

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.

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.

Bell & Howell Information and Learning
300 North Zeeb Road, Ann Arbor, MI 48106-1346 USA
800-521-0600

UMI[®]

**LAGRANGIAN MEASUREMENTS AND
LOW-DIMENSIONAL MODELS FOR OCEANOGRAPHIC
AND ATMOSPHERIC DATA ASSIMILATION**

**By
Mark Buehner**

**SUBMITTED IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY
AT
DALHOUSIE UNIVERSITY
HALIFAX, NOVA SCOTIA
MARCH, 2000**

© Copyright by Mark Buehner, 2000



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.

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 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.

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.

0-612-57360-5

Canada

DALHOUSIE UNIVERSITY

FACULTY OF GRADUATE STUDIES

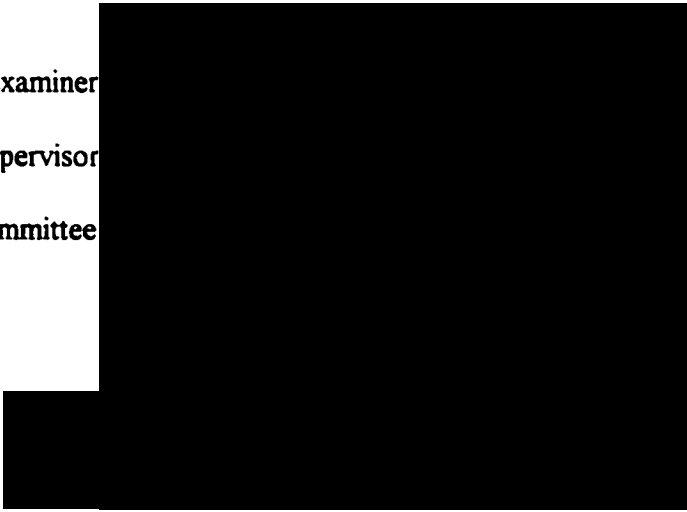
The undersigned hereby certify that they have read and recommend to the Faculty of Graduate Studies for acceptance a thesis entitled "Lagrangian Measurements and Low-Dimensional Models for Oceanographic and Atmospheric Data Assimilation"

by Mark Buehner

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

Dated: March 13, 2000

External Examiner
Research Supervisor
Examining Committee



DALHOUSIE UNIVERSITY

Date: March, 2000

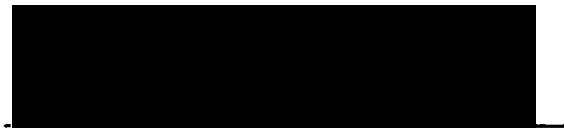
Author: Mark Buehner

Title: Lagrangian Measurements and Low-Dimensional
Models for Oceanographic and Atmospheric Data
Assimilation

Department: Oceanography

Degree: Ph.D. Convocation: May Year: 2000

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	ix
List of Figures	x
Abstract	xiii
List of Symbols	xiv
Acknowledgements	xxi
1 Introduction	1
1.1 Maximum Likelihood Estimation	5
1.2 Linear Regression as a Framework for Data Assimilation	6
1.3 Assimilation Methods for Time-Dependent Problems	10
1.3.1 Kalman Filter and Smoother	11
1.3.2 Variational Method of Data Assimilation	14
1.4 Low-Dimensional Representation of the Model State and Dynamics .	16
1.5 Outline of Thesis	17
2 A Low-Dimensional Ocean Model	20
2.1 Dynamical Theory	21
2.2 Scaling Analysis and Model Description	24
2.3 Rectification Mechanism	30

2.4	Resonance Shifting	32
2.5	Snapshots of Vorticity	32
2.6	Typical Drifter Trajectories and Pseudo-SST Images	34
2.7	Conclusions	38
3	Assimilation of Ocean Drifter Trajectories	41
3.1	Introduction	41
3.2	Ocean Drifter Data	42
3.3	Model for Drifter Observations	44
3.3.1	Trajectory Models	45
3.3.2	Error in the Model Trajectory	46
3.3.3	Error in the Observations	50
3.3.4	Model for the Observed Trajectory	50
3.4	Estimating the Model Parameters	54
3.4.1	Estimation Problem	54
3.4.2	Calculating the Optimal Estimate	55
3.4.3	Uncertainty in the Estimate	57
3.5	Practical Issues for Assimilating Trajectories	57
3.5.1	Experimental Setup	58
3.5.2	Nonlinearity of Ocean/Drifter Model	60
3.5.3	Assimilation of Sub-trajectories	63
3.5.4	Mis-specification of Observation Error Statistics	69
3.5.5	Mis-specification of Model Error Statistics	70
3.5.6	Diagnosing Mis-specified Error Statistics	73
3.6	Discussion and Conclusions	75
4	Assimilation of Sequential Satellite Images	82
4.1	Introduction	82
4.2	Existing Methods	83
4.2.1	Area Correlation Methods	84

4.2.2	Feature Matching Methods	85
4.2.3	Data Insertion	86
4.2.4	Inverse Methods	86
4.3	The Proposed Method	87
4.3.1	Stochastic Advection Model	88
4.3.2	Cost Function	90
4.4	Application to Simulated SST Images	93
4.5	Application to AVHRR Ice Images	95
4.5.1	Ice Advection Model	98
4.5.2	Results	100
4.5.3	Interpretation of Estimated Current Field	104
4.5.4	Sensitivity Studies	108
4.6	Discussion and Conclusions	110
5	Estimation of 3D-Var Background Error Covariances using Empirical Orthogonal Functions	113
5.1	Introduction	113
5.2	Overview of 3D-Var	115
5.2.1	Incremental Formulation	116
5.2.2	Dynamical Importance of the Background Error Covariances .	117
5.2.3	Sources of Information on Background Error	119
5.2.4	Background Error Covariances with Homogeneous and Isotropic Correlations	121
5.3	Representing the Background Error Covariances with EOFs	128
5.3.1	Blending EOFs with \mathbf{B}_{hi}	129
5.3.2	Horizontal Localisation	131
5.4	Results Using EOFs for Stationary Correlations	138
5.4.1	Details of Implementation	138
5.4.2	The Structure Functions	139
5.5	Discussion and Conclusions	150

5.5.1	Limitations	150
5.5.2	EOFs in the Ensemble KF and Integration with 4D-Var	151
6	A Sub-Optimal Assimilation Scheme for Nonlinear Models	153
6.1	Introduction	153
6.2	Sub-Optimal Assimilation Schemes	155
6.3	Approximate Adjoint Model	158
6.3.1	Numerical Linearisation of the Ocean Model	159
6.3.2	Reduced Dimension Subspace	160
6.3.3	Reduced Dimension Adjoint Model	162
6.4	Identical Twin Experiment	164
6.4.1	Description of the Experiment	164
6.4.2	Results	172
6.5	Discussion and Conclusions	173
7	Concluding Remarks	178
7.1	Summary of Results	178
7.2	Error Statistics	181
7.3	Sub-optimal Schemes	182
7.4	Operational Ocean Prediction	183
A	Kalman Filter Algorithm	185
B	Adjoint Method	187
C	Low-Dimensional Model Parameters	189
D	AR(1) Correlated u^s	192
E	Adjoint Model for Image Advection	194
F	Calculating EOFs by Singular Value Decomposition	198

G Localisation Using an Iterative Eigendecomposition Algorithm	201
Bibliography	203

List of Tables

2.1	Scales used to derived the simplified model	25
2.2	Relative scales of terms in the nonlinear barotropic vorticity equation	27
2.3	Units for prognostic variables and structure functions	28
2.4	Specified and derived model parameters	30
3.1	Estimation error when using a full trajectory or sub-trajectories . . .	69

List of Figures

1.1	Schematic of filtering and smoothing approaches	13
2.1	Conceptual model for tidal rectification	23
2.2	Dependence of resonant frequency and prognostic variables on tidal amplitude	33
2.3	Relative vorticity fields for sub-resonant, resonant, and super-resonant forcing	35
2.4	Typical trajectories for sub-resonant, resonant, and super-resonant forcing	36
2.5	Dependence of final drifter position on tidal amplitude	37
2.6	Map of net Lagrangian displacement after one tidal cycle	39
2.7	Pseudo-SST images after advection by the simplified ocean model . .	40
3.1	Schematic of trajectory definitions	48
3.2	True and modelled trajectories used in experiments	59
3.3	Distributions of drifter locations	61
3.4	Error ellipses along true and modelled trajectory	63
3.5	Error ellipses along true and model sub-trajectories	64
3.6	Error in γ along full and sub-trajectory	65
3.7	Horizontal divergence and water depth along trajectory	66
3.8	Cost functions for perfectly observed trajectory	68
3.9	Estimation error as a function of σ^o	71
3.10	Estimation error as a function of σ^u	72

3.11	Statistical distance of estimated trajectory from observations when including or neglecting observation and model errors	76
4.1	Schematic diagram of advection scheme for calculating J_I	92
4.2	Cost function for pair of pseudo-SST images	94
4.3	A complete ice image from over the Labrador shelf	96
4.4	Pair of extracted subimages used for assimilation	97
4.5	Same as Figure 4.4 after processing	97
4.6	Contour plot of J_I as a function of A and θ	101
4.7	Optimal ocean currents from assimilation and manual tracking	102
4.8	Optimal ocean current streamfunction superimposed on images	104
4.9	Scatter plot of optimal velocities vs. manually tracked velocities . . .	105
4.10	Histograms of pixel intensity for differenced images	106
4.11	Optimal streamfunction compared with bathymetry	107
5.1	Streamfunction variance calculated from NMC method	122
5.2	Latitudinal and pressure dependence of unbalanced temperature variance	126
5.3	Latitudinal and pressure dependence of unbalanced velocity potential variance	127
5.4	Idealised example of effect of localising masks	137
5.5	Horizontal structure function of geopotential height using \mathbf{B}_{hi}	142
5.6	Same as Figure 5.5, but using EOF-based covariance matrices	143
5.7	Horizontal structure functions of geopotential height using \mathbf{B} calculated from masked error samples	145
5.8	Vertical cross-section showing impact of blending with \mathbf{B}_{hi}	146
5.9	Vertical cross-sections of the zonal wind showing effect of baroclinic forcing	147
5.10	Same as Figure 5.9, except for geopotential height	148
5.11	Horizontal structure function of geopotential height and wind showing effect of orographic forcing	149

6.1	Mean state of model used in the identical twin experiment	165
6.2	Schematic diagram showing three phases of experiment	166
6.3	Pattern and time series of first two EOFs	168
6.4	Same as Figure 6.3, but for modes 3 and 4.	169
6.5	Same as Figure 6.3, but for modes 10 and 20.	170
6.6	Power spectra, coherence, and phase of model 2 and 3 amplitudes . .	171
6.7	Plot of J as a function of iteration number	173
6.8	Snapshots of “true” and optimal ocean states	174
6.9	Time differences for the “true” and the optimal solutions	175

Abstract

Oceanographic observations are typically too sparse to provide a continuous picture of the evolving ocean state. However, the ability to accurately estimate the past, present, and future state of the ocean has many important applications including climate change research, fisheries management, weather forecasting, and marine pollution management. Data assimilation methods utilise knowledge of the ocean's governing physical processes to estimate the complete time-dependent ocean state.

The goals of the thesis are to provide effective new approaches for assimilating Lagrangian measurements and to examine sub-optimal assimilation schemes based on a low-dimensional representation of the model state and dynamics. Four related studies address these goals: (1) Several issues pertaining to the assimilation of ocean drifter data are examined, including the effects of the velocity component unresolved by ocean models and the nonlinearity of the advection equation. For illustration, experiments are performed with a simplified ocean model developed to capture the basic nonlinear response to a tidal current over isolated coastal topography. (2) A method for extracting surface currents from sequential satellite images of an advected quantity (such as ice or sea surface temperature) is presented. The problem is formulated in a data assimilation context and successfully applied to both artificial data and a pair of real sea ice images from a region over the Labrador shelf. (3) An approach is developed for incorporating a low-dimensional representation of the forecast error statistics in a sequential assimilation system such that several of the typically imposed assumptions can be relaxed. Within the context of an operational numerical weather prediction system, the approach is shown to effectively resolve dynamical influences on the stationary error statistics. Certain aspects of this study may also be applicable to the newer field of operational ocean prediction. (4) A low-dimensional linear approximation of a nonlinear ocean model is obtained to formulate a sub-optimal assimilation scheme. The method avoids the manual coding of the linearised model and its adjoint by treating the model as a "black box". The effectiveness of the method is demonstrated with an identical twin experiment using an idealised configuration of a nonlinear primitive equation model.

Taken together, the approaches examined in the thesis allow realistic ocean models to be effectively combined with remotely sensed Lagrangian data. This may represent a path for the future development of operational ocean prediction systems.

List of Symbols

symbol	description
Latin Symbols	
a_i	parameters for simplified ocean model
A	factor for scaling wind speed to obtain ice speed
a^1	component of state vector projected into subspace spanned by EOFs
a^2	component of state vector projected into neglected subspace
\mathcal{A}	linearised ocean/drifter model
B	background (b.g.) covariance matrix (for errors in s^b)
B_b	b.g. cov. matrix constructed by blending B_e and B_{hi}
B_e	b.g. cov. matrix constructed with EOFs
B_{hi}	b.g. cov. matrix with homogeneous, isotropic correlations
B_l	localised b.g. cov. matrix
B_t	true b.g. cov. matrix
\mathcal{B}	linear operator that produces optimal increment to the controls
C	temporal response of Rossby wave component
\tilde{C}	complex amplitude of periodic response Rossby wave component
\tilde{C}	spectral representation of global correlations
\mathcal{C}	cov. matrix from long model run used to obtain the EOFs
d_n	statistical distance between estimated and observed positions
D	linear numerical model

symbol	description
$\mathcal{D}()$	time-stepping form of nonlinear numerical model
\mathbf{E}	set of empirical orthogonal functions (EOFs)
\mathbf{E}_1	EOFs retained for reduced dimension basis
\mathbf{E}_2	EOFs of neglected subspace
f	coriolis parameter
f_X	probability distribution function of the random variable X
\mathbf{f}_n	forcing vector at timestep n
\mathbf{F}	EOFs scaled by their singular values
\mathbf{G}	transformation of forcing into effect on the model state
h	water depth
h'	height of bank scaled by water depth
h_∞	water depth away from bank in simplified ocean model
\mathbf{H}	linear (or linearised) observation operator
$\mathcal{H}()$	nonlinear observations operator
i	square-root of -1
I	satellite image: pixel intensity as a function of position
I^t	“true” satellite image observed by a perfect instrument
$\tilde{I}_{N/2}[\mathbf{x}_{N/2}^k I_0]$	image advected from $n = 0$ to $n = N/2$
$\tilde{I}_{N/2}[\mathbf{x}_{N/2}^k I_N]$	image advected from $n = N$ to $n = N/2$
\mathbf{I}	identity matrix
J	cost function
J_α	cost function corresponding to prior estimate for controls
J_d	cost function for assimilating drifter trajectories
J_I	cost function for assimilating images without regularization
$\mathbf{J}_{\alpha\alpha}$	Hessian matrix of J with respect to the controls
L	Lagrange function

symbol	description
L_b	hor. length scale of topographic feature in simplified model
L_c	local horizontal length scale of L
L_e	local horizontal length scale of EOF-based cov. function
L_l	local horizontal length scale of localised cov. function
L	correlation matrix used for localising background cov. matrix
\tilde{L}	spectral representation of L
\mathcal{L}	geostrophic balance operator
n	timestep, i.e. $n\Delta t$ seconds after start of assimilation period
N	total number of time-steps
N_α	dimension of control vector
N_b	total number of background error samples
N_e	number of retained EOFs
N_l	number of localising masks used
N_s	total dimension of the model state
N_y	number of observations
N	diagonal scaling matrix for calculating EOFs from long model run
$p()$	probability of event in brackets
P	dimension of ocean model, i.e. number of variables at all locations
P_s	surface pressure
P'_s	unbalanced component of surface pressure
$(P_s)_w$	balanced component of surface pressure
$\ln(q)$	natural logarithm of specific humidity
$\ln(q)'$	unbalanced component of natural logarithm of specific humidity
r	distance from origin in polar co-ordinate system
R	Ridge parameter
S	temporal response of Rossby wave component

symbol	description
$S(\mathbf{x}, t)$	source/sink term in the pixel intensity conservation equation
\mathbf{S}	spherical spectral transform
\mathbf{s}^b	prior state estimate used in assimilation (background state)
\mathbf{s}_n	model state vector at timestep n , i.e. all prognostic variables
$\Delta \mathbf{s}$	state increment added to background state to obtain full state
$\Delta \mathbf{s}^a$	optimal estimate of state increment
$\Delta \mathbf{s}_u$	unbalanced component of state increments
T	temperature
T'	unbalanced component of temperature
T_\downarrow	balanced component of temperature
\mathcal{T}	amplitude of forcing term for Rossby wave
Δt	time increment between model timesteps
U_∞	maximum tidal current in far-field for simplified ocean model
\mathbf{u}^l	large scale velocity field
\mathbf{u}^m	velocity field from the ocean model, assumed equal to \mathbf{u}^l
\mathbf{u}^o	the ocean-driven component for ice advection model
\mathbf{u}^s	random model error, i.e. the unmodelled component of \mathbf{u}^t
\mathbf{u}^t	the true velocity field
\mathbf{u}^w	the direct wind-driven component for ice advection model
\mathbf{U}	set of singular vectors spanning the model state space
\mathbf{V}	set of singular vectors spanning the sample index space
\mathcal{V}	empirical inverse hydrostatic operator
W_b	weighting term applied to observed ice beacon trajectories in J
W_r	weighting term applied to regularisation term
\mathbf{w}	arbitrary state vector
\mathcal{W}	diagonal matrix of standard deviations of background error

symbol	description
\mathbf{x}_0	position of deployment of drifter
\mathbf{x}^b	observed trajectory of an ice beacon
\mathbf{x}^i	modelled trajectory of an ice floe
\mathbf{x}^k	model trajectory of image pixels for assimilating images
\mathbf{x}^m	model produced trajectory, i.e. model counterpart to \mathbf{x}_n^{obs}
\mathbf{x}^o	observed trajectory data, i.e. horizontal position at timestep n
\mathbf{x}^t	true trajectory that is observed with noise
\mathbf{X}^o	stochastic model for observed drifter positions
\mathcal{X}	linear model between controls and observations
\mathbf{y}	observation vector containing all observations
\mathbf{y}'	difference between observations and $\mathcal{H}(\mathbf{s}^b)$
Z	amplitude of the mean current in the simplified ocean model
z	a single forecast (background) error sample
\mathbf{Z}	all forecast (background) error samples in matrix form
Greek Symbols	
α	scalar control parameter
$\boldsymbol{\alpha}$	vector containing the controls
$\tilde{\alpha}$	spectral representation of $\boldsymbol{\alpha}$
$\hat{\alpha}$	optimal estimate for the controls
$\boldsymbol{\alpha}_0$	prior estimate for $\boldsymbol{\alpha}$
γ	linearised advection model
Γ	linearised advection model along an entire trajectory
Δ	small perturbation for calculating finite difference approximations
δ_j	perturbation vector (dimension N_e) with j th element equal to Δ
Δ_j	perturbation vector (dimension N_s) with j th element equal to Δ
ϵ^i	error in pixel intensity for an observed satellite image

symbol	description
ϵ^o	random observation error vector
ϵ^m	vector of random errors in the model solution
ϵ^{tot}	total error between the modelled trajectory and the observations
ϵ^x	vector of random errors in the modelled drifter trajectory
ϵ	Rossby number
ζ	relative vorticity
θ	angle to the right of wind vector of resulting wind-driven ice motion
ϑ	angle between streamlines and wind vector
λ	linear friction coefficient
λ	friction coefficient scaled by f
λ	adjoint vector
λ^1	adjoint vector corresponding to low-dimensional subspace
λ^2	adjoint vector corresponding to neglected subspace
Λ	diagonal matrix of singular values
Λ_e	diagonal matrix of only the singular values of retained EOFs
μ	fraction of vorticity dissipated in model of tidal rectification
σ^a	cov. matrix of errors optimal estimate of a scalar control
σ^i	std. dev. of errors in the observed pixel intensity of an image
σ^o	standard deviation of a single component of observation error
σ^u	std. dev. of small scale velocities unresolved by ocean model
Σ^a	cov. matrix of errors in prior estimate of controls
$\Sigma^{\hat{a}}$	cov. matrix of errors in the optimal estimate of the controls
Σ^m	cov. matrix of errors in model solutions
Σ^o	cov. matrix of observation errors
Σ^s	b.g. cov. matrix (same as \mathbf{B})
Σ^{tot}	cov. matrix for total error between modelled traj. and observations

symbol	description
$\tilde{\Sigma}^{tot}$	true cov. matrix Σ^{tot}
Σ^u	cov. matrix of small scale velocities unresolved by ocean model
Σ^x	cov. matrix of the errors in the modelled drifter trajectory
ϕ_i	spatial structure function used to construct ψ
φ	angle for polar co-ordinate system
χ	velocity potential
χ'	unbalanced component of velocity potential
χ_ψ	balanced component of velocity potential
ψ	stream function
ψ_∞	streamfunction of tidal component in simplified ocean model
ω_l	frequency of linear topographic Rossby wave in simplified model
ω_{res}	resonant frequency of simplified ocean model
ω_t	tidal frequency in simplified ocean model
ω'	tidal frequency scaled by f

Acknowledgements

Firstly, I would like to thank my supervisor, Keith Thompson, for his guidance and constant encouragement. The remaining members of the supervisory committee: Pierre Gauthier, Marlon Lewis and Dan Wright must also be thanked for many helpful discussions and for providing valuable suggestions that improved the thesis. I am grateful to Paola Malanotte-Rizzoli, the external examiner, for taking the time to evaluate the thesis and provide constructive comments. Several individuals are thanked for providing the early encouragement that helped convince me to transfer into the oceanography department: Michael Dowd, Marlon Lewis, Chris Taggart, and Keith Thompson. Lastly, the flexibility and support from the management of the Data Assimilation group at the Meteorological Service of Canada during the final two years of the thesis research is acknowledged.

Chapter 1

Introduction

Observations of the ocean are typically too sparse in space and time to provide a continuous picture of the evolving ocean state (consisting of the pressure, velocity, temperature, and salinity fields). Taking into account the different characteristic time and length scales, *Ghil and Malanotte-Rizzoli* (1991) estimated that there are effectively 10^4 times fewer observations of the ocean as compared to the atmosphere. Also, the observations are highly concentrated near the ocean surface, in the northern hemisphere, and close to the continents. This is due to the logistics involved with using traditional, ship-based, methods of data collection. Satellite-borne instruments are a promising source of oceanographic data, providing extensive horizontal coverage. However, because the ocean is essentially opaque to electro-magnetic waves, these sensors are restricted to observing only the ocean surface. Currently, extensive observations from space of the sea-surface temperature and height in addition to surface winds over the ocean are routinely acquired (*Ikeda and Dobson*, 1995). In the future, acoustic tomography arrays and other advanced systems will increase the number of observations of the interior somewhat (*Munk et al.*, 1995). A global network of autonomous drifters that periodically measure the ocean's vertical structure is also planned (*Argo Science Team*, 1999). However, the global ocean will probably remain highly under-sampled for the foreseeable future.

Conversely, the atmosphere is reasonably well sampled, especially in the northern

hemisphere over the continents. A global network of temperature, wind, humidity, and surface pressure measurements has been in place for many years in support of weather forecasting. Traditional sources of data include radiosondes, land and ship-based surface stations, aircraft, and drifting or moored buoys. Remotely sensed observations include cloud-drift from geostationary satellites, marine surface winds from scatterometers, and information on vertical temperature and water vapour profiles from radiometers.

The ability to accurately estimate the past, present, and future oceanic or atmospheric state has many important applications. These include climate change research, fisheries management, weather forecasting, and marine and air pollution management. Accurate estimates of the present state (nowcasting) are necessary to successfully use forecast models to predict the future behaviour of the ocean or atmosphere. For the atmosphere, most of the effort is focussed on improving the accuracy of weather forecasts. Oceanographic research has tended to focus on the study of the ocean's past behaviour, known as hindcasting, to advance our knowledge of important oceanographic phenomena and enable the estimation of uncertain physical parameters. Data assimilation methods address the problem of optimally estimating the past or present state of the ocean or atmosphere when the observations are limited in both space and time.

To extract the maximum amount of information out of limited observations, advanced data assimilation methods attempt to utilise knowledge of geophysical fluid dynamics. Numerical models of the ocean and atmosphere contain much information on the expected relationships between geophysical variables at different locations and times, and between the different types of variables. Data assimilation involves combining a numerical model with the observations and estimates of the uncertainties to obtain a dynamically consistent estimate of the true state of the ocean or atmosphere. Statistical information on the errors in both the observations and the model dynamics can also be utilised. By complementing direct observations with dynamical and statistical information, a better estimate of the true state can be obtained than by

using the observations alone. This thesis focuses on several important issues related to the assimilation of oceanographic and atmospheric data.

The development and application of assimilation methods to geophysical fluids began in meteorology, motivated by the need to improve short-range weather forecasts. A group of “objective analysis” methods succeeded the earlier approach of manually interpolating observed values. These empirical methods rely on limited statistical information and approximate solution methods to nowcast the three-dimensional atmospheric state. The particular method of statistical interpolation (*Lorenc*, 1981), also known as Optimal Interpolation (OI), came into common use at several numerical weather prediction (NWP) centres. The OI scheme attempts to provide the best linear unbiased estimate of the present state based on observations and a short-term forecast produced by a numerical model. The solution method, however, involves certain approximations required to make the algorithm computationally feasible. Essentially, the observations are used to periodically correct the state predicted by a dynamical model. Recently, sophisticated assimilation schemes that rely more on the model dynamics have progressed from the research stage (*Talagrand and Courtier*, 1987) to operational implementation (*Rabier et al.*, 1999) for NWP. These schemes make better use of the observations and provide solutions that are consistent with the model dynamics. However, they also require substantially more computer power.

It is only during the last decade that models have been more frequently combined with data in oceanographic research. This delayed development is due to both the lack of routinely acquired observations and also the absence of an immediate societal demand for ocean prediction. Even with recent increases in the quantity of remotely sensed oceanographic data, assimilation schemes are still required to use models as dynamic interpolators and extrapolators to a greater extent than in meteorology. This is due to the need to propagate information from observations concentrated at the surface and near the coast into the ocean interior. Partly for this reason, research in oceanography has quickly moved to schemes more advanced than, for instance, OI that was used in operational NWP for many years. Also, specialised approaches have

been developed in an attempt to deal with the problem of propagating information from the surface into the interior (*Oschlies and Willebrand, 1996*). Recently some effort has been directed towards establishing operational forecast systems for the coastal ocean for applications such as storm surge prediction and oil spill forecasting (*Heemink and Metzelaar, 1995; Griffin and Thompson, 1996; Aikman et al., 1996; Malanotte-Rizzoli and Young, 1997*).

Our ability to accurately hindcast, nowcast, and forecast the oceanic or atmospheric state, depends on the following three factors:

1. the availability of sufficient data;
2. the ability of numerical models to accurately represent real geophysical processes; and
3. the application of effective methods for combining the data with a model.

The major contributions of this research are to item 3. The goals of the thesis are to: (1) provide effective new approaches for assimilating Lagrangian data, and (2) examine sub-optimal assimilation schemes based on a low-dimensional representation of the state and dynamics of a sophisticated numerical model. The two goals are closely related since the successful utilisation of more sophisticated models depends on having sufficient data to constrain all of the model's degrees of freedom.

The next two sections provide brief introductions to the theory of maximum likelihood estimation and data assimilation using generalised linear regression as the framework. In Section 1.3, a brief description of several standard assimilation methods for time-dependent models is presented. The use of a low-dimensional representation of the model state and dynamics, which is a common theme throughout the thesis, is introduced in Section 1.4. The final section gives an overview of the thesis.

1.1 Maximum Likelihood Estimation

The theory of maximum likelihood estimation (MLE) is quite general and is applicable to many statistical estimation problems. For example, this method can be applied to problems involving nonlinear models and non-Gaussian errors. It also provides a general framework that helps make explicit all simplifying assumptions when formulating an estimation problem. The goal of the method is to determine the most likely set of model parameters of a stochastic model given one observed realization. This provides a useful context in which to consider many approaches for assimilating geophysical data.

To illustrate the method, consider the estimation of a single model parameter from a single observed quantity. Following *Priestley* (1981), the probability density function (pdf) for the observed quantity, denoted by X , is specified as a function of the unknown model parameter, denoted by α . The pdf, $f_X(x, \alpha)$, is defined as

$$\int_x^{x+\delta} f_X(x, \alpha) dx = p(x < X < x + \delta), \quad (1.1)$$

where $p()$ is the probability of the event in the parentheses. Therefore, given the correct value of the model parameter, $f_X(x, \alpha)$ provides a measure of how “likely” the random variable X will be close to x . The distribution of a set of realizations from the stochastic model will be consistent with this pdf. The estimation problem is addressed by considering this function for a fixed value of the observed variable, $X = x$, as the model parameter is varied. This gives the relative likelihood that the random variable X is “close” to the specific value x as a function of α . The maximum likelihood estimator is the value of α that maximises the likelihood that X is close to the observed value, denoted x° . That is, we seek the value of α that maximises $f_X(x^\circ, \alpha)$. *Lorenc* (1986) used MLE and a Bayesian approach to examine several assimilation approaches for NWP.

1.2 Linear Regression as a Framework for Data Assimilation

While MLE is very general and can be applied to many types of estimation problems, it is more practical to discuss various issues related to data assimilation within the simpler context of generalised linear regression. Linear regression is a special case of MLE that is applicable to linear models with Gaussian errors. *Thacker* (1988a) highlighted the connections between standard regression theory and some advanced data assimilation methods.

To illustrate some important issues related to linear regression that apply to both oceanographic and atmospheric data assimilation, consider the following system of linear equations:

$$\begin{aligned} \mathbf{y} &= \mathcal{X}\boldsymbol{\alpha} + \boldsymbol{\epsilon}^\circ \\ &= [\mathcal{X} \ \mathbf{I}] \begin{bmatrix} \boldsymbol{\alpha} \\ \boldsymbol{\epsilon}^\circ \end{bmatrix}. \end{aligned} \tag{1.2}$$

In the notation used here, and throughout the thesis, bold lower-case variables represent vectors, bold upper-case variables represent matrices, and variables in regular type are scalars. The vector \mathbf{y} is the set of N_y observations and $\boldsymbol{\epsilon}^\circ$ are the associated unknown observation errors. The vector $\boldsymbol{\alpha}$ is the set of N_α unknown model parameters, or controls. The $N_y \times N_\alpha$ matrix \mathcal{X} is the linear model, assumed to be error free, that maps the controls into the model counterparts of the observations. This is referred to as the forward model.

Many important problems in oceanographic and atmospheric data assimilation can, in theory, be expressed in the form given in (1.2). As an example, assume we wish to estimate the ocean state from time t_0 to t_3 given observations at times t_n , for $n = 1, 2, 3$. Also, assume that a perfect linear model that describes the evolution of the ocean state from one time-step to the next is given as

$$\mathbf{s}_n = \mathbf{D}\mathbf{s}_{n-1}, \tag{1.3}$$

where the subscripts denote the time index. The vector \mathbf{s}_n is the ocean state vector containing all of the prognostic model variables at all locations on the model grid. The matrix \mathbf{D} is the discretized form of the linear model dynamics. The observations are related to the ocean state according to

$$\mathbf{y}_n = \mathbf{H}\mathbf{s}_n + \boldsymbol{\epsilon}_n^o. \quad (1.4)$$

The vector \mathbf{y}_n contains the observations at time-step n . The matrix \mathbf{H} is the linear observation operator that maps the state vector into the model counterparts of the observations. Here, we have assumed the observation operator is time-independent, corresponding to a fixed observing array. However, the extension to a time-varying observing array, with \mathbf{H} being a function of time, is straightforward. Since the model is assumed perfect, the ocean state at any time can be written in terms of the initial state, \mathbf{s}_0 , which is taken as the set of unknowns. Therefore, this problem can be concisely written in the form

$$\begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \\ \mathbf{y}_3 \end{bmatrix} = \begin{bmatrix} \mathbf{H} & 0 & 0 \\ 0 & \mathbf{H} & 0 \\ 0 & 0 & \mathbf{H} \end{bmatrix} \begin{bmatrix} \mathbf{D} \\ \mathbf{D}^2 \\ \mathbf{D}^3 \end{bmatrix} \mathbf{s}_0 + \begin{bmatrix} \boldsymbol{\epsilon}_1^o \\ \boldsymbol{\epsilon}_2^o \\ \boldsymbol{\epsilon}_3^o \end{bmatrix}. \quad (1.5)$$

This is of the same form as (1.2) with the forward model $\boldsymbol{\chi}$ equal to the product of the two matrices in the first term on the right side of (1.5) and the control vector $\boldsymbol{\alpha}$ equal to the initial state \mathbf{s}_0 .

Assuming that the observation errors are distributed normally as

$$\boldsymbol{\epsilon}^o \sim N(0, \boldsymbol{\Sigma}^o),$$

then using (1.2) leads to the following normal distribution for the pdf of \mathbf{y} :

$$\mathbf{y} \sim N(\boldsymbol{\chi}\boldsymbol{\alpha}, \boldsymbol{\Sigma}^o). \quad (1.6)$$

Under the assumptions of a linear model and Gaussian errors the problem of estimating the most likely values for $\boldsymbol{\alpha}$ reduces to generalised linear regression (see e.g. *Draper and Smith*, 1981). The maximum likelihood estimate of $\boldsymbol{\alpha}$ maximises the pdf

for \mathbf{y} when evaluated at the actual observed values. This is equivalent to minimising the following cost function:

$$J = \frac{1}{2} \boldsymbol{\epsilon}^{\circ T} \boldsymbol{\Sigma}^{\circ-1} \boldsymbol{\epsilon}^{\circ} = \frac{1}{2} (\boldsymbol{\mathcal{X}} \boldsymbol{\alpha} - \mathbf{y})^T \boldsymbol{\Sigma}^{\circ-1} (\boldsymbol{\mathcal{X}} \boldsymbol{\alpha} - \mathbf{y}), \quad (1.7)$$

obtained by taking the $-\log()$ of the likelihood function. The superscript T represents matrix transposition. By setting the gradient of J with respect to $\boldsymbol{\alpha}$ to zero, the optimal estimate is obtained:

$$\hat{\boldsymbol{\alpha}} = (\boldsymbol{\mathcal{X}}^T \boldsymbol{\Sigma}^{\circ-1} \boldsymbol{\mathcal{X}})^{-1} \boldsymbol{\mathcal{X}}^T \boldsymbol{\Sigma}^{\circ-1} \mathbf{y}. \quad (1.8)$$

This estimator is unbiased and, because the likelihood function is Gaussian, it also minimises the estimation error variance. The covariance matrix of the error in $\hat{\boldsymbol{\alpha}}$ is obtained by taking the outer product of (1.8) with itself after removing the mean and taking expectations, as follows:

$$\overline{(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha})(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha})^T} = (\mathbf{J}_{\boldsymbol{\alpha}\boldsymbol{\alpha}})^{-1} = (\boldsymbol{\mathcal{X}}^T \boldsymbol{\Sigma}^{\circ-1} \boldsymbol{\mathcal{X}})^{-1}, \quad (1.9)$$

where $\boldsymbol{\alpha}$ is the true value for the controls and the matrix $\mathbf{J}_{\boldsymbol{\alpha}\boldsymbol{\alpha}}$ is the Hessian of J with respect to $\boldsymbol{\alpha}$.

If the product $\boldsymbol{\mathcal{X}}^T \boldsymbol{\Sigma}^{\circ-1} \boldsymbol{\mathcal{X}}$ in (1.8) is singular, additional information is required to obtain a unique solution. This occurs when $N_y < N_{\boldsymbol{\alpha}}$, but can otherwise occur whenever the data are insufficient to uniquely estimate all of the model parameters. Since the matrix product $\boldsymbol{\mathcal{X}}^T \boldsymbol{\Sigma}^{\circ-1} \boldsymbol{\mathcal{X}}$ is equal to the Hessian of J , its singularity means that the cost function is flat (that is, has zero curvature) in one or more directions in model parameter space. Consequently, adjusting the model parameters along these directions has no effect on J and therefore an infinite number of solutions minimise J . These directions are said to be “unobservable” (*Bennett, 1992*). Even when the matrix product $\boldsymbol{\mathcal{X}}^T \boldsymbol{\Sigma}^{\circ-1} \boldsymbol{\mathcal{X}}$ is not strictly singular, it may be poorly conditioned. The result is that one or more linear combinations of the model parameters are only weakly determined by the data. The least squares estimate can yield unrealistic values along such directions. To obtain a unique or simply more realistic solution, one option is

to introduce prior information on α . If a prior estimate for α is given as α_0 with the errors in this estimate having the covariance matrix Σ^α , then the cost function becomes

$$J = \frac{1}{2} (\alpha - \alpha_0)^T \Sigma^{\alpha-1} (\alpha - \alpha_0) + \frac{1}{2} (\mathcal{X}\alpha - \mathbf{y})^T \Sigma^{\circ-1} (\mathcal{X}\alpha - \mathbf{y}). \quad (1.10)$$

This result can be obtained using a Bayesian approach, as described by *Lorenc* (1986), or by treating α_0 as if it were a set of direct observations of the controls with errors that are uncorrelated with the actual observation errors. The new term penalises departures of the model parameters from their prior estimates. Therefore, model parameters that are completely undetermined by the data take on the value given in α_0 . Now, the maximum likelihood estimate for α is

$$\hat{\alpha} = (\mathcal{X}^T \Sigma^{\circ-1} \mathcal{X} + \Sigma^{\alpha-1})^{-1} (\mathcal{X}^T \Sigma^{\circ-1} \mathbf{y} + \Sigma^{\alpha-1} \alpha_0). \quad (1.11)$$

Note that the introduction of this prior information causes the matrix to be inverted to have full rank, if Σ^α is assumed to be full rank. The approach from the field of statistical estimation known as Ridge regression (see e.g. *Draper and Smith*, 1981) can be related to (1.11). It is equivalent to assuming $\Sigma^\alpha = R^{-1} \mathbf{I}$ and $\alpha_0 = \mathbf{0}$, where R is the ridge parameter. In the limit as the ridge parameter goes to zero, the estimator is equivalent to using the generalised inverse such that the solution is set to zero along those directions that are unspecified by the data.

When the observations are not sufficient to provide a unique estimate of α , additional information can also be supplied in the form of a smoothness constraint such that differences between elements of α that represent spatially adjacent quantities are penalised (*Thacker*, 1988b). Terms that penalise the total kinetic energy or enstrophy may also be introduced to the cost function. These types of additional constraints are generally referred to as regularization terms. A scalar weighting parameter associated with each regularization term must be provided to specify their importance relative to the other sources of information. These weighting factors are often determined by ad hoc methods or by using a cross-validation study in which the factor is adjusted to optimise the fit to withheld observations.

In summary, the approach of generalised linear regression provides a convenient framework for introducing the main concepts of data assimilation. The basic approach of minimising a cost function, J , that depends on the data, a set of unknowns, and any *a priori* information on the unknowns is at the heart of most assimilation methods. The main differences in the assimilation methods reviewed in the next subsection are in the assumptions made in specifying the estimation problem to be solved, and the solution method. When applied to real geophysical problems, some important practical issues include the effect of nonlinear model dynamics and the estimation of error covariances for *a priori* information.

1.3 Assimilation Methods for Time-Dependent Problems

An example was given in the previous section of formulating a simple time-dependent problem so that it could be solved using linear regression. However, for most realistic applications the size of the matrices would be so large that performing the required matrix inversions in (1.8) or (1.11) is impractical. Also, the numerical model or observation operator may be nonlinear or it may be necessary to allow for error terms, ϵ^m , in the model equations to account for inadequacies in the numerical model. To facilitate the discussion, the time-dependent model is expressed in the following general form:

$$\mathbf{s}_n = \mathcal{D}(\mathbf{s}_{n-1}) + \mathbf{G}\mathbf{f}_n + \epsilon_n^m. \quad (1.12)$$

The state vector dimension, denoted by N_s , is typically $O(10^5)$ or greater. The operator $\mathcal{D}()$ is a set of N_s nonlinear functions of the state vector. The vector \mathbf{f}_n contains the forcing variables which can include the surface wind stress and boundary conditions and are assumed to be linearly related and additive to the state vector. The matrix \mathbf{G} provides the necessary linear transformation of the forcing variables to give the effect on the state vector. The observations are related to the model state

according to the nonlinear form of (1.4)

$$\mathbf{y}_n = \mathcal{H}(\mathbf{s}_n) + \boldsymbol{\epsilon}_n^o. \quad (1.13)$$

The covariance matrix of the observation error is denoted by $\boldsymbol{\Sigma}^o$ and of the model error is denoted by $\boldsymbol{\Sigma}^m$. The observation and model errors are often assumed to be Gaussian, uncorrelated in time, and have zero mean.

The goal of the methods described in this section is to estimate the time-dependent state vector, \mathbf{s}_n , using the observations, \mathbf{y}_n , and the model dynamics (1.12) along with estimates of the error statistics. The books of *Bennett* (1992) and *Wunsch* (1996) give descriptions and comparison of these and other approaches. In contrast with standard linear regression, the methods are designed for use with time-dependent models. Some of these methods are, however, not computationally feasible for realistic ocean or atmospheric models. A benefit of these methods is that the optimal solution can often be found regardless of the source or type of available data as long as the model state can be related to the observed quantity through (1.13). As a result, data from several sources with varied spatial and temporal resolutions, such as satellite images, ocean drifters, moored current meters and sea level gauges, can be used simultaneously to obtain the optimal state estimate.

1.3.1 Kalman Filter and Smoother

One of the better known methods for combining data with time-dependent models is the Kalman filter (*Kalman*, 1960). The Kalman filter (KF) algorithm (see Appendix A) sweeps through the data once, forward in time (see Figure 1.1). The estimated state at any given time is statistically optimal with respect to past and present data, for the case that the models for the dynamics and the observations are linear and the errors are normally distributed and serially uncorrelated. Consequently, this method is most appropriate for nowcasting and forecasting applications. The algorithm is sequential in that the model and observations are used in separate steps: the forecast and analysis steps, respectively. The optimal estimate for the state vector at each

analysis time is computed by minimising a cost function, similar to (1.10), that is the sum of the squared statistical distance from the current observations and from the prior estimate of the state. This prior estimate is a model forecast obtained by integrating the model (1.12) forward from the optimal state estimated at the previous analysis time with ϵ^m set to zero. When the KF is first begun, a prior estimate for the state at the first time must be supplied for the initial analysis along with its associated error covariance matrix. The remaining steps of the KF algorithm are simply necessary to update the error covariance matrices associated with both the model forecast and the optimal state estimate.

Since the covariances of the errors in the estimated state vector at each time are calculated as a part of the KF algorithm, the uncertainties of all the estimates are available. However, the need to propagate the error covariances through time usually renders the KF algorithm computationally infeasible for practical oceanographic and atmospheric applications. Several attempts have been made to simplify this part of the KF algorithm (e.g. *Fukumori and Malanotte-Rizzolli*, 1995; *Cane et al.*, 1996; *Dowd and Thompson*, 1997; *Fisher*, 1998).

For observation and model errors that are Gaussian with zero mean, the optimal estimates will be unbiased and have Gaussian error. In the case of nonlinear model dynamics the extended Kalman filter (EKF) can be used (*Jazwinski*, 1970). Nonlinearity, however, can lead to error statistics for the resulting forecast that are significantly non-Gaussian even when the observation and model errors are Gaussian. In this case, the optimal estimates may become biased and the covariance matrix is not sufficient to describe the distribution of the errors; higher order moments are required. *Miller et al.* (1994) demonstrated the need to include higher order moments in an EKF applied to the Lorenz model.

The Kalman smoother is an extension of the filter that produces an improvement to the KF state estimates (*Gelb*, 1974). After applying the KF algorithm, an additional sweep is made through the data, backwards in time. This sweep allows the KF estimates to be modified by incorporating information from future observations.

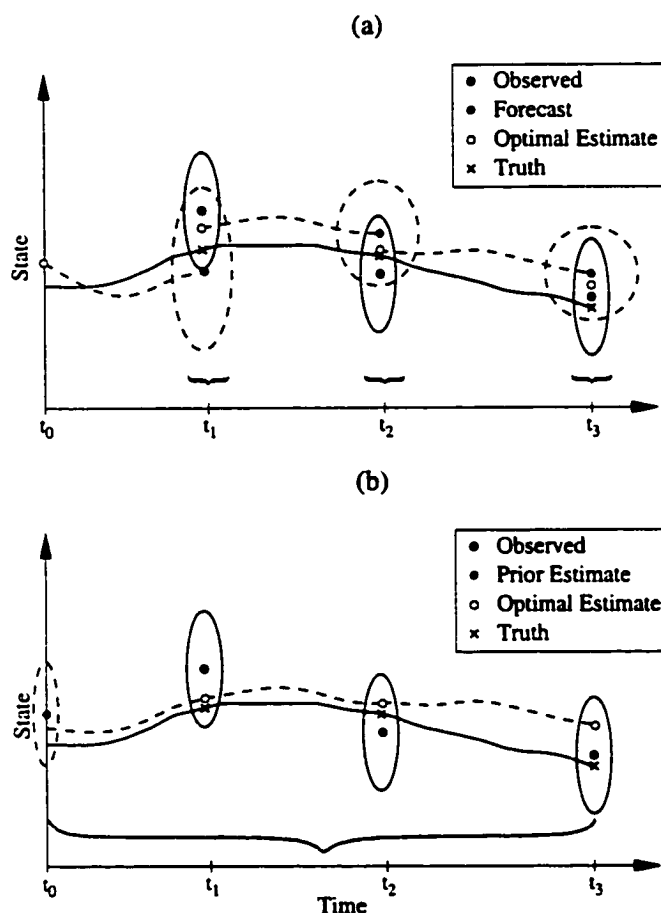


Figure 1.1: Schematic diagram of (a) a filtering (sequential) algorithm, such as OI or the KF, and (b) a strong constraint smoothing algorithm, such as the four-dimensional variational approach. In both panels the dashed line represents a model solution and the solid line is the evolution of the true state. The ellipses represent the uncertainty on the observations and prior estimates (forecasts). In the filtering algorithm the optimal estimate at each time is calculated using only the observations and forecast at a single time along with their uncertainties (denoted by the small braces). For the smoothing approach, all of the data within the assimilation window and a prior estimate of the initial conditions are used simultaneously to estimate the optimal states over the entire period (denoted by the large brace). This figure shows the strong constraint smoother, but perturbations (with specified statistics) can be introduced along the estimated state trajectory to allow some divergence from a perfect model solution (weak constraint).

The result is a set of state estimates that is optimal with respect to all of the data. Therefore, this approach is suitable also for hindcasting applications. The optimal smoother estimates minimise the cost function

$$\begin{aligned}
 J = & \frac{1}{2} (\mathbf{s}_0 - \mathbf{s}_0^{\mathbf{b}})^T \boldsymbol{\Sigma}^{\mathbf{s}-1} (\mathbf{s}_0 - \mathbf{s}_0^{\mathbf{b}}) + \\
 & \frac{1}{2} \sum_{n=1}^N \left[(\mathbf{H}\mathbf{s}_n - \mathbf{y}_n)^T \boldsymbol{\Sigma}^{\mathbf{o}-1} (\mathbf{H}\mathbf{s}_n - \mathbf{y}_n) \right] + \\
 & \frac{1}{2} \sum_{n=1}^N \left[(\mathbf{s}_n - \mathbf{D}\mathbf{s}_{n-1} - \mathbf{G}\mathbf{f}_n)^T \boldsymbol{\Sigma}^{\mathbf{m}-1} (\mathbf{s}_n - \mathbf{D}\mathbf{s}_{n-1} - \mathbf{G}\mathbf{f}_n) \right].
 \end{aligned} \tag{1.14}$$

The first two terms in (1.14) are analogous to the cost function (1.10), except that the time dependence is explicitly stated and a slightly different notation is used. The state vector $\mathbf{s}_0^{\mathbf{b}}$ is the prior estimate for the initial state (sometimes referred to as the background state) and $\boldsymbol{\Sigma}^{\mathbf{s}}$ is the covariance matrix for the error in this estimate. However, the smoother suffers from the same computational limitations as the filter due to the need to explicitly propagate the error covariances.

1.3.2 Variational Method of Data Assimilation

The variational method of data assimilation (also referred to as the adjoint method) can be used to address the same problem as the Kalman smoother, that is, to minimise the cost function (1.14). However, the approach is quite different. *Thacker and Long* (1988) give a good overview of the method along with one of the first applications to an oceanographic problem. The method enables the gradient of J with respect to all of the unknown variables to be efficiently calculated (as described in Appendix B). The optimal set of controls can then be found by minimising J through an iterative process using a gradient descent optimisation algorithm. In general, the model errors could be included in the set of unknowns to be estimated. However, since the number of iterations required for the minimisation to converge is proportional to the number of unknowns (*Gill et al.*, 1981), the adjoint method is used most often in conjunction with the assumption that the model error terms are negligible (that is,

$\Sigma^m = 0$). Now, the state at any time can be expressed as a function of the initial conditions, and open boundary conditions for limited area models. This assumption can often be justified on the grounds that the level of uncertainty in the initial and open boundary conditions is much greater than that in the model dynamics. Therefore, the initial and boundary conditions are taken as the controls. In addition, any uncertain model parameters or forcing variables can also be included in the set of controls. The method can be used with nonlinear models and does not require the burden of propagating large error covariance matrices using the model dynamics. However, for nonlinear models J may contain multiple extrema and therefore convergence to the global minimum is not guaranteed. In practice, the number of iterations required for convergence is typically much less than the control space dimension since most of the reduction in the cost function can be obtained by modifying only the large scales (*Tanguay et al.*, 1995).

Applications of this approach in oceanography include those described by *Thacker and Long* (1988), *Griffin and Thompson* (1996), and *Weaver and Vialard* (1999). Within the field of operational meteorology, this approach has become known as 4D-Var (*Talagrand and Courtier*, 1987). Examples of recent NWP applications are those of *Laroche and Gauthier* (1998), *Thépaut et al.* (1999), *Zupanski et al.* (1999), and *Rabier et al.* (1999). Once the adjoint model is obtained, it can be used for other applications such as sensitivity analysis with respect to uncertain model parameters or initial conditions (e.g. *Thompson and Sykes*, 1990; *Errico and Vukicevic*, 1992) or the calculation of initial perturbations that result in optimal growth (*Ehrendorfer and Tribbia*, 1997).

A common assimilation method used in NWP is the so-called 3D-Var scheme (*Parish and Derber*, 1992; *Courtier et al.*, 1998; *Gauthier et al.*, 1999). This approach is similar to the Kalman filter and OI in that it is sequential and therefore the cost function only includes data for a specific time over the three spatial dimensions. This is in contrast to the 4D-Var approach that simultaneously incorporates data through space and time for the estimation. To make the algorithm feasible for realistic NWP

applications, the error statistics are not propagated in time, but assumed to be stationary (or, at least, slowly evolving in a way that is independent of the flow). Also, unlike OI and the KF, an iterative minimisation procedure employing the cost function gradient is used to obtain the optimal solution. Thus, the approach is identical to OI except for differences in the solution method. With OI, the estimation problem is simplified by limiting the number of observations that influence the optimal state estimate at a given location. This approach is known as “data selection” (*Lorenc, 1981*). Conversely, the iterative solution method of 3D-Var allows all observations to be included in calculating the state estimate at each location. Unlike OI, the 3D-Var approach also allows the use of observation types that are only indirectly related to the model variables, such as measurements of atmospheric radiance.

1.4 Low-Dimensional Representation of the Model State and Dynamics

Assimilation methods applied to time-dependent problems can be computationally expensive, often to the extent of being infeasible for large, sophisticated oceanic and atmospheric models. However, due to sparse observations and limited knowledge of the required statistical information, the use of simpler sub-optimal assimilation schemes is often justified. A common theme of many sub-optimal assimilation schemes is the use of a reduced dimension subspace for representing the state vector and the model dynamics. Throughout the thesis a low-dimensional representation of the state vector and/or model dynamics is employed to achieve practical means for estimating the state of the ocean or atmosphere from limited and diverse types of data.

As an example, consider the KF algorithm described above. The most computationally expensive part of the KF is the propagation through the model dynamics of the error covariance matrix associated with the forecast. Since the covariance structures of the errors may be quite smooth relative to the computational grid of the model, most of the variance will be concentrated within a relatively low-dimensional

subspace. Therefore, by somehow representing and propagating the error covariance matrix within a suitably chosen subspace the KF may be made feasible, even for large models. Also, applications of the so-called incremental approach to variational assimilation problems (described in detail in Chapters 5 and 6) often employ a low-dimensional representation to speed up the convergence to the cost function minimum.

1.5 Outline of Thesis

The thesis is comprised of several distinct subprojects that can be loosely divided according to two themes. The first theme is the optimal use of Lagrangian data, addressed in Chapters 3 and 4. The second theme, which runs throughout the thesis, is the use of a low-dimensional representation of the model state or dynamics to provide practical sub-optimal assimilation schemes.

A detailed outline of the thesis follows:

- In Chapter 2, a low-dimensional idealised model of the barotropic response to a tidal current flowing over isolated topography is presented. The model captures the basic nonlinear dynamics of a forced topographic Rossby wave and a rectified mean current. These include the effects of the mean current on the resonant frequency of the system. Simulated drifter trajectories and pseudo sea surface temperature (SST) images are generated using the model. These artificial observations are used in the subsequent two chapters along with the simplified ocean model to demonstrate how such data can be used to extract information on the ocean current field.
- In Chapter 3, several issues pertaining to the assimilation of ocean drifter data are addressed. These include the effect of the small scale velocities that are unresolved by ocean models. The combination of this error and the nonlinearity of the advection equation can lead to difficulties when applying linear estimation theory. Practical approaches to overcome these problems are proposed and

evaluated. For illustration, experiments are performed with the simplified ocean model presented in the previous chapter.

