# The Maintenance of Conservative Physical Laws within Data Assimilation Systems*

G. A. JACOBS

*Naval Research Laboratory, Stennis Space Center, Mississippi*

H. E. NGODOCK

*Department of Marine Sciences, University of Southern Mississippi, Stennis Space Center, Mississippi*

(Manuscript received 9 July 2002, in final form 28 March 2003)

ABSTRACT

In many data assimilation applications, adding an error to represent forcing to certain dynamical equations may be physically unrealistic. Four-dimensional variational methods assume either an error in the dynamical equations of motion (weak constraint) or no error (strong constraint). The weak-constraint methodology proposes the errors to represent uncertainties in either forcing of the dynamical equations or parameterizations of dynamics. Dynamical equations that represent conservation of quantities (mass, entropy, momentum, etc.) may be cast in an analytical or control volume flux form containing minimal errors. The largest errors arise in determining the fluxes through control volume surfaces. Application of forcing errors to conservation formulas produces nonphysical results (generation or destruction of mass or other properties), whereas application of corrections to the fluxes that contribute to the conservation formulas maintains the physically realistic conservation property while providing an ability to account for uncertainties in flux parameterizations. The results suggest that advanced assimilation systems must not be liberal in applying errors to conservative equations. Rather systems must carefully consider the points at which the errors exist and account for them correctly. Though careful accounting of error sources is certainly not an entirely new idea, this paper provides a focused examination of the problem and examines one possible solution within the 4D variational framework.

## 1. Introduction

The purpose of data assimilation systems is to make an estimate of the state of the world given the knowledge at hand and expectations of errors in that knowledge (Talagrand 1997). Knowledge at hand includes the many observations returned in near real time in addition to prior research that has developed dynamical equations built on the fundamental concept of conservation of quantities and representation of these equations within a numerical computer framework. It is usual to begin the optimization problem by minimizing a cost function, which expresses the total weighted sum of errors to available knowledge. Let the dynamical equations be given as

$$\mathbf{A}_D\mathbf{x}_{\rm bg} = \mathbf{b}_D, \tag{1}$$

where $\mathbf{A}_D$ is the dynamical operator, and $\mathbf{b}_D$ is a forcing to the dynamical equations. Similarly, the boundary conditions may be written as

$$\mathbf{A}_B\mathbf{x}_{\rm bg} = \mathbf{b}_B, \tag{2}$$

initial conditions may be written as

$$\mathbf{A}_I\mathbf{x}_{\rm bg} = \mathbf{b}_I. \tag{3}$$

Let

$$\mathbf{A}_N = \begin{bmatrix}\mathbf{A}_D\\ \mathbf{A}_B\\ \mathbf{A}_I\end{bmatrix} \quad \text{and} \quad \mathbf{b} = \begin{bmatrix}\mathbf{b}_D\\ \mathbf{b}_B\\ \mathbf{b}_I\end{bmatrix}, \tag{4}$$

Assume that the matrix $\mathbf{A}_N$ is invertible (i.e., the forward problem is well posed and has a unique solution) and let $\mathbf{x}_{\rm bg}$ be the solution to the forward problem

$$\mathbf{A}_N\mathbf{x}_{\rm bg} = \mathbf{b}. \tag{5}$$

A correction $\mathbf{x}$ to the background solution $\mathbf{x}_{\rm bg}$ is constructed so that it minimizes the expected error to dynamical equations as well as measurements. First, let the matrix $\mathbf{A}$ be the tangent linearization of the nonlinear dynamical operator $\mathbf{A}_N$ about the background solution $\mathbf{x}_{\rm bg}$. The cost function to minimize is then written as

* Naval Research Laboratory Contribution Number JA/7323-99-0030.

*Corresponding author address:* Dr. G. A. Jacobs, NRL Code 7323, Stennis Space Center, MS 39529.
E-mail: jacobs@nrlssc.navy.mil

$$J(\mathbf{x}) = (\mathbf{A}\mathbf{x})^{\mathrm{T}}\mathbf{W}(\mathbf{A}\mathbf{x}) + [\mathbf{A}_M(\mathbf{x} + \mathbf{x}_{\mathrm{bg}}) - \mathbf{b}_M]^{\mathrm{T}} \times \mathbf{W}_M[\mathbf{A}_M(\mathbf{x} + \mathbf{x}_{\mathrm{bg}}) - \mathbf{b}_M] \quad (6)$$

with

$$\mathbf{W} = \begin{bmatrix} \mathbf{W}_D & 0 & 0 \\ 0 & \mathbf{W}_B & 0 \\ 0 & 0 & \mathbf{W}_I \end{bmatrix}, \quad (7)$$

where the weights $\mathbf{W}_D$, $\mathbf{W}_B$, $\mathbf{W}_I$, and $\mathbf{W}_M$ are the inverses of the cross covariances of the dynamical equations, boundary conditions, initial conditions, and measurements. The error cross covariances between these quantities are assumed to be zero. The operator $\mathbf{A}_M$ provides measurements of the solution [which is $(\mathbf{x} + \mathbf{x}_{\mathrm{bg}})$], and the observations are included in $\mathbf{b}_M$.

The minimizing solution of the cost function may be constructed through a variety of techniques. As a concrete example, consider the oceans. An excellent overview of many of the methods applied to ocean environment estimation has been provided by Robinson et al. (1998). The most general of these is the four-dimensional variational data assimilation (4DVAR) solution. In the derivation of 4D variational techniques, a set of control variables is proposed. These usually include initial conditions and boundary conditions. In weak-constraint inverse problems (Sasaki 1970), an additional forcing error is proposed as a control variable for each dynamical equation. The unknown forcing (or correction or residual) varies in time and space, and the assimilation system computes an optimal estimate of the unknown forcing. The nonconservation of properties occurs when this forcing error is applied to conservative dynamical equations. In these cases, quantities such as mass, momentum, heat, or salt may be created or destroyed.

Sequential techniques such as the Kalman filter or its many variants such as the ensemble Kalman filter or suboptimal methods such as data insertion or nudging may be derived as special cases of the general 4DVAR solution by applying certain assumptions to the error covariances in the cost function. All these methods have been applied to the ocean estimation problem, and the particular aspect considered within this examination is the conservation of properties on which the original dynamical equations are founded. For example, conservation of momentum and mass lead to analytic equations that describe the relationship between velocity and sea level height in the ocean. Builders of numerical ocean models often go to great lengths to ensure that quantities are conserved, and this has led to flux-conservative numerical models that maintain the conservation relation even within the numerical representation of the analytic equations. If the ocean model is initialized with a given mass, and there are no fluxes into or out of the model, the model maintains the mass through time to the precision of numerical round-off errors.

It is often the case that the assimilation problem is set up in such a manner that it does not maintain conservation. For example, within the tide estimation work of Kantha (1995), a nudging technique is used in which the assimilation innovation at the analysis time is proportional to the difference between the observed sea level and the model forecast state. If the observed sea level is higher than the model forecast then the analysis model sea level is increased. This implies a creation of mass as the sea level is raised. The data insertion technique employed by Smedstad et al. (1997) ensures that at the analysis time, the sea level integrated globally does not change. Given one sea level measurement higher than the model sea level at the same point, the model sea level at the point is raised and simultaneously the sea level throughout the globe is lowered. Thus, while no mass is created or destroyed globally, mass is created and destroyed locally. Another way to view this is that mass is transferred throughout the globe instantaneously in violation of conservation of momentum. The 4DVAR implementation used by Ngodock et al. (2000) to examine the equatorial Pacific Ocean includes a correction to the sea level. Since the sea level in the model at each time step is determined through conservation of mass, any correction to the equation results in the nonconservation of properties.

