Variational Assimilation of Altimeter Data in a Multilayer Model of the Tropical Pacific Ocean

ANTHONY T. WEAVER* AND DAVID L. T. ANDERSON+

Atmospheric, Oceanic and Planetary Physics, Department of Physics, Clarendon Laboratory, Oxford, United Kingdom

(Manuscript received 19 December 1995, in final form 26 August 1996)

ABSTRACT

A four-dimensional variational method is used to examine the extent to which a time sequence of altimeter measurements can determine the subsurface flow in a linear multilayer model of the tropical Pacific Ocean. The experiments are all of the identical-twin type. Complete maps of sea level extracted from the model in a control integration play the role of the altimeter observations in the assimilation experiments. The results of the experiments indicate that, over timescales of months, the sea level information can be effectively propagated into the subsurface, particularly in the dynamically active equatorial region. Several degrees off the equator, however, where waves propagate more slowly, the recovery of the subsurface flow in models containing more than two vertical modes is significantly more difficult. The sensitivity of these results to the lengths of the data sampling and assimilation periods is discussed.

1. Introduction

Sea surface height, as measured by satellite altimeters, is one of the few ocean dynamic variables for which global, time-continuous measurements are available. Since large-scale variations in the sea surface are in effect caused by large-scale subsurface dynamical processes, there has been considerable interest in trying to use a time sequence of altimeter measurements to determine the subsurface circulation in three-dimensional (3D) ocean models (e.g., see review papers Ghil and Malanotte-Rizzoli 1991 or Anderson et al. 1996). The problem is similar in many ways to the early attempts made by meteorologists to determine the 3D circulation of the atmosphere from a time sequence of surface pressure measurements (Bengtsson 1979). Extracting 3D information from surface measurements, however, is a much more important issue in oceanography since direct observations of the fluid's vertical structure are much more sparse for the ocean than for the atmosphere.

The fundamental problem for altimeter assimilation is how to assimilate the information in the vertical so as to correct the subsurface fields (density and velocity) simultaneously with the directly observed surface fields. Viewed as a static assimilation (i.e., an adjustment of the model fields at a single time), there is clearly an infinite number of ways to make a correction that is consistent with the given altimeter data. The problem is thus mathematically underdetermined and can only have a unique solution if additional (prior) constraints are imposed. The type of constraints and manner in which they are imposed are what distinguish the various assimilation methods.

Haines (1994) classifies three of the most common types of sequential methods used for assimilating altimeter data. First, there are methods which make adjustments to the deeper layers on the basis of a priori statistics, derived in most cases directly from the model output (Hurlburt et al. 1990; Mellor and Ezer 1993; Ezer and Mellor 1994) but in some exceptional cases from independent hydrographic data (De Mey and Robinson 1987). For the altimeter problem, the relevant statistics are the vertical correlations between anomalies of sea surface height (or sea surface pressure) and the subsurface density field. Once computed, these correlations are then held fixed in the analysis equation that is used at each assimilation step to update the model's density field. The velocity field is usually updated simultaneously with the density field using correction terms that geostrophically balance the statistically derived density field increments.

The second class of methods comprises simple dynamical assimilation methods, such as direct insertion and nudging, which rely solely on the model's internal dynamics to transfer the surface information into the

* Current affiliation: Laboratoire d'Océanographie Dynamique et de Climatologie, (CNRS/ORSTOM/UPMC), Université Paris VI, Paris, France.

+ Current affiliation: European Centre for Medium-Range Weather Forecasts, Reading, United Kingdom.

Corresponding author address: Dr. Anthony T. Weaver, Laboratoire d'Océanographie Dynamique et de Climatologie, (CNRS/ORSTOM/UPMC), Université Paris VI, boîte 100, 4 place Jussieu, 75252 Paris Cedex 05, France.

deeper ocean (Hurlburt 1986; Berry and Marshall 1989; Holland 1989; Holland and Malanotte-Rizzoli 1989; Verron 1990; Haines et al. 1993). No explicit vertical projection scheme is imposed in the analysis equation as only the directly observed (surface) field is updated; only by integrating the model forward to the next analysis time can an adjustment in the deeper layers take place. For example, the simple technique of nudging either the upper-layer vorticity field in quasigeostrophic models or the upper-layer geostrophic velocity field in primitive equation models to the corresponding quantities derived directly from altimetry, has been shown to be quite effective in eddy-intense regions such as the Gulf Stream and Agulhas Current.

Guided by physical insight, Haines (1994) has suggested a third class of methods for assimilating altimeter data that emphasizes the preservation of subsurface water mass information during each assimilation step. For quasigeostrophic and shallow-water models, this leads to a constraint that the potential vorticity in the deeper layers be conserved during the assimilation step (Haines 1991; Haines et al. 1993). For primitive equation models with active thermodynamics, the constraint is extended to conserve other water mass properties such as temperature and salinity on isopycnal surfaces (Haines 1994; Cooper and Haines 1996; Oschlies and Willebrand 1996). In effect, this procedure leads to a lifting or lowering of the isopycnal surfaces by an amount that can be determined from the altimeter data (i.e., the observed surface pressure) and an additional constraint of a zero change in pressure gradient at some specified level.

From a practical point of view, these three methods are attractive because of their conceptual simplicity and relatively low cost of implementation. These advantages, however, come at the expense of making poor use of the temporal information contained in the evolving sea surface height fields since each assimilation step is essentially independent of the previous ones. The temporal link between the observations is established through the dynamical constraint. However, with the simpler 4D assimilation methods, the dynamics do not enter directly into the analysis equations; they are used to propagate the updated model state vector from one analysis time to the next, but have little or no influence on the weights in the assimilation scheme. Furthermore, they are applied strictly in a forward sense so that information can only be propagated into the future, which makes these methods less appealing for hindcast studies.

Haines's classification of altimeter data assimilation methods deliberately excludes the class of so-called ''optimal'' methods, most notably Kalman filtering/smoothing (KFS) and four-dimensional variational assimilation (4D-Var), which are founded on the principles of statistical linear estimation. While these methods are generally much more expensive to implement, which indeed often precludes their application to models other than those of very limited complexity, they do in theory make better use of time-dependent dynamical constraints. In the sequential process of KFS the dynamics are used to propagate (exactly for linear models but only approximately for nonlinear models) the model state error covariance matrix from one observation time to the next. The predicted state error covariance matrix is then combined with the observation error covariance matrix to define optimal weights for the analysis equation (i.e., optimal in the sense that the norm of the expected analysis error variance is minimized). Thus, unlike the stationary correlation functions used in the simpler statistical assimilation methods, those used in the Kalman procedure will evolve with the flow. In 4D-Var, on the other hand, the dynamics are imposed as constraints (either weak or strong) in a weighted-least-squares minimization problem. The observations at different times are thus assimilated together to produce a dynamically consistent 4D analysis. Indeed, it is now a well-established result that 4D-Var and fixed-interval KFS are theoretically equivalent in the linear regime assuming a perfect model and normally distributed observation errors (Ghil and Malanotte-Rizzoli 1991; Talagrand 1993).

Another serious drawback with the simpler methods when applied to equatorial assimilation studies is that they usually rely on a geostrophic balance condition to adjust the velocity field simultaneously each time the altimeter data are assimilated. A straightforward geostrophic constraint is clearly not possible near the equator, and what alternative balance constraint(s) could be applied effectively there is by no means obvious. Without a balance constraint, however, the adjustment of the velocity field will be left entirely up to the model. This may lead to a large initialization shock, the implications of which may be particularly severe near the equator where spurious equatorial Kelvin and mixed Rossby–gravity waves can be excited (Moore 1990). With KFS or 4D-Var, however, the reliance on balance constraints such as geostrophy is a priori unnecessary. In KFS the velocity field is adjusted using dynamically computed correlations between the model variables; in 4D-Var, these correlations are not actually computed but are implicit in the variational adjustment of the model fields to satisfy the model dynamics.

In this paper, we explore the possibility of using 4D-Var to project altimeter data into the deeper layers of a model of the tropical Pacific Ocean. It is in this region that ocean data assimilation is likely to contribute significantly to improving short-term climate forecasts using coupled atmosphere–ocean models, and indeed altimeter data are viewed as a potentially important ingredient in the assimilation system. Through a series of identical twin experiments using a multilayer reduced-gravity model, we address the following questions. How much information about the ocean state can we extract given a time sequence of perfect altimeter data, an accurate numerical model and forcing field, and an assimilation scheme that makes optimal use of the dynamics?

In particular, how well are the subsurface fields constrained in models with different vertical resolution, and how sensitive are the results to the lengths of the assimilation and data sampling periods?

The paper is organized as follows. In section 2, a general formulation of the problem is given, and the solution technique using the variational-adjoint approach is outlined. The ocean model and assimilation procedure are described in section 3, and the numerical experiments are presented in section 4. A general discussion and conclusions are given in section 5.

2. General outline of the variational problem

Consider an ocean described by a linear, discrete model of the form

$$\mathbf{x}_{n+1} = \mathbf{M}_n\mathbf{x}_n + \mathbf{C}\mathbf{f}_n, \tag{1}$$

where $\mathbf{x}_n$ is the N-dimensional vector of the ocean state at time $t_n = n\Delta t$ and $\mathbf{f}_n$ an S-dimensional forcing vector on the interval $[t_n, t_{n+1}]$. The $N \times N$ model operator $\mathbf{M}_n$ evolves the state vector on $[t_n, t_{n+1}]$ and the $N \times S$ matrix $\mathbf{C}$ maps the forcing onto this state. Assuming that this system is free of model and forcing error, the state at any time t_n is uniquely determined by its initial and boundary conditions. We further assume that the boundary conditions are known so that the only control parameter in (1) is the initial state $\mathbf{x}_0$.

Now let $\mathbf{y}_n^o$ represent a P-dimensional vector ($P < N$) of perfect observations at t_n. The link between the model state vector and these observations is assumed to be linear and defined by a $P \times N$ observation operator $\mathbf{H}$;

$$\mathbf{y}_n^o = \mathbf{H}\mathbf{x}_n \tag{2}$$

Assuming a time sequence of observations of the form (2) is available on $[t_0, t_L]$, can this information be effectively combined with the dynamical constraint (1) to reconstruct $\mathbf{x}_n$ on $[t_0, t_L]$? This is the fundamental problem we address here. As the evolution of the state is controlled by the initial conditions, it is equivalent to asking whether $\mathbf{x}_0$ can be recovered from the observations, or in the language of estimation theory, whether the state is observable on $[t_0, t_L]$ (Gelb 1974).

One way of tackling the problem is to reformulate it as a constrained 4D-Var problem in which the model and observations are imposed as strong and weak constraints respectively (Sasaki 1970). The (weak) observation constraint is embedded in a cost function, which is minimized with respect to the control vector (in this case the initial state) subject to the (strong) constraint that the solution trajectory exactly satisfy the model dynamics. The cost function takes the general form

$$J = \frac{1}{2}\sum_{0\le n\le L}(\mathbf{H}\boldsymbol{\xi}_n - \mathbf{y}_n^o)^{\mathrm{T}}\mathbf{W}^{-1}(\mathbf{H}\boldsymbol{\xi}_n - \mathbf{y}_n^o), \tag{3}$$

where $\mathbf{W}^{-1}$ is a positive definite, symmetric matrix that defines the metric on the P-dimensional space of observation errors at t_n and $\boldsymbol{\xi}_n$, $n = 0, \cdots, L$, is a sequence of state vectors satisfying (1). In a real data application, $\mathbf{W}^{-1}$ should reflect the accuracy and representativeness of the data and thus be defined by the inverse of the sum of the observation and representativeness error covariance matrices (Lorenc 1988). However, when the data are perfect and when an iterative method is used to minimize J, the role of this matrix is simply to provide a reasonable scaling of the model/observation differences taking into account their physical units. Providing these differences are small on the final iteration (defined here by the fractional reduction of J from its initial value), the weak observation constraint will have approached, for all practical purposes, the strong constraint limit.

The minimization of J may be achieved efficiently using gradient descent methods, for which a specific algorithm is needed for computing the gradient of the cost function with respect to the control vector ($\boldsymbol{\nabla}_{\xi_0}J$). The adjoint equations provide the most economical tool when J is a highly implicit function of the control variable as is the case with most numerical models of the ocean and atmosphere (Le Dimet and Talagrand 1986; Lewis and Derber 1985; Talagrand and Courtier 1987; Courtier and Talagrand 1987; Thacker and Long 1988). To derive the adjoint equations, an inner product, $\langle\ ,\ \rangle_{E^{-1}}$, must first be defined on N-dimensional state space. Let the positive definite, symmetric matrix $\mathbf{E}^{-1}$ define the metric for this inner product; that is, for two state vectors $\mathbf{a}$ and $\mathbf{b}$, we define $\langle \mathbf{a}, \mathbf{b}\rangle_{E^{-1}} = \mathbf{a}^{\mathrm{T}}\mathbf{E}^{-1}\mathbf{b}$. In the context of gradient descent minimization, the inverse matrix $\mathbf{E}$ is usually referred to as the preconditioning matrix, and the quantity $\boldsymbol{\nabla}^*_{\xi_0}J = \mathbf{E}\boldsymbol{\nabla}_{\xi_0}J$ as the preconditioned gradient or direction of steepest ascent (Gill et al. 1980; Tarantola 1987). Mathematically, $\mathbf{E}^{-1}$ defines a mapping between state space and its associated dual; that is, it transforms a state vector, whose components may consist of a variety of physical variables, into a dual vector, whose components have physical units that are inverse to those of the corresponding components of the state vector. Since the gradient vector is by definition a dual vector, the action of $\mathbf{E}$ can be viewed as a transformation of $\boldsymbol{\nabla}_{\xi_0}J$ into state space. This transformation is fundamental when the state variables contain a mixture of physical units.

