

## ChapterF

# WHAT IS DATA ASSIMILATION REALLY SOLVING, AND HOW IS THE CALCULATION ACTUALLY DONE?

Ichiro Fukumori

*Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA 91109*

**Abstract** Data assimilation is reviewed in the context of an inverse problem. The mathematical nature of the problem is examined and some of its common solutions are described, clarifying some of the implicit assumptions that underlie both problem and solution. For instance, Kalman filtering and Rauch-Tung-Striebel smoothing can be identified as recursive least-squares inversions of the assimilation problem but of different parts of the problem. The temporal evolution of a filtered solution is not physically consistent, but that of a smoothed solution is. Understanding these characteristics is essential in effectively assimilating observations as well as in utilizing and further improving the assimilated solution. Practical steps in implementing a filtering and smoothing algorithm are illustrated by examples from the Consortium for "Estimating the Circulation and Climate of the Ocean" (ECCO).

**Keywords:** data assimilation, Kalman filter, smoother, consistency, ECCO

## 1. INTRODUCTION

Data assimilation is a procedure in which observations are combined with models. The observations correct model errors on the one hand, and the models extrapolate the data information in space, time, and among different properties on the other. The result of assimilation is generally a more complete and more accurate description of the state of the modeled system than those obtained by either observations or model simulations alone. However, data assimilation is not a panacea for correcting every model error or for compensating all deficiencies of observations.

Because ocean models have finite degrees of freedom, model estimates are inherently different from observations regardless of errors in measurements. Moreover, most data assimilation schemes incorporate approximations and/or simplifications that dictate what is being solved and how the results could be utilized. For instance, because of data increments, budgets of heat and other properties cannot be closed in a physically

consistent manner for many sequential data assimilation estimates while for other estimates budgets can be closed. Understanding what is being solved and how it is done so are fundamental to utilizing data assimilated estimates and to devising means of improving them further.

The nature of the data assimilation problem and some of its solutions are reviewed to clarify these underlying properties and to elucidate their implications. The data assimilation problem is mathematically identified in Section 2. In Section 3, the Kalman filter and Rauch-Tung-Striebel smoother are compared in the context of a least-squares solution to this mathematical problem. The nature of data and model errors that are utilized as weights in assimilation is reviewed in Section 4. Practical issues in implementing these solutions are described in Section 5, using examples from the near real-time data assimilation system of the Consortium for “Estimating the Circulation and Climate of the Ocean” (ECCO; Stammer *et al.* 2002.) The discussion is summarized in Section 6.

## 2. DATA ASSIMILATION AS AN INVERSE PROBLEM

Mathematically, data assimilation can be identified as an inverse problem; the state of a dynamic system (model),  $\mathbf{x}$ , and its controls,  $\mathbf{u}$ , (non-state variables of the model) are estimated given a set of observations,  $\mathbf{y}$ , and a model; e.g.,

$$\begin{pmatrix} \vdots \\ \mathbf{H}\mathbf{x}_t \\ \mathbf{H}\mathbf{x}_{t+1} \\ \vdots \\ \mathbf{x}_{t+1} - \mathbf{A}\mathbf{x}_t - \mathbf{G}\mathbf{u}_t \\ \mathbf{x}_{t+2} - \mathbf{A}\mathbf{x}_{t+1} - \mathbf{G}\mathbf{u}_{t+1} \\ \vdots \end{pmatrix} \approx \begin{pmatrix} \vdots \\ \mathbf{y}_t \\ \mathbf{y}_{t+1} \\ \vdots \\ \mathbf{0} \\ \mathbf{0} \\ \vdots \end{pmatrix} \quad (1)$$

where ... denote similar equations at different instances,  $t$ , indicated by the subscripts. Variable  $\mathbf{x}$  consists of all the model’s prognostic variables and  $\mathbf{u}$  includes forcing, boundary condition, and sources of model error. Terms that include quantities to be solved ( $\mathbf{x}$  and  $\mathbf{u}$ ) are on the left hand side and the rest are placed on the right hand side. The upper part of Eq (1) relates the model state to the observations by the observation operator  $\mathbf{H}$ . The lower part describes the model’s temporal evolution by operators  $\mathbf{A}$  and  $\mathbf{G}$  that embody the model physics and dynamics. The right hand side of the model equations (lower part of Eq 1) is identically zero as all terms of the model are generally

uncertain and are placed on the left hand side. (Sources of model error are included in  $\mathbf{u}$ .)

For simplicity, we assume a linear model for most of this discussion. The problem above and the solutions discussed below can be extended to non-linear models with suitable linearization. Bold upper and lower case characters represent matrices and column vectors, respectively. The time increment from  $t$  to  $t+1$  above denotes an arbitrary increment, as opposed to a single model time-step, and corresponds to instances at which observations are available.