- In Chapter 4, a method for assimilating sequential satellite images of a quantity that is advected by the surface currents (such as ice or sea surface temperature) is presented. Existing methods for automatically extracting velocity fields from sequential images are first described. Then, the general estimation problem is formulated for this problem in a data assimilation context. The method is first tested using the pseudo-SST images obtained in Chapter 2. It is then applied to the recovery of surface ocean currents using a pair of real ice images from a region over the Labrador Shelf.
- In Chapter 5, an approach is developed for incorporating a low-dimensional representation of the forecast error statistics in a sequential assimilation system. The approach allows several of the simplifying (but often unrealistic) assumptions typically imposed when formulating these statistics to be relaxed. The importance of the forecast error statistics in sequential assimilation schemes is first illustrated. The effectiveness of the proposed approach to resolve dynamical influences on the error structures is then evaluated within the context of an operational NWP system. A possible extension of this approach for incorporating temporally evolving error statistics is also suggested.
- In Chapter 6, a low-dimensional linear approximation of a numerical ocean model is developed for the purpose of data assimilation. The model is used in formulating a sub-optimal four-dimensional variational assimilation scheme. The method avoids the need to formulate the adjoint model manually by treating the ocean model as a “black box”. The effectiveness of the method is demonstrated with an identical twin experiment using an idealised configuration of the CANDIE ocean model (*Sheng et al.*, 1998).

The thesis chapters are interrelated in several ways. The method for assimilating sequential satellite images is a direct extension of the method for assimilating

Lagrangian trajectories. Each image pixel is treated as a drifter whose most likely trajectory is determined from a sequence of images. Also, the use of Lagrangian trajectories and satellite images represents a large increase in the amount of available oceanographic data. This increase in the quantity of data potentially means that more complex ocean models may be used. This provides the motivation for the use of sub-optimal assimilation methods such as those described in Chapters 5 and 6. Chapters 5 and 6 both use a similar method to calculate a reduced dimension subspace to represent the model state within an assimilation system. Though mostly focused on the NWP problem, some aspects of Chapter 5 may also be applicable to the newer field of operational ocean prediction. Taken together, the oceanographic-related chapters allow realistic ocean models to be efficiently combined with a large amount of remotely sensed data. This represents a possible path for the future development of operational ocean prediction systems. The thesis ends by summarising the overall conclusions of the research and discussing the implications for the future of operational ocean prediction.

Chapter 2

A Low-Dimensional Ocean Model

A model of periodically forced barotropic flow over an isolated topographic feature is presented in this chapter. The model is a simplified version of the nonlinear finite-difference model of *Yingshuo and Thompson (1997)*. The flow consists of a periodic tidal current, a single topographic Rossby wave mode propagating around the bank, and an along-isobath current with a nearly constant steady-state velocity generated by the process of tidal rectification. The model captures the complex nonlinear behaviour resulting from the interaction of the tide, Rossby wave, and along-isobath current including a relationship between the amplitude of the along-isobath current and the resonant frequency of the Rossby wave.

In the next section the oceanographic theory related to the mechanisms of tidal rectification and resonance shifting captured by the simplified model are reviewed. The derivation of the prognostic equations and the specific model configuration are presented in Section 2.2. The rectification mechanism captured by the model is described in Section 2.3. Section 2.4 discusses the effect of the along-isobath current amplitude on the resonant frequency of the system. The effect of the system's resonance on the vorticity field is shown in Section 2.5. In Section 2.6, the model is used to simulate a series of drifter trajectories and also to produce a pair of pseudo-SST fields that are advected by the model flow field. These sources of data are used in the subsequent two chapters to illustrate how information on the ocean current

field can be extracted from these types of observations. The final section gives some conclusions.

2.1 Dynamical Theory

The early studies of *Huthnance* (1973) and *Loder* (1980) provide the theory and observational evidence for the generation of a steady along-isobath current caused by the rectification of a tidal current over a ridge or bank. Following *Pingree and Maddock* (1985), the steady current around a bank can be related to the advection of relative vorticity by considering the nonlinear vorticity equation

$$\frac{\partial \zeta}{\partial t} + \nabla \cdot (\mathbf{u}\zeta) + f\nabla \cdot \mathbf{u} + \lambda\zeta = 0, \quad (2.1)$$

where \mathbf{u} is the horizontal velocity vector, f is the Coriolis parameter (assumed constant), and λ is a linear friction coefficient. In deriving (2.1), the simplifying assumption has been made that the friction coefficient is independent of water depth. If the terms in the above equation are integrated through time and the system is assumed to be in a periodic steady state, the first term is eliminated. Integrating the terms horizontally over a disc centred on the bank and bounded by a depth contour eliminates the third term, assuming a rigid lid. The remaining terms give the relationship

$$\langle v \rangle = -\frac{1}{\lambda} \langle u\zeta \rangle, \quad (2.2)$$

where u and v are now the across and along-isobath velocities, respectively. The braces represent averages through time and around the boundary of the disc (along a depth contour). Therefore, the along-isobath component will have non-zero mean whenever the mean flux of relative vorticity across a depth contour is non-zero.

A persistent relative vorticity flux across a depth contour can be sustained by the combined effect of the Earth's rotation and friction, as described by *Huthnance* (1973). To illustrate, consider the conceptual model of a column of water advected up and down a linear slope while conserving potential vorticity $(\zeta + f)/h(\mathbf{x})$. Therefore,

anywhere along its path the relative vorticity of the column is a linear function of the across-isobath position and equal to

$$\zeta = \frac{f}{h(0)} \frac{dh}{dx} x, \quad (2.3)$$

where the relative vorticity is assumed to be zero at the mean position, $x = 0$, for convenience.

A plot of the relative vorticity as a function of position is shown in Figure 2.1a (assuming dh/dx is positive). At any point along the path of the water column, the net flux of relative vorticity across a depth contour is zero. Next, consider a column that is allowed to dissipate relative vorticity due to bottom friction and, for the sake of this conceptual model, the dissipation occurs only at the end points of the path. The resulting evolution of relative vorticity as a function of across-bank position (assuming $\zeta(0) \ll f$) is shown in Figure 2.1b, once a periodic steady-state is reached. The corresponding plot of ζ as a function of u , assuming u is constant along the on-bank and off-bank trajectories, is shown in 2.1c. Now, the net flux of relative vorticity by this column at any position along the path is positive. For example, at the mean position the net flux is

$$\overline{u\zeta} = \mu \frac{u\zeta_{max}}{2} + \mu \frac{(-u)(-\zeta_{max})}{2} = \mu u \zeta_{max}, \quad (2.4)$$

where ζ_{max} is the magnitude of the maximum relative vorticity attained at both end-points and μ is the fraction of this relative vorticity that is dissipated. Therefore, at any point along the path, the water column is responsible for a net positive flux of relative vorticity down the slope (or negative flux up the slope). Consequently, in accordance with (2.2) a mean along-isobath flow is generated, oriented with shallow water to the right of the direction of motion. For a more realistic model, the relationship between ζ and u would be modified in detail, but with similar orientation as the polygon in Figure 2.1c, and would therefore generate a mean current in the same direction.

A net across-isobath flux of relative vorticity also occurs when a topographic Rossby wave is excited by tidal currents over a bank. If the wave is very energetic,

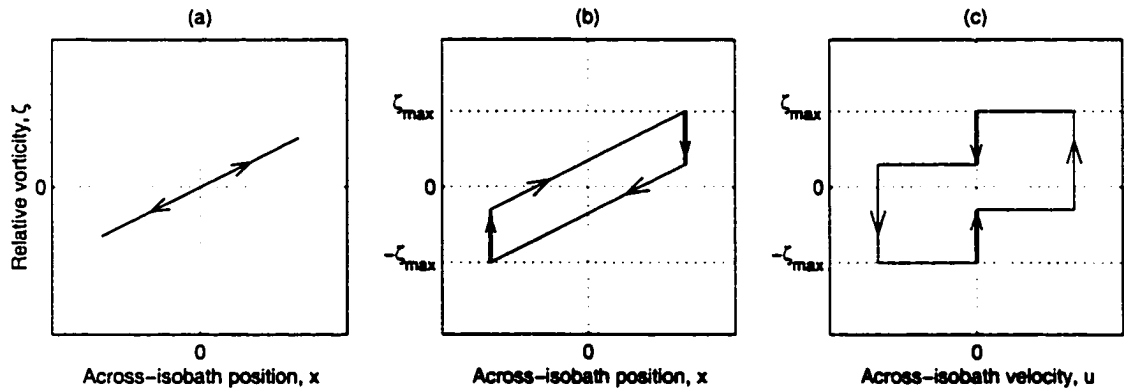


Figure 2.1: Conceptual model for tidal rectification over a linear slope. (a) Case with no friction leading to no net across-isobath flux of relative vorticity. (b) Case with friction included causing a net positive vorticity flux down the slope at each location along the path. (c) Plot of vorticity as a function of the across-isobath velocity. Note: thick arrows denote dissipation and water depth increases in the positive x -direction.

the dominant forcing of the along-isobath current is the advection by the tide of the relative vorticity associated with the topographic Rossby wave. When the frequency of the forcing equals the frequency of the wave propagating around the bank, the system becomes resonant resulting in a strong clockwise mean flow around the bank. This mechanism will be further discussed later in the context of the simplified model.

In addition to the nonlinear interaction of the topographic Rossby wave and the tide in producing a mean current, the mean current can also affect the wave. *Hart* (1990) presented a theory of how such nonlinear effects can modify the resonant frequency of a topographic Rossby wave away from the linear solution. Empirical evidence for this effect was shown by *Pratte and Hart* (1991) from laboratory experiments. *Yingshuo and Thompson* (1997) discuss two possible mechanisms for this effect of the mean flow on the wave frequency. Since the mean flow and the phase speed of the wave are both clockwise around the bank, the frequency of the wave is increased by a Doppler shift. Conversely, because negative relative vorticity over the top of the bank is associated with the mean flow, this also has the effect of reducing

the total background vorticity in which the wave is propagating, thus decreasing its frequency. They argue that this latter effect dominates, thus explaining the decrease in the system's resonant frequency with increasing along-isobath flow, as observed in their modelling study.

2.2 Scaling Analysis and Model Description

The simplified model introduced below is an attempt to capture the dominant barotropic response to an oscillating tidal flow over a topographic feature, such as the banks found on the Eastern Canadian continental shelf. The approach, suggested by *Yingshuo and Thompson (1997)*, is to use a simplified low-dimensional model. Output from their nonlinear finite-difference model shows that the flow is primarily composed of a topographic Rossby wave with azimuthal wave number one and a nearly steady clockwise along-isobath current superimposed on the large scale tidal current. Therefore the main assumption of the simplified model derived in this chapter is that the solution can be parameterised in terms of only these flow components: the largest scale topographic Rossby wave mode, an along-isobath current, and the tide. By prescribing the fixed spatial structure of the flow components, only a few prognostic variables remain. The goal of seeking a model solution of this form is to capture the basic nonlinear interactions of the wave, along-isobath current and tide components.

The validity of the simplified model is assessed below using scaling analysis. The dynamics are governed by five nondimensional parameters. The Rossby number defined with respect to the tidal velocity away from the bank (U_∞) and the horizontal scale of the bank (L_b) is $\varepsilon = U_\infty/(fL_b)$. The ratios of the tidal frequency and linear damping coefficient to the Coriolis parameter are defined as $\omega' = \omega_t/f$ and $\lambda' = \lambda/f$, respectively. The vertical scale of the bank is represented by the ratio of the height of the bank to the water depth away from the bank, denoted by $\Delta h' = \Delta h/h_\infty$. Finally, the ratio of the external Rossby radius to the horizontal scale of the bank is denoted as $L'_r = \sqrt{gh_\infty}/(fL_b)$. The scales used for deriving the model are given in Table 2.1.

Table 2.1: Scales used to derived the simplified model based on $f = 10^{-4} \text{ s}^{-1}$, $U_\infty = 10^{-2} \text{ m s}^{-1}$, $h_\infty = 10^3 \text{ m}$, $\Delta h = 10^2 \text{ m}$ and $L_b = 10^4 \text{ m}$.

Variable	Scale
ε	10^{-2}
ω'	10^{-1}
λ'	10^{-2}
$\Delta h'$	10^{-1}
L'_r	10^2

The external forcing of the vorticity dynamics consists of planetary vorticity tube stretching by the tide impinging on the bank, $f(\nabla \cdot \mathbf{u}_t)$. If this forcing is normalised by f^2 it is $O(\varepsilon \Delta h')$. Using a polar coordinate system (r, φ) with the origin at the centre of the bank, the water depth is taken to depend only on r . As a consequence, the azimuthal dependence of the external forcing varies as $\cos(\varphi)$ and only the topographic Rossby mode with azimuthal wave number one is excited directly. Modes with higher azimuthal wave numbers can only be excited through nonlinear interactions. Similarly, the along-isobath current (with no azimuthal variation) is only forced by the nonlinear interaction of the directly forced topographic Rossby mode with the tide.

Since $\varepsilon \ll 1$ according to the present scaling, one would expect the dynamics to remain nearly linear. However, if the periodic tidal forcing equals the free-wave frequency of the topographic Rossby wave, the wave component can grow until limited by friction due to resonance. Depending on the strength of the friction, the amplitude of the wave could be amplified at resonance such that nonlinearity becomes important. Away from resonance the amplitude of the wave scales as the forcing term divided by the tidal frequency.

The three flow components (including only the directly forced topographic Rossby mode) are substituted into the vorticity equation (2.1). The resulting terms have either no azimuthal dependence or a wave number one or two azimuthal dependence.

All non-zero terms in the vorticity equation are listed in Table 2.2, where the subscripts w , a , and t denote the Rossby wave, along-isobath current, and tide components, respectively. The scaling for each term with respect to the nondimensional parameters in Table 2.1 is determined for both resonant and non-resonant forcing. The ε^2 dependence of the along-isobath current (quadratic in the tidal amplitude) is consistent with the theory of *Huthnance* (1973). According to the relative amplitudes (also shown in Table 2.2), when the system is away from resonance all of the nonlinear terms are at least an order of magnitude smaller than the external forcing term, $f(\nabla \cdot \mathbf{u}_t)$. Therefore, the forcing for the harmonic Rossby mode (azimuthal wave number 2) is negligible. However, because the forcing of the along-isobath current has a non-zero temporal mean, the growth of this component is only limited by friction and attains a similar amplitude as the directly forced Rossby mode. Conversely, at resonance several of the nonlinear terms are of at least the same order of magnitude as the external forcing. According to this scaling the harmonics become sufficiently large that they should be included for the simplified model to be valid at resonance. However, since the purpose of developing the model is for idealised assimilation experiments with passively advected drifters or tracers, the resulting displacements from these flow components should be considered. The along-isobath current is expected to be nearly steady and therefore the displacements from this component in isolation are bounded by the horizontal scale of the bank, $L_b \sim 10^4$ m. The temporal and spatial variation of the harmonic Rossby modes causes displacements from these components alone that are bounded by $U_\infty/(\omega_t k) \sim 10^3/k$ m, where k is the azimuthal wave number. Therefore, the importance of the Rossby harmonics on drifter displacement decreases with increasing azimuthal wave number. The nonlinear terms involving the relative vorticity of the tide, ζ_t , can be safely neglected since they are all smaller than the external forcing when the system is at resonance.

By assuming a rigid lid (since $L'_r \gg 1$), the flow field is represented by the parameterised transport streamfunction

$$\psi = \psi_\infty(r, \varphi, t) + [C(t) \cos(\varphi) + S(t) \sin(\varphi)] \phi_1(r) + Z(t) \phi_0(r), \quad (2.5)$$

Table 2.2: Scales of the non-zero terms in the nonlinear vorticity equation after decomposition of the velocity into a topographic Rossby wave, along-isobath current and tide component. Only the Rossby mode with azimuthal wave number one is included. The scales are given when the system is at and away from resonance. The scales and numerical values are normalised relative to the external forcing term, $f(\nabla \cdot \mathbf{u}_t)$.

Term	Scale (at resonance)		Scale (off resonance)	
No azimuthal dependence:				
$\nabla \cdot (\mathbf{u}_t \zeta_w)$	ε/λ'	10^0	ε/ω'	10^{-1}
$\nabla \cdot (\mathbf{u}_w \zeta_t)$	$\varepsilon \Delta h' / \lambda'$	10^{-1}	$\varepsilon \Delta h' / \omega'$	10^{-2}
$\lambda \zeta_a$	ε/λ'	10^0	ε/ω'	10^{-1}
Azimuthal wave number 1:				
$\partial \zeta_t / \partial t$	ω'	10^{-1}	same	
$\partial \zeta_w / \partial t$	ω' / λ'	10^1	1	10^0
$\lambda \zeta_t$	λ'	10^{-2}	same	
$\lambda \zeta_w$	1	10^0	λ' / ω'	10^{-1}
$\nabla \cdot (\mathbf{u}_t \zeta_a)$	$\varepsilon^2 / \lambda'^2$	10^0	$\varepsilon^2 / (\lambda' \omega')$	10^{-1}
$\nabla \cdot (\mathbf{u}_a \zeta_t)$	$\varepsilon^2 \Delta h' / \lambda'^2$	10^{-1}	$\varepsilon^2 \Delta h' / (\lambda' \omega')$	10^{-2}
$\nabla \cdot (\mathbf{u}_w \zeta_a)$	$\varepsilon^2 \Delta h' / \lambda'^3$	10^1	$\varepsilon^2 \Delta h' / (\lambda' \omega'^2)$	10^{-1}
$\nabla \cdot (\mathbf{u}_a \zeta_w)$	$\varepsilon^2 \Delta h' / \lambda'^3$	10^1	$\varepsilon^2 \Delta h' / (\lambda' \omega'^2)$	10^{-1}
$f(\nabla \cdot \mathbf{u}_t)$	1	10^0	same	
$f(\nabla \cdot \mathbf{u}_w)$	$1/\lambda'$	10^2	$1/\omega'$	10^1
Azimuthal wave number 2:				
$\nabla \cdot (\mathbf{u}_t \zeta_t)$	ε	10^{-2}	same	
$\nabla \cdot (\mathbf{u}_w \zeta_w)$	$\varepsilon \Delta h' / \lambda'^2$	10^1	$\varepsilon \Delta h' / \omega'^2$	10^{-1}
$\nabla \cdot (\mathbf{u}_t \zeta_w)$	ε/λ'	10^0	ε/ω'	10^{-1}
$\nabla \cdot (\mathbf{u}_w \zeta_t)$	$\varepsilon \Delta h' / \lambda'$	10^{-1}	$\varepsilon \Delta h' / \omega'$	10^{-2}

Table 2.3: Units for the prognostic variables and horizontal structure functions in the simplified ocean model of barotropic flow over an isolated topographic feature.

Variable	Units
C, S	$\text{m}^3 \text{s}^{-1}$
Z	$\text{m}^5 \text{s}^{-1}$
ϕ_0	m^{-2}
ϕ_1	1

where ψ_∞ is the tidal component, and $\phi_0(r)$ and $\phi_1(r)$ are the radial shapes of the along-isobath current and topographic Rossby wave components, respectively. The only time-dependent prognostic variables are $C(t)$ and $S(t)$ that together determine the amplitude and phase of the topographic Rossby wave, and $Z(t)$, the amplitude of the along-isobath component. The units for these variables are given in Table 2.3. The streamfunction for the tidal current, assumed to be aligned in the east-west direction, is prescribed to have the following form:

$$\begin{aligned} \psi_\infty &= U_\infty h_\infty r \sin(\varphi) \sin(\omega_t t) \\ &= U_\infty h_\infty y \sin(\omega_t t). \end{aligned} \tag{2.6}$$

To derive the simplified model, the streamfunction (2.5) is substituted into the vorticity equation (2.1) and only those nonlinear terms with no azimuthal dependence or only a sine or cosine dependence on φ are retained. Based on the scaling analysis above, terms involving the relative vorticity of the tide are also neglected. All of the remaining terms are then grouped according to their dependence on φ into three equations corresponding to the prognostic variables C , S , and Z .

The fixed radial shape for the wave component, ϕ_1 , is obtained by neglecting the nonlinear and friction terms in the vorticity equation and assuming the free wave solution

$$C(t) = \cos(\omega_t t), \quad S(t) = \sin(\omega_t t). \tag{2.7}$$

The result, given in Appendix C, is an eigenvalue problem with a frequency (eigenvalue) corresponding to each possible radial shape of the wave (eigenfunction). The solution with the highest frequency, ω_l , corresponds to the radial shape with no zero crossings. Because the external forcing projects well onto this mode (it also has no zero crossings) and the numerical solutions of *Yingshuo and Thompson* (1997) suggest it is the dominant mode, this radial shape is chosen for ϕ_1 . The radial dependence in the wave equations is then eliminated by projecting the radial structures of each term onto the radial structure of the wave using the orthogonality relationship defined by the eigenvalue problem. This choice of ϕ_1 is only strictly valid if the nonlinear terms are negligible, which occurs when the forcing frequency is not close to the resonant frequency.

From the scaling analysis above, the dominant balance for the terms with no azimuthal variation is between the friction term and the advection of the relative vorticity of the Rossby wave by the tide. The radial shape of the along-isobath current, ϕ_0 , is chosen such that the radial shapes in the dominant balance are equal (that is, the forcing projects completely onto the response; see Appendix C). This is consistent with assuming that the vorticity balance is satisfied locally in the radial direction.

As a result of these assumptions and manipulations, the following set of nonlinear coupled ODEs for the prognostic variables C, S , and Z is obtained:

$$dC/dt = -\lambda C - \omega_l [1 + (a_1 + a_2)Z] S - \omega_l U_\infty h_\infty \sin(\omega_l t) [a_3 Z + a_4] \quad (2.8)$$

$$dS/dt = -\lambda S + \omega_l [1 + (a_1 + a_2)Z] C \quad (2.9)$$

$$dZ/dt = -\lambda Z - \frac{U_\infty h_\infty f}{2\omega_l} \sin(\omega_l t) C. \quad (2.10)$$

The coefficients a_k , ($k=1, \dots, 4$), derived in Appendix C, are functions of the bottom topography and the radial shapes of the along-isobath current and Rossby wave. The radial shape functions and model coefficients are calculated using topography defined by $h = 1000 - 750 \exp[-(r/L_b)^2]$ where L_b is 10 km. Table 2.4 gives the values for the remaining specified and derived parameters.

Table 2.4: Specified and derived parameters for the simplified ocean model.

Parameter	Value
U_∞	$5 \times 10^{-3} \text{ m s}^{-1}$
f	10^{-4} s^{-1}
λ	$0.02 f$
ω_t	$0.3 f$
ω_l	$-0.345 f$
a_1	$-1.8801 \times 10^{-13} \text{ s m}^{-5}$
a_2	$1.0438 \times 10^{-13} \text{ s m}^{-5}$
a_3	$-6.4110 \times 10^{-9} \text{ s m}^{-4}$
a_4	$6.5349 \times 10^4 \text{ m}$

In summary, the assumptions that allow the model to be simplified down to only three ODEs are only valid under certain conditions. The radial shapes of the topographic Rossby wave and along-isobath current and the neglect of modes with azimuthal wave number greater than one are strictly valid only away from resonance when the nonlinear terms in the wave equation are negligible. However, as the system approaches resonance the simplified model should exhibit the correct qualitative nonlinear behaviour with respect to the interactions of the retained components of the solution.

2.3 Rectification Mechanism

As seen from (2.10), the along-isobath current is forced by the nonlinear interaction of the tide and the wave. To illustrate this mechanism, the nonlinear terms in the wave equations, (2.8) and (2.9), are neglected and the notation of the forcing term simplified, resulting in

$$dC/dt = -\lambda C - \omega_l S + \mathcal{T} \sin(\omega_l t) \quad (2.11)$$

$$dS/dt = -\lambda S + \omega_l C, \quad (2.12)$$

where \mathcal{T} is the amplitude of the tidal forcing term (for $\mathcal{T} > 0$, the maximum westward tide occurs at $t = \pi/2\omega_t$). The wave's dipole structure is oriented such that $(C, S) = (1, 0)$ corresponds with negative vorticity on the eastern side of the bank and $(C, S) = (0, 1)$ corresponds with negative vorticity on the northern side. Assuming a periodic response at the forcing frequency, the transfer function between the tidal forcing and the C component of the wave is given by

$$\mathcal{C} = \left[\frac{\lambda(\lambda^2 + \omega_l^2 + \omega_t^2) + i\omega_t(\lambda^2 + \omega_l^2 - \omega_t^2)}{(\lambda^2 + \omega_l^2 - \omega_t^2)^2 + 4\omega_t^2\lambda^2} \right] \mathcal{T}, \quad (2.13)$$

where \mathcal{C} is the complex amplitude of the periodic wave response.

By neglecting friction and assuming the forcing frequency is not close to the resonant frequency, (2.13) reduces to

$$\mathcal{C} = \left[\frac{i\omega_t}{\omega_l^2 - \omega_t^2} \right] \mathcal{T}. \quad (2.14)$$

Therefore, for sub-resonant forcing ($\omega_l^2 > \omega_t^2$) the phase of the response leads the tidal forcing by $\pi/2$, that is, the negative lobe of the Rossby wave is on the southern side of the bank at the time of maximum westward tidal current. For super-resonant forcing, the response lags the forcing by $\pi/2$. From (2.10), the forcing of the mean flow is proportional to the product of \mathcal{C} with the tide. Therefore, since without friction \mathcal{C} and \mathcal{T} are in quadrature, the wave is aligned with the tidal current such that there is no net flux of relative vorticity generated by the wave across depth contours. Conversely, with λ non-zero the real part of (2.13) is also non-zero and therefore \mathcal{C} and \mathcal{T} are no longer in quadrature. Consequently, with friction included, the wave and tide are partially in phase leading to a net negative flux of relative vorticity onto the bank. Near resonance, the transfer function is dominated by its real part (assuming $\lambda \ll \omega_l$) causing the tide and wave response to be nearly in phase. When in phase, the tidal flow acts to optimally advect the negative vorticity of the wave onto the bank and the positive vorticity off the bank. The result is a large net across-isobath flux of relative vorticity and a strong clockwise mean flow around the bank.

2.4 Resonance Shifting

Inspection of (2.8) and (2.9) shows that the resonant frequency for the system is

$$\omega_{res} = \omega_l [1 + (a_1 + a_2)Z], \quad (2.15)$$

where the frequency for the linear Rossby wave, ω_l , is modified by the amplitude of the along-isobath flow, Z . The coefficients a_1 and a_2 correspond to two mechanisms by which the along-isobath current affects the frequency of the Rossby wave (*Yingshuo and Thompson, 1997*). Since the magnitude of a_1 is larger than a_2 , in the present application, and positive Z corresponds to a clockwise mean current, the net effect is a decrease in the resonant frequency of the system.

Figure 2.2 shows the dependence of ω_{res} (scaled by f) and the typical size of the three prognostic variables on U_∞ after the simplified model is spun-up from rest for 20 tidal cycles. The plotted quantities are the root-mean-squared (rms) values of the prognostic variables averaged over four points of the tidal cycle ($0, \pi/2, \pi, 3\pi/2$) and over the last four tidal periods. The Rossby wave and along-isobath current coefficients were scaled to remove the effect of increasing tidal amplitude in the absence of resonance. Consideration of the dominant balances in the prognostic equations (2.8-2.10) suggest scaling the Rossby wave and along-isobath current by $(U_\infty h_\infty L_b)$ and $(U_\infty^2 h_\infty^2 L_b / \lambda)$, respectively. As expected, the resonant frequency decreases with increasing tidal current amplitude. As the resonant frequency approaches the forcing frequency, the scaled amplitudes of the rms Rossby wave and along-isobath current of the flow exhibit a rapid increase. After the resonant frequency drops below the forcing frequency (super-resonant forcing), the scaled rms values of the prognostic variables decrease to a level similar to those at sub-resonant forcing.

2.5 Snapshots of Vorticity

Figure 2.3 shows the relative vorticity field (scaled by U_∞ / L_b) at the time of maximum westward tidal current for increasing values of U_∞ , corresponding to sub-resonant,

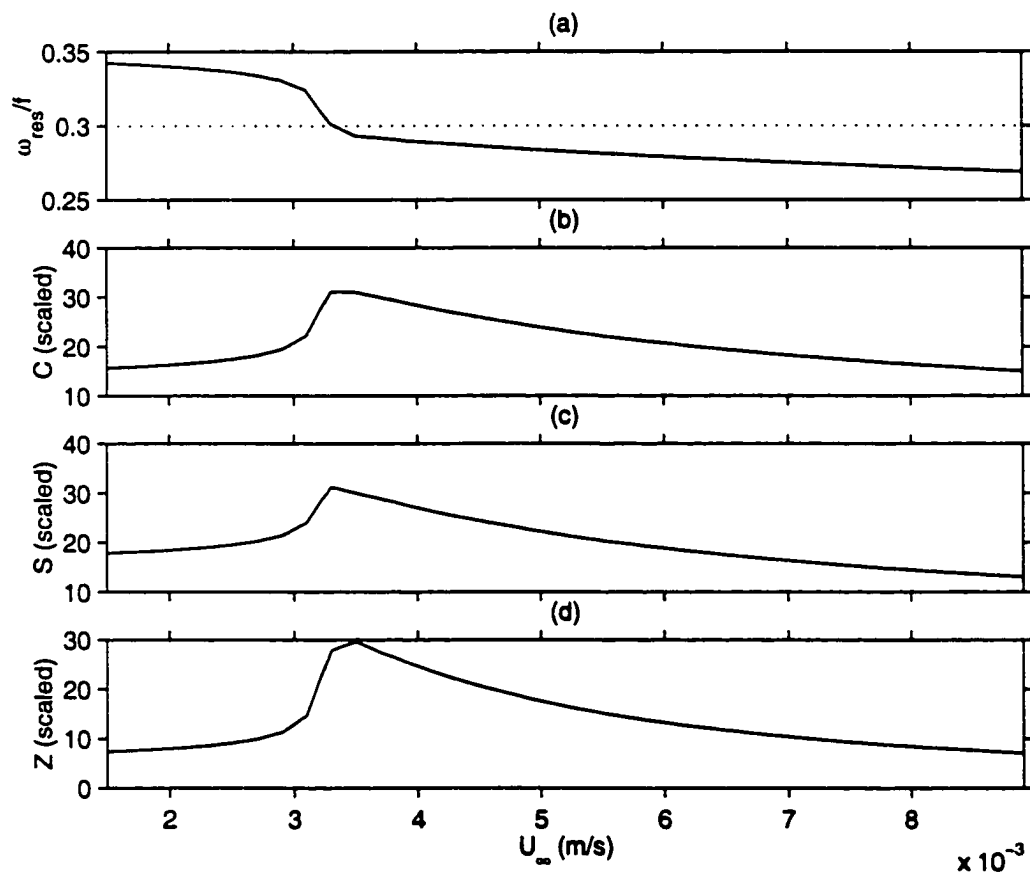


Figure 2.2: Dependence on U_∞ of (a) the resonant frequency, ω_{res} , as given in (2.15), and the scaled rms value of (b) $C(t)$, (c) $S(t)$, and (d) $Z(t)$ averaged over four tidal periods. The largest amplitudes are seen where $U_\infty = 3.35 \times 10^{-3} \text{ m s}^{-1}$ at which point the resonant frequency, ω_{res} (solid curve in (a)), equals the forcing frequency, ω_t (dotted line in (a)).

resonant, and super-resonant forcing. This figure agrees reasonably well with the patterns in Figure 2 of *Yingshuo and Thompson (1997)* and with the discussion in Section 2.3. At sub-resonant forcing, the vorticity field is predominately positive on the northern flank of the bank and negative on the southern flank. The situation is reversed at super-resonant frequencies. At resonance, the vorticity dipole is oriented with negative vorticity to the east where the tide is incident. Also, with increasing U_∞ and the corresponding increase in the strength of the clockwise mean current, which is contributing negative vorticity over the entire bank, the positive lobe of the Rossby wave appears to be increasingly displaced from the centre of the bank.

2.6 Typical Drifter Trajectories and Pseudo-SST Images

Figure 2.4 shows drifter trajectories over two tidal periods deployed at several locations around the bank at the time of maximum westward tidal current. Over the centre of the bank the initial direction of motion is generally to the east for sub-resonant forcing and to the west for super-resonant forcing, corresponding to the reverse in the pattern of relative vorticity shown in Figure 2.3. The overall displacement is clearly greater for trajectories that begin near the top of the bank, where conservation of volume leads to increased velocities.

Starting from a position near the centre of the bank (400 m east, 100 m north of centre) a trajectory was computed over four tidal cycles for a range of values of U_∞ . The resulting positions after two and four tidal cycles are shown in Figure 2.5. The dependence on U_∞ is highly nonlinear as the system passes through resonant forcing. The nonlinear behaviour away from resonance that is apparent after four tidal cycles will be discussed in the next chapter. Because of the interesting relationship between the trajectories and the tidal amplitude, this parameter will be used for examples in the next two chapters as the sole model control parameter to be estimated from simulated observations.

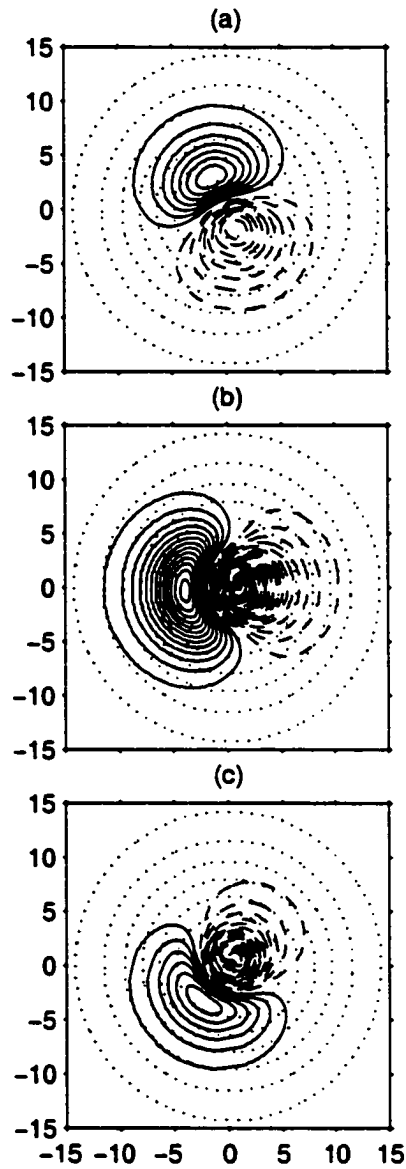


Figure 2.3: The relative vorticity fields (scaled by U_∞/L_b) at the time of maximum westward tidal flow for values of U_∞ corresponding to (a) sub-resonant ($2 \times 10^{-3} \text{ m s}^{-1}$), (b) resonant ($3.35 \times 10^{-3} \text{ m s}^{-1}$) and (c) super-resonant ($7.5 \times 10^{-3} \text{ m s}^{-1}$) forcing. Negative values have dashed contours and the contour interval is 10. Bathymetry contours are also shown at intervals of 100 m. Note how the wave is always aligned with negative vorticity somewhat to the east when the tide is at maximum strength to the west. At resonance the tide and wave are in phase such that the negative vorticity of the wave is optimally advected onto the bank over the entire tidal cycle.

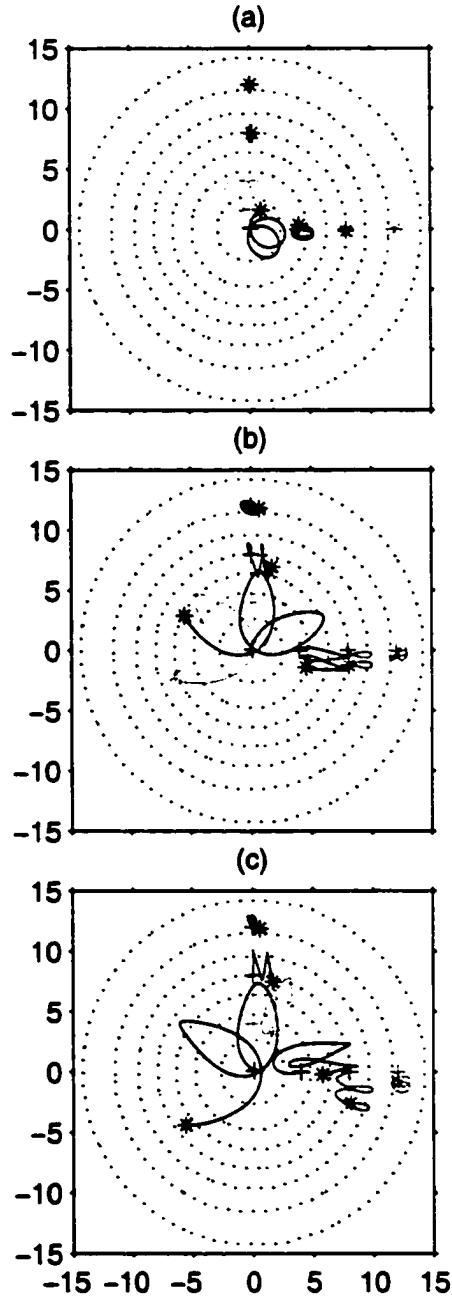


Figure 2.4: Trajectories over two tidal cycles from various initial locations corresponding to the same values of U_∞ as in the previous figure and beginning at the time of maximum westward tidal current. Note the reverse in initial direction over the top of the bank as the system passes through resonance. (+ = starting locations, * = final locations).

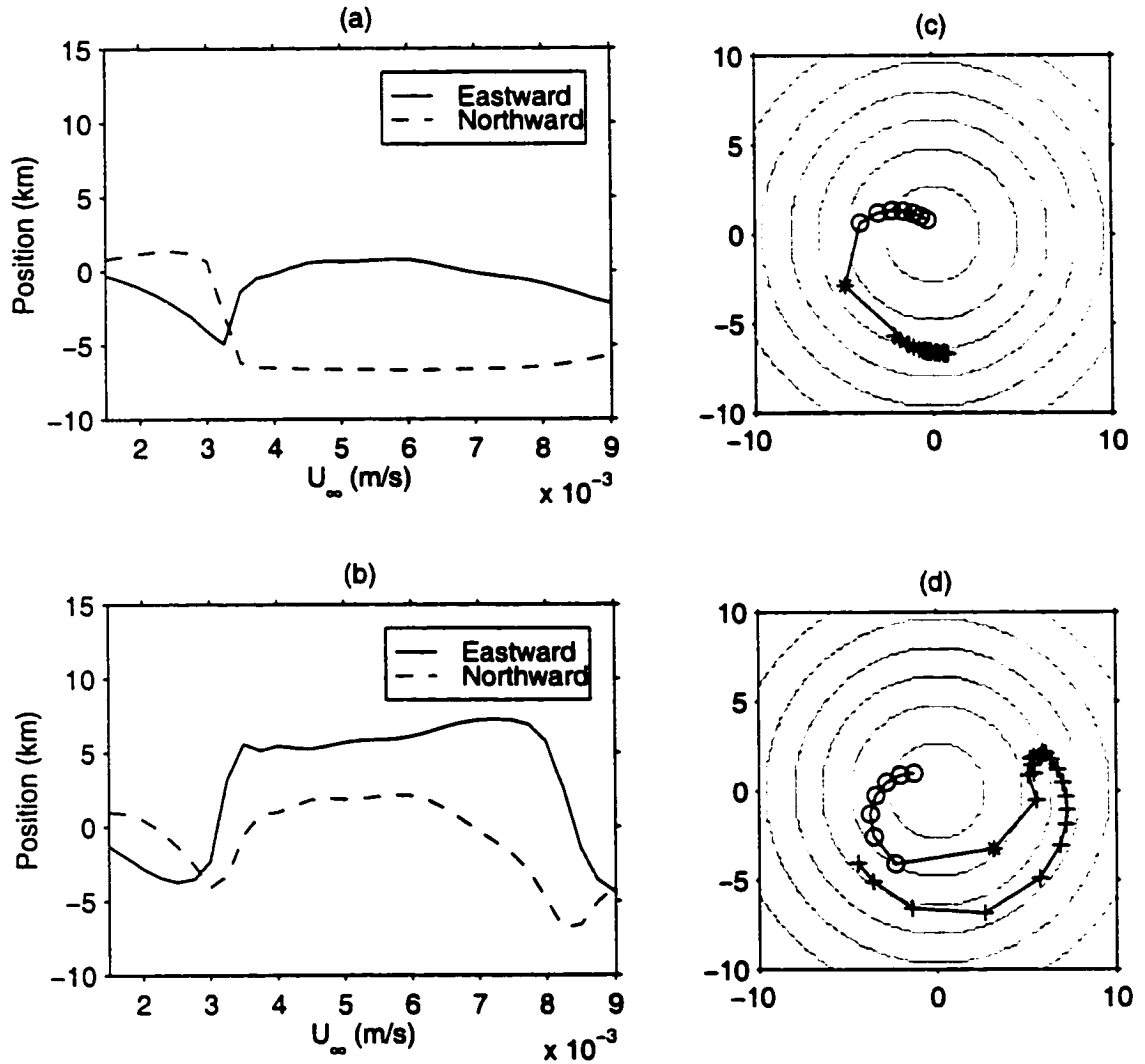


Figure 2.5: Drifter position after (a) two and (b) four tidal cycles as a function of U_∞ starting from an initial location near the centre of the bank. Note the nonlinearity of the relationship, especially near resonance (when $U_\infty = 3.35 \times 10^{-3}$ m s⁻¹). Panels (c) and (d) show the same positions corresponding to sub-resonant (o), resonant (*), and super-resonant (+) forcing superimposed on the bathymetry.

The net Lagrangian displacements over one tidal cycle were calculated over the bank for $U_\infty = 5.0 \times 10^{-3} \text{ m s}^{-1}$. The drifters were released at the time of maximum westward tidal current. The results, shown in Figure 2.6, suggest that over most of the bank net drifter displacement follows isobaths in the clockwise direction. In one region, however, to the south-west of the bank centre, more intensive stirring is occurring. When the drifters are released at the time of maximum eastward tidal current this intense stirring region is instead located to the north-east of the bank centre. Figure 2.7 shows the result of advecting a pseudo-SST field over the same period starting from an idealised field with a linear gradient in temperature between the north-west and the south-east corner of the image.

2.7 Conclusions

This chapter describes a low-dimensional ocean model that captures most of the interesting nonlinear behaviour of the more sophisticated model of *Yingshuo and Thompson (1997)*. By parameterising the solution in terms of only a few flow components with fixed spatial structures, a model is obtained with only three prognostic variables. This simple model, however, can produce complex flow fields and maps of net Lagrangian displacement that depend on the specified tidal amplitude in a highly nonlinear way. These features make this model an ideal tool for idealised assimilation experiments.

Drifter trajectories and sequential tracer images, such as those shown above, are two types of Lagrangian data that could be used to provide information on the ocean current field. Chapter 3 addresses the use of data from drifters. Several key issues are illustrated by numerical examples that use the simplified model presented above. In Chapter 4, a method for extracting velocity information from a sequence of satellite images is presented. The method is first tested using the pair of pseudo-SST images in Figure 2.7 and the simplified ocean model before application to real images.

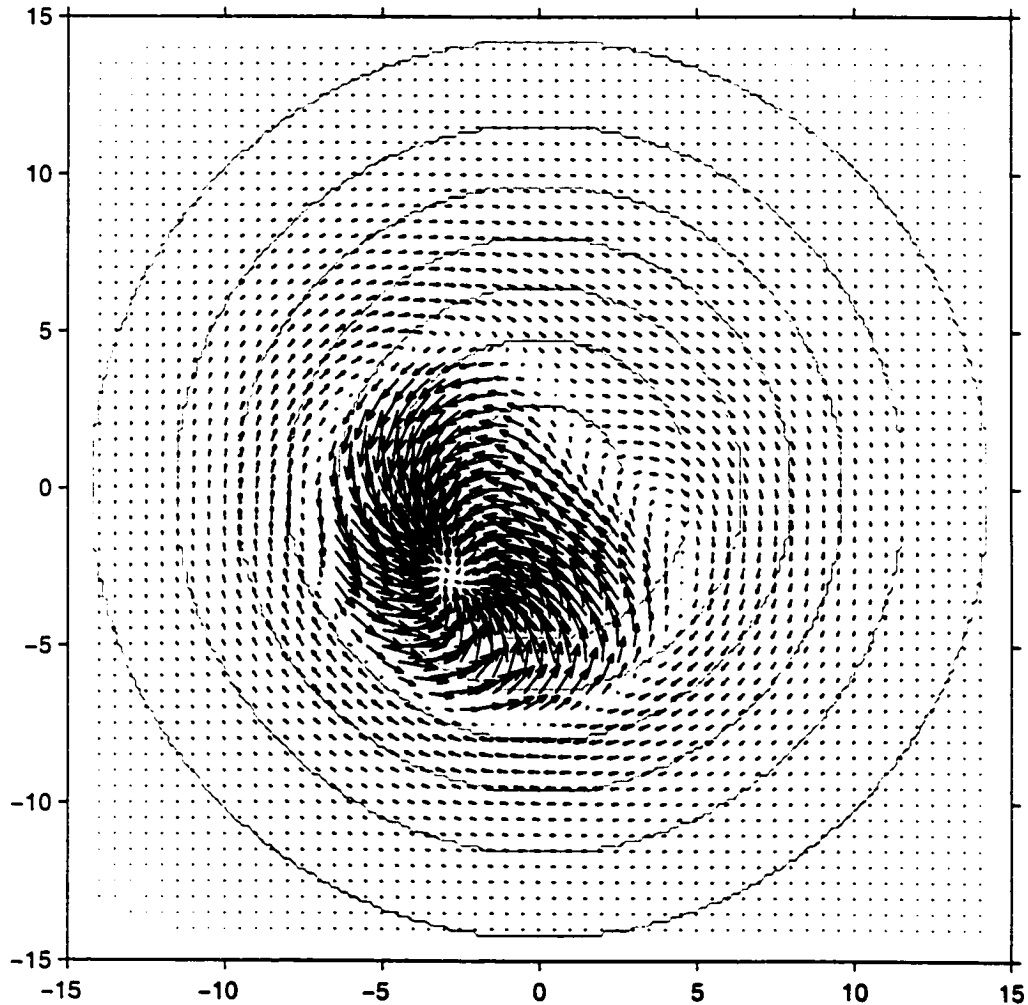


Figure 2.6: Net Lagrangian displacement over one tidal cycle scaled by 0.25. The drifters are released at the time of maximum westward tidal current for $U_\infty = 5.0 \times 10^{-3} \text{ m s}^{-1}$. Note that because the actual displacements are four times larger than the plotted vectors, the drifters do not converge near 2.5 km west and 2.5 km south of the centre of the bank, as it appears in this figure, but are actually displaced passed that point and nearly exchange places with particles initially on the opposite side of the point.

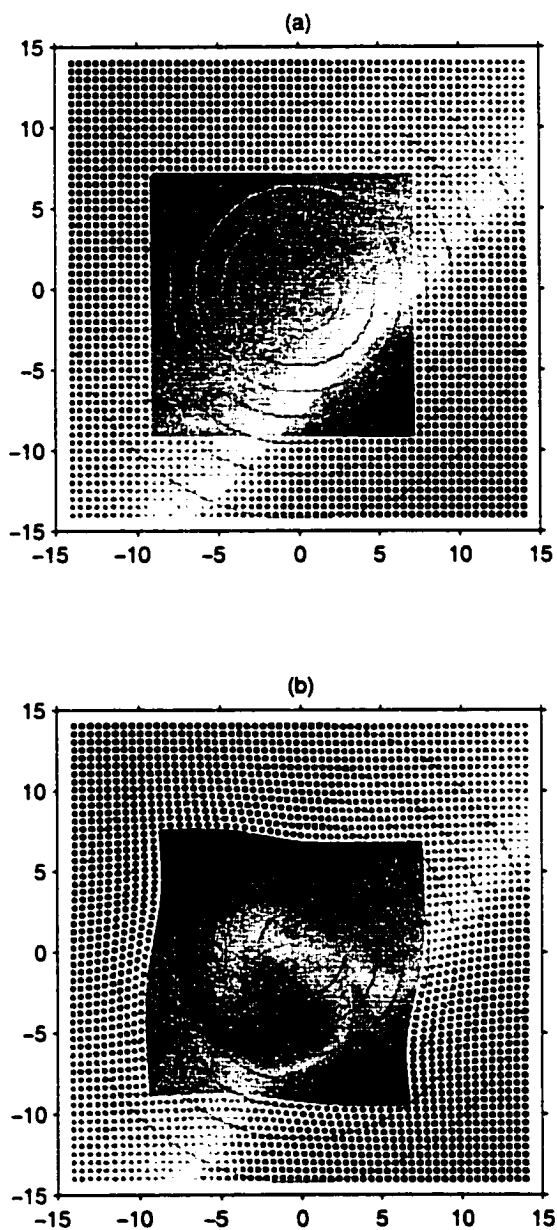


Figure 2.7: Pseudo-SST images over the model domain. The original field (a) has a linear variation in SST between the north-west and south-east corners of the domain. The second image (b) is produced by advecting the pixels in the first image for one tidal cycle with $U_\infty = 5.0 \times 10^{-3} \text{ m s}^{-1}$. The image resolution is 500 m, except in the region of intense stirring where it is 125 m.

Chapter 3

Assimilation of Ocean Drifter Trajectories

3.1 Introduction

Lagrangian trajectory data are widely used in the study of ocean currents. These data are collected by, for example, drogued drifters that track a water mass at a specified depth, sub-surface floats that follow water parcels along constant pressure or density surfaces, or beacons fixed to sea ice floes that track their position over time. With all types of instrument, a time series of locations is transmitted by satellite or stored for subsequent recovery providing information about the Lagrangian character of the flow over a wide range of scales. For example, a major program to establish a global network of approximately 3000 profiling drifters is planned for the near future (*Argo Science Team*, 1999). This will provide oceanographers with a vastly enhanced source of Lagrangian measurements in addition to temperature and salinity profiles.

Ocean models are generally formulated in the Eulerian framework. Possibly for this reason, Lagrangian trajectories have not been widely assimilated into ocean models. In this chapter the assimilation of this type of data into an ocean model is examined. The ultimate goal is to extract the maximum amount of information on the evolving large scale flow field. The framework is, however, also suited to the study of

the statistical properties of the small scale flows. In this context, the actual drifter positions are not of direct interest as they might be, for example, when tracking icebergs.

The following section provides a general discussion of ocean drifters, including a survey of previous studies. A stochastic model for simulating the trajectory and its inherent errors is presented in Section 3.3. Section 3.4 provides a framework for assimilating data into this model. In Section 3.5, the low-dimensional ocean model presented in the previous chapter is used to illustrate several issues that arise when assimilating drifter trajectories in practice. These focus on the nonlinearity of the advection equation and its effect on the estimation problem. Section 3.6 concludes with a discussion of some limitations, practical implications and extensions of the results.

3.2 Ocean Drifter Data

Some ocean drifters are neutrally buoyant and can therefore follow flows along constant density surfaces; others are designed to follow the flow at a predetermined depth, possibly using a drogue suspended from a surface buoy. Depending on the size of the drifter or drogue, the smallest scale of motion affecting the drifter motion may vary from $O(1\text{m})$ to $O(10\text{m})$. Ideally, the trajectory of the drifter is the path integral along a constant depth or density surface of the velocity field at all spatial scales down to about the size of the drifter or drogue. In practice, however, the inability of many drifter designs to follow vertical water motion can lead to a generalised Stokes drift (*Davis*, 1991). The direct effect of the wind on a near-surface drifter can also contribute to the motion (*Niiler et al.*, 1987; *O'Donnell et al.*, 1997).

Previous studies utilising drifter trajectory data generally fall into two categories. The first focuses on estimating average statistical properties of the flow. The second includes attempts to optimally extract the large scale component of the flow. In both categories, most studies convert the time series of observed positions into equivalent

“observed” velocities, usually through first-differencing.

Studies in the first category have applied purely statistical methods to estimate quantities such as Eulerian or Lagrangian time and length scales (e.g. *Middleton and Garrett*, 1986; *Sanderson*, 1995), and dispersion coefficients (e.g. *Thomson et al.*, 1990; *Sanderson and Booth*, 1991) from drifter data. A requirement for most of these calculations is that the mean component of the flow (defined with respect to an appropriate time/length scale) is somehow first estimated and removed. *Thomson et al.* (1990) estimated eddy diffusivity using trajectories from drogued drifters in the Northeast Pacific. They first calculated daily drifter velocities from first-differenced drifter positions in their study of sub-inertial variability. The Lagrangian mean velocity for each trajectory was then removed before calculating the eddy diffusivity. The authors point out that it would be more appropriate to account for a spatially varying mean flow. *Sanderson* (1995) fit a simple linear kinematic eddy model to first-differenced drifter trajectories from the Scotian Shelf after the dominant tidal periods were removed. Although the errors in the regression were assumed to be uncorrelated, analysis of the residual motion showed them to be correlated at large distances in the direction parallel to the velocity vector. This prevented the estimation of the eddy diffusivity. One possible reason for these long correlation lengths is the presence of a slowly evolving component of the flow not resolved by the simple kinematic model. In summary, approaches used to calculate flow statistics from drifter data are strongly dependent on the removal of the large scale, slowly evolving component of the flow.

Plots of filtered drifter trajectories have been used in studies to provide information on regional ocean circulation (e.g. *Poulain and Warn-Varnas*, 1996). Alternatively, some recent attempts to obtain quantitative estimates of the large scale flow have involved the use of a numerical ocean model. The data are assimilated into a model that only resolves the large scales. The resulting “fitted” velocity field is taken to be the estimated large scale flow. For example, *Griffin and Thompson* (1996) converted drifter position data from the Scotian Shelf into velocities by first-differencing.

These velocities were then assimilated along with other sources of data assuming the errors were serially uncorrelated. A similar approach was used by *Morrow and DeMey* (1995) in a study of the Azores Current, except they first filtered the drifter data to remove inertial oscillations. *Kamachi and O'Brien* (1995) assimilated drifter data into an equatorial Pacific model. They started a model-simulated drifter each day from the position of each observed drifter. Then the squared distances between the model simulated and true drifters were calculated once, 24 hours later. The sum of these squared distances was then minimised using an unconventional method incorporating an adjoint model and simple error statistics. The work of *Carter* (1989) describes the use of a Kalman filter to assimilate velocity measurements from drifting neutrally buoyant buoys. That study focuses on a practical approach that enables the Kalman filter to be implemented in a feasible manner. By using this assimilation method, the small scale flow is treated as a part of the serially uncorrelated forecast error that is estimated at each time-step and for which a simple spatial covariance structure is specified. In each of the above studies, only relatively simple error statistics have been used in previous studies when assimilating drifter data.