With respect to $\langle\ ,\ \rangle_{E^{-1}}$, it is straightforward to show that the discrete adjoint equations are

$$\boldsymbol{\lambda}_n^* = \mathbf{M}_n^*\boldsymbol{\lambda}_{n+1}^* - \mathbf{H}^*(\mathbf{H}\boldsymbol{\xi}_n - \mathbf{y}_n^o), \tag{4}$$

where $\boldsymbol{\lambda}_n^*$ is the adjoint state vector, and

$$\mathbf{M}_n^* = \mathbf{E}\mathbf{M}_n^{\mathrm{T}}\mathbf{E}^{-1}, \tag{5}$$

$$\mathbf{H}^* = \mathbf{E}\mathbf{H}^{\mathrm{T}}\mathbf{W}^{-1} \tag{6}$$

are the adjoint operators of $\mathbf{M}_n$ and $\mathbf{H}$, respectively. Here $\mathbf{M}_n^*$ maps state errors at t_{n+1} into state errors at t_n, while $\mathbf{H}^*$ maps observation errors at t_n into state errors at t_n. The ''initial'' condition of (4) is $\boldsymbol{\lambda}_{L+1}^* = 0$ and the value

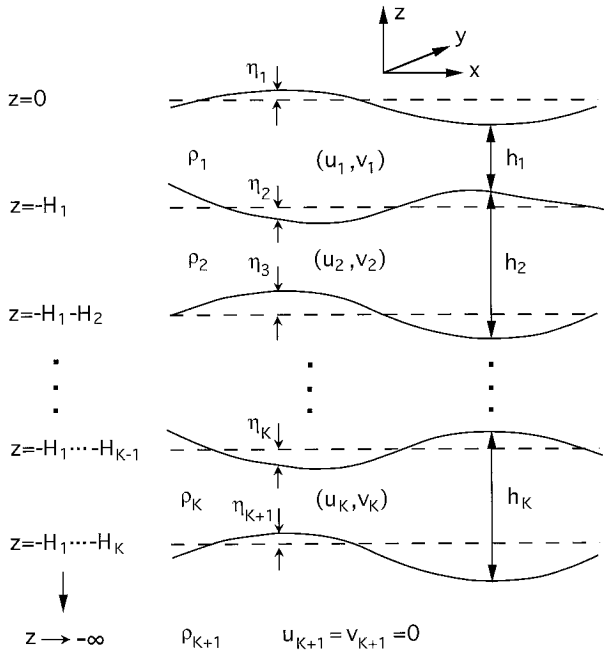

FIG. 1. Schematic illustration of the multilayer reduced-gravity model. The model consists of $K + 1$ layers, each layer k having density ρ_k. The $(K + 1)$th layer is infinitely deep and passive. The prognostic variables in layers $k = 1, \cdots, K$ are the horizontal velocities (u_k, v_k) and layer depth perturbations $h'_k = h_k - H_k = \eta_k - \eta_{k+1}$, where H_k is the mean (i.e., undisturbed) depth of layer k, and η_k is the interface displacement between layers $k - 1$ and k.

of the adjoint variables at $t = 0$ gives the preconditioned gradient $\boldsymbol{\lambda}_0^* = \nabla_{\xi_0}^* J$. Thus, for any initial estimate of $\boldsymbol{\xi}_0$, $\nabla_{\xi_0}^* J$ can be obtained by first integrating the model equations forward from t_0 to t_L, and then by integrating the adjoint equations backward from t_{L+1} to t_0 using the weighted model data misfits at t_n $(0 \le n \le L)$ as discrete forcing terms. Note that by multiplying (4) by $\mathbf{E}^{-1}$ and introducing new (i.e., dual) variables $\boldsymbol{\lambda}_n = \mathbf{E}^{-1}\boldsymbol{\lambda}_n^*$, the adjoint equations can be written independently of the choice of metric on state space. In other words, it is sufficient to integrate the dual equation and to introduce the metric only at the end of the integration by applying the inverse transformation $\boldsymbol{\lambda}_0^* = \mathbf{E}\boldsymbol{\lambda}_0$.

Within the minimization routine, the control variables are updated each iteration using an equation of the form

$$\boldsymbol{\xi}_{0,r+1} = \boldsymbol{\xi}_{0,r} - \alpha_r \mathbf{K}_r^{*-1} \nabla_{\xi_{0,r}}^* J, \tag{7}$$

where r is the iteration number, α_r is a nondimensional scalar (the step size), and $\mathbf{K}_r^{-1} = \mathbf{K}_r^{*-1}\mathbf{E}$ is an approximation to the inverse of the Hessian matrix (the matrix of second-order derivatives of J). On the first iteration, $r = 0$, $\mathbf{K}_0^{-1}$ is approximated by the preconditioning matrix $\mathbf{E}$ (i.e., $\mathbf{K}_0^{*-1}$ is taken to be the identity matrix). Minimization routines of the quasi-Newton (QN) type then update $\mathbf{K}_r^{-1}$ each iteration using some variant of the Broyden–Fletcher–Goldfarb–Shanno formula (Gill et al. 1980). The application of $\mathbf{K}_r^{-1}$ to the negative of the preconditioned gradient provides the direction of search for the current iteration. The control variables are subsequently modified by ''stepping'' in this direction. For a quadratic cost function, the optimal step size is straightforward to compute using an optimal line search (Gill et al. 1980).

3. The ocean model and assimilation procedure

a. The model equations

The model is similar to the 1½-layer reduced-gravity model of Sheinbaum and Anderson (1990a,b) but has been extended in this study to include an arbitrary number of active layers K. As in the 1½-layer model, the bottom layer in the K½-layer model [i.e., the $(K + 1)$th layer] is taken to be infinitely deep and passive as a way of isolating the dynamically active upper ocean from the much more slowly evolving deep ocean. This has the effect of eliminating the fast barotropic mode while retaining K baroclinic modes. The equatorial waves supported by the model equations, most importantly baroclinic Rossby and Kelvin waves, play a major role in the seasonal and interannual variability of the tropical ocean circulation. As a result, when forced by realistic winds, simple models of this type are capable of reproducing much of the observed tropical ocean variability on these timescales (Cane 1979; Busalacchi and O'Brien 1981). Furthermore, close variants of this model have been used successfully as the ocean component in simple coupled atmosphere–ocean models aimed at studying El Niño–Southern Oscillation (McCreary 1983; McCreary and Anderson 1984).

The governing equations are based on a linearized version of the shallow-water equations. They are formulated on an equatorial β plane in Cartesian coordinates (x, y) where x and y are the zonal and meridional directions, respectively. Corresponding to these directions are velocities (u_k, v_k) where the subscript k corresponds to the kth active layer $(k = 1, \cdots, K)$. Each layer has a mean (i.e., undisturbed) thickness H_k and density ρ_k. The variation of the layer depth is denoted by h_k and the perturbation displacement from the mean layer depth by $h'_k = h_k - H_k = \eta_{k-1} - \eta_k$, where η_k is the interface displacement between layers $k - 1$ and k. A schematic representation of the layer configuration is given in Fig. 1.

The equations for the prognostic variables (u_k, v_k, h'_k) are

$$\frac{\partial u_k}{\partial t} - \beta y v_k + \frac{1}{\rho_0}\frac{\partial p'_k}{\partial x} = \nu\nabla^2 u_k - r u_k + \frac{\delta_{k1}\tau^x}{\rho_0 H_k}, \tag{8}$$

$$\frac{\partial v_k}{\partial t} + \beta y u_k + \frac{1}{\rho_0}\frac{\partial p'_k}{\partial y} = \nu\nabla^2 v_k - r v_k + \frac{\delta_{k1}\tau^y}{\rho_0 H_k}, \tag{9}$$

$$\frac{\partial h'_k}{\partial t} + H_k\left(\frac{\partial u_k}{\partial x} + \frac{\partial v_k}{\partial y}\right) = 0, \tag{10}$$

where β is the variation of the Coriolis parameter with

respect to meridional distance y, ρ_0 is the mean density of the layers, and

$$p_k' = \rho_0 \sum_{l=1}^{K} G_{kl} h_l' \tag{11}$$

is the perturbation pressure. The parameters G_{kl} are the elements of a $K \times K$ symmetric matrix of reduced gravities:

$$G_{kl} = \begin{cases} \sum_{j=k}^{K} g'_{jj+1} & \text{if } 1 \leqslant l \leqslant k \\ \sum_{j=l}^{K} g'_{jj+1} & \text{if } k < l \leqslant K, \end{cases} \tag{12}$$

where $g'_{kk+1} = (\rho_{k+1} - \rho_k)g/\rho_0$. Equation (11) follows from the hydrostatic equation and the additional requirement that the horizontal pressure gradients vanish in the infinitely deep bottom layer. The pressure gradient in the top layer is provided by the displacement of the sea surface $\eta_1 = p_1'/\rho_0 g$. Equation (11) will thus be used to diagnose the model equivalent of the altimeter measurement.

Diffusion of momentum is introduced through the Laplacian term where the coefficient of horizontal eddy viscosity is given by ν. The viscous boundary condition applied at the solid walls is no-slip. The second term on the right-hand side of (8) and (9) is an additional (Rayleigh) damping term with a meridionally dependent coefficient $r = r(y)$, which is introduced to suppress spurious coastal Kelvin waves that propagate along the artificial northern and southern boundaries of the model. The numerical value of $1/r$ increases linearly from 0 to 5 days within a 5° ''sponge layer'' along the boundary. The Kronecker delta δ_{kl} ($\delta_{kl} = 1$ if $k = l$; $\delta_{kl} = 0$ if $k \neq l$) is used to indicate that the zonal and meridional wind stresses denoted by τ^x and τ^y, respectively, are acting as depth-independent body forces in the upper layer only.

Equations (8)–(10) are solved numerically on an Arakawa C-grid using standard finite difference methods (for details see Weaver 1994). The model domain extends from 30°S to 30°N, 122°E to 68°W. Realistic geometry is included on the western and eastern boundaries and to close the box model solid zonal walls are imposed at the northern and southern boundaries. A horizontal resolution of 1° is used in both the zonal and meridional directions.

The standard values for the reduced gravities and mean layer depths for the different layer versions of the model are listed in Table 1. These particular values are chosen so that the estimated speed of the first baroclinic mode Kelvin wave c_1 remains roughly the same when the number of layers in the model is varied. A value of $c_1 = 2.8$ m s^{-1} is chosen to be consistent with typical estimates of this wave speed from observations (Wunsch and Gill 1976). The values of the other model parameters are $\beta = 2.28 \times 10^{-11}$ m^{-1} s^{-1}, $\nu = 2.0 \times 10^3$ m^2 s^{-1}, and $\rho_0 = 1.0 \times 10^3$ kg m^{-3}.

b. The forcing field

Realistic forcing is included in the form of monthly mean climatological pseudostresses (i.e., vectors of the form $|\mathbf{U}|\mathbf{U}$ where $\mathbf{U}$ is the wind vector) compiled at The Florida State University (FSU) for the period 1965–92 (Stricherz et al. 1992). The pseudostresses are linearly interpolated to the model space–time grid and then converted to wind stress $\boldsymbol{\tau} = (\tau^x, \tau^y)^{\mathrm{T}}$ using the drag equation $\boldsymbol{\tau} = \rho_a C_D |\mathbf{U}|\mathbf{U}$, where $C_D = 1.15 \times 10^{-3}$ is the drag coefficient, and $\rho_a = 1.2 \times 10^{-3}$ kg m^{-3} is the density of air.

c. Preconditioning

As pointed out in section 2, a metric for state space must be defined a priori for preconditioning the gradient. When a background constraint is included in the cost function, the weighting matrix for the background error, which is usually some estimate of the inverse of the background error covariance matrix, provides a natural metric (Tarantola 1987; Lorenc 1988). However, in the absence of a background constraint, an alternative metric must be defined. Following Courtier and Talagrand (1990) and Thépaut and Courtier (1991), we choose a physical metric based on energy. This metric is a natural by-product of the equations of motion and thus can be expected to provide an adequate scaling of the components of the gradient.