As in most geophysical inverse problems<sup>1</sup>, Eq (1) is rank deficient. In particular, there are generally more unknowns than the number of constraints. For example, the dimension of  $\mathbf{x}$ , excluding the temporal dimension, is of order several million for typical general circulation models, whereas there are only about 20,000 hydrographic profiles during the entire World Ocean Circulation Experiment. Consequently, there are an infinite number of solutions that could satisfy Eq (1). Different criteria are used to derive particular solutions. One such criterion is least-squares, and is reviewed below.

### 3. KALMAN FILTER AND RAUCH-TUNG-STRIEBEL SMOOTHER AS LEAST SQUARES INVERSIONS

The least squares solution (cf. Chapter 5) provides a general solution to inverse problems such as Eq (1). Namely, the least squares solution  $\hat{\mathbf{a}}$  ( $\hat{\mathbf{a}}$  denotes an estimate) and its error covariance matrix  $\mathbf{R}_{\hat{\mathbf{a}}\hat{\mathbf{a}}}$  for a general linear inverse problem,

$$\mathbf{E}\mathbf{a} = \mathbf{b} \quad (2)$$

when the right hand side  $\mathbf{b}$  is given (known), are,

$$\hat{\mathbf{a}} = \mathbf{a}_0 + \mathbf{R}_{\mathbf{aa}}\mathbf{E}^T(\mathbf{E}\mathbf{R}_{\mathbf{aa}}\mathbf{E}^T + \mathbf{R}_{\mathbf{bb}})^{-1}(\mathbf{b} - \mathbf{E}\mathbf{a}_0) \quad (3)$$

$$\mathbf{R}_{\hat{\mathbf{a}}\hat{\mathbf{a}}} = \mathbf{R}_{\mathbf{aa}} - \mathbf{R}_{\mathbf{aa}}\mathbf{E}^T(\mathbf{E}\mathbf{R}_{\mathbf{aa}}\mathbf{E}^T + \mathbf{R}_{\mathbf{bb}})^{-1}\mathbf{E}\mathbf{R}_{\mathbf{aa}} \quad (4)$$

where  $\mathbf{a}_0$  is a prior estimate of  $\mathbf{a}$ , and  $\mathbf{R}_{\mathbf{aa}}$  and  $\mathbf{R}_{\mathbf{bb}}$  are prior error covariance matrices of  $\mathbf{a}_0$  and  $\mathbf{b}$ , respectively. (The latter includes representation error for  $\mathbf{E}$ . See Section 4.1 for further discussion.) Filtering

---

<sup>1</sup> Basic matrix algebra is fundamental to mathematical discussions below and data assimilation in general. See, for instance, Wunsch (1996) for a general discussion of inverse methods that includes a brief summary of matrix and vector algebra relevant to the subject.

and smoothing algorithms can be identified as such least squares inversions and are reviewed below focusing on what they respectively solve.

A least-squares solution is identical to a minimum variance estimate when weights used in least-squares are suitable inverse error covariance matrices. These solutions are optimal in the sense that they optimize a given criteria (function) and that the expected error variance of  $\hat{\mathbf{a}}$  is minimum among all (linear) estimates. Least-squares, as well as filtering and smoothing described below, do not necessarily assume Gaussian statistics. When the statistical distribution of  $\mathbf{a}$  is Gaussian, the least-squares estimate is also a maximum likelihood estimate. Otherwise, the least-squares solution and the maximum likelihood estimate are distinct.

### 3.1 What do Kalman Filters Solve?

The Kalman filter (e.g., ChapterB) corrects model forecasts  $\hat{\mathbf{x}}_t^f$  and its error covariance matrix  $\mathbf{P}_t^f$  by,

$$\hat{\mathbf{x}}_t^a = \hat{\mathbf{x}}_t^f + \mathbf{P}_t^f \mathbf{H}^T (\mathbf{H} \mathbf{P}_t^f \mathbf{H}^T + \mathbf{R}_t)^{-1} (\mathbf{y}_t - \mathbf{H} \hat{\mathbf{x}}_t^f) \quad (5)$$

$$\mathbf{P}_t^a = \mathbf{P}_t^f - \mathbf{P}_t^f \mathbf{H}^T (\mathbf{H} \mathbf{P}_t^f \mathbf{H}^T + \mathbf{R}_t)^{-1} \mathbf{H} \mathbf{P}_t^f \quad (6)$$

using notation defined in Eq (1).  $\mathbf{R}$  is the data error covariance matrix (cf Section 4.1). Superscripts  $f$  and  $a$  denote model forecast and filter analysis, and  $\hat{\mathbf{x}}_t^a$  and  $\mathbf{P}_t^a$  are the Kalman filter's state analysis and its corresponding error covariance matrix, respectively. The correspondence between Eqs (3) and (5) and between (4) and (6) shows that the Kalman filter can be regarded as a least squares inversion of operator  $\mathbf{H}$ .

