbf{z}_t - \mathbf{H}_t\bar{\mathbf{x}}_t), \quad (2.22)$$

where the model forecast is determined through

$$\bar{\mathbf{x}}_t = \mathbf{D}_t\hat{\mathbf{x}}_{t-1}. \quad (2.23)$$

The following sequence of matrix operations determines the error covariance matrices for $\bar{\mathbf{x}}_t$ and $\hat{\mathbf{x}}_t$ as well as the gain matrix:

$$\text{var}(\bar{\mathbf{x}}_t - \mathbf{x}_t) \equiv \mathbf{M}_t = \mathbf{D}_t\mathbf{P}_{t-1}\mathbf{D}_t^T + \mathbf{G}_t\mathbf{Q}_t\mathbf{G}_t^T. \quad (2.24)$$

$$\text{var}(\hat{\mathbf{x}}_t - \mathbf{x}_t) \equiv \mathbf{P}_t = (\mathbf{M}_t^{-1} + \mathbf{H}_t^T\mathbf{R}_t^{-1}\mathbf{H}_t)^{-1} \quad (2.25)$$

$$\mathbf{K}_t = \mathbf{P}_t\mathbf{H}_t^T\mathbf{R}_t^{-1}. \quad (2.26)$$

These equations require knowledge of the initial state $\mathbf{x}_0 \sim WS(\hat{\mathbf{x}}_0, \mathbf{M}_0)$, the system noise $\mathbf{e}_t^m \sim WS(0, \mathbf{Q}_t)$, and the observation noise $\mathbf{e}_t^o \sim WS(0, \mathbf{R}_t)$ (WS designates a ‘wide-sense’ distribution, i.e. one characterized by its mean and covariance).

Estimates of the state are then produced sequentially. For the linear system with Gaussian noise, the Kalman filter is optimal in the sense that it is a maximum likelihood estimate of the system state.

It is useful to comment on the minimization approach which leads to the Kalman filter estimates. *At any given time step*, the optimal estimate $\hat{\mathbf{x}}_t$ can be obtained by minimizing the following cost function

$$J_f = (\mathbf{x}_t - \bar{\mathbf{x}}_t)^T \mathbf{M}_t^{-1} (\mathbf{x}_t - \bar{\mathbf{x}}_t) + (\mathbf{z}_t - \mathbf{H}_t \mathbf{x}_t)^T \mathbf{R}_t^{-1} (\mathbf{z}_t - \mathbf{H}_t \mathbf{x}_t) \quad (2.27)$$

with respect to \mathbf{x}_t (Diderrich 1985). The forecast and observation errors are combined in a single scalar quantity J_f which measures the squared deviations of the model forecast and observations from the true state weighted by the inverse of their respective covariance matrices.

2.3.2 Extensions and Applications

For a nonlinear model or observation equation, the optimality properties of the Kalman filter no longer hold. Linear systems and Gaussian error statistics imply that the mean and covariance offer a complete description of the probability structure of the system. The presence of nonlinearities implies that higher order probability moments exist (see Appendix B). In addition, a host of other dynamical and estimation issues arise in nonlinear filtering (see Jazwinski 1970).

The extended Kalman filter is a direct extension of the linear filter which is suitable for application with nonlinear models. The extended Kalman filter uses the nonlinear model (2.13) as the basis for the forecast equation (2.23). The nonlinear observation equation (2.14) is used in the updating equation (2.22). The matrix equations (2.24-2.26) retain their same form but with \mathbf{D}_t and \mathbf{H}_t being the linearized versions of their nonlinear counterparts \mathbf{d} and \mathbf{h} about the current state estimate $\hat{\mathbf{x}}_t$. Ideally, the model is re-linearized at every time step but, in practice, this can occur less frequently as the model state may not vary significantly over a given time interval.

Monte Carlo methods offer a possible means to evaluate the forecast error statistics. Evensen (1994) uses such an approach with a two layer nonlinear quasi-geostrophic ocean model which proceeds as follows. An ensemble of initial states is defined which are representative of the initial probability density. The model is then integrated forward from these states until the first observation time, the results recorded, and the model forecast error variance estimated. It is then updated using available observations in the same way as the usual Kalman filter. The procedure is then repeated using a new ensemble of model states derived from the updated statistics. The results of Evensen's application were encouraging and suggested that $O(10^2)$ model integrations were adequate to specify the forecast error statistics in that case.

Filters of the stochastic approximation type (Gelb 1974) offer an alternative solution to the nonlinear filtering problem. Hoang et al. (1994) apply such a nonlinear adaptive filter which estimates the time evolution of gain matrix \mathbf{K}_t by minimizing the model prediction errors $\mathbf{z}_t - \mathbf{H}\bar{\mathbf{x}}_t$. This minimization involves the adjoint of the dynamical operator \mathbf{d} , and entirely avoids the use of (2.24)-(2.26) in determining the gain matrix.

The main advantage of the Kalman filter for oceanographic applications is that it produces explicit error estimates and can account for model errors. Its widespread use has been hindered by two important issues. First, the computational burden imposed by the large dimension of most ocean models makes implementation of the full Kalman filter problematic (Ghil and Malanotte-Rizzoli 1991). Second, it is difficult to accurately specify the suite of necessary input statistics, including observation errors (instrument noise and errors in interpolating the observations to the model variables and to the model grid) and system noise (neglected dynamics, numerical approximations, and stochastic forcing). These are often poorly known and yet important to filter performance (Jiang and Ghil 1992, Daley 1992). It is important to note, however, that all assimilation methods must make assumptions about these input statistics, but often this is not done explicitly.

2.4 Smoothing

Data assimilation methods based on filtering are suitable mainly for real time applications where nowcasts and forecasts are carried out as new data becomes available. However, in the case where an oceanographic field experiment has been completed, or archived data sets are available, it is *hindcasts* of the ocean state which are of interest. The key feature of a hindcast is the availability of future information relative to the analysis time. This allows inferences about the system noise process to be made, and it is this feature that distinguishes the filtering problem from the smoothing problem.

2.4.1 Regression Solution

Consider combining the linear model (2.15) and observation equation (2.16) using a regression approach. Following Duncan and Horn (1972) the regression model $\mathbf{z} = \mathbf{D}\mathbf{b} + \mathbf{e}$ corresponding to the filtering problem (hereafter assuming, for simplicity, that $\mathbf{G}_t \equiv \mathbf{I}$) is

$$\begin{pmatrix} \hat{\mathbf{x}}_0 \\ 0 \\ \mathbf{z}_1 \\ 0 \\ \mathbf{z}_2 \\ \vdots \\ 0 \\ \mathbf{z}_N \end{pmatrix} = \begin{pmatrix} \mathbf{I} & 0 & 0 & \dots & 0 & 0 \\ -\mathbf{D}_1 & \mathbf{I} & 0 & \dots & 0 & 0 \\ 0 & \mathbf{H}_1 & 0 & \dots & 0 & 0 \\ 0 & -\mathbf{D}_2 & \mathbf{I} & \dots & 0 & 0 \\ 0 & 0 & \mathbf{H}_2 & \dots & 0 & 0 \\ \vdots & \vdots & & & \vdots & \\ 0 & 0 & 0 & \dots & -\mathbf{D}_N & \mathbf{I} \\ 0 & 0 & 0 & \dots & 0 & \mathbf{H}_N \end{pmatrix} \begin{pmatrix} \mathbf{x}_0 \\ \mathbf{x}_1 \\ \mathbf{x}_2 \\ \vdots \\ \mathbf{x}_{N-1} \\ \mathbf{x}_N \end{pmatrix} + \begin{pmatrix} -\mathbf{e}_0 \\ -\mathbf{e}_1^m \\ \mathbf{e}_1^o \\ -\mathbf{e}_2^m \\ \mathbf{e}_2^o \\ \vdots \\ -\mathbf{e}_N^m \\ \mathbf{e}_N^o \end{pmatrix} \quad (2.28)$$

where $\hat{\mathbf{x}}_0$ is a prior estimate of \mathbf{x}_0 . The generalized least squares regression solution for (2.28) reveals that the last rows of the solution vector $\hat{\mathbf{b}}$ associated with \mathbf{x}_N are equivalent to the Kalman filter estimate $\hat{\mathbf{x}}_N$ (Duncan and Horn 1972). The estimates $\hat{\mathbf{x}}_t$ for $t < N$ (denoted $\hat{\mathbf{x}}_{t|N}$) are not the same as the Kalman filter estimates since future information with respect to the analysis time is available. These are the smoothing estimates and identical to the filtering estimates only at the end of

the observation interval. (Recall that this property was immediately evident in the probabilistic approach).

Due to the large size of the arrays in the regression equation (2.28), direct solution using the approaches of Section 2.1 is not feasible for realistic cases. The structure of \mathbf{D} in (2.28) is such that it is sparse and banded in blocks, which correspond to time steps. It is sensible to approach the smoothing problem in an optimization framework which explicitly takes into account the temporal structure of the problem.

2.4.2 The Kalman Smoother

The regression problem given by (2.28) is equivalent to minimizing the cost function

$$J_s = \frac{1}{2} \mathbf{e}_0^T \mathbf{P}_0^{-1} \mathbf{e}_0 + \frac{1}{2} \sum_{t=1}^N \left\{ \mathbf{e}_t^o T \mathbf{R}_t^{-1} \mathbf{e}_t^o + \mathbf{e}_t^m T \mathbf{Q}_t^{-1} \mathbf{e}_t^m \right\} \quad (2.29)$$

with respect to \mathbf{x}_t for $t = 0, \dots, N$. The quantity J_s is the sum of the squares of the initial error, the observation error, and the model error weighted by the inverse of their respective covariance matrices.

To minimize (2.29) one could use the chain rule and directly differentiate this equation with respect to the \mathbf{x}_t , set the result to zero, and solve for \mathbf{x}_t . An equivalent, but more convenient, approach is to treat the problem as a constrained optimization problem. That is, we seek to minimize

$$J_s = \frac{1}{2} (\mathbf{x}_0 - \hat{\mathbf{x}}_0)^T \mathbf{P}_0^{-1} (\mathbf{x}_0 - \hat{\mathbf{x}}_0) + \frac{1}{2} \sum_{t=1}^N \left\{ (\mathbf{z}_t - \mathbf{H}_t \mathbf{x}_t)^T \mathbf{R}_t^{-1} (\mathbf{z}_t - \mathbf{H}_t \mathbf{x}_t) + \mathbf{e}_t^{mT} \mathbf{Q}_t^{-1} \mathbf{e}_t^m \right\} \quad (2.30)$$

with respect to the unknown \mathbf{x}_t using the constraint

$$\mathbf{e}_t^m = \mathbf{x}_t - \mathbf{D}_t \mathbf{x}_{t-1}. \quad (2.31)$$

The solution to the constrained optimization problem can be found using the method of Lagrange multipliers (*e.g.* Bertsekas 1982). The Lagrange function is

$$L = J_s + \sum_{t=1}^N \boldsymbol{\gamma}_{t-1}^T (\mathbf{x}_t - \mathbf{D}_t \mathbf{x}_{t-1} - \mathbf{e}_t^m)$$

where γ_t represents the unknown Lagrange multipliers. The minimum in J_s is found by differentiating L with respect to the unknowns $\gamma_t, \mathbf{x}_t, \mathbf{e}_t^m$ and \mathbf{x}_0 and setting the result to zero. This results in the following system of equations:

$$\frac{\partial L}{\partial \gamma_t} = 0 \Rightarrow \mathbf{x}_t = \mathbf{D}_t \mathbf{x}_{t-1} + \mathbf{e}_t^m \quad (2.32)$$

$$\frac{\partial L}{\partial \mathbf{x}_t} = 0 \Rightarrow \gamma_{t-1} = \mathbf{D}_{t+1}^T \gamma_t + \mathbf{H}_t^T \mathbf{R}_t^{-1} (\mathbf{z}_t - \mathbf{H}_t \mathbf{x}_t) \quad (2.33)$$

$$\frac{\partial L}{\partial \mathbf{e}_t^m} = 0 \Rightarrow \mathbf{e}_t^m = \mathbf{Q}_t \gamma_{t-1} \quad (2.34)$$

$$\frac{\partial L}{\partial \mathbf{x}_0} = 0 \Rightarrow \mathbf{x}_0 = \hat{\mathbf{x}}_0 + \mathbf{P}_0 \mathbf{D}_0^T \gamma_0. \quad (2.35)$$

The Kalman smoother algorithm is based on the above system of equations and proceeds as follows:

1. The first equation describes a forward sweep through the model equations. The initial conditions \mathbf{x}_0 and \mathbf{e}_t^m are unknown or uncertain, but observations \mathbf{z}_t are available. As demonstrated in Section 2.3.1, the optimal estimates of the state for the forward sweep, $\hat{\mathbf{x}}_t$, are the Kalman filter estimates. Note that the estimates $\hat{\mathbf{x}}_N$ at the final time are equal to the smoother estimates $\hat{\mathbf{x}}_{N|N}$.
2. The second equation describes a backward sweep of the Lagrange multipliers γ_t , starting from $\gamma_N = 0$. The Lagrange multiplier equations are forced by the discrepancy between the observations \mathbf{z}_t and the Kalman filter estimates $\hat{\mathbf{x}}_t$ from (2.32), which take the place of \mathbf{x}_t in this equation.
3. The third equation allows estimates of the system noise \mathbf{e}_t^m to be made using the Lagrange multipliers.
4. The final equation updates the original estimate of the initial state \mathbf{x}_0 using the Lagrange multiplier obtained at the end of the backward integration.

Updated estimates of the state can also be obtained at any intermediate time. Note that every Kalman filter estimate $\hat{\mathbf{x}}_t$ has an associated error covariance matrix

\mathbf{P}_t . Therefore, without loss of generality, \mathbf{x}_t and $\hat{\mathbf{x}}_t$ could be substituted for \mathbf{x}_0 and $\hat{\mathbf{x}}_0$ in (2.30). Estimating these using the smoother equations yields

$$\hat{\mathbf{x}}_{t|N} = \hat{\mathbf{x}}_t + \mathbf{P}_t \mathbf{D}_t^T \boldsymbol{\gamma}_t, \quad (2.36)$$

where $\hat{\mathbf{x}}_{t|N}$ denotes the estimate of the analyzed field using all the information available to time N . This is the smoothing estimate, and the above equation clearly shows that it is a further refinement on the Kalman filter estimate $\hat{\mathbf{x}}_t$.

To summarize, the Kalman smoother requires a forward sweep through the observations (the Kalman filter) and a subsequent backward sweep which determines the Lagrange multipliers. Future information relative to the analysis time is now used to update the Kalman filter estimates of the state and make inferences about the system noise (forcing). Only one forward and backward sweep is needed since an optimal estimate of the state at the final time ($\hat{\mathbf{x}}_{N|N} = \hat{\mathbf{x}}_N$) is available to begin the forcing of the Lagrange multipliers in the backward integration. A set of recursive relations are also available for determining the error covariances for its state estimates of the Kalman smoother (Bryson and Ho 1969, Section 13.2)

2.4.3 The Adjoint Method

An important subclass of smoothing problems occurs when the model state can be parameterized in terms of a small number of unknowns such as initial or boundary conditions or internal parameters such as friction coefficients. This implies that the model dynamics are treated as a strong constraint and can be expressed in terms of these unknowns. The gradient of J_s with respect to these unknowns can be determined and the minimization carried out based on this information. We motivate this approach using the Kalman smoother as illustrated below.

Suppose that Kalman filter estimates were not available for the forward sweep in the smoothing algorithm given by (2.32)-(2.35). One could, in principle, proceed by guessing the initial conditions \mathbf{x}_0 and the system noise \mathbf{e}_t^m and integrate forward the model (2.32). Clearly, the resulting estimates for \mathbf{x}_t would be suboptimal to a

degree depending on the validity of these guesses. However, the Lagrange multiplier equation (2.33) could still be integrated backwards using the estimated \mathbf{x}_t . The updated estimates of \mathbf{x}_t from (2.36) are clearly suboptimal and imply that J_s is not at a minimum and therefore the gradient of J_s with respect to \mathbf{e}_t^m and \mathbf{x}_0 , is not equal to zero. Importantly, the gradient does provide information about the direction in which the guesses for \mathbf{e}_t^m and \mathbf{x}_0 should be adjusted in order to achieve a smaller value of the cost function J_s . Once the \mathbf{e}_t^m and \mathbf{x}_0 are suitably modified using this gradient information, the procedure can be repeated and new estimates obtained. This iterative approach to minimizing J_s is the essence of the control methods discussed in this section.

The number of iterations required for the above procedure to converge to a minimum in J_s typically scales with the number of variables which must be estimated (Gill *et al.* 1981). Such control methods thus prove useful in cases where the model can be parameterized in terms of a relatively small number of unknowns, or where suitable regularization terms exist to accelerate the rate of convergence. This fact, together with the practical difficulty in accurately specifying the error statistics of the system noise, has resulted in the widespread application of these adjoint based smoothing methods in oceanographic data assimilation. To illustrate the method in more detail we consider the special case of an initial value problem.

Now consider the case where the system noise \mathbf{e}_t^m is negligible (or the error covariance of the system noise \mathbf{Q}_t goes to zero). Inspection of (2.29) reveals that the dynamics become a strong constraint which must be satisfied exactly. Under the above assumptions, the smoothing problem of (2.29) now seeks to minimize

$$J = \frac{1}{2} \sum_{t=1}^N (\mathbf{z}_t - \mathbf{H}_t \mathbf{x}_t)^T \mathbf{R}_t^{-1} (\mathbf{z}_t - \mathbf{H}_t \mathbf{x}_t) \quad (2.37)$$

with respect to \mathbf{x}_t , subject to the constraint

$$\mathbf{x}_t - \mathbf{D} \mathbf{x}_{t-1} = 0. \quad (2.38)$$

Suppose the initial conditions \mathbf{x}_0 are taken as the unknown control variables of the

problem. With the dynamics now describing an initial value problem, the initial conditions can be adjusted in such a way that the model best matches the observations. The goal of (2.37)-(2.38) is then to choose an estimate for \mathbf{x}_0 which minimizes the observation/model misfit as given by J . We illustrate the solution to this problem first using regression and then using the so-called adjoint method.

Consider an alternative representation of (2.38) as a matrix equation

$$\mathbf{x} = \mathbf{D}\mathbf{b}, \quad (2.39)$$

where $\mathbf{x} = (\mathbf{x}_1^T, \dots, \mathbf{x}_N^T)^T$, $\mathbf{D} = \text{diag}(\mathbf{D}, \mathbf{D}^2, \dots, \mathbf{D}^N)$, and \mathbf{b} is the vector of unknown control variables such that $\mathbf{b} = \mathbf{x}_0$. The model has been parameterized in terms of the initial conditions, whose specification is sufficient to determine (or control) all subsequent \mathbf{x}_t . Given observations $\mathbf{z} = (\mathbf{z}_1^T, \dots, \mathbf{z}_N^T)^T$, the goal is to minimize

$$J = (\mathbf{z} - \mathbf{H}\mathbf{x})^T \mathbf{R}^{-1} (\mathbf{z} - \mathbf{H}\mathbf{x}),$$

where $\mathbf{H} = \text{diag}(\mathbf{H}_1, \dots, \mathbf{H}_N)$. Substituting (2.39) into the above equation, differentiating J with respect to \mathbf{b} and setting the result to zero yields

$$\hat{\mathbf{b}} = (\mathbf{H}^T \mathbf{D}^T \mathbf{R}^{-1} \mathbf{D} \mathbf{H})^{-1} \mathbf{H}^T \mathbf{D}^T \mathbf{R}^{-1} \mathbf{z}$$

which is a generalized least-squares regression estimate $\hat{\mathbf{b}}$ for \mathbf{b} in terms of the observations \mathbf{z} (see Section 2.1.1) and a solution to the constrained optimization problem of (2.37)-(2.38). Note that the error covariance matrix for $\hat{\mathbf{b}}$ is given by the quantity in brackets. In the general case, the control vector \mathbf{b} could contain not only the initial conditions but any quantity which controls the evolution of the state. These include the forcing and any uncertain internal parameters of the problem.

The above analysis reveals the main strength of the optimal control method: the high dimensional state space can be projected onto a much smaller dimension control variable space by treating a portion of the dynamics as exact. However it would be a difficult task to express any realistic ocean model in the form of (2.39), and the array sizes are sufficiently large to make the above matrix manipulations problematic. A

more usual approach to (2.37)-(2.38) is to retain the time-stepping model and use a Lagrange multiplier technique.

Following Section 2.4.2, the Lagrange function for the problem is obtained by appending to J in (2.37) the constraint (2.38) multiplied by Lagrange multipliers γ_t . Differentiating with respect to the unknowns γ_t and \mathbf{x}_t and setting the result to zero yields a modified version of (2.32) and (2.33) where $\mathbf{e}_t^m = 0$. Differentiating with respect to the controls yields

$$\frac{\partial L}{\partial \mathbf{x}_0} = \frac{\partial J}{\partial \mathbf{x}_0} = -\mathbf{D}^T \boldsymbol{\gamma}_0, \quad (2.40)$$

which provides the required information for the minimization of J . The minimization algorithm proceeds as follows: (i) Guess a value for \mathbf{x}_0 , (ii) Run the model forward N time steps and evaluate J , (iii) Run the backward model for γ_t ($\gamma_N = 0$), and evaluate $\partial J / \partial \mathbf{x}_0$ as above, (iv) Use a gradient descent algorithm to adjust \mathbf{x}_0 such that J is decreased, and (v) Continue the procedure until $|\partial J / \partial \mathbf{x}_0| < \epsilon$ where ϵ is some threshold value which designates that a minimum has been found. A linear, conjugate gradient technique (Gill *et al.* 1981, section 4.8) can be shown to converge to a minimum in J in m iterations where m is less than or equal to the number of control variables p . (Note that the effective convergence rate can be accelerated through judicious use of regularization). The computational requirements in this case (for each iteration) are one forward integration of the model, one backward integration of its adjoint, and the calculation of gradients. This iterative smoothing technique has come to be known as the *adjoint method* of data assimilation.

2.4.4 The Representer Solution

Recall the form of the cost function given by (2.37)-(2.38) and, for simplicity, let $\mathbf{R}_t = \mathbf{I}$, implying that the observations all have equal variance. The method of representer is based on expressing the state \mathbf{x}_t as

$$\mathbf{x}_t = \sum_{k=1}^K \beta^k \mathbf{c}_t^k, \quad (2.41)$$

where \mathbf{c}_t^k are the representer functions, β^k are time invariant coefficients and K refers to the total number of data points $\mathbf{z} = (\mathbf{z}_1^T, \dots, \mathbf{z}_N^T)^T$. (Note that the usual development of the representer method assumes a prior estimate of \mathbf{x}_t in which case the LHS of (2.41) is interpreted as a correction term on this prior. We have chosen to ignore this for clarity and notational simplicity).

The representer field \mathbf{c}_t^k , $t = 1, \dots, N$ represents the influence of the k th observation on the overall model state. It satisfies the model equations

$$\mathbf{c}_t^k = \mathbf{D}_t \mathbf{c}_{t-1}^k + \boldsymbol{\gamma}_t^k \quad (2.42)$$

with an initial condition $\mathbf{c}_0^k = \boldsymbol{\gamma}_0^k$. The $\boldsymbol{\gamma}_t^k$ satisfy the adjoint equations

$$\boldsymbol{\gamma}_{t-1}^k = \mathbf{D}_{t+1}^T \boldsymbol{\gamma}_t^k + H_t^{kT} z_t^k, \quad \boldsymbol{\gamma}_N^k = 0, \quad (2.43)$$

where z_t^k is the k th observation and is obtained at time t , and H_t^k represents the corresponding row of \mathbf{H}_t .

Consider an interpretation of the representers. Examining the adjoint equations (2.43) reveals that for each representer, a single observation drives the adjoint equations. In this sense, $\boldsymbol{\gamma}_t^k$ may be interpreted as an influence function for that observation on the ocean state. Since the integration of (2.43) runs backward in time from an initial state of rest, $\boldsymbol{\gamma}_t^k$ will be zero at times greater than t . The $\boldsymbol{\gamma}_t^k$ then provide a forcing term for the forward integration of (2.42) and thus serve to define the representer field \mathbf{c}_t^k associated with the k th observation. The net effect of sequentially applying (2.43) and (2.42) is to spread the influence of a single observation over the entire integration period. The set of \mathbf{c}_t^k may be interpreted as Green's functions for each of the observations.

Given that the representer functions have been obtained, an estimate for state \mathbf{x}_t requires that the weighting coefficients β^k in (2.41) be determined. Define $\boldsymbol{\beta} = (\beta^1, \dots, \beta^K)^T$ and $\mathbf{C}_t = (\mathbf{c}_t^1, \dots, \mathbf{c}_t^K)$ whereupon (2.41) is expressed in matrix form as

$$\mathbf{x}_t = \mathbf{C}_t \boldsymbol{\beta}. \quad (2.44)$$

Pre-multiplying both sides by H_t^k provides a model prediction of the data point z_t^k at time t , i.e.

$$\hat{z}_t^k = H_t^k C_t \beta. \quad (2.45)$$

Stacking the complete set of \hat{z}_t^k in the vector $\hat{\mathbf{z}}$ transforms (2.45) to

$$\hat{\mathbf{z}} = \mathbf{HC}\beta \quad (2.46)$$

where the product \mathbf{HC} is determined via (2.45). This provides model predictions of the observations in terms of the coefficients β . By replacing $\hat{\mathbf{z}}$ with the actual observations \mathbf{z} and performing the inversion we obtain the estimate for the representer coefficients

$$\hat{\beta} = (\mathbf{HC})^{-1} \mathbf{z}. \quad (2.47)$$

Given $\hat{\beta}$, and knowledge of the representer functions, the state at any time can be estimated through (2.41). Some properties of representer functions are presented in Bennett (1992, chapter 5).

The representer functions are determined by solving $2K$ initial value problems (an integration of the forward model (2.42) and its adjoint (2.43) are carried out for each representer function). The representer coefficients are then obtained by a matrix inversion involving a $k \times k$ matrix. These computational requirements can be contrasted with the coupled, two point boundary value problem that governs the adjoint method of Section 2.4.3. That technique uses an iterative solution to obtain the minimum of the cost function, where the number of iterations varies roughly as the number of control variables p (depending on the regularization terms). As a result, the representer technique may prove efficient in the case where $k < p$.

2.4.5 Application and Extensions

The major motivation for the use of these smoothing techniques in oceanography has been to extract the maximum amount of information about the ocean state from existing data sets. For practical purposes, adjoint methods, which treat the dynamics

as a strong constraint, have generally been preferred. Such an approach allows an ocean model to be parameterized in terms of a (relatively) small number of unknowns, typically the initial/boundary conditions and internal parameters, thereby aiding the optimization procedure. Ill-conditioning in the problem, usually due to the sparse nature of the observations, is alleviated by adding regularization terms to the cost function (see Section 2.2). These are often based on physical knowledge of the problem at hand. The strong similarity between the forward and backward models allows efficient, and familiar, numerical schemes to be used together with well developed gradient descent algorithms designed for large scale problems (Navon and Legler 1987).

The optimization framework of the smoothing problem allows for a conceptually straightforward extension to the case of nonlinear models and observation equations. However, in practice, nonlinear optimizations can prove difficult. Nonlinearities introduce the possibility of multiple minima in the cost function. Even the task of finding a local minimum using gradient descent techniques is hindered as a result of the non-quadratic nature of the cost function. For weakly nonlinear problems, optimization methods applicable for linear systems often prove adequate. However, for strongly nonlinear systems, techniques such as simulated annealing and Monte-Carlo methods must be considered (e.g. Miller *et al.* 1994).

2.5 Summary

The primary character of oceanographic data is its irregular distribution as well as its diversity of forms, ranging from satellite images to coastal tide gauges. Data assimilation methods have the potential to map these data both temporally and spatially in a manner consistent with our knowledge of ocean dynamics, as well as test hypotheses about ocean dynamics and thermodynamics. In this chapter, we have introduced a number of generic methods for combining ocean models with data. However, as with any estimation problem, successful applications will rely on the

appropriateness of the data and the model relative to the requirements of the physical problem at hand.

Data assimilation in oceanography includes general inverse, filtering and smoothing methods. General inverse methods prove useful mainly to highlight the important elements of the data assimilation problem. Most oceanographic interest lies in assimilation techniques based on time dependent models for the purpose of nowcasting (filtering) and hindcasting (smoothing) ocean circulation. In the linear case, the optimal filter and smoother can be readily identified. Still, the computational burden imposed by the large dimension of most ocean models, as well as incomplete knowledge of the required error statistics, often prevents practical applications of these methods. This points to the need for suboptimal, or approximate, data assimilation techniques.

In the case of nonlinear models, the optimality properties of the Kalman filter and smoother no longer hold as the mean and covariance are insufficient to provide a complete description of the ocean state. The validity of nonlinear extensions of the filter and smoother are then questionable in many instances. Higher order filters and smoothers, which keep track of additional moments of the probability density function, are available but computationally infeasible for most realistic problems. This argues strongly for the use of optimization techniques for treating large scale problems in nonlinear data assimilation. The estimation problem is treated from an optimization perspective, i.e. minimizing a deterministic cost function. In addition to the attendant technical and numerical difficulties, there are also a number of outstanding issues, such as demonstrating one has found a true global minimum, and the form and weighting given to the regularization terms.

Finally, we mention the issue of regularization, which is common to most problems in oceanographic data assimilation (in fact to all problems if one considers that the ocean is an infinite dimensional system). Underdetermined systems arise as a result of insufficient data available for estimation of the unknown parameters. If one is not prepared to change the postulated model, ill-conditioning can be alleviated either by

reformulating the problem in terms of a smaller number of parameters or by introducing prior information into the problem. Even in the apparently well posed cases, regularization may prove necessary to achieve estimates which fall within acceptable limits.

This overview serves as a basic primer on data assimilation. The remainder of this thesis concerns the application of these methods to problems in coastal and continental shelf oceanography and frequent reference to the material presented in this chapter will be made throughout.

Chapter 3

Extraction of Tidal Streams from a Ship-Borne ADCP

The ship-mounted acoustic Doppler current profiler (ADCP) has become an important instrument for the study of coastal and continental shelf circulation. Its basic measurement is a vertical profile of the horizontal current obtained at regular time intervals along the cruise track of the ship. The flexibility of a moving observation platform allows for flexible sampling strategies and full areal coverage of many circulation features. It has been used for such diverse purposes as measuring flow around headlands (Geyer and Signell 1990), identifying internal wave structure (Marmorino and Trump 1992), determining the flow in a channel (Simpson *et al.* 1990) and as an independent check on numerical models (Howarth and Proctor 1992).

Ready interpretation of ship-borne ADCP data in coastal and continental shelf regions is confounded by the presence of tides in the record. For a current time series from a fixed location, harmonic analysis (e.g. Godin 1972) is adequate to remove the tides. However, the movement of the ship precludes a straightforward procedure for tide extraction from the ADCP record. In such a case, tide removal must take into account not only the periodic variation through time but also the spatially varying amplitude and phase resulting from the progression of the tide through the region.

A number of methods have been proposed for removing tides from ADCP records.

The simplest involves designing the sampling strategy so that repeat measurements are made at fixed locations at regular time intervals. Conventional harmonic analysis can then be used for de-tiding the series (Geyer and Signell 1990, Simpson *et al.* 1990). This method is not always feasible, because sampling is often dictated by other considerations and the ship-borne ADCP is only a supplemental source of information. Another method uses prior estimates of the tides from observations or numerical modeling which allows the tidal signal to be removed directly from the ADCP record (Howarth and Proctor 1992, Foreman and Freeland 1991). Clearly, this approach is restricted to areas in which the tides are well known. A final method fits arbitrary basis functions to the ADCP record to describe the spatially varying tidal amplitude and phase (Candela *et al.* 1992). The use of these interpolation functions, which ignore the dynamics, is problematic in regions with relatively complicated flow fields (Foreman and Freeland 1991).

The problem of tide extraction is readily considered in the context of oceanographic data assimilation. Bennett and McIntosh (1982) first treated open ocean tidal modeling as an inverse problem. Their approach was based on the variational analysis of fixed tide gauge and current meter time series using a shallow water model as a weak constraint. More recently, sea level and current meter time series have been analyzed with a tidal model, enforced as a strong constraint, for the purpose of estimating various model parameters (Das and Lardner 1990, Lardner *et al.* 1993) and boundary conditions (Lardner 1993). This study addresses the problem of tidal analysis of a time-space series of velocity (a ship-borne ADCP record) using simple tidal dynamics.

In this chapter, the problem of tide extraction from the ADCP record is treated as a discrete inverse problem as in Section 2.1. The proposed method is straightforward, robust, and computationally efficient. These qualities make it suitable for operational shipboard use. The focus is on limited area models with one or more open boundaries and spatial scales of 10 to 1000 km. Tidal flows across the open boundaries of the model are treated as unknown quantities and estimated from ADCP data in the

interior. In this manner, tidal maps over the model domain are obtained directly from the ADCP data. These can then be used for a more detailed examination of the subtidal circulation.

The chapter is organized as follows. Section 3.1 describes the Western Bank region of the Scotian Shelf off Canada's east coast and presents current meter and ship-borne ADCP data collected on the cruise of April 1992. Tidal dynamics and the tidal model are then introduced in Section 3.2. Section 3.3 sets up the problem of tidal analysis of ADCP data as a regression problem in the frequency domain. Finally, application of these methods to the ADCP data collected on the Scotian Shelf is carried out in Section 3.4. A summary and discussion follows in Section 3.5.

3.1 Observations

A ship-borne ADCP records a time-space series of the horizontal components of current over depth. The focus of this study is on estimating the barotropic tide from this record, and henceforth it will be assumed that each of these vertical profiles has been depth averaged. This produces a pair of horizontal velocity components at each observation time and location. In a typical case, observation times occur at regular intervals but with some missing values or gaps. The observation locations correspond to the cruise track of the ship and represent irregular, but quasi-continuous, transects through the region of interest. Clearly, these data can contain a great deal of information on coastal circulation. In this section, we motivate the tidal extraction procedure by examining a set of current meter data and a ship ADCP record from a shelf region and discuss the tidal and subtidal variability contained in these records.

As part of the Ocean Production Enhancement Network's (OPEN) cod recruitment study, a number of cruises have been made recently to the Western Bank region of the Scotian Shelf off Canada's east coast. Figure 3.1 shows the bathymetry of this region and details the study area of the April 1992 cruise. Overall, the physics of the Scotian Shelf are relatively complex, particularly on the outer shelf which interacts

strongly with the open ocean (for a review, see Smith and Schwing 1991). For the Western Bank region, the Rossby number is less than 0.1 and a hydrographic survey taken during the cruise (CTD stations shown in Figure 3.1) indicated that this region is generally well mixed in the vertical down to about 70m. Some density structure becomes evident below this level and along the edge of the Bank (Griffin and Lochmann 1992). We now introduce current meter and ship-borne ADCP data collected on the April 1992 Western Bank cruise.

Current data were obtained from 3 current meters deployed at the locations shown in Figure 3.1 at the depths given in Table 3.1. The corresponding time series plots of the horizontal components of velocity are given in Figure 3.2. Density profiles indicated a relatively barotropic region near these moorings (Griffin and Lochmann 1992), suggesting that the measured currents reflect the depth-mean flow. The tidal signal associated with the M_2 , S_2 , K_1 and O_1 constituents was extracted from each of the current meter time series using a least squares regression (these constituents were separable given the length of the record). Table 3.1 indicates that the tides in this region accounted for approximately 75% of the total variance in the current meter records. The main purpose of this current meter data, in the present investigation, is to provide an independent check on the ADCP tidal extraction procedure carried out in Section 3.4.

The wind driven component of each de-tided current meter series was isolated by expressing it as a linear combination of the wind vector (shown in Figure 3.2) at lags 0, 16, 32 and 48 hours. This procedure is a reasonable one provided the current meter is far from coastal boundaries (Pollard and Millard 1970). In fact, analysis of the transfer function between the wind and the three current time series indicated a flow to the right of the wind with a damped resonant response centered on the inertial frequency, consistent with simple Ekman dynamics. Table 3.1 indicates the nontidal variance was split between the wind and a residual associated with low frequency (sub-synoptic) variability.

Ship-borne ADCP data were also available for the cruise period. The original

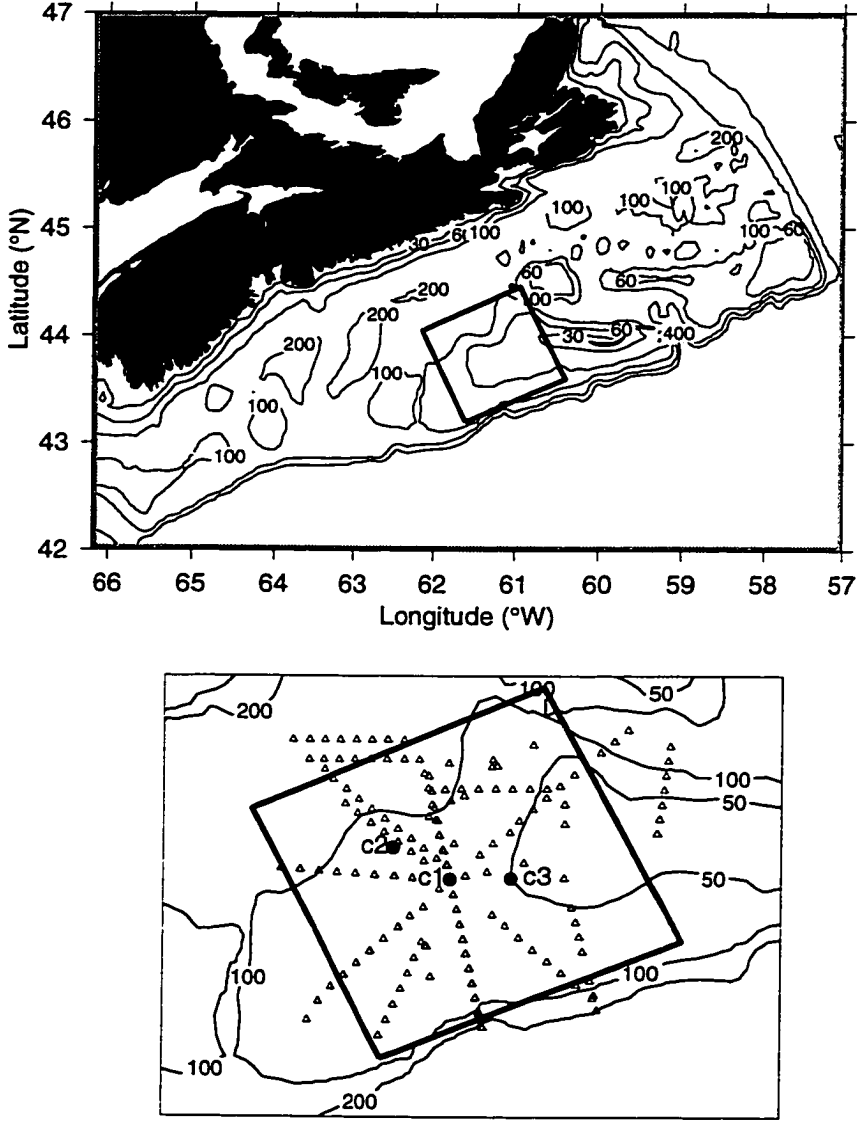


Figure 3.1: (Upper panel) Coastline of Nova Scotia and the bathymetry, in meters, of the Scotian Shelf. The rectangular box indicates the study region, and model domain, centered on Western Bank. (Lower panel) Detail of the boxed region in the upper panel. The locations of the three current meters (c1, c2, c3, represented by solid dots) deployed on the April 1992 cruise and CTD stations (triangles) are also shown.

