Variational data assimilation and parameter estimation in an equatorial Pacific Ocean model

OLE MARTIN SMEDSTAD* and JAMES J. O’BRIEN

Mesoscale Air-Sea Interaction Group, The Florida State University, Tallahassee, FL 32306, USA

Abstract—A variational data assimilation and parameter estimation method for a reduced gravity model is developed. The method is applied to the equatorial Pacific Ocean. In the variational formalism a cost function measuring the “distance” between the model solution and the observations is minimized. The phase speed in the model is used as a control parameter and the optimal spatial structure giving the best fit of the model to the observations is determined. In the minimization algorithm a conjugate gradient descent direction is used. The method is computationally effective, and for the experiments considered convergence is achieved in ten iterations or less.

Several experiments are performed using the model solutions as observations. It is shown that the assimilation algorithm is able to determine the large scale spatial structure of the phase speed, even if observations are available at only three stations. The estimated phase speed is not sensitive to errors in the observations, and the algorithm gives a unique solution to the problem.

Real sea level observations from three stations are assimilated for two different periods. The year 1979 was chosen to represent a year without an El Niño, while 1982/83 was chosen to represent an El Niño year. For 1979 the assimilation gave a phase speed with higher values in the west and lower values in the east compared to the initial guess of a constant phase speed. Assimilation of observations in 1982/83 gave the opposite picture, with lower values in the west and higher values in the east. This result is consistent with observations. The phase speed is proportional to the depth of the thermocline, and during normal conditions the basic stratification consists of a deep thermocline in the west and a shallow thermocline in the east. During an El Niño the picture is reversed. Calculating the correlation coefficient between the model results and the observations shows that the correlation increased for all the stations during the assimilation, even at stations which were not a part of the assimilation.

CONTENTS

1. Introduction 180
2. Variational Formalism 185
   2.1. Description of the method 185
   2.2. The cost function J 188
   2.3. Outline of solution procedure 189
   2.4. Computational aspects 190
3. Descent Methods 191
   3.1. Introduction 191
   3.2. The conjugate-gradient method 192
   3.3. Newton, quasi-Newton and limited-memory quasi-Newton methods 195
   3.4. Computing the step length 197
4. Parameter Estimation 199
5. Specification of the Problem 202
   5.1. The numerical model 202
   5.2. The adjoint equations 204
   5.3. The wind forcing 207
   5.4. The observations 208

* Present affiliation: Sverdrup Technology Inc., Stennis Space Center, MS 39529, USA.
1. INTRODUCTION

Oceanographers will experience an explosion in the number of observations available in the next decade. Several new observational techniques are being developed. Satellite measurements (altimetry, scatterometry) around 1990 will give oceanographers a large amount of new data. New techniques for observing the interior of the ocean (such as tomography) will also create new data and it will be important to be able to extract as much information as possible from this new data set. The use of numerical models will play a crucial role in this work, and methods for four-dimensional data assimilation have to be developed. From a meteorologist’s point of view, data assimilation is the process through which observations distributed in time and space are treated in order to specify the initial conditions of a numerical forecast. The large number of stations taking meteorological observations at synoptic times make the initializing process possible. Even with the new observational techniques in oceanography mentioned earlier, the amount of data will not be large enough for initialization of the oceanographic models in the same sense as the meteorological models.

Oceanographers have until now had little experience with four-dimensional data assimilation. Meteorologists have on the other hand worked with this problem for a long time. For a review of data assimilation methods used in meteorology, see e.g. Bengtsson, Ghil and Källén (1981), Lorenc (1986), Navon (1986) or Le Dimet and Navon (1989). A short summary of three different classes of assimilation algorithms is given below, as well as a description of some of the results published in data assimilation in oceanography.

Over the last decade, the development and implementation of four-dimensional data assimilation techniques have dramatically improved the accuracy and dynamical consistency of meteorological analysis. Over the past 30 years of development of numerical analysis and assimilation schemes, most presented schemes belong to one of basically three different classes of algorithms:

(1) Local polynomial interpolation methods.
(2) Statistical (optimal) interpolation methods.
(3) Variational numerical analysis methods.

In the polynomial interpolation method, polynomial functions are adjusted to the observed data in the close neighborhood of each gridpoint. These methods were first introduced by Panofsky (1949) and also by Gilchrist and Cressman (1954), and have
the great advantage of being both very simple and economical to use. The polynomial
interpolation methods have been used for a number of years and are still often used for
various purposes.

Past experience on atmospheric scales, for example, does not enter into the analysis in
the local polynomial interpolation methods. In the second class of assimilation schemes,
the statistical (optimal) interpolation schemes, past experience about the behaviour of
the atmosphere is used as the main source of information for determination of the
interpolation weights. Eliassen (1954) first proposed the method based on spatial
autocorrelation functions of the pressure field, while Gandin (1965) developed the
method for operational use. In most of the implemented schemes, weights are assigned
to a linear combination of observed departures from a guess field in an attempt to
minimize the mean square analysis error. The statistical information needed is the first
and second moments of the observed and unknown variables. These methods are now
widely used, especially as part of assimilation procedures for large-scale prediction
models, see e.g. Lorenc (1981). One major advantage of statistical interpolation is that
it produces a practical and internally consistent approach for treating a large set of
heterogeneous observations, and it is at present the technique which produces the best
results for operational weather forecasting. However, there are certain difficulties in
using statistical interpolation. One is that it tends in some cases to smooth the analyzed
fields excessively. Excessive smoothing may be particularly troublesome for mesoscale
forecasting models, since it may inhibit developments which are unlikely from a
statistical point of view, but are very important to predict correctly, just because of the
rarity of their occurrences.

The statistical (optimal) interpolation methods are computationally expensive to use.
Large systems of linear equations have to be solved to find the interpolation weights.
Successive correction methods introduced by Bergthorsson and Doos (1955) and
Cressman (1959) and modified by Bratseth (1986) may be looked upon as an empirical
approximation to the statistical interpolation method. In these methods the interpolation
weights are computed explicitly without solving a system of linear equations, so the
number of computations is relatively low compared with the statistical interpolation
method.

For linear models an extension of the optimal interpolation methods is the Kalman or
Kalman–Bucy (K–B) filtering (Kalman, 1960; Kalman and Bucy, 1961). The K–B
filter is even more computationally expensive to use than the optimal interpolation
methods. In the latter method an estimate of the error covariance function (the
covariance of the model/data differences) is used to compute a correction to the model
solution. In the K–B filter the estimate of the covariance function is updated each time
step. If a sophisticated model is used, the computational requirements may be unafford-
able. The K–B filter has appeared in meteorology, e.g. Ghil, Cohn, Tavantis and
Isaacson (1981). Lately it has been discussed in the oceanographic literature: Miller
(1986), Budgell (1986, 1987) and Bennett and Budgell (1987). In all these appli-
cations a simple model has been used.

Kindle (1986) used a direct replacing of observations into an eddy resolving
numerical model and he found that the model integration would not converge given
observed data unless the data had a time/space sampling rate equal to the time/space
decorrelation scale of the model eddy activity. Bennett and Budgell (1987) got similar
results using a K–B filter. Miller (1986) used a Kalman filter to assimilate data into an
eddy resolving open ocean model. He used the full Kalman filter, but in order to implement the method, the filter was applied to simplified systems designed to capture some of the properties of open ocean modelling.

The third class of assimilation schemes consists of variational analysis methods. In these schemes a given measure of the “distance” between the analysis and the observations is minimized. The analyzed field must at the same time (approximately) satisfy an explicit dynamical constraint. The constraint will normally be expressed by one (or more) differential equation(s). Variational schemes have a high mathematical technicality and also a high computational cost. Although the usefulness of variational methods for meteorological problems was pointed out very early by SASAKI (1955, 1958, 1970), and in spite of a fairly large number of various studies, e.g. THOMPSON (1969), LEWIS and BLOOM (1978) and BLOOM (1983), these methods have not been fully utilized until recently. This is especially true for assimilation studies which address problems containing explicit time derivatives, see e.g. LEWIS and BLOOM (1978).

LEDIMET and TALAGRAND (1986) studied different variational algorithms for analysis and assimilation of meteorological observations. They discussed two different approaches to data assimilation. In the first approach the original constrained problem is transformed into an unconstrained problem or a sequence of unconstrained problems. Three different classical algorithms are presented, the penalty, the duality and the augmented Lagrangian algorithm. The latter is a generalization of the penalty and the duality algorithms. The second approach uses optimal control techniques, and is based on the use of an adjoint dynamical equation. This algorithm seems to work well in the assimilation of observations distributed in time. The main advantage of looking at unconstrained problems is that these problems can be solved by classical descent algorithms such as the steepest descent or the conjugate-gradient method.

In the last few years the variational approach has been studied for meteorological application by several authors, e.g. DERBER (1985), LEWIS and DERBER (1985), HARLAN and O'BRIEN (1986), HOFFMAN (1986), LORENC (1986, 1988), TALAGRAND and COURTIER (1987) and COURTIER and TALAGRAND (1987). In many of these investigations the models used are similar to the oceanographic models and from the success reported in these investigations it may be inferred that the variational techniques should be considered for oceanographic data assimilation.

A few remarks about the different classes of data assimilation should be made. Although they were conceived and developed independently, they do have mutual relationships. For example the interpolation method of GILCHRIST and CRESSMAN (1954) can be described as a simple case of statistical interpolation. KIMELDORF and WAHBA (1970) have shown that the statistical interpolation method produces fields which are the solution of a variational problem in which the function to be minimized is the sum of two terms with one term representing the distance to the observations and the other term some measure of smoothness of the fields. The forward-backward data assimilation introduced by MOREL, LEFÈVRE and RABREAU (1971) can be related to the variational assimilation. In their approach, the model is integrated forward and backward repeatedly over time to obtain an adjustment of the model to the observations. In the variational approach the model itself is integrated forward, but the adjoint of the model is used in the backward integration. THACKER (1986) discussed the connection between Kalman filtering and the variational approach for data assimilation using a linear model. Kalman filtering can be thought of as an algorithm to solve the variational equations.
As mentioned earlier oceanographers have not had much experience with data assimilation. If the phrase data assimilation is used in a broader sense, the first attempts of assimilating data into models for the ocean were the dynamical methods used to produce maps of the currents from hydrographic observations as in Sandstrøm and Helland-Hansen (1903). Geostrophic shear can be directly established from the density data. These early attempts at data assimilation were followed by more sophisticated models e.g. Stommel and Schott's (1977) beta spiral or Wunsch's (1977, 1978) inverse method. Recently there have been some developments in the area of data assimilation in oceanographic numerical models. Marshall (1985) used estimation theory to assimilate simulated data from satellite altimetry into an ocean model, and at the same time improving the geoid. Webb and Moore (1986) made use of the projection methods of estimation theory, but they assumed that the measurements were error free, and that altimeter measurements were available everywhere. They also assumed that the ocean currents were weak so that the ocean could be represented by a superposition of linear Rossby waves. A result of this study was that the determination of the deeper structure of the ocean was limited by the phase separation that develops over each assimilation cycle between modes of the ocean with the same horizontal wavenumber but differing vertical structure. Hurlburt (1986) used a two-active-layer, free surface primitive equation model on a β-plane to investigate the dynamic transfer of surface to subsurface information. Perfect altimeter data were simulated by the free surface of the two-layer model. The results showed that the maximum update interval that provided success was about half the shortest major time scale in the model.

Bennett and McIntosh (1982) used a variational method in the investigation of tidal motion. Their results show that the choice of data weights are of great importance. Provost (1893) and Provost and Salmon (1986) have used a variational technique to assimilate hydrographic station data to estimate the three-dimensional field of geostrophic velocities. The method of weak constraints (Sasaki, 1970) was used. They determined the smoothest velocity field which was consistent with the data and at the same time approximately satisfied the dynamical constraints. The way they assured a smooth solution was to penalize kinetic energy as well as enstrophy. Malanotte-Rizzoli and Holland (1986) used a quasi-geostrophic general circulation model to investigate the effect of data insertion into a numerical model, and how data insertion in local, limited regions affect the dynamics of the model, and also how the dynamics spread the inserted information to different regions. Their model is steady, weakly nonlinear and highly frictional. The method they use is a naive approach with direct insertion of “observations” in the model. The “observations” used in this study are identical-twin data, i.e. data generated by the model. The results show that the region influenced by the insertion and the improvement of the results depend upon the location of the “observed” data, the orientation of the section inserted and the length of the section. Malanotte-Rizzoli and Holland (1988) extended their results from the 1986 paper to the transient eddy-resolving case. One of their major results is that in the transient case a single data section is very ineffective to drive the model towards the reference ocean over time scales of ~100 days. All the different sections they tried were equally ineffective. If instead the data are inserted over a period longer than the model equilibrium time, a single section is quite effective in driving the model to the reference ocean.

Holland and Malanotte-Rizzoli (1989) studied the effect of along-track assimila-
tion of altimeter data into a three layer eddy-resolving quasi-geostrophic model. Again identical twin data were used. They reached the conclusion that if perfect data are available (i.e. data at every gridpoint and every \( \frac{1}{2} \) day with linear interpolation in time), a simple (nudging) technique can be very successful in driving the model towards the control solution. If altimetric data are available only along actual satellite tracks with repeat periods of either 10 or 20 days, the results are not as promising. The residual root mean square (rms) errors are close to 70% at the end of the assimilation period. In MALANOTTE-RIZZOLI, YOUNG and HAIDVOGEL (1989) a primitive equation model was used to study a series of idealized initialization/assimilation experiments. The first question they ask is whether the initialization shocks are equally important in ocean models as they are in the atmospheric models. Their conclusion is that the ocean models do not seem to be as sensitive to unbalanced initializations. All that is necessary to ensure a smooth evolution, is a geostrophically balanced initialization. They also studied data assimilation and the effect of different types of data. A simple insertion technique was used as the assimilation scheme, and identical twin data represented the observations. The results of this investigation show that the knowledge of the interior density field is the most effective data in reducing the rms errors.