However, given that the data assimilation problem is a combined inversion of observations and model equations (Eq 1), the Kalman filter does not solve (invert) the entire data assimilation problem, in particular, the model equations (lower part of Eq 1). In fact, combining Eq (5) with the model forecasting step,  $\hat{\mathbf{x}}_t^f = \mathbf{A} \hat{\mathbf{x}}_{t-1}^a + \mathbf{G} \hat{\mathbf{u}}_t^0$ , where  $\hat{\mathbf{u}}_t^0$  is the a priori estimate of the control, the temporal evolution of the Kalman filter analysis satisfies,

$$\hat{\mathbf{x}}_t^a = \mathbf{A} \hat{\mathbf{x}}_{t-1}^a + \mathbf{G} \hat{\mathbf{u}}_{t-1}^0 + \mathbf{P}_t^f \mathbf{H}^T (\mathbf{H} \mathbf{P}_t^f \mathbf{H}^T + \mathbf{R}_t)^{-1} (\mathbf{y}_t - \mathbf{H} \hat{\mathbf{x}}_t^f) \quad (7)$$

Eq (7) is different from the model equations (lower half of Eq 1) due to the filter's data increment (third term of Eq 7). As illustrated in Figure 1, the data increment (black line) is not ascribed to particular processes as are the first two terms of Eq (7) (dotted black curve), and thus the temporal evolution between  $\hat{\mathbf{x}}_{t-1}^a$  and  $\hat{\mathbf{x}}_t^a$  is physically inconsistent. For instance, budgets of heat and other properties cannot be closed between the two instances.

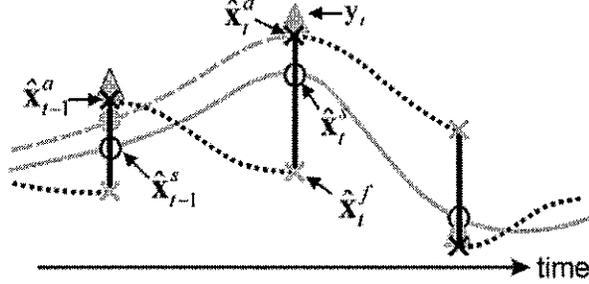


Figure 1. Schematic of a state element's temporal evolution in a typical sequential assimilation. Abscissa is time and ordinate is the state's value. Filtering progresses by a model forecasting step integrating the model along the dotted black curve from an analysis  $\hat{\mathbf{x}}_{t-1}^a$  (black cross) to a forecast  $\hat{\mathbf{x}}_t^f$  (gray cross). At time  $t$  the Kalman filter corrects the forecast  $\hat{\mathbf{x}}_t^f$  to another analysis  $\hat{\mathbf{x}}_t^a$  (black cross), bringing the model state closer to the observations  $y_t$  (gray triangle) along the solid black line. This filter correction is inverted by a smoother that corrects the model's prior evolution (dotted black curve) and the prior analysis  $\hat{\mathbf{x}}_{t-1}^a$  (black cross) as depicted by the dashed gray curve and gray circle, respectively. In turn, differences at earlier times can be further inverted backwards in time. A general smoothed estimate and its temporal evolution initiated at some future instant is depicted by the white circles (e.g.,  $\hat{\mathbf{x}}_{t-1}^s$  and  $\hat{\mathbf{x}}_t^s$ ) and the solid gray curve, respectively.

### 3.2 What is a Smoother?

Whereas filters solve only the upper part of Eq (1), smoothers invert the entire data assimilation problem identified by Eq (1). The data increment in Eq (7) represents errors in the model that is being corrected by the assimilated data. These errors include those of the prior model evolution (dotted black curve in Figure 1) as well as those of the state at the previous assimilation instant from which the model forecasting step was taken (black crosses). The correspondence between the data increment and these errors can be recognized as another inverse problem defined by the lower half of Eq (1) that has not been solved by the Kalman filter (Eq 5). The sequential smoother described below employs the filtered solution to invert the model's temporal evolution that defines this lower half of Eq (1).

Namely, given the data assimilated analysis at time  $t$ ,  $\hat{\mathbf{x}}_t^a$ , the model equations of Eq (1) define another inverse problem,

$$\hat{\mathbf{x}}_t^a = \mathbf{A}\hat{\mathbf{x}}_{t-1}^s + \mathbf{G}\hat{\mathbf{u}}_{t-1}^s = (\mathbf{A} \quad \mathbf{G}) \begin{pmatrix} \hat{\mathbf{x}}_{t-1}^s \\ \hat{\mathbf{u}}_{t-1}^s \end{pmatrix} \quad (8)$$