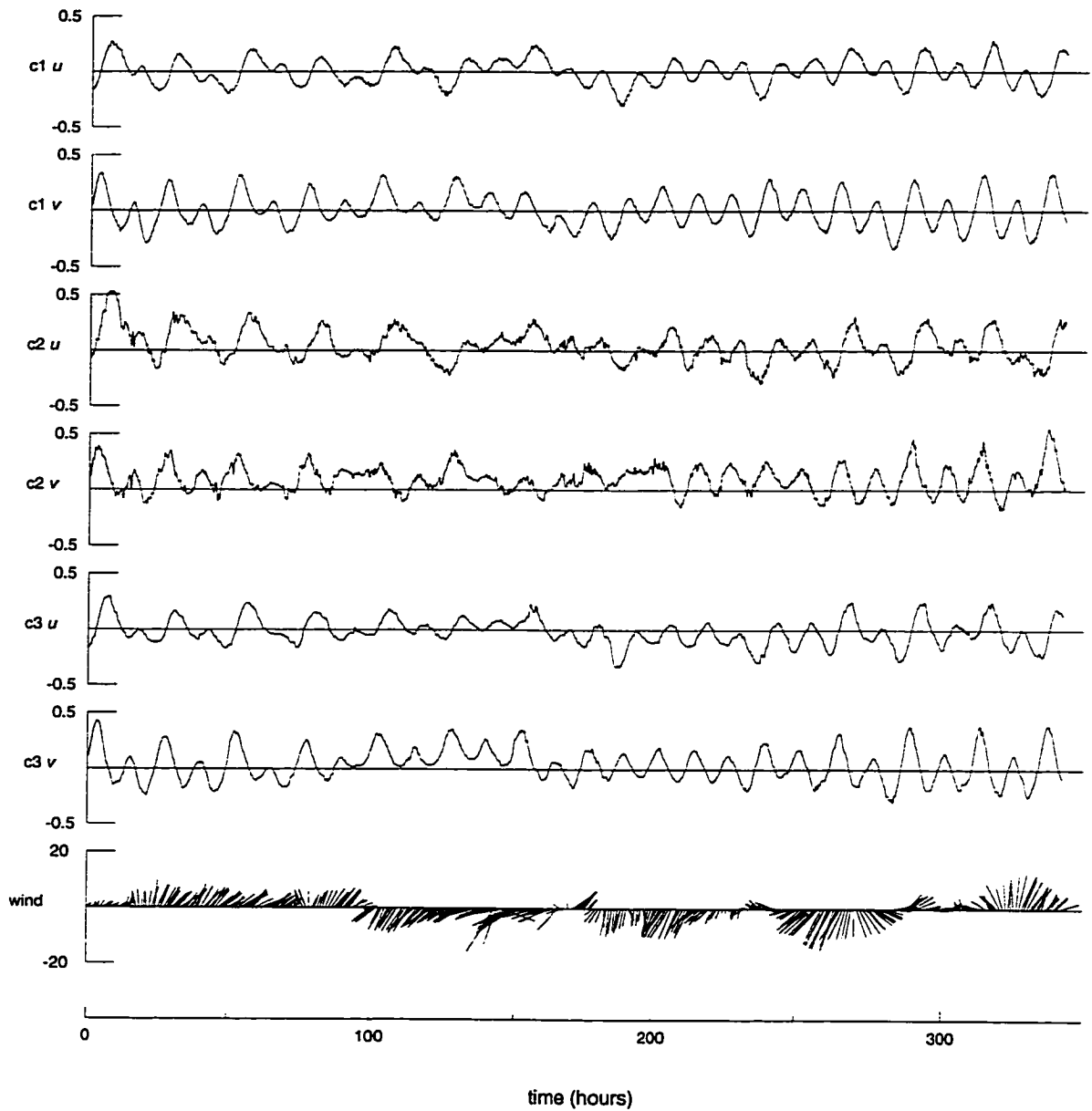


Figure 3.2: Time series plots of the east-west u and north-south v velocity components for the three current meters ($c1$, $c2$, $c3$, in m s^{-1}). The time series of the wind vector (in m s^{-1}) is also shown in the bottom panel. The series begins at 2115, April 20, 1992.

ADCP data consisted of a time-space series of vertical profiles of the horizontal current, bottom referenced and output every 3 minutes. Since the focus here is on extracting the barotropic tide, the vertical profiles were depth averaged from the surface to 60m (or bottom) and re-sampled at 15 minute time intervals. The end result was a data set consisting of a pair of horizontal velocity components, each with a corresponding observation time and location.

A time series representation of this ADCP data is shown in Figure 3.3. The time series show strong diurnal and semi-diurnal oscillations in the velocity components. Power spectra of the data, shown in Figure 3.4 confirm that most of the energy in the series is contained in these bands. Figure 3.4 also indicates that the north-south component v is dominated by the semi-diurnal oscillations, while the east-west component u has its energy equally split between the diurnal and semi-diurnal bands. The lower panel in Figure 3.4 shows the rotary spectrum of the complex time series $u + iv$. It indicates a clockwise sense of rotation for the current vector in both the diurnal and semi-diurnal bands.

Due to the observation location changing over time, tidal constituents derived from the ADCP velocity might be expected to have an amplitude and phase which varies through time, thereby precluding application of the usual harmonic analysis techniques. Figure 3.5 shows the results of a complex demodulation of the u component of velocity about the M_2 tidal frequency and suggests that the amplitude and phase of the tide vary within the individual tidal frequencies. Part of this variation in amplitude might be accounted for by the spring-neap cycle, but the tidal amplitude also strongly correlates with the water depth beneath the ship at the recording time. The apparent phase of the M_2 tide also varies as a result of the movement of the ship. One fairly extreme example is evident in Figure 3.5 at approximately 240 hours when irregularities in the velocity signal are found with the movement of the ship over the shelf break. The source of this phase shift is unclear; it might be due to physical processes such as internal tides contaminating the barotropic signal, or be due to the instrument errors such as loss of bottom tracking.

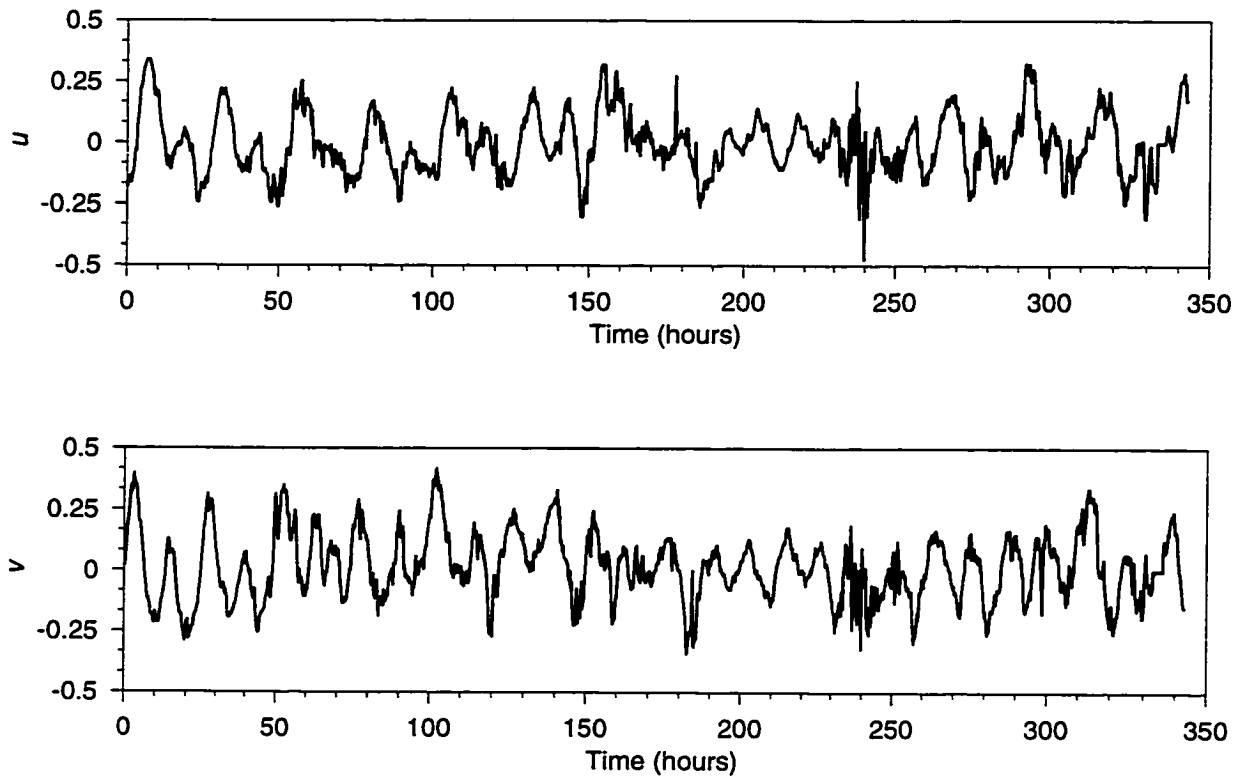


Figure 3.3: Time series plots of the east-west (u) and north-south (v) components of the depth-averaged velocity recorded by the ship ADCP (in m s^{-1}). The series begins at 21:15, April 20, 1992.

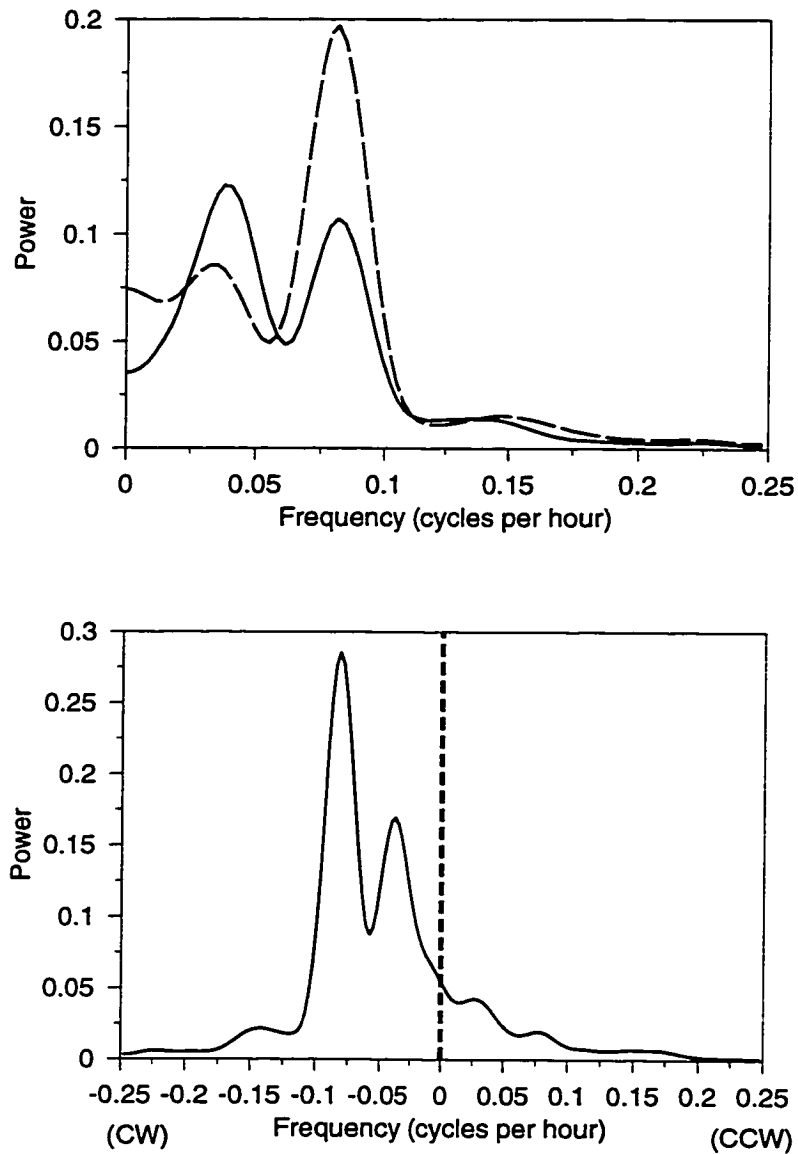


Figure 3.4: (Upper Panel) Power spectrum (m^2s^{-2}) of the east-west u (solid line) and north-south v (dashed line) components of the depth averaged current from the ship-ADCP. (Lower Panel) Rotary spectrum (m^2s^{-2}) of the depth-averaged ADCP velocity. The labels CW and CCW below the x-axis indicate clockwise and counter-clockwise rotation, respectively.

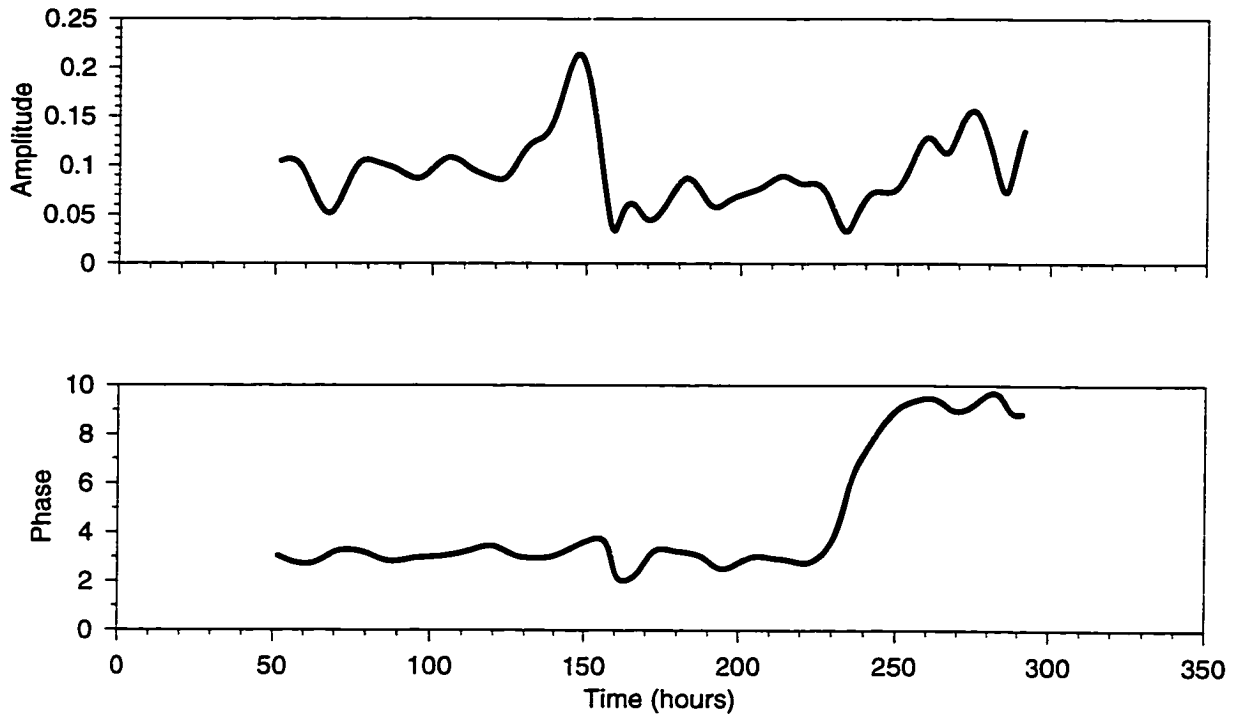


Figure 3.5: Results from complex demodulation of the u component of the depth averaged ADCP velocity centered about the M_2 tidal frequency. The upper panel shows variation in the M_2 amplitude (ms^{-1}) over time, and the lower panel shows the M_2 phase (radians) over time. The time series is truncated at either end as a result of the M_2 filter.

Table 3.1: Variance (cm^2s^{-2}) partitioning for the 3 current meter time series (c1, c2, c3, with mooring depths indicated) and the depth-averaged ADCP currents. The variance is decomposed into tides (M_2, S_2, K_1, O_1), wind, and a residual.

current meter	tides	wind	residual	total
c1 (33m)	267	29	47	341
c2 (20m)	232	24	81	335
c3 (20m)	275	27	52	353
ADCP	183	14	116	342

To represent the spatial dimension of the ADCP data, Figure 3.6 shows vectors of hourly ADCP data for various time windows during the cruise period. Clearly the data are difficult to interpret in this case. The rotary nature of the current vector in the presence of tides is evident, resulting in what appears to be a number of apparent convergences and divergences in the flow field. A meaningful residual circulation can only be obtained by careful removal of the tides from the record.

3.2 Tidal Model

3.2.1 Governing Equations

A model describing the spatial and temporal variation of the barotropic tide is an integral part of our tidal extraction procedure. A suitable model for the Western Bank region is based on the linearized, depth averaged shallow water equations

$$\begin{aligned} \frac{\partial \vec{u}}{\partial t} + f\vec{k} \times \vec{u} + g\nabla\eta + \frac{r}{h}\vec{u} &= \vec{0} \\ \frac{\partial \eta}{\partial t} + \nabla \cdot (\vec{u}h) &= 0, \end{aligned} \quad (3.1)$$

where t is time, $\vec{u}(\vec{x}, t)$ contains the east-west and north-south components of the depth averaged current and $\eta(\vec{x}, t)$ represents the sea surface elevation above some undisturbed depth of water $h(\vec{x})$. The horizontal gradient operator is denoted $\nabla =$

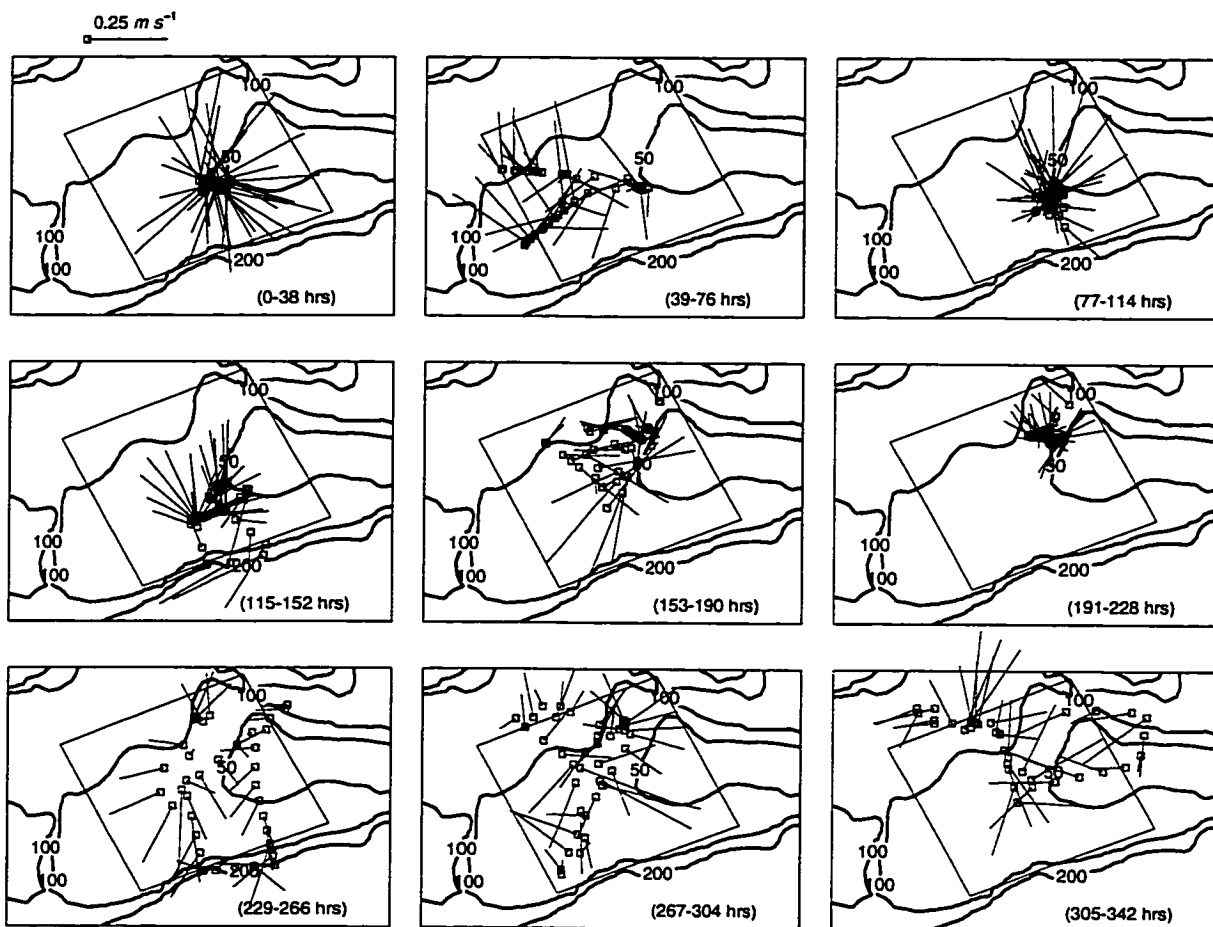


Figure 3.6: Vector plots of the observed depth averaged currents from the ADCP using a time interval of 1 hour beginning 21:15, April 20, 1992. The contours represent the bathymetry of the region, and the large box is the study area (model domain). The small open squares represent the location of the ship, and the associated straight lines are the current vectors.

$(\partial/\partial x, \partial/\partial y)$. The Coriolis parameter is represented by f and \vec{k} is the unit vertical vector. The remaining parameters are the acceleration of gravity g and the bottom friction coefficient r . For simplicity, a linear friction based on the depth-averaged velocity is used.

The shallow water equations (3.1) require initial and boundary conditions to be properly posed. Consider the initial conditions for \vec{u} and η . The ocean tides are observed to be in a periodic steady state, and the model must be integrated until this condition is satisfied. As a result, the velocity and sea level fields are independent of the initial conditions, and they need not be considered further.

For the boundary conditions, we denote the flow normal to the boundary by $\vec{u} \cdot \vec{n}$, where \vec{n} is the unit outward normal at the boundary of the domain. Open boundary flows of the form

$$\vec{u} \cdot \vec{n} = \sum_{k=1}^K \Re\{a_k e^{i\omega_k t}\} \quad (3.2)$$

are postulated to describe the tidal flows across the open boundary. Here, ω_k is the frequency of the k th tidal constituent, a_k is its complex amplitude, \Re denotes the real part of a complex quantity and the summation is over the K tidal constituents of interest. Although the tidal frequencies are well known, the boundary amplitudes ($|a_k|$) and phases ($\arg a_k$) constitute a major source of uncertainty.

The flows across the open boundary for each tidal constituent are parameterized in terms of spatial structure functions. For a north-south oriented boundary, the spatial variation in the complex amplitude a_k of the normal velocity at the boundary is expressed as

$$a_k(y) = \sum_{p=1}^P h^{-1} b_{kp} \zeta_p(y), \quad (3.3)$$

where $h = h(y)$. The structure functions are represented by ζ_p and their coefficients are given by the complex constants b_{kp} . For a typical model domain, a set of P structure functions would be used to represent the boundary flows along each open boundary. Note that the expansion is expressed in terms of transport since the barotropic tidal velocities scale inversely with depth. The purpose of (3.3) is to allow

the boundary flows to be parameterized by a smaller number of coefficients than the number of boundary grid points.

We have assumed that the tidal flows across the open boundary are periodic in time, with known frequency. As the shallow water equations (3.1) define a linear system, the interior sea level and velocities are also periodic in time with the same frequency as the boundary forcing but with a different amplitude and phase. This fact allows the shallow water equations to be posed sensibly in the frequency domain. If it is assumed that \vec{u}, η have a time dependence of the form $e^{i\omega t}$, the shallow water equations (3.1) can be written

$$\begin{aligned} (i\omega + r/h)\vec{u} + f\vec{k} \times \vec{u} + g\nabla\eta &= 0 \\ i\omega\eta + \nabla \cdot (\vec{u}h) &= 0, \end{aligned} \quad (3.4)$$

where \vec{u}, η are now defined as the complex amplitudes of their respective variables. The result is a boundary value problem which requires specification of the complex amplitudes of the boundary flows. This transformation of the shallow water equations into the frequency domain allows the explicit temporal dependence to be removed from the equations.

3.2.2 Discrete Form

Tidal dynamics are governed by the shallow water equations posed in the frequency domain, i.e. the boundary value problem of (3.2)-(3.4). Consider the general form for a discrete (numerical) implementation of this boundary value problem. For the k th tidal constituent, it can be expressed as the matrix equation

$$\mathbf{D}_k^{\text{LHS}} \mathbf{x} = \mathbf{D}_k^{\text{RHS}} \mathbf{b}_k. \quad (3.5)$$

The vector \mathbf{b}_k contains the coefficients b_{kp} of (3.3) which describe the tidal flows across the open boundary. The vectors \mathbf{x} contain the complex amplitudes of the velocity and sea level defined on the interior model grid points. The matrices $\mathbf{D}_k^{\text{LHS}}$ and $\mathbf{D}_k^{\text{RHS}}$ are derived from the shallow water dynamics (3.4).

Premultiplying both sides by the inverse of $\mathbf{D}_k^{\text{LHS}}$ and retaining only the rows corresponding to the horizontal velocity yields

$$\mathbf{x}_k = \mathbf{D}_k \mathbf{b}_k, \quad (3.6)$$

where \mathbf{x}_k no longer contains the sea level, only horizontal velocity. The resulting matrix \mathbf{D}_k has complex entries and represents a mapping of the model boundary to the interior for a single tidal constituent.

To include the full set of K tidal constituents into the above form, we stack the equations in (3.6) as follows:

$$\begin{pmatrix} \mathbf{x}_1 \\ \vdots \\ \mathbf{x}_K \end{pmatrix} = \begin{pmatrix} \mathbf{D}_1 & \cdots & \mathbf{0} \\ \vdots & \ddots & \vdots \\ \mathbf{0} & \cdots & \mathbf{D}_K \end{pmatrix} \begin{pmatrix} \mathbf{b}_1 \\ \vdots \\ \mathbf{b}_K \end{pmatrix}$$

where the left hand side contains the complex amplitudes of the tidal velocities at the model grid points. The right hand side is composed of a block diagonal matrix, the blocks being the dynamics matrix for each of the different tidal frequencies. This matrix multiplies the set of complex amplitudes of the boundary flows parameterized in terms of spatial structure functions. The above equation may be expressed more concisely as

$$\mathbf{x} = \mathbf{D} \mathbf{b} \quad (3.7)$$

which represents the complete tidal system in the frequency domain and corresponds to the time domain shallow water equations (3.1) and the boundary conditions (3.2) with a_k defined through (3.3). The dimensions of these quantities in relation to the model grid and number of tidal constituents is summarized in Table 3.2.

Note that the dynamics matrix \mathbf{D} can be obtained in at least two different ways. One method involves finite differencing the shallow water equations in the frequency domain (3.4) and writing the equations directly into the matrix form leading to (3.5). (3.6) is then a solution to the boundary value problem. An alternative method is based on using an existing tidal model. Suppose that a suitable time-stepping numerical

tidal model, based on (3.1), exists for the region. The dynamics operator \mathbf{D} may be generated as follows. Impose periodic flows across a portion of the open boundaries corresponding to one element of \mathbf{b} and furthermore use suitable radiation conditions on the remaining open boundaries. Integrate the model to a periodic steady state and perform a harmonic analysis of the interior velocities to yield the response \mathbf{x} to a single element of \mathbf{b} and therefore a column of the matrix \mathbf{D} . By continuing this process with respect to each element of \mathbf{b} , the full matrix \mathbf{D} can be sequentially obtained. Note that the two methods are not equivalent. The latter method relies on the use of radiation boundary conditions, while the former solves the boundary value problem directly.

3.3 Inverse Analysis

Tidal analysis involves fitting the shallow water equations to the time-space series of ADCP velocity using generalized least squares regression (Section 2.1). The tidal model is treated as a strong constraint inasmuch as the velocity in the model interior is determined exactly by specifying the boundary state and solving the dynamic equations. As a result, the boundary conditions can be treated as the unknown quantities to be estimated from the ADCP record.

To justify the choice of a strong constraint formalism for this case, note that the analysis is not intended to explain the full variability of the observations, only that part specifically due to the barotropic tide. (From the current meter data in Table 1, we might expect the tides to explain at most 75% of the variance). As a result, it is reasonable to assume that the observation error (the tidal residual) is much larger than the model error (the dynamical uncertainty in describing the barotropic tide) provided that the model is a reasonable one for the region, i.e. the assumptions of the tidal model (3.1) are satisfied.

Table 3.2: Definition of selected matrices and their dimensions. In this table, N is the total number of u and v points on the model grid, P is the number of spatial structure functions needed to describe the complex amplitudes of velocity points at the model boundaries for a single tidal constituent, K is the number of tidal constituents and L is the number of u and v observations.

Matrix	Size	Description
\mathbf{x}	$NK \times 1$	Complex amplitude of interior velocity at the model grid points including all K tidal constituents.
\mathbf{b}	$PK \times 1$	Coefficients of structure functions for the complex amplitude of velocity normal to the model boundary for the K tidal constituents.
\mathbf{D}	$NK \times PK$	Dynamics matrix mapping from \mathbf{b} to \mathbf{x} and derived from the shallow water equations in the frequency domain.
\mathbf{x}_k	$N \times 1$	Complex amplitude of interior velocity at the model grid points for the k th tidal constituent.
z_t	1×1	Observation of a single velocity component at time t .
\mathbf{z}	$L \times 1$	Vector containing l observed horizontal velocity components over time.
\mathbf{s}_t	$N \times 1$	Spatial interpolator to map n velocities on model grid at time t to a single measurement location.
\mathbf{H}	$L \times NK$	Overall interpolation matrix producing model counterparts to observations.

3.3.1 Comparing Model to Data

The tidal extraction method requires a means of comparing the model predictions, given in the frequency domain, to ADCP data obtained in the time domain. (In the terms of Chapter 2, this is the problem of constructing the observation operator \mathbf{H}). In other studies this comparison of model to data has been straightforward. McIntosh and Bennett (1984) used a frequency domain model similar to (3.4) to analyze current and sea level time series from fixed moorings. By performing a simple harmonic analysis on these data, the complex amplitudes were readily obtained and directly

comparable to the model variables. Lardner *et al.* (1993) used a time-stepping tidal model to assimilate time series of tidal velocities and sea level. In the case of a ship-borne ADCP, the comparison of model to data is more complicated.

Consider a single observation z_t measuring one velocity component at a particular location at time t . The model counterpart \hat{z}_t to the observation is

$$\hat{z}_t = \mathbf{s}_t^T \sum_{k=1}^K \Re\{\mathbf{x}_k e^{i\omega_k t}\}. \quad (3.8)$$

Here, the spatial interpolation vector \mathbf{s}_t represents a linear operator which maps from the model grid to the observation location at time t . Note that, in general, this vector changes through time, reflecting the fact that the measurement location is not fixed. The remainder of the equation describes the conversion from the frequency domain to the time domain.

Alternatively, one can introduce the following equivalent matrix form

$$\hat{z}_t = \Re \left\{ \mathbf{s}_t^T \left(\mathbf{I}e^{i\omega_1 t} \dots \mathbf{I}e^{i\omega_N t} \right) \begin{pmatrix} \mathbf{x}_1 \\ \vdots \\ \mathbf{x}_N \end{pmatrix} \right\} \quad (3.9)$$

where \mathbf{I} is an identity matrix. This can be written more concisely as

$$\hat{z}_t = \Re\{\mathbf{s}_t^T \mathbf{T}_t \mathbf{x}\}$$

where \mathbf{T}_t is the temporal interpolation matrix whose form is given above.

Extending the above analysis to an observation vector \mathbf{z} containing velocity components measured over time gives model counterparts $\hat{\mathbf{z}}$ to the data

$$\begin{aligned} \hat{\mathbf{z}} &= \Re\{\mathbf{S}\mathbf{T}\mathbf{x}\} \\ &= \Re\{\mathbf{H}\mathbf{x}\} \end{aligned} \quad (3.10)$$

where \mathbf{S} and \mathbf{T} are spatial and temporal interpolation matrices whose components are derived from \mathbf{s}_t , \mathbf{T}_t and \mathbf{H} is their product and corresponds to an overall interpolation operator. (Note that for implementation purposes, it makes sense to evaluate \mathbf{H}

directly using (3.8), or (3.9), rather than explicitly forming \mathbf{S} and \mathbf{T}). This clarifies the process of comparing model output in the frequency domain with observations in the time domain, when those observations include a spatial dimension.

3.3.2 Regression Solution

The regression equation that forms the basis of the inverse analysis is

$$\mathbf{z} = \hat{\mathbf{z}} + \mathbf{e},$$

where the error \mathbf{e} is the tidal residual and measures the discrepancy between the observations and the model predictions. This error is of unknown character, and its second-order properties are described by the error covariance matrix Σ . Using (3.7) and (3.10), and converting all matrices and vectors to their real forms (e.g. Brillinger 1981, section 3.7), gives the regression equation

$$\begin{aligned} \mathbf{z} &= \mathbf{H}\mathbf{x} + \mathbf{e} \\ &= \mathbf{H}\mathbf{D}\mathbf{b} + \mathbf{e}. \end{aligned} \tag{3.11}$$

The first equation of (3.11) contains no dynamics and the problem of minimizing the error would involve finding estimates of the interior tidal velocities \mathbf{x} from the ADCP observations \mathbf{z} . This ignores the dynamical links between the \mathbf{x} and the resulting solution may have little physical meaning. The second equation of (3.11) includes shallow water dynamics which imposes spatial structure on the interior and reduces the number of unknowns by expressing the interior velocities in terms of the boundary flows.

Generalized least squares regression (Section 2.1) minimizes the weighted sum of squares of the error to yield an optimal estimate $\hat{\mathbf{b}}$ for the unknown boundary flows \mathbf{b} in terms of the ADCP observations \mathbf{z} , i.e.

$$\hat{\mathbf{b}} = (\mathbf{D}^T \mathbf{H}^T \Sigma^{-1} \mathbf{H} \mathbf{D})^{-1} \mathbf{D}^T \mathbf{H}^T \Sigma^{-1} \mathbf{z}. \tag{3.12}$$

An estimate of the interior field $\hat{\mathbf{x}}$ is then obtained using (3.7) as

$$\hat{\mathbf{x}} = \mathbf{D}\hat{\mathbf{b}}, \quad (3.13)$$

and the estimate for the model counterparts to the data is

$$\hat{\mathbf{z}} = \mathbf{H}\mathbf{D}\hat{\mathbf{b}}. \quad (3.14)$$

Error estimates are also easily obtained as a part of the regression calculation. The covariances of $\hat{\mathbf{b}}$, $\hat{\mathbf{x}}$ and $\hat{\mathbf{z}}$ are (e.g. Sen and Srivastava 1990, chapter 2).

$$\begin{aligned} \text{var}(\hat{\mathbf{b}} - \mathbf{b}) &= \sigma^2(\mathbf{D}^T\mathbf{H}^T\boldsymbol{\Sigma}^{-1}\mathbf{H}\mathbf{D})^{-1} \\ \text{var}(\hat{\mathbf{x}} - \mathbf{x}) &= \sigma^2\mathbf{D}(\mathbf{D}^T\mathbf{H}^T\boldsymbol{\Sigma}^{-1}\mathbf{H}\mathbf{D})^{-1}\mathbf{D}^T \\ \text{var}(\hat{\mathbf{z}} - \mathbf{z}) &= \sigma^2\mathbf{H}\mathbf{D}(\mathbf{D}^T\mathbf{H}^T\boldsymbol{\Sigma}^{-1}\mathbf{H}\mathbf{D})^{-1}\mathbf{D}^T\mathbf{H}^T. \end{aligned}$$

These quantities are used to determine confidence intervals for the estimates $\hat{\mathbf{b}}$, $\hat{\mathbf{x}}$ and $\hat{\mathbf{z}}$.

3.3.3 Extensions

Suppose that a weak constraint formalism (i.e. including model errors) was required for the problem. In this case, model errors would be added to (3.7), which would modify the error covariance structure in (3.11). (An analogous procedure was carried out in Section 2.3.1 in the derivation of the Kalman filter). In other words, not only would the tidal residual need to be parameterized but so also would the dynamical errors in modeling the barotropic tide. Given the uncertainty in this and the expected (small) magnitude of its effect, we have chosen to remain with the strong constraint method.

It is also straightforward to introduce into the analysis any additional regularization terms, such as smoothing operators. They enter as prior information, or “bogus” data (Thacker 1988), on the boundary or interior flows. Mathematically, regularization adds another matrix (usually positive definite) to the bracketed term in (3.12),

thereby better conditioning the required matrix inversion (see Section 2.1.2). This technique closely resembles ridge regression in statistics, and the weight given the regularization term can be chosen through cross validation techniques (Golub *et al.* 1979).

To summarize, we have posed the tidal analysis of ADCP data as a highly over-determined regression problem. Careful attention has been paid to reducing the number of parameters to be estimated by writing the interior flows in terms of suitably parameterized boundary flows. The spatial structure functions of (3.3) also ensure that the number of unknowns remains constant irrespective of grid resolution. As a result, the need for introducing additional regularization terms (e.g. spatial smoothing) is not anticipated.

3.4 Application

The tidal and subtidal variability of Western Bank circulation are now examined using the ADCP data described in Section 3.1. Guided by the current meter results, the tidal constituents chosen for the analysis were M_2 , S_2 , K_1 and O_1 . The domain of the limited area tidal model is shown in Figure 3.1. Its dimensions are approximately 120km by 120km and it is inclined so as to be roughly parallel to the shelf break. The model had four open boundaries and used an Arakawa C-grid. The interior dimensions of the grid includes 14×14 η points, 14×13 u points and 13×14 v points. The open boundary is composed of 14 u points on the east and west boundaries and 14 v points on the north and south boundaries. The grid spacing is 8km.

Note that although the model is adequate for the Western Bank case, it is a relatively simple model and has small dimensions. The method is, however, easily extended and computationally feasible for larger problems with complex geometry.

Consider the specification of the unknown boundary flows \mathbf{b} . Chebyshev Type II polynomials were chosen for the basis set ζ_p . Numerical experiments indicated that only the first basis function, the mean transport across each boundary, was

required in this application. For our case, including higher order structure functions gave only a marginally better fit to the ADCP data and estimates in data sparse regions. Moreover, the flows were unrealistic, having a relatively large magnitude and small scale structure. This suggests model overfitting, since these flow features only appear in regions unconstrained by data and only in cases where additional degrees of freedom are introduced into the estimation. Higher order structure functions would be expected to be important in cases where the amplitude and phase of the tide vary significantly relative to the size of the model domain.

The finite difference form of the frequency domain model (3.4) was determined for the above grid and written directly as the required matrix equation (3.7). Given the grid described above and the 4 tidal constituents, the resulting matrix dimensions were the following: the state vector in the model interior \mathbf{x} was 2912×1 ; the dynamics matrix \mathbf{D} was 2912×32 ; the data vector \mathbf{z} was 2744×1 ; and the unknown boundary flow vector \mathbf{b} was 32×1 . (See Table 3.2 for matrix dimensions). The regression analysis of (3.12) thus involved inverting a 32×32 matrix.

Error estimates require that the covariance of the tidal residual Σ be determined. For simplicity, it was assumed that errors in the north-south v and east-west u components of the current were independent allowing treatment of the two components of the tidal residual separately. (In fact, their correlation coefficient was 0.16). The serial correlation of the tidal residuals for both u and v was found to be adequately described by a first-order autoregressive (AR(1)) process, i.e. fitting an AR(1) model resulted in serially uncorrelated errors with a constant variance. The estimated autoregressive coefficients implied a decorrelation time of about 1.5 hours for the residuals. Given this AR(1) error structure, the error covariance matrix Σ and its inverse were easily determined (Morrison 1967, section 8.11). Clearly, more complex statistical models could be considered for modelling the tidal residual. However, it was decided to use the most parsimonious model which adequately described the residual variability.

Tidal ellipses for M_2 , S_2 , K_1 and O_1 estimated by this analysis are shown in Figure 3.7. (These represent the estimates $\hat{\mathbf{x}}$ in (A)). The ellipses scale roughly inversely

with depth, being generally larger in the shallow water and smaller in the deep water. The M_2 tide is seen to dominate the region, and along with the two diurnal constituents, accounts for most of the variation in the tidal signal. The sense of rotation in all cases is clockwise, consistent with the results shown in Figure 3.4. Overall, the ellipse maps agree with previous calculations for the Scotian Shelf region (Gregory 1988). Table 3.1 indicates that the tides accounted for about 60% of the variance in the ADCP record. This was a lower proportion than for the current meters suggesting that the ADCP record contained additional nontidal processes, or more background noise, due to the ship's movement.