There are many examples of optimal estimation experiments within which the authors have worked to ensure conservation. The ensemble Kalman filter scheme employed by Evensen and van Leeuwen (1996) constructs a set of ocean states each of which is determined through the application of a randomly perturbed wind field. Because the wind forcing is external to the conservation equations, the numerical model conserves all the internal properties. The tide inversion work of Egbert and Ray (2001) enforces mass conservation and provides evidence that such a strong constraint results in a more realistic solution. The work assumes that all the errors occur in the momentum conservation equations. Thus, while mass is maintained, momentum is not.

Note that these assimilation problems are often quite different from the usual atmospheric prediction problem in which the initial state contains the major errors and the dynamics are assumed to be exact strong constraints. In the strong-constraint case, the assimilation system generates corrections only to the initial state, and the dynamics conserve all properties.

The examination presented here uses a 4DVAR approach since the many methods for assimilation may be derived as special cases from this. In particular, the adjoint solution to the cost function provides a flexible system to examine alterations in the dynamics as well as weak and strong constraints within the system. In order to clearly illuminate the problem, a very simple problem is set up (section 2). This problem is rather unrealistic and contrived, but it does allow a clear demonstration of the nonconservative properties that occur when applying weak constraints to conservative equa-

tions (section 3). One possibility to solve the conservation problem has been suggested by A. Bennett. This approach prescribes the conservation equations as strong constraints, but the fluxes used within the conservation equations are prescribed as weak constraints (section 4). A proof is then provided to demonstrate that a weak-constraint conservation equation cannot conserve properties (section 5). The examination of the simple problem leads to the discussion of the momentum equation and errors within its terms (section 6).

## 2. A simple dynamical system

For the sake of clarity, this discussion makes several simplifying assumptions. These assumptions could be relaxed without altering the fundamental point. Assume that the ocean density is constant and uniform, variations occur only along one coordinate axis, advection and diffusion of momentum are negligible, and there is no external input of momentum from surface or bottom stress. The equations describing the simplified 1D barotropic ocean circulation are

$$\frac{\partial \eta}{\partial t} = -\frac{\partial Hu}{\partial x} \tag{8}$$

$$\frac{\partial u}{\partial t} = -g\frac{\partial \eta}{\partial x}, \tag{9}$$

where $\eta$ is the sea level deviation from its rest state, $u$ is the vertically averaged ocean velocity, $H$ is the spatially varying ocean depth (the depth if the ocean were at rest), and $g$ is the gravitational acceleration constant. Equation (8) is a result of the conservation of mass. The equation is linearized by ignoring the contribution of $\eta$ to the total depth so that only $H$ appears in the equation instead of $(H + \eta)$. Equation (9) represents conservation of momentum. This equation ignores the effects of nonlinear advection, horizontal diffusion, bottom friction, and momentum input by wind stress through the ocean surface. In certain situations, the effects not included in these equations are small. Let conditions be such that the following assumptions are met. Assume Eqs. (8) and (9) accurately represent the conservation of mass and momentum and that a numerical representation is made that accurately models the differential equations (a flux-conservative formulation). Assume the major source of uncertainty in the dynamical equations is the ocean rest depth $H(x)$. This is often the case in shallow ocean areas where bathymetry is poorly known or sediments are resuspended and deposited causing a temporally changing ocean depth. Certainly additional terms could be included in both these equations to more accurately represent the conservation of quantities. However, including these extra terms at the moment would cloud the main issue, which is the effects of taking the conservative equations as weak constraints.

The situation at hand implies that the inverse solution treat momentum [Eq. (9)] as a strong constraint while treating continuity [Eq. (8)] as a weak constraint since the main uncertainty lies in the bathymetry. For the sake of simplicity, assume that the initial and boundary conditions are known exactly. Thus the unknowns (control variables) are the corrections to the continuity equation, $C(x, t)$, due to errors in the ocean rest depth.

This example may appear to be slightly contrived, and it is admittedly so. It is arguable that the depth $H(x)$ is itself an unknown parameter for which the assimilation system must solve. If the system were to estimate $H(x)$, the problem would be rendered nonlinear. This nonlinear inverse problem can be solved by optimal control methods (parameter estimation) using descent algorithms and can easily be likened to the strong-constraint approach. However, estimation of $H(x)$ is not the objective. The objective is to determine the sea level and velocity that best satisfy the equations of motion given that the ocean depth is uncertain. The goal here is not to argue that (8) and (9) accurately represent a particular physically realizable situation. Rather, Eq. (8) serves only as an example of a conservation equation for a quantity (mass in this case) in which the flux through a control volume may be uncertain [in this particular case due to the uncertainty in $H(x)$].

Though the discussion centers on this one example, additional examples are immediate. Consider the vertical diffusion equation in certain numerical ocean models in which the vertical flux of a quantity $S$ is provided by

$$\text{flux}_S = K\frac{\partial S}{\partial z}. \tag{10}$$

The uncertainty in the vertical flux is due to uncertainty in its parameterization in terms of the vertical diffusivity $K$ and the representation of the flux as being proportional to the vertical gradient. At the ocean surface or bottom, there may exist external inputs that describe the flux, and these may contain large errors as well. The equation describing the conservation of the quantity $S$ is

$$\frac{\partial S}{\partial t} = \frac{\partial \text{flux}_S}{\partial z}. \tag{11}$$

The traditional method for weak-constraint assimilation is to add an unknown forcing to the right-hand side of (11), and the assimilation process would provide an optimal estimate of the unknown forcing. However, the estimated forcing applied to (11) appears as a source or sink of the quantity. If the quantity $S$ were the salt content in the ocean, the resulting optimal solution for the forcing in (11) would be applying salt sources and sinks throughout the water column. The optimal solution seemingly violates physical laws because of the manner in which the problem is set up. The uncertainties in the conservation equation (11) are so small that (11) may be regarded as a strong constraint. The uncertainties lie within the flux parameterization given by (10). The issue

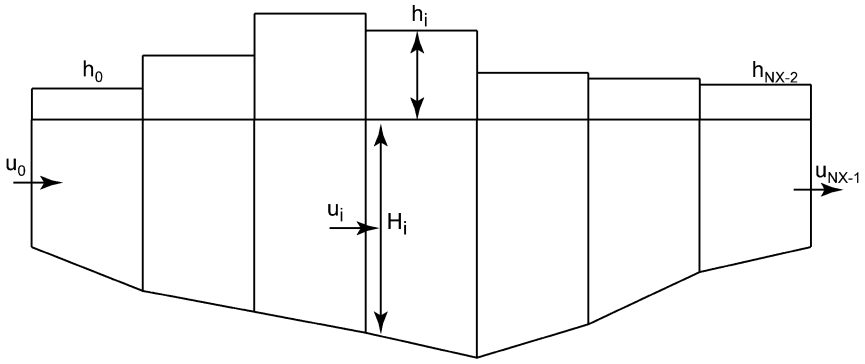


FIG. 1. A one-dimensional C grid is used to represent the dynamical equations. The sea level height $\eta$ exists at points centered between velocity points $u$. The problem is well posed by specification of the initial conditions for all values of $\eta$ and $u$ along with boundary conditions for $u$.

comes to properly determining the point at which lies the expected error in knowledge of physical dynamics.

The example provided by Eqs. (8) and (9) is chosen to examine how an assimilation system may create an egregious violation of a physical law, which is conservation of mass. The flux errors within (8) are due to uncertainties in $H(x)$. The violation of conservation of other quantities is just as egregious, but the creation or destruction of matter raises a more immediate outcry to the violation of basic principles.