to estimate the model state and control at time  $t-1$ , denoted by superscript  $s$ . Eq (8) can also be solved by least squares (Eq 3). In particular, as model error sources (process noise, See Section 4.2) are explicitly included in the formulation ( $\mathbf{u}$ ), an exact solution can be sought that would satisfy model constraints (e.g., closed heat budget, etc) by estimating these errors. This

amounts to setting  $\mathbf{R}_{bb} = 0$  in Eq (3). The filtered estimate  $\hat{\mathbf{x}}_{t-1}^a$  and the a priori control  $\hat{\mathbf{u}}_{t-1}^0$  provide the prior solutions in (3), and their error covariance matrix defines the equivalent of  $\mathbf{R}_{aa}$ ;

$$\begin{pmatrix} \mathbf{P}_{t-1}^a & \mathbf{0} \\ \mathbf{0} & \mathbf{Q}_{t-1} \end{pmatrix} \quad (9)$$

where  $\mathbf{Q}$  denotes the error covariance of  $\hat{\mathbf{u}}^0$ . Standard Kalman filtering assumes temporally uncorrelated process noise that makes errors in  $\hat{\mathbf{x}}_{t-1}^a$  and  $\hat{\mathbf{u}}_{t-1}^0$  uncorrelated to each other, and thus off-diagonal blocks are zero in Eq 9.

Substitution of these elements in Eq (3) yields new estimates  $\hat{\mathbf{x}}_{t-1}^s$  and  $\hat{\mathbf{u}}_{t-1}^s$  such that,

$$\begin{pmatrix} \hat{\mathbf{x}}_{t-1}^s \\ \hat{\mathbf{u}}_{t-1}^s \end{pmatrix} = \begin{pmatrix} \hat{\mathbf{x}}_{t-1}^a \\ \hat{\mathbf{u}}_{t-1}^0 \end{pmatrix} + \begin{pmatrix} \mathbf{P}_{t-1}^a \mathbf{A}^T (\mathbf{A} \mathbf{P}_{t-1}^a \mathbf{A}^T + \mathbf{G} \mathbf{Q}_{t-1} \mathbf{G}^T)^{-1} \\ \mathbf{Q}_{t-1} \mathbf{G}^T (\mathbf{G} \mathbf{Q}_{t-1} \mathbf{G}^T + \mathbf{A} \mathbf{P}_{t-1}^a \mathbf{A}^T)^{-1} \end{pmatrix} (\hat{\mathbf{x}}_t^a - \mathbf{A} \hat{\mathbf{x}}_{t-1}^a - \mathbf{G} \hat{\mathbf{u}}_{t-1}^0) \quad (10)$$

Previous filtered estimates at time  $t-2$  can be improved and be made consistent with this estimate using these results ( $\hat{\mathbf{x}}_{t-1}^s$  as opposed to the filter analysis  $\hat{\mathbf{x}}_{t-1}^a$  in Eq 8) in another inversion. By induction, other filtered estimates at earlier instances can be improved by such inversion recursively back in time such that,

$$\begin{pmatrix} \hat{\mathbf{x}}_{t-1}^s \\ \hat{\mathbf{u}}_{t-1}^s \end{pmatrix} = \begin{pmatrix} \hat{\mathbf{x}}_{t-1}^a \\ \hat{\mathbf{u}}_{t-1}^0 \end{pmatrix} + \begin{pmatrix} \mathbf{P}_{t-1}^a \mathbf{A}^T (\mathbf{A} \mathbf{P}_{t-1}^a \mathbf{A}^T + \mathbf{G} \mathbf{Q}_{t-1} \mathbf{G}^T)^{-1} \\ \mathbf{Q}_{t-1} \mathbf{G}^T (\mathbf{G} \mathbf{Q}_{t-1} \mathbf{G}^T + \mathbf{A} \mathbf{P}_{t-1}^a \mathbf{A}^T)^{-1} \end{pmatrix} (\hat{\mathbf{x}}_t^s - \mathbf{A} \hat{\mathbf{x}}_{t-1}^a - \mathbf{G} \hat{\mathbf{u}}_{t-1}^0) \quad (11)$$

in general. (Note the use of  $\hat{\mathbf{x}}_t^s$  in the last term instead of  $\hat{\mathbf{x}}_t^a$ , thus defining a recursion.)

The recursive relation Eq (11) can be recognized as the Rauch-Tung-Striebel (RTS) fixed-interval smoother. The RTS smoother can be shown to provide estimates of the state and control using all observations within a fixed time interval and is a general solution to the assimilation problem (Eq 1). (The smoother alters all filtered estimates except that at the end of the fixed time-interval; i.e., The Kalman filter estimate at the end of the time-interval is a least-squares solution of Eq (1) but not at intervening times.)