In a study by SCHROTER and WUNSCH (1986) the effect of observational errors in the driving of the models is investigated. The procedure is based upon nonlinear optimization methods. From their algorithm it is possible to calculate the qualitative sensitivity of the objective function to change in the data errors and finding an optimization technique capable of dealing with data uncertainty. MOORE, COOPER and ANDERSON (1987) studied the initialization and data assimilation in two different models of the Indian Ocean. The models they used were a reduced-gravity model and the GCM of BRYAN (1969) in the SEMTNER (1974) coding. The method they used was direct updating of the whole temperature and/or the velocity fields. In an attempt to simulate assimilation of data from the proposed TOGA XBT network, they used the method of successive correction (BERGTHORSSON and Döös, 1955). All their tests were performed using an "identical twin" approach. The result of this study showed that the use of temperature data (mass) gave better results than if velocity observations were used. The effect of using velocity data could be increased if the ratio between the potential and kinetic energy in the model were changed. An increase of the kinetic energy made velocity data more useful in the assimilation process.

In two papers, THACKER and LONG (1988) and THACKER (1988), a variational data assimilation procedure is described. The method is illustrated using a simple model of the wind-driven equatorial ocean. It is shown how surface elevation and wind stress observations can be used to recover the model state. In THACKER (1988) the process of fitting a model to inadequate data is discussed. Results show that for the simple three wave model of THACKER and LONG (1988), a reasonable fit can be obtained even if the number of observations is less than the number of degrees of freedom of the model. THACKER (1989) discussed the role of the Hessian matrix (the second derivatives of the cost function) in fitting models to measurements and he gave an example of how the inverse Hessian can be identified as the covariance matrix that establishes the accuracy to which the model state is determined by the observations. Calculating the Hessian matrix can be an impossible task for a complicated model using asynoptic observations.

The approach followed here is the variational data assimilation method. SASAKI'S (1970) method of strong constraints is used. The assimilation procedure is formulated for
an equatorial Pacific Ocean model. The goal will be to determine the solution of the model which best fits the observations. The best fit solution could be determined in different ways. The meteorological approach of finding the initial conditions which give the best solution is one method. When the variational data assimilation method is used, it is possible to adjust other variables than the initial conditions. Parameters in the model can easily be incorporated in the procedure, so that they can be used as control parameters. The variational algorithm is formulated here using the phase speed of the Pacific Ocean model as the control parameter. The phase speed is then estimated so that the model solution is as close as possible to the observations. Observations of sea level from island stations will be assimilated into the model.

The variational method is described in Section 2. A general description of the variational formalism is given. Similarities between different approaches to the variational data assimilation method are discussed. The choice of the cost function and different ways of dealing with insufficient data are also discussed. An outline of the solution procedure and some computational aspects of the method are described. A general description of the conjugate gradient method is given in Section 3, while some aspects of parameter estimation is discussed in Section 4. In Section 5 the problem of data assimilation in a Pacific Ocean model is specified. The model equations are described and the corresponding adjoint equations are derived. The wind stress used as forcing in the Pacific model is presented. A description of the sea level observations from the island stations in the Pacific Ocean is given. The method for calculating the data misfit used as forcing in the adjoint equations is also described. Numerical results obtained using the data assimilation algorithm are presented in Section 6. Both simple experiments using the height and velocity field of a Kelvin wave as initial conditions and realistic results using observed wind stress as forcing, are discussed. A summary of the results and the conclusions are given in Section 7.

2. VARIATIONAL FORMALISM

2.1 Description of the method

The assimilation scheme used is a variational method. As mentioned in the introduction, this scheme minimizes a given measure of the distance between the model and the observations. The primary field of interest is time dependent models of the ocean and a mathematical model may be written as

\[
\frac{\partial x}{\partial t} = F(x; c)
\]

where the components of the vector \( x \) are the dependent variables, \( t \) is time and the vector \( c \) contains the parameters of the model. The components of \( x \) denote the various oceanographical fields (velocity, height field, \ldots) under consideration. \( F \) may be a linear or non-linear operator. It is assumed that the system (1) is not closed, i.e. additional information has to be provided in order to obtain a unique solution to (1). The additional information will be given by \( y \)—a control variable. \( y \) can for instance consist of the initial
conditions, boundary conditions, some of the parameters in the problem, or a combination of these. When \( y \) is defined, a unique solution \( x(y) \) of (1) can be found. It is important to realize that the control variable \( y \) must belong to a set of admissible control, \( \mathcal{Y} \). To determine \( \mathcal{Y} \) one may use physical information about the initial conditions or the parameters used as a control variable.

One is interested in obtaining a solution of (1) which is close to the observations, \( x' \). Closeness may be defined in different ways, and it is usual to define a cost function \( J \), which measures the distance between the solution of (1) and the observations. The definition of \( J \) will be discussed later. The observation, \( x' \), will consist of measurements of oceanographic fields, e.g. the height field or the velocity field. A variational problem can now be stated as a problem (P):

\[
\text{Find } y^* \text{ which belongs to } \mathcal{Y} \text{ and minimizes the cost function } J
\]

where \( y^* \) represents the optimal \( y \).

The problem stated above is a constrained minimization problem with the model eqn (1) representing the constraint. The problem is solved by redefining it so that it becomes a problem of unconstrained minimization. Standard procedures for solving problems of this type may be found in textbooks on optimization (e.g. Gill, Murray and Wright (1981), Luenberger, (1984)) or in the meteorological literature (e.g. Le-Dimet and Talagrand, 1986; Navon and LeGler, 1987).

Sasaki (1970) introduced two different methods in variational analysis, the method of weak constraint related to the penalty method and the method of strong constraint. The weak constraint formalism consists of minimizing the functional \( E \) defined by

\[
E(x, y) = J(x, y) + \alpha \| G(x; c) \|^2
\]

where \( \alpha \) is a prespecified weight and \( G \) is given by

\[
G(x; c) = \frac{\partial x}{\partial t} - F(x; c).
\]

It is important to note that in this formulation \( E \) is quadratic in \( G \) and the weights are prespecified. It therefore follows that the constraint is only approximately satisfied. In some cases this may be justified since the model used as constraint is an approximation to the real world and should only be satisfied to its own accuracy.

The approach that will be followed here is based on the classical Lagrange multiplier technique. A Lagrangian \( L(x, \lambda, y) \) can be constructed

\[
L(x, \lambda, y) = J(x, y) + (\lambda, G(x; c))
\]

where the components of the vector \( \lambda \) are the Lagrangian multipliers. \((,)\) is an inner product defined on the functional space to which \( G(x; c) \) belongs. This is the strong constraint formalism according to Sasaki (1970). The constrained minimization problem (P) is thus replaced by an unconstrained problem with respect to the variables \( x, \lambda \) and \( y \). Using this formalism it is insured that the observations will satisfy the constraints exactly. It can be shown that the problem of determining the stationary points of the functional
$J(x, y)$ under the constraint $G(x; c) = 0$ is equivalent to the problem of determining the stationary points of (4) with respect to the variables $x, y$ and $\lambda$ (see e.g. BERTSEKAS, 1982). The equations which express that the Lagrangian is stationary, are called the Euler-Lagrange equations of problem (P). The Euler-Lagrange optimality condition is given by

$$\frac{\partial L}{\partial \lambda}(x^*, \lambda^*, y^*) = 0$$

$$\frac{\partial L}{\partial x}(x^*, \lambda^*, y^*) = 0$$

$$\frac{\partial L}{\partial y}(x^*, \lambda^*, y^*) = 0$$

and determines $x^*$, $\lambda^*$ and $y^*$. Equation (5) gives the original model equation back. The operator in (6) is the adjoint of the operator in (5). It is therefore customary to call eqn (6) the adjoint equation of (5). The model equation will propagate information forward in time, while the adjoint equation will propagate information backward in time. When the appropriate model equations have been derived, the interpretation of the adjoint equation will be discussed in more detail. The last equation, eqn (7), yields the gradient of $L$ with respect to the control variable. The gradient plays an important role in determining the best fit solution, since it is one of the major constituents in the descent algorithms used to find the minimum of the cost function (the Lagrangian). As can be seen from these equations the whole variational analysis depends on the choices of the functional $J$ and the constraint $G$. The method of solution is also an important part of the analysis. For most choices of $J$ and $G$ an iterative method has to be used and the goal is to keep the number of iterations as low as possible.

There are other classical algorithms for solving the constrained problem (P). One algorithm introduces the penalized functional

$$J_\epsilon(x, y) = J(x, y) + \frac{1}{\epsilon} \|G(x; c)\|^2$$

where $\epsilon$ represents what is called the penalty parameter and is a sequence of positive numbers tending to zero. It can be shown that under certain hypotheses the solution $x^*_\epsilon$ tends to the solution $x^*$ of the original problem when $\epsilon$ tends to zero. The similarity between this method and what SASAKI (1970) called the weak constraint formalism (2) is evident. However, there is an advantage of using the penalty method. By varying $\epsilon$ instead of using only one constant prespecified weight $\alpha$, one can determine the dependence of the solution upon the choice of coefficient.

Numerical problems may arise when the penalty algorithm is used. For small values of $\epsilon$ the method may lead to ill-conditioning and inaccuracy in the determination of the
minimum. This can be dealt with by introducing the augmented Lagrangian

\[ L(x, y) = J(x, y) + \frac{1}{\varepsilon} \|G(x; c)\|^2 + (\lambda, G(x; c)) \]

where as before \( \lambda \) represents the Lagrangian multipliers (see Bertsekas, 1982). The augmented Lagrangian method would in Sasaki’s terminology be a combination between weak and strong constraints. The major advantage of the augmented Lagrangian algorithm is that it is computationally more efficient than the penalty method. The augmented Lagrangian alleviates the problems connected with using the penalty or the multiplier method alone. Navon and de Villiers (1983) used an augmented Lagrangian algorithm to enforce integral invariants in a shallow water equation model of the atmosphere.

It is possible to formulate the method of adjoint equations without introducing the Lagrange multipliers. Optimal control techniques which in fact are techniques for minimizing a functional as the cost function, can be used to calculate the gradient of the cost function with respect to the control variable(s). Using these techniques it is necessary to solve the model eqn (1) and then solve the adjoint of the linearized form of the same equation. This approach is somewhat more difficult to formulate. It is necessary to write a well-posed problem and carefully specify the functional framework of the variational problem. Since the optimal control techniques do not introduce the Lagrange multipliers it might look like this is a large reduction in the number of unknowns. But, in fact, in the calculation of the gradient of the cost function it is necessary to compute the same numbers of variables. The method of optimal control techniques is discussed in Le Dimet and Talagrand (1986) (see also Lions, 1971). A review of these techniques used in meteorology can be found in Le Dimet and Navon (1989).

2.2. The cost function J

The objective of the data assimilation is to find a solution which is close to the observations. The cost function \( J \) is constructed so it measures the distance (the misfit) between the model and the observations, and it must therefore have the property of a norm. There are several different functional forms of \( J \) which might be considered, and each one of them will give a different result for the best fit model solution. The variational method makes use of the derivative of \( J \), and the cost function must therefore be differentiable. It is usual to choose \( J \) as the square of a norm. \( J \) can then be defined as

\[
J(x, c) = \frac{1}{2}(x - x')^T K_s (x - x') + \frac{1}{2}(c - c')^T K_c (c - c')
\]

where as above \( ' \) denotes an observed or estimated quantity and the \( K_s \) are specified validity matrices. They depend on the error variance of each observational point, and give information about the quality of the data. If the observational errors are uncorrelated, the validity matrices are diagonal. The \( K_s \) are always symmetric positive definite matrices. Observations may not be available everywhere. For points where there are no observations, the validity matrices are set equal to zero. The first term in eqn (10) is called the data misfits and will be the forcing for the adjoint equation. The last term in eqn (10) is added to the cost function because the goal of this research is to estimate the
parameters in the model. By adding this term, the new estimate of the parameters will not be too far from the initial guess. In other words minimizing the cost function results in a solution which is close to the observations and new values of the parameters which are close to the estimate. The effect of the last term will be discussed further in Section 4.

For numerical models of the ocean or the atmosphere, the number of degrees of freedom is related to the number of gridpoints used, or for a spectral model, the number of terms included in the spectral expansion. Most models already have more degrees of freedom than the available data bases. One way to supplement insufficient data is by adding a “penalty” or smoothing term to the cost function. SASAKI (1970) used this idea and he pointed out that adding a smoothing term to the cost function suppressed high frequencies and wave numbers in the solution. WAHBA and WENDELBERGER (1980) have shown that the degree of the highest derivative in the smoothing term must be at least two orders greater than the highest derivative in the data constraint if a mathematically well behaved solution is going to be obtained. This idea has been used by e.g. PROVOST (1983), PROVOST and SALMON (1986), HARLAN and O’BRIEN (1986), THACKER (1988) and LEGLER, NAVON and O’BRIEN (1989). there are many choices of these penalty terms. For instance the addition of a term of the form

\[ K_p \nabla^2 x \]  

will minimize the curvature of the assimilated field. In other words it tends to make the resulting field smoother.

Another way to add more information would be to use a solution of a previous model run as observations, or if no such run is available climatological values for the control parameters could be used as observations (LORENC, 1986). As long as the weights, K, used for the “observations” from previous forecasts or climatology are small compared to the Ks of the real observations, the effect of these data points will have a negligible effect on the solution except for the part that is not determined by the real data.

2.3. Outline of solution procedure

The procedure for using the variational method for data assimilation and parameter estimation can be formulated in the following way.