The representer approach is used to examine the impact of changing the equation to which the weak constraint is applied. As demonstrated by Bennett (1992) the optimal solution $\hat{\mathbf{x}}$ of the weak-constraint problem [Eq. (6)] is provided by a linear combination of representer functions added to the background. A simplified discussion of the representer solution is presented in the appendix. The optimal solution may be written as

$$\mathbf{x} = \sum_i \beta_i r_i + \mathbf{x}_{\text{bg}}, \tag{12}$$

where $\beta_i$ are the weights of the representer functions $\mathbf{r}_i$, and $\mathbf{x}_{\text{bg}}$ is the background estimate [the solution to Eq. (5)]. A representer function is constructed for each measurement by

$$r_i = \mathbf{A}\mathbf{W}^{-1}\mathbf{A}^*\mathbf{A}_{Mi}, \tag{13}$$

where $\mathbf{A}$ is the linearized dynamical operator that provides the model state for all space and time including the appropriate initial and boundary conditions, $\mathbf{W}^{-1}$ is the matrix providing the dynamical equation, boundary, and initial condition error covariance estimates, $\mathbf{A}^*$ is the adjoint of the model operator, and $\mathbf{A}_{Mi}$ is a measurement function that provides the $i$th measurement of the state. The examples here are based on finite-difference representations of the dynamical equations. In this case the dynamical operator $\mathbf{A}$ is a matrix and the adjoint operator is the transpose of the matrix. Equation (13) is solved for the representer function by first solving the adjoint equations forced by the measurement functional, next performing the covariance multiplication, and then using the result to force the forward dynamics. Assume that the background field $\mathbf{x}_{\text{bg}}$ satisfies conservation properties. This would be expected as the background field is usually the result of a model forecast without any corrections applied to the dynamical equations. Because the optimal solution is a linear combination of representers, the conservation (or nonconservation) properties of the representer functions will be passed on to the final inverse solution. For example, if there were one measurement of height and a mass-conserving background field estimate with a representer function that is not mass conserving, then the conservation of mass in the optimal solution estimate will be destroyed. Thus, the representer functions for a given measurement may be examined to demonstrate the conservative properties of the optimal solution.

## 3. Nonconservative assimilation

A finite-difference scheme is introduced to discretely model the analytic equations (8) and (9) with a correction estimate to the mass conservation equation. The numerical grid (Fig. 1) is similar to an Arakawa C grid (Mesinger and Arakawa 1976):

$$\frac{\eta_{i,n+1} - \eta_{i,n-1}}{2dt} = -\frac{H_{i+1}u_{i+1,n} - H_i u_{i,n}}{dx} + C_{i,n}, \tag{14}$$

$$\frac{u_{i,n+1} - u_{i,n-1}}{2dt} = -g\frac{\eta_{i,n} - \eta_{i-1,n}}{dx}. \tag{15}$$

The first and second subscripts provide the spatial and temporal indices, respectively. A leapfrog forward stepping in time is used for both $u$ and $\eta$. The leapfrog scheme produces an unrealistic or parasitic mode (Kowalik and Murty 1993). An Asselin filter in time provides a discrete time-smoothing operator and suppresses the parasitic mode (Asselin 1972). While more accurate methods are available (such as Runge–Kutta), a simple formulation provides a clearer example, and the in-

creased computational time for an accurate solution to the simple equations is bearable. The Asselin filter is considered to be a dynamical equation used to propagate the state forward in time, and this particular dynamical equation is taken as a strong constraint when computing the representer functions. For the experiments here, the depth $H_i$ is set to 100 m throughout all space and time. The domain length is 1000 km with a grid spacing of 5 km. The experiment time range is 5.5 h with a time step of 50 s.

The initial conditions applied are

$$u_{i,0} = u\_\text{input}_{i,0} \tag{16}$$

$$\eta_{i,0} = \eta\_\text{input}_{i,0}, \tag{17}$$

and the boundary conditions applied are

$$u_{0,n} = u\_\text{input}_{0,n} \tag{18}$$

$$u_{NX-1,n} = u\_\text{input}_{NX-1,n}. \tag{19}$$

The initial and boundary covariance amplitudes relative to the continuity correction have a large impact on the representer functions. Because the intent here is to examine the conservative properties of the representer function due to the correction to the conservative equations, the boundary and initial condition covariances cloud the issue at hand. Therefore, the initial and boundary conditions are taken to be strong constraints in the problems considered here. This removes questions concerning error covariances associated with the boundary and initial condition inputs. Thus, all equations including boundary and initial conditions are strong constraints except for the conservation equation (14), and the control variables are the values of the function $C_{i,n}$.

Two representer functions are constructed. The measurement position for each representer function is the same: centered in space at 500 km and at the 3-h time point. The representer functions for a velocity measurement (Fig. 2) and a height measurement (Fig. 3) indicate the effect each measurement would have on the optimal solution. Solution of the adjoint equations of (14)–(19) are convoluted by the covariance function $\text{cov}_{i,j,n,m}$ representing the spatial and temporal error covariances. The covariance function used here is

$$\text{cov}_{i,j,n,m} = e^{-(x_i-x_j)^2/L^2-(t_n-t_m)^2/T^2} \tag{20}$$

with $L$ taken to be 100 km and $T$ taken to be 1 h. This covariance function assumes that errors have no bias (zero mean) and are Gaussian distributed with a variance of one. The same covariance function is applied to all residuals within this examination.

The representer function sea level and velocity integrated over space demonstrate the conservation of properties within the representers. For the velocity measurement (Fig. 2) integrated volume does not change in time. However, the integrated velocity (proportional to integrated momentum) does indicate an increase over time. There is no requirement that the integrated velocity be maintained in the representer, but rather that the balance between the sea level and velocity field be maintained since (15) is taken as a strong constraint.

For the height measurement (Fig. 3), total volume increases around the measurement time period and levels off by the end of the 5-h representer function. The reason for the increasing volume (and thus generated mass) is that the correction added to the continuity equation $C_{i,n}$ in (14) is positive throughout all space and time. The spatially integrated velocity for this representer function remains constant throughout time. The correction to the background provided by this representer function depends on the dynamical error covariance in Eq. (20), the measurement error covariance, and the actual difference between the background and the measurement. Thus the color bar range in the figures does not indicate the actual correction to the background field. However, the nonconservation properties will be included in the solution.

## 4. Conservative assimilation

The finite difference equations are slightly altered to explicitly compute the mass flux. In addition, the continuity equation is taken as a strong constraint by removing the error $C_{i,n}$. The control variables are now the mass flux corrections. Thus the mass flux is made a weak constraint by including the error estimate $F_{i,n}$. The mass flux error estimate in this case accounts for the uncertainty in mass flux due to uncertainty in the ocean depth:

$$f_{i,n} = H_i u_{i,n} + F_{i,n} \tag{21}$$

$$\frac{\eta_{i,n+1} - \eta_{i,n-1}}{2dt} = -\frac{f_{i+1,n} - f_{i,n}}{dx} \tag{22}$$

$$\frac{u_{i,n+1} - u_{i,n-1}}{2dt} = -g\frac{\eta_{i,n} - \eta_{i-1,n}}{dx}. \tag{23}$$

The equations for initial and boundary equations remain the same as Eqs. (14)–(15). The conservative representer functions for the velocity and height measurements (Figs. 4 and 5) indicate similar spatial patterns as in the nonconservative case with the influence of the measurements propagating from the measurement point both back and forward in time with phase speed of $\sqrt{gH}$. The spatial scales of the conserved representer functions are slightly smaller than the nonconserved functions because the same spatial and temporal length scales are used for all covariances of the correction fields and the flux correction appears as a derivative in the continuity equation. Both total mass and velocity are conserved in both the velocity and height measurement representers.