In Eq (11), past data information is contained in the Kalman filtered analysis  $\hat{\mathbf{x}}_{t-1}^a$  while information of formally future observations is carried backward in time by the smoothed estimate  $\hat{\mathbf{x}}_t^s$ . Owing to the additional information from formally future observations, the smoothed estimates are generally more accurate (has smaller error) than corresponding filtered estimates. The error covariance matrix of the smoothed estimates  $\hat{\mathbf{x}}_{t-1}^s$  and  $\hat{\mathbf{u}}_{t-1}^s$  (Eq 11),  $\mathbf{P}_{t-1}^s$  and  $\mathbf{Q}_{t-1}^s$ , respectively, is given by,

$$\begin{pmatrix} \mathbf{P}_{t-1}^s \\ \mathbf{Q}_{t-1}^s \end{pmatrix} = \begin{pmatrix} \mathbf{P}_{t-1}^a \\ \mathbf{Q}_{t-1}^a \end{pmatrix} - \begin{pmatrix} \mathbf{P}_{t-1}^a \mathbf{A}^T (\mathbf{A} \mathbf{P}_{t-1}^a \mathbf{A}^T + \mathbf{G} \mathbf{Q}_{t-1}^a \mathbf{G}^T)^{-1} \mathbf{A} \mathbf{P}_{t-1}^a \\ \mathbf{Q}_{t-1}^a \mathbf{G}^T (\mathbf{G} \mathbf{Q}_{t-1}^a \mathbf{G}^T + \mathbf{A} \mathbf{P}_{t-1}^a \mathbf{A}^T)^{-1} \mathbf{G} \mathbf{Q}_{t-1}^a \end{pmatrix} \quad (12)$$

$$+ \begin{pmatrix} \mathbf{L}_{t-1} \mathbf{P}_t^a \mathbf{L}_{t-1}^T \\ \mathbf{M}_{t-1} \mathbf{P}_t^a \mathbf{M}_{t-1}^T \end{pmatrix}$$

where,

$$\begin{aligned} \mathbf{L}_{t-1} &\equiv \mathbf{P}_{t-1}^a \mathbf{A}^T (\mathbf{A} \mathbf{P}_{t-1}^a \mathbf{A}^T + \mathbf{G} \mathbf{Q}_{t-1}^a \mathbf{G}^T)^{-1} = \mathbf{P}_{t-1}^a \mathbf{A}^T \mathbf{P}_t^{f-1} \\ \mathbf{M}_{t-1} &\equiv \mathbf{Q}_{t-1}^a \mathbf{G}^T (\mathbf{G} \mathbf{Q}_{t-1}^a \mathbf{G}^T + \mathbf{A} \mathbf{P}_{t-1}^a \mathbf{A}^T)^{-1} = \mathbf{Q}_{t-1}^a \mathbf{G}^T \mathbf{P}_t^{f-1} \end{aligned} \quad (13)$$

are the coefficient matrices in Eq (11) (smoother gain matrices) introduced for shorthand notation.

The correspondence between Eqs (11) and (3) and between (12) and (4) shows that the RTS smoother is a recursive inversion of the model (Eq 8). In particular, the smoothed state estimate (upper part of Eq 11) and smoothed control estimate (lower part of Eq 11) can be identified as inversions of  $\mathbf{A}$  and  $\mathbf{G}$ , respectively. Moreover, as illustrated above, the smoother solution was derived to exactly satisfy the model equation, which can also be found by substituting results of Eq (11) to the right hand side of Eq (8) to yield,

$$\hat{\mathbf{x}}_t^s = \mathbf{A} \hat{\mathbf{x}}_{t-1}^s + \mathbf{G} \hat{\mathbf{u}}_{t-1}^s \quad (14)$$

The additional (last) term in Eq (12) relative to (4) reflects the uncertainties of the left hand side of Eq (8), and similar equations at other instances, while the smoother solves for such exact solution. The inversion and the physical consistency of the smoothed estimates are illustrated by the gray curves in Figure 1.

Although smoothed solutions satisfy model equations (Eq 14), smoothing should not be confused with the so-called "strong constraint" estimation (Sasaki, 1970) that assumes that models have no errors except in initial condition. In fact, the model solution by itself does not satisfy the model; viz.,  $\hat{\mathbf{x}}_t^s \neq \mathbf{A} \bar{\mathbf{x}}_{t-1}^s + \mathbf{G} \hat{\mathbf{u}}_{t-1}^0$ . Smoothing is generally a "weak constraint" inversion that allows for model errors, but one that explicitly provides estimates of these inaccuracies. The explicit estimation of these model error

sources as opposed to leaving them unknown ( $\hat{\mathbf{u}}_t^s$  in Eq 14 instead of  $\hat{\mathbf{u}}_t^0$ ), is what allows for the temporal evolution of the smoothed solution to be physically consistent.