1. Choose a first guess for the control parameters.
2. Integrate the model forward for the period over which the observations are assimilated.
3. Calculate the data misfits \((x - x')\).
4. Integrate the adjoint equations backward in time forced by the data misfits.
5. Calculate the gradient of \(L\) with respect to the control variables.
6. Use the gradient of \(L\) in a descent algorithm to find a new value of the control parameters which make the cost function move towards its minimum.
7. Check if the optimal solution has been found. This can be done by checking the norm of \(\nabla J\) or the value of \(J\) to see if it is less than a prescribed tolerance.
8. If this is not the case, the procedure described above can be repeated using the new values of the control parameters as a new guess and continue the iterative process until a satisfactory solution has been determined.
When trying to find the minimum of the cost function (the Lagrangian) it is important to realize that one cannot expect the minimum to be exactly zero. Experience from meteorology has shown that the decrease in the value of the cost function or its gradient is at most a couple of orders of magnitude. Hydrologist's experience in estimating parameters in groundwater flow using the variational technique, have shown that relying on the gradient norm as the only measure of convergence to the optimal solution may lead to an unnecessarily large number of iterations. A better way of determining if the optimal solution has been found would be to check and see if \( \| \nabla J \|_2 \) is less than a prescribed tolerance, where \( \| \nabla J_0 \|_2 \) denotes the norm of the gradient during the first iteration. In view of the meteorological and hydrological experience the reduction of the tolerance can be expected to be of the order \( 10^{-2} \). Similarly, instead of checking the absolute value of \( J \), the ratio \( J/J_0 \) should be less than a specified tolerance. The decrease in the cost function itself may not be a good measure of how close the solution is to the minimum. Trying to fit a model to real observations is a difficult task. Depending on the complexity of the model and its ability to predict the real world, it may not be possible to reduce the value of the cost function several orders of magnitude. If the initial guess for the parameters is close to the optimal value, one cannot expect the cost function or its gradient to experience a large decrease. This does not mean that the minimization has not been successful. Other means of checking the behavior of the model should be used to see if the results are improved. In the experiments described later, a correlation coefficient is calculated to show that after the assimilation there is an increase in the correlation between the model and the observations, even if the decrease in the cost function is less than an order of magnitude.

During the first few iterations the reduction in the value of the cost function is fast. After this initial decrease, the reduction per iteration will be slower. The goal should be to keep the number of iterations below about 15-20. If it is necessary to continue with further iterations, the descent algorithm may have to be changed. Another choice may give a better convergence rate. The Hessian matrix plays an important role in the rate at which the algorithm converges. The condition number (the ratio of the largest eigenvalue to the smallest) and the distribution of the eigenvalues themselves determines the rate of convergence (see Thacker, 1989, for a discussion of the role of the Hessian matrix in data assimilation).

2.4. Computational aspects

The variational approach to data assimilation implies the solving of the system of eqns (5)–(7). This may be computationally challenging. For an ocean model the number of variables can easily exceed \( 10^6 \). The system (5)–(7) consists of the model equation and their adjoints. The adjoint equations are of similar complexity as the original model equations. Each iteration requires one forward integration of the model itself and one backward integration of the corresponding adjoint equations. Depending on how many iterations are necessary for the solution to converge, the method may require several times the computer time used by the integration of the ocean model. The necessity of using a descent algorithm which will find the minimum of the cost function in as few iterations as possible is therefore an important part of the variational formalism.

There are several important aspects which have to be considered when a descent
algorithm is chosen. Some of the simpler algorithms, e.g., the steepest descent method, are easy to implement, but their slow convergence rate may make them impractical to use. More sophisticated methods such as Newton or quasi-Newton methods have quadratic convergence rates, but these algorithms require the storage of Hessian matrices. For problems in oceanography with a large number of variables, the dimension of these matrices may be too large to fit on available computers. Another aspect which is important in the descent algorithm, is the determination of the step length. The step length determines the distance which the descent algorithm is moving down the gradient of the cost function. There are different ways of calculating this parameter. Most of the methods require an additional integration of the forward model. One iteration of the data assimilation procedure may therefore result in 3 times the computer time required by integrating the model itself.

3. DESCENT METHODS

3.1. Introduction

There are several different descent methods available for minimizing the function in (4). The oldest and also the simplest way to find the minimum of a function of several variables, is to use the method of steepest descent. One of the reasons why this method is still important is that it is one of the simplest for which a satisfactory analysis exists. Several of the more advanced algorithms are a result of a modification of the basic steepest descent method in such a way that the new algorithm has improved convergence properties.

The models used in oceanographic studies often have a large number of variables $N$, which can easily exceed $10^4$, and for such cases the only choice for a descent method is the conjugate-gradient algorithm with its better convergence rate than the steepest descent. There are methods with better rates of convergence than the conjugate-gradient algorithm, e.g. Newton and quasi-Newton methods. These methods have quadratic and superlinear rates of convergence, respectively, but the disadvantage is that they require storage of Hessian matrices of second derivatives of size $(N \times N)$. Conjugate-gradient algorithms require storage of only a few vectors of length $N$.

The conjugate-gradient algorithm is an iterative method for unconstrained minimization of a general function of $N$ variables. The method produces a better approximation of the minimum of the unconstrained nonlinear function with each iteration. During each iteration an estimate is made of the best way to change each component of the vector $x$ in order to produce the maximum reduction in the function. The descent direction is found by combining information about the gradient of the function with information from earlier iterations to produce a new search direction. The algorithm also estimates an optimal step size which must be used to find the magnitude of the changes along the search direction.

The conjugate-gradient method has been successfully applied in meteorology to minimize the cost function used in variational analysis, e.g. Hoffman (1982, 1984), Navon and de Villiers (1983) and Derber (1985). Navon and Legler (1987) compared different conjugate-gradient algorithms by applying them to two different
meteorological problems, and their conclusion was that the most consistent method was the SHANNO and PHUA (1980) quasi-Newton limited-memory (memoryless) conjugate-gradient algorithm. An introduction to the conjugate-gradient method will be presented in the next section. The SHANNO and PHUA (1980) implementation of the SHANNO (1978a, b) algorithm will be used. This algorithm contains both a limited-memory quasi-Newton conjugate gradient method and a quasi-Newton method with a BFGS update (BROYDEN, 1970; FLETCHER, 1970; GOLDFARB, 1970; SHANNO, 1970). Detailed description of different minimization algorithms can be found in e.g. GILL, MURRAY and WRIGHT (1981) or Luenberger (1984). The conjugate-gradient method was initiated by HESTENES and STIEFFEL (1952) and the historical development of the method can be found in HESTENES (1980).

3.2. The conjugate-gradient method

The conjugate-gradient method, or more generally the conjugate-direction method is analyzed for a purely quadratic function. The problem which will be considered is the problem of minimizing the cost function in eqn (10). This is a problem of unconstrained minimization and it can be written in the following form

\[
\text{minimize } J(x) = \frac{1}{2} x^T Q x + b^T x + c
\]

where \( J(x) \) represents the cost function and \( x \) is a vector representing the \( N \) variables, \( x_0, \ldots, x_{N-1} \), of the function. It is assumed that \( J(x) \) can be written as a quadratic function, where \( Q \) is a positive definite symmetric matrix, \( b \) is a vector and \( c \) is a constant. \( x^T \) represents the transpose of \( x \). In general an iterative algorithm for solving eqn (12) takes the form

\[
x_{k+1} = x_k + \alpha_k d_k
\]

where \( d_k \) is a descent direction and \( \alpha_k \) is a positive step size parameter. Equation (13) is referred to as a generalized gradient method (or just a gradient method).

Let

\[
g_k = \nabla J(x_k)
\]

(14)

denote the gradient of \( J \) with respect to \( x_k \). If the method of steepest descent is used, the descent direction \( d_k \) is simply given by \( -g_k \), the negative gradient of the function.

A couple of useful properties of conjugate-directions will be stated. The definition of conjugacy can be formulated as

**Definition:** Given a symmetric matrix \( Q \), two vectors \( d_i \) and \( d_j \) are said to be conjugate with respect to \( Q \) if \( d_i^T Q d_j = 0 \) for \( i \neq j \).

If \( Q = I \), conjugacy is equivalent to the usual notion of orthogonality.

**Theorem 1:** If \( Q \) is positive definite and the vectors \( d_i \) are mutually conjugate, then these vectors are linearly independent.
The next step will be to construct a set of mutually conjugate directions. For a given set of \( N \) linearly independent vectors \( v_0, v_1, \ldots, v_{N-1} \), it is possible to construct a set of conjugate directions \( d_0, d_1, \ldots, d_{N-1} \) in the following way. Let

\[
   d_0 = v_0
\]

and then define

\[
   d_i = v_i + \sum_{j=0}^{i-1} a_{ij} d_j
\]

for \( i = 1, \ldots, N-1 \). The coefficients \( a_{ij} \) are chosen so that \( d_i \) is conjugate to the previous directions \( d_{i-1}, \ldots, d_0 \). Multiplying eqn (16) by \( Qd_l \) for \( l = 0, \ldots, i-1 \) gives

\[
   d_i^T Qd_i = v_i^T Qd_i + \sum_{j=0}^{i-1} a_{ij} d_j^T Qd_i = 0.
\]

If all the previous \( a_{ij} \)s have been chosen so that \( d_0, \ldots, d_{i-1} \) are conjugate, then

\[
   d_j^T Qd_l = 0, j \neq l
\]

and from eqn (17) it follows

\[
   a_{ij} = \frac{v_i^T Qd_j}{d_j^T Qd_j}
\]

for all \( i = 0, \ldots, N-1 \) and \( j = 0, \ldots, i-1 \).

The set of vectors \( d_0, \ldots, d_{N-1} \) defined by eqns (15)-(19) is conjugate with respect to \( Q \).

The above procedure can now be used to develop the conjugate-gradient method. The first step is to let

\[
   v_0 = -g_0
\]

i.e. the initial step is in the direction of the negative gradient of \( J \) and it is identical to a steepest descent direction. The rest of the \( v \)s are chosen as \( v_1 = -g_1, \ldots, v_{N-1} = -g_{N-1} \). Using eqn (13) one gets

\[
   x_1 = x_0 + \alpha_0 d_0.
\]

The next conjugate direction \( d_1 \) can be found by using eqns (16) and (19) with \( v_1 = -g_1 \).

\[
   d_1 = -g_1 + \frac{g_1^T Qd_0}{d_0^T Qd_0} d_0
\]
From eqn (12)

\[ g_{k+1} - g_k = Q(x_{k+1} - x_k) \]

and using eqn (13)

\[ g_1 - g_0 = Q(x_1 - x_0) = \alpha_0 Q d_0. \]

eqn (12) can now be written as

\[ d_1 = -g_1 + \frac{g_1^T (g_1 - g_0)}{d_1^T (g_1 - g_0)} d_0. \]

The process can be repeated and the result is

\[ d_k = -g_k + \sum_{j=0}^{k-1} \frac{g_k^T (g_{k+1} - g_j)}{d_j^T (g_{k+1} - g_j)} d_j. \]

This expression can be simplified using the fact that

\[ g_j^T d_j = 0 \]

for \( j = 0, \ldots, k - 1 \). The subspace spanned by \( d_0, \ldots, d_{k-1} \) is also the subspace spanned by \( g_0, \ldots, g_{k-1} \) and hence

\[ g_j^T g_j = 0 \]

for \( j = 0, \ldots, k - 1 \). Equation (26) now reduces to

\[ d_k = -g_k + \beta_k d_{k-1} \]

where

\[ \beta_k = \frac{g_k^T (g_k - g_{k-1})}{d_{k-1}^T (g_k - g_{k-1})}. \]

Again using eqns (27), (28) and (29) for \( d_{k-1} \), the expression for \( \beta \) can be rewritten as

\[ \beta_k = \frac{g_k^T g_k}{d_{k-1}^T d_{k-1}}. \]

It is important to note that in order to calculate the new direction \( d_k \), it is only necessary to know the current and the previous gradients of \( J \) and the previous direction \( d_{k-1} \).

There is one problem with using the conjugate-gradient algorithm described above. If
the number of variables is large as in oceanographical problems, the conjugate-gradient method can produce a direction of search after a few iterations which is not very efficient due to loss of conjugacy. There have been several attempts to avoid this effect. One of the first applications used periodic restarts with the steepest descent direction. However, there is a disadvantage using this method. The reduction at the restart iteration is often poor compared with the reduction that would have occurred without restarting. BEALE (1972) proposed another restart procedure. He showed that using the direction

$$d_k = -g_k + \beta_k d_{k-1} + \gamma_k d_r$$

which is similar to the conjugate-gradient direction with the addition of a multiple of the restart direction $d_r$. In this method $\beta_k$ and $\gamma_k$ are given by

$$\beta_k = \frac{(g_{k-1} - g_k)^T g_{k+1}}{(g_{k+1} - g_k)^T g_k}$$

$$\gamma_k = \frac{(g_{k-1} - g_k)^T g_{k+1}}{(g_{k+1} - g_k)^T d_r}.$$

$d_r$ is the last direction of the previous conjugate-gradient cycle along which a linear search was made. POWELL (1977) suggested a condition for when a restart should be performed. If the inequality

$$|g_k^T g_{k+1}| \geq 0.2 \|g_{k+1}\|$$

holds or if there have been $N$ linear searches, the conjugate gradient should be restarted. A restart should also take place if the direction of search is not sufficiently downhill, where an adequate downhill direction is one that satisfies the two inequalities

$$g_{k+1}^T d_{k+1} \leq -0.8 \|g_{k+1}\|^2$$

$$g_{k+1}^T d_{k+1} \geq -1.2 \|g_{k+1}\|^2.$$

This means that frequent restarts will take place.

