

**GLOBAL BAROTROPIC VARIABILITY OF THE OCEAN  
IN RESPONSE TO ATMOSPHERIC FORCING BASED ON  
MULTICHANNEL REGRESSION AND KALMAN FILTER  
TECHNIQUES**

by

**Michael Chechelnitsky**

B.Sc. Upsala College, East Orange, NJ  
(1993)

Submitted in partial fulfillment of the  
requirements for the degree of

Master of Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

and the

WOODS HOLE OCEANOGRAPHIC INSTITUTION

May 1996

© Michael Chechelnitsky 1996

The author hereby grants to MIT and to WHOI permission to reproduce  
and to distribute copies of this thesis document in whole or in part.

Signature of Author .....

.....  
Joint Program in Physical Oceanography  
Massachusetts Institute of Technology  
Woods Hole Oceanographic Institution  
May 28, 1996

Certified by .....

Carl Wunsch  
Cecil and Ida Green Professor of Physical Oceanography  
Thesis Supervisor

Accepted by .....

.....  
Chairman, Joint Committee for Physical Oceanography  
Massachusetts Institute of Technology  
Woods Hole Oceanographic Institution

MASSACHUSETTS INSTITUTE  
OF TECHNOLOGY  
**WITHDRAWN**  
JUN 03 1996  
**FROM**  
LIBRARIES  
**MIT LIBRARIES**



**GLOBAL BAROTROPIC VARIABILITY OF THE OCEAN IN RESPONSE  
TO ATMOSPHERIC FORCING BASED ON MULTICHANNEL  
REGRESSION AND KALMAN FILTER TECHNIQUES.**

by

Michael Chechelnitsky

Submitted in partial fulfillment of the requirements for the degree of  
Master of Science at the Massachusetts Institute of Technology  
and the Woods Hole Oceanographic Institution

May 28, 1996

**Abstract**

TOPEX/POSEIDON altimetry data are employed in the analysis of the global ocean response to atmospheric forcing. We use two different approaches to test the hypothesis that the global sea surface height variability can be adequately described by linear barotropic ocean models: the multichannel regression and the optimal smoothing techniques.

We start with the simplest linear vorticity balance and continue by building a hierarchy of more complicated models by including effects of topography and time dependence. We use auto-regressive external (ARX) time-series models to test the hypothesis in all of the Pacific Ocean. We also test whether any significant residual regression on the atmospheric loading is left after the inverted barometer effect is corrected for. We find that no linear barotropic model is consistent with the data.

We provide a check on the results of the multichannel regression by using a Kalman filter and optimal smoother. We use sequential estimation in the form of filtering-smoothing algorithm. We run the estimate for an area of 4000 km by 2000 km in the Northeast Pacific. We analyze model and data error structures by simulating the model without data assimilation. The results show that the model forecast on average explains 33% of the data variability. The Kalman filter updates the model very efficiently and produces an estimate which explains 76% of the data variance. The optimal smoother estimate is very similar to the Kalman filter estimate. Running the model in other regions of the Pacific produced worse fits of the model to the data. This supports the conclusion that the linear barotropic dynamics fails to describe the SSH variability.

Thesis Supervisor: Carl Wunsch,

Title: Cecil and Ida Green Professor of Physical Oceanography

## Acknowledgements

I am thankful to my thesis advisor, Carl Wunsch, not only for suggesting this topic, but also for allowing me to explore the subject freely and independently, while also providing guidance and support whenever it was needed. I would also like to thank my many fellow students who spent endless hours in often fruitless discussions. I am especially indebted to Francois Primeau who taught me the basics of numerical integration and read through most of this manuscript. Charmaine King kindly provided me with the data. Last, but by no means least, I would like to thank my wife, Natasha, for her love and care.

This research was partially funded by a NASA Global Change Fellowship.

# Contents

<b>Abstract</b>	<b>3</b>
<b>Acknowledgements</b>	<b>4</b>
<b>Introduction</b>	<b>14</b>
<b>1 Physical model.</b>	<b>18</b>
1.1 Notation . . . . .	20
<b>2 The Data</b>	<b>21</b>
2.1 Variability of the T/P data for the northwestern Pacific. . . . .	22
2.2 Comparison of variability of inputs and outputs. . . . .	24
<b>3 Model validation through system identification.</b>	<b>26</b>
3.1 Literature review. . . . .	26
3.2 Statistical Model. . . . .	27
3.2.1 Averaging . . . . .	30
3.3 Simplified Physical Model . . . . .	32

3.4	Results . . . . .	33
3.5	Results for more complicated models. . . . .	37
3.6	Conclusions. . . . .	40
<b>4</b>	<b>State-space approach.</b>	<b>42</b>
4.1	Overview of the method. . . . .	42
4.1.1	Kalman filter and RTS smoother. . . . .	44
4.2	Implementation of the model. . . . .	45
4.3	Literature review. . . . .	47
4.4	Model and data error estimates. . . . .	49
4.5	Results. . . . .	51
4.6	Conclusions. . . . .	55
	<b>Appendix 1.</b>	<b>57</b>
	<b>Bibliography</b>	<b>59</b>

## List of Figures

1	A system. . . . .	15
2	SSH from T/P at 200° E, 40° N in cm. Time mean has been removed. . .	62
3	SSH from T/P at 200° E, 40° N in cm. Annual, semi-annual cycles, and time mean have been removed. . . . .	62
4	Mean SSH spatial variability from gridded T/P data as a function of time.	63
5	Contours of mean SSH spatial variability from gridded T/P data. . . . .	63
6	Spatial distribtuion of SSH from T/P at 78th, 79th, and 80 T/P cycles. Time mean has been removed. . . . .	64
7	gridded SSH from T/P at 182° E, 50° N in cm. . . . .	65
8	Mean SSH spatial variability from averaged T/P data as a function of time.	66
9	Contours of mean SSH spatial variability from averaged T/P data. . . . .	66
10	Mean SSH spatial variability from filtered T/P data as a function of time.	67
11	Contours of mean SSH spatial variability from filtered T/P data. . . . .	67
12	Mean SSH spatial variability from filtered but not averaged T/P data as a function of time. . . . .	68
13	Contours of mean SSH spatial variability from filtered but not averaged T/P data. . . . .	68

14	Contours of percentage of explained variance and the histogram of explained variance. . . . .	69
15	Distribution of regression coefficients for normalized variables for the filtered data. . . . .	70
16	Distribution of regression coefficients for normalized variables after the coefficients have been averaged in space. . . . .	71
17	Contours of percentage of explained variance and the histogram of explained variance for the band-passed data. . . . .	72
18	Distribution of regression coefficients for normalized variables for the band-passed data. . . . .	73
19	Distribution of regression coefficients for normalized variables after the coefficients have been averaged in space and the data has been band-passed. . . . .	74
20	Contours of percentage of explained variance and the histogram of explained variance for the model with differenced $\eta$ as the output. . . . .	75
21	Distribution of regression coefficients for normalized variables for $(\nabla \times \tau)_z$ (upper left), $\tau^x$ (upper right), and $\partial\zeta_a/\partial x$ (lower left). The values below minus one standard deviation are denoted by black, above one standard deviation by white, and in between by gray. The underlying contours show bathymetry at the contour spacing of 1 km. . . . .	76
22	Distribution of regression coefficients for normalized variables for $(\nabla \times \tau)_z$ (upper left), $\tau^x$ (upper right), and $\partial\zeta_a/\partial x$ (lower left). The values below minus one standard deviation are denoted by black, above one standard deviation by white, and in between by gray. The coefficients have been averaged in space. . . . .	77
23	Contours of percentage of explained variance and the histogram of explained variance for the model with differenced $\eta$ as the output and the band-passed data. . . . .	78

24	Distribution of regression coefficients for normalized variables for $(\nabla \times \tau)_z$ (upper left), $\tau^x$ (upper right), and $\partial\zeta_a/\partial x$ (lower left). The values below minus one standard deviation are denoted by black, above one standard deviation by white, and in between by gray. The data has been band-passed.	79
25	Distribution of regression coefficients for normalized variables for $(\nabla \times \tau)_z$ (upper left), $\tau^x$ (upper right), and $\partial\zeta_a/\partial x$ (lower left). The values below minus one standard deviation are denoted by black, above one standard deviation by white, and in between by gray. The coefficients have been averaged in space. The data has been band-passed. . . . .	80
26	Contours of percentage of explained variance and the histogram of explained variance. The data has been averaged in space. . . . .	81
27	Distribution of regression coefficients for normalized variables for the data averaged in space. . . . .	82
28	Contours of percentage of explained variance and the histogram of explained variance for the model with topographic beta. . . . .	83
29	Distribution of regression coefficients for normalized variables for $(\nabla \times \tau)_z$ (upper left), $\tau^x$ (upper right), $\partial\zeta_a/\partial x$ (lower left), and $\partial\zeta_a/\partial y$ (lower right). The values below minus one standard deviation are denoted by black, above one standard deviation by white, and in between by gray. The underlying contours show bathymetry at the contour spacing of 1 km. . . . .	84
30	Contours of percentage of explained variance and the histogram of explained variance for the model with topographic beta without the SLP. . .	85
31	Distribution of regression coefficients for normalized variables for $(\nabla \times \tau)_z$ (upper) and $\tau^x$ (lower). The values below minus one standard deviation are denoted by black, above one standard deviation by white, and in between by gray. The underlying contours show bathymetry at the contour spacing of 1 km. . . . .	86

32	Contours of percentage of explained variance and the histogram of explained variance for the model with topographic beta without the SLP for the northeastern Pacific only. . . . .	87
33	Distribution of regression coefficients for normalized variables for $(\nabla \times \tau)_z$ (upper) and $\tau^x$ (lower)for the northeastern Pacific only. The values below minus one standard deviation are denoted by black, above one standard deviation by white, and in between by gray. The underlying contours show bathymetry at the contour spacing of 1 km. . . . .	88
34	Contours of root mean square variability for $\frac{\partial \eta}{\partial x} \left( \frac{\beta}{f} - \frac{h_y}{h} \right) + \frac{\partial \eta}{\partial y} \left( \frac{h_x}{h} \right)$ (upper), $\frac{f}{\rho gh} (\nabla \times \tau)_z$ (middle), and $\frac{\beta}{\rho gh} \tau^x$ (lower) in $m^2$ for the northeastern Pacific. . . . .	89
35	Contours of log of rms( $\frac{f}{\rho gh} (\nabla \times \tau)_z$ ) over rms( $\frac{\partial \eta}{\partial x} \left( \frac{\beta}{f} - \frac{h_y}{h} \right) + \frac{\partial \eta}{\partial y} \left( \frac{h_x}{h} \right)$ ) (upper) and log of rms( $\frac{\beta}{\rho gh} \tau^x$ ) over rms( $\frac{\partial \eta}{\partial x} \left( \frac{\beta}{f} - \frac{h_y}{h} \right) + \frac{\partial \eta}{\partial y} \left( \frac{h_x}{h} \right)$ ) (lower) for the northeastern Pacific. . . . .	90
36	Contours of root mean square variability for $\frac{\partial \eta}{\partial x} \left( \frac{\beta}{f} - \frac{h_y}{h} \right) + \frac{\partial \eta}{\partial y} \left( \frac{h_x}{h} \right)$ (upper), $\frac{f}{\rho gh} (\nabla \times \tau)_z$ (middle), and $\frac{\beta}{\rho gh} \tau^x$ (lower) in $m^2$ for all of the Pacific. . . . .	91
37	Contours of log of rms( $\frac{f}{\rho gh} (\nabla \times \tau)_z$ ) over rms( $\frac{\partial \eta}{\partial x} \left( \frac{\beta}{f} - \frac{h_y}{h} \right) + \frac{\partial \eta}{\partial y} \left( \frac{h_x}{h} \right)$ ) (upper) and log of rms( $\frac{\beta}{\rho gh} \tau^x$ ) over rms( $\frac{\partial \eta}{\partial x} \left( \frac{\beta}{f} - \frac{h_y}{h} \right) + \frac{\partial \eta}{\partial y} \left( \frac{h_x}{h} \right)$ ) (lower) for all of the Pacific. . . . .	92
38	Contours of percentage of explained variance and the histogram of explained variance for the model with the time derivative of the Laplacian without the $\tau^x$ term. . . . .	93
39	Distribution of regression coefficients for normalized variables for $(\nabla \times \tau)_z$ (upper). The values below minus one standard deviation are denoted by black, above one standard deviation by white, and in between by gray. The underlying contours show bathymetry at the contour spacing of 1 km. . . . .	94

40	Distribution of regression coefficients for normalized variables for $(\nabla \times \tau)_z$ (upper). The values below minus one standard deviation are denoted by black, above one standard deviation by white, and in between by gray. The underlying contours show bathymetry at the contour spacing of 1 km.	95
41	Contours of percentage of explained variance and the histogram of explained variance for the model with the time derivative of the Laplacian.	96
42	Distribution of regression coefficients for normalized variables for $(\nabla \times \tau)_z$ (upper) and $\tau^x$ (lower). The values below minus one standard deviation are denoted by black, above one standard deviation by white, and in between by gray. The underlying contours show bathymetry at the contour spacing of 1 km.	97
43	Contours of root mean square variability for $\frac{\partial}{\partial t} \nabla^2 \eta + \left(\beta - f \frac{h_y}{h}\right) \frac{\partial \eta}{\partial x} + \left(f \frac{h_x}{h}\right) \frac{\partial \eta}{\partial y}$ (upper), $\frac{f}{\rho g h} (\nabla \times \tau)_z$ (middle), and $\frac{\beta}{\rho g h} \tau^x$ (lower) in the Pacific.	98
44	Contours of log of rms( $\frac{f}{\rho g h} (\nabla \times \tau)_z$ ) over rms( $\frac{\partial}{\partial t} \nabla^2 \eta + \left(\beta - f \frac{h_y}{h}\right) \frac{\partial \eta}{\partial x} + \left(f \frac{h_x}{h}\right) \frac{\partial \eta}{\partial y}$ ) and log of rms( $\frac{\beta}{\rho g h} \tau^x$ ) over rms( $\frac{\partial}{\partial t} \nabla^2 \eta + \left(\beta - f \frac{h_y}{h}\right) \frac{\partial \eta}{\partial x} + \left(f \frac{h_x}{h}\right) \frac{\partial \eta}{\partial y}$ ) for the Pacific.	99
45	Contours of percentage of explained variance and the histogram of explained variance for the auto-regressive model.	100
46	Distribution of regression coefficients for normalized variables $-hJ(\eta, \frac{f+\beta y}{h})$ (upper left), $(\nabla \times \tau)_z$ (upper right), and $\tau^x$ (lower). The values below minus one standard deviation are denoted by black, above one standard deviation by white, and in between by gray. The underlying contours show bathymetry at the contour spacing of 1 km.	101
47	Distribution of regression coefficient on $\nabla^2 \eta(i, j, t - 1)$ .	102
48	Contours of percentage of explained variance and the histogram of explained variance for the auto-regressive model in the south eastern Pacific.	103

49 Estimates of the mean SSH variance ( $\text{cm}^2$ ) computed for each T/P cycle as an initial condition showing (a) data error and )b) simulation error. The dashed curves are the data and simulation variance, respectively. . . . . 104

50 Contours of the estimate of (a) data error covariance  $R$  and (b) simulation error covariance  $EPE^T$ . . . . . 104

51 Contours of the SSH at  $t=0$  (2nd T/P cycle) and 10 days later (3rd T/P cycle) from the data. Annual, semi-annual cycles, and time mean have been removed. . . . . 105

52 Contours of potential vorticity. . . . . 105

53 Comparison of a) model forecast  $q(25, -)$ , b) Kalman filter assimilation . . 106

54 The amount of variance of the data ( $E_h$ , solid), accounted for by the forecast sea level( $E_f$ ,dotted), by the Kalman filter update ( $E_d$ ,dashed), the RTS smoother estimate ( $E_s$ , dash-dotted), and the initial conditions ( $E_i$ , circles), for the model in Northeastern Pacific. The values in the box represent mean values for the respective variables. . . . . 107

55 Evolution of mean variance of the (a) the model forecast, (b) the Kalman filter update, and (c) the smoothed estimate. The time is in T/P cycles. The steady filter was employed 5 T/P cycles, and therefore, the lines are constant for  $t > 6\text{T/P}$  cycles. The units are  $m^2$ . . . . . 108

56 Comparison of SSH variability at  $200^\circ$  E and  $40^\circ$  N for the Nort east Pacific assimilation: the data (solid line), the model forecast (dashed line), the kalman filter assimilation (dotted line), and the smoother estimate (dash-dotted line). T/P data was averaged over  $10^\circ$  by  $10^\circ$  and had annual and semi-annual cycles removed. . . . . 109

57 Comparison of SSH variability at  $220^\circ$  E and  $20^\circ$  N for the central east Pacific assimilation: the data (solid line), the model forecast (dashed line), the kalman filter assimilation (dotted line), and the smoother estimate (dash-dotted line). T/P data was averaged over  $10^\circ$  by  $10^\circ$  and had annual and semi-annual cycles removed. . . . . 109

58 Longitude versus time plot of SSH anomalies relative to the mean at  $40^\circ$  N from [Left to right, top to bottom]: the observed field (filtered T/P data)  $\mathbf{h}(t)$ , the model forecast  $\mathbf{q}(t, -)$ , the Kalman filter update  $\mathbf{q}(t)$ , and the smoother estimate  $\mathbf{q}(t, +)$ . The resolution of the plots is 1 T/P cycle by  $2^\circ$ . T/P data was in addition averaged over  $10^\circ$  by  $10^\circ$  and had annual and semi-annual cycles removed. . . . . 110

59 The amount of variance of the data ( $E_h$ , solid), accounted for by the forecast sea level ( $E_f$ , dotted), by the Kalman filter assimilation ( $E_d$ , dashed), the data update ( $E_d$ , dash-dotted), and the initial conditions ( $E_i$ , circles), for the model in south east Pacific. The values in the box represent mean values for the respective variables. . . . . 111

60 The amount of variance of the data ( $E_h$ , solid), accounted for by the forecast sea level ( $E_f$ , dotted), by the Kalman filter assimilation ( $E_d$ , dashed), the data update ( $E_d$ , dash-dotted), and the initial conditions ( $E_i$ , circles), for the model in central east Pacific. The values in the box represent mean values for the respective variables. . . . . 112

## Introduction.

In recent years it became clear that to understand human induced climate change we first need to understand natural variability of the world climate. The world ocean is one of the parts of the climate system which we understand least. The spatial scales of the large scale ocean circulation are grand, and time scales are very long. To date, the dynamics of this enormous system has been grossly undersampled and a very elegant linear theory of general circulation variability has not been tested adequately by observations. The ultimate limitation was the absence of a truly global reliable dataset. Such a dataset is now available from the TOPEX/POSEIDON (T/P) altimetry satellite.

The field of physical oceanography aims at describing the ocean dynamics. There is a well developed linear theory of the global circulation which starts from the Navier-Stokes equations and mass conservation and develops a dynamic equation for time evolution of the sea surface height (SSH). In this work we will only consider the part of the theory which assumes constant density of the ocean. Although it is well known that stratification is of primary importance, it has been long argued that at large spatial and long time scales the dynamics can be well approximated by barotropic models. The atmosphere is considered as known and provides a boundary condition at the sea surface. We test the hypothesis that we can predict time variability of the ocean, namely the SSH, from the knowledge of the atmospheric fields by using linear barotropic models. We study skill of different linear models in describing time variability of the SSH. We start with the simplest linear balance, the balance between the advection of the planetary vorticity and the curl of the wind stress, and proceed by including effects of bottom topography and time dependence. The derivation of the equations is presented in Chapter 1.

The data consist of altimetric measurements averaged over  $2^\circ$  by  $2^\circ$  by 10 days. The altimetry data measure sea surface height (SSH) relative to geoid, an equipotential surface of the actual gravitational field. These data are described in Chapter 2.

We chose two different approaches: system identification analysis and state-space models. The difference between these two methods can be described as follows. The ultimate goal is to establish a functional relationship between the output, or the field which we aim to predict, and the inputs, or the forcing. In this work, the inputs stand for atmospheric data and the outputs stand for oceanic data. The system is subject

to random shock, or noise, which we cannot observe and predict. The system can be described by a diagram shown on Figure 1.

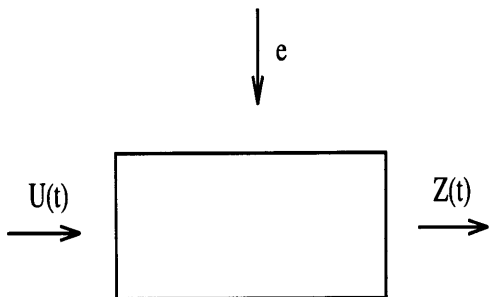


Figure 1: A system.

The main idea can be illustrated as follows. Suppose we would like to test a very simple dynamical system which can be described by the following equation:

$$\frac{\partial z(t, x)}{\partial t} = \frac{\partial u(t, x)}{\partial x} + e(t, x), \quad t \geq 0 \quad (1)$$

where  $z(t, x)$  is the output;  $u(t, x)$  is the input, assumed to be known;  $e(t, x)$  is the noise, commonly assumed to be white;  $t$  and  $x$  are independent variables denoting temporal and spatial coordinates respectively. The aim is to compare the observed field  $z_{\text{obs}}(t, x)$  with the output field  $z(t, x)$  produced by forcing the dynamical system (1) with observed data  $u_{\text{obs}}(t, x)$ . From a purely deterministic point of view, or, equivalently, neglecting the unknown noise  $e(t, x)$ , we can solve the equation (1) given some set of initial and boundary conditions (the latter absent in this very simple model), say

$$z(0, x) = z_1(x) \quad (2)$$

and forcing  $u(t, x) = u_{\text{obs}}(t, x)$  for all  $t$  and  $x$ . For this example, we can write an explicit solution for  $z(t, x)$

$$z(t, x) = \int_0^t \frac{\partial u_{\text{obs}}(t', x)}{\partial x} dt' + z_1(x)$$

and then compare it with the observed data  $z_{\text{obs}}(t, x)$ . Depending on the results of the comparison we can accept or reject the model (1) as compatible with the observed data.

However, we can address the problem from a different point of view. We can check whether the balance (1) holds by doing simple finite differencing of both input and output:

$$\frac{z_{\text{obs}}(t, x) - z_{\text{obs}}(t - \delta t, x)}{\delta t} = \frac{u_{\text{obs}}(t, x) - u_{\text{obs}}(t, x - \delta x)}{\delta x} \quad (3)$$

We then investigate whether the two sides satisfy that relationship, e.g. whether the residual of the two sides of equation (3) satisfies the assumption of white noise and has sufficiently low variance.

There are advantages and disadvantages to both methods. The first method involves integrating the observed forcing field, which, to our benefit, is an error reducing procedure for white noise. In the second method we difference the observed fields, thus, to our disadvantage, enhancing the error of the predicted fields. However, the second approach does not require knowledge of the solution to the analog of equation (1). Such a solution may be very complicated or simply unavailable in analytic form and require numerical integration. Neither do we need to know the initial and boundary conditions and the forcing everywhere in the second method if we are only interested in validity of the model (1) in some part of the domain. Those two reasons indicate benefits of the second method. Using the first method we can consider the model in a smaller domain with the possible complications arising along the “artificial” boundaries. For example, it is a standard practice to use open boundary conditions and so-called “sponge layers”.

In this work we use both approaches. We believe that this provides a good check for the consistency of the two methods. We start with the time-series approach because the simplest balance has no time derivatives and can only be tested by the second method. It is described in Chapter 3. We treat it as a multichannel regression problem. We choose linear models with a small number of degrees of freedom (parameters). In order to do that we construct a local ARMAX (auto-regressive moving average external) model for the sea surface height as an output, and wind forcing (either wind velocities, or wind stress) and sea level pressure (SLP) as inputs. Such a model explains local variability of output in terms of that of inputs. The model can be auto-regressive both in space and time. We present the results for an hierarchy of linear barotropic models. The results indicate that these simple models fail to adequately describe the variability of the SSH, with the possible exception of some isolated areas. However, because of the deficiencies of the time-series analysis, we turn to the first approach for the support of our conclusions.

The first approach is pursued in the framework of the control theory, namely the Kalman filter and optimal smoother. It is described in Chapter 4. Both the model and the observations are assumed noisy. We provide an overview of the approach and review

of the literature, and describe model and data errors. We then describe results of a run in the northeastern Pacific. The model is propagating the signal in a way which agrees poorly with the data. Running the model in similar size regions in south and central east Pacific showed that in those regions the model is not at all consistent with the data. The results support the conclusion of the Chapter 3 that these simple models fail to explain global variability of the SSH.

## Chapter 1

### Physical model.

Let us assume linear, homogeneous,  $\beta$ -plane dynamics based on the reasoning that it describes most of variability on low-frequency and low-wavenumber scales. This hypothesis has been argued by means of theoretical studies for a long time. The oceanic response to time-dependent forcing can either be trapped (vertically or horizontally) in the neighborhood of the region that is forced, or propagate away from the region in form of the waves. Frankignoul and Müller (1979) described the quasi-geostrophic response of stratified flat bottom ocean to stochastic atmospheric forcing on the beta-plane. They argued that wind stress is the dominant forcing mechanism under the quasi-geostrophic approximation. In the case of flat bottom and under linear approximation one can separate the dynamics into an infinite number of vertical modes, the so-called barotropic, and infinite number of baroclinic modes. The corresponding Rossby radius of deformation is about 3200 km for the barotropic mode, 46 km for the first baroclinic mode, 20 km for the second baroclinic mode, etc. Because the model is linear, the oceanic response spectra is completely determined by the atmospheric spectra. Thus, the authors were able to describe oceanic response to idealized wind forcing. The response can be resonant and off-resonant. The off-resonant behavior corresponds to trapping of the response. The resonant response is identified as propagating waves. The barotropic response is resonant at practically all frequencies. The baroclinic response is resonant at frequencies corresponding to periods larger than 150 days for the idealized atmospheric forcing. Willebrand et al. (1980) described more realistic forcing fields obtained from NMC data. Based on the analysis of realistic atmospheric spectra they argued that at periods between 1 and 300 days the oceanic response to large-scale forcing is essentially barotropic.

For our modeling, we do *not* assume that the bottom is flat. The system of equations we use is equivalent to so-called Laplace tidal equations:

$$\frac{\partial u}{\partial t} - fv = -g \frac{\partial(\eta - \zeta_a)}{\partial x} + \frac{\tau^x}{\rho h} \quad (1.1)$$

$$\frac{\partial v}{\partial t} + fu = -g \frac{\partial(\eta - \zeta_a)}{\partial y} + \frac{\tau^y}{\rho h} \quad (1.2)$$

$$\frac{\partial \eta}{\partial t} + \frac{\partial(hu)}{\partial x} + \frac{\partial(hv)}{\partial y} = 0. \quad (1.3)$$

The notation is conventional and described in section (1.1). If all dynamic variables have a time dependence proportional to  $\exp(i\sigma t)$  one can readily derive an equation for SSH from the system above:

$$\begin{aligned} \nabla^2 \eta + \frac{\partial \eta}{\partial x} \left( \frac{-\beta(\sigma^2 + f^2)}{i\sigma(\sigma^2 - f^2)} + \frac{h_x}{h} - \frac{fh_y}{i\sigma h} \right) + \frac{\partial \eta}{\partial y} \left( \frac{2f\beta}{(\sigma^2 - f^2)} + \frac{fh_x}{i\sigma h} + \frac{h_y}{h} \right) + \frac{(\sigma^2 - f^2)}{gh} \eta = \\ \frac{f}{i\sigma g} \left( \frac{\partial(\tau^y/\rho h)}{\partial x} - \frac{\partial(\tau^x/\rho h)}{\partial y} \right) + \frac{(\tau^x/\rho h)}{g} \left( \frac{-\beta(\sigma^2 + f^2)}{i\sigma(\sigma^2 - f^2)} + \frac{h_x}{h} - \frac{fh_y}{i\sigma h} \right) \\ + \frac{(\tau^y/\rho h)}{g} \left( \frac{2f\beta}{(\sigma^2 - f^2)} + \frac{fh_x}{i\sigma h} + \frac{h_y}{h} \right) - \frac{1}{g} \left( \frac{\partial(\tau^x/\rho h)}{\partial x} + \frac{\partial(\tau^y/\rho h)}{\partial y} \right) + \nabla^2 \zeta_a \\ + \frac{\partial \zeta_a}{\partial x} \left( \frac{-\beta(\sigma^2 + f^2)}{i\sigma(\sigma^2 - f^2)} + \frac{h_x}{h} - \frac{fh_y}{i\sigma h} \right) + \frac{\partial \zeta_a}{\partial y} \left( \frac{2f\beta}{(\sigma^2 - f^2)} + \frac{fh_x}{i\sigma h} + \frac{h_y}{h} \right). \end{aligned} \quad (1.4)$$

The equation (1.4) is very general and is valid for any set of parameters for which linearity is a good assumption. We denote the Earth radius by  $R$ . If  $\sigma \ll f$ , i.e. the Rossby number  $\sigma/f$  is small, a spatial scale of the motions is such that  $R\sigma/Lf \ll 1$ , i.e.  $\frac{\partial}{\partial t} \frac{\partial}{\partial x} \ll \beta$ , and the bottom topography is constant, the equation (1.4) becomes

$$\beta \frac{\partial \eta}{\partial x} = \frac{f}{\rho gh} \left( \frac{\partial \tau^y}{\partial x} - \frac{\partial \tau^x}{\partial y} + \frac{\beta \tau^x}{f} \right) \quad (1.5)$$

If we drop the assumptions of flat bottom and that  $R\sigma/Lf \ll 1$ , and substitute back  $\partial/\partial t$  for  $i\sigma$  we obtain a following balance:

$$\frac{\partial}{\partial t} \nabla^2 \eta + hJ \left( \eta, \frac{f + \beta y}{h} \right) = \frac{f}{\rho gh} \left[ \left( \frac{\partial \tau^y}{\partial x} - \frac{\partial \tau^x}{\partial y} \right) + \frac{\beta \tau^x}{f} \right]; \quad (1.6)$$

Other useful approximations and resulting equations are considered in Chapter 3.

## 1.1 Notation

We use the following notation throughout the paper:

---

$\eta$ , sea surface height relative to temporal mean

$\tau = (\tau^x, \tau^y)$ ,  $x$  and  $y$  components of wind stress at the sea level

$\zeta_a$ , “height” corresponding to atmospheric pressure,  $\zeta_a = -p_a/\rho g$

$f$ , mean value of Coriolis parameter,  $f = 2\Omega \sin \theta_0$

$\beta$ , coefficient of latitudinal variation of Coriolis parameter,  $\beta = 2\Omega \cos \theta_0/R$

$g$ , gravity

$L$ , spatial scale of the motions, assumed to represent both  $x$  and  $y$  scales

$R$ , radius of the Earth

$T$ , temporal scale of the motions

$T_{inertial}$ , inertial time scale

$h$ , bottom relief, or mean depth of homogeneous water model depending on context

$\rho$ , density of homogeneous fluid.

$J$ , the Jacobian:  $J(f, g) = \frac{\partial f}{\partial x} \cdot \frac{\partial g}{\partial y} - \frac{\partial f}{\partial y} \cdot \frac{\partial g}{\partial x}$ .

---

## Chapter 2

### The Data

Altimetric data provide estimates of sea surface height (SSH) relative to the geoid. Other oceanic variables can be obtained by employing various approximations, such as geostrophy for horizontal velocities. All data were mapped on a  $2^\circ \times 2^\circ \times 10$  days grid. Sea surface height was obtained from the T/P experiment and processed to remove tidal and inverted barometer effects, as described in (King et al., 1994). The atmospheric data consisted of sea level wind velocity and stress fields and sea level pressure as provided by ECMWF. One can consult Trenberth et al. (1989), for the details of the computation of the wind stress field. After a series of experiments, we decided to use the wind stress data rather than wind velocity data since model fits were basically the same, but wind stress was easier to interpret as forcing. The bathymetry was obtained from a National Geophysical Data Center gridded bathymetry database and smoothed to  $2^\circ$  by  $2^\circ$  to provide a smoothed version which influences the large-scale flow of interest. This particular choice of smoothing scale is based on a study by Koblinsky et al. (1989), who reported that a smoothing scale of 175 km resulted in the best agreement with the theoretical analysis based on the topographic Sverdrup relation.

We subtracted the time mean at every grid point and interpolated over the missing values by setting them to zero after the subtraction of the time mean. These data are henceforth referred to as “gridded data”. We analyze the variability of the northwestern Pacific, the region which will be shown to be one of the most likely candidates for the validity of the barotropic dynamics. The analysis of variability after different filters are applied is similar in other regions.

## 2.1 Variability of the T/P data for the northwestern Pacific.

We consider the gridded data in the area between  $180^\circ$  E and  $224^\circ$  E and  $30^\circ$  N and  $50^\circ$  N. The gridded data are variable in three dimensions: two spatial and one temporal. We reshape the two spatial coordinates into one coordinate by reindexing. We can assume that data is stationary in time and/or in space. Depending on which assumption we use we can substitute the ensemble average by averaging over either space or time domain. If we take the gridded T/P data the mean spatial variance is  $24 \text{ cm}^2$ . The variability is shown on Figure 4 as a function of time. We observe that there are two very big spikes in the data at 31st and 79th T/P cycles, and additional spikes at cycles 10, 20, 41, and 55. Otherwise, the distribution is fairly uniform.

The spikes which appear on the graph of the spatial variance, Figure 4, are present in the gridded data as well. In fact, if we plot the spatial distribution of the data as a vector (converting two dimensional space into one dimensional by reindexing) at the cycles 78, 79, and 80 we see that the spikes appear only at cycle 79, see Figure 6. Furthermore, big spikes appear only at three locations in the northwestern corner of the domain:  $180^\circ$  E and  $48^\circ$  N,  $180^\circ$  E and  $50^\circ$  N,  $182^\circ$  E and  $50^\circ$  N, which are on the boundary of the domain. We plot the raw gridded data at  $182^\circ$  E and  $50^\circ$  N on Figure 7. We believe that the source of these very significant outliers is the fact that the data comes from two sources: TOPEX and POSEIDON. Most of the data comes from TOPEX. However, the POSEIDON observations are used at six cycles: 20, 31, 41, 55, 65, and 79, which are the cycles where we observe the spikes. The POSEIDON data have a lot of missing data and spurious data. Because the TOPEX data are not available at these times we can either interpolate in time between cycles 30 and 32, 78 and 80, and so on, or proceed with the analysis using those “bad” data. We choose the latter because spikes disappear after the averaging and the fact that the quality of the POSEIDON data is satisfactory over most of the domain. Wunsch and Stammer (1995) used the POSEIDON data and it did not lead to any inconsistencies, although it did increase the noise level at some locations.

The mean temporal variance is  $42 \text{ cm}^2$  and the contours of variability are plotted on Figure 5 in a horizontal plane. We see that the southwestern corner of the region is most energetic. The values in that region reach  $120 \text{ cm}^2$ , which is much larger than the mean values over the domain. Overall, the variability is diminishing northward. The

northwestern corner is also very energetic as it is directly influenced by Kuroshio. This is seen very clearly from maps of global variability shown in Wunsch and Stammer (1995).

After averaging the data over  $10^\circ$  by  $10^\circ$  boxes the variability changes significantly. The mean spatial variability reduces to  $3.4 \text{ cm}^2$  and its temporal evolution is shown on Figure 8. We immediately see that the two big spikes are gone completely and the values dropped by almost an order of magnitude. However, a new spike at the very end of the time series is produced. The aerial averaging reduces small scale variability. The mean temporal variance is reduced much less and is equal to  $20 \text{ cm}^2$ . Its spatial distribution is shown on Figure 9. The variability in the western part of the region has reduced a lot because we averaged the data with much less energetic central part. However, the mean gradient of the variability is still south ward and westward, although it is much smaller than for the gridded data.

If we then eliminate the annual and semi-annual cycles we reduce the temporal variability a great deal. We used a simple band-stop filter to remove annual and semi-annual frequencies. It was shown to be a necessary step in the analysis of the inverted barometer effect (Gentemann, 1995), as otherwise the inverted barometer effect was dominated by very strong annual and semi-annual cycles in the SSH and SLP data. To do the filtering we removed annual and semi-annual harmonics obtained from the discrete Fourier transform of exactly 2 years of data at every grid point. Example of time series of the data used in the analysis at  $200^\circ \text{ E}$ ,  $40^\circ \text{ N}$  is presented on Figure 2 and after the removal of the annual and semi-annual cycles on Figure 3. In Figure 2 one can clearly see a strong annual cycle which is absent from Figure 3.

The mean value of temporal variability drops to  $6.6 \text{ cm}^2$ . The contours are shown on Figure 11. Now the western part is no longer the most energetic, which shows that the annual and semi-annual cycles dominated the variability in that region more than in any other region. Variability increases southward. The spatial variability changes very little to  $3.1 \text{ cm}^2$ . The time evolution is shown on Figure 10, and is very similar to that of averaged data, Figure 8. Again, we observe a large spike at 55th cycle caused by the "bad" POSEIDON observations, large increase of variance toward the last observation, and fluctuations around the mean value rest of the time. However, the spike is not a significant outlier and we do not interpolate over the POSEIDON observations. Thus,

the eliminating annual cycle independently at each location does not change time evolution of mean spatial variability.

If we eliminate the annual and semi-annual cycles without the spatial averaging the mean temporal variance drops to 25.0 cm<sup>2</sup> but the mean spatial variance is reduced only to 21.1 cm<sup>2</sup>. The time evolution of the mean spatial variance, shown on Figure 12, is very close to that of the gridded data, Figure 4. The time series is shifted downward but no spikes or other significant features are removed. The contours of temporal variability are shown 13 in the horizontal plane. We see again that the annual cycle is the strongest in the western part of the region and the field is very inhomogeneous.

## 2.2 Comparison of variability of inputs and outputs.

We did a comparison of variability of inputs and outputs for the time-dependent linear vorticity balance (1.6) in the northeastern Pacific. We convert all quantities to sec<sup>-2</sup> units by multiplying both sides of the equation by  $g/f$ . A similar analysis was done in Fu and Pihos, (1995) for the global variability. They used the data averaged over 10° by 10° boxes, centered on a 5° by 5° grid. The variability of the terms on the left hand side of the linear vorticity balance (1.6) (the outputs) decreases by an order of magnitude if we perform averaging. We obtain that at 40° N, 200° E rms variability of the planetary vorticity term is equal to 4.7e-14 sec<sup>-2</sup> for averaged over 10° by 10° data, and 2.2e-13 sec<sup>-2</sup> for the gridded data. On the other hand, the rms variability of the wind forcing is only slightly changed: 2.1e-14 sec<sup>-2</sup> for averaged over 10° by 10° data and 3.3e-14 sec<sup>-2</sup> for the gridded data. The same is true all over the model domain, and the mean rms variability of the planetary vorticity is 5.0e-14 sec<sup>-2</sup> for averaged over 10° by 10° data, as compared to 3.0e-13 sec<sup>-2</sup> for the gridded data. The planetary vorticity term is just one of the terms on the left hand side of the balance. Therefore, we conclude that for the gridded data variability of the LHS is more than an order of magnitude larger than that of RHS. We define “averaged” data as the data where averaging over 10° by 10° has been performed. Although we start in Chapter 3 with gridded data we show that most of the variability cannot be explained by the model. We then continue with the

analysis of the averaged data. For the analysis of the model based on Kalman filter and optimal smoother in Chapter 4 we use only the averaged data.

## Chapter 3

### Model validation through system identification.

#### 3.1 Literature review.

In a recent paper Wunsch (1991) performed an analysis of large scale oceanic variability based on altimetric Geosat satellite observations. Atmospheric data used in the analysis included NMC wind data and sea level pressure. The paper demonstrated the application of multichannel regression models to time-dependent sea surface height observations and atmospheric fields. It addressed the important question of whether sea surface height variability in the ocean can be explained by that of atmospheric forcing. This is a particular aspect of ocean atmosphere coupling. It was pointed out that one of the major disadvantages was the poor quality of the data. With the availability of much better quality data from the TOPEX/POSEIDON (T/P) experiment we decided to carry out a more extensive analysis in the same framework. With lower orbital error, down to approximately 4 cm from about 25 cm, one can hope to see improvements in the analysis of linear models on the long spatial and temporal scales.