While the discussion above has focused on sequential smoothing, there are other equally effective smoothing algorithms. In particular, when model error sources are made part of the estimate, the so-called adjoint method or 4dVAR (Chapter 5) is equivalent to the RTS smoother (Eq 11). The adjoint estimation directly solves for the smoothed solution ( $\hat{\mathbf{x}}^s$  and  $\hat{\mathbf{u}}^s$ ) without deriving intermediate filter estimates.

#### 4. WHAT ARE DATA ERRORS AND MODEL ERRORS?

“Data” error covariance  $\mathbf{R}$  and “model” error covariance  $\mathbf{Q}$  in effect define the solution to the data assimilation problem (e.g., Eqs 5 and 11). ( $\mathbf{P}$  is a function of  $\mathbf{R}$  and  $\mathbf{Q}$ .) Their understanding and specification are, therefore, fundamental to assimilation and in utilizing their results. In fact, as described below, a part of what is commonly regarded as “model” error should in fact be considered “data” error.  $\mathbf{R}$  and  $\mathbf{Q}$  are better considered error covariances of the “data constraint” and the “model constraint”, respectively.

##### 4.1 “Data” Error

“Data” and “model” errors can be best understood by considering the true nature of the model vis-à-vis that of the observations and the ocean. The following discussion follows that of Cohn (1997). For instance, the model’s true state  $\bar{\mathbf{x}}_t$  (overbar denotes true solution) can be recognized as representing the ocean in finite dimension,

$$\bar{\mathbf{x}}_t \equiv \mathbf{\Pi} \mathbf{w}_t \quad (15)$$

Function  $\mathbf{\Pi}$  defines the model state given the complete state of the ocean  $\mathbf{w}_t$  (which has infinite degrees of freedom). Observations  $\mathbf{y}_t$  are samples of this ocean  $\mathbf{w}_t$  that could be written as,

$$\mathbf{y}_t = \mathbf{E} \mathbf{w}_t + \varepsilon \quad (16)$$

where  $\mathbf{E}$  describes the sampling operation and  $\varepsilon$  denotes measurement errors (e.g., instrument error).

In terms of the model, Eq (16) can be rewritten as,

$$\mathbf{y}_t = \mathbf{H} \bar{\mathbf{x}}_t + \{\mathbf{E} \mathbf{w}_t - \mathbf{H} \mathbf{\Pi} \mathbf{w}_t\} + \varepsilon \quad (17)$$

The last two terms of Eq (17),

$$\{\mathbf{E}\mathbf{w}_t - \mathbf{H}\mathbf{I}\mathbf{w}_t\} + \varepsilon \quad (18)$$

can be identified as the error of the observation equation  $\mathbf{y}_t = \mathbf{H}\mathbf{x}_t$ , that defines the data assimilation problem (Eq 1), i.e., the covariance of Eq (18) is  $\mathbf{R}$ . The first part of (18) is the difference between error free observations ( $\mathbf{E}\mathbf{w}_t$ ) and error-free equivalent of the model ( $\mathbf{H}\mathbf{I}\mathbf{w}_t$ ). The two are generally different because the model does not simulate the entire spectrum of the ocean but only parts of it (Eq 15).

For instance, a coarse resolution model of  $1^\circ$  horizontal resolution does not simulate meso-scale variability, and a 1.5-layer reduced-gravity model does not simulate barotropic motion. What a model cannot simulate constitutes part of the errors of the observation equation as described by Eq (18) and is termed “representation error.” Other common examples of representation error include,

- Baroclinic variability for a barotropic model
- External gravity waves for a rigid-lid model
- Skin temperature for most models with thick surface layers
- Micro-structure for most large-scale models

In numerical weather forecasting, meteorologists typically employ larger “data” error than the measurement accuracy of the observations so as to maximize the skill of their forecasts. Forcing models to agree with observations that the models cannot simulate, result in models propagating the data correction incorrectly in time, causing larger errors in the model evolution than otherwise.

Some observations are dominated by representation error, making them difficult to utilize. For instance, individual drifter and float trajectories can depend on small-scale variabilities of the ocean, such that two floats deployed a short distance away from each other have dramatically different trajectories (e.g., Paduan and Niiler, 1993). Such measurements (Lagrangian trajectory as opposed to Eulerian velocities along the trajectory) that are dominated by representation errors do not provide strong data constraints, and cannot be used effectively.

## 4.2 “Model” Error

The nature of model process noise  $\mathbf{Q}$  can be deduced in a similar fashion as data error above. The model can be written in shorthand as,

$$\mathbf{x}_{t+1} = A(\mathbf{x}_t, \hat{\mathbf{u}}_t^0) \quad (19)$$

where  $A$  denotes a general function describing the evolution of the model state.  $\hat{\mathbf{u}}_t^0$  denotes the model’s particular control that includes its forcing, boundary condition, and parameters. For generality,  $\hat{\mathbf{u}}_t^0$  also includes other

sources of process noise as discussed below that are zero a priori. The ocean evolution can be thought of similarly as,

$$\mathbf{w}_{t+1} = L(\mathbf{w}_t, \mathbf{v}_t) \quad (20)$$

where  $L$  describes the evolution of the ocean and  $\mathbf{v}_t$  is the forcing and boundary conditions of the ocean.