3.3. Newton, quasi-Newton and limited-memory quasi-Newton methods

Newton’s method is based on the idea that the function $J$ which is to be minimized is approximated by a quadratic function. If $x^*$ is the minimum of the function, then a Taylor series expansion of $J$ near a point $x_k$ is given by

$$J(x_k + p) \sim J(x_k) + g_k^T p + \frac{1}{2} p^T Q(x_k) p$$

where as before $g_k = \nabla J(x_k)$ and $p = x^* - x_k$, the distance to the minimum. The right hand
side is minimized if \( p_k \) satisfies the system

\[
Q_k p_k = -g_k
\]

which gives the Newton method. A descent algorithm using Newton's method is therefore given by

\[
x_{k+1} = x_k - \alpha_k Q_k^{-1} g_k
\]

where \( Q_k^{-1} \) is the inverse Hessian matrix. It is also assumed that \( Q_k^{-1} \) exists and that the direction \( p_k \) determined from eqn (39) is a descent direction, i.e. \( p_k^T g_k < 0 \). If the Hessian matrix \( Q_k \) is positive definite, the direction \( p_k \) is clearly a descent direction, since

\[
p_k^T g_k = -g_k^T Q_k^{-1} g_k < 0.
\]

On the other hand if \( Q_k \) is not positive definite, the quadratic function need not have minimum or even a stationary point. To be able to use this method it is necessary to know the zero, first and second derivatives of \( f(x) \) at any point.

Quasi-Newton methods are based on the assumption that information about the curvature (the Hessian) of the quadratic function can be obtained as the iterations of a descent method proceed using the observed behavior of \( f \) and \( g \). The whole idea is therefore to compute an approximation to the curvature of the function without actually forming the Hessian matrix. A quasi-Newton iteration method can be described by

\[
x_{k+1} = x_k + \alpha_k d_k
\]

\[
B_k d_k = -g_k
\]

where \( B_k \) is an approximation to the Hessian matrix. At the first iteration it is usual to take \( B_0 \) as the identity matrix. After each iteration \( B_k \) is updated to obtain a new approximation of the Hessian, \( B_{k+1} \)

\[
B_{k+1} = B_k + U_k
\]

where \( U_k \) is the update matrix. The new approximation \( B_{k+1} \) is required to satisfy the quasi-Newton condition

\[
B_{k+1} s_k = y_k
\]  \hspace{1cm} (44)

where

\[
s_k = x_{k+1} - x_k
\]

the change in \( x \) during the \( k \)-th iteration, and

\[
y_k = g_{k+1} - g_k
\]

the change in the gradient.

There are several different ways of updating the Hessian matrix. The update formula
used by Shanno and Phua (1980) is the BFGS formula
\[
B_{k+1} = B_k - \frac{B_k s_k y_k y_k^T B_k}{s_k^T B_k s_k} + \frac{y_k y_k^T}{y_k^T s_k}
\]
where \(s_k\) and \(y_k\) are given by eqns (45) and (46), respectively. The derivation of eqn (47)
can be found in e.g. Luenberger (1984).

The idea behind the limited-memory (memoryless) quasi-Newton methods is that the
Hessian matrix is not computed, but a limited number of quasi-Newton corrections are
added to the identity matrix. The additional storage required is only the vectors defining
the updates. Different methods exist depending on the number of vectors used in the
updating. The method used by Shanno (1978a, b) is defined by
\[
d_{k+1} = -g_{k+1} \left[ \left(1 + \frac{y_k^T y_k}{p_k^T y_k} \right) \frac{y_k^T g_{k+1}}{p_k^T y_k} - \frac{y_k^T g_k + 1}{p_k^T y_k} \right] p_k + \frac{p_k^T g_{k+1}}{p_k^T y_k} y_k
\]
where
\[
p_k = x_{k+1} - x_k.
\]

3.4. Computing the step length

In order for the descent method to converge, it is important that the step length
produces a “sufficient decrease” in \(J\) at each iteration. The “sufficient decrease” may be
satisfied by several different conditions on \(\alpha_k\). An exact line search would have given
\[
\min_{\alpha} J(x_k + \alpha d_k)
\]
but in general the solution of the nonlinear function \(\partial J/\partial \alpha_k = 0\) cannot be implemented in
a finite number of operations.

The Goldstein-Armijo principle is one method for choosing the step length \(\alpha_k\) so that
a sufficient decrease in \(J\) is achieved (Goldstein and Price, 1967). The principle states
that \(\alpha_k\) should be chosen so that the following is true
\[
0 < -\mu_1 \alpha_k g_k^T d_k \leq J(x_k) - J(x_k + \alpha_k d_k) \leq -\mu_2 \alpha_k g_k^T d_k
\]
where \(\mu_1\) and \(\mu_2\) are scalars satisfying
\[
0 < \mu_1 \leq \mu_2 \leq 1.
\]
The upper and lower bounds of (51) ensure that \(\alpha_k\) is neither “too large” nor “too small”. It
is important to realise that eqn (51) alone does not guarantee a good value of \(\alpha_k\). For
instance, choosing \(\alpha_0\) as \(10^{-5}\) would satisfy eqn (51) for almost all functions encountered
in practice, if \(\mu_1\) and \(\mu_2\) are appropriate small values. This would be an “efficient” way to
find the \(\alpha_k\) so that a suitable \(\alpha_k\) would be found with only a single function evaluation per
iteration, but the descent method would be extremely inefficient if such a step length was chosen. In choosing the step length it is therefore essential to consider the performance of the algorithm not only in terms of function evaluations per iteration, but also in terms of the decrease in $J$ at each iteration.

Another way of interpreting $\alpha_k$ is in terms of univariate minimization which requires that the magnitude of the directional derivative at $x_k + \alpha_k d_k$ be sufficiently reduced from that at $x_k$

$$|g(x_k + \alpha_k d_k)^T d_k| \leq -\eta g^T d_k$$

where

$$0 \leq \eta \leq 1$$

Since eqn (53) does not involve the change in $J$, it is not adequate to ensure a sufficient decrease. Many step length algorithms therefore include the following conditions (GILL, SAUNDERS and WRIGHT, 1982)

$$J(x_k) - J(x_k + \alpha_k d_k) \geq -\mu g^T d_k$$

where

$$0 < \mu \leq \frac{1}{2}.$$  

The method for finding the optimal step length which will be followed here is a method used by DERBER (1985). The first step is to use an initial guess for the step length and then update the parameters of interest according to

$$c_{k+1} = c_k + \alpha_k d_k$$

where $\alpha_k$ is an initial guess for the step length. The model can then be integrated forward with the new estimate of the parameters. The goal will be to maximize the reduction in the cost function between the two model runs.

$$\Delta J = J(x_k, c_k) - J(x_k, c_k + \alpha_k d_k).$$

Previous experience will give some idea of what the initial value of the step length should be.

The new value of the parameters produce a perturbation in the model representation of the observations. Let $\delta x_k$ denote the perturbation. Using the definition of the cost function in eqn (10), the perturbation of the cost function as a result of the perturbation to the model representation, leads to

$$\delta J = \frac{1}{2}(\delta x_k^T K_x \delta x_k + 2(x_k - x')^T K_x \delta x_k).$$

Assume that another choice of the step length, $\alpha_k$, would result in a perturbation $\delta x_k'$ of
the model field

\[ \delta x_k = \frac{\alpha_k}{\alpha_i} \delta x_k \]

For linear dynamics (as for the case studied here) this would be exactly true, but for nonlinear dynamics it may not be a good approximation. Derber (1985) used the approximation for a nonlinear atmospheric model, and for the cases he studied, eqn (60) worked satisfactorily. Equation (59) now becomes

\[ \delta J = \frac{1}{2} \left( \frac{\alpha_k^2}{\alpha_i} \delta x_k^T K_x \delta x_k + \frac{\alpha_k}{\alpha_i} (x_k - x')^T K_x \delta x_k \right) \]

where the prime on \( \delta x_k \) has been dropped. The goal is to maximize \( \delta J \), i.e. the derivative of \( \delta J \) with respect to \( \alpha_k \) is set equal to zero.

\[ \frac{\partial \delta J}{\partial \alpha_k} = 0 \]

which gives

\[ \alpha_k = -\alpha_i \frac{(x_k - x')^T K_x \delta x_k}{\delta x_k^T K_x \delta x_k} \]

for the optimal step length. In later iterations \( \alpha_i \) is replaced by \( \alpha_{k-1} \).

The limited-memory quasi-Newton conjugate gradient algorithm of Shanno and Phua (1980) is implemented in the experiments described later. A few modifications of the routine have been performed. The step length is no longer computed as in the original code, but instead eqn (63) is used to calculate the new step length. In the original algorithm the cost function and its gradient are given in a function call. In the modified version the values of the cost function and its gradient are calculated and written to a file, which is read in a subroutine.

4. PARAMETER ESTIMATION

If the parameter estimation is going to be successful, an important question must be addressed; under what condition can one expect the proposed estimation method to give unique and stable results? Similar problems have been studied in several papers concerning groundwater flow, e.g. Chavent, Dupuy and Lemonnier (1975), Neuman (1980), Yakowitz and Duckstein (1980), Carrera and Neuman (1986a, b, c) and Yeh (1986). Panchang and O'Brien (1988) studied the estimation of the friction factor in tidal rivers. In these papers which consider both time-dependent and steady state flows, important aspects about data assimilation and parameter estimation are discussed. There are four important terms which must have a clear definition: well-posedness, uniqueness, identifiability and stability. These definitions will be given in the following paragraphs.
Strictly speaking, if the method is going to work, the problem should be well-posed. The inverse problem is often ill-posed, but there are situations under which a meaningful solution can be found, although in a limited sense. It will therefore be important to be able to recognize the circumstances which allow a solution of the problem to be found. The inverse problem will then be to determine $c^2$ on the basis of $p(x, y, t)$ and the inverse relationship $c^2(x, y, t) = R(p(x, y, t))$. This problem is said to be well-posed if and only if (1) to every $p$ there corresponds a solution $c^2(x, y, t)$; i.e. a solution exists; (2) the solution is unique for any given $p(x, y, t)$; and (3) the solution depends continuously on $p(x, y, t)$; i.e. the solution is stable. If the inverse problem fails to satisfy one or more of these three requirements, it is said to be ill-posed.

Uniqueness and parameter identifiability (Kitamura and Nikagiri, 1977) can be defined in the following way. Let $p_1(x, y, t) = F(c^2_1(x, y, t))$ and $p_2(x, y, t) = F(c^2_2(x, y, t))$ be two solutions of the forward problem, and $c^2_1(x, y, t) = R(p_1(x, y, t))$ and $c^2_2(x, y, t) = R(p_2(x, y, t))$ be two solutions of the inverse problem, then

$$\|p_1(x, y, t) - p_2(x, y, t)\|_0 \Leftrightarrow \|c^2_1(x, y, t) - c^2_2(x, y, t)\| = 0$$

where $\| \|$ represents a norm over the appropriate space. In practical problems $p$ is only given at discrete points in space and time, and $R$ represents a minimization of a functional $J$ as given in Section 2. While uniqueness refers to the inverse problem, $R$, identifiability refers to the forward problem, $F$. If two sets of parameters lead to the same function $p(x, y, t)$, the parameters are said to be unidentifiable. Uniqueness on the other hand is concerned with the problem whether different parameters may be found from a given $p(x, y, t)$, if so the parameters are nonunique.

Stability can be defined in the following way. For every $\varepsilon > 0$ there exists a $\delta$ such that for $c^2_1(x, y, t) = R(p_1(x, y, t))$ and $c^2_2(x, y, t) = R(p_2(x, y, t))$ one has

$$\|p_1(x, y, t) - p_2(x, y, t)\| < \delta \Rightarrow \|c^2_1(x, y, t) - c^2_2(x, y, t)\| < \varepsilon$$

Equation (65) states that small errors in the variables must not lead to large changes in the computed parameter. Tarantola (1987) discusses the concepts of uniqueness and stability in more detail.

Parameter estimation in groundwater flow considers parabolic partial differential equations, since this is the type of equations hydrologists like to believe governs groundwater flow. Kitamura and Nikagiri (1977) studied the identifiability of continuously varying and constant parameters in a linear one-dimensional parabolic partial differential equation

$$\frac{\partial u}{\partial t} = \frac{\partial}{\partial x} \left[ a(x) \frac{\partial u}{\partial x} \right] + b(x)u + f(x, t)$$

where $a(x)$ and $b(x)$ represent the parameters. They assume that the parameters in the system equation are unknown but continuous, together with their first (and for $a(x)$, second) derivatives.
second) derivatives, over the whole flow domain. They also assume that the boundary conditions are known. Their results show that the parameters are only identifiable under certain fairly restrictive conditions.

The problems of identifiability and stability can be solved by reducing the number of parameters to be estimated. In the hydrological literature the most common way to do this is to approximate the parameters by a known class of functions depending on a finite number of parameters. If the parameter dimension is not reduced, it may be difficult or even impossible to determine the spatial structure of the parameter field. Instability is often characterized by the fact that during the solution process the parameter values are bouncing back and forth between high and low values.

As mentioned in Section 2.2, the last term in the cost function

\[ J(x, c) = \frac{1}{2}(x - x')^T K_x (x - x') + \frac{1}{2}(c - c')^T K_c (c - c') \]

is added because the interest of this research is parameter estimation, and the term makes sure that the new estimate of the parameters is not too far from the initial guess. In the hydrologist’s nomenclature the term represents prior information about the parameters. Uniqueness of the solution implies that the cost function \( J \) is convex. The requirements for this condition is that the Hessian matrix is positive definite, see e.g. GILL, MURRAY and WRIGHT (1981). Writing the cost function as

\[ J = J_x + J_c \]

where \( J_x \) represents the first term in eqn (67), while \( J_c \) represents the last term or the term representing prior information about the parameters. The Hessian is represented by

\[ \frac{\partial^2 J}{\partial c^2} = \frac{\partial^2 J_x}{\partial c^2} + \frac{\partial^2 J_c}{\partial c^2} \]

The first term gives

\[ \frac{\partial^2 J_x}{\partial c^2} = K_x \left[ \left( \frac{\partial x}{\partial c} \right)^T \left( \frac{\partial x}{\partial c} \right) + (x - x')^T \frac{\partial^2 x}{\partial c^2} \right] \]

which can be positive or negative. So the first term in the minimization criterion does not guarantee that the cost function is convex. The last term in eqn (69) gives

\[ \frac{\partial^2 J_c}{\partial c^2} = K_c \]

which is clearly a positive term. Adding prior information about the parameters therefore increases the chance that the cost function will be convex. Of course there is no guarantee that the term in eqn (71) will make the Hessian positive definite. CARRERA and NEUMAN (1986b) discuss the effect of prior information in a few simple examples of
estimation of aquifer parameters. Their examples clearly show that the addition of prior information lead to unique solutions.