In this chapter, a framework is presented for assimilating drifter data into an ocean model while accounting for the errors due to the unresolved scales of motion. This approach will be useful in the study of large scale flow or to remove this component for the purpose of estimating the statistical properties of the unresolved scales of motion.

3.3 Model for Drifter Observations

A model for simulating the observed positions of a drifter is presented in this section. This model is used in a later section to assimilate drifter observations. It is composed of a drifter model and a model of the observation process. An important component of the model is the consideration of the unavoidable errors involved in simulating drifter motion. A major source of error is the component of the flow field at scales that

are large enough to influence the drifter motion, but too small to be resolved by the ocean model. An additional source of error originates in the observation process due to imperfect measurement instruments. Incorporating both types of error results in a stochastic model for the observed drifter trajectory. The deterministic component (from the ocean model) and two random components (due to model and observation errors) are discussed in the following three sub-sections. The complete stochastic model is presented in the final sub-section.

3.3.1 Trajectory Models

The true drifter trajectory is the result of integrating the (unknown) ocean velocity field, denoted by \mathbf{u}^t , beginning from the location and time of drifter deployment (given by \mathbf{x}_0 and $t = 0$, respectively). For the sake of clarity, it is assumed that the drifter is perfectly advected by \mathbf{u}^t and therefore does not experience any relative motion through the water. Thus, the true trajectory, \mathbf{x}^t , is the solution of

$$\frac{d\mathbf{x}^t}{dt} = \mathbf{u}^t(t, \mathbf{x}^t) \quad ; \quad \mathbf{x}^t(0) = \mathbf{x}_0. \quad (3.1)$$

In general, this is a nonlinear model due to the dependence of \mathbf{u}^t on \mathbf{x}^t , that is, the spatial variability in the velocity field.

A simple forward difference integration is used to approximate (3.1):

$$\mathbf{x}_n^t = \mathbf{x}_{n-1}^t + \mathbf{u}_{n-1}^t(\mathbf{x}_{n-1}^t)\Delta t \quad ; \quad \mathbf{x}_0^t = \mathbf{x}_0, \quad (3.2)$$

where $t = n\Delta t$. Similarly, we can define a model trajectory, \mathbf{x}^m , derived from the velocity field produced by the ocean model using

$$\mathbf{x}_n^m = \mathbf{x}_{n-1}^m + \mathbf{u}_{n-1}^m(\mathbf{x}_{n-1}^m, \boldsymbol{\alpha})\Delta t. \quad (3.3)$$

To integrate this model, the initial drifter position and the model flow field, \mathbf{u}^m , are required. It is assumed that the model flow field and possibly \mathbf{x}_0 depend on a set of unknown controls, denoted by the vector $\boldsymbol{\alpha}$.

3.3.2 Error in the Model Trajectory

Ocean models cannot perfectly model the true currents that advect a drifter. This is because such models are designed, out of necessity, to capture only the large scale variability of ocean currents. The grid spacing of such models is generally greater than 10 km with the sub-grid-scale processes being parameterised, often using a constant eddy viscosity. This relatively coarse resolution is often adequate to model the evolution of the large scale velocity field. In this chapter it is assumed that, given the correct initial and boundary conditions, the ocean model can perfectly reproduce the larger scales of Eulerian motion. As illustrated below, however, the effect of the unresolved small scale motions on drifter position can accumulate and be amplified over long trajectories.

The true velocity at time-step n and position \mathbf{x} is represented as the sum of the large and small scale components,

$$\mathbf{u}_n^t(\mathbf{x}) = \mathbf{u}_n^l(\mathbf{x}) + \mathbf{u}_n^s(\mathbf{x}). \quad (3.4)$$

The ocean model is assumed to be capable of reproducing the large scale flow field given the correct value for the controls

$$\mathbf{u}^m(\boldsymbol{\alpha}) = \mathbf{u}^l. \quad (3.5)$$

Therefore, $\mathbf{u}_n^s(\mathbf{x})$ accounts for the velocity errors due to the scales of motion unresolved by the ocean model. Since the velocity error is assumed to be unknown and random, its effect can be appropriately accounted for only if we can reasonably estimate its joint probability density function (pdf) through space and time. In reality, the statistics of \mathbf{u}^s should depend on the large scale flow and additional factors such as the surface wind stress and the vertical and horizontal density structure. In this chapter, however, the distribution of \mathbf{u}^s is assumed to be Gaussian with zero mean and covariance matrix at time-step n and position \mathbf{x} defined by

$$\boldsymbol{\Sigma}_n^u(\mathbf{x}) = \overline{\mathbf{u}_n^s(\mathbf{x}) [\mathbf{u}_n^s(\mathbf{x})]^T}. \quad (3.6)$$

The overbar denotes expectation over realisations. The covariance between \mathbf{u}^s at two times along a trajectory is defined as

$$\Sigma_{n,m}^u = \overline{\mathbf{u}_n^s(\mathbf{x}_n^t) [\mathbf{u}_m^s(\mathbf{x}_m^t)]^T}. \quad (3.7)$$

The process will often be correlated in time and space along a true trajectory. To simplify the following development, however, the following is assumed:

$$\Sigma_{n,m}^u = 0 \text{ for } n \neq m. \quad (3.8)$$

This assumption of uncorrelated velocity errors is not a necessary part of the formulation, but it simplifies the analysis and is often convenient since accurate estimates of $\Sigma_{n,m}^u$ are difficult to obtain.

Given the true values for the controls, the error in the modelled drifter trajectory at time-step n is defined as (see Figure 3.1)

$$\boldsymbol{\epsilon}_n^x = \mathbf{x}_n^t - \mathbf{x}_n^m(\boldsymbol{\alpha}). \quad (3.9)$$

The major assumption is now made that this positional error is sufficiently small that between the true and modelled positions the spatial derivatives of the model velocity field can be considered constant. The validity of this assumption partly depends on the length of the trajectory. In Section 3.3.4 and the discussion approaches are given for limiting the growth of $\boldsymbol{\epsilon}^x$. Using (3.9) we then obtain the following expression for the difference in the velocities used in (3.2) and (3.3):

$$\mathbf{u}_n^t(\mathbf{x}_n^t) - \mathbf{u}_n^m(\mathbf{x}_n^m) = \left(\frac{\partial \mathbf{u}_n^m}{\partial \mathbf{x}} \right)^T \boldsymbol{\epsilon}_n^x + \mathbf{u}_n^s(\mathbf{x}_n^t), \quad (3.10)$$

where the derivative is evaluated at \mathbf{x}_n^m (see Appendix B for definitions of vector derivatives). The first term on the right side of (3.10) is the velocity error due to evaluating \mathbf{u}^m at the wrong location (see Figure 3.1). The second term is the velocity error, \mathbf{u}^s , due to the unresolved scales of motion. The following equation for the evolution of the error between the model and true trajectory is then obtained by subtracting (3.3) from (3.2) and using (3.9) and (3.10):

$$\boldsymbol{\epsilon}_n^x = \gamma_{n-1} \boldsymbol{\epsilon}_{n-1}^x + \mathbf{u}_{n-1}^s \Delta t, \quad (3.11)$$

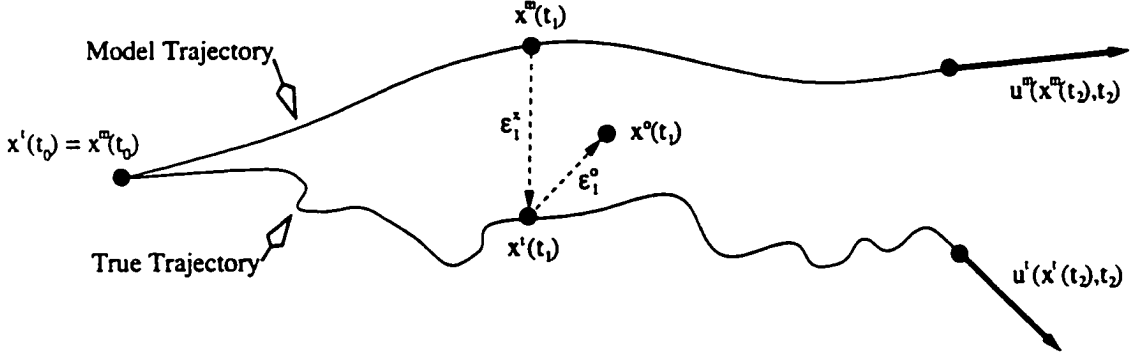


Figure 3.1: Schematic showing relationships between observed and modelled trajectories assuming the initial position in the model trajectory is error free. Symbol definitions: \mathbf{x}^t is the true trajectory; \mathbf{x}^m is the model trajectory produced by advecting the drifter with the large scale velocity field ($\mathbf{u}^l = \mathbf{u}^m$); $\boldsymbol{\epsilon}^x$ is the error in the model trajectory, with covariance matrix $\boldsymbol{\Sigma}^x$, caused by the ocean model not resolving \mathbf{u}^s ; \mathbf{x}^o is the observed drifter position; $\boldsymbol{\epsilon}^o$ is the error in the observed position, with covariance matrix $\boldsymbol{\Sigma}^o$.

where $\boldsymbol{\gamma}_n$ is defined as

$$\boldsymbol{\gamma}_n = \mathbf{I} + \left(\frac{\partial \mathbf{u}_n^m}{\partial \mathbf{x}} \right)^T \Delta t. \quad (3.12)$$

Therefore, $\boldsymbol{\gamma}$ accounts for the linearised effect of the large scale velocity field on the evolution of errors in the modelled trajectory. The initial condition for $\boldsymbol{\epsilon}^x$ in (3.11) is determined by the uncertainty in the location of drifter deployment.

The model for $\boldsymbol{\epsilon}^x$ is forced at each time-step by \mathbf{u}^s . This forcing term is assumed to represent the net effect of the small scale velocities on the drifter position over the time-step. In the case that the small scale velocities are uncorrelated along the trajectory, the resulting errors in drifter position exhibit the first order Markov property: the error at time-step n depends only on the error at $n - 1$ and the random perturbation \mathbf{u}_{n-1}^s . For AR(p) correlated velocity errors (see Appendix D for $p = 1$), (3.11) becomes a $(p + 1)$ order Markov process.

The response of $\boldsymbol{\epsilon}^x$ to an impulsively applied \mathbf{u}^s depends on $\boldsymbol{\gamma}$. For a uniform large scale velocity field (that is, $\boldsymbol{\gamma} = \mathbf{I}$) the impulse response is simply a step function.

Therefore, the errors in position are the sum of all previous errors in velocity (that is, a random walk). In general, over more than one time-step the errors will either be amplified or attenuated in the directions of the eigenvectors of γ with eigenvalues that are greater or less than one in magnitude, respectively. Consequently, the evolution of the positional errors depends on the spatial gradients of the large scale component of the flow field: the region of uncertainty surrounding the model drifter position can become stretched, squashed, or rotated at later time-steps by the linear variation in the large scale velocity field. Compared to previous studies (e.g. *Sanderson, 1995; Griffin and Thompson, 1996*), inclusion of the effect of the linearised velocity field on the evolution of errors in drifter position represents a first step in modelling the interaction between the resolved and unresolved scales of motion.

Assuming \mathbf{u}^s is Gaussian with zero mean, ϵ^x will also be Gaussian with zero mean. Therefore, the evolving distribution of ϵ^x is fully determined by its covariance matrix, denoted by Σ^x . Assuming \mathbf{u}^s is serially uncorrelated along the trajectory, the covariance matrix of ϵ^x evolves according to

$$\Sigma_n^x = \gamma_{n-1} \Sigma_{n-1}^x \gamma_{n-1}^T + \Delta t^2 \Sigma_{n-1}^u. \quad (3.13)$$

This equation is similar to the step in the Kalman filter algorithm that propagates the analysis error covariance matrix from the previous observation time according to the linear model dynamics to produce the forecast error covariance matrix at the current observation time. The covariance of ϵ^x at time-steps n and $n - p$ for $p > 0$ is given by

$$\Sigma_{n,n-p}^x = \gamma_{n-1} \cdots \gamma_{n-p+1} \gamma_{n-p} \Sigma_{n-p}^x. \quad (3.14)$$

Together, (3.13) and (3.14) can be used to specify the entire joint distribution of all of the ϵ^x along a trajectory.

The major assumption made in deriving (3.11), that \mathbf{u}^m is linear between the true and model trajectories, is fundamental to the analysis in the remainder of this chapter. If the model velocity field can not be linearised because the model error is too large, the errors in drifter position will be nonlinear in \mathbf{u}^s . Consequently, the

statistics for ϵ^x may be non-Gaussian and have non-zero mean. The implications of this are discussed in Section 3.6.

3.3.3 Error in the Observations

The observation process introduces a distinct source of error that should be considered when combining drifter data with models. This source of error originates from imperfections in the instruments used to measure, store and transmit the drifter positions. Since the source of this error is often quite well understood, specifying the error statistics is usually more straightforward than for \mathbf{u}^s . The observation error for an observed drifter position at the n th time-step, denoted by \mathbf{x}_n^o , is defined as

$$\epsilon_n^o = \mathbf{x}_n^o - \mathbf{x}_n^t. \quad (3.15)$$

The observation errors are assumed to have zero mean with covariance matrix Σ^o . These errors can usually be assumed to be Gaussian and uncorrelated through time.

3.3.4 Model for the Observed Trajectory

Combining (3.9) and (3.15) we obtain the following stochastic model for the observed drifter position:

$$\mathbf{X}_n^o(\boldsymbol{\alpha}) = \mathbf{x}_n^m(\boldsymbol{\alpha}) + \epsilon_n^x + \epsilon_n^o. \quad (3.16)$$

The modelled trajectory is a deterministic function of $\boldsymbol{\alpha}$ and the error terms are treated as random variables with zero mean. A capital letter is used for the stochastic model of the observed positions to distinguish it from the actual observations, denoted \mathbf{x}^o , which represent a single realisation.

To obtain a single general expression for \mathbf{X}^o , the two-dimensional position and error vectors for all time-steps along the trajectory are stacked to obtain a single vector for each, denoted by the vector arrow. Using the stacked notation, the model is written as

$$\vec{\mathbf{X}}^o(\boldsymbol{\alpha}) = \vec{\mathbf{x}}^m(\boldsymbol{\alpha}) + \Gamma(\boldsymbol{\alpha})\vec{\mathbf{u}}^s + \vec{\boldsymbol{\epsilon}}^o, \quad (3.17)$$

where multiplication by the matrix $\Gamma(\alpha)$ is equivalent to using (3.11) to obtain ϵ^x . For example, assuming the first position is known without error, and therefore is not included in \bar{X}^o , the appropriate form of Γ for a trajectory with five observed positions is

$$\Gamma = \begin{bmatrix} \mathbf{I} & 0 & 0 & 0 \\ \gamma_1 & \mathbf{I} & 0 & 0 \\ \gamma_2\gamma_1 & \gamma_2 & \mathbf{I} & 0 \\ \gamma_3\gamma_2\gamma_1 & \gamma_3\gamma_2 & \gamma_3 & \mathbf{I} \end{bmatrix} \Delta t, \quad (3.18)$$

where the observations are assumed to occur at each time-step. From (3.17) and (3.18), it is clear that while the dependence of the observed position on the errors has been linearised, the dependence on the controls may be nonlinear. Since the observation and velocity errors are assumed Gaussian with zero mean, the linear dependence on the errors allows the distribution of \bar{X}^o to be written as

$$\bar{X}^o \sim N(\bar{x}^m, \Gamma \bar{\Sigma}^u \Gamma^T + \bar{\Sigma}^o). \quad (3.19)$$

The covariance matrices $\bar{\Sigma}^u$ and $\bar{\Sigma}^o$ correspond with the stacked vectors \bar{u}^s and $\bar{\epsilon}^o$, respectively. From (3.17), it is also clear that the errors due to u^s are correlated between observation times (due to the off-diagonal elements of Γ), whereas the observation errors are not.

The model in (3.16) can be used to help understand the effect of first-differencing the observed drifter positions. For simplicity, it is assumed that only a single model time-step is required between observation times. Then by first-differencing (3.16), one obtains

$$X_{n+1}^o - X_n^o = x_{n+1}^m - x_n^m + \epsilon_{n+1}^x - \epsilon_n^x + \epsilon_{n+1}^o - \epsilon_n^o, \quad (3.20)$$

where x^m is the full model trajectory. Using (3.3), (3.11) and (3.12) this can be rewritten as

$$\frac{X_{n+1}^o - X_n^o}{\Delta t} = u_n^m(x_n^m) + \left(\frac{\partial u_n^m}{\partial x} \right)^T \epsilon_n^x + u_n^s + \frac{\epsilon_{n+1}^o - \epsilon_n^o}{\Delta t}. \quad (3.21)$$

This expression shows that the model counterpart to the first-differenced trajectory is the model velocity evaluated along the modelled trajectory. The second term appears because of the possible difference in \mathbf{u}^m when evaluated along the true and modelled trajectories. Also, the observation error is now correlated between adjacent first-differenced positions which also include either ϵ_n^o or ϵ_{n+1}^o .

The same result is obtained by multiplying (3.17) by

$$\Gamma^{-1} = \begin{bmatrix} \mathbf{I} & 0 & 0 & 0 \\ -\gamma_1 & \mathbf{I} & 0 & 0 \\ 0 & -\gamma_2 & \mathbf{I} & 0 \\ 0 & 0 & -\gamma_3 & \mathbf{I} \end{bmatrix} (\Delta t)^{-1} \quad (3.22)$$

and rearranging terms. It can be shown that the estimator for α obtained using (3.17) is unaffected by any such nonsingular transformation. Therefore, it is equivalent to assimilate the observations as a full trajectory or as velocities after first-differencing.

In the special case that the observation error is sufficiently small relative to \mathbf{u}^s that it can be neglected (i.e. $|\epsilon^o| \ll |\epsilon^x|$), the model velocity in (3.21) can instead be evaluated at the observed position. This is equivalent to resetting the model trajectory to each observed (true) position, that is,

$$\mathbf{x}_{n+1}^m = \mathbf{x}_n^o + \mathbf{u}_n^m(\mathbf{x}_n^o)\Delta t. \quad (3.23)$$

As a consequence, ϵ_n^x is zero in (3.21) leading to

$$\frac{\mathbf{X}_{n+1}^o - \mathbf{x}_n^o}{\Delta t} = \mathbf{u}_n^m(\mathbf{x}_n^o) + \mathbf{u}_n^s. \quad (3.24)$$

It is clear that, under the assumption of serially uncorrelated \mathbf{u}^s , the error is now uncorrelated between these ‘‘observed’’ velocities. In the case of correlated \mathbf{u}^s only the covariance matrix Σ^u and the serial correlation, which are independent of α , are required to assimilate the first-differenced trajectory. Also (3.24) shows that a model trajectory is no longer required since the model counterpart is simply the model velocity at the observed (true) drifter positions.

If the large scale flow field varies linearly through space, the velocities are correctly modelled whether or not the modelled trajectory is reset to the observed positions. This can be seen by substituting the model velocity at the true position into (3.21) using the first order Taylor series expansion

$$\mathbf{u}^m(\mathbf{x}^t) = \mathbf{u}^m(\mathbf{x}^o) = \mathbf{u}^m(\mathbf{x}^m) + \left(\frac{\partial \mathbf{u}^m}{\partial \mathbf{x}} \right)^T \boldsymbol{\epsilon}^x, \quad (3.25)$$

where the observation error is still assumed to be negligible and therefore $\boldsymbol{\epsilon}^x$ is known at the observation times. Due to nonlinear spatial variation in the model velocity field, however, the approaches will generally not use equivalent velocities when $\boldsymbol{\epsilon}^x$ becomes sufficiently large such that (3.25) no longer holds.

If the observation error can not be neglected, the error covariance matrix $\boldsymbol{\Sigma}^x$ is not reset to zero when the model trajectory is reset to an observed position. Instead, $\boldsymbol{\Sigma}^x$ is set to the observation error covariance because this now represents the uncertainty in the initial position of the trajectory. Also, the observation error contributes not only to the diagonal of the total covariance matrix (i.e. corresponding to $\boldsymbol{\epsilon}^o + \text{epsilon}onbf{xnb}$), but also to a band of off-diagonal blocks. This can be seen by considering the form of (3.16) appropriate for a model trajectory reset to the n th observation:

$$\mathbf{X}_{n+1}^o = \mathbf{x}_n^o + \mathbf{u}_n^m(\mathbf{x}_n^o)\Delta t + \mathbf{u}_n^s\Delta t - \gamma_n\boldsymbol{\epsilon}_n^o + \boldsymbol{\epsilon}_{n+1}^o, \quad (3.26)$$

where the positions \mathbf{X}_n^o and \mathbf{X}_{n+1}^o would both contain $\boldsymbol{\epsilon}_n^o$ and therefore be correlated. The diagonal and off-diagonal blocks of the covariance matrix of the total error, denoted $\boldsymbol{\Sigma}^{tot}$, are

$$\boldsymbol{\Sigma}_{n+1}^{tot} = \gamma_n\boldsymbol{\Sigma}_n^o\gamma_n^T + \boldsymbol{\Sigma}_{n+1}^o + \boldsymbol{\Sigma}_n^u(\Delta t)^2 \quad (3.27)$$

$$\boldsymbol{\Sigma}_{n,n+1}^{tot} = -\boldsymbol{\Sigma}_n^o\gamma_n^T. \quad (3.28)$$

The results given above for the case of a single time-step between the observations generalises to cases when multiple time-steps are required. The model trajectory is reset to each observed position, producing a set of model sub-trajectories. Between

observation times, the trajectory and error covariance are simply propagated according to (3.3) and (3.13), respectively. The accumulated errors in position due to \mathbf{u}^s are still uncorrelated between observation times, for serially uncorrelated \mathbf{u}^s , and the observation error causes the modelled random variable \mathbf{X}^o at adjacent observation times to be correlated. This is again contrary to the case of modelling the entire trajectory, for which the contribution to \mathbf{X}^o from observation error at adjacent observation times is uncorrelated, but the errors due to \mathbf{u}^s are correlated.

3.4 Estimating the Model Parameters

This section describes how maximum likelihood estimation (MLE, introduced in chapter 1) can be used to obtain an estimate of α from an observed trajectory, \mathbf{x}^o , along with specified statistics for the random errors ϵ^o and \mathbf{u}^s . The most direct method of addressing this problem, as implemented by *Griffin and Thompson (1996)*, is to calculate “observed” velocities by first-differencing the observed drifter positions that are adjacent in time. However, a more general approach can be taken using the stochastic model (3.17).

3.4.1 Estimation Problem

To illustrate how MLE is applied to the assimilation of drifter trajectories, consider the general case of estimating the controls using observed positions and the stochastic model (3.17). Given the full covariance matrices of the errors $\bar{\mathbf{u}}^s$ and $\bar{\epsilon}^o$ and a first guess for α , the joint distribution of the stacked vector $\bar{\mathbf{X}}^o$ can be obtained as given by (3.19). This pdf, evaluated at the observed drifter positions, $\bar{\mathbf{x}}^o$, is the required likelihood function, defined as

$$f_{\bar{\mathbf{x}}^o}(\bar{\mathbf{x}}^o, \alpha). \quad (3.29)$$

Since the pdf is assumed to be Gaussian, the problem can be simplified by minimising the following cost function:

$$J_d(\boldsymbol{\alpha}) = \frac{1}{2} [\bar{\mathbf{x}}^o - \bar{\mathbf{x}}^m(\boldsymbol{\alpha})]^T [\boldsymbol{\Sigma}^{tot}(\boldsymbol{\alpha})]^{-1} [\bar{\mathbf{x}}^o - \bar{\mathbf{x}}^m(\boldsymbol{\alpha})] + \log \left[(2\pi)^N |\boldsymbol{\Sigma}^{tot}(\boldsymbol{\alpha})|^{1/2} \right] , \quad (3.30)$$

where N is the number of observed drifter positions and $\boldsymbol{\Sigma}^{tot}$ is the covariance matrix of the total error $\boldsymbol{\Gamma}(\boldsymbol{\alpha})\bar{\mathbf{u}}^s + \bar{\boldsymbol{\epsilon}}^o$. The second term in (3.30) appears due to the dependence of $\boldsymbol{\Sigma}^{tot}$ on $\boldsymbol{\alpha}$. The subscript d indicates that this cost function is for drifter data. The $\boldsymbol{\alpha}$ that minimises this function is the maximum likelihood estimator for the controls, denoted $\hat{\boldsymbol{\alpha}}$. In practice, information from other observation types will usually be available and will contribute additional terms to the total cost function (e.g. *Morrow and DeMey, 1995*).

As described in Chapter 1, some independent information may be supplied in the form of an *a priori* estimate for the value of $\boldsymbol{\alpha}$, denoted $\boldsymbol{\alpha}_o$. If the errors in $\boldsymbol{\alpha}_o$ are Gaussian, with covariance matrix $\boldsymbol{\Sigma}^a$, then this prior estimate can be incorporated into the assimilation problem by adding the following term to J_d :

$$J_a(\boldsymbol{\alpha}) = \frac{1}{2} (\boldsymbol{\alpha} - \boldsymbol{\alpha}_o)^T (\boldsymbol{\Sigma}^a)^{-1} (\boldsymbol{\alpha} - \boldsymbol{\alpha}_o). \quad (3.31)$$

3.4.2 Calculating the Optimal Estimate

The adjoint method outlined in the first chapter can be used to efficiently find the controls that minimise the cost function. To apply this method, the adjoint of the ocean model must be obtained to provide the sensitivity (that is, the derivatives) of the model velocities with respect to the controls. The adjoint of the drifter model is also required to provide the sensitivity of the model drifter positions with respect to each of the model velocities used for the advection. These two adjoint models together, through the chain rule, provide the sensitivity of the model drifter positions with respect to the controls. This information allows the gradient of the cost function to be calculated and consequently the value of $\hat{\boldsymbol{\alpha}}$ to be efficiently found through the

use of an iterative optimisation algorithm. Since the model is likely nonlinear, the derivatives and also Σ^{tot} must be recalculated at each iteration of the minimisation.

Alternatively, if the model consists of a small number of controls, the Gauss-Newton method (see e.g. *Shumway*, 1988) can be employed to minimise the cost function. This procedure involves solving a series of linearised estimation problems until the solution converges to the nonlinear solution. The ocean/drifter model is linearised with respect to the optimal controls at the j th iteration

$$\bar{\mathbf{X}}^o(\boldsymbol{\alpha}_{j+1}) = \bar{\mathbf{x}}^m(\hat{\boldsymbol{\alpha}}_j) + \mathcal{A}_j(\boldsymbol{\alpha}_{j+1} - \hat{\boldsymbol{\alpha}}_j) + \bar{\boldsymbol{\epsilon}}^{tot}, \quad (3.32)$$

where \mathcal{A}_j is the linearised model. If the number of controls is sufficiently small, the linearisation may be obtained numerically by performing a series of ocean/drifter model integrations, each with a different control variable perturbed from the reference value. The linearised estimation problem at iteration $(j+1)$ is solved using generalised regression (see Chapter 1)

$$\hat{\boldsymbol{\alpha}}_{j+1} = \hat{\boldsymbol{\alpha}}_j + \mathcal{B}_j[\bar{\mathbf{x}}^o - \bar{\mathbf{x}}^m(\hat{\boldsymbol{\alpha}}_j)], \quad (3.33)$$

where

$$\mathcal{B}_j = [\mathcal{A}_j^T (\Sigma_j^{tot})^{-1} \mathcal{A}_j]^{-1} \mathcal{A}_j^T (\Sigma_j^{tot})^{-1}. \quad (3.34)$$

Since the covariance matrix of the total error depends on the controls, Σ_j^{tot} is also calculated using $\hat{\boldsymbol{\alpha}}_j$. Because the problem is non-linear, the linearisation and regression must be repeated with respect to the previous optimal solution (j is incremented by one) until convergence.

As in linear regression, a linear relationship between the controls and the model counterparts to the observations results in a cost function that is quadratic with respect to the controls. This guarantees convergence to the globally optimal solution, independent of the first-guess for the controls. Conversely, if the relationship is non-linear the cost function may be non-quadratic, possibly with multiple minima and maxima. In this case, for either assimilation scheme to converge to the global minimum, the initial guess for the controls must be sufficiently close to the global solution

such that the cost function gradient is directed towards this solution and not towards a secondary minimum.

3.4.3 Uncertainty in the Estimate

It is often desirable to quantify the uncertainty in the estimated controls by calculating the covariance matrix of their errors. In cases where the estimation procedure converges, linear theory can be applied to analytically calculate the error covariance matrix of the estimated controls, even when the *a priori* error statistics are mis-specified. The following expression gives the true error covariance matrix of the estimated controls, even when the prior error statistics used for the estimation are not equal to the true statistics

$$\Sigma^{\hat{a}} = \mathbf{B}_{\infty} \Sigma^{tot} \mathbf{B}_{\infty}^T. \quad (3.35)$$

The matrix \mathbf{B}_{∞} is obtained using the linearised model and *a priori* error covariance matrix evaluated using (3.34) at the converged estimate $\hat{\alpha}_{\infty}$. The true (unknown) error covariance matrix is denoted by Σ^{tot} . When the covariance matrix and the linearised model used for the estimation correspond to the true ones, (3.35) simplifies to

$$\Sigma^{\hat{a}} = \left[\mathcal{A}^T (\Sigma^{tot})^{-1} \mathcal{A} \right]^{-1}. \quad (3.36)$$

In real applications, Σ^{tot} is unknown. In idealised experiments, however, such as those presented later in this chapter, (3.35) can be used to predict the increase in the estimation error due to mis-specification of the *a priori* error statistics.

3.5 Practical Issues for Assimilating Trajectories

In this section some important issues are examined that arise when assimilating drifter trajectory data with the models and approaches presented earlier. These issues are illustrated through idealised experiments using the low-dimensional model of tidally

driven barotropic flow over a bank presented in the previous chapter. First, the impact of nonlinearity in the relationship between the model controls and the trajectory positions is discussed. Then, in an attempt to reduce the effects of this nonlinearity, the trajectory is modelled as a set of sub-trajectories. Finally, the impact of mis-specifying the *a priori* error statistics is discussed and a method for diagnosing mis-specified statistics proposed.

3.5.1 Experimental Setup

In all the following numerical examples the true value for the control, U_∞ , is equal to $5.0 \times 10^{-3} \text{ m s}^{-1}$. A single trajectory, four tidal periods in length, and originating near the top of the bank at the time of maximum westward tidal current is used.

The data consists of eight observed positions that are equally separated in time, including the initial location (open circles in Figure 3.2). In much of the following discussion the observations are assumed to be error free (as in Figure 3.2). For the discussion of mis-specified error statistics (Sections 3.5.4 to 3.5.6), however, a non-negligible observation error is introduced. This observation error has a standard deviation $\sigma^o = 500 \text{ m}$ and the covariance matrix is of the form

$$\Sigma^o = \begin{bmatrix} \sigma^o & 0 \\ 0 & \sigma^o \end{bmatrix}^2. \quad (3.37)$$

It is assumed, however, that the initial position of the drifter is known without error, since this may be the location of drifter deployment.

The covariance matrix for the velocity error, \mathbf{u}^s , is assumed to be of the form

$$\Sigma^u = \begin{bmatrix} \sigma^u & 0 \\ 0 & \sigma^u \end{bmatrix}^2. \quad (3.38)$$

The standard deviation for the components of \mathbf{u}^s is 0.05 m s^{-1} with the time-step equal to 450 s. With no velocity gradients (that is, $\boldsymbol{\gamma} = \mathbf{I}$) this corresponds to an error in the modelled drifter position with standard deviation of 360 m at the time of the second observation. The velocity errors are assumed to be uncorrelated along the trajectory.

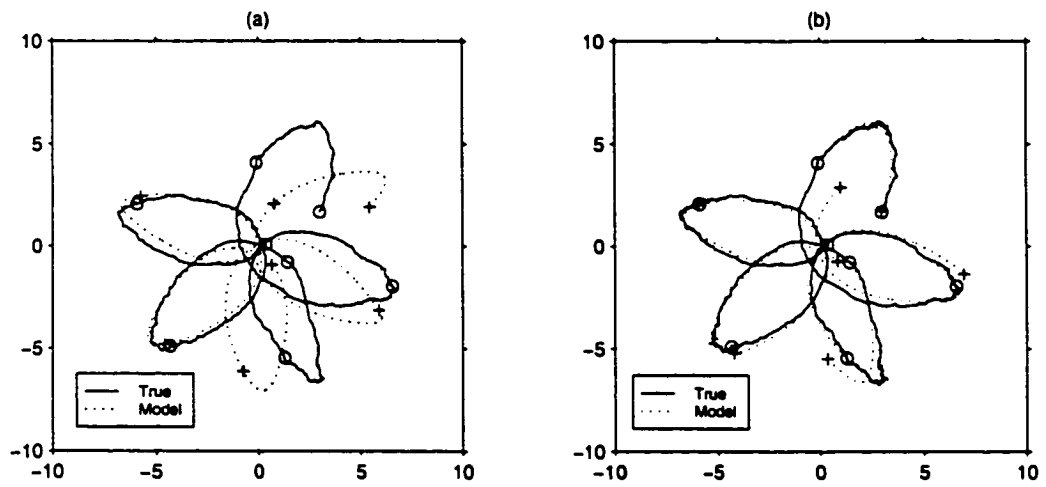


Figure 3.2: Plots of true and modelled trajectories as (a) full model trajectory and (b) model sub-trajectories that are reset to the observations. The observations (o) and corresponding locations along the modelled trajectory (+) are also shown. The observations are error free and the unresolved scales of motion account for the divergence between the two trajectories.

Figure 3.2a shows the true and modelled trajectories, both starting from the same location and employing the same large scale flow field. The difference between the two trajectories is due solely to the effect of a randomly generated realisation of \mathbf{u}^s that is included when simulating the true trajectory. Figure 3.2b shows the set of sub-trajectories obtained by resetting the trajectory to each perfectly observed position.

Figure 3.3 shows the observed drifter positions produced from a large number of randomly generated realisations of the small scale velocity component. Note how the actual distribution of locations becomes spread approximately along depth contours.

3.5.2 Nonlinearity of Ocean/Drifter Model

When assimilating drifter trajectories, nonlinearity in the ocean/drifter model can have important consequences. One source of nonlinearity is the relationship between the controls and the model velocity field due to nonlinearity in the ocean model. Another important source is the spatial variation of the partial derivatives $\partial \mathbf{u}^m / \partial \mathbf{x}$ and $\partial \mathbf{u}^m / \partial \alpha$. To illustrate the latter, consider the derivative of \mathbf{x}_{n+1}^m with respect to α using (3.3)

$$\begin{aligned} \frac{d\mathbf{x}_{n+1}^m}{d\alpha} &= \frac{d\mathbf{x}_n^m}{d\alpha} \left(\mathbf{I} + \frac{\partial \mathbf{u}_n^m}{\partial \mathbf{x}} \Delta t \right) + \frac{\partial \mathbf{u}_n^m}{\partial \alpha} \Delta t \\ &= \frac{d\mathbf{x}_n^m}{d\alpha} \boldsymbol{\gamma}_n^T + \frac{\partial \mathbf{u}_n^m}{\partial \alpha} \Delta t. \end{aligned} \tag{3.39}$$

If any term in (3.39) depends on α , then the model trajectory position may be a nonlinear function of α . Nonlinearity from the ocean model would cause the partial derivative $\partial \mathbf{u}_n^m / \partial \alpha$ to depend on α at a given location. However, the partial derivatives in (3.39) are evaluated along the model trajectory which may, in turn, depend on α . Therefore, spatial variation of the partial derivatives $\partial \mathbf{u}_n^m / \partial \alpha$ and $\partial \mathbf{u}_n^m / \partial \mathbf{x}$ causes these terms to vary when evaluated along the model trajectories for different values of α . In general, a given pair of model trajectories will become increasingly separated as their length increases and, as a consequence, the partial derivatives become more

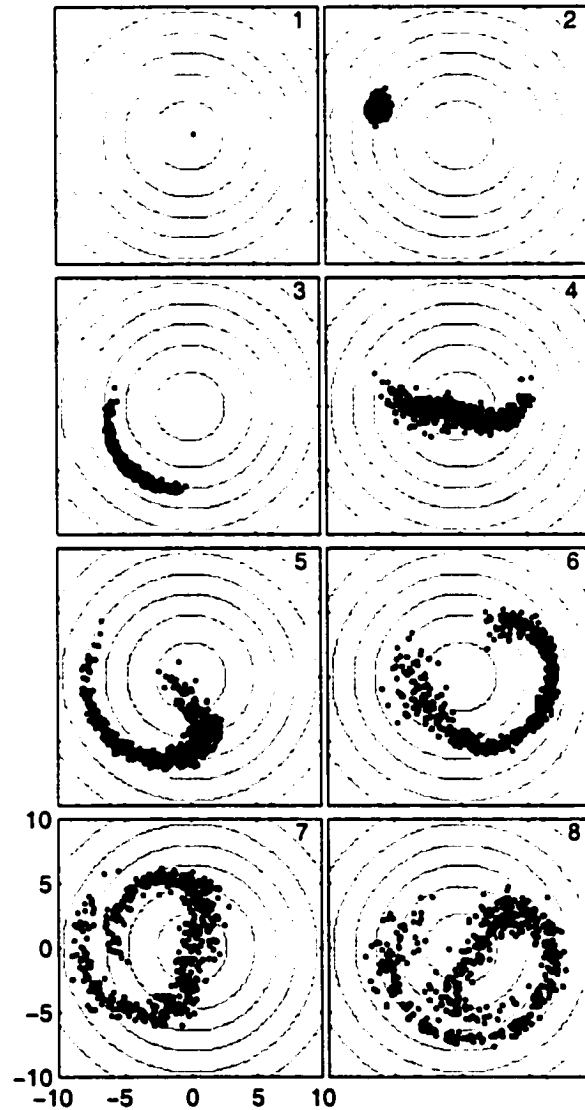


Figure 3.3: Distributions of drifter location at the eight observation times. Each of the 500 points in the eight frames was calculated by advecting a drifter through a flow field composed of the true large scale field derived from the simplified model and a randomly generated realisation of the small scale field at each time-step.

sensitive to changes in α . Therefore, over a given range of values for the controls this source of nonlinearity will tend to be more important for longer trajectories.

To illustrate, Figure 2.5 shows the dependence of the final trajectory position on U_∞ for two trajectory lengths. Note that the longer trajectory is more nonlinear for high values of U_∞ . This is consistent with nonlinearity originating from spatial variation in the derivatives in (3.39). As discussed in the previous chapter, the nonlinearity for low values of U_∞ is due to the nonlinearity of the ocean model (associated with a dependence of $\partial \mathbf{u}_n^m / \partial \alpha$ on α) that occurs at near-resonant forcing. This source of nonlinearity is unrelated to the trajectory length.

Nonlinearity between the model trajectory and the controls results in a non-quadratic cost function. As discussed earlier, this can lead to difficulties in obtaining the optimal estimate.

Another important consequence for the estimation problem occurs when the separation between the true and optimal model trajectories is sufficiently large that the spatial variation in \mathbf{u}^m is nonlinear between them. In this case, the linearisation of \mathbf{u}^m that lead to (3.13) is not valid. Therefore Σ^x is not correctly calculated and, as a consequence, the value of the cost function is incorrect. These errors in the calculation of Σ^x will accumulate along the trajectory.

Figure 3.4 shows a set of ellipses representing the covariance matrix of ϵ^x calculated using (3.13) along both the true trajectory and the model trajectory evaluated using the true values for the control. The ellipses in Figure 3.4 are shown at the times of the observations. The observations (circles) and model counterparts (plus signs) are indicated along the trajectories. Note how the error covariances along the true and model trajectories eventually diverge and become quite dissimilar as the trajectories become increasingly separated. The ellipses at the final two times undergo a completely opposite transformation: along the true trajectory the ellipse appears to rotate clockwise, whereas along the model trajectory the rotation is counterclockwise. This represents a mis-specification of the error statistics that occurs because the assumption that the errors remain small compared to the spatial variation in the

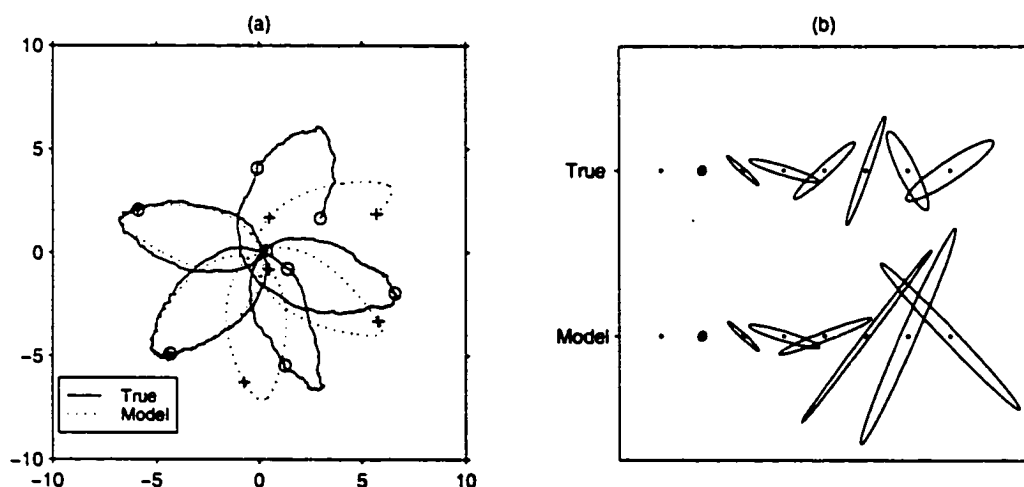


Figure 3.4: (a) True drifter trajectory (solid) and the model trajectory (dotted) corresponding to the true value for the control, $U_\infty = 5.0 \times 10^{-3} \text{ m s}^{-1}$. Observations (circle) and model counterparts (plus sign) are shown. Note how the trajectories diverge solely due to neglecting the small scale component of the velocity field when calculating the model trajectory (which also explains why the model trajectory appears smoother than the true trajectory). (b) Gaussian error ellipses corresponding to the covariance matrix Σ^x calculated along both trajectories using (3.13). Note that they eventually become very different due to differences in the matrix γ along the two trajectories.

model flow field has been violated. Consequently, the actual distribution of the error in the drifter position is no longer Gaussian and, therefore, simply using covariances is not sufficient to model the errors. This can be seen in Figure 3.3 where only the distribution at the second observation time appears truly Gaussian. By the fifth observation time, it appears that approximation by a Gaussian distribution can no longer represent the true distribution.

3.5.3 Assimilation of Sub-trajectories

The problems of multiple minima in the cost function and the departure from a Gaussian error distribution result in part from using a model-derived counterpart for

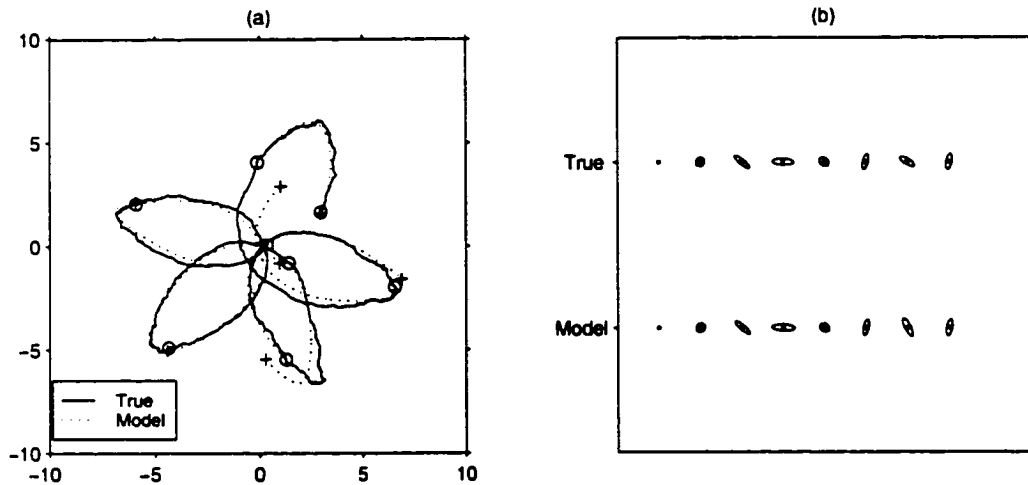


Figure 3.5: Same as Figure 3.4, except model trajectory is reset to each observed position. Note how trajectories remain closer together and the evolution of the error in the model trajectory is very similar along the two trajectories.

the entire trajectory. In an attempt to reduce these problems, the approach of using sub-trajectories, as described in Section 3.3.4, is examined.

Effect on error pdf

A set of ellipses similar to those in Figure 3.4, representing the evolution of the error covariance matrix of ϵ^x , was calculated using a set of model sub-trajectories (Figure 3.5). Since the observation error is assumed negligible, the error in the model trajectory is reset to zero at the time of each observation. Consequently, the error does not grow to the same magnitude as when using the full model trajectory. It is also apparent that the evolution of the ellipses is much more similar along the true and model trajectories. This implies that the statistics are nearly Gaussian, making the use of a simple quadratic cost function valid. These results would equally apply if the u^s were correlated along the trajectory.

To illustrate the effect of using sub-trajectories on the evolution of the error covariance matrix Σ^x , a measure of the variation in the spatial derivatives of the velocity

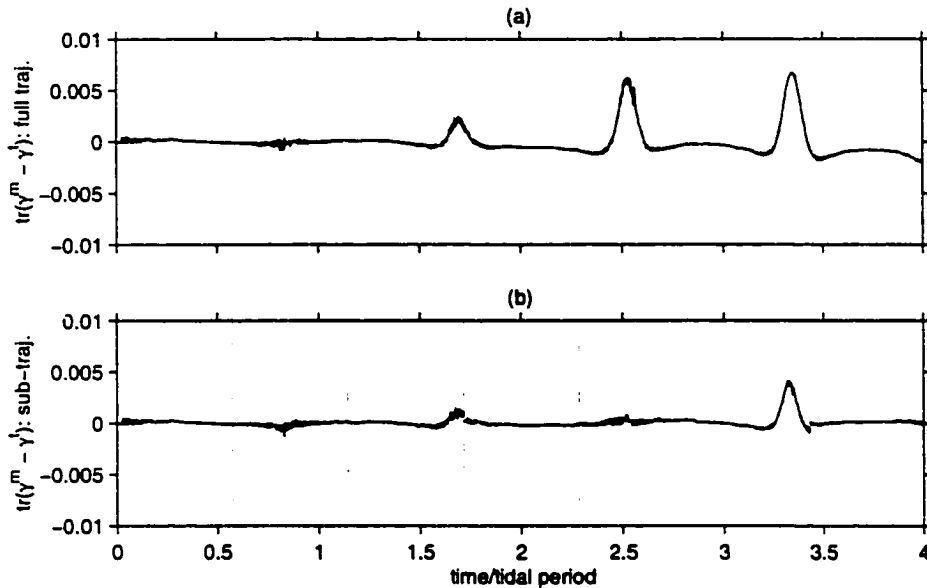


Figure 3.6: Measure of the difference in the spatial derivatives of the model velocity field (horizontal divergence) between the true and estimated model trajectories (trace of the difference in γ normalised by the trace of γ along the true trajectory) corresponding to (a) the full model trajectory from Figure 3.4 and, (b) the model trajectory that is periodically reset to each observed position from Figure 3.5 (sub-trajectories). Observation times are indicated in (b) by vertical dashed lines.

field was computed. The trace of the difference between the γ matrices corresponding to the true and model trajectories normalised by the trace of γ along the true trajectory is shown in Figure 3.6. This was calculated using both the full model trajectory (Figure 3.6a) and the model trajectory reset to the perfectly observed positions (Figure 3.6b). The figure shows that the γ matrices are consistently more similar when using sub-trajectories. The peaks in both graphs correspond to times when trajectories are over the top of the bank.

The horizontal divergence of the velocity field is a measure of overall growth (positive divergence) or decay (negative divergence) of the error along the model trajectory. Since the flow field is parameterised in terms of a transport streamfunction, horizontal divergence can only occur as a result of variation in the water depth along

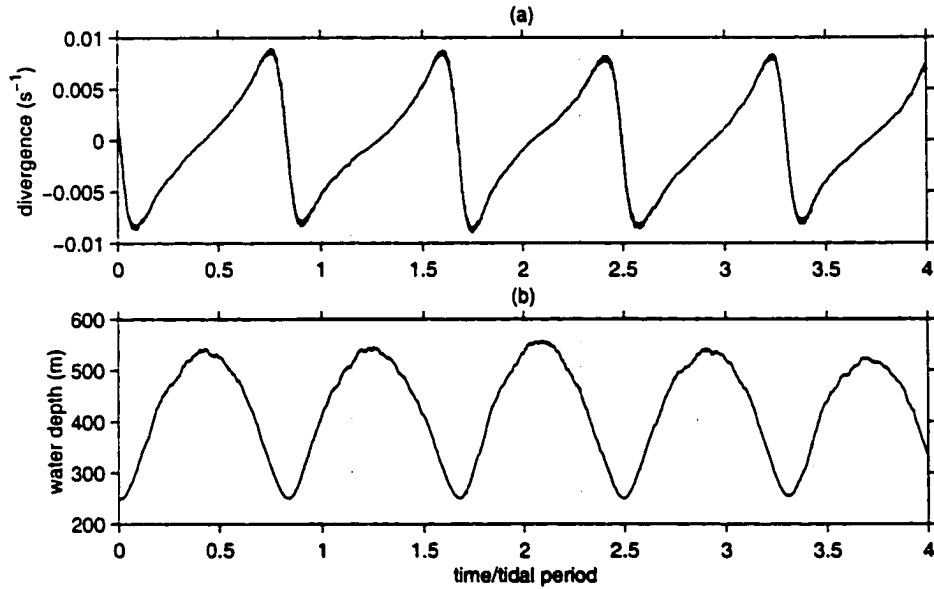


Figure 3.7: The (a) horizontal divergence and (b) water depth along the length of the true trajectory shown in Figure 3.5. Note how the magnitude of the divergence is greatest when on top of the bank, corresponding to the times of greatest growth in the difference of γ in Figure 3.6.

a streamline, given by

$$\nabla \cdot \mathbf{u} = -\frac{\mathbf{u} \cdot \nabla h}{h}. \quad (3.40)$$

Figure 3.7 shows the divergence and water depth along the true trajectory. The largest magnitudes of divergence correspond to the shallow areas on top of the bank, consistent with the inverse dependence on water depth in (3.40). Therefore these are the locations where errors are greatly amplified as the drifter is advected onto the bank. Conversely, errors are attenuated as the drifter moves into deeper water.

Effect on cost function

Still assuming negligible observation error, the cost function (3.30) was calculated for a single perfectly observed trajectory as a function of U_∞ using both the full model

trajectory and sub-trajectories (Figures 3.8a and b, respectively). Both cost functions exhibit a maximum near the value of U_∞ that corresponds to resonant forcing. This departure from a quadratic shape can be explained by the strong nonlinearity of the ocean model at resonance and is consistent with the nonlinear dependence of the prognostic model variables shown in Figure 2.2. For higher values of U_∞ , however, the cost function computed using the full model trajectory exhibits a second maximum, whereas when the model trajectory is reset to the observations, the shape is closer to a quadratic function. This maximum is due to the nonlinearity in the drifter model that is enhanced as the trajectory length is increased. As discussed above, the use of sub-trajectories reduces this nonlinearity by reducing the displacement in the model trajectory resulting from a given change in U_∞ . Hence, the cost function calculated using sub-trajectories has a more quadratic shape in this region. The global minimum for Figure 3.8a is located at $U_\infty = 4.75 \times 10^{-3} \text{ m s}^{-1}$ and for Figure 3.8b is located at $U_\infty = 5.1 \times 10^{-3} \text{ m s}^{-1}$.

Effect on estimator

Using (3.35), the standard error for the estimates of U_∞ were obtained. The true error covariance matrix, Σ^{tot} , is calculated by propagating the errors along the true trajectory. The results (Table 3.1) are only an indication of the actual error in the estimator since the errors along the full trajectory are non-Gaussian. They show that resetting the model trajectory to the observations reduces the standard deviation of the estimation error by a factor of almost three. This is mostly due to the fact that the error covariance matrices calculated along the model and true trajectories are more similar when sub-trajectories are used. This is seen by comparing Figures 3.4 with 3.5.

If the true trajectory were known prior to the estimation for the purpose of linearising the model and propagating the *a priori* error covariance matrix Σ^* (that is, $\Sigma_\infty^{tot} = \Sigma^{tot}$), the standard error would be as shown in the second column of Table 3.1. Using the true trajectory in this way is artificially equivalent to the case where

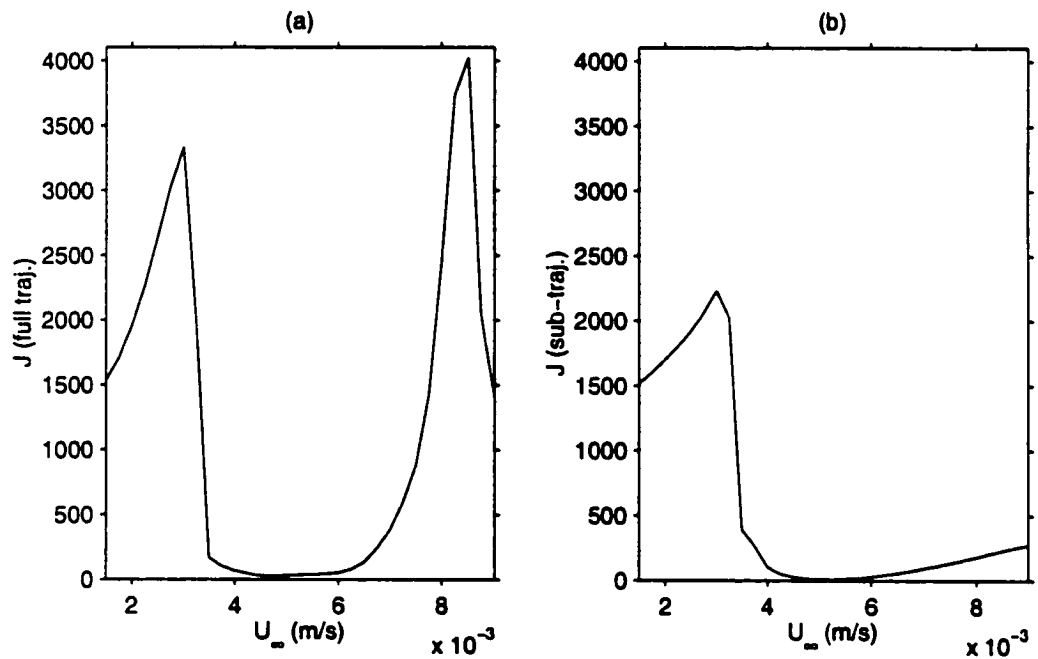


Figure 3.8: The cost function for a single perfectly observed drifter trajectory computed using: (a) the full model trajectory or (b) model sub-trajectories. Note how the local maximum around $U_{\infty} = 8 \times 10^{-3} \text{ m s}^{-1}$ is removed by using sub-trajectories.