There are a number of more recent publications which deserve attention. Fu and Pihos (1994) investigated the validity of the so-called inverted barometer (IB) effect from T/P data. The IB effect is characterized by the rise and fall of sea level to compensate for the change of atmospheric pressure at the rate of -1 cm/bar. It is a purely static phenomenon. Fu and Pihos used the same framework of multivariate time series analysis as Wunsch (1991). First, they performed regression of SSH on sea level pressure alone. The results were unsatisfactory even in mid-latitudes. Then they included effects of the

wind forcing on the SSH and concluded that in regions outside of the tropics the IB effect is valid on time scales between 20 and 300 days. This result was quite different from the Geosat result of vanDam and Wahr (1993), who reported a regression coefficient of about  $-0.7$  cm/bar. The results were believed to be different due to inaccuracy of Geosat data. However, the results for wind regression coefficients from multichannel regression in Fu and Pihos (1994) were inconsistent with the underlying physical model and significantly different from those of Wunsch (1991). We try to resolve these issues in this work. Ponte (1994) used a shallow water numerical model to establish relation between wind- and pressure-driven sea level variability. He proposed that multivariate statistical models would not be very useful in analysis of large scale variability because of remote wind forcing and non-linear effects. Our present findings seem to support these ideas. Fu and Davidson, (1995) looked at the validity of the time-dependent linear vorticity balance in Pacific Ocean by using a single channel model. Their results were largely unsatisfactory, and the balance was rejected except for a few isolated points in mid- and high-latitudes. We investigate the validity of this balance through a number of multichannel statistical models. Our results support their findings mostly at the same isolated locations.

### 3.2 Statistical Model.

The problem falls under the category of system identification. It is a well developed subject with an extensive literature, for example, Ljung (1987), Box and Jenkins (1976), and a number of specialized programs, such as Matlab System Identification Toolbox, Ljung (1994). The fields, or functions, we are dealing with change in time and in space, e.g. SSH observations  $h(x, y, t)$ . The SSH observations are provided in a finite number of spatial locations  $(x_i, y_j)$ , and we denote the observations as multichannel time series  $\mathbf{h}(t) = \{h(i, j, t)\}$ . It is most common to have  $t$  in regularly spaced intervals, which after setting the initial time to zero, take the form:  $t_k = k\Delta t, k = 0, 1, \dots, N$ . In this analysis,  $\Delta t$  is given by the T/P return period of 9.91 days, denoted as 10 days, and as 1 in non-dimensional units. Atmospheric forcing is given on the same spatial and temporal grid as the oceanic data. We can write the forcing field in the same form of multichannel time series  $\mathbf{u}(t) = \{\mathbf{u}(i, j, t)\}$ . The indices  $i, j$  denote a grid point. Note that in the analysis the forcing field  $\mathbf{u}(t)$  is a *vector* for each pair of  $i_0, j_0$ . A simple example of such time

series model without the deterministic input is random walk, or “drunkard”, model:

$$z(t) = z(t - 1) + e(t),$$

where  $e(t)$  is random forcing. This system is not stationary and the drunkard eventually makes arbitrarily large excursions from the initial point, i.e. the variance of  $z(t)$  is infinite. For the modeling of physical processes, unlike the “drunkard” model, it is common to assume that the process is stationary, which means that its first and second moments, the mean and the variance respectively, are finite and are not changing over time. Under these assumptions we substitute time means for the expectation operator  $E(\cdot)$ . An example of a system with deterministic inputs would be the following model:

$$z(t) - \alpha z(t - 1) = u(t) + e(t), \tag{3.1}$$

or in a multichannel case

$$z(t) - \alpha z(t - 1) = u_1(t) - u_2(t) + e_1(t), \tag{3.2}$$

which states that the output is equal to  $\alpha$  times the last value plus the input plus some random shock. This system is stationary provided the input is stationary and  $|\alpha| < 1$ . It is also common to assume that the inputs and the noise are independent. In practice, we do not have sufficient information to establish independence of two stochastic variables, and since for the normally distributed variables independence is identical to zero correlation, and the central limit theorem states that in the limit all distributions tend to normal, we substitute correlation for dependence.

From now on, all time series are assumed to be stationary. First we consider the single input and output case. The input is denoted by  $u(t)$  and output by  $z(t)$ . The time and space intervals are taken to be unity for simplicity. We introduce a back-shift operator  $q$ , which maps  $y(t)$  to the previous value:  $q(y(t)) = y(t - 1)$ . Then the model, conventionally denoted as the ARMAX model, is

$$A(q)z(t) = B(q)u(t - nk) + C(q)e(t), \tag{3.3}$$

where

$$A(q) = \sum_{k=0}^{k=na} a_k q^k, a_0 = 1,$$

$$B(q) = \sum_{k=0}^{k=nb} b_k q^k,$$

$$C(q) = \sum_{k=0}^{k=nc} c_k q^k$$

where  $e(t)$  denotes white noise process. The polynomials  $A$ ,  $B$ , and  $C$  together with orders  $na, nb, nc$  and  $nk$  specify the model, and the model should be chosen on some physical grounds. The order  $na$  denotes number of time lags into the past and it is equal to zero if the output has no memory. For the “drunkard” model it is 1. The order  $nk$  denotes the number of delay periods from input to output. For example, the effect of regulatory laws need to be modeled with delay as it takes time to implement the laws into practice. The order  $nb$  stands for number of time lags of inputs used in the analysis, and when only the present forcing is believed to play a role it is set to zero. The order  $nc$  stands for number of terms in the noise regression on itself, commonly denoted as the moving average part of the model. The name “average” is misleading as the sum of coefficients  $c_k$  need not to be equal to 1, but is well established nonetheless. There are other equivalent forms of the model. For example, the so-called Box-Jenkins model is obtained by dividing through by the polynomial  $A(q)$ . We reason that by including neighboring channels we can introduce a finite difference analog of a partial differential equation.

In this notation the system (3.1) would be modeled as follows:

$$na = 1; nb = 0; nc = 0; nk = 0; A(q) = 1 + a_1 q; B(q) = b_0; C(q) = 1.$$

If the model indeed describes the observations and finite differencing provides a perfect estimate of the derivatives, we would obtain “ideal” values  $a_1 = -\alpha, b_0 = 1$ . However, because the observations are not perfect and finite differencing introduces error we expect estimated values to differ from the “ideal” values. The procedure is then as follows: given observations of inputs and outputs and the model orders  $na = 1, nb = 0, nc = 0$  and  $nk = 0$  we can estimate the matrix polynomials  $A$  and  $B$  at every grid point  $(x_i, y_j)$  by using least squares. The least square estimate provides us with the uncertainties of parameter estimates, as described in Appendix 1. The full treatment is given in Ljung (1987). If the values of the coefficients of  $A$  and  $B$  are within one standard deviation from the “ideal” values we cannot reject the model.

The model can be rewritten in the same form (3.3) for the multivariable case. Then each of the coefficients  $a_1, b_0$ , etc., is a matrix. For the analysis we considered only the case of single output model, i.e.  $z(t)$  is a scalar measurement of SSH. Because our knowledge of

the noise process was very limited, we assumed  $C = 1$ , i.e. dropping the moving average part of the model and reducing the model to a so-called ARX form. However, if we were to believe that the noise is red we could model that easily by setting  $nc = 1$ . In the case of example (3.2), the model becomes:

$$\begin{aligned}
 A(q)z(t) &= B(q)[u_1(t), u_2(t)]^T + C(q)e(t); & (3.4) \\
 na &= [1]; nb = [0, 0]; nc = [0]; nk = [0, 0]; \\
 A_1 &= [a_1]; B_0 = [b_{11}, b_{12}]; C = 1; \\
 A(q) &= 1 + A_1q; B(q) = B_0; C(q) = 1.
 \end{aligned}$$

### 3.2.1 Averaging

The model estimation is done purely locally at every spatial location. This produces a noisy parameter estimate as the data vary significantly from one grid point to the next. We have a choice between 1) averaging parameters estimated for the model on the original grid, or 2) averaging the data and then applying the model on a reduced grid. We refer to this as data averaging and parameter averaging respectively. For parameter averaging we use a maximum likelihood estimate of the mean and its uncertainty, as described in Appendix 1. Data averaging is done by computing integrals over finite areas.

To illustrate the difference between the two methods of averaging we consider the example of a single channel linear regression. We want to estimate an average regression coefficient of  $y_i(t)$  on  $u_i(t)$ , given time series at  $N$  locations. The subscript  $i$  denotes a spatial location. We can compute the regression coefficient at each location using a standard least-squares estimate and then average them, i.e.

$$y_i(t) = \theta_i u_i(t), \quad i = 1, \dots, N, t = 1, \dots, M; \quad (3.5)$$

$$\hat{\theta}_i = \frac{\sum_{t=1}^{t=M} y_i(t)u_i(t)}{\sum_{t=1}^{t=M} u_i(t)u_i(t)} \quad (3.6)$$

$$\hat{\theta} = \frac{1}{N} \sum_{i=1}^{i=N} \hat{\theta}_i = \frac{1}{N} \sum_{i=1}^{i=N} \frac{\sum_{t=1}^{t=M} y_i(t)u_i(t)}{\sum_{t=1}^{t=M} u_i(t)u_i(t)}. \quad (3.7)$$

We denote this procedure as the parameter averaging. For the data averaging we first average the data and then compute the regression coefficient, i.e.

$$y(t) = \theta u(t), \quad y(t) = \frac{1}{N} \sum_{i=1}^{i=N} y_i(t), \quad u(t) = \frac{1}{N} \sum_{i=1}^{i=N} u_i(t); \quad (3.8)$$

$$\hat{\theta} = \frac{\sum_{t=1}^{t=M} y(t)u(t)}{\sum_{t=1}^{t=M} u(t)u(t)} = \frac{\sum_{t=1}^{t=M} \left( \sum_{i=1}^{i=N} y_i(t) \sum_{i=1}^{i=N} u_i(t) \right)}{\sum_{t=1}^{t=M} \left( \sum_{i=1}^{i=N} u_i(t) \sum_{i=1}^{i=N} u_i(t) \right)} \quad (3.9)$$

The two estimates (3.7) and (3.9) are clearly different. The reason is that the estimate for  $\hat{\theta}$  is a linear function of  $y(t)$  but is a non-linear function of  $u(t)$ . When we do data averaging we average *both* the outputs  $y(t)$  and the inputs  $u(t)$ . Analogous analysis can be done for the multichannel regression model and the auto-regressive external model parameter estimates.

In addition, we weight parameters proportional to the percentage of variance explained by the model at that spatial point, thus putting more weight on better fits of the model to the data. The MLE procedure requires assessment of independence in the data from one grid point to another. The independence can be approximated by estimating correlations between output time series in different spatial points. The correlation drops off with increasing distance between two time-series, or channels, and we take the distance at which it on average drops off to 0.3 (95% confidence interval for normally distributed variables) as the measure of independence. The reasons for substituting correlation for the measure of dependence have been explained above. Recent research in time-series analysis has proposed use of “mutual information” between two time series as the measure of independence for data coming from non-linear systems, Abarbanel et al. (1994), as it has a desirable property that zero mutual information implies independence, and zero correlation does not, except for the special case of normally distributed variables. However, since it does not make a significant difference for the present analysis we use the more classical notion of correlation.

First, we address some of the questions raised in Wunsch (1991), which could not have been examined properly with Geosat data. We show that one needs to be careful in the choice of input (forcing) variables. Rejecting the simplest linear model as inadequate, we consider more complicated linear models including higher order derivatives.

### 3.3 Simplified Physical Model

We use the model as described in Chapter 1. In the case of flat bottom and under the assumptions of

$$\sigma \ll f, \text{ i.e. small Rossby number, and } R\sigma/Lf \ll 1, \quad (3.10)$$

the equation for SSH becomes :

$$\beta \frac{\partial \eta}{\partial x} = \frac{f}{\rho g h} \left( \frac{\partial \tau^y}{\partial x} - \frac{\partial \tau^x}{\partial y} \right) + \frac{\beta \tau^x}{\rho g h} + \beta \frac{\partial \zeta_a}{\partial x} \quad (3.11)$$

$\tau_x, \tau_y$  are the wind stress components, and  $\zeta_a = -p_a/\rho g$  is the IB height corresponding to atmospheric pressure. The value for  $\rho$  is taken from climatological data. Since this value changes by less than 2 percent over the global ocean the barotropic model is insensitive to a particular choice. The IB effect was accounted for in the original data but we leave it in nonetheless. It was shown (Fu and Pihos, 1994) that the IB correction was not valid everywhere and thus we would expect a residual regression on  $\zeta_a$  left in the model. We are interested in the wind forcing related variability as well as pressure related variability, and therefore, we want to get rid of any possible residual effect of the SLP on the SSH. For the analysis of barotropic dynamics  $h$  is taken to represent the entire ocean depth, but can be taken as the equivalent ocean depth for analysis of baroclinic dynamics. Of course, all baroclinic and barotropic modes are present in the signal, but some of them have much larger amplitudes than the rest.

If we discretize the equation (3.11) and use the finite difference form for the derivatives we obtain:

$$\frac{\beta}{L} (\eta(i, j, t) - \eta(i-1, j, t)) = \frac{f}{\rho g h L} \hat{\mathbf{k}} \cdot \text{curl } \tau + \frac{\beta}{\rho g h} \tau^x(i, j, t) + \frac{\beta}{L} \frac{\partial \zeta_a}{\partial x}. \quad (3.12)$$

where the curl and  $x$  derivative are differenced as follows:

$$\begin{aligned} \hat{\mathbf{k}} \cdot \text{curl } \tau(i, j, t) &= \frac{1}{2} [\tau^y(i+1, j, t) - \tau^y(i-1, j, t) - \tau^x(i, j+1, t) + \tau^x(i, j-1, t)] \\ \frac{\partial \zeta_a(i, j, t)}{\partial x} &= \frac{1}{2} [\zeta_a(i+1, j, t) - \zeta_a(i-1, j, t)]. \end{aligned} \quad (3.13)$$

Indices  $i, j$  denote  $i^{\text{th}}, j^{\text{th}}$  longitudinal and latitudinal grid points respectively. For the points on the boundary of the grid we use next point differencing instead of central differencing. We now rewrite (3.12) as:

$$\eta(i, j, t) = \frac{L}{\beta} \left( \frac{f}{\rho g h L} \hat{\mathbf{k}} \cdot \text{curl} \tau + \frac{\beta}{\rho g h} \tau^x(i, j, t) + \frac{\beta}{L} \frac{\partial \zeta_a}{\partial x} + \frac{\beta}{L} \eta(i-1, j, t) \right) \quad (3.14)$$

### 3.4 Results

Once we have a physical model in mind we can test it. By doing this for model (3.14) we provide estimates of the coefficients  $f/\rho g h \beta, L/\rho g h$ , etc., from the data. We ran the model for the following choice of inputs and output:

$$\mathbf{u}(t) = \left[ \hat{\mathbf{k}} \cdot \text{curl} \tau(i, j, t), \tau^x(i, j, t), \frac{\partial \zeta_a}{\partial x}(i, j, t), \eta(i-1, j, t) \right]^T ; \quad z(t) = \eta(i, j, t) \quad (3.15)$$

The chosen model was the simplest one, without any time lags:

$$z(t) = \mathbf{b}^T \cdot \mathbf{u}(t) + e(t) \quad (3.16)$$

where  $\mathbf{b} = [b_{\text{curl} \tau}, b_{\tau^x}, b_{\zeta_a}, b_{\eta(i-1)}]^T$  is the vector of parameters. If the IB effect is perfectly valid there would be no correlation between the original SSH and sea level pressure and accordingly the coefficient  $b_{\zeta_a}$  would be equal to 0. This model directly corresponds to (3.14). The analysis was done on normalized (i.e. variance equal to unity) variables. Doing the analysis on normalized variables puts all variables on equal footing and eliminates any dependence on the choice of physical units. We estimated the coefficients  $\mathbf{b}$  at all grid points and then plotted them in longitude - latitude plane. First we establish where the parameters are statistically different from zero. The dimensional units for the corresponding values of physical quantities, e.g.  $h$ , are invoked whenever we believe the results to be sufficiently good.

We present results for the model (3.15 - 3.16) estimated from the original data. The spatial distribution and histogram of explained variance are presented on Figure 14. The bathymetry contours are shown on the background. The model explains on average 28 percent of variability in the output. In about half of the domain, the model explains less than 20 percent of the variability. At about 20 percent of all grid points, the model

explains more than 50 percent of the variability. The estimates of the model parameters are shown on Figure 15. For simplicity, we denoted positive statistically significant values by white, negative statistically significant values by black, and statistically insignificant values by gray. The estimate is said to be statistically significant if it is larger than one standard deviation in absolute value. The estimates of the  $b_{\eta(i-1)}$  coefficient are positive and significant over almost all of the domain. This is consistent with the underlying physical model and is to be expected because time series of SSH at two adjacent grid points are highly correlated. They significantly deviate only in the southwestern Pacific where there is very large topographic variation. The values of the  $b_{curl\tau}$ ,  $b_{\tau^*}$ , and  $b_{\zeta_a}$  are insignificant over most of the domain and where they are significant there are no clear zones of one color. This shows that parameter estimates are very unstable, changing drastically from one point to another. The values of the  $b_{curl\tau}$  are mostly negative in mid-latitudes in the Northern hemisphere and mostly positive in mid-latitudes in the Southern hemisphere, and we denote this as an “anti- $f$ ” effect. This is exactly opposite to what one would expect if this were to represent  $f/\rho gh\beta$  because  $f$  changes sign from positive to negative moving from the Northern hemisphere to the Southern hemisphere. The coefficient  $b_{\tau^*}$  is insignificant over most of the domain, but is otherwise positive as we would anticipate from the physical model. The value for  $b_{\zeta_a}$  is insignificant over most of the domain and shows no preference for either positive or negative values. This indicates that remaining effect of SLP on SSH variability is not important for this model.

We did parameter averaging as described in section 2.1 over  $6^\circ \times 6^\circ$  boxes assuming independence at  $4^\circ$  distance. The results are displayed on Figure 16 in the same format as on Figure 15. Averaging makes the results look smoother but the same conclusions remain valid. Because of the “anti- $f$ ” effect, insignificance of the estimates of  $b_{\tau^*}$  and  $b_{\zeta_a}$ , and low percentage of explained variance we reject model (3.16) as being inadequate for explaining variability in the original data.

We performed an additional analysis for a model corresponding to equation (3.14). We included the two components of the curl, namely the respective partial derivatives of the zonal and meridional wind stress, separately creating a model with a different choice of inputs. The results were nearly the same as described above, that is we rejected the model. It has to be noted that increasing the number of degrees of freedom (number of parameters) in the model we improve the fit. The reason is that we have more parameters

to adjust to accommodate variability of the output. In the very extreme case of a model with the number of degrees of freedom equal to the length of time series we can explain all the variability of the output. But we could do it with *any* set of randomly generated time series and it is a completely useless result. This fact is well understood, and to distinguish which statistical model provides a better description of the data, one needs to account for the number of degrees of freedom as well as for the explained variance. For example, the Akaike Final Prediction Error (FPE) is equal to  $[(1 + n/N)/(1 - n/N) * V]$  instead of  $V$ , where  $V$  is unexplained variance in the output,  $n$  is the number of degrees of freedom,  $N$  is the number of temporal observations (see Ljung (1987) for more details). That is, the error prediction parameter is increased when we increase the number of degrees of freedom. This says that the model leaves a larger proportion of variance unexplained than one would predict by simply taking the value of  $V$ .

The next thing to notice is that there are scales present in the data which are shorter than required by assumption (3.10). For spatial scales longer than 200 km, the condition (3.10) implies that  $T \gg 20$  days. Thus, we applied the same model (3.16) to data low passed at the frequencies corresponding to periods longer than 100 days. It gave a much better fit in terms of the proportion of explained variance as is shown on Figure 17. The mean value for explained variance jumped to 60 percent. We note again that all the analysis is performed on normalized data, that is we normalize the data after the filtering. We may expect that the data had a lot of high frequency variability and the model was unsuccessfully trying to account for it. The estimates of the parameters  $\mathbf{b}$  are shown on Figure 18 in the same format as on Figure 15. Comparing Figures 15 and 18, we notice that the gray color has almost disappeared, i.e. estimated values became significant. They are, however, very non-uniform. This indicates that the model does not describe well the dynamics, even though the time series at each location can be well described by the regression on 4 inputs. It may be the case that we could approximate the data well by a cubic polynomial, with four degrees of freedom, whose coefficients change a lot from one grid point to another. This would not mean that we can describe the dynamics by a cubic dependence as the coefficients have to be constant for that to be true. Note that we included the coefficients which change in space, such as  $f$ , in the definition of the inputs and outputs. On Figure 19 we display parameters averaged in the same way as we did for the original data on Figure 16. We see that  $b_{curl\tau}$  in mid-latitudes is clearly negative in the Northern hemisphere and positive in the Southern hemisphere,

the same "anti- $f$ " effect. The coefficients  $b_{\tau^x}$  and  $b_{\zeta_\alpha}$  are not significant over most of the domain after the averaging. The results for the coefficient  $b_{\eta(i-1)}$  are nearly identical to that for the low-passed data. Thus, we have to reject the model (3.15 - 3.16) based on low-passed data even though the explained variance is high because the coefficients change a lot and some of the estimates are not significant over most of the domain.

It is known that high correlation of one of the inputs with the output can contaminate the results. The time series of  $\eta(i, j, t)$  and  $\eta(i-1, j, t)$  are correlated at 0.6 on average, which is a statistically significant value. We decided to fix the coefficient  $b_{\eta(i-1)}$  to be 1 and estimate only the coefficients  $b_{\text{curl}\tau}$ ,  $b_{\tau^x}$ , and  $b_{\zeta_\alpha}$ . This is equivalent to using

$$\mathbf{u}(t) = \left[ \hat{\mathbf{k}} \cdot \text{curl} \tau(i, j, t), \tau^x(i, j, t), \frac{\partial \zeta_\alpha}{\partial x}(i, j, t) \right]^T ; \quad z(t) = \eta(i, j, t) - \eta(i-1, j, t), \quad (3.17)$$

i.e. taking finite differenced version of  $\frac{\partial \eta}{\partial x}$ , the meridional component of geostrophic velocity, as output, returning to the original (3.12). The disadvantage of doing this is that the meridional component of geostrophic velocity has much higher uncertainty than the SSH and we decrease the signal-to-noise ratio. We repeated the same analysis as before for this model and display the results on Figure 20-25 in the same sequence as on Figures 14-19. The results for the original data are not significant since the percentage of explained variance is less than 10 percent on average, and the estimates of parameters are not significant in most of the domain. The results for the low-passed data are significant over most of the domain but are very non-uniform with no clear regions where the model would be valid. Thus, we must reject this model as well.

We did analysis for the model (3.16) for the original data averaged over  $6^\circ \times 6^\circ$  boxes as described in section 2.1. The variance explained by the model (3.16) is shown on Figure 26. Comparing it to Figure 14 we see that the averaged data provides a much better fit. The estimates for the parameters are shown on Figure 27 in the same format as on Figure 15. Coefficient  $b_{\eta(i-1)}$  has an average value of 0.56. This value is closer to the expected 1 than for the earlier results. Coefficient  $b_{\text{curl}\tau}$  is still largely insignificant and the "anti- $f$ " effect seems to be present. Coefficient  $b_{\tau^x}$  is mostly positive. The values of  $b_{\zeta_\alpha}$  are insignificant and may indicate that inclusion of  $\partial \zeta_\alpha / \partial x$  into the regression is not necessary. Thus, we reject this model for the averaged data.

### 3.5 Results for more complicated models.

Once we realize that the simplest linear vorticity balance does not seem to account for variability in the data, we can try to include more terms into the equation for SSH. Let us begin by including effects of topography, i.e. dropping the assumption of a flat bottom. Thus, with the other assumptions still in place, we obtain the following balance:

$$\begin{aligned} \frac{\partial \eta}{\partial x} \left( \frac{\beta}{f} - \frac{h_y}{h} \right) + \frac{\partial \eta}{\partial y} \left( \frac{h_x}{h} \right) = & \quad (3.18) \\ \frac{1}{\rho g h} \left( \frac{\partial \tau^y}{\partial x} - \frac{\partial \tau^x}{\partial y} \right) + \frac{1}{\rho g h} \frac{\beta}{f} \tau^x + \frac{\partial \zeta_a}{\partial x} \left( \frac{\beta}{f} - \frac{h_y}{h} \right) + \frac{\partial \zeta_a}{\partial y} \left( \frac{h_x}{h} \right) \end{aligned}$$

We took the ocean depth  $h$  as the whole depth of the water column, and thus spatial gradients of  $h$  are given by spatial gradients by bathymetry. We have to make a choice of whether to run the model for two output time series  $\partial \eta / \partial x$  and  $\partial \eta / \partial y$  or a combined time series  $\partial \eta / \partial x (\beta / f - h_y / h) + \partial \eta / \partial y (h_x / h)$ . We decided to do the analysis for a combined time series as it makes the result easier to interpret. After finite differencing for the derivatives, we ran the analysis for this model with four inputs, i.e. one input for each term on the RHS of (3.18). The curl  $\tau$  was taken as a single input. The original data was averaged over a  $10^\circ \times 10^\circ$  boxes on a  $5^\circ \times 5^\circ$  grid as described in section 2.1. In addition we filtered the data to remove variability on the shortest scales of less than 40 days and the longest scales of longer than 200 days. It was shown in Fu and Davidson (1995) that this is the band where we would expect the linear vorticity balance with topography to be applicable. Our experiments with the gridded data produced results worse than the ones presented below. The estimated variance of the model is presented on Figure 28 in the same format as on Figure 14 and shows no significant improvement. The estimated coefficients are still highly inhomogeneous as seen on Figure 29. There are two regions where parameter estimates for coefficient of curl  $\tau$  are positive and significant, the north east and the south east Pacific. However, estimates of all other coefficients in these regions are not significant. In addition, we note that estimates of coefficients on  $\zeta_a$  are negative and statistically significant over most of the domain which clearly contradicts the underlying physical equation. The analysis of IB effect presented in Fu and Pihos (1994) was done without taken into effect topographic variations. Thus, we need to conclude that by including atmospheric loading and topography we obtain statistically significant negative regression of IB corrected SSH on the atmospheric load. However, because

the model fails as a whole we cannot state that this implies that the IB correction was excessive. We analyzed the model without regression on  $\zeta_a$  corresponding to equation

$$\frac{\partial \eta}{\partial x} \left( \frac{\beta}{f} - \frac{h_y}{h} \right) + \frac{\partial \eta}{\partial y} \left( \frac{h_x}{h} \right) = \frac{1}{\rho g h} \left( \left( \frac{\partial \tau^y}{\partial x} - \frac{\partial \tau^x}{\partial y} \right) + \frac{\beta}{f} \tau^x \right) \quad (3.19)$$

including each of the two terms on the RHS as an input. The results show some improvement as seen on Figures 30 and 31. First, there are no statistically significant negative values of coefficient on curl  $\tau$  in two large areas from 170E to 230E and 40N to 60N and 220E to 250E and 60S to 30S. The results for coefficient on  $\tau^x$  are positive and statistically significant over most of north Pacific region. If we look closer at the results of the analysis in the North Pacific region 170E to 230E and 40N to 60N displayed on Figure 32 we see that the explained variance is sufficiently high in this region only in the band from 40° N to 50° N. The estimates of the coefficients are varying in a wide range as seen from Figure 33, but are over 0.3 in non-dimensional units and statistically significant in most of the domain. However, the results of regression on normalized variables provide information only about the correlation of inputs and output. For the model to be acceptable we need variability of the inputs to be on the same scale as variability of the output. The RMS variability of the inputs and output are displayed on Figure 34 and log of the ratio of RMS variability of the inputs to RMS variability of the output are displayed on Figure 35, with values for the log of the ratio below (-5) and above 5 denoted by white color. We see that this model still needs to be rejected except at a few isolated points as the variability of the inputs is not sufficient to explain variability of the output. If we look at the global map of RMS variability of inputs and output and the log of the ratio of RMS variability of inputs to output displayed on Figures 36 and 37 respectively we can see that RMS variability of the inputs is at least a few orders of magnitude lower over all of the Pacific, with the exception of North Pacific and areas of significant topographic slope.

The data spacing does not allow for time scales shorter than 20 days and therefore  $\sigma \ll f$ . However,  $R\sigma/Lf$  is less than 1, but not much less than 1 even for the filtered data we analyze. This assumption can be inappropriate if  $L$  is sufficiently small, i.e. there is significant variability on shorter scales. Then, dropping this assumption we are left with:

$$\frac{\partial}{\partial t} \nabla^2 \eta + \left( \beta - f \frac{h_y}{h} \right) \frac{\partial \eta}{\partial x} + \left( f \frac{h_x}{h} \right) \frac{\partial \eta}{\partial y} = \frac{f}{\rho g h} \left[ \left( \frac{\partial \tau^y}{\partial x} - \frac{\partial \tau^x}{\partial y} \right) + \frac{\beta \tau^x}{f} \right], \quad (3.20)$$

identical to equation (1.6). This is different from the equation (3.18) in the time derivative of the Laplacian term. This is the same equation as the equation (1) in Fu and Davidson, (1995). However, the right hand side is corrected and is consistent with Pedlosky, (1987), equation (5.2.2). Other distinctions include NMC atmospheric data in Fu and Davidson, (1995) instead of the ECMWF data and filtering techniques used here. This provides an independent check on their results based on a different atmospheric data set. We can modify the inputs by neglecting the  $\tau^x$  term and treating the curl of the wind stress as a single time series. The results are presented on Figures 38 and 39. They confirm the previous conclusion that the balance seems to hold only in the north east Pacific and the south east Pacific, if at all. The map of the ratio of the input and the output RMS variability is shown on Figure 40 for comparison with Fu and Davidson, (1995), Figure 2. The estimates of Fu and Davidson show larger variability of the wind stress curl, apparently because of the error on the RHS of their equation (1). If we now add back the  $\tau^x$  term, we explain a higher proportion of the variance, Figure 41, but we change estimates of the curl coefficients only at a few points, for example, at  $255^\circ$  E,  $55^\circ$  S, as seen on Figure 42. The map of the RMS variability is shown on Figure 43. The map of log of ratios of RMS variability is shown on Figure 44. It shows that the RMS variability of the  $\tau^x$  is not large enough to account for the difference in input and output variability. Overall, the results displayed on Figures 41-43 show that the equation (3.20) needs to be rejected as the estimates are not consistent and the variability of the forcing is inadequate to explain variability of the output. Inclusion of the “IB” height does not change the results in any significant respect. Again, we have to reject the model.

So far we only used linear regression models, which are a subset of ARX models. We can use the ARX modeling setup to use time auto-regressive structure as well. This leads to estimating the coefficients of a following model:

$$\nabla^2\eta(i, j, t) + a_1\nabla^2\eta(i, j, t - 1) = \mathbf{b}^T \left[ - \left( \left( \beta - f \frac{h_y}{h} \right) \frac{\partial\eta}{\partial x} + f \frac{h_x}{h} \frac{\partial\eta}{\partial y} \right), \frac{f}{\rho gh} \left( \frac{\partial\tau^y}{\partial x} - \frac{\partial\tau^x}{\partial y} \right), \frac{\beta}{\rho gh} \tau^x \right]; \mathbf{b} = [b_1, b_2, b_3]^T. \quad (3.21)$$

The results for this model are shown on Figures 45-47. The explained variance is displayed on Figure 45 and shows that we explain on average only 22 percent of variability with this 4 degrees of freedom model. The estimates for  $\mathbf{b}$  are shown on Figure 46 and we immediately note that there is a large area in south Pacific where the estimates for  $b_2$  and

$b_3$  are positive statistically significant. Everywhere else we reject the model. However, the estimates of  $b_1$  are not positive statistically significant in this area. The estimate of  $a_1$  coefficient is shown on Figure 47 and we see that the coefficient is negative statistically significant as is expected because output time series are highly correlated at 10 days. If we look closer at the South Pacific region  $190^\circ$  E to  $240^\circ$  E and  $60^\circ$  S to  $20^\circ$  S we see that though the estimates are positive and significant the explained variance of the model still varies widely within the region, Figure 48. Thus, although this model can provide good explanation of variability at some isolated points it fails to provide it on a larger scale.

### 3.6 Conclusions.

This work shows that we have to reject linear homogeneous shallow water models; they are inadequate to describe SSH variability in the Pacific. Some models are able to explain a large proportion of SSH variability in small isolated regions, such as the north east Pacific and south east Pacific, but none can do it globally. The results are negative but consistent with the findings of other researchers, e.g. Fu and Davidson (1995). There can be many reasons why these models fail. First, one may think that inclusion of other terms in the original equation (1.4) can provide for a better agreement. We ran several experiments with more inputs and did not observe any significant improvement apart from that due to increased number of degrees of freedom. Second, one would look for “missing” physics, such as effects of stratification, non-linearity, heat fluxes, etc. Some of these can be modeled in this framework, and some cannot. For example, stratification could be modeled by using an expansion into barotropic and baroclinic modes, i.e. inclusion of several baroclinic modes into the model. Modeling of a system of equations can be easily done using vector ARX models, but one needs to specify the correlation structure for the outputs a priori. Non-linearity is difficult to handle in the framework of regression analysis. Overall, it seems that the methods presented in this chapter can give a clear answer on the question of consistency of the models with the data if one believes that the errors small. But since more complicated models require higher order derivatives the finite differencing error can ruin the analysis. The approach of the next section seem to be better suited for such models, at the expense of greatly increasing requirements for

computer time and storage and much more detailed analysis of the data and model error structure.

## Chapter 4

### State-space approach.

#### 4.1 Overview of the method.

At this point we turn our attention to state space modeling. The column vectors are written as bold lower case characters, and matrices as upper case characters. We need to rewrite our dynamical equation in the following way:

$$\mathbf{q}(t+1) = A(t)\mathbf{q}(t) + B(t)\mathbf{w}(t) + \Gamma(t)\mathbf{u}(t), \quad t = 1, \dots, T, \quad (4.1)$$

where  $\mathbf{q}$  denotes state space vector,  $\mathbf{w}$  denotes the known forcing, and  $\mathbf{u}$  denotes the unknown part of the forcing, or *controls*.  $A$  is the “state transition matrix”.  $B$  and  $\Gamma$  map the forcing onto the state. The time unit is taken to be one for simplicity. The state space vector  $\mathbf{q}$  includes all physical quantities necessary to describe the system at time  $t$ . Forcing includes boundary conditions, and everything else externally prescribed to the model. The reason for including the controls is that one cannot expect to have perfect knowledge of the forcing. For example, the curl of the wind stress, the forcing in model 1.6, is itself a product of noisy measurements.

It is a well known fact that any time dependent model can be rewritten in this form by a proper choice of the state vector  $\mathbf{q}$ . For example, the equation

$$\frac{D^2\eta}{Dt^2} = f(t)$$

can be rewritten in the following way:

$$\mathbf{q}(t+1) = A\mathbf{q}(t) + \mathbf{f}(t),$$

$$\text{where } \mathbf{q}(t) = \begin{pmatrix} \eta(t) \\ \eta(t-1) \end{pmatrix}; \mathbf{f}(t) = \begin{pmatrix} f(t) \\ 0 \end{pmatrix}; A = \begin{pmatrix} 2 & -1 \\ 1 & 0 \end{pmatrix}.$$

The important difference with the conventional mathematical physics formulation is that we need to provide estimates of the uncertainties of the state variables. The controls  $\mathbf{u}$  are assumed to have zero mean and covariance  $Q$ :

$$\langle \mathbf{u}(t) \rangle = 0; \langle \mathbf{u}(t)(\mathbf{u}(t))^T \rangle = Q(t). \quad (4.2)$$

One of the main difficulties of the method is that one *rarely* has a reliable estimate of the covariance  $Q$ . It includes uncertainties of the forcing and any physics which the model fails to describe accurately.

We complete the description of the model by providing initial conditions,  $\tilde{\mathbf{q}}(\mathbf{0}) = \mathbf{q}_0$  with the corresponding uncertainty covariance

$$P(0) = \langle (\mathbf{q}(0) - \tilde{\mathbf{q}}(0))(\mathbf{q}(0) - \tilde{\mathbf{q}}(0))^T \rangle, \quad (4.3)$$

where tilde denotes the hypothetical *true* value. In the same way we define the covariance at any time step  $t$

$$P(t) = \langle (\mathbf{q}(t) - \tilde{\mathbf{q}}(t))(\mathbf{q}(t) - \tilde{\mathbf{q}}(t))^T \rangle. \quad (4.4)$$

We rarely have good estimates of  $P(0)$ . However, it is often true that results are insensitive to the choice of  $P(0)$  and very large values are chosen. The mean value of the diagonal of  $P(t)$ , i.e. the mean variance of the model simulation, rapidly decreases with time. Now, we have a full description of the model and we can integrate the initial conditions forward. However, one can expect that because of the unknown controls the uncertainty of the estimate  $P(t)$  will grow with time and very soon the estimate will become completely useless. We rescue the situation by assimilating data when it becomes available. This prevents the uncertainty of the estimate from growing with time.

To complete the system we relate observations  $\mathbf{h}$  to the state space vector  $\mathbf{q}$ :

$$\mathbf{h}(t) = E(t)\mathbf{q}(t) + \mathbf{r}(t), \quad (4.5)$$

where  $\mathbf{r}$  stands for the observational noise and any physics which is present in the data but not in the model.  $E$  is the ‘‘observation matrix’’. The noise is assumed to have zero

mean, a *known* covariance matrix  $R(t)$ , and be white in time:

$$\langle \mathbf{r}(t) \rangle = 0; \langle \mathbf{r}(t)\mathbf{r}(t')^T \rangle = R(t)\delta_{t,t'}. \quad (4.6)$$

For the sake of completeness we present the algorithm for the Kalman filter and optimal smoother. We use the Rauch-Tung-Sriebel (1965) smoother algorithm which allows one to start from the final estimate made by the Kalman filter and run the smoothing backward in time. After completing the smoother we obtain an optimal estimate at every time step which uses all available information. Full treatment is available in the literature, e.g. Wunsch (1996), Anderson and Moore (1979).

#### 4.1.1 Kalman filter and RTS smoother.

We start with the Kalman filter which takes the initial state to the final time  $t_f$ . The smoother then takes the final assimilated state back to time 0 producing optimal in the least-squares sense estimate of the state using all available information.

One of the many equivalent Kalman filter formulations taking the state from time  $t$  to time  $t + 1$  is

$$\mathbf{q}(t + 1, -) = A\mathbf{q}(t) + B\mathbf{w}(t) \quad (4.7)$$

$$P(t + 1, -) = AP(A)^T + \Gamma Q(\Gamma)^T \quad (4.8)$$

$$K(t + 1) = P(t + 1, -)E^T(EP(t + 1, -)E^T + R)^{-1} \quad (4.9)$$

$$\mathbf{q}(t + 1) = \mathbf{q}(t + 1, -) + K(t + 1)(\mathbf{h}(t + 1) - E\mathbf{q}(t + 1, -)) \quad (4.10)$$

$$P(t + 1) = P(t + 1, -) - K(t + 1)EP(t + 1, -) \quad (4.11)$$

We will use the assumption that  $A, B, \Gamma, E, R$  and  $Q$  are time independent from now on. The RTS smoother algorithm takes the Kalman filter estimates at time  $t + 1$  and  $t$  and produces an optimal estimate at time  $t$ :

$$\mathbf{q}(t, +) = \mathbf{q}(t) + L(t + 1)(\mathbf{q}(t + 1, +) - \mathbf{q}(t + 1, -)) \quad (4.12)$$

$$L(t + 1) = P(t)(A)^T(P(t + 1, -))^{-1} \quad (4.13)$$

$$P(t, +) = P(t) + L(t + 1)(P(t + 1, +) - P(t + 1, -))(L(t + 1))^T \quad (4.14)$$

$$\mathbf{u}(t, +) = \mathbf{u}(t) + M(t+1)(\mathbf{q}(t+1, +) - \mathbf{q}(t+1, -)) \quad (4.15)$$

$$M(t+1) = W(t)A^T(P(t+1, -))^{-1} \quad (4.16)$$

$$Q(t, +) = Q(t) + M(t+1)(P(t+1, +) - P(t+1, -))(M(t+1))^T \quad (4.17)$$

$$\mathbf{q}(t_f, +) = \mathbf{q}(t_f); P(t_f, +) = P(t_f); \quad (4.18)$$

We use a convenient notation where “-” signs represents estimates after the model simulation, or forecast, lack of a sign represents estimates after the Kalman filter assimilation, and “+” sign represents estimates after the optimal smoothing.