Then, the model evolution in terms of the true model state can be written as,

$$\begin{aligned} \bar{\mathbf{x}}_{t+1} &= \mathbf{\Pi} \mathbf{w}_{t+1} = \mathbf{\Pi} L(\mathbf{w}_t, \mathbf{v}_t) \\ &= A(\bar{\mathbf{x}}_t, \hat{\mathbf{u}}_t^0) + \{ \mathbf{\Pi} L(\mathbf{w}_t, \mathbf{v}_t) - A(\mathbf{\Pi} \mathbf{w}_t, \hat{\mathbf{u}}_t^0) \} \end{aligned} \quad (21)$$

using Eqs (15), (19) and (20). The last term in  $\{ \}$  mathematically describes what model error (process noise) is; process noise is the difference between the true evolution of the ocean projected to the model space  $\mathbf{\Pi} L(\mathbf{w}_t, \mathbf{v}_t)$  and the model evolution given the true model state and its particular control  $A(\mathbf{\Pi} \mathbf{w}_t, \hat{\mathbf{u}}_t^0)$ .

As shown by Eq (21), process noise could be due to errors in the given control ( $\mathbf{v}_t$  versus its equivalent in  $\hat{\mathbf{u}}_t^0$ ) and to differences in model algorithm  $A$  and the true model evolution  $L$  and their interaction with operator  $\mathbf{\Pi}$ . The former includes, for example, errors in the particular external forcing, boundary condition, and model parameters used by the model. The latter includes errors due to finite differencing, truncation, and interaction with scales and processes ignored by the model. The two types of error sources could be considered external and internal errors of the model algorithm, respectively, and are both identified as elements of the model control vector.

### 4.3 Specification of Data Error and Model Error

While their principles are understood, the actual specification of data and model error covariances are not trivial. For instance, it is not entirely clear what operator  $\mathbf{\Pi}$  that defines these errors is for different models, let alone the errors' statistical properties. However, there are some practical means of quantifying these errors prior to assimilation. Here we describe the so-called "covariance matching" method described by Fu et al. (1993).

Observations  $\mathbf{y}$  and their model simulation's equivalent  $\mathbf{m}$  could be written as the sum of the true signal  $\mathbf{s}$  (1<sup>st</sup> term on the right hand side of Eq 17) and their respective uncertainties  $\mathbf{r}$  and  $\mathbf{p}$ ,

$$\begin{aligned} \mathbf{y} &= \mathbf{s} + \mathbf{r} \\ \mathbf{m} &= \mathbf{s} + \mathbf{p} \end{aligned} \quad (22)$$

To first approximation, we may assume  $\mathbf{s}$ ,  $\mathbf{r}$ , and  $\mathbf{p}$  to have zero means and to be uncorrelated with each other. (See section 5.2.5 for dealing with non-zero means.) Then, the covariance among these elements can be written as,

$$\begin{aligned}\langle \mathbf{y}\mathbf{y}^T \rangle &= \langle \mathbf{s}\mathbf{s}^T \rangle + \langle \mathbf{r}\mathbf{r}^T \rangle \\ \langle \mathbf{m}\mathbf{m}^T \rangle &= \langle \mathbf{s}\mathbf{s}^T \rangle + \langle \mathbf{p}\mathbf{p}^T \rangle \\ \langle \mathbf{y}\mathbf{m}^T \rangle &= \langle \mathbf{s}\mathbf{s}^T \rangle\end{aligned}\quad (23)$$

where brackets denote statistical expectation. Assuming ergodicity and stationarity, quantities on the left hand side could be estimated by averaging the data and model estimates in time. Then,

$$\begin{aligned}\langle \mathbf{r}\mathbf{r}^T \rangle &= \langle \mathbf{y}\mathbf{y}^T \rangle - \langle \mathbf{y}\mathbf{m}^T \rangle \\ \langle \mathbf{p}\mathbf{p}^T \rangle &= \langle \mathbf{m}\mathbf{m}^T \rangle - \langle \mathbf{y}\mathbf{m}^T \rangle\end{aligned}\quad (24)$$

The former is a direct estimate of data error covariance matrix  $\mathbf{R}$ , while the latter provides an indirect estimate of process noise covariance  $\mathbf{Q}$ . Namely, given a process noise model ( $\mathbf{u}$  in Eq 1) and its covariance  $\mathbf{Q}$ , the corresponding model simulation error can be estimated using standard methods. In particular, using the notation defined in Eq (1), the stationary limit of such error  $\mathbf{P}^{sim}$  is the solution to the Lyapunov Equation,

$$\mathbf{P}^{sim} = \mathbf{A}\mathbf{P}^{sim}\mathbf{A}^T + \mathbf{G}\mathbf{Q}\mathbf{G}^T \quad (25)$$

which is related to the empirical estimate Eq (24) by,

$$\mathbf{H}\mathbf{P}^{sim}\mathbf{H}^T = \langle \mathbf{p}\mathbf{p}^T \rangle \quad (26)$$

Eqs (24), (25) and (26) provide a means to calibrate the process noise estimate  $\mathbf{Q}$ .