Table 3.1: Standard error of the estimated value of U_∞ when the full model trajectory (first row) and model sub-trajectories (second row) are used. In the second column are the results obtained using the true trajectory for propagating the errors and linearising the model, that is, when $\Sigma_\infty^{tot} = \Sigma^{tot}$.

	Trajectory used to calculate Σ_∞^{tot} :	
	Model trajectory	True trajectory
Full model trajectory	4.30×10^{-4}	1.38×10^{-4}
Model sub-trajectories	1.62×10^{-4}	1.38×10^{-4}

$\partial \mathbf{u}_n^m / \partial \mathbf{x}$ (and γ) is spatially invariant. Therefore, the equivalence of the estimation standard error when full and sub-trajectories are used is consistent with the equivalence of the two approaches when \mathbf{u}^m varies linearly through space, discussed earlier. This value represents the theoretical minimum for the standard error that would be obtained with perfectly specified error statistics. From these results, it appears that most of the increase in estimation error due to mis-specification of the error statistics can be eliminated by using sub-trajectories.

In summary, the use of sub-trajectories reduces the problems associated with the nonlinearity of the drifter model by reducing the separation between the true and modelled trajectories. This conclusion is valid only when the observation error is small relative to the error in the modelled trajectory accumulated between subsequent observations. Other cases are considered in Section 3.6.

3.5.4 Mis-specification of Observation Error Statistics

In the cases discussed above, it was assumed that the observation error could be neglected. Under this assumption, it was shown that using the full model trajectory could lead to a mis-specification of the error statistics due to nonlinear spatial variation of \mathbf{u}^m , even if the statistics of the velocity errors were correctly specified. In this and the following sections, the effect of deliberately mis-specifying the observation or velocity error statistics, when neither is actually negligible, is evaluated. The true

variance for the small scale velocity field and the observation error are as described in Section 3.5.1. The observation error must now be taken into account when the model trajectory is reset to the observations. This causes the total errors at adjacent observation times to be correlated when assimilating sub-trajectories, as described in Section 3.3.4.

Figure 3.9 shows the standard error of the estimated U_∞ as σ^o is varied, calculated using (3.35). The dependence is shown for the case using the full model trajectory (Figure 3.9a) and sub-trajectories (Figure 3.9b). The error standard deviations were also calculated assuming the true trajectory and model statistics were known for the estimation (dashed line). The results for the full model trajectory again only give an indication of the actual distribution of estimation error since higher order moments are required for the error in the modelled trajectory (due to its non-Gaussian distribution shown in Figure 3.3).

In both cases the minimum variance does not correspond exactly with the true observation error standard deviation. Within the entire range of the assumed observation error standard deviation, however, use of sub-trajectories gives a lower standard error than when using the full model trajectory. The lower estimation error obtained by assimilating sub-trajectories is due to the reduced separation between the true and modelled trajectories when using sub-trajectories. This occurs even though the observations do not lie perfectly on the true trajectory because the error in the full model trajectory quickly becomes greater than the observation error.

3.5.5 Mis-specification of Model Error Statistics

The effect of mis-specifying the statistics of the small scale velocity field, σ^u , was also evaluated (Figure 3.10). Similar to the case of mis-specified observation error statistics, the minimum standard error for U_∞ is not located at the true value for σ^u . Also, it appears that neglecting model error completely when it is present has a large impact on errors of the estimated model parameter: when using a full model trajectory, the standard error is $18.0 \times 10^{-4} \text{ m s}^{-1}$ and with the model trajectory reset

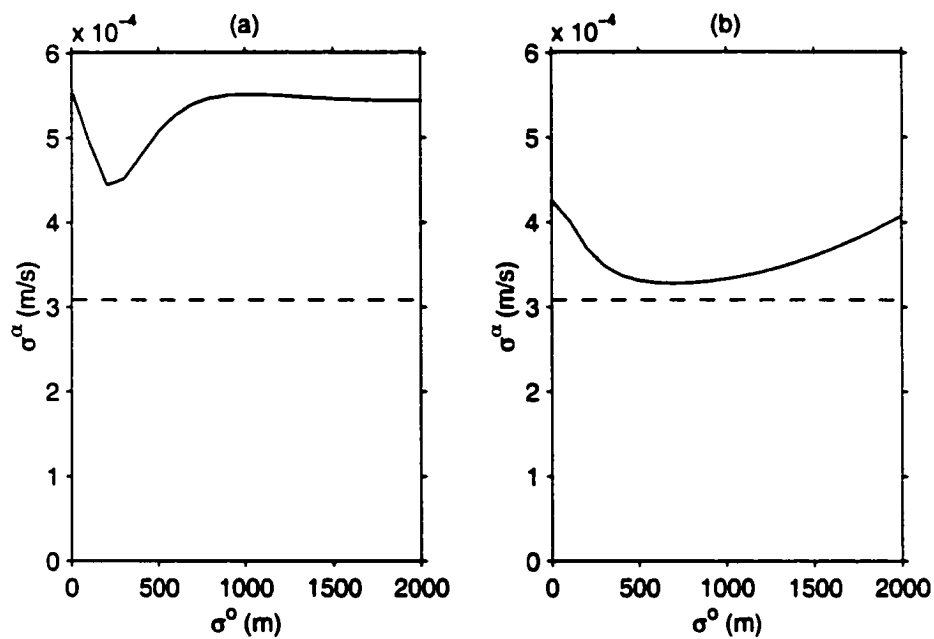


Figure 3.9: Standard error of the estimated value of U_∞ as a function of the *a priori* assumed observation error standard deviation (solid). Error standard deviation also shown when true value of σ^o is used (500 m) and the error Σ^x is evaluated along true trajectory (dashed). Results shown for (a) the full model trajectory and (b) the model sub-trajectories.

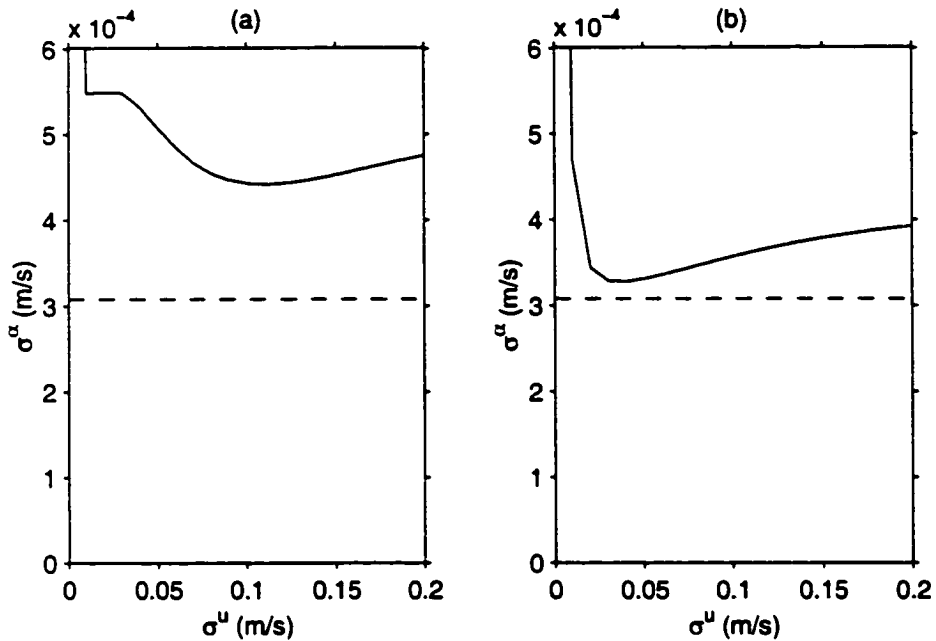


Figure 3.10: Similar to Figure 3.9, but for dependence on standard deviation of the unresolved small scale velocities, σ^u . True value of σ^u is 0.05 m s^{-1} .

to the observations, the standard error is $30.0 \times 10^{-4} \text{ m s}^{-1}$.

As already noted, for the case of assimilating the full model trajectory, the accumulated error in position due to \mathbf{u}^* is correlated between observation times, whereas the observation error is uncorrelated. Conversely, for sub-trajectories the observation error is correlated between observation times according to (3.28) and the model error is uncorrelated (assuming \mathbf{u}^* is serially uncorrelated). With this in mind, comparison of Figures 3.9 and 3.10 shows that the standard error of U_∞ is more sensitive to mis-specification of the uncorrelated errors (that is, it is more sensitive to mis-specifying observation error for full trajectories and model error for sub-trajectories). However, the use of sub-trajectories still gives a lower standard error over almost the entire range of σ^u and σ^ρ for which it was calculated.

3.5.6 Diagnosing Mis-specified Error Statistics

In the field of statistical estimation, it is common to analyse the residuals between the observations and the estimated model counterparts to determine the validity of the *a priori* statistics used in the estimation (Talagrand, 1998; Menard et al., 1999). In the approach presented here for assimilating observations of drifter trajectories into an ocean model, the *a priori* specified parameters that determine the error statistics are σ^u and σ^o . A method for checking the consistency of these specified standard deviations with the residuals would be of significant value.

One approach for checking the consistency of the *a priori* error statistics is to calculate the squared statistical distance between the observations and the estimated trajectory (Menard et al., 1999)

$$(\bar{\mathbf{x}}^o - \bar{\mathbf{x}}^m)^T (\Sigma^{tot})^{-1} (\bar{\mathbf{x}}^o - \bar{\mathbf{x}}^m), \quad (3.41)$$

which is similar to the cost function, J_d . Assuming Gaussian errors and Σ^{tot} is correctly specified, the squared statistical distance has a chi-square probability distribution with $N_y - N_\alpha$ degrees of freedom, where N_y is the number of independent observations and N_α is the number of independent controls. Therefore, a simple test can be made to determine if the calculated quantity is consistent with this distribution. However, the value of (3.41) is sensitive to an arbitrary scaling of the total error covariance matrix. Because such a scaling is equivalent to applying a constant scaling and offset to J_d , the estimated values and their error variance are unaffected. Therefore, this diagnostic can not be used to determine when either σ^u or σ^o are mis-specified.

Alternatively, the squared statistical distance between the observed and the predicted positions can be calculated at each observation time using the corresponding diagonal block of the *a priori* error covariance model as the norm. The squared distance at the time of the n th observation is

$$d_n^2 = (\mathbf{x}_n^o - \mathbf{x}_n^m)^T (\Sigma_n^{tot})^{-1} (\mathbf{x}_n^o - \mathbf{x}_n^m), \quad (3.42)$$

where Σ_n^{tot} is the n th 2×2 diagonal block of the total error covariance matrix. The variables d_n^2 are also approximately chi-square random variables with two degrees of freedom, but are not independent due to the off-diagonal blocks of the total covariance matrix.

Plots of d_n^2 using both the full and sub-trajectories may be useful tools in diagnosing the mis-specification of the error statistics. Ignoring the influence of the spatial gradients of the model velocity field (that is, $\gamma = \mathbf{I}$), the standard error of the modelled trajectory grows as $\sqrt{t} \sigma^u$. Therefore, for trajectories that are sufficiently long, the error due to the unresolved velocity field eventually dominates the diagonal blocks of the total error covariance matrix. Consequently, any obvious growth or attenuation of d_n^2 calculated from the full model trajectory should give an indication that σ^u is mis-specified. This is similar to the findings of *Menard et al.* (1999) for the case of assimilating atmospheric tracer constituents with a Kalman Filter. Conversely, use of sub-trajectories prevents the error along the model trajectory from accumulating, thus preventing it from dominating the diagonal blocks of the total covariance matrix. Instead, depending on the temporal separation of the observations, the observation error may dominate the diagonal blocks along the entire length of the trajectory. Consequently, mis-specification of observation error statistics should be apparent from plots of d_n^2 calculated using sub-trajectories.

Figure 3.11 shows d_n^2 along the full model trajectory (left panels) and the model trajectory that has been reset to the observations (right panels) where the observed positions were obtained using a random realisation of both the observation and velocity errors. For both cases they are plotted using the diagonal blocks of the true covariance matrix (where both observation and velocity error statistics are included; top panels) and also using a covariance model that neglects observation error (middle panels) and one that neglects velocity error (bottom panels). Using the true covariance matrix, the values of d_n^2 mostly remain below 5.991, the probability of which is 95% according to the chi-square distribution, for both the full and sub-trajectories (Figure 3.11a,d). This suggests that the residuals are consistent with this *a priori*

covariance matrix. When the observation error is ignored, the values of d_n^2 exceed the 95% level by a large amount, most noticeably in the case of using sub-trajectories (Figure 3.11e), as expected. The neglect of model error manifests itself as a growth in the d_n^2 along the full model trajectory that eventually exceeds the 95% level (Figure 3.11c). Therefore, it appears that the use of d_n^2 , as defined in (3.42), simultaneously applied to both the full and sub-trajectories may be useful as a diagnostic tool for determining when σ^u or σ^o have been incorrectly specified.

When a large amount of trajectory data is available, but the error statistics are unknown, an iterative approach may prove useful. First, some initial estimates for the error standard deviations are used to produce a set of “optimal” controls. Then, the approach just outlined for diagnosing the mis-specification of the error statistics is applied to the residuals to improve the error standard deviations. The process is iterated until the residuals become consistent with the *a priori* statistics. By applying this approach, statistical information on the small scale motions is obtained. Unlike previous approaches (as outlined in the introduction to this chapter), these motions are defined as deviations from a spatially and temporally varying “mean” flow given by the solution of an oceanographic model.

3.6 Discussion and Conclusions

In summary, a framework for assimilating drifter trajectories was examined in the context of MLE. It was found that the nonlinearity of $\mathbf{dx}/dt = \mathbf{u}(\mathbf{x})$ leads to two distinct problems. The first is that multiple minima appear in the cost function, thus making the estimation problem more difficult to solve. Also, the statistics of the errors in the modelled trajectory may be non-Gaussian, leading to mis-specified error statistics when only the first two statistical moments are calculated. The use of sub-trajectories (that is, model trajectories whose positions are continually reset to the observed positions) can reduce this problem by reducing the separation between the true and modelled trajectories. The standard error of the estimator was also shown

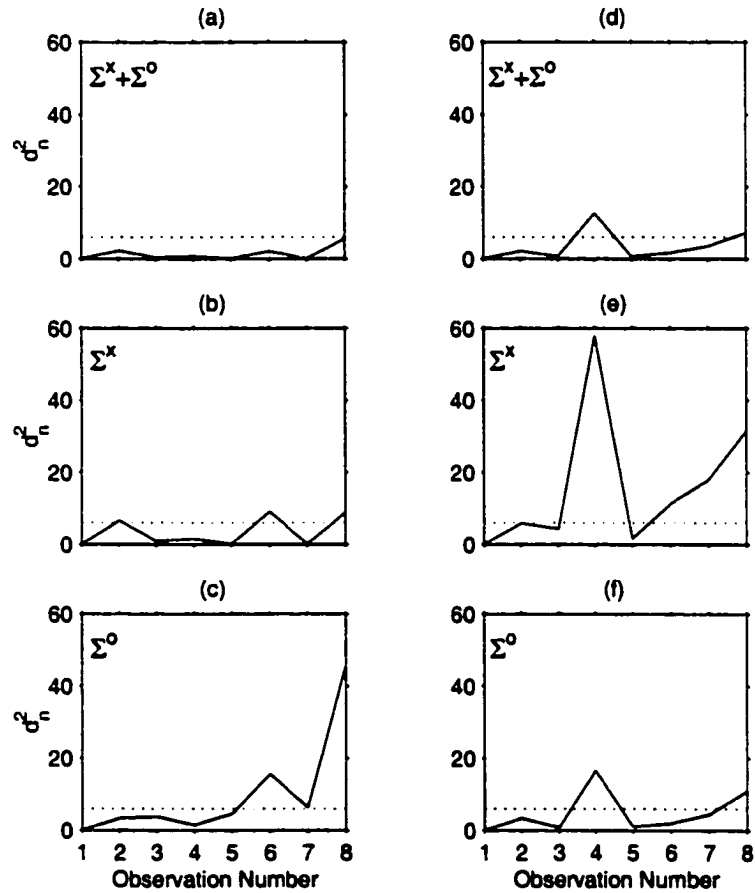


Figure 3.11: The squared statistical distance d_n^2 , as defined by (3.42), between the observed and estimated full model trajectory (a)-(c) and sub-trajectories (d)-(f) using the diagonal blocks of various covariance matrices as the norm (shown in upper-left corner of each plot). The true covariance matrix for the total error is used in (a) and (d); the observation error component is neglected in the covariance matrix used for (b) and (e); and the velocity error component is neglected for (c) and (f). The probability is 95% that a chi-square variable is less than the value shown by the dotted line.

to increase when the *a priori* error statistics are deliberately mis-specified. Finally, a simple method for analysing the residuals calculated using both the full trajectory and sub-trajectories appears to give a good indication of when the statistics of either error source have been mis-specified.

An important assumption made in this chapter is that errors in the modelled trajectory remain sufficiently small to allow the use of the linearised model velocity field for propagating the error statistics. Violation of this assumption can lead to non-Gaussian errors. In that case, if the distribution could be parameterised in some feasible way to enable it to be propagated by the actual (nonlinear through space) model velocity field, then the need for sub-trajectories would presumably no longer exist. A quadratic cost function would, however, no longer be appropriate. *Miller et al.* (1994) found it necessary to propagate higher order moments when using an EKF applied to the Lorenz model. Propagation of the full distribution of the errors along the trajectory and the need to use the more general form of MLE would substantially increase the complexity and computational expense of assimilating drifter trajectories. Such a general approach is presented by *Miller et al.* (1999) for assimilating data with several idealised, but highly nonlinear models formulated as stochastic differential equations. Alternatively, a Monte Carlo approach, similar to that used to obtain the distributions in Figure 3.3, could be used to estimate the distribution at each iteration of the estimation procedure. The use of sub-trajectories, however, appears to be sufficient for the case examined in this chapter to maintain a Gaussian distribution for the errors. In general, the effectiveness of using sub-trajectories will depend on the degree of spatial nonlinearity of the model flow field relative to the error in the model trajectory accumulated between observation times.

A major assumption in the formulation of the stochastic trajectory model (3.17) is that the errors in the ocean model are at the small spatial scales and do not affect the evolution of the large scale flow field. If the errors are only at small scales, it is easier to justify that they are uncorrelated along the trajectory. Frequently, however, this assumption will not be valid. Even when the errors are mostly due to

the unresolved small scale motions, many of the less sophisticated parameterisation schemes are incapable of accurately accounting for their effect on the resolved scales. Sources of ocean model error that directly affect the large scale motions may result from model bathymetry errors, surface flux errors, inaccurate parameterisation of mixing and bottom friction, or the neglect of nonlinear or baroclinic effects. These can cause velocity errors that are time-correlated, large scale, non-homogeneous and anisotropic, thus making them difficult to model statistically. The significance of these errors may depend on the length of the assimilation window relative to the temporal growth rate of the resulting error in the resolved velocity field. When the growth rate is high, it may be necessary to increase the size of the control vector by adding correction terms to the prognostic equations of the ocean model that account for the error and thus reduce the error growth in the model velocity field. The benefit of using such correction terms would rely on having both the sufficient data required to estimate the increased degrees of freedom in the control vector and a suitably specified spatial (and possibly temporal) covariance matrix for the corrections.

Sub-trajectories were suggested as a way of assimilating drifter data to reduce the separation between the true and simulated trajectories, thus improving the specification of the error statistics. Alternatively, the sub-trajectories could be calculated backward in time by advecting an observed position to the time of the previous observed position. Error growth along the forward trajectory (due to horizontally divergent flow) will decay in the backward direction and vice-versa. Therefore, using the backward trajectory reduces the separation between the trajectories in horizontally divergent flow. Conversely, the forward trajectory produces less separation in horizontally convergent flow. In general, when it is not known *a priori* if the flow field in a given region is convergent or divergent it seems reasonable to speculate that the best approach is to advect adjacent observed positions to the time half-way between. This approach should reduce large separations between the true and modelled trajectories caused by divergent flow.

The approach and examples in this chapter focused on the assimilation of a single

trajectory. Frequently, data from multiple drifters will be simultaneously available along with data from other sources. In this case the error correlation between nearby pairs of drifters or between a drifter and another observation type due to spatial correlation of the velocity errors may need to be considered. If the errors in a pair of modelled trajectories are correlated at some time, they will continue to be correlated along the remainder of the trajectories, regardless of their subsequent separation distance. *Daley* (1996b) performed an idealised study of the propagation of such covariances using idealised, non-divergent flow fields in an Eulerian framework. In those experiments, the initial covariance matrix was specified and the velocity error set to zero. It was found that application of different discretisation schemes to the advection equation affected the propagation of the error covariances.

Within any realistic data assimilation application, a pragmatic approach should be applied to the assimilation of drifter trajectory data. By considering the relative magnitudes of observation error and the error in the modelled trajectory due to the unresolved scales of motion, it may often be justifiable to adopt a simplified approach. For example, in cases where the trajectory observations are widely spaced in time, it may be found that between consecutive observation times, the error variance in the modelled trajectory grows to a level much larger than the observation error variance. In such cases, it may be appropriate to neglect observation error and assimilate the observations as sub-trajectories or as velocities after first-differencing, as in *Griffin and Thompson* (1996). Figure 3.9 showed that neglecting observation error and assimilating with sub-trajectories gave a lower estimation error, in that particular case, than using the correct error statistics with the full model trajectory.

In other cases, the observations may be closely spaced in time such that the error in the modelled trajectory remains small relative to the observation error. In that case it would be desirable to assimilate with a full model trajectory or long sub-trajectories that span several observation times. If the first observed position also contains observation error (contrary to the assumption made throughout the chapter) then the initial position of the modelled drifter can be included as part of

the control vector, α . The appropriate cost function is the same as (3.19), except an added term penalises the difference between the observed and estimated initial drifter position (weighted by the inverse of the observation error covariances) and the model trajectory is now computed using the initial position from the control vector. The advantage of controlling the initial position of the modelled trajectory is that the separation between the model and true trajectories should be reduced over a long model trajectory. This approach would likely be successful in cases where the velocity changed slowly between the observed positions such that several of the subsequent observed positions could provide useful information on both the initial position and the velocity field. In other words, there must still be fewer degrees of freedom in the control vector after addition of the initial drifter position than in the observations along the full trajectory or long sub-trajectory.

Another possible simplification may be made when the error in the modelled trajectory is too large to be neglected, but small relative to the scale of variation in the large scale flow field. In this case the model and true trajectory would experience the same velocities from the large scale flow field. Consequently, $\partial \mathbf{u} / \partial \mathbf{x}$ can be assumed to be zero for the purpose of propagating the error in the model trajectory (that is, $\gamma = \mathbf{I}$ in (3.13)). This removes the dependence of Σ^{tot} on α and therefore Σ^{tot} only needs to be calculated once during the minimisation. This corresponds to advection with simple diffusion, that is, an isotropic, Gaussian pdf with standard deviation growing as $\sqrt{t} \sigma^u$.

Finally, the approach presented in this chapter can be extended to assimilate sequential images of a concentration-like quantity that is advected with the fluid. Such types of data from both the ocean and atmosphere are usually remotely sensed by satellite-based instruments. They include measurements of sea-ice concentration, sea surface temperature, ocean colour, cloud imagery, and concentration of atmospheric chemical constituents, such as ozone. Because of the close relationship between the assimilation of drifter and image data, many of the issues examined in this chapter apply to the assimilation of images. Additional issues related to the assimilation

of sequential images have also been discussed by *Daley* (1996a) and *Kelly* (1989). The next chapter is a study of the surface currents over a region of the Labrador Shelf using a pair of real sea-ice images and an estimation procedure related to the developments in this chapter.

Chapter 4

Assimilation of Sequential Satellite Images

4.1 Introduction

Satellite images of sea ice distribution and sea surface temperature (SST) spanning large areas ($\sim 1000 \text{ km} \times 1000 \text{ km}$) of the ocean with spatial resolution from tens of metres to a kilometre are routinely acquired. The period between passes over the same portion of the ocean is on the order of a day. These spatial and temporal scales are often suitable for resolving the velocity field responsible for advecting features in the images. In this chapter a new method is presented for estimating velocity information from sequential images of, for example, sea ice or SST. The method is applied to the estimation of surface currents from images of sea ice over the Labrador shelf. As in the previous chapter, the goal of the method is the estimation of the large scale flow field and not the actual distribution of the observed quantity.

Several satellites with advanced very high-resolution radiometers (AVHRR) orbit the Earth in a way that provides excellent spatial coverage of the polar and mid-latitude regions. The AVHRR sensor passively measures the infrared or visible portion of the spectrum and therefore cannot be used to view ice or thermal features on the ocean surface in the presence of cloud cover. Synthetic aperture radar (SAR)

images have been used extensively in ice research. This is an active sensor that can produce high resolution images of sea ice distribution independent of weather conditions (*Carsey and Holt, 1987*). SAR images are now routinely acquired by the new Canadian satellite, RADARSAT. The relative advantages of AVHRR, SAR, and other satellites used for ice imaging are discussed by *Emery et al. (1995)*.

Several methods have been developed for automatically estimating velocity fields from a sequence of satellite images. Some of these are used for operationally tracking sea ice, for example at the Alaska SAR facility (*Kwok et al., 1990*). Similar types of data are used to recover atmospheric wind fields. For example, wind vectors inferred from cloud imagery are used routinely for NWP (e.g. *Tomassini et al., 1999*). *Daley (1996a)* used an extended Kalman filter and a simple wind model in an idealised setting to recover the two-dimensional wind field from observations of an atmospheric chemical tracer constituent. The purpose of this chapter is to introduce a practical method that allows the addition of sequential satellite images as a new source of data in an oceanographic data assimilative scheme. A variational approach is used to fit an ocean/advection model to the data.

An outline is given in section 4.2 of the methods previously developed for automatically obtaining velocity information from sequential satellite images. In section 4.3 the proposed method for estimating the velocity field from satellite images is introduced. This method is first tested in section 4.4 using the pair of pseudo-SST images shown at the end of Chapter 2. An application suited for tracking ice motion in the marginal ice zone using a pair of real satellite images is then presented in section 4.5. Section 4.6 is a discussion of the proposed method, including a comparison with existing methods and the final section presents some conclusions.

4.2 Existing Methods

The methods described in this section range from the purely statistical to those that include complex model dynamics. Area correlation and feature matching methods are

two statistical methods which have been used widely and successfully in operational ice tracking. Several applications of these techniques have also been carried out using SST images. The methods, however, commonly produce some clearly incorrect vectors, or “fliers”, that are a consequence of spurious correlation maxima. This is partly because the methods make no use of physical models to constrain the solutions. Another drawback of these methods is that other types of data related to ocean currents, ice floe trajectories, water density, or surface wind stress can not be readily incorporated.

Data insertion is a simple method of combining satellite images with model dynamics. Inverse methods are based on estimating the optimal values of some model parameters, treated as unknown. One such method uses a simple advection model to track a single ice floe. Another method uses a one-step Eulerian advection equation to estimate the surface current field using SST images. The advantage of these methods is that the estimated velocity fields satisfy the assumed model equations while optimally fitting the observations.

4.2.1 Area Correlation Methods

The most common method for using sequential images is based on correlations between sub-areas of a pair of images (*Ninnis et al.*, 1986; *Tokmakian et al.*, 1990). The first image is divided into equal sub-areas with the size of the sub-areas determined by the size of the image features. For each sub-area the linear translation is then found that maximises the correlation between the translated square from the first image and the corresponding area in the second image. The method is often referred to as the Maximum Cross-Correlation, or MCC, method. The result is a field of displacement vectors which are converted into velocity vectors by dividing by the time period between the images.

These methods assume a linear displacement of common features between images. It has been shown that the correlation peak broadens and eventually becomes statistically insignificant as rotation of the image features increases (*Kwok et al.*, 1990).

Therefore it is not possible to detect small-scale translational motion that involves a strong rotation (Vesecky *et al.*, 1988). This is particularly limiting in the marginal ice zone, where piece-wise rotation and translation of the ice floes are often observed. Emery *et al.* (1992) also encountered difficulties in extracting velocities from SST images of Gulf Stream rings. However, this method may be modified to include a rotation parameter for each sub-area. This extension of the method often results in a parameter space that is prohibitively large to search.

4.2.2 Feature Matching Methods

Feature matching methods are fundamentally different from area correlation methods. These methods involve a feature identification procedure followed by the application of a feature matching algorithm.

The Ψ - S method (Kwok *et al.*, 1990; McConnell *et al.*, 1991) is useful for tracking ice floes that undergo a strong rotation, but do not significantly deform. First, a set of individual features are characterised in both images using the so-called Ψ - S curve (see McConnell *et al.*, 1991, section II). This method of characterising features allows rotation to be easily detected. The correlation is calculated between pairs of these curves, one from each image. Pairs that are highly correlated are considered a match. The operational system used at the Alaska SAR Facility for tracking in the marginal ice zone involves first using a feature matching method to obtain displacements and rotations for a finite number of distinctive features in the images (Kwok *et al.*, 1990). These displacements are then used as initial estimates in an area correlation algorithm with both translation and rotation parameters, resulting in a vastly reduced search space. A similar method has been developed that is useful when deformation is significant, but rotation is not (McConnell *et al.*, 1991). The studies of Holland and Yan (1992) and Kuo and Yan (1994) present similar feature tracking methods developed for application to SST images.

4.2.3 Data Insertion

The U.S. Navy operational ice forecasting systems use satellite imagery and a complex ice-ocean model. The ice model includes the processes of ice advection, deformation, formation, and melting. The oceanic forcing is derived from a coupled ice-ocean model (*Hibler and Bryan, 1987*). The complete model is run daily to provide five day ice forecasts with the 24 hour forecast from the run of the previous day used for initialisation. Forecast fields include the ice motion, thickness, and concentration.

A simple method is used for assimilating observations of ice concentration into this model (*Preller and Posey, 1989; Preller, 1992*). An ice analysis is produced weekly by blending many data sets (including several types of remotely sensed images). The analysis is then used to completely replace the previous day's model-derived ice concentration field. Other fields, including ice thickness and mixed layer temperature, are then adjusted to make them consistent with the inserted concentration field. The effect of this is to abruptly adjust, once per week, the model state for the ice to agree with the observations. With this approach, however, there is no direct effect on the velocity field in the ocean model due to the ice observations, only the secondary effects due to changes in the other model variables.

4.2.4 Inverse Methods

Larouche and Dubois (1990) assimilated a pair of SAR ice images into a simple ice advection model driven by the surface wind and ocean current. The initial and final positions of an individual ice floe were manually derived from the two SAR images. Starting from the initial observed floe position, the model was used to predict the final positions. The control parameters were the speed and direction of the uniform and steady surface current. These parameters were adjusted to minimise the difference between the final position predicted by the model and the position observed in the second image.

Kelly (1989) used a direct inverse method to derive flow fields from consecutive SST images. This study demonstrates the utility of assimilating satellite images into

a simple model to obtain a globally fitted velocity field. The approach is based on the following simple Eulerian advection equation:

$$\frac{\partial I}{\partial t} + u \frac{\partial I}{\partial x} + v \frac{\partial I}{\partial y} = \text{Sources} - \text{Sinks}, \quad (4.1)$$

where $I(t, x, y)$ is the image pixel intensity, in this case representing surface temperature. The spatial and temporal derivatives of I were estimated directly from the pair of images. This approach enables one to estimate the advective velocities, (u, v) , and the right hand side of (4.1) under certain conditions. However, any velocity component normal to the gradient of pixel intensity in the images lies in the null space of the solution. To determine a realistic solution with flow components both parallel and normal to the spatial gradient of the tracer field, *Kelly* (1989) augmented the problem with additional constraints to produce an over-determined problem. The solution was then obtained by minimising the weighted sum of the mean squared misfit of the advection equation (4.1) and three regularization terms related to the velocity field: mean squared divergence, mean squared vorticity, and total kinetic energy.

The fact that velocities normal to the gradient of pixel intensity (that is, parallel to contours of pixel intensity) can not be resolved by the images is an important issue for all methods, including the proposed method presented below. To illustrate, consider typical SST images of the Gulf Stream: temperature contours tend to be aligned parallel with the flow. Therefore, a sequence of these images would provide little direct information on the surface velocity in the Gulf Stream. Such images would, however, provide excellent information on the position and meandering of the current (motion parallel to the SST gradient).

4.3 The Proposed Method

This section outlines the proposed method for estimating the ocean's surface velocity field from a sequence of satellite images. It is closely related to the approach for assimilating drifter trajectories presented in the previous chapter. An observed image, denoted $I^o(\mathbf{x})$, is assumed to have pixel intensities related to a quantity that is

advected by the surface velocity field, such as SST or ice. As in the previous chapter, an ocean model that depends on a set of unknown controls is assumed to accurately represent the large scale dynamics. To estimate the surface velocities using the sequence of images and ocean model, the maximum likelihood estimate for the controls is obtained. These controls are used with the ocean model to give the estimated velocity fields.

To calculate the maximum likelihood estimate for the controls, a separate stochastic model for the true image at a given time is calculated from adjacent observed images. The optimal controls maximise the likelihood that the modelled images are equal. For example, modelled images can be produced at the time of the second of an image pair. In this case, the first image is advected to the time of the second and the controls are sought that maximise the likelihood that the advected image equals the second observed image.

In general, the information from the satellite images can be combined with *a priori* estimates for the controls and other sources of data. The use of an ocean model also introduces valuable information concerning the ocean dynamics. These factors make this approach more general and potentially more powerful than the existing approaches described in the previous section.

4.3.1 Stochastic Advection Model

The advection model accounts for the underlying physical processes that cause the observed quantity in the satellite images, either sea ice concentration or SST, to change between images. It must also stochastically model the errors due to inadequacies in the ocean model and in the observing instrument. In general, the controls may include the initial and boundary conditions for a numerical ocean model used to drive the advection and additional parameters governing any sources or sinks included in the model.

In the case of sea ice, many processes other than passive advection may be important. In most instances, the motion of the ice through the water due to the direct effect

of the surface wind must be included. Additionally, a complex model, such as that of *Hibler and Bryan* (1987), may be required to account for melting or growth of the ice and also the effects of internal ice pressure. Over short time scales and in regions with low ice concentration, however, a simple model is often adequate (*Thorndike and Colony*, 1982). To be able to assimilate a sequence of thermal images, a model for the evolving SST field is required. *Kelly* (1989) used a simple model in which the local time rate of change of temperature included a term that varied linearly in space in addition to advection. This was used to represent a source term to account for the combined effect of surface heat flux and vertical entrainment. Horizontal diffusion is another process often included in such models (e.g. *Qiu and Kelly*, 1993).

Assuming the observed quantity is perfectly advected with the flow, but possibly subject to a source/sink term, denoted by $S(\mathbf{x}, t)$, leads to

$$\frac{\partial I^t}{\partial t} + \mathbf{u}^t \cdot \nabla I^t = S(\mathbf{x}, t), \quad (4.2)$$

where I^t is the true image (that is, the image that would be observed by a perfect instrument) and \mathbf{u}^t is the true velocity field including all scales of motion. The source/sink term can include the effects of many of the factors mentioned above. Consequently, images evolve through time according to the discrete-time model

$$I_n^t(\mathbf{x}_n^t) = I_0^t(\mathbf{x}_0^t) + \sum_{j=1}^n S_j(\mathbf{x}_j^t) \Delta t, \quad (4.3)$$

for any time-step $n > 0$ where Δt is the separation between time-steps. The pixel intensity at a given location is observed subject to error, according to

$$I_n^t(\mathbf{x}_n^t) = I_n^o(\mathbf{x}_n^t + \boldsymbol{\epsilon}_n^o) + \epsilon_n^i, \quad (4.4)$$

where ϵ^i is the error in observed pixel intensity and $\boldsymbol{\epsilon}^o$ is the error in the position assigned to the image pixel (navigational error).

By combining (4.3) and (4.4), a stochastic model for the true pixel intensity at time-step n at a given location can be expressed in terms of the observed image at time-step 0

$$\tilde{I}_n^t(\mathbf{x}_n^t | I_0^o) = I_0^o[\mathbf{x}_0^m(\mathbf{x}_n^t, \boldsymbol{\alpha}) + \boldsymbol{\epsilon}_0^o + \boldsymbol{\epsilon}_0^i] + \epsilon_0^i, \quad (4.5)$$

where the tilde indicates an advected image and the source/sink term has been eliminated for clarity. The model position $\mathbf{x}_0^m(\mathbf{x}_n^t, \boldsymbol{\alpha})$ is obtained by advecting the given position \mathbf{x}_n^t to time-step 0 using the large scale velocity field from the ocean model. The error $\epsilon_0^{\tilde{x}}$ is due to the accumulation of error during the advection from the small scale velocities not resolved by the model (see Chapter 3). Assuming the navigational error and errors due to the unresolved velocities are small relative to the scale of variation in the observed images, the following linear approximation is made:

$$\tilde{I}_n^t(\mathbf{x}_n^t | I_0^o) \approx I_0^o[\mathbf{x}_0^m(\mathbf{x}_n^t, \boldsymbol{\alpha})] + \frac{\partial I_0^o}{\partial \mathbf{x}}(\epsilon_0^{\tilde{x}} + \epsilon_0^o) + \epsilon_0^i. \quad (4.6)$$

For AVHRR images with ~ 1 km pixel size that are separated by ~ 10 h, ϵ^o is less than 1 km and assuming $\sigma^u \sim 10$ cm s $^{-1}$ gives $\epsilon^x \sim 1$ km. According to (4.6), an observed image that is advected according to the velocities resolved by the ocean model differs from the true image by three sources of error. The error term ϵ^x serves the same role, but in a statistical way, as the diffusion term in the more typical advection-diffusion models used for evolving tracer fields. The use of a diffusion term is, in fact, equivalent to computing the mean advected image given by (4.6) after averaging over realisations of ϵ^x and also including the quadratic in ϵ^x term (*Bennett, 1996*). The advantage of this statistical approach is, however, that the model in (4.6) can be integrated forward or backward through time, whereas a deterministic advection-diffusion equation cannot.

4.3.2 Cost Function

The general approach for computing the cost function for a sequence of images, denoted by J_I , involves advecting the images to a common time using the advection model (4.6). Assuming the linearisation in this equation is valid and the errors are Gaussian with zero mean, the error between the two modelled images is also Gaussian with zero mean. Consequently, a standard cost function can be formulated by taking the $-\log()$ of the likelihood function. For the sake of clarity, the method for evaluating J_I is only described in detail for the case of two images. Also, all errors are

assumed to be negligible, except for the error in measured pixel intensity, ϵ^i , which is assumed to be uncorrelated and have equal variance across the images. Therefore, the velocities from the ocean model are equal to the true velocities and the resulting displacement equal to the true displacement. This assumption may not be valid when using a coarse resolution model to assimilate SST images in areas with sharp gradients in surface temperature such as the Gulf Stream. The time between the images is discretized as $n\Delta t$ with $n = 0, \dots, N$.

The time half-way between the images, where $n = N/2$, is used as the common time. This choice should result in the least deformation of the images. A regularly spaced grid is defined at the middle time with a resolution equal to that of the images. Then, the k th position in this grid, denoted as $\mathbf{x}_{N/2}^k$, is advected backwards for $N/2$ time-steps by numerically integrating the velocity field (see Figure 4.1). Use of the superscript k identifies the resulting trajectory \mathbf{x}_n^k that passes through the k th position on the grid at $n = N/2$. The pixel intensity in the first image at the advected position \mathbf{x}_0^k is calculated using a Gaussian weighted interpolation scheme. Also, the pixel intensity may be modified by the integral of the source/sink term, $S_n(\mathbf{x})$, along the trajectory, however, this term is again neglected for clarity. This intensity is assigned to the k th grid position at $n = N/2$. When the pixel intensity is defined in this way for each element in the grid, the result is an image advected to the mid-point using the advection model and starting from the first image. This image is denoted by

$$\tilde{I}_{N/2}(\mathbf{x}_{N/2}^k | I_0) = I_0(\mathbf{x}_0^k) + \epsilon_0^i, \quad (4.7)$$

where the spatial interpolation is assumed in evaluating the pixel intensity of the observed image I_0 (the superscripts are dropped for clarity).

A similar treatment is applied to the second image. The positions in the grid at $n = N/2$ are advected forwards for $N/2$ time-steps to obtain \mathbf{x}_N^k . The pixel intensities in the second image are interpolated to the resulting positions. This second advected image at $n = N/2$ is obtained using the advection model and starting from the second

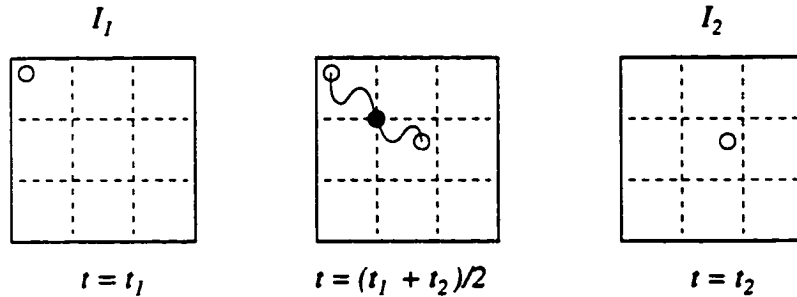


Figure 4.1: This schematic diagram shows how pixel intensities from two images, I_1 and I_2 (unfilled circles), are advected to a common time and location (filled circle). The procedure is repeated for each location on the grid at the common time.

observed image. It is denoted by

$$\tilde{I}_{N/2}(\mathbf{x}_{N/2}^k | I_N) = I_N(\mathbf{x}_N^k) + \epsilon_N^i. \quad (4.8)$$

The cost function is calculated as the sum of the squared differences in pixel intensity between the two advected images at time index $N/2$

$$J_I(\boldsymbol{\alpha}) = \frac{1}{4(\sigma^i)^2} \sum_k \left\{ \tilde{I}_{N/2}(\mathbf{x}_{N/2}^k | I_0) - \tilde{I}_{N/2}(\mathbf{x}_{N/2}^k | I_N) \right\}^2, \quad (4.9)$$

where σ^i is the standard deviation of ϵ^i . The dependence of J_I on $\boldsymbol{\alpha}$ is through the model velocity field used to obtain the two advected images. The summation in (4.9) is over all k for which both $\tilde{I}_{N/2}(\mathbf{x}_{N/2}^k | I_0)$ and $\tilde{I}_{N/2}(\mathbf{x}_{N/2}^k | I_N)$ are defined. This form of the cost function is based on the assumption that errors in the advected images are Gaussian, spatially homogeneous, and uncorrelated between pixels. (As discussed in Chapter 3, the framework of maximum likelihood estimation could also allow for non-Gaussian errors.)

The advection model is numerically integrated between the times of the images using a sufficient number of time-steps to approximate the solution to the continuous model, $d\mathbf{x}/dt = \mathbf{u}(\mathbf{x}, t)$. This differs in several ways from the methods of *Kelly* (1989) and area correlation. Firstly, by advecting each pixel in the images according to the advection model this method makes use of features down to the scale of a pixel.

The use of many time-steps in the integration allows the trajectories of each pixel to have curvature, thus supporting the rotation of features of any size. Also, temporal variation in the velocity field and processes which affect the pixel intensity during advection can be incorporated in the model.

The assimilation problem reduces to finding the values for the controls that minimise J_I . When incorporated into the advection model, these values for the controls optimally account for the motion between the two observed satellite images. The cost function is minimised using its gradient. Development of the adjoint model required to calculate the gradient of J_I is presented in Appendix E. The Gaussian-weighted interpolation scheme is necessary to guarantee that the pixel intensity at each of the advected locations is a continuous function of location. This ensures that the cost function is a continuous function of the controls, assuming the velocities depend on the controls in a continuous manner. Simple interpolation schemes, such as nearest-neighbour or bi-linear interpolation, produce a cost function with a step-like shape making it very difficult to find the minimum using a standard optimisation algorithm.

4.4 Application to Simulated SST Images

In this section, the method for assimilating sequential images described above is tested using the pair of pseudo-SST images shown at the end of Chapter 2. The first image is an idealised SST field that varies linearly between the north-west and south-east corners. The second image is produced by advecting the pixel locations from the first image according to flow fields from the low-dimensional model of flow over an isolated topographic feature described in Chapter 2. The first image corresponds with the time of maximum westward tidal velocity and the second image is from one tidal cycle later. As discussed in Chapter 2, the net Lagrangian velocity over a tidal cycle is small over most of the domain except for the region to the south-west of the bank centre. To illustrate how these images may be used to obtain information on the velocity field, the cost function (4.9) is evaluated over a range of values for the

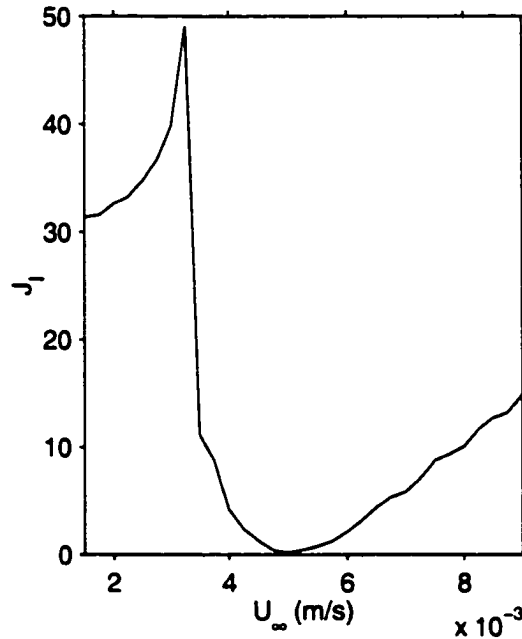


Figure 4.2: Cost function for the pair of pseudo-SST images shown in Figure 2.7. No model or observation errors were introduced. Only a subset of 64 pixels were advected to calculate J . The maximum near $U_\infty = 3.3 \times 10^{-3} \text{ m s}^{-1}$ is due to the non-linearity of the low-dimensional ocean model when the resonant frequency of the wave equals the tidal frequency.

tidal amplitude, U_∞ , assuming both the observations and the model are error free. The “observed” images were produced using $U_\infty = 5 \times 10^{-3} \text{ m s}^{-1}$.

To reduce the computational expense, only a subset of the image pixels was advected to calculate J : a square region subsampled at 1 km intervals beginning at the centre of the bank and extending 7 km to the south and west (a total of 64 pixels). The resulting cost function is shown in Figure 4.2. Since no model or observation error was introduced, the correct value of U_∞ produces almost a perfect match to the observed images (except for errors due to interpolating between the grids of the advected and observed images).

Note the similarity between the shape of this cost function and the cost function

in Figure 3.8b obtained when assimilating a series of drifter sub-trajectories. The maximum near $U_\infty = 3.3 \times 10^{-3} \text{ m s}^{-1}$ is again due to the non-linearity of the low-dimensional ocean model when the resonant frequency of the topographic Rossby wave equals the tidal frequency. If the images were separated by a longer period, the pixel intensities in the second image would become more mixed and, as in the case of long drifter trajectories, the cost function would likely exhibit more extrema. If errors in the ocean model or the observed images were non-negligible, then the minimum value for J_I would be larger and the optimal value for U_∞ would not exactly equal the true value. Also, incorporating these errors would significantly increase the computational expense of obtaining the estimate.

4.5 Application to AVHRR Ice Images

The method described in section 4.3 was applied to a pair of thermal AVHRR satellite images (the first from NOAA-12, the second from NOAA-11) from March 7, 1994 with a separation time of approximately 7.5 hours (Figure 4.3). A cloud free area of 100 by 100 pixels was extracted from both images corresponding to an area in the marginal ice zone off the coast of southern Labrador over Hawke Saddle and the northern half of Belle Isle Bank (Figure 4.4). The spatial resolution of the images is 1300 m in the zonal direction and 1100 m in the meridional direction. The extracted area covers approximately 130 km by 110 km.

The images were processed to reduce the large scale spatial variability in pixel intensity caused by differential illumination across the image plane. A simple “levelling” procedure was used (*Russ*, 1995). Then, the images were normalised using histogram equalisation (*Robinson*, 1985) to reduce the effects on pixel intensity of temporal variations in surface temperature and calibration differences between satellites. The resulting pixel intensities were normalised to have a range of 0 to 10 (Figure 4.5).

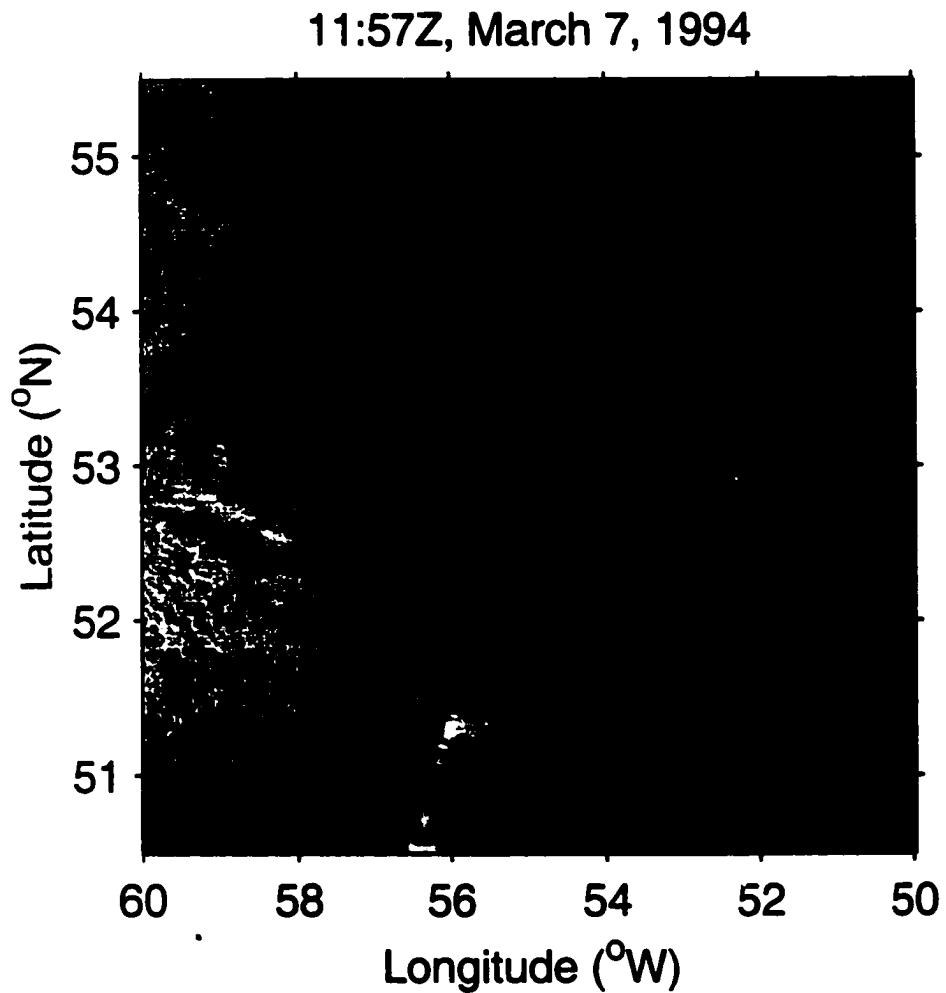


Figure 4.3: One of the complete 512 by 512 pixel AVHRR images showing the pack ice along the coast of southern Labrador and the northern tip of Newfoundland (left), the marginal ice zone (centre), and the cloud covered open water off of the shelf (top-right). The values along the axes are in degrees of latitude and longitude.

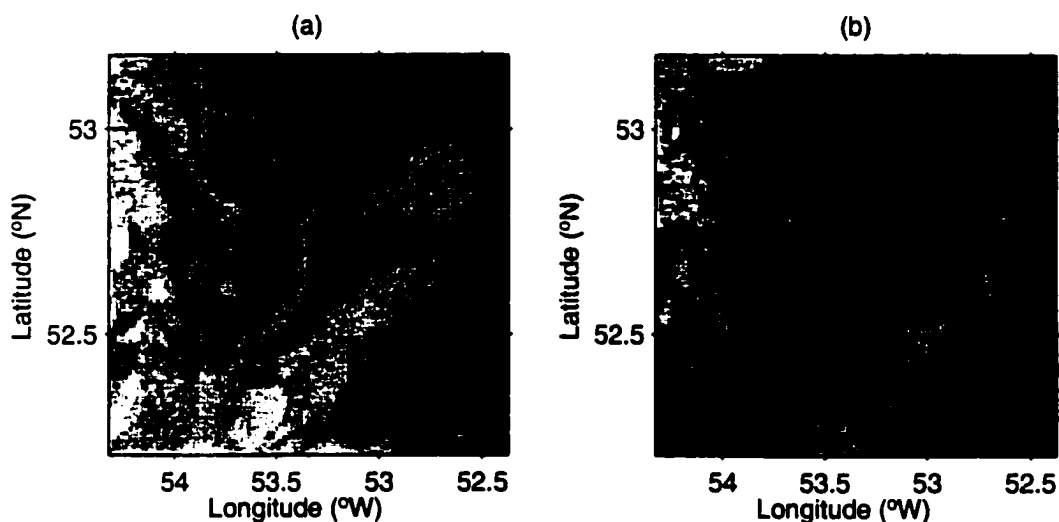


Figure 4.4: The pair of extracted 100 by 100 pixel subimages. The image in (a) is from 11:57, March 7, 1994 and (b) is from approximately 7.5 hours later at 19:33, March 7, 1994.