Let $\xi = (u_1, \cdots, u_K, v_1, \cdots, v_K, h_1', \cdots, h_K')^{\mathrm{T}}$ and $\Lambda = (\lambda_1, \cdots, \lambda_K, \mu_1, \cdots \mu_K, \kappa_1', \cdots, \kappa_K')^{\mathrm{T}}$ be two state vectors satisfying the continuous equations (8)–(10). It is straightforward to show that the quadratic quantity

$$\langle \xi, \Lambda \rangle_{E^{-1}} = \int \int \sum_{k=1}^{K} \left(H_k u_k \lambda_k + H_k v_k \mu_k + \sum_{l=1}^{K} G_{kl} h_k' \kappa_l' \right) dx\, dy \tag{13}$$

TABLE 1. The standard values of the reduced gravities g'_{kk+1} ($\times 10^{-2}$ m s^{-2}) and mean layer depths H_k (m) used in the K½-layer model ($K = 1, \cdots, 4$), and the estimated mode speeds c_n (m s^{-1}).

Model	g'_{12}	g'_{23}	g'_{34}	g'_{45}	H_1	H_2	H_3	H_4	c_1	c_2	c_3	c_4
1½ layer	3.92	—	—	—	200	—	—	—	2.80	—	—	—
2½ layer	3.00	3.00	—	—	100	100	—	—	2.80	1.07	—	—
3½ layer	1.56	1.56	1.56	—	100	100	100	—	2.80	1.00	0.69	—
4½ layer	0.95	0.95	0.95	0.95	100	100	100	100	2.81	0.97	0.64	0.52

is an invariant of the system (8)–(10) in the absence of forcing and dissipation. Equation (13) defines an inner product between ξ and Λ, the square of its associated norm being related to the total perturbation energy, E, in the nondissipative unforced model; that is, $\|\xi\|^2_{E^{-1}} = \langle \xi, \xi \rangle_{E^{-1}} = 2E/\rho_0$.

The discrete analog of (13) is used as the inner product for the numerical experiments. Let $\boldsymbol{\xi} = (\mathbf{u}_1^{\mathrm{T}}, \cdots, \mathbf{u}_K^{\mathrm{T}}, \mathbf{v}_1^{\mathrm{T}}, \cdots, \mathbf{v}_K^{\mathrm{T}}, \mathbf{h}_1'^{,\mathrm{T}}, \cdots, \mathbf{h}_K'^{,\mathrm{T}})^{\mathrm{T}}$ and $\boldsymbol{\Lambda} = (\boldsymbol{\lambda}_1^{\mathrm{T}}, \cdots, \boldsymbol{\lambda}_K^{\mathrm{T}}, \boldsymbol{\mu}_1^{\mathrm{T}}, \cdots, \boldsymbol{\mu}_K^{\mathrm{T}}, \boldsymbol{\kappa}_1'^{,\mathrm{T}}, \cdots, \boldsymbol{\kappa}_K'^{,\mathrm{T}})^{\mathrm{T}}$ denote two N-dimensional state vectors of the discrete model, where $N = I$ horizontal grid points $\times$ K layers $\times$ 3 fields = 24 595 $\times$ K, excluding boundary points. The inner product $\langle \boldsymbol{\xi}, \boldsymbol{\Lambda} \rangle_{E^{-1}} = \boldsymbol{\xi}^{\mathrm{T}} \mathbf{E}^{-1} \boldsymbol{\Lambda}$ where the matrix $\mathbf{E}^{-1} = \mathrm{diag}(\mathbf{E}_u^{-1}, \mathbf{E}_v^{-1}, \mathbf{E}_{h'}^{-1})$ is block-diagonal, with blocks defined by

$$\mathbf{E}_u^{-1} = \mathbf{E}_v^{-1} = \mathrm{diag}(H_1\mathbf{I}, \cdots, H_K\mathbf{I}) \quad \text{and}$$

$$\mathbf{E}_{h'}^{-1} = \begin{pmatrix} G_{11}\mathbf{I} & & G_{1K}\mathbf{I} \\ & \ddots & \\ G_{K1}\mathbf{I} & & G_{KK}\mathbf{I} \end{pmatrix}, \tag{14}$$

where $\mathbf{I}$ is the $I \times I$ identity matrix. Note that each term in the discrete analogue of (13) should also be weighted by the corresponding area element $\Delta x \Delta y$ over which it is defined. However, since the horizontal resolution in the model is uniform ($\Delta x = \Delta y = 1°$), these additional weights can be omitted as only the relative weighting of each term in the inner product is of importance for preconditioning.

d. The cost function

The cost function is defined by (3), where $\mathbf{y}_n^o$ consists of complete maps of periodic sea level (SL) fields extracted from the model during a so-called truth run (see section 4). From Eq. (11), the nonzero elements in each row of $\mathbf{H}$ can be seen to be the coefficients $G_{1k}/\rho_0 g$ multiplying each h_k'. The weighting matrix is defined by $\mathbf{W}^{-1} = g\mathbf{I}$, implying that each SL observation is given equal weight. The multiplicative constant g clearly has no effect on the minimizing solution; it has been introduced to give the cost function dimensions of energy in order to be consistent with the energy preconditioner. With these choices of $\mathbf{W}^{-1}$ and $\mathbf{E}^{-1}$, the columns of the adjoint matrix $\mathbf{H}^*$ [Eq. (6)] are simply unit vectors in the direction of $h_1'(i, j)$ where (i, j) is the gridpoint location of the SL observation. In other words, only the adjoint of the upper-layer height field is directly forced in the adjoint equations implying that, in the limit of a null-length assimilation period, the descent algorithm would make a correction to $h_1'(i, j)$ to fit the SL observation but would leave all other fields unchanged. In this respect, the preconditioning matrix also defines how the SL information would be assimilated in a 3D variational analysis.

We have deliberately omitted a background constraint from the cost function since it is our intent to concentrate on the efficiency of the dynamics for improving the subsurface state estimate in the absence of prior information on background error statistics (most notably the vertical covariances). Some form of background constraint, whether it be a direct penalty on weighted high-order derivatives in the background error or the fields themselves (Thacker and Long 1988; Sheinbaum and Anderson 1990b), or a statistical penalty using estimates of the background error covariances (Tarantola 1987; Lorenc 1988), is usually required in practice to contain unrealistic ''noise'' in the minimizing solution, although a priori such noise can be expected to be less problematic in a perfect linear model where dissipation will act as an efficient filter during the forward model integration. Moreover, given that the SL data are supplied everywhere on the model grid, the data constraint itself will provide a large amount of spatial smoothness information which should be sufficient to ensure a smooth analysis.

e. Minimization

The gradient descent algorithm used in this study is the variable-storage QN routine (M1QN3) of Gilbert and Lemaréchal (1989). Each iteration, the algorithm builds an approximation to the inverse Hessian matrix using gradient and search directions from previous iterations. The inverse Hessian is initially approximated by the identity matrix [in the space whose metric is defined by (14)] and is then updated each iteration using the 10 most recent gradients and search directions. The optimal step length is computed using an exact line-search. In all experiments, the convergence criterion for minimization is taken to be a three-order-of-magnitude reduction in the norm of the preconditioned gradient.

4. The assimilation experiments

The assimilation experiments are conducted using the K½-layer model with K ranging from 1 to 4 (Table 1). The model is spunup for 5½ years from rest using the climatological FSU winds. The fields produced at this time (day 0) define the initial conditions of the truth run from which complete maps of SL observations are subsequently extracted. Unless otherwise stated, these SL maps will consist of the complete SL field on the 15th day of each month. The background initial conditions are taken to be the fields produced after 5 years of the spinup run. A forced integration from the background state using the ''correct'' wind field subsequently defines the ''incorrect'' model simulation (hereafter referred to as the background or control run). In all assimilation experiments, the background state is also used to define the starting point for the optimization.

a. The 1½-layer model

In the first experiment (expt A), SL data are assimilated into a 1½ layer model over a 6-month assimilation

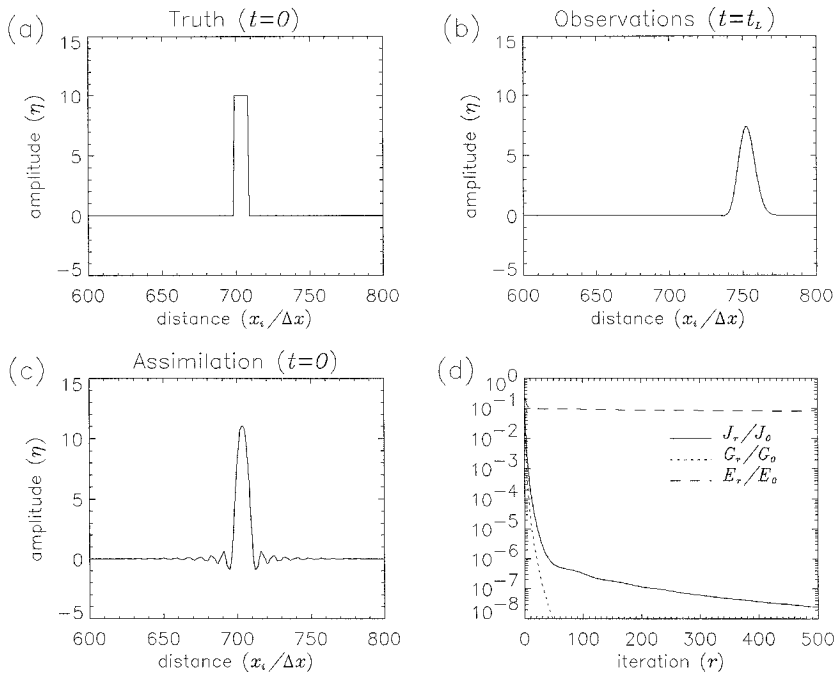

FIG. 2. The optimization experiment with the diffusive wave equation [Eq. (15) with $c = 1$ and $\nu = 0.01$]. (a) The true wave at $t = 0$; (b) the observed wave at $t = t_L = 250\Delta t$; and (c) the recovered wave at $t = 0$ obtained by optimizing the initial conditions to fit the observation of the wave in (b). (d) The cost function, J_r (solid); the squared norm of the gradient, $G_r = (\nabla J_r)^T(\nabla J_r)$ (dotted); and the squared norm of the wave error at $t = 0$, E_r (dashed), as a function of iteration number r. Each quantity has been normalized by its respective value at $r = 0$ and plotted with a $\log_{10}$ vertical scale.

period. The problem of determining the ocean state from SL data is not terribly challenging in this experiment since, in the 1½ layer model, SL is directly proportional to the perturbation height field. One complete field of SL is thus sufficient to specify unambiguously the height (mass) field, although it provides no direct information on the velocity field. The optimization converged after 25 iterations. The results (not shown) indicate an almost perfect reconstruction of the initial state (the velocity field as well as the directly observed height field) except in regions closely confined to the boundaries where difficulties can be expected because of the importance of dissipative processes there. Dissipation has the effect of removing information about the initial state during the course of the model integration, the implication in an inverse experiment being that the data will poorly resolve those features in the initial conditions that are effectively damped out by the observation time. Geometrically, the contours of constant cost will be strongly elongated in the directions in initial state space associated with these strongly damped features. This implies that the gradient of the cost function will only be slightly perturbed in these directions, which in turn implies that the rate of convergence of the optimization will be slow in these directions (Thacker 1989).

As a simple illustration of this point, we consider an experiment with the 1D wave equation

$$\frac{\partial \eta}{\partial t} + c\frac{\partial \eta}{\partial x} = \nu\frac{\partial^2 \eta}{\partial x^2}, \tag{15}$$

where $\eta = \eta(x, t)$ is the wave amplitude, c is the wave speed, and ν is the diffusion coefficient. The parameters are taken to be known with (nondimensional) values of $c = 1$ and $\nu = 0.01$. Equation (15) is solved numerically using finite difference techniques similar to those used for the ocean model. The spatial domain is taken to be the infinite interval $-\infty < x < \infty$ with boundary conditions such that $\eta \to 0$ as $x \to \pm\infty$. The corresponding adjoint of (15) is used to calculate the gradient of the cost function with respect to the initial conditions. The basic experiment we consider is that of trying to reconstruct an initial square wave (see Fig. 2a) by minimizing the squared difference from a perfect and complete observation of this wave at the end of the assimilation period, $t_L = 250\Delta t$ (Fig. 2b), where the first-guess initial condition for the model (i.e., the initial point for optimization) is taken to be zero.