5. SPECIFICATION OF THE PROBLEM

In this section the variational formalism developed in Section 2 is applied to a model of the equatorial Pacific Ocean. The governing equations are presented and the adjoint equations derived. The set of equations are able to assimilate sea level observations into the Pacific Ocean model, and the focus of this study is on the ability of sea level observations to adjust (estimate) parameters in the model. The parameter which is estimated is the phase speed. In fact it is the square of the phase speed which is estimated, but the term phase speed will be used. The method is not restricted to the use of only sea level observations, but could easily be modified so it could assimilate other types of observations e.g. velocity data. However, this will not be considered in the present study. It is also possible to adjust other parameters than the phase speed, e.g. the eddy viscosity or even adjusting the windstress, i.e., the forcing function, so the model solution fits the observations in an optimal way. In the initial tests of the data assimilation algorithm, it is important to keep the problem under consideration as simple as possible, so the focus of this research will be on estimating only one parameter, the phase speed. Earlier work has shown that it is sometimes necessary to reduce the number of degrees of freedom in order to be able to find unique and stable parameters. The important physics in the model under consideration consist of equatorial waves propagating across the Pacific Ocean, and to reduce the number of degrees of freedom it is assumed that the parameter is a function of longitude only.

5.1. The numerical model

The model used is a reduced-gravity, linear transport model for the equatorial Pacific Ocean. Due to the large latitudinal extent, spherical coordinates are used, with \( \phi \) (longitude) increasing eastward and \( \theta \) (latitude) increasing northward. In the reduced gravity model the ocean is assumed to consist of two layers of slightly different density \( (\rho_1, \rho_2) \), with the interface being an approximation of the pycnocline. The lower layer of the model is dynamically inactive, while the motion of the upper layer represents the first baroclinic mode. Similar models (but in an equatorial \( \beta \)-plane) have been successfully used to simulate the ocean circulation in the equatorial regions (e.g. ADAMEC and O'BRIEN, 1978; CANE, 1979; BUSALACCHI and O'BRIEN, 1980).

Let \( U = uh \) and \( V = vh \) represent the eastward and northward component of upper-layer transport, respectively, where \( (u, v) \) are the depth-independent \( (\phi, \theta) \) velocity components in the upper layer and \( h \) the thickness of the upper layer. Figure 1 shows a schematic picture of the reduced gravity model. The model equations are:

\[
\frac{\partial U}{\partial t} - fV = -
\]
Variational data assimilation

Fig. 1. The geometry of the reduced gravity model.

\[
\begin{align*}
\frac{\partial V}{\partial t} + fU &= -\frac{c^2}{a} \frac{\partial H_{06}^2}{\partial \theta} \tau^\theta + A \nabla^2 V \\
\frac{\partial h}{\partial t} + \frac{1}{a \cos \theta} \left[ \frac{\partial U}{\partial \phi} + \frac{\partial}{\partial \theta} (V \cos \theta) \right] &= 0
\end{align*}
\]

where \( \nabla^2 \) is given by the simplified form

\[
\nabla^2 = \frac{1}{a^2 \cos^2 \theta} \left[ \frac{\partial^2}{\partial \phi^2} + \cos \theta \frac{\partial}{\partial \theta} \left( \cos \theta \frac{\partial}{\partial \theta} \right) \right]
\]

and \( c^2 = gH [(\rho_2 - \rho)/\rho] \) is the barotropic gravity wave speed, which is the parameter to be estimated. It is assumed that \( c^2 \) is a function of longitude only, \( c^2 = c^2(\phi) \). The densities of the lower and upper layer are \( \rho_2 \) and \( \rho \), respectively, while \( a \) is the radius of the earth. The Coriolis parameter \( f = 2\omega \sin \theta \) where \( \omega \) is the rotation rate of the earth and \( A \) is a horizontal kinematic eddy viscosity. The wind stress \( \tau = (\tau^x, \tau^\theta) \), is applied as a body force over the upper layer. The values used for some of the different parameters are: \( A = 2000 \text{ m}^2 \text{s}^{-1} \), \( \omega = 0.729 \times 10^{-4} \text{ s}^{-1} \), \( a = 6378 \text{ km} \). The initial thickness of the upper layer is 300 m.

The model geometry extends from 120°E to 75°W and from 20°S to 25°N as in Kubota and O'Brien (1988), see Fig. 2. The effect of the coast at the eastern and the western end of the basin is included in the model by simulating the coastline as a no-slip boundary, i.e. \( U = V = 0 \). The northern and the southern boundaries are open and the open boundary condition described by Camerlengo and O'Brien (1980) is applied along these boundaries.
Fig. 2. The Pacific Ocean model domain. The southern and the northern boundaries are open. The position of the sea level stations used in the assimilation is also shown.

\[ h_{i-1,j+1} u_{i-1,j+1} h_{i,j+1} u_{i,j+1} h_{i+1,j+1} u_{i+1,j+1} \]

\[ v_{i-1,j+1} v_{i,j+1} \]

\[ h_{i-1,j} u_{i,j} h_{i+1,j} u_{i+1,j} \]

\[ \Delta \phi \]

\[ \Delta \theta \]

\[ h_{i-1,j-1} u_{i-1,j-1} h_{i,j-1} u_{i,j-1} h_{i+1,j-1} u_{i+1,j-1} \]

\[ v_{i-1,j-1} v_{i,j-1} \]

Fig. 3. The Arakawa C grid.

In solving eqns (72)–(74) the staggered C-grid of ARAKAWA (1966) is used, see Fig. 3. The gridspacing is \( \frac{1}{2}^\circ \) in longitude and \( \frac{1}{4}^\circ \) in latitude. The equations are integrated in time using a leapfrog scheme, with a forward scheme used every 99th time step to eliminate the computational mode. The timestep is 40 min. A Dufort-Frankel scheme is used for the diffusive term.
5.2. The adjoint equations

The formalism developed in Section 2 can now be followed using the equations for the Pacific Ocean model described above as constraints. It is necessary to define a cost function $J$, which measures the distance between the observations and the model solution. $J$ will be chosen as the quadratic function

$$J(h, c^2) = \int_{\mathcal{Z}} \left[ \frac{K_h}{2}(h-h')^2 + \frac{K_c}{2}(c^2-c'^2)^2 \right] d\sigma$$

where $h'$ represents an observation of the upper layer thickness and $c'^2$ is an estimate of the phase speed. It has been assumed that the observational errors are uncorrelated. The coefficients $K_h$ and $K_c$ are validity coefficients and they were assumed to be constants in the experiments described later. The spatial and temporal domain over which the Pacific model is integrated is represented by $\mathcal{Z}$. The last term in (76) makes sure that the final value of the phase speed is not too far from the initial estimate. This is important since there are certain physical limits for $c^2$. The initial estimate should be a reasonable one so that the minimization process just has to make a small correction to the initial guess.

The next step is to form the associated Lagrange function

$$L(U, V, h, \lambda_u, \lambda_v, \lambda_h, c^2) =$$

$$\int \left[ \frac{\partial U}{\partial t} - fV + c^2 ight]$$

$$+ \int_{\mathcal{Z}} \lambda_u \left[ \frac{\partial V}{\partial t} + fU + \frac{c^2}{a} \frac{\partial h}{\partial \theta} - \frac{\tau^\theta}{\rho} - A \nabla^2 V \right] d\sigma$$

$$+ \int_{\mathcal{Z}} \lambda_h \left[ \frac{\partial h}{\partial t} + \frac{1}{a \cos \theta} \left( \frac{\partial U}{\partial \phi} + \frac{\partial V}{\partial \theta} (V \cos \theta) \right) \right] d\sigma$$

$$+ \int_{\mathcal{Z}} \frac{K_h}{2} (h-h')^2 d\sigma + \int_{\mathcal{Z}} \frac{K_c}{2} (c^2-c'^2)^2 d\sigma$$

where $\lambda_u, \lambda_v$ and $\lambda_h$ are the Lagrangian multipliers for $U, V,$ and $h$, respectively. The stationary points of the Lagrange function (which correspond to the minimum of the cost function) can be found by letting the first variation of $L$ with respect to the variables $U, V, h, \lambda_u, \lambda_v, \lambda_h,$ and $c^2$ vanish. The first variation of $L$ with respect to $\lambda_u, \lambda_v$ and $\lambda_h$ gives the following equations:

$$-f = \frac{\partial}{\partial t} \left( \frac{c^2}{a \cos \theta} \frac{\partial h}{\partial \phi} + \frac{\tau^\theta}{\rho} + A \nabla^2 U \right)$$

$$\frac{\partial V}{\partial t} + fU = -\frac{c^2}{a} \frac{\partial h}{\partial \theta} + \frac{\tau^\theta}{\rho} + A \nabla^2 V$$
\[ \frac{\partial h}{\partial t} + \frac{1}{a \cos \theta} \left[ \frac{\partial U}{\partial \phi} + \frac{\partial}{\partial \theta} (V \cos \theta) \right] = 0 \]

which are identical to the original model eqns (72)-(74). Letting the first variation of \( L \) with respect to \( U, V \) and \( h \) vanish, gives the adjoint equations:

\[ \frac{\partial \lambda_u}{\partial t} + f \lambda_v = \frac{c^2}{a \cos \theta} \frac{\partial \lambda_h}{\partial \phi} + A \nabla^2 \lambda_u \]

\[ \frac{\partial \lambda_v}{\partial t} - f \lambda_u = \frac{c^2}{a \cos \theta} \frac{\partial \lambda_h}{\partial \phi} + A \nabla^2 \lambda_v \]  \hspace{1cm} (82)

\[ \frac{\partial \lambda_h}{\partial t} \frac{1}{a \cos \theta} \left[ \frac{\partial \lambda_u}{\partial \phi} + \frac{\partial}{\partial \theta} (\lambda_v \cos \theta) \right] + K_A(h - h') = 0 \]  \hspace{1cm} (83)

It is worth noting that the adjoint equations are forced by the data misfits represented by the last term in eqn (83). The windstress does not enter into these equations. Comparing eqns (81)-(83) to the original model eqns (78)-(80) one can see they have a similar form, except that the adjoint equations correspond to an evolution backward in time. The Lagrangian multipliers will therefore propagate information about the data misfits back to the initial time of the data assimilation period. Since the system of equations governing the Lagrangian multipliers are similar to the model equations, the information will propagate in the form of equatorial Kelvin and Rossby waves. However, in the adjoint equations the equatorial Kelvin wave will have a westward phase propagation, while the Rossby waves will have an eastward phase propagation.

In the derivation of the adjoint equations, the natural boundary conditions for the adjoint variables follow. The “initial” conditions at \( t = T \) for the adjoint variables are homogeneous ones, namely \( \lambda_u = \lambda_v = \lambda_h = 0 \). The adjoint equations are forced by the data misfits, and after the last computational time level for the physical model there are no data. \( \lambda_u = \lambda_v = \lambda_h = 0 \) is therefore the natural initial condition for the adjoint variables. At the solid boundaries in the western and eastern part of the basin, the boundary conditions are given by \( \lambda_u = \lambda_v = 0 \). These conditions are identical to the no-slip conditions used in the model equations, \( U = V = 0 \) at the solid boundaries. Using the open boundary conditions of CAMERLENGO and O'BRIEN (1980) for \( U \) and \( V \), the conditions for the adjoint variables can be derived. It turns out that identical open boundary conditions must be used for \( \lambda_u \) and \( \lambda_v \).

Letting the first variation of \( L \) with respect to \( c^2 \) vanish

\[ \int_\vartheta \int_\phi \left[ \frac{\lambda_u}{c^2 a \cos \theta} \frac{\partial h}{\partial \phi} + \frac{\lambda_v}{c^2 a \cos \theta} \frac{\partial h}{\partial \theta} \right] d\theta dt + TL \cdot K_A(c^2 - c^2) = 0 \]  \hspace{1cm} (84)
a solution which is not the one which is sought. The model equations are solved using finite differences. One has to be careful when going from the continuous equations to the finite difference representation. Deriving the adjoint of the continuous equations and then taking the finite difference of the result may not give the desired adjoint, which is the adjoint of the finite difference version of the model. Details of the derivation of the adjoint equations can be found in Smedstad (1989). In the numerical solution of the problem, the equations were rescaled so that only the ratio between $K_s$ and $K_c$ occurred. Experiments using acceptable values for this ratio did not change the results (the model under consideration is linear). As a result the ratio $K_s/K_c$ was chosen to be unity.

The temporal domain in which the cost function (76) is minimized must be determined. The spatial domain is given by the spatial domain of the model under consideration. It is not so clear what to choose for the time interval. Numerical models are not perfect representations of the real world, and it would be meaningless to try to assimilate data over too long time intervals. Different models will require different assimilation periods. The time scales of the model are an important factor. For instance in a meteorological forecast model it would be natural to choose an assimilation period of a few days, since the forecast is not very accurate after this period. On the other hand, for an oceanographic model such as the reduced gravity model used in this study, it would not be correct to choose a period of a few days. The major time scales of the motion in this type of model are of the order of months. A natural assimilation period would therefore be of the order of months too. In the initial experiments described in the next section a period of six months was chosen. When the model is driven by real wind forcing, it is important that the information have time to propagate throughout the model. A Kelvin wave takes about 3 months to cross the Pacific Ocean, while the first symmetric mode Rossby waves need 9 months to cross. A data assimilation period of 12 months was therefore chosen in this case.