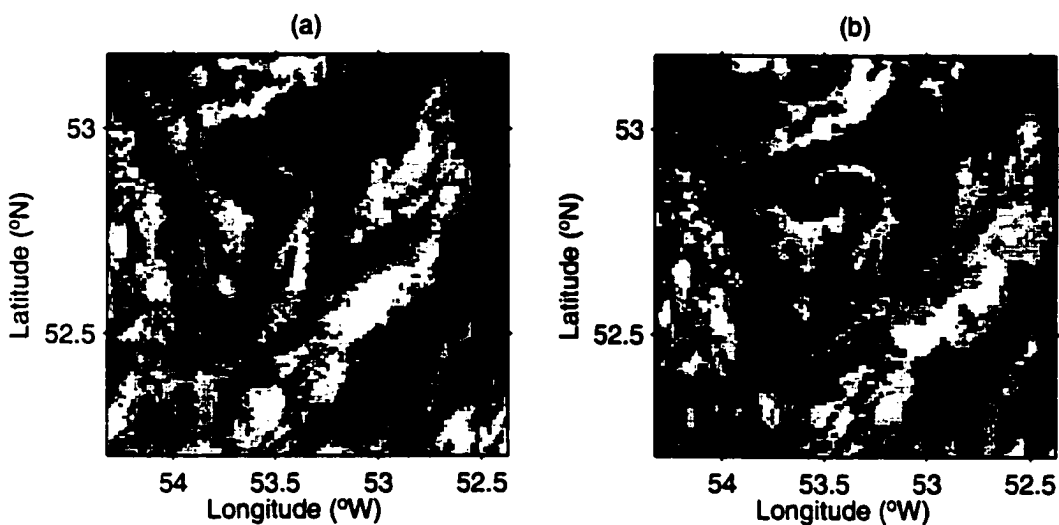


Figure 4.5: The same images as shown in Figure 4.4 after processing. The images were 'levelled' to reduce spatial variability in pixel intensity due to spatially varying illumination. Histogram equalisation was used to maximise contrast.

4.5.1 Ice Advection Model

First, a physical model for the advection of the ice is formulated. Following the analysis of ice motion by *Thorndike and Colony* (1982) it is assumed that the wind stress is in balance with the water stress on the lower surface of the ice. This assumption should be accurate for moderate to high wind speeds and variations in the forcings of several hours or longer. This leads to the following simple model (*Thorndike and Colony*, 1982) for ice motion:

$$\mathbf{u} = \mathbf{u}^w + \mathbf{u}^o. \quad (4.10)$$

The vectors \mathbf{u}^w and \mathbf{u}^o are the wind-driven and ocean-driven components of the ice motion, respectively. Internal ice stress is ignored and floes are allowed to freely deform. However, the resulting deformation should be reasonable even without internal ice stress because both the wind and the surface current fields are expected to have low divergence and shear on the scale of ice floes in the marginal ice zone. This model may be appropriate for such regions where ice drifts almost freely and the wind and ocean currents have a major effect on ice displacement (*Wadhams*, 1986). However, internal ice stress greatly constrains the movement of pack ice thus making this assumption invalid for such regions. However, as discussed in section 4.6, this assumption is not a necessary part of the general approach.

The wind-driven component of the ice velocity, \mathbf{u}^w , is defined as the wind speed multiplied by a constant, A , with the direction of the motion at a fixed angle, θ , to the right of the wind velocity. The two parameters, A and θ , are related to the drag coefficients between the upper surface of the ice and the overlying air and between the lower surface of the ice and the ocean. Values for these drag coefficients vary significantly with many factors including floe size and roughness, stability of the atmospheric boundary layer, turbulence in the oceanic surface boundary layer, and wave action (*Guest and Davidson*, 1987). The great uncertainty in these drag coefficients leads to the choice of treating A and θ as controls.

Six hourly wind fields on a 1° grid from the Canadian Meteorological Centre were

used to drive the ice advection model. These data were interpolated both in time, using four-point Lagrange interpolation, and in space, using bi-linear interpolation. The mean and standard deviation were calculated over the 48 hour period preceding the second image. The standard deviation of the wind speed is less than 25% of the mean speed (12.1 m s⁻¹). This implies relatively steady conditions and is consistent with the simplifying assumptions used in deriving the ice motion model.

The effect of the surface ocean current, \mathbf{u}^o , on the ice motion is assumed to be additive. The current field is taken to be horizontally non-divergent with a streamfunction, ψ . Therefore, the horizontal velocity field is defined as

$$\mathbf{u}^o(t, x, y) = \left(-\frac{\partial\psi(t, x, y)}{\partial y}, \frac{\partial\psi(t, x, y)}{\partial x} \right)^T. \quad (4.11)$$

(A better choice may be to use a transport streamfunction, consistent with a non-divergent depth-averaged transport, but using (4.11) allows the use of bathymetry as an independent check for the solution in section 4.5.3.) The streamfunction is defined using a linear combination of generalised structure functions, denoted by $\phi_i(x, y)$. These structure functions form the set of basis functions for the streamfunction which is expanded as follows:

$$\psi(t, x, y) = \sum_i \alpha_i(t) \phi_i(x, y). \quad (4.12)$$

Polynomials in x and y up to degree seven are used to define these structure functions (e.g. $\phi_i = x^n y^m$). Additionally, the current field is assumed to be constant between the images. The coefficients, denoted by the vector α , are the controls which define the streamfunction of the unknown current field.

The appropriate adjoint model was then formulated following the procedure presented in Appendix E to obtain the cost function gradient with respect to A , θ , and α .

4.5.2 Results

The cost function was first evaluated with all controls set to zero, representing the case of no ice motion. This value is proportional to the variance of the difference between the observed images.

Optimal values for the wind-driven model parameters were first found independent of the current field. The current field was set to zero everywhere ($\alpha = 0$) and a quasi-Newton minimisation routine was used to find the values of A and θ that minimise the cost function. Figure 4.6 shows a contour plot of the cost function surface as a function of these controls. The optimal values for the parameters are $A = 0.0285$ and $\theta = 36.5^\circ$. As a result of advection by only the wind-driven component J_I is reduced to 43% of its original value. From visual inspection of Figure 4.6 this is clearly the global minimum within the range of parameter values considered. These values are consistent with the slightly lower values ($A = 0.0205$ and $\theta = 25^\circ$) found from ice beacon data during a 60 day period in 1992, but closer to shore in pack ice where a higher ice concentration and proximity to the coast restrict the ice motion (*Carrieres et al.*, 1996). The predicted wind-driven ice velocity has an average speed of 34 cm s^{-1} directed 64° south of east. The parameters A and θ were fixed to these values for all subsequent analyses, but this is not necessary as discussed in section 4.5.4.

The image pair, with the wind effect removed, was then analysed both manually and using an area correlation algorithm to obtain a rough approximation of the ocean current field. The velocity field from the area correlation algorithm was similar to the manually tracked vectors except that the area correlation method gave several “fliers”. Therefore only the manually tracked velocity vectors, shown in Figure 4.7, were used in this application. The manually tracked vectors are only determined where ice features are clearly identifiable in both images and are therefore irregularly spaced. Other researchers have used such manually tracked vectors to study surface circulation. For example, *Ikeda and Tang* (1992) used the vectors to fit a stream-function for an area in the marginal ice zone over the Labrador and Newfoundland shelves.

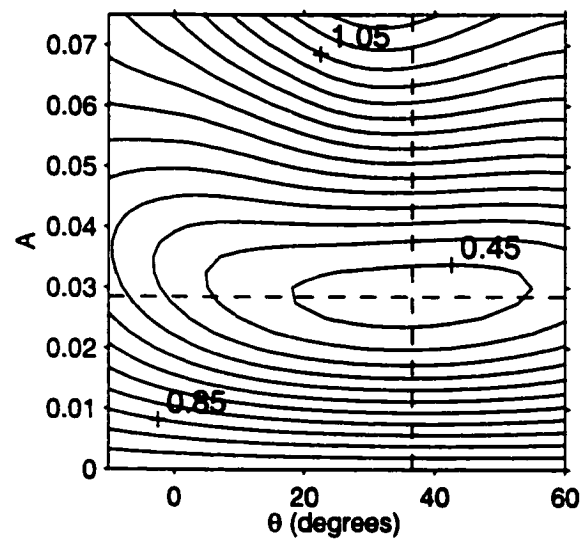


Figure 4.6: Contour plot of the cost function with α set to zero. The only controls are A and θ . Note the global minimum located at the intersection of the two dashed lines defined by $A = 0.0285$ and $\theta = 36.5^\circ$. J_I has been normalised by its original value with no ice motion. Contour values are separated by 0.05.

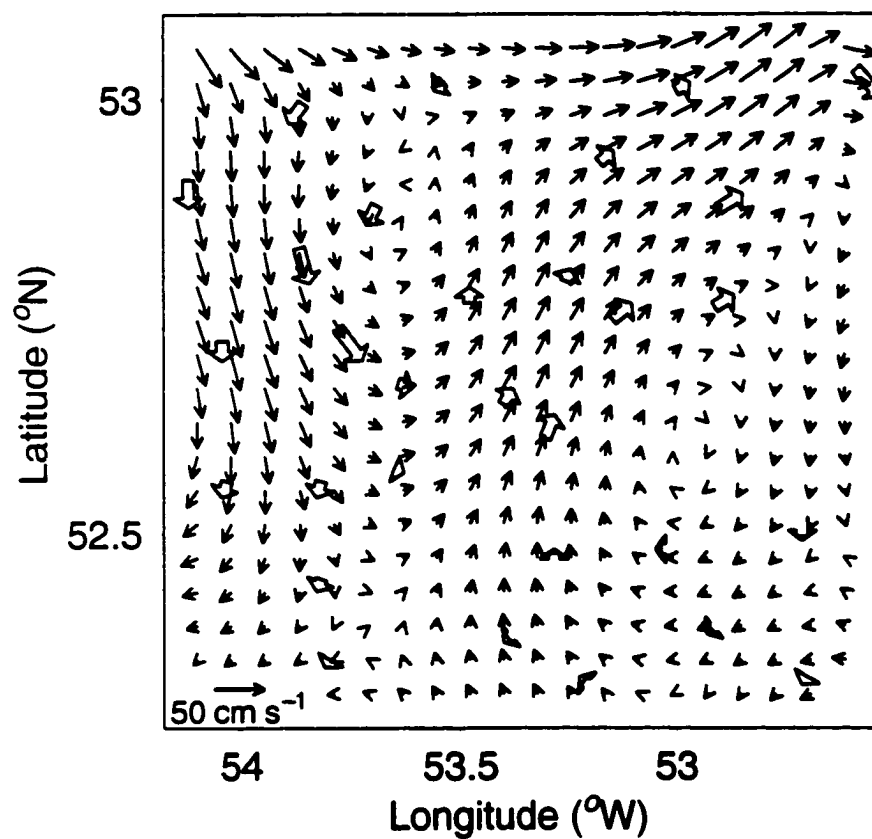


Figure 4.7: Comparison of the optimal ocean current field from the assimilation method and velocities from manual tracking. The optimal velocity field is represented by the narrow arrows distributed on a regular grid. The wide arrows are those derived from manual tracking.

Initial assimilation experiments without regularization gave unrealistic velocity vectors in areas with inadequate information. Therefore, a term which penalises enstrophy was included in the cost function to smooth out strong shears in the velocity field. The modified cost function is

$$J = J_I + W_r \sum_{t,x,y} [\nabla^2 \psi(t, x, y)]^2, \quad (4.13)$$

where J_I is the cost function (4.9). Employing an appropriate value for W_r , the method gave a velocity field which appears more realistic. The methods used for determining the maximum polynomial degree for the streamfunction and the best value of W_r are outlined in section 4.5.4. The optimal value of J_I is 21.9% of its initial value. Therefore the optimal ice velocity field is able to account for 78.1% of the variance between the original image pair. Figure 4.8 shows the resulting optimal streamfunction superimposed on the pair of images after the wind effect was removed. The relationship between the manually tracked velocities and the corresponding velocities from the optimal velocity field is shown in Figure 4.9. The optimal flow field and the manually tracked velocities, presented as vector plots in Figure 4.7, compare well over the region.

The computational time to reach the optimal solution for a pair of images using the adjoint model is about 30 minutes (running on a SPARC station 10, model 70). In general, the overall computational demand is proportional to the effort required to run the ice-ocean model for the period between the images and also the number of iterations required to find the optimal controls. The first factor depends on the complexity of the model and the size of the images. The latter depends both on the number of controls and the degree to which the optimisation problem is well-posed.

Figure 4.10 shows the frequency distribution of pixel intensity for the difference between the images. The frequency distribution of the difference between two random, uncorrelated images with the same distribution of pixel intensities as the ice images is shown in Figure 4.10(c). The quartiles of these distributions are identified in the figures by the vertical dashed lines. The distribution corresponding to the original

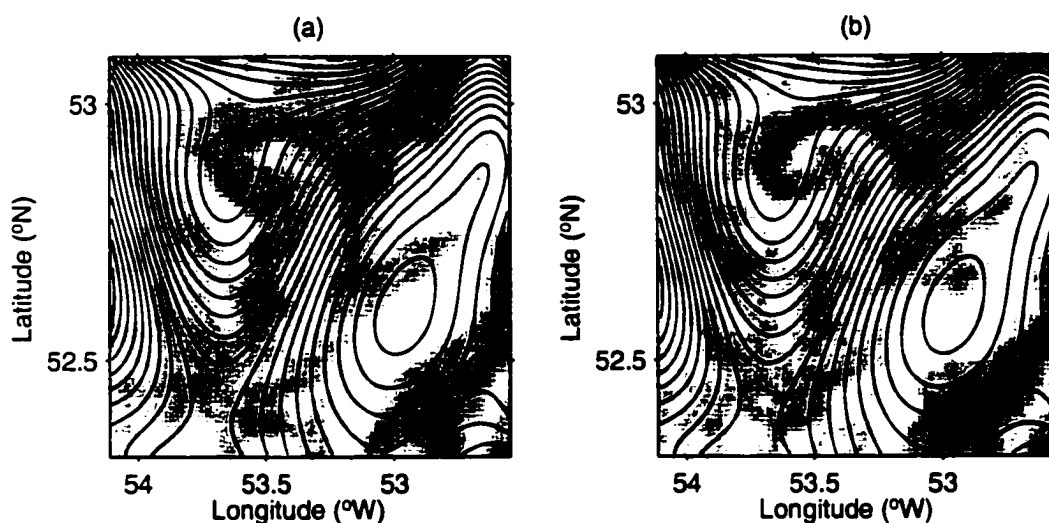


Figure 4.8: Contour plot of the optimal ocean current streamfunction superimposed on the two images (a) and (b) after removal of the wind effect. Note that the direction of flow is clockwise in the gyre to the east.

images (Figure 4.10(a)) closely matches that corresponding to the random images with the same nearly triangular form and the tails stretching out to the extremes of the range. However, the distribution corresponding to the optimally advected images (Figure 4.10(b)) has a much reduced spread with a corresponding high peak centred at zero. The interquartile range corresponding to the advected images is almost three times less than that of the original images.

4.5.3 Interpretation of Estimated Current Field

It is generally accepted that circulation on the Labrador shelf is largely governed by bathymetry (*De Young et al.*, 1995). Figure 4.11(b) is a contour plot of the bathymetry for the region. The shallow eastern portion of the region is the western half of Belle Isle Bank; the deep northern portion of the region is a feature called Hawke Saddle; and the southern area, which is also relatively deep, is the northern tip of St. Anthony Basin. One of the main features of the optimal current field is a relatively strong

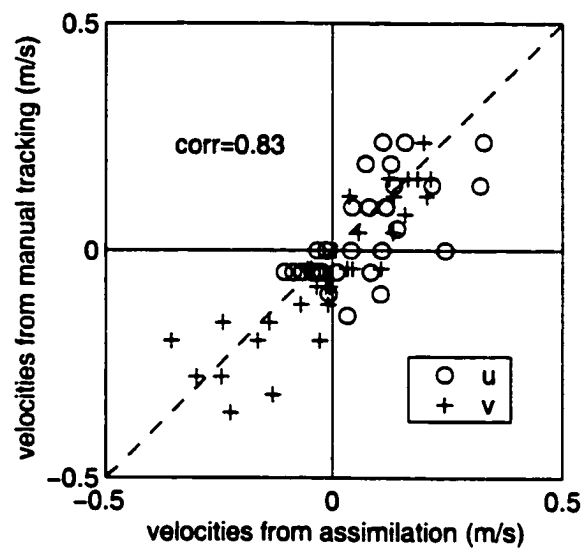


Figure 4.9: Scatter plot of velocity components from manual tracking against the corresponding velocity components from the assimilation procedure. Note the dashed line represents the perfect 1:1 relationship. The correlation coefficient is 0.83.

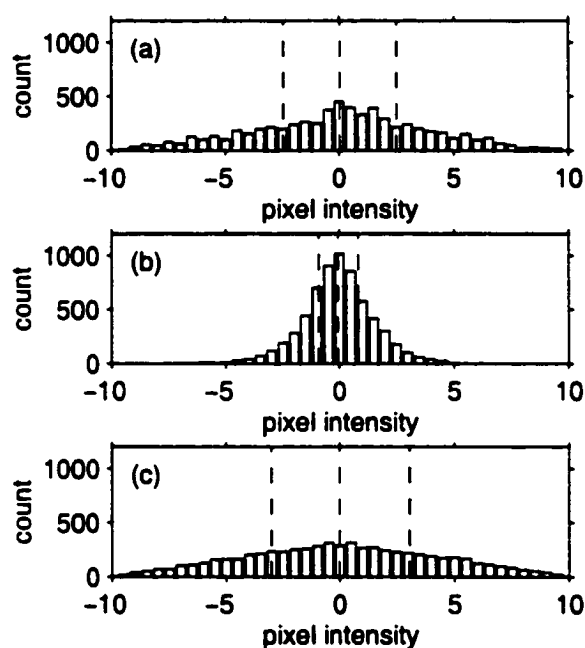


Figure 4.10: Histograms of the pixel intensity corresponding to the differenced images. (a) is the histogram corresponding to the difference between the original two images. (b) corresponds to the difference between the two images after advection by the optimal ice advection model. (c) shows the histogram corresponding to the difference between two uncorrelated random images with the same pixel intensity distribution (uniform) as the original images. The vertical dashed lines identify the quartiles. Therefore 50% of the data falls in between the two outer dashed lines.

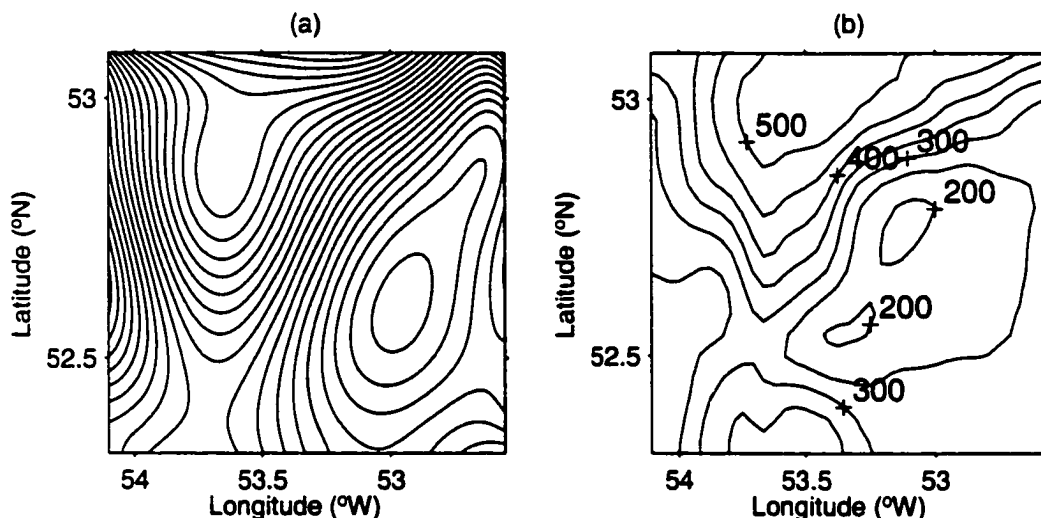


Figure 4.11: Comparison of the optimal streamfunction and the bathymetry of the region. (a) shows the streamfunction and (b) shows the bathymetry with contour increments of 50 m. Note the close correspondence of streamlines with the bathymetry contours.

current (up to 40 cm s^{-1}) which flows southward along the western boundary of the region, turns eastward, and flows north-eastward out of the north-east corner of the region. This path closely corresponds to the southern boundary of Hawke Saddle, a region with a strong gradient in water depth. Between St. Anthony Basin and Hawke Saddle is a well defined saddle point in bathymetry (52.5° N, 53.7° W) which corresponds well with the saddle point in the streamfunction. Comparing the bathymetry with the streamfunction shows that the streamfunction closely follows lines of constant depth.

The optimal flow field also agrees qualitatively with observations of the mean circulation in the region. *Tang et al.* (1996) observed from ice beacon data between 1985 and 1989, after removal of the wind effect, that the mean flow is eastward along the northern flank of Belle Isle Bank. *Greenberg and Petrie* (1988), who summarise current meter observations from Belle Isle Bank, observed north-eastward flow over

the northern portion of the bank and southward flow over the eastern side. *Peter-son* (1987) observed an eastward flow over southern Hawke Saddle using sequential satellite ice images after removal of the wind effect.

4.5.4 Sensitivity Studies

Even though the values of the wind-driven parameters obtained from the satellite images are reasonable, there is still a possible problem of co-linearity of the wind-driven motion and the ocean current. That is, the component of the ocean current which is spatially correlated with the wind-driven ice motion will contribute to the wind-driven parameter values in the procedure that was followed. The minimisation was therefore repeated with the wind-driven parameters and the streamfunction parameters allowed to vary simultaneously. The result was a flow field very similar to the original field, but with the centre of the gyre on Belle Isle Bank slightly offset to the east.

Due to the nearly uniform wind field, both temporally and spatially, any errors in the wind data would affect the optimal values of A and θ , but not the resulting \mathbf{u}^w . Large scale errors in image navigation may also result in errors in these parameters causing a uniform error in \mathbf{u}^w of about 6 cm s^{-1} . Relative navigational errors over the model domain are negligible.

The choice of basis functions used for the streamfunction model was also explored. The simple polynomial terms used to specify the streamfunction are not orthogonal which may cause some inefficiency in the minimisation procedure. A truncated Fourier series expansion was also attempted. The computational cost, however, of evaluating trigonometric functions increased the time of evaluating the cost function by about a factor of four. The optimal streamfunction for the case of 24 Fourier components was very similar to the one obtained using simple polynomials. Experiments with Legendre polynomials also provided similar results.

The degree of the polynomial was determined by examining the fit of the polynomials to the manually tracked velocities using linear regression. Through this

approach, it was found that a polynomial of degree seven (consisting of 35 terms) was sufficient to model the manually tracked velocities. However, in general such additional analyses will not be possible, especially if the method is fully automated. Therefore one may take an approach analogous to statistical model building. In this case, a simple model is fit first. Then, after an analysis of the residuals, it is decided whether or not to increase the complexity of the model, and the process is iterated. At some point the question arises: "How far should one proceed in adding complexity to the model?" Additional complexity increases the computation time and may only result in a small decrease in the value of the cost function. Some statistical measures, such as Akaike's information criteria (AIC), are available for such uses (*Priestley, 1981*). Initial experiments with polynomials of degree 4 and 5 reduced J_I to 28% and 23% of its original value, respectively. In general, this approach can be used to test which modelling assumptions are most consistent with the images by observing which model gives the best fit taking into account model complexity. This appears to be a more promising way to use sequential images to study ice-ocean physical processes compared with analysing displacement fields derived from purely statistical tracking methods that already have certain constraints on the displacement field.

The form and relative weighting of the regularization terms also requires careful consideration. Experiments were conducted where only mean kinetic energy was penalised. The resulting flow fields did not sufficiently damp out unrealistic flows without affecting the overall field significantly. In this application, an ad hoc approach was taken to determine the weighting factor, W_r , in (4.13). Several weighting factors were attempted and the one which gave the velocity field with the highest correlation to the manually tracked velocities was chosen. However, in the absence of a manually derived analysis the approach of cross validation can be adopted (e.g. *Griffin and Thompson, 1996*). Following this approach, a portion of the grid at $n = N/2$ is initially withheld in the calculation of the cost function. Then several values of W_r are used and the one resulting in the velocity field that minimises the misfit of the withheld portion is chosen. This value of W_r is then used in the final assimilation

with all of the grid.

4.6 Discussion and Conclusions

The proposed method for using sequential images represents a significant improvement over existing methods in the following ways:

1. The proposed method enables the utilisation of all available data in determining the velocity field;
2. By making prior assumptions about the form of the solution, the set of feasible solutions is constrained;
3. The method may be extended to include model dynamics; and
4. The model can be subsequently used to forecast ice movement or SST evolution using the estimated flow field as the initial conditions for the ocean model.

Unlike existing methods of ice tracking, it is possible to include other sources of data with the proposed method by simply adding terms to the cost function. For example, if additional data are available in the form of an ice beacon trajectory, given as the time series \mathbf{x}_n^b , then the additional term in the cost function would take the form

$$J_b = \frac{1}{2} (\bar{\mathbf{x}}^b - \bar{\mathbf{x}}^i)^T (\Sigma^{tot})^{-1} (\bar{\mathbf{x}}^b - \bar{\mathbf{x}}^i), \quad (4.14)$$

where the arrow denotes the stacked vector containing all two-dimensional observed beacon positions. The model counterpart to the beacon positions, denoted $\bar{\mathbf{x}}^i$, is obtained by integrating the modelled velocity field for ice motion starting from the time and location of beacon deployment. The covariance matrix of the total error between these vectors, Σ^{tot} , is calculated as described in the previous chapter for drifter trajectories.

The proposed method inherently constrains the set of feasible solutions to those velocity fields that are consistent with the ocean model. Area correlation and feature

matching methods often result in erratic vectors. Several approaches have been developed which attempt to reduce the number of these “fliers”. For example, *Emery et al.* (1991) imposed some spatial coherence by filtering the resulting vector field. Another approach is to automatically identify those individual vectors that are incorrect using statistical measures derived from the cross-correlation surface and replace them by interpolating between the neighbouring vectors (*Vesecky et al.*, 1988). In the simple application presented in the previous section, the use of a truncated polynomial expansion as well as the regularization term ensures that the ice motion field will vary spatially in a reasonably smooth way. The truncated expansion of the velocity field also greatly reduces the number of control parameters and therefore reduces the computation time. This is arguably superior to the approach of filtering the resulting high resolution velocity field, which actually increases the computation time of the overall procedure.

The proposed method can be extended to include model dynamics. The simple ocean model used in this application does not include any constraints other than horizontal non-divergence and spatial smoothness. In general, the streamfunction may be allowed to vary in time in a manner consistent with expected ocean dynamics. By employing a numerical circulation model, the controls can be the initial conditions, time-varying open boundary conditions, or system noise of the model. A benefit of using a circulation model is that the scale of variation in the flow field is determined by the model and therefore does not need to be estimated, as in the application presented. Model dynamics can also be incorporated into the ice model. The assumption of no interaction between ice floes makes the application presented here particularly suitable to tracking ice in the marginal ice zone. However, if a model accounting for internal ice stress and/or thermodynamic processes was available, it would be relatively straight forward to include it in a modified ice advection model making the application of the method more general. Naturally, the incorporation of more complex dynamics may result in an increased number of controls, thus requiring more data (additional images or other sources) to determine the optimal solution.

To extend the proposed method for forecasting ice motion, the ice advection model could be used. For the application described in section 4.5, the wind-driven component of the ice motion accounted for about three-quarters of the total variance that the ice advection model was able to explain. Therefore, by solely using the wind-driven model parameters determined by the inverse method and forecast winds, a significant amount of future ice motion should be accurately forecasted. If the method is extended to include ocean dynamics, these dynamics can be incorporated when forecasting ice motion. The ocean model would be integrated forward in time using the optimal flow field derived from the images as the initial conditions. The resulting flow field, along with predicted wind data and optimal wind-driven model parameters, would drive the ice advection model to predict the ice motion beyond the final satellite image.

In conclusion, the applications presented in this chapter were successful using a very simple model for the advection of the observed quantity. A more sophisticated approach may be necessary for images that are more widely separated in time. In that case the error in the trajectories due to the unresolved velocities may become significant. Consequently, the statistics of this error, denoted by ϵ^* in (4.6), would need to be calculated (following the approach described in the previous chapter) at the common time chosen for comparing the images. As mentioned earlier, this term serves the same role as the diffusion term in the more typical advection-diffusion models used for evolving tracer fields.

Further development and application of the approach presented in this chapter could potentially improve current ice tracking and forecasting capabilities. The method provides a framework in which an appropriate dynamical model is used to assimilate all possible data collected at different times by different types of sensors. This is quite different from the techniques typically applied to the tracking of sea ice. More generally, application of this method provides oceanographers a valuable new source of data on surface currents by applying the method to ice, SST or ocean colour imagery where appropriate.

Chapter 5

Estimation of 3D-Var Background Error Covariances using Empirical Orthogonal Functions

5.1 Introduction

In the field of numerical weather prediction (NWP), the main role of data assimilation is to produce an optimal estimate of the present state of the atmosphere for initialising a forecast model. This is typically done by using the information from a set of observations to correct a short-term forecast, referred to as the background state. Since atmospheric data consists primarily of a sparse network of point measurements, the data assimilation scheme must spatially interpolate the information from the observations and also spread the information from the observed geophysical variables to the other variables when producing the corrections. These spatial and between variable relationships are governed by the covariances of the errors in the background state used in the assimilation scheme.

Advanced four-dimensional variational assimilation schemes, such as 4D-Var or the various approaches based on the Kalman filter (KF), to some extent use the

forecast model to propagate the background error covariances. The model is used to evolve the background statistics through time, thereby causing the statistics to be non-stationary and depend on the specific meteorological situation. For example, situations such as rapid cyclogenesis can have a strong influence on the background error statistics, and therefore dramatically change the way the observations are used in producing the estimated atmospheric fields (*Rabier and coauthors, 1997*). Several NWP centres have recently implemented 4D-Var in either pre-operational (*Thépaut et al., 1999; Zupanski et al., 1999*) or operational (*Rabier et al., 1999*) mode.

Three-dimensional assimilation schemes, such as OI or 3D-Var, do not account for the temporal evolution of the background statistics. The statistics are instead assumed to be stationary and representative of the climatological background error. This simplification substantially reduces the computational expense of the data assimilation step in NWP. Additional constraints also typically applied to the background error statistics include assuming the correlations are homogeneous and isotropic in the horizontal and that they can be described by a simple functional form, such as a Gaussian function. While these assumptions are often necessary due to the limited information from which the statistics are computed, they often result in the information from the observations being interpolated in a less than realistic manner. For instance, due to the assumption of homogeneous and isotropic horizontal correlations made in most 3D-Var and OI schemes, the full influence of the atmosphere's dynamical response to baroclinic or orographic forcing can not be captured.

The goal of this chapter is to explore a new approach for representing the background error covariances within a typical 3D-Var system. The approach is based on using a truncated set of empirical orthogonal functions (EOFs) that span the most important subspace of background error. The EOFs are calculated from a set of samples representative of background error without imposing the typical constraints on the form of the covariances. However, without these constraints the limited number of error samples causes the covariance matrix to be highly rank-deficient and significantly affected by estimation error. Several approaches are examined to overcome

these problems. The first approach is to constrain the horizontal correlation functions to have limited horizontal extent. Another approach is to increase the rank of the estimated covariance matrix by blending a small number of EOFs with the conventional full-rank covariance matrix estimated by assuming homogeneous and isotropic correlations. In a specific application, an EOF-based covariance matrix is used to evaluate the impact of relaxing the assumptions of homogeneity and isotropy of the background error correlations within the Canadian 3D-Var.

The following section gives a brief overview of the 3D-Var algorithm, the specific approach used to represent the background error statistics in a typical system, and the possible sources of information on background error. In Section 5.3, an approach for formulating a more general background covariance matrix using EOFs is presented. Preliminary results, given in Section 5.4, demonstrate the impact of the various approaches on the analysis increment. Section 5.5 concludes with a discussion of the limitations and possibly valuable applications of the proposed approach.

5.2 Overview of 3D-Var

Several operational NWP centres currently employ, or have employed in the recent past, a 3D-Var system (*Parrish and Derber, 1992; Courtier et al., 1998; Gauthier et al., 1999*). Like the KF, the analysis and forecast steps of 3D-Var occur sequentially through time. This is unlike other assimilation methods discussed in this thesis that simultaneously fit an entire time-dependent model solution to data. Within a typical NWP system, the role of 3D-Var, like OI, is to combine the current set of atmospheric observations, \mathbf{y} , with a short-term forecast valid for the same time, referred to as the background state and denoted by \mathbf{s}^b , to produce an estimate of the complete atmospheric state. The optimal estimate is the state vector, \mathbf{s} , that minimises the cost function

$$J = \frac{1}{2} [\mathbf{s} - \mathbf{s}^b]^T \mathbf{B}^{-1} [\mathbf{s} - \mathbf{s}^b] + \frac{1}{2} [\mathcal{H}(\mathbf{s}) - \mathbf{y}]^T \boldsymbol{\Sigma}^o^{-1} [\mathcal{H}(\mathbf{s}) - \mathbf{y}], \quad (5.1)$$

where Σ^o is the observation error covariance matrix, \mathbf{B} is the background error covariance matrix (where \mathbf{B} is used in place of Σ^b to simplify notation) and $\mathcal{H}()$ is the possibly nonlinear observation operator that maps the state vector into the model counterpart to the observations.

Except for a difference in notation and the possibility of a nonlinear forward model, (5.1) is identical to the cost functions (1.10) used to introduce generalised linear regression and (A.3) used to describe the KF algorithm. However unlike the KF, the background error covariances are assumed to be stationary in the 3D-Var algorithm. This avoids the costly propagation of this matrix with the linearised model dynamics that makes the original KF algorithm infeasible for realistic NWP applications. Also, as opposed to the explicit inversion used in the KF algorithm (which requires inverting large matrices), the optimal solution for 3D-Var is found by minimising the cost function with an iterative optimisation algorithm employing the gradient of the cost function (as with the approaches in the previous two chapters).

5.2.1 Incremental Formulation

Unlike linear regression and the standard KF algorithm, some of the observed data types in (5.1) may be nonlinearly related to the state vector. To maintain the linearity of the estimation problem, the observation operator is linearised leading to

$$J = \frac{1}{2} \Delta \mathbf{s}^T \mathbf{B}^{-1} \Delta \mathbf{s} + \frac{1}{2} (\mathbf{H} \Delta \mathbf{s} - \mathbf{y}')^T \Sigma^{o-1} (\mathbf{H} \Delta \mathbf{s} - \mathbf{y}'), \quad (5.2)$$

where \mathbf{H} is the observation operator linearised with respect to the background state and the increment, or correction, to the background state is defined as

$$\Delta \mathbf{s} = \mathbf{s} - \mathbf{s}^b. \quad (5.3)$$

The initial misfit between the observations and the background state projected into observation space is defined as

$$\mathbf{y}' = \mathbf{y} - \mathcal{H}(\mathbf{s}^b). \quad (5.4)$$

Typically, the increment is assumed to be sufficiently small that the solution of this linearised estimation problem equals the nonlinear solution. In general, however, $\mathcal{H}(\cdot)$ could be re-linearised with respect to the updated estimate, the new solution found, and these steps iterated until the solution converges to the nonlinear solution.

The estimated $\Delta\mathbf{s}$, referred to as the analysis increment, is essentially an estimate for the error in the background state. The first term of (5.2) constrains the spatial structure of the increment according to the background error covariances that are typically spatially smooth. Consequently, the analysis increment will be spatially smooth. This justifies the use of a lower resolution to represent $\Delta\mathbf{s}$, \mathbf{B} , and \mathbf{H} to decrease the computational cost of minimising (5.2). The highest resolution, however, is maintained for calculating \mathbf{y}' , since this calculation is performed only once. Therefore, the information from the corrections is subjected to certain approximations (lower resolution and linearised observation operator), whereas the information from the background state (in which the model dynamics may have generated small scale information) is treated at full resolution. This approach of simplifying the representation of $\Delta\mathbf{s}$ and its relationship to the observation space while maintaining the highest accuracy for calculations involving \mathbf{s}^b is known as the incremental approach (*Courtier et al.*, 1994). Recent oceanographic applications of this approach include the studies by *Weaver and Vialard* (1999) and *Thompson et al.* (1998).

5.2.2 Dynamical Importance of the Background Error Covariances

The analysis increment generated by a single observation, the so-called structure function, is a useful diagnostic for evaluating how the assimilation system spreads information from individual observations both spatially and between variables. For illustration, consider the exact solution for the analysis increment (see (A.6) and (A.7) in Appendix A):

$$\Delta\mathbf{s}^a = \mathbf{B}\mathbf{H}^T (\Sigma^o + \mathbf{H}\mathbf{B}\mathbf{H}^T)^{-1} \mathbf{y}'. \quad (5.5)$$

For a single observation, the linearised observation operator \mathbf{H} is a row vector and, therefore, the quantity in brackets is a scalar. Furthermore, if the observation is of the same type as a variable in the analysis increment and is located at a grid-point, then the resulting analysis increment is proportional to the corresponding column of \mathbf{B} . Since the covariance matrix \mathbf{B} is taken to be stationary in 3D-Var, the way in which information from the observations is spread through space and to the other variables is fixed. In estimating the background error covariance matrix, constraining assumptions are often placed on its structure, such as assuming the correlations to be homogeneous and isotropic. However, the atmosphere's dynamic response to, for instance, spatially varying orography or baroclinic forcing can be significantly anisotropic, nonhomogeneous and non-stationary. The magnitude and spatial structure of the real background error will be significantly influenced by such dynamic responses, as described below.

Baroclinic instabilities in the atmosphere play a key role in daily weather events and also in the global poleward heat transport (*Gill, 1982*). They occur frequently below the sub-polar jet stream in areas with strong horizontal temperature gradients. For example, in winter along the eastern edge of the continents cold air from the land meets air that has been warmed by the ocean's western boundary currents creating strong horizontal temperature gradients. When the vertical shear in the zonal wind associated with the meridional temperature gradient becomes sufficiently large, wave-like disturbances can grow spontaneously. These unstable waves grow by converting the available potential energy from the horizontal temperature gradient into kinetic energy. The resulting decrease in available potential energy is associated with a net poleward heat flux that tends to decrease the meridional temperature gradient. This continues until a new equilibrium is reached consisting of a finite amplitude wave and a modified zonal flow. Disturbances with spatial scales similar to the baroclinic Rossby radius, which in the atmosphere is $O(1000 \text{ km})$, tend to be most unstable. Furthermore, linear stability analysis shows that the pressure and wind fields of the unstable waves (modes) tilt in the vertical against the zonal shear (*Holton, 1992*).

Within an assimilation system, any error in the initial conditions that project onto these unstable waves would grow rapidly. Therefore, the errors in the background state near the sub-polar jet stream would likely exhibit a westward vertical tilt in their spatial covariance structure when conditions are favourable for rapid growth of disturbances. Several studies with four-dimensional assimilation systems have confirmed that, in regions with baroclinic activity, structure functions tend to have a tilted structure (*Rabier and coauthors, 1997; Houtekamer and Mitchell, 1998*). In regions where such disturbances play a dominant role in the atmospheric dynamics, the stationary background statistics would also be expected to have tilted vertical covariance structures.

Isolated orography located in the path of a mean zonal wind can have significant effects on the low-level temperature and wind fields. *Ringler and Cook (1999)* discuss the seasonality of orographically forced stationary waves. Using a quasi-geostrophic model, they determined that the nonlinear interaction of mechanical and thermal forcing due to orography can produce a stationary response unlike the superposition of the individual linear responses. Over the Rockies, low-level diabatic heating occurs during the summer (on the order of 3 K/day) and cooling during winter (1-2 K/day), causing seasonal variation in the response. Both model results and observations show that in winter the average low-level streamfunction, after removing the zonal mean flow, has a positive maximum over the Rockies. In the summer, a small minimum is located over the Rockies with positive maxima to the west and east. Within an assimilation system, errors in the zonal wind initial conditions would either enhance or weaken this response to orography in the forecast. Therefore a correlation between the zonal wind and the orographic response should be present in the seasonally averaged background error statistics.

5.2.3 Sources of Information on Background Error

Estimating the statistical properties of background error in NWP systems is a difficult task since the true state of the atmosphere is, of course, not available for direct

comparison against the background state. An approach, described by *Hollingsworth and Lönnberg* (1986), of comparing forecasts against accurate observations has been widely used. The approach attempts to separate observation from background error by making the assumption that observation error is uncorrelated between closely situated measurements. The effectiveness of this method, however, is limited by the sparsity of observations, especially over the oceans and the southern hemisphere. As a consequence, only the horizontal correlation length scale (assuming isotropic correlations), vertical correlations, and variances of the stationary background error can be reasonably well estimated. These statistics must also be assumed to be horizontally homogeneous over large regions of the globe. *Gauthier et al.* (1999) used horizontal length scales and variances estimated by separately averaging over the northern extratropics and the tropics. They then used the summer (winter) statistics from the north for the southern extratropical summer (winter) statistics.

Another approach, suggested by *Parrish and Derber* (1992) and referred to as the NMC method, uses only a set of forecasts from an existing assimilation system to estimate the stationary background error statistics. The differences are calculated between forecasts with two distinct lead times, but valid for the same time (usually 24 h and 48 h forecasts). These forecast differences are taken to be representative of background error and their statistics are estimated from a two to three month ensemble. Some NWP centres have found a noticeable improvement in their forecasts after switching to a background error covariance matrix estimated using this method (*Parrish and Derber*, 1992; *Derber and Bouttier*, 1999). The method is also very convenient since the ensemble of forecast differences provide information for all variables over the entire globe on the same grid as the forecast model.

Assuming 24 h and 48 h forecasts valid for the same time are used, the basis of the NMC method can be understood by considering how the two forecasts differ at the initial time of the 24 h forecast. At that time, the initial state of the 24 h forecast has been influenced by all of the assimilated observations over the previous 24 hours, (that is, the initial time of the 48 h forecast) and therefore should be closer to the

truth than the 48 h forecast. Therefore, the difference at this time should resemble the forecast error. This initial perturbation is then propagated through the model dynamics for 24 hours, allowing additional growth in areas of unstable dynamics and decay in areas of stable dynamics. This additional propagation time ensures that the imposed multi-variate and spatial structure from the assimilation system used to produce these forecasts does not strongly influence the new background statistics being estimated.

The NMC approach, however, has several drawbacks. For instance, in areas with few observations the initial perturbation may be small and the resulting error variance consequently underestimated (see Figure 5.1). Close to the surface some fields may be dominated by the external forcing. If the same external forcing is used for both forecasts, this can also result in the underestimation of the error variances for certain variables. Furthermore, use of 24 h and 48 h forecasts appears to allow too much growth in unstable regions to be representative of the error in the background state (typically a 6 h forecast). This leads to the overestimation of the error variance, for example, in the northern sub-polar jet region (*M. Fisher, ECMWF, 1999, personal communication*). Some of these problems can be mitigated, for example, by using zonally averaged variances and scaling these variances according to information derived from direct comparison of forecasts with observations following the approach mentioned above.

5.2.4 Background Error Covariances with Homogeneous and Isotropic Correlations

The background error statistics in the Canadian 3D-Var system have recently been reformulated, as described in detail by *Gauthier et al. (1998)*, following the approach of the European Centre for Medium-range Weather Forecasting (ECMWF) (*Derber and Bouttier, 1999*). The background error statistics are estimated from a set of error samples obtained using the NMC method (described in the previous section). Since only a small number (~ 100) of error samples are used, a set of assumptions must be

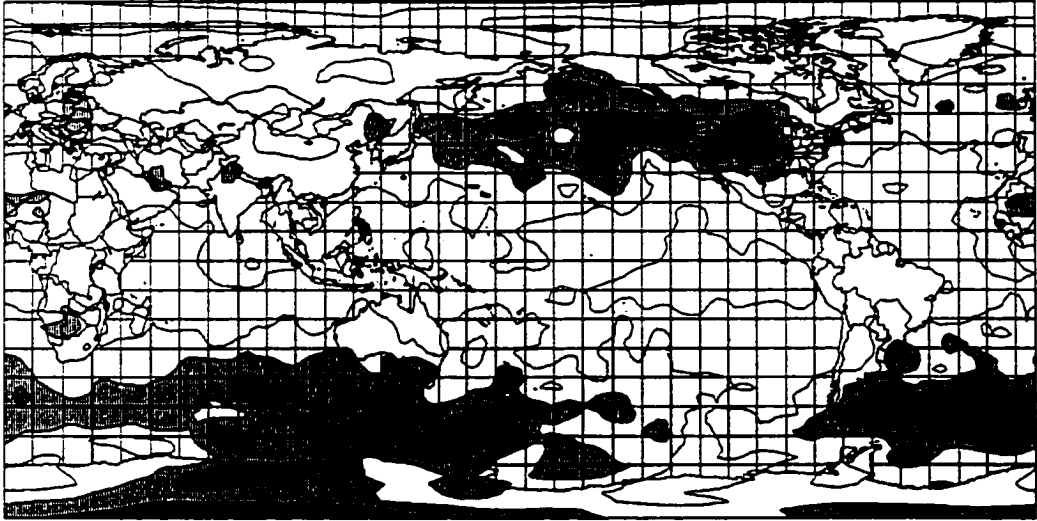


Figure 5.1: Streamfunction variance near 500 hPa obtained directly from a set of 91 lagged forecast differences (48 h minus 24 h forecasts). The contour interval is $5 \times 10^5 \text{ m}^2 \text{ s}^{-1}$. Note the low values over the tropics and the North Atlantic where lack of observations would suggest relatively high background error should occur.

imposed on the form of \mathbf{B} , which has a rank of $O(10^6)$, to improve its estimation. This section gives an overview of the aspects of the formulation relevant to the subsequent description of using EOFs to represent the background error covariances.

The full covariance matrix is not explicitly calculated within the 3D-Var, since it would be inefficient to store and manipulate this matrix. Instead, the control vector used for the minimisation, denoted by $\boldsymbol{\alpha}$, is defined according to

$$\Delta \mathbf{s} = \mathbf{B}_{hi}^{1/2} \boldsymbol{\alpha}, \quad (5.6)$$

where $\mathbf{B}_{hi} = \mathbf{B}_{hi}^{1/2} \left(\mathbf{B}_{hi}^{1/2} \right)^T$ and the subscript hi refers to the fact that the horizontal correlations are assumed to be homogeneous and isotropic. Using this definition allows the cost function (5.2) to be rewritten in terms of $\boldsymbol{\alpha}$ as

$$J = \frac{1}{2} \boldsymbol{\alpha}^T \boldsymbol{\alpha} + \frac{1}{2} \left(\mathbf{H} \mathbf{B}_{hi}^{1/2} \boldsymbol{\alpha} - \mathbf{y}' \right)^T \boldsymbol{\Sigma}^{-1} \left(\mathbf{H} \mathbf{B}_{hi}^{1/2} \boldsymbol{\alpha} - \mathbf{y}' \right). \quad (5.7)$$

This definition of the control vector therefore results in pre-conditioning the minimisation with respect to the background term. As a consequence, the minimisation of

(5.7) can be performed with significantly fewer iterations compared with using $\Delta\mathbf{s}$ directly as the control vector. Also, assuming the starting point in the minimisation is the background state (that is, $\boldsymbol{\alpha} = 0$ initially), the inverse of \mathbf{B}_{hi} is not required. The square-root of \mathbf{B}_{hi} is defined in terms of the following sequence of operators that are applied to $\boldsymbol{\alpha}$ whenever the value of $\Delta\mathbf{s}$ is required to calculate the observation part of the cost function (or its gradient):

$$\Delta\mathbf{s} = \mathcal{M}\mathcal{W}\mathcal{S}^{-1}\tilde{\mathbf{C}}^{1/2}\boldsymbol{\alpha}, \quad (5.8)$$

The remainder of this section provides a brief description of the operators in (5.8).

The operator, \mathcal{M} , in (5.8) accounts for the statistical relationships between the different geophysical variables in $\Delta\mathbf{s}$. At each grid point, the variables used for the increment and the background error covariance matrix are

$$\Delta\mathbf{s} = [\Delta\psi, \Delta\chi, \Delta T, \Delta p_s, \Delta \ln(q)]^T, \quad (5.9)$$

where the wind field is represented by $\Delta\psi$ and $\Delta\chi$, the streamfunction and velocity potential, respectively; the mass field by ΔT and Δp_s , the temperature and surface pressure, respectively; and the moisture field is represented by the natural logarithm of specific humidity, $\Delta \ln(q)$. Due to the importance of the geostrophic balance between the wind and mass variables in the extratropics, the forecast errors between these fields will also be in approximate geostrophic balance. In the extratropics near the surface, the effect of bottom friction causes a coupling between the rotational and divergent components of the wind field, and also between their forecast errors, due to the Ekman balance (*Polavarapu, 1995*). These multi-variate relationships are modelled using balance operators and result in covariances between the background error of the mass and wind fields.

The relationship between the mass and wind increments due to geostrophy is defined by

$$\begin{bmatrix} \Delta T \\ \Delta p_s \end{bmatrix} = \mathcal{V}\mathcal{L}\Delta\psi + \begin{bmatrix} \Delta T' \\ \Delta p_s' \end{bmatrix}, \quad (5.10)$$

where primes denote the “unbalanced” component of the variable. The linear operator \mathcal{L} is the geostrophic balance operator that transforms each horizontal field of streamfunction into a field of the balanced mass variable. The balanced mass variable is defined as

$$R \left(\sum_l T_l \Delta \ln p_l + T_r \ln p_s \right),$$

where the summation is from the surface to the vertical level of interest, R is the gas constant of dry air, T_r is a reference temperature and $\Delta \ln p$ is the difference in the natural logarithm of pressure at adjacent vertical levels. The local balance operator, where $\Delta \psi$ is simply multiplied by the local value of the Coriolis parameter, is used for reasons discussed in *Gauthier et al. (1999)*. The linear operator \mathcal{V} is an empirical inverse hydrostatic operator that transforms vertical profiles of the balanced mass variable into temperature profiles. This operator is estimated using a regression analysis over the ensemble of error samples between temperature profiles and profiles of $\mathcal{L}\Delta\psi$, in grid-point space. This approach is used to avoid problems of increased noise in the vertical structure and the null space associated with using a theoretically based inverse hydrostatic operator. Since not all vertical modes of the wind field are expected to be coupled with the mass field, this operator can also act to filter out these modes.

The divergent wind component is modelled as the sum of a balanced and unbalanced part according to

$$\Delta\chi = -\tan(\vartheta)\Delta\psi + \Delta\chi', \quad (5.11)$$

where ϑ is the turning angle between the streamlines and the wind vectors that is allowed to vary with latitude and vertical level. This differs from the homogeneous balance operator relating balanced mass to divergence in spectral space used by *Derber and Bouttier (1999)*. The turning angle is determined using a regression analysis between the error samples of χ and ψ .

The balance operators are first used to transform the error samples into the unbalanced variables

$$\Delta \mathbf{s}_u = [\Delta \psi, \Delta \chi', \Delta T', \Delta p', \Delta \ln(q)]^T, \quad (5.12)$$

by removing the balanced components. The remaining correlations between these variables were found to be small, thus supporting the assumption that the two balance operators described above are sufficient to account for the between-variable relationships. Therefore, to simplify the formulation of \mathbf{B}_{hi} , the correlations between the different variables in $\Delta \mathbf{s}_u$ are set to zero. Application of the operator \mathcal{M} is equivalent to applying (5.10) and (5.11) such that

$$\Delta \mathbf{s} = \mathcal{M} \Delta \mathbf{s}_u. \quad (5.13)$$

The background error covariance matrix in terms of the variables in $\Delta \mathbf{s}_u$ is denoted by \mathbf{B}_u and is related to \mathbf{B}_{hi} according to

$$\mathbf{B}_{hi} = \mathcal{M} \mathbf{B}_u \mathcal{M}^T. \quad (5.14)$$

Figure 5.2 shows the ratio of the unbalanced temperature variance to the total temperature variance obtained from a set of error samples. The low values in the extratropics demonstrate the importance of the geostrophic balance. Values larger than one are likely due to the assumption of homogeneity for the empirical inverse hydrostatic operator. Similarly, Figure 5.3 shows the ratio of the variances of unbalanced velocity potential to total velocity potential near the surface. In the extratropics, the balance operator is able to account for nearly half of the total variance at the surface.

The estimation of the covariance matrix \mathbf{B}_u requires estimating the correlation matrix and the variances for each variable in $\Delta \mathbf{s}_u$. The assumption is made that the horizontal correlations for each variable are horizontally homogeneous and isotropic. The vertical correlations are assumed to be homogeneous in the horizontal. As a consequence of these assumptions, the spectral representation of the correlations take on a block diagonal form, where the diagonal blocks depend only on total wavenumber,

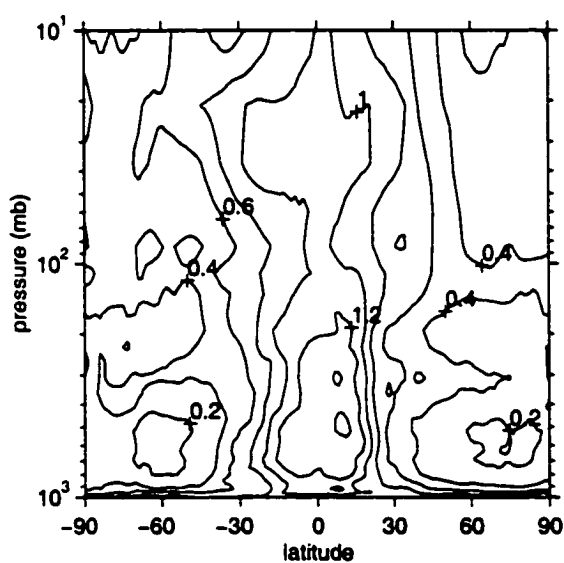


Figure 5.2: Ratio of the variance of unbalanced temperature to the total temperature variance as a function of pressure and latitude. Note that in the tropics the geostrophic balance is not able to explain any of the temperature variance and therefore the unbalanced component equals the total temperature. In the extratropics, however, the temperature variance is explained mostly by its balanced component above the planetary boundary layer.