To start the analysis we need to specify: the model and data transformation matrices  $A, B, \Gamma$ , and  $E$ ; error covariance matrices  $P(0), Q$ , and  $R$ ; initial conditions  $\mathbf{q}(0)$  and the forcing  $\mathbf{f}(t)$ . We then compute a model forecast  $\mathbf{q}(1, -)$  by using the known forcing  $\mathbf{f}(0)$  through (4.7). Next, we compute error covariance  $P(1, -)$  of this estimate through (4.8). Now we can compute the Kalman gain  $K(1)$ , the assimilated state  $\mathbf{q}(1)$  and its error covariance  $P(1)$  by using (4.9)-(4.11) respectively. We then start with  $\mathbf{q}(1)$  and compute  $\mathbf{q}(2, -)$  making use of  $\mathbf{f}(1)$  and so on. We run this algorithm until we produce  $\mathbf{q}(t_f)$  and its error covariance  $P(t_f)$ . At this point we are finished with the Kalman filter estimation. We then proceed backwards in time with the smoother. The smoother is initialized through (4.18) and then ran backward until we compute  $\mathbf{q}(0, +)$ . If we do not need to produce an estimate of the unknown controls, the smoother algorithm reduces to only three steps (4.12)-(4.14).

To recapitulate, the problem is to rewrite our model and data as defined in Chapters 1 and 2 into the framework defined through equations (4.1 - 4.6) and then apply the Kalman filter and RTS smoother.

## 4.2 Implementation of the model.

We take the time dependent linear vorticity balance (1.6) as the dynamical model:

$$\begin{aligned} \frac{\partial}{\partial t} \nabla^2 \eta + hJ(\eta, \frac{f + \beta y}{h}) &= w(t) \\ w(t) &= \frac{f}{\rho g h} \left[ \left( \frac{\partial \tau^y}{\partial x} - \frac{\partial \tau^x}{\partial y} \right) + \frac{\beta \tau^x}{f} \right], 0 \leq x \leq L, 0 \leq y \leq L_1 \end{aligned} \quad (4.19)$$

subject to boundary conditions  $\eta(t) = \eta_B(t)$  on the boundary  $\mathcal{G}$ . Topography is denoted by small  $h$ . Introducing non-dimensional variables

$$(x, y) = L(x', y'), t = Tt', \eta = N_0\eta', (\tau^x, \tau^y) = \tau((\tau^x)', (\tau^y)') \quad (4.20)$$

and conventionally dropping the primes, we rewrite the equation as:

$$\frac{\partial \nabla^2 \eta}{\partial t} + \omega h J(\eta, \frac{1 + \beta_0 y}{h}) = \frac{\omega \tau L}{\rho g h N_0} \left( \frac{\partial \tau^y}{\partial x} - \frac{\partial \tau^x}{\partial y} + \beta_0 \tau^x \right) \equiv w(t) \quad (4.21)$$

where  $\omega = fT, \beta_0 = \beta L/f$ . We assume that the model is given on a regular rectangular grid

$$x_i = (i - 1)/(N - 1), i = 1, \dots, N; y_j = L_1(j - 1)/(L(M - 1)), j = 1, \dots, M - 1.$$

Next, we introduce the relative vorticity  $\zeta(t) = \Delta \eta(t) \equiv \nabla^2 \eta(t)$  and the potential vorticity  $\zeta_0 = (\beta_0 + y)/h$ . The Laplacian of the SSH  $\nabla^2 \eta$  has been neglected in the definition of the potential vorticity since we linearized the model. We can solve the equation (4.21) as a five-step scheme for the relative vorticity.

- Start with the initial condition for the relative vorticity  $\zeta(0)$ .
- Solve  $\Delta \eta_P(0) = \zeta(0)$  with b.c.  $\eta(0) = 0$  on  $\mathcal{G}$ , the Poisson problem:

$$\eta_P(0) = \Delta_P^{-1} \zeta(0). \quad (4.22)$$

- Solve  $\Delta \eta_L(0) = 0$  with b.c.  $\eta(0) = \eta_B$  on  $\mathcal{G}$ , the Laplace problem:

$$\eta_L(0) = \Delta_L^{-1} \eta_B. \quad (4.23)$$

- Sum them up  $\eta(0) = \eta_P(0) + \eta_L(0)$ .
- Step forward  $\partial \zeta / \partial t = -\omega h J(\eta(0), \zeta_0) + w(t)$

We then continue with this process forward in time. However, this scheme can be rewritten as a one step dynamic equation for the relative vorticity by finite differencing the time derivative:

$$\zeta(t + \delta t) = \zeta(t) + \delta t \left[ -\omega \cdot h \cdot J(\Delta_P^{-1} \zeta(t), \zeta_0) \right] + F(t), \quad (4.24)$$

$$\text{where } F(t) = \delta t \left[ -\omega \cdot h \cdot J(\Delta_L^{-1} \eta_B(t), \zeta_0) + w(t) \right] \quad (4.25)$$

where  $h, \eta_B(t), \zeta$ , and  $w(t)$  are assumed known. The forcing in the latter equation (4.25) can be interpreted as the sum of the source of potential vorticity (PV) due to advection of PV through the domain boundary and the source of PV due to the wind.

We define the state vector  $\mathbf{q}(t)$  as

$$\mathbf{q}(t) = [\eta(x_1, y_1, t), \eta(x_2, y_1, t), \dots, \eta(x_N, y_M, t)]^T. \quad (4.26)$$

The relative vorticity  $\zeta(t)$  is given by the Laplacian of the state vector  $\mathbf{q}$ :  $\zeta(t) = \Delta \mathbf{q}(t)$ , where  $\Delta$  is discretized using centered difference scheme. The Jacobian is also discretized using centered differences. We employ a direct solution method by LU decomposition for the Laplace and Poisson problems. Since the potential vorticity  $\zeta_0$  is known a priori, we can rewrite the advection term as a linear operator:  $J \equiv J(\cdot, \zeta_0)$ . The actual algorithm for the model dynamics consists of three steps:

- Start by converting the initial condition for  $\eta(t - 1)$  into relative vorticity  $\zeta(t - 1)$ ,
- Time stepping for  $\zeta(t - 1)$  with the small increment  $\delta t$  until we obtain the desired  $\zeta(t)$ ,
- Converting the relative vorticity  $\zeta(t)$  back into the SSH  $\eta(t)$ .

Because we have chosen SSH as the state vector, the observation matrix which relates the data to the state vector is simply the identity matrix, i.e.  $E(t) = I$ . The observations, given by the filtered and averaged data, are organized into a vector  $\mathbf{h}(t)$ . It is ordered in exactly the same way as  $\mathbf{q}(t)$ .

### 4.3 Literature review.

Before we start discussing the results we present a short review of previous publications which provide the necessary background to the problem. The optimal estimation approach has been used in a number of investigations since altimetric measurements providing global coverage have become available. Gaspar and Wunsch (1989) presented an analysis of Geosat data in the very energetic part of the Northwestern Atlantic based on

free linear barotropic Rossby wave equation

$$\frac{\partial}{\partial t} \nabla^2 \eta + \beta \frac{\partial \eta}{\partial x} = 0. \quad (4.27)$$

The paper contains an excellent discussion of the estimation procedure and sets the framework for assimilating the oceanic data with the Kalman filter and optimal smoother. The authors solved equation (4.27) by using an expansion in terms of sinusoidal basis functions, so that the state transition matrix  $A$  could be written out explicitly. Thus, the algorithm (4.7)-(4.18) could be directly applied to the system without concern for numerical instabilities. On average the model forecast was consistent with only a small fraction of the variability in the data (6 % to 15 % depending on the type of estimator used). The model was very simple and did not include any atmospheric forcing, so that one should not expect a large portion of the variance to be consistent with the model. It is important to realize that the results only tell whether the model is *consistent* with the data, and depend on the choice of model process noise and measurement noise (error) which are often only approximately known.

Fu et al. (1991) concentrated on the analysis of forced wave propagation in the tropical Pacific based on Geosat data. The tropical ocean is known to be well described by equatorial Kelvin and Rossby wave dynamics. The authors used a free sinusoidal wave model allowing for time variation of the amplitude and phase, and included barotropic as well as first and second baroclinic modes. For the optimal value of the process error covariance, the model accounted for 23 % of the signal variance. The authors were able to qualitatively describe the most prominent features in the data. They reported that introducing the smoother only marginally improved the solution performance. They postponed any discussion of sensitivity of the results to the choice of the error covariance to their second paper.

In the continuation of this paper Fu et al. (1993) described a linear wind-driven model of the tropical Pacific. This model was taken from Miller and Cane (1989). Overall, the approach was very similar to the first paper. However, in this study the authors used a steady state filter which makes the computations much more efficient. It eliminates the need to propagate the error-covariance matrix  $P(t)$  (the most computationally expensive part of the algorithm), and the need to store the error-covariance matrix  $P(t)$  at every time step. For this more realistic model more than 60 % of the variance could be accounted

for by the model – a remarkable achievement. However, the results were sensitive to the choice of model and observation error-covariance matrices, pointing to the need for careful estimates of data and model errors.

In a very recent paper Fukumori (1995) described the use of approximate Kalman filter and optimal smoother to a reduced-gravity shallow water model in the Tropical Pacific. Nearly 80 percent of the expected signal variance was accounted for by the model. The analysis of the results involved a comparison with the data from the TAO current meter measurements, which provided an independent check on the estimation. The smoothed estimates were the most consistent with the TAO data. This confirms usefulness of the Kalman filter and the optimal smoother estimation procedure, because this procedure provides a correction for both the model and the data.

#### 4.4 Model and data error estimates.

We use an approach similar to that of Fu et al. (1993) to make estimates of process and measurement noise structures. The measurement noise is the sum of the measurement errors, e.g. orbital errors, tidal errors, etc., and the oceanic signal not associated with the dynamical model at hand. The measurement error covariance can be estimated by comparing observations  $\mathbf{h}$  to model simulations  $\mathbf{q}_m$  *without data assimilation*. We can separate both model state and data into the true signal and their respective errors:

$$\mathbf{h}(t) = \tilde{\mathbf{h}}(t) + \mathbf{r}(t), \quad (4.28)$$

$$\mathbf{q}_m(t) = \tilde{\mathbf{q}}(t) + \mathbf{p}(t). \quad (4.29)$$

Both  $\mathbf{r}$  and  $\mathbf{p}$  are assumed to be stochastic with zero mean and covariance  $R$  and  $P$  respectively. Furthermore, the errors are assumed to be independent of each other and of the true state. We can then compute covariance for the data and the model by using (4.5).:

$$\langle \mathbf{h}'\mathbf{h}'^T \rangle = \langle \tilde{\mathbf{h}}'\tilde{\mathbf{h}}'^T \rangle + R, \quad (4.30)$$

$$\langle E\mathbf{q}'_m(E\mathbf{q}'_m)^T \rangle = \langle E\tilde{\mathbf{q}}(E\tilde{\mathbf{q}})^T \rangle + EPE^T = \langle \tilde{\mathbf{h}}'\tilde{\mathbf{h}}'^T \rangle + EPE^T, \quad (4.31)$$

where primes denote that the mean has been subtracted. Using the assumption that the data and the model errors are independent we obtain:

$$\langle (\mathbf{h}' - E\mathbf{q}'_{\mathbf{m}})(\mathbf{h}' - E\mathbf{q}'_{\mathbf{m}})^T \rangle = EPE^T + R. \quad (4.32)$$

Then, combining (4.30), (4.31), and (4.32), we obtain

$$R = \frac{1}{2} \left( \langle (\mathbf{h}' - E\mathbf{q}'_{\mathbf{m}})(\mathbf{h}' - E\mathbf{q}'_{\mathbf{m}})^T \rangle - \langle E\mathbf{q}'_{\mathbf{m}}(E\mathbf{q}'_{\mathbf{m}})^T \rangle + \langle \mathbf{h}'\mathbf{h}'^T \rangle \right) \quad (4.33)$$

$$EPE^T = \frac{1}{2} \left( \langle (\mathbf{h}' - E\mathbf{q}'_{\mathbf{m}})(\mathbf{h}' - E\mathbf{q}'_{\mathbf{m}})^T \rangle + \langle E\mathbf{q}'_{\mathbf{m}}(E\mathbf{q}'_{\mathbf{m}})^T \rangle - \langle \mathbf{h}'\mathbf{h}'^T \rangle \right) \quad (4.34)$$

This provides a constraint on the process error covariance  $Q$  through equation (4.8).

We performed analysis of the errors by using data at each T/P cycle as initial conditions. We then propagated the signal for 10 days and compared the results with the data at the next cycle, by analogy with the Kalman filter time stepping from  $t = 0$  to  $t = 1$ . The time averaged data and simulation errors are 1.3 and 1.4  $\text{cm}^2$ , whereas the averages of the data and simulation variance are 2.6 and 2.7  $\text{cm}^2$ , respectively. Figure 49 shows mean diagonal of the respective data and simulation error covariance as function of time. We obtain estimate of  $R$  and  $EPE^T$  by substituting the bracketed ensemble averages with time averages over 82 T/P cycles. The contours of the resulting covariance are shown on Figure 50. A few resulting negative values for the variance are due to the fact that the estimators are not necessarily positive definite, the averaging is done over very small samples, and the assumption of independence may be incorrect. However, since we are not using this information in the actual implementation of the Kalman filter but rather as a guideline it does not create any problems.

After running the experiments we chose Gaussian spatial covariance for both  $R$  and  $P(0)$ .

$$R_{i,j} = \langle n(x_1, y_1, t), n(x_2, y_2, t) \rangle = \sigma_M^2 \delta_{i,j} \exp \left[ -(x_1 - x_2)^2 / L_x^2 - (y_1 - y_2)^2 / L_y^2 \right]$$

and where  $i$  is the index corresponding to  $(x_1, y_1)$  and  $j$  is the index corresponding to  $(x_2, y_2)$ , and  $L_x$  and  $L_y$  represent zonal and meridional scales of the error, respectively. The same formula is used for  $P(0)$  with  $\sigma_M$  replaced by  $\sigma_P$ . We used  $L_x = 200\text{km}$ ,  $L_y = 200\text{km}$ ,  $\sigma_M = 2\text{cm}^2$ , and  $\sigma_P = 4\text{cm}^2$ .

The process covariance is modeled in the form of a stationary wind error, with Gaussian spatial covariance among the forcing components. Again, we assume that the process noise is uncorrelated in time. That is,

$$\langle u_i(x_1, y_1), u_j(x_2, y_2) \rangle = \sigma^2 \delta_{i,j} \exp \left[ -(x_1 - x_2)^2 / D_x^2 - (y_1 - y_2)^2 / D_y^2 \right] \quad (4.35)$$

where subscripts denote  $i$  and  $j$  denote the time, and  $D_x$  and  $D_y$  represent zonal and meridional scales of the process error, respectively. This defines  $Q$ , see (4.2). We used  $D_x = 1000\text{km}$ ,  $D_y = 400\text{km}$ ,  $\sigma = 2\text{cm}^2$ .

To summarize, we have made several simplifying assumptions about  $Q$  and  $R$ . We assumed, as in most practical applications, that the two error covariances are uncorrelated in time and between themselves. We have also used a simple Gaussian structure rather than the more complicated form obtained from equations (4.33) - (4.34). The process noise was assumed to be solely due to wind stress error, although we chose an estimate which should be an overestimate of the wind error. This should include effects of other noise processes. Given the simplicity of the model we do not think that more accurate estimates of the error covariances are worth the effort.

## 4.5 Results.

At this point we present the results of the analysis. We chose a “quiet” region in the Northeastern Pacific from  $180^\circ$  E to  $224^\circ$  E and from  $30^\circ$  N to  $50^\circ$  N. The time series of the data after the removal of the annual and semi-annual cycles at  $200^\circ$  E,  $40^\circ$  N is shown onn Figure 3. We ran the experiments for the original averaged data, as defined in Chapter 2.

We ran the experiments with linear interpolation in time over 10 day intervals for both wind forcing and boundary conditions. We integrate the model as described above with the step size of 0.5 days over a 10 day interval. We used a 0.5 days time step for numerical stability. The units for the Kalman filter are such that 1 corresponds to 1 T/P cycle, or 10 days. We take the initial conditions to be equal to T/P observations at the 2nd T/P cycle, since the first T/P cycle provided bad data. We use data from 82 T/P cycles, or 812 days.

The initial state of the system,  $\mathbf{q}_0$ , and the data 10 days later are shown on Figure 51. Both fields are rather complicated. Although both display some vortex like behavior there are more differences than similarities between the two. The correlation coefficient between the two fields is equal to 0.06, and is not significant. It is not evident that a simple model like the one we are using is going to describe this complicated dependence.

The potential vorticity,  $\zeta_0$ , due to the beta effect and the topography, is shown on Figure 52. We observe that it represents a significant perturbation to the flat bottom case where lines of constant PV coincide with lines of constant  $y$ . Note the closed PV contours in the western part of the region. We know that in the absence of forcing the steady model would drive the flow along the lines of constant potential vorticity. Due to the presence of wind forcing and time dependence in our application this is no longer true.

The comparison of forecast, data, and assimilation fields after the 25th T/P cycle, i.e. 240 days of applying the Kalman filter estimation, is shown on Figure 53. We can see that the model forecast  $\mathbf{q}(25, -)$  is quite different from the observations, and that the Kalman filter assimilation step brings the estimated field  $\mathbf{q}(25)$  closer to the observations. At some other points in time, the forecast and the observations are much closer. However, it is not feasible to judge the quality of the model forecast by looking at the longitude-latitude plots at every time step.

Following the procedure outlined in Gaspar and Wunsch (1989), we evaluate these quantities at each time step:

$$E_h(t) = (\mathbf{h}'(t))^T(\mathbf{h}'(t))/(N - 1) \quad (4.36)$$

$$\epsilon_f(t) = (\mathbf{h}'(t) - E\mathbf{q}'(t, -))^T(\mathbf{h}'(t) - E\mathbf{q}'(t, -))/(N - 1) \quad (4.37)$$

$$\epsilon_a(t) = (\mathbf{h}'(t) - E\mathbf{q}'(t))^T(\mathbf{h}'(t) - E\mathbf{q}'(t))/(N - 1) \quad (4.38)$$

$$\epsilon_s(t) = (\mathbf{h}'(t) - E\mathbf{q}'(t, +))^T(\mathbf{h}'(t) - E\mathbf{q}'(t, +))/(N - 1) \quad (4.39)$$

$$\epsilon_i(t) = (\mathbf{h}'(t) - E\mathbf{q}'(t - 1))^T(\mathbf{h}'(t) - E\mathbf{q}'(t - 1))/(N - 1). \quad (4.40)$$

where  $N$  stands for the length of the state vector. The first quantity  $E_h$  stands for the variance of the data, the second  $\epsilon_f$  for the variance of the misfit between the data and the forecast, the third  $\epsilon_a$  for the variance of the misfit between the data and the Kalman filter assimilation, the fourth  $\epsilon_s$  for the variance of the misfit between the data and the

smoothed estimate, and the last  $\epsilon_i$  for the variance of the misfit between the data and the initial condition for the forecast. We remind the reader that the Kalman filter estimate serves as an initial condition for the next step, see equation (4.7). The last quantity allows us to compare results with the purely persistent model without forcing, i.e.  $\mathbf{q}(t+1) = \mathbf{q}(t)$ . We then define

$$\begin{aligned}
 E_f(t) &= E_h(t) - \epsilon_f(t); \\
 E_a(t) &= E_h(t) - \epsilon_a(t); \\
 E_s(t) &= E_h(t) - \epsilon_s(t); \\
 E_d(t) &= E_a(t) - E_f(t); \\
 E_i(t) &= E_h(t) - \epsilon_i(t).
 \end{aligned}
 \tag{4.41}$$

At each time step these quantities show how much of the variance has been explained by:  $E_f$  – the model forecast,  $E_a$  – the Kalman filter assimilation, and  $E_s$  – the smoother estimate.  $E_d$  shows how much of the variance is explained by the model update.  $E_i$  shows how much variance would be explained if the model would have no dynamics at all. If the model and the data were in perfect agreement the first value  $E_f$  would be equal to the model variance  $E_h$ . However, because both the model and the data contain errors it is never equal to  $E_h$ . Since the Kalman filter brings the model forecast into a better agreement with the data we are guaranteed that  $E_f \leq E_a \leq E_s \leq E_h$ .

These quantities are plotted as a function of time on Figure (54). The mean values for each of the time series are shown in the box inside the plot. On average the model forecast explains 33 percent of the data variability. The model is doing much better than a pure persistence model. The Kalman filter assimilated field is much closer to the observations than is the forecast. On average it explains 77 percent of the variability in the data. Thus, the update is very efficient and the mean model update  $E_d$  is equal to 44 percent of the data variance. The smoother estimates are very similar to the Kalman filter assimilation, even though they assimilate additional information. This is similar to the results obtained by Fu et al. (1993), who also noted that the smoother results were only slightly different from the Kalman filter results. This is consistent with the fact the forecast is very poor, and thus assimilation of new information about the observations does not produce any noticeable changes. One of the main features of the Kalman filter and optimal smoother is that it provides the uncertainties of the estimates. Figure 55 displays

evolution of the mean state error variance as a function on time for the model forecast, the Kalman filter update and the smoother estimate. The model forecast has the largest uncertainty at all times because in the Kalman filter assimilated and smoothed estimates we use additional information and therefore reduce the uncertainty. The uncertainty of the forecast and the assimilated state are reduced with time until they reach a steady limit after about 50 days of assimilation. The steady filter is used after that to reduce the computational time and storage requirements. A control run was performed and it showed that employing a steady filter does not change the results in any significant manner, in agreement with other authors, e.g. Fu et al. (1993), Fukumori (1995), etc. Therefore, the lines are constant after the 6th T/P cycle. The smoother is reducing the error at its initial step from  $t_f$  to  $t_f - 1$  and is then steady because of the steady Kalman filter counterpart. It starts to change again at the 6th cycle and the mean error increases toward the initial time  $t_0 = 2\text{nd T/P cycle}$ .

The results of the assimilation at a particular location are shown on Figure 56. They show that the model forecast drives the SSH away from the data. Recall that the rms time variability of the observations used in the analysis is only 2.5 cm, and the disagreement larger than 2cm is significant. The Kalman filter assimilation brings the forecast into agreement with the data at every assimilation time step. At some points, e.g. at the 25th cycle, the model carries the signal away from the data, but then successive assimilation brings it again close to the data. For comparison we make a similar plot at the center of the domain, where we expect the best forecast, for the central East Pacific region. It is shown on Figure 57. The results are very different. Now the forecast makes very big excursions away from the observations. Only because we continuously assimilate the data the model simulation stays within an acceptable range.

We produced time-longitude plots of the observations, the model forecast, the Kalman filter assimilation, and the smoothed estimate. We display the results along  $40^\circ$  N latitude on Figure 58. Due to the lack of strong wave signal these plots are not nearly as informative as they are for the tropical Pacific. All the fields are given on the same grid. There is a qualitative agreement between the forecast and the observed fields because we assimilate the data every 10 days. Even so, there are significant differences, e.g. strong negative anomaly near the 10th cycle in the western part of the domain in the forecast field. This misfit is much less pronounced in the Kalman filter and the optimal smoother

estimates. Overall, it seems that the best way to judge the quality of the assimilation in the absence of additional data sets is by doing the analysis of the variance described above. Still, we conclude that the model is not consistent with the data.

We ran similar analysis in other parts of the Pacific and we present the results for the south east Pacific from  $220^{\circ}$  E to  $270^{\circ}$  E and from  $50^{\circ}$  S to  $30^{\circ}$  S. We chose this region according to the results of the multichannel regression analysis of Chapter 3 it . These results indicated that together with the northeastern region the variability in the southeastern region can be described by the barotropic dynamics. The model produces very poor estimate of the data in this region. We present analysis of the variance explained by the assimilation on Figure 59 in the manner similar to that shown on Figure 54. The forecast explains less variability than even the “pure persistence model”. On average the variance of the misfit between the forecast and the data is larger than the variance of the data itself. The successive use of the Kalman filter is very efficient and keeps the model from being driven completely away from the observations. The analogous run for the central east Pacific from  $208^{\circ}$  E to  $242^{\circ}$  E and from  $10^{\circ}$  N to  $30^{\circ}$  N produced very bad results. The analysis of the variance is presented on Figure 60 in the same manner as before. The variance of the misfit is now 5 times as large as that of the data. That is, the model propagates the signal in the totally wrong direction. This is, of course, not surprising as the model physics is clearly not adequate for this tropical region.

## 4.6 Conclusions.

In this chapter we presented an assimilation of over two years of T/P data with the linear vorticity balance approximated by equation (1.6). The model has no friction. The spatial resolution of the model was chosen to be that of the gridded data, i.e. very coarse. The assimilation is performed by using a steady Kalman filter after the state vector error settles to a steady value and we implement the steady filter after 50 days – a time period similar to that used by Jiang and Ghil (1993). In the north east Pacific, where topography is relatively smooth, and stratification is weak, we observe some agreement between the model and the data. This is reflected in the fact that on average the model forecast describes 33 percent of the data variability. The model propagates the signal in

the direction of the data, but it does not reproduce the data well. Moreover, there are significant qualitative differences between the model simulation and the data. Regions near the boundaries show the most difference. We conclude that in this “quiet” region of the Pacific we cannot adequately describe the variability of the data by the model.

Similar analysis in other areas, such as the central east Pacific and south east Pacific, showed that the model was inadequate for describing the variability there as well. The Kalman filter data update was very efficient in assimilating information in both cases. In the central Pacific, the forecast misfit has a much larger variance than the data itself, and in the south Pacific the forecast misfit has comparable variability to that of the data. The conclusion is clearly negative and the model is strongly rejected. The results of chapter 3 showed that the model cannot be rejected in several isolated locations and the results of this chapter show that no large area of the Pacific can be adequately described by the barotropic dynamics, i.e. the linear barotropic dynamics is not capable of describing variability on truly large scales. We believe that further analysis along the lines of this research using more complicated models should bring a better understanding of the global ocean variability.

The analysis can be carried forward in several directions. First, one can provide a better estimate of the error coming from mis-specification of the boundary conditions. This can be done in the manner outlined in Section 4.4. The model integration could be done on a finer grid than the observations, which would mean that the matrix  $E$  of equation (4.5) would become non-identity. Assimilation of the along track data which have much finer resolution and smaller error is possible by making the matrix  $E$  time-dependent. Next, more complicated dynamics including effects of stratification can be added. This would change of dynamic equation (4.1) and the process error structure.

## Appendix 1.

### 1. Model Properties.

First, we write the predicted value of  $y(t)$  in the form  $\hat{y}(t|\theta)$  which is in general a complicated expression involving past values of  $y$  and  $u$  and vector of parameters  $\theta$ . However, for the ARX model with  $A(q) = C(q) = 1$  it is simply  $\hat{y}(t|\theta) = \sum_{k=nb}^{k=ny+nb} \mathbf{B}^t(k)\mathbf{u}(t-k)$ . The prediction error is then defined as

$$\epsilon(t|\theta) = y(t) - \hat{y}(t|\theta) \quad (1)$$

After we collected the data it is natural to choose value of the parameter vector  $\theta$  which minimizes the cost function

$$\lambda_N(\theta) = \frac{1}{N} \sum_{t=1}^{t=N} \epsilon^2(t|\theta)$$

$$\hat{\lambda}_N = \min_{\theta} \lambda_N; \hat{\theta}_N = \arg \min_{\theta} \lambda_N. \quad (2)$$

It can be shown (see Ljung [1987]) that

$$\sqrt{N} (\hat{\theta}_N - \theta_0) \in \text{As}N(0, \lambda \bar{R}^{-1}) \quad \text{and} \quad (3)$$

$$\frac{\hat{\lambda}_N}{\lambda} \in \text{As} \frac{1}{N-d} \chi(N-d), \quad (4)$$

$$\text{where } d = \text{length}(\theta), \lambda = E[\epsilon^2(t, \theta)], \psi(t, \theta) = \frac{d\hat{y}(t|\theta)}{d\theta}, \bar{R} = E\psi(t, \theta)\psi^T(t, \theta)$$

where estimate of  $\lambda$  is  $\hat{\lambda}_N$  (2) and  $R$  is estimated by

$$\hat{R}_N = \frac{1}{N} \sum_{t=1}^{t=N} \psi(t, \hat{\theta}_N)\psi^T(t, \hat{\theta}_N) \quad (5)$$

Thus, minimisation of the cost function  $\lambda_N(\theta)$  provides us with the uncertainties of the estimates.

We can also formulate a hypothesis of whether all parameters are zeroes. This gives us Fisher's test on variances.

## 2. Maximum likelihood estimator of the mean

If the vector of parameters is denoted by  $\mathbf{z}_i$  and has uncertainty  $\mathbf{P}_i$ , where  $i$  subscript denotes spatial location where estimates were obtained, the mean value of these parameters and its uncertainty can be estimated as

$$\bar{\mathbf{z}} = \left( \sum_{i=1}^M \mathbf{P}_i^{-1} \right)^{-1} \left( \sum_{i=1}^M \mathbf{P}_i^{-1} \mathbf{z}_i \right) \quad (6)$$

$$\bar{\mathbf{P}} = \left( \sum_{i=1}^M \mathbf{P}_i^{-1} \right)^{-1} \quad (7)$$

## Bibliography

- [1] Abarbanel, H. D. I., R. Brown, J. J. Sidorowich, and L. Sh. Tsimring, 1993. Analysis of observed chaotic data. *Rev. Mod. Phys.*, 55, 1331-1392.
- [2] Box G. E. P. and G. M. Jenkins, 1976. Time series analysis: forecasting and control. Holden-Day.
- [3] Frankignoul, C., and P. Muller, 1979 Quasi-geostrophic response of an infinite beta-plane ocean to stochastic forcing by the atmosphere. *J. Phys. Oceanogr.*, 9, 104-127.
- [4] Fu L.-L. and R. Davidson, 1995 A note on the barotropic response of sea level to time-dependent wind-forcing. *J. Geophys. Res.*, 100, C12, 24955-24,963.
- [5] Fu L.-L., I. Fukumori and R. Miller, 1993 Fitting dynamics to the Geosat sea level observations in the tropical pacific ocean. Part II: a linear, wind-driven model. *J. Phys. Oceanogr.*, 23, 2162-2181.
- [6] Fu L.-L. and G. Pihos, 1994 Determining the response of sea level to atmospheric pressure forcing using TOPEX/POSEIDON data. *J. Geophys. Res.*, 99, C12, 24,619-24,632.
- [7] Fu L.-L., J. Vazquez and C. Perigaud, 1991 Fitting dynamics to the Geosat sea level observations in the tropical pacific ocean. Part I: a free wave model. *J. Phys. Oceanogr.*, 21, 798-809.
- [8] Fukumori I., 1995 Assimilation of TOPEX sea level measurements with a reduced-gravity, shallow water model of the tropical Pacific Ocean. *J. Geophys. Res.*, 100, C12, 25,027-25,039.

- [9] Gaspar P. and C. Wunsch, 1989 Estimates from altimeter data of barotropic Rossby waves in the Northwestern Atlantic Ocean. *J. Phys. Oceanogr.*, **19**, 1821-1844.
- [10] Gentemann, C. 1995. *An investigation of sea level response to atmospheric pressure forcing using Topex/Poseidon data*. Thesis (B.S.)—Massachusetts Institute of Technology, Dept. of Earth, Atmospheric, and Planetary Sciences, 1995.
- [11] Jiang. S. and M. Ghil, 1993 Dynamical properties of error statistics in a shallow-water model. *J. Phys. Oceanogr.*, **23**, 2541-2566.
- [12] King, C., D. Stammer, and C. Wunsch, The CMPO/MIT TOPEX/POSEIDON altimetric data set, Rep. 30, 33 pp. plus color plates, Cent. for Global Change Sci., Mass. Inst. of Technol., Cambridge, 1994 .
- [13] Koblinsky C. J., J. P. Niller, and W. J. Schmitz, 1989. Observations of wind-forced deep ocean currents in the North Pacific. *J. Geophys. Res.*, **94**, 10,773-10,790.
- [14] Ljung, L. 1987. *System identification - theory for the user*. Prentice Hall, Englewood Cliffs, NJ.
- [15] Ljung, L. 1994. *System identification toolbox, user's guide*, 4th ed. The MathWorks, Inc., Natick, Mass.
- [16] Pedlosky, J. 1987. *Geophysical Fluid Dynamics*, 2nd ed. Springer-Verlag.
- [17] Ponte, R. M. 1994: Preliminary assesment of the accuracy and precision of TOPEX/POSEIDON altimeter data with respect to the large scale ocean circulation. *J. Geophys. Res.*, **99**, C4, 8,033–8,039.
- [18] Willebrand J., S. G. H. Philander and R. C. Pacanowski, 1980 The oceanic response to large scale atmospheric disturbances. *J. Phys. Oceanogr.* , **10**, 411-429.
- [19] Wunsch, C. 1991. Large scale response of the ocean to atmospheric forcing at low frequencies. *J. Geophys. Res.*, **96**, 15,083–15,092.
- [20] Wunsch, C. and D. Stammer, 1995. The global frequency-wavenumber spectrum of oceanic variability from TPOEX/POSEIDON. *J. Geophys. Res.*, **100**, C12, 24955-24,963.

- [21] Wunsch, C. 1996. *The ocean circulation inverse problem.*, Cambridge University Press, New York, NY.

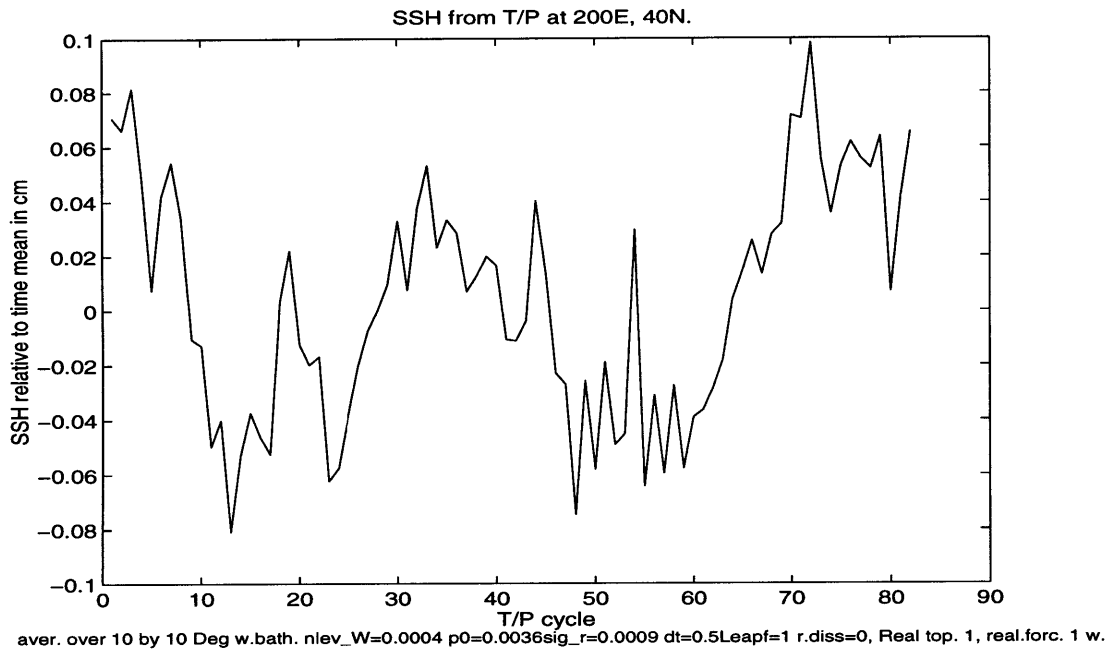


Figure 2: SSH from T/P at 200° E, 40° N in cm. Time mean has been removed.

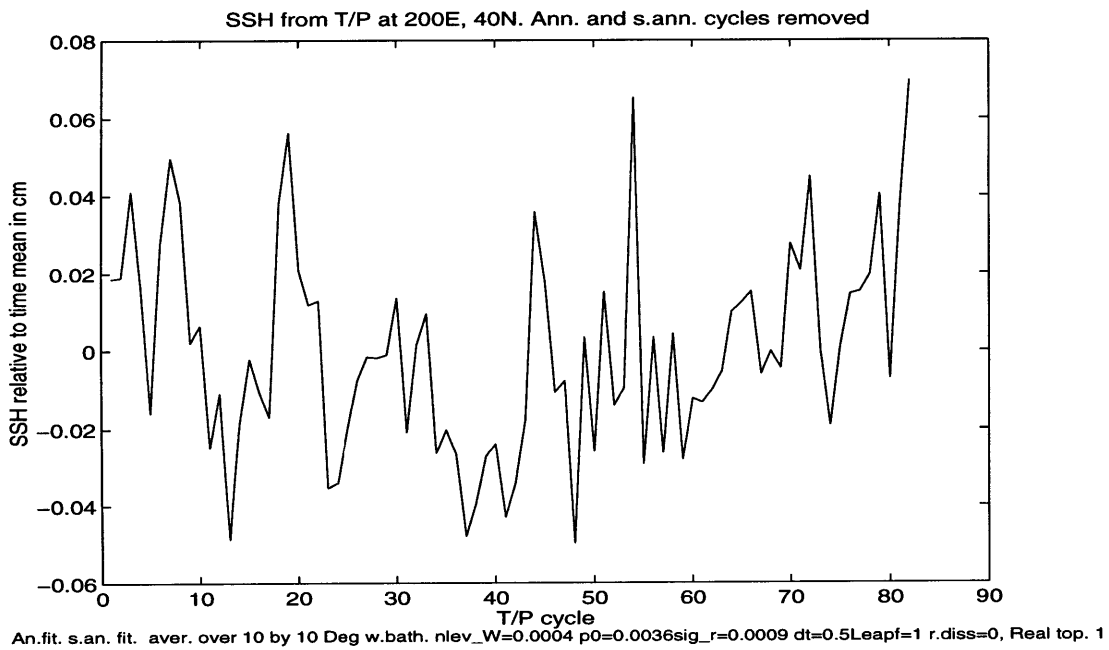


Figure 3: SSH from T/P at 200° E, 40° N in cm. Annual, semi-annual cycles, and time mean have been removed.

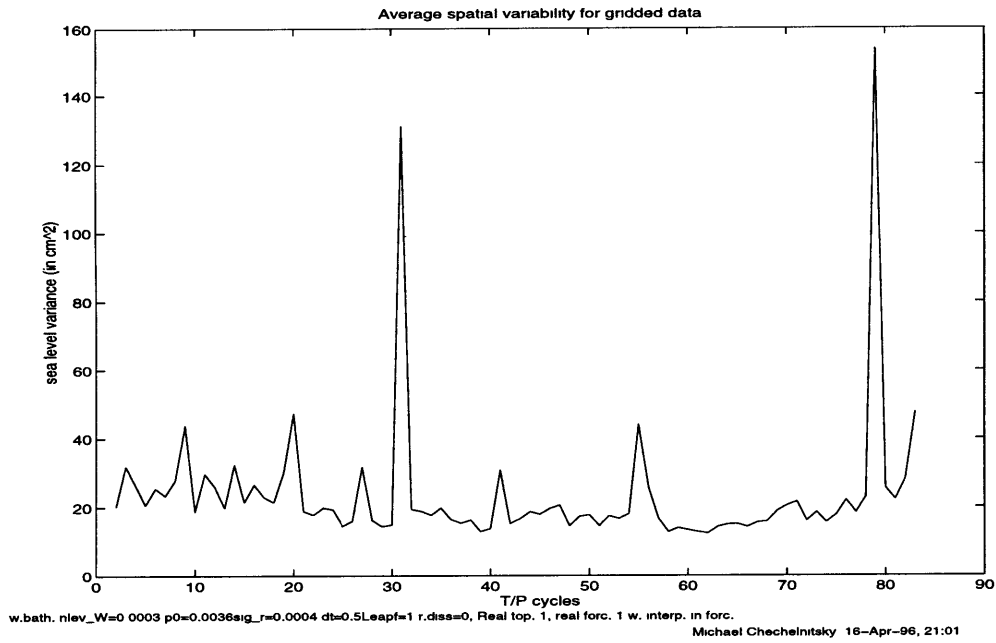


Figure 4: Mean SSH spatial variability from gridded T/P data as a function of time.