5.3. The wind forcing

The wind stress data used as forcing of the Pacific Ocean model come from ship observations. The individual wind observations have been grouped into monthly values on a $2 \times 2^\circ$ grid (see Legler and O'Brien, 1986, for more details). From this coarse grid the data are interpolated to the model grid using a cubic spline interpolation. Figure 4 shows an example of the interpolated wind stress for June 1979. The quantity plotted is the pseudostress, i.e., the magnitude of the wind velocity times the component. Windstress is determined by multiplying the pseudostress by the density of air and a drag coefficient.

5.4. The observations

Sea surface elevation (SSE) observations from island stations (provided by the University of Hawaii) will be assimilated into the model. The model does not include sea level as one of the dependent variables, and it will therefore be necessary to transform SSE to upper layer thickness (ULT). The SSE data will be transformed to ULT by
dividing by the density difference between the two layers of the model, i.e

\[
\text{ULT} = \frac{\text{SSE}}{\Delta \rho / \rho}
\]

Time series of daily observed SSE from the island stations are used. These observations will contain both high and low frequencies. The main interest of this study is to be able to determine the basic stratification of the Pacific Ocean, and the high frequencies will not give important information for this purpose. The time series are therefore filtered using a low pass filter before the data are assimilated into the model. As mentioned above the wind stress used to force the Pacific Ocean model, is a monthly average. The shortest time scale resolved by the model is therefore 2 months. The sea level data are as a consequence low pass filtered with a pass period of 60 days. As an example of the time series of sea level observation converted into ULT, Fig. 5a shows ULT as a function of time for the station at Santa Cruz. The figure shows the daily observation for the period 1978–1983. Figure 5b is a plot of the filtered time series for the same period, using a pass period of 60 days. Missing data have been filled in by linear interpolation or if data are missing at the beginning or end of the time series, persistence is used. This will not influence the assimilation in this research, since the periods considered here did not have any missing data points. Before the difference between ULT from the model and the observations was calculated, the mean was removed from both time series.

The Pacific Ocean model saves time series of the height field once a day for the island stations in Fig. 2. The \( U, V \) and \( H \) fields are saved every 6 days for the whole domain. Similarly for the adjoint equations, \( \lambda_u, \lambda_v \) and \( \lambda_h \) fields are saved every 6 days. The \( h, \lambda_u \) and \( \lambda_v \) fields are used when the gradient of \( L \) with respect to \( c^2 \) is calculated by eqn (84). Saving the variables every 6 days may seem not to be often enough. But as mentioned earlier, the main interest of this study is to be able to determine the large scale variation in the Pacific Ocean and the time integral in eqn (84) will be adequately approximated using data every 6 days.

The adjoint equations are forced by the difference between the model solution and the observations, the data misfits. The timestep in the adjoint model is the same as in the
Fig. 5. Time series of upper layer thickness (ULT) for the station at Santa Cruz. ULT in meters is plotted as a function of time. (a) Daily observations of ULT. (b) The filtered time series using a pass period of 60 days.
original model, 40 min. In order to obtain a smooth forcing of the adjoint equations, the data misfits are linearly interpolated so that the forcing is a smooth function of time.

6. DATA ASSIMILATION IN THE PACIFIC OCEAN MODEL

As initial experiments of data assimilation in the Pacific Ocean model, a few simple examples are discussed. The primary objective of this research is to be able to determine the phase speed in the model. Adjustment to the external forcing is not considered here, and in the first experiments there is no wind stress applied. The Pacific Ocean model is instead initialized by the height and velocity field of a Kelvin wave. It is important to be able to check the performance of the data assimilation algorithm. Identical twin data are therefore used in the preliminary experiments, that is, the "observations" are results from the model. The simple set up of an initial Kelvin wave gives a nice picture of how the adjoint equations propagate information throughout the model domain. The number of waves present is limited, and it is relatively easy to explain what is happening.

![Plot of the upper layer thickness (ULT) and velocity field from the Pacific Ocean model. A phase speed of $c = 6.0 \text{ m s}^{-2}$ was used. The initial ULT was 300 m. Contour interval is 10 m. The figure shows the results at (a) day 0, (b) day 60, (c) day 90, (d) day 180.](image-url)

Fig. 6.
After the initial test of the algorithm, the estimation of the phase speed in "real" model runs is considered. The Pacific Ocean model was spun up from rest using the wind stress described in Section 5.3. The integration started in 1972, and the model was run through 1983. As an initial guess for the parameter $c^2$, a value of 6.0 m$^2$/s$^{-2}$ was used. This value corresponds to the one used by Kubota and O'Brien (1988). A few experiments using identical twin data are considered initially. Since $c^2$ gives information about the basic stratification of the Pacific Ocean, it would be interesting to be able to determine $c^2$ for two different periods. First, a "normal" year is used as observations. The year 1979 was chosen to represent what is called the normal situation. Second, the situation in the Pacific Ocean changes dramatically during an El Niño event and to investigate the effect of a major climatic event on the estimated value of the parameter, 1982/1983 was chosen as an example of an El Niño year. The El Niño of 1982/83 is the strongest on record in this century.

6.1 The Pacific Ocean model initialized by a Kelvin wave

In the first experiment the Pacific Ocean model was initialized by the height and velocity field of a Kelvin wave. The model was integrated forward for 180 days using a
constant phase speed $c^2 = 6.0 \text{ m}^2 \text{s}^{-2}$. The results from this run will represent the “observations”. The value of the phase speed was then changed to $c^2 = 4.0 \text{ m}^2 \text{s}^{-2}$ and the model was again integrated forward for 180 days to create the model results. In Fig. 6a–d the results of the model using $c^2 = 6.0 \text{ m}^2 \text{s}^{-2}$ are shown. The figures show the upper layer thickness (ULT) and the velocity field. Upper layer thickness is represented by contour lines, while the velocity components $U$ and $V$ are represented by arrows. The initial Kelvin wave at day 0 can be seen in Fig. 6a. Figure 6b shows the wave at day 30. The wave has now hit the eastern boundary of the model, and coastal Kelvin waves are propagating northward and southward along the coastline. The open boundaries at the south and north boundary let the coastal Kelvin waves leave the model domain. The Rossby waves created by the reflection can also be observed. These waves can be more clearly seen in Fig. 6c, which shows the ULT and velocity field at day 90. The model results after 180 days are shown in Fig. 6d. The Rossby waves have now propagated into the middle of the basin. Similar results are obtained when a phase speed of $c^2 = 4.0 \text{ m}^2 \text{s}^{-2}$ is used. These results are not shown. Experiments using 360 days instead of 180 days as the assimilation period did not change the results in these initial experiments.

6.1.1. Experiment 1: Constant phase speed and observations everywhere. The experiment considered assumes the best possible situation which can occur, that is, obser-
Variational data assimilation

The observations of ULT are available everywhere in space. The observations are assumed to be perfect, i.e. they are not contaminated by observational errors. The question of what effect such errors will have on the estimation process will be addressed later.

The next step in the variational data assimilation algorithm is to calculate the data misfits between the two model runs described above. The linearly time interpolated data misfits (in meters) for the model gridpoint corresponding to the island station at Santa Cruz are shown as a function of time (days) in Fig. 7. The figure shows the misfits during the first iteration. Note that day 180 corresponds to the end of the integration period of the model, which is the "initial" time for the integration of the adjoint equations.

The adjoint equations are integrated "backward" for 180 days using the data misfits as forcing. In this case there is a forcing term at every gridpoint of the model. Figure 8a–d shows the time evolution of the fields of the Lagrangian variables $\lambda_u$, $\lambda_v$ and $\lambda_h$. In Figure 8a, the three fields are shown after the adjoint equations have been integrated for 30 days. This corresponds to day 150 in the forward model. As described earlier, in the adjoint model Kelvin and Rossby waves move in opposite directions compared to the model itself. From the figures it is apparent that the Rossby waves indeed move eastward (Fig. 8a and b), while the Kelvin wave caused by the reflection of the Rossby waves is seen to move westward, see Fig. 8c and d. It is interesting to note that the higher model Rossby waves propagate in front of the gravest mode Rossby wave in Fig. 8a.

The next step in the assimilation procedure is to calculate the gradient of the cost function with respect to the phase speed using eqn (84). In Fig. 9 the time and longitudinal variation are shown. The latitudinal dependence has been integrated out. The figure shows the results after the first iteration. It is worth noting how the information about the gradient propagates in time and space. Initially, that is at the end of the assimilation period, the information propagates from west to east with the Rossby wave speed. Calculating the phase speed of the propagation gives a value of $c = 0.64 \text{ ms}^{-1}$, which is very close to the mode 1 Rossby wave phase speed. The Rossby waves connected with the eastward propagation can be seen in Fig. 8a and b. After the information hits the eastern boundary, the information propagates westward as a Kelvin wave. Again calculating the phase speed from Fig. 9, the westward propagation can be found in Fig. 8c and d. The Kelvin wave is about to leave the eastern end of the basin in Fig. 8c, which shows the results at day 60. Looking at Fig. 9 again, day 60 is exactly when the information about the gradient starts to propagate westward. At day 0 the information has come all the way to 160°E. The initial conditions for the model, see Fig. 6a, show that the gradient of the Kelvin wave height field is close to zero west of 160°E. This gradient is a part of the integral determining the gradient of the cost function with respect to $c^2$. The information does not have a chance to propagate further west because of the chosen initial conditions. as will be discussed later, this has an important effect on the ability to determine the spatial structure of the phase speed. Of course, in a real situation this limitation will not occur.

The simplest test is to determine a constant phase speed. The space and time dependence of the gradient of $L$ is removed by integrating over the remaining space direction (longitudinal) and over the time interval of the data assimilation. Since $c^2$ and the gradient of the cost function is a constant, a steepest descent minimization algorithm was used. The step length was calculated according to eqn (63). After a new value of the phase speed was determined, the iterative procedure continued. The model was integrated forward using the new value of the phase speed and the adjoint model was
FIG. 8. Plot of the Lagrangian multipliers for the case where observations are available everywhere. The Lagrangian multipliers for the $h$ equation is shown using contours, while the Lagrangian multipliers for the $U$ and $V$ equations are represented by vectors. The scale of the vectors are shown in the lower left corner of each figure. Contour interval is 10.0. The figure shows the results at (a) day 150, (b) day 120, (c) day 60, (d) day 0.
integrated backward so that a new value of the gradient of the cost function could be calculated. A new minimization was performed and the convergence of the algorithm could be checked. As a convergence criterion \( \| g \| / \| g_0 \| < 1.0 \cdot 10^{-2} \) was used in this and the following experiments. The results of the iterative process can be seen in Fig. 10a-c. Figure 10a shows the value of the cost function, normalized by its initial value, as a function of the number of iterations. As can be observed during the first few iterations, there is a rapid decrease in the cost function. After 3 iterations \( J/J_0 \) has dropped to about \( 10^{-3} \), i.e. the cost function has been reduced by 3 orders of magnitude. In Fig. 10b the value of the gradient of the cost function, normalized by its initial value, is shown. The gradient also has a rapid decrease in magnitude during the first couple of iterations. After 4 iterations the value of the gradient has decreased by 3 orders of magnitude, just as the cost function itself. The change in the value of the parameter \( c^2 \) during the iterations is shown in Fig. 10c. After 4 iterations the value of \( c^2 \) is 6.0 m\(^2\) s\(^{-2}\), which is equal to the exact value.

Setting \( c^2 \) and thus the gradient of the cost function to be a constant, is the simplest scenario which can occur. To determine \( c^2 \) as a function of longitude is a much more difficult problem. Instead of treating the gradient of the cost function as a constant, the gradient is allowed to vary with longitude, but not with time. The time dependence of the gradient was therefore removed by integrating over the time interval of the assimilation. To find the optimal \( c^2 \) the modified version of the limited-memory conjugate gradient algorithm of Shanno and Phua (1980) described in Section 3, was used. This algorithm gave very good convergence rates for the experiments performed in this research. The algorithm converged in ten iterations or less for all the cases considered. Compared to the simple method of steepest descent, the number of iterations dropped by a factor of about three. Using the quasi-Newton method with BFGS-updates did not change the convergence rate appreciably. For most cases the limited-memory version of the algorithm converged faster. The difference between the two methods was never more than two iterations. The step length was calculated using the expression in eqn (63).

The main interest of this research is to determine the large scale variation of the phase speed which will be a representation of the basic stratification of the Pacific Ocean. The
short length scale variations in the estimated values of $c^2$ are therefore removed by applying a filter to the solution for $c^2$. If these short scale variations were not removed, the estimated values of $c^2$ varied rapidly with longitude. This kind of instability can be overcome by reducing the parameter dimension. Restricting the attention to the large scale variation of $c^2$ is one way of reducing the dimension of the parameter. The filter
applied to the phase speed reduced the resolution of the parameter to $\frac{1}{4}$ of the original resolution. The resolution is therefore 4° in longitude.

A smoothing penalty term was tested in the model, requiring a smooth $h$ field. The effect on the estimated phase speed was however minimal. The field of the Lagrangian multipliers became smoother, but when new values of the parameter were calculated, they still needed to be filtered. Another choice of penalty term may have given a different result, e.g. adding a penalty term acting on the parameter itself.

The results of the iterative process are shown in Fig. 11a and b and Fig. 12a–d. The algorithm converged after 7 iterations with a convergence criterion of $\|\epsilon\|/\|\theta_0\| < 1.0 \times 10^{-2}$. The cost function normalized by its initial value as a function of iterations, is shown in Fig. 11a. There is a rapid decrease during the first three iterations. After four iterations the value of the cost function has decreased an order of magnitude. The value does not decrease much during the next couple of iterations. In Fig. 11b the normalized value of the gradient of the cost function is shown. The gradient also experiences a rapid decrease in the beginning. After seven iterations the value of the gradient has decreased below the convergence criterion.