In the absence of diffusion ($\nu = 0$), the result of the optimization is a perfect reconstruction of the initial conditions in only 4 iterations. When the experiment is repeated with nonzero diffusion, many more iterations are required to reduce the squared norm of the gradient (dotted curve in Fig. 2d) to a level of convergence comparable to the nondiffusive case. Moreover, a complete

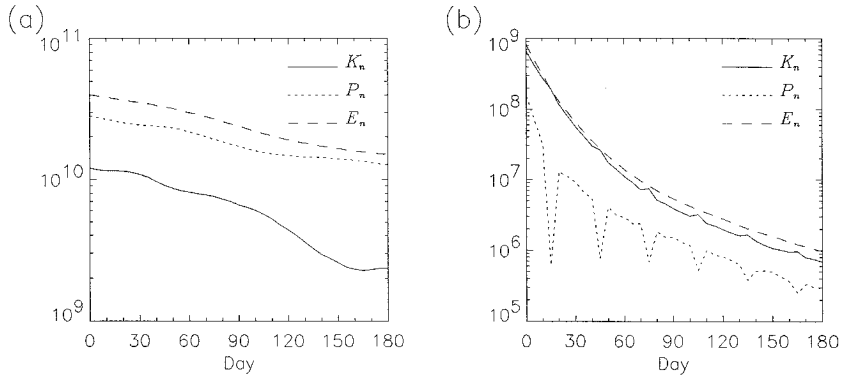

FIG. 3. Expt A: A 6-month experiment with SL assimilated once per month into the 1½-layer model. The evolution of the squared norm of the state error over the 6-month period for (a) the control run and (b) the assimilation run. The dashed curves in (a) and (b) show the total error, while the solid and dotted curves show the contributions to the total error from the velocity and height errors, respectively.

reconstruction of the square wave appears impossible as illustrated in Fig. 2c, which shows the optimized initial conditions after 500 iterations. Consequently, there are directions in initial state space in which these data contain little or no information. These directions are associated with eigenvectors of the Hessian matrix, which have zero, or nearly zero, eigenvalues. The Hessian in this case is ill-conditioned (i.e., the ratio of its largest and smallest eigenvalues—the condition number —is large relative to one), the practical implication of which is to slow the rate of convergence of the optimization as shown in Fig. 2d.

During gradient descent minimization, those linear combinations of control variables well constrained by the data will be optimized in the early stages of minimization, while those poorly constrained will be optimized in the later stages. In this example, it is the strongly damped small-scale components of the state vector that are poorly constrained by the data and are being adjusted in the later iterations. The large-scale signal, on the other hand, is recovered after only a few iterations. This is apparent in Fig. 2d, which shows the squared norm of the error in the initial conditions (dashed curve) essentially stabilizing after about 10 iterations. In fact, the optimized solution after 10 iterations resembles very closely the solution after 500 iterations indicating that nothing much is being gained through the additional iterations.

Returning to experiment A, the diffusive time scale of equatorially trapped waves can be estimated by considering the balance $\partial u/\partial t \sim \nu\partial^2 u/\partial x^2$ in (8). Approximating the spatial derivative with centered differences and assuming damped wave solutions of the form $u \sim e^{-\delta t}e^{ikx}$, where δ is the decay rate (δ^{-1} is the diffusion timescale) and $k = 2\pi/L_s$ the wavenumber (L_s is the wavelength), yields $\delta = 2\nu(1 - \cos k\Delta x)/(\Delta x)^2$. For the shortest resolvable wave, $L_s = 2\Delta x$, this corresponds to a timescale of 18 days, which for experiment A is about half the data sampling interval and one-tenth the assimilation period of 6 months. The waves most strongly affected by diffusion are short Rossby waves along the western boundary and coastal Kelvin waves along the eastern boundary. Further damping of coastal Kelvin waves occurs in the vicinity of the artificial northern and southern boundaries where Rayleigh friction is present to prevent these spurious waves from propagating back into the central Pacific. The maximum damping time scale associated with the Rayleigh friction is 5 days.

In the identical-twin experiments, the model and forcing are perfect so that even without data assimilation, any error in the initial state will gradually diminish with time by virtue of the linear dissipative dynamics of the model. This is evident in Fig. 3a, which shows the time evolution of the squared norm of the state error (dashed curve) for the control run. Splitting this quantity into height (dotted curve) and velocity (solid curve), error contributions indicates that height errors are dominant throughout the integration period.

The corresponding errors for the assimilation run are shown in Fig. 3b. The dramatic decrease in the errors relative to those of the control occurs since Laplacian diffusion is very efficient at dissipating the small scales where most of the error is confined after assimilation. Figure 3b also indicates a greater error reduction in the height term (dotted curve) compared to the velocity term (solid curve), which is opposite to their relative error reduction in the control run. This is not surprising, however, as height, which is one-to-one with SL in the 1½-layer model, is the directly assimilated quantity. The inverted spikes that appear in the height error occur at the observation times where a very close fit to the SL data has been achieved.

b. The 2½-layer model

The 6-month assimilation experiment is now repeated for the 2½-layer model (expt B). There is no longer a

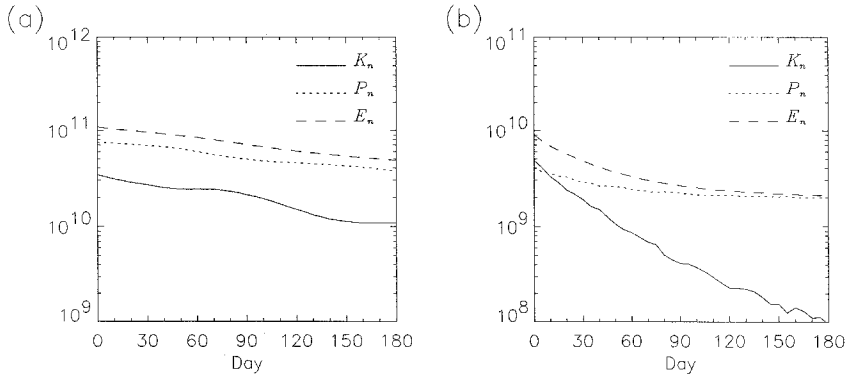

FIG. 4. As in Fig. 3 but for expt B with the 2½-layer model.

one-to-one correspondence between SL and the height field in the 2½-layer model. Therefore, contrary to experiment A, the height (mass) field at each point is no longer unambiguously specified by one SL measurement; there is an infinite combination of height fields, h_1' and h_2', which is consistent with a given SL observation. The problem is then not only to determine the unobserved velocity field but also to find the correct vertical distribution of the height field since the altimeter provides only integral (i.e., indirect) information on the latter.

The optimization converged after 48 iterations. The cost function was reduced by over four orders of magnitude and the squared norm of the initial state error by over one order of magnitude. The time evolution of the squared norm of the state error in the control and assimilation runs is illustrated in Figs. 4a and 4b respectively. Both height and velocity fields are significantly improved by the assimilation (notice the difference in the vertical scales in Figs. 4a and 4b). In contrast to experiment A, however, it is the height error (dotted curve) not the velocity error (solid curve) that remains dominant after assimilation.

Figure 5a shows the error in the initial (i.e., day 0) height field in layer 2 before assimilation, while Fig. 5b shows the corresponding error after assimilation. Comparing these figures, it can be seen that most of the coherent error structure (primarily associated with Rossby waves) has been eliminated as a result of the assimilation. The error reduction is particularly good in the upper layer (not shown), which is expected since SL is a direct measurement of the perturbation pressure in that layer. More encouraging is the error reduction in Fig. 5b, which implies that information has been effectively propagated into the lower layer.

c. The 3½-layer model

In experiment C, the basic 6-month assimilation experiment using SL data from day 15 of each month is repeated using the 3½-layer model. Convergence of the optimization was reached after 50 iterations. Although the SL data are fitted very closely, the assimilation was not nearly as successful in reducing the state error as in the 1½-layer and 2½-layer experiments; the relative reduction in the squared norm of the initial state error is 86% greater in experiment A and 44% greater in experiment B. Indeed, the rate at which the error is reduced over the assimilation interval is comparable to that in the control run (not shown), indicating that the errors are only weakly affected by diffusion in the forward integration and thus predominantly large scale. Contour plots of the height field errors illustrate this more clearly. Figure 6 shows the error in the initial height field in layer 3 after assimilation. The error before assimilation (i.e., the background error) has similar structure to Fig. 5a for the 2½-layer model. The assimilation has had very little effect in many regions of the model. It appears to have been more effective in breaking up the error structure within the central equatorial region, and less effective in the eastern equatorial Pacific at higher latitudes where the errors, although of somewhat smaller amplitude than those in the background field, are still prominent.

Although the global error (as measured by the energy inner product) is still decreasing on the final iteration (not shown), indicating that useful (large scale) information is still being extracted in these later iterations, locally the error has actually intensified in pockets near the western boundary. In this experiment, the local error amplification is greatest in the lower layer near the western boundary where frictional effects are important and, hence, where the state is likely to be least sensitive to the SL observations.

The fact that there are far fewer SL observations than control variables in experiment C (6 SL maps $\times$ 8288 SL grid points = 49 728 observations compared to N = 73 785 control variables) indicates that a priori we cannot expect to constrain all aspects of the model state during assimilation. To do so, there must be at least as many observations as control variables, although, as illustrated in the next section, such a requirement is generally not sufficient. Here we can increase the quantity of SL data either by increasing the sampling rate or by

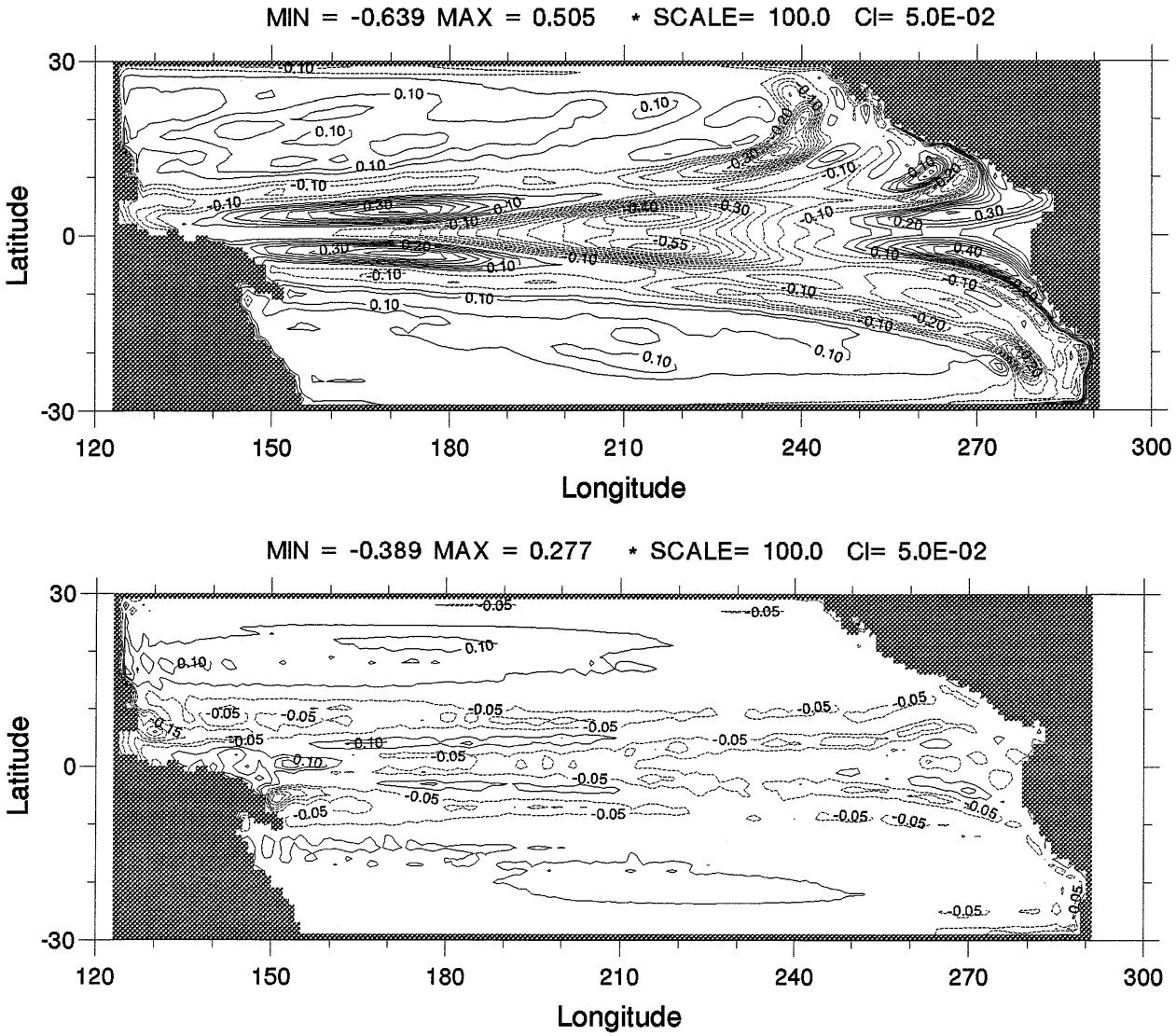

FIG. 5. Expt B: As for expt A (see Fig. 3 caption) but with the 2½-layer model. (a) The control run error (i.e., difference from truth) in the height field in layer 2 at the beginning of the integration period (day 0). (b) As in (a) but for the assimilation run. The contour interval is 5.0 m in both plots.

extending the assimilation period. It is thus of interest to examine to what extent the relatively poor performance in experiment C is a consequence of the sampling interval being too long or the assimilation period being too short.

1) SENSITIVITY TO THE SAMPLING INTERVAL

In experiment D, the same 6-month assimilation period of experiment C is used but now the data sampling rate is increased from one full field every month to one full field of SL data every 10 days. This corresponds to a threefold increase in the total number of observations and is more than double the number of control variables. This generous supply of SL data is not unrealistic when compared to the data coverage from current satellite altimeters: for example, the 10-day repeat-orbit period of TOPEX/Poseidon is capable of mapping the surface with a cross-track resolution at the equator of approximately 2.5°. This is slightly coarser than the 1° resolution used in the experiments but is still adequate for resolving the important large-scale features in the tropical oceans.