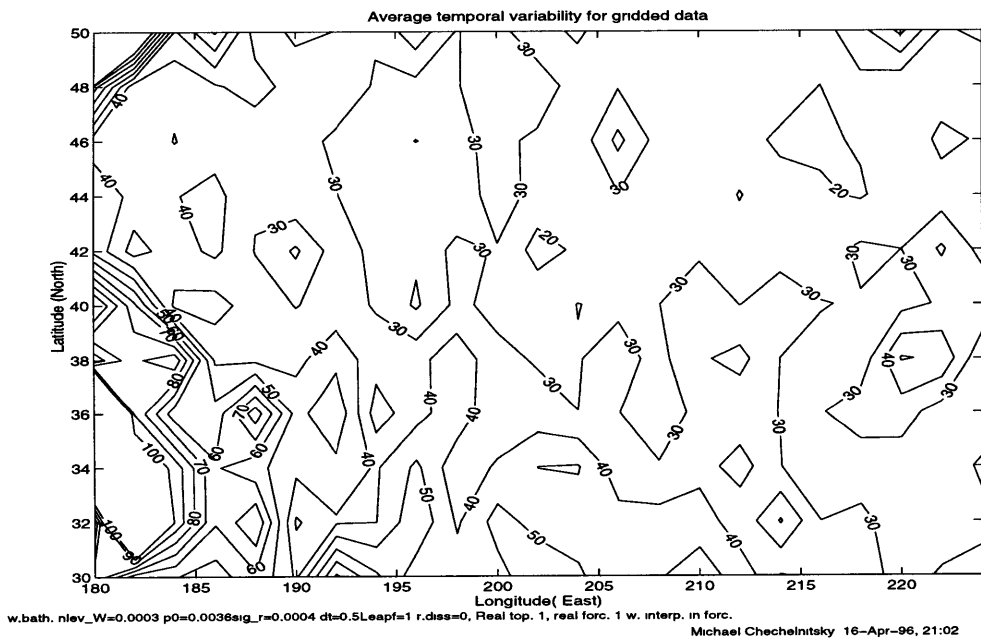
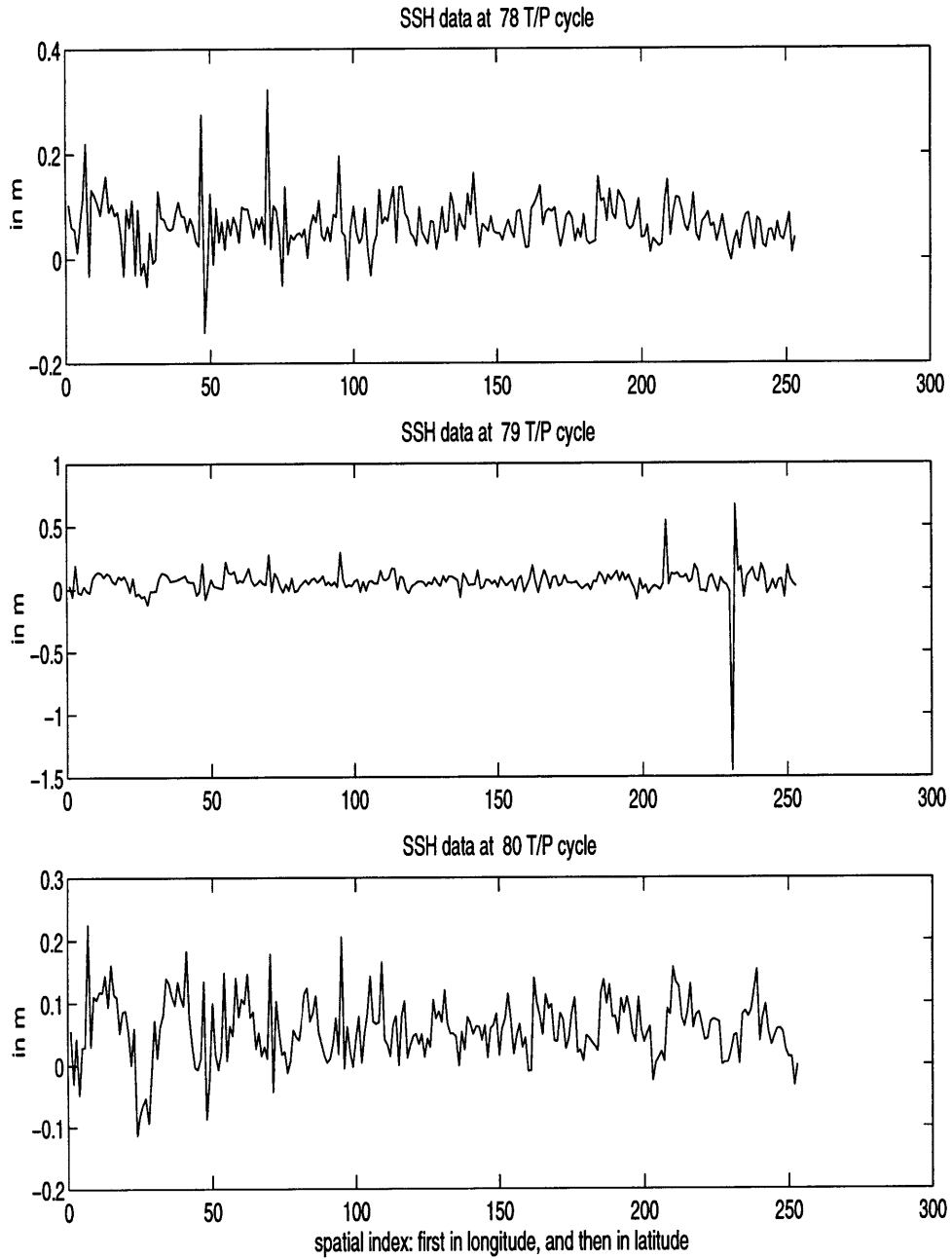


Figure 5: Contours of mean SSH spatial variability from gridded T/P data.



X from 180E to 224E; Y from 30N to 50N

Michael Chechelnitzsky 24-Apr-96, 1:25

Figure 6: Spatial distribution of SSH from T/P at 78th, 79th, and 80 T/P cycles. Time mean has been removed.

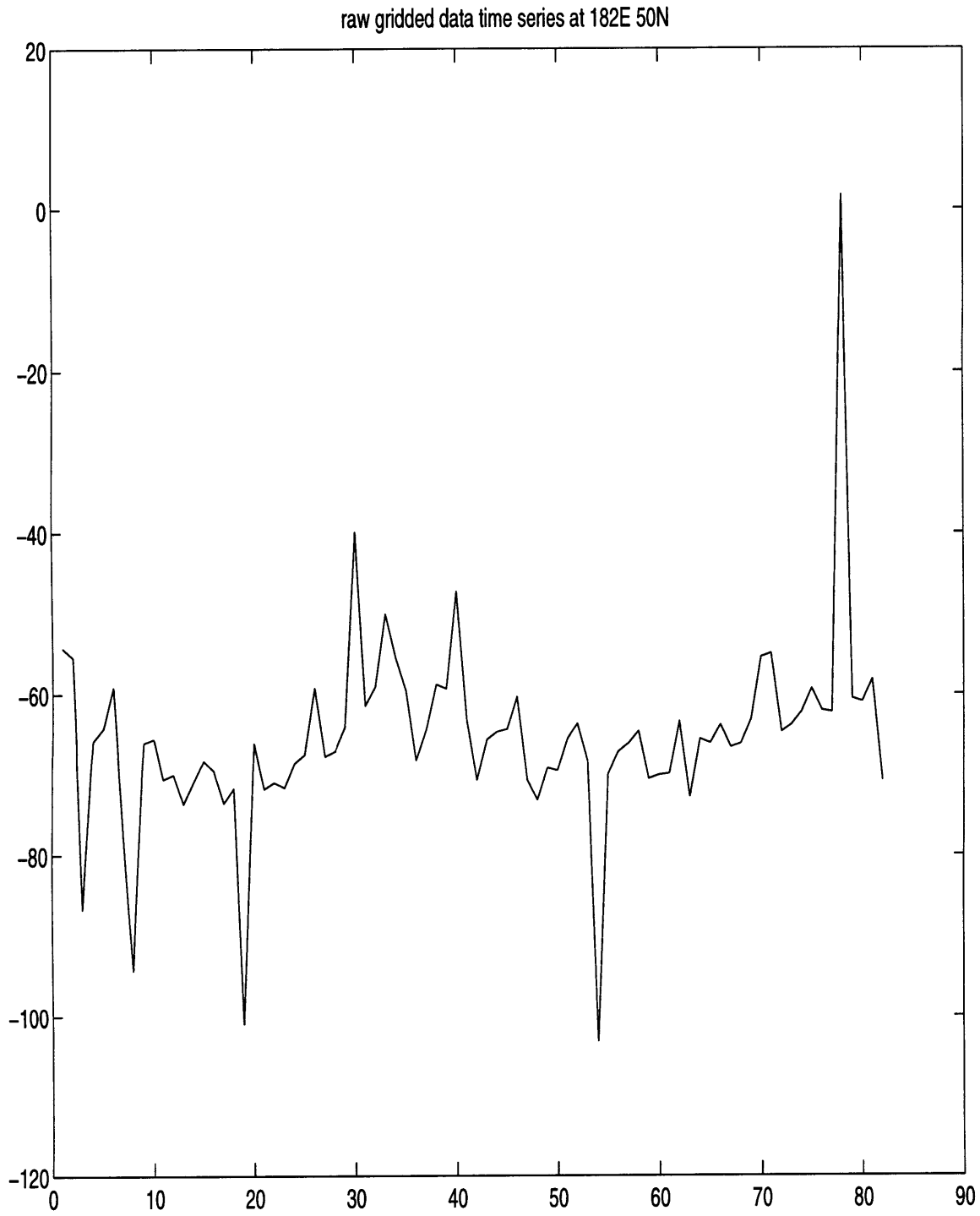


Figure 7: gridded SSH from T/P at 182° E, 50° N in cm.

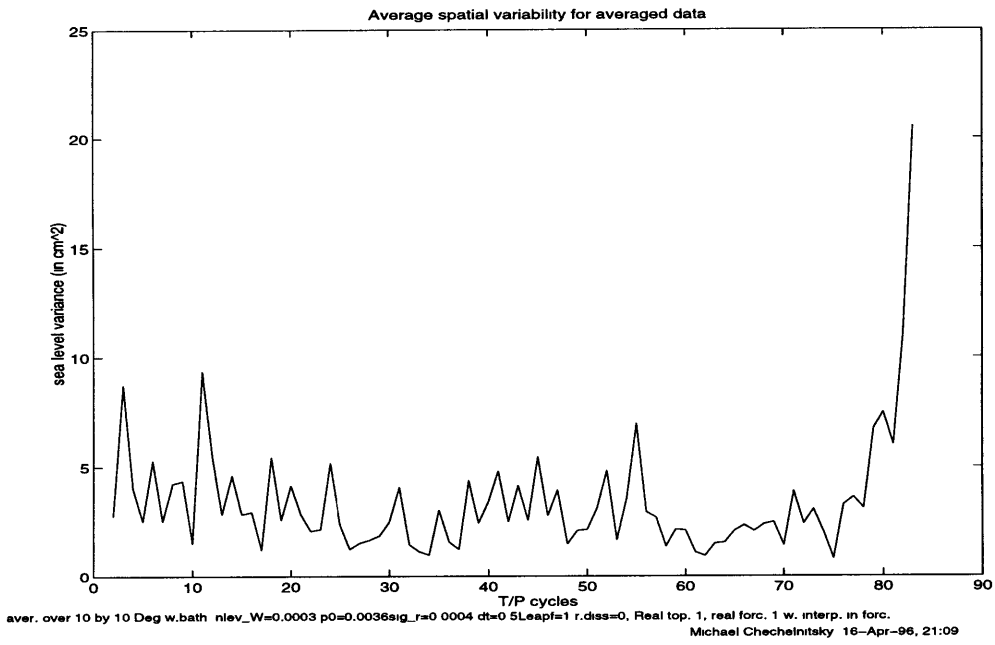


Figure 8: Mean SSH spatial variability from averaged T/P data as a function of time.

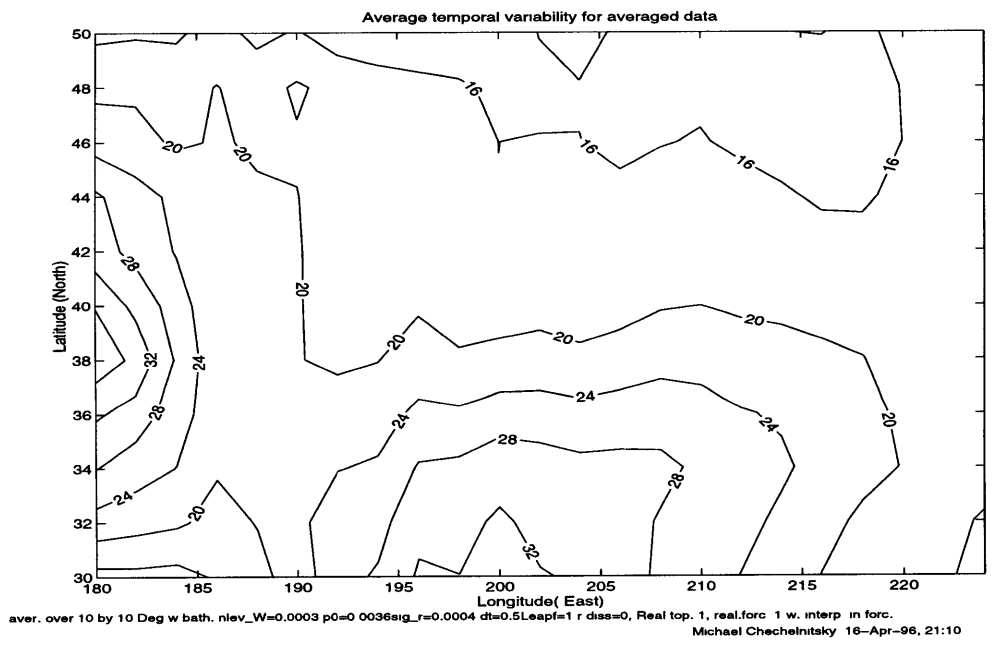


Figure 9: Contours of mean SSH spatial variability from averaged T/P data.

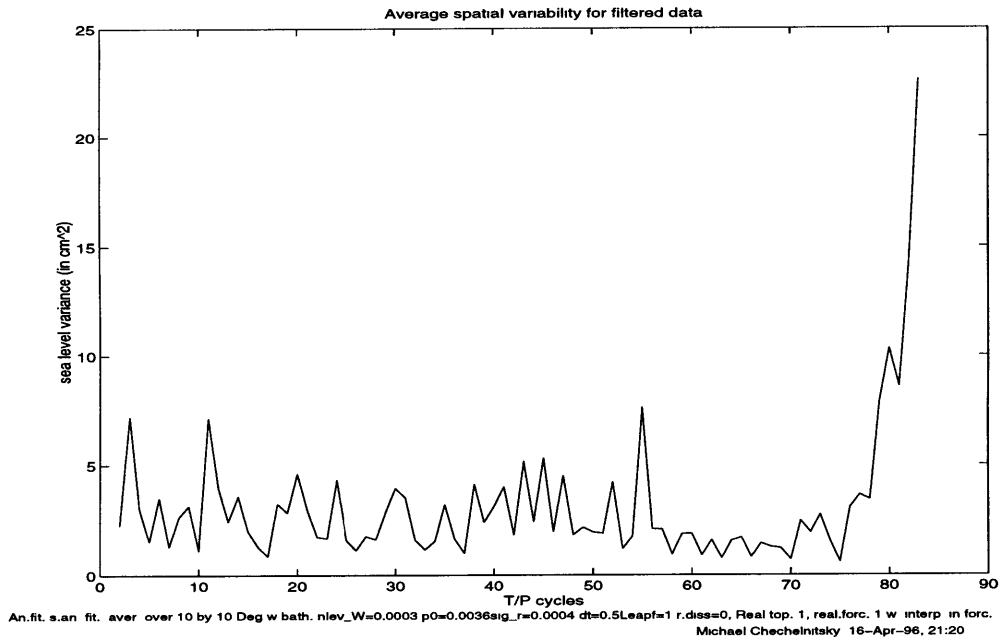


Figure 10: Mean SSH spatial variability from filtered T/P data as a function of time.

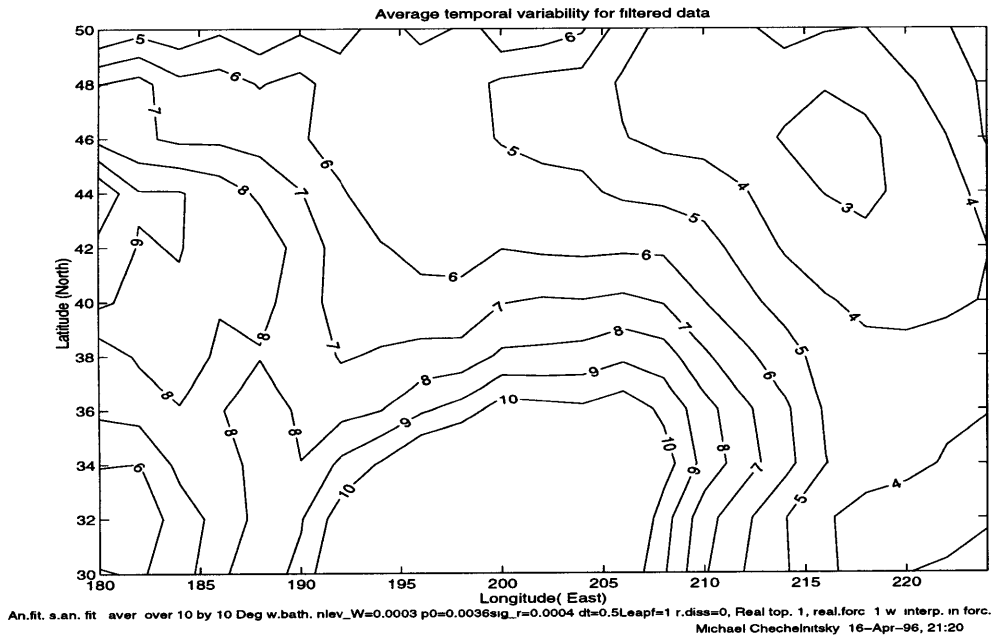


Figure 11: Contours of mean SSH spatial variability from filtered T/P data.

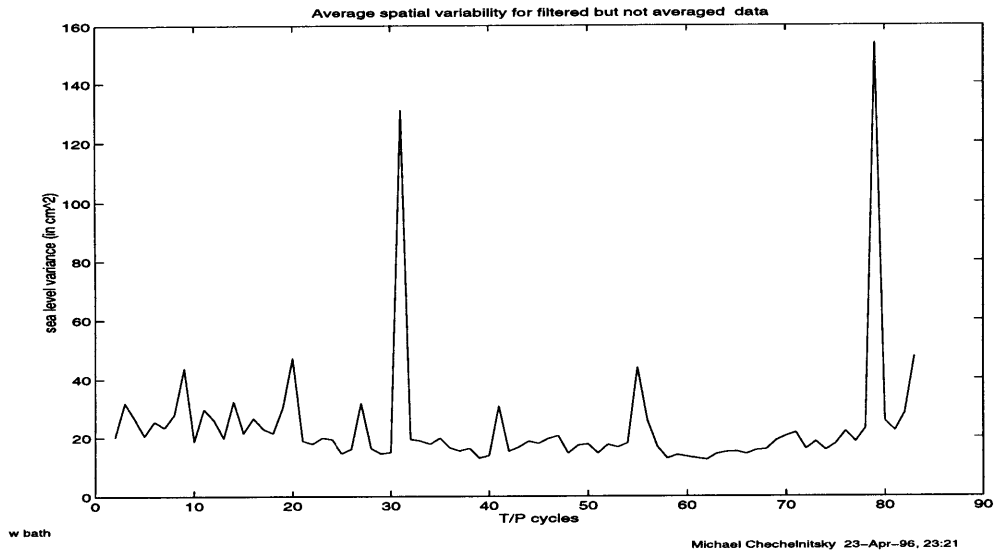


Figure 12: Mean SSH spatial variability from filtered but not averaged T/P data as a function of time.

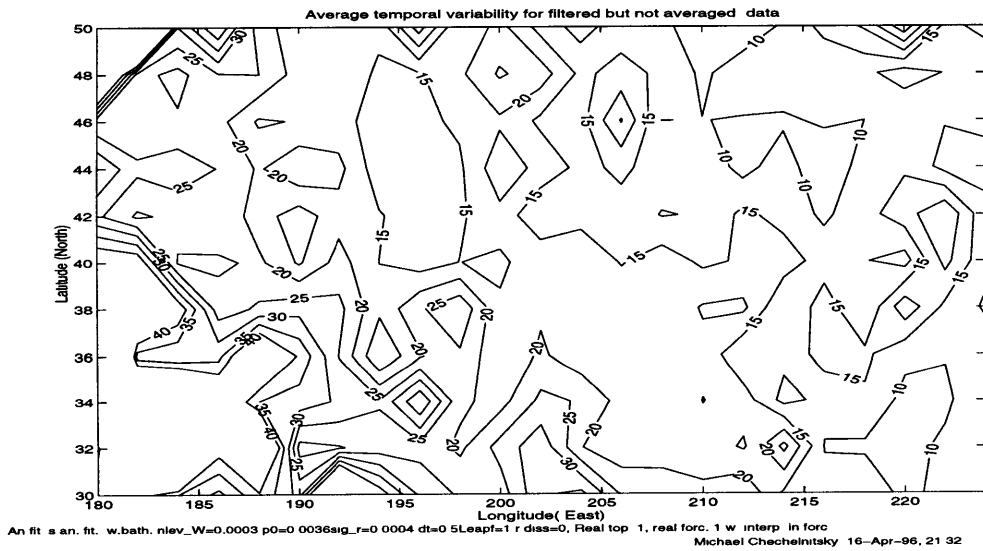


Figure 13: Contours of mean SSH spatial variability from filtered but not averaged T/P data.

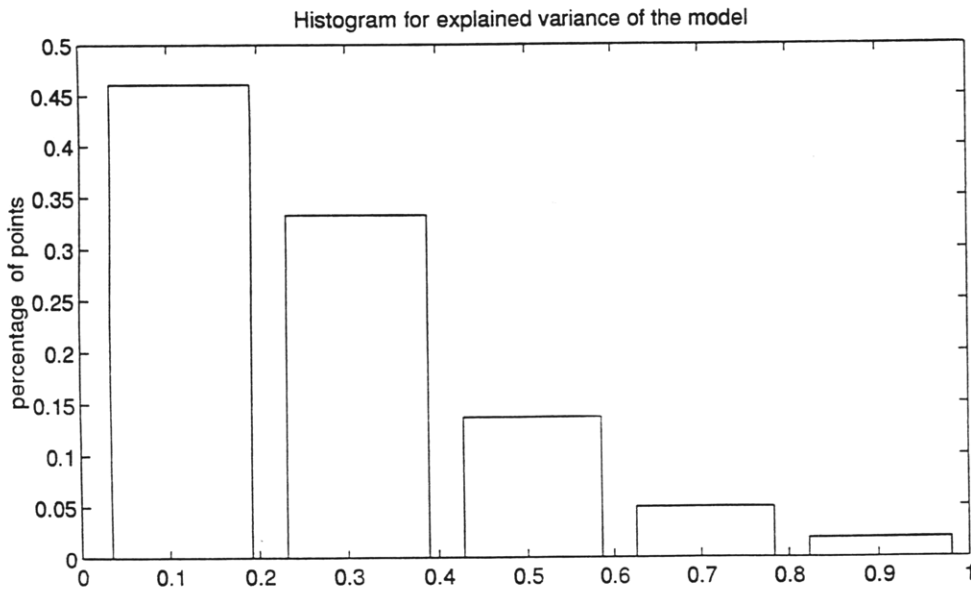
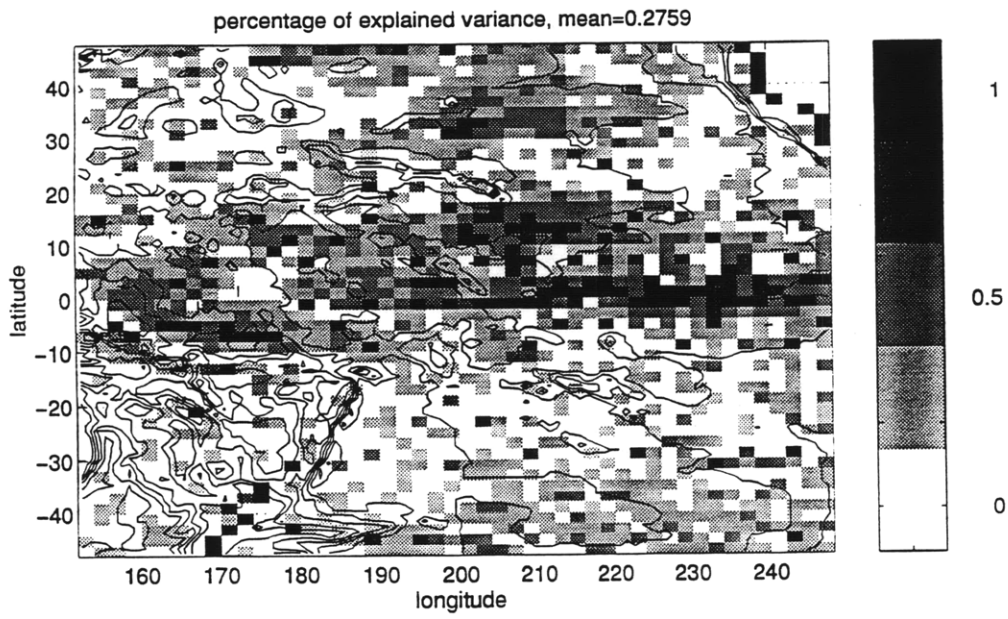


Figure 14: Contours of percentage of explained variance (upper part) and the histogram of explained variance (lower part) for the model (3.11). On the upper plot the underlying contours show bathymetry at the contour spacing of 1 km.

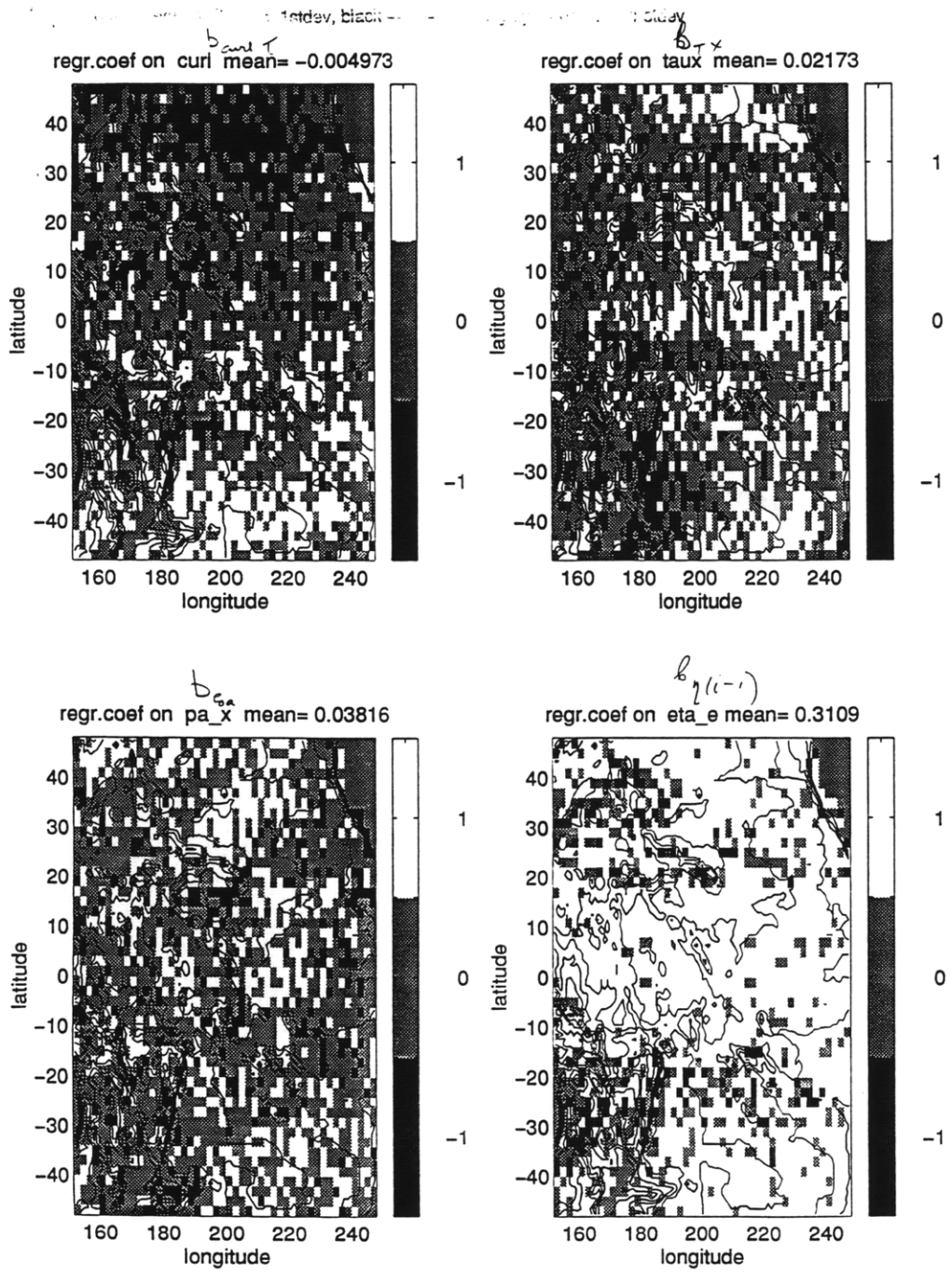


Figure 15: Distribution of regression coefficients for normalized variables for  $(\nabla \times \tau)_z$  (upper left),  $\tau^x$  (upper right),  $\partial \zeta_a / \partial x$  (lower left), and  $\eta(x-1, y, t)$  (lower right). The values below minus one standard deviation are denoted by black, above one standard deviation by white, and in between by gray. The underlying contours show bathymetry at the contour spacing of 1 km. The data are filtered.

output eta: white -- > 1stdev, black -- < -1 stdev, gray -- betw. -1 1stdev, indep\_factor=0.3333 x\_aver=4 y\_

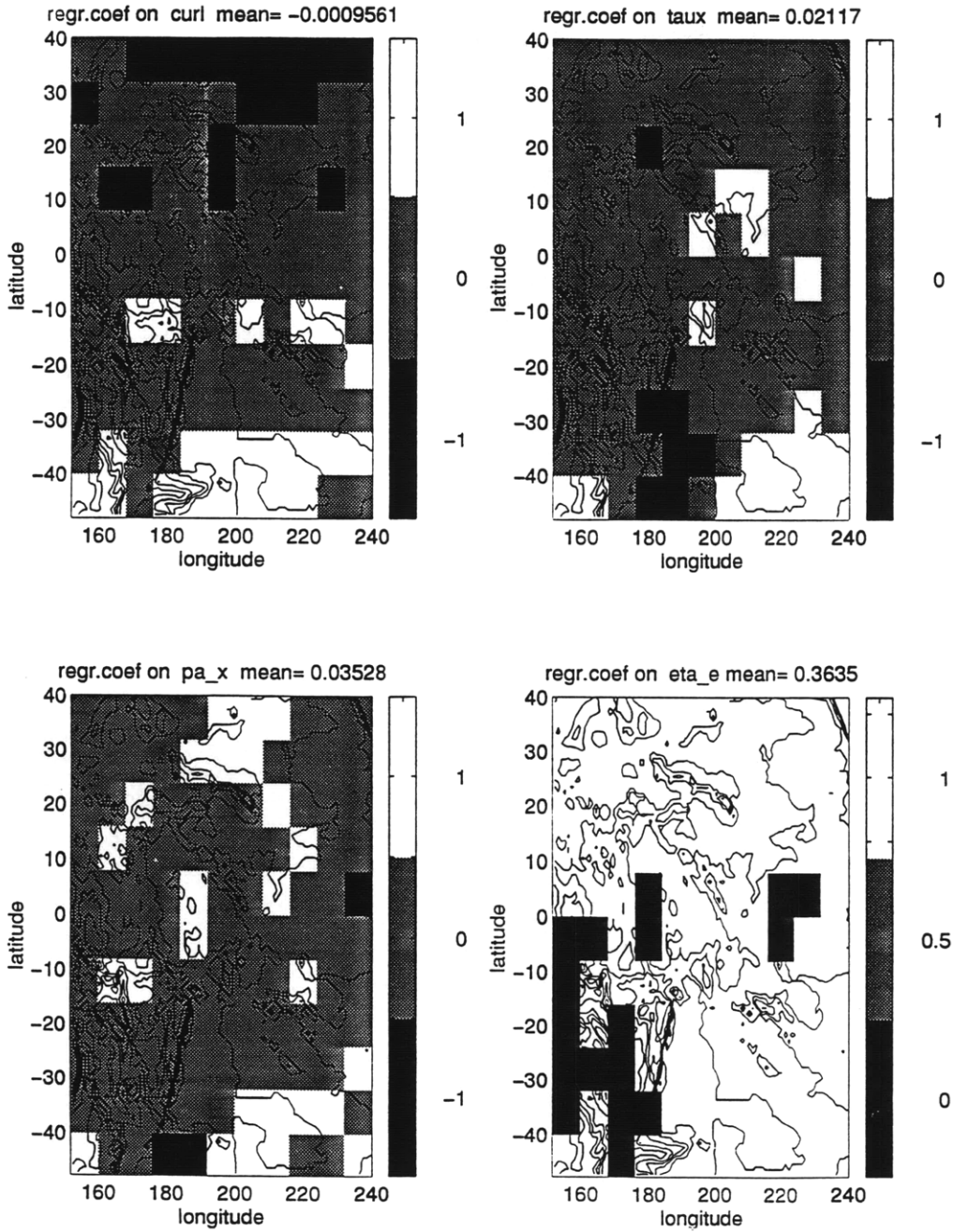


Figure 16: Same as in Figure 15 but the coefficients have been averaged in space.

output: eta inputs: tau\_x curl pa\_x eta\_e

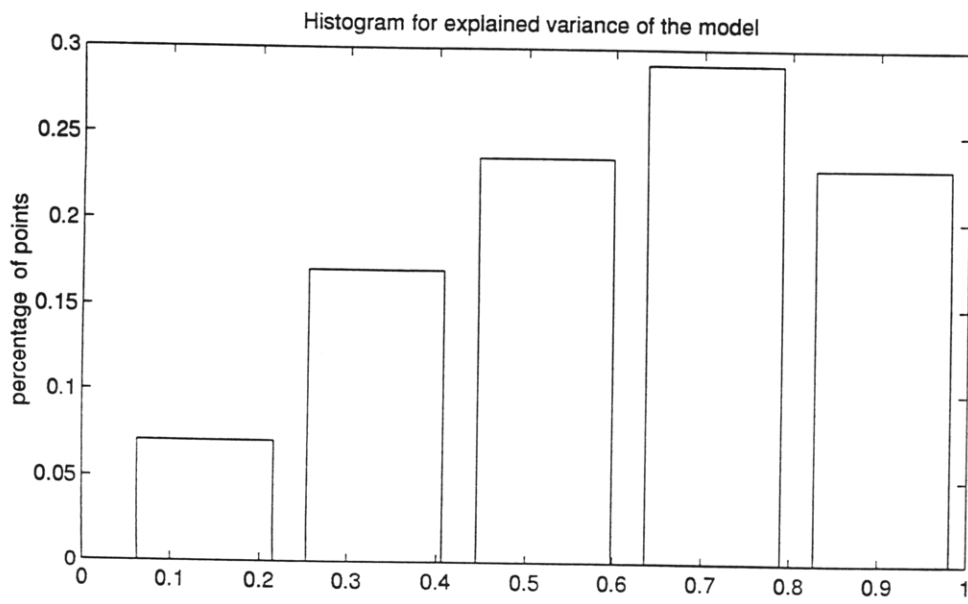
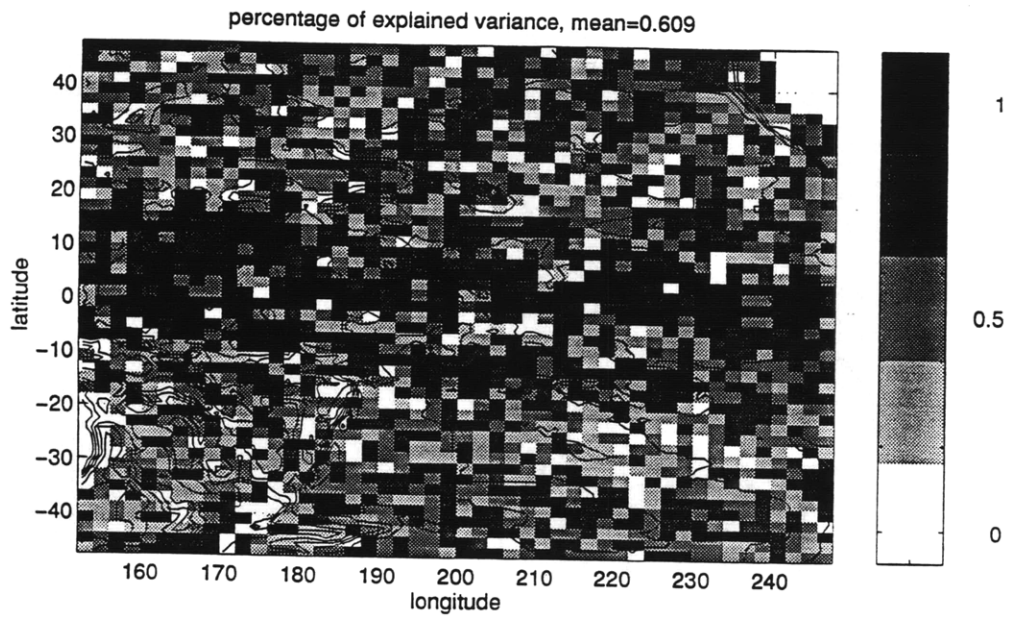
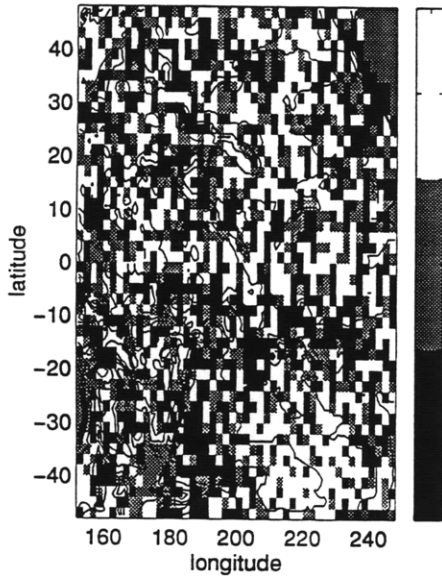