The evolution of the estimated value of the phase speed is shown in Fig. 12a–d. Figure 12a is a plot of the longitudinal variation of the phase speed after the first iteration. As can be seen there is a relatively strong correction in the eastern part of the basin, while the correction is getting smaller and smaller as the western end of the basin is
approached. West of about 160°E there is hardly any correction at all. The reason is as mentioned above, the initial conditions of this experiment. In the area west of 160°E the data misfits are basically equal to zero during the period of the data assimilation, which means that there is no direct forcing of the adjoint equations in the area. The dynamics of
the model can however propagate information into the region. The Lagrangian multipliers in Fig. 8d are not equal to zero west of 160°E. This does not on the other hand make the gradient of the cost function much different from zero, since the expression for the gradient (84) also contains the terms ∂h/∂φ and ∂h/∂θ which both are small in the area. It is therefore to be expected that the correction in the western part of the basin is smaller than in the eastern part.

Figure 12b shows $c^2$ after the second iteration. The correction in the eastern part of the basin has now given a maximum value of $c^2$ slightly higher than the “correct” value. In the western part the correction is still very small. After the fifth iteration, Fig. 12c, a flat area has developed in the eastern part with $c^2$ very close to the value used to create the observations ($c^2 = 6.0 \text{ m}^2 \text{s}^{-2}$). Because of the reason explained above, the correction in the western end is much smaller than in the eastern part. Figure 12d shows the results after seven iterations. The area of “correct” values for $c^2$ has now spread to nearly half the basin, but there is still a discrepancy between the real and estimated $c^2$ in the eastern part of the basin. After seven iterations the gradient of the cost function has decreased about 2 orders of magnitude and the algorithm has converged.

6.1.2. Verification of the gradient of the cost function. As mentioned earlier, it is important to obtain the right gradient of the cost function. One way to check if the correct gradient has been found is described below. Let

$$J(h, c_0^2 + ah) = J(h, c_0^2) + ah \nabla_{c_0^2} J(h, c_0^2) + \text{h.o.t.}$$

be a Taylor expansion of the cost function in eqn (76). $a$ is a small scalar and $h$ is an arbitrary vector of unit length. Rewriting (86) one can define a function of $a$ by

$$f(a) = \frac{J(h, c_0^2 + ah) - J(h, c_0^2)}{ah \nabla_{c_0^2} J(h, c_0^2)} \approx 1 + \text{h.o.t.}$$

The result of the calculation in (87) is only valid for values of $a$ sufficiently large so that the numerator in (87) can be calculated accurately. If $a$ is chosen close to the machine zero one cannot expect to be able to verify that the correct gradient has been found. For values of $a$ which are not too close to the machine zero one should expect to obtain a value for $f(a)$ which is close to 1. For the CYBER 205 the machine zero is about $10^{-15}$ and if $a$ is chosen between say $10^{-3}$ to $10^{-10}$ one would expect to find $f(a) \sim 1$.

To check if the calculation of the cost function gives consistent answers one can calculate $J(h, c_0^2 - ah)$ and check if

$$J(h, c_0^2 + ah) - J(h, c_0^2) = \frac{1}{2} (J(h, c_0^2 + ah) - J(h, c_0^2 - ah)).$$

The experiment performed in Section 6.1.1 was used to check the gradient. The gradient of the cost function with respect to the phase speed was calculated using eqn (84). Different values of $a$ were chosen and eqn (88) was verified. In Fig. 13 a plot of $f(a)$ is shown. It is clearly seen that for $a$ between $10^{-3}$ and $10^{-10}$, eqn (87) is verified. The correct gradient is therefore found.
6.2. Assimilation of identical twin data in the Pacific Ocean model forced by real winds

As mentioned above two different scenarios will be investigated, but first a few test cases are studied. In these initial experiments the results from the year 1979 are used, with identical twin data as observations.

6.2.1. Experiment 2: Constant phase speed and observations everywhere. In this experiment the results from the original model integration \( (c^2 = 6.0 \text{ m}^2 \text{s}^{-2}) \) are used as observations. A new integration of the model for the year 1979 was performed, changing the constant \( c^2 \) to \( 4.0 \text{ m}^2 \text{s}^{-2} \). A data assimilation and parameter estimation procedure as described in experiment 1 was performed. Perfect observations were available at every gridpoint of the model. In the experiments with the model forced by real winds, the model is restarted using the end of the previous year as initial conditions. The phase speed determined by the algorithm is a time averaged phase speed over the period of the assimilation. After each iteration the model was integrated for two years, the year prior to the year of the assimilation plus the year of the actual assimilation. This was done in order for the model to adjust to the new phase speed even in the beginning of the assimilation period.

In Fig. 14a and b, the normalized cost function and its gradient is shown, respectively. As before there is a rapid decrease in both the cost function and the gradient during the first few iterations. After seven iterations the values have dropped about two orders of magnitude and the algorithm converged. Figure 15a–c shows the corresponding evolution of the spatial structure of \( c^2 \). In Fig. 15a the results after the first iteration are shown. There is a correction all over the basin, with a slightly larger correction in the central region. After the third iteration, Fig. 15b, the “correct” value of \( 6.0 \text{ m}^2 \text{s}^{-2} \) has been found in the western part of the basin, while there is still a discrepancy between the estimated and the correct value in the eastern part. Figure 15c which is a plot of \( c^2 \) after the seventh iteration, shows that the correct value of \( c^2 \) has been determined. \( c^2 \) is now very close to a constant over the whole domain, and there are only small departures from the value of \( c^2 = 6.0 \text{ m}^2 \text{s}^{-2} \) used to create the “observations”.

![Plot of the function $f(\alpha)$ in eqn (87)](image)

**FIG. 13.** Plot of the function $f(\alpha)$ in eqn (87)
6.2.2. **Experiment 3: Variable phase speed and observations everywhere.** The experiments described so far have estimated a constant $c^2$. Even if $c^2$ were allowed to vary with longitude during the iterative process, the end result should give a constant phase speed. In a real ocean this is not very realistic, and an important question arises if the method is able to determine a spatially varying phase speed. In this experiment the algorithm is therefore tested on a more realistic case, one in which the phase speed is a function of longitude. Figure 16 is a plot of the phase speed used in the model integration creating the “observations”. The variation of $c^2$ is supposed to model a realistic variation, since one would expect to observe higher values in the west compared to the east. This is at least correct under “normal” conditions, i.e. when the El Niño event is not dominating the Pacific Ocean. The upper layer thickness is shallow in the east and deep in the west, and consequently $c^2$ would have a longitudinal distribution as in Fig. 16. The model was restarted in 1977 and integrated through 1979 using the new phase speed in order to let the model adjust.

As an initial guess, $c^2$ was assumed to be constant with a value of 4.0 m$^2$ s$^{-2}$. Again perfect observations were available at every gridpoint of the model. The results of the assimilation are shown in Fig. 17a–b and in Fig. 18a–d. The normalized cost function is plotted as a function of the number of iterations in Fig. 17a, while the normalized gradient is plotted in Fig. 17b. The rapid decrease during the first few iterations can be observed. After ten iterations the decrease in the gradient is more than two orders of
Fig. 15. The longitudinal variation of the phase speed during the iterative process in experiment 2. (a) after one iteration, (b) after three iterations, (c) after seven iterations.

magnitude and the algorithm converged. Figure 18a–d shows the evolution of the spatial variation of $c^2$. After the first iteration there is a strong adjustment in the western part of the basin and a smaller correction further to the east. Comparing the results after the first iteration to the optimal $c^2$ in Fig. 16, it is difficult to see if the estimation will work. Figure 18b which is a plot of $c^2$ after three iterations, shows a promising development of the spatial structure. A slope has become evident with higher values in the western part compared to the values further east, but there is still a rather large discrepancy if the results from the third iteration are compared to the real $c^2$. After six iterations, Fig. 19c, the spatial variation of $c^2$ is beginning to look very much like the spatial variation of $c^2$ in Fig. 16. The maximum value in the west is slightly lower than the real values, while in the
eastern part \( c^2 \) has a slightly higher value than it is supposed to. Finally, in Fig. 18d the results after ten iterations is shown. The spatial variation is now very close to the variation of the \( c^2 \) used to create the “observations”, see Fig. 16. As a conclusion one can say that the algorithm has been successful in determining the spatially varying phase speed.

Figure 16. The longitudinal variation of the phase speed used to create the observations.

Figure 17. The results from experiment 3 are shown. (a) The normalized cost function \( J/J_0 \) is plotted as a function of the number of iterations. (b) The normalized gradient of the cost function \( \|g\|/\|g_0\| \) is plotted as a function of the number of iterations.
FIG. 18. The longitudinal variation of the phase speed during the iterative process in experiment 3. (a) after one iteration, (b) after three iterations, (c) after six iterations, (d) after ten iterations.
6.2.3. Experiment 4: Variable phase speed and observations at 3 stations. In all the experiments described above, observations have been assumed to be available everywhere. This will not be the case in a real data assimilation. Even with the new satellites, the coverage will not be good enough to give observation at every gridpoint of the model. The time resolution will also be quite different from what has been used in the above experiments. The TOPEX/POSEIDON project is still trying to decide what the repeat period of the satellite should be. Periods of 10 or 20 days have been suggested, which mean that for a certain position a new observation will be available every 10(20) days. These repeat periods will give a spatial resolution of about $320 \times 320$ km$^2$ (10 days) and $160 \times 160$ km$^2$ (20 days) in midlatitudes.

The goal of this research is to assimilate sea level observations from island stations. The number of stations is limited and the number of observations is therefore drastically reduced compared to the cases discussed above. In this experiment it is assumed that data are only given at 3 stations (the same stations will be used when real sea level observations are assimilated). The stations are, Santa Cruz, Jarvis, and Truk (see Fig. 2). These three stations are distributed over the Pacific Ocean, with Santa Cruz representing the eastern part of the ocean, Jarvis the central region and Truk the western region. The
Fig. 20. The longitudinal variation of the phase speed during the iterative process in experiment 4, (a) after one iteration, (b) after three iterations, (c) after six iterations, (d) after ten iterations.
Variational data assimilation

experiment is identical to experiment 3, with the phase speed in Fig. 16 creating the observations and the initial guess for $c^2$ is 4.0 m$^2$s$^{-2}$. The observations are again assumed to be perfect.

Figure 19a–b and Fig. 20a–d show the results of the iterative process. The normalized cost function is shown in Fig. 19a, while the normalized gradient is shown in Fig. 19b. The algorithm converged after ten iterations. Figure 20a–d show the evolution of the spatial variation of $c^2$. After the first iteration, Fig. 20a, there is a small adjustment over most of the basin, with slightly higher values in the west. Figure 20b which is the result after the third iteration starts to show the right spatial structure. A slope from west to east has developed. The variation of $c^2$ after six iterations is shown in Fig. 20c. Now a definite resemblance to the "real" $c^2$ has developed. The values in the western and eastern part are too high, but overall there is a quite good agreement. After ten iterations the algorithm converged and the results are shown in Fig. 20d. Using observations from only three stations, it is still possible to estimate the correct spatial structure of $c^2$.

6.2.4. **Uniqueness and stability of the solution.** An important question is whether the solution is unique or not. If the solution is unique, the cost function $J$ should be convex, i.e. only one global minimum of the function should exist. An experiment with observations in only one region of the Pacific Ocean, showed that the estimated structure of the phase speed could look completely different from the case when observations were available in more than one region. If observations are available at only one station, the algorithm will adjust the parameter so that the cost function has a minimum at this station. Adding observations in other regions will change the spatial structure of the estimated $c^2$, in order for the total cost function to decrease. For a given number of stations, the estimated structure of $c^2$ is unique. Given observations at only one station it is difficult and in most cases impossible to determine a spatially varying $c^2$. To be able to determine a spatially varying phase speed it turned out to be necessary to have observations in the different dynamical regions of the model. Observations should be available in the western region where the model is dominated by the westward propagating Rossby waves, in the central region where there is a mixture between Kelvin and Rossby waves influencing the model and in the eastern region where for the most part Kelvin waves dominate the solution.

The final solution is not sensitive to the initial guess of $c^2$. Several experiments starting the assimilation with different guesses for the initial phase speed, showed that the algorithm always converged to the correct solution. The number of iterations needed for the algorithm to converge, did not depend strongly on the initial guess either. Changing the initial guess from 4.0 m$^2$s$^{-2}$ to 3.0 m$^2$s$^{-2}$ in experiment 3, the algorithm converged in 11 iterations instead of 10.

Another question is whether the estimated parameter is stable, i.e. will small errors in the variables lead to large changes in the computed parameter. To investigate this, the same scenario as in experiment 4 was used, but the "observations" were contaminated by what would represent errors in the observations. The errors consisted of normally distributed random numbers added to the time series of the observations. The maximum error would represent an accuracy of measuring the sea surface elevation of ±0.05 m. Even with this large error in the observations, the algorithm was able to find the correct spatial structure of the phase speed. The algorithm converged in ten iterations and the
final result of the assimilation was nearly identical to the result obtained for the case with perfect observations shown in Fig. 20d.

6.3. Assimilation of real sea level observations in the Pacific Ocean model

6.3.1. Experiment 5: Real sea level observations from 3 stations in 1979. Real sea level observations from three stations are used in this experiment. The stations are the same as in the experiment described above, Santa Cruz, Jarvis and Truk. The initial guess for $c^2$ was a constant value of 6.0 m$^2$s$^{-2}$. The results of the assimilation procedure are shown in Fig. 21a–b and Fig. 22a–c. Figure 21a is a plot of the normalized cost function. The cost function for the three stations dropped to about 35% of its initial value after six iterations. Looking at the cost function at the three different stations separately, showed that for the station at Santa Cruz the decrease was much larger. After six iterations the cost function had dropped to about 10% of its initial value. For the station in the central region, Jarvis, there was basically no change at all, while for Truk in the western region,
there was a decrease of about 60%. The cost function in the central region had a low value during the whole iterative process. The initial value of the cost function at Jarvis was less than 20% of the value at the two other stations, and at the end of the assimilation the cost function at Santa Cruz and Jarvis had about the same value. Since

Fig. 22. The longitudinal variation of the phase speed during the iterative process in experiment 5, (a) after one iteration, (b) after three iterations, (c) after six iterations.
the data misfits (the cost function is a measure of the misfits) are the forcing for the
adjoint equations, a station where the misfits are small will not have as much influence on
the final solution. This could be changed by letting the validity coefficients be functions
of space and not as here constants. On the other hand, there is no reason to assume that
the observations at Jarvis should have a larger influence on the solution than the other
stations. Figure 21b shows the corresponding normalized gradient of the cost function.
The gradient dropped to about 5% of its initial value. Increasing the number of iterations
did not result in a further decrease in the gradient or the cost function. Even if the
decrease is not two orders of magnitude as in the previous identical twin experiments, the
norm of the gradient after six iterations has about the same value as in the identical twin
examples. The conclusion is therefore that the minimum of the cost function has been
found.