The squared norm of the initial state error in the experiment D analysis, though slightly smaller than in experiment C, was still found to be unacceptably large (only a 3% improvement). Increasing the sampling rate even further (expt E), to an extremely generous one complete field of SL data every day (equal to 180 datasets) did not significantly improve the analysis either ($<$5% improvement). In fact, the errors at the end of the assimilation period had very similar magnitudes to

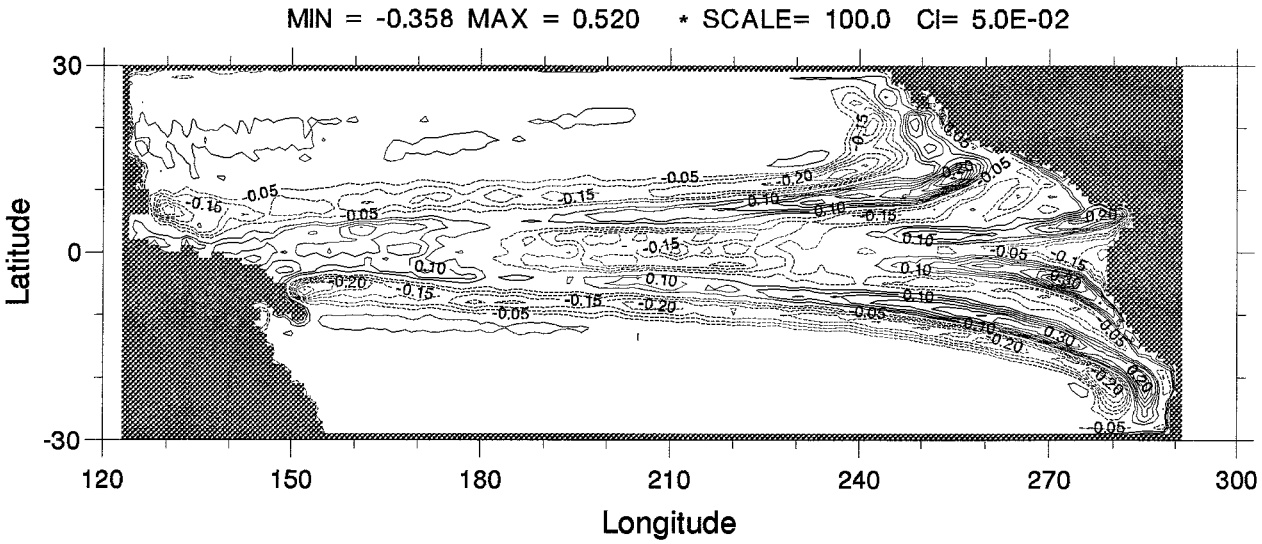

FIG. 6. Expt C: As for expt A (see Fig. 3 caption) but with the 3½-layer model. The error in the initial (i.e., day 0) height field in layer 2 after assimilation. The contour interval is 5.0 m.

those in experiment D. The most noticeable impact of the extra SL observations was to smooth the analysis. Increasing the temporal density of the observations has thus primarily helped to sort out some of the small-scale components of the flow. On the other hand, those large-scale components of the flow that were poorly constrained by the observations with the longer sampling interval have not benefited much from the additional observations, while those that were already relatively well constrained have only been supplied redundant information.

The pronounced latitudinal dependence in the height errors suggests that zonally averaged root-mean-square (rms) errors should provide a useful diagnostic for assessing the impact of the assimilation. The rms errors in the optimized (dashed curve) and background (solid curve) initial conditions for the velocities and heights in all layers are shown in Fig. 7 for experiment C. Both the height and zonal velocity fields in each layer show noticeable improvements from assimilating SL, particularly over the equatorial region where the background errors are largest. When the sampling rate is increased (expt D and expt E), the rms errors in these fields are only slightly reduced compared to those in experiment C. The rms errors in the meridional velocity field, on the other hand, are much more sensitive to changes in the sampling rate. In experiment C, the errors are comparable to those in the background field, while in experiment D and, in particular, experiment E these errors are markedly reduced, especially in layer 1.

Closer examination of the background errors in the meridional velocity field reveals that the greatest contribution to the error is in the neighborhood of the western boundary where meridional currents are strongest due to western intensification. These errors remain prominent even after assimilation in experiment C. In experiment E, however, they are largely suppressed possibly because, with the shorter sampling interval, the assimilation algorithm is able to distinguish many of the short Rossby waves that dominate the variability in this area. As these waves are strongly affected by dissipation, one would expect difficulty in reconstructing them when the observation sampling interval is greater than their diffusive timescale. For example, the diffusive timescale of 18 days for the $2\Delta x$ waves is greater than the 1-day sampling interval in experiment E but less than the 30-day sampling interval in experiment C.

2) SENSITIVITY TO THE ASSIMILATION PERIOD

In experiment F the assimilation period is extended from 6 months to 1 year, and a 4D-Var inversion is performed given SL observations every 10 days. The reduction in the initial state error is greater than in all previous 3½-layer experiments conducted with the 6-month assimilation period. Greatest improvements are observed at lower latitudes as shown in the zonally averaged rms errors for the height field (compare Figs. 8a–c with Figs. 7a–c). The pronounced dip in the errors over the equatorial region suggests that the assimilation has been particularly effective in reconstructing equatorially trapped waves.

The theoretical results of Webb and Moore (1986) may be brought to bear on the experimental results observed here. They showed that successful reconstruction of the model's subsurface fields from altimetry is crucially linked to the phase separation that develops over the observation interval between the different vertical wave modes. First, we illustrate this point in a simplified framework by considering an assimilation experiment in a dynamical system consisting of two independent waves, whose evolution is governed by

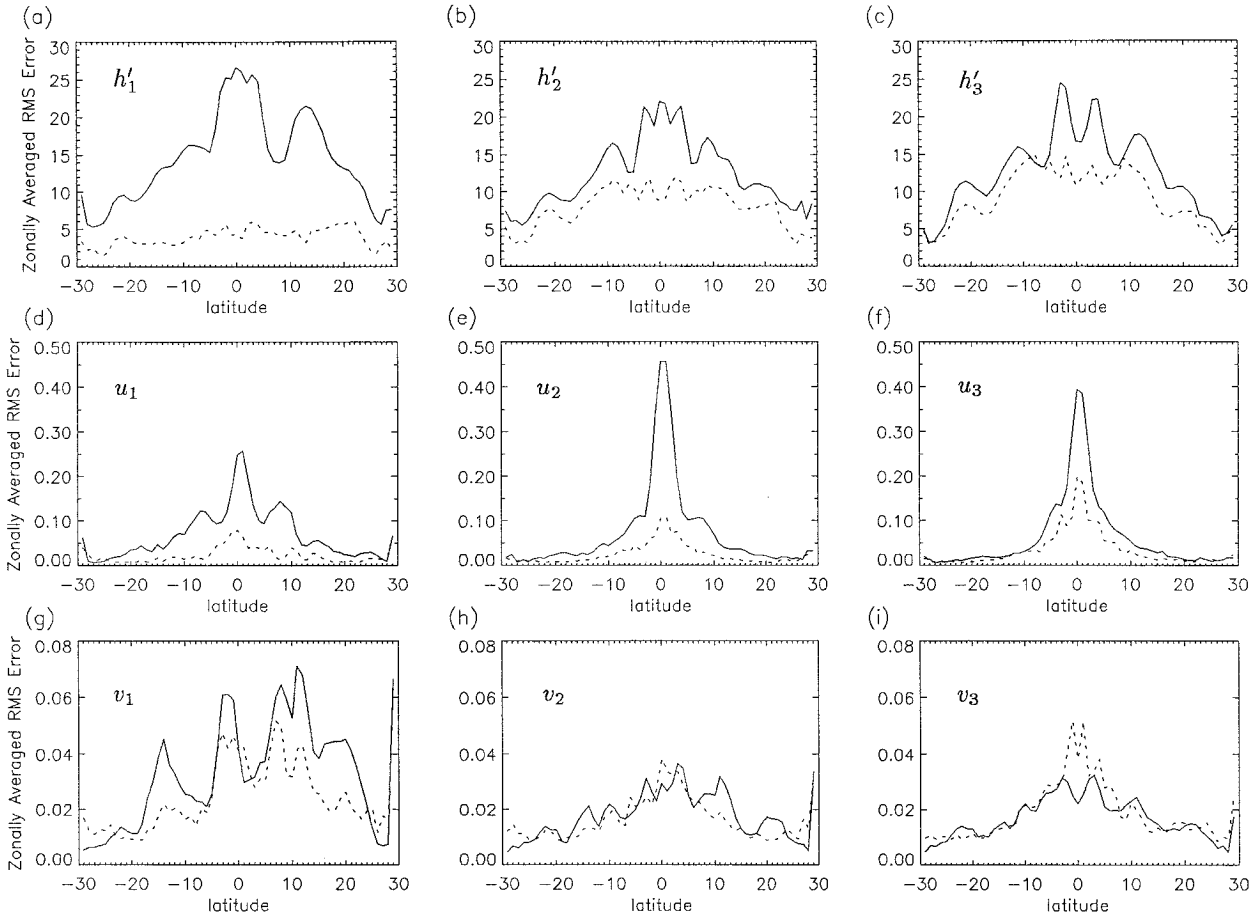

FIG. 7. Expt C: As for expt A (see Fig. 3 caption) but with the 3½-layer model. The zonally averaged rms errors in the initial conditions: (a)–(c) the height errors in layers 1 to 3; (d)–(f) the zonal velocity errors in layers 1 to 3; and (g)–(i) the meridional velocity errors in layers 1 to 3. The solid curves are the rms errors in the control (background) initial conditions, and the dotted curves are the rms errors in the optimized initial conditions. The labels for the vertical axes are in MKS units.

$$\frac{\partial \eta}{\partial t} + c_1 \frac{\partial \eta}{\partial x} = 0; \qquad \frac{\partial \zeta}{\partial t} + c_2 \frac{\partial \zeta}{\partial x} = 0, \tag{16}$$

where η and ζ are the (unknown) wave amplitudes, and c_1 and c_2 the (known) wave speeds. We are loosely thinking of η and ζ as the perturbations to the top surface from two vertical modes with wave speeds c_1 and c_2. In the first experiment, the wave speeds are chosen to be $c_1 = 0.5$ and $c_2 = 1.0$. The control variables consist of the initial conditions $\eta(x, 0)$ and $\zeta(x, 0)$. The true initial conditions to be reconstructed by the assimilation are square waves, which for simplicity are taken to be of equal amplitude (Figs. 9a and 9b). The first-guess initial conditions for the optimization are taken to be zero.

We suppose that the observing system is of an altimeter type, providing perfect measurements of the superposition $(\eta + \zeta)$ of these waves, and that two complete fields are available at the beginning and end of the assimilation period. The ''altimeter'' observations at t_0 are illustrated in Fig. 9c; the true wave solutions at t_L are shown in Figs. 9d and 9e; and the altimeter observations at this time are illustrated in Fig. 9f. The waves do not maintain their shape as they evolve because of numerical dispersion arising from the centered differencing of the spatial term. [The solution can in fact be shown analytically to be a linear combination of Bessel functions of the first kind (Mesinger and Arakawa 1976).] Note that neither of these two datasets is sufficient by themselves to determine the individual wave amplitudes uniquely; at any given time there is an infinite combination of wave amplitudes consistent with the given altimeter observation. The ambiguity can be removed only if the data are supplemented with additional (prior) information. In 4D-Var the extra information comes from the time-dependent dynamics incorporated directly into the assimilation scheme. Consequently, both datasets can be propagated back to a common reference time (the initial conditions) and then used simultaneously in the fitting procedure. The problem is then, in principle at least, fully determined. The result of the minimization after 100 iterations (the num-