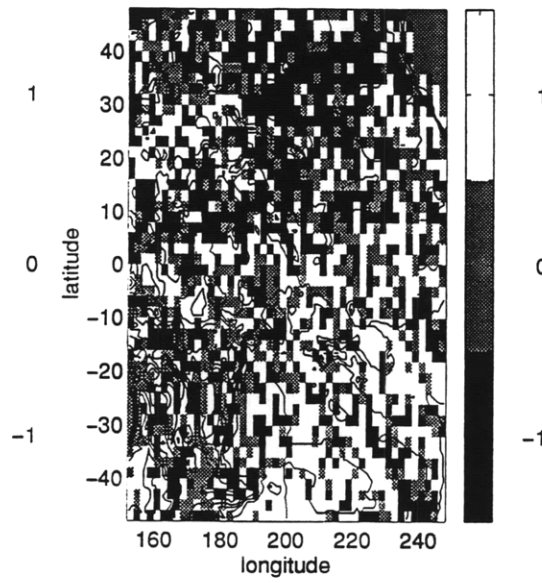
Figure 17: Same as in Figure 14 but for the band-passed data.

output eta 1 --> gt. 1 stdev, -1 --> less (-1) stdev, 0 --> betw. -1 1 stdev

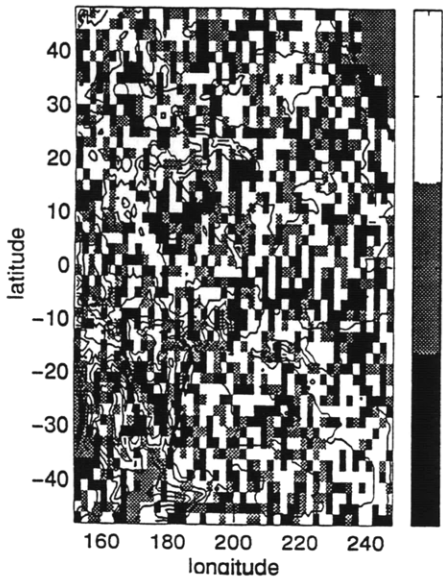
regr.coef on tau\_x mean= 0.00142



regr.coef on curl mean= -0.02566



regr.coef on pa\_x mean= 0.03815



regr.coef on eta\_e mean= 0.324

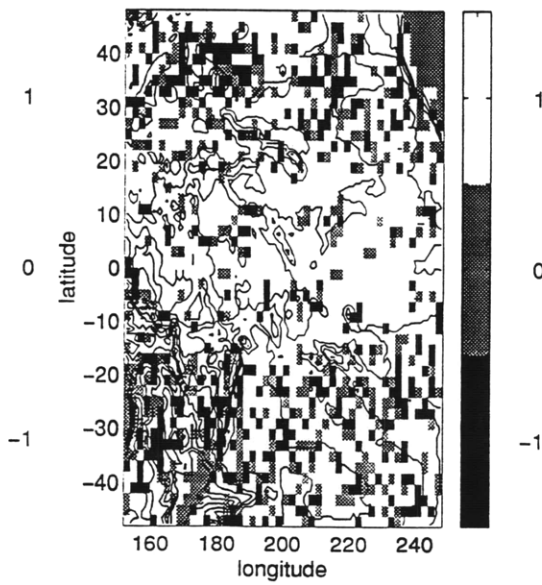


Figure 18: Same as in Figure 15 but the data in addition has been band-passed.

output eta white -- > 1stdev, black -- < -1 stdev, gray -- betw. -1 1stdev, indep\_factor=0.3333 x\_aver=4 y\_ε

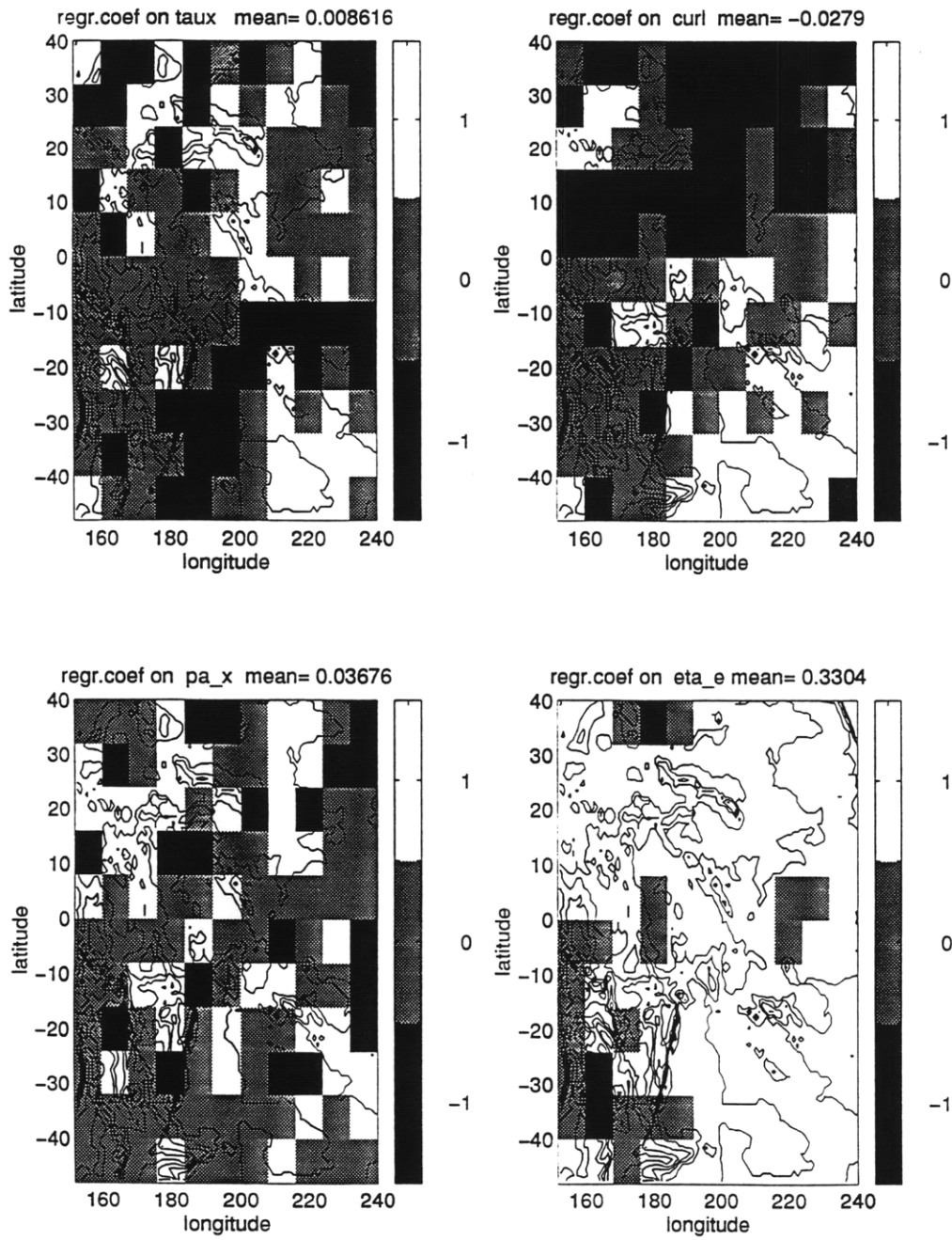


Figure 19: Same as in Figure 16 but the data in addition has been band-passed.

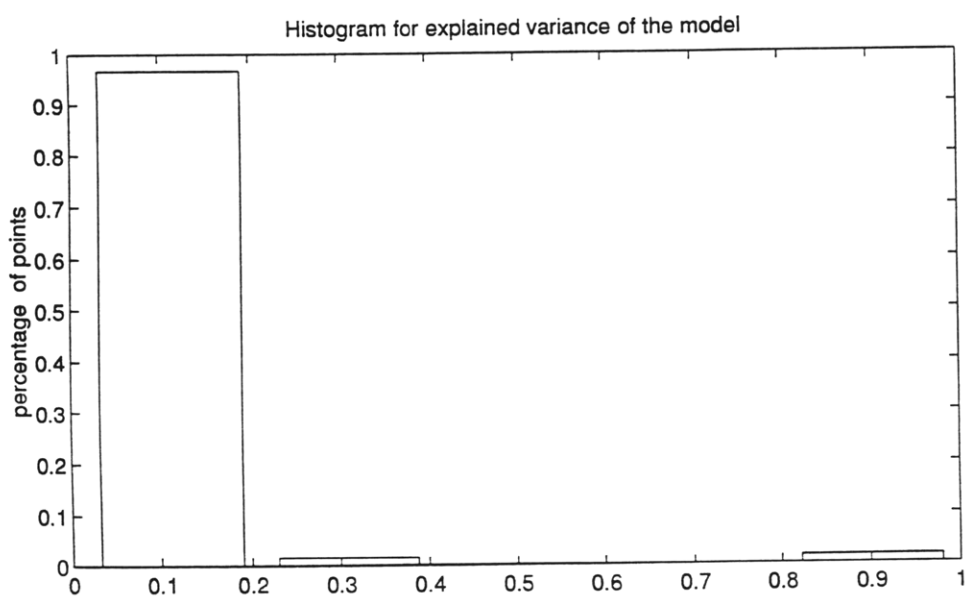
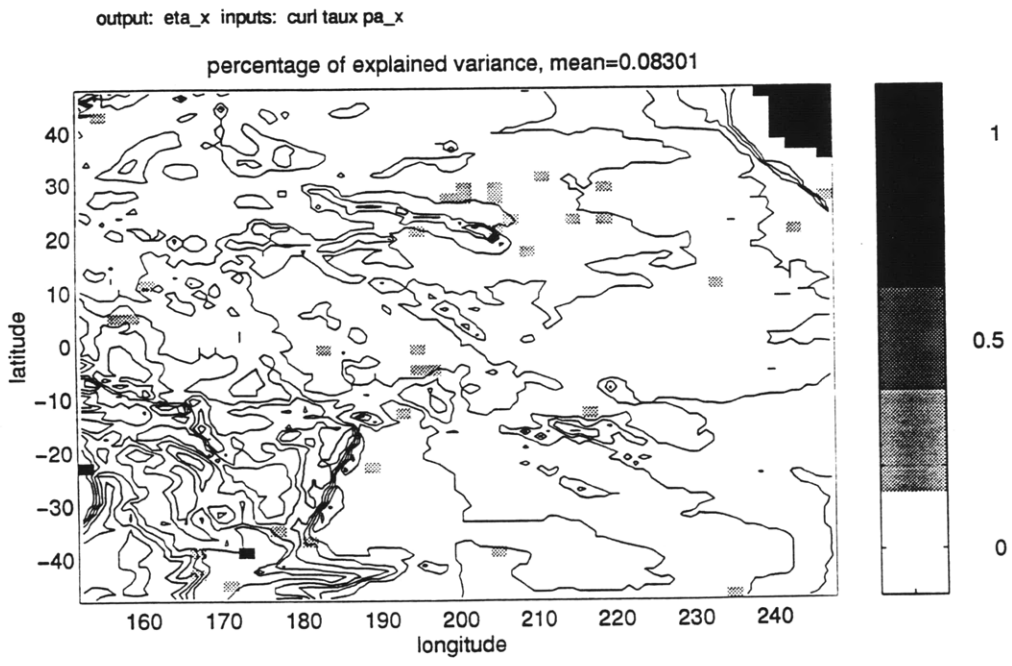


Figure 20: Same as in Figure 14 but for the model 3.17.

output eta\_x 1 --> gt. 1 stdev, -1 --> less (-1) stdev, 0 --> betw. -1 1 stdev

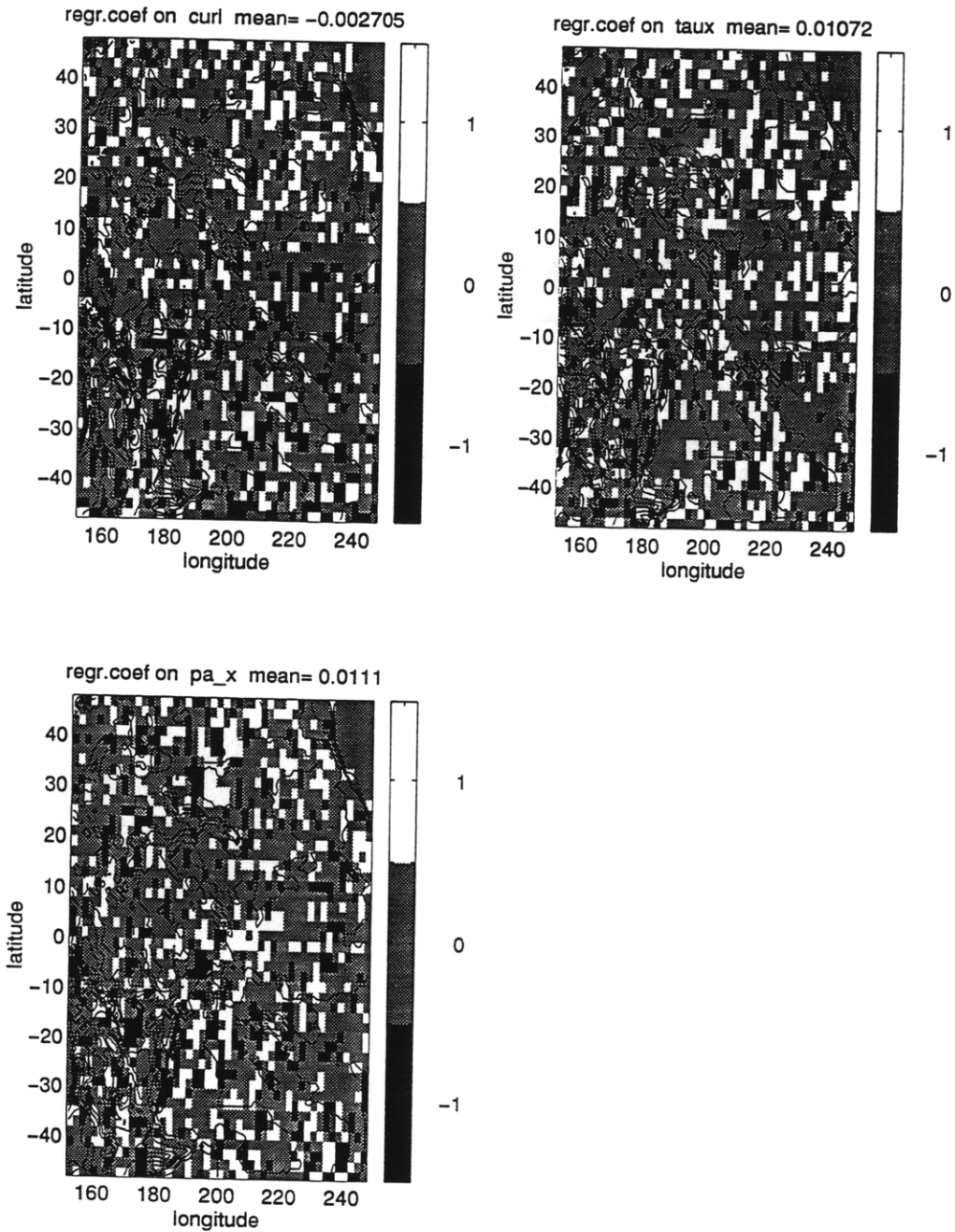


Figure 21: Distribution of regression coefficients for normalized variables for  $(\nabla \times \tau)_z$ (upper left),  $\tau^x$ (upper right), and  $\partial\zeta_a/\partial x$ (lower left). The values below minus one standard deviation are denoted by black, above one standard deviation by white, and in between by gray. The underlying contours show bathymetry at the contour spacing of 1 km.

output eta\_x white --> 1stdev, black --< -1 stdev, gray -- betw. -1 1stdev, indep\_factor=0.3333 x\_aver=4,

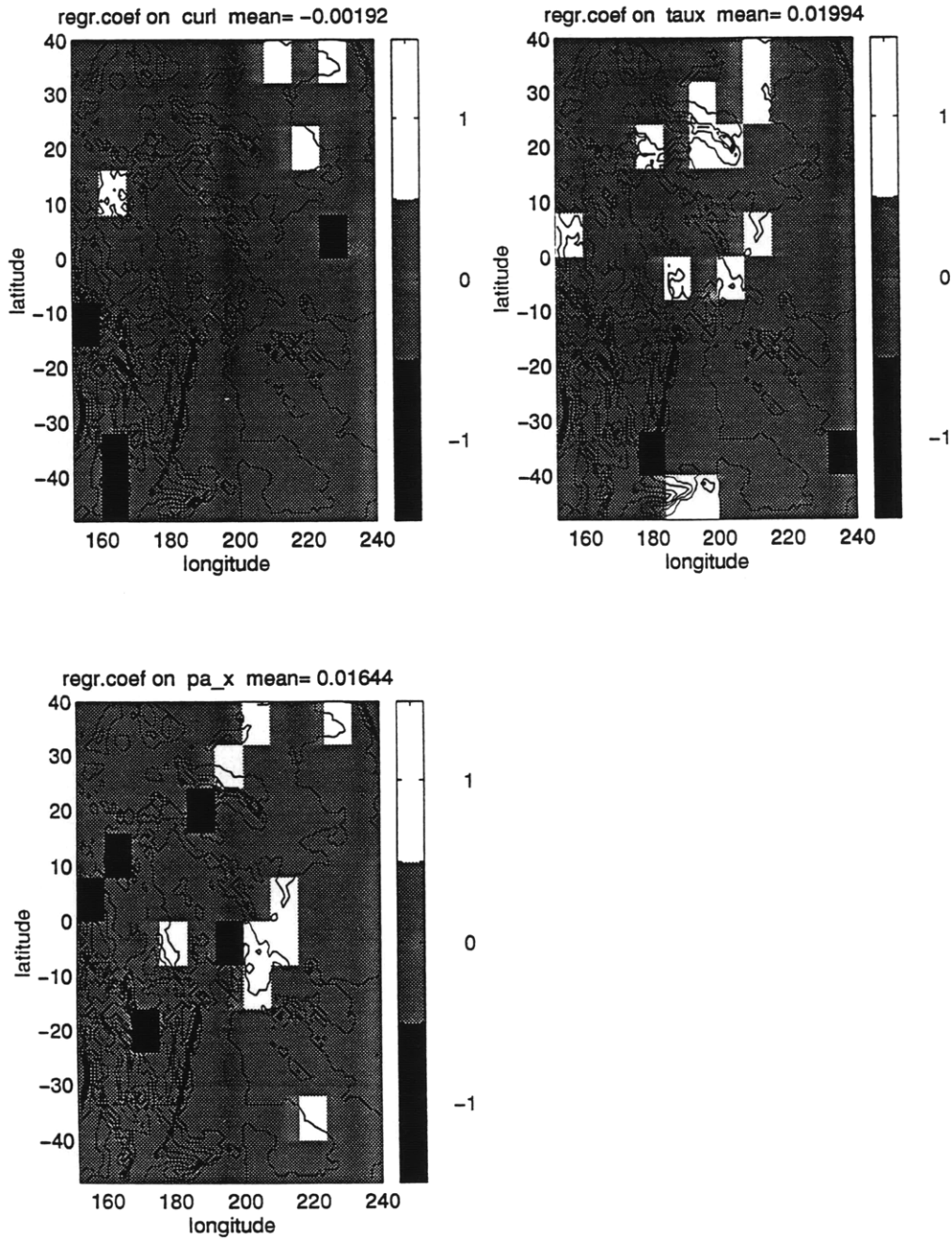
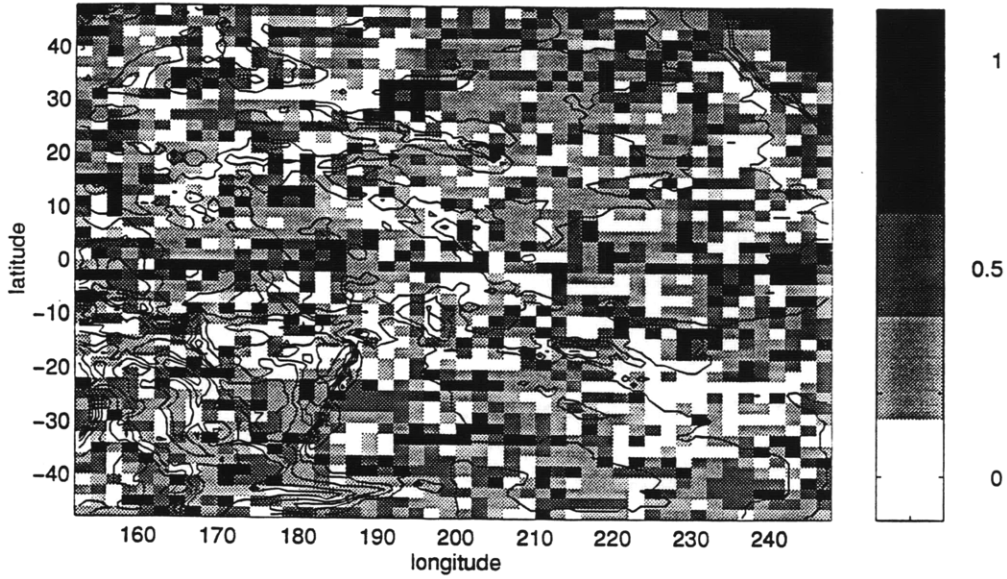


Figure 22: Same as in Figure 21 but the coefficients have been averaged in space.

output: eta\_x inputs: curl tau\_x pa\_x

percentage of explained variance, mean=0.3806



Histogram for explained variance of the model

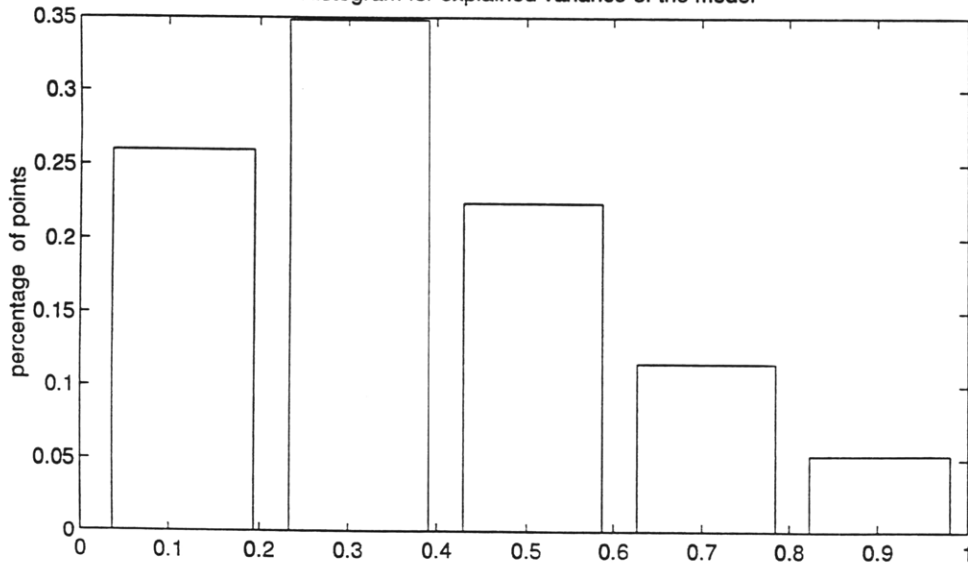


Figure 23: Same as in Figure 20 but for the band-passed data.

output eta\_x 1 --> gt. 1 stdev, -1 --> less (-1) stdev, 0 --> betw. -1 1 stdev

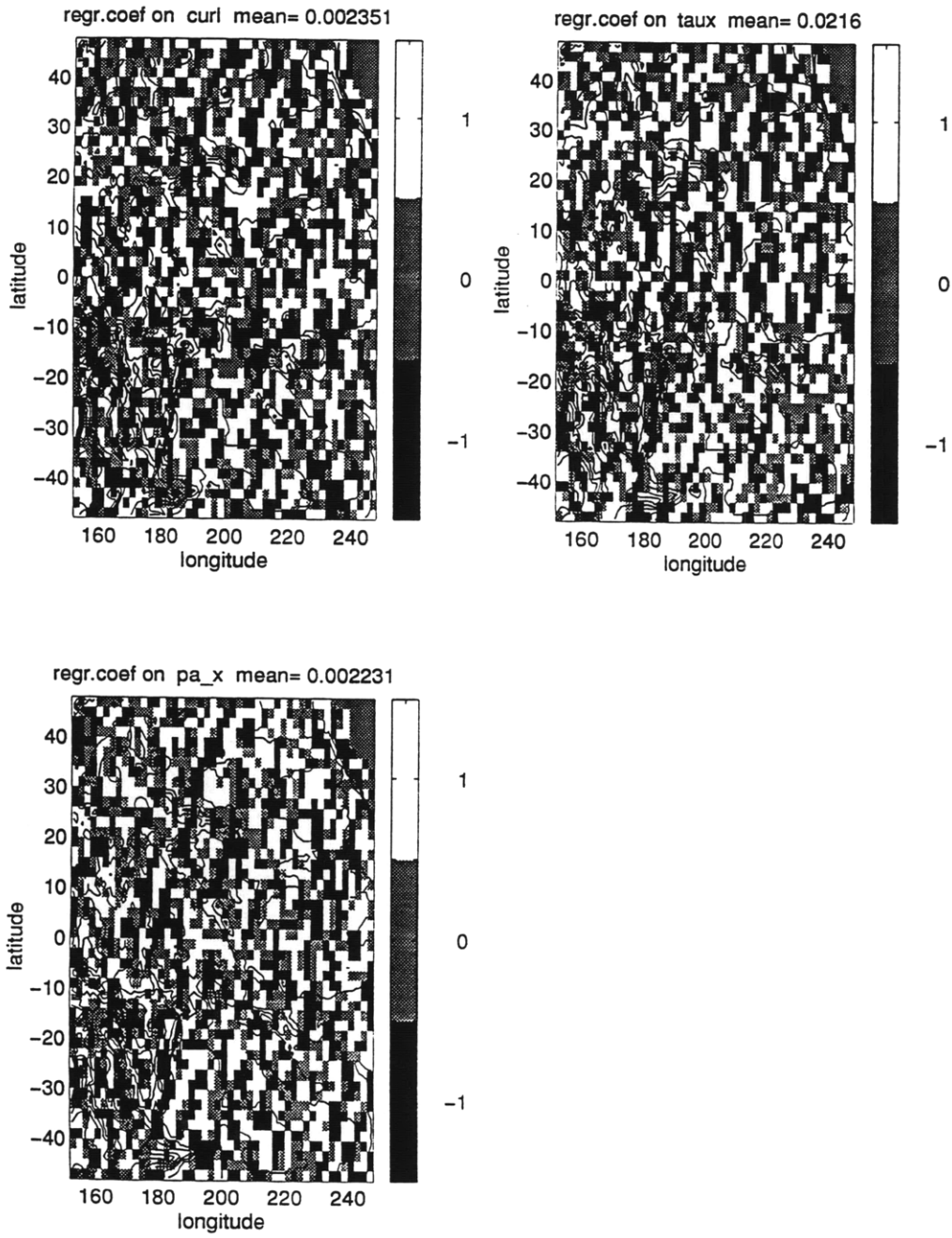


Figure 24: Same as in Figure 21 but for the band-passed data.

output eta\_x white -- > 1stdev, black -- < -1 stdev, gray -- betw. -1 1stdev, indep\_factor=0.3333 x\_aver=4 y

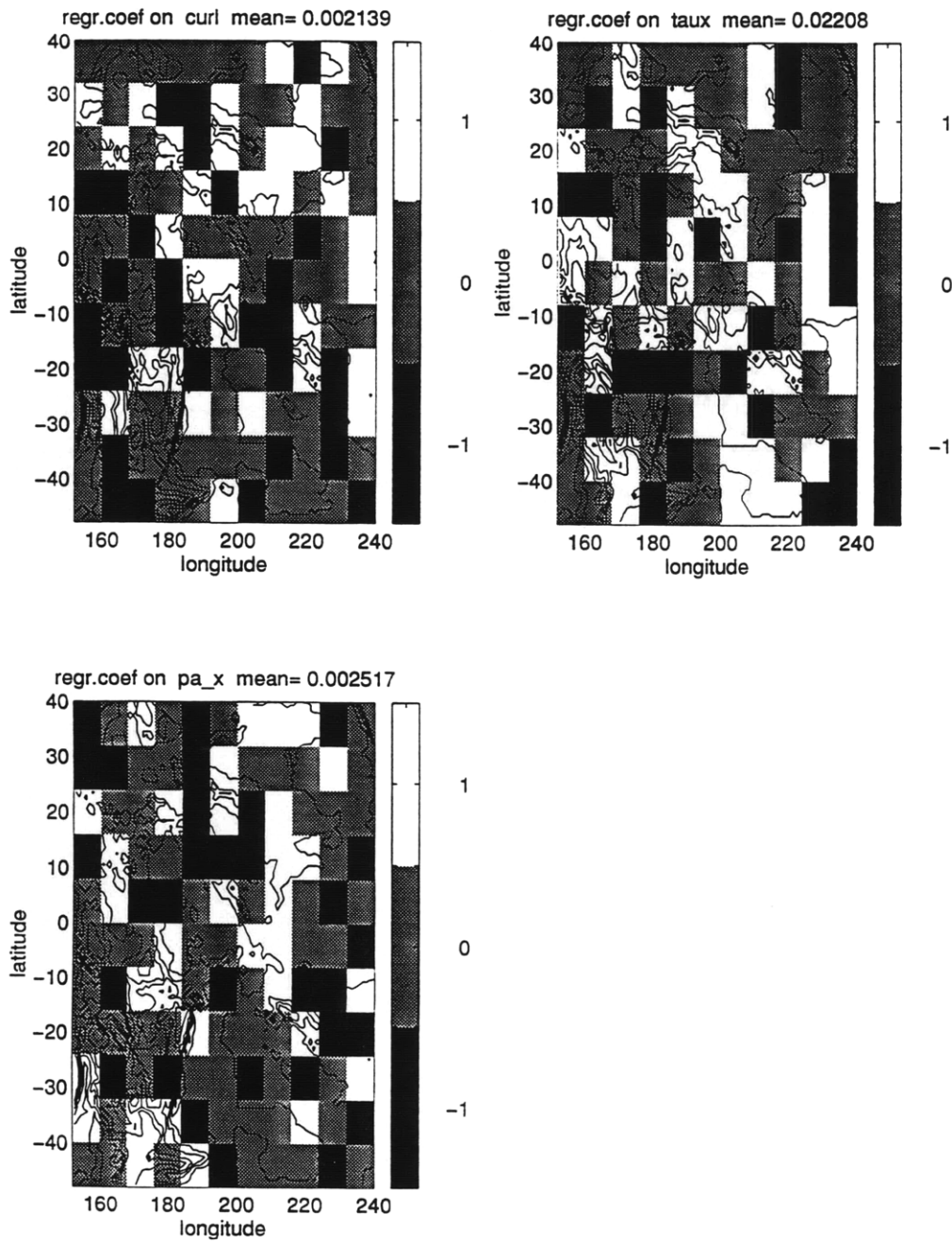


Figure 25: Same as in Figure 22 but for the band-passed data.

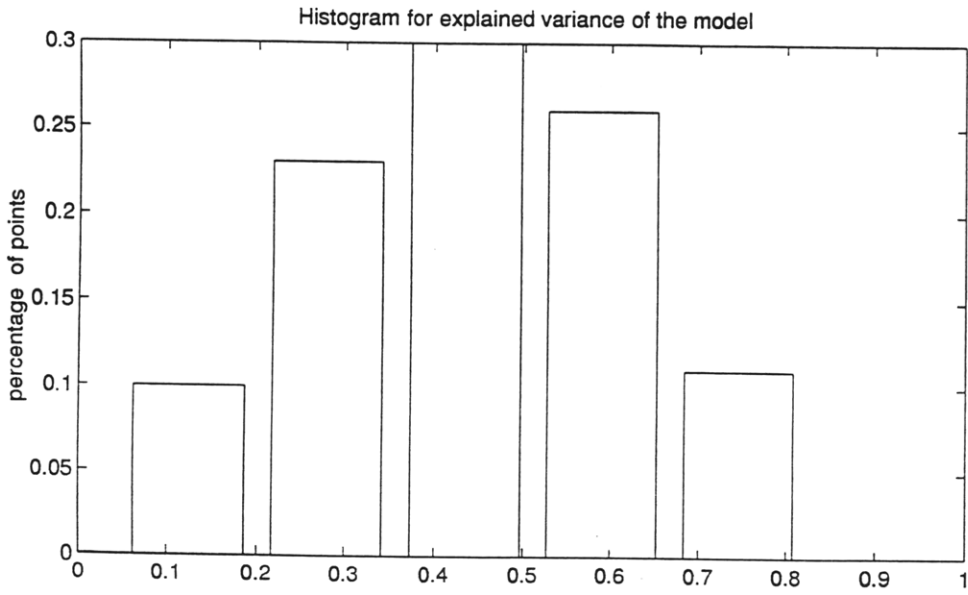
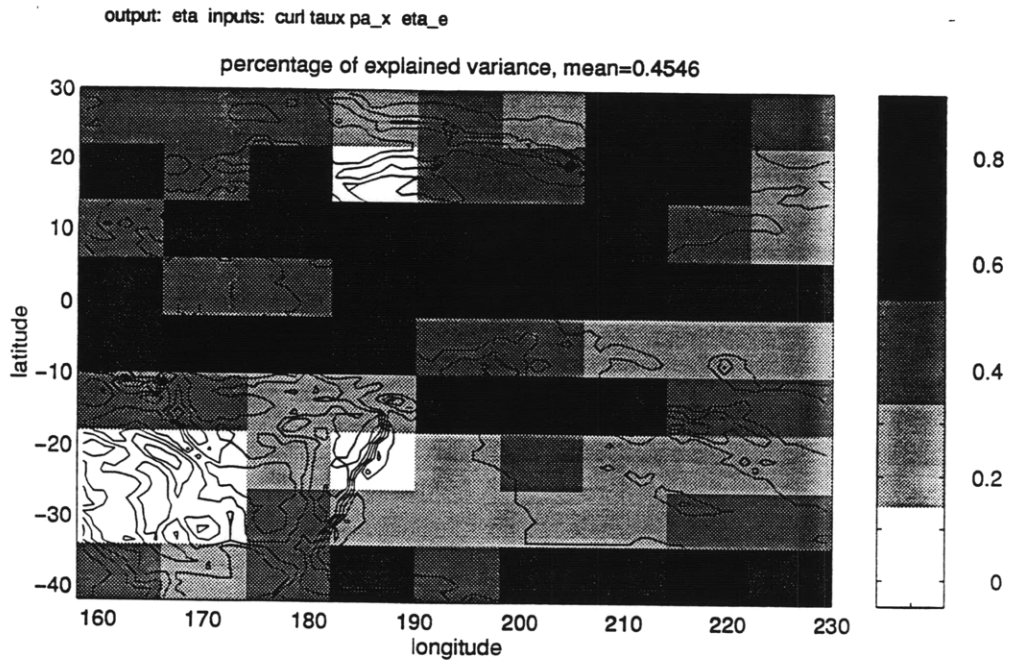


Figure 26: Same as in Figure 14 but the data has been averaged in space.

output eta white -- > 1stdev, black -- < -1 stdev, gray -- betw. -1 1 stdev

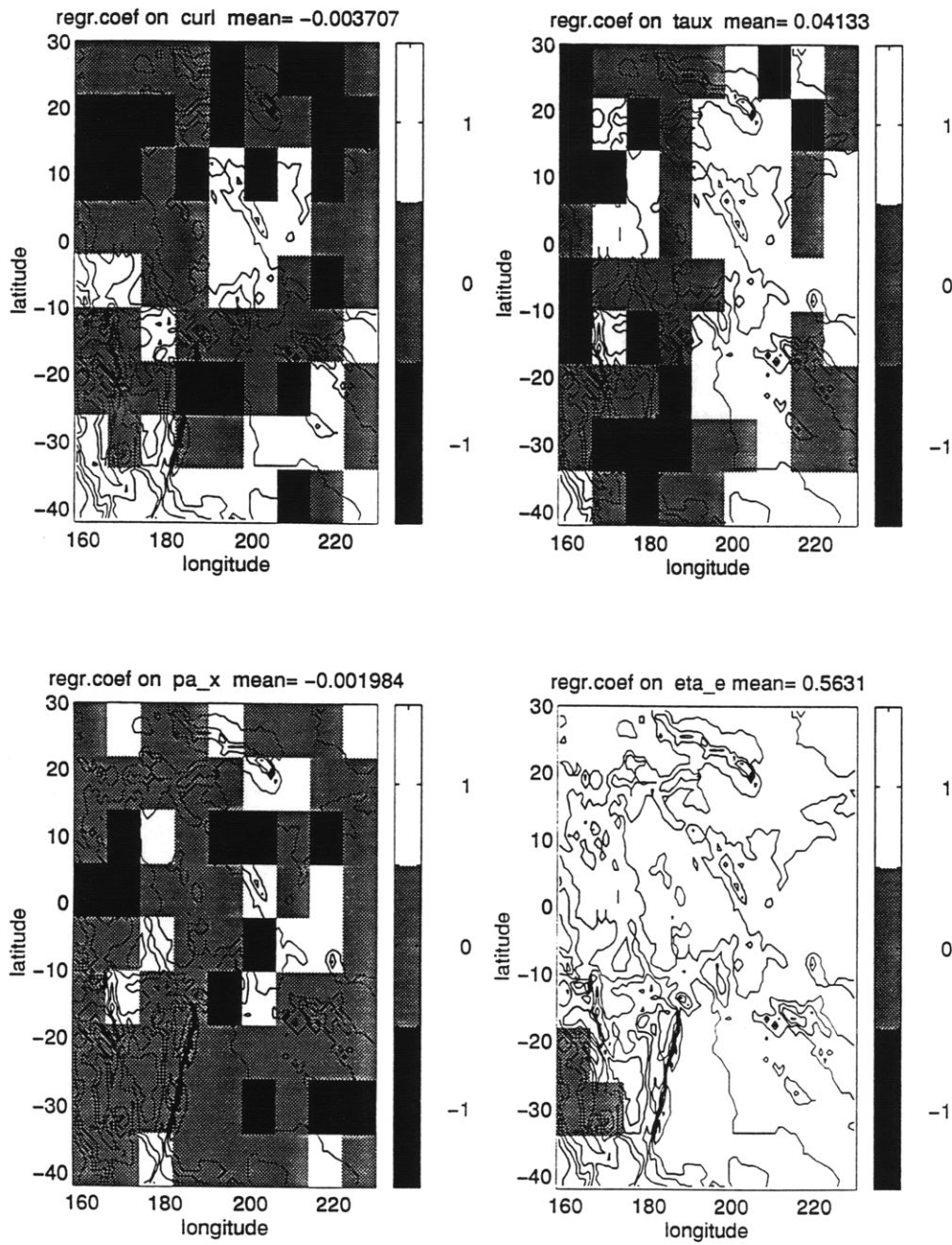


Figure 27: Same as in Figure 15 but for the data averaged in space.

nr\_out: (eta\_x)\*(beta-f\*d\_y/h)+(eta\_y)\*(f\*d\_x/h) inputs: f/rho/g/h.\*curl 1/rho/g/h.\*taux.\*beta zeta\_x.\*(beta-f\*d\_y/h) zeta\_y.