Figure 22a is a plot of the spatial structure of $c^2$ after the first iteration. There is an
adjustment in the central and western region, with higher values in the west and lower
values in the middle of the basin. After three iterations, Fig. 22b, there is a strong
adjustment. A slope has developed with high values in the west and low values in the
east. Finally, the result after six iterations is shown in Fig. 22c. There is a slope from west
to east with most of the slope confined to the western end of the Pacific Ocean. From
about 160°W, $c^2$ is more or less constant with a value of 5.2 m$^2$ s$^{-2}$. The maximum value
in the west is about 6.5 m$^2$ s$^{-2}$. The spatial structure of $c^2$ with high values in the west and
lower values in the east, is in good agreement with observations from the Pacific. During
normal conditions the basic stratification consists of a deep thermocline in the west and a
shallow thermocline in the east, which corresponds exactly to the results obtained ($c^2$ is
proportional to the depth of the thermocline (ULT)).

A question now arises; did the new $c^2$ improve the model results? In terms of the cost
function the answer is yes. To investigate this issue further, the correlation coefficient
between the model results and the observations is calculated for the different stations.
The coefficient is calculated using the formula

$$
\frac{\sum_{i=1}^{360} hh'}{\left(\sum_{i=1}^{360} h^2 \sum_{i=1}^{360} h'^2\right)^{1/2}}
$$

where $h$ and $h'$ represents the time series of the model results and the observations,
respectively. The mean value is removed from both $h$ and $h'$. Figure 23 is a plot of the
correlation coefficient as a function of iterations for the stations at Santa Cruz, Jarvis,
Truk, Nauru and Guam. The stations at Santa Cruz, Jarvis and Truk were used in the
data assimilation, while Nauru and Guam are stations which were not included in the
assimilation. The correlation coefficient increases for all the stations, with a dramatic
increase for the station at Santa Cruz. The model did not do a good job in predicting the
ULT at Guam prior to the assimilation, but at the end of the assimilation there is a
significant increase in the correlation coefficient. It is interesting to note that the new
estimate of $c^2$ was able to improve the model results at this location, even if Guam was
not one of the stations used in the assimilation. An increase in the correlation coefficient
6.3.2. Experiment 6: Real sea level observations from 3 stations in 1982/83. As in experiment 5, real sea level observations from the 3 stations at Santa Cruz, Jarvis and Truk were used. In this case the assimilation started in June 1982 and continued for one year. The El Niño which occurred in 82/83 is therefore a part of this assimilation. Again the initial guess of $c^2$ was 6.0 m$^2$s$^{-2}$. Figure 24a is a plot of the normalized cost function. As in the previous experiment the cost function decreased to almost 35% of its initial value after five iterations. The station at Santa Cruz experienced the largest decrease in the cost function. After five iterations the value had dropped to about 10% of the initial value. As in the previous experiment the cost function at Jarvis was more or less constant during the iterative process, with a value of about the same magnitude as the final result at Santa Cruz. The reduction of the cost function at Truk was about 50%. Figure 24b shows the normalized gradient of the cost function, and as in the previous experiment the gradient dropped to about 5% of its initial value. Further iterations did not improve the results in this case either. The norm of the gradient after five iterations had about the same value as when the identical twin experiments converged, and the minimum of the cost function has therefore been found.

The corresponding evolution of the spatial structure of $c^2$ is shown in Fig. 25a–c. After
the first iteration, Fig. 25a, there is an adjustment in the western and central region of the basin. Close to the western boundary the values of $c^2$ have dropped, while higher values can be found in the central area. Figure 25b shows the results after the third iteration. Now there is a large decrease in the western area, while higher values than the initial can be found east of about 160°E. A maximum occurs around 160°W. After five iterations, Fig. 25c, a steep slope has developed with the lowest values close to the western boundary where $c^2 = 3.3 \text{ m}^2\text{s}^{-2}$. The maximum value of $c^2 = 7.0 \text{ m}^2\text{s}^{-2}$ occurs around 160°W with slowly decreasing values as the eastern boundary is approached. As in the previous experiment, the spatial structure of the estimated phase speed is in good agreement with observations. An El Niño event is recognized by the fact that the eastern Pacific Ocean becomes warmer, which corresponds to an increase in the upper layer thickness.

The correlation coefficient for the stations are shown in Fig. 26. The general picture is that the model gives better correlation for all the stations during an El Niño year. During the iterative process the correlation coefficient increases for all the stations. The stations at Guam and Nauru are not shown, since continuous observations were not available for the assimilation period. The observations at Ponape were not used in the assimilation,
Fig. 25. The longitudinal variation of the phase speed during the iterative process in experiment 6, (a) after one iteration, (b) after three iterations, (c) after five iterations.
and as in the previous experiment, the new estimate of $c^2$ is able to improve the model results even at locations which are not part of the assimilation.

The reason why the correlation coefficients are higher in 82/83 is that this type of model is doing a good job of predicting the strong signal in the Pacific Ocean during an El Niño year (BUSALACCHI and O'BRIEN, 1980). In 1979 there is not a strong signal and one cannot expect the model to give as good results as in 82/83, and the lower correlation coefficients in the first experiment are therefore to be expected.

7. SUMMARY AND CONCLUSIONS

A variational data assimilation method for a reduced gravity model has been developed. The method was applied to the equatorial Pacific Ocean. The phase speed of the model was chosen to be the control parameter of the problem, and the optimal spatial structure of the parameter was determined. Knowledge of the spatial structure of the phase speed gives information about the basic stratification of the Pacific Ocean. In the variational formalism a cost function measuring the “distance” between the observations and the model results is minimized. The method consists of integrating the model equations forward in time over the period which data are assimilated. Data misfits between the model and the observations are then calculated and the adjoint equations of the model are integrated backward using the data misfits as forcing. The gradient of the cost function with respect to the control parameter is found using the model and the adjoint variables and it is used in a minimization algorithm to determine a new value of the phase speed. A conjugate gradient method was used to determine the search direction. The quasi-Newton limited-memory conjugate gradient method of SHANNO and PHAUP (1980) was implemented with a few modifications. A step length was calculated using the method of DERBER (1985). The convergence criterion for the algorithm required the value of the cost function and/or its gradient to have decreased by a
specified tolerance. If this was not the case the iterative procedure was continued. The new value of the parameter was then used in the procedure described above, and it was repeated until a satisfactory decrease in the cost function and its gradient was achieved.

Several different experiments have been performed to test the assimilation algorithm. The first example used a simple set up of the Pacific Ocean model. The model was initialized by the height and velocity field of a Kelvin wave and the model was integrated forward in time using a value of $c^2 = 6.0 \text{ m}^2 \text{s}^{-2}$ to create what were called the observations. The assimilation algorithm was started using an initial guess for $c^2 = 4.0 \text{ m}^2 \text{s}^{-2}$ and the optimal value of $c^2$ was estimated. This simple example clearly demonstrated how information propagated backward in time through the adjoint variables. Perfect observations were assumed to be available everywhere, and it was shown that it was possible to determine the original phase speed. The algorithm converged in seven iterations. Due to the initial conditions of this experiment, the gradient of the cost function was small in certain areas of the domain and the final result was not as close to the “real” solution as was the case in the next experiments.

The Pacific Ocean model was forced with real winds in the experiments which followed. The model was integrated from 1972 through 1983, and in the initial experiments, identical twin observations from the year 1979 were used. A similar experiment as the one described for the Kelvin wave case, was performed. The algorithm converged in seven iterations and the final result of the assimilation gave a $c^2$ which was constant with longitude. Since a constant $c^2$ is not very realistic as the optimal solution when real data are assimilated, another experiment was performed. In this case the $c^2$ used to create the “observations” was a function of longitude, with high values in the western part of the basin and then decreasing values as the eastern part of the Pacific Ocean was approached. Again perfect observations were used, and the assimilation was started with an initial guess of a constant $c^2 = 4.0 \text{ m}^2 \text{s}^{-2}$. The algorithm converged after ten iterations and the final result of the estimation was a spatially varying $c^2$ which had the correct structure with high values in the west, and low values in the east.

In a real data assimilation, observations will not be available at every gridpoint of the numerical model. Since the goal of this research was to assimilate sea level observations from island stations, the next experiment used observations at only three stations as forcing for the adjoint equations. The stations which was chosen were Santa Cruz, Jarvis and Truk. These three stations represent the three different regions of the Pacific Ocean. Santa Cruz representing the eastern part of the ocean, Jarvis the central region and Truk the western region. The experiment was identical to the previous experiment except for the number of available observations. Even in this case the algorithm converged after ten iterations and the spatial variation of the estimated $c^2$ was again very close to the correct structure.

Real sea level observations were assimilated during two different periods. In the first experiment, observations from 1979 were used. This is a normal year, i.e. there was no El Niño event during 1979. In the second experiment observations from 1982/83 were assimilated. A strong El Niño occurred during this period. In both experiments sea level observations from the three stations used in the previous experiment were assimilated. As an initial guess for the phase speed a constant value of $6.0 \text{ m}^2 \text{s}^{-2}$ was chosen. The results from the assimilation of observations from 1979 gave a spatial structure of the phase speed with higher values in the west and then a gradual slope towards the east where the values had decreased compared to the initial guess. This is in good agreement
with observations. The general picture of the upper layer thickness in the Pacific Ocean is a thick ULT in the west and then a slope towards a thinner ULT in the east. Assimilating the observations during 1982/83 gave a completely reversed picture. Now the phase speed had lower values in the west and higher values in the east, which corresponds to a decrease in the ULT in the west and an increase in the east. This confirms our knowledge of an El Niño. A warming of the eastern Pacific Ocean occurs, which corresponds to an increase in the ULT and in the west there is a cooling corresponding to a decrease in the upper layer thickness.

A correlation coefficient between the model results and the observations was also calculated. The correlation increased for all the stations considered during the assimilation period. An important result is that the new estimate of the phase speed was able to improve the correlation not only at the stations used in the assimilation, but also at stations which were not a part of the assimilation.

The results have shown that the estimated phase speed was a unique solution to the problem. It has also been shown that to be able to determine the spatial structure of the phase speed, it was necessary to have observations in the three different dynamical regions of the model. Observations should be available in the eastern region where Kelvin waves dominate the dynamics, in the central region where there is a mixture between Kelvin and Rossby waves influencing the solution and in the western region where Rossby waves play the most important role.

The estimated parameter is stable, i.e. errors in the observations do not result in large variations in the parameter. Normally distributed random numbers were added to the observation in an identical twin experiment. The errors corresponded to an uncertainty in the sea level measurement of ±0.05 m. The final result of the assimilation gave a solution which was very close to the “real” spatial variation.

The variational data assimilation algorithm can be computationally expensive to use. It is necessary to be able to find the optimal solution in as few iterations as possible. The results from the different experiments have shown that the algorithm is computationally efficient. Convergence was achieved in ten iterations or less for all the cases considered. One of the crucial parts of the algorithm is the choice of the minimization method. Several different conjugate gradient algorithms exist. NAVON and LEGLER (1987) tested some of the available subroutines and their conclusion was that the limited-memory quasi-Newton conjugate gradient method of SHANNO and PHUA (1980) (which was used here) was the one that gave the best convergence rates. There is a constant development in the area of optimization, and one way of improving the convergence rate could be to implement another subroutine. NAVON (personal communication) has tested a new subroutine based on an algorithm of LIU and NOCEDAL (1988) which gave even better convergence rates than SHANNO and PHUA (1980). A saving of up to 25% in the number of iterations was found. For problems which require a large number of iterations this is an important improvement. The two assimilations of real sea level observations studied here, required only five or six iterations to converge. A saving of one or maybe two iterations would not mean much in terms of saved computer time.

Another crucial part of the method is the way the step length is calculated. It is important to have a routine which gives the optimal step length. The method used here (DERBER, 1985), worked satisfactorily for the experiments considered but may not be adequate for nonlinear models. The initial guess had some influence on the convergence rate. It turned out that the initial guess should not be too small.
The estimated parameter has been assumed to be a function of longitude only. Letting the phase speed be an arbitrary function of both latitude and longitude would not be a feasible problem to solve given the limited number of observations available from sea level stations. Satellite observations may give enough information to make it possible to determine the spatial structure of a phase speed which is a function of both latitude and longitude. At least it should be possible to determine the large scale structure of the phase speed. Short scale variations will probably have to be removed from the solution. As mentioned earlier experience has shown that it is sometimes necessary to reduce the dimension of the parameter in order to obtain stable results. Restricting the attention to the large scale variation is one way of reducing the dimension of the parameter.

8. ACKNOWLEDGEMENTS

This research was supported by NASA Oceanic Processes Branch and in part by the Florida State University through time granted on its Cyber 205 supercomputer. Special thanks goes to Klaus Wyrtki and Gary Mitchum, University of Hawaii for providing the sea level data, and to I. M. Navon for many helpful discussions concerning conjugate gradient methods. Assistance in the form of facilities and technical support were provided by the Supercomputer Computations Research Institute (SCRI) at Florida State University. Assistance from Mimi Burbank using TEx is appreciated. This is contribution number 295 to the Geophysical Fluid Dynamics Institute at the Florida State University.

9. REFERENCES