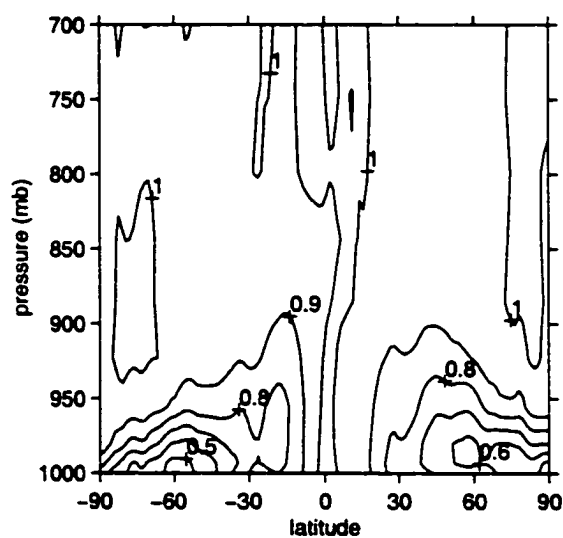


Figure 5.3: Ratio of the variance of unbalanced velocity potential to the total velocity potential variance near the surface as a function of pressure and latitude. Note that in the extratropics, the variance explained by unbalanced component is reduced to almost one half of the total variance at the surface.

as shown by *Gauthier et al.* (1998). This results in an efficient and convenient way of representing the correlation matrices. The variances of \mathbf{B}_u are represented in grid-point space and assumed to depend only on vertical level and latitude. Therefore, the background error covariance matrix for the unbalanced variables can be written as

$$\mathbf{B}_u = \mathbf{W} \mathbf{S}^{-1} \tilde{\mathbf{C}} \mathbf{S} \mathbf{W}^T, \quad (5.15)$$

where $\tilde{\mathbf{C}}$ is the spectral representation of the correlation matrix, \mathbf{W} is the diagonal matrix of standard deviations in grid-point space and \mathbf{S} is the horizontal spectral transform.

Ultimately, the full covariance matrix

$$\mathbf{B}_{hi} = \mathbf{M} \mathbf{W} \mathbf{S}^{-1} \tilde{\mathbf{C}} \mathbf{S} \mathbf{W}^T \mathbf{M}^T \quad (5.16)$$

has nonhomogeneous and anisotropic covariances, even though the horizontal correlations are homogeneous and isotropic. This is due to both the modulation as a function of latitude by the standard deviations, in \mathbf{W} , and the nonhomogeneous character of the balance operators, in \mathbf{M} . Because of the assumptions used to formulate \mathbf{B}_{hi} , this matrix will likely be full-rank and reasonably well estimated even if only a relatively small number of error samples are used. As discussed previously, however, the simplifying assumptions made in the formulation of \mathbf{B}_{hi} , especially that of homogeneous and isotropic correlations, often can not be justified when the influence of the real atmosphere on the background error is considered.

5.3 Representing the Background Error Covariances with EOFs

In this section, a method of using EOFs to represent the background error covariance matrix is presented. A covariance matrix represented by the leading N_e EOFs

calculated directly from a set of N_b error samples is given by

$$\mathbf{B}_e = \mathbf{E}\mathbf{\Lambda}_e^2\mathbf{E}^T, \quad (5.17)$$

where the columns of \mathbf{E} are the EOFs and the elements of the diagonal matrix $\mathbf{\Lambda}_e^2$ are the corresponding eigenvalues. The rank of \mathbf{B}_e equals the number of retained EOFs, the maximum being $(N_b - 1)$. As a result, the analysis increment obtained using such a covariance matrix would only span the N_e -dimensional subspace spanned by the EOFs.

Alternatively, an EOF-based covariance matrix can be calculated in conjunction with some of the assumptions used to estimate the conventional covariance matrix, \mathbf{B}_{hi} . For example, if the error samples are obtained by the NMC method, it may be preferable to retain the use of balance operators to model the multi-variate relationships and also use the spatially averaged variances to overcome some of the limitations of the error samples, as described earlier. Therefore, only the assumptions of homogeneity and isotropy of the correlations are relaxed. This is accomplished by calculating the EOFs from error samples that have been transformed into the unbalanced variables and normalised by the three-dimensional standard deviations of the samples. Then, the resulting covariance matrix is given by

$$\mathbf{B}_e = \mathcal{M}\mathcal{W}(\mathbf{E}\mathbf{\Lambda}_e^2\mathbf{E}^T)\mathcal{W}^T\mathcal{M}^T. \quad (5.18)$$

An approach for obtaining the EOFs and eigenvalues from a small set of error samples is presented in Appendix F. Two proposed approaches are described below for increasing the rank and improving the estimation of an EOF-based covariance matrix.

5.3.1 Blending EOFs with \mathbf{B}_{hi}

One method of increasing the rank of \mathbf{B}_e beyond the number of retained EOFs, is to use the full-rank covariance matrix \mathbf{B}_{hi} projected into the subspace orthogonal to the EOFs, that is, the null space of \mathbf{B}_e . To blend the EOFs with \mathbf{B}_{hi} , it is first assumed that the retained EOFs are accurate estimates of the leading eigenvectors of

the true background error covariance matrix, \mathbf{B}_t . Then, the eigenspace of this matrix is partitioned into two subspaces: one spanned by the EOFs, \mathbf{E}_1 , and the remaining subspace spanned by \mathbf{E}_2 ,

$$\mathbf{B}_t = [\mathbf{E}_1 \ \mathbf{E}_2] \begin{bmatrix} \Lambda_1^2 & 0 \\ 0 & \Lambda_2^2 \end{bmatrix} \begin{bmatrix} \mathbf{E}_1^T \\ \mathbf{E}_2^T \end{bmatrix}. \quad (5.19)$$

The eigenvectors and eigenvalues of the orthogonal subspace, \mathbf{E}_2 and Λ_2^2 , respectively, are unknown. To use the covariance matrix \mathbf{B}_{hi} to describe the errors in this subspace, Λ_2^2 is first replaced by

$$\Lambda_2^2 = \mathbf{E}_2^T \mathbf{B}_t \mathbf{E}_2, \quad (5.20)$$

obtained by pre- and post-multiplying (5.19) by \mathbf{E}_2^T and \mathbf{E}_2 , respectively. This is the projection of the true covariance matrix into the subspace orthogonal to the EOFs. The true (and unknown) covariance matrix in (5.20) is then replaced with the full-rank covariance matrix, \mathbf{B}_{hi} . Also, by using the relationship

$$\mathbf{E}_2 \mathbf{E}_2^T = \mathbf{I} - \mathbf{E}_1 \mathbf{E}_1^T, \quad (5.21)$$

due to the orthogonality of the two subspaces, the following covariance matrix is obtained:

$$\mathbf{B}_b = [\mathbf{E}_1 \ (\mathbf{I} - \mathbf{E}_1 \mathbf{E}_1^T)] \begin{bmatrix} \Lambda_1^2 & 0 \\ 0 & \mathbf{B}_{hi} \end{bmatrix} \begin{bmatrix} \mathbf{E}_1^T \\ (\mathbf{I} - \mathbf{E}_1 \mathbf{E}_1^T) \end{bmatrix}, \quad (5.22)$$

where \mathbf{B}_b denotes the result of blending the EOFs with \mathbf{B}_{hi} .

The full covariance matrix \mathbf{B}_b is not calculated explicitly, but the optimisation problem is again pre-conditioned according to \mathbf{B}_b . First, the square root of the covariance matrix is defined as

$$\mathbf{B}_b^{1/2} = [\mathbf{E}_1 \ (\mathbf{I} - \mathbf{E}_1 \mathbf{E}_1^T)] \begin{bmatrix} \Lambda_1 & 0 \\ 0 & \mathbf{B}_{hi}^{1/2} \end{bmatrix}, \quad (5.23)$$

such that $\mathbf{B}_b = \mathbf{B}_b^{1/2} (\mathbf{B}_b^{1/2})^T$. Then, the relationship between the control vector and the increment, $\Delta\mathbf{s}$, is defined in a similar way as in Section 5.2.4 as

$$\Delta\mathbf{s} = \mathbf{B}_b^{1/2} \begin{bmatrix} \boldsymbol{\alpha}_1 \\ \boldsymbol{\alpha}_2 \end{bmatrix}. \quad (5.24)$$

As a consequence, the increment is made up of two components:

$$\Delta\mathbf{s} = \mathbf{E}_1 \boldsymbol{\Lambda}_1 \boldsymbol{\alpha}_1 + (\mathbf{I} - \mathbf{E}_1 \mathbf{E}_1^T) \mathbf{B}_{hi}^{1/2} \boldsymbol{\alpha}_2 \quad (5.25)$$

The first component controls the increment in the space spanned by the EOFs. The second component controls the increment in the subspace that is orthogonal to the EOFs (the null space of \mathbf{B}_e) and has a covariance structure supplied by \mathbf{B}_{hi} .

5.3.2 Horizontal Localisation

Preliminary results using various EOF truncations calculated from a set of $O(100)$ error samples indicate that while the vertical covariances of \mathbf{B}_e are well estimated, the horizontal covariances contain large spurious values between widely separated locations. Similarly, *Houtekamer and Mitchell (1998)* found significant covariances at large separation distances within a low-resolution ensemble Kalman filter (EnKF). As a consequence observations over, for example, Australia would have a significant impact on the analysis increment over Canada. Such a long-range influence of an observation is unrealistic and is caused by sampling error due to the relatively small sample size (~ 100) used to estimate the covariance matrix, \mathbf{B} , with rank of $O(10^6)$. To reduce this problem, the assumption is imposed that the horizontal covariances beyond a specified separation distance should become very small. This assumption is used in place of the more constraining assumptions that the correlations should fit a specified functional form and be horizontally homogeneous and isotropic. The assumption that the horizontal correlations are local constrains the form of \mathbf{B} , thus decreasing the total number of degrees of freedom to be estimated. This results in increasing the maximum rank of the covariance matrix that can be estimated from

a given set of error samples. In the following sections, two methods of limiting the horizontal extent of the background error covariances are examined (an additional method is described in Appendix G).

Localisation of the Global EOFs with a Localising Correlation Function

The first method of damping covariances at large separation distances is similar to a localisation example presented in Section 4d of *Gaspari and Cohn* (1999). They state that the product of the estimated horizontal covariance function with another covariance function is a valid (that is, positive semi-definite) covariance function. Therefore, the desired localisation can be obtained by multiplying the estimated horizontal covariances by a space-limited correlation function. An ideal correlation function for localisation would preserve the local structure of the original covariances while completely suppressing the covariances at distances beyond which the original estimates are believed to be significantly different from zero. A top-hat function would be ideal, except that it is not positive semi-definite.

For illustration, this approach is applied to the case of localising the covariance matrix of a one-dimensional function with the truncated EOF expansion

$$\mathbf{B}_e = \mathbf{E}\Lambda_e^2\mathbf{E}^T. \quad (5.26)$$

For convenience, the EOFs, after scaling by the square root of their corresponding eigenvalues, are rewritten as

$$\mathbf{F} = \mathbf{E}\Lambda_e. \quad (5.27)$$

Then, if the i th column of \mathbf{F} is denoted by \mathbf{f}_i , (5.26) can be rewritten as

$$\mathbf{B}_e = \mathbf{F}\mathbf{F}^T = \sum_{i=1}^{N_e} \mathbf{f}_i\mathbf{f}_i^T. \quad (5.28)$$

The EOF representation of the covariance matrix can therefore be expressed as a sum of matrices, each the outer product of the scaled EOF with itself. As described in Appendix G, this localisation approach could equally be applied to the full covariance

matrix estimated directly from the error samples instead of using a truncated EOF expansion.

Assuming the correlation matrix used for localisation is homogeneous and isotropic, it is denoted as

$$\mathbf{L} = \begin{bmatrix} l_1 & l_2 & l_3 & \dots \\ l_2 & l_1 & l_2 & \dots \\ l_3 & l_2 & l_1 & \dots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix}, \quad (5.29)$$

where $l_1 = 1$ and all off-diagonal elements are less than one. For convenience, the following definition of the $\text{diag}()$ operator is introduced:

$$\text{diag}(\mathbf{f}_j) \equiv \begin{bmatrix} F_{1j} & 0 & 0 & \dots \\ 0 & F_{2j} & 0 & \dots \\ 0 & 0 & F_{3j} & \dots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix}, \quad (5.30)$$

where F_{ij} is the element from the i th row and j th column of \mathbf{F} . The localised background error covariance matrix can then be defined as

$$\begin{aligned} \mathbf{B}_l &= \sum_{i=1}^{N_e} \text{diag}(\mathbf{f}_i) \mathbf{L} \text{diag}(\mathbf{f}_i) \\ &= \begin{bmatrix} \text{diag}(\mathbf{f}_1)\mathbf{L}^{1/2} & \text{diag}(\mathbf{f}_2)\mathbf{L}^{1/2} & \dots \end{bmatrix} \begin{bmatrix} \mathbf{L}^{T/2}\text{diag}(\mathbf{f}_1) \\ \mathbf{L}^{T/2}\text{diag}(\mathbf{f}_2) \\ \vdots \end{bmatrix}, \end{aligned} \quad (5.31)$$

which when expanded gives the desired result

$$\mathbf{B}_l = \sum_{i=1}^{N_e} \begin{bmatrix} l_1 F_{1i}^2 & l_2 F_{1i} F_{2i} & l_3 F_{1i} F_{3i} & \dots \\ l_2 F_{2i} F_{1i} & l_1 F_{2i}^2 & l_2 F_{2i} F_{3i} & \dots \\ l_3 F_{3i} F_{1i} & l_2 F_{3i} F_{2i} & l_1 F_{3i}^2 & \dots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix} = \begin{bmatrix} l_1 B_{11} & l_2 B_{12} & l_3 B_{13} & \dots \\ l_2 B_{21} & l_1 B_{22} & l_2 B_{23} & \dots \\ l_3 B_{31} & l_2 B_{32} & l_1 B_{33} & \dots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix}, \quad (5.32)$$

where B_{ij} is the element from the i th row and j th column of the covariance matrix \mathbf{B}_e . Therefore, the definition of \mathbf{B}_l in (5.31) results in a covariance matrix where each column of the original covariance matrix has been multiplied element-wise by the localising correlation function centred on the diagonal. Since the maximum of the correlation function is aligned with the diagonal, the variance of the original covariance matrix is unaffected. Conversely, the local horizontal length scale (defined as $(-c)^{-1/2}$ where c is the curvature of the correlation function evaluated at the origin) of the localised covariance matrix, L_l , is reduced according to

$$L_l = \frac{L_e L_c}{(L_e^2 + L_c^2)^{1/2}}, \quad (5.33)$$

where L_e is the length scale for the EOF-based covariance matrix and L_c is the length scale of the localising correlation matrix. Because the EOF-based covariances are anisotropic, L_e and L_l depend on the direction in which the curvature is calculated.

The square root of \mathbf{B}_l can then be used to define the relationship between the control vector and the increment as before

$$\Delta \mathbf{s} = \mathbf{B}_l^{1/2} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \end{bmatrix} = \left[\text{diag}(\mathbf{f}_1) \mathbf{L}^{1/2} \quad \text{diag}(\mathbf{f}_2) \mathbf{L}^{1/2} \quad \dots \right] \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \end{bmatrix}, \quad (5.34)$$

such that $\mathbf{B}_l = \mathbf{B}_l^{1/2} (\mathbf{B}_l^{1/2})^T$. Now the vector of controls, α_i , determines the spatially varying amplitude of the i th EOF, \mathbf{f}_i . Each vector α_i has dimension equal to the horizontal grid and controls the i th EOF. Therefore, the total dimension of the control vector is equal to the number of EOFs times the dimension of the horizontal grid. The EOFs, however, do not need to be explicitly localised and then stored, which would likely be infeasible (for example, in the configuration described in Section 5.4, 47520 localised EOFs would need to be stored at full resolution). Instead, only the non-localised EOFs and the square root of the localising correlation matrix need to be stored and then applied to the control vector according to (5.34) during the minimisation to construct the increment.

A modification of this approach that employs an iterative eigendecomposition algorithm is described in Appendix G.

Localisation of Error Samples with a Discrete set of Masks

Another approach for damping the covariances at large horizontal separation distances is based on defining a set of discrete localised regions over the globe between which the covariance should be zero. These regions are separated by transition zones in which the background error is correlated, to some extent, with the errors in the neighbouring regions. To impose zero correlation between the regions, the individual error samples are each multiplied by several two-dimensional over-lapping localisation masks that correspond to each of the localised regions. The result is that several localised error samples, each with nonzero values only in a single region and its neighbouring transition zones, are produced from each original sample. This, in effect, expands the size of the ensemble and therefore also increases the maximum number of EOFs that may be calculated from a given set of error samples. Any problems of negative eigenvalues that may arise when modifying a covariance matrix directly are also avoided, regardless of the shape of the masks. The localisation masks are chosen to have limited horizontal extent and such that the original sample variance is unaffected.

To illustrate, assume a one-dimensional random function of space, $g(x)$, has a spatial covariance function from which a set of random samples is denoted $g_i(x)$, $i = 1, \dots, N_b$. Each sample is individually multiplied by each of the localisation masks $l_j(x)$, $j = 1, \dots, N_l$, to produce $N_b \times N_l$ "localised" samples from the original set of N_b samples. To conserve variance, the variance of the localised function, $\sigma_l(x)^2$, given by

$$\sigma_l(x)^2 = \frac{1}{(N_b - 1) N_l} \sum_{i=1}^{N_b} \sum_{j=1}^{N_l} [l_j(x)g_i(x)]^2 = \frac{\sigma(x)^2}{N_l} \sum_{j=1}^{N_l} l_j(x)^2. \quad (5.35)$$

must equal the original sample variance, $\sigma(x)^2$. Therefore, any set of masks that

satisfies

$$\sum_{j=1}^{N_l} l_j(x)^2 = N_l \quad (5.36)$$

for all values of x will not affect the sample variance.

A set of masks, that each vary from a region of constant value to a region of zero through a transition zone as the square root of position, were applied in an idealised example. Figure 5.4 shows that the horizontal extent of the localised covariance function is limited by the extent of the non-zero region of the masks. Also, within the region of constant mask value the covariance is unchanged.

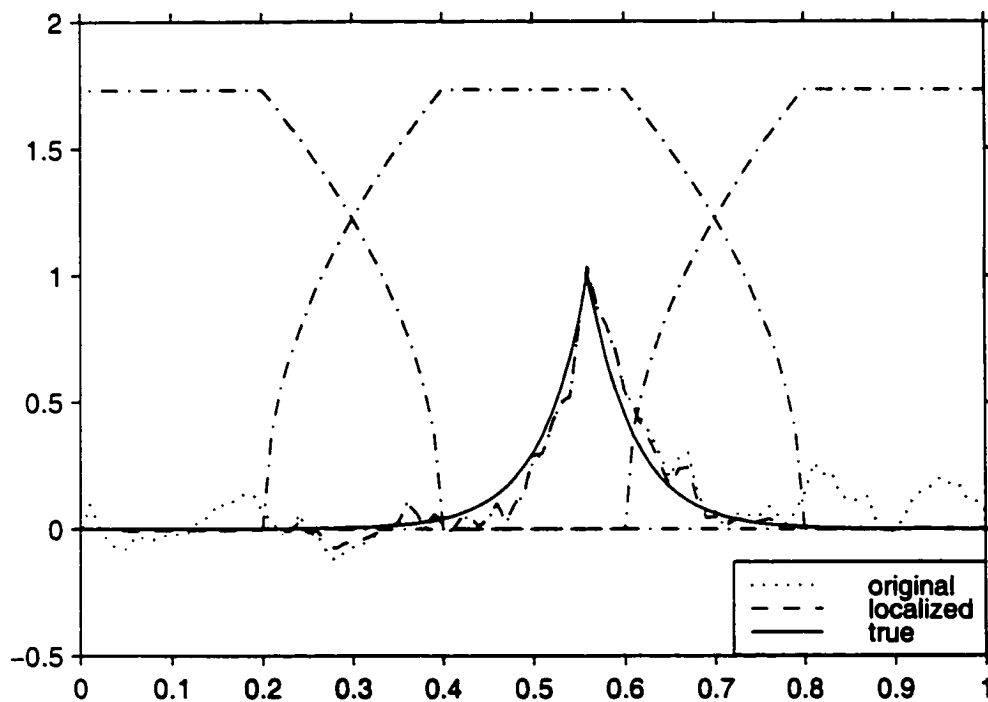


Figure 5.4: Example of the effect of localising masks (dash-dotted lines) on estimating spatial covariances. The localised covariance (dashed) is equal to the original (dotted) within the region of constant mask value (0.4-0.6) and goes to zero through transition zone. The true covariance function used to generate the 500 realizations used for the calculation is also shown (solid). The localised covariance function is clearly more similar to the truth than the original sample covariance function which has large spurious values in the tails.

5.4 Results Using EOFs for Stationary Correlations

A set of numerical experiments was performed with a version of the Canadian 3D-Var that is still under development. As an initial attempt, EOFs of the stationary correlation matrix were calculated from the same set of error samples as used for calculating \mathbf{B}_{hi} . The only difference between the EOF-based covariance matrix and \mathbf{B}_{hi} is that the horizontal correlations were not constrained to be homogeneous and isotropic. The structure functions obtained using the EOFs and \mathbf{B}_{hi} are presented to evaluate the effect of relaxing the assumption of homogeneous and isotropic background error correlations and instead localising the horizontal correlations and/or using correlations that are homogeneous and isotropic only in the null-space of the EOFs.

5.4.1 Details of Implementation

First, 120 forecast differences from December 1998 and January 1999 were calculated from the operational global assimilation cycle at the Canadian Meteorological Centre using the GEM forecast model (*Côté et al.*, 1998). These error samples were horizontally interpolated from the 400 by 200 point grid of the forecast model to a 120 by 60 point Gaussian grid and the sample mean removed for all variables. Then, some of the same steps used for calculating the covariance matrix \mathbf{B}_{hi} were followed:

- the balanced components of the original error samples were removed to obtain the unbalanced variables (χ', T', p') ; and
- the unbalanced variables were normalised by the three-dimensional sample standard deviations.

The Euclidean norm was chosen for calculating the EOFs of the correlation matrix for the unbalanced variables. Some problems with using this norm are discussed

below. Since the unbalanced variables are assumed to be uncorrelated, the EOFs could be calculated separately for each variable.

The full set of 119 EOFs were calculated from the normalised ensemble using singular value decomposition, as described in Appendix F. To examine the effectiveness of localising these EOFs, a correlation matrix with homogeneous and isotropic Gaussian correlations was calculated in spectral space at the spectral truncation T31 (about half the resolution of the EOFs). The length scale of these Gaussian correlation functions was set to 2000 km for the streamfunction and unbalanced velocity potential, 1000 km for the unbalanced temperature and humidity, and 1500 km for the unbalanced surface pressure. These length scales were chosen to be significantly larger than those given by *Gauthier et al.* (1998) for the isotropic correlations of the same variables. Consequently, the local shape of the correlations should not be significantly affected.

To evaluate the use of discrete localisation masks, the normalised error samples were each multiplied by three masks producing a set of 360 localised samples. One mask each is active over the northern and southern extratropics with constant regions poleward of 45° latitude. The third mask has a constant region equator-ward of about 15° . The regions in between vary as the square-root of position and act as transition zones (see Figure 5.4). The effect of these masks is to eliminate covariances between the extratropics (poleward of 45°) and the tropics (equator-ward of 15°). The first 150 EOFs from these localised error samples were calculated.

5.4.2 The Structure Functions

Several single observation experiments were performed to evaluate the impact on the structure functions of varying the EOF truncation, applying the localisation methods, and blending the EOFs with \mathbf{B}_{hi} . The resulting structure functions were also compared with those obtained using only \mathbf{B}_{hi} with respect to their ability to capture the effects of baroclinicity and orography.

Effect of EOF truncation

As expected, truncation of the EOF expansion generally results in structure functions that are spatially smoother and give less weight to the observation, due to a decrease in the background error variance. This overall reduction in the background error variance could be compensated for by scaling the variances. However, truncation also appears to reduce the variance of the wind field significantly more than the mass field. This leads to an imbalance between the weight given to wind and height observations, when compared with the full covariance matrix, that cannot be eliminated by a simple scaling factor. For this reason most of the results shown below were obtained using the full set of $(N_b - 1)$ EOFs.

The imbalance between the weight given to wind and height observations can be explained by considering the relationship between the mass and wind increments. In the extratropics, the analysis increment of streamfunction contributes substantially to both the wind and the mass fields (through the balance operators). Since the Euclidean norm was chosen for calculating the EOFs of the correlation matrix for the unbalanced variables (which includes the streamfunction), they optimally account for the correlation structure of the errors in streamfunction. Because the balanced mass field is proportional to the streamfunction ($f\psi$, where f is the Coriolis parameter) the error correlation of mass in the extratropics is also well represented. The wind field increment, however, is primarily composed of the spatial derivatives of the streamfunction and thus has relatively larger contributions from the high wave numbers of streamfunction, as compared with the mass field. The leading EOFs, therefore, will not optimally capture the correlation structure of the wind field errors. As a consequence, when the EOF expansion is truncated ($N_e < (N_b - 1)$) the resulting error variance for the wind field is underestimated relative to that for the balanced mass field and simple scaling of the streamfunction eigenvalues can not correct this problem.

The signal captured by a truncated set of EOFs depends on the norm used for their calculation. Because the EOFs could be calculated separately for each variable

type (due to the use of balance operators for the between-variable relationships) it was originally thought that the Euclidean norm would be sufficient. However, consider the the energy norm, which treats each variable in terms of the common unit of energy. Using this norm, the inner product of the terms involving the three-dimensional streamfunction field is given by

$$\langle \mathbf{s}(\psi), \mathbf{s}(\psi) \rangle_E = \int_S \left\{ \frac{RT_r}{p_r^2} (p_s)_\psi^2 + \int_0^{p_s} \left[\nabla\psi \cdot \nabla\psi + \nabla\chi_\psi \cdot \nabla\chi_\psi + \frac{c_p}{T_r} T_\psi^2 \right] dp \right\} dS, \quad (5.37)$$

where T_r and p_r are a constant reference temperature and pressure, respectively; c_p and R are the specific heat at constant pressure and gas constant of dry air, respectively; and S is the horizontal domain. The variables χ_ψ , T_ψ , and $(p_s)_\psi$ are the balanced components derived from ψ using the balance operators from Section 5.2.4. If this inner product was used to define the norm for calculating the EOFs, then the EOFs would optimally capture this weighted sum of the rotational and balanced divergent wind components and the balanced temperature and surface pressure error correlations. By placing more weight on the higher wave numbers of ψ (through the terms involving $\nabla\psi$), use of the energy norm would likely correct, to some extent, the underestimation of the background error variance for wind that occurs when the Euclidean norm for only streamfunction is used. However, since the spectrum for streamfunction (and therefore also balanced mass) is more red than the wind spectrum, any truncation of the EOF expansion would likely still decrease the background error variance for wind more than for balanced mass. Also, note that the energy norm involves the integral over pressure. Therefore, the samples are weighted according to the mass contained between adjacent vertical levels.

Impact of applying localisation

Structure functions were obtained after applying each of the methods described in Section 5.3.2 for localising the horizontal covariances. Figure 5.5 shows the structure function of geopotential height of the 850 hPa pressure level resulting from a height

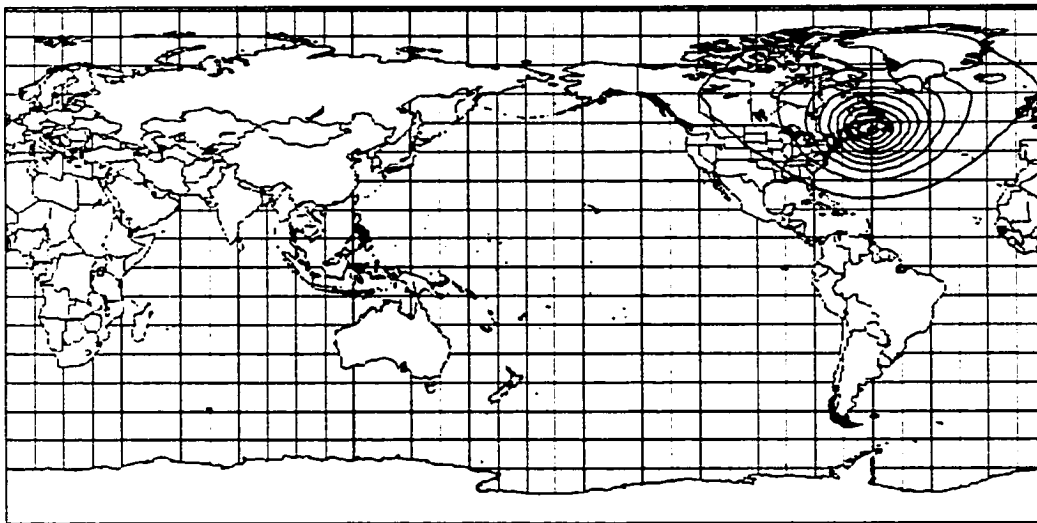


Figure 5.5: Horizontal structure function of geopotential height at 850 hPa resulting from a height observations at the same level and located at 45°N, 60°W. This figure shows the result of using the background error covariance matrix \mathbf{B}_{hi} , based on homogeneous and isotropic correlations. Plotted results in this and subsequent figures were normalised to give a maximum value of one with the contour interval equal to 0.1.

observation at the same level in the northern extratropics using \mathbf{B}_{hi} . In Figure 5.6, results are shown using the original EOFs and the EOFs after localising with the large scale Gaussian correlation matrix. The results from the original EOFs show significant height increments over the entire globe, especially over the southern oceans. The localised EOFs effectively damp these horizontal covariances while preserving the structure in the vicinity of the observation.

Similarly, Figure 5.7 shows the effect of using a set of discrete localisation masks applied directly to the error samples. The masks effectively dampen the analysis increment in the tropics and the southern hemisphere, which was the goal of choosing masks located over the tropical and extratropical regions. As already stated, the maximum number of EOFs are tripled by using three localisation masks. Due to the overlapping transition zones (as shown in Figure 5.4), the resulting covariance

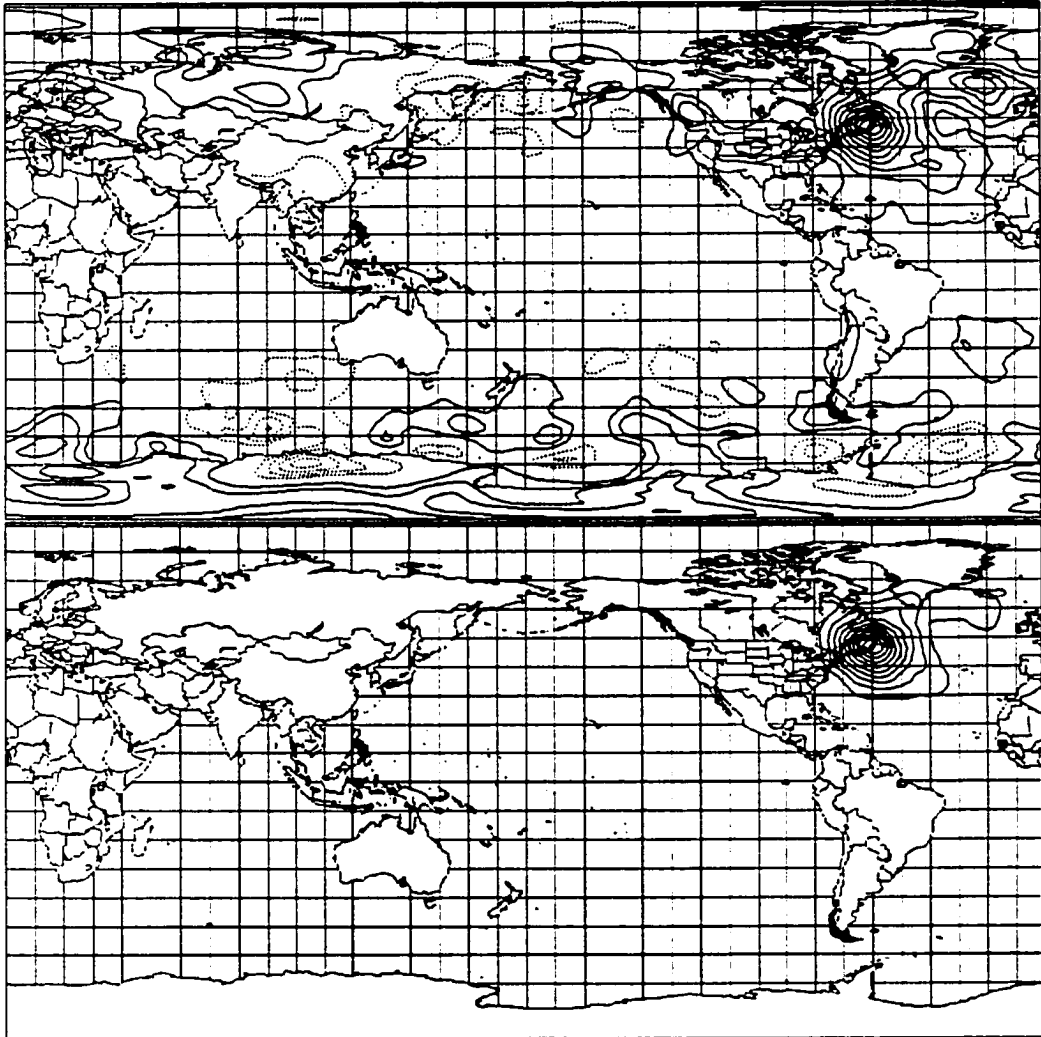


Figure 5.6: Same as Figure 5.5, but using EOF-based background error covariance matrices. Top: covariance matrix obtained using full set of EOFs for the correlation matrix. Bottom: same as top panel, but with Gaussian localisation masks.

matrix still possesses off-diagonal blocks that link neighbouring masked regions. As a consequence, the EOFs themselves are not partitioned according to the localised regions, but generally still span the entire globe. Consequently, the number of EOFs required to explain a given percentage of the total variance will be approximately three times greater than for the EOFs of the unmasked error samples. This explains why 150 EOFs from the masked error samples still give broader horizontal covariance structure within the northern extratropics than 60 EOFs from the unmasked samples.

Effect of blending EOFs with \mathbf{B}_{hi}

Figure 5.8 shows the effect of blending the EOFs with the projection of \mathbf{B}_{hi} in the orthogonal subspace, as described in Section 5.3.1. A zonal-vertical cross-section is shown for the structure function of geopotential height resulting from a geopotential height observation at 500 hPa. The vertical coordinate is roughly logarithmic in pressure (actually logarithmic in the eta value that defines the vertical levels of the analysis grid). The structure function obtained using only \mathbf{B}_{hi} exhibits a perfectly isotropic shape in the horizontal, as expected, whereas the EOF-based covariance matrix produces a significantly anisotropic shape. Blending of these two covariance matrices produces a structure function that appears to be slightly less noisy than using the EOFs alone while still retaining some of the anisotropy from the EOFs.

Baroclinic effects

Figure 5.9 shows the zonal-vertical cross-section of the structure functions for wind resulting from a wind observation at 850 hPa. The observation is located off the east coast of Canada in a region of frequent baroclinic activity. The structure function obtained using \mathbf{B}_{hi} again has a horizontally isotropic shape, whereas the result obtained using the EOF-based covariance function has a significantly tilted shape. A similar tilted shape is shown in Figure 5.10 for the structure function of geopotential height resulting from a height observation at 850 hPa. When using the EOFs, the structure functions, especially for wind, extend much higher into the atmosphere. These results

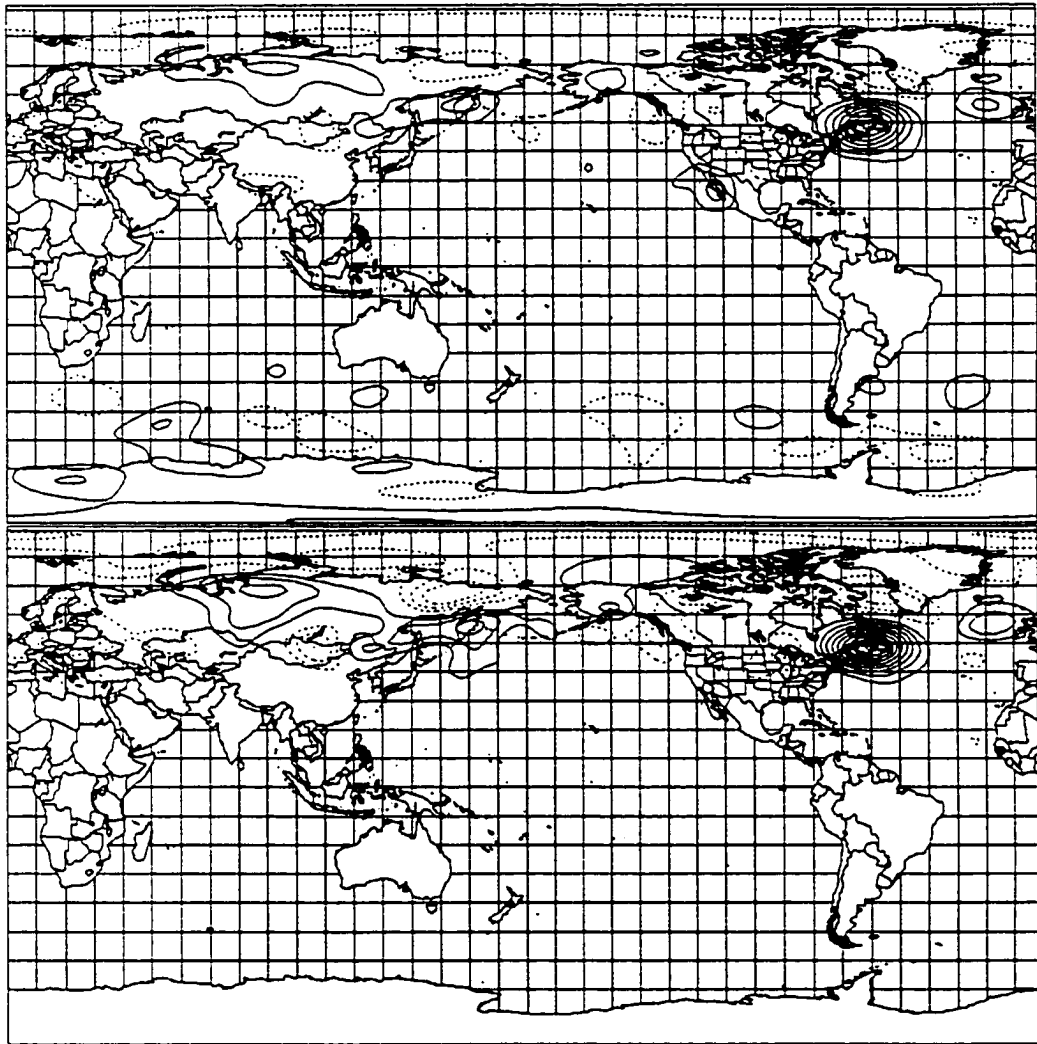


Figure 5.7: Horizontal structure functions of geopotential height at 500 hPa for a height observations at the same pressure level and located at 50°N , 60°W . Top: result using 60 EOFs for the auto-correlation matrix. Bottom: result using 150 EOFs after applying three discrete localisation masks to the error samples (localising the extratropics from the tropics).

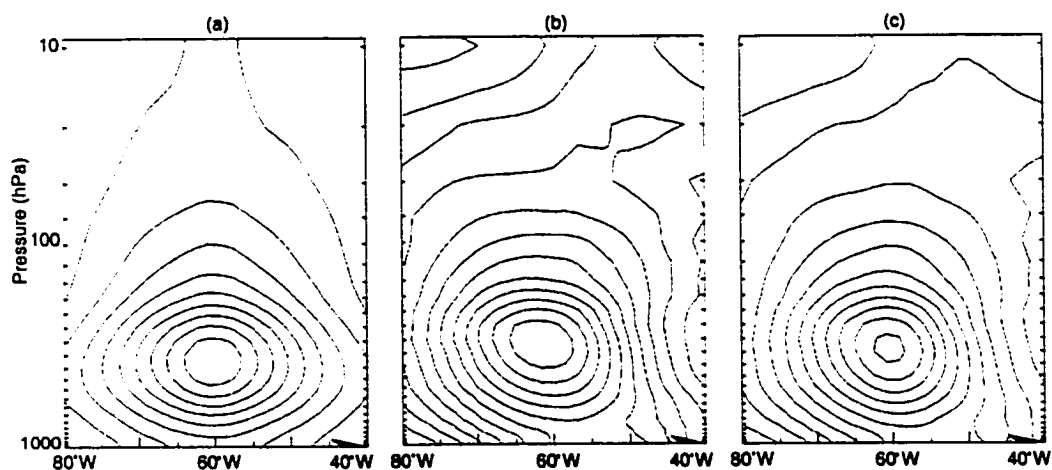


Figure 5.8: West to east vertical cross-section of the structure function of geopotential height corresponding to a height observation at the 500 hPa and located at 50°N , 60°W . The cross-sections span 40° longitude. Left: background error covariance matrix based on homogeneous and isotropic correlations. Centre: EOF-based covariance matrix. Right: Covariance matrix resulting from blending EOFs with homogeneous and isotropic correlations.

are consistent with those obtained using four-dimensional assimilation systems and predicted from the theory of baroclinically unstable modes discussed earlier.

The consequence of such tilted structure functions is that observations at a single level will result in analysis increments that will either enhance, attenuate, or simply shift the phase of the unstable component in the background state. On the contrary, with isotropic increments the changes to the background state due to single level observations would not project well onto the unstable modes and therefore would not be as effective in producing appropriate corrections in baroclinically active situations.

Orographic effects

Figure 5.11 shows the effects of orography on the structure function of wind and geopotential height fields near the surface resulting from a zonal wind observation at 500 hPa over the Rockies. Use of the conventional covariance matrix results in a horizontal structure with symmetric positive and negative maxima in geopotential

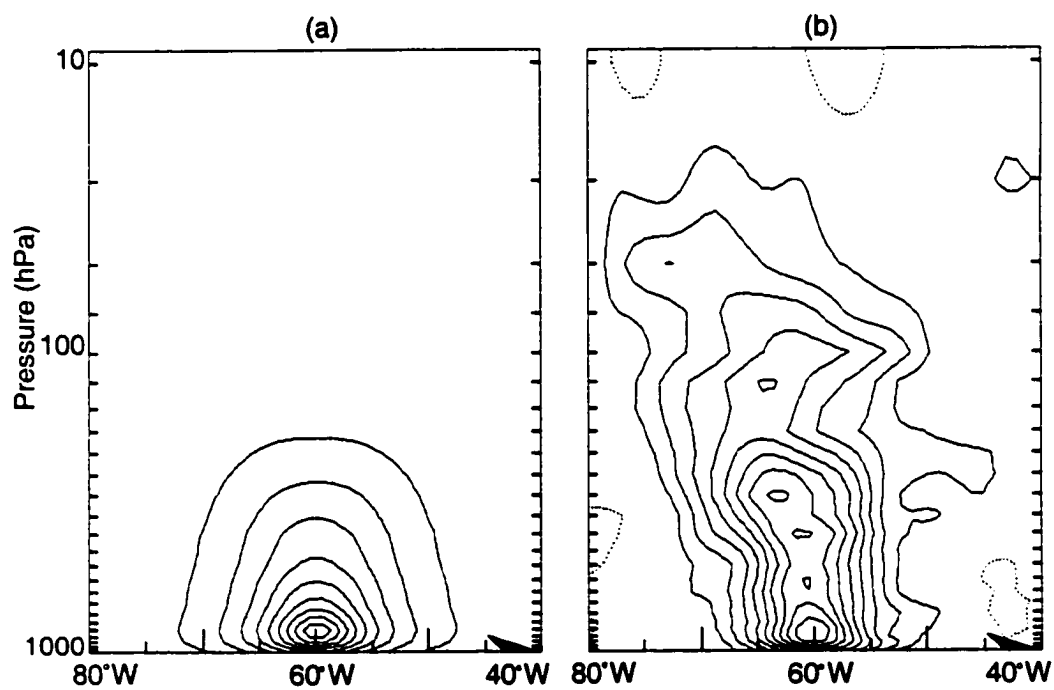


Figure 5.9: Zonal-vertical cross-sections of the zonal wind for a wind observations at 850 hPa near the east coast of Canada (50°N , 60°W). (a) Result using \mathbf{B}_{hi} , (b) result using EOF-based covariance matrix showing anisotropic effect of baroclinic forcing.

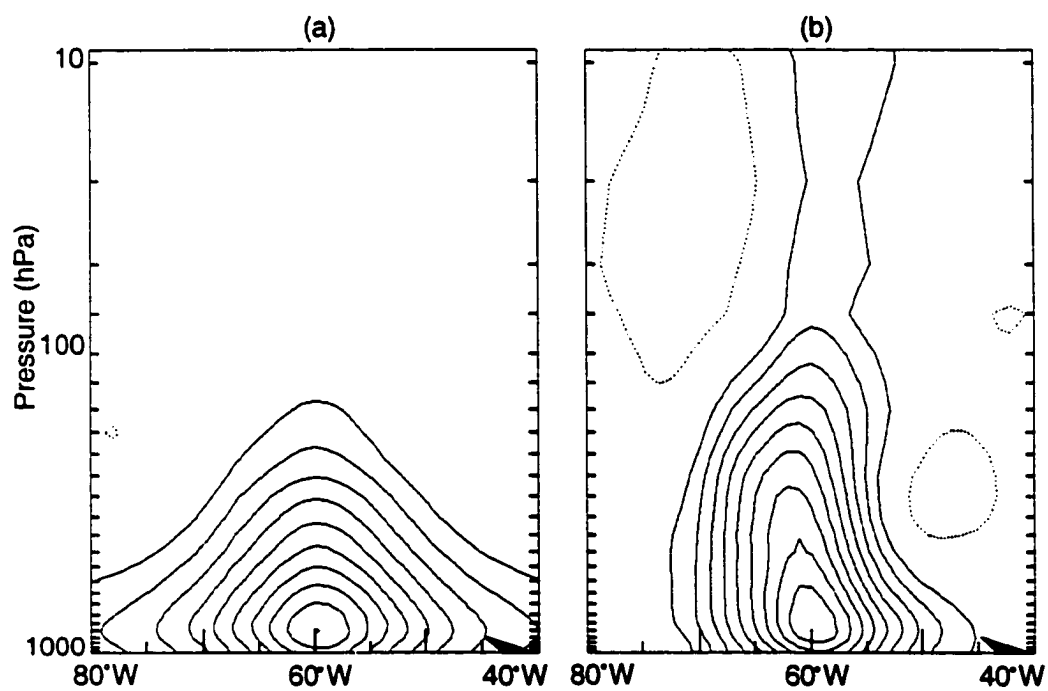


Figure 5.10: Same as Figure 5.9, except showing the structure function of geopotential height corresponding to a height observation.

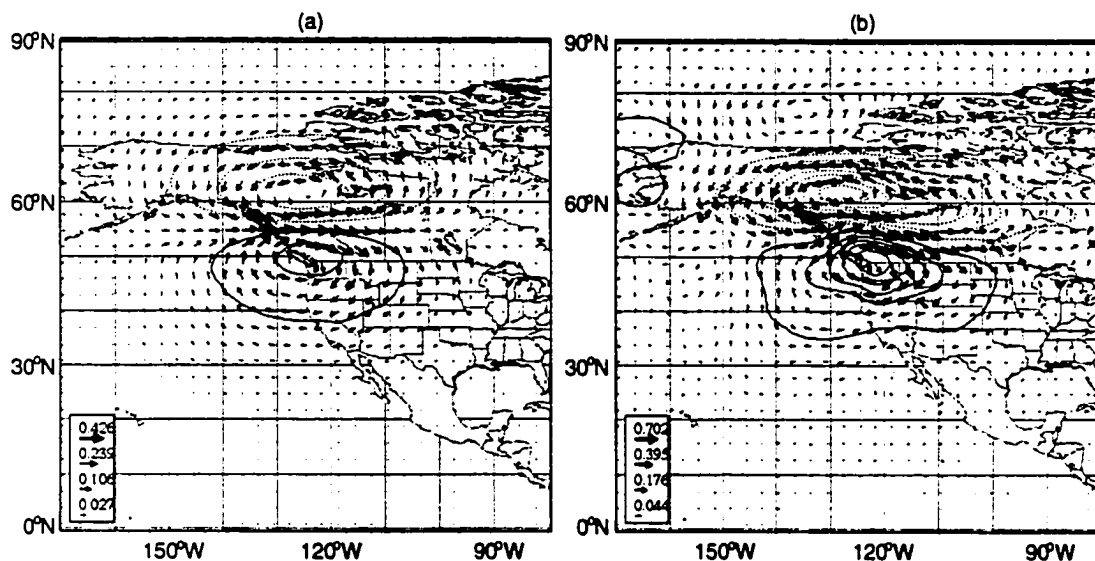


Figure 5.11: Structure function of the 850 hPa wind and geopotential height (contours) fields resulting from a zonal wind observation at 500 hPa over the Rockies (125° W 55° N). (a) Result using \mathbf{B}_{hi} , (b) result using EOF-based covariance matrix showing the effect of orography.

height and streamfunction increments to the south and north of the wind observation, respectively. The use of the EOF-based covariance matrix leads to a similar result, however, it appears that a ridge of high pressure aligned with the mountain range is superimposed on the positive and negative geopotential height maxima. This is consistent with the winter results of *Ringler and Cook* (1999) discussed earlier. Because of this ridge, the wind field increment along the same latitude as the observation has a northward component on the windward slope and a southward component on the leeward slope. The results in Figure 5.11b also show apparently spurious wind increments away from the observation (especially to the north and west), whereas by assuming homogeneous and isotropic correlations the pattern in Figure 5.11a diminishes in a predictable manner away from the observation.

5.5 Discussion and Conclusions

5.5.1 Limitations

The use of EOFs to modify the representation of the background error covariance matrix presented in the previous section suffers from some limitations.

The most obvious limitation is the source of forecast error information used to estimate the EOFs. The size of the ensemble may be too small to accurately estimate the correlation functions, especially for wind, or the NMC method may not be capable of providing accurate information on the nonhomogeneous and anisotropic correlations. Specifically, the temporal mismatch of assuming 48 h minus 24 h forecast differences are representative of the errors in the 6 h forecasts may limit the use of the correlation information. An alternative approach, recently implemented at ECMWF (*M. Fisher, ECMWF, 1999, personal communication*), may provide more appropriate information on the correlations. This approach uses an ensemble of assimilation cycles that each use independently perturbed values for the observations and uncertain model parameters, similar to the EnKF. The conventional stationary background error covariance matrix is used to perform the analysis step for each ensemble member. The new background error statistics are then estimated from the spread in the 6 h forecasts over a period of several weeks.

The method of localising with a large scale correlation function also suffers from some limitations. Since Gaussian correlations were used to damp the correlations at large separation distances, the correlations will not be forced completely to zero. It may be possible to use a more effective damping correlation function, such as those used by *Gaspari and Cohn (1999)*. Also, because this localisation method does not produce an orthogonal set of basis functions, the blending approach can not easily be applied. Using the alternative approach of *Zupanski (1999)*, but with a positive-definite localisation function as described in Appendix G, would allow an orthogonal basis to be calculated. Then, by blending the truncated EOF expansion with the covariance matrix \mathbf{B}_{hi} , it would be possible, for example, to rely more on the isotropic

and homogeneous correlations for the small scales, while using the EOFs primarily to represent the large-scale correlation structure. This may improve the covariance matrix, especially with respect to the wind structure functions.

5.5.2 EOFs in the Ensemble KF and Integration with 4D-Var

An alternative to assuming completely stationary background error statistics is to incorporate the 3D-Var algorithm within an EnKF scheme. The EnKF, introduced by *Evensen (1994)*, produces information on forecast error using a Monte Carlo approach. The algorithm is based on running an ensemble of assimilation cycles in parallel, each using observations that have been independently and randomly perturbed from their observed values (*Burgers et al., 1998*). These perturbations are generated such that they are consistent with the observation error statistics. Other uncertain parameters within the model can be perturbed according to their associated uncertainties. It is also possible to include the effect of model error, however, this requires an estimate of the model error statistics. *Mitchell and Houtekamer (1999)* describe an adaptive approach for incorporating model error in an EnKF. The background error covariance matrix is then estimated at each analysis time from the spread in the ensemble of short-term forecasts. The same covariance matrix is used to perform the assimilation for each ensemble member. This results in non-stationary background error statistics that depend on the recent history of the flow field. A limitation of this approach, and the reason it has not yet been implemented operationally, is the computational cost of running enough parallel assimilation cycles to sufficiently span the background error space. The minimum number of ensemble members required for a realistic NWP system has yet to be determined.

The advantage of using the 3D-Var and an EOF-based covariance matrix within the EnKF is that the methods for horizontal localisation and blending with a full-rank covariance matrix presented earlier could also be used, if necessary. The blending could be performed with a slowly evolving, yet full-rank covariance matrix, \mathbf{B}_{hi} . Instead of using the NMC method, this covariance matrix could be estimated from,

say, the spread of all the 6 h forecasts over the previous two weeks. The blended covariance matrix would be flow-dependent, nonhomogeneous, and anisotropic in the subspace spanned by the ensemble members and slowly evolving, homogeneous, and isotropic in the orthogonal subspace.

The standard implementation of 4D-Var relies on the error covariance matrix associated with the background state at the beginning of the assimilation window. Because of the relatively short assimilation window typically used for 4D-Var, the structure functions at the end of the window can be significantly influenced by the covariance matrix used at the beginning of the window. However, the flow-dependent covariance matrix of the errors in the background state can not be easily calculated explicitly as part of the 4D-Var algorithm. *Fisher (1998)* found that use of a simplified KF to provide a low-dimensional, flow-dependent estimate for the background error statistics within a 4D-Var system provided a small, yet statistically significant improvement to the subsequent forecasts over using the conventional stationary covariance matrix. In a similar fashion, the EOF-based covariance matrix calculated as part of an EnKF system could be used to provide flow-dependent error statistics for the initial conditions in a 4D-Var system. If the EnKF is used primarily for this purpose, it may be sufficient to implement the forecast model and analysis at a degraded horizontal resolution for the ensemble members to increase efficiency. Therefore, the approaches presented in this chapter for using EOFs to model nonhomogeneous and anisotropic background covariances could be used to improve the background error statistics used in a 4D-Var system.