To further test the ADCP tidal estimates, the results were also compared to ellipses derived from the three current meters deployed near the center of the model domain. Figure 3.8 shows this comparison and indicates that the current meter ellipses match well those derived from the ADCP and in all but one case ($c3$, S_2) fall within the estimated 95% confidence region. For the relatively weak flows associated with S_2 and O_1 , Figure 3.8 suggests that the estimated ellipses were not significantly different from zero. To further refine the estimates, ADCP data sets widely separated in time could easily be used in the analysis of Section 3.3. Furthermore, the current meter data itself could be included in the inverse calculation but, in this case, it has been withheld for model validation purposes.

The estimated tidal currents along the cruise track of the ship \hat{z} are shown in Figure 3.9 (compare with Figure 3.3). This estimated tidal time series shows a spring-neap cycle and semidiurnal and diurnal oscillations. These periodic oscillations exhibit small discrete jumps over the length of the record, and become more pronounced toward the end. These are a result of the changes in depth along the ship track. The model grid further accentuates this when interpolating to the observation locations. The approximate width of the 95% confidence intervals for the estimated series was 0.07ms^{-1} . However, it did vary slightly over the length of the record, depending on the measurement location, and ranged from 0.06ms^{-1} to 0.09ms^{-1} .

The estimated error in the overall tidal flow field generally fell within this same

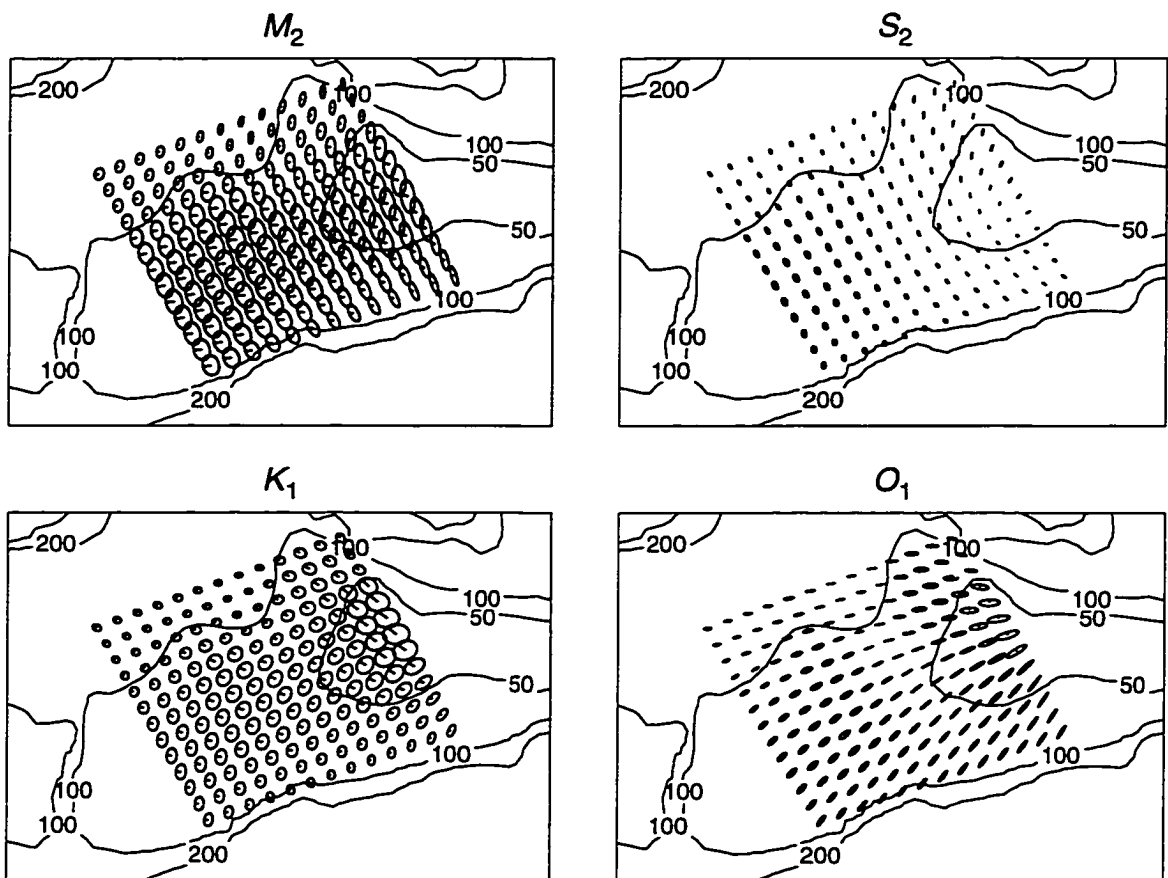


Figure 3.7: Tidal ellipse maps for M_2, S_2, K_1, O_1 estimated from the ADCP data and plotted at the model grid points. The ellipses are scaled to represent half the tidal excursion of the M_2 tide. The initial phase is represented by the straight line beginning at the center of each ellipse. The sense of rotation is clockwise.

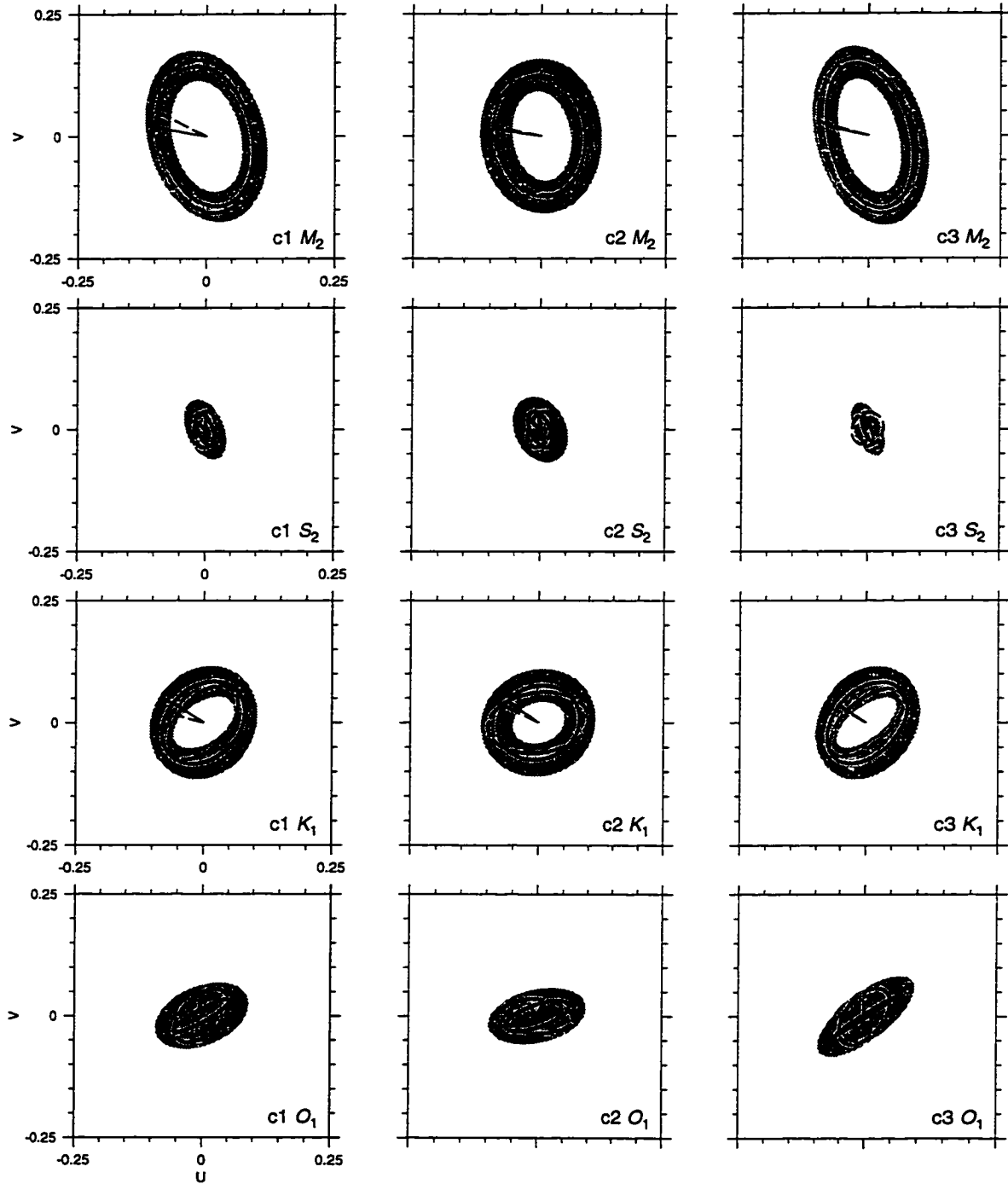


Figure 3.8: Comparison of M_2 , S_2 , K_1 , O_1 tidal ellipses obtained from current meters c1, c2, and c3 (dashed lines) and those estimated from the ADCP at corresponding locations (solid lines). The initial phase is represented by the straight line beginning at the center of each ellipse. The shaded area represents the 95% confidence region for the estimated ADCP tidal ellipses; u and v denote the respective east-west and north-south components of current velocity in m s^{-1} .

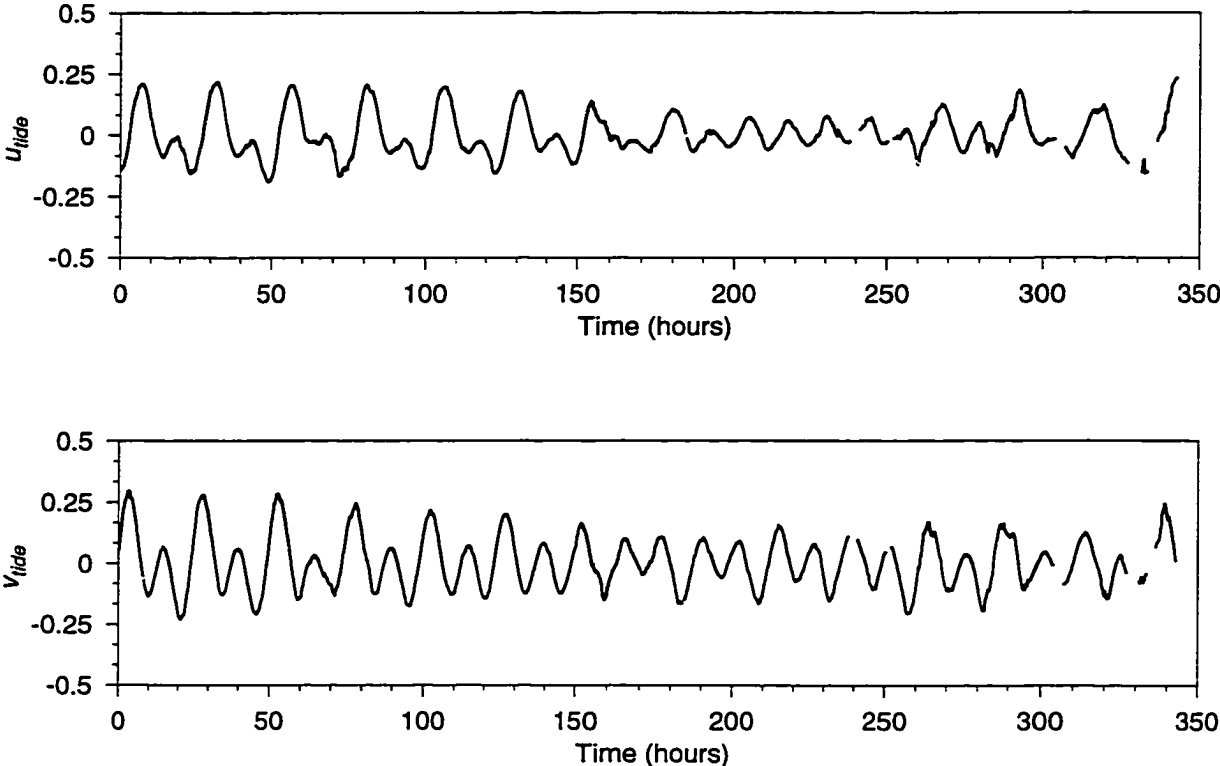


Figure 3.9: Time series plots of the east-west (u) and north-south(v) components (in $m s^{-1}$) of the barotropic tidal velocity estimated from the ship ADCP data . The series begins at 21:15, April 20, 1992. Gaps in the record indicate that the ship was outside the model domain.

range. Its magnitude depends on the duration of the ADCP record together with the (tidal) signal to noise ratio. The spatial variability of the error closely reflected the data distribution; it is lower in the central region and higher in the data sparse areas near the model boundary. The error field was also nearly identical for each of the 4 tidal constituents. Error ellipses were oriented mainly in the southwest-northeast direction. The major axes of the tidal ellipses often fell approximately perpendicular to this direction implying that the overall amplitude of the tidal constituent is well captured while its orientation is subject to some uncertainty.

Figure 3.10 shows vector plots of the tidal residual series (compare these with the vector plots of the original ADCP data in Figure 3.6). The residual currents are generally weaker than in the original series, especially in the central region near the crest of Western Bank. There is also some suggestion of a persistent northward current in the western part of the region. However, this is difficult to see due to the irregular cruise track of the ship and presence of a wind driven component to the circulation.

To remove the wind effect, the simple wind-driven model of Section 3.1 was fit (using the same lagged values of the wind) to the ADCP tidal residual series. In this case, the wind model explained only about 10% of the variability in the de-tided ADCP record, in contrast to the nearly 30% explained for the de-tided current meter records. Again, the transfer function between the wind and ADCP record had a broad peak centered near the inertial frequency but the corresponding direction of the flow was inconsistent with a simple Ekman flux to the right of the wind. Information on the wind driven circulation appears to be buried in the multitude of other signals recorded by the ADCP as the ship moved throughout the region.

To isolate the steady part of the circulation, the ADCP residual (tides and wind removed) were averaged into 0.2° latitude/longitude blocks. The result of this is shown in Figure 3.11 together with surface currents obtained from a diagnostic calculation (Sheng and Thompson 1996) using density data from 177 CTD casts obtained during the cruise (Figure 3.1). (Surface currents were used for comparison since the upper

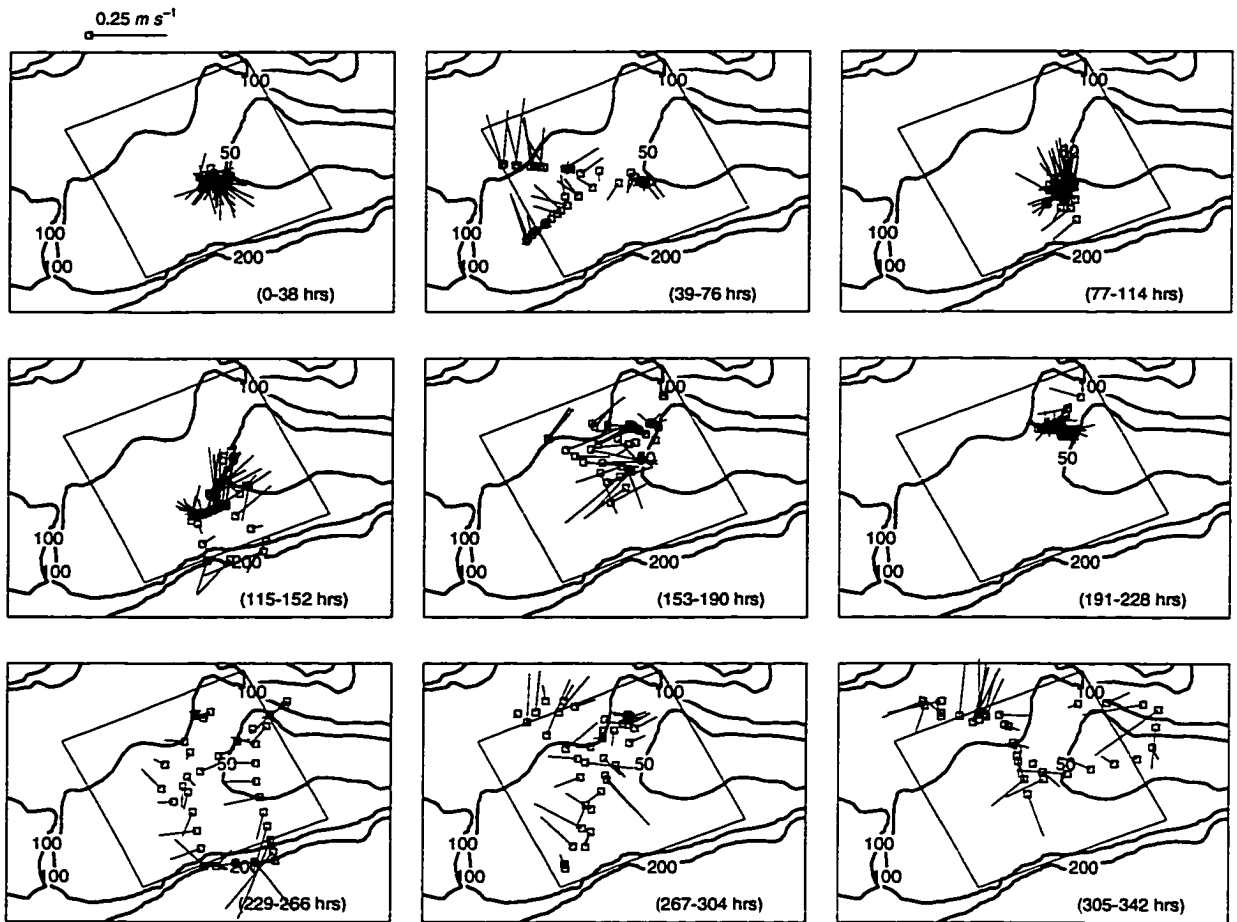


Figure 3.10: Vector plots of the estimated tidal residuals for the hourly ADCP data beginning 2115, April 20, 1992. Contours represent the bathymetry of the region and the large box is the study area (model domain). The small open squares represent the location of the ship, and the associated straight lines are the current vectors.

30m is relatively barotropic and therefore free of the sheared flows due to horizontal density gradients). The general flow patterns obtained from the two sources are similar, both show an anticyclonic gyre centered on the crest of Western Bank and a persistent northward flow to the west of this gyre. The existence of this gyre has been confirmed with drifter deployments made during the cruise (Sanderson 1995). Note that a corresponding averaging of the original ADCP data showed a series of almost randomly oriented vectors indicating that the tides could not simply be averaged to obtain the residual.

3.4.1 Sensitivity Analysis

A sensitivity analysis was carried out to determine the influence of grid resolution and the possible need for additional regularization. The tidal analysis of the ADCP data was carried out as described above for grid sizes ranging from a 14×14 grid (8km resolution) through to a 20×20 grid (5.6km resolution) inclusive. In every case, the horizontal extent of the model domain remained identical. The original bathymetry data, with a resolution of about 8km, was mapped to each of the new grids using bilinear interpolation. As a result, no formerly unresolved bathymetric features appear as grid size is increased, and any sensitivity to grid size cannot be attributed to this effect.

The results of this sensitivity analysis showed that the coefficient of determination R^2 measuring the model fit to the ADCP data (that is, the tidal part of the series) varied less than $\pm 1\%$ around a value of 61% over the specified range of grid sizes. Neither did the R^2 increase or decrease systematically. Typical ellipse fields for M_2 for 14×14 and 20×20 grids are shown in Figure 3.12. They are not markedly different from one another nor are the other constituents over the range of grid resolutions. This suggests that extrapolation to data sparse areas was not adversely affected. The discrepancies found close to the model boundaries are likely due to the the mapping of the observations to the grid points (since a nearest neighbor approach is used, this mapping is slightly different for each new grid). Finally, a 28×28 grid (4km resolution)

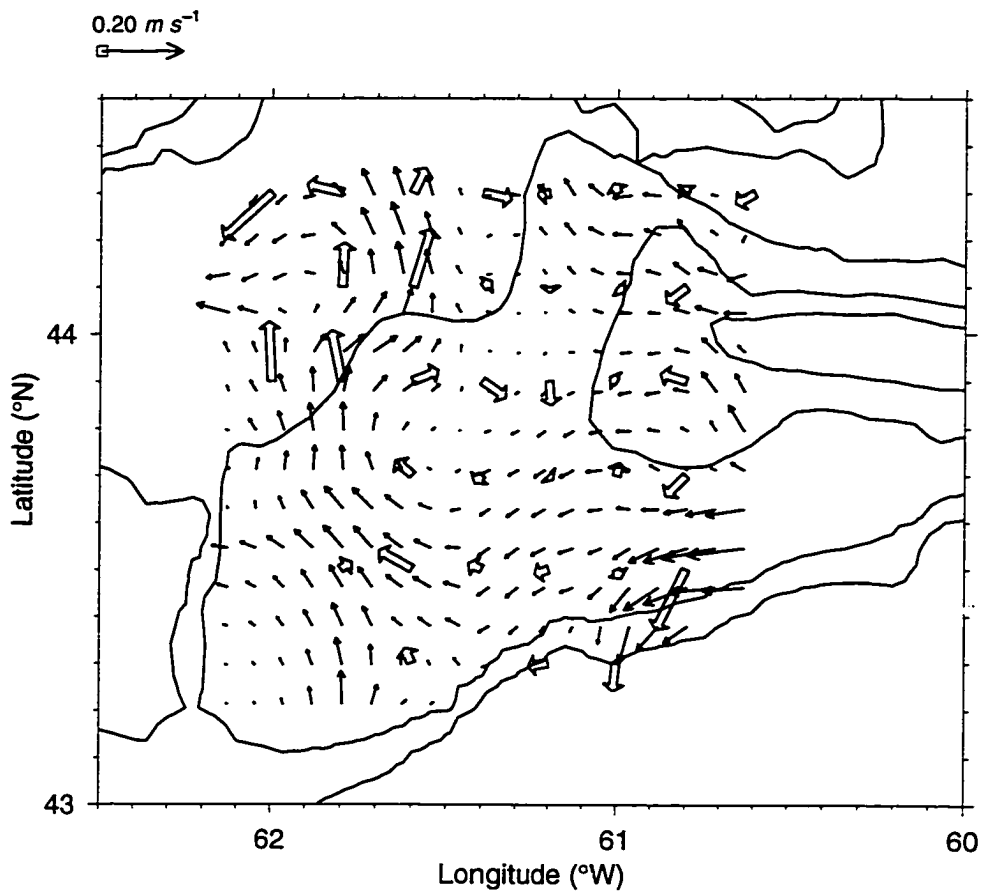


Figure 3.11: Comparison of the residual circulation on Western Bank derived from the ADCP (open arrows) and a diagnostic calculation of the flow field (solid arrows). The ADCP residual is prepared by binning the residual (tide and wind effects removed) into 0.2° latitude/longitude boxes. The diagnostic calculation (Sheng and Thompson 1996) estimates the surface circulation using 177 density profiles collected during the April 1992 cruise (Figure 3.1).

was tested. Some small changes are evident: the R^2 is slightly higher than the control (but less than 2%), and the magnitude of the ellipses is slightly larger in the data sparse regions (but only by a maximum of 4%). Overall, we conclude that the results are robust over a range of grid resolutions.

3.5 Summary and Discussion

A conceptually straightforward method for de-tiding the velocity series obtained from a ship-borne ADCP has been presented. The method fits a limited area tidal model, based on shallow water dynamics, to the ADCP observations of depth averaged velocity. More generally, the method provides a means to carry out harmonic analysis on a time-space series of velocity. This is offered as an alternative to the procedure of Candela *et al.* (1992), where arbitrary spatial interpolation functions were fitted to the ADCP velocities.

Application of the method to estimating tidal flows from ship ADCP data from the Western Bank region of the Scotian Shelf has been successful. The estimated tidal ellipses obtained from the ADCP record match those of the current meter records and regional tidal maps. The weaker tidal constituents are less consistent with these independent sources but within estimated errors limits. The residual field, after removal of the tides and the wind-driven circulation, clearly showed a gyre centered on Western Bank and a persistent northward flow to its west. These features are also confirmed using other data sources obtained on the April 1992 cruise.

Some additional considerations arise in the tidal estimation of a time-space series. One concern is the ability of the analysis to separate closely spaced frequency bands when a spatial dimension is included in the series. Another concern is that the tides may be aliased by the movement of the ship. To test for this possibility in the Western Bank case, a simulated velocity time series was generated by sampling the flow field of the diagnostic calculation in Figure 3.11 along the cruise track of the ship. Performing the tidal analysis on this series showed no evidence of any spurious tidal signals.

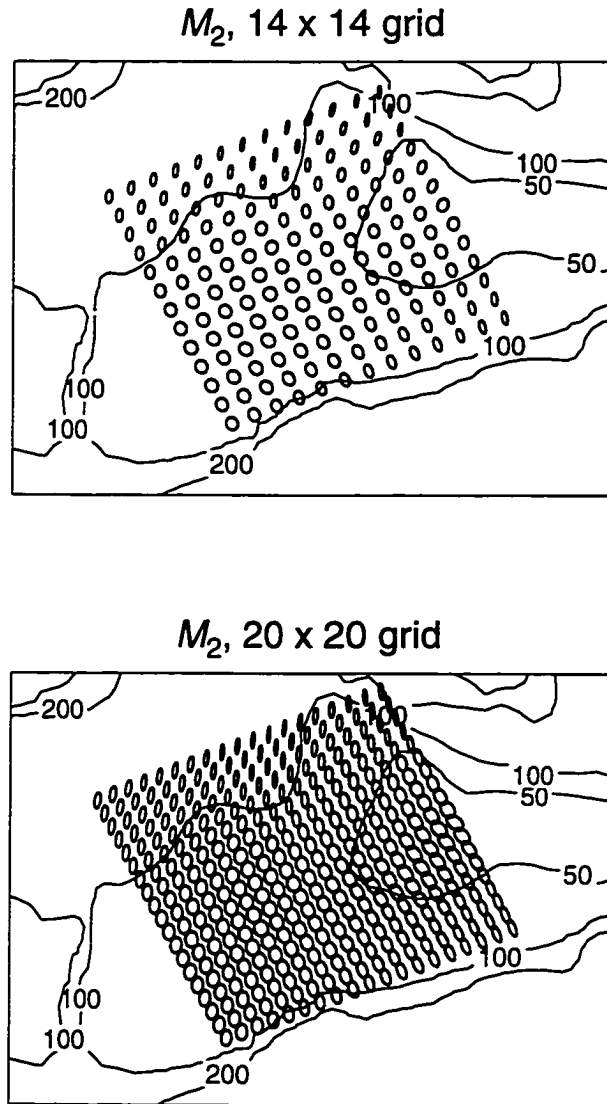


Figure 3.12: Comparison of estimated M_2 tidal ellipses for a 14×14 grid and a 20×20 grid. Ellipses are scaled to represent one-third of the M_2 tidal excursion.

The use of generalized least squares regression has a correspondence with other commonly used assimilation techniques. The general problem solved here is one of minimizing the weighted squared observation/model discrepancy

$$J = (\mathbf{z} - \hat{\mathbf{z}})^T \Sigma^{-1} (\mathbf{z} - \hat{\mathbf{z}}). \quad (3.15)$$

The model counterparts to the data $\hat{\mathbf{z}}$ are constrained to satisfy the shallow water equations (3.1). The condition making this applicable to tides is that the normal flows across the open boundary are further constrained to be periodic in time with known frequency but an unknown complex amplitude \mathbf{b} . In this paper, the estimate $\hat{\mathbf{b}}$ which minimizes J is found by transforming the shallow water equations into the frequency domain and expressing them as a discrete boundary value problem in matrix form. The model counterparts to the data are then written in terms of the unknown \mathbf{b} . Differentiation of J with respect to the unknown \mathbf{b} and setting the result equal to zero yields an estimate for the boundary state $\hat{\mathbf{b}}$.

Consider another case in which it is desired to use a time stepping numerical tidal model (e.g. Lardner 1993). To minimize J subject to such a model posed in the time domain, optimal control/adjoint approaches can prove useful (see Section 2.4.3). The minimum of J with respect to \mathbf{b} can be found iteratively using an adjoint based smoothing method (Section 2.4.3). Each iteration involves a cycle comprising a forward integration of the model equations, a backward integration of the adjoint equations and calculation of the gradient of J with respect to the unknown quantities. The gradient information is used along with a descent algorithm to converge on the minimum (see, for example, Thacker and Long 1988). We have implemented the above variational procedure for the present problem and found it to be quite inefficient. This is illustrated below.

The above variational procedure was applied to the tidal analysis of the ADCP data using a fully explicit finite difference form of the shallow water equations (3.1) on the identical model grid used for the frequency domain model. The time step, set by the CFL condition, was 120 seconds. A forward integration of this model required about 5 model days to reach a periodic steady state and a further 14 model

days to produce model counterparts to the data. The adjoint equations required a similar amount of computation time. A minimum of 20 iterations were required to achieve convergence and the final estimates for \mathbf{b} and \mathbf{x} were comparable to those found by the statistical-dynamical approach in Section 3.4. Error analysis of the estimates would involve computing the second derivatives of J with respect to \mathbf{b} , i.e. the Hessian matrix. In addition to the large computational cost of this compared to the our approach, numerical experimentation also pointed to additional concerns related to proper radiation of wave energy in such a boundary control problem.

Other solution techniques may become necessary if the dynamical assumptions relating to the use of the linearized, depth averaged shallow water equations for describing the barotropic tide are not satisfied. These include such processes as tidal rectification and interactions between the barotropic and baroclinic tides. One possible approach in these cases is to retain the linear model and introduce model errors explicitly into the analysis. Another possibility is to use a more complicated model. For instance, to analyze the internal tides in the deep water regions surrounding Western Bank from the current profiles would require at least a two layer model. Similarly, including nonlinear advective terms would transform the problem into one of nonlinear regression and suggest the use of more general variational data assimilation techniques.

In summary, the method for de-tiding ADCP data appears robust. The generalized least squares regression of Section 2.1 has been applied directly by taking advantage of the fact that the governing equations can be posed sensibly in the frequency domain leading to manageable dimensions for the constituent matrices. Tidal extraction relies solely on the ADCP data and requires no external sources of tidal information. The dynamics provide interpolation functions and are thus preferable to other basis functions. Realistic error estimates are also produced as an intermediate step in the regression calculation. In addition, for the application shown in this chapter no arbitrary regularization terms, such as spatial smoothness, were required to get well-conditioned solutions. Overall, the method should allow an ADCP tidal

residual to be obtained efficiently and be minimally contaminated by the de-tiding procedure.

Chapter 4

Forecasting Coastal Circulation using an Approximate Kalman Filter

In recent years, operational prediction of coastal circulation has received increased attention. Accurate estimates of circulation are required to address such coastal zone problems as storm surge forecasting, predicting oil spill and iceberg trajectories, and assessing the impact of point source pollutants. Currently, a number of efforts are underway to develop operational nowcasting and forecasting systems for the coastal ocean (e.g. Heemink and Van-Stijn 1993, Griffin and Thompson 1995, Aikman *et al.* 1996).

The under-sampled nature of coastal and continental shelf regions makes circulation estimates difficult. Data from remote sensing, coastal tide gauges, and field measurements, including current meters, CTDs, drifters, and ship borne acoustic Doppler current profilers (ADCP), provide a basis on which to infer flow fields. However, to sensibly temporally and spatially map these data it is also necessary to consider ocean dynamics in the form of numerical models. These models of the coastal ocean must cope with irregular coastlines, highly variable bathymetry, and extensive open boundaries.

As illustrated in Chapter 2, data assimilation techniques are divided into two basic categories, filtering and smoothing. The choice of a method depends on whether one is interested in nowcasts/forecasts or hindcasts, respectively. The Kalman filter (Kalman 1960) is potentially well suited to the problem of nowcasting and forecasting circulation in the coastal ocean. Estimates of the ocean state through time are obtained using a recursive algorithm and have the property of being maximum likelihood (and minimum variance) estimates in the case of linear models with Gaussian noise. A notable example of Kalman filter application is the operational storm surge prediction scheme implemented in the North Sea (Heemink and Kloosterhuis 1990).

Oceanographic application of the Kalman filter must address two important issues. First, the computational burden imposed by the large dimension of most ocean models makes implementation of the full Kalman filter problematic. Section 2.3.1 reveals that the Kalman filter algorithm involves multiplying and inverting matrices with a size corresponding to the number of elements in the ocean state vector. Second, it is difficult to accurately specify the statistics required by the Kalman filter. Information on the measurement errors (instrument noise and errors in interpolating the observations to the model variables and to the model grid) and system noise (neglected dynamics, numerical approximations, and stochastic forcing) are often poorly known and yet important to filter performance (Jiang and Ghil 1992, Daley 1992).

An assumption often made to reduce the computational burden of the Kalman filter is that the system under consideration is in a statistical steady state. This requires an observation array with fixed measuring locations and time-invariant dynamics, whereupon an asymptotic limit for the Kalman gain matrix may be efficiently pre-computed (Anderson and Moore 1979, chapter 4). This approximation has been used by Heemink and Kloosterhuis (1990) in their storm surge forecasting scheme. Fukumori *et al.* (1993) also found that the approximation gave good results when using an observation array designed to mimic the assimilation of altimetric data into a general circulation model.

Few general simplification schemes exist in the case where the steady state approximation is not valid. The calculation of the forecast error covariance (2.24) can sometimes be simplified. For example, Dee (1991) presents a scheme whereby the forecast errors are advected by the estimated flow fields. Evensen (1994) uses Monte Carlo methods to evaluate the forecast error in what has come to be known as an ensemble Kalman filter. Another approach is the coarse grid approximation suggested by Fukumori and Malanotte-Rizzoli (1995) for evaluating the covariance and gain equations (2.24)-(2.26). An alternative to Kalman filtering is found in filters of the stochastic approximation type (Gelb 1974), such as the nonlinear adaptive filters used by Hoang *et al.* (1994).

In this chapter, an approximate Kalman filter is proposed for the nowcasting and forecasting of coastal ocean circulation. Representation of a circulation model in terms of its dynamical modes is used to reduce the effective dimension of the model. The main premise is that important dynamical features of the system can be captured with a relatively small number of modes. Such a simplification greatly enhances the computational efficiency of the Kalman filter, yet retains the aspects of the dynamics necessary for state estimation and error propagation. Using modes may circumvent some of the issues associated with the spatial regularity of the Kalman filter (Bennett and Budgell 1987) and might also prove useful for nonlinear assimilation problems.

The outline of the chapter is as follows. Section 4.1 reviews the background material on the basic linear theory of the Kalman filter in the context of coastal forecasting. In Section 4.2, the modal representation of the dynamics and Kalman filter is derived and interpreted. Section 4.3 addresses the issue of selecting an appropriate subset of the modes on which to base a reduced ocean model and corresponding Kalman filter. Controllability and observability are first discussed and a selection criterion is then proposed to identify those modes preferentially excited by the forcing. Section 4.4 illustrates the application of the approximate filter to a prototype shallow water model of the Scotian Shelf region off the east coast of Canada. The performance of the approximate filter is assessed using wind and lateral boundary forcing with both fixed

and moving observation arrays. Discussion and conclusions follow in Section 4.5.

4.1 Background

In this section, the linear theory of the Kalman filter is briefly reviewed in the context of coastal forecasting. Chapter 2 and Appendix B offer a more complete development of the filtering theory and repetition has been avoided where possible.

The state space representation of a linearized, numerical coastal model takes the form of a difference equation

$$\tilde{\mathbf{x}}_t = \tilde{\mathbf{D}}_t \tilde{\mathbf{x}}_{t-1} + \mathbf{G}_t \mathbf{e}_t^m. \quad (4.1)$$

The state vector $\tilde{\mathbf{x}}_t$ contains the prognostic variables defined on the model grid with subscripts denoting time. The operator $\tilde{\mathbf{D}}_t$ corresponds to the discretized governing equations and allows the model fields to be stepped forward in time. The system noise, or forcing, is represented by \mathbf{e}_t^m and is projected onto the model state by the matrix \mathbf{G}_t . This forcing can include both deterministic and stochastic components, e.g.

$$\mathbf{e}_t^m = \boldsymbol{\mu}_t + \mathbf{w}_t,$$

where $\boldsymbol{\mu}_t$ represents a time varying mean and \mathbf{w}_t is a stochastic process. Heemink (1986) outlines the development of such a Kalman filter for storm surge prediction where the tide is considered deterministic and the effect of the surge is assumed to be an additive stochastic process. While $\boldsymbol{\mu}_t$ could easily be included in the development in this chapter, we have chosen, for simplicity, to assume that it is equal to zero.

For most oceanographic problems, the dynamics are time-invariant. The form of the governing equations are not a function of time. A time index has been included on $\tilde{\mathbf{D}}_t$ in the linear model (4.1) to allow for the possibility of a dynamics operator which varies through time. In oceanographic problems this occurs mainly in the case where a nonlinear model is successively linearized about a time varying base state (Section 2.2). In practice, the matrix $\tilde{\mathbf{D}}_t$ can be determined either by writing the finite

difference form of the governing equations directly in the required matrix/vector, or through an impulse response technique using an existing model. (The former approach was used in Chapter 3, while the latter approach is outlined in Section 4.4).

For realistic oceanographic applications, the stochastic forcing is often serially correlated. For example, forcing functions for limited-area coastal models include surface wind stress and flows across the open boundaries, both of which are correlated in time. Optimality of the Kalman filter can be severely compromised if serial correlation is not properly accounted for in the specification of the system noise (Daley 1992). A possible approach to this problem is to represent the stochastic process \mathbf{e}_t^m as an autoregressive, moving-average (ARMA) process. This can be represented in the Markovian form (4.1) and driven by a purely white noise process (e.g. Anderson and Moore 1979, chapter 2). For the simple case in which the forcing takes the form of an auto-regressive process of order 1, we may modify (4.1) as follows:

$$\begin{pmatrix} \tilde{\mathbf{x}} \\ \mathbf{e}^m \end{pmatrix}_t = \begin{pmatrix} \tilde{\mathbf{D}}_t & \mathbf{G}_t \\ \mathbf{0} & \mathbf{A} \end{pmatrix} \begin{pmatrix} \tilde{\mathbf{x}} \\ \mathbf{e}^m \end{pmatrix}_{t-1} + \mathbf{e}_t, \quad (4.2)$$

or, more concisely,

$$\mathbf{x}_t = \mathbf{D}_t \mathbf{x}_{t-1} + \mathbf{e}_t. \quad (4.3)$$

In the above, \mathbf{A} is a matrix containing the autoregressive coefficients. The system noise is represented by \mathbf{e}_t which is assumed zero-mean and serially uncorrelated. The elements of \mathbf{e}_t corresponding to \mathbf{e}_t^m drive the AR(1) forcing; elements of \mathbf{e}_t corresponding to $\tilde{\mathbf{x}}_t$ account for model error.

Oceanographic measurements available for assimilation are related to the state vector through the equation

$$\mathbf{z}_t = \mathbf{H}_t \mathbf{x}_t + \mathbf{e}_t^o. \quad (4.4)$$

Here, \mathbf{H}_t transforms the observed variables to the prognostic variables of the model and interpolates the observation array to the model grid. The observation error term \mathbf{e}_t^o includes errors in these procedures, in addition to instrument noise. The time dependence of \mathbf{H}_t is introduced mainly to allow for time varying observation arrays

such as the ship-borne ADCP which measures currents along the cruise track of the ship. Note that \mathbf{H}_t operates on \mathbf{x}_t , rather than $\bar{\mathbf{x}}_t$, and allows for the possibility of directly observing the forcing.

The Kalman filter combines the model (4.3) and observations (4.4) to give an optimal nowcast of the ocean state. At time t , the Kalman filter estimate of the ocean state $\hat{\mathbf{x}}_t$ is

$$\hat{\mathbf{x}}_t = \bar{\mathbf{x}}_t + \mathbf{K}_t(\mathbf{z}_t - \mathbf{H}_t\bar{\mathbf{x}}_t), \quad (4.5)$$

where the model forecast $\bar{\mathbf{x}}_t$ is given by

$$\bar{\mathbf{x}}_t = \mathbf{D}_t\hat{\mathbf{x}}_{t-1}. \quad (4.6)$$

The model forecasts are nudged towards the observations according to the weight given by the gain matrix \mathbf{K}_t . The gain matrix is determined in parallel with (4.5) and (4.6) according to (2.24)-(2.26) in Section 2.3.1. These equations also determine $\text{var}(\hat{\mathbf{x}}_t - \mathbf{x}_t)$ and $\text{var}(\bar{\mathbf{x}}_t - \mathbf{x}_t)$. k -step ahead forecasts are obtained by application of (4.6) using the current estimate $\hat{\mathbf{x}}_t$ as the initial condition. The Kalman filter algorithm is solved in a recursive nature, which is an important feature for its application to operational coastal forecasting.