The amplitudes of the representer functions in Figs. 2 and 3 are quite different from the amplitudes of the representer functions in Figs. 4 and 5. This is because the adjoint variables for the flux and for the sea level

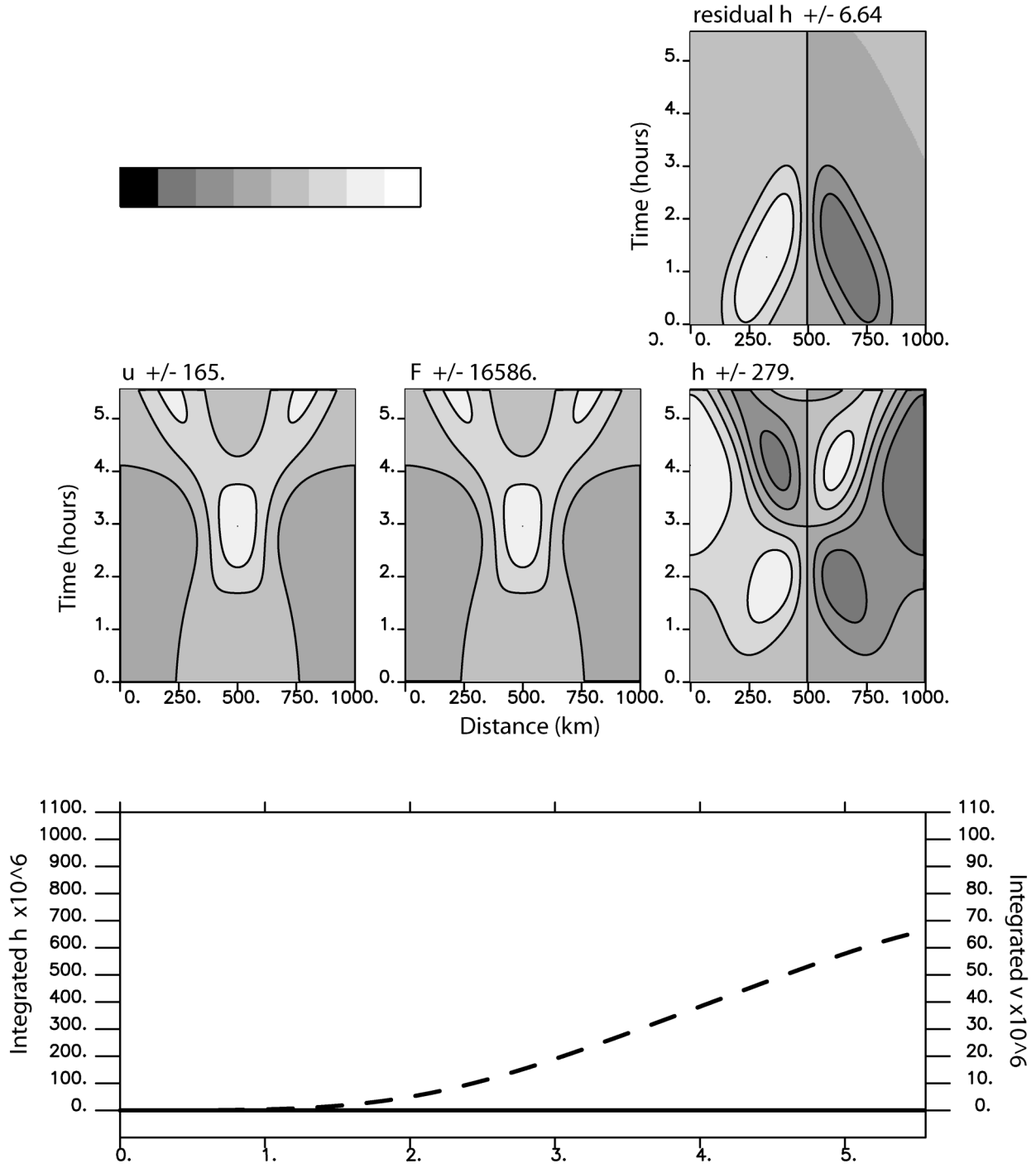


FIG. 2. The representer function for a weak constraint on continuity and a measurement of velocity indicates (top right) the correction applied to continuity and how the measurement will affect the solution of (middle left) velocity, (middle center) mass flux, and (middle right) sea level. (bottom) The spatially integrated sea level as a function of time indicates that a velocity measurement conserves mass (solid line) but does not conserve momentum (dashed line). The numbers above each shaded plot indicate the shade bar range.

have different magnitudes. In the conservative case, a measurement functional forces the adjoint dynamics by directly adding to the sea level residual. The sea level residual then contributes to the flux residual by a factor of $1/dx$ (which is taken to be 5000 m in these examples). Thus the same measurement will contribute differently to the different residuals. The covariance functions applied should account for this fact. Because the sea level