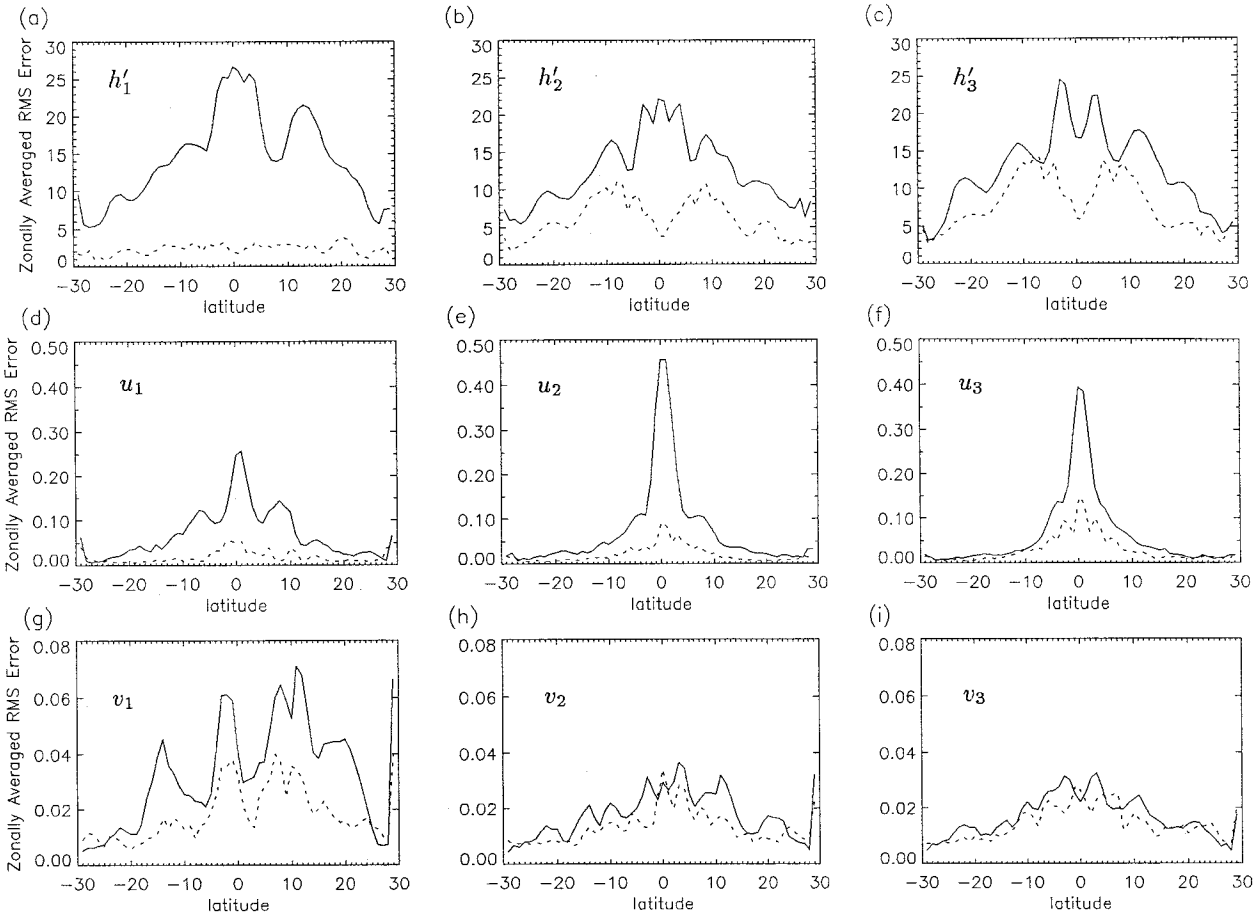

FIG. 8. As in Fig. 7 but for expt F with the 1-yr assimilation period.

ber required to satisfy the convergence criterion of an eight order of magnitude reduction in the squared norm of the gradient) is shown in Figs. 9g and 9h. Both waves are well recovered at this point and in fact can be fully recovered by performing several more iterations.

Now consider the dynamical system (16) consisting of two waves with identical phase speeds, $c_1 = c_2 = 1.0$. As in the previous example, the altimeter observations are taken to be at t_0 and t_L. The optimization converges very rapidly as shown in Fig. 10c but to a solution that is inconsistent with truth (Figs. 10a and 10b). The wave solutions are nevertheless completely consistent with the altimeter observations as can be verified visually by adding the waves in Figs. 10a and 10b and comparing the result with the data shown in Fig. 9c.

That waves with identical phase speeds cannot be distinguished from measurements of their superposition is obvious when the dynamical equations are rewritten in terms of new variables, $M = \eta + \zeta$ and $D = \eta - \zeta$, describing the superposition and difference of the wave amplitudes respectively. Adding and subtracting the equations in (16) and taking $c_1 = c_2 = c$ yields two independent equations for M and D. Estimating the individual amplitudes, η and ζ, in (16) is equivalent to estimating M and D in this new system. As the altimeter measurement provides direct information on M but no information on D, the cost function is strictly a function of the additive field M. The gradient of the cost function with respect to the difference field is thus identically zero ($\partial J/\partial D \equiv 0$) implying that, within the optimization process, the difference field will remain unchanged from its value prior to assimilation. Indeed, the optimized difference of the recovered wave solutions in Figs. 11a and 11b is consistent with the zero difference field of the first guess.

The sensitivity of the results to the choice of wave speeds is studied further by repeating the experiment with wave speeds that are very similar to each other ($c_1 = 0.99$ and $c_2 = 1.0$). In contrast to the previous example, the cost function will no longer be completely "flat" in the direction in parameter space associated with the difference field; the phase separation that develops between the waves over the assimilation cycle (i.e., between observation times t_0 and t_L), though very slight, will be sufficient to perturb the cost function in this direction such that $\partial J/\partial D \neq 0$. Graphically, it is very difficult to distinguish between the waves at the

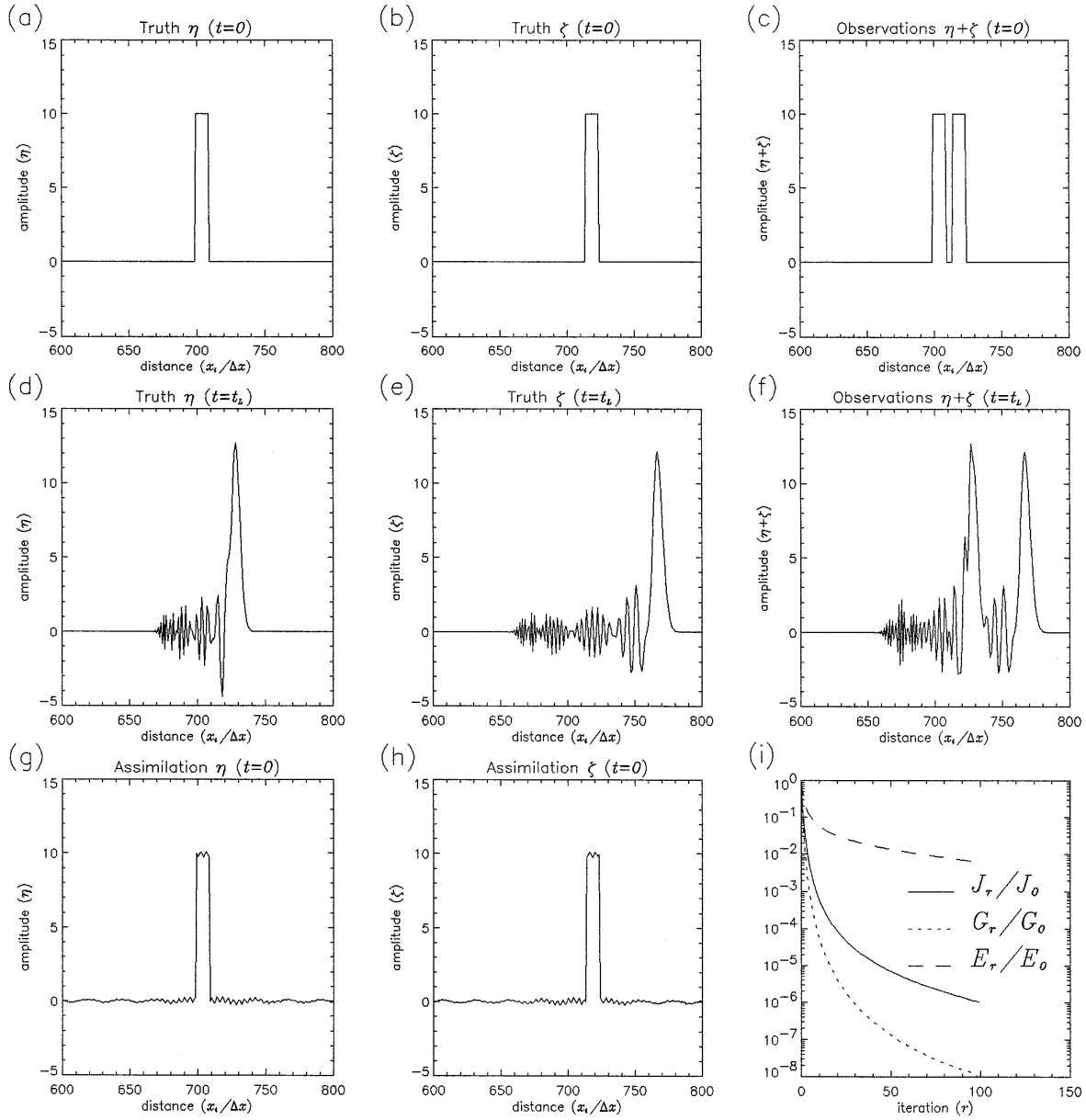

FIG. 9. The optimization experiment with the dynamical system consisting of two independent waves with different phase speeds [Eq. (16) with $c_1 = 0.5$ and $c_2 = 1.0$]. (a)–(b) The true waves at $t = 0$. (c) The observed superposition of the waves at $t = 0$. (d)–(f) As in (a)–(c) but at $t = t_L = 250\Delta t$. (g)–(h) The recovered waves at $t = 0$ obtained by optimizing the initial conditions to fit the observations at $t = 0$ and $t = t_L$. (i) The optimization quantities as defined in Fig. 2d.

end of the assimilation period. Remarkably, however, the optimization is successful in reconstructing most of the wave signal after a little more than 100 iterations (Figs. 11a and 11b). There is, nevertheless, considerable difficulty in extracting this information as evident in the nonmonotonic reduction of the cost gradient (Fig. 11c). In fact, only after many more iterations could we achieve a level of error reduction comparable to the first example.

The phase separation that develops between the waves over the assimilation cycle is thus a crucial factor in determining the conditioning of the Hessian. In particular, a small phase separation leads to a poorly conditioned Hessian (i.e., one with small eigenvalues), whereas a large phase separation leads to a well-conditioned Hessian (i.e., one with eigenvalues that are roughly of the same order of magnitude). In this example, the eigenvalues are essentially clustered into two well-separated groups; one group containing relatively large eigenvalues is associated with the well-constrained additive field, while the other group containing very small eigenvalues is associated with the poorly constrained

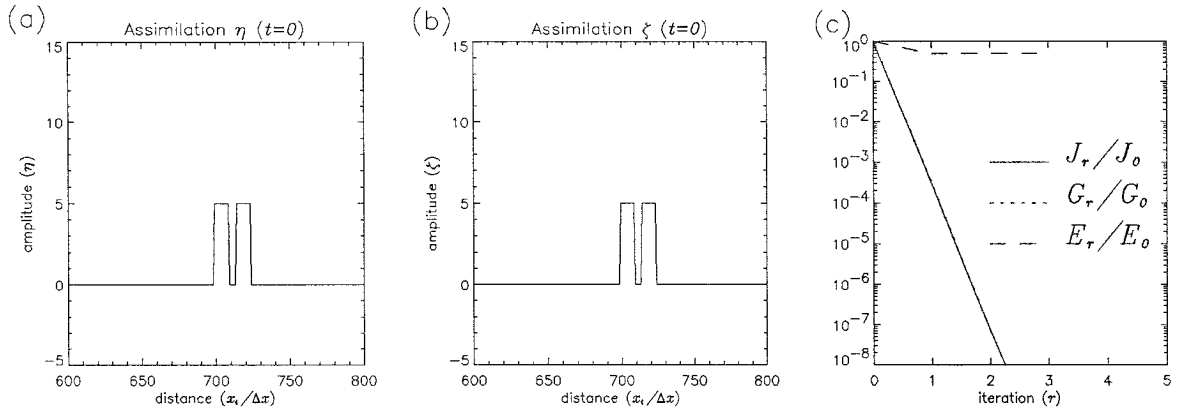

FIG. 10. The optimization experiment with the dynamical system consisting of two independent waves with identical phase speeds ($c_1 = c_2 = 1.0$). (a)–(b) The recovered waves at $t = 0$ obtained by optimizing the initial conditions to fit the observations at $t = 0$ and $t = t_L$. (c) The optimization quantities as defined in Fig. 2d.

difference field. The practical consequence of this clustering effect is to lead to a relatively quick convergence of the additive field but to a much slower convergence of the difference field as the eigenvalues associated with the latter are very close to zero. Even so, these experiments demonstrate that, with perfect data, an acceptable solution can eventually be reached given enough computing effort and provided that the problem is not fundamentally degenerate as in the previous example.