The information required for implementation of the Kalman filter are the mean and covariance of the initial state $(\bar{\mathbf{x}}_0, \mathbf{M}_0)$, the system noise $(\bar{\mathbf{e}}_t, \mathbf{Q}_t)$, and the measurement noise $(\bar{\mathbf{e}}_t^o, \mathbf{R}_t)$. As mentioned, it can be difficult to specify \mathbf{Q}_t and \mathbf{R}_t . The influence of the initial conditions diminishes as the system moves away from the initial state (i.e. within a few spindown times of the model), and so the specification of \mathbf{M}_0 may not be crucial in some circumstances. If it is assumed that these statistics are known, and the processes \mathbf{e}_t^o and \mathbf{e}_t are Gaussian, independent of one another and uncorrelated in time, the Kalman filter estimates are maximum likelihood estimates of the state.

Implementation of a full Kalman filter is difficult for realistic oceanographic problems, mainly due to its computational requirements. For an ocean model, the number of elements n in the model state vector is roughly the number of grid points in the

model multiplied by the number of prognostic variables. A relatively small coastal model with grid size 30×30 with 4 levels (or spectral coefficients) in the vertical and prognostic in horizontal velocity and sea level would require a state vector with $O(10^4)$ elements. If we assume a grid spacing of 10 km and a maximum depth of 400 m, the time step of the model, set by the CFL condition, is about 2.5 minutes. At every time step, model forecasts and the computationally costly matrix calculations associated with (2.24)-(2.26) must be carried out. These facts, coupled with the uncertainty in specifying the input statistics, argues strongly for the use of approximate, or suboptimal, filters in oceanographic data assimilation.

4.2 Modal Representation

In this section, a reduced dimension Kalman filter is proposed for operational coastal data assimilation. In the case of a finite difference ocean model, significant reduction in the model dimension is not usually possible by direct means. The grid resolution of the model is chosen to resolve the scales of interest, and the model time step is then set by stability considerations. The coarse grid approximation of Fukumori and Malanotte-Rizzoli (1995) for the variance and gain calculations (2.24)-(2.26) offers one possible simplification. In their case, a model of an idealized Gulf Stream, which is linearized about the mean jet trajectory, is used to test their approach, and encouraging results were achieved. An alternative method of scale truncation (and consequent dimension reduction) is offered based on a suitable subset of the eigenvectors (modes) of $\tilde{\mathbf{D}}_t$. An approximate Kalman filter can then be based on these modes.

4.2.1 Dynamics

To describe the approach, we make the simplifying assumption that $\tilde{\mathbf{D}}_t$ and \mathbf{G}_t in (4.1) are time-invariant and drop their subscripts. The next step is to define a new basis for the vector space that contains the model state $\tilde{\mathbf{x}}_t$. Suppose that an alternative basis set, represented by the columns of $\tilde{\Phi}$ exists. Furthermore, assume that these

basis sets are related by the linear transformation

$$\tilde{\mathbf{x}}_t = \tilde{\Phi} \tilde{\boldsymbol{\alpha}}_t. \quad (4.7)$$

where $\tilde{\boldsymbol{\alpha}}_t$ represents the coordinates in the new basis.

Suppose we choose the columns $\tilde{\boldsymbol{\phi}}$ of $\tilde{\Phi}$ to be the right-hand eigenvectors of $\tilde{\mathbf{D}}$, i.e.

$$\tilde{\mathbf{D}} \tilde{\Phi} = \tilde{\Phi} \tilde{\Lambda}, \quad (4.8)$$

where $\tilde{\Lambda}$ is a diagonal matrix of its eigenvalues $\tilde{\lambda}$. Substituting (4.7) into (4.1) yields

$$\tilde{\Phi} \tilde{\boldsymbol{\alpha}}_t = \tilde{\mathbf{D}} \tilde{\Phi} \tilde{\boldsymbol{\alpha}}_{t-1} + \mathbf{G} \mathbf{e}_{t-1}^m.$$

Using (4.8) and pre-multiplying by $\tilde{\Phi}^{-1}$ gives an equation for the evolution of the modal coefficients

$$\tilde{\boldsymbol{\alpha}}_t = \tilde{\Lambda} \tilde{\boldsymbol{\alpha}}_{t-1} + \tilde{\Phi}^{-1} \mathbf{G} \mathbf{e}_{t-1}^m. \quad (4.9)$$

This is an alternative description of the model dynamics in (4.1). In fact, (4.9) is identical to the original system since only a change in coordinate systems has occurred.

This modal decomposition represents a diagonalization of the dynamics matrix $\tilde{\mathbf{D}}$, i.e.

$$\tilde{\Phi}^{-1} \tilde{\mathbf{D}} \tilde{\Phi} = \tilde{\Lambda}.$$

This similarity transformation is possible provided that the eigenvalues of $\tilde{\mathbf{D}}$ are distinct, which implies that the eigenvectors are linearly independent and $\tilde{\Phi}^{-1}$ exists (Noble and Daniel 1977, Theorem 8.2). However, distinct eigenvalues are not guaranteed for a general $n \times n$ matrix, such as the non-symmetric $\tilde{\mathbf{D}}$ associated with the discretized dynamics. If $\tilde{\mathbf{D}}$ has repeated eigenvalues, the Jordan canonical form of a matrix allows for transformation of $\tilde{\mathbf{D}}$ into a block diagonal matrix, and the existence of a modal evolution equation of the form (4.9) (Barnett 1990, section 8.1). The main difference is that modes, or generalized eigenvectors, associated with each repeated eigenvalue are coupled (hence a block diagonal structure for $\tilde{\Lambda}$). In this chapter we assume, for simplicity, that the eigenvalues of $\tilde{\mathbf{D}}$ are distinct, but recognize that a slightly modified development is necessary if repeated eigenvalues are present.

4.2.2 Interpretation of the Modes

The eigenvalue-eigenvector pairs $(\lambda, \tilde{\phi})$ represent the dynamic modes, or principal oscillation patterns (Hasselmann 1988), of the model (4.1). For a typical coastal ocean model, the dynamics matrix $\tilde{\mathbf{D}}$ is obtained by finite differencing of the governing partial differential equations. It has real-valued elements and is non-symmetric. This form is typical of models with multiple variables and/or more than one spatial dimension. In this case, the eigenvalues and eigenvectors are both real and complex (in which case they occur in conjugate pairs). The free modes considered here are time-evolving patterns supported by a numerical circulation model and depend on the discretized governing equations as well as coastal and open boundary conditions. For simple dynamics, analytic modes can often be derived. These would be comparable to dynamical modes derived from a numerical model provided that the model is a good approximation of the (continuous) system. Gallagher *et al.* (1991) provides a detailed review of modal properties, some of which are described below.

To illustrate some properties of the dynamical modes, consider the evolution of the one component of $\tilde{\alpha}$ in (4.9) in the absence of any forcing, i.e

$$\alpha_t = \lambda \alpha_{t-1}. \quad (4.10)$$

Furthermore, assume that $\alpha_0 = 1$ which describes an initial value problem with unit forcing at the initial time. In this case,

$$\alpha_t = \lambda^t,$$

and the modal behaviour depends on whether the eigen-pairs are real or complex.

For real eigen-pairs, the contribution of a single mode to the state is then given by

$$\begin{aligned} \tilde{\mathbf{x}}_t &= \alpha_t \tilde{\phi} \\ &= \lambda^t \tilde{\phi}. \end{aligned}$$

Therefore, real modes represent spatial patterns in the prognostic variables (described by the eigenvectors) which (i) persist if $|\lambda| = 1$, or (ii) decay if $|\lambda| < 1$, with an e-folding timescale $\tau = -\Delta t / \ln(|\lambda|)$ (where Δt is the time step of the model).

For complex modes, the set of $(\alpha, \lambda, \tilde{\phi})$ occur in conjugate pairs. The contribution to the state by a mode is given by

$$\begin{aligned}\tilde{\mathbf{x}}_t &= \alpha_t \tilde{\phi} + [\alpha_t \tilde{\phi}]^* \\ &= 2|\lambda|^t [\Im\{\tilde{\phi}\} \cos(\theta t) + \Re\{\tilde{\phi}\} \sin(\theta t)]\end{aligned}$$

where $*$ represents complex conjugation and

$$\theta = \tan^{-1} \left(\frac{\Im(\lambda)}{\Re(\lambda)} \right).$$

Complex modes therefore represent patterns which oscillate smoothly between the real and imaginary parts of $\tilde{\phi}$ in the following order

$$\dots \rightarrow \Re\{\tilde{\phi}\} \rightarrow -\Im\{\tilde{\phi}\} \rightarrow -\Re\{\tilde{\phi}\} \rightarrow \Im\{\tilde{\phi}\} \rightarrow \Re\{\tilde{\phi}\} \rightarrow \dots$$

such that a complete cycle is completed in a period given by $T = 2\pi/\theta$ and the decay rate is determined by $|\lambda|$. (As an example, consider the mode shown in Figure 4.2)

We now briefly examine the spectral properties of the modes. The spectrum of a single modal coefficient α is determined, using (4.9), as

$$\Gamma^\alpha(\omega) = \frac{\Gamma^f(\omega)}{|e^{i\omega} - \lambda|^2}, \quad (4.11)$$

where a time-step of unity is assumed. Here, Γ^α represents the spectral density of the modal coefficient and Γ^f is the spectral density of the corresponding element in the forcing vector $\tilde{\Phi}^{-1} \mathbf{G} \mathbf{e}_i^m$. If the forcing is white ($\Gamma_f(\omega) = 1$, for all ω), the spectrum has a maximum value of $(1 - |\lambda|)^{-2}$ at $\omega = \theta$, the resonant frequency. The width of the spectrum is controlled by $|\lambda|$ indicating that increased damping (decay) of a mode results in a broader spectrum. The overall implication is that the spectral properties of the forcing can preferentially excite certain modes.

One final point to note is the relationship between the eigenvectors and eigenvalues of $\tilde{\mathbf{D}}$ of the original model (4.1), and \mathbf{D} in the augmented model (4.3). The eigen-equation for the augmented system

$$\mathbf{D}\phi = \phi\lambda$$

takes the form

$$\begin{pmatrix} \tilde{\mathbf{D}} & \mathbf{G} \\ \mathbf{0} & \mathbf{A} \end{pmatrix} \begin{pmatrix} \tilde{\phi} \\ \phi_f \end{pmatrix} = \lambda \begin{pmatrix} \tilde{\phi} \\ \phi_f \end{pmatrix}. \quad (4.12)$$

This reduces to the system of equations

$$\tilde{\mathbf{D}}\tilde{\phi} + \mathbf{G}\phi_f = \lambda\tilde{\phi} \quad (4.13)$$

$$\mathbf{A}\phi_f = \lambda\phi_f. \quad (4.14)$$

By inspection, we see that

$$\phi = \begin{pmatrix} \tilde{\phi} \\ 0 \end{pmatrix}$$

satisfies (4.13)-(4.14) and therefore is an eigenvector of \mathbf{D} . Note that it is also an eigenvector of $\tilde{\mathbf{D}}$ augmented with a zero vector. The number of eigenvectors of this form corresponds to the number of rows (or columns) of $\tilde{\mathbf{D}}$. Therefore, carrying out an analysis of the modal properties in terms of either the original or augmented models proves to be equivalent. The remaining eigenvectors are associated with the model forcing.

4.2.3 Kalman Filter Equations

It is straightforward to use the Kalman filter with the modal representation given by (4.9). As a first step, following (4.3), we augment (4.9) with the AR(1) forcing which gives

$$\begin{pmatrix} \tilde{\alpha} \\ \mathbf{e}^m \end{pmatrix}_t = \begin{pmatrix} \tilde{\Lambda} & \tilde{\Phi}^{-1}\mathbf{G} \\ \mathbf{0} & \mathbf{A} \end{pmatrix} \begin{pmatrix} \tilde{\alpha} \\ \mathbf{e}^m \end{pmatrix}_{t-1} + \mathbf{e}_t. \quad (4.15)$$

This is represented more concisely as

$$\alpha_t = \Lambda \alpha_{t-1} + e_t. \quad (4.16)$$

To determine the Kalman filter equations, the original model in (4.3) is replaced with (4.16). The observation operator (4.4) and Kalman gain are modified accordingly. The Kalman filter equations given in Section 2.3.1 then becomes

$$\hat{\alpha}_t = \bar{\alpha}_t + \mathbf{K}_t^\alpha (z_t - \mathbf{H}\Phi_t \bar{\alpha}_t) \quad (4.17)$$

where

$$\bar{\alpha}_t = \Lambda_t \hat{\alpha}_{t-1} \quad (4.18)$$

with

$$\text{var}(\bar{\alpha}_t - \alpha_t) \equiv \mathbf{M}_t^\alpha = \Lambda_t \mathbf{P}_{t-1}^\alpha \Lambda_t^T + \mathbf{Q}_t^\alpha \quad (4.19)$$

$$\text{var}(\hat{\alpha}_t - \alpha_t) \equiv \mathbf{P}_t^\alpha = (\mathbf{M}_t^{\alpha^{-1}} + \Phi^T \mathbf{H}_t^T \mathbf{R}_t^{-1} \mathbf{H}_t \Phi)^{-1} \quad (4.20)$$

$$\mathbf{K}_t^\alpha = \mathbf{P}_t^\alpha \Phi^T \mathbf{H}_t^T \mathbf{R}_t^{-1}. \quad (4.21)$$

The use of complex numbers (characterizing modal coefficients) as elements of the state vector in the Kalman filter presents no conceptual difficulty (Peterson 1968). In addition, the derivation of the Kalman filter in terms of generalized least squares regression given in Chapter 2 holds for the complex case (see Brillinger 1981, section 3.7 for the details of converting the real-valued matrices in the regression equation to their complex counterparts).

4.3 Selection of the Modes

The following properties of the modal representation (4.9) are useful for designing an approximate, or reduced dimension, Kalman filter:

1. The modal coefficients $\tilde{\alpha}_t$ evolve independently since $\tilde{\Lambda}$ is diagonal. The evolution of each mode satisfies both the governing equations and the boundary conditions as well as being dynamically uncoupled from other modes.

2. Each modal coefficient in $\tilde{\alpha}_t$ describes a spatial pattern with the resolution of the original model (4.1).
3. The model time step now depends on the time scale on which the significant variation in the forcing occurs and the spindown time (decay scale) associated with the modes. It is no longer a direct consequence of stability considerations, such as the CFL condition, which limit the time step in the original model. Larger time steps may be possible.

Each of these properties offer advantages for designing and implementing an approximate Kalman filter. The first property allows for a straightforward reduction in the model based on choosing a subset of the modes. That is, any modes deemed unimportant can simply be neglected in (4.9) by truncating $\tilde{\alpha}_t$ and $\tilde{\Lambda}$. In a finite difference implementation this would not be possible. The second property implies that truncating the model maintains the spatial resolution. Finally, the third property offers the possibility of increased computational efficiency in the numerical integration of the model through time.

To be more specific about model reduction, suppose that the modes (eigenvalues and eigenvectors) are divided into 2 sets, denoted $(\tilde{\Lambda}_1, \tilde{\Phi}_1)$ and $(\tilde{\Lambda}_2, \tilde{\Phi}_2)$. If the first set is chosen as a basis for a reduced dimension ocean model, then the truncated second set comprises the unresolvable modes of the system and defines the null space of the reduced ocean model. The uncoupled equations for $\tilde{\alpha}_1$ and $\tilde{\alpha}_2$ implies that any energy present in the truncated subspace remains there. This satisfies the criteria for system reduction put forth by Fukumori and Malanotte-Rizzoli (1995).

The key property of this modal subset $(\tilde{\Lambda}_1, \tilde{\Phi}_1)$ is that it should describe both the ocean state at a given time, as well as its temporal evolution. In the remainder of this section, we present a means for choosing a suitable subset of the modes based on preserving these properties. First, the ability to observe and control the modes is discussed. Second, the partitioning of the energy from the forcing amongst the modes is considered. These considerations allow modes preferentially excited by the forcing to be identified, and an approximate dynamical model to be constructed.

4.3.1 Observability and Controllability

As a first step in choosing a subset of the modes, we eliminate from our consideration any modes that cannot be observed or controlled. Observability and controllability are important properties of any data assimilation problem and take on a particularly simple form when using a modal representation of the dynamics. They are introduced and discussed below.

The ability to observe the dynamic system (4.1) using a given observational array can be assessed on a quantitative basis. If $\tilde{\phi}$ is an eigenvector of $\tilde{\mathbf{D}}$ with a non-zero eigenvalue, then that mode is said to be observable if $\tilde{\mathbf{H}}\tilde{\phi} \neq 0$. (Here, $\tilde{\mathbf{H}}$ is that portion of the \mathbf{H} in (4.4) which corresponds to $\tilde{\mathbf{x}}_t$). This states that to observe a mode, it must have a non-zero projection on the observation space. In other words, the nodes of a mode must not coincide with the observing locations.

The ability of the forcing to drive the dynamic system (4.1) to an arbitrary state is referred to as the controllability of the system. This is stated as follows. Suppose that ϕ^a is an eigenvector of $\tilde{\mathbf{D}}^T$ with a non-zero eigenvalue. If $\mathbf{G}^T\phi^a \neq 0$, the mode is controllable. If all the modes are controllable, so is the overall system. To further interpret the controllability condition we consider below the problem of controlling the model forcing such that the system is driven to a state which best fits the observations.

Recall the control problem of Section 2.4.3. Define a cost function J which measures the squared error of the observation/model misfit and treat the model (4.1) as a strong constraint. This leads to the Lagrange function

$$L = J + \sum_t \gamma_{t-1}^T (\tilde{\mathbf{x}}_t - \tilde{\mathbf{D}}_t \tilde{\mathbf{x}}_{t-1} - \mathbf{G} \mathbf{e}_{t-1}^m). \quad (4.22)$$

Differentiating L with respect to $\tilde{\mathbf{x}}_t$ yields the adjoint model

$$\gamma_{t-1} = \tilde{\mathbf{D}}^T \gamma_t - \frac{\partial J}{\partial \tilde{\mathbf{x}}_t} \quad (4.23)$$

and differentiating with respect to \mathbf{e}_t^m gives the gradient in terms of the γ_{t-1} , i.e.

$$\frac{\partial J}{\partial \mathbf{e}_t^m} = \mathbf{G}^T \gamma_t. \quad (4.24)$$

In the control problem this gradient information is used to adjust \mathbf{e}_t^m such that J is made a minimum.

Now, the ϕ^a are modes of the adjoint operator $\tilde{\mathbf{D}}^T$ and provide an alternative basis for the space spanned by γ_t , in exactly the same way that $\check{\phi}$ does for $\tilde{\mathbf{x}}_t$. Suppose that only a single mode ϕ^a of the adjoint operator is excited by $\partial J/\partial \tilde{\mathbf{x}}_t$ (the forcing for the adjoint model and representing the observation/model misfit). In this case, the gradient (4.24) becomes

$$\frac{\partial J}{\partial \mathbf{e}_t^m} = \mathbf{G}^T \phi^a \alpha^a, \quad (4.25)$$

where α^a represents the coefficient of the adjoint mode. Now if $\mathbf{G}^T \phi^a = 0$, the gradient (4.25) is zero and there is no way to adjust the controls such that J is made smaller. However, an observation/model misfit, represented in $\partial J/\partial \tilde{\mathbf{x}}_t$, clearly exists and ϕ^a defines a null space in the adjoint model. The system cannot be controlled, in the sense of adjusting the model to rectify the misfit to observations, with an adjoint mode having no projection on the forcing.

In summary, elimination of the modes that cannot be controlled or observed corresponds to neglecting modes that are in the null space of the observations or forcing, respectively. All remaining modes describe some potentially important part of the overall dynamics and cannot be eliminated without further *a priori* assumptions. The next section discusses a selection procedure in which the modes are ordered based on an energy criterion.

4.3.2 Energetics

For conventional limited-area coastal models, surface wind or lateral boundary forcing is prescribed based on data, climatology and regional knowledge. We assume that the dominant source of randomness in the model (4.1) or (4.9) is related to the stochastic nature of this forcing \mathbf{e}_t^m . In other words, the system described by the coastal model is primarily driven by wind or boundary forcing. This does not imply that other model error processes are zero (they should, in fact, be included in the specification of \mathbf{Q}_t), but rather that it is convenient to neglect them for the purposes of modal

selection. Outlined below is a procedure for choosing the modes preferentially excited by this forcing.

Suppose that we are given, or have approximated, the stochastic forcing as an ARMA, or a general linear, process. The cross-covariance of the forcing is then known and the power spectrum is determined by taking its Fourier transform. Denote the spectral matrix of the forcing \mathbf{e}_t^m by $\Gamma^f(\omega)$ with ω being the frequency. The spectral matrix of the modal coefficients $\Gamma^\alpha(\omega)$ is determined through (4.9) to be

$$\Gamma^\alpha(\omega) = \mathbf{C}\Gamma^f(\omega)\mathbf{C}^T, \quad (4.26)$$

where

$$\mathbf{C} = (\mathbf{I}e^{i\omega} - \tilde{\mathbf{\Lambda}})^{-1}\tilde{\mathbf{\Phi}}^{-1}\mathbf{G}.$$

The diagonal elements $\Gamma_{ii}^\alpha(\omega)$ reflect the partitioning of the spectral density (energy) contained in the forcing amongst the dynamical modes at frequency ω . Note that (4.26) is simply a multivariate generalization of the auto-spectrum given in (4.11).

Given that a statistical description of the forcing (i.e. $\Gamma^f(\omega)$) is known, we propose a criterion for selecting a subset of the modes based on the total energy contained in each mode. Define

$$s_i = \sum_j \Gamma_{ii}^\alpha(\omega_j), \quad (4.27)$$

where s_i designates the sum over all frequencies of the spectral density of the i th mode, and measures the total variance (energy) contained in that mode. Sorting the s_i in descending order allows the modes preferentially excited by the forcing to be identified. By specifying a cutoff (either in terms of number of modes, or cumulative energy), a suitable subset of modes can be chosen. This selection criterion is further discussed in the application of Section 4.5.

4.4 Application

A simple shelf circulation model for the Scotian Shelf off the east coast of Canada was used to test the reduced dimension Kalman filter. The overall goal was to assess its

suitability for an operational nowcasting/forecasting scheme. Figure 4.1 shows a map of the Scotian Shelf region. The area has an irregular coastline, and numerous offshore basins and banks. The overall physics of the Scotian Shelf is relatively complex and reviewed in Smith and Schwing (1991). We note that linear dynamics have proven adequate for describing the subtidal variability associated with the wind and boundary driven circulation (Thompson and Sheng 1996). As a result, we use a linear model on which to test the approximate Kalman filter.

The model was based on the linearized, depth-averaged shallow water equations

$$\begin{aligned} \frac{\partial \vec{u}}{\partial t} + f \vec{k} \times \vec{u} &= -g \nabla \eta - \frac{r}{h} \vec{u} + \frac{\vec{\tau}}{\rho_0 h} \\ \frac{\partial \eta}{\partial t} + \nabla \cdot (\vec{u} h) &= 0 \end{aligned} \quad (4.28)$$

where t is time, \vec{u} contains the components of the depth averaged current and η represents the sea surface elevation. With the exception of the surface wind stress $\vec{\tau}$ and the reference density ρ_0 , the remainder of the notation is defined in Section 3.2. The value of the friction coefficient was $r = 0.001 \text{ms}^{-1}$.

The model domain is also shown in Figure 4.1. The shallow water equations (4.28) were finite differenced following Heaps (1969) based on an Arakawa C-grid with a spacing between adjacent η points of 25 km. The model was implemented on a rectangular domain, and points falling on the land were assigned shallow depths (1m). This approximately satisfies the coastal boundary condition is no normal transport. The remaining boundaries are open with a simple gravity wave radiation condition. The model is subject to forcing by surface wind stress and inflows normal to the north-eastern open boundary

The resulting model state vector is of dimension 661. Its elements correspond to velocity components and sea levels defined on the model grid. Clearly, the overall size of this problem is relatively small but is adequate both to illustrate and to provide a preliminary test of the approximate Kalman filter. Extension to larger problems is straightforward.

Using the finite difference form of the numerical model, the dynamics matrix \tilde{D}

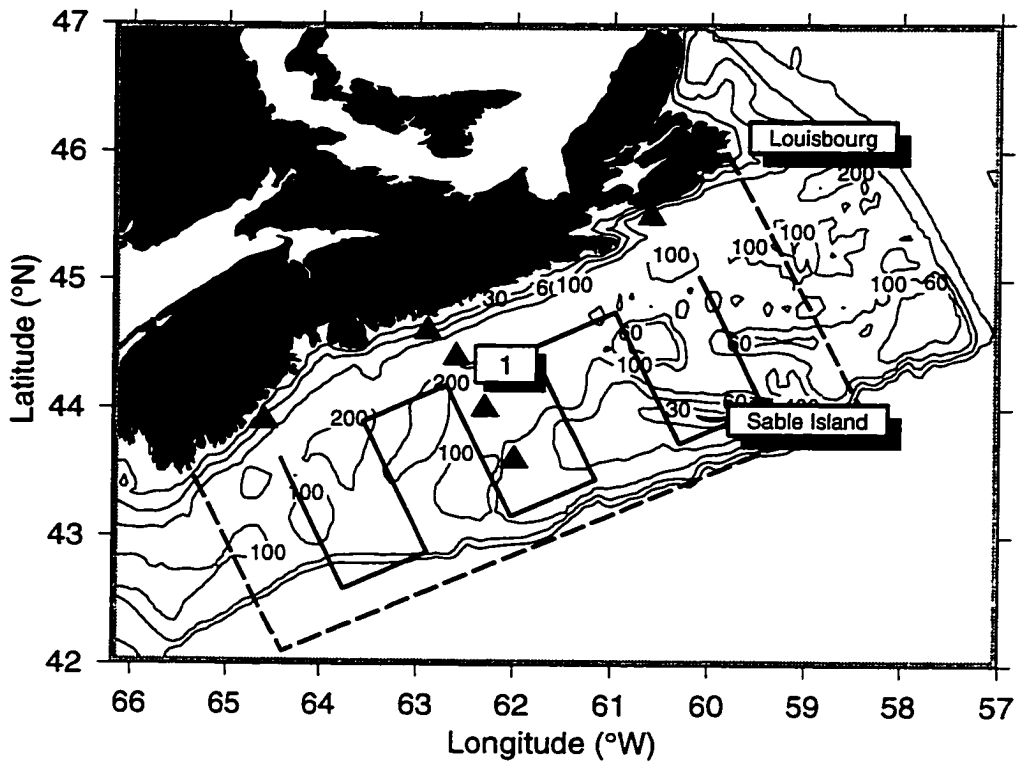


Figure 4.1: Coastline of Nova Scotia and the bathymetry, in meters, of the Scotian Shelf. The rectangular box (dashed) indicates the model domain. Also shown are the locations of the fixed (CASP) observation array (triangles) and the moving (ship ADCP) array (solid line).

was determined using the following algorithm: (i) Set the i th element of the state vector equal to one and all other elements to zero. (ii) Run the model forward one time step from this initial state (with zero forcng) and record the response. This gives the i th column of $\tilde{\mathbf{D}}$. (iii) Repeat this procedure for all i thereby building up $\tilde{\mathbf{D}}$ column by column. This algorithm corresponds to determining a series of impulse response functions for each of the prognostic variables at each of the grid points of the model.

The resulting dynamics matrix $\tilde{\mathbf{D}}$ is real-valued and non-symmetric. Some of the eigenvalues and eigenvectors of $\tilde{\mathbf{D}}$ are real and others are complex. There are 60 modes where $|\lambda| < 10^{-8}$. There are 25 real-valued modes (standing patterns), 8 of which have decay scales less than 10 hours, with the rest being greater than 50 hours. The remaining 576 modes are complex and occur in conjugate pairs. An example of one of these propagating modes for the Scotian Shelf is shown in Figure 4.2. This pattern oscillates between its real and imaginary parts as given in Section 4.2. Figure 4.3 shows a scatter plot of the decay time versus the oscillation period for all propagating modes and indicates that these span 3-4 orders of magnitude and appear to cluster into two groups, with small and large decay-oscillation periods, respectively. It was also found that long oscillation periods were associated with spatially smooth patterns.

It is notable that the dynamical modes, as derived from a numerical model, are not easily interpreted in terms of analytic modes, such as continental shelf waves. The dynamical modes generally had complex spatial structure, and frequencies that were higher than would be expected from simple theory based on idealized systems. This result is likely due to complex bathymetry and coastlines, as well as the relatively coarse resolution of the model.

We experimented with two types of forcings important on the Scotian Shelf: (i) flows across the open boundary at the northeastern end of the model domain, and (ii) a spatially uniform surface wind stress. Data from the Canadian Atlantic Storms Program (CASP) (Lively, 1988) was used to determine the appropriate statistical

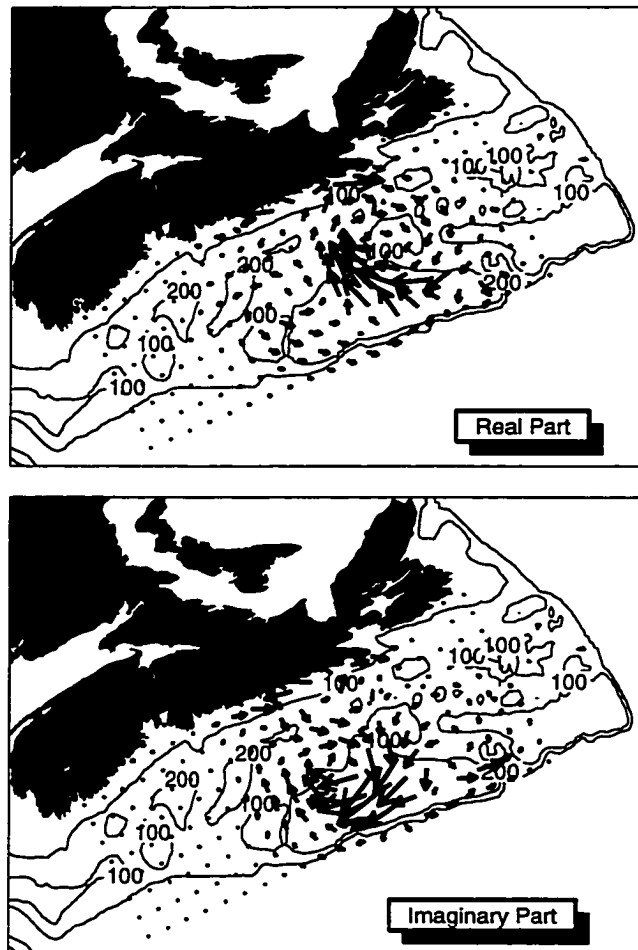


Figure 4.2: The flow field associated with one example of a propagating mode of the Scotian Shelf model. This mode was found to be important for both the wind and boundary forced cases. It has a decay time of 484 hours and an oscillation period of 84 hours.

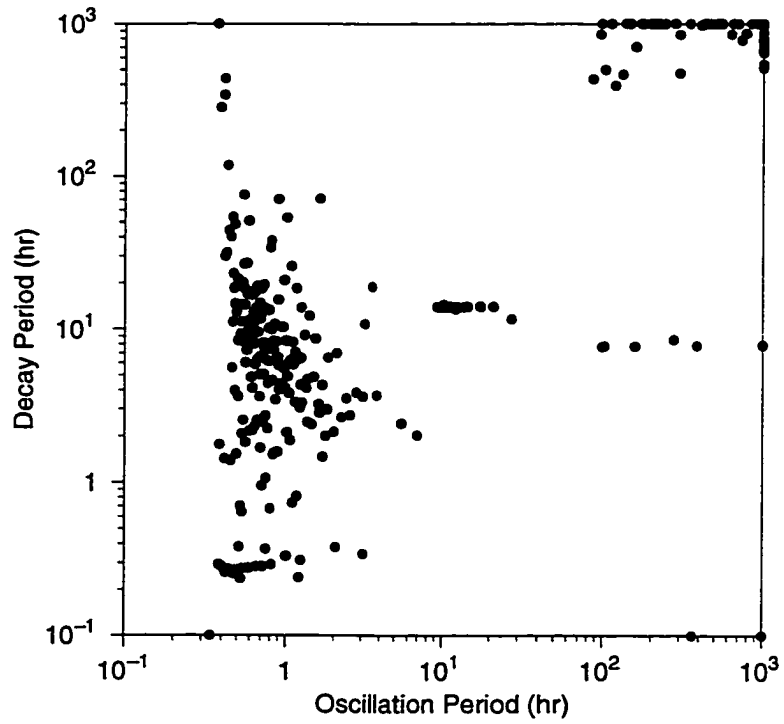


Figure 4.3: Scatter plot of the decay time versus oscillation period for the propagating (complex) modes. For completeness, the modes that would fall outside of the indicated range are shown at the limits of the plot.

form for these forcings. The temporal evolution of the flows across the northeastern boundary was diagnosed from the coastal sea level record at Louisbourg (location in Figure 4.1). We assume a sea level decay to an offshore value of zero. The cross-shelf shape of the sea level profile is defined so that a spatially uniform transport normal to this boundary can be obtained geostrophically (Thompson and Sheng 1996). The hourly Louisbourg sea-level record during CASP (Figure 4.4) was represented well as an autoregressive process of first-order with coefficient $a = 0.9922$. Similarly, the wind record at Sable Island was used as a proxy for a spatially uniform wind stress over the shelf (location in Figure 4.1, time series in Figure 4.4). The temporal evolution of the east-west (x) and north-south (y) components of hourly wind stress were found to evolve according to a multivariate AR(1) process with coefficients $a_x = 0.9903, a_y = 0.9819$ and cross terms near zero ($< 10^{-3}$). The high values of the autoregressive coefficients for the wind and sea level are indicative of the high correlation of the values between adjacent hours.

Both fixed and moving observation arrays (Figure 1) were used to test the approximate Kalman filter in a series of identical twin experiments. The fixed observation array was based on the CASP array and consisted of three coastal sea level stations, and a cross-shelf array of three moorings, each measuring bottom pressure (sea level) and the two components of the depth averaged flow. The moving observation array was designed to mimic the sampling of currents on the shelf by a ship-borne acoustic Doppler current profiler (ADCP). It simulates a ship traveling at about 12 knots following (and repeating) the transect shown in Figure 1. The fixed array had an observation operator constant in time, while the moving array had one which changes through time.

Controllability and observability conditions are shown in Figure 4.5. For both wind and boundary forcing, the system was found to be controllable for all modes, with the possible exception of those near-zero modes in which $|\lambda| < 10^{-8}$. The system was found to be observable using the fixed array, again excluding those modes with $|\lambda| < 10^{-8}$. For the moving array, the assessment of observability is more difficult.

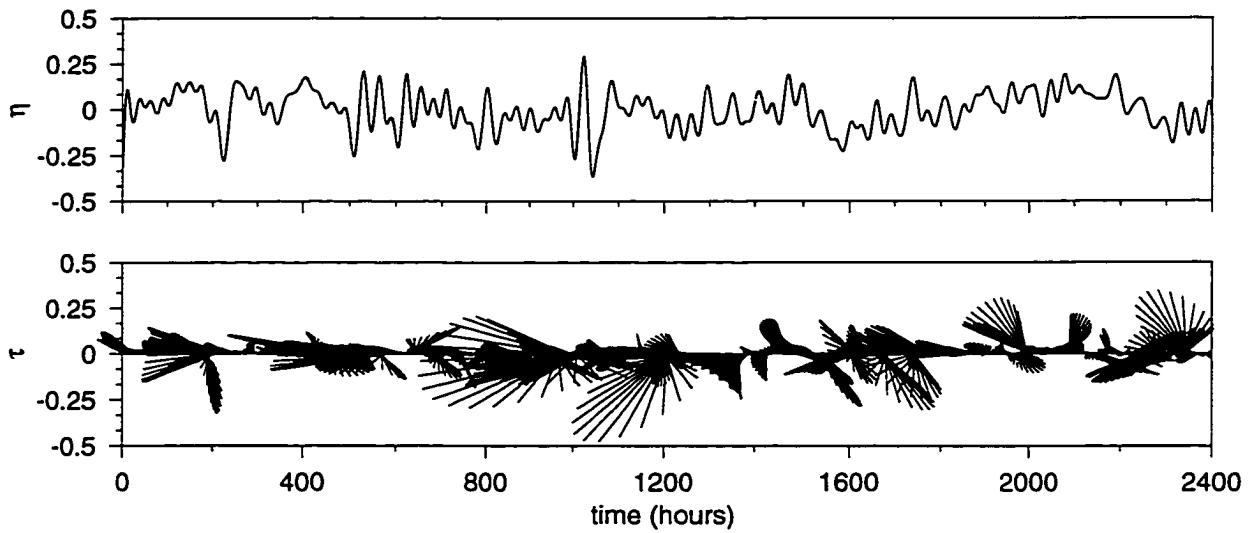


Figure 4.4: (Upper Panel) Time series of Louisbourg sea level (meters) recorded during CASP. (Lower Panel) Time series of the wind stress vector (Pa) recorded at Sable Island during CASP. Both series have been low pass filtered with a cutoff of 2 cycles per day and begin at 16:00, November 24, 1985.

\mathbf{H}_t in this case describes a point measurement of velocity whose location changes over time. The observability condition for some typical values of \mathbf{H}_t suggested that they were being satisfied for most modes with $|\lambda| > 10^{-8}$. A more general statement about observability cannot be made for this case.

Mode selection was carried out separately for both the wind and boundary forcing. The AR(1) representation of the forcing determined from the CASP data were used to specify $\Gamma^f(\omega)$. Based on this, the energy (variance) in each of the propagating modes was calculated and the results are shown in Figure 4.6. Modes were selected such that greater than 99% of the total energy was accounted for by those modes chosen. Selecting the real modes proved to be more problematic as many have eigenvalues nearly indistinguishable from one (likely associated with the coastal grid points). This implies that any energy present at zero frequency would cause these modes to resonate and dominate the energy spectrum. Due to this difficulty, the effect of adding the real modes was tested by comparing the output from the reduced model, with certain of the real modes included, and the full model (for various realizations of the forcings). It was found that only a relatively small number of real modes were required. The resulting selection procedure chose 46 modes (4 real) for the boundary forcing, and 44 modes (4 real) for the wind forcing. An example of one of the modes included in both subsets is shown in Figure 4.2. These modes form the basis for the reduced ocean model and the approximate Kalman filter.

Synthetic data were generated using the full model for both the fixed and moving observation arrays. The forcing was prescribed based on a single realization for both the stochastic wind and boundary forcing. A zero-mean, serially uncorrelated, Gaussian white noise with a standard deviation equal to the signal was then added to these synthetic velocity and sea level series. This serves to specify the observation error covariance matrix \mathbf{R} . The synthetic data for these four cases, provided the basis for the numerical experiments testing the approximate Kalman filter based on dynamical modes.