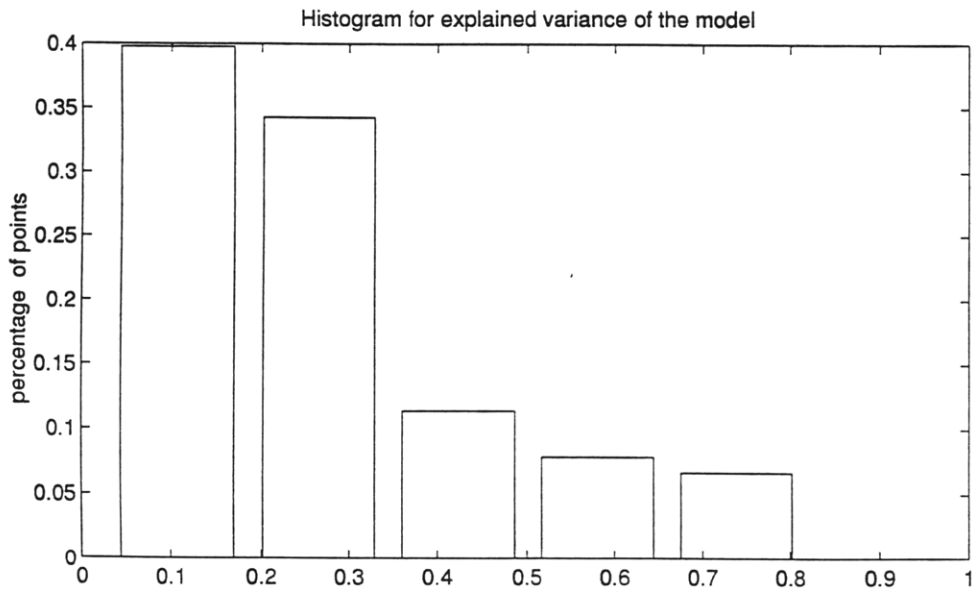
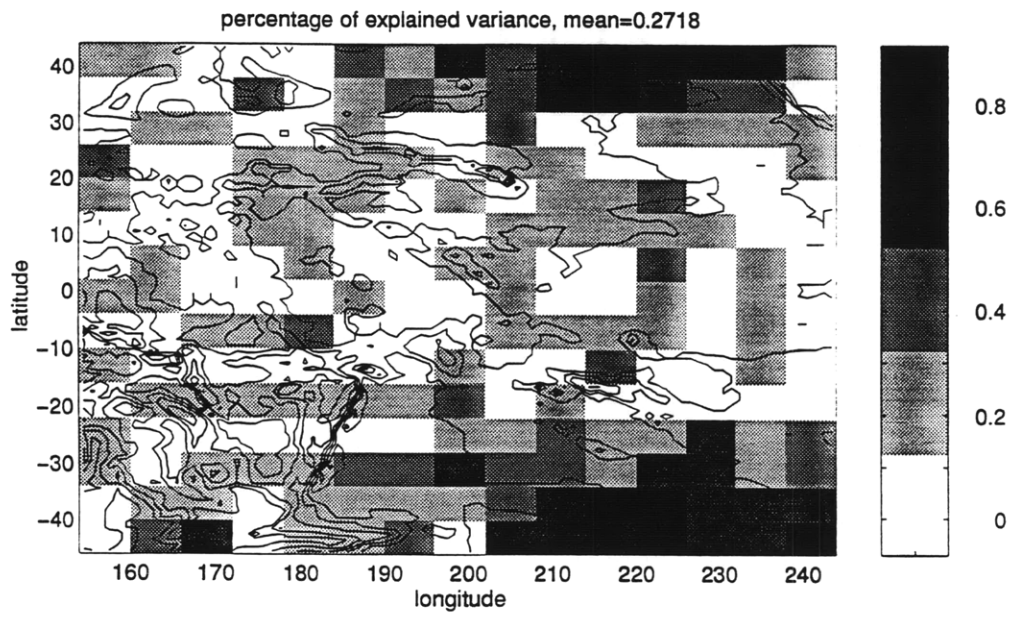
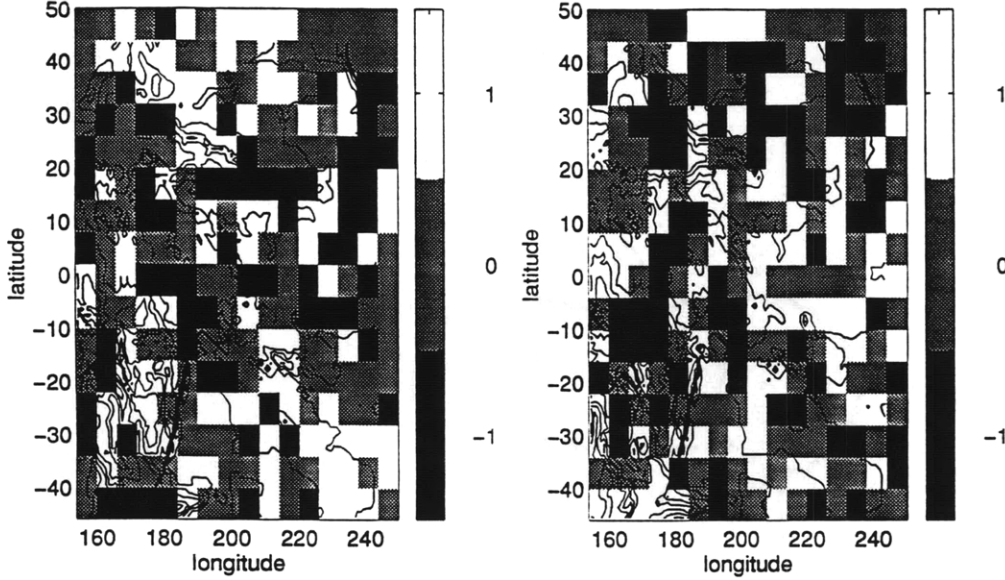


Figure 28: Same as in Figure 14 but for the model 3.18.

output (eta\_x).\*(beta-f\*d\_y/h)+(eta\_y).\*(f\*d\_x/h) 1 --> gt. 1 stdev, -1 --> less (-1) stdev, 0 --> betw. -1 1 s

regr.coef on f/rho/g./h.\*curl mean= 0.0005568 regr.coef on 1/rho/g./h.\*taux.\*beta mean= 0.01172



regr.coef on zeta\_x.\*(beta-f\*d\_y/h) mean= -2.115 regr.coef on zeta\_y.\*(f\*d\_x/h) mean= -0.866

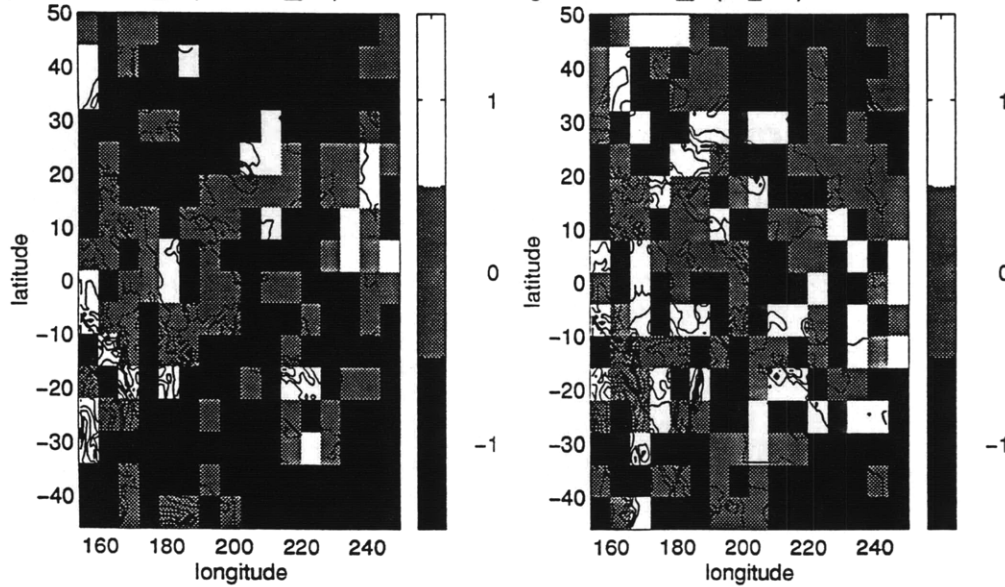


Figure 29: Distribution of regression coefficients for normalized variables for  $(\nabla \times \tau)_z$  (upper left),  $\tau^x$  (upper right),  $\partial\zeta_a/\partial x$  (lower left), and  $\partial\zeta_a/\partial y$  (lower right). The values below minus one standard deviation are denoted by black, above one standard deviation by white, and in between by gray. The underlying contours show bathymetry at the contour spacing of 1 km.

output:  $(\eta_x) \cdot (\beta - f'_d y/h) + (\eta_y) \cdot (f'_d x/h)$  inputs:  $f/\rho/g/h \cdot \text{curl } 1/\rho/g/h \cdot \text{taux} \cdot \beta$

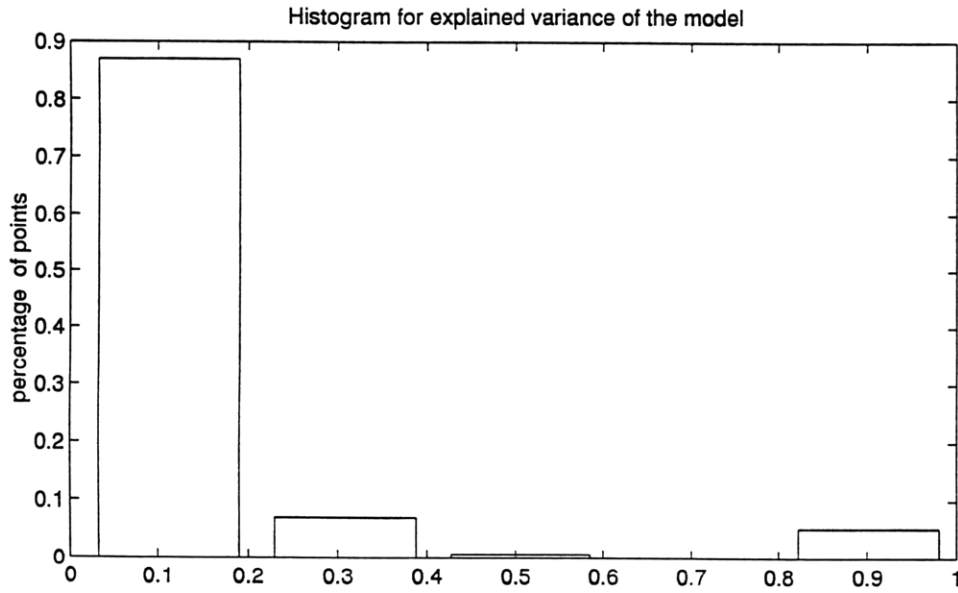
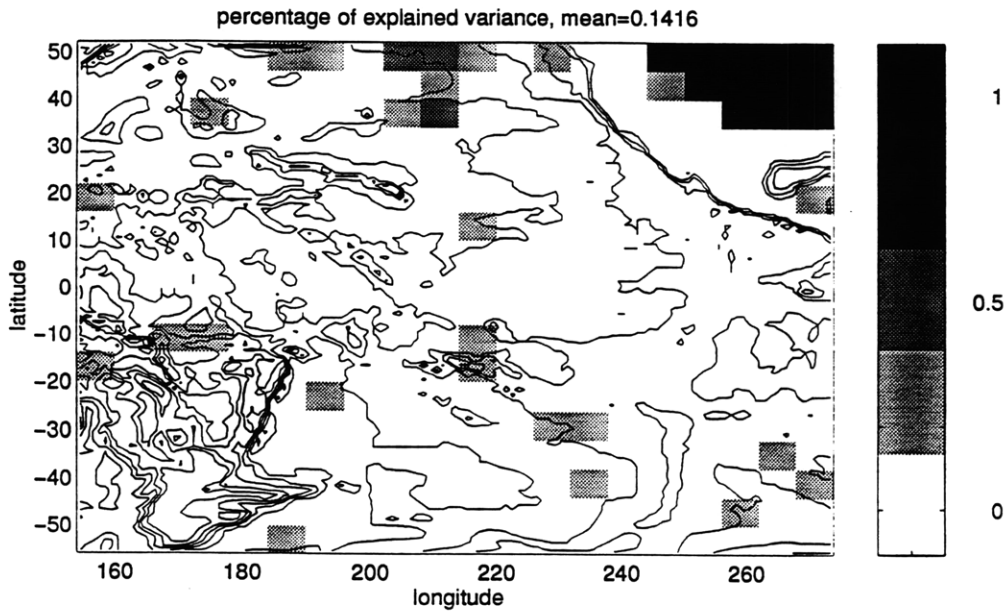


Figure 30: Same as in Figure 14 but for the model 3.19.

output (eta\_x)\*(beta-f\*d\_y/h)+(eta\_y)\*(f\*d\_x/h) 1 --> gt. 1 stdev, -1 --> less (-1) stdev, 0 --> betw. -1 1 s

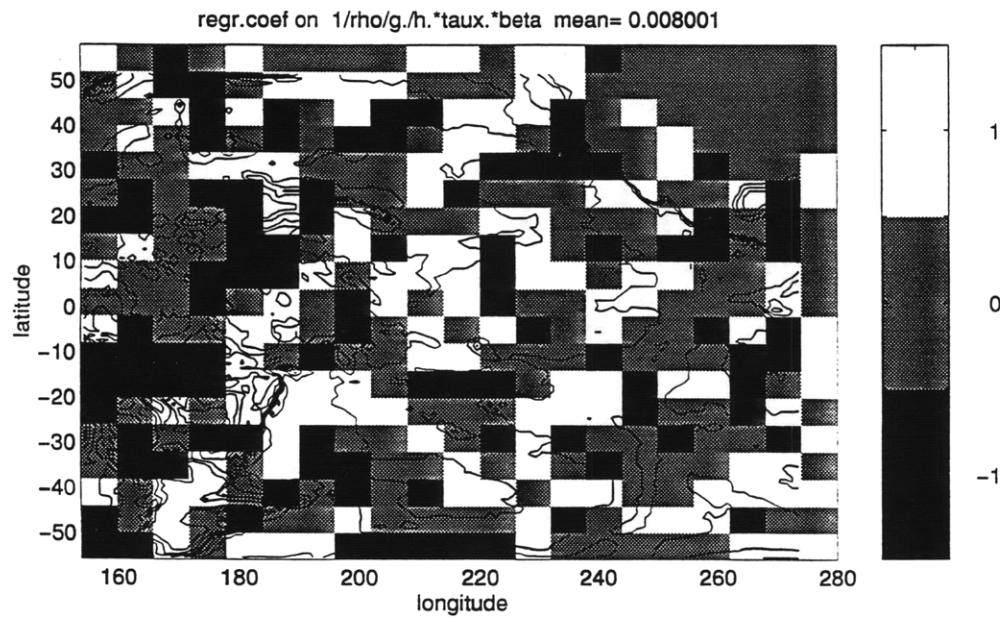
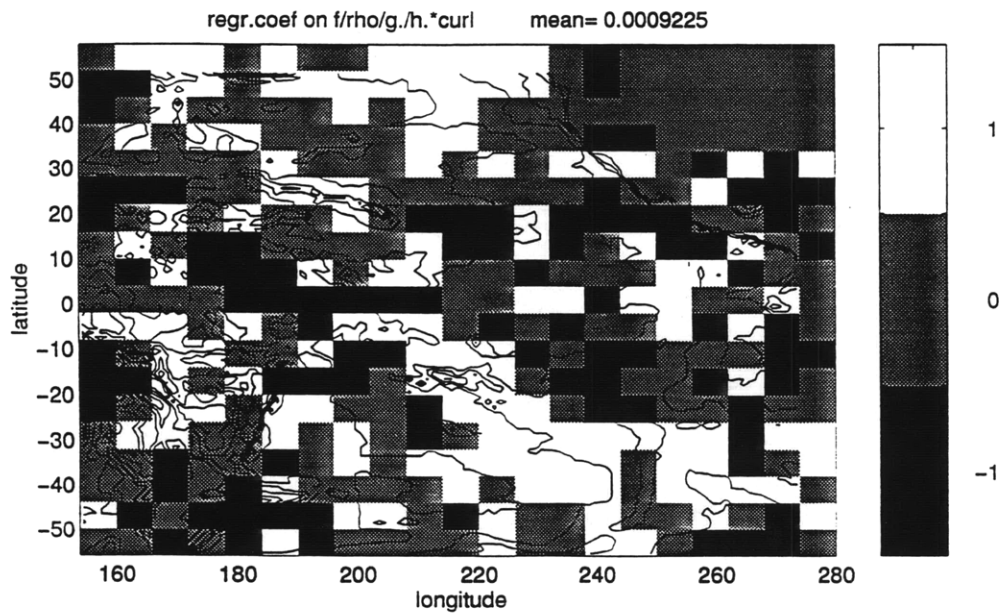


Figure 31: Distribution of regression coefficients for normalized variables for  $(\nabla \times \tau)_z$ (upper) and  $\tau^x$ (lower). The values below minus one standard deviation are denoted by black, above one standard deviation by white, and in between by gray. The underlying contours show bathymetry at the contour spacing of 1 km.

output:  $(\eta_x) \cdot (\beta - f \cdot d_y / h) + (\eta_y) \cdot (f \cdot d_x / h)$  inputs:  $f/g/\rho/h \cdot \text{curl } \beta / \rho/g/h \cdot \tau_{\text{aux}}$

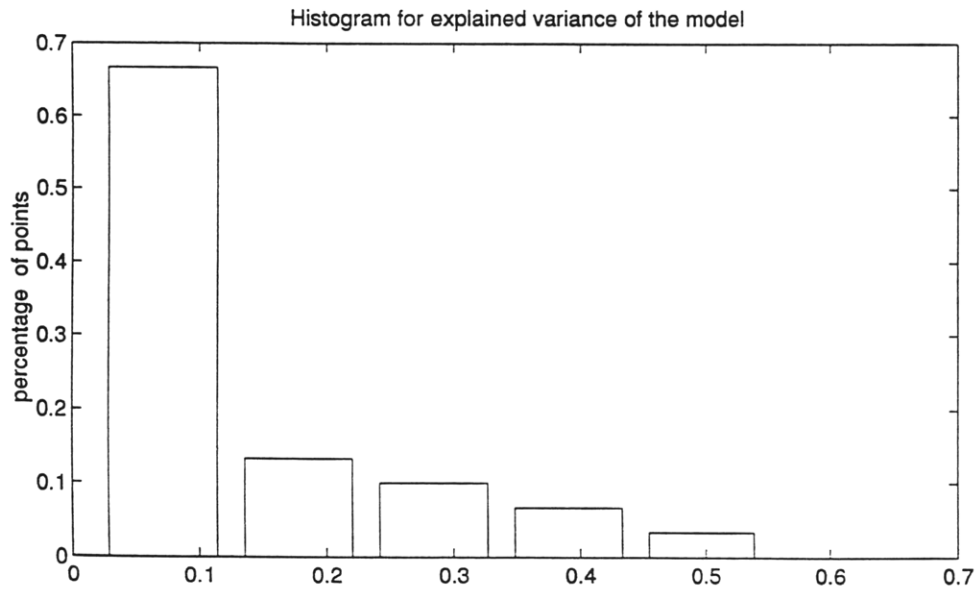
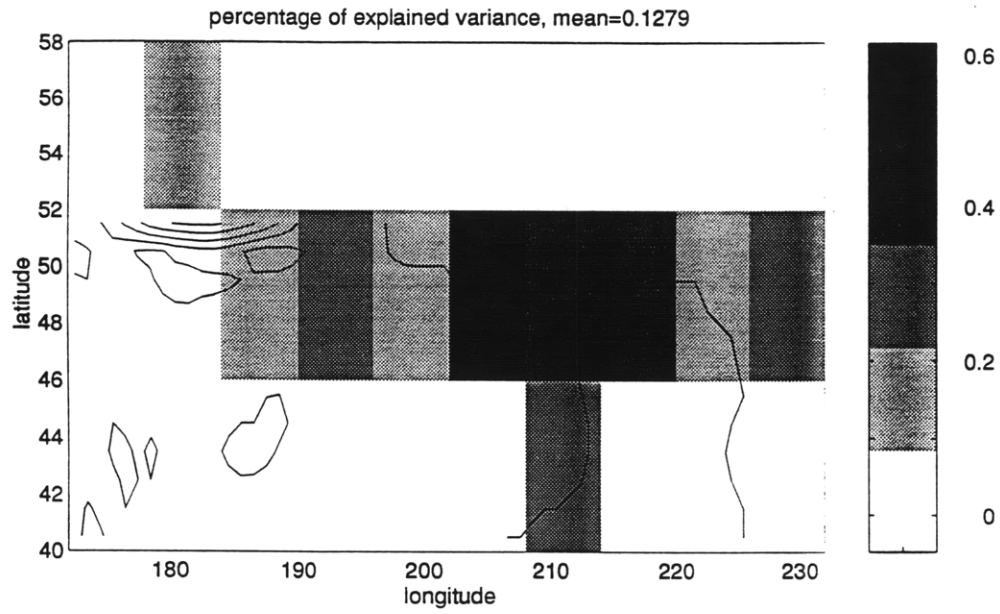


Figure 32: Same as in Figure 30 but for the northeastern Pacific only.

output (eta\_x).\*(beta-f\*d\_y./h)+(eta\_y).\*(f\*d\_x./h) 1 --> gt. 1 stdev, -1 --> less (-1) stdev, 0 --> betw. -1 1 s

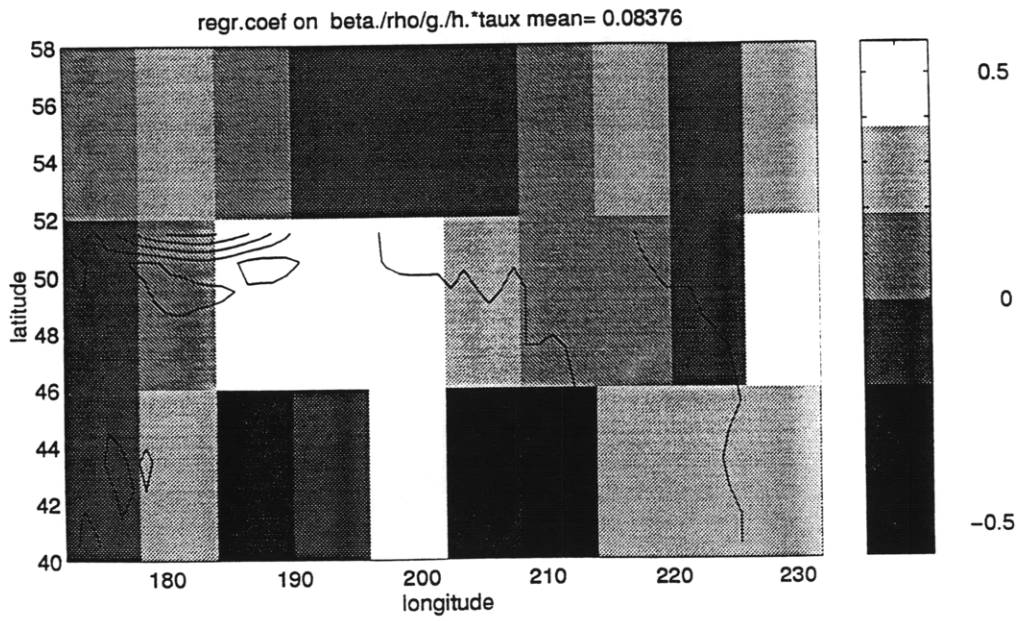
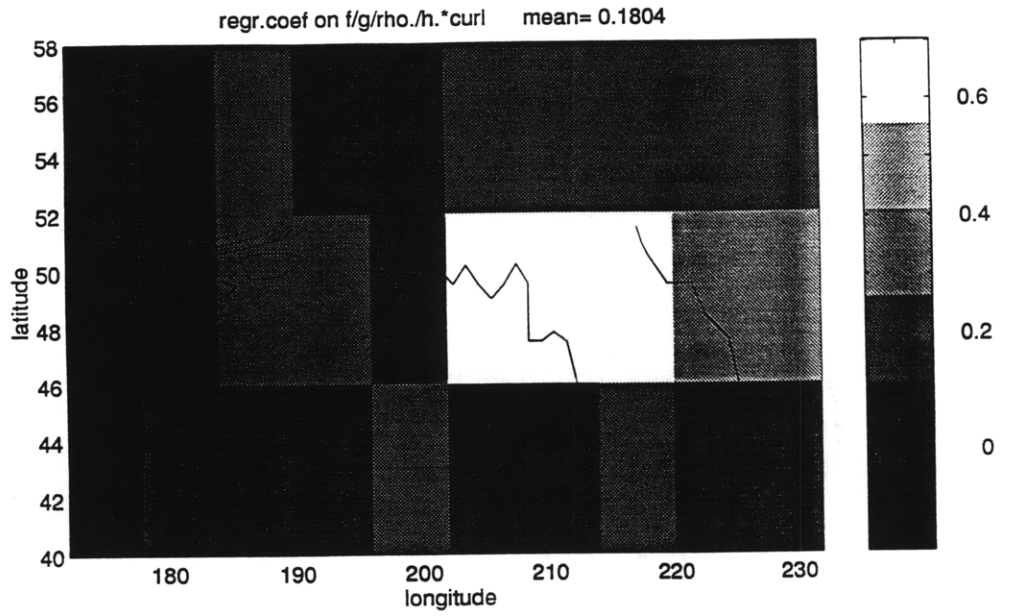
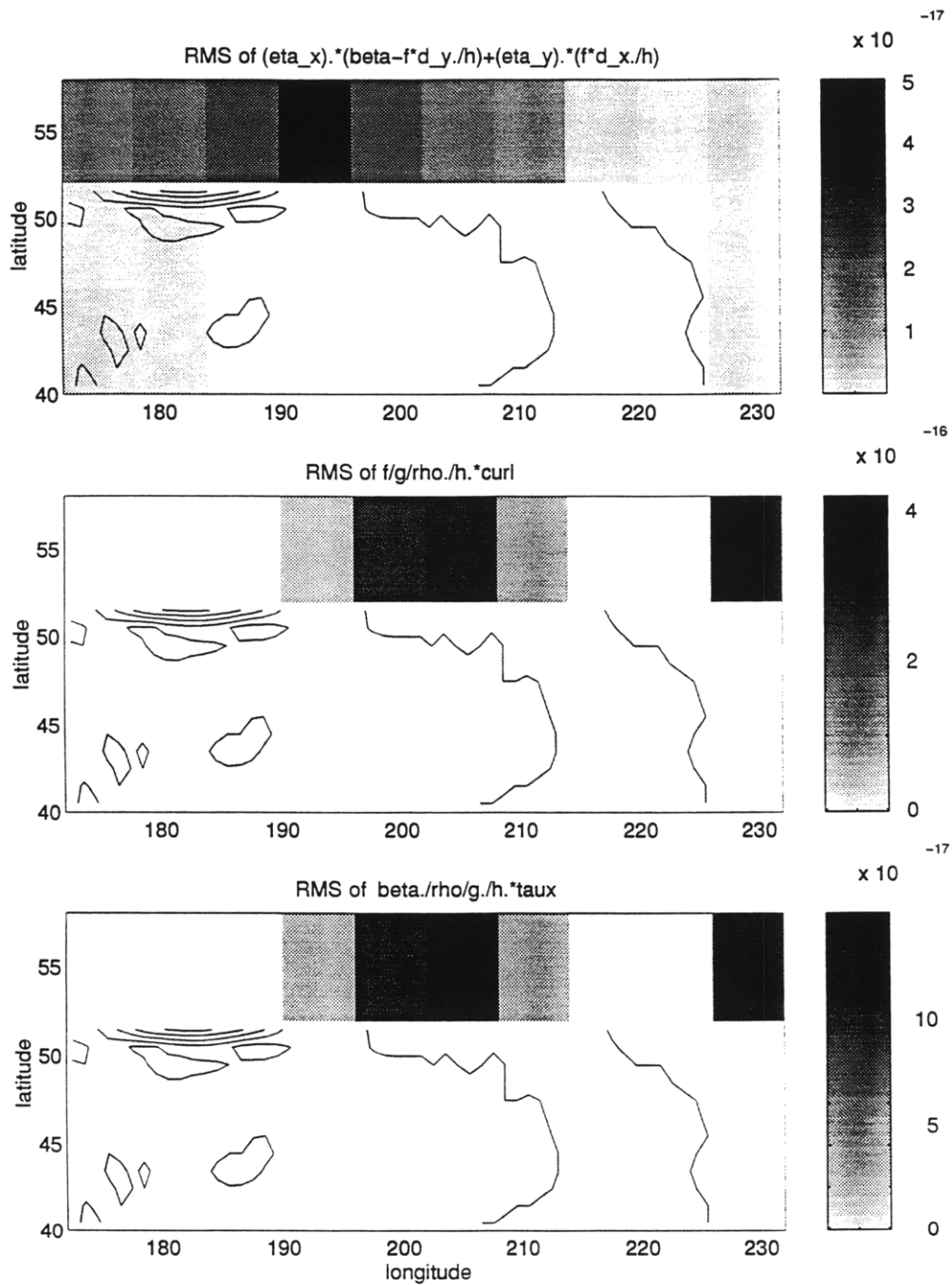


Figure 33: Same as in Figure 31 but for the northeastern Pacific only.



5

2

Figure 34: Contours of root mean square variability for  $\frac{\partial \eta}{\partial x} \left( \frac{\beta}{f} - \frac{h_y}{h} \right) + \frac{\partial \eta}{\partial y} \left( \frac{h_x}{h} \right)$  (upper),  $\frac{f}{\rho g h} (\nabla \times \tau)_z$  (middle), and  $\frac{\beta}{\rho g h} \tau^x$  (lower) in  $m^2$  for the northeastern Pacific.

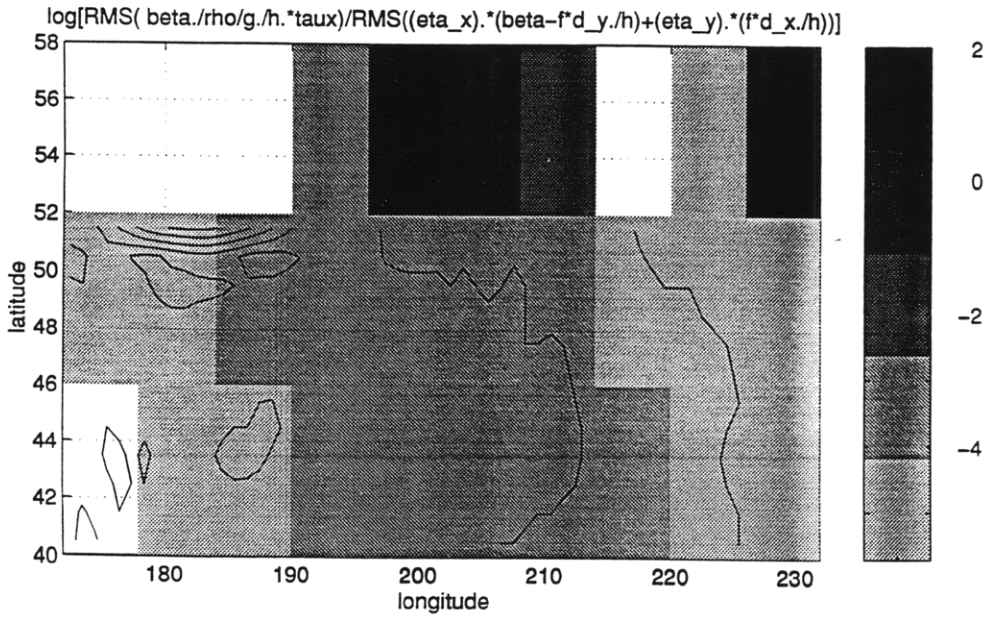
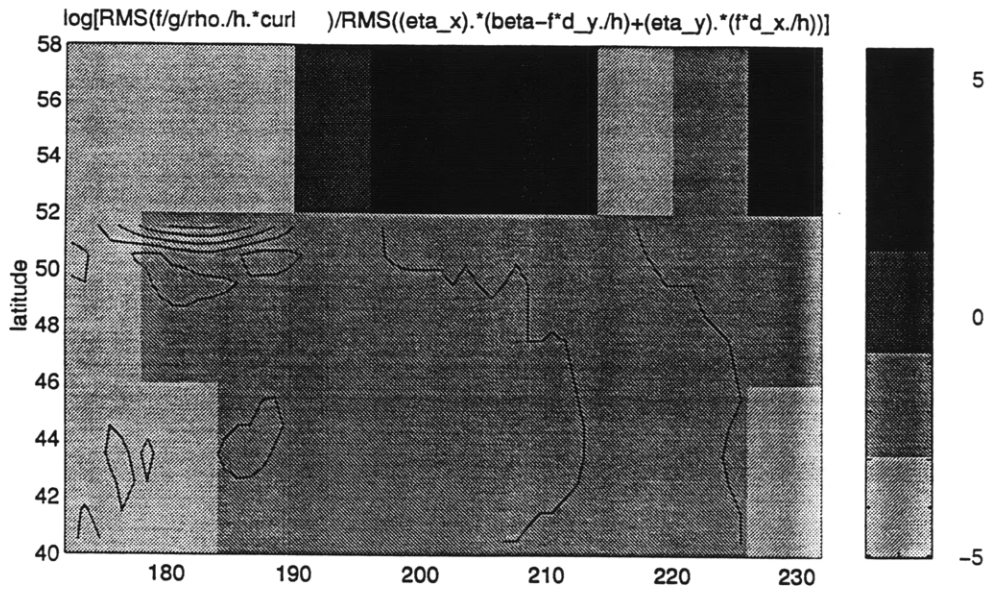


Figure 35: Contours of  $\log$  of  $\text{rms}\left(\frac{f}{\rho_0 g h} (\nabla \times \tau)_z\right)$  over  $\text{rms}\left(\frac{\partial \eta}{\partial x} \left(\frac{\beta}{f} - \frac{h_y}{h}\right) + \frac{\partial \eta}{\partial y} \left(\frac{h_x}{h}\right)\right)$  (upper) and  $\log$  of  $\text{rms}\left(\frac{\beta}{\rho_0 g h} \tau_x\right)$  over  $\text{rms}\left(\frac{\partial \eta}{\partial x} \left(\frac{\beta}{f} - \frac{h_y}{h}\right) + \frac{\partial \eta}{\partial y} \left(\frac{h_x}{h}\right)\right)$  (lower) for the northeastern Pacific.

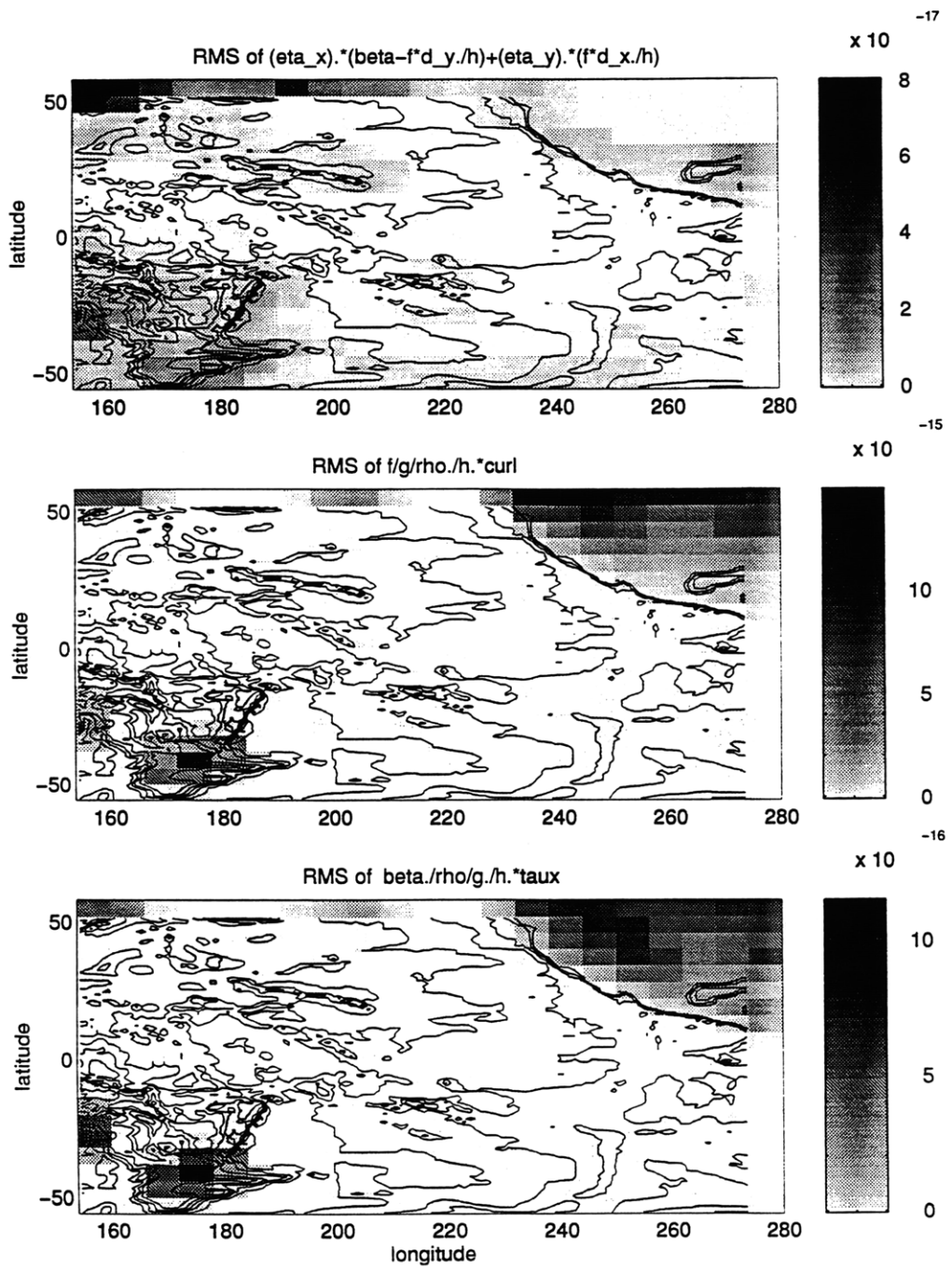


Figure 36: Same as in Figure 34 for all of the Pacific.