In the ocean model, the success of the assimilation at lower latitudes can thus be attributed to the faster dynamical adjustment of the equatorial ocean in comparison to the adjustment of the ocean at higher latitudes. As a consequence, there is greater opportunity for the different vertical wave modes to develop significant phase separations on much shorter timescales. The propagation speeds of low-frequency long Rossby waves can be estimated using the mode speeds c_n listed in Table 1. At low latitudes, long Rossby waves exist as a set of discrete meridional modes that have westward phase propagation and speeds $c_{n,m}^{\mathrm{LRW}} \approx c_n/(2m + 1)$ where m is the meridional mode number ($m = 1, 2, \cdots$) (Gill 1982). For example, for the first meridional, second baroclinic mode and first meridional, third baroclinic mode the phase speeds are approximately $c_{2,1} = 0.33$ m s^{-1} and $c_{3,1} = 0.23$ m s^{-1} respectively. Thus, these waves travel distances of approximately 2.6° and 1.8° over the 10-day observation interval, and 93.6° and 64.8° over the one-year assimilation period. On the other hand, the one-year assimilation period is still quite short for long Rossby waves of the second and third baroclinic mode to propagate more than a few degrees at higher latitudes. The propagation speeds of off-equatorial long Rossby waves are approximated by $c_n^{\mathrm{LRW}} \approx \beta c_n^2/f_0^2$, where f_0 gives the local value of the Coriolis parameter (Gill 1982). At 20° latitude, $c_2 = 0.9$ cm s^{-1} for the second baroclinic mode and $c_3 = 0.4$ cm s^{-1} for the third baroclinic mode, implying that in one year these waves travel distances of only 2.6° and 1.2°. Therefore, one would expect difficulty in distinguishing these higher modes from altimetry since their phase separation over the assimilation period is only very slight. Indeed, these ideas are further supported by Figs. 12a and 12b, which show respectively the background and analysis

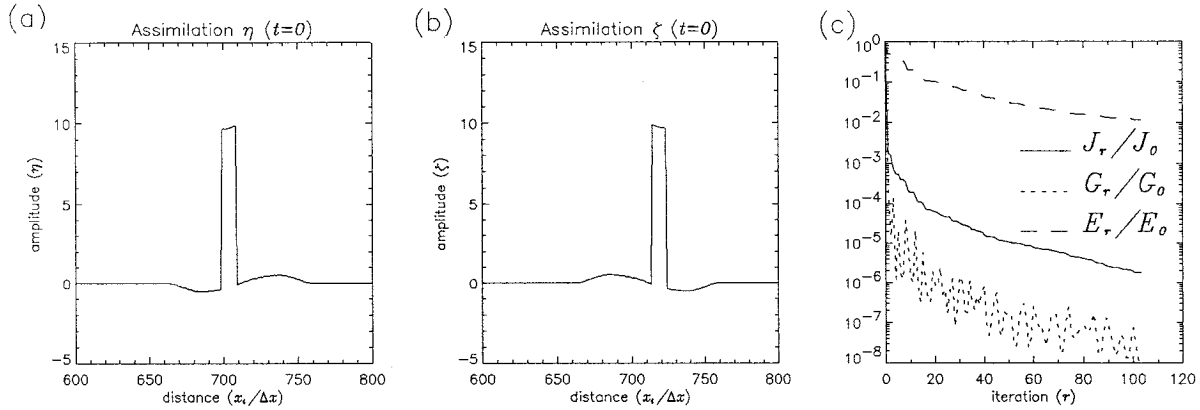

FIG. 11. As in Fig. 10 but for two independent waves with similar phase speeds ($c_1 = 0.99$ and $c_2 = 1.0$).

error in the layer-3 height field at the end of the assimilation period (day 360). The errors in the equatorial region, where the wave speeds are fastest, are small. At 10°–15°N/S, however, 1 year is not long enough to allow reconstruction of the waves. In the region 20°–30°N/S, the error is also small because there the signal was small and the background was accurate (Fig. 12a).

d. The 4½-layer model

In view of the results with the 3½-layer model, one could expect greater difficulty in reconstructing the subsurface fields in a 4½-layer model given SL data alone. Figure 13 shows the zonally averaged rms errors in the initial conditions after a 1-yr optimization experiment using the 4½-layer model with SL data assimilated every 10 days (expt G). As in the 3½-layer experiments, the error reduction relative to the control (the difference between the solid and dotted curves in the panels in Fig. 13) indicates, though somewhat less impressively, that SL information is more effectively communicated into the subsurface at lower latitudes than at higher latitudes (compare the relative reductions in the subsurface height field errors in Figs. 8b–c for expt. F with those in Figs. 13a–c for expt G). Overall, however, the recovery of the subsurface flow is worse than in experiment F, which suggests that additional features in the state are now poorly constrained by the data. This would seem reasonable in light of the previous discussion since the inclusion of another active layer introduces a slower baroclinic mode of propagation, which should be difficult to distinguish from the other slow modes over this relatively short assimilation period.

5. Discussion and conclusions

A 4D variational method has been used to assimilate simulated altimeter data into a multilayer linear model of the tropical Pacific Ocean. The experiments have all been of the identical-twin type. Complete SL maps extracted from the model in a control integration played the role of the altimeter observations in the assimilation experiments. As SL provides only single-layer information (it is directly proportional to the pressure in the upper layer), the issue of primary interest has been the recovery of the flow structure in the deeper layers. This has been explored using models with up to four active layers.

Two main points have emerged from the experiments. First, concerning the usefulness of altimeter data, the experiments have demonstrated quite convincingly that information about the 3D ocean state is more easily extracted near the equator than in off-equatorial regions. The crucial factor governing the success of the assimilation is the phase separation that develops between the different baroclinic modes over the observation interval (a result previously demonstrated in a theoretical framework by Webb and Moore 1986). This phase difference manifests itself in the evolving SL fields. In off-equatorial regions, where the variability is dominated by slowly propagating baroclinic Rossby waves, the highest modes can only produce significant phase separations after several years or even decades depending on how many modes are present in the system. As a result, the higher modes cannot be easily distinguished for an assimilation period on the order of one year, as was evident in the 3½-layer and 4½-layer model experiments which showed very distinctive patterns of Rossby wave error persisting after assimilation. On the other hand, SL information is much more easily transferred to depth near the equator where the faster propagation speeds of large-scale waves enable greater phase separations to develop between the higher modes on a much shorter timescale.

These findings are in general agreement with those of Moore (1986), who studied the altimeter problem in a 2½-layer model of the Indian Ocean using a simpler assimilation strategy. In that work, a sequential, univariate scheme was used to project the data directly onto the dynamical modes of the model at each analysis time. Like many other simpler assimilation schemes, the projection operator was fixed in time. Moreover, as the scheme was univariate, only the modal coefficients of the pressure field (i.e., only the modal coefficients which directly contribute to the SL displacement) were updated during each analysis. Information from the updated pressure field was then transferred to the velocity field as the model adjusted during its subsequent integration to the next analysis time. This differs very much from the 4D-Var approach, which produces a dynamically balanced analysis over the entire assimilation period by adjusting the initial conditions to fit the observations. In 4D-Var, the data are projected onto the initial state (the velocity field as well as the pressure field) via a projection operator, which is defined implicitly in the algorithm in terms of the time-dependent model and its adjoint.

This leads to the second main point worth emphasizing, which is the importance of fully exploiting equatorial wave dynamics within the assimilation procedure in view of their role as efficient information propagators. Indeed, 4D-Var has been shown here, albeit in highly favorable conditions, to be a potentially effective method for dynamically projecting SL data to depth within the equatorial belt. However, the results also indicate that the task becomes significantly more difficult as soon as the vertical resolution is extended to include more than a couple of layers, especially outside the equatorial belt where a close fit to the SL data could be achieved while leaving the subsurface fields grossly in error. Successful reconstruction of the subsurface flow in the linear baroclinic model depends on our ability to capture the phase separation between the higher modes within the assimilation procedure, a problem that will be further complicated with real data if the model phase speeds, as determined by prespecified stratification parameters,

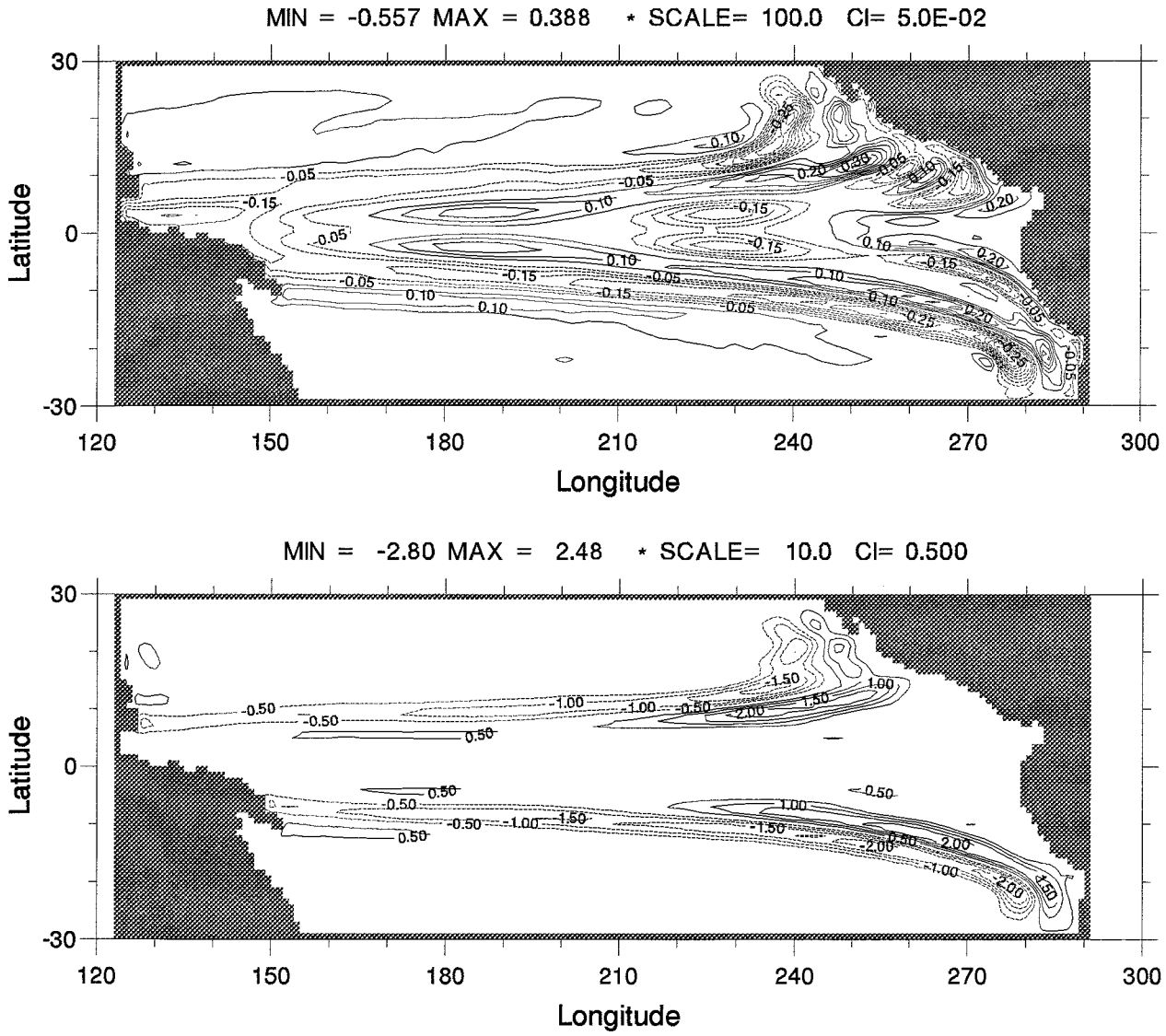

FIG. 12. Expt F: A 1-yr experiment with SL assimilated once every 10 days into the 3½ layer model. (a) The control run error in the height field in layer 3 at the end of the integration period (day 360). (b) As in (a) but for the assimilation run. The contour interval is 5.0 m in both plots.

are inconsistent with those observed, or if the phase separations are masked by sufficiently large observation errors. Without additional information, assimilation experiments of considerably longer duration than a year must be considered. This in turn raises the technical issue of how to cycle 4D-Var since the length of an inversion period will be limited by, for example, dissipation when only initial conditions are adjusted. Of fundamental importance within the cycling procedure is the estimation of the inverse of the Hessian matrix (the analysis error covariance matrix of the initial conditions) and the propagation of this matrix forward in time (e.g., as in the Kalman filter) to the beginning of the next assimilation cycle where it can be used, together with the estimate of the state at this time, in a background term for the next assimilation cycle. Major advances in this important area are currently being made at NWP centers (e.g., for recent advances at ECMWF see Fisher and Courtier 1995) and the possibility of carrying these techniques over to the ocean assimilation problem should be explored in future studies.