To satisfy the assumption of serially uncorrelated system noise, we followed (4.3)

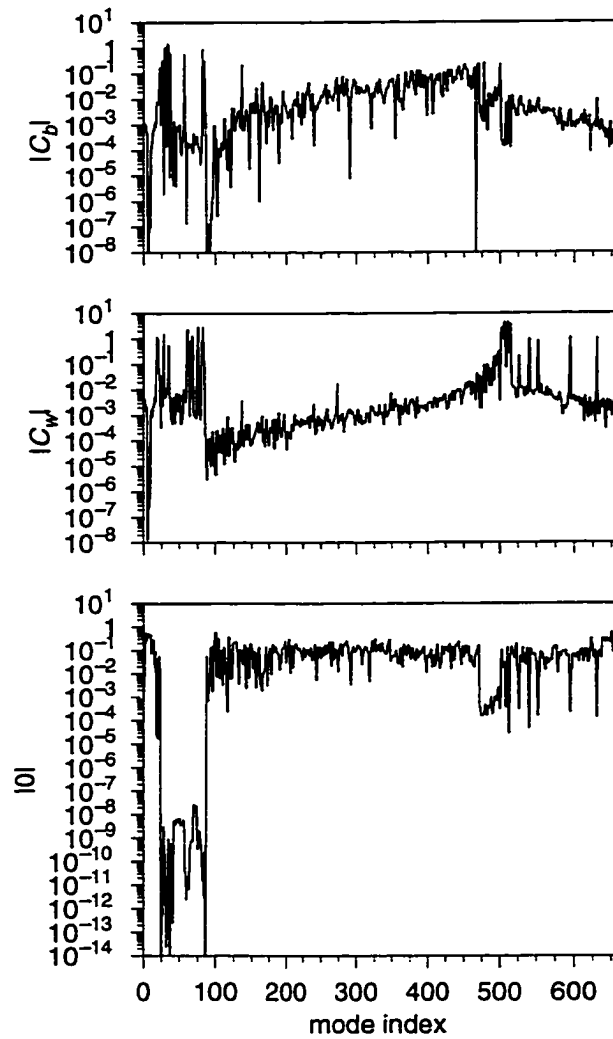


Figure 4.5: Controllability $|C|$ and observability $|O|$ indices for the various experiments. Here, $C = \mathbf{G}^T \phi_a$ and $O = \mathbf{H} \phi$ (see Section 3). Subscripts b, w on C refer to boundary and wind forcing, respectively. O is based on the fixed (CASP) array. The mode index is sorted in terms of frequency, starting with the zero frequency (real) modes.

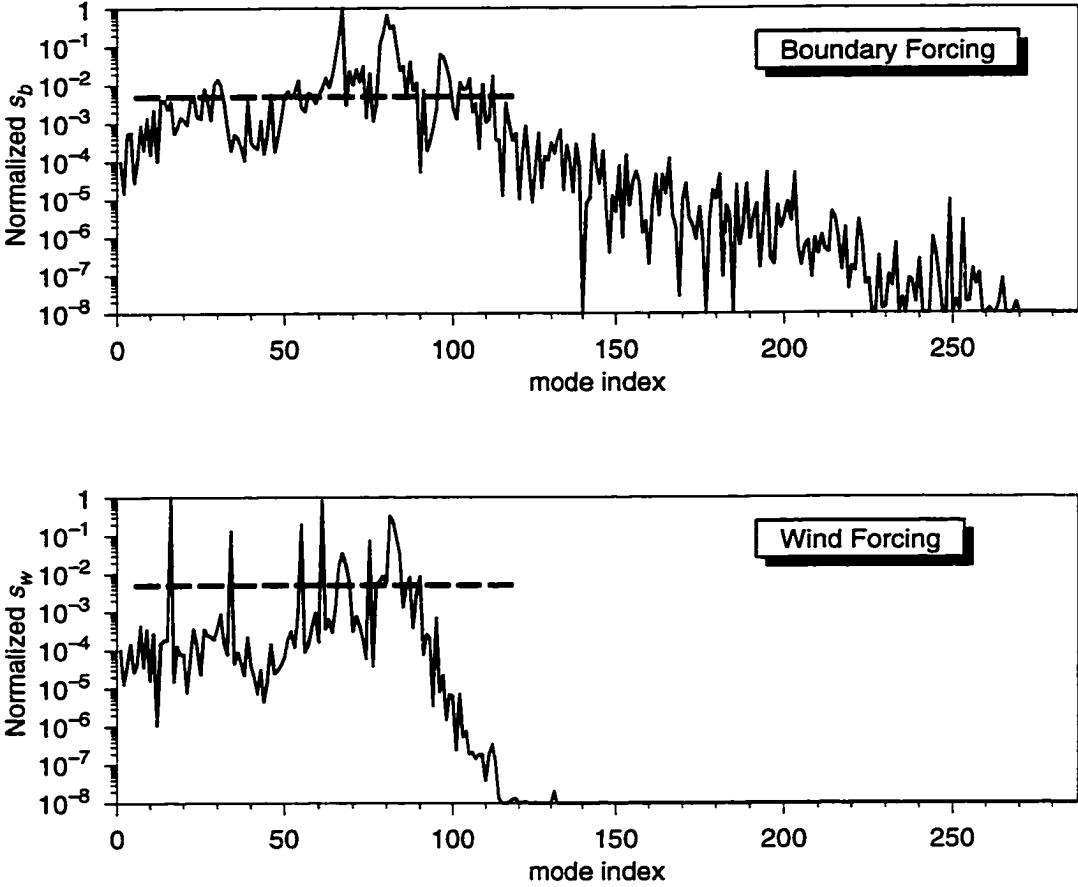


Figure 4.6: The normalized selection criteria s for the propagating (complex) modes. This measures the relative energy (variance) in each mode. (Note that only one of each complex conjugate pair is plotted in each case). The subscripts b, w on s refer to the boundary and wind forced cases, respectively. The modal index is sorted on the basis of frequency, starting with the low frequency modes. The dashed line indicates the cutoff for inclusion of the modes in the reduced model.

and augmented the reduced ocean model to include the AR(1) forcing,

$$\begin{pmatrix} \alpha_1 \\ e^m \end{pmatrix}_t = \begin{pmatrix} \Lambda_1 & \{\tilde{\Phi}^{-1}\mathbf{G}\}_1 \\ \mathbf{0} & \mathbf{A} \end{pmatrix} \begin{pmatrix} \alpha_1 \\ e^m \end{pmatrix}_{t-1} + \mathbf{e}_t, \quad (4.29)$$

where the subscript 1 denotes the reduced model and when applied to $\tilde{\Phi}^{-1}\mathbf{G}$ it describes the extraction of the appropriate rows of this matrix. As in (4.3), \mathbf{A} contains the auto-regressive coefficients and \mathbf{e}_t represents the system noise. The elements of \mathbf{e}_t driving e_t^m were assumed Gaussian with a variance determined so that $\text{var}(e_t^m)$ matched that of the CASP data shown in Figure 4.4. The elements of \mathbf{e}_t associated with α_1 were chosen to be an independent, zero-mean Gaussian white noise process with variance equal to that of the observations. This assumes that both the model and observations are considered to have a similar level of uncertainty. These assumptions fully specify the model error covariance \mathbf{Q} .

The approximate Kalman filter was implemented based on the reduced model (4.29) for a fixed and moving observation array using both wind and boundary forcing. The specification of \mathbf{R} and \mathbf{Q} is outlined above, and for simplicity, the initial covariance \mathbf{M}_0 was set equal to \mathbf{Q} . Based on these input statistics, the filter was not unduly biased towards either the model or the observations.

Figure 4.7 shows results from the application of the approximate Kalman filter to the fixed observing array using both wind and boundary forcing. In these cases, the filter estimates were within $\pm 1\%$ of their true values and indistinguishable from one another on the plots of Figure 4.7. The Kalman gain matrix converged to a steady state value in about 12 hours (Figure 4.8). Similarly, Figure 4.9 shows the application of the approximate Kalman filter for the moving array. The estimates of the velocity along the simulated ship track, from both the wind and boundary forcing, were again within $\pm 1\%$ their true values. The estimated rms error determined from \mathbf{P}_t for u, v at the observation locations was about 5 cm/s for both the fixed and moving observation arrays. The noisy observations, together with the knowledge that the stochastic character of the forcing is AR(1) with known coefficients, appears sufficient to recover the model state at the observation locations.

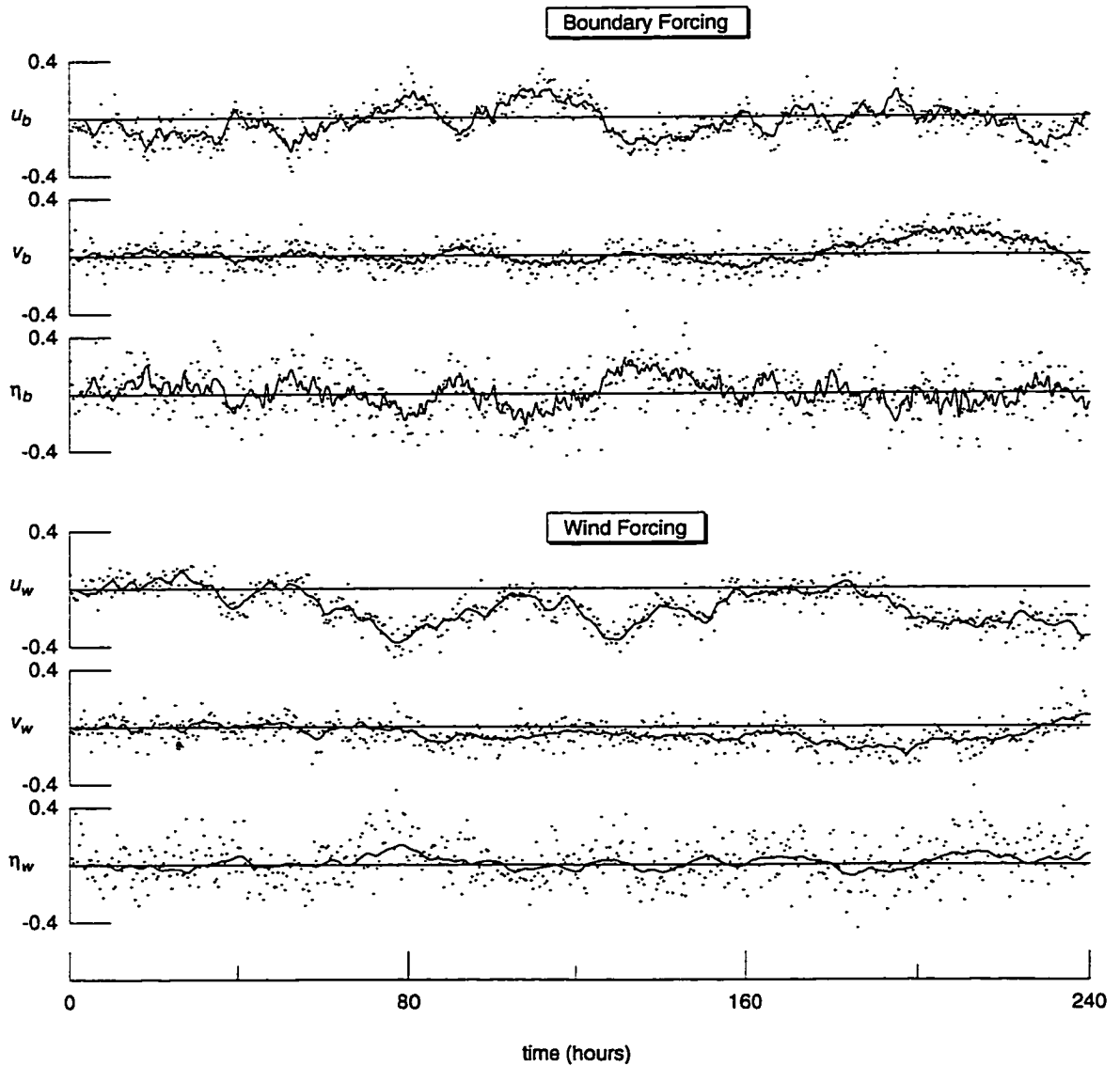


Figure 4.7: Results from the approximate Kalman filter at mooring 1 for the fixed observation array. u, v are the horizontal components of velocity in ms^{-1} and η represents the sea level in m. The dots indicate the observations and the solid line represents the Kalman filter estimates at that location (which coincide with true values). The upper three panels corresponding to boundary forcing, and the lower three to wind forcing.

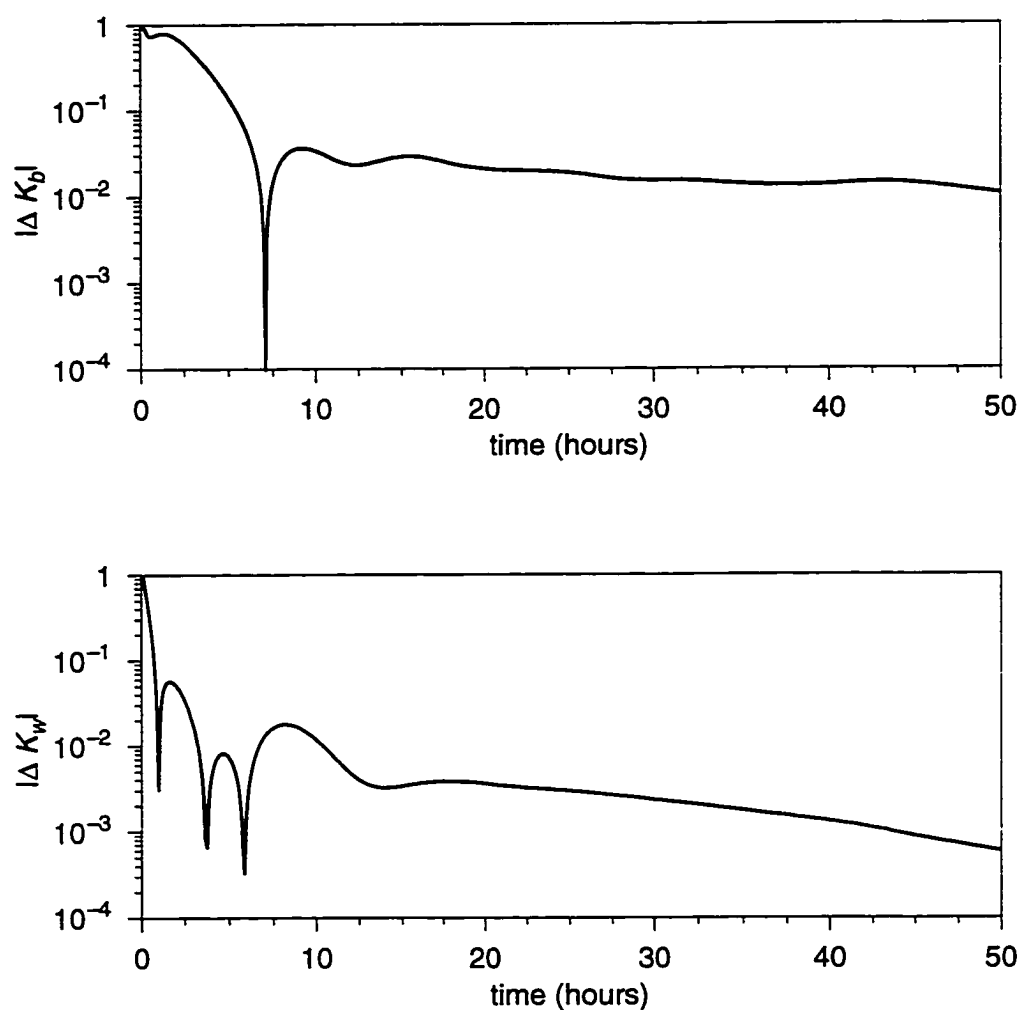


Figure 4.8: Convergence of the Kalman gain matrix beginning at the start of the assimilation period. Here, $|\Delta K|$ denotes the (summed) absolute value of the difference between corresponding elements of Kalman gain matrices at successive time steps (scaled by its initial value). The subscripts b and w refer to the boundary and wind forced cases, respectively.

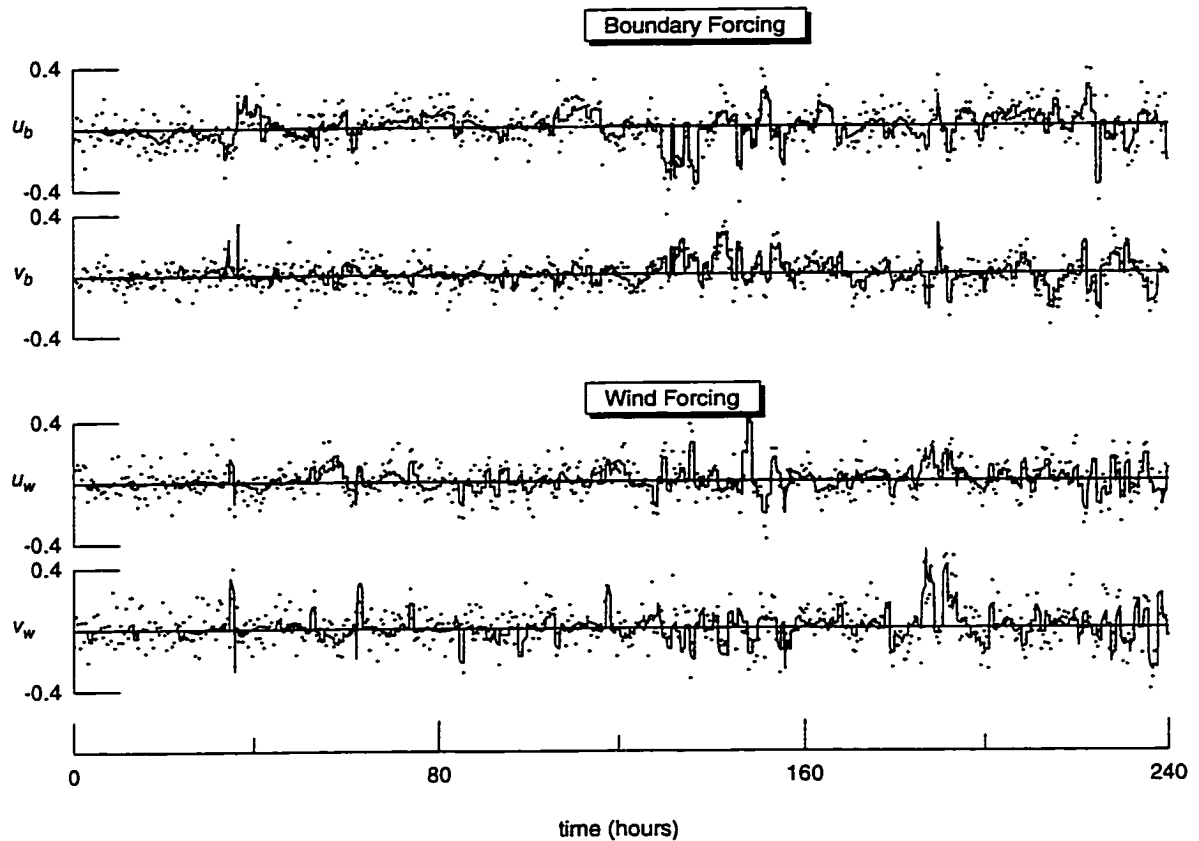


Figure 4.9: Results from the approximate Kalman filter for the moving observation array (ship ADCP). u, v are the horizontal components of velocity in ms^{-1} at the observing locations at the given time. The dots indicate the observations and the solid line represents the Kalman filter estimates at that location (which coincide with true values). The upper two panels corresponding to boundary forcing, and the lower two to wind forcing.

Figure 4.10 shows a typical snapshot of the flow field for an intermediate time in the assimilation using the fixed array and wind forcing and compares this to its true value. In this case, the important features of the actual flow field were recovered by the approximate filter which suggests a reasonable ability to extrapolate to locations other than those observed. Detailed examination confirms this conclusion for the fixed array, excepting the region in the northeast quarter of the model domain. There, the large distance from the observation locations as well as the irregular bathymetry, which results in shelf wave scattering, combine to give poor estimates of the actual flow field. The moving observation array provided good estimates of the actual flow field near the observation locations but had much less ability to extrapolate into data poor regions.

The recovery of the actual wind and boundary forcing by the Kalman filter is shown in Figure 4.11. The fixed observing array was able to estimate fairly well the important low frequency components of both the wind and boundary forcing. Again, the moving observation array does a poor job in recovering the forcing functions, suggesting that the model is overfitting to compensate for the sparse data provided by the ship ADCP at a given time.

4.5 Discussion and Conclusions

In this chapter, we have proposed an approximate, or reduced dimension, Kalman filter and tested its suitability for use in a prototype coastal circulation model. Reduction in the dimension of the state vector necessary to describe the ocean was achieved by a representation in terms of the dynamical modes supported by the numerical model. Given the statistical properties of the model forcing, the important modes can be identified and a suitable subset chosen. These modes then provide a basis for a reduced ocean model. If this subset is relatively small, a Kalman filter based on these modes will have greatly enhanced computational efficiency and no stability restrictions on its time step.

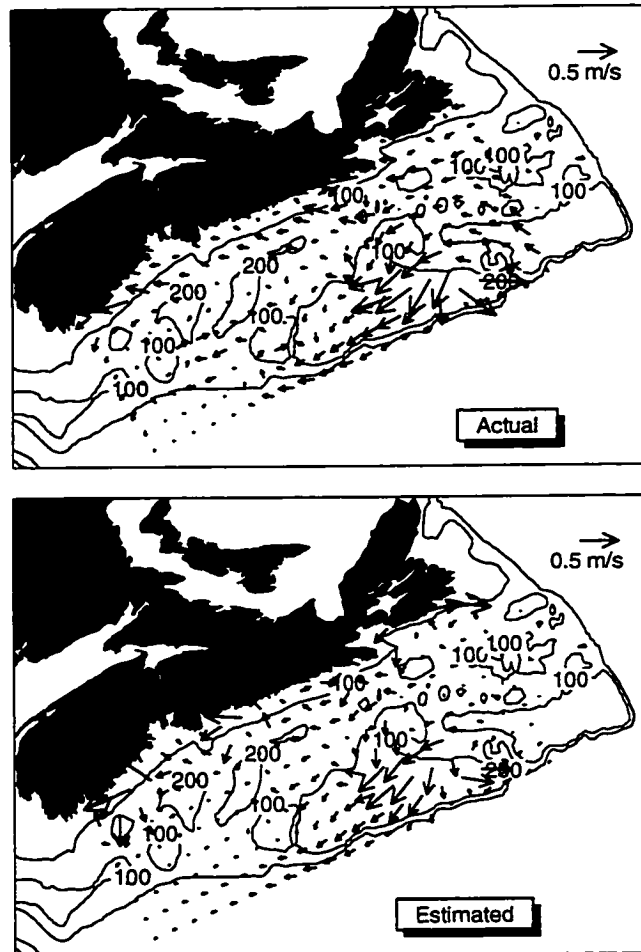


Figure 4.10: A comparison of the actual (upper panel) versus estimated (lower panel) flow field at ninety hours into the assimilation period. These flow fields correspond to the wind forced case and are based on the fixed observation array.

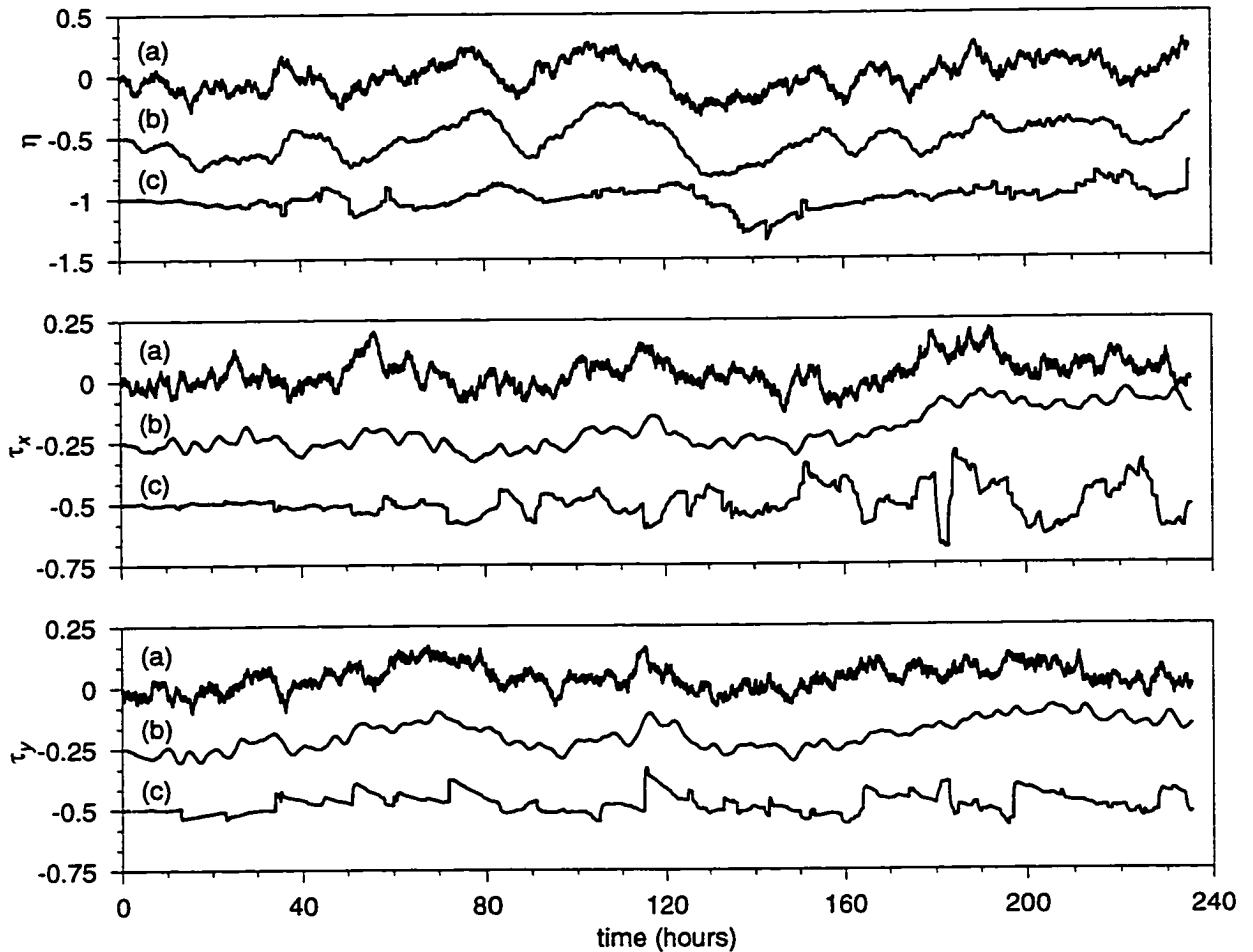


Figure 4.11: A comparison of the actual versus estimated forcing. The upper panel corresponds to the boundary forced case with η being the coastal sea level (m) at the upstream boundary. The middle and lower panels correspond to the wind forced case, with τ_x, τ_y being the east-west and north-south components of the spatially uniform wind stress (Pa). In each panel, (a) is the actual forcing, (b) is the estimate (offset) using the fixed array, and (c) is the estimate (offset) using the moving array. The offset used is 0.5m for η and 0.25 Pa for τ .

Our application to a prototype circulation model of the Scotian Shelf proved to be encouraging. Using both wind and upstream boundary forcing, the model dimension n was reduced to less than 1/10 its value for the original finite difference implementation. With the computational burden of the Kalman filter varying as n^3 (Gelb 1974), this implies a computational reduction of a factor of 1000. In all cases tested, the approximate Kalman filter proved capable of recovering the true model state at the observation locations to within $\pm 1\%$ using the information in the noisy measurements and knowledge of the statistical character of the forcing. The filter was able to recover the entire flow field (extrapolation) and the forcing functions in the case of the fixed observations array. However, the extrapolation ability proved poor for the moving observation array. This suggests, not surprisingly, that the fixed array contains more information relevant for the estimation of flow fields than does the moving array. The increased spatial coverage of the moving array is not enough to compensate for its ability to measure only velocity at a single point at any given time.

The controllability and observability conditions are important diagnostics for any assimilation problem, and straightforward to determine using a modal representation. To observe the system, the observation locations must not coincide with nodal points of the modes. To control the system, the modes of the adjoint operator are used to assess the ability of the forcing to drive the system to an arbitrary state (for instance, a state which best matches the observations). This suggests the possibility that a modal representation of the model, or its adjoint, may allow for efficient variational data assimilation schemes. Since controllability and observability are addressed on a mode by mode basis, ordering of the modes in terms of energy facilitates a practical assessment of whether or not these conditions are satisfied, i.e. if a mode is unimportant, the ability to control or observe it is of little interest.

The selection criterion based on the energy contained in each mode is not without its drawbacks. The modes $\tilde{\phi}$ provide a non-orthogonal basis for the vector space spanned by $\tilde{\mathbf{x}}$. The variance explained by a particular subset of the modes cannot

then be simply assessed without taking into account the off-diagonal terms of $\Gamma^\alpha(\omega)$. This contrasts with the basis provided by (complex) empirical orthogonal functions (e.g. Brillinger 1981) which are orthogonal and have a natural ordering in terms of variance based on their eigenvalues, but no dynamical meaning. On balance, it is felt that the ordering of the modes based on energy is a straightforward and reasonable basis on which to select the modes. However, the general question of mode selection remains an area for future research.

A major premise of this work is that the important dynamical and statistical properties of the model are preserved with a relatively small number of modes. This is also a key assumption when using principal oscillation patterns which are, in fact, identical to our modal decomposition. This technique is, however, usually applied to data (Latif and Flugel 1991) or model output (von Storch *et al.* 1988), with the 'dynamics' matrix being derived empirically. Our application uses a dynamics matrix equivalent to a finite difference implementation of the model. It can be determined either through the impulse response technique used in Section 4.4, or a finite difference form of the model written directly into matrix form.

The added cost associated with the approximate Kalman filter lies in the determination of the dynamics matrix $\tilde{\mathbf{D}}$, calculation of its eigenvalues and eigenvectors, and the selection of a suitable subset of modes. The important computational savings result from simplification of the calculation of the gain and error covariance matrices in (2.24)-(2.26), as well as the possibility of longer time steps. For a linear model in which the form of the dynamic equations do not vary in time, the eigen-pairs can be determined and the important modes selected prior to any data assimilation exercise. In this case, the dimension of the full model is of little significance in an operational system based on the approximate filter. For nonlinear systems, or ones with time varying dynamics, this is not the case.

The extended Kalman filter is a direct extension of the linear filter which is suitable for application with nonlinear models. The extended Kalman filter uses the

nonlinear model in the forecast equation (2.23). The remainder of the equations retain their form but with the dynamics matrix \mathbf{D}_t now being a linearization of the nonlinear model about the current state estimate $\hat{\mathbf{x}}_t$ (we ignore nonlinear measurement operators). Ideally, the model is re-linearized at every time step, but in practice this can occur less often as the model state may not vary significantly over a given time interval.

To directly apply the extended Kalman filter using the method of this chapter, the eigenvalues and eigenvectors of \mathbf{D}_t would have to be calculated every time the model was linearized. (There is no means to pre-compute \mathbf{D}_t , or its eigenpairs, since they are a function of the current estimate). Once a suitable subset of modes has been chosen, the approximation could be applied to the gain and error covariance calculation (2.24)-(2.26). Note that the full nonlinear model would be retained for the forecast equation (see Fukumori and Malanotte-Rizzoli, 1995). The usefulness in this case rests on whether the computational savings in the gain calculation justifies the additional computational cost. This cost depends mainly on the dimension of the model and the interval over which the linearization is valid. Allowing for interacting modes, i.e. a non-diagonal Λ , may allow the length of this interval to be extended.

These speculations lead to the possibility that an important application of dynamic modes might lie in their ability to facilitate suboptimal, yet robust, data assimilation schemes for nonlinear models. As is evident from Chapter 2, many data assimilation methods rely on essentially linear methods. For the filtering problem, consider the local linearization of the dynamics for the purpose of error propagation. This linearization might be carried out efficiently, say, based on analysis of the principal oscillation patterns of certain of the model output. If one could empirically identify the important dynamic patterns, an efficient Kalman filter could be constructed. Similarly, for smoothing methods, identification of approximate adjoint operators might follow an analogous procedure.

In summary, the synthesis of data and ocean models through data assimilation

involves complex algorithms. For operational schemes, approximate yet robust methods are desirable. Our reduced dimension Kalman filter is one such method directly suitable for linear coastal circulation models and potentially extensible to more complex models. Its important feature is an approximate dynamical operator suitable for error propagation and determination of time varying gain matrices. In this way, the important physics of the model is retained while data synthesis is simplified. It is felt that many practical data assimilation schemes might benefit using ideas and approaches of this kind.

Chapter 5

Determining Dynamically Consistent Density and Velocity

Inferring ocean circulation from temperature and salinity data is a well studied problem in oceanography. Hydrographic fields are less affected by energetic small scale motions than are point measurements of velocity or transport. As a result, they contain useful information on large scale circulation features. In fact, much of our general picture of ocean circulation has been obtained by tracing the temperature-salinity (T-S) characteristics of water masses (e.g. Wüst 1935). A quantitative description of these water mass tracing techniques is to be found in modern tracer inverse methods (Martel and Wunsch 1993, Schlitzer 1993). Another approach uses large-scale dynamic balances to relate observed T-S (or density) to velocity or transport. These dynamic methods are useful for generating and testing hypotheses about the physical origins of circulation, and ultimately seek a detailed and accurate picture of the general circulation. It is these dynamic methods, and their application to coastal regions, which provides the focus for this chapter.

While dynamic methods are potentially useful tools, their application frequently proves difficult. Hydrographic data consists mainly of depth profiles which are characterized by their asynchronous and irregular distribution. It is often argued (e.g.

Ghil and Malanotte-Rizzoli 1991) that the ocean is grossly under-sampled with respect to important time and space scales. Archived data have been used to produce statistical maps of hydrography (e.g. Levitus 1982). These seasonally averaged and spatially smoothed fields reflect the important climatological features found in the ocean. However, errors resulting from excessive smoothing make it difficult to obtain dynamically consistent circulation from climatological fields.

A number of methods are available to relate circulation to ocean hydrography. Perhaps the most commonly used assumption is that the measured or mapped density field provides an accurate representation of its true value. In this case, knowledge of the spatial variations in the density field implies a fixed, time-invariant (baroclinic) pressure gradient from which flow fields may be diagnosed. These diagnostic methods have a long history in physical oceanography (see Defant 1961 for a review) and modern diagnostic methods have become quite sophisticated incorporating vertical mixing, turbulence closure and a variety of boundary conditions (e.g. Lynch *et al.* 1992, deYoung *et al.* 1993). Of particular importance when applying these methods is proper treatment of the bottom torque generated by the misalignment of the density field and bathymetry (Sarkisyan and Ivanov 1971, Holland and Hirschman 1972, Huthnance 1984, Mertz and Wright 1992). The numerical calculation of these terms based on smoothed density fields frequently results in unrealistic circulation features (Rattray 1982).

A more realistic approach is to recognize explicitly that the observed density field is only an approximation of its true value and inevitably will contain unresolved variability over a range of time and space scales. Therefore it is more appropriate to treat density as a random variable, rather than a fixed input quantity. β -spiral techniques (Stommel and Schott 1978, Olbers *et al.* 1985) adopt this assumption to estimate deep reference velocities from observed density and tracer data. Similarly, the section method of Wunsch (1978) determines reference velocities using an inverse approach. Robust diagnostic approaches (Ezer and Mellor 1994, Sarmiento and Bryan 1982) also alter diagnostic results by allowing the input density field to partially adjust

to the circulation.

Although these approaches allow for errors in the observed density they do not ensure dynamically balanced density and velocity fields. (Roughly speaking, dynamic balance here refers to the density and velocity fields simultaneously satisfying both the equations for advection of density and thermal wind). Clearly, to satisfy this condition, the evolution equation governing the density distribution must be explicitly taken into account. Wunsch (1994) and Bogden *et al.* (1993) are notable in recognizing this need. Both of these studies focus on North Atlantic circulation and enforce conservation of density along streamlines (interpreted as a minimum mixing of density) as a weak constraint in order to select the appropriate barotropic mode.

The goal of this chapter is to investigate the issues surrounding the determination of dynamically consistent density and velocity fields from hydrographic data, with an emphasis on coastal regions. In contrast to other methods, the focus is mainly on the consequences of treating the density equation as a strong constraint. The development is pedagogical and intended both to illustrate problems associated with the diagnostic method, as well as demonstrate the need to include the density equation when determining circulation from density data.

The chapter is outlined as follows. Section 5.1 discusses some dynamical issues in estimating circulation from density data. Section 5.2 treats the estimation problem in an optimization framework and reviews some common approaches in this context. Section 5.3 then illustrates the joint estimation of density and velocity with a specific example: a simple, depth-averaged steady-state model applicable to coastal regions is proposed which includes the advection of density as well as the interaction of the density field and the bathymetry. An estimation technique is then outlined for this boundary value problem whereby the buoyancy fluxes across the open boundaries into the model domain are determined from interior point observations of the density field. Numerical experiments are carried out to illustrate the method and explore some generic problems. Discussion and conclusions follow in Section 5.4.

5.1 Dynamics

In this section, we discuss some of the dynamical issues involved in determining ocean circulation from the density field. The implication of diagnosing transport from density is first addressed for a case which does not make use of an evolution equation for density. The consequences of adding the density equation are then investigated.

Dynamical equations are considered which describe the steady-state balance of a hydrostatic, incompressible, Boussinesq fluid. The area of interest is coastal and continental shelf regions which are well-mixed in the vertical due to bottom friction. The governing equations are given by the following:

$$f\vec{k} \times \vec{u} + \frac{1}{\rho_0} \nabla P - \frac{1}{\rho_0} \frac{\partial \vec{\tau}}{\partial z} = 0 \quad (5.1)$$

$$\frac{\partial P}{\partial z} + \rho g = 0 \quad (5.2)$$

$$\nabla \cdot \vec{u} + \frac{\partial w}{\partial z} = 0 \quad (5.3)$$

$$\vec{u} \cdot \nabla \rho + w \frac{\partial \rho}{\partial z} - K \nabla^2 \rho - K_v \frac{\partial^2 \rho}{\partial z^2} = 0. \quad (5.4)$$

Here, \vec{u} and w represent the horizontal and vertical components of velocity, respectively. The density is given by ρ , with ρ_0 being a spatially invariant reference value. Pressure is denoted by P , frictional stresses by $\vec{\tau}$. Horizontal mixing of density parameterized by K , with K_v being its vertical counterpart. The remainder of the notation is standard and given in Section 3.2. The horizontal momentum balance is described by (5.1). If $\vec{\tau}$ is negligible (a reasonable assumption away from boundary layers) it reduces to the familiar geostrophic relation. The hydrostatic balance in the vertical is given by (5.2) and the incompressibility condition implies (5.3). Finally, (5.4) governs the density distribution and includes both advection and diffusion.

Some important simplifications in (5.1)-(5.4) include the following. First, advective nonlinearities in momentum are neglected in (5.1). Scaling arguments (e.g. Pedlosky 1979) show this to be a valid assumption provided the Rossby number

$R_o = u/fL \ll 1$, where u and L are horizontal velocity and length scales, respectively. We consider a mid-latitude system ($f \sim 10^{-4}\text{s}^{-1}$) with typical current speeds of less than 0.1ms^{-1} and horizontal dimensions greater than 100km. This implies a R_o of 10^{-2} and justifies neglect of the advective terms. Second, we have chosen in (5.4) to use density as a state variable rather than explicitly resolving its constituent elements: temperature and salinity (which would require separate tracer equations for each of these elements as well as a nonlinear equation of state to relate them to density). This is a reasonable assumption in regions provided the T-S properties of the water remain stable (e.g. Gill 1982, Section 4.10).

We now consider some properties of (5.1)-(5.3) with a focus on the estimation of transport from density. Let $\rho = \rho_0(1 + \varepsilon)$ where ε represents the density anomaly about the fixed reference value ρ_0 . Integrating the hydrostatic relation (5.2) with respect to depth gives

$$P(z) = P_{\text{atm}} + \rho_0 g \int_z^\eta (1 + \varepsilon) dz. \quad (5.5)$$

where the pressure at any depth z is seen to be a sum of the atmospheric pressure P_{atm} and that imposed by the overlying water column. If P_{atm} is assumed constant, the horizontal pressure gradient as a function of depth is determined from (5.5) to be

$$\nabla P = \rho_0 g \nabla \eta + \rho_0 g \int_z^0 \nabla \varepsilon dz \quad (5.6)$$

where the sea level is represented by η . (The usual approximation has been used to set the limits on the integration e.g. Csanady (1979)). The pressure gradient is composed of a barotropic component due to the sea level gradient, and a baroclinic component due to the horizontally variable density field.