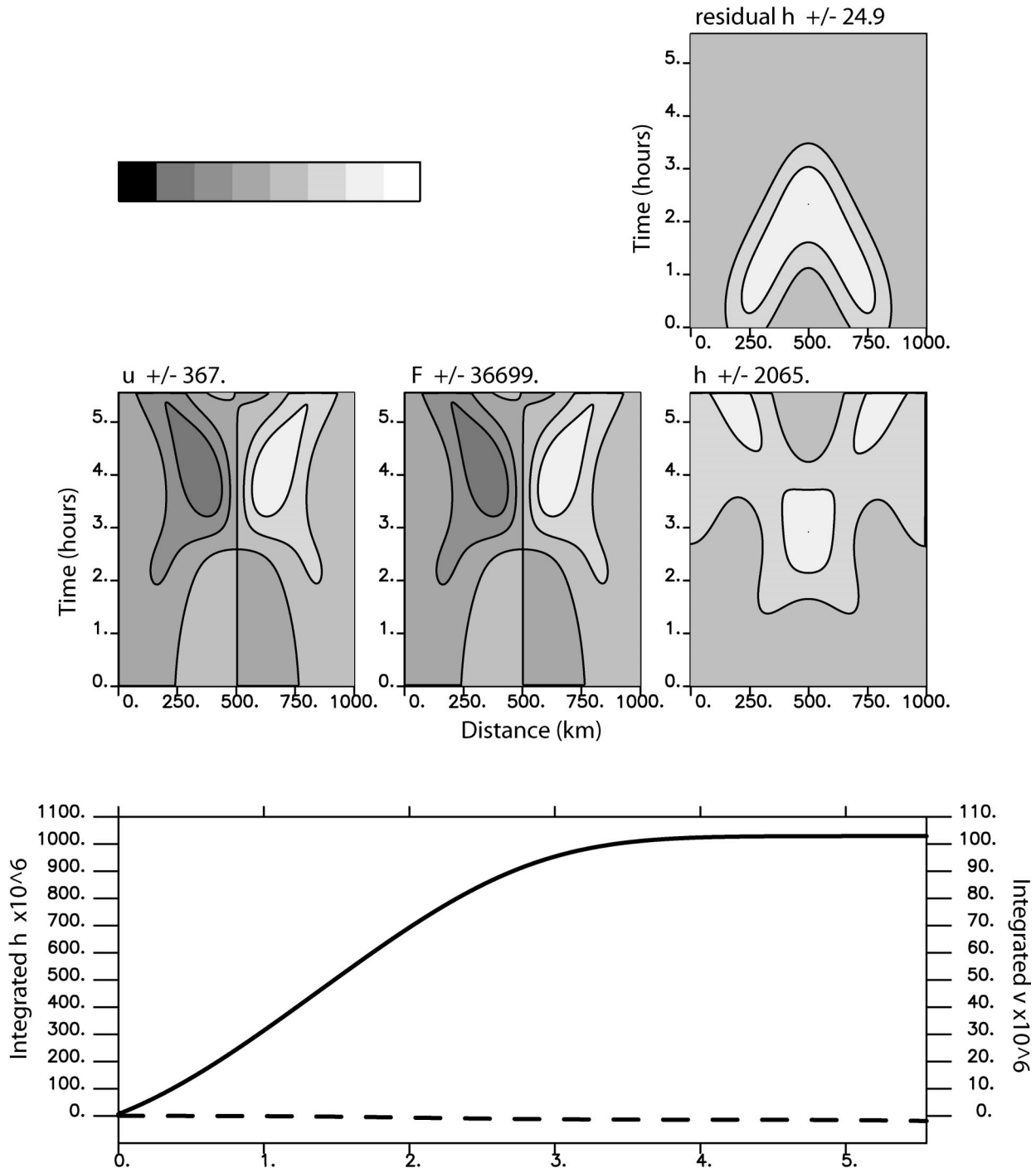


FIG. 3. The same as in Fig. 2 except for a measurement of sea level. (bottom) The spatially integrated sea level as a function of time indicates that a velocity measurement does not conserves mass (solid line) but does conserve momentum (dashed line). Shaded areas indicated as in Fig. 2.

and flux are quite different variables, it should be expected that they should have quite different error covariance amplitudes and structures. The examples here use the same error covariance for each.

## 5. Nonconserving solutions cannot conserve

The traditional approach to handle model error in data assimilation (both in variational and sequential) is to

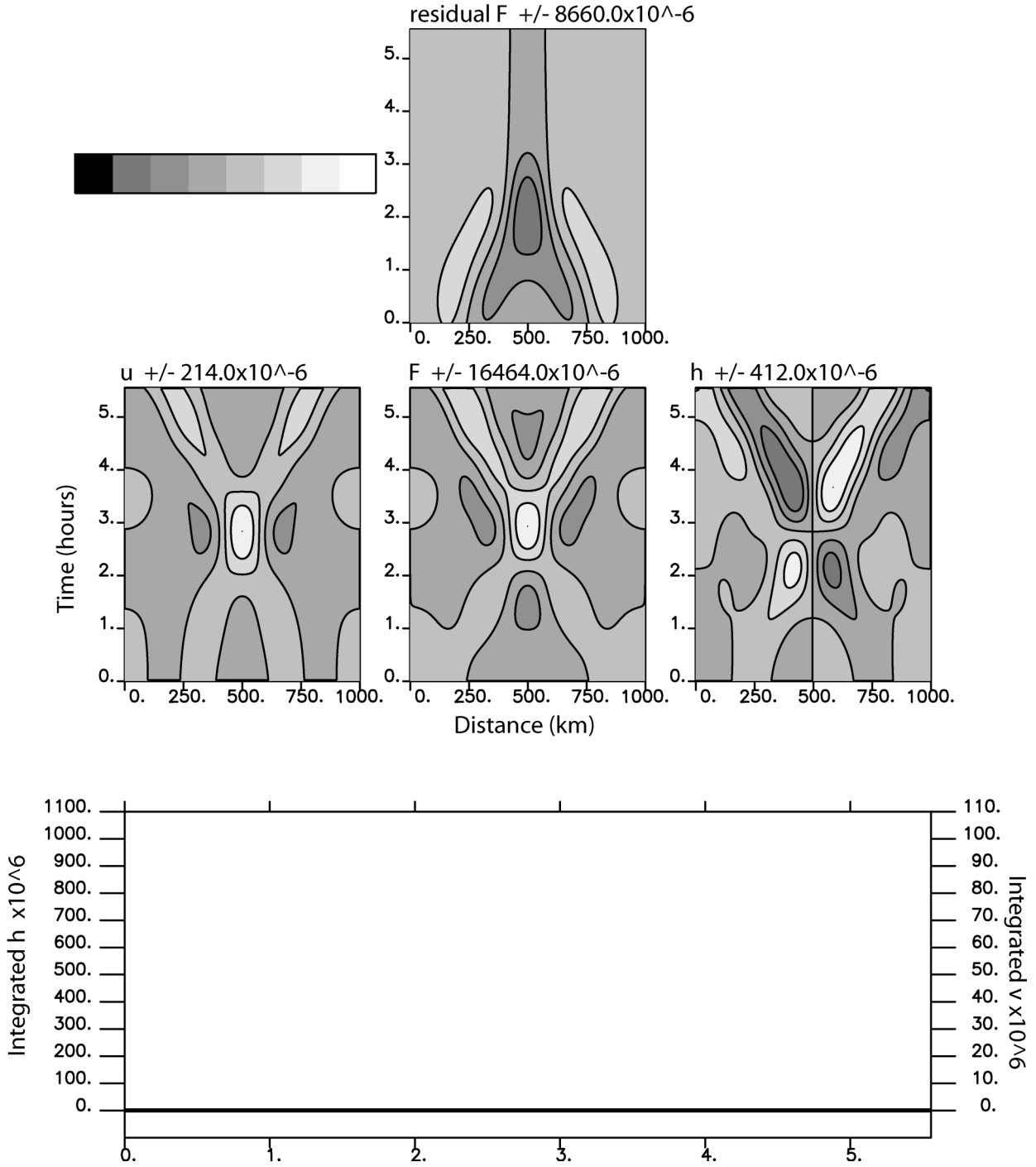


FIG. 4. The representer function for a weak constraint on mass flux and a measurement of velocity indicates (top center) the correction applied to mass flux and how the measurement will affect the solution of (middle left) velocity, (middle center) mass flux, and (middle right) sea level. (bottom) The spatially integrated sea level as a function of time indicates that a velocity measurement conserves mass (solid line) and conserves momentum (dashed line). Shaded areas indicated as in Fig. 2.