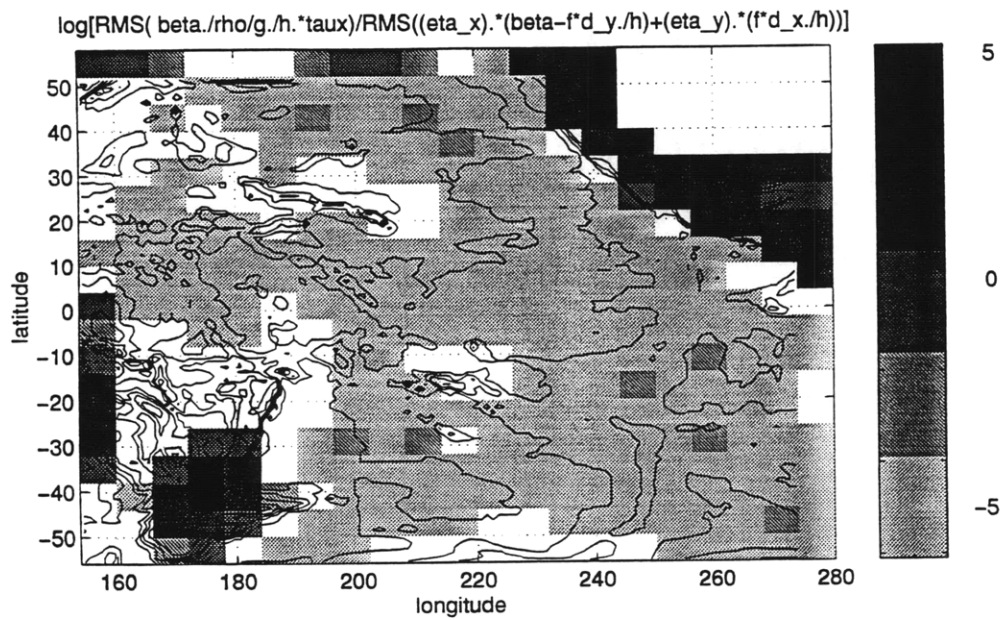
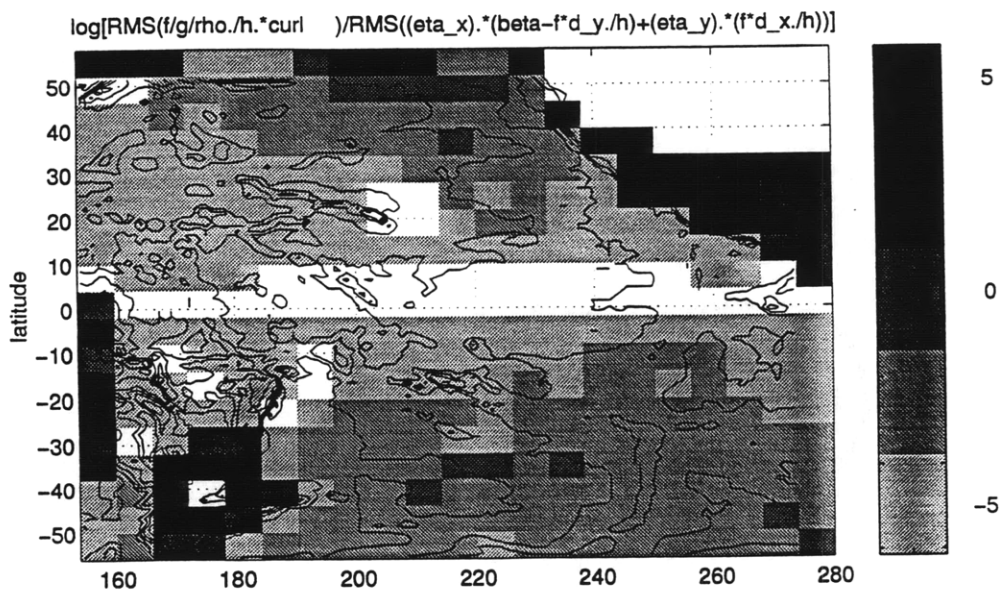


Figure 37: Same as in Figure 35 for all of the Pacific.

output: laplace\_eta\_t+(eta\_x).\*(beta-f\*d\_y/h)+(eta\_y).\*(f\*d\_x/h) inputs: t/rho/g./h.\*curl

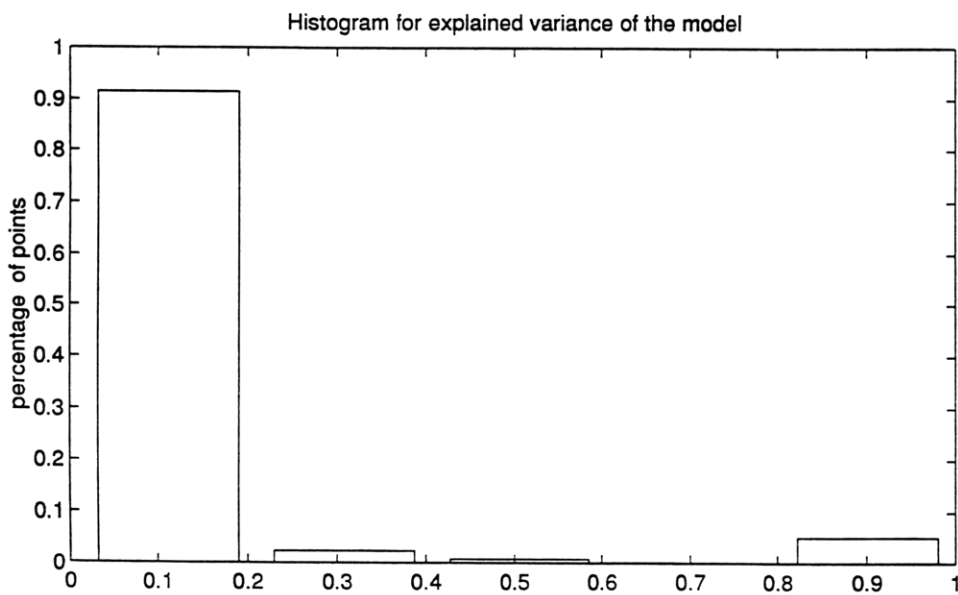
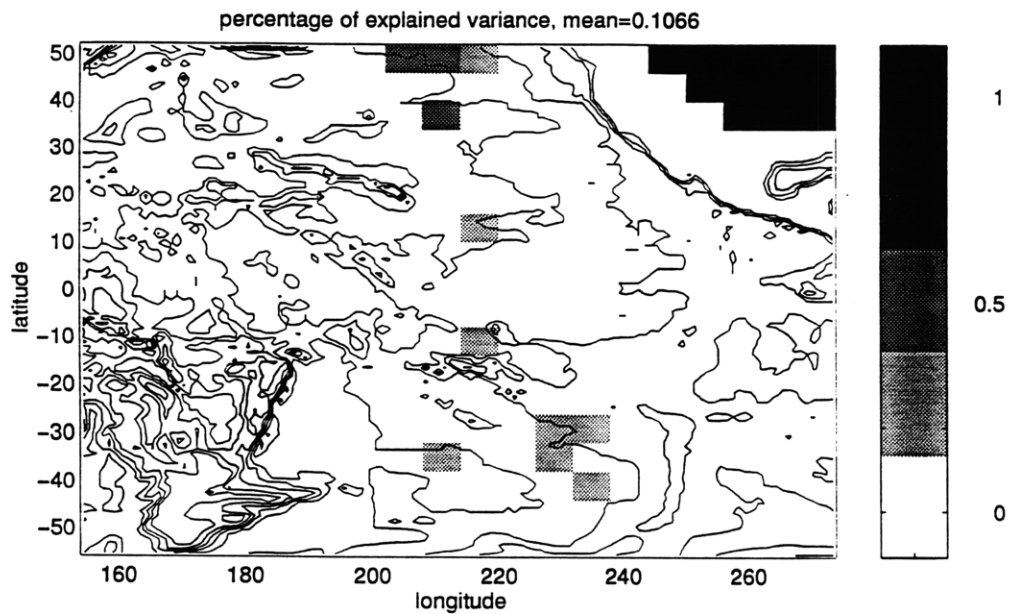


Figure 38: Same as in Figure 14 but for the model 3.20 without the  $\tau^x$  term.

output laplace\_eta\_t+(eta\_x).\*(beta-f\*d\_y/h)+(eta\_y).\*(f\*d\_x/h) 1 --> gt. 1 stdev, -1 --> less (-1) stdev, 0 --

regr.coef on f/rho/g./h.\*curl mean= 0.001451

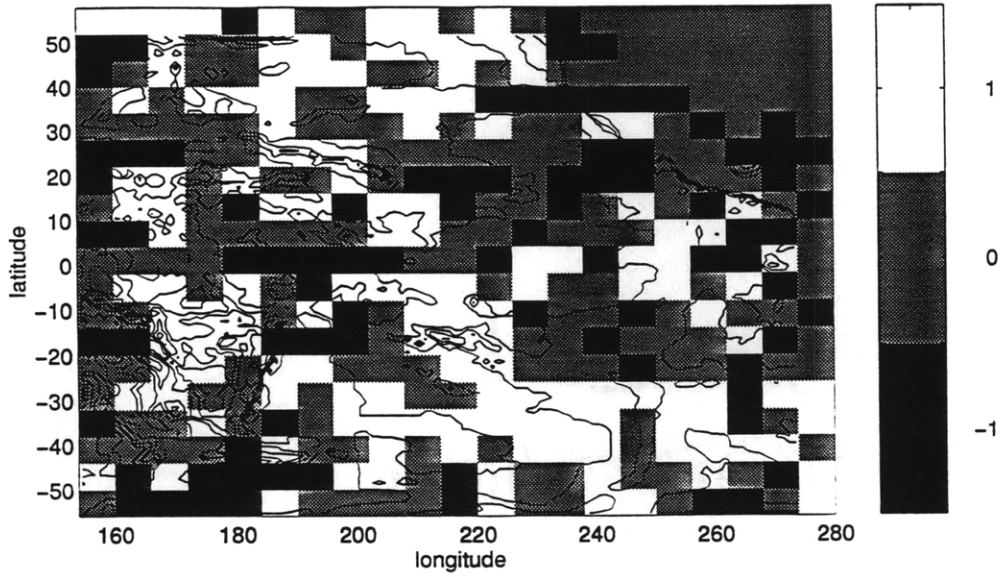


Figure 39: Distribution of regression coefficients for normalized variables for  $(\nabla \times \tau)_z(\text{upper})$ . The values below minus one standard deviation are denoted by black, above one standard deviation by white, and in between by gray. The underlying contours show bathymetry at the contour spacing of 1 km.

$\log[\text{RMS}(f/\rho_0/g./h.*\text{curl})/\text{RMS}(\text{laplace\_eta\_t}+(\text{eta\_x}).*(\text{beta}-f*d\_y./h)+(\text{eta\_y}).*(f*d\_x./h))]$

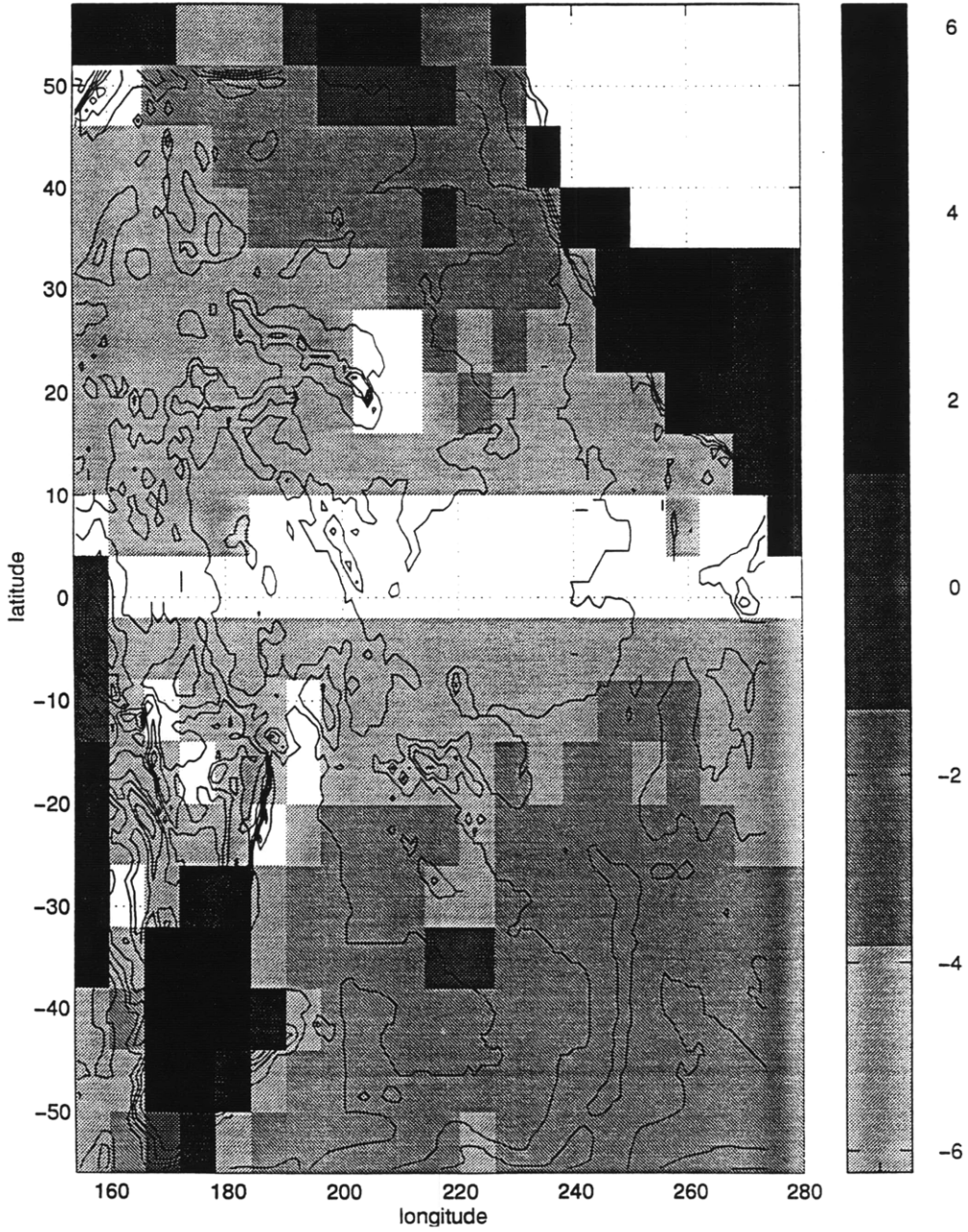


Figure 40: Contours of  $\log$  of  $\text{rms}(\frac{f}{\rho_0 g h} (\nabla \times \tau)_z)$  over  $\text{rms}(\frac{\partial}{\partial t} \nabla^2 \eta + (\beta - f \frac{h_y}{h}) \frac{\partial \eta}{\partial x} + (f \frac{h_x}{h}) \frac{\partial \eta}{\partial y})$  for the Pacific.

output:  $\text{laplace\_eta\_t} + (\text{eta\_x}) \cdot (\text{beta} - f^*d\_y/h) + (\text{eta\_y}) \cdot (f^*d\_x/h)$  inputs:  $f/g/\rho/h \cdot \text{curl } \text{beta}/\rho/g/h \cdot \text{taux}$

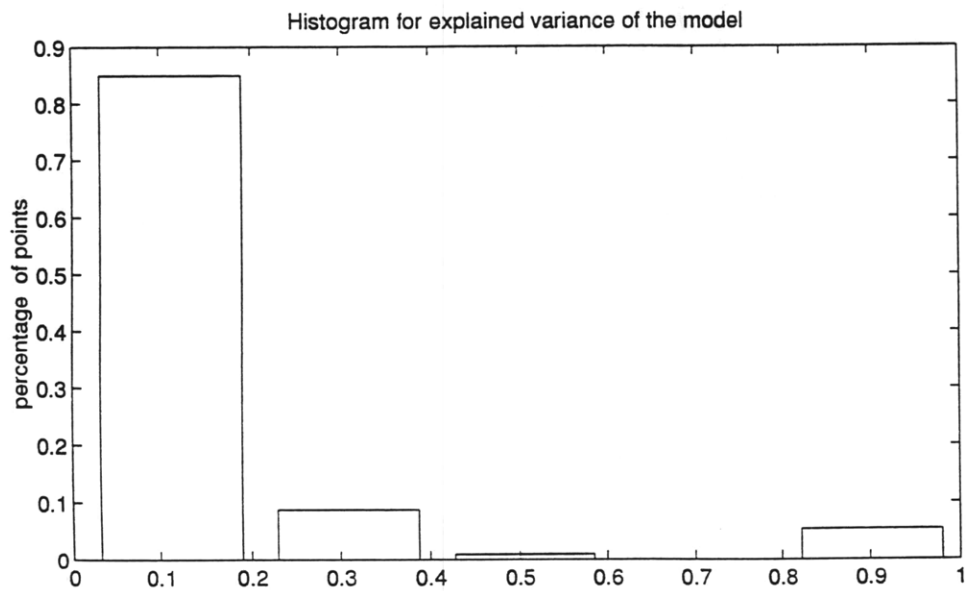
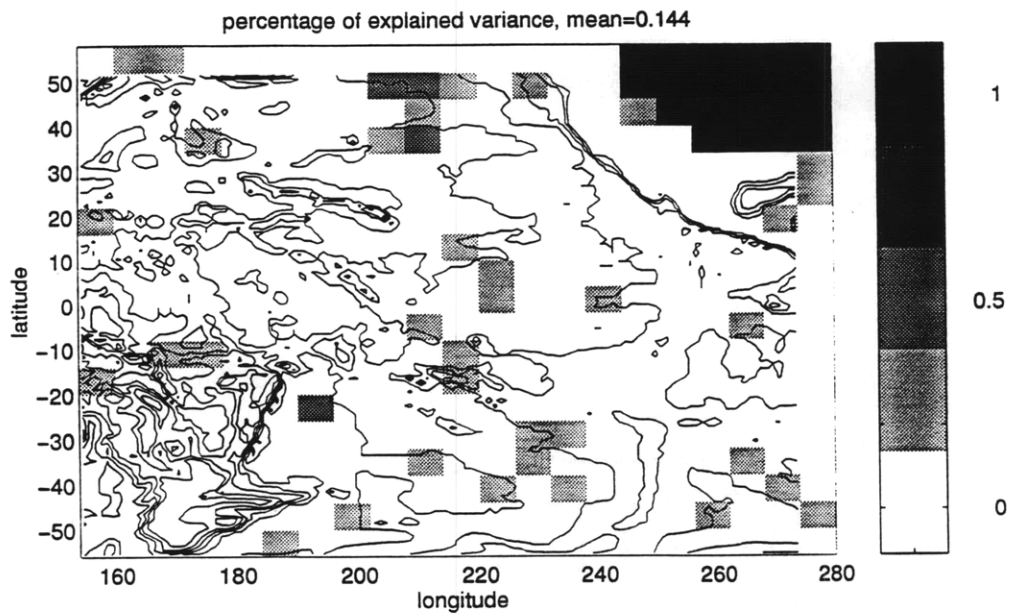
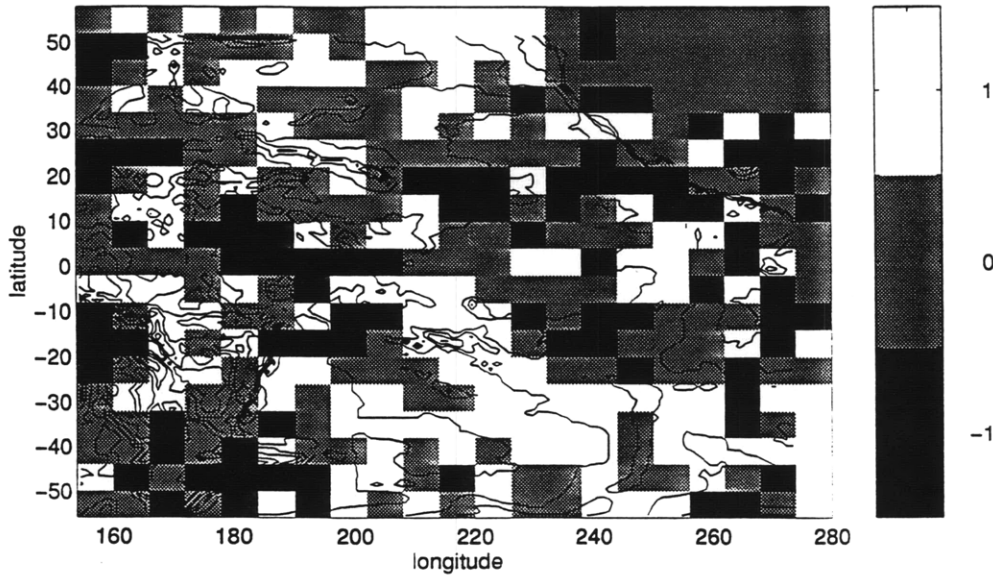


Figure 41: Same as in Figure 14 but for the model 3.20.

output laplace\_eta\_t+(eta\_x).\*(beta-f\*d\_y/h)+(eta\_y).\*(f\*d\_x/h) 1 --> gt. 1 stdev, -1 --> less (-1) stdev, 0 --

regr.coef on f/g/rho./h.\*curl mean= 0.03754



regr.coef on beta./rho/g./h.\*taux mean= 0.02147

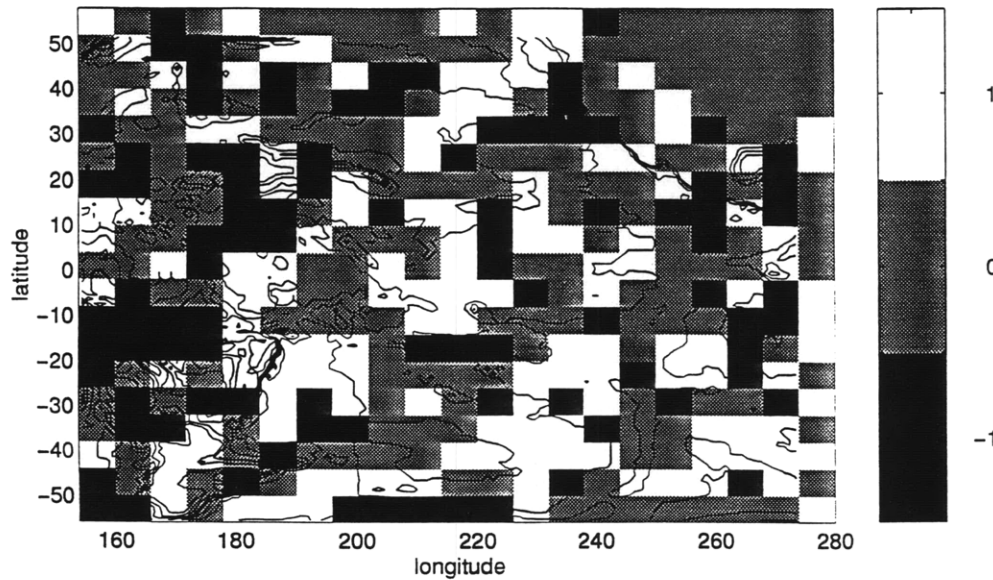


Figure 42: Distribution of regression coefficients for normalized variables for  $(\nabla \times \tau)_z$ (upper) and  $\tau^x$ (lower). The values below minus one standard deviation are denoted by black, above one standard deviation by white, and in between by gray. The underlying contours show bathymetry at the contour spacing of 1 km.

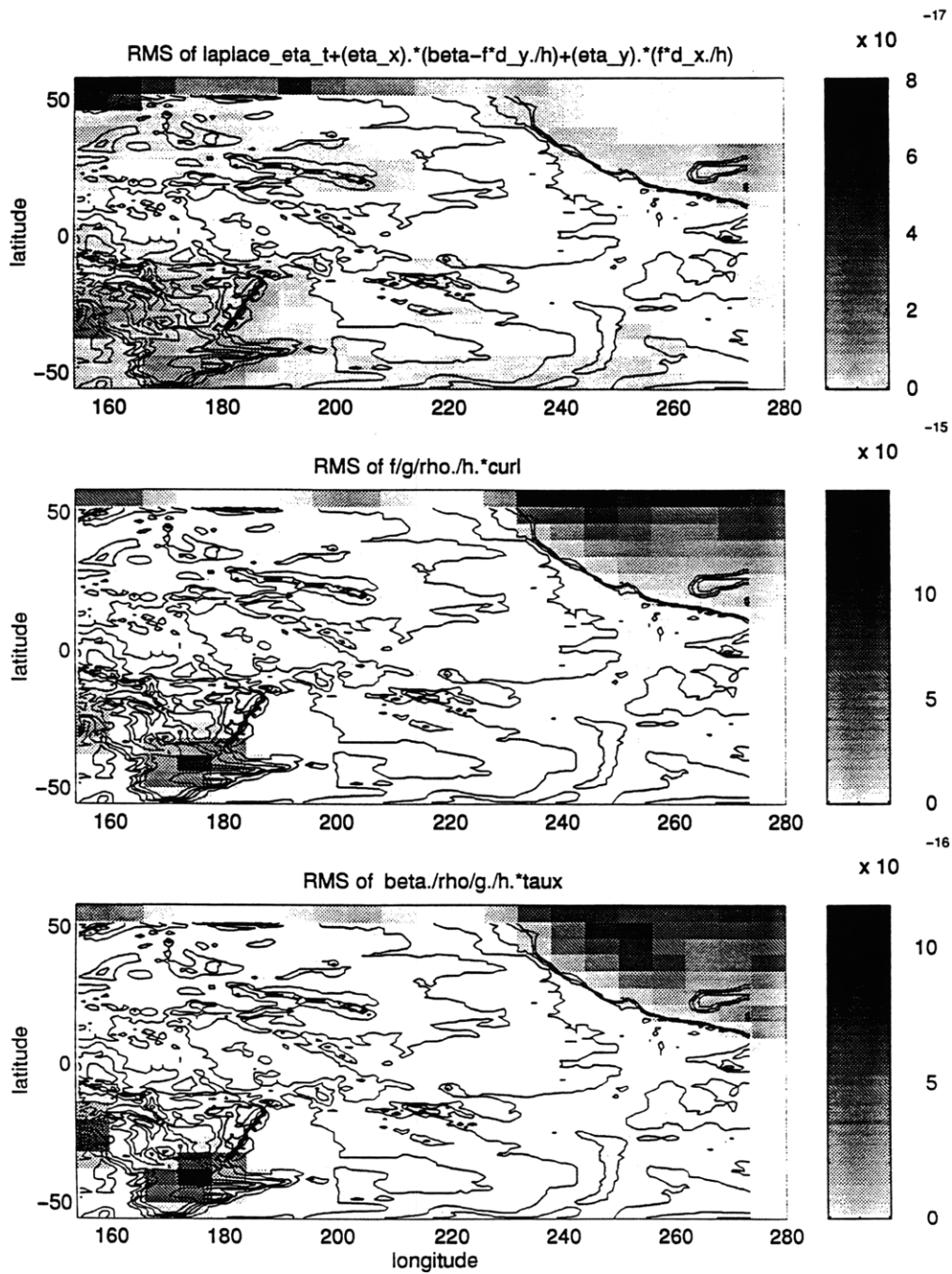


Figure 43: Contours of root mean square variability for  $\frac{\partial}{\partial t} \nabla^2 \eta + \left( \beta - f \frac{h_y}{h} \right) \frac{\partial \eta}{\partial x} + \left( f \frac{h_x}{h} \right) \frac{\partial \eta}{\partial y}$  (upper),  $\frac{f}{\rho g h} (\nabla \times \tau)_z$  (middle), and  $\frac{\beta}{\rho g h} \tau^x$  (lower) in the Pacific.

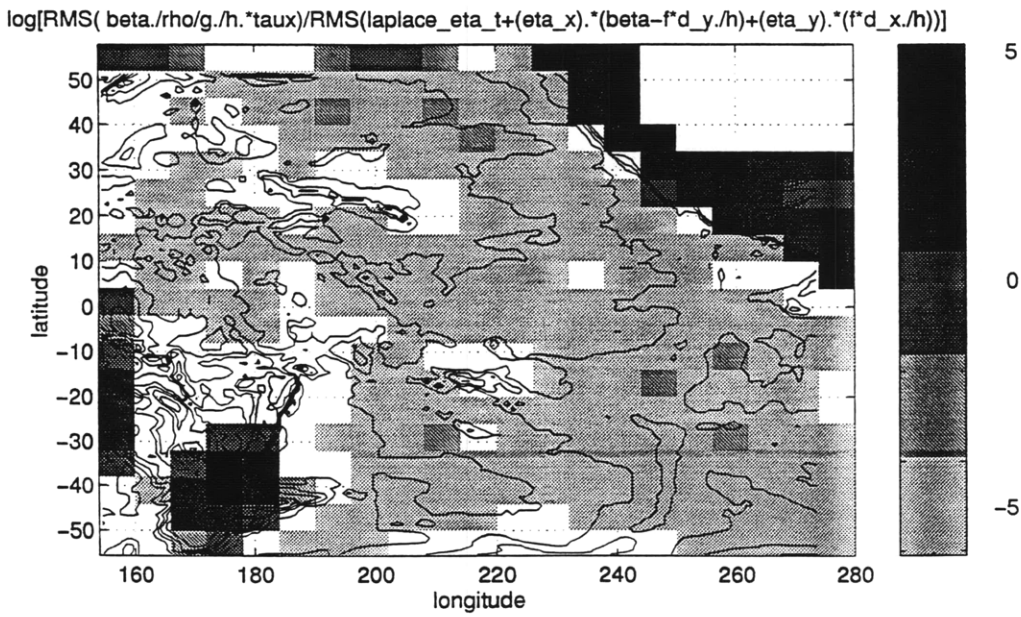
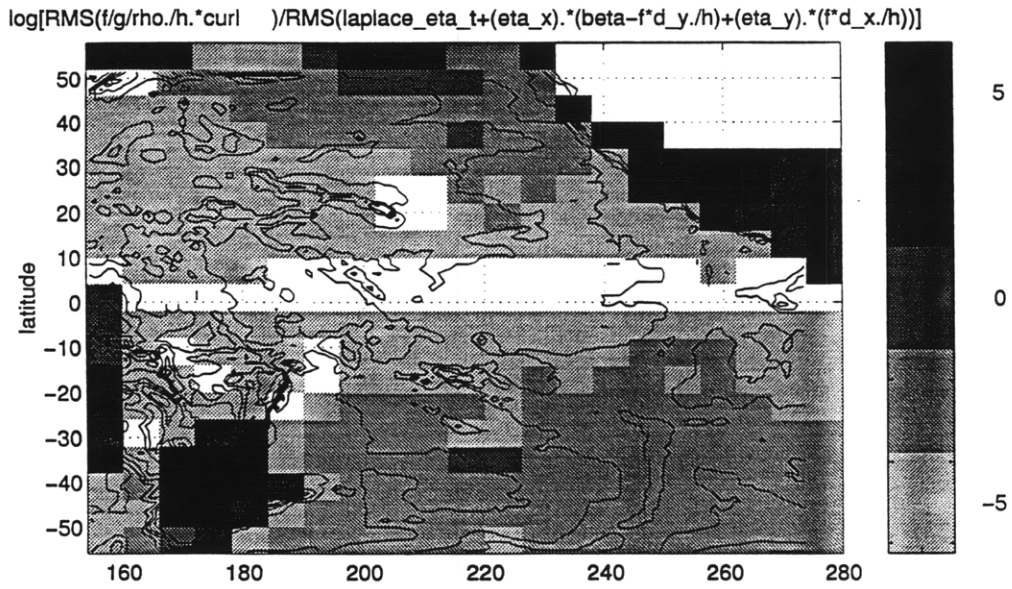


Figure 44: Contours of  $\log$  of  $\text{rms}(\frac{f}{\rho_0 g h} (\nabla \times \tau)_z)$  over  $\text{rms}(\frac{\partial}{\partial t} \nabla^2 \eta + (\beta - f \frac{h_y}{h}) \frac{\partial \eta}{\partial x} + (f \frac{h_x}{h}) \frac{\partial \eta}{\partial y})$  and  $\log$  of  $\text{rms}(\frac{\beta}{\rho_0 g h} \tau^x)$  over  $\text{rms}(\frac{\partial}{\partial t} \nabla^2 \eta + (\beta - f \frac{h_y}{h}) \frac{\partial \eta}{\partial x} + (f \frac{h_x}{h}) \frac{\partial \eta}{\partial y})$  for the Pacific.

output: laplace\_eta inputs:  $-\left((\eta_x) \cdot (\beta - f'_d y/h) + (\eta_y) \cdot (f'_d x/h)\right) / g/\rho/h \cdot \text{curl } \beta / \rho/g/h \cdot \tau_{\text{aux}}$

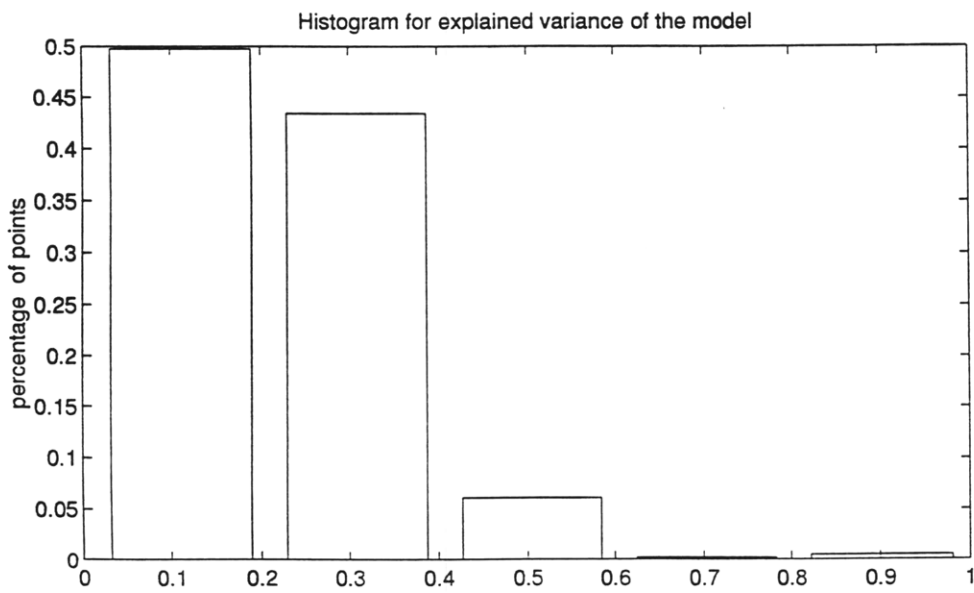
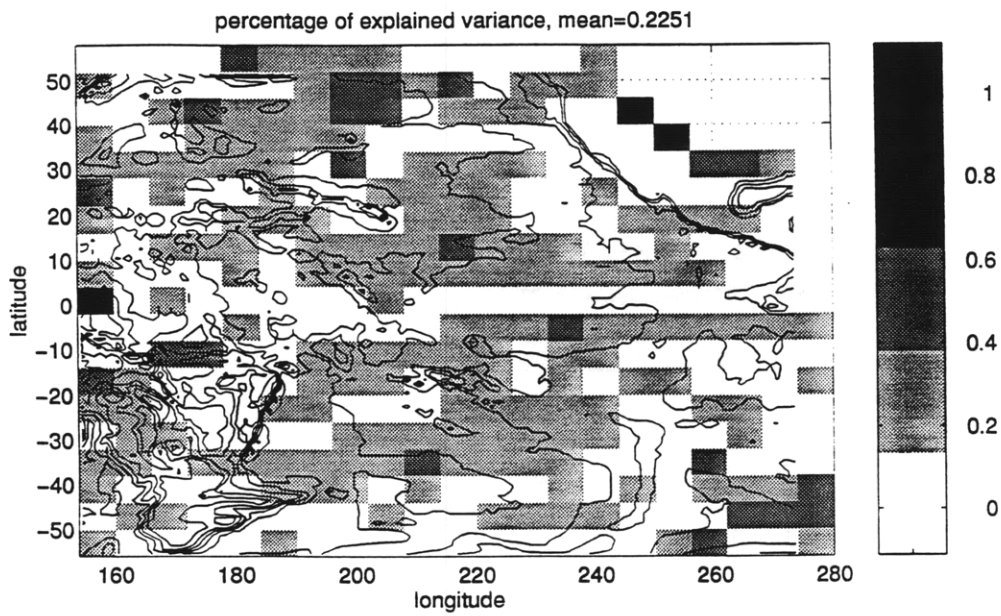


Figure 45: Same as in Figure 14 but for the model 3.21.

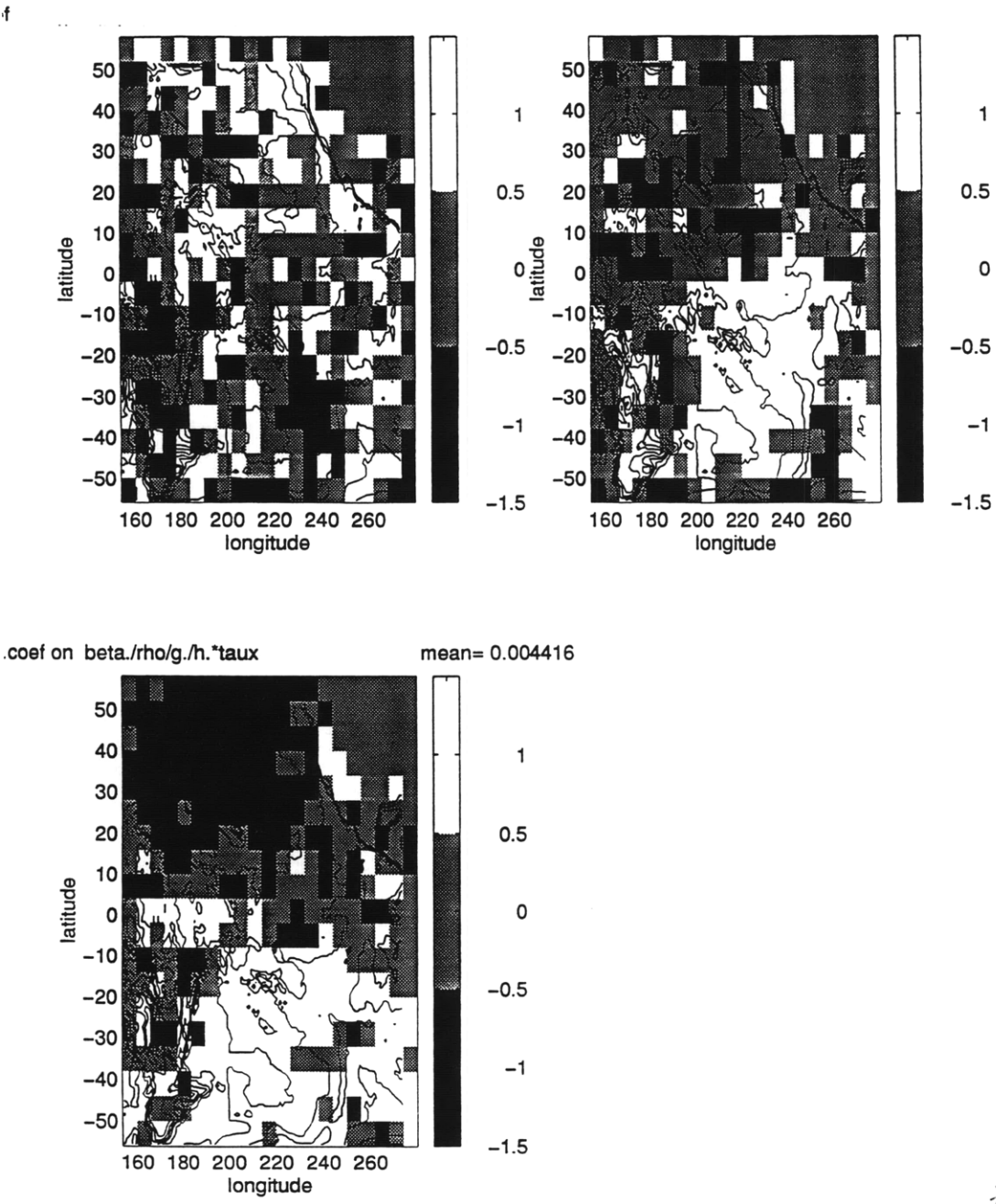


Figure 46: Distribution of regression coefficients for normalized variables for  $-\left(\left(\beta - f\frac{h_y}{h}\right)\frac{\partial\eta}{\partial x} + f\frac{h_x}{h}\frac{\partial\eta}{\partial y}\right)$  (upper left),  $(\nabla \times \tau)_z$ (upper right), and  $\tau^x$ (lower). The values below minus one standard deviation are denoted by black, above one standard deviation by white, and in between by gray. The underlying contours show bathymetry at the contour spacing of 1 km.

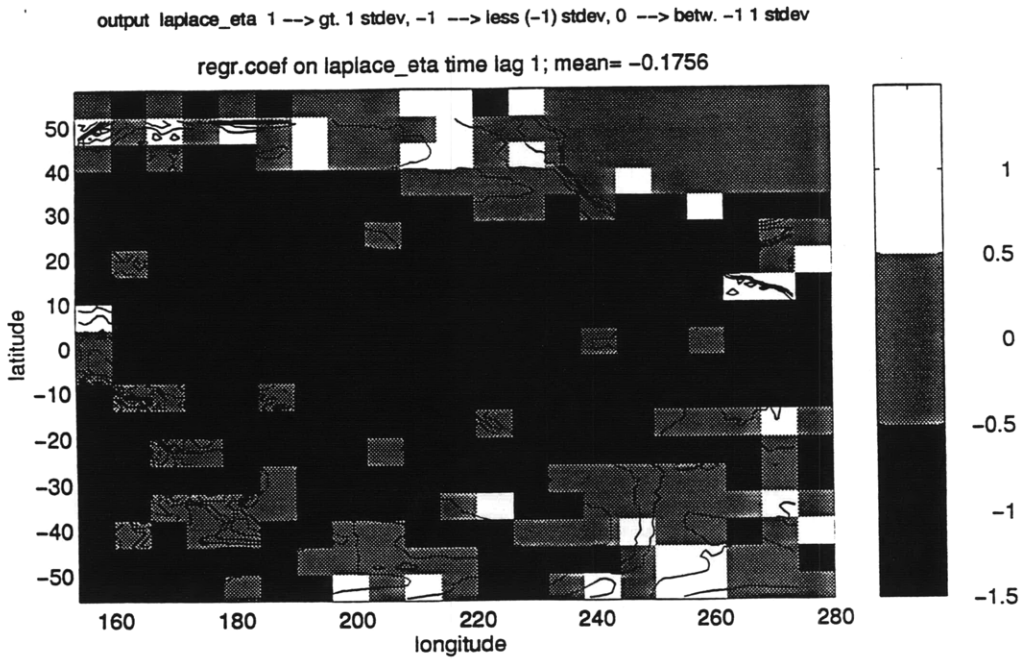


Figure 47: Distribution of regression coefficient on  $\nabla^2\eta(i, j, t - 1)$ .