Consider the inviscid version of (5.1). Ignoring the terms involving $\vec{\tau}$, the geostrophic limit of the momentum equations is obtained. The total velocity can be decomposed as

$$\vec{u} = \vec{u}_b + \vec{u}_s \quad (5.7)$$

where the velocity components satisfy

$$f\vec{k} \times \vec{u}_b = -g\nabla\eta \quad (5.8)$$

$$f\vec{k} \times \vec{u}_s = -g \int_z^0 \nabla\epsilon dz. \quad (5.9)$$

This decomposition serves to isolate the influence of each of the two terms in the horizontal pressure gradient in (5.6) on the overall flow. The pressure gradient in (5.8) due to sea level is independent of depth and drives the barotropic flow \vec{u}_b . In (5.9), the velocity field is related to the horizontal density gradient and is a statement of the thermal wind relation, i.e.

$$\frac{\partial \vec{u}}{\partial z} = \frac{g}{f} \vec{k} \times \nabla \epsilon. \quad (5.10)$$

The barotropic transport is defined as

$$\vec{U}_b = \int_{-h}^0 \vec{u}_b dz \quad (5.11)$$

where $h = h(\vec{x})$ is the bottom depth. Using (5.8), it is found that \vec{U}_b is related to sea level according to

$$\vec{U}_b = \frac{gh}{f} \vec{k} \times \nabla \eta. \quad (5.12)$$

Suppose that the barotropic transport satisfies the continuity equation $\nabla \cdot \vec{U}_b = 0$. (The continuity condition is obtained by vertically integrating the incompressibility condition (5.3) and applying kinematic boundary conditions (e.g. Pedlosky 1979)). This leads to the well-known property

$$\vec{U}_b \cdot \nabla \left(\frac{f}{h} \right) = 0 \quad (5.13)$$

which indicates that in a geostrophically balanced flow the barotropic component follows contours of f/h .

The baroclinic transport \vec{U}_s associated with (5.9) is defined analogously to (5.11). This component of transport satisfies

$$\vec{U}_s = \frac{gh}{f} \vec{k} \times \int_{-h}^0 \left(1 + \frac{z}{h} \right) \nabla \epsilon dz. \quad (5.14)$$

If \vec{U}_s satisfies the continuity equation $\nabla \cdot \vec{U}_s = 0$, it follows that

$$\vec{U}_s \cdot \nabla \left(\frac{f}{h} \right) = g \int_{-h}^0 z \vec{k} \cdot \left\{ \nabla \varepsilon \times \nabla \left(\frac{1}{h} \right) \right\} dz. \quad (5.15)$$

This demonstrates that the sheared flows associated with the horizontally variable density field can drive transport across f/h contours. Specifically, it is the misalignment of the density field and the bathymetry which forces this cross-isobath transport. This is the so-called JEBAR (joint effect of baroclinicity and bottom relief) effect (Huthnance 1984, Mertz and Wright 1992). It has important dynamic consequences in coastal and continental shelf regions which exhibit strong variations in bottom topography, and is explored in more detail in Section 5.3.

An important element of the estimation problem can now be highlighted. Suppose the density field ε is known and the total transport,

$$\vec{U} = \vec{U}_b + \vec{U}_s \quad (5.16)$$

as given by (5.11) and (5.14), is to be determined. The total transport satisfies the continuity condition

$$\nabla \cdot \vec{U} = 0. \quad (5.17)$$

Given that ε is specified, our governing equations (5.16)-(5.17) constitute a linear system of three equations in the three unknowns: \vec{U} and η . According to the general inverse methods of Section 2.1, a solution is possible provided that no null space exists.

However, examining (5.11) reveals that density contains no information on the barotropic transport and a null space does indeed exist. This indeterminacy is a statement of the so-called "level of no motion" problem which is directly evident from the thermal wind relation (5.10). That is, using only the geostrophic, hydrostatic and incompressibility conditions implies that sea level is in no way related to the density field and the baroclinic and barotropic modes are uncoupled. It is clear that to obtain the complete transport field additional information on \vec{U}_b is required. Indeed, specifying $\vec{U}_b \times \nabla(f/h)$ at a single point along each of the f/h contours removes the

indeterminacy (i.e. \vec{U}_b is then known at all points on the f/h isoline). Note that this procedure is equivalent to specifying lateral boundary conditions on either \vec{U}_b or η in the case where no closed f/h contours exist.

Another more practical problem in determining even the resolvable baroclinic component of transport \vec{U}_s is that, in reality, the input density field is only imperfectly known. Estimated density fields are generally obtained from irregularly distributed observations and statistically mapped into what is assumed to be a representative (climatological) description of the density field. Errors arise due to smoothing of features in the mapping procedure as well as the neglect of any temporal variability. According to (5.14), this estimated density field must be differentiated in order to obtain the baroclinic pressure gradient, a procedure which tends to amplify any small scale errors. The dynamical consequence of this is evident from (5.15): apparent misalignments in the density field and bathymetry will drive spurious flows across f/h contours.

Including the Density Equation

Thus far we have used only the geostrophic, hydrostatic and incompressibility conditions to relate density to transport. In doing so, a null space has been identified associated with the barotropic transport. Consider the density equation (5.4). Scaling arguments indicate that we can neglect this equation only when the barotropic pressure gradient dominates over the baroclinic one ($|\vec{u}_b| \gg |\vec{u}_s|$), in other words in cases where density acts as a passive tracer. Otherwise density serves as a dynamically active tracer which couples the evolution of the density and velocity fields and results in a nonlinear system. This fact also implies that density can no longer be directly obtained from observations since mapped fields will not, in general, satisfy the governing equations. Clearly, adding a prognostic density equation has important consequences for both the dynamics and the estimation problem. These are examined in more detail below.

Anticipating the model used in Section 5.3, we assume, for simplicity, that density

does not vary vertically, only horizontally, i.e.

$$\varepsilon = \varepsilon(\vec{x}).$$

Furthermore, consider the density equation in the limit that mixing is negligible ($K, K_v = 0$). In terms of the density anomaly ε , (5.4) now takes the form

$$\vec{u} \cdot \nabla \varepsilon = 0 \quad (5.18)$$

and is a statement of conservation of density along streamlines. It is seen that density is advected by the total flow field and hence depends on the barotropic component of the flow. To show this more explicitly, expand \vec{u} as in (5.7) and vertically integrate (5.18) from the bottom $z = -h$ to the surface $z = 0$. This yields

$$\vec{U}_b \cdot \int_{-h}^0 \nabla \varepsilon dz + h \int_{-h}^0 \vec{u}_s \cdot \nabla \varepsilon dz = 0. \quad (5.19)$$

Now, \vec{u}_s can be expressed entirely in terms of the density field ε using (5.9) giving

$$\begin{aligned} \int_{-h}^0 \vec{U}_b \cdot \nabla \varepsilon dz &= \frac{gh}{f} \int_{-h}^0 \left\{ \vec{k} \cdot \left(\nabla \varepsilon \times \int_z^0 \nabla \varepsilon dz \right) \right\} dz' \\ &= 0. \end{aligned} \quad (5.20)$$

The integrand on the LHS describes the advection of the density field by the barotropic flow. On the RHS, the integrand represents the advection of density by the sheared flow associated with thermal wind. This is manifested as a measure of the misalignment of the density field at any level with its integral throughout the overlying water column. This term is zero in our case since we have assumed that $\varepsilon = \varepsilon(\vec{x})$. The equation demonstrates that \vec{U}_b can, indeed, be related to information contained in the density field.

However, a null space still exists in the determination of barotropic transport from density. Consider the case where $\varepsilon = \varepsilon(f/h)$. Then any barotropic transport of the form $\vec{U}_b = \vec{U}_b(f/h)$ satisfies (5.20) and lies in the null space. However, it might be argued that this constitutes, in some sense, a weaker condition than for the previous

case. There, a null space associated with barotropic transport existed for any density fields.

It is seen that adding the density equation (5.20) to the previous geostrophic system (5.16)-(5.17) results in four governing equations in four unknowns: \vec{U} , η , and ε . Density is a prognostic variable and is now treated as an unknown since, in general, observed density is inconsistent with the model equations. That is, measurements can only provide guidance for the solution in the sense that the model equations are solved in such a way that predicted density is chosen as a best fit to the observations. This approach leads to dynamically balanced density and circulation fields and provides the essence of the joint estimation problem considered in this chapter.

5.2 Estimation

The estimation problem associated with determining circulation from density data can be approached from the perspective of data assimilation. Consider the cost function

$$J = \int_V \kappa_o \{ \varepsilon - \varepsilon_{\text{obs}} \}^2 + \kappa_1 \{ 5.1, 5.2, 5.3 \}^2 + \kappa_2 \{ 5.4 \}^2 dV. \quad (5.21)$$

Here, J is a scalar quantity measuring the squared observation and model errors, integrated over the volume V of interest (the model domain). The first term in the integrand measures the squared deviations of the observed density anomaly ε_{obs} from its true value ε and is weighted by κ_o . (A continuous representation of the measurements has been used here for convenience, although their discrete nature is recognized. If observations are mapped to a grid, or expanded in terms of structure functions, the above form provides a reasonable representation). The second term measures the squared errors in the governing equations (5.1)-(5.3) or, in other words, the extent to which the equality is satisfied for each of these relations. Similarly, the final term represents the extent to which the density equation (5.4) is satisfied. These latter two terms are weighted by κ_1 and κ_2 , respectively. Note that in the cost function the dynamics (5.1)-(5.3) have been considered separately from (5.4).

This split was used in the previous section and provides a means of distinguishing the methods reviewed in this section.

A data assimilative approach to combining density data with the governing equations chooses density and velocity fields such that J is made a (global) minimum. The form of J has already been defined and it remains only to specify the weighting functions κ_o , κ_1 , and κ_2 . Ideally, these weights are the inverse of the appropriate covariance functions, but in practice usually reflect subjective belief in the validity of each of the equations. Following Tziperman *et al.* (1992) and Thacker (1992) some existing methods of combining hydrographic data with dynamics are now reviewed in the context of the cost function J .

Diagnostic Calculations

In the diagnostic limit, observed density is treated as a fixed input field ($\kappa_o \rightarrow \infty$). The momentum, incompressibility and hydrostatic equations are considered exact ($\kappa_1 \rightarrow \infty$) and the advection of density is ignored ($\kappa_2 = 0$). It is clear from the results of Section 5.1 that by using these equations and prescribing the baroclinic pressure gradient, the barotropic mode remains undetermined. Additional information must be imposed externally, either as a deep reference velocity or as flows across the open lateral boundaries of the model domain. An important advantage of the diagnostic case is that a linear calculation often suffices to determine the baroclinic transport.

Early applications of the diagnostic method which included topographic effects identified the interaction of the density field and the bathymetry as an important dynamic effect (Sarkisyan and Ivanov, 1971; Holland and Hirschman, 1972). As discussed previously, these JEBAR terms are, in practice, often difficult to evaluate. Errors in the observed density result in apparent misalignments of the bottom density field with the topography leading to erroneous circulation features. Applications of the diagnostic method have dealt extensively with this problem. One early approach was to cast the governing equations in a form such that the JEBAR terms need not be calculated explicitly (Mellor *et al.* 1982, Rattray 1982). An alternative (Sheng and Thompson 1996) is to avoid using gridded density fields and instead determine steric

height directly from density profiles (the rationale being that gridding these leads to a less noisy baroclinic pressure gradient). Such approaches have eliminated many of the problems associated with using a fixed input density field in the diagnostic calculation.

β -Spiral Methods

The β -Spiral method was first introduced by Stommel and Schott (1977). It provides velocity fields from hydrographic data based on identifying unknown deep reference velocities. (This is a solution to the "level of no motion" problem outlined in Section 5.1 and, in some sense, an alternative to explicitly considering the dynamic influence of bottom topography). The governing equations of the β -spiral are based on geostrophy, tracer conservation, and conservation of potential vorticity. The method is so named because the gradient β of planetary vorticity constrains the horizontal velocity to spiral with depth (Bryden 1980). The β -spiral method provides a closed form expression for velocity at any depth in terms of the local properties of the density (tracer) distribution. These velocities can be related to an underlying deep reference flow through the thermal wind relation (5.10). Application of the method uses observations of density at each level to provide an estimate for the reference velocities. A unique value is then chosen from this set of estimates in a least-squares sense (taking into account the spatial covariance structure of the estimates (Olbers *et al.* 1985)).

In the framework of the cost function J , the observed density field is treated as a random variable (κ_o is finite). The geostrophic, hydrostatic, and incompressibility conditions are used to reconstruct the total flow field from the estimated reference velocities ($\kappa_1 \rightarrow \infty$). The density equation is used to obtain the reference velocities, but no attempt is made to update the density field or to ensure a dynamic balance with the recovered velocity (κ_2 is finite). The important simplification of the method results from using the governing equations to express the unknown reference velocities in terms of the density field. This results in a linear estimation problem and drastically

reduces the number of unknowns. Wunsch (1994) offers an extension of the β -spiral approach which updates density and ensures it is in dynamic balance with velocity.

Robust Diagnostic Approach

The robust diagnostic method uses existing ocean models to determine circulation estimates from hydrographic data while avoiding the inverse formalism. The important feature is that the observed density field is allowed to vary (κ_o is finite). Since an ocean model with fairly complete physics is used, the form of J in (5.21) is not strictly applicable (temperature and salinity equations are included). However, the essence of the approach is that errors are permitted in the density specification (κ_2 is finite) while the remainder of the dynamics are satisfied exactly ($\kappa_1 \rightarrow \infty$). Note that these weights are identical to that of the β -spiral, but the approach is quite distinct.

Applications of the robust diagnostic method are not widespread. The method was first introduced by Sarmiento and Bryan (1982). Their approach was to adjust the model predicted temperature and salinity fields towards climatology by allowing for artificial heat and salt sources. More recently, Ezer and Mellor (1994) use an approach whereby they carry out a diagnostic calculation using observed hydrography, followed by prognostic calculation in which those density and velocity fields are allowed to dynamically adjust to one another. The adjustment procedure is not carried out to completion so the resultant fields match neither the purely diagnostic nor the purely prognostic case.

The main advantage of these methods for analyzing hydrographic data is that they are straightforward modifications on existing (and validated) ocean circulation models. However, as is clear from the above examples, the methodology is varied (gridded climatological fields are arbitrarily introduced into the calculations) and errors are difficult to quantify.

Inverse Methods

Inverse methods are able to take into account the stochastic nature of the data and errors in the dynamics. By interpreting the diagnostic, robust diagnostic and β -spiral

methods in the context of a cost function J , it is clear that all these techniques may be viewed as inverse methods whose solution may be approached via an optimization framework. A pioneering application of inverse methods is that of Wunsch (1978) who estimates deep reference velocities from data. It is distinct from the β -spiral method mainly in that it utilizes section data in a singular value decomposition technique (Section 2.1.2) to determine a unique value for the absolute velocity field. Wunsch's section method illustrates an important strength of inverse techniques, the ability to obtain error estimates.

None of the techniques outlined thus far have explicitly addressed the joint estimation of both density and velocity from hydrographic data in order to obtain dynamically balanced fields. A few inverse techniques have, however, considered the issue. The ambitious study of Tziperman *et al.* (1992) attempts to fit a primitive equation model to hydrography, buoyancy fluxes, and eddy mixing parameters to both update these inputs as well as estimate ocean circulation. The results illustrate mainly the difficulty in obtaining a solution (poor conditioning, existence of local minima) in such a large-scale nonlinear inverse problem. Wunsch (1994) and Bogden *et al.* (1993) adjust the input density fields to better satisfy the condition of conservation of density along streamlines (minimum mixing of density) in order to provide for a dynamic balance with the circulation and choose a unique value for the barotropic mode. The study of Bogden *et al.* (1993) is notable since it takes careful account of the influence of bathymetry on circulation. Both of these studies also consider the limit where κ_0, κ_2 are finite and $\kappa_1 \rightarrow \infty$.

5.2.1 Summary

A number of issues have been discussed thus far pertaining to the estimation of circulation from hydrographic data. In a system based only on the geostrophic, hydrostatic and incompressibility conditions, a dynamical null space exists in the sense that the barotropic component of the flow cannot be determined from density information alone. As a result, diagnostic calculations which prescribe the baroclinic

pressure gradient are incomplete. A further problem is that they may also produce unrealistic flow features due to numerical evaluation of JEBAR terms from density data. The root cause of both these defects is the inability of the diagnostic method to allow the density field to adjust to the flow field.

Introducing the equation governing the distribution of density modifies the problem substantially. The nonlinear model couples the density and velocity fields, providing a relation between the barotropic flow and the density field. The dynamic balance between these variables also eliminates potential problems with non-representative JEBAR terms. Density observations must now be integrated into the model using an estimation procedure, since density is now a prognostic variable. β -spiral and robust diagnostic techniques, as well as the inverse methods of Wunsch (1994) and Bogden *et al.* (1993), all treat the density equation as a weak constraint yet implement the techniques in very different ways.

5.3 A Strong Constraint Approach

In this section, we consider a particular limit for the estimation problem in which the dynamics, including the density equation, are treated as a strong constraint. A specific example is used to explore the consequences of this approximation and compare and contrast it to the more usual diagnostic approach.

Coastal and continental shelf regions are characterized by strong topographic variations and extensive open boundaries. The governing equations used in this example provide a simple description of the physics governing the co-evolution of density and velocity in coastal regions, including both JEBAR effects and the advection of density. A boundary control approach is taken whereby the buoyancy fluxes across the open lateral boundaries of the model domain control the density and velocity in the interior (buoyancy fluxes across the air-sea interface are neglected). The assimilation method fits point measurements of density to the model predicted values by adjusting the boundary inflows while satisfying the governing equations exactly.

The model is based on the depth integrated versions of (5.1)-(5.4). The major simplification lies in the use of a vertically well mixed water column in which density is assumed to vary only in the horizontal. Hendershott and Rizzoli (1976) use this assumption in modelling the circulation of the Adriatic Sea where convective processes act to maintain the vertical homogeneity. We instead consider a relatively shallow system where the water column is well mixed due to frictional, or turbulent, processes. While this is recognized as an idealization, it greatly simplifies the problem and yet does not unduly compromise the essential physics. (The physics of the adjustment to steady state are somewhat unrealistic, but are not considered here). The remaining simplifications are that the Coriolis parameter is constant (f -plane), wind stress is neglected, and bottom friction is parameterized in terms of transport.

5.3.1 Model Formulation

Assuming a vertically homogeneous water column and retaining the stress terms in (5.1) implies that the total transport takes the form

$$\vec{U} = \frac{gh}{f} \vec{k} \times \left\{ \nabla \eta + \frac{h}{2} \nabla \varepsilon - \frac{1}{\rho_0 gh} (\vec{\tau}^s - \vec{\tau}^b) \right\}, \quad (5.22)$$

where the baroclinic pressure gradient follows from (5.14). Surface and bottom stresses are denoted by $\vec{\tau}_s$ and $\vec{\tau}_b$, respectively.

Consider these stress terms. The surface stress $\vec{\tau}_s$ is due to the momentum transfer from the wind to the water column and is neglected here for simplicity. Bottom stress can be parameterized in terms of the depth-mean velocity in cases where internal friction is high, such as in a well-mixed water column. Following Csanady (1979) we adopt the formula

$$\frac{\vec{\tau}_b}{\rho_0} = r \frac{\vec{U}}{h}, \quad (5.23)$$

which assumes that bottom friction varies linearly with the depth averaged velocity.

To further simplify the governing equations, the transport streamfunction is introduced, and defined as

$$\frac{\partial \Psi}{\partial x} = -V, \quad \frac{\partial \Psi}{\partial y} = U. \quad (5.24)$$

Substituting the transport streamfunction in (5.22) and taking the curl of the resulting equations yields the vorticity equation

$$J\left(\Psi, \frac{f}{h}\right) + \frac{g}{2}J(\varepsilon, h) + r\nabla \cdot \left(\frac{\nabla\Psi}{h^2}\right) = 0 \quad (5.25)$$

where J denotes the Jacobian operator

$$J(A, B) = \frac{\partial A}{\partial x} \frac{\partial B}{\partial y} - \frac{\partial A}{\partial y} \frac{\partial B}{\partial x}.$$

The first term in (5.25) represents vortex stretching associated with flow across isobaths. The second term is the JEBAR term and describes the vorticity generated by bottom torque associated with a misalignment of the density field and the bathymetry. The final term represents the dissipative effects of bottom friction.

To allow the density field to co-evolve with the transport, the density equation (5.4) is required. Vertically integrating this equation under the assumption of a depth-independent density anomaly yields

$$J(\varepsilon, \Psi) - hK\nabla^2\varepsilon = 0. \quad (5.26)$$

Here, the first term represents the advection of density and the second term represents horizontal mixing.

As a result of these manipulations, our governing equations now consist of a vorticity equation for the transport (5.25) and a tracer equation for density (5.26). These compose a system of nonlinear, elliptic partial differential equations in two spatial dimensions and require boundary conditions on Ψ and ε . This may be contrasted to the inviscid limit of (5.1)-(5.4) which give hyperbolic equations, whose properties were explored in Section 5.1. The conclusions of that section will be expected to hold in an approximate sense, provided that the frictional and dissipation terms do not dominant the system.

5.3.2 Assimilation

The data assimilation procedure used here matches point observations of density to the model predictions. The adjustable parameters (control variables) are chosen to be

the values of Ψ and ε at inflow boundaries of the model. These define buoyancy fluxes across the open boundaries into the model interior. The goal is then to minimize the cost function

$$J = \sum_{k=1}^K (\varepsilon_{\text{obs}}^k - \varepsilon_{\text{mod}}^k)^2 \quad (5.27)$$

with respect to the controls while satisfying the governing equations (5.25) and (5.26) exactly. In the above, $\varepsilon_{\text{obs}}^k$ represents the k th observation and $\varepsilon_{\text{mod}}^k$ is its model counterpart. In the context of (5.21), we have allowed κ_o to remain finite while $\kappa_1, \kappa_2 \rightarrow \infty$. This limit is distinct from any of the methods surveyed in Section 5.2.

The optimization associated with (5.27) is carried out in a discrete vector space and, for clarity, we discuss our assimilation method in this context. Suppose estimates for the unknown Ψ, ε at the inflow boundaries are given, and appropriate conditions are specified at the outflow and solid boundaries. Interior values of Ψ, ε can then be determined through a discrete version of the system (5.25)-(5.26) given by

$$\mathbf{x} = \mathbf{d}(\mathbf{b}). \quad (5.28)$$

Here, the vector \mathbf{x} contains the values of Ψ, ε at the n model interior points and \mathbf{b} contains parameters which define Ψ, ε at the inflow boundary points. The nonlinear operator \mathbf{d} contains the discretized dynamics of the boundary value problem given by (5.25)- (5.26). In other words, it represents a numerical model which serves to map from the boundary state to the interior.

To compare point observations of the density field, contained in the $k \times 1$ vector \mathbf{z} , to the values at the model grid points we introduce the $k \times n$ interpolation matrix \mathbf{H} such that

$$\mathbf{z} = \mathbf{H}\mathbf{x} + \mathbf{e}, \quad (5.29)$$

where \mathbf{e} represents the observation error. The interpolation scheme is assumed linear since, in this case, a model variable (density) is observed directly. Substituting from (5.28) into (5.29) yields the regression equation

$$\mathbf{z} = \mathbf{H}\mathbf{d}(\mathbf{b}) + \mathbf{e}, \quad (5.30)$$

where the first term on the right hand side is a nonlinear operator which now includes both the dynamics and the spatial interpolation scheme and relates the unknown \mathbf{b} to the observations \mathbf{z} .

The overall goal of the estimation is to choose \mathbf{b} such that the squared error

$$J = \mathbf{e}^T \mathbf{e}$$

is minimized. Once the optimal boundary inflows are determined, the interior values \mathbf{x} may be obtained using (5.28). Overall, the problem outlined is one of nonlinear regression.

5.3.3 Application

Numerical solution of the system of nonlinear partial differential equations (5.25)-(5.26), or equivalently the evaluation of $\mathbf{d}(\mathbf{b})$, was based on the algorithm of Melgaard and Sinovec (1981). It uses the method of lines technique (Madsen and Sinovec 1974) in which the partial differential equations are mapped into a set of ordinary differential equations and solved using a standard ODE integrator. Dirichlet, Neumann and mixed boundary conditions are supported and the algorithm solves the time dependent initial-boundary value problem in two dimensions. To obtain the steady-state solution appropriate for (5.25)-(5.26), artificial time derivative terms were introduced and the model integrated forward from a zero initial state until these terms effectively vanished. This is analogous to the method of successive over-relaxation (e.g. Press *et al.* 1989).

The governing equations (5.25)-(5.26) were scaled both to facilitate interpretation of the results, as well as allow efficient implementation of the PDE solver and optimization routines. The scaling is carried out as follows:

$$(x, y, h, \Psi, \varepsilon) = (Lx^*, Ly^*, \bar{h}h^*, U\bar{h}L\Psi^*, \bar{\varepsilon}\varepsilon^*),$$

where the * superscripts denote the scaled variables. The scaling factors were the horizontal length scale $L = 250\text{km}$ which is the east-west dimension of the model

domain, the vertical scale $\bar{h} = 100\text{m}$ representing a maximum depth, a horizontal density variation $\bar{\varepsilon} = 10^{-3}$, and a typical velocity $U = 0.1\text{ms}^{-1}$. The remaining parameters in the problem are gravity $g = 10\text{ms}^{-2}$, a typical mid-latitude value of the Coriolis parameter $f = 10^{-4}\text{s}^{-1}$, a friction coefficient $r = 10^{-3}\text{ms}^{-1}$, and horizontal diffusivity $K = 10^2\text{m}^2\text{s}^{-1}$. The bathymetry in Figure 1 and all results shown hereafter are scaled based on these values. Given this scaling, the dissipation terms in (5.25) and (5.26) are about one-tenth the values of the other terms in their respective equations. For the remainder of the chapter, we drop the * notation but the above scaling is to be understood.

The model geometry was designed to mimic a limited-area coastal model and is shown in Figure 5.1. The bathymetry reflects an idealized shelf with a central gully, bounded by one coastal and three open boundaries. The northern boundary is chosen as the inflow boundary and $\Psi(x)$ and $\varepsilon(x)$ are specified there. (The rationale for this placement of the inflow boundary is that in the northern hemisphere long shelf waves propagate both phase and energy with shallow water to their right). The coastal boundary to the west has a no flow condition and the streamfunction is constant. The corresponding condition on density is $\partial\varepsilon/\partial x = 0$ allowing for density variations to occur along the coast. At the offshore boundary to the east, the streamfunction and density are fixed at specified deep ocean values. The remaining southern boundary is an outflow boundary and $\partial\Psi/\partial y = \partial\varepsilon/\partial y = 0$ there. The model is based on a 20×10 grid with uniform spacing.

The buoyancy fluxes across the northern boundary into the model domain constitute the control variables of the problem. In keeping with our philosophy of maximum simplicity, we assume a linear slope in Ψ and ε from coastal values of zero to specified offshore values. (To account for the condition $\partial\varepsilon/\partial x = 0$ at the coast, a zero-slope region, comprising one-tenth the east-west dimension, is added piecewise to the sloped region of ε .) The streamfunction and density at the northern boundary are given by

$$\Psi = \mathbf{b}_\Psi x, \quad \varepsilon = \mathbf{b}_\varepsilon x$$

and only the slopes \mathbf{b}_Ψ , \mathbf{b}_ε need to be specified, thereby reducing the number of control

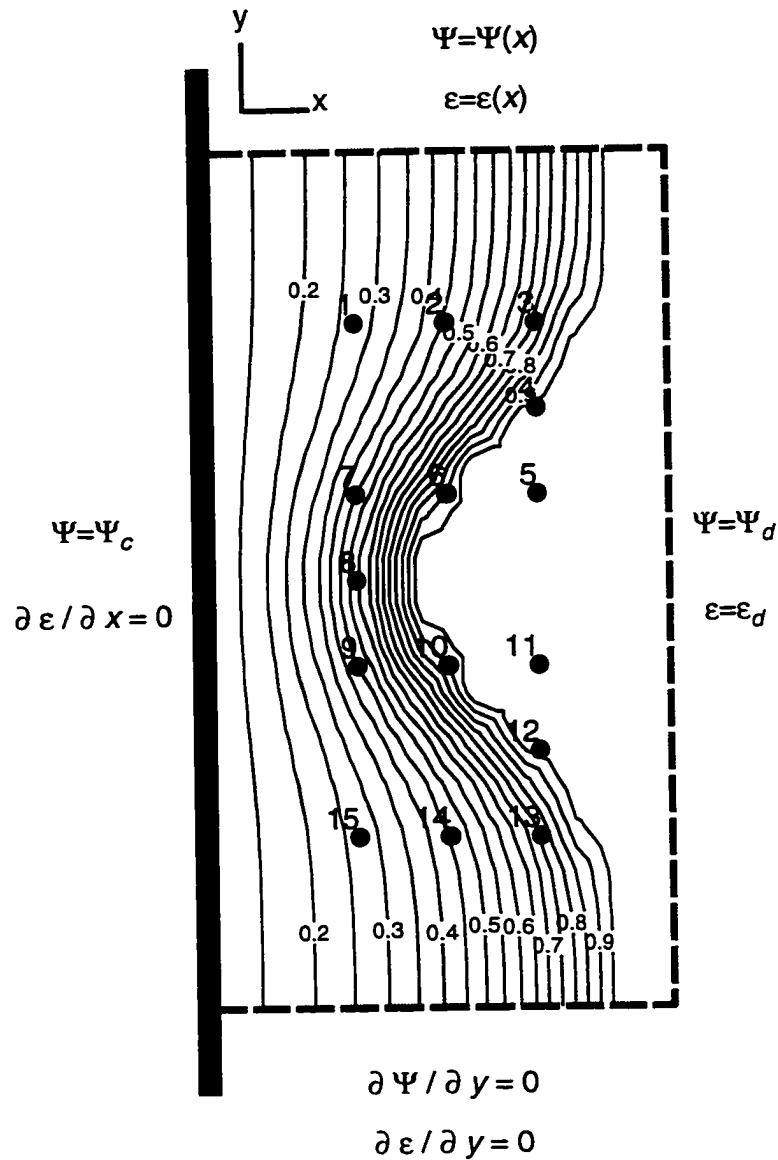


Figure 5.1: Geometry of the idealized coastal model. The contours represent the bottom topography and are scaled by a depth of 100m. The thick solid line is the coastal boundary and dashed lines represent open boundaries. The solid dots indicate measurement locations and are identified by their associated numbers. The boundary conditions used in the model are indicated adjacent to each boundary and the coordinate axes are shown.

variables to two.

The northern boundary is an inflow boundary. The streamfunction along this boundary is given by

$$\Psi = \Psi_{bc} + \Psi_{bt}$$

where the subscripts bc and bt on Ψ are the baroclinic and barotropic contributions to the total streamfunction, respectively. The total streamfunction Ψ is governed by \mathbf{b}_Ψ , which is constrained to be greater than one in order to ensure an inflow into the model domain. Specifying the density gradient through \mathbf{b}_ε then designates the relative strength of the baroclinic and barotropic components of the flow. We consider cases where \mathbf{b}_ε is greater than zero, which corresponds to fresher water near the coast. Recall the relation between transport and the baroclinic pressure gradient as given by (5.22). If the frictional stress terms and sea level gradients are neglected, the baroclinic component of the transport streamfunction (denoted by the s subscript) in terms of the scaled variables is given by

$$\frac{\partial \Psi_{bc}}{\partial x} = - \left[\frac{g\bar{h}\bar{\varepsilon}}{2UfL} \right] h^2 \frac{\partial \varepsilon}{\partial x}.$$

With the chosen values for the scale parameters, the value of the square bracketed term is $1/5$. Based on this result, we consider two cases by which to illustrate the joint estimation of transport and density: (i) a case dominated by the barotropic pressure gradient ($\mathbf{b}_\Psi = 0.5, \mathbf{b}_\varepsilon = 0.1$), and (ii) a case dominated by the baroclinic pressure gradient ($\mathbf{b}_\Psi = 0.1, \mathbf{b}_\varepsilon = 0.5$). These both correspond to inflows into the model domain, due to the barotropic pressure gradient associated with the sea level gradient.

In the barotropically dominated case, shown in Figure 5.2, the density field is governed by $\mathbf{b}_\varepsilon = 0.1$ which, according to the above relation, implies that the baroclinic portion of \mathbf{b}_Ψ has a value of only 0.02. The actual value of \mathbf{b}_Ψ used for this case is 0.5 and thus the baroclinic pressure gradient is much smaller than the barotropic pressure gradient. As a result, the transport streamlines roughly follow isobaths. There is some evidence of cross-isobath transport due to bottom friction but the requirement

of conservation of mass (fixed coastal and offshore streamfunction values) minimizes this effect. Examining the vorticity equation (5.25) reveals that it is dominated by the stretching term. The density field effectively acts as a passive tracer and is advected so that it reflects the pattern found in the streamlines, while satisfying the necessary boundary conditions. Note that the offshore boundary condition results in a diffusive flux of density into the model domain. This weakens the cross-shelf density gradient in the southern part of the domain.

The baroclinically dominated case, shown in Figure 5.3, is distinctly different. The inflow boundary conditions indicate an approximate balance between the transport and the baroclinic pressure gradient. This corresponds to a flux of freshwater at the coast which is advected downstream while at the same time its cross-shelf gradient is gradually eroded by the diffusive flux of density from offshore. The JEBAR term is a significant component of the overall vorticity balance and misalignments in the density field and the bathymetry drive a variety of gyres in regions of steep topography. The gyre in the southern portion of the model domain is a combination of the JEBAR effect and a return flow required to satisfy conservation of mass. Note that the density is only minimally affected by these small scale circulation features. Moreover, the cross-shelf density gradient is eroded as we move downstream due to advective and diffusive fluxes of buoyancy, and the system gradually becomes more dominated by the barotropic pressure gradient.

A diagnostic calculation is now illustrated in which a prescribed density field is used to determine the baroclinic pressure gradient which, in turn, drives the transport. Note that by fixing the density field ϵ , we eliminate (5.26), and treat the JEBAR terms as a known forcing in (5.25). In practice, this input density field would be obtained from available profiles of density. To mimic this procedure, density measurements were extracted from the baroclinically dominated situation at the 15 observation locations shown in Figure 5.1. (The barotropic case is of little interest here due to the inability of density to capture the barotropic part of the flow). These error-free measurements were statistically mapped to the grid points of the model using

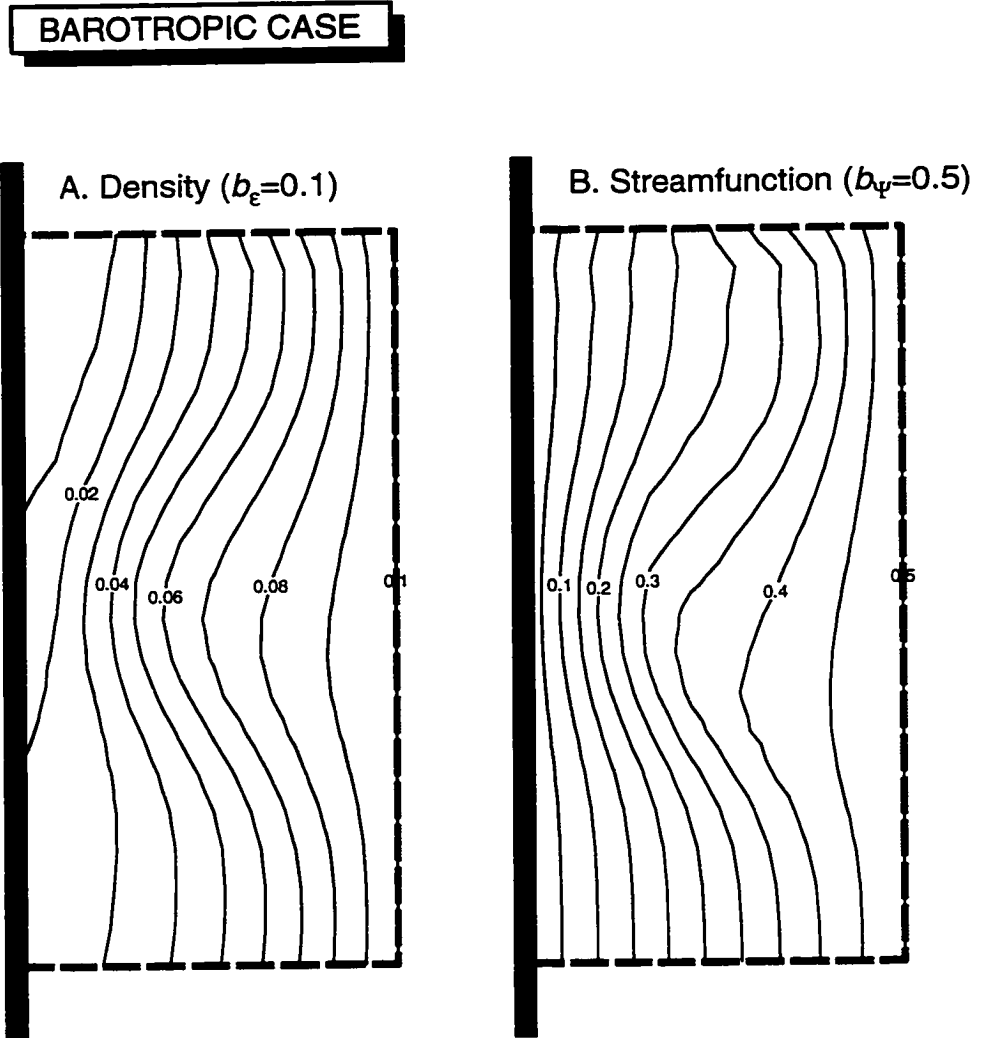


Figure 5.2: Baseline results for the barotropically dominated case. Panel A shows the density anomaly and Panel B shows the transport streamfunction. The control parameters governing the problem are indicated above each panel. Results are scaled as indicated in the text.

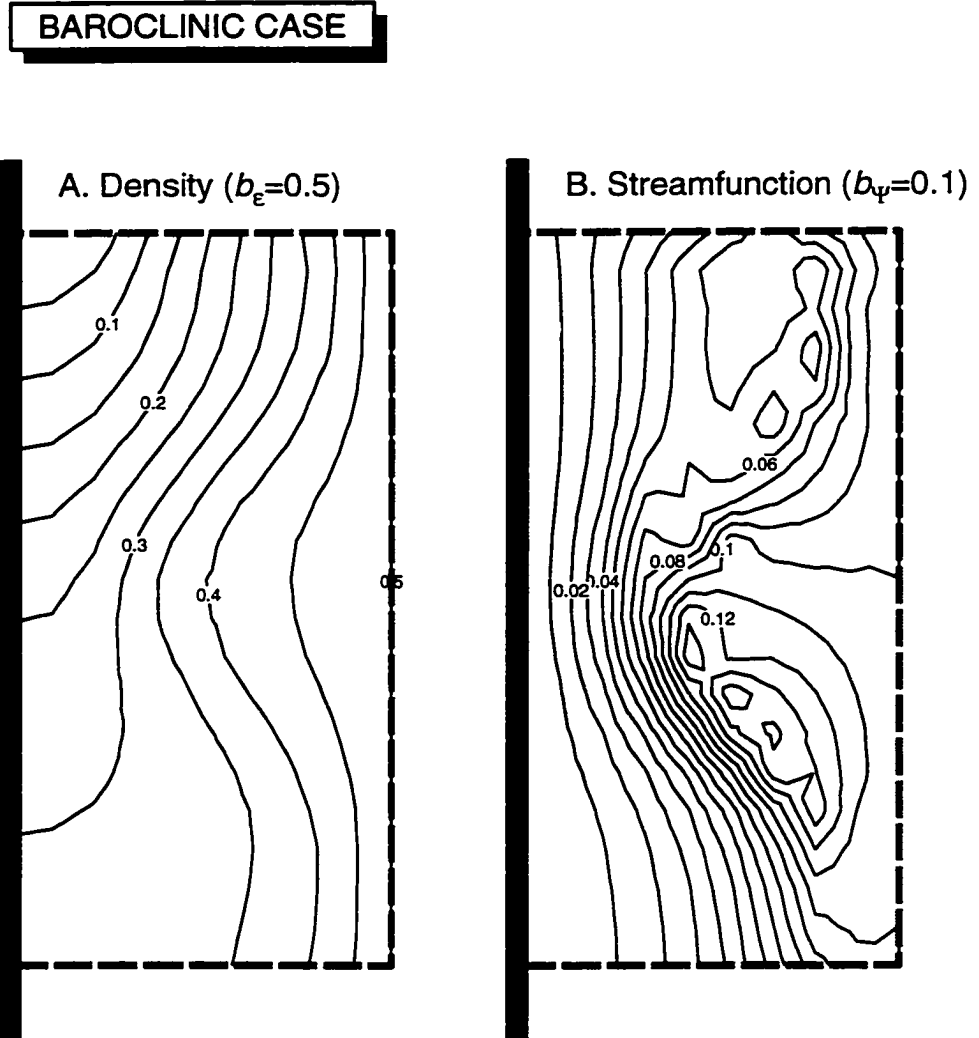


Figure 5.3: Baseline results for the baroclinically dominated case. Panel A shows the density anomaly and Panel B shows the transport streamfunction. The control parameters governing the problem are indicated above each panel. Results are scaled as indicated in the text.