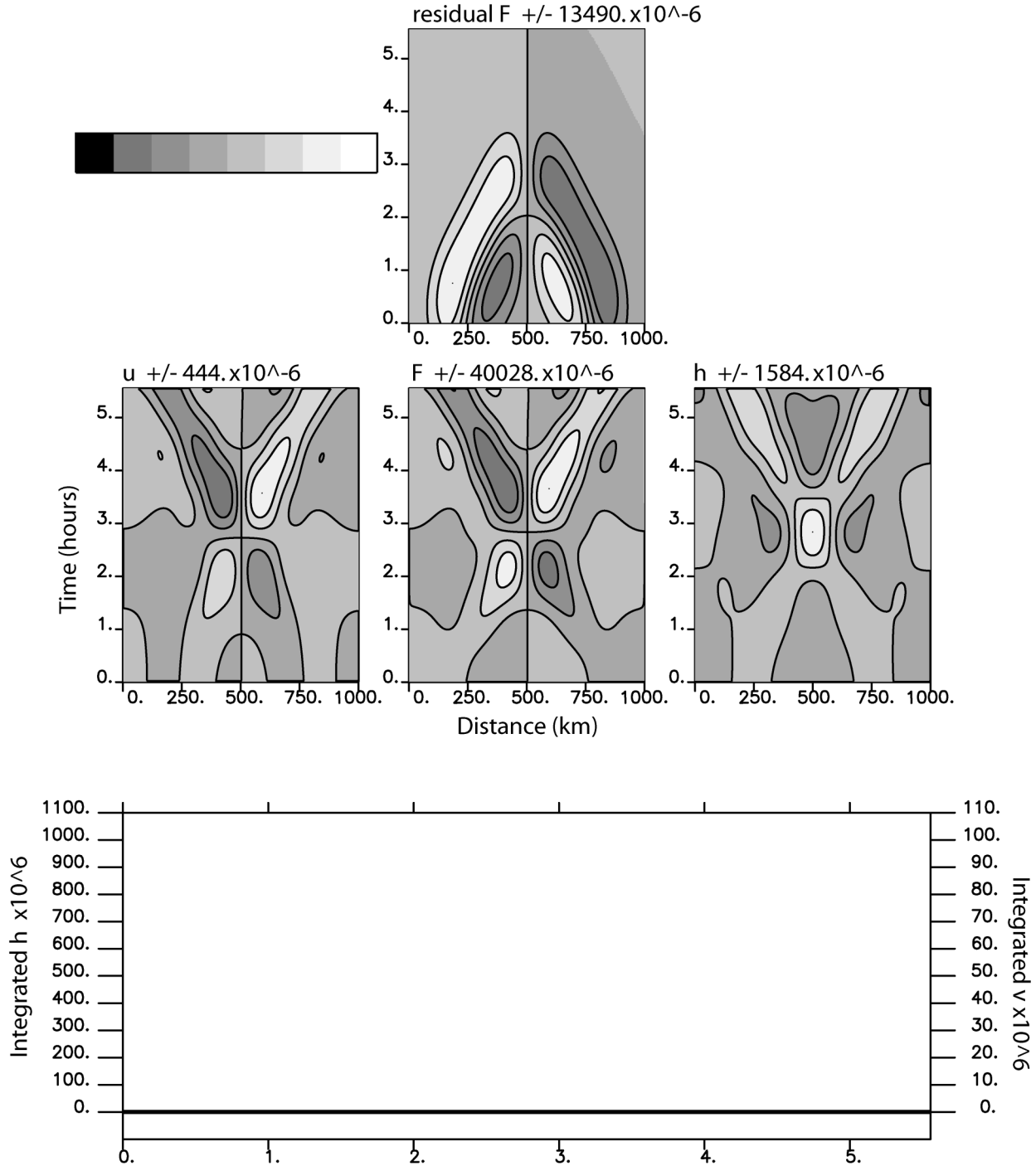


FIG. 5. The same as in Fig. 4 except for a measurement of sea level. Shaded areas indicated as in Fig. 2.

include an error term in the right-hand side of the model equations. The generalized weak-constraint system for the dynamics used here is given by

$$f - Hu = F, \tag{24}$$

$$\frac{\partial \eta}{\partial t} + \frac{\partial f}{\partial x} = C, \tag{25}$$

$$\frac{\partial u}{\partial t} + g\frac{\partial \eta}{\partial x} = M, \tag{26}$$

where unknown forcing functions or residuals are added to each dynamical equation. The residuals are $F$, $C$, and $M$. In variational data assimilation the residuals are computed through integration of the adjoint equations,

which is accomplished by integrating from the final time to the initial time:

$$-\frac{\partial^* C}{\partial t} - g\frac{\partial^* M}{\partial x} = 2\sum_{m=1}^{N_M}(d_m - \eta_m)\delta(t - t_m) \times \delta(x - x_m), \quad (27)$$

$$F - \frac{\partial^* C}{\partial x} = 0, \quad (28)$$

$$-\frac{\partial^* M}{\partial t} - HF = 0, \quad (29)$$

where the adjoint final and boundary conditions have been omitted for the sake of clarity. The adjoint boundary conditions are homogeneous if the boundary conditions are considered strong constraints in the forward model (as in the examples here). The superscript * denotes the adjoint operator (i.e., $-(\partial^*/\partial t)$ is the adjoint of $\partial/\partial t$ in the forward model).

In computing the optimal solution, a residual provides forcing to a forward equation. If an equation is taken as a strong constraint, the corresponding residuals computed by (27)–(28) become Lagrange multipliers, and no forcing is applied to that equation in the forward model. Thus, in both the examples covered in sections 3 and 4, the computation of the residuals is the same. However, different equations are taken as strong constraints within each example. To affect this, all that is required is that the appropriate residual be added to the forward equation within each example.

As demonstrated by the experiments in section 3, treating the continuity equation as a weak constraint does not conserve mass. When the continuity equation is taken as a strong constraint and the mass flux as a weak constraint, the solution of this system conserves mass because no forcing or correction is applied to the forward equation for continuity. In fact, any expression for $F$ would lead to mass conservation because the continuity equation is a strong constraint. It may be demonstrated that both solutions are equal under the condition $C = -(\partial F/\partial x)$ [by substituting Eq. (24) into (25)]. In practice, the residuals $C$ and $F$ are specified by the adjoint equations, and the adjoint equation (28) implies $F = \partial^* C/\partial x$. In order to obtain the same solution from both the approaches considered in sections 3 and 4, it would be required that

$$C = -\frac{\partial}{\partial x}\left(\frac{\partial^* C}{\partial x}\right). \quad (30)$$

This last condition (on the continuous derivative and its continuous adjoint) is satisfied only for a special form of the residual $C$. If the continuous derivative is self-adjoint (with suitable boundary conditions), then $C$ could be a sinusoid function in space, or any function that can be expanded into a Fourier series. This would be rather difficult to achieve (actually impose) on $C$ since we do not have control on the residual.

In practice, the derivatives are represented by discrete operators. We should investigate if the earlier condition on $C$ could be satisfied in the discrete derivative and its adjoint. Let the derivative operator be represented by a matrix with real coefficients $\mathbf{D}$, so that the requirement (30) becomes $C = -\mathbf{DD}^* C$, where the superscript * denotes the (matrix) transposition. This condition should hold for any $C$ since there is no control or constraint imposed on $C$. To satisfy the condition in the earlier numerical model, the discrete operator $\mathbf{D}$ must be such that $\mathbf{DD}^* = -\mathbf{I}$, where $\mathbf{I}$ is the unit $N \times N$ matrix, $N$ being the number of grid points. Let the general term of $\mathbf{D}$ be $d_{ij}$. Then the $ij$th term of $\mathbf{DD}^*$ is

$$\sum_{k=1}^{N} d_{ik} d_{jk} = -\delta_{ij}, \quad (31)$$

where $\delta_{ij}$ is the Kronecker symbol:

$$\delta = \begin{cases} 1 \text{ if } i = j \\ 0 \text{ if } i \neq j \end{cases}.$$

The diagonal terms in Eq. (31) are

$$\sum_{k=1}^{N} d_{ik}^2 = -1, \quad (32)$$

which has no real solution.

There is a second case that must be considered in general, and this involves the problem in which the derivative of flux occurs on the right-hand side of the conservation equation. In such a case the requirement that the two methods in sections 3 and 4 produce the same solution is that $C = \mathbf{DD}^* C$ for any residual $C$, which is translated in discrete notation by $\mathbf{DD}^* = \mathbf{I}$. This implies that the discretization of the continuous first derivative operator should be provided by an orthonormal matrix, which is not the case. Thus, there is no (real) discretization that could provide the solutions from both approaches to be the same. There is clearly one approach that displays conservation properties and one that cannot by construction.

Note that the developments of this section assume that the covariance function for the continuity and fluxes is an identity operator. Generalization to include covariance implies that $C = -\mathbf{DVD}^* C$ must be satisfied, where $\mathbf{V}$ is the covariance function. The covariance $\mathbf{V}$ must be positive definite, and thus it may be split into two matrixes so that the requirement becomes $C = -(\mathbf{DU})(\mathbf{DU})^* C$. Because $(\mathbf{DU})(\mathbf{DU})^*$ is positive definite, it is not possible for $C = -(\mathbf{DU})(\mathbf{DU})^* C$.