$\alpha \cdot \text{put: laplace\_eta inputs: } -((\text{eta\_x}) \cdot (\text{beta} - f \cdot d\_y / h) + (\text{eta\_y}) \cdot (f \cdot d\_x / h)) / g / \rho / h \cdot \text{curl beta} / \rho / g / h \cdot \text{tau}$

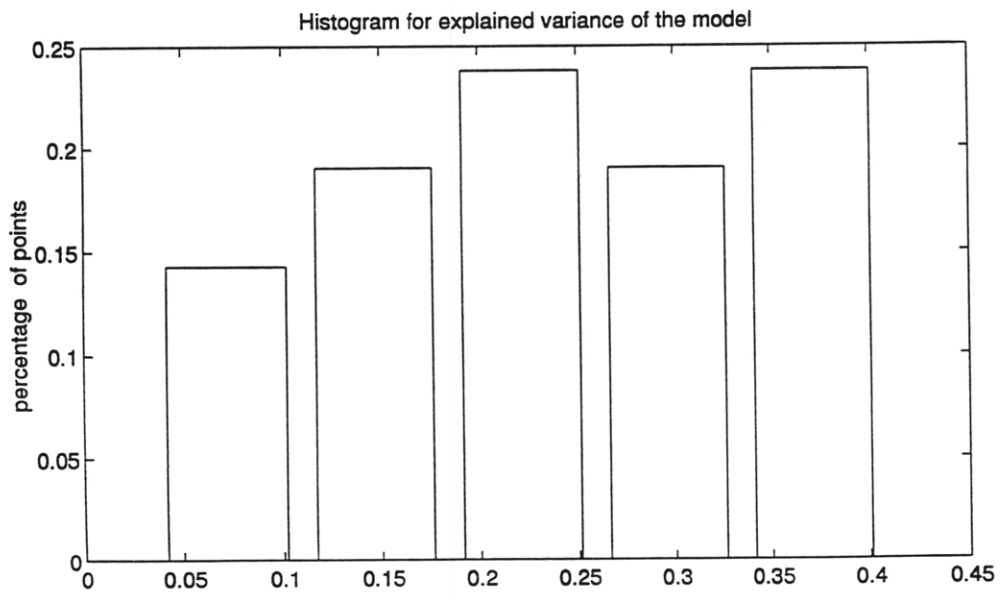
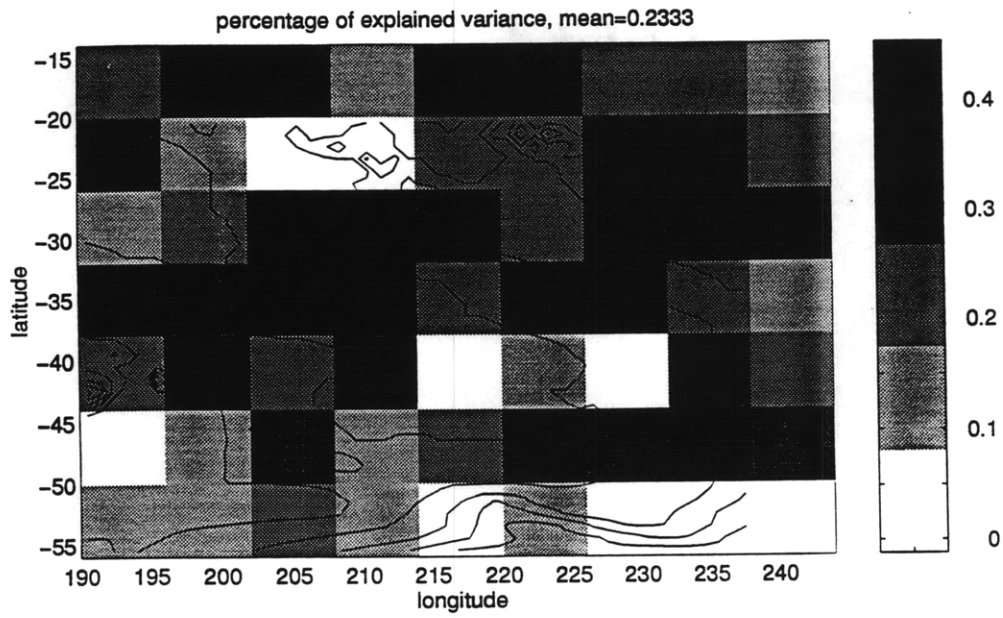


Figure 48: Same as in Figure 45 but for the south eastern Pacific.

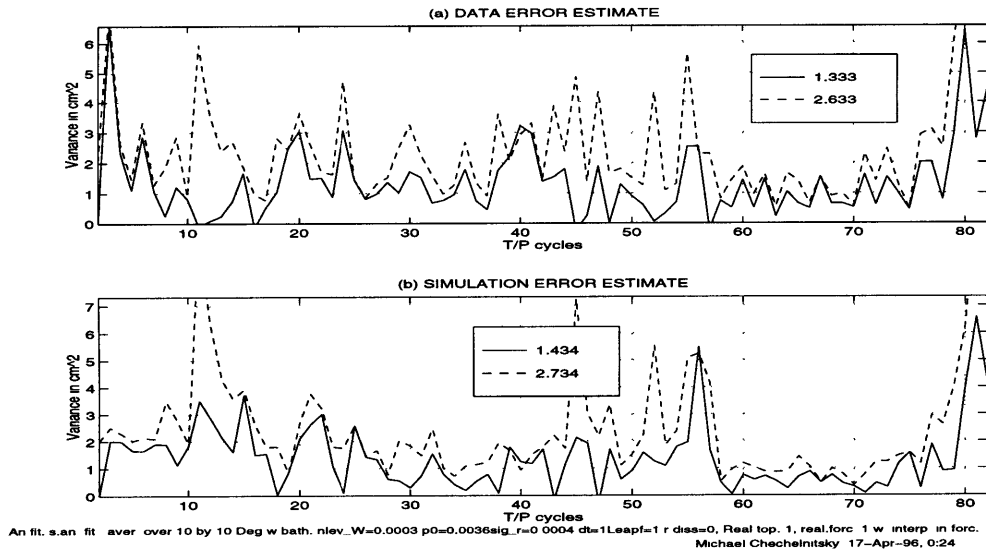


Figure 49: Estimates of the mean SSH variance ( $\text{cm}^2$ ) computed for each T/P cycle as an initial condition showing (a) data error and (b) simulation error. The dashed curves are the data and simulation variance, respectively.

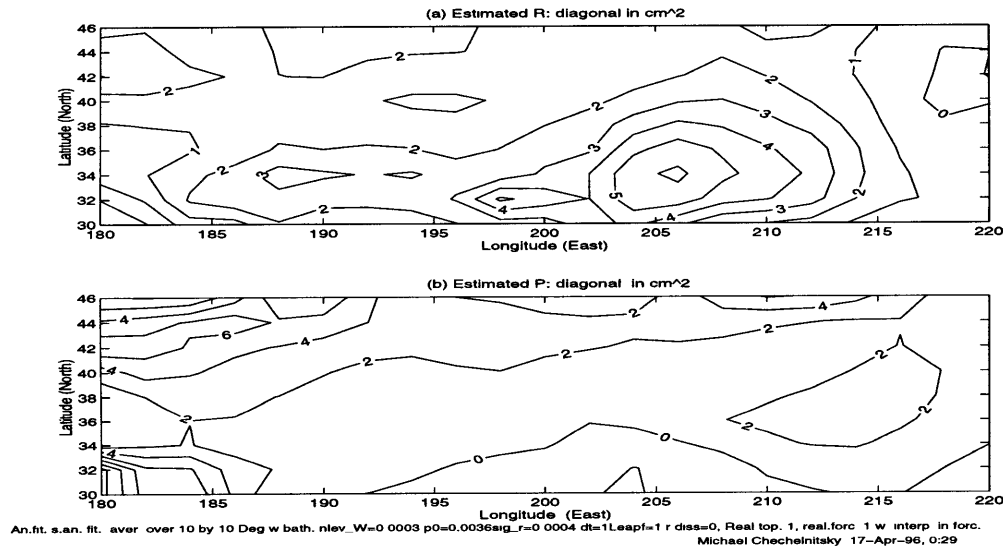


Figure 50: Contours of the estimate of (a) data error covariance  $R$  and (b) simulation error covariance  $EPE^T$ .

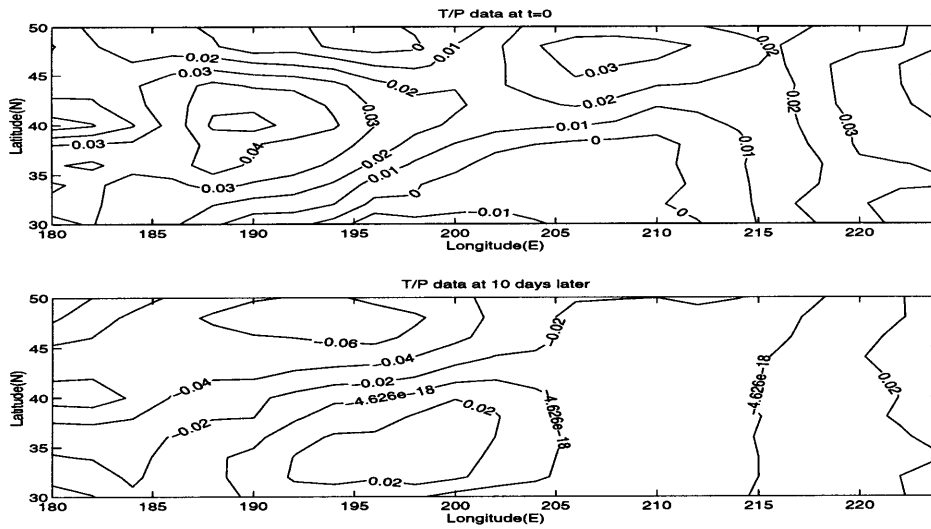


Figure 51: Contours of the SSH at t=0 (2nd T/P cycle) and 10 days later (3rd T/P cycle) from the data. Annual, semi-annual cycles, and time mean have been removed.

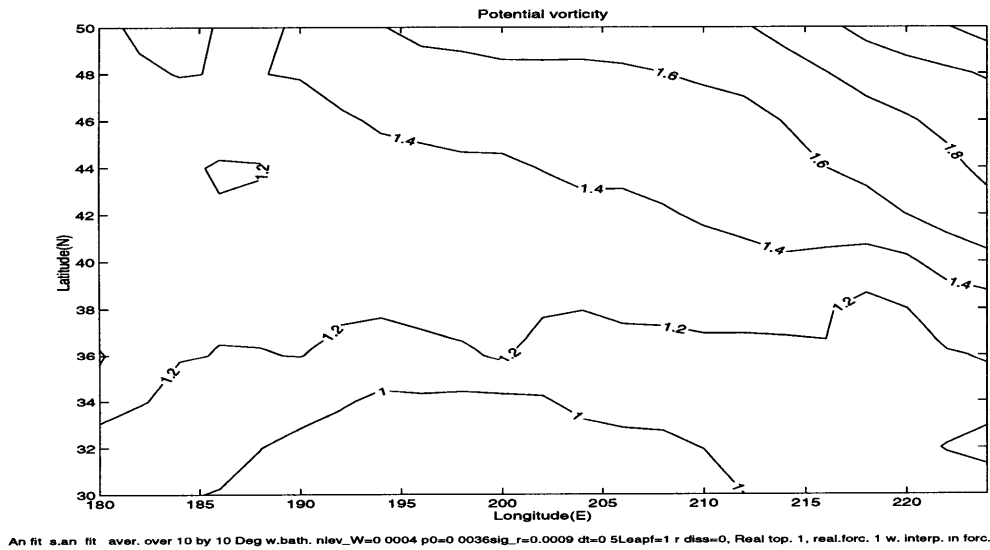
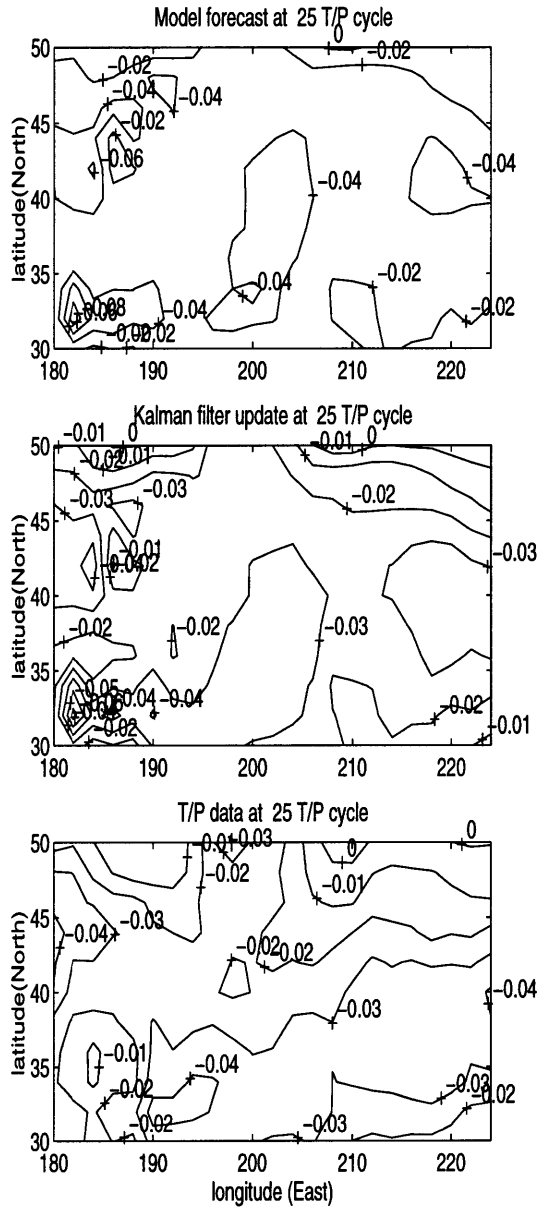


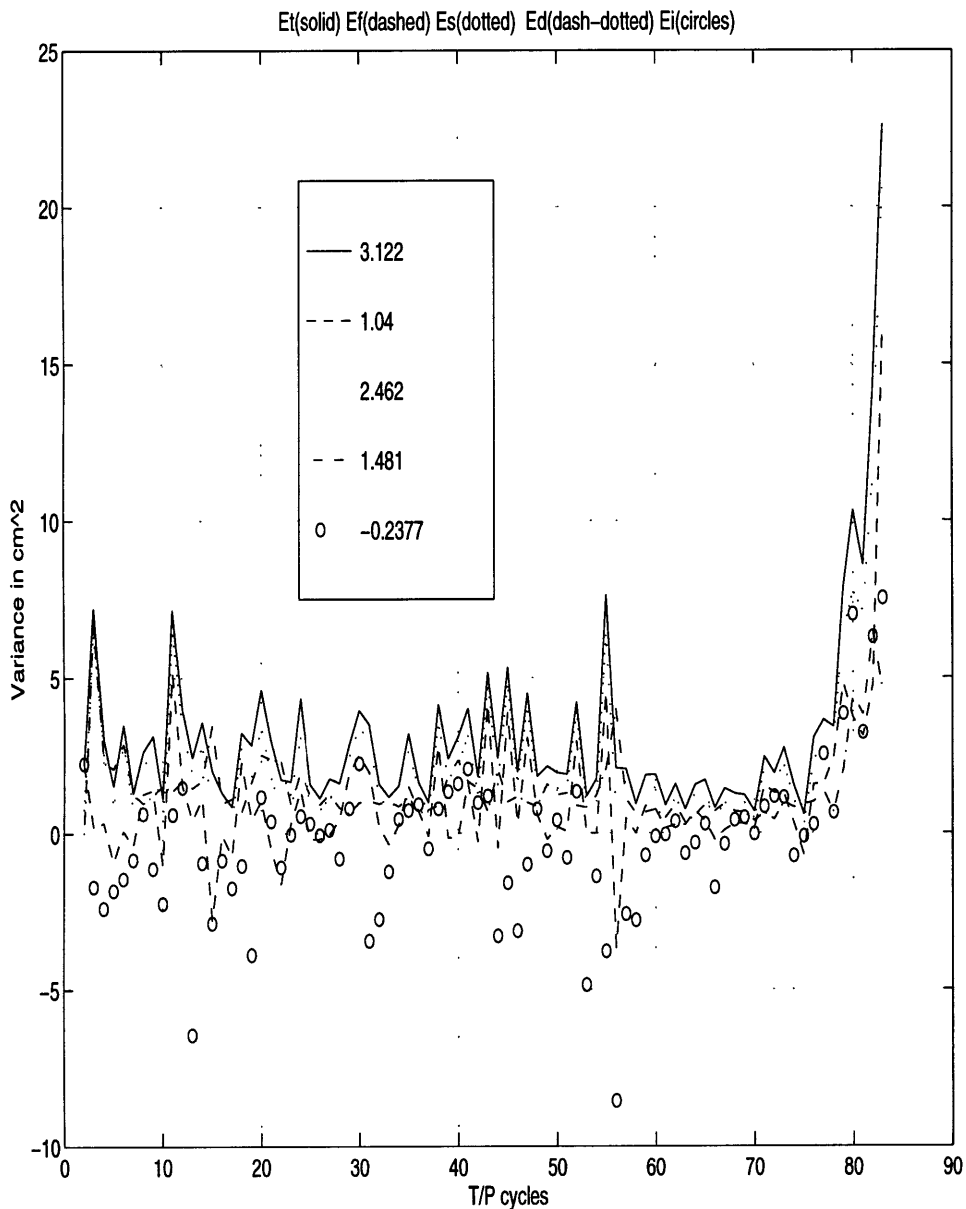
Figure 52: Contours of potential vorticity.



An.fit.s.an. fit. aver. over 10 by 10 Deg w.bath. nlev\_W=0.0003 p0=0.0036sig\_r=0.0004 dt=0.5Leapf=1 r.diss=0, Real top. 1, real.forc. 1 w. interp. in forc.

Michael Chechelnitsky 11-Apr-96, 3:22

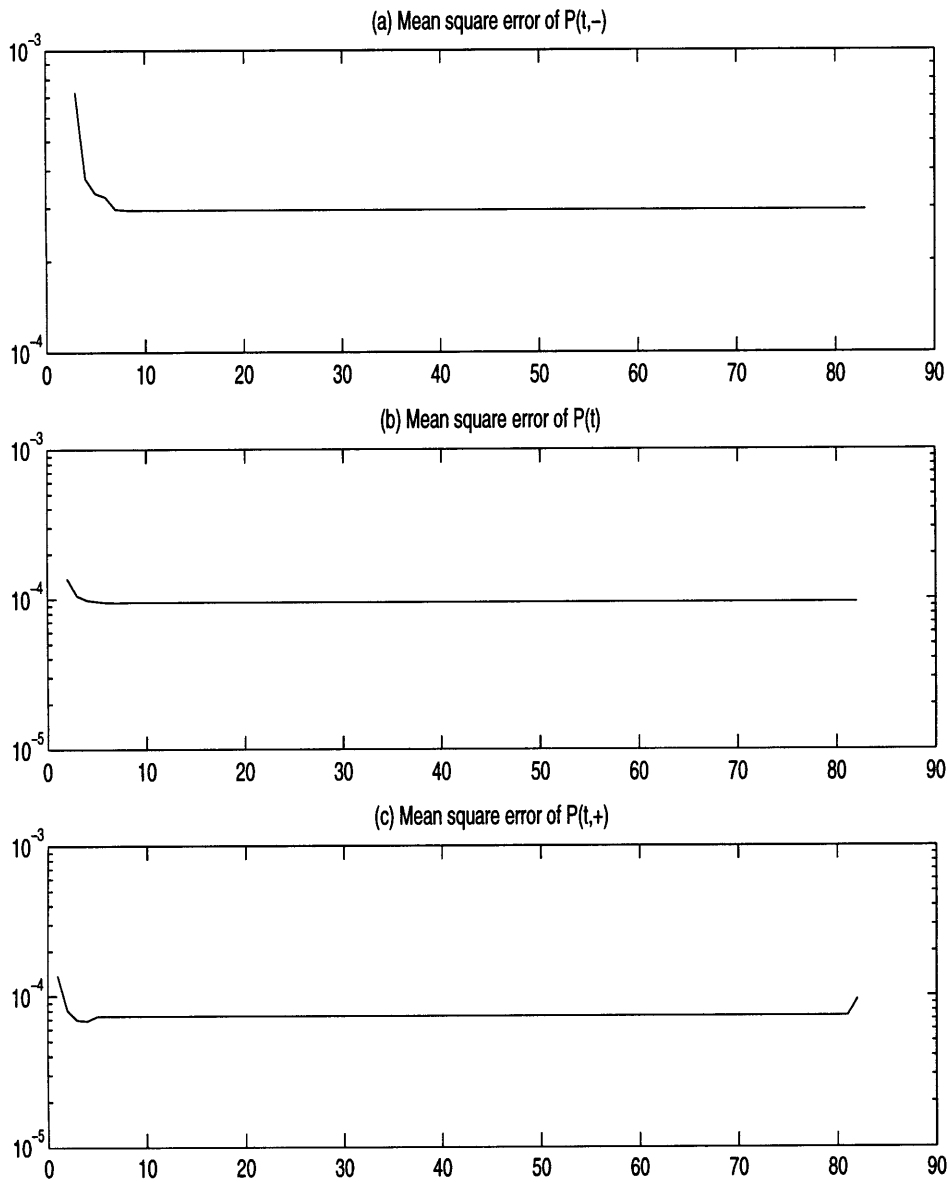
Figure 53: Comparison of a) model forecast  $q(25, -)$ , b) Kalman filter assimilation  $q(25)$ , and c) the observations  $h(25)$  after 2 applying the Kalman filter for 240 days.



An.fit. s.an. fit. aver. over 10 by 10 Deg w.bath. nlev\_W=0.0001 p0=0.0004sig\_r=0.0002 dt=0.5Leapf=1 r.diss=0, Real top. 1, real.forc. 1 w. interp. in forc.

Michael Chechelnitsky 19-Apr-96, 2:34

Figure 54: The amount of variance of the data ( $E_h$ , solid), accounted for by the forecast sea level ( $E_f$ , dotted), by the Kalman filter update ( $E_d$ , dashed), the RTS smoother estimate ( $E_s$ , dash-dotted), and the initial conditions ( $E_i$ , circles), for the model in Northeastern Pacific. The values in the box represent mean values for the respective variables.



An.fit. s.an. fit. aver. over 10 by 10 Deg w.bath. nlev\_W=0.0001 p0=0.0004sig\_r=0.0002 dt=0.5Leapf=1 r.diss=0, Real top. 1, real.forc. 1 w. interp. in forc.  
 Michael Chechelnitzsky 19-Apr-96, 2:49

Figure 55: Evolution of mean variance of the (a) the model forecast, (b) the Kalman filter update, and (c) the smoothed estimate. The time is in T/P cycles. The steady filter was employed 5 T/P cycles, and therefore, the lines are constant for  $t > 6T/P$  cycles. The units are  $m^2$ .

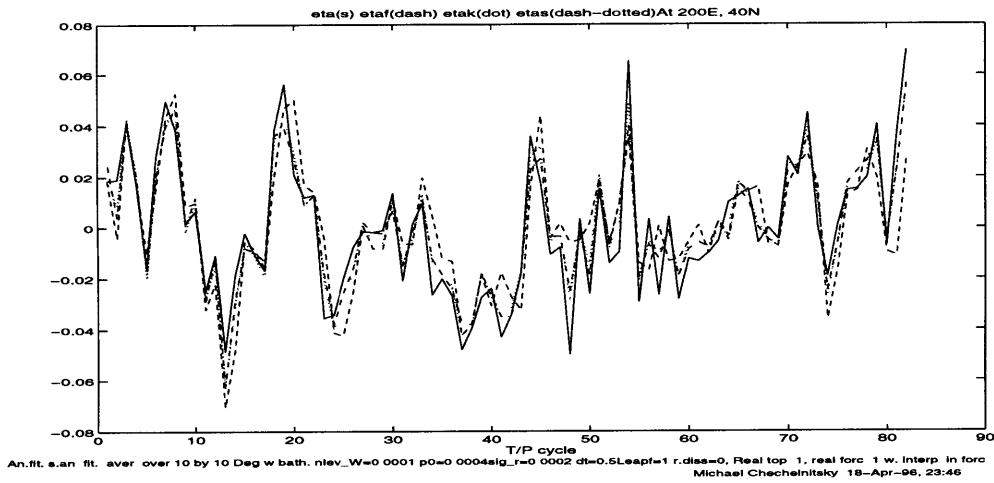


Figure 56: Comparison of SSH variability at 200° E and 40° N for the North east Pacific assimilation: the data (solid line), the model forecast (dashed line), the Kalman filter assimilation (dotted line), and the smoother estimate (dash-dotted line). T/P data was averaged over 10° by 10° and had annual and semi-annual cycles removed.

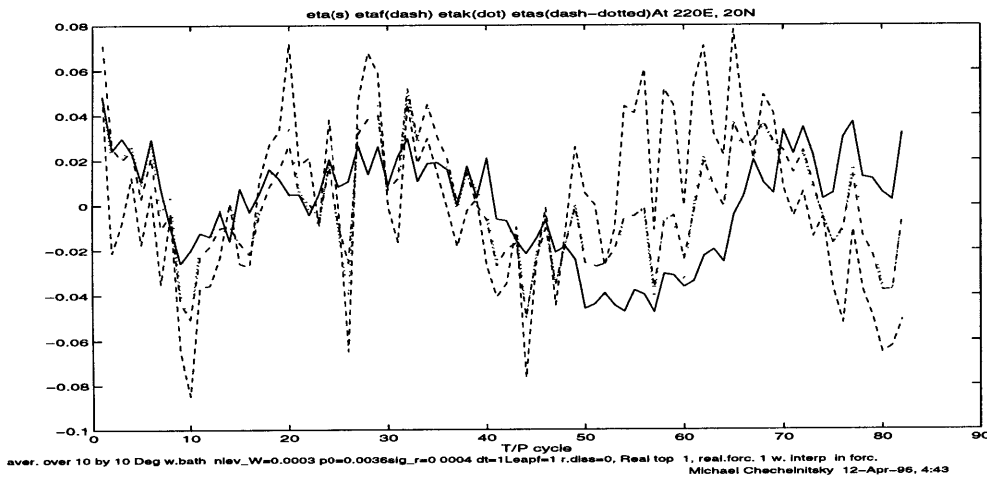
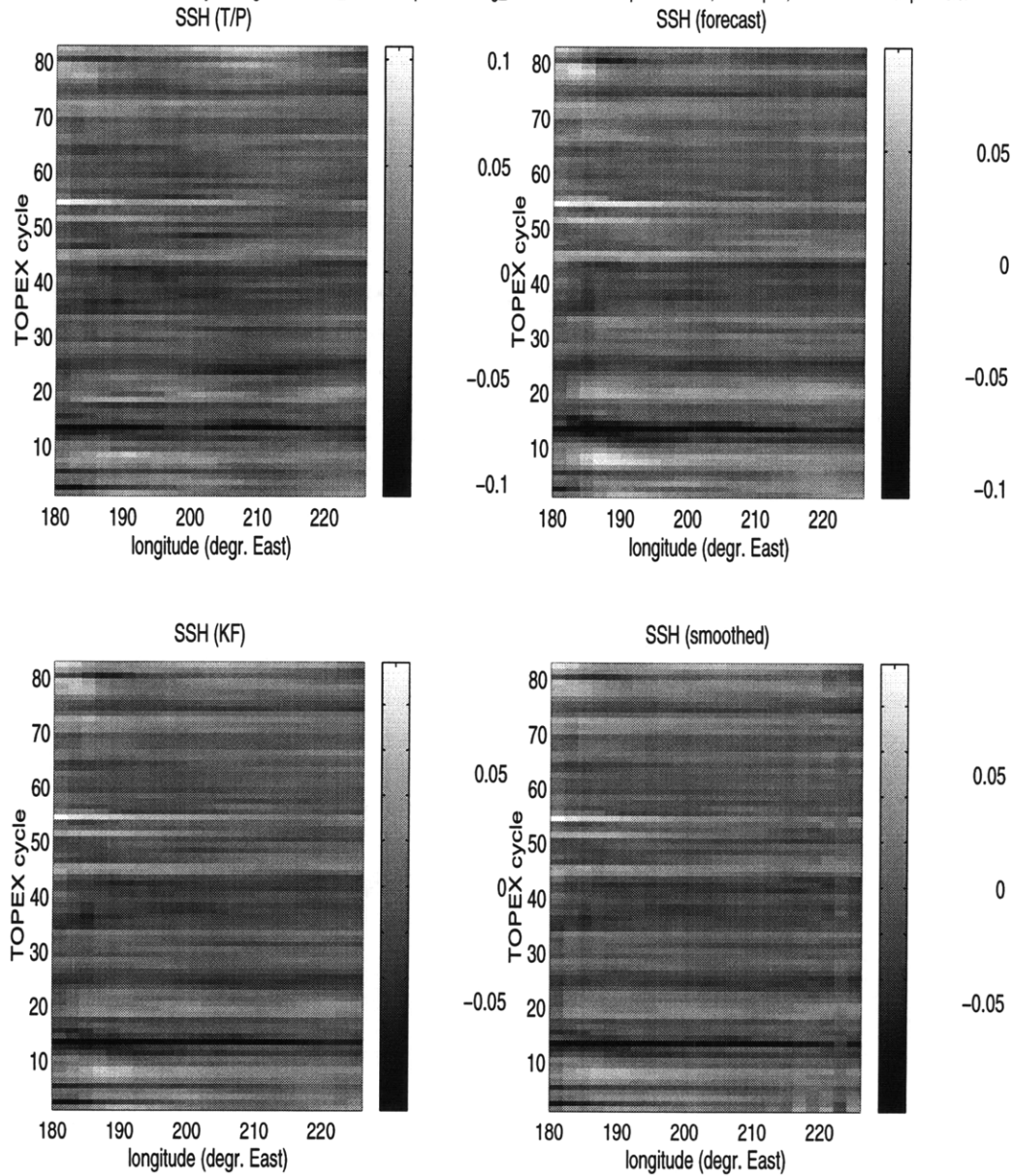


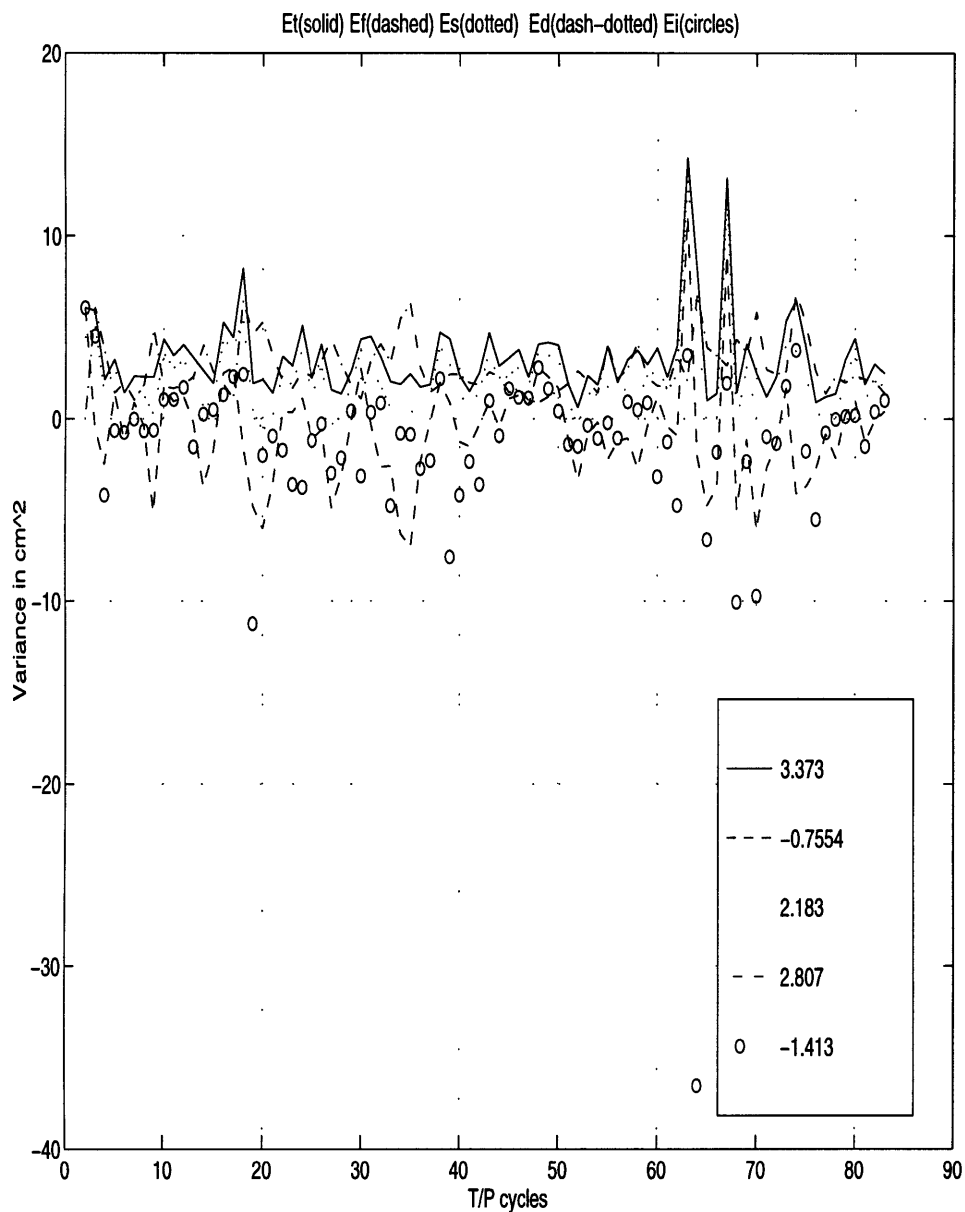
Figure 57: Comparison of SSH variability at 220° E and 20° N for the central east Pacific assimilation: the data (solid line), the model forecast (dashed line), the Kalman filter assimilation (dotted line), and the smoother estimate (dash-dotted line). T/P data was averaged over 10° by 10° and had annual and semi-annual cycles removed.

at 40E An.fit. s.an. fit. aver. over 10 by 10 Deg w.bath. nlev\_W=0.0001 p0=0.0004sig\_r=0.0002 dt=0.5Leapf=1 r.diss=0, Real top. 1, real.forc. 1 w. interp. in forc.



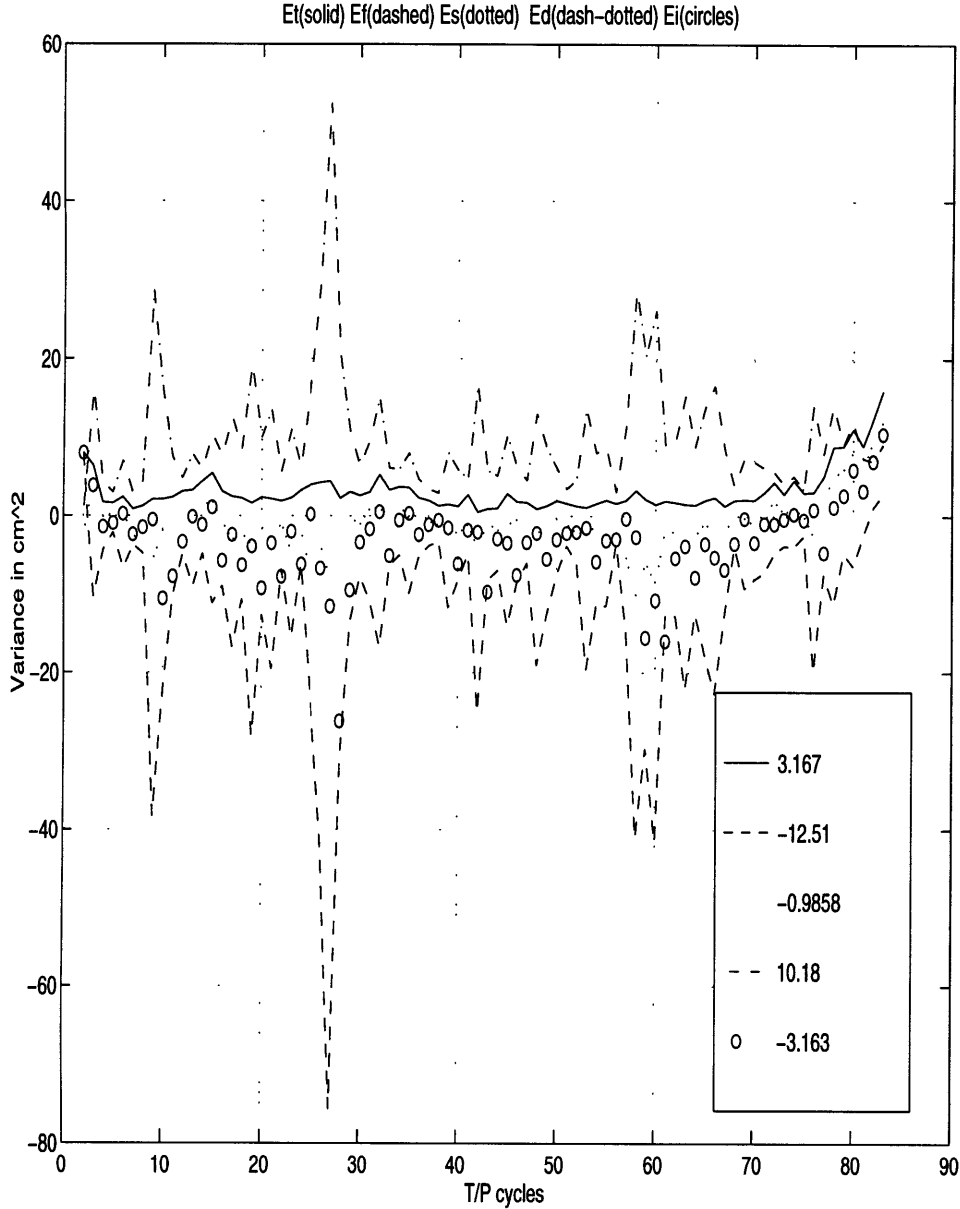
Michael Chechelnitsky 18-Apr-96, 23:47

Figure 58: Longitude versus time plot of SSH anomalies relative to the mean at  $40^\circ$  N from [Left to right, top to bottom]: the observed field (filtered T/P data)  $\mathbf{h}(t)$ , the model forecast  $\mathbf{q}(t, -)$ , the Kalman filter update  $\mathbf{q}(t)$ , and the smoother estimate  $\mathbf{q}(t, +)$ . The resolution of the plots is 1 T/P cycle by  $2^\circ$ . T/P data was in addition averaged over  $10^\circ$  by  $10^\circ$  and had annual and semi-annual cycles removed.



An.fit. s.an. fit. aver. over 10 by 10 Deg w.bath. nlev\_W=0.0001 p0=0.0004sig\_r=0.0002 dt=0.5Leapf=1 r.diss=0, Real top. 1, real.forc. 1 w. interp. in forc.  
 Michael Chechelnitsky 18-Apr-96, 23:36

Figure 59: The amount of variance of the data ( $E_h$ , solid), accounted for by the forecast sea level ( $E_f$ , dotted), by the Kalman filter assimilation ( $E_d$ , dashed), the data update ( $E_d$ , dash-dotted), and the initial conditions ( $E_i$ , circles), for the model in south east Pacific. The values in the box represent mean values for the respective variables.



An.fit s.an. fit. aver. over 10 by 10 Deg w.bath. nlev\_W=0.0003 p0=0.0036sig\_r=0.0004 dt=0.5Leapf=1 r.diss=0, Real top. 1, real.forc. 1

Michael Chechelnitzky 19-Apr-96, 3:08

Figure 60: The amount of variance of the data ( $E_n$ , solid), accounted for by the forecast sea level ( $E_f$ , dotted), by the Kalman filter assimilation ( $E_d$ , dashed), the data update ( $E_d$ , dash-dotted), and the initial conditions ( $E_i$ , circles), for the model in central east Pacific. The values in the box represent mean values for the respective variables.