Barne's algorithm (Daley 1991, section 3.6) and the result is shown in Figure 5.4. This recovered density field resembles that of Figure 5.3 and is reasonably accurate in the region where the observations are located.

The streamfunction associated with the diagnostic calculation is shown in Panel B of Figure 5.4. Its pattern is very different from the actual streamfunction value (Figure 5.3, Panel B), particularly with regard to the strong gyre present over the gully region. To isolate the source of this spurious gyral circulation, the actual and diagnostic JEBAR forcing was calculated and the results are shown in Figure 5.5. The actual JEBAR associated with the balanced density field is of much smaller magnitude than its diagnostic counterpart and has its largest values widely distributed over the regions of steep bottom topography. The diagnostic JEBAR forcing is very large in the gully region and matches the gyral circulation pattern. The erroneous circulation in the diagnostic calculation results entirely from what appears to be a rather slight misalignment of the input density field and the bathymetry over this region of rapidly varying topography. This example clearly illustrates the sensitivity of the diagnostic method to JEBAR and motivates the need for dynamically balanced density and transport.

This joint estimation of density and transport was investigated using a series of numerical experiments based on both the barotropic and baroclinic cases. The nonlinear regression technique outlined in Section 5.3.2 provided the basis for the assimilation scheme with $\mathbf{b} = (\mathbf{b}_\psi, \mathbf{b}_e)$ and $\mathbf{d}(\mathbf{b})$ given by (5.25) and (5.26). The observation array is shown in Figure 5.1 and the interpolation matrix \mathbf{H} in (5.30) is defined on this basis. It serves simply to pick the appropriate values of predicted density from the model grid. The squared observation/model discrepancy $J = \mathbf{e}^T \mathbf{e}$ is minimized using a nonlinear optimization based on NAG routine E04FCF which is outlined in Gill *et al.* (1981, Section 4.7.5). This algorithm corresponds to the iterative Gauss-Newton procedure for nonlinear regression given in Sen and Srivastiva (1990, Appendix C). It is based on a series of linear regressions which, starting from an initial guess, converges on the optimal value of the unknown \mathbf{b} . Due to the nonlinear

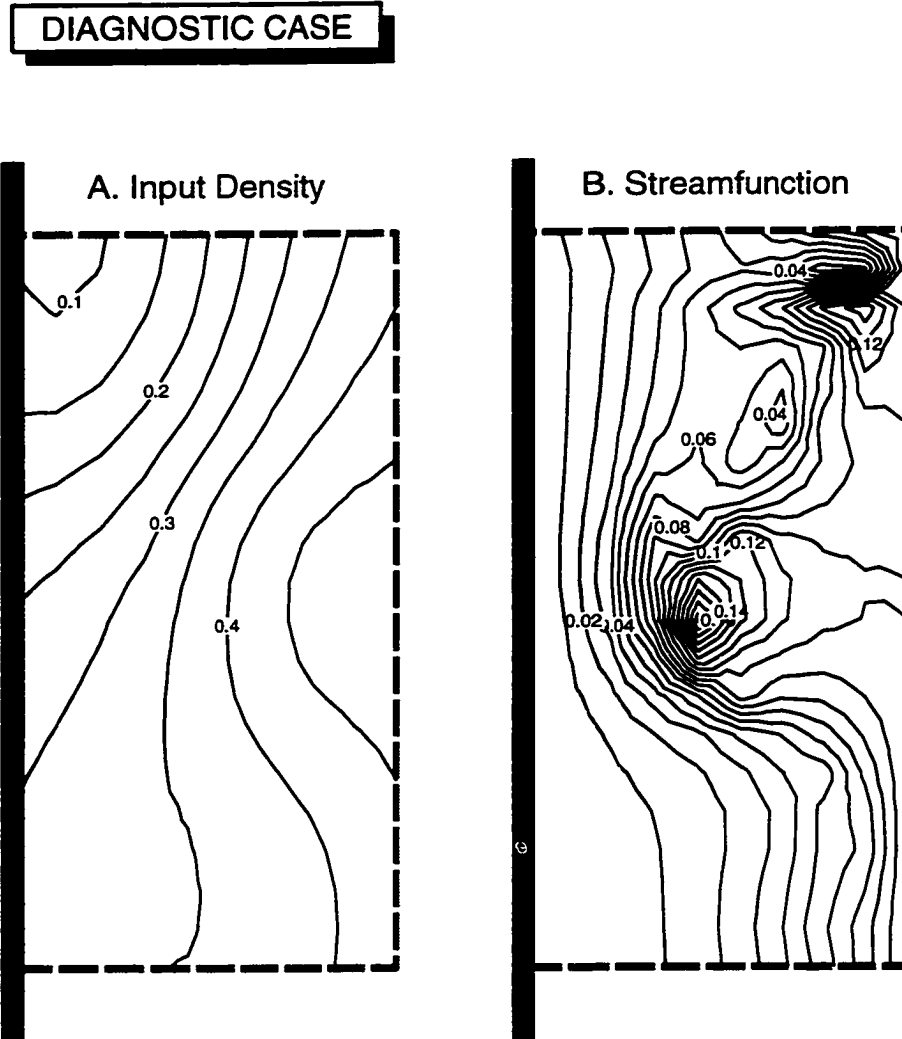


Figure 5.4: Sample diagnostic calculation. Panel A represents the input density field derived from analysis of the 15 density observations in the baroclinic case using Barne's algorithm. Panel B shows the transport streamfunction derived from this density field using the diagnostic method. Results are scaled as indicated in the text.

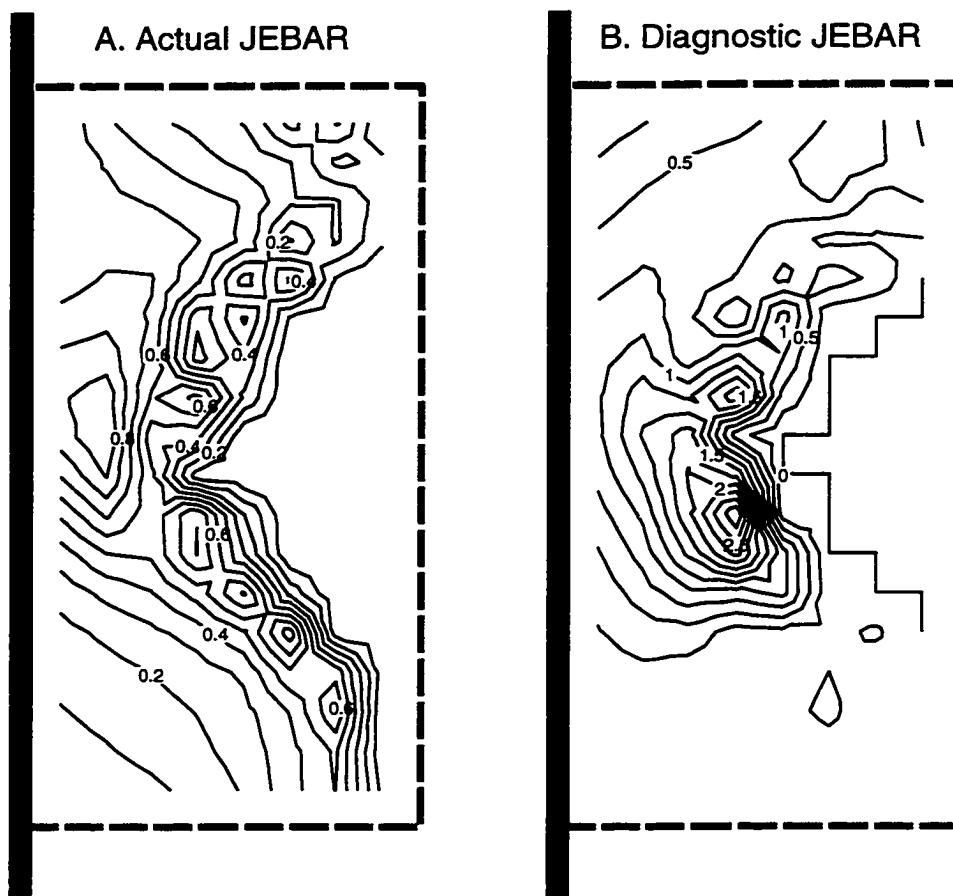


Figure 5.5: Panel A. Actual JEBAR forcing calculated from the dynamically balanced density field in the baroclinic case. Panel B. JEBAR forcing as determined from the input density field used in the diagnostic calculation. In both cases, the results have been normalized by the maximum value of the actual JEBAR forcing. The region adjacent to the boundary is blank since the spatial derivatives of ϵ and Ψ required to calculate JEBAR are not defined there.

model, J is non-quadratic and therefore local minima are possible.

Assimilation experiments for the barotropic and baroclinic situations used both exact and noisy point measurements of density as shown in Figure 5.6. The noisy observations were obtained by adding normally distributed random noise of a magnitude indicated in the plots and which sets the signal to noise ratio. These are intended to test the robustness of the minimization in the presence of observation error. The ability of the assimilation to recover the true values of buoyancy fluxes at the inflow boundary ($\mathbf{b}_\psi, \mathbf{b}_\epsilon$) is outlined below for the barotropic and baroclinic cases, with and without noise.

(i) Barotropic Case

Performance diagnostics for the assimilation based on the barotropic case are shown in Figure 5.7. Using exact observations of density, the true inflow boundary conditions were successfully recovered. It can be seen that the squared error and the magnitude of the gradient of the cost function decrease somewhat erratically, reflecting the fact that the assimilation first adjusted \mathbf{b}_ϵ to near its optimal value, then more slowly converged towards \mathbf{b}_ψ in the remaining iterations.

For noisy measurements, the assimilation recovered the actual value of \mathbf{b}_ϵ , but failed to determine \mathbf{b}_ψ . Figure 5.7 shows that the squared error asymptotes at about 10^{-2} while the gradient of the cost function decreases as it attempts to adjust \mathbf{b}_ψ to further minimize J , but with little effect. The final value of \mathbf{b}_ψ obtained by the assimilation is greater than the baroclinic contribution to the flow, but less than the value of \mathbf{b}_ψ which gave rise to the observed density field. The use of density information for recovering both the streamfunction and density fields appears to be very sensitive to noise in the observations.

Examining the cost function in the vicinity of the minimum further illustrates the sensitivity of the assimilation to noise. Panels A and B in Figure 5.8 show contour plots of the cost function for the barotropic case. For exact observations, the cost function shows a distinct minimum at $\mathbf{b}_\psi = 0.5, \mathbf{b}_\epsilon = 0.1$. However, it also reveals the insensitivity of the cost function to changes in \mathbf{b}_ψ at regions outside $0.08 < \mathbf{b}_\epsilon < 0.12$.

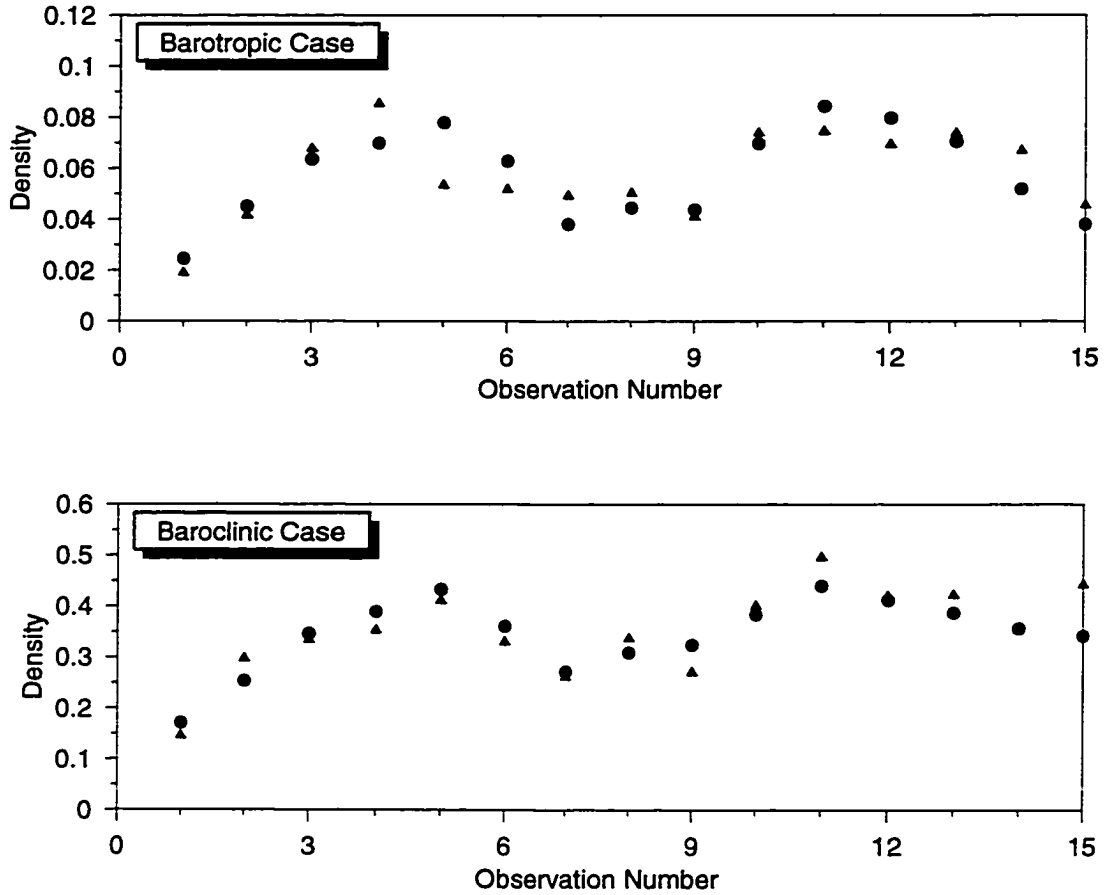


Figure 5.6: Point observations of density at the observing locations shown in Figure 5.1 for both the barotropic and baroclinic cases. Solid dots refer to the actual values, and triangles represent those same observations with random noise added.

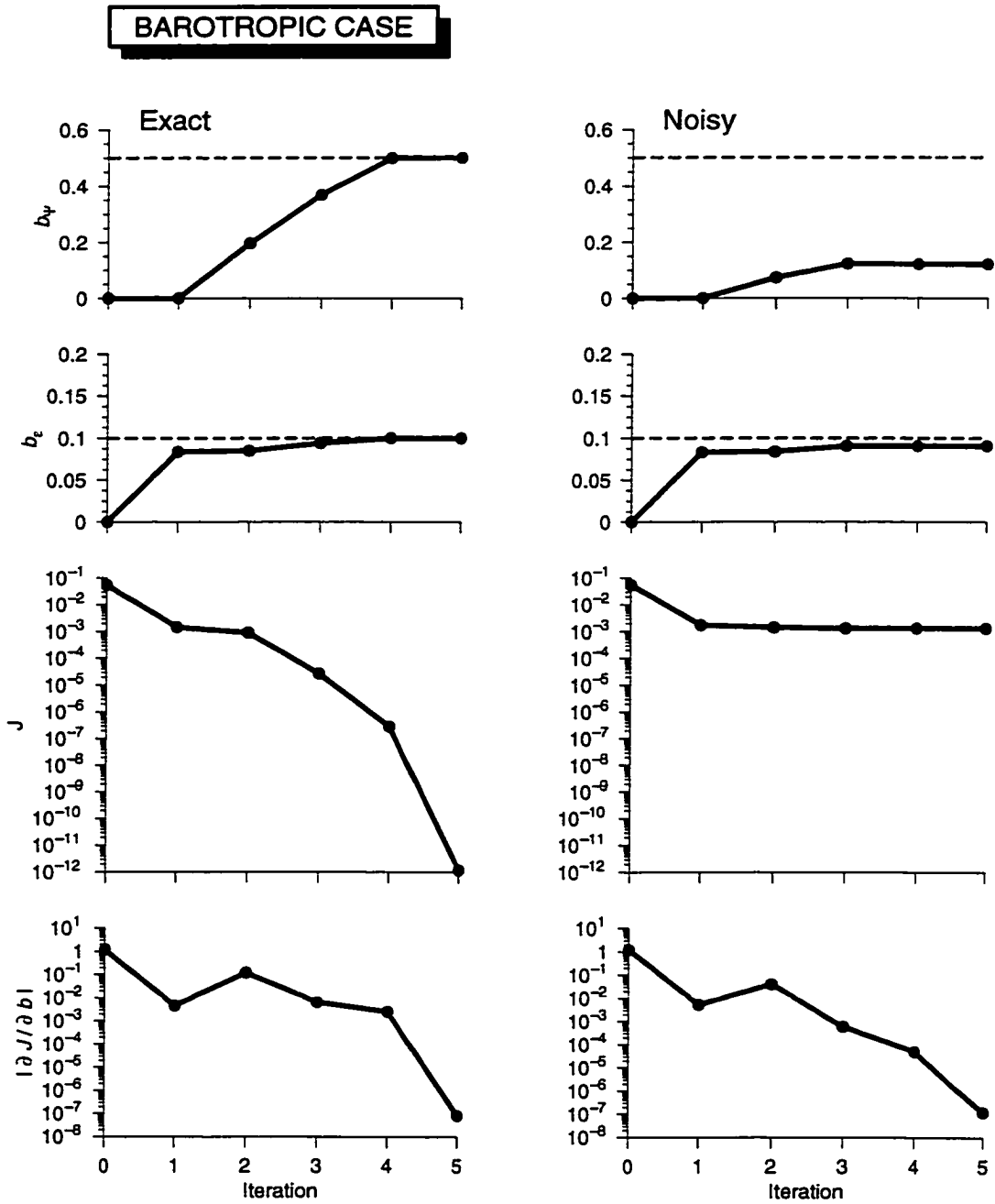


Figure 5.7: Performance diagnostics for the joint estimation problem in the barotropic case for both exact and noisy density observations. Here, b_ψ and b_ϵ refer to values of the control variables, J is the sum of squares of the errors and $|\partial J/\partial b|$ refers to its gradient. Iteration number corresponds to the minimization associated with the nonlinear regression.

(This provides a rationale for the convergence path of the minimization). For noisy observations, the minimum of the cost function is indistinct, being found near $\mathbf{b}_\epsilon = 0.1$ but ranging over \mathbf{b}_ψ . Clearly, the cost function is insensitive to changes in \mathbf{b}_ψ in the presence of noise, at least in this region of parameter space.

The failure of the assimilation to recover the minimum in \mathbf{b}_ψ in the presence of noise appears to be related to the null space associated with the barotropic component of transport. This assertion was examined in more detail. Define \mathbf{D} as the matrix associated with the linearization of $\mathbf{d}(\mathbf{b})$ in (5.30) about the optimal \mathbf{b} . The eigenvalues and eigenvectors of $\mathbf{D}^T \mathbf{D}$ give us information on the conditioning of this linear regression problem and is locally valid near the minimum (see Section 2.1). It was found that the eigenvector associated with the \mathbf{b}_ψ direction has a near-zero eigenvalue, confirming the existence of a null space associated with \mathbf{b}_ψ (in the vicinity of the minimum) in the presence of noise.

It is concluded that the strong barotropic component coupled with the errors in the data creates an indeterminacy in the estimation problem. Density in this case is only weakly related to the barotropic transport and, in effect, acts as a passive tracer. For a successful assimilation, either exact information on the tracer distribution is required or additional information on the transport streamfunction would need to be introduced to the problem.

(ii) Baroclinic Case

The assimilation of limited density observations in the baroclinic case appears to be more robust than its barotropic counterpart. The performance diagnostics for the assimilation, shown in Figure 5.9, indicate that it converges efficiently to the actual values of \mathbf{b}_ψ and \mathbf{b}_ϵ for both exact observations. For the noisy observations, \mathbf{b}_ϵ is recovered almost exactly, while \mathbf{b}_ψ is over-estimated somewhat but consistent with the density errors. The cost function and its gradient with respect to the controls decrease rapidly. Like the barotropic case, the assimilation first adjusts density to near its optimal value and then adjusts the streamfunction, a strategy that arises due to the fact that only density is being observed.

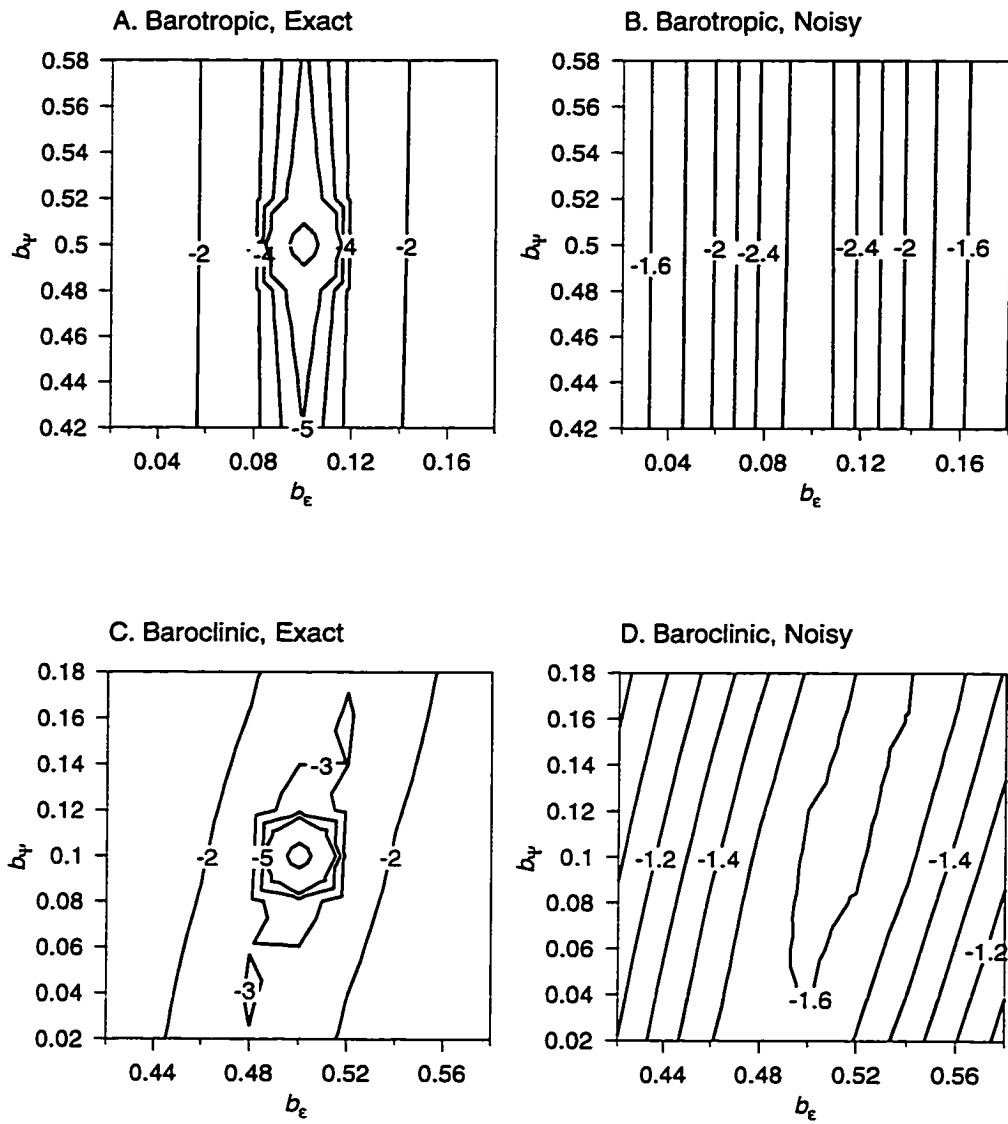


Figure 5.8: Contour plots of the cost function (base 10) in control variable space for the various test cases.

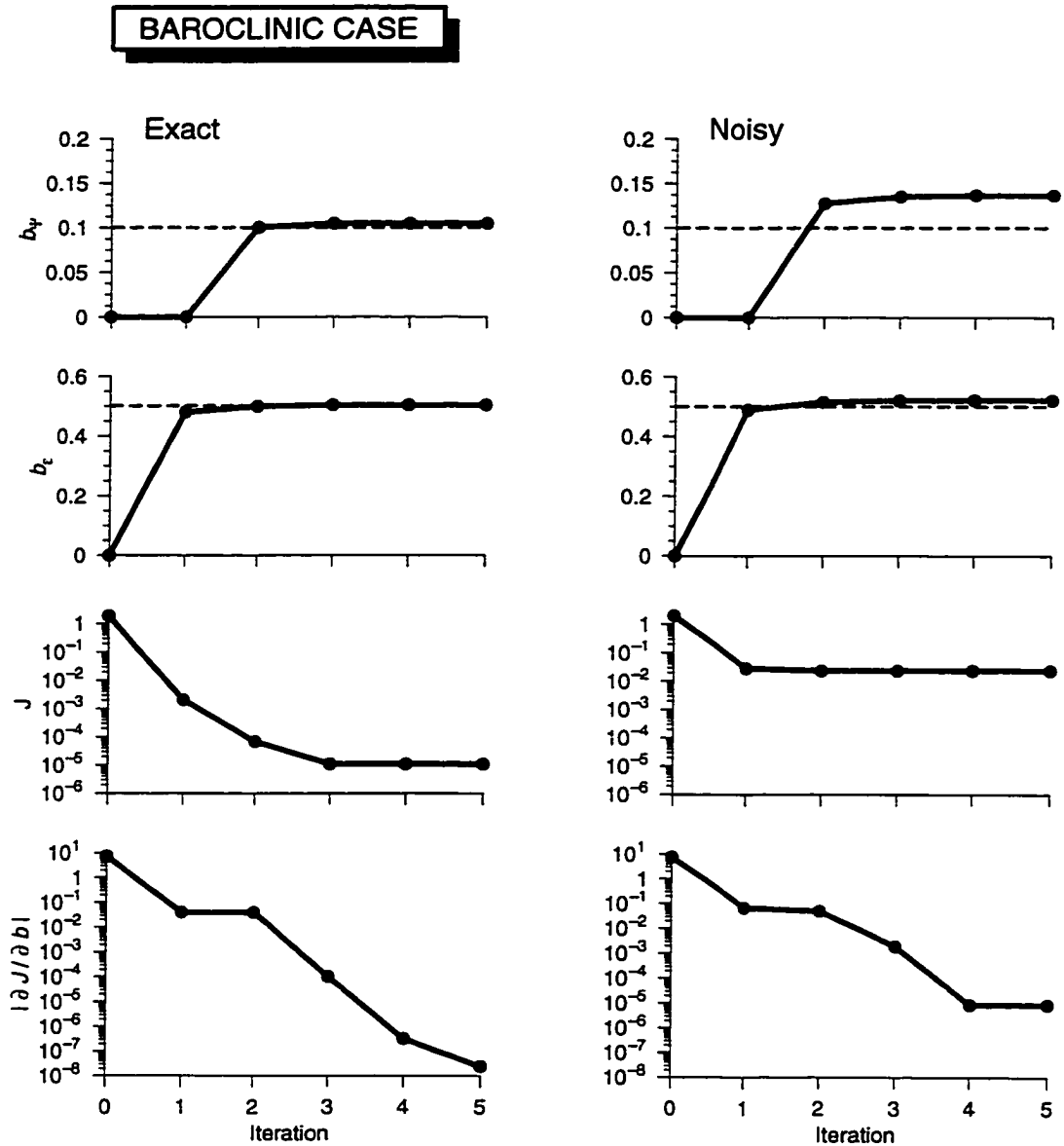


Figure 5.9: Performance diagnostics for the joint estimation problem in the baroclinic case for both exact and noisy density observations. Here, b_ψ and b_e refer to values of the control variables, J is the sum of squares of the errors and $|\partial J/\partial b|$ refers to its gradient. Iteration number corresponds to the minimization procedure associated with the nonlinear regression.

Panels C and D of Figure 5.8 show the shape of the cost function in the vicinity of the minimum for the baroclinic case. For exact observations, a sharp minimum in the cost function is found. But when noise is added to the observations, the minimum becomes less distinct. However, an important feature is revealed by the shape of J . The major axis is tilted with respect to \mathbf{b}_ψ and \mathbf{b}_ϵ implying that J is sensitive to variations in both \mathbf{b}_ψ and \mathbf{b}_ϵ . The eigenvectors of $\mathbf{D}^T\mathbf{D}$ confirm this tilt of 75 degrees, and it is notable that the associated eigenvalues differ only by a single order of magnitude.

The influence of each of the individual observations on the baroclinic assimilation was also examined. Density measurements which occurred over regions of steep topography were found to be the most important in the estimation of \mathbf{b}_ψ . This is sensible since they are of primary importance in accurately determining the JEBAR part of the overall vorticity balance. It is concluded that the joint estimation of density and transport for the baroclinically dominated case appears to be well conditioned even in the presence of noise. The reason is that density acts as a dynamically active tracer and thus has a strong dynamic link to transport.

5.4 Discussion and Conclusions

In this chapter, we have investigated the problem of estimating circulation from density data. Using the geostrophic, hydrostatic, and incompressibility relations, the well known indeterminacy in obtaining the barotropic mode from density data was demonstrated. To determine an absolute velocity field in this case, external information is required on the barotropic flow or sea level gradient. An alternative is to include the density equation explicitly into the analysis thereby dynamically linking the density to the flow. The result is a system of nonlinear governing equations giving balanced density and velocity fields. An estimation procedure is required since, with density a prognostic variable, the model must be fit to the density data.

The general problem of determining circulation from density can be approached

in the framework of variational data assimilation. A general cost function was proposed and existing estimation procedures were examined in this context. An example was then provided to investigate the particular limit where the dynamics, including the density equation, were considered a strong constraint. This example focused on coastal regions and was based on an idealized model which included JEBAR and the advection of density. A boundary control approach was used whereby the buoyancy fluxes across the inflow boundaries of the model domain were estimated from point observations of density in the interior. This simple model demonstrated that in a barotropically dominated case, the transport was very sensitive to noise in the density field. Density in this case acted as a passive tracer of which accurate measurements were required to recover transport. For the baroclinically dominated case, density was dynamically active and thus contained additional useful information on the flow. As a result, the assimilation was robust in the presence of observational noise.

The idealized example given in Section 5.3 not only illustrates the issues involved in the joint estimation of density and transport, it also serves to identify the features which must be considered for any practical application. Methods based on the approach outlined here might provide a possible alternative to diagnostic calculations commonly carried out in shelf regions. Importantly, these methods produce dynamically balanced density and flow fields which take into account the coupling between the barotropic and baroclinic components of the flow. This allows more accurate determination of the bottom torques associated with JEBAR. Below, we further speculate on issues arising in more practical applications.

The model given by (5.25)-(5.26) has a number of unrealistic assumptions. The most important of these is the use of a vertically homogeneous density field. While this assumption is rarely satisfied in practice, it has allowed us to obtain a very simple model with which to investigate the joint estimation problem. A number of dynamic factors must be considered to account for depth dependence in density. Examining (5.14) indicates that we must depth integrate over the vertically varying horizontal

density gradient in order to obtain the baroclinic transport. The advection of density given by (5.19) also reveals that the interaction between the vertical profiles of density and the sheared velocity associated with thermal wind must be taken into account.

Allowing for vertical variations in density introduces complicated three-dimensional structure to the density and velocity fields. While implementing such a model may be tedious, its application presents no conceptual difficulty for numerical methods. Clearly, the same conclusion also applies to the addition of surface wind stress and buoyancy fluxes, relaxing the f -plane assumption, and adding realistic coastlines and bottom topography. In this regard, the work of Salmon (1994) is notable in that it formulates a relatively simple, generalized two-layer model which is able to account for the important physics associated with horizontally and vertically variable density.

We have chosen to use a boundary control approach as the basis for the assimilation problem. This implies that the model equations are treated as strong constraints and can be parameterized in terms of the boundary conditions. In Section 5.3, advection of density across the lateral open boundaries was assumed to control the distribution of density and velocity in the model interior. In realistic cases, buoyancy fluxes may also occur through the surface boundary (heat transfer and evaporation/precipitation) and the coastal boundary (freshwater input from rivers and runoff). If these sources are deemed significant, they must either be prescribed, or estimated, as part of the overall modelling and assimilation procedure.

One area in which the boundary control procedure may not be suitable is for cases where the steady state assumption breaks down. Advection of density in the governing nonlinear equations offers the possibility of (baroclinic) instabilities. This was evident in the successive over-relaxation procedure used to achieve a steady state. When the density gradients were large, a steady state could not be achieved (or achieved in only a statistical sense) due to what appeared to be an instability process. This implies some limits for the relevant range of \mathbf{b}_ψ and \mathbf{b}_e . It also suggests some problems with extending the approach to the boundary control of time dependent models with a prognostic density equation. If the model generates internal instabilities which are

only weakly linked to the boundary conditions, boundary control will prove difficult. An approach which updates the entire field (such as the Kalman filter) may therefore be more suitable.

In spite of these issues, a simple application using a realistic nonlinear ocean model might be feasible. The key issue is keeping the number of unknowns small, so that a minimization without explicit gradient information could be carried out (the numerical evaluation of gradients is only feasible for a small number of unknowns). Consider a semi-enclosed basin where the circulation is controlled by buoyancy fluxes from a small number of well defined locations (including rivers), and where surface buoyancy fluxes were small. With a good first guess of the circulation and buoyancy inputs (e.g. from a diagnostic calculation using climatological density), the boundary control method might provide a means to refine these initial density and velocity fields in a dynamically consistent manner. If the number of unknowns is large, gradient information, say from an adjoint model, would be needed to facilitate the minimization.

In summary, we suggest that some of the shortcomings of the commonly used diagnostic method can be overcome by explicitly considering the evolution equation for density. In practice this will likely involve assimilating hydrographic data into contemporary ocean models based on a fairly complete set of dynamic and thermodynamic equations. A simple boundary control approach based on such models might prove to be a useful first step. However, identifying a workable approach to assimilating data into a complex, nonlinear ocean with a prognostic density equation remains as an area for future research.

Chapter 6

Concluding Remarks

In this thesis, some aspects of assimilating data into coastal models have been investigated. The approach was based on isolating the essential physical processes involved in each of the assimilation problems examined. This facilitates a detailed exploration of the dynamical and estimation issues involved in analyzing data using simple models. Specific conclusions for each of the studies carried out in this thesis are given at the end of their respective chapters, and will not be repeated here. Instead, the following general remarks are offered:

1. Data analysis techniques based on simple dynamics offer a viable alternative to more generic statistical schemes such as optimal interpolation. Both the tidal analysis of the time-space series of velocity from a ship-ADCP (Chapter 3), as well as the inversion of point observations of density for transport in Chapter 5, are suggestive of the diverse array of data types and models which can be treated with such assimilation methods.
2. Process-oriented approaches that isolate the essential physics are a useful precursor to the development of more complete data assimilation schemes. Their role is to isolate specific dynamic properties and examine these in the context of available observations and estimation schemes. This highlights elements of the problem likely to be important in applications based on more complex models.

This feature was particularly evident in Chapter 5 where a detailed examination of the dynamics governing the density and transport fields provided considerable insight into the joint estimation of these quantities.

3. The identification of inflow boundary conditions using interior data is possible using data assimilation techniques. This is an outstanding issue in regional modelling studies and assimilation methods provide an alternative to the commonly used practice of specifying inflow boundaries based on climatology. Estimation of boundary conditions has been a common theme throughout this work and has been examined in the context of both weak and strong constraint methods and using a variety of dynamics.
4. Straightforward, but effective, data assimilation schemes are needed for practical applications. Optimal schemes are desirable but often prove difficult to implement, particularly in the case of operational systems which require near real-time data acquisition and assimilation. This provided the motivation for the approximate Kalman filter in Chapter 4.

For the remainder of this chapter, we examine the implications of this final point with regard to assimilating data into realistic ocean models using suboptimal schemes.

6.1 Suboptimal Data Assimilation

This thesis, having a data assimilation perspective, argues that analysis of oceanographic data is best organized around a numerical circulation model. It is evident that suitable data (i.e. in terms of its quantity, quality, type, and distribution) provides one important component of any successful estimation scheme. It also seems sensible that a realistic ocean model should be used to assimilate these data for a detailed and accurate picture of circulation to be obtained. For coastal regions a number of models with fairly complete physics are presently available (see Haidvogel

and Beckmann 1996). These models are nonlinear and support a variety of complex dynamics including, for example, baroclinic instabilities.

In principle, it is possible to use any ocean model with the methods which were outlined in Chapter 2. Of these, the extended Kalman filter and adjoint-based smoothing are perhaps the most widely used nonlinear data assimilation techniques. However, a number of outstanding issues arise when using these techniques to assimilate data into complex, nonlinear ocean models. These include:

- **Computational load.** Filtering and smoothing algorithms often have formidable computational requirements. For instance, the extended Kalman filter requires, at every time step, the model to be linearized and matrices of dimension comparable to the ocean state vector to be multiplied and inverted. Similarly, adjoint-based smoothing methods require multiple integrations of the model and its adjoint over the interval of interest. The required number of integrations can vary widely depending on the conditioning of the estimation problem.
- **Optimality.** As demonstrated in Chapter 2 and Appendix A, the optimality properties of the Kalman filter and smoother, such as maximum likelihood, no longer hold for nonlinear models. Nonlinear dynamics generate higher order probability moments which often prove significant for atmospheric and oceanic systems (Miller *et al.* 1994). These moments are not accounted for in the commonly used nonlinear extensions of the filtering and smoothing algorithms.
- **Input Statistics.** For practical application of any assimilation algorithm, the uncertainty associated with the observations and model must be specified. While this is often possible for the observation errors, it is not clear how to obtain the complex error fields associated with realistic ocean models. While this fact is made explicit in the extended Kalman filter it is, in fact, equivalent to specifying the form and strength of the regularization terms used in adjoint based smoothing.
- **Matrix or Operator Representation.** In practice, ocean models are implemented

based on discretized dynamics, while assimilation methods are introduced using operator notation or vector functions. As a result, difficulties in implementing some techniques are often not entirely obvious. For instance, constructing, implementing and verifying the adjoint code associated with the discretized equations of a complex nonlinear ocean model is not a trivial task.

For practical purposes, it is evident that a trade-off exists between the complexity of the model used, and the sophistication of the assimilation scheme. Computational requirements of the extended Kalman filter and, to a lesser extent, adjoint-based smoothing, are prohibitive for forecasting systems based on contemporary ocean models operating in near real-time. Moreover, these schemes cannot be justified on the basis of statistical optimality due to nonlinearities and incomplete prior information. It is clear that the practical difficulties and expertise required to both apply and maintain such schemes may detract somewhat from their theoretical advantages (e.g. Thacker 1992). These considerations argue strongly for robust and efficient suboptimal data assimilation schemes for use with existing nonlinear ocean models.