## 6. Discussion

It may be argued that in the presence of many data, a system could use a combination of some conserving as well as some nonconserving representer functions. On one hand, the weighted sum of conserving representer functions will be conserving. On the other hand,

the weighted sum of two (or more) nonconserving representer functions might possibly be conserving. This would be the case if there were two (or more) measurements of, say, sea level at the same location in space and time, whose average equals the first guess (background) field at the particular location, and the sum of the representer weights would be zero. Essentially the representer functions of the measurements would cancel one another. In this case the measurements are bringing no new information about the ocean, and could be discarded in the assimilation process.

The weights used in the cost function (6) describe not only the expected errors of each dynamical equation or state value but also the cross covariances. In the example provided it is assumed that there is no cross-correlated error between the equation for sea level (continuity) and velocity (momentum). It would be possible to include the cross correlation so that when a correction to sea level is made, it would also produce an according change in velocity so that it would induce the mass flux needed to change the sea level. Specifying such covariances for this simple system would be possible. However, for a larger system containing many more variables, it is not as easily determined. The application of independent corrections to each state variable is a result of our inability to accurately determine the true error covariance structure. By simplifying each dynamical equation into flux components and conservation components, the cross-covariance error determination is simplified.

It may also be argued that there should be no strong constraints within an optimal solution. The weights applied to the dynamical equations, boundary conditions, and initial conditions should reflect the expected error levels. The strong-constraint assumption for a particular equation reflects our expectation that the equation error levels are much less than the error levels of other equations. The strong constraint is used as a method to save computer memory space in the optimization process and to allow the solution process to focus on the major sources of errors.

The example provided by the mass flux and momentum equations is purposely simplified to present more clearly the basic idea. The momentum equation is taken to be a strong constraint in section 3 and both the momentum and continuity equations are taken to be strong constraints in section 4. However, it is not intended to advocate strong-constraint assimilation systems. Every dynamical equation has uncertainties within it. The uncertainties may be the result of neglected terms in equations, errors in parameters, truncated tangent linearizations of equations, truncated Taylor series used to construct finite-difference representations of analytical derivatives, or spatial averages that may be required (e.g., to compute a $v$ velocity value at a $u$ velocity point on a C grid). Thus each dynamical equation used in a numerical model contains some level of uncertainty. However, the accuracy in some equations may be relatively so high that the equations could be taken as a strong constraint without significant degradation to the optimal solution.

As an example, the momentum flux equation (for a hydrostatic and Bousinnesque fluid) is

$$\frac{\partial u}{\partial t} + \frac{\partial}{\partial x}(F_x) + \frac{\partial}{\partial y}(F_y) + \frac{\partial}{\partial z}(F_z) - fv + \frac{1}{\rho_o}\frac{\partial p}{\partial x} = 0 \tag{33}$$

with pressure determined by

$$p(x, y, z, t) = \int_z^0 g\rho(z')\, dz' + g\rho_o\eta(x, y, t) + p_{\text{atm}}(x, y, t) \tag{34}$$

and with the horizontal fluxes provided by advection and tangential stresses given by

$$F_x = uu - \frac{1}{\rho_o}K_h\frac{\partial u}{\partial x} \tag{35}$$

$$F_y = uv - \frac{1}{\rho_o}K_h\frac{\partial u}{\partial y} \tag{36}$$

$$F_z = uw - \frac{1}{\rho_o}K_v\frac{\partial u}{\partial z}. \tag{37}$$

Note that different diffusivities are provided in the horizontal ($K_h$) and vertical ($K_v$) directions.

In the momentum conservation equation (33), the largest errors are due to the numerical representation of the equation. On a finite-difference $C$ grid, values of the $v$ velocity exist at grid points different from grid points of $u$ velocity. The model must spatially average values of $v$ to compute the Coriolis force at a $u$ point, and this induces errors. The pressure gradient is computed by vertical integration of the density in combination with the sea level (34). Sigma coordinate models are known to contain large errors in this term due to the manner in which Eq. (34) is represented numerically (Stellings and Vankester 1994). Discretization errors are reduced as the numerical model resolution is increased. If a numerical model is able to properly resolve the dynamical processes, the discretization errors in (33) are minimal and the equation may be regarded as a strong constraint.

However, the horizontal fluxes described by Eqs. (35)–(37) contain substantial errors. In the construction of a variational data assimilation system, the dynamical equations are usually linearized, a tangent linear model constructed, and the adjoint of the tangent linear model used to determine the optimal values of control variables. For example, the gradient of the advection of $u$ momentum due to $u$ velocity becomes

$$\frac{\partial}{\partial x}(uu) = \frac{\partial}{\partial x}[(u^{\text{bg}} + u')(u^{\text{bg}} + u')]$$

$$= \frac{\partial}{\partial x}(u^{\text{bg}}u^{\text{bg}} + u'u^{\text{bg}} + u^{\text{bg}}u' + u'u')$$

$$\approx \frac{\partial}{\partial x}(u^{bg}u^{bg} + u'u^{bg} + u^{bg}u'), \quad (38)$$

where $u^{bg}$ is a prior or background field estimate for the optimal solution of $u$, and $u'$ is a perturbation of the background field that the assimilation system will optimally determine. In constructing the tangent linear model, the nonlinear terms involving perturbations from the background field ($u'u'$) are ignored, and this leads to the final approximation of (37). By ignoring the nonlinear perturbation terms, an error is induced into the fluxes. Thus linearization errors from advection should not contribute to the conservation equation (33).

The horizontal and vertical diffusivities contain errors, and the errors are expected to have substantially different covariance amplitudes and length scales. The vertical boundary conditions for the momentum flux equation are the surface wind stress and bottom stress. In the adjoint model, residuals from the momentum equation would be passed to the wind stress field. This is very important since the wind field is expected to contain dramatically different spatial and temporal covariance scales than errors in the internal vertical flux.

Assimilation systems should not assume all conservative equations to be near strong constraints. Conservation equations for different parameters such as chlorophyll may have sources and sinks distributed throughout the ocean interior. Thus a direct error to the conservation equation may be warranted in certain circumstances in addition to possible flux and forcing errors.

While the nonconservative approach may not explicitly compute the fluxes and their adjoints, roughly the same number of operations would be required under the nonconservative and conservative approaches. However, the improved conservative properties do not come without cost. If the conservation equations are treated as strong constraints in the system, then residual values become Lagrange multipliers and still must be computed even though the Lagrange multipliers need not be saved. For the momentum equation, the nonconservative approach would require only one residual to be saved in computer storage (either in memory or on hard disk). Under the conservation approach, three residuals would be required for the $u$ momentum equation [one for each flux term in (35)–(37)]. While computational requirements are only slightly higher, storage requirements are much higher. An advantage of the conservative approach is that determination of covariance errors is slightly simplified since the covariance errors represent smaller groups of terms with fewer derivatives. It is arguably easier to determine the covariance error of the few terms in the flux equation than the covariance error of all the entire group of terms in the momentum equation at once.

Certainly the impact of the conservative versus nonconservative approach must be evaluated within a more realistic assimilation system. Systems are presently being constructed with the flexibility to test both the conservative and nonconservative approaches. However, definitive conclusions as to the influence and importance in operational systems are still some years away.

## 7. Conclusions

Caution should be exercised when determining the accuracy of our knowledge of dynamical equations. Conservation equations and the numerical flux form representation of such equations do not contain large errors. The simplified test presented here has demonstrated that solving for unknown forcing within conservation equations can produce the spurious generation or destruction of properties. To avoid this, applying errors to flux parameterization leads to conservation. The conservation approach may also be considered to alleviate some of the difficulty in specifying the dynamical error covariances. Based on these considerations, it would be worthwhile to implement and test this method in a realistic operational system.