Chapter 6

A Sub-Optimal Assimilation Scheme for Nonlinear Models

6.1 Introduction

Variational approaches to data assimilation rely on the adjoint of the linearised forward model for calculating the gradient of the cost function, J , with respect to the controls. This allows J to be efficiently minimised using a standard optimisation algorithm. The full adjoint model is the transpose of the tangent linear model (TLM) of the nonlinear model equations (see Appendix B). Typically, the required derivatives are derived analytically from the discrete form of the model equations. For sequential assimilation systems, such as 3D-Var described in the previous chapter, the relationship between the controls and the model counterpart to the observations may be quite simple. For an interior pressure observation, for example, this may include only a spatial interpolation operator and possibly integration of the hydrostatic operator to transform the temperature and surface pressure included in the control vector into pressure. For such systems, formulation of the adjoint model is relatively straight forward. In four-dimensional assimilation schemes, however, formulation of the adjoint model for large nonlinear time-dependent models is time consuming and

subject to errors (*Thacker, 1992*). This is due to the large number of state variables and corresponding nonlinear equations. Iterative solvers employed as part of the temporal model integration can cause particular difficulties since the values of the variables at each iteration need to be stored to calculate the adjoint. Also, without careful consideration, the linearisation of highly nonlinear processes, such as cloud and precipitation parameterisations in atmospheric models, can cause unrealistic results (*Janiskova et al., 1999*). As numerical models become more complex these sources of difficulty will become increasingly important.

The quadratically nonlinear terms that dominate the equations of motion (neglecting the often highly nonlinear parameterisation schemes) have to be linearised with respect to each of the variables and therefore produce twice the number of terms in the TLM and adjoint model. Consequently, the linearised models can be about twice as computationally expensive to run as the nonlinear version. For large models the expense of evaluating the cost function gradient may limit the period over which data can be assimilated or reduce the number of allowable iterations used to reach the cost function minimum. Also, especially with models used in research applications, the numerical schemes or parameterisations may be subject to frequent modifications. Such changes would require corresponding, and often non-trivial, changes to the adjoint code.

To avoid the effort required in formulating the adjoint model, some effort has gone into developing automatic adjoint generators (e.g. *Giering and Kaminski, 1996*). These are essentially symbolic program manipulators. They accept the original ocean model coded in a programming language (such as Fortran) and process it to produce the code of the corresponding adjoint model. However, they have not been widely used and their success still largely depends on how the original numerical model is coded.

For the above reasons, a method for obtaining a sub-optimal adjoint model which is efficient to run and avoids the manual process of deriving the full adjoint model code would be desirable. The approximate adjoint model would be useful in other

applications where adjoint models are used, such as model tuning, sensitivity analysis, and determination of singular vectors (see e.g. *Moore and Farrell, 1993; Buizza et al., 1993; Errico and Vukicevic, 1992*). The goal of this chapter is to develop and apply an approach for quickly obtaining a sub-optimal adjoint-based assimilation scheme. The adjoint model is calculated within a reduced dimension subspace spanned by a truncated set of empirical orthogonal functions (EOFs). The next section gives an overview of existing studies related to reduced dimension models and sub-optimal assimilation schemes. In Section 6.3 the proposed scheme for obtaining an approximate adjoint model is presented. The results from an identical twin experiment are presented in Section 6.4 to demonstrate the application and effectiveness of the method. The final section concludes with a discussion of some of the limitations and possible extensions to the method.

6.2 Sub-Optimal Assimilation Schemes

With the incremental approach to variational data assimilation, introduced in the previous chapter, a simpler numerical model is used for calculating the correction to a nonlinear model solution that is obtained from an *a priori* estimate of the controls (*Courtier et al., 1994*). Typically this involves using a linearised version of the original numerical model at a degraded resolution and possibly with simplified parameterisations. The result is an increase in computational efficiency and a decrease in the effort required to formulate the adjoint model. The increment to the controls is obtained by using the linear model to project the correction into observation space during minimisation of the cost function, but the initial model-data misfits are calculated with respect to the nonlinear model solution. This inner loop may be followed by an iteration of the outer loop in which the increment is used to calculate an updated nonlinear model solution. The process is repeated, usually with several iterations of the outer loop, until convergence. For example, *Thompson et al. (1998)* used a linear

tidal model to estimate the open boundary conditions for a more sophisticated nonlinear model for the Gulf of St. Lawrence. The results were significantly improved over those obtained using the linear model alone, but the effort and expense of using the adjoint model of the nonlinear model was avoided. Many sub-optimal assimilation schemes follow this basic approach of using a simpler model for estimating a correction to a nonlinear model solution.

Statistical approaches have also been applied for obtaining simplified dynamical models, often for climate prediction applications. *Selten* (1997) used a statistical approach in experiments with a barotropic atmospheric model. Using output from long model runs, he determined the optimal coefficients for the linear and quadratic terms of a low-dimensional statistical model. The model variables were the amplitudes of a truncated set of EOFs. Alternatively, *DaCosta and Vautard* (1997) used an approach based on estimating the mean of the individual terms in the potential vorticity equation using a low-dimensional representation of the state vector. The state vector was based on a truncated set of principal components calculated from a large data set of meteorological analyses. To estimate the dependence of each vorticity tendency term on the low-dimensional state, the meteorological analyses were binned according to discrete regions in the low-dimensional state space and the average of each term in the potential vorticity equation calculated for each region. The model was then integrated using the average of the tendency terms corresponding to the binned low-dimensional states close to the present model state. Unlike other statistical approaches this approach can capture the basic dynamics of a system that may alternate between multiple quasi-stationary flow regimes.

Assimilating data into large, nonlinear models with the standard Kalman filter is often computationally infeasible. Therefore several studies have focused on developing approximate methods. *Fisher* (1998), *Fukumori and Malanotte-Rizzolli* (1995), *Cane et al.* (1996), *Dowd and Thompson* (1997), and *Verron et al.* (1999) each attempted to make the Kalman filter feasible by reducing the dimension of the state vector. They did so by choosing a reduced dimension subspace in which to represent

the state vector. The forecast error statistics are propagated only within this subspace. The ensemble Kalman filter described in the previous chapter uses a Monte Carlo approach to maintain a low-dimensional representation of the error statistics. However, nonlinearities in the model equations still pose difficulties for some of the assimilation schemes based on a reduced dimension Kalman filter.

In a study of the large scale Pacific ocean circulation, *Stammer and Wunsch* (1996) used a set of localised and geostrophically balanced vortices as the basis functions for a reduced dimension model. The linear dynamics of perturbations from the mean state in this reduced dimension subspace was determined using “model Green’s functions”. These are calculated from short model runs using initial conditions that are perturbed according to each of the basis functions that defines the subspace. The linear model is used in assimilation experiments with both model-simulated and real data.

A common theme of many of the sub-optimal assimilation schemes mentioned above is the use of a reduced dimension subspace for representing the state vector. Some of the approaches simply use a degraded spatial resolution for the state vector. A frequent choice is EOFs which are an optimal representation in that they capture the maximum variance of a multivariate time series (usually long model runs or data sets are used), according to some prescribed norm. Typically the subspace spanned by the EOFs is considered fixed in time. The singular vectors used to formulate a simplified Kalman filter (KF) by *Fisher* (1998) are similar to EOFs, but they are evolved through time according to the model dynamics and optimally capture the variance of the state at some future time. This appears to be a good choice for the KF algorithm since after each analysis the reduced subspace can be chosen to optimally resolve the forecast error covariance matrix at the time of the subsequent analysis or beyond. This approach is, however, computationally expensive and actually requires the TLM and adjoint of the forecast model to calculate the singular vectors. For a linear KF application, *Dowd and Thompson* (1997) used normal modes of the dynamical model. These have the advantage that their temporal evolution is completely decoupled from each other. Consequently, a model based on a

truncated set of normal modes perfectly models the evolution of the retained modes.

In the context of data assimilation, controllability and observability are important criteria in the selection of a reduced dimension bases. *Dowd and Thompson (1997)* used these criteria when selecting normal modes for a simplified KF. If a basis vector is not controllable, the controls are unable to change the component of the state vector that projects onto this basis vector. When no observations provide information on the component of the state vector projected onto a given basis vector, then that vector is said to be unobservable. Basis vectors that are either uncontrollable or unobservable do not play a role when assimilating data and therefore can be safely neglected.

In this chapter, a method is presented for quickly obtaining an approximate adjoint model. The adjoint model is projected into a reduced dimension subspace. The best choice of subspace for an assimilation scheme based on an adjoint model is not obvious because the observations and controls may be arbitrarily spread throughout the assimilation period and the application may be for hindcasting or forecasting. In this chapter a truncated set of EOFs are used to span the reduced dimension subspace. The method treats the ocean model as a “black box”, similar to the approach of *Stammer and Wunsch (1996)*. The result is a sub-optimal assimilation scheme obtained using only output from the nonlinear ocean model.

6.3 Approximate Adjoint Model

In this chapter, a method is presented for quickly obtaining an approximate adjoint model. It is closely related to the approach used by *Stammer and Wunsch (1996)*. The method treats the ocean model (repeated here from Appendix B for convenience),

$$\mathbf{s}_n = \mathcal{D}(\mathbf{s}_{n-1}) + \mathbf{G} \mathbf{f}_n, \quad (6.1)$$

as a “black box”. The state vector \mathbf{s} has dimension N_s and $\mathcal{D}()$ represents the discrete time-stepping form of the nonlinear model dynamics. The vector \mathbf{f} represents any external forcing and \mathbf{G} transforms the forcing into their effect on the state vector. The method presented below avoids the need to analytically derive all of the N_s^2

partial derivatives of the model equations required for the TLM and adjoint models at each time-step. Therefore, it is a relatively easy task to establish a system for assimilating data using a new model and also to accommodate changes in the model.

6.3.1 Numerical Linearisation of the Ocean Model

As an alternative to calculating the linearised model coefficients by analytically deriving all of the partial derivatives of the model equations, they can be approximated by the finite difference form

$$\frac{\partial \mathcal{D}_j(\mathbf{s})}{\partial s_i} \approx \frac{\mathcal{D}_j(\mathbf{s} + \Delta^i) - \mathcal{D}_j(\mathbf{s})}{\Delta}. \quad (6.2)$$

This is the partial derivative of the j th model equation with respect to the i th state variable, evaluated at the state \mathbf{s} . The N_s -dimensional vector Δ^i has the i th element equal to the small perturbation Δ and the remaining elements set to zero. To evaluate (6.2), the ocean model (6.1) is simply initialised with the state \mathbf{s} and integrated for one time-step with the forcing set to zero. The resulting state vector, $\mathcal{D}(\mathbf{s})$, is stored. Next, the model is initialised with the perturbed state $(\mathbf{s} + \Delta_i)$ and the same procedure performed with the result being $\mathcal{D}(\mathbf{s} + \Delta_i)$. The j th component of the partial derivative with respect to s_i is calculated by taking the difference between the j th element of these vectors and dividing by the magnitude of the perturbation, Δ .

This basic approach allows the model to be linearised in a way that is independent of the specific model, $\mathcal{D}()$. However, for a nonlinear model these derivatives are functions of the model state and therefore, strictly speaking, must be recalculated at each time-step. Since it requires $O(N_s)$ time-steps of model integration to evaluate the linearised model coefficients at a single time-step, this approach is infeasible unless the following two approximations can be made while still retaining enough useful information on the gradient of J . Firstly, the frequency of recalculating the coefficients must somehow be decreased. Secondly, the dimension of the adjoint model must be decreased to reduce the number of required partial derivatives. The method can be made feasible if the number of time-steps between recalculating the coefficients is of

the same order as the reduced dimension of the adjoint model.

For a nonlinear atmospheric model, *Errico and Vukicevic* (1992) showed that it was sufficient in a particular case to update the TLM coefficients every 3 hours. For the case of an ocean model, the rate of change of the coefficients has not yet been determined. Consequently, tests were employed as part of the identical twin experiments to determine an appropriate update frequency. The frequency depends on the rate of change of the model state and also the relative importance of the nonlinear terms in the model. It is expected that due to the slower evolution of the ocean as compared to the atmosphere, the coefficients will possibly only need to be recalculated every few days.

6.3.2 Reduced Dimension Subspace

Realistic ocean models can often have state vectors with dimension, N_s , that is $O(10^5)$ or greater. However, due to constraining dynamical relationships (such as geostrophy) and the limited variation in the external forcing, the output from such models does not realise all of the possible degrees of freedom. Therefore, it should be possible to significantly reduce the model dimension when formulating the adjoint model while preserving most of the dynamical information.

A truncated set of EOFs are used to define the subspace. These modes are convenient since they are easy to calculate from a long run of the numerical model, and their orthogonality simplifies the formulation of the adjoint model. The modes span the subspace that optimally accounts for the variance in the state vector. They may not, however, be the best choice for optimally resolving the relationship between the controls and the observations. The dimension of the subspace is denoted as N_e . The basis vectors are the columns of \mathbf{E}_1 . The basis vectors of the remaining $(N_s - N_e)$ -dimensional subspace, which is neglected, is denoted by the matrix \mathbf{E}_2 . The parameter N_e is chosen so that the subspace \mathbf{E}_1 contains most of the information on both the mean and time varying components of the model state vector.

To generate the EOFs, the model is integrated using the best available set of a

priori estimates for the controls. After the model state reaches a statistical steady state, if one exists, output from a long model run is used to calculate the EOFs. The $N_s \times N_s$ matrix, \mathbf{C} , which is similar to the covariance matrix, except that the mean has not been removed from the samples, is calculated as

$$\mathbf{C} = \mathbf{N} \left(\overline{\mathbf{s}_n \mathbf{s}_n^T} \right) \mathbf{N}. \quad (6.3)$$

The overbar denotes an average over time. The matrix \mathbf{C} contains all of the contemporaneous sample covariances of the model state with itself plus a contribution from the mean. The diagonal scaling matrix, \mathbf{N} , normalises each type of variable (e.g. velocity, pressure, and density) by the square root of its spatially averaged variance. As in the previous chapter this norm is chosen mostly out of convenience. A more appropriate norm may be total energy. In the case here, the basis vectors are used to represent the full model state. However, if the subspace is used to represent a correction with respect to a background state, as in the incremental approach described previously, it may be more appropriate to first multiply the samples by an estimate of the standard deviation of the background error divided by the sample standard deviation. This would give improved representation of the correction in areas with high background error where these corrections should be relatively large.

The eigenvectors of this matrix are the multi-variate EOFs which are patterns in all of the state variables that have temporally uncorrelated amplitudes. The mean state, because it is uncorrelated with the time varying patterns, is mostly isolated in one mode. This may not be an optimal representation if the state vector tends to alternate between multiple quasi-stationary states. The corresponding eigenvalues are the mean squared amplitude of each mode. The modes are partitioned so that \mathbf{E}_1 contains the first N_e leading eigenvectors. The projection of the state vector onto the subspaces \mathbf{E}_1 and \mathbf{E}_2 are proportional to the principal components and are denoted \mathbf{a}_n^1 and \mathbf{a}_n^2 , respectively. These are related to the original state vector by

$$\mathbf{s}_n = \mathbf{N}^{-1} (\mathbf{E}_1 \mathbf{a}_n^1 + \mathbf{E}_2 \mathbf{a}_n^2). \quad (6.4)$$

6.3.3 Reduced Dimension Adjoint Model

To formulate the reduced dimension adjoint model, the subspace \mathbf{E}_2 is simply neglected. The assumption is made that if enough EOFs can be retained, the effect on the evolution of the retained modes from the remaining subspace is negligible. Therefore, the approximate substitution

$$\mathbf{s}_n = \mathbf{N}^{-1} \mathbf{E}_1 \mathbf{a}_n^1 \quad (6.5)$$

is made into the nonlinear model equation (6.1). The following equation results after pre-multiplying by $\mathbf{E}_1^T \mathbf{N}$ to project each term into the reduced dimension subspace:

$$\mathbf{a}_n^1 = \mathbf{E}_1^T \mathbf{N} \mathcal{D} (\mathbf{N}^{-1} \mathbf{E}_1 \mathbf{a}_{n-1}^1) + \mathbf{E}_1^T \mathbf{N} \mathbf{G} \mathbf{f}_n. \quad (6.6)$$

The corresponding adjoint model is then obtained in the normal way (see Appendix B) treating \mathbf{a}_n^1 as the model state

$$\boldsymbol{\lambda}_n^1 = \frac{\partial \mathcal{D}}{\partial \mathbf{a}_n^1} \mathbf{N} \mathbf{E}_1 \boldsymbol{\lambda}_{n+1}^1 - \frac{\partial J}{\partial \mathbf{a}_n^1}. \quad (6.7)$$

The reduced dimension adjoint model is used to obtain an estimate of the gradient of J with respect to \mathbf{a}_n^1 . This can be related to the gradient with respect to the full state, \mathbf{s}_n , using the chain rule

$$\frac{\partial J}{\partial \mathbf{s}_n} = \frac{\partial \mathbf{a}_n^1}{\partial \mathbf{s}_n} \frac{\partial J}{\partial \mathbf{a}_n^1} = \mathbf{N} \mathbf{E}_1 \frac{\partial J}{\partial \mathbf{a}_n^1}. \quad (6.8)$$

Any additional relationships between the controls and \mathbf{s}_n are then used to obtain the gradient with respect to the controls (in the experiments described below, only the initial state, \mathbf{s}_0 , is used as the controls). This gradient information is used together with the nonlinear model to iteratively minimise the cost function.

The coefficient matrix in this modified adjoint model is now only $N_e \times N_e$. This matrix is the transpose of the TLM for the forward model projected onto the EOF subspace, \mathbf{E}_1 . Therefore, (6.7) propagates information on the gradient of J with respect to the state vector projected onto the subspace \mathbf{E}_1 , only. Consequently, the value $-\boldsymbol{\lambda}_0^1$ is the gradient of J with respect to the initial state projected onto \mathbf{E}_1 .

The reduced dimension adjoint model coefficients can be evaluated numerically in a manner analogous to that described above for the full adjoint model. This is similar to the method used by *Stammer and Wunsch* (1996). The elements of the first matrix in the product in (6.7) are approximated by

$$\left[\frac{\partial \mathcal{D}(\mathbf{s})}{\partial \mathbf{a}^1} \right]_{ij} \approx \frac{\mathcal{D}_j(\mathbf{s} + \mathbf{N}^{-1} \mathbf{E}_1 \Delta^i) - \mathcal{D}_j(\mathbf{s})}{\Delta}. \quad (6.9)$$

The vector Δ^i is now a perturbation with the i th element equal to Δ and the remaining $N_e - 1$ elements zero. Again, to evaluate this expression requires only the results of one-step model integrations with the forcing set to zero. All the required coefficients for the reduced dimension adjoint model can be calculated from $(N_e + 1)$ one step model integrations. Then the resulting $N_e \times N_s$ matrix is post-multiplied by $\mathbf{N} \mathbf{E}_1$ to obtain the $N_e \times N_e$ reduced dimension adjoint model coefficients.

A result of the above formulation is an adjoint model with no interaction between the retained subspace, \mathbf{E}_1 , and the neglected subspace, \mathbf{E}_2 . However, unlike the case for a linear model (see e.g. *Dowd and Thompson*, 1997), it is impossible to choose a subspace for a nonlinear model that is dynamically uncoupled from the remaining state space. This can be illustrated by considering the full adjoint model after the bases are rotated to be aligned with the full set of EOFs using the transformation $\lambda = \mathbf{N} \mathbf{E}_1 \lambda^1 + \mathbf{N} \mathbf{E}_2 \lambda^2$. The full adjoint model is

$$\begin{bmatrix} \lambda_n^1 \\ \lambda_n^2 \end{bmatrix} = \begin{bmatrix} \mathbf{E}_1^T \mathbf{N}^{-1} \mathbf{D}^T \mathbf{N} \mathbf{E}_1 & \mathbf{E}_1^T \mathbf{N}^{-1} \mathbf{D}^T \mathbf{N} \mathbf{E}_2 \\ \mathbf{E}_2^T \mathbf{N}^{-1} \mathbf{D}^T \mathbf{N} \mathbf{E}_1 & \mathbf{E}_2^T \mathbf{N}^{-1} \mathbf{D}^T \mathbf{N} \mathbf{E}_2 \end{bmatrix} \begin{bmatrix} \lambda_{n+1}^1 \\ \lambda_{n+1}^2 \end{bmatrix} - \begin{bmatrix} \partial J / \partial \mathbf{a}_n^1 \\ \partial J / \partial \mathbf{a}_n^2 \end{bmatrix}, \quad (6.10)$$

where $\mathbf{D}^T = \partial \mathcal{D} / \partial \mathbf{s}$ is evaluated at \mathbf{s}_n . In general, no bases $[\mathbf{E}_1 \ \mathbf{E}_2]$ can make the upper-right block of this coefficient matrix zero for all times. Therefore, the backwards evolution of the reduced dimension adjoint vector λ_n^1 , in \mathbf{E}_1 , depends on the adjoint vector λ_n^2 , in \mathbf{E}_2 . Consequently, errors in the approximate adjoint model, which simply neglects this dependence, may accumulate over a long assimilation period. These errors result from the failure to account for the effect of adjusting the state in \mathbf{E}_1 on the state in \mathbf{E}_2 at later time-steps and the subsequent effect of this on J . However, by choosing \mathbf{E}_1 to span the subspace with highest variance, \mathbf{E}_2 contains little energy.

Therefore, it is assumed that little energy will also be transferred from \mathbf{E}_1 to \mathbf{E}_2 as a result of adjustments in \mathbf{E}_1 during the minimisation of J . This is a fundamental assumption of the approach that will be further explored in the discussion.

6.4 Identical Twin Experiment

An identical twin experiment was performed using an idealised ocean model to demonstrate the method described above. This type of experiment is a simulation of a real data assimilation exercise. The values of the controls are specified and pseudo-observations are derived from the solution of the model. This set of controls and the associated model solution represent the “true” ocean. Then, an independent set of controls is taken as the initial estimate. The assimilation scheme is applied in an attempt to recover the true controls using only a limited set of observations of the “true” ocean. The degree to which these controls are recovered provides a measure of the information content of the observations used and the effectiveness of the assimilation method.

6.4.1 Description of the Experiment

The model used for the identical twin experiment is an idealised version of the CANDIE model (*Sheng et al.*, 1998). This is a nonlinear primitive equation model with a rigid lid. The domain of the model is a 1600 km by 1600 km box with a flat bottom and vertical walls. The resolution in the horizontal is 40 km. The model has four levels in the vertical. The total dimension of the state vector, including horizontal velocity, density, and surface pressure, is 5360. Since the model is hydrostatic and incompressible, the remaining variables of vertical velocity and pressure below the surface are diagnosed from the state vector. Because the domain is an enclosed box, no open boundary conditions need to be specified. Wind forcing is steady and mimics the average wind stress pattern over the North Atlantic. Not surprisingly, the resulting mean circulation pattern consists of two large gyres concentrated at the

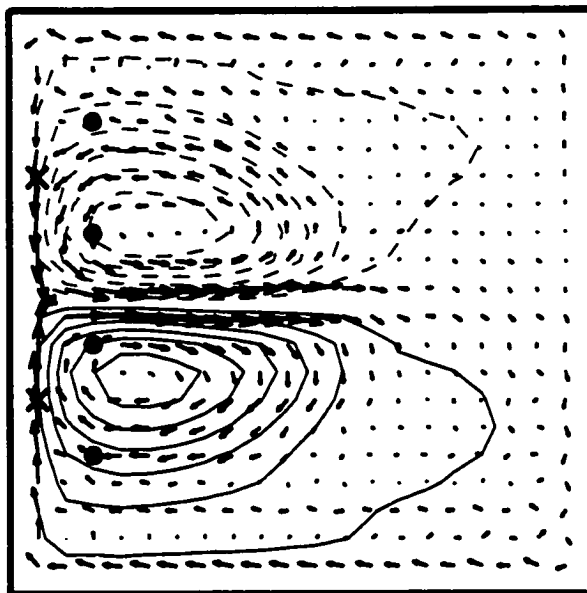


Figure 6.1: Mean surface velocity and pressure fields from 1000 day integration of the model used in the identical twin experiment. The four dots represent locations where velocity, density, and pressure profiles are observed. The two crosses represent locations where surface pressure is observed. The domain is 1600 km by 1600 km with a horizontal resolution of 40 km and 4 levels in the vertical.

western boundary with a strong jet separating them (Figure 6.1). Because of weak initial stratification and the lack of surface buoyancy fluxes, the resulting circulation pattern is strongly barotropic. The model supports instabilities in the jet and the generation of eddies. Because of the steady forcing, all of the variability in the model state after spin-up results from these internal sources.

The experiment itself consists of three phases as shown in Figure 6.2. The first is the spin-up phase where the model is integrated from a state of rest for 1000 days to allow the statistics of the model state to stabilise. The second phase is the next 1000 days of model integration where the output of the model is used to generate the EOFs. Also, the mean model state is calculated from this 1000 day period (shown in Figure 6.1) and is referred to as the ocean climate. The third phase consists

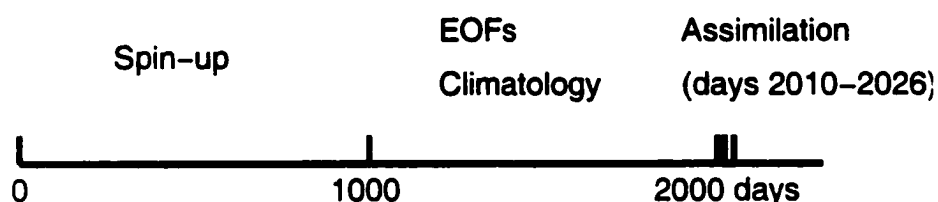


Figure 6.2: Schematic diagram showing the three phases of the assimilation experiment: the spin-up phase, the period from which the EOFs and the model climatology are calculated, and the assimilation period from which the pseudo-observations are taken.

of days 2010 to 2026 of the model integration which are taken as the assimilation period. The pseudo-observations are derived from the model state during this period. The initial conditions (model state on day 2010) projected on the reduced dimension subspace are the controls. The initial estimate for these controls are taken as the model climatology projected on the subspace. A background term is also included in the cost function with a low weighting factor that penalises departures from the model climatology. This term improves the conditioning of the problem. The close temporal proximity between the assimilation period and the period used to estimate the EOFs helps ensure that the retained EOFs will effectively span the true model state within the assimilation period.

The pseudo-observations were chosen to mimic realistic oceanic data. At four locations (solid circles in Figure 6.1) near the western boundary the horizontal velocity, density, and pressure were observed at all depths once per day. These are meant to

mimic a fixed observing array near the coast. Surface pressure was observed daily at two locations (crosses in Figure 6.1) on the western boundary to mimic a pair of coastal sea level gauges. Finally, a single snapshot of surface pressure over the entire domain in the middle of the assimilation period (day 2018) was also observed to mimic data supplied by satellite altimetry. The observation errors were assumed to be uncorrelated and have variance approximately related to the total variance for each variable type in the long model run (days 1000 to 2000).

The EOFs were generated following the method described in section 6.3.2 from 1000 days of model output sampled twice daily. The resulting modes with large eigenvalues are coherent patterns of gyres and currents. The modes also satisfy the model's boundary conditions. The pressure, density, and velocity fields appear to be balanced, in that flows within each mode are nearly geostrophic. The first EOF is shown in the top panel of Figure 6.3. From the pattern of this mode and the evolution of its corresponding amplitude (principal component) during the 1000 day period (also shown in Figure 6.3) this mode is clearly an approximation to the mean ocean state. The subsequent modes in order of decreasing eigenvalues correspond to large scale variations in the circulation patterns (Figures 6.3 and 6.4). The modes with lower variance show progressively smaller scale patterns of variation (Figure 6.5).

Modes 2 and 3 appear to be associated with regular periodic fluctuations in the location of the jet. Figure 6.6 shows the power spectra for the amplitudes of these modes. The coherence and phase spectra are also shown. The two time series are highly coherent at the frequency where almost all of the energy is concentrated, around 0.01 cycles/day. At this frequency the amplitude of mode 3 leads mode 2 by 90° , which is about 25 days. Therefore, the amplitudes of these modes are in quadrature and correspond with the downstream propagation of a large meander in the jet with a period of about 100 days. Mode 4 is concentrated where the western boundary currents separate from the coast. This mode appears to be associated with meridional fluctuations in the location of this separation.

The first 200 EOFs were sufficient to account for over 99% of the variance of the

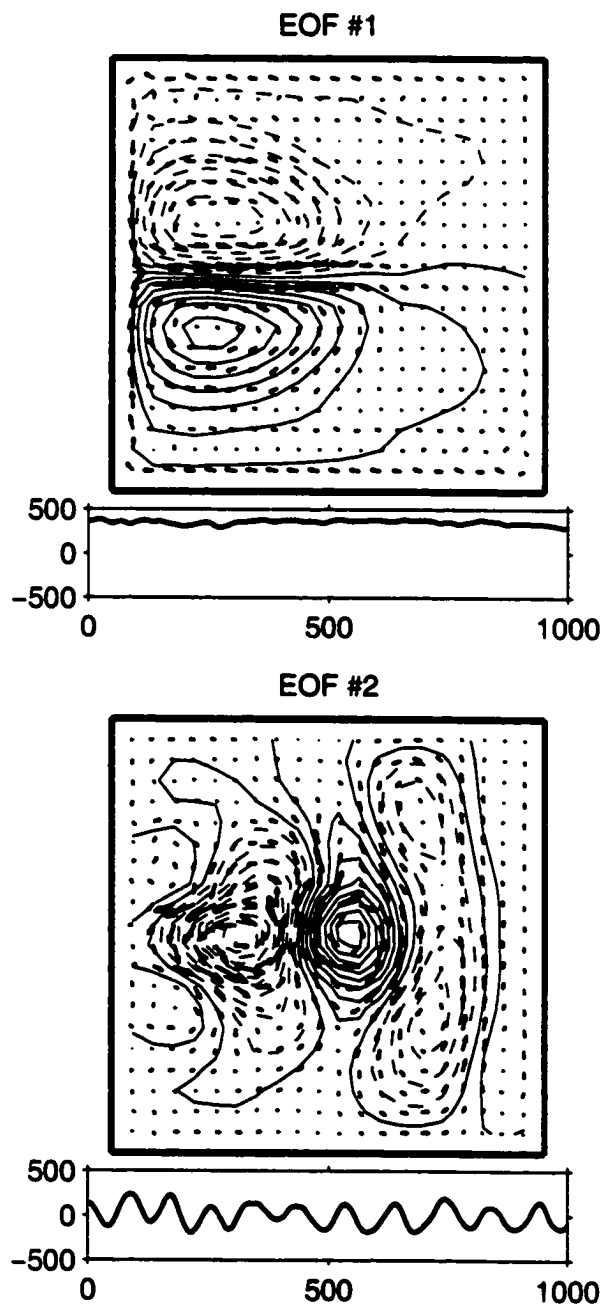


Figure 6.3: Plots show the surface velocity and pressure fields of the first two EOFs. The time component of the modes during the 1000 day period from which they were calculated is also shown.

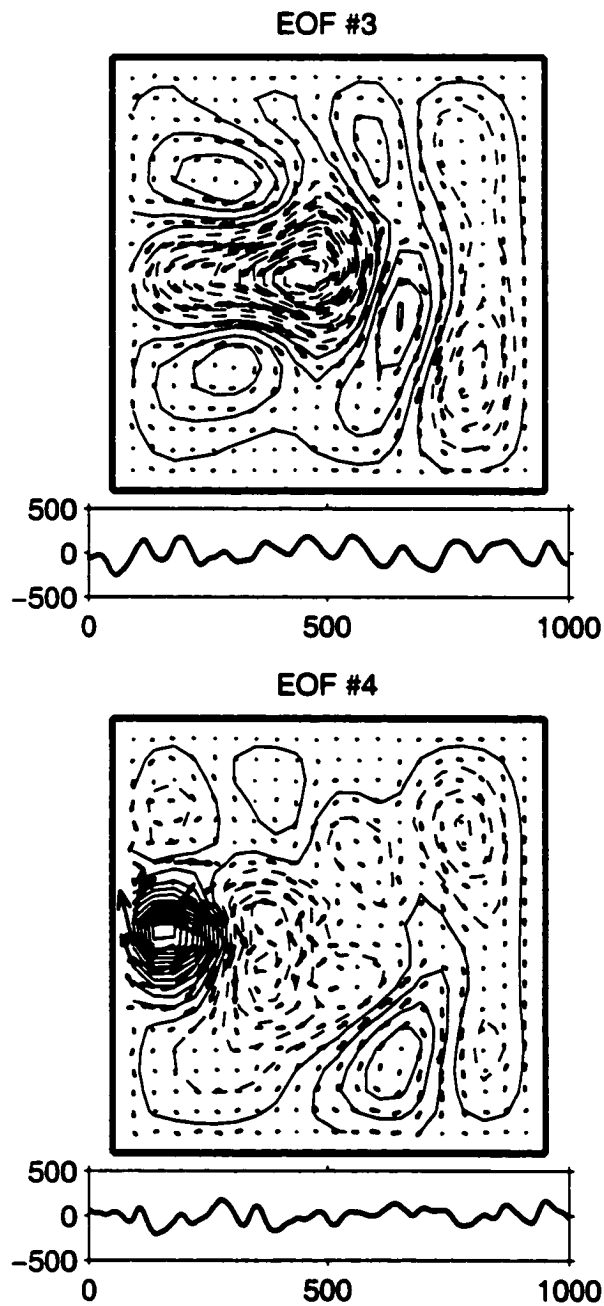


Figure 6.4: Same as Figure 6.3, but for modes 3 and 4.

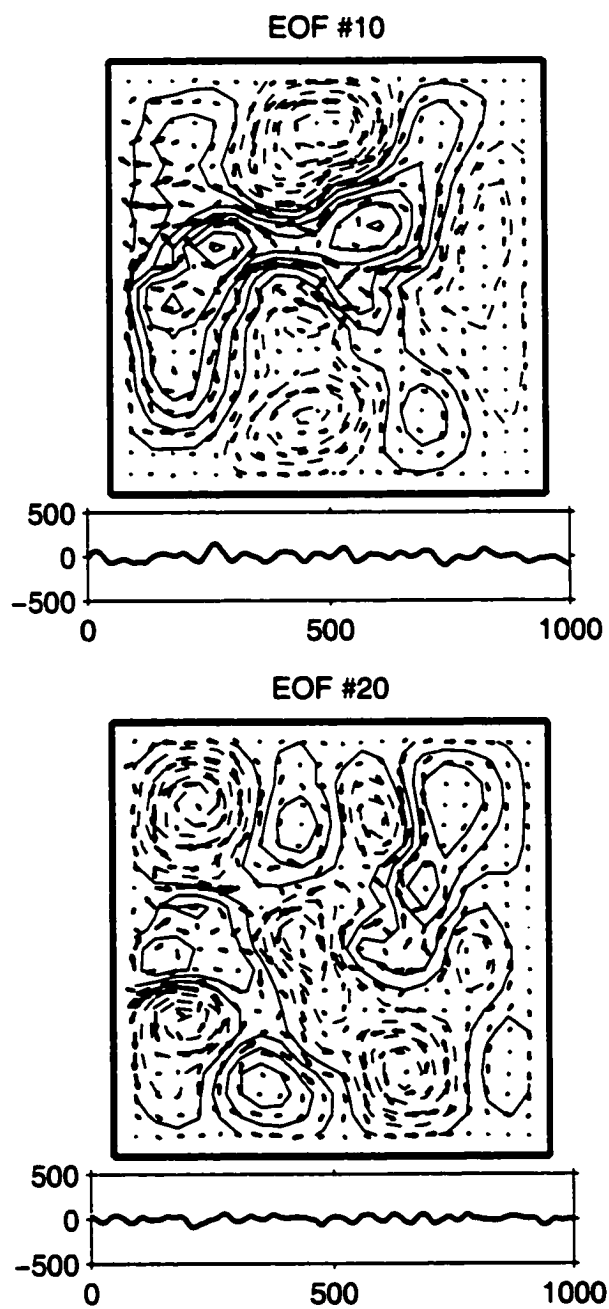


Figure 6.5: Same as Figure 6.3, but for modes 10 and 20.

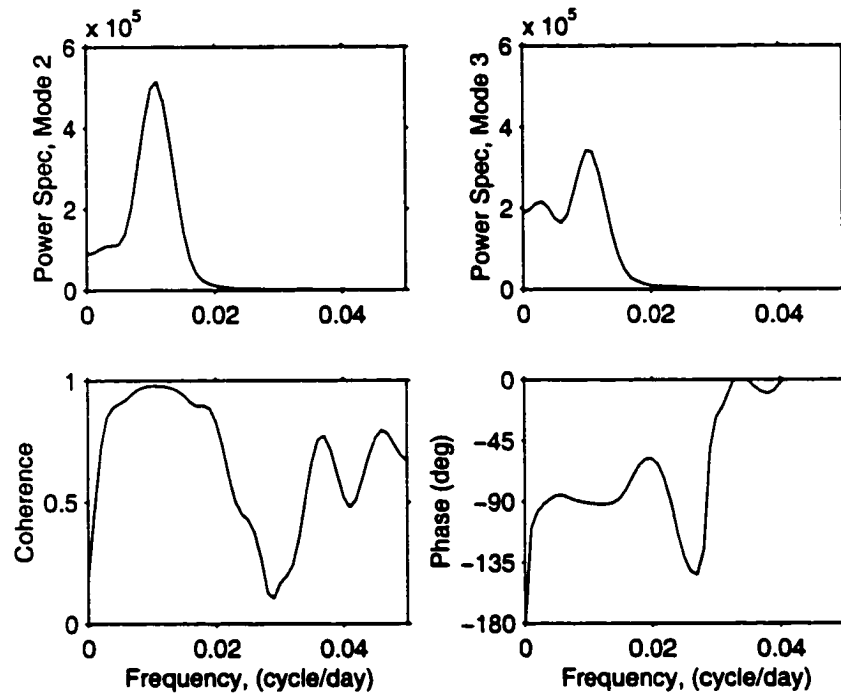


Figure 6.6: Upper plots show the power spectra for the amplitudes of the second and third EOFs. Both spectra have a peak at about 0.01 cycles/day. At this frequency, the coherence (bottom left panel) is 0.97 and the phase (bottom right panel) is -90° . Therefore, the amplitudes are in quadrature with mode 3 leading mode 2. This corresponds to a periodic downstream propagation of the dominant meander in the jet with a period of about 100 days.

model state during the 1000 day period from which they were calculated. Since the initial amplitudes of the EOFs are taken as the controls, the issue of controllability does not need to be considered. Due to the weak baroclinic forcing in the model, the modes are mostly barotropic. Therefore, all of the modes will also be observed since the observations include a complete snapshot of the surface pressure. These 200 modes were used as the basis for the reduced dimension adjoint model in the assimilation.

A series of trials showed that the adjoint model coefficients could be evaluated every four days (equal to 400 time-steps) without a significant loss of accuracy. Therefore, the reduced dimension adjoint model (6.7) was calculated using 200 EOFs with the coefficients evaluated every four days. Consequently, it is computationally less expensive to calculate the coefficients of the reduced dimension adjoint model than to simply integrate the nonlinear ocean model over the assimilation period.

6.4.2 Results

Using the reduced dimension adjoint model and starting with the model climatology as the first guess for the initial conditions, the cost function was minimised. The quasi-Newton minimisation algorithm known as BFGS was used to minimise J by adjusting the initial conditions. The value of J along with its components corresponding to each type of observation are shown in Figure 6.7 as a function of iteration number. The value of J corresponding to the first guess of the initial conditions was reduced by 85% by assimilating the limited set of observations.

The left three panels in Figure 6.8 show the surface velocity and pressure fields of the “true” ocean during the assimilation period. The right three panels show the corresponding fields that resulted after the initial conditions were found that minimised J . The overall pattern in the fields are in good agreement. Figure 6.9 shows the result of subtracting the fields at the end of the assimilation period from the initial fields for both the “true” ocean and the assimilation results. The good agreement between these plots illustrates that the movement of the jet and the gyres

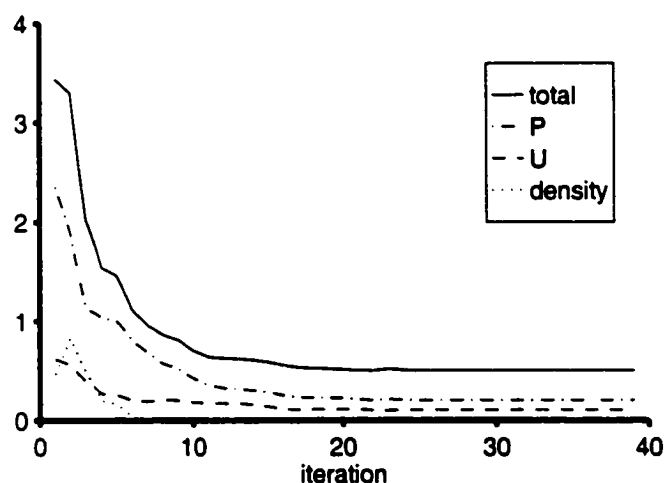


Figure 6.7: The value of J as a function of iteration number in the minimisation. The contributions to J from the pressure, velocity, and density observations are also shown.

are well reproduced in the assimilation results.

As expected, the “true” ocean state was not reproduced exactly. However, without the full adjoint model, it cannot be easily determined how much this is a function of the approximations in the assimilation scheme or a result of the limited number of observations.

6.5 Discussion and Conclusions

The results of the identical twin experiment are encouraging and confirm the results from several other studies that schemes based on reduced dimension models can provide useful results. For the model used in the experiment, the assimilation scheme could be made computationally feasible. The approximate adjoint model was able to propagate information backwards within the reduced dimension subspace to the initial time and supply useful information on the sensitivity of J to the initial conditions.

In this study 200 modes were retained for the adjoint model. A sensitivity study

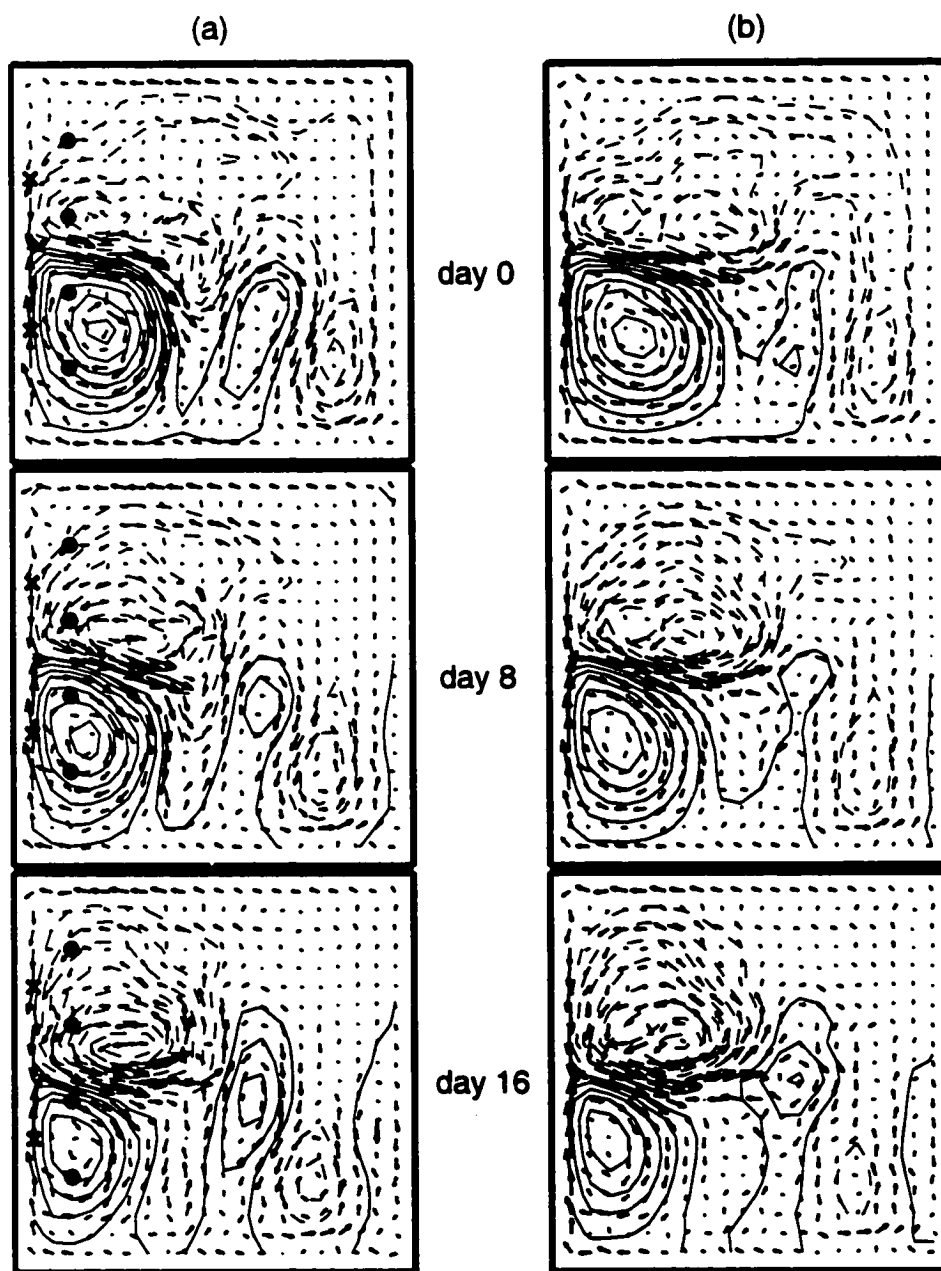


Figure 6.8: Three snapshots of the surface velocity and pressure fields during the assimilation period for (a) the “true” ocean and (b) the recovered ocean states from the assimilation.

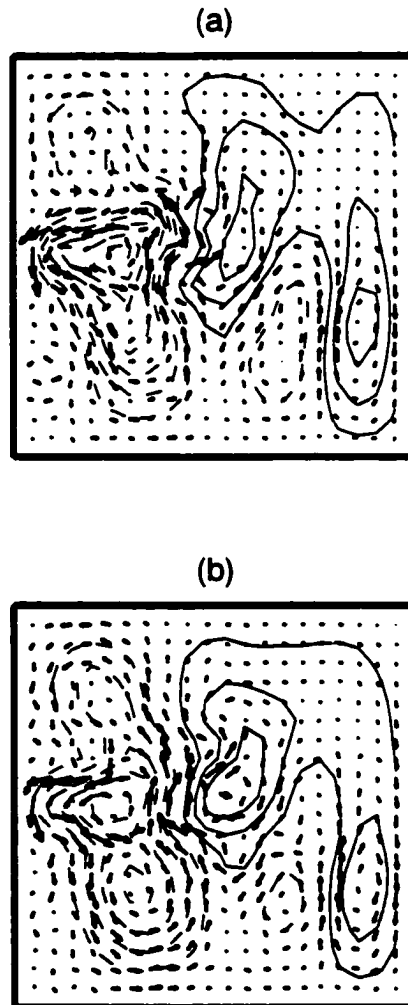


Figure 6.9: The difference in surface velocity and pressure fields between the beginning and end of the assimilation period for (a) the “true” ocean and (b) the recovered ocean states from the assimilation.

could be carried out to determine the effect of using a greater or lesser number of modes. Even though the major features of the flow field can be represented by very few modes, it is likely that the correct dynamical evolution of these features would rely on some of the less energetic modes. The low-dimensional adjoint model would fail to give useful gradient information with respect to controlling the main features when such dynamically important modes are neglected.

A possible drawback of the method is the inconsistency between the full nonlinear model used for the forward run and the reduced dimension adjoint model used to calculate the cost function gradient. This inconsistency may cause convergence problems with the minimisation algorithm, however, no such problems were found in the identical twin experiment. As an alternative, the reduced dimension TLM model could be used, thus making the forward and adjoint models perfectly consistent. In that case the incremental approach, described earlier, would be used and the controls would be defined as a correction to the previous solution of the nonlinear model. By simply restarting the minimisation algorithm after each iteration, the approach used in this chapter can be considered as a special case of the incremental approach where only one iteration is allowed for the minimisation of each linearised problem.

One of the weaknesses of the method is the way the reduced dimension subspace is calculated. The choice of this subspace is of major importance for the success of the method. The EOFs depend on the *a priori* estimates of the controls, since these controls determine the model solution used to generate the EOFs. In general, one may not have much prior information on the controls. If the control estimates result in a model solution that is not representative of the true ocean state, then the low-dimensional subspace will be inefficient in representing the true state. In the context of an operational ocean forecasting system, however, experience gained over time from comparing the low-dimensional model state with observations should result in a better estimate of the possible range of values of the controls that could then be used to improve the EOFs.

The approach outlined in this chapter uses a fixed set of basis functions. This may

be appropriate when the system is expected to remain within a single flow regime with a well defined mean state. In cases where the system may undergo a transition to a distinct quasi-stationary regime, however, it may be more effective to use a set of basis vectors that either evolves according to the model dynamics or is periodically recalculated. The simplified KF schemes described by *Verron et al. (1999)* employs time-evolving basis vectors. These basis functions are propagated continually through the model dynamics starting from a set of EOFs calculated from a long model run. As for the method presented in this chapter, there is a risk that the subspace may fail to capture significant modes that have evolved from perturbations that are initially small and outside the low-dimensional subspace. However, since the basis functions themselves satisfy the model dynamics, this should reduce the coupling through the dynamics between the resolved and unresolved subspaces.

As ocean models become increasingly large and complex, the implementation of conventional four-dimensional data assimilation schemes will become an even more daunting task than today. However, the relative success of the method applied in this chapter and other recently published oceanographic studies demonstrate that sub-optimal approaches may hold much promise for the future of oceanographic data assimilation.

Chapter 7

Concluding Remarks

This chapter provides a summary of the results from each of the preceding chapters that dealt with a diverse range of issues related to the assimilation of data in oceanography and meteorology. Two important themes, however, were repeatedly touched upon throughout: (1) the requirement of *a priori* statistical information for the error in a prior estimate of the controls or in the model dynamics; and (2) the use of a simplified representation of the model dynamics, state vector or error statistics to obtain a sub-optimal scheme. These are two areas where subjective judgement must be used when designing the assimilation scheme appropriate for a specific oceanographic or meteorological application. Sections 7.2 and 7.3 provide a discussion of these themes. The final section is a brief discussion contrasting the utility of idealised studies, such as those employed in this thesis, and the development of a comprehensive operational ocean forecasting system.

7.1 Summary of Results

Chapter 2: Analysis of a low-dimensional ocean model for tidal flow over a bank.

- the topographic Rossby wave mode with azimuthal wave number one is directly forced by tidal advection of planetary vorticity across isobaths

- a steady along-isobath current is forced by the tidal advection of the relative vorticity associated with the Rossby wave
- the resonant frequency of the Rossby wave decreases with increasing along-isobath current and the system becomes highly nonlinear when the tidal frequency approaches the resonant frequency
- the simple model produces realistic simulated drifter and image data that are used in idealised assimilation studies in Chapters 3 and 4

Chapter 3: Examination of a framework for assimilating drifter data that focuses on the effects of nonlinearity of the advection equation.

- the nonlinearity of the advection equation causes non-Gaussian error in the modelled trajectory and also results in a non-quadratic cost function (both of these effects increase with the trajectory length)
- using sub-trajectories reduces these problems and allows standard estimation techniques to be applied
- mis-specified error statistics can be diagnosed from normalised residuals to the full and sub-trajectories allowing statistical information on the unresolved velocities to be inferred from drifter data

Chapter 4: Evaluation of a new approach for extracting velocity information from sequential images, such as ice or SST.

- by allowing dynamical and statistical constraints in addition to other types of observations to be used, the proposed approach is more general than the purely statistical techniques usually applied to these data
- assimilation of simulated SST images results in a similar cost function as when using trajectories

- the method was applied successfully to a pair of ice images from the Labrador shelf (7.5 h separation) using a simple non-dynamical model and a regularization term
- most of the variation between the images (57%) is due to wind drift and comparison with bathymetry shows that the estimated stationary surface stream-function (which explains an additional 21% of the between-image variance) is consistent with along-isobath flow

Chapter 5: Improved representation of the forecast error statistics within an NWP application (3D-Var).

- the leading EOFs of the stationary forecast error covariances capture dynamically important features unresolved by standard covariance models (baroclinic tilt and orographic forcing)
- new localisation techniques were developed to suppress noisy correlations at large horizontal separation distances
- a practical approach was developed for blending EOFs with a standard full-rank estimate to enable the optimal correction to the forecast to span more than just the EOF subspace

Chapter 6: Development of an effective and efficient scheme for assimilation with a nonlinear ocean model that avoids the need to obtain the full adjoint model.

- using an idealised double gyre model, the leading EOFs calculated from a long model run capture most of the variance of the state vector: 200 modes account for over 99% of the variance (full state dimension is 5360)
- the two most energetic time-varying modes capture a 100 day period meander in the position of the jet
- useful information on the cost function gradient given by the linearised model in the EOF subspace calculated by treating model as a “black box”

- in an identical twin experiment, the evolution of the state is well recovered from limited observations

7.2 Error Statistics

An important part of many assimilation schemes is the calculation of the error statistics for a background state, usually resulting from a short-term forecast. When these statistics are poorly specified, the effectiveness of the scheme can be limited. For example, if the background error statistics have a null space, this possibly can lead to uncontrolled error growth. The KF provides an approach for calculating the background error covariances for the case of linear model dynamics. The extended KF can be applied to nonlinear dynamics, but it may fail in cases where the nonlinearity is sufficiently strong (*Miller et al.*, 1994). However, application of the KF or extended KF is often not possible due to the large dimension of the state vector, thus making the propagation of the error statistics through the model dynamics infeasible. Consequently, approximate methods are usually required. For example, in Chapter 5 several approaches were described for obtaining an estimate of the stationary background error statistics within the context of an operational numerical weather prediction (NWP) system. Alternatively, the Monte Carlo approach of the ensemble KF can be used to estimate the background error statistics (*Evensen*, 1994). Similar approaches may be appropriate for an operational ocean prediction system.

Model error can result from unresolved physical processes, errors in the model geometry or from inadequate parameterisations of the sub-grid scale processes. Estimation of this source of error is usually difficult. In Chapter 3, an approach was suggested for tuning the standard deviation of the velocity errors using information from the residuals to the fitted trajectories. This approach would, however, not be effective for estimating more detailed statistical information of the errors. In Chapter 5, the error in the short-term forecast used as the background state contains both model error and error from the imperfect initial state used to make the forecast. An

ad hoc method was used to estimate the stationary statistics for the net effect of both these sources of error. *Daley* (1992) proposed an approach for separating these two sources of error to estimate the homogeneous component of the model error statistics.