6.2 Future Directions

Based on the issues outlined in the previous discussion, I will speculate briefly on some possible approaches to assimilating data into nonlinear ocean models. First, since the focus here lies mainly in operational prediction, I restrict myself to considering filtering algorithms based on sequential estimation schemes. Second, let it be assumed that adjoint code, and associated gradient information, is unavailable. Clearly, adjoint code is extremely valuable. It provides the basis for filtering techniques such as optimal nudging (Zou *et al.* 1992, Stauffer and Bao 1993), nonlinear adaptive filters (Hoang *et al.* 1994), and stochastic optimal control (Heemink and Metzelaar 1995). In addition, the sequential variational methods of Pires *et al.* (1996) offers an extension of adjoint-based smoothing to what is, in fact, a filtering problem. Therefore, while this second assumption may appear rather restrictive, it is

justified in many cases due the difficulty of deriving adjoint code from existing ocean models. Below, three approaches are suggested which might hold some promise in the development of efficient, yet effective, operational schemes.

Recently, the suboptimal technique of nudging has received renewed attention (Oschlies and Willebrand 1996, Fischer and Latif 1995, Ezer and Mellor 1994). While this a relatively unsophisticated data assimilation method, it has been reconsidered in light of the difficulties encountered in implementing the more optimal techniques, not a small part of which is their computational burden. The study of Oschlies and Willebrand (1996) is particularly notable in that it recognizes that while the dynamically consistent nudging matrices provided by the (extended) Kalman filter are desirable, they are hard to obtain for realistic ocean models. As an alternative, they construct an approximate nudging matrix which ensures that data are introduced into the model integration in a dynamically balanced manner, taking into account the relative uncertainties. They demonstrate that this nudging procedure eliminates spurious adjustment processes and improves the overall model estimates. I feel that such intelligent nudging approaches, built on a foundation of ocean dynamics, are extremely useful and may allow the present generation of ocean models to be used for data assimilation with a minimum of effort.

A probabilistic framework might also prove to be valuable in developing practical data assimilation schemes. Appendix A demonstrates that integrating an ocean model may be viewed as a transformation on the input probability density function describing the ocean state. Motivated by this fact, Evensen (1994) has proposed a Monte Carlo method to evaluate this transitional density function in the context of the Kalman filter (referred to as the ensemble Kalman filter). This approach allows for error growth associated with the nonlinear dynamics and accounts for the existence of higher order probability moments. It also eliminates the direct calculation of the forecast error variance, the most computationally demanding and difficult step in the extended Kalman filter. Evensen suggests that that this ensemble Kalman filter may be computationally feasible for realistic problems and therefore may prove useful

in nonlinear data assimilation.

Finally, the reduction of complex dynamical systems is considered as a means to efficiently implement data assimilation using existing ocean models. This approach is based on constructing approximate dynamical models with greatly reduced dimension and therefore enhanced computational efficiency (see Chapter 4). These reduced dynamics might allow suboptimal nudging matrices or approximate adjoint operators to be obtained. Hasselmann (1988) presents some intriguing ideas about nonlinear system reduction. The basic idea is that a small set of patterns can often be identified which account for the important components of the spatial and temporal evolution of the system. These notions might be extended to the reduction of a complex, nonlinear ocean model. Based on the observation that the realized degrees of freedom in an ocean model are much smaller than the actual degrees of freedom, such a system reduction might be possible.

To conclude, the increased availability of oceanographic data and advances in numerical ocean modelling have brought us to a stage where a realistic description of the ocean circulation is feasible. The development of data assimilation has also moved beyond the research stage to a point where operational circulation models can now be considered. Without a doubt, a large number of outstanding issues still remain. Many of these are fundamental questions on nonlinear systems theory and statistical estimation, and difficult to address even in simple models. However, given the experience gained thus far, it is felt that a reasonable path is to begin developing workable (suboptimal) data assimilation schemes which can be implemented with contemporary ocean models. It is hoped that the simple, process-oriented models of this thesis provide an initial step in that direction.

Appendix A

Least-Squares Regression

The classic regression equation is given by

$$\mathbf{z} = \mathbf{D}\mathbf{b} + \mathbf{e}$$

where

\mathbf{z} – $n \times 1$ vector of observations

\mathbf{D} – $n \times p$ matrix representing the model

\mathbf{b} – $p \times 1$ vector of unknown parameters

\mathbf{e} – $n \times 1$ vector of errors.

The goal is to choose the unknown \mathbf{b} such that the model is a best fit to the data.

If it is assumed that the error $\mathbf{e} \sim N(0, \sigma^2 \mathbf{I})$, the probability density function of \mathbf{e} is

$$p(\mathbf{e}) = \frac{1}{\sigma^n (2\pi)^{n/2}} \exp \left\{ -\frac{1}{2\sigma^2} \mathbf{e}^T \mathbf{e} \right\}.$$

Now, since \mathbf{e} is determined by the observations (which are known) and the unknown parameters \mathbf{b} , we may view $p(\mathbf{e})$ as a function of \mathbf{b} . The maximum value of $p(\mathbf{e})$ is, in some sense, the most likely value. It is determined by specifying \mathbf{b} such that the quantity in brackets, the squared error, is minimized.

Least squares regression then minimizes the sum of squares of the error

$$J = \mathbf{e}^T \mathbf{e}$$

to yield an estimate for \mathbf{b} in terms of the data \mathbf{z} . Under the distributional assumptions given above, this corresponds to minimizing the error variance. Differentiating J with respect to \mathbf{b} and setting the result to zero yields

$$\hat{\mathbf{b}} = (\mathbf{D}^T \mathbf{D})^{-1} \mathbf{D}^T \mathbf{z}$$

where $\hat{\mathbf{b}}$ is the estimate for \mathbf{b} which minimizes J . An estimate of the observations $\hat{\mathbf{z}}$ is then obtained as

$$\hat{\mathbf{z}} = \mathbf{D} \hat{\mathbf{b}}.$$

Error estimates are also easily obtained as a part of the regression calculation. The error covariances of $\hat{\mathbf{b}}$ and $\hat{\mathbf{z}}$ are (e.g. Sen and Srivastiva 1990)

$$\text{var}(\hat{\mathbf{b}} - \mathbf{b}) = \sigma^2 (\mathbf{D}^T \mathbf{D})^{-1}$$

$$\text{var}(\hat{\mathbf{z}} - \mathbf{z}) = \sigma^2 \mathbf{D} (\mathbf{D}^T \mathbf{D})^{-1} \mathbf{D}^T$$

These quantities can be used to determine confidence intervals for the estimates $\hat{\mathbf{b}}$ and $\hat{\mathbf{z}}$.

Generalized least squares regression assumes $\mathbf{e} \sim N(0, \Sigma)$ which leads to different estimates than those given above. The principles are, however, identical.

Appendix B

Probabilistic Approach

This appendix outlines a probabilistic derivation of the general solution to the filtering and smoothing problems. The approach clearly illustrates how the dynamics and measurement processes alter the probability density function which describes the ocean state. In doing so, it suggests some of the important issues in estimation problems involving nonlinear and non-Gaussian systems which are not entirely obvious from an optimization perspective.

B.1 Filtering

A complete description of the state \mathbf{x}_N at time $t = N$ based on observations up to and including time N is contained in the conditional probability density function (pdf)

$$p(\mathbf{x}_N|\mathbf{Z}_N) \tag{B.1}$$

where $\mathbf{Z}_N = \{\mathbf{z}_0, \dots, \mathbf{z}_N\}$ represents the set of observations in the interval $(0, N)$. Suppose \mathbf{Z}_N and $p(\mathbf{x}_N|\mathbf{Z}_N)$ are known. The model dynamics (2.13) allow this pdf to be stepped forward in time. This model forecast is referred to as the *prediction process*. If observations then become available at the forecast time, a *measurement process* combines this additional information with the estimate from the prediction

process. The measurement and prediction processes together yield the general filtering solution.

B.1.1 The Prediction Process

The prediction process determines the conditional probability density function $p(\mathbf{x}_{N+1}|\mathbf{Z}_N)$. Expressing $p(\mathbf{x}_{N+1}|\mathbf{Z}_N)$ in terms of the marginal density function we have

$$\begin{aligned} p(\mathbf{x}_{N+1}|\mathbf{Z}_N) &= \int p(\mathbf{x}_{N+1}|\mathbf{x}_N, \mathbf{Z}_N)p(\mathbf{x}_N|\mathbf{Z}_N)d\mathbf{x}_N \\ &= \int p(\mathbf{x}_{N+1}|\mathbf{x}_N)p(\mathbf{x}_N|\mathbf{Z}_N)d\mathbf{x}_N \end{aligned} \quad (\text{B.2})$$

where the Markov property of the dynamics has been used in the second equality. The prediction equation (B.2) gives the updated density function $p(\mathbf{x}_{N+1}|\mathbf{Z}_N)$ in terms of a transition density $p(\mathbf{x}_{N+1}|\mathbf{x}_N)$ and the known prior density function $p(\mathbf{x}_N|\mathbf{Z}_N)$. The prediction process given here yields the best estimate of the system state before measurement.

B.1.2 The Measurement Process

Suppose that a new observation \mathbf{z}_{N+1} becomes available. This additional information can be used to produce an updated estimate of the state. This new estimate is summarized by the conditional density function $p(\mathbf{x}_{N+1}|\mathbf{z}_{N+1}, \mathbf{Z}_N)$. For notational convenience let

$$p(\mathbf{x}_{N+1}|\mathbf{z}_{N+1}, \mathbf{Z}_N) = p(\mathbf{x}_{N+1}|\mathbf{Z}_{N+1}).$$

Now, using Bayes' theorem

$$p(\mathbf{x}_{N+1}|\mathbf{Z}_{N+1}) = \frac{p(\mathbf{z}_{N+1}|\mathbf{x}_{N+1}, \mathbf{Z}_N)p(\mathbf{x}_{N+1}|\mathbf{Z}_N)}{p(\mathbf{z}_{N+1}|\mathbf{Z}_N)} \quad (\text{B.3})$$

$$= \frac{p(\mathbf{z}_{N+1}|\mathbf{x}_{N+1})p(\mathbf{x}_{N+1}|\mathbf{Z}_N)}{p(\mathbf{z}_{N+1}|\mathbf{Z}_N)}. \quad (\text{B.4})$$

Examination of this result shows that the desired quantity $p(\mathbf{x}_{N+1}|\mathbf{Z}_{N+1})$ can be expressed in terms of the pdf of the prediction process $p(\mathbf{x}_{N+1}|\mathbf{Z}_N)$ multiplied by a

transfer function which depends on the properties of the observations. The measurement process gives the best estimate of the system state after measurement.

B.1.3 The General Filtering Solution

The general solution of the filtering problem is obtained using the results from the prediction and measurement processes. Substituting for $p(\mathbf{x}_{N+1}|\mathbf{Z}_{N+1})$ in (B.4), using (B.2), gives

$$\begin{aligned} p(\mathbf{x}_{N+1}|\mathbf{Z}_{N+1}) &= \frac{p(\mathbf{z}_{N+1}|\mathbf{x}_{N+1}) \int p(\mathbf{x}_{N+1}|\mathbf{x}_N) p(\mathbf{x}_N|\mathbf{Z}_N) d\mathbf{x}_N}{p(\mathbf{z}_{N+1}|\mathbf{Z}_N)} \\ &= \frac{\int p(\mathbf{z}_{N+1}|\mathbf{x}_{N+1}) p(\mathbf{x}_{N+1}|\mathbf{x}_N) p(\mathbf{x}_N|\mathbf{Z}_N) d\mathbf{x}_N}{p(\mathbf{z}_{N+1}|\mathbf{Z}_N)}. \end{aligned}$$

The denominator in this equation can be rewritten as

$$\begin{aligned} p(\mathbf{z}_{N+1}|\mathbf{Z}_N) &= \int p(\mathbf{z}_{N+1}|\mathbf{x}_{N+1}, \mathbf{Z}_N) p(\mathbf{x}_{N+1}|\mathbf{Z}_N) d\mathbf{x}_{N+1} \\ &= \int \int p(\mathbf{z}_{N+1}|\mathbf{x}_{N+1}) p(\mathbf{x}_{N+1}|\mathbf{x}_N, \mathbf{Z}_N) p(\mathbf{x}_N|\mathbf{Z}_N) d\mathbf{x}_N d\mathbf{x}_{N+1} \\ &= \int \int p(\mathbf{z}_{N+1}|\mathbf{x}_{N+1}) p(\mathbf{x}_{N+1}|\mathbf{x}_N) p(\mathbf{x}_N|\mathbf{Z}_N) d\mathbf{x}_N d\mathbf{x}_{N+1} \end{aligned}$$

and the equation becomes

$$p(\mathbf{x}_{N+1}|\mathbf{Z}_{N+1}) = \frac{\int p(\mathbf{z}_{N+1}|\mathbf{x}_{N+1}) p(\mathbf{x}_{N+1}|\mathbf{x}_N) p(\mathbf{x}_N|\mathbf{Z}_N) d\mathbf{x}_N}{\int \int p(\mathbf{z}_{N+1}|\mathbf{x}_{N+1}) p(\mathbf{x}_{N+1}|\mathbf{x}_N) p(\mathbf{x}_N|\mathbf{Z}_N) d\mathbf{x}_N d\mathbf{x}_{N+1}}. \quad (\text{B.5})$$

The equation (B.5) provides a general solution to the filtering problem. It specifies a recursive relation which governs the temporal evolution of the conditional probability density function describing the state for a Markov process with measurement. This updating involves the transitional pdf $p(\mathbf{x}_{N+1}|\mathbf{x}_N)$, which is based on the dynamics, as well as information on the pdf of the observations. The general solution to the filtering problem can be difficult to solve (Kitagawa 1987). For linear models with Gaussian noise processes, the mean and covariance of the governing pdfs are adequate to completely specify the filtering problem, which reduces to the Kalman filter of Section 2.3.1

B.2 Smoothing

The smoothing problem is governed by the conditional pdf $p(\mathbf{x}_0, \dots, \mathbf{x}_N | \mathbf{Z}_N)$. Using Bayes' theorem, this conditional density function can be represented,

$$p(\mathbf{x}_0, \dots, \mathbf{x}_N | \mathbf{Z}_N) = \frac{p(\mathbf{Z}_N | \mathbf{x}_0, \dots, \mathbf{x}_N) p(\mathbf{x}_0, \dots, \mathbf{x}_N)}{p(\mathbf{Z}_N)}. \quad (\text{B.6})$$

For an interpretation of (B.6) in terms of the model (2.13) and measurement (2.14) equations, note the following:

1. In $p(\mathbf{Z}_N | \mathbf{x}_0, \dots, \mathbf{x}_N)$, the sequence $\{\mathbf{x}_0, \dots, \mathbf{x}_N\}$ is given. A measurement relation of the form (2.14) then implies that the \mathbf{z}_t are independent if \mathbf{e}_t^o is a white noise process. This gives

$$\begin{aligned} p(\mathbf{Z}_N | \mathbf{x}_0, \dots, \mathbf{x}_N) &= p(\mathbf{e}_1^o) p(\mathbf{e}_2^o) \dots p(\mathbf{e}_N^o) \\ &= \prod_{t=1}^N p(\mathbf{e}_t^o) \end{aligned}$$

2. Expanding $p(\mathbf{x}_0, \dots, \mathbf{x}_N)$ and applying the Markov property of the model (2.13) yields

$$\begin{aligned} p(\mathbf{x}_0, \dots, \mathbf{x}_N) &= p(\mathbf{x}_N | \mathbf{x}_{N-1}, \dots, \mathbf{x}_0) p(\mathbf{x}_{N-1}, \dots, \mathbf{x}_0) \\ &= p(\mathbf{x}_N | \mathbf{x}_{N-1}) p(\mathbf{x}_{N-1}, \dots, \mathbf{x}_0) \\ &= p(\mathbf{x}_N | \mathbf{x}_{N-1}) P(\mathbf{x}_{N-1} | \mathbf{x}_{N-2}, \dots, \mathbf{x}_0) p(\mathbf{x}_{N-2}, \dots, \mathbf{x}_0) \\ &= p(\mathbf{x}_N | \mathbf{x}_{N-1}) P(\mathbf{x}_{N-1} | \mathbf{x}_{N-2}) \dots p(\mathbf{x}_1 | \mathbf{x}_0) p(\mathbf{x}_0) \\ &= p(\mathbf{x}_0) \prod_{t=1}^N p(\mathbf{x}_t | \mathbf{x}_{t-1}). \end{aligned}$$

The final expression in this equation shows that the joint density function is the product of the initial density function and some transition density function. Note that the transition density function $p(\mathbf{x}_N | \mathbf{x}_{N-1})$ is equivalent to the system

noise term $\mathbf{G}\mathbf{e}_{N-1}^m$ of (2.13) in view of the fact that \mathbf{x}_{N-1} is given and \mathbf{e}_t^m is a white noise process.

Using the results above, eqn. (B.6) can be rewritten as

$$p(\mathbf{x}_0, \dots, \mathbf{x}_N | \mathbf{Z}_N) = \frac{p(\mathbf{x}_0) \prod_{i=1}^N p(\mathbf{G}\mathbf{e}_i^m) \prod_{i=1}^N p(\mathbf{e}_i^o)}{p(\mathbf{Z}_N)} \quad (\text{B.7})$$

This gives the probability density of the complete state $\{\mathbf{x}_0, \dots, \mathbf{x}_N\}$ conditioned on the complete set of observations $\{\mathbf{Z}_N\}$ as a function of the dynamic and measurement equations. It illustrates that $P(\mathbf{x}_0, \dots, \mathbf{x}_N | \mathbf{Z}_N)$ involves information on the initial conditions and the probability distributions of the model and observation errors. A linear, Gaussian system reduces the maximum likelihood estimate associated with (B.7) to the minimization problem (2.29) which governs the Kalman smoother. Clearly, both nonlinear and non-Gaussian systems lead to complex expressions for the pdf governing any filtering or smoothing problem.

References

- Aikman, F., G.L. Mellor, T. Ezer, D. Sheinin, P. Chen, L. Breaker, K. Bosley and D.B. Rao. 1996. Towards an operational nowcast/forecast system for the U.S. East Coast. (submitted manuscript).
- Anderson, B.D.O. and J.B. Moore. 1979. *Optimal Filtering*. Prentice-Hall. 357pp.
- Backus, G.E. and J.F. Gilbert. 1967. Numerical applications of a formalism for geophysical inverse problems. *Geophys. J. R. Astron. Soc.* 13:247-276.
- Backus, G.E. and J.F. Gilbert. 1968. The resolving power of gross earth data. *Geophys. J. R. Astron. Soc.* 16:169-205.
- Barnett, S. 1990. *Matrices: Methods and Applications*. Clarendon Press, Oxford, 450pp.
- Bennett, A.F. 1992. *Inverse Methods in Physical Oceanography*. Cambridge University Press, New York, 346pp.
- Bennett, A.F. and W.P. Budgell. 1989. The Kalman smoother for a linear quasi-geostrophic model of ocean circulation. *Dyn. Atmos. Ocean.* 13(3-4):219-268.
- Bennett, A.F. and W.P. Budgell. 1987. Ocean data assimilation and the Kalman Filter: spatial regularity. *J. Phys. Oceanogr.* 17(10):1583-1601.
- Bennett, A.F. and P.C. McIntosh. 1982. Open ocean modelling as an inverse problem: Tidal theory. *J. Phys. Ocean.* 12:1004-1018

- Bertsekas, D.P. 1982. *Constrained Optimization and Lagrange Multiplier Methods*. New York: Academic Press.
- Bogden, P.S., R.E. Davis and R. Salmon. 1993. The North Atlantic circulation: combining simplified dynamics with hydrographic data. *J. Mar. Res.* 51:1-52.
- Brillinger, D.R. 1981. *Time Series. Data Analysis and Theory*. Holden-Day, San Francisco, 540pp.
- Bryden, H.L. 1980. Geostrophic vorticity balance in midocean. *J. Geophys. Res.* 85:2825-2828.
- Bryson, A.E. and Y. Ho. 1969. *Applied Optimal Control*. Blaisdell, Waltham, Mass., 481pp.
- Candela, J., R.C. Beardsley and R. Limeburner. 1992. Separation of tidal and subtidal currents in ship mounted acoustic Doppler current profiler observations. *J. Geophys. Res.* 97:769-788.
- Craven, P. and G. Wahba. 1979. Smoothing noisy data with spline functions: Estimating the correct degree of smoothing by the method of generalized cross validation. *Numer. Math.* 31:377-403.
- Csanady, G.T. 1979. The pressure field along the western margin of the North Atlantic. *J. Geophys. Res.* 84(C8):4905-4915.
- Das, S.K. and R.W. Lardner. 1990. On the estimation of parameters of hydraulic models by assimilation of periodic tidal data. *J. Geophys. Res.* 96:15187-15196.
- Daley, R. 1992. The effect of serially correlated observation and model error on atmospheric data assimilation. *Monthly Weather Review.* 120:165-177.
- Daley, R. 1991. *Atmospheric Data Analysis*. New York: Cambridge University Press. 457pp.

- Dee, D.P. 1991. Simplification of the Kalman filter for meteorological data assimilation. *Q. J. R. Meteorol. Soc.* 117:365-384.
- Defant, A. 1961. *Physical Oceanography*, Vol 1., New York: Pergamon. 729pp.
- deYoung, B., R.J. Greatbatch and K.B. Forward. 1992. A diagnostic coastal circulation model with application to Conception Bay, Newfoundland. *J. Phys. Oceanogr.* 23:2617-2635.
- Diderrich, G.T. 1985. The Kalman filter from the perspective of Goldberger-Theil estimators. *The American Statistician.* 39(3):193-198.
- Duncan, D.B. and Horn, S.D. 1972. Linear dynamic recursive estimation from the viewpoint of regression analysis. *Journal of the American Statistical Society.* 67:815-821.
- Evensen, G. 1994. Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte-Carlo methods to forecast error statistics. *J. Geophys. Res.* 99:10143-10162.
- Ezer, T. and G.L. Mellor. 1994. Diagnostic and prognostic calculations of the North Atlantic circulation and sea level using a sigma coordinate ocean model. *J. Geophys. Res.* 99:14159-14171.
- Foreman, M.G.G. and H.J. Freeland. 1991. A comparison of techniques for tide removal from ship mounted acoustic Doppler measurements along the southwest coast of Vancouver Island. *J. Geophys. Res.* 96:17007-17021.
- Fukumori, I., and P. Malanotte-Rizzoli. 1995. An approximate Kalman filter for ocean data assimilation: An example with an idealized Gulf Stream model. *J. Geophys. Res.* 100:6777-6793.
- Fukumori, I., J. Benveniste, C. Wunsch, D.B. Haidvogel. 1993. Assimilation of sea surface topography into an ocean circulation model using a steady-state smoother. *J. Phys. Oceanogr.* 1831-1855.

- Gallagher, F., H. von Storch, R. Schnur, and G. Hannoschock. 1991. The Pop Manual. Deutsches KlimaRechenZentrum. Technical report no. 1.
- Gelb, A. 1974. *Applied Optimal Estimation*. Cambridge: MIT Press, 374 pp.
- Geyer, W.R. and R. Signell. 1990. Measurements of tidal flow around a headland with a shipboard acoustic Doppler current profiler. *J. Geophys. Res.* 95:3189-3197.
- Ghil, M. and P. Malanotte-Rizzoli. 1991. Data assimilation in meteorology and oceanography. *Adv. Geophys.* 33:141-266.
- Gill, A.E. 1982. *Atmosphere-Ocean Dynamics*. Academic Press, San Diego, 662pp.
- Gill P.E., W. Murray, and M.H. Wright. 1981. *Practical Optimization*. London: Academic Press. 401pp.
- Godin, G. 1972. *The Analysis of Tides*. London. University of Liverpool Press.
- Golub, G., M. Heath and G. Wahba. 1979. Generalized cross validation as a method for choosing a good ridge parameter. *Technometrics*. 21:215-223.
- Gregory, D.N. 1988. Tidal Current Variability on the Scotian Shelf and Slope. *Can. Tech. Rep. Hydrogr. Ocean Sci. No. 109*.
- Griffin, D.A. and K.R. Thompson. 1996. The adjoint method of data assimilation used operationally for shelf circulation. *J. Geophys. Res.* 101:3457-3478.
- Griffin, D. and S. Lochmann. 1992. Petrel V Cruise 22 to Western Bank. OPEN Report 1992/6. Department of Oceanography. Dalhousie University.
- Haidvogel, D.B. and Beckmann. 1996. Numerical models of the coastal ocean. submitted manuscript.

- Hasselmann, K. 1988. PIPs and POPs: The reduction of complex dynamic systems using principal interaction and oscillation patterns. *J. Geophys. Res.* 93:11015-11021.
- Heaps, N.S. 1969. A two dimensional numerical sea model. *Phil Trans. Roy. Soc. Lond.* 265:93-137.
- Heemik, A.W. 1986. Storm surge prediction using Kalman filtering. Ph.D. Thesis. Twente University of Technology. The Hague.
- Heemink, A.W. and I.D.M. Metzelaar. 1995. Data assimilation into a numerical shallow water flow model: a stochastic optimal control approach. *J. Marine Systems.* 6:145-158.
- Heemink, A. and T. Van-Stijn. 1993. Optimization of shallow sea models. *Cray Channels.* 15:20-22.
- Heemik, A.W. and H. Kloosterhuis. 1990. Data assimilation for non-linear tidal models. *Int. J. Numer. Meth. Fluids* 11:1097-1112.
- Hendershott, M.C. and P. Rizzoli. 1976. The winter circulation of the Adriatic Sea. *Deep Sea Research.* 23:353-370.
- Hoang, S., P. DeMey and O. Talagrand. 1994. A simple adaptive algorithm of the stochastic approximation type for system parameter and state estimation. *Proceeding of the 33rd Conference on Decision and Control.* Buena Vista, Florida, USA. December.
- Holland, W.R. and A.D. Hirschman. 1972. A numerical calculation of the circulation in the North Atlantic Ocean. *J. Phys. Oceanogr.* 2:336-354.
- Howarth, M.J. and R. Proctor. 1992. Ship ADCP measurements and tidal models of the North Sea. *Cont. Shelf Res.* 12:601:623.

- Hsueh, Y. and Peng, C.Y. 1978. A diagnostic model of continental shelf circulation. *J. Geophys. Res.* 83(C6):3033-3041.
- Huthnance, J.M. 1984. Slope currents and "JEBAR". *J. Phys. Oceanogr.* 14:795-810.
- Jackson, D.D. 1972. Interpretation of inaccurate, insufficient and inconsistent data. *Geophys. J. R. Astr. Soc.* 28:97-109.
- Jazwinski, A.H. 1970. *Stochastic processes and filtering theory*. New York: Academic Press.
- Jiang, S. and M. Ghil. 1993. Dynamical properties of error statistics in a shallow water model. *J. Phys. Oceanogr.* 23:2541-2566.
- Johannessen, J.A., L.P. Roed, O.M. Johannessen, G. Evensen, B. Hackett, L.H. Petterson, P.M. Haugan, S. Sandven, and R. Shuchman. 1993. Monitoring and modeling of the marine coastal environment. *Photogrammetric Eng. and Remote Sensing.* 59:351-361.
- Kalman, R.E. 1960. A new approach to linear filtering and prediction problems. *Journal of Basic Engineering.* 35-45.
- Kitagawa, G. 1987. Non-Gaussian state-space modeling of nonstationary time series. *Journal of the American Statistical Association.* 82:1032-1041.
- Lanczos, C. 1961. *Linear Differential Operators*. London: Van Nostrand.
- Lardner, R.W. 1993. Optimal control of open boundary conditions for a numerical tidal model. *Computer Methods in Applied Mechanics and Engineering.* 102:367-387.
- Lardner, R.K., A.H. Al-Rabeh and N. Gunay. 1993. Optimal estimation of parameters for a two-dimensional hydrodynamical model of the Arabian Gulf. *J. Geophys. Res.* 98:18229-18242.

- Latif, M. and M. Flugel. 1991. An investigation of short-range climate predictability in the tropical Pacific. *J. Geophys. Res.* 96:2661-2643.
- LeDimet, F-X. and O. Talagrand. 1986. Variational algorithms for analysis and assimilation of meteorological observations: theoretical aspects. *Tellus.* 38A:97-110.
- Levitus, S. 1982. Climatological Atlas of the World Ocean, *NOAA Tech. Pap.*, 3, 173pp.
- Lively, R.R. 1988. Current meter, meteorological, sea-level and hydrographic observations for the CASP experiment, off the coast of Nova Scotia, November 1985 to April 1986. *Can. Tech. Rep. Hydrogr. Ocean Sci.* No. 100: vii + 428 p.
- Lynch, D.R., F.E. Werner, D.A. Greenberg, and J.W. Loder. 1992. Diagnostic model for baroclinic, wind-driven and tidal circulation in shallow seas. *Cont. Shelf Res.* 12:37-64.
- Madsen, N.K. and R.F. Sinovec. 1974. The numerical method of lines for the solution of nonlinear partial differential equations. In: *Computational Methods in Nonlinear Mechanics*. J.T. Oden et al. (Eds.), Texas Institute for Computational Mechanics, Austin, Tex., pp. 371-380.
- Marmorino, G.O. and C.L. Trump. 1992. Acoustic Doppler current profiler measurements of possible lee waves south of Key West, Florida. *J. Geophys. Res.* 97:7271-7275.
- Martel, F. and C. Wunsch. 1993. The north Atlantic circulation in the early 1980's - an estimate from inversion of a finite difference model. *J. Phys. Oceanogr.* 23:898-924.
- McIntosh, P.C. and A.F. Bennett. 1984. Open ocean modeling as an inverse problem: M_2 tides in Bass Strait. *J. Phys. Oceanogr.* 14(3): 601-614.

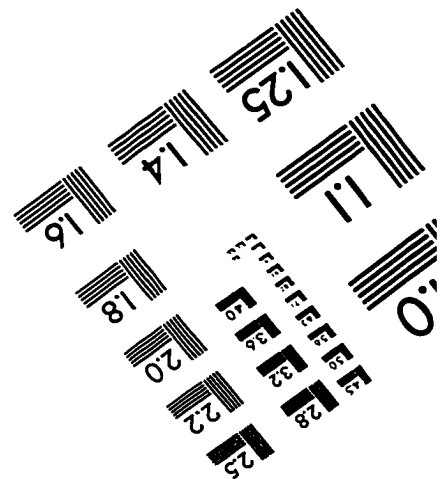
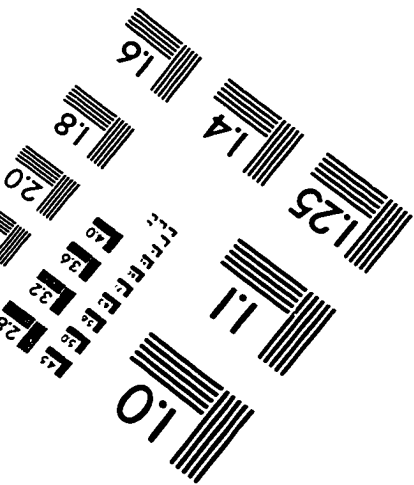
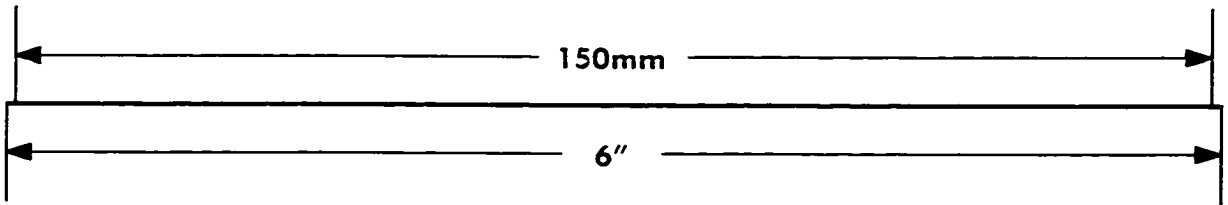
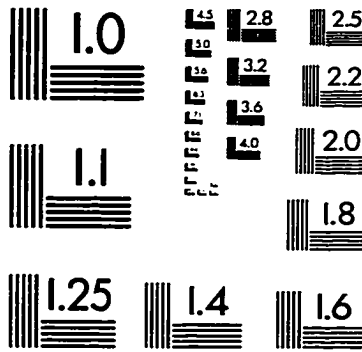
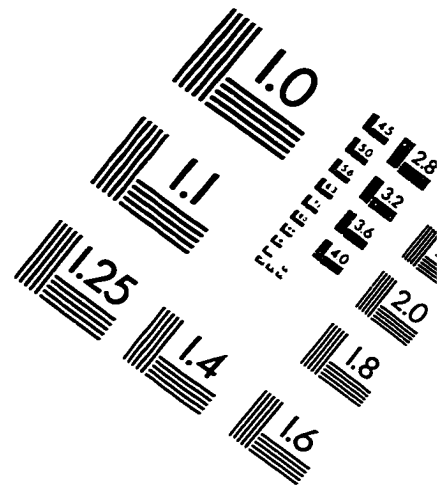
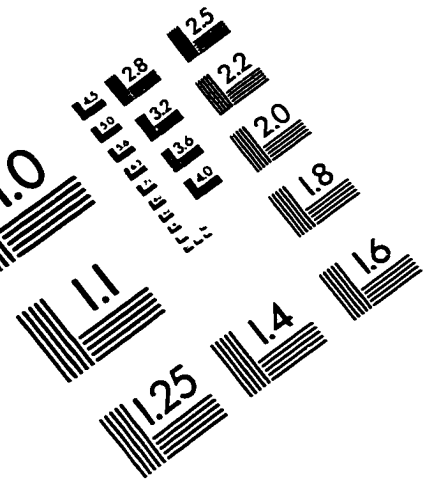
- McIntosh, P.C. and G. Veronis. 1993. Solving underdetermined tracer inverse problems by spatial smoothing and cross validation. *J. Phys. Oceanogr.* 23:716-730.
- Melgaard, D.K. and R.F. Sinovec. 1981. General software for two-dimensional nonlinear partial differential equations. *ACM Transactions in Mathematical Software.* 7(1):106-125.
- Mellor, G.L., C.R. Mechoso, and E. Keto. 1982. A diagnostic calculation of the general circulation of the Atlantic Ocean. *Deep Sea Research.* 20:1171-1192.
- Mertz, G. and D.G. Wright. 1992. Interpretations of the JEBAR term. *J. Phys. Oceanogr.* 22:301-313.
- Miller, R.N., M. Ghil and F. Gauthieuz. 1994. Advanced data assimilation in strongly nonlinear dynamical systems. *J. Atmos. Sci.* 51:1037-1056.
- Morrison, D.F. 1967. *Multivariate Statistical Methods.* McGraw-Hill, New York, 338 pp.
- Navon, I.M. and D.M. Legler, 1987: Conjugate-gradient methods for large-scale minimization in meteorology. *Monthly Weather Review.*, 115:1479-1502.
- Noble, B. and J.W. Daniel. 1977. *Applied Linear Algebra.* Prentice-Hall, Toronto, 477pp.
- Olbers, D.J., M. Wenzel and J. Willerbrand. 1985. The inference of North Atlantic circulation patterns from climatological hydrographic data. *Rev. Geophys.* 23:313-356.
- Oschlies, A., and J. Willebrand. 1996. Assimilation of Geosat altimeter data into an eddy resolving primitive equation model of the North Atlantic Ocean. *J. Geophys. Res.* 101:14175-14190.
- Pedlosky, J. 1979. *Geophysical Fluid Dynamics.* New York: Springer-Verlag.

- Peterson, D.P. 1968. On the concept and implementation of sequential analysis for linear random fields. *Tellus*. 673-686.
- Pires, C., R. Vautard, and O. Talagrand. 1996. On extending the limits of variational assimilation in nonlinear chaotic systems. *Tellus*. 48A:96-121.
- Pollard, R.T. and R.C. Millard. 1970. Comparison between observed and simulated wind-generated inertial oscillations. *Deep Sea Research*. 17:813-821.
- Press, W.H., B.P. Flannery, S.A. Teukolsky, and W.T. Vetterling. 1989. *Numerical Recipes*. New York: Cambridge University Press.
- Rattray M. 1982. A simple exact treatment of baroclinicity-bathymetry interaction in a frictional, iterative, diagnostic ocean model. *J. Phys. Oceanogr.* 12:997-1003
- Salmon, R. 1994. Generalized two-layer models of ocean circulation. *J. Marine Res.* 52:865-908.
- Sanderson, B.G. 1995. Structure of an eddy measured with drifters. *J. Geophys. Res.* 100:6761-6776.
- Sarkisyan, A.S. and V.F. Ivanov. 1971. The joint effect of baroclinicity and bottom relief as an important factor in the dynamics of ocean currents. *Izv. Acad. Sci. USSR, Atmos. Oceanic. Phys., Engl. Trans.* 7:173-188.
- Sarmiento, J.L. and K. Bryan. 1982. An ocean transport model for the North Atlantic. *J. Geophys. Res.* 87:394-408.
- Sasaki, Y. 1970. Some basis formalisms in numerical variational analysis. *Monthly Weather Review*. 98:875-883.
- Schlitzer, R. 1993. Determining the mean, large-scale circulation of the Atlantic with the adjoint method. *J. Phys. Oceanogr.* 23:1935-1952.

- Sen, A. and M. Srivastava. 1990. *Regression analysis: Theory, methods, and applications*. Springer-Verlag, New York, 347 pp.
- Sheng, J. and K.R. Thompson. A robust method for diagnosing regional shelf circulation from scattered profiles. *J. Geophys. Res.* In press.
- Simpson, J.H., E.G. Mitchelson-Jacob, A.E. Hill. 1990. Flow structure in a channel from an acoustic Doppler current profiler. *Continental Shelf Research*. 10:589-603.
- Smith, P.C. and F.B. Schwing. 1991. Mean circulation and variability on the eastern Canadian continental shelf. *Cont. Shelf Res.* 11:977-1012.
- Stauffer, D.R. and J. Bao. 1993. Optimal determination of nudging coefficients using the adjoint equations. *Tellus*. 45A:358-369.
- Stommel, H and F. Schott. 1977. The beta spiral and the determination of the absolute velocity field from hydrographic station data. *Deep Sea Research*. 24:325-329.
- Stroch, H. v., T. Bruns, I. Fisher-Bruns and K.H. Hasselmann. 1988. Principal oscillation pattern analysis of the 30 to 60 day oscillation in a GCM. *J. Geophys. Res.* 93:11022-11036.
- Tarantola, A. 1987. *Inverse Problem Theory: Methods for Data Fitting and Model Parameter Estimation*. Elsevier, New York, 613 pp.
- Thacker, W.C. 1992. Oceanographic inverse problems. *Physica D*. 60:16-37.
- Thacker, W.C. 1988. Fitting models to inadequate data by enforcing spatial and temporal smoothness. *J. Geophys. Res.* 93(C9):10655-10665.
- Thacker, W.C. and R.B. Long. 1988. Fitting dynamics to data. *J. Geophys. Res.* 93(C2):1227-1240.

- Thompson, K.R. and J. Sheng. 1996. Assessing the predictive skill of a 3D barotropic model of subtidal current variability on the Scotian Shelf. *J. Geophys. Res.* (in press).
- Tziperman, E., W.C. Thacker, R.B. Long and S-H Hwang. 1992. Oceanic data analysis using a general circulation model. Part 1: Simulations. *J. Phys. Oceanogr.* 22:1434-1457.
- Wiggins, R.A. 1972. The general linear inverse problem: Implications for surface waves and free oscillations for the earth's structure. *Rev. Geophys. Space Phys.* 10:251-285.
- Wunsch, C. 1978. The North Atlantic general circulation west of 50°W determined by inverse methods. *Rev. Geophys. Space Phys.* 16(4):583-620.
- Wunsch, C. 1994. Dynamically consistent hydrography and absolute velocity in the eastern North Atlantic Ocean. *J. Geophys. Res.* 99:14071-14090.
- Wust, G., 1935. Schichtung and Zirkulation des Atlantischen Ozeans: Das Bodenwasser und die Stratosphäre. *Wiss. Ergeb. Dtsch. Atl. Exped. Meteor 1925-1927*, 6, 288pp.
- Zou, X, I.M. Navon, and F.X. LeDimet. 1992. An optimal nudging data assimilation scheme using parameter estimation. *Q. J. R. Meteor. Soc.* 118:1163-1186.

IMAGE EVALUATION TEST TARGET (QA-3)



APPLIED IMAGE, Inc
1653 East Main Street
Rochester, NY 14609 USA
Phone: 716/482-0300
Fax: 716/288-5989

© 1993, Applied Image, Inc., All Rights Reserved