*Acknowledgments.* We would like to thank two anonymous reviewers, whose comments greatly helped to improve this text. Discussions with Andrew Bennett helped to set the background on which this work has been built. This work was sponsored by the Office of Naval Research (program element 601153N) as part of the project ''Error Propagation in the Continental Shelf.''

## APPENDIX

### Representer Solution to the Linear Minimization Problem

This appendix outlines a simplified derivation of the representer solution for a discrete problem. A more general derivation for nonlinear problems may be found in Uboldi and Kamachi (2000). Both these derivations compute the optimal perturbation $\mathbf{x}$ instead of the total solution ($\mathbf{x} + \mathbf{x}_{bg}$) as in Bennett (1992). Given the definition of the dynamical operators and cost function within the introduction,

$$\mathbf{A}_D\mathbf{x}_{bg} = \mathbf{b}_D \quad (A1)$$

$$\mathbf{A}_B\mathbf{x}_{bg} = \mathbf{b}_B, \quad (A2)$$

$$\mathbf{A}_I\mathbf{x}_{bg} = \mathbf{b}_I, \quad (A3)$$

$$\mathbf{A}_N = \begin{bmatrix}\mathbf{A}_B\\ \mathbf{A}_B\\ \mathbf{A}_I\end{bmatrix}, \mathbf{b} = \begin{bmatrix}\mathbf{b}_D\\ \mathbf{b}_B\\ \mathbf{b}_I\end{bmatrix}, \quad (A4)$$

$$J(\mathbf{x}) = (\mathbf{A}\mathbf{x})^T\mathbf{W}(\mathbf{A}\mathbf{x}) + [\mathbf{A}_M(\mathbf{x} + \mathbf{x}_{bg}) - \mathbf{b}_M]^T \times \mathbf{W}_M[\mathbf{A}_M(\mathbf{x} + \mathbf{x}_{bg}) - \mathbf{b}_M]. \quad (A5)$$

Define the residual variables as the weighted residuals to the dynamical, boundary, and measurement errors,

$$\boldsymbol{\lambda}_D = \mathbf{W}_D(\mathbf{A}_D\mathbf{x}) \tag{A6}$$

$$\boldsymbol{\lambda}_B = \mathbf{W}_B(\mathbf{A}_B\mathbf{x}) \tag{A7}$$

$$\boldsymbol{\lambda}_I = \mathbf{W}_I(\mathbf{A}_I\mathbf{x}). \tag{A8}$$

Then the cost function may be written as

$$J(\mathbf{x}) = (\mathbf{A}\mathbf{x})^{\mathrm{T}}\boldsymbol{\lambda} + [\mathbf{A}_M(\mathbf{x} - \mathbf{x}_{\mathrm{bg}}) - \mathbf{b}_M]^{\mathrm{T}} \times \mathbf{W}_M[\mathbf{A}_M(\mathbf{x} - \mathbf{x}_{\mathrm{bg}}) - \mathbf{b}_M], \tag{A9}$$

where

$$\boldsymbol{\lambda} = \begin{bmatrix} \boldsymbol{\lambda}_D \\ \boldsymbol{\lambda}_B \\ \boldsymbol{\lambda}_I \end{bmatrix}. \tag{A10}$$

For the $i$th row of $\mathbf{A}_M$ (the $i$th measurement), solve $\mathbf{A}\boldsymbol{\lambda}_i = \mathbf{A}_{Mi}^{\mathrm{T}}$ for the weighted residuals $\boldsymbol{\lambda}_i$. This is the solution of the adjoint model operator (the transpose of the forward problem) forced by the $i$th measurement functional. The $i$th representer function $\mathbf{r}_i$ is then a solution of

$$\mathbf{A}\mathbf{r}_i = \begin{bmatrix} \mathbf{W}_D & 0 & 0 \\ 0 & \mathbf{W}_B & 0 \\ 0 & 0 & \mathbf{W}_I \end{bmatrix}^{-1} \boldsymbol{\lambda}_i. \tag{A11}$$

Note there is a representer function for each measurement. Assume the correction to the background is a linear combination of the representer functions

$$\mathbf{x} = \mathbf{r}\boldsymbol{\beta}, \tag{A12}$$

where $\boldsymbol{\beta}$ is a column vector of the amplitudes of each representer function, and the matrix $\mathbf{r} = [\mathbf{r}_i]$ has the representers as columns. If this form of solution is placed into the cost function, the values of $\boldsymbol{\beta}$ that provide a minimization are given by

$$(\mathbf{W}_M + \mathbf{A}_M\mathbf{r})\boldsymbol{\beta} = \mathbf{b}_M. \tag{A13}$$

It is also possible to demonstrate that the first variation of the cost function at the solution $\mathbf{x} = \mathbf{r}\boldsymbol{\beta}$ is 0 if the weights $\boldsymbol{\beta}$ are given as in (A13). One important point to note is that the measurement weights $\mathbf{W}_M$ do not alter the representer functions but rather influence the amplitude of each representer contributing to the optimal solution.

Thus, the optimizing solution to the cost function is a linear combination of representer functions. Each representer function reflects the impact of a single measurement. The properties of the total solution are a direct result of the properties of the representer functions composing it. If the representer functions are not conservative, the total solution will not be conservative.

## REFERENCES


Asselin, R., 1972: Frequency filter for time integrations. *Mon. Wea. Rev.,* **100,** 487–490.

Bennett, A. F., 1992: *Inverse Methods in Physical Oceanography*. Cambridge University Press, 346 pp.

Egbert, G. D., and R. D. Ray, 2001: Estimates of M2 tidal energy dissipation from TOPEX/POSEIDON altimeter data. *J. Geophys. Res.,* **106,** 22 475–22 502.

Evensen, G., and P. J. van Leeuwen, 1996: Assimilation of Geosat altimeter data for the Agulhas Current using the ensemble Kalman filter with a quasigeostrophic model. *Mon. Wea. Rev.,* **124,** 85–96.

Kantha, L. H., 1995: Barotropic tides in the global oceans from a nonlinear tidal model assimilating altimetric tides. 1. Model description and results. *J. Geophys. Res.,* **100,** 25 283–25 308.

Kowalik, Z., and T. S. Murty, 1993: *Numerical Modeling of Ocean Dynamics.* World Scientific, 481 pp.

Mesinger, F., and A. Arakawa, 1976: *Numerical Methods Used in Atmospheric Models.* GARP Publication Series, No. 17, World Meteorological Organization, 64 pp.

Ngodock, H. E., B. S. Chua, and A. E. Bennett, 2000: Generalized inverse of a reduced gravity primitive equation ocean model and tropical atmosphere–ocean data. *Mon. Wea. Rev.,* **128,** 1757–1777.

Robinson, A. R., P. F. J. Lermusiaux, and N. Q. Sloan III, 1998: Data assimilation. *The Sea,* K. H. Brink and A. R. Robinson, Eds., The Global Coastal Ocean, Vol. 10, John Wiley and Sons, 541–593.

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

Smedstad, O. M., D. N. Fox, H. E. Hurlburt, G. A. Jacobs, E. J. Metzger, and J. L. Mitchell, 1997: Altimeter data assimilation into a 1/8 degrees eddy resolving model of the Pacific Ocean. *J. Meteor. Soc. Japan,* **75,** 429–444.

Stellings, G. S., and J. A. T. M. Vankester, 1994: On the approximation of horizontal gradients in sigma coordinates for bathymetry with steep bottom slopes. *Int. J. Numer. Methods Fluids,* **18,** 915–935.

Talagrand, O., 1997: Assimilation of observations: An introduction. *J. Meteor. Soc. Japan,* **75,** 191–209.

Uboldi, F., and M. Kamachi, 2000: Time–space weak-constraint data assimilation for nonlinear models. *Tellus,* **52A,** 412–421.