Estimation of the model or background error statistics would be most effective within the context of an operational ocean forecast system. This would allow an updated estimate of the statistics to be made from past results of the system using approaches such as those described in Chapters 3 and 5. The statistics used for the assimilation can also be continually checked for consistency against the misfits between the resulting forecasts and the observations.

7.3 Sub-optimal Schemes

As an alternative approach to specifying the error statistics for the background state, the controls can be defined as the correction to the background state parameterised in terms of a smaller number of specified basis functions. This approach simplifies the estimation problem by reducing the dimension of the control vector. The use of spatially smooth basis functions serves a similar role as specifying background error statistics with smooth horizontal correlations. It may, however, be more convenient to use a reduced set of basis functions in certain applications. For example, in Chapter 4 the flow field was parameterised in terms of a truncated polynomial expansion to eliminate spurious small scale variations. Similarly, in Chapter 6 the ocean state vector and model dynamics were projected into a low-dimensional sub-space spanned by a truncated set of EOFs. Such low-dimensional approximations can be justified when either insufficient observations or statistical information for the background error are available to effectively estimate the model state at full resolution.

As mentioned above it is often difficult to estimate the true error statistics for a large dimension system. To make this possible, assumptions such as homogeneity and stationarity often must be imposed on the error covariances, as described in Chapter 5. For highly nonlinear systems, the actual statistics may have higher order moments

above the first two that are even more difficult to estimate. In general, assimilation schemes neglect statistical moments beyond covariances since the storage and manipulation of this statistical information can become computationally infeasible for large dimension systems. In fact, proper treatment of the error covariances for NWP and many realistic oceanographic applications is even beyond today's technical limits. Therefore, the use of sub-optimal approaches employing either a simplified representation of the error statistics or a Monte Carlo approach is often necessary due to computational limitations or a lack of statistical information.

7.4 Operational Ocean Prediction

Many of the results in this thesis were only possible to obtain because assimilation approaches were studied within an idealised context. By working with artificial data produced by the model and errors generated from a known distribution, the effect of any simplifying assumptions can be precisely known. These types of experiments can also be used to evaluate the effect of introducing measurements that do not yet exist in reality as long as the expected observation error statistics are known. The studies of assimilating Lagrangian measurements (Chapters 3 and 4) used the very simple, yet dynamically interesting ocean model developed in Chapter 2. Studies with such low-dimensional nonlinear models allow the effects of the nonlinearity on the estimation problem to be examined in detail. It is hoped that such studies can provide guidance when applying similar assimilation schemes to realistic nonlinear ocean models.

In combination with such idealised studies that attempt to isolate certain aspects of a data assimilation problem, the establishment of an operational forecast system can also benefit research on practical assimilation approaches. The forecasts from such a system are evaluated on a routine basis through a broad range of situations. Then, any new assimilation approaches or data types are tested in a parallel system that is otherwise identical to the operational system and their effectiveness evaluated

relative to the operational system. A new approach or type of data is introduced to the operational system only after it is shown to consistently improve the forecast skill over a sufficient range of cases. This is the development approach used at most centres responsible for NWP. In oceanographic applications, it is common for various assimilation approaches to be applied to such a variety of situations with varying spatial and temporal scales, geographic domains, and types of data that direct comparison of the results is often difficult. The performance of an operational forecast system for a given region can be used as the benchmark against which alternative approaches are compared.

After an operational forecast system is shown to have significant forecast skill, the output can be made available on a routine basis for a variety of applications including climate prediction, marine search and rescue, ice forecasting, and oil spill modelling. It is hoped that the approaches for incorporating Lagrangian measurements and applying sub-optimal assimilation schemes presented in this thesis can contribute to the development of such operational systems.

Appendix A

Kalman Filter Algorithm

The Kalman filter (KF) is a sequential assimilation method that produces state estimates after a single sweep through the data, forward in time. The estimated state, referred to as the “analysis” or “analysed state”, at a given time is only influenced by the data up to that time. For linear problems, and depending on the extent to which the error covariance matrices are correctly specified, the resulting estimated states are statistically optimal with respect to past and present data. It is typically assumed that the observation and model errors are Gaussian with zero mean and serially uncorrelated through time. In the case of nonlinear dynamics, the extended Kalman filter (EKF) can be applied. The only difference between the KF and EKF is that the linearised form of the nonlinear $\mathcal{D}()$ and $\mathcal{H}()$ operators must be used for the EKF. However, as discussed in the introduction, nonlinearity can lead to statistical suboptimality unless the algorithm is extended to include the higher order moments.

There are four steps in the KF algorithm that are performed whenever observations are available (assumed to be each time-step here):

1. Forecast the model state at the current time-step, denoted by \mathbf{s}_n^b , using the model and the optimal estimate from the previous analysis, with the model error terms set to zero:

$$\mathbf{s}_n^b = \mathbf{D}\hat{\mathbf{s}}_{n-1} + \mathbf{G}\mathbf{f}_n \quad (\text{A.1})$$

2. Calculate the covariance matrix, denoted by Σ_n^s , of the error in this forecast by propagating the error covariances of the previous analysis (Σ_{n-1}^s) through time using the linear model dynamics and adding the covariances of model error (Σ^m):

$$\Sigma_n^s = \mathbf{D}\Sigma_{n-1}^s\mathbf{D}^T + \Sigma^m. \quad (\text{A.2})$$

3. Blend the information from the forecast and the observations (taking into account their respective error covariances) to obtain the current analysed state, denoted by $\hat{\mathbf{s}}_n$. The optimal estimate is the state, \mathbf{s}_n , that minimises the cost function

$$J = \frac{1}{2} (\mathbf{s}_n - \mathbf{s}_n^b)^T \Sigma_n^{s-1} (\mathbf{s}_n - \mathbf{s}_n^b) + \frac{1}{2} (\mathbf{H}\mathbf{s}_n - \mathbf{y}_n)^T \Sigma^o{}^{-1} (\mathbf{H}\mathbf{s}_n - \mathbf{y}_n). \quad (\text{A.3})$$

4. Compute the covariance of this estimate, given by

$$\Sigma_n^{\hat{s}} = (\mathbf{H}^T \Sigma^o{}^{-1} \mathbf{H} + \Sigma_n^{s-1})^{-1}. \quad (\text{A.4})$$

Note that step three is the same as the regression approach used in (1.10)-(1.11) when prior estimates for the model parameters (in this case the forecast from the previous analysis) are given. The other steps in the KF algorithm are simply necessary to update the covariance matrix associated with the errors in the model forecast (Σ_n^s) and the errors in the resulting optimal state vector estimate ($\Sigma_n^{\hat{s}}$).

The optimal state that minimises (A.3) can be written explicitly in a form similar to (1.11) as

$$\hat{\mathbf{s}}_n = (\mathbf{H}^T \Sigma^o{}^{-1} \mathbf{H} + \Sigma_n^{s-1})^{-1} (\mathbf{H}^T \Sigma^o{}^{-1} \mathbf{y} + \Sigma_n^{s-1} \mathbf{s}_n^b). \quad (\text{A.5})$$

More frequently, the explicit solution is written in the equivalent form

$$\hat{\mathbf{s}}_n = \mathbf{s}_n^b + \mathbf{K}_n (\mathbf{y}_n - \mathbf{H}\mathbf{s}_n^b), \quad (\text{A.6})$$

where the so-called Kalman gain matrix is given by

$$\mathbf{K}_n = \Sigma_n^s \mathbf{H}^T (\Sigma^o + \mathbf{H}\Sigma_n^s \mathbf{H}^T)^{-1}. \quad (\text{A.7})$$

Appendix B

Adjoint Method

The adjoint method is essentially an efficient way to calculate the gradient of J with respect to a set of controls. To illustrate, assume a perfect nonlinear ocean model ($\Sigma^m = 0$) is given as

$$\mathbf{s}_n = \mathcal{D}(\mathbf{s}_{n-1}) + \mathbf{G} \mathbf{f}_n. \quad (\text{B.1})$$

If the controls are specified (usually including the initial and any open boundary conditions), the model (B.1) can be integrated forward in time to produce an estimate for the ocean state at all times. Therefore, one can think of the time-dependent state vector as being a function of the controls. Given a prior estimate for the controls, denoted by $\boldsymbol{\alpha}_0$, the goal is to find the $\boldsymbol{\alpha}$ that minimises the following cost function:

$$J = \frac{1}{2} (\boldsymbol{\alpha} - \boldsymbol{\alpha}_0)^T \boldsymbol{\Sigma}^{\boldsymbol{\alpha}-1} (\boldsymbol{\alpha} - \boldsymbol{\alpha}_0) + \frac{1}{2} \sum_{n=0}^N (\mathbf{H} \mathbf{s}_n - \mathbf{y}_n)^T \boldsymbol{\Sigma}^{\mathbf{o}-1} (\mathbf{H} \mathbf{s}_n - \mathbf{y}_n), \quad (\text{B.2})$$

subject to the strong constraint (B.1).

To calculate the gradient of (B.2) with respect to $\boldsymbol{\alpha}$, the constrained optimisation problem given by (B.1)-(B.2) is transformed into an unconstrained problem using Lagrange multipliers to form the Lagrange function

$$L = J + \sum_{n=1}^N \boldsymbol{\lambda}_n^T [\mathbf{s}_n - \mathcal{D}(\mathbf{s}_{n-1}) - \mathbf{G} \mathbf{f}_n]. \quad (\text{B.3})$$

The vector of Lagrange multipliers, λ_n , has the same dimension as the state vector and is referred to as the adjoint vector. The Lagrange function is differentiated with respect to the state vector and the results equated to zero. The Lagrange multipliers, in effect, account for the dependencies of the state vector through time. Therefore, L can be differentiated with respect to \mathbf{s}_n as if it were independent of future states. The result is the following adjoint model:

$$\lambda_n = \frac{\partial \mathcal{D}}{\partial \mathbf{s}_n} \lambda_{n+1} - \frac{\partial J}{\partial \mathbf{s}_n}, \quad (\text{B.4})$$

for $n = N, \dots, 0$ and with $\lambda_{N+1} = 0$. This is a time-stepping model with time running backwards. The model is linear with the matrix of coefficients being equal to the transpose of (that is, adjoint to) the tangent linear dynamics evaluated with respect to the solution of (B.1). In this notation the derivative $\partial \mathcal{D} / \partial \mathbf{s}_n$ is the $N_s \times N_s$ Jacobian matrix of $\mathcal{D}(\cdot)$ and $\partial J / \partial \mathbf{s}_n$ is the gradient of J with respect to the state vector. These are defined as follows where i and j are row and column indices:

$$\left[\frac{\partial \mathcal{D}}{\partial \mathbf{s}} \right]_{ij} = \frac{\partial \mathcal{D}_j}{\partial s_i} \quad ; \quad \left[\frac{\partial J}{\partial \mathbf{s}} \right]_i = \frac{\partial J}{\partial s_i}. \quad (\text{B.5})$$

The gradient of J with respect to the controls is then determined by taking the gradient of (B.3). For example, the gradient of J with respect to the initial state, \mathbf{s}_0 , is

$$\frac{dJ}{d\mathbf{s}_0} = -\frac{\partial \mathcal{D}}{\partial \mathbf{s}_0} \lambda_1 + \frac{\partial J}{\partial \mathbf{s}_0} = -\lambda_0, \quad (\text{B.6})$$

where the partial derivative with respect to \mathbf{s}_0 represents the sensitivity of changes in the initial conditions only to the part of the cost function corresponding to the initial time. The full derivative of J with respect to \mathbf{s}_0 is used to denote the dependency on the initial state, through the linearised dynamics, of the parts of J that depend on the model state at all times.

Appendix C

Low-Dimensional Model Parameters

By neglecting the nonlinear and friction terms in the vorticity equation (2.1) and assuming the periodic harmonic solution (2.7), the following generalised eigenvalue problem is obtained:

$$r \left[\Phi_1 - \frac{\phi_1}{hr^2} \right] = -\frac{1}{\omega_l} \left(\frac{f}{h} \right)_r \phi_1, \quad (\text{C.1})$$

where the subscript r denotes $\partial/\partial r$ and Φ_i is defined as

$$\Phi_i = \frac{1}{r} \frac{\partial}{\partial r} \left(\frac{r}{h} \frac{\partial \phi_i}{\partial r} \right), \quad (\text{C.2})$$

for $i = 0, 1$. The eigenvalues of (C.1) correspond with $(1/\omega_l)$, the inverse of the frequency of the linear Rossby wave, and the eigenfunctions correspond with the radial structure of the wave, ϕ_1 . The remaining term on the right side of (C.1), $(-f/h)_r$ is a weighting function that defines the orthogonality relationship of the eigenfunctions.

Following the derivation outlined in chapter 2, the equations corresponding with a sine and cosine dependence on the azimuthal angle are given by:

$$\left(\frac{f}{h} \right)_r \phi_1 (S_t + \lambda S) - C\omega_l \phi_1 \left\{ Z \left[\left(\frac{\Phi_0}{h} \right)_r + \left(\frac{f}{h} \right)_r \frac{(\phi_0)_r}{rh\omega_l} \right] + \left(\frac{f}{h} \right)_r \right\} = 0 \quad (\text{C.3})$$

$$\left(\frac{f}{h}\right)_r \phi_1 (C_t + \lambda C) + S\omega_l \phi_1 \left\{ Z \left[\left(\frac{\Phi_0}{h}\right)_r + \left(\frac{f}{h}\right)_r \frac{(\phi_0)_r}{rh\omega_l} \right] + \left(\frac{f}{h}\right)_r \right\} = -\frac{\omega_l U_\infty h_\infty \sin(\omega_l t) r}{2} \left[Z \left(\frac{\Phi_0}{h}\right)_r + \left(\frac{f}{h}\right)_r \right]. \quad (\text{C.4})$$

The final form of these equations given in the chapter are obtained by projecting each term onto the linear topographic Rossby wave mode corresponding with no zero crossings in the radial direction. This projection is performed using the weighting function given above. This results in the following definitions for the four parameters in the simplified model:

$$a_1 = \frac{\int \phi_1^2 \left(\frac{\Phi_0}{h}\right)_r dr}{\int \phi_1^2 \left(\frac{f}{h}\right)_r dr} \quad (\text{C.5})$$

$$a_2 = \frac{\int \frac{\phi_1^2 (\phi_0)_r}{r\omega_l h} \left(\frac{f}{h}\right)_r dr}{\int \phi_1^2 \left(\frac{f}{h}\right)_r dr} \quad (\text{C.6})$$

$$a_3 = \frac{\int r \phi_1 \left(\frac{\Phi_0}{h}\right)_r dr}{\int \phi_1^2 \left(\frac{f}{h}\right)_r dr} \quad (\text{C.7})$$

$$a_4 = \frac{\int r \phi_1 \left(\frac{f}{h}\right)_r dr}{\int \phi_1^2 \left(\frac{f}{h}\right)_r dr} \quad (\text{C.8})$$

where the integrals are over $(0, \infty)$. The values for the parameters were evaluated numerically using discrete approximations to the above integrals and derivatives.

Grouping the terms without a dependence on the azimuthal angle gives

$$\Phi_0 (Z_t + \lambda Z) = -\frac{U_\infty h_\infty}{2r\omega_l} \left[\left(\frac{f}{h}\right)_r \frac{\phi_1}{h} \right]_r C \sin(\omega_l t). \quad (\text{C.9})$$

The radial dependence of this equation is eliminated by simply equating the radial dependence of the mean current response (left side of (C.9)) with the radial dependence of the forcing (right side of (C.9)). This maximises the response by assuming it projects completely onto the forcing and leads to the following relationship between the radial shapes of the Rossby wave and the mean current:

$$\Phi_0 = \frac{1}{r} \left[\left(\frac{f}{h} \right)_r \frac{\phi_1}{h} \right]_r \quad (\text{C.10})$$

from which ϕ_0 is calculated.

Appendix D

AR(1) Correlated \mathbf{u}^s

The general case with serially correlated velocity errors is mathematically complex. Here we only consider velocity errors governed by the discrete-time AR(1) process

$$\mathbf{u}_n^s = \beta \mathbf{u}_{n-1}^s + \mathbf{u}_n^{\text{ar}}, \quad (\text{D.1})$$

where the model errors \mathbf{u}_n^s are understood to be along the model trajectory, that is, located at \mathbf{x}_n^m . The matrix, β , has eigenvalues less than one and the Gaussian evolution process, \mathbf{u}^{ar} , is serially uncorrelated with covariance matrix

$$\Sigma^{\text{ar}} = \overline{\mathbf{u}_n^{\text{ar}} (\mathbf{u}_n^{\text{ar}})^T}. \quad (\text{D.2})$$

Some algebraic manipulation of (3.11) and (D.1) at time-steps n and $n - 1$ leads to the following expression for ϵ_n^x in terms of \mathbf{u}^{ar} :

$$\epsilon_n^x = (\gamma_{n-1} + \beta) \epsilon_{n-1}^x - \beta \gamma_{n-1} \epsilon_{n-2}^x + \Delta t \mathbf{u}_{n-1}^{\text{ar}}. \quad (\text{D.3})$$

This form is similar to an AR(2) stochastic model except that the coefficients may vary through time along the trajectory. The stochastic model similar to (3.17), but for AR(1) correlated velocity error is

$$\tilde{\mathbf{X}}^o(\alpha) = \tilde{\mathbf{x}}^m(\alpha) + \Gamma_2(\alpha) \tilde{\mathbf{u}}^{\text{ar}} + \tilde{\epsilon}^o, \quad (\text{D.4})$$

where multiplication by the matrix Γ_2 is equivalent to applying the model (D.3) to obtain the $\tilde{\epsilon}^x$. The total error covariance matrix can therefore be written as

$$\Sigma^{tot}(\alpha) = \Gamma_2(\alpha)\Sigma^{ar}\Gamma_2(\alpha)^T + \Sigma^o. \quad (\text{D.5})$$

Appendix E

Adjoint Model for Image Advection

In this appendix, the adjoint model is developed corresponding to the cost function (4.9) and the advection model (4.6) presented in chapter 4. For simplicity the velocity field is considered to be a known analytic function of the controls.

To obtain the trajectories required to calculate J_I , a simple finite difference approximation is used. To advect the position at $n = N/2$ backwards in time, the following time-stepping model is used:

$$\mathbf{x}_n^k = \mathbf{x}_{n+1}^k - \mathbf{u}_{n+1}(\mathbf{x}_{n+1}^k)\Delta t, \quad (\text{E.1})$$

for $n = N/2 - 1, \dots, 0$. Similarly, the position at time index $N/2$ is advected forwards with the time-stepping model:

$$\mathbf{x}_n^k = \mathbf{x}_{n-1}^k + \mathbf{u}_{n-1}(\mathbf{x}_{n-1}^k)\Delta t, \quad (\text{E.2})$$

for $n = N/2 + 1, \dots, N$.

In this simple case, with no dynamical ocean model, the advection equation is the only set of constraints in the problem. Therefore, the Lagrange function (obtained

following the approach outlined in Appendix A) is

$$\begin{aligned}
 L = J_I + & \sum_{n=0}^{N/2-1} [\boldsymbol{\lambda}_n^k]^T [\mathbf{x}_n^k - \mathbf{x}_{n+1}^k + \mathbf{u}_{n+1}(\mathbf{x}_{n+1}^k)\Delta t] \\
 & + \sum_{n=N/2+1}^N [\boldsymbol{\lambda}_n^k]^T [\mathbf{x}_n^k - \mathbf{x}_{n-1}^k - \mathbf{u}_{n-1}(\mathbf{x}_{n-1}^k)\Delta t],
 \end{aligned} \tag{E.3}$$

where $\boldsymbol{\lambda}_n^k$ is the so-called adjoint vector at time-step n corresponding to the trajectory \mathbf{x}_n^k . The resulting adjoint model propagates information towards the intermediate time along both trajectory segments, that is, from $n = 0$ to $N/2$ and from $n = N$ to $N/2$. This is the opposite direction as the integration of the advection model used to calculate J_I . The adjoint model corresponding to the advection equation between $n = 0$ and $N/2$ is

$$\boldsymbol{\lambda}_n^k = \left(\mathbf{I} - \frac{\partial \mathbf{u}_n}{\partial \mathbf{x}_n^k} \Delta t \right) \boldsymbol{\lambda}_{n-1}^k - \frac{\partial J_I}{\partial \mathbf{x}_n^k}, \tag{E.4}$$

for $n = 0, \dots, (N/2) - 1$ with $\boldsymbol{\lambda}_{-1}^k$ set to zero. Similarly, the adjoint for the advection equation between $n = N$ and $N/2$ is:

$$\boldsymbol{\lambda}_n^k = \left(\mathbf{I} + \frac{\partial \mathbf{u}_n}{\partial \mathbf{x}_n^k} \Delta t \right) \boldsymbol{\lambda}_{n+1}^k - \frac{\partial J_I}{\partial \mathbf{x}_n^k}, \tag{E.5}$$

for $n = N, \dots, (N/2) + 1$ with $\boldsymbol{\lambda}_{N+1}^k$ set to zero. The vector derivatives (notation defined in Appendix B) are evaluated at \mathbf{x}_n^k , obtained by integrating (E.1) and (E.2). This is the same form of the adjoint model that would be obtained for the problem of assimilating trajectory data presented in Chapter 3.

The forcing terms of (E.4) and (E.5) are the partial derivatives of J_I with respect to \mathbf{x}_n^k with all other positions along the trajectory held constant. However, due to the advection model, a change in \mathbf{x}_n^k will change all of the subsequent positions along the trajectory. The adjoint model accounts for these relationships. The resulting values of $\boldsymbol{\lambda}_n^k$ are the derivatives of J_I (multiplied by -1) with respect to \mathbf{x}_n^k with all temporal dependencies from the advection model included. Therefore, integration of the adjoint model is equivalent to applying the chain rule along the trajectory. For

example, with $N = 6$ the derivative (gradient) of J_I with respect to \mathbf{x}_4 obtained from (E.5) is

$$\begin{aligned} \frac{dJ_I}{d\mathbf{x}_4} &= \frac{\partial J_I}{\partial \mathbf{x}_4} + \frac{\partial \mathbf{x}_5}{\partial \mathbf{x}_4} \frac{\partial J_I}{\partial \mathbf{x}_5} + \frac{\partial \mathbf{x}_5}{\partial \mathbf{x}_4} \frac{\partial \mathbf{x}_6}{\partial \mathbf{x}_5} \frac{\partial J_I}{\partial \mathbf{x}_6} \\ &= -\lambda_4, \end{aligned}$$

where the superscript is dropped for clarity and derivatives of the form $\partial \mathbf{x}_n / \partial \mathbf{x}_{n-1}$ are obtained from (E.2). Note that this is not strictly a full derivative since J_I also depends on the positions prior to \mathbf{x}_4 and may also depend on the velocities or the source/sink term directly. However, the full derivative notation is used here to denote that all dependencies due to the advection equation are included.

The forcing terms for the adjoint model at $n = 0$ and N depend on the spatial gradients of the observed images. They are obtained by differentiating the cost function after substituting in the observed images using (4.6). The forcing term at $n = 0$ is

$$\frac{\partial J_I}{\partial \mathbf{x}_0^k} = \frac{\partial I_0}{\partial \mathbf{x}_0^k} \left[\tilde{I}_{N/2}(\mathbf{x}_{N/2}^k | I_0) - \tilde{I}_{N/2}(\mathbf{x}_{N/2}^k | I_N) \right]. \quad (\text{E.6})$$

At the final time the forcing term is

$$\frac{\partial J_I}{\partial \mathbf{x}_N^k} = -\frac{\partial I_0}{\partial \mathbf{x}_N^k} \left[\tilde{I}_{N/2}(\mathbf{x}_{N/2}^k | I_0) - \tilde{I}_{N/2}(\mathbf{x}_{N/2}^k | I_N) \right]. \quad (\text{E.7})$$

These are each a product of two terms: the difference in pixel intensity after advection to the intermediate time; and the gradient in the image at the advected position. The image gradients are clearly important for determining how much the value of J_I will change as a result of varying the value of a control parameter. Since a change in the controls will change the two endpoints of the trajectory, the value of the cost function will only change if the pixel intensity is different at the new positions. The value of J_I will be unchanged for small changes in the two endpoints in the direction normal to the local gradient of pixel intensity. For $n \neq 0, N/2, N$ the only contribution to the forcing terms comes from the dependency of the source term on \mathbf{x}_n^k

$$\frac{\partial J_I}{\partial \mathbf{x}_n^k} = \frac{\partial S}{\partial \mathbf{x}_n^k} \Delta t \left[\tilde{I}_{N/2}(\mathbf{x}_{N/2}^k | I_0) - \tilde{I}_{N/2}(\mathbf{x}_{N/2}^k | I_N) \right]. \quad (\text{E.8})$$

To evaluate the gradient of the cost function with respect to the controls, the gradient of the Lagrange function with respect to α is evaluated. Assuming the source/sink terms do not depend directly on α , the result is

$$\frac{dJ_I}{d\alpha} = \sum_k \left[\sum_{n=1}^{N/2} \frac{\partial \mathbf{u}_n(\mathbf{x}_n^k)}{\partial \alpha} \lambda_{n-1}^k \Delta t - \sum_{n=N-1}^{N/2} \frac{\partial \mathbf{u}_n(\mathbf{x}_n^k)}{\partial \alpha} \lambda_{n+1}^k \Delta t \right]. \quad (\text{E.9})$$

This is equivalent to simply using the chain rule. This can be seen, for example, by substituting $(\lambda_{n+1}^k = -dJ_I/d\mathbf{x}_{n+1}^k)$ and, from (E.2), $(\Delta t \mathbf{I} = \partial \mathbf{x}_{n+1}^k / \partial \mathbf{u}_n)$ into the second sum. Again, the full derivative notation is used to indicate that all dependencies from the advection equation are included. Because the contribution to the gradient of J_I from each grid element at $n = N/2$ is additive, each trajectory can be computed separately and the contributions to both J_I and its gradient accumulated.

Appendix F

Calculating EOFs by Singular Value Decomposition

The EOFs are the leading eigenvectors of the sample covariance matrix

$$\mathbf{B} = \frac{1}{N_b - 1} \sum_{i=1}^{N_b} \mathbf{z}_i \mathbf{z}_i^T, \quad (\text{F.1})$$

where \mathbf{z}_i is the i th error sample after removal of the sample mean and N_b is the number of background error samples. It is usually necessary to define a norm to be used in calculating the EOFs. If the error samples simultaneously span several geophysical variables, then it is necessary to use a norm, such as total energy, to transform the variables into appropriate common units. Alternatively, the error samples may be normalised by their sample standard deviations if the EOFs are used to only represent the correlations. In the following formulation, however, it is assumed that the norm is the identity or that the error samples have already been appropriately normalised.

Since N_b is generally much smaller than the dimension of the analysis increment, denoted N_s , \mathbf{B} is highly rank-deficient and therefore it is practical to use singular value decomposition (SVD) to obtain the eigenvectors and corresponding eigenvalues. If all the error samples are combined to form a single N_s by N_b matrix, \mathbf{Z} , then applying

SVD to \mathbf{Z} gives (Gollub and VanLoan, 1983)

$$\mathbf{Z} = \mathbf{U} \begin{bmatrix} \mathbf{\Lambda} \\ \mathbf{0} \end{bmatrix} \mathbf{V}^T, \quad (\text{F.2})$$

where \mathbf{U} and \mathbf{V} are both orthogonal square matrices of dimension N_s and N_b , respectively. The $(N_b - 1) \times (N_b - 1)$ diagonal matrix $\mathbf{\Lambda}$ contains the singular values. The EOFs sought are the first $(N_b - 1)$ columns of \mathbf{U} , and the corresponding eigenvalues are the diagonal elements of $\mathbf{\Lambda}^2$ after division by $(N_b - 1)$. This can be seen by post-multiplying (F.2) by the transpose of itself and comparing with (F.1)

$$\mathbf{Z}\mathbf{Z}^T = \sum_{i=1}^{N_b} \mathbf{z}_i \mathbf{z}_i^T = \mathbf{U} \begin{bmatrix} \mathbf{\Lambda}^2 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \mathbf{U}^T. \quad (\text{F.3})$$

At most, $(N_b - 1)$ EOFs can be calculated (assuming $N_b < N_s$) since one degree of freedom is removed from the samples by subtracting the mean. Some of these modes may, however, mostly be composed of sampling noise. Therefore, one may expect there to exist an optimal number of EOFs to retain, denoted N_e , for representing \mathbf{B} .

To obtain \mathbf{U} and $\mathbf{\Lambda}^2$, the eigenvectors and eigenvalues of the following smaller eigenvalue problem of dimension $(N_b - 1)$ are first calculated:

$$\mathbf{Z}^T \mathbf{Z} = \mathbf{V} \mathbf{\Lambda}^2 \mathbf{V}^T. \quad (\text{F.4})$$

From these results, the EOFs can be obtained by solving the SVD equation (F.2) for the columns of \mathbf{U} corresponding to the N_e largest eigenvalues

$$\mathbf{u}_i = \frac{\mathbf{Z} \mathbf{v}_i}{\Lambda_i}, \quad (\text{F.5})$$

where \mathbf{u}_i and \mathbf{v}_i are the i th columns of \mathbf{U} and \mathbf{V} , respectively, and Λ_i is the corresponding singular value. The matrices \mathbf{E} and $\mathbf{\Lambda}_e$ are used to denote the first N_e retained columns of \mathbf{U} and the diagonal matrix with the corresponding singular values divided by $\sqrt{(N_b - 1)}$, respectively. Therefore, the covariance matrix formed from a truncated set of EOFs can be efficiently represented as

$$\mathbf{B}_e = \mathbf{E} \mathbf{\Lambda}_e^2 \mathbf{E}^T. \quad (\text{F.6})$$

It should be noted that the vector \mathbf{v}_i is a time series that is proportional to the projection of the i th EOF on each of the error samples. By examining these time series, it is possible to evaluate if a given mode is associated with a stationarity component of the error or with sporadic events.

Appendix G

Localisation Using an Iterative Eigendecomposition Algorithm

Several iterative algorithms exist for computing a limited number of eigenvectors and eigenvalues of a large matrix. Most of these, such as the Lanczos algorithm, do not explicitly require the full covariance matrix in memory, but instead only require that the product of the covariance matrix with a series of supplied vectors be calculated. As pointed out by *Zupanski* (1999), localisation of the horizontal covariances estimated from a set of error samples can be incorporated in the calculation of the eigenvectors and eigenvalues when using such an iterative algorithm. In that study, the horizontal covariances were localised by applying a cut-off radius beyond which the covariances were set to zero. This is equivalent to multiplying the original covariance functions by top-hat functions centred on the diagonal. Referring to *Gaspari and Cohn* (1999), because the top-hat is not a valid covariance function, the resulting localised covariance functions are not guaranteed to be positive definite. To deal with this situation, *Zupanski* (1999) simply removed eigenvectors from the truncation for which the eigenvalues were negative (over 20% of the computed leading modes).

Alternatively, a valid localising correlation function could be used, as in the previous method, within an iterative eigendecomposition algorithm. Due to the similarity of (5.28) and (F.1), the error samples can replace the scaled EOFs in the expression

for the localised covariance matrix given by (5.31). Consequently, the product of this localised matrix with an arbitrary vector \mathbf{w} is

$$\mathbf{B}_l \mathbf{w} = \frac{1}{N_b - 1} \sum_{i=1}^{N_b} \text{diag}(\mathbf{z}_i) \mathbf{L} \text{diag}(\mathbf{z}_i) \mathbf{w}. \quad (\text{G.1})$$

The resulting EOFs from this approach will be somewhat different from those calculated using the first approach described in Section 5.3.2. Here, the leading eigenvectors of a localised covariance matrix are obtained, forming an orthogonal basis. In the previous approach, the EOFs calculated from the original covariance matrix were localised, forming a non-orthogonal basis with many modes originating from each of the original EOFs. With this approach, therefore, many times more EOFs must be calculated (and stored) to account for a given percentage of the total variance as compared with the approach from the chapter. Also, if the EOF expansion is truncated, the covariances at large separation distances may not be as damped as one would expect. If the full set of EOFs are retained, then the two approaches would give equivalent results, though the first method still would not produce an orthogonal basis.

Bibliography

- Aikman, G., G. L. Mellor, T. Ezer, D. Shenin, P. Chen, L. Breaker, K. Bosley, and D. B. Rao, *Modern Approaches to data assimilation in ocean modeling*, P. Malanotte-Rizzoli Ed., pp. 347–376, Elsevier Publ., 1996.
- Argo Science Team, Argo: The global array of profiling floats, in *OCEANOBS99: The ocean observing system for climate*, Saint-Raphaël, France, 1999.
- Bennett, A., *Inverse methods in physical oceanography*, Cambridge University Press, 1992.
- Bennett, A. F., *Particle displacements in inhomogeneous turbulence*, pp. 1–46, Stochastic Modelling in Physical Oceanography, Birkhäuser Boston, 1996.
- Buizza, R., J. Tribbia, F. Molteni, and T. Palmer, Computation of optimal unstable structures for a numerical weather prediction model, *Tellus*, **45A**, 388–407, 1993.
- Burgers. G., P. J. VanLeeuwen, and G. Evensen, Analysis scheme in the ensemble Kalman filter, *Mon. Wea. Rev.*, **126**, 1719–1724, 1998.
- Cane, M., A. Kaplan, R. N. Miller, B. Tang, E. C. Hackert, and A. J. Busalacchi, Mapping tropical Pacific sea level: Data assimilation via a reduced state space Kalman filter, *J. Geophys. Res.*, **101**, 22,599–22,617, 1996.
- Carrieres, T., B. Greenan, S. Prinsenberg, and I. K. Peterson, Comparison of Canadian daily ice charts with surface observations off Newfoundland, winter 1992, *Atmosphere-Ocean*, **34**, 207–226, 1996.

- Carsey, F., and B. Holt, Beaufort-Chukchi ice margin data from Seasat: ice motion, *J. Geophys. Res.*, **92**, 7163–7172, 1987.
- Carter, E. F., Assimilation of Lagrangian data into a numerical model, *Dyn. of Atm. and Ocean*, **13**, 335–348, 1989.
- Côté, J., S. Gravel, A. Méthot, A. Patoine, M. Roch, and A. Staniforth, The operational CMC-MRB global environmental multiscale (GEM) model. part I: Design considerations and formulation, *Mon. Wea. Rev.*, **126**, 1373–1395, 1998.
- Courtier, P., J.-N. Thépaut, and A. Hollingsworth, A strategy for operational implementation of 4D-var using an incremental approach, *Q. J. R. Meteorol. Soc.*, **120**, 1367–1387, 1994.
- Courtier, P., E. Andersson, W. Heckley, J. Pailleux, D. Vasiljevic, M. Hamrud, A. Hollingsworth, F. Rabier, and M. Fisher, The ECMWF implementation of three dimensional variational assimilation (3D-Var). part I: formulation, *Q. J. R. Meteorol. Soc.*, **124**, 1783–1808, 1998.
- DaCosta, E., and R. Vautard, A qualitatively realistic low-order model of the extratropical low-frequency variability built from long records of potential vorticity, *Journal of the Atmospheric Sciences*, **54**, 1064–1084, 1997.
- Daley, R., Estimating model-error covariances for application to atmospheric data assimilation, *Mon. Wea. Rev.*, **120**, 1735–1746, 1992.
- Daley, R., Recovery of the one and two dimensional windfields from chemical constituent observations using the constituent transport equation and an extended Kalman filter, *Meteorol. Atmos. Phys.*, **60**, 119–136, 1996a.
- Daley, R., Error propagation and observability for the constituent transport equation in steady, non-divergent, two-dimensional flow, *Atmosphere-Ocean (Andre Robert Memorial Volume)*, pp. 323–351, 1996b.

- Davis, R. E., Observing the general circulation with floats, *Deep-Sea Research*, **38**, S531–S571, 1991.
- Derber, J., and F. Bouttier, A reformulation of the background error covariance in the ECMWF global data assimilation system, *Tellus*, **51A**, 195–221, 1999.
- DeYoung, B., Y. Lu, and R. Greatbatch, Synoptic bottom pressure variability on the Labrador and Newfoundland continental shelves, *J. Geophys. Res.*, **100**, 8639–8653, 1995.
- Dowd, M. G., and K. R. Thompson, Forecasting coastal circulation using an approximate Kalman filter based on dynamical modes, *Continental Shelf Research* (in press).
- Draper, N. R., and H. Smith, *Applied Regression Analysis*, Wiley series in probability and mathematical statistics, Wiley, 1981.
- Ehrendorfer, M., and J. J. Tribbia, Optimal prediction of forecast error covariances through singular vectors, *Journal of the Atmospheric Sciences*, **54**, 286–313, 1997.
- Emery, W. J., C. W. Fowler, J. Hawkins, and R. H. Preller, Fram Strait satellite image-derived ice motions, *J. Geophys. Res.*, **96**, 4751–4768, 1991.
- Emery, W. J., C. W. Fowler, and C. A. Clayson, Satellite-image-derived Gulf Stream currents compared with numerical model results, *J. Atmos. Oceanic Technol.*, **9**, 286–304, 1992.
- Emery, W. J., C. W. Fowler, and J. A. Maslanik, Satellite remote sensing of ice motion, in *Oceanographic Applications of Remote Sensing*, edited by M. Ikeda and F. W. Dobson, chap. 23, pp. 367–379, CRC Press, 1995.
- Errico, R. M., and T. Vukicevic, Sensitivity analysis using an adjoint of the PSU-NCAR mesoscale model, *Monthly Weather Review*, **120**, 1644–1660, 1992.

- Evensen, G., Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics, *J. Geophys. Res.*, **99**, 10,143–10,162, 1994.
- Fisher, M., Development of a simplified Kalman filter, **Tech. Rep. 260**, ECMWF, 1998.
- Fukumori, I., and P. Malanotte-Rizzoli, An approximate Kalman filter for ocean data assimilation: An example with an idealized Gulf Stream model, *J. Geophys. Res.*, **100**, 6777–6793, 1995.
- Gaspari, G., and S. Cohn, Construction of correlation functions in two and three dimensions, *Q. J. R. Meteorol. Soc.*, **125**, 723–757, 1999.
- Gauthier, P., M. Buehner, and L. Fillion, Background-error statistics modelling in a 3D variational data assimilation scheme, in *Proceedings of the ECMWF workshop on diagnosis of data assimilation systems, 2-4 November, 1998, Reading, UK*, 1998.
- Gauthier, P., C. Charette, L. Fillion, P. Koclas, and S. Laroche, Implementation of a 3D variational data assimilation system at the Canadian meteorological centre. part 1: The global analysis, *Atmosphere-Ocean*, **37**, 103–156, 1999.
- Gelb, A., ed., *Applied Optimal Estimation*, M.I.T. Press, 1974.
- Ghil, M., and P. Malanotte-Rizzoli, Data assimilation in meteorology and oceanography, *Adv. in Geophys.*, **33**, 141–266, 1991.
- Giering, R., and R. Kaminski, Recipes for adjoint code construction, **Tech. Rep. 212**, Max-Planck-Institut for Meteorology, 1996.
- Gill, A. E., *Atmosphere-Ocean Dynamics*, vol. 30 of **International Geophysics**, Academic Press, New York, 1982.
- Gill, P., W. Murray, and M. Wright, *Practical Optimization*, Academic Press, 1981.

- Gollub, H. G., and C. F. VanLoan, *Matrix computations*, The John Hopkins University Press., Baltimore, U.S.A., 476 pp., 1983.
- Greenberg, D. A., and B. D. Petrie, The mean barotropic circulation on the Newfoundland shelf and slope, *J. Geophys. Res.*, **93**, 15,541–15,550, 1988.
- Griffin, D. A., and K. R. Thompson, The adjoint method of data assimilation used operationally for shelf circulation, *J. Geophys. Res.*, **101**, 3457–3477, 1996.
- Guest, P. S., and K. L. Davidson, The effect of observed ice conditions on the drag coefficient in the summer east greenland sea marginal ice zone, *J. Geophys. Res.*, **92**, 6943–6954, 1987.
- Hart, J. E., On oscillatory flow over topography in a rotating fluid, *J. Fluid Mech.*, **214**, 437–454, 1990.
- Heemink, A. W., and I. D. M. Metzelaar, Data assimilation into a numerical shallow water flow model: a stochastic optimal control approach, *J. Mar. Syst.*, **6**, 145–158, 1995.
- Hibler, W. D., and K. Bryan, A diagnostic ice-ocean model., *J. Phys. Oceanogr.*, **17**, 987–1015, 1987.
- Holland, J. A., and X. H. Yan, Ocean thermal feature recognition, discrimination, and tracking using infrared satellite imagery, *IEEE Trans. Geosci. Remote Sensing*, **30**, 1992.
- Hollingsworth, A., and P. Lönnberg, The statistical structure of short-range forecast errors as determined from radiosonde data. part i: the wind field, *Tellus*, **38A**, 111–136, 1986.
- Holton, J. R., *An introduction to dynamic meteorology*, Academic Press, 1992.
- Houtekamer, P. L., and H. L. Mitchell, Data assimilation using an ensemble Kalman filter technique, *Mon. Wea. Rev.*, **126**, 796–811, 1998.

- Huthnance, J. M., Tidal current asymmetries over the Norfolk Sandbanks, *Estuar. Coast. Mar. Sci.*, **1**, 89–99, 1973.
- Ikeda, M., and F. W. Dobson, eds., *Oceanographic applications of remote sensing*, CRC Press, 1995.
- Ikeda, M., and C. L. Tang, Detection of the Labrador current using ice-floe movement in synthetic aperture radar imagery and ice beacon trajectories, *Atmosphere-Ocean*, **30**, 223–245, 1992.
- Janiskova, M., J.-N. Thépaut, and J.-F. Geleyn, Simplified and regular physical parameterizations for incremental four-dimensional variational assimilation, *Mon. Wea. Rev.*, **127**, 26–47, 1999.
- Jazwinski, A. H., *Stochastic processes and filtering theory*, Academic Press, 1970.
- Kalman, R. E., A new approach to linear filtering and prediction problems, *Journal of Basic Engineering*, pp. 35–45, 1960.
- Kamachi, M., and J. J. O'Brien, Continuous data assimilation of drifting buoy trajectory into an equatorial Pacific Ocean model, *J. Mar. Syst.*, **6**, 159–178, 1995.
- Kelly, K., An inverse model for near-surface velocity from infrared images, *J. Phys. Oceanogr.*, **19**, 1845–1864, 1989.
- Kuo, N. J., and X. H. Yan, using the shape-matching method to compute sea-surface velocities from AVHRR satellite images, *IEEE Trans. Geosci. Remote Sensing*, **32**, 724–728, 1994.
- Kwok, R., J. C. Curlander, R. McConnell, and S. S. Pang, An ice-motion tracking system at the Alaska SAR facility, *IEEE J. Ocean. Eng.*, **15**, 44–54, 1990.
- Laroche, S., and P. Gauthier, A validation of the incremental formulation of 4D variational data assimilation in a nonlinear barotropic flow, *Tellus*, **50A**, 557–572, 1998.

- Larouche, P., and J. M. Dubois, Dynamical evaluation of the surface circulation using remote sensing of drifting ice floes, *J. Geophys. Res.*, **95**, 9755–9764, 1990.
- Loder, J. W., Topographic rectification of tidal currents on the sides of Georges Bank, *J. Phys. Oceanogr.*, **10**, 1399–1416, 1980.
- Lorenc, A., A global three-dimensional multivariate statistical interpolation scheme, *Mon. Wea. Rev.*, **109**, 701–721, 1981.
- Lorenc, A. C., Analysis methods for numerical weather prediction, *Q. J. R. Meteorol. Soc.*, **112**, 1177–1194, 1986.
- Malanotte-Rizzoli, P., and R. E. Young, Gulf Stream system assimilation experiments: a sensitivity study, *J. Atmos. Oceanic Technol.*, **14**, 1392–1408, 1997.
- McConnell, R., R. Kwok, J. C. Curlander, W. Kober, and S. S. Pang, Ψ -S correlation and dynamic time warping: two methods for tracking ice floes in SAR images, *IEEE Trans. Geosci. Remote Sensing*, **29**, 1004–1012, 1991.
- Menard, R., L.-P. Chang, and J. Larson, Application of a robust chi-square validation diagnostic in PSAS and Kalman filtering assimilation experiments, in *Proceedings of the Third WMO Symposium on Assimilation of Observations in Meteorology and Oceanography, Quebec City, June 7-10, 1999*, 1999.
- Middleton, J. F., and C. Garrett, A kinematic analysis of polarized eddy fields using drifter data, *J. Geophys. Res.*, **91**, 5094–5102, 1986.
- Miller, R. N., M. Ghil, and R. Gauthiez, Advanced data assimilation in strongly nonlinear dynamical systems, *Journal of the Atmospheric Sciences*, **51**, 1037–1056, 1994.
- Miller, R. N., E. F. Carter, and S. T. Blue, Data assimilation into nonlinear stochastic models, *Tellus*, **51**, 167–194, 1999.

- Mitchell, H. L., and P. L. Houtekamer, Adaptive model-error estimation for an ensemble Kalman filter, in *Proceedings of the third WMO international symposium on assimilation of observations in meteorology and oceanography*, 1999.
- Moore, A. M., and B. F. Farrell, Rapid perturbation growth on spatially and temporally varying oceanic flows determined using an adjoint method: Application to the Gulf Stream, *J. Phys. Oceanogr.*, **23**, 1682–1702, 1993.
- Morrow, R., and P. DeMey, Adjoint assimilation of altimetric, surface drifter, and hydrographic data in a quasi-geostrophic model of the Azores Current, *J. Geophys. Res.*, **100**, 25,007–25,025, 1995.
- Munk, W., P. Worcester, and C. Wunsch, *Ocean Acoustic Tomography*, Cambridge University Press, 1995.
- Niiler, P. P., R. E. Davis, and H. J. White, Water-following characteristics of a mixed layer drifter, *Deep-Sea Research*, **34**, 1867–1881, 1987.
- Ninnis, R. M., W. J. Emery, and M. J. Collins, Automated extraction of pack ice motion from advanced very high resolution radiometer imagery, *J. Geophys. Res.*, **91**, 10,725–10,734, 1986.
- O'Donnell, J., A. A. Allen, and D. L. Murphy, An assessment of the errors in Lagrangian velocity estimates obtained by FGGE drifters in the Labrador current, *J. Atmos. Oceanic Technol.*, **14**, 292–307, 1997.
- Oschlies, A., and J. Willebrand, Assimilation of Geosat altimeter data into an eddy-resolving primitive equation model of the North Atlantic Ocean, *J. Geophys. Res.*, **101**, 14,175–14,190, 1996.
- Parrish, D. F., and J. C. Derber, The national meteorological center's spectral statistical interpolation analysis system, *Mon. Wea. Rev.*, **120**, 1747–1763, 1992.
- Peterson, I. K., A snapshot of the Labrador Current inferred from ice-floe movement in NOAA satellite imagery, *Atmosphere-Ocean*, **25**, 402–415, 1987.

- Pingree, R. D., and L. Maddock, Rotary currents and residual circulation around banks and islands, *Deep Sea Res.*, **32**, 929–947, 1985.
- Polavarapu, S. M., Divergent wind analyses in the oceanic boundary layer, *Tellus*, **47A**, 221–239, 1995.
- Poulain, P.-M., and A. Warn-Varnas, Near-surface circulation of the Nordic seas as measured by Lagrangian drifters, *J. Geophys. Res.*, **101**, 18,237–18,258, 1996.
- Pratte, J. M., and J. E. Hart, Experiments on periodically forced flow over topography in a rotating fluid, *J. Fluid Mech.*, **229**, 87–114, 1991.
- Preller, R. H., Sea ice prediction: The development of a suite of sea-ice forecasting systems for the northern hemisphere, *Oceanography*, **5**, 64–68, 1992.
- Preller, R. H., and P. G. Posey, *Present methods of data assimilation in the U.S. Navy's sea ice forecasting models*, Naval Oceanographic and Atmospheric Research Lab., Stennis Space Center, MS, 1989.
- Priestley, M. B., *Spectral Analysis and Time Series*, Academic Press, 1981.
- Qiu, B., and K. A. Kelly, Upper-ocean heat balance in the Kuroshio extension region, *J. Phys. Oceanogr.*, **23**, 2027–2041, 1993.
- Rabier, F., and coauthors, Recent experimentation on 4D-var and first results from a simplified Kalman filter, **Tech. Rep. 240**, ECMWF, 1997.
- Rabier, F., J.-F. Mafouf, and E. Klinker, The ECMWF operational implementation of four-dimensional variational assimilation, in *Proceedings of the third WMO international symposium on assimilation of observations in meteorology and oceanography*, 1999.
- Ringler, T. D., and K. H. Cook, Understanding the seasonality of orographically forced stationary waves: interaction between mechanical and thermal forcing, *Journal of the Atmospheric Sciences*, **56**, 1154–1174, 1999.

- Robinson, I. S., *Satellite Oceanography*, Ellis Horwood Limited, Chichester, 1985.
- Russ, J. C., *The Image Processing Handbook*, CRC Press, 1995.
- Sanderson, B. G., Structure of an eddy measured with drifters, *J. Geophys. Res.*, **100**, 6761–6776, 1995.
- Sanderson, B. G., and D. A. Booth, The fractal dimension of drifter trajectories and estimates of horizontal eddy-diffusivity, *Tellus*, **43A**, 334–349, 1991.
- Selten, F. M., A statistical closure of a low-order barotropic model, *Journal of the Atmospheric Sciences*, **54**, 1085–1093, 1997.
- Sheng, J., D. G. Wright, R. J. Greatbatch, and D. E. Dietrich, CANDIE: A new version of the DieCAST ocean circulation model, *J. Atmos. Oceanic Technol.*, **15**, 1414–1432, 1998.
- Shumway, R. H., *Applied statistical time series analysis*, Prentice Hall, 1988.
- Stammer, D., and C. Wunsch, The determination of the large-scale circulation of the Pacific ocean from satellite altimetry using model Green's functions, *J. Geophys. Res.*, **101**, 18,409–18,432, 1996.
- Talagrand, O., A posteriori evaluation and verification of analysis and assimilation algorithms, in *Proceedings of the Workshop on Diagnosis of Data Assimilation Systems, ECMWF, Reading November 2-4, 1998*, 1998.
- Talagrand, O., and P. Courtier, Variational assimilation of meteorological observations with the adjoint vorticity equation. i-theory, *Q. J. R. Meteorol. Soc.*, **113**, 1311–1328, 1987.
- Tang, C., Q. Gui, and I. Peterson, Modeling the mean circulation of the Labrador Sea and the adjacent shelves, *J. Phys. Oceanogr.*, 1996, (accepted).
- Tanguay, M., P. Bartello, and P. Gauthier, four-dimensional data assimilation with a wide range of scales, *Tellus*, **47A**, 974–997, 1995.

- Thacker, W. C., Three lectures on fitting numerical models to observations, **Tech. rep.**, GKSS- Forschungszentrum Geesthacht GmbH, 1988a.
- Thacker, W. C., Fitting models to inadequate data by enforcing spatial and temporal smoothness, *J. Geophys. Res.*, **93**, 10,655–10,665, 1988b.
- Thacker, W. C., Oceanographic inverse problems, *Physica*, **D**, 16–37, 1992.
- Thacker, W. C., and R. B. Long, Fitting dynamics to data, *J. Geophys. Res.*, **93**, 1227–1240, 1988.
- Thépaut, J.-N., M. Janiskova, P. Gauthier, G. Desroziers, G. Hello, B. Pouponneau, and F. Veerse, Towards an operational 4D-Var assimilation system at METEOFRANCE, in *Proceedings of the third WMO international symposium on assimilation of observations in meteorology and oceanography*, 1999.
- Thompson, K. R., M. Dowd, and Y. Lu, Oceanographic data assimilation and non-linear regression, in *Proceedings of the Section on Statistics and the Environment of the American Statistical Association*, 1998.
- Thompson, N. R., and J. F. Sykes, Sensitivity and uncertainty analysis of a short-term sea ice motion model, *J. Geophys. Res.*, **95**, 1713–1739, 1990.
- Thomson, R. E., P. H. LeBlond, and W. J. Emery, Analysis of deep-drogued satellite-tracked drifter measurements in the northeast Pacific, *Atmosphere-Ocean*, **28**, 409–443, 1990.
- Thorndike, A. S., and R. Colony, Sea ice motion in response to geostrophic winds, *J. Geophys. Res.*, **87**, 5845–5852, 1982.
- Tokmakian, R., P. T. Strub, and J. McClean-Padman, evaluation of the maximum cross-correlation method of estimating sea surface velocities from sequential satellite images, *J. Atmos. Oceanic Technol.*, **7**, 852–865, 1990.

- Tomassini, M., G. Kelly, and R. Saunders, Use and impact of satellite atmospheric motion winds on ECMWF analyses and forecasts, *Mon. Wea. Rev.*, **127**, 971–986, 1999.
- Verron, J., L. Gourdeau, D. T. Pham, R. Murtugudde, and A. J. Busalacchi, An extended Kalman filter to assimilate satellite altimeter data into a nonlinear numerical model of the tropical Pacific Ocean: Method and validation, *J. Geophys. Res.*, **104**, 5441–5458, 1999.
- Vesecky, J. F., R. Samadani, M. P. Smith, J. M. Daida, and R. N. Bracewell, Observation of sea-ice dynamics using synthetic aperture radar images: automated analysis, *IEEE Trans. Geosci. Remote Sensing*, **26**, 38–48, 1988.
- Wadhams, P., The seasonal ice zone, in *The Geophysics of Sea Ice*, edited by N. Untersteiner, vol. 146 of **NATO ASI Series (B)**, chap. 14, Plenum Press, 1986.
- Weaver, A., and J. Vialard, Development of an ocean incremental 4D-Var scheme for seasonal forecasting, in *Proceedings of the third WMO international symposium on assimilation of observations in meteorology and oceanography*, 1999.
- Wunsch, C., *The ocean circulation inverse problem*, Cambridge University Press, 1996.
- Yingshuo, S., and K. R. Thompson, Oscillating flow of a homogeneous fluid over an isolated topographic feature, *Atmosphere-Ocean*, **35**, 229–255, 1997.
- Zupanski, M., Anisotropic and nonhomogeneous background error covariance formulation based on eigendecomposition, (submitted to *Q. J. R. Meteorol. Soc.*,) 1999.
- Zupanski, M., D. Zupanski, E. Rogers, D. Parrish, and G. DiMego, The NCEP's new regional 4D-Var data assimilation system, in *Proceedings of the third WMO international symposium on assimilation of observations in meteorology and oceanography*, 1999.