On timescales of a year or less, improvements to the assimilation scheme and extensions of the assimilation database must be envisaged in order to obtain an accurate initialization of ocean models that support more than two baroclinic modes. A major area of development needed for the assimilation scheme is in the background constraint, which in this study was ignored altogether. It is in the formulation of the background constraint where some of the ideas developed for the simpler as-

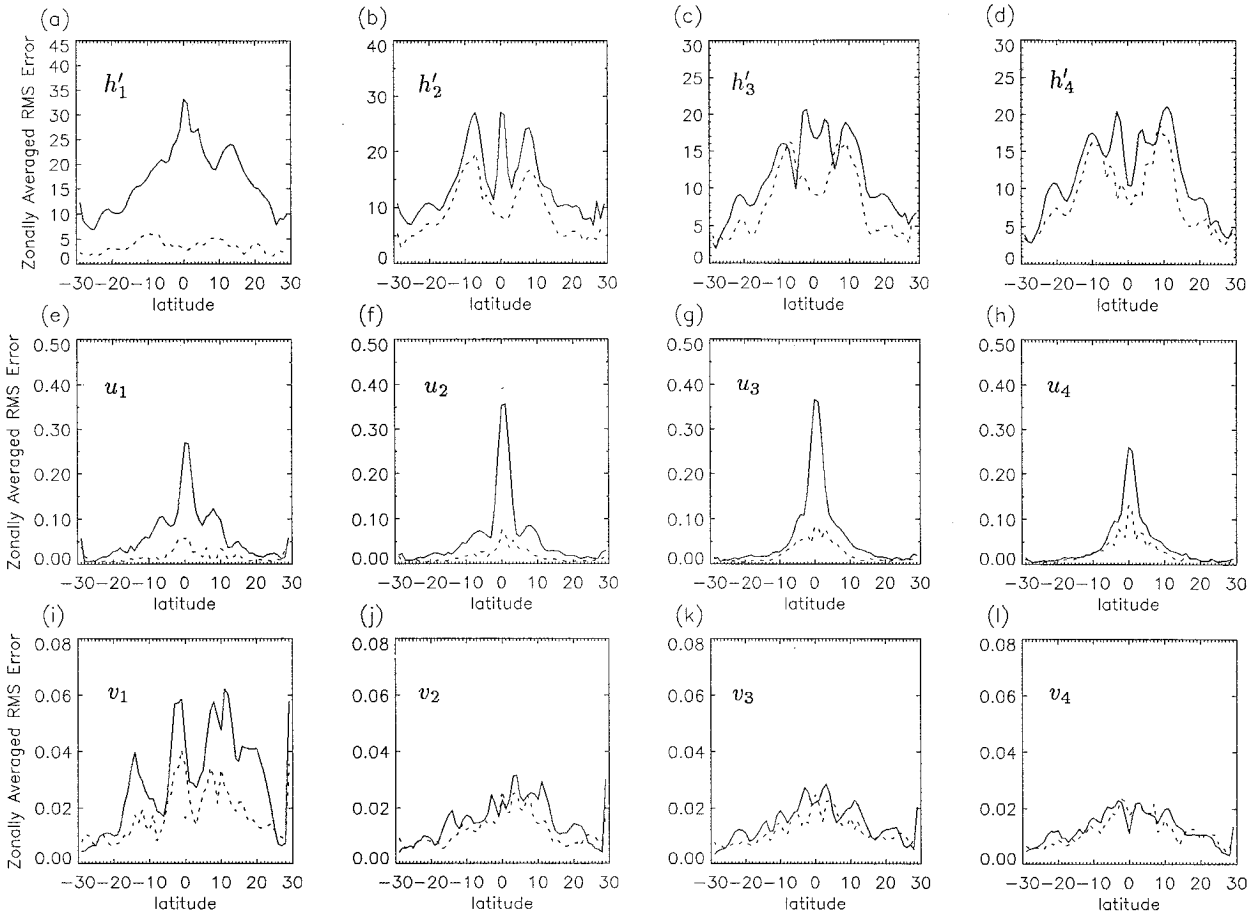

FIG. 13. Expt G: As for expt F (see Fig. 12 caption) but with the 4½ layer model. The zonally averaged rms errors in the initial conditions: (a)–(d) The height errors in layers 1 to 4; (e)–(h) the zonal velocity errors in layers 1 to 4; and (i)–(l) the meridional velocity errors in layers 1 to 4. The solid curves are the rms errors in the the control (background) initial conditions, and the dotted curves are the rms errors in the optimized initial conditions. The labels for the vertical axes are in MKS units.

similation schemes may be tested in the 4D-Var framework. For example, vertical correlation models similar to those of Ezer and Mellor (1994) and Mellor and Ezer (1993) may be incorporated into the background term, the hope being that this statistical information would complement the dynamical information imposed by the model constraint and thus improve the recovery of the subsurface flow from altimeter data. The work of Haines (1991, 1993, 1994) suggests that another possible way of improving the information exchange might be to include an explicit penalty on changes to the background potential vorticity in the lower layers. While this constraint has been shown to work well for sequential assimilation schemes applied to eddy-resolving models of the midlatitude oceans, it remains to be tested in a 4D-Var framework with tropical ocean models such as the one used in this study.

Extending the assimilation database to include, for example, direct measurements of the ocean's subsurface thermal structure is the most obvious next step in the development of a 3D ocean initialization scheme for the tropical Pacific. In this respect, the altimeter assimilation problem should be viewed within the more general context of operational data assimilation, which will ultimately combine data of all types to produce the best possible ocean analysis. In the tropical Pacific, subsurface thermal measurements are routinely available through a comprehensive observing system comprising the Tropical Atmosphere–Ocean (TAO) array of moored buoys and the ship-of-opportunity network of expendable bathythermographs (XBTs). The extent to which the combined assimilation of TAO, XBT, and altimeter data can determine the vertical structure of the ocean model is an important question that will be addressed in a future paper.

Acknowledgments. This work has been conducted as part of the first author's Ph.D. thesis. He wishes to thank the Canadian Natural Sciences and Engineering Research Council for providing financial support in the form of a postgraduate research scholarship. The minimization routine used in the experiments was kindly

provided by J.-C. Gilbert at INRIA. Helpful comments on the manuscript were received from Andy Moore and Julio Sheinbaum.

REFERENCES

Anderson, D. L. T., J. Sheinbaum, and K. Haines, 1996: Data assimilation in ocean models. *Rep. Prog. Phys.*, **59,** 1–58.

Bengtsson, L., 1979: On the use of a time sequence of surface pressures in four-dimensional data assimilation. *Tellus,* **32,** 189–196.

Berry, P., and J. Marshall, 1989: Ocean modelling studies in support of altimetry. *Dyn. Atmos. Oceans,* **13,** 269–300.

Busalacchi, A., and J. J. O'Brien, 1981: Interannual variability of the equatorial Pacific in the 1960's. *J. Geophys. Res.*, **86,** 10 901–10 907.

Cane, M. A., 1979: The response of an equatorial ocean to simple wind stress patterns: I Model formulation and analytic results. *J. Mar. Res.,* **57,** 233–252.

Cooper, M., and K. Haines, 1996: Altimetric assimilation with water property conservation. *J. Geophys. Res.*, **101,** 1059–1077.

Courtier, P., and O. Talagrand, 1987: Variational assimilation of meteorological observations with the adjoint vorticity equation, II: Numerical results. *Quart. J. Roy. Meteor. Soc.*, **113,** 1329–1347.

——, and ——, 1990: Variational assimilation of meteorological observations with the direct and adjoint shallow-water equations. *Tellus,* **42A,** 531–549.

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

Ezer, T., and G. Mellor, 1994: Continuous assimilation of Geosat altimeter data into a three-dimensional primitive equation Gulf Stream model. *J. Phys. Oceanogr.*, **24,** 832–847.

Fisher, M., and P. Courtier, 1995: Three algorithms for estimating the covariance matrix of analysis error in incremental variational data assimilation. *Proc. Second WMO Int. Symp. on Assimilation of Observations in Meteorology and Oceanography,* Vol. 1, Tokyo, Japan, World Meteor. Org., 229–234.

Gelb, A., 1974: *Applied Optimal Estimation.* Academic Press, 374 pp.

Ghil, M., and P. Malanotte-Rizzoli, 1991: Data assimilation in meteorology and oceanography. *Advances in Geophysics,* Vol. 33, Academic Press, 141–266.

Gilbert, J. C., and C. Lemaréchal, 1989: Some numerical experiments with variable-storage quasi-Newton algorithms. *Math. Program B,* **45,** 407–435.

Gill, A. E., 1982: *Atmosphere–Ocean Dynamics.* Academic Press, 662 pp.

Gill, P. E., W. Murray, and M. H. Wright, 1980: *Practical Optimization.* Academic Press, 401 pp.

Haines, K., 1991: A direct method for assimilating sea surface height data into ocean models with adjustments to the deep circulation. *J. Phys. Oceanogr.,* **21,** 843–868.

——, 1994: Dynamics and data assimilation in oceanography. *Data Assimilation: Tools for Modelling of the Ocean in a Global Change Perspective,* P. P. Brasseur and J. C. Nihoul, Eds., NATO ASI Series, Springer-Verlag, 1–32.

——, P. Malanotte-Rizzoli, R. E. Young, and W. R. Holland, 1993: A comparison of two methods for the assimilation of altimeter data into a shallow water model. *Dyn. Atmos. Oceans,* **17,** 89–133.

Holland, W. R., 1989: Altimeter data assimilation into ocean circulation models—Some preliminary results. *Oceanic Circulation Models: Combining Data and Dynamics,* D. L. T. Anderson and J. Willebrand, Eds., Kluwer Academic, 203–232.

——, and P. Malanotte-Rizzoli, 1989: Assimilation of altimeter data into an ocean model with adjustment to the deep circulation. *J. Phys. Oceanogr.,* **19,** 1507–1534.

Hurlburt, H. E., 1986: Dynamic transfer of simulated altimeter data into subsurface information by a numerical model. *J. Geophys. Res.,* **91,** 2371–2400.

——, D. N. Fox, and E. J. Metzger, 1990: Statistical inference of weakly correlated subthermocline fields from satellite altimeter data. *J. Geophys. Res.*, **95,** 11 375–11 409.

Le Dimet, F. X., and O. Talagrand, 1986: Variational algorithms for analysis and assimilation of meteorological observations: Theoretical aspects. *Tellus,* **38A,** 97–110.

Lewis, J. M., and J. C. Derber, 1985: The use of adjoint equations to solve a variational adjustment problem with convective constraints. *Tellus,* **37A,** 309–322.

Lorenc, A. C., 1988: Optimal nonlinear objective analysis. *Quart. J. Roy. Meteor. Soc.*, **114,** 205–240.

McCreary, J. P., 1983: A model of tropical ocean–atmosphere interaction. *Mon. Wea. Rev.*, **111,** 370–387.

——, and D. L. T. Anderson, 1984: A simple model of El Niño and the Southern Oscillation. *Mon. Wea. Rev.*, **112,** 934–946.

Mellor, G. L., and T. Ezer, 1991: A Gulf Stream model and an altimetry assimilation scheme. *J. Geophys. Res.*, **96,** 8779–8795.

Mesinger, F., and A. Arakawa, 1976: Numerical methods used in atmospheric models. Vol. 1. GARP Publ. 17, 64 pp. [Available from WMO, Case Postale No. 5, CH-1211 Geneva 20, Switzerland.]

Moore, A. M., 1986: Data assimilation in ocean models. Ph.D. thesis, University of Oxford, U.K., 168 pp.

——, 1990: Linear equatorial wave mode initialization in a model of the tropical Pacific Ocean: An initialization scheme for tropical ocean models. *J. Phys. Oceanogr.,* **20,** 423–445.

Oschlies, A., and J. Willebrand, 1996: Assimilation of Geosat satellite data into an eddy-resolving primitive equation model of the North Atlantic Ocean. *J. Geophys. Res.,* **101,** 14 175–14 190.

Sasaki, Y. K., 1970: Some basic formalisms in numerical variational analysis. *Mon. Wea. Rev.*, **98,** 875–883.

Sheinbaum, J., and D. L. T. Anderson, 1990a: Variational assimilation of XBT data. Part I. *J. Phys. Oceanogr.*, **20,** 672–688.

——, and ——, 1990b: Variational assimilation of XBT data. Part II. *J. Phys. Oceanogr.,* **20,** 689–704.

Stricherz, J. N., J. J. O'Brien, and D. M. Legler, 1992: Atlas of Florida State University tropical Pacific winds for TOGA: 1966–1985. Mesoscale Air-Sea Interaction Group Tech. Rep. 92-073698, 275 pp. [Available from MS B-174 MASIG, The Florida State University, Tallahassee, FL 32306.]

Talagrand, O., 1993: Data assimilation problems. *Energy and Water Cycles in the Climate System,* E. Raschke and D. Jacob, Eds., NATO ASI Series, Springer-Verlag, 187–213.

——, and P. Courtier, 1987: Variational assimilation of meteorological observations with the adjoint vorticity equation, I: Theory. *Quart. J. Roy. Meteor. Soc.,* **113,** 1311–1328.

Tarantola, A., 1987: *Inverse Problem Theory: Methods for Data Fitting and Model Parameter Estimation.* Elsevier, 613 pp.

Thacker, W. C., 1989: The role of the Hessian matrix in fitting models to measurements. *J. Geophys. Res.*, **94,** 6177–6196.

——, and R. B. Long, 1988: Fitting dynamics to data. *J. Geophys. Res.,* **93,** 1227–1240.

Thépaut, J. N., and P. Courtier, 1991: Four-dimensional variational data assimilation using the adjoint of a multilevel primitive equation model. *Quart. J. Roy. Meteor. Soc.*, **117,** 1225–1254.

Verron, J., 1992: Nudging satellite altimeter data into quasi-geostrophic ocean models. *J. Geophys. Res.*, **97,** 7479–7491.

Weaver, A. T., 1994: Variational data assimilation in numerical models of the ocean. Ph.D. dissertation, University of Oxford, 244 pp.

Webb, D. J., and A. M. Moore, 1986: Assimilation of altimeter data into ocean models. *J. Phys. Oceanogr.*, **16,** 1901–1913.

Wunsch, C., and A. E. Gill, 1976: Observations of equatorially trapped waves in Pacific sea level variations. *Deep-Sea Res.,* **23,** 371–390.