Figure 2 illustrates an example of such estimate for assimilating altimetric sea level data with a coarse ( $1^\circ$ ) resolution model. Because of the model's limited spatial resolution, the data error estimate (a) is dominated by meso-scale variability that constitutes the model's representation error (Section 4.1), as evidenced by large values in western boundary regions. Wind error (c) is estimated to be the dominant source of model error for simulating large-scale sea level variability. Note the first order correspondence between the empirical (b) and theoretical (d) error estimates of model simulated sea level.

## 5. EXAMPLES OF IMPLEMENTING ASSIMILATION; HOW IS ASSIMILATION ACTUALLY DONE?

While the theory of data assimilation is well understood, implementing assimilation is often nontrivial owing to its large computational requirements

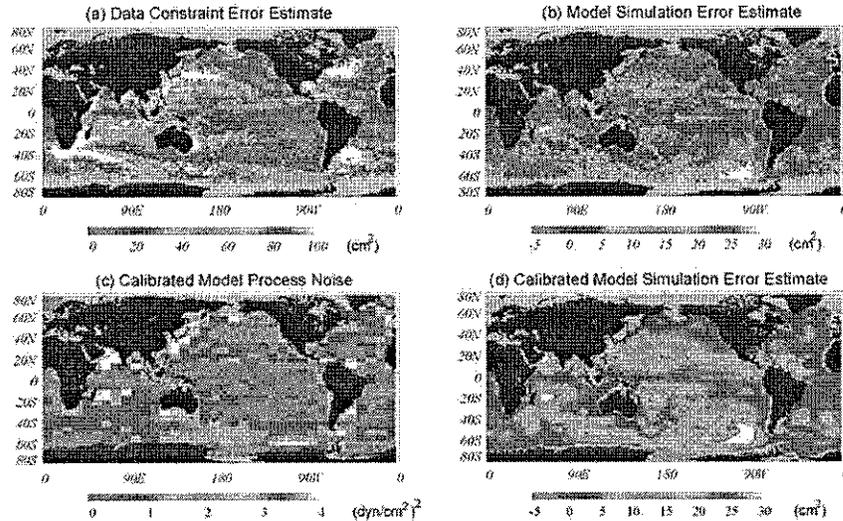


Figure 2. An example of prior error (variance) calibration; (a), (b) error estimates of altimetric sea level constraint and model simulated sea level, respectively, based on a model-data comparison (Eq 24), (c) calibrated wind stress error estimate (zonal component), (d) model simulated sea level error estimate based on (c). Note the first order consistency between (b) and (d) (Eq 26). The model is an ocean general circulation model with a  $1^\circ$  spatial resolution. (Adapted from Fukumori, *et al.*, 1999.)

and to a number of approximations and simplifications that are necessary to make the calculations tractable. An example specifying prior error estimates was described in section 4.3. Other examples of actually carrying out assimilation are described below to further elucidate some practical steps employed in an assimilation system. The examples are taken from the assimilation system of the Consortium for “Estimating the Circulation and Climate of the Ocean (ECCO).”

### 5.1 Consortium for “Estimating the Circulation and Climate of the Ocean” (ECCO)

The ECCO Consortium focuses on advancing data assimilation from an experimental tool to an operational means to study ocean circulation (Stammer *et al.*, 2002.) ECCO estimates are characterized by their physical consistency (Section 3.2) owing to smoothing algorithms (RTS smoother and adjoint method). The estimates are based on a state-of-the-art primitive equation model (MITgcm; Marshall *et al.* 1997) and employ a diverse suite of in situ and satellite remote sensing observations including temperature and salinity profiles and sea level.

The ECCO estimates are available from its data server at <http://www.ecco-group.org/las>. In particular, ECCO has established a near real-time analysis producing estimates of large-scale global ocean circulation (73°S~73°N) on a regular basis (<http://ecco.jpl.nasa.gov/external>). The model employed is of moderate resolution (1° telescoping to 1/3° within 10° of the equator, 10m layers within 150m of the surface with a total of 46 vertical levels) with its parameters adjusted by a Green's function estimation (Menemenlis, *et al.*, 2004.) The near real-time analysis is conducted by an approximate Kalman filter and RTS smoother. Aspects of this near real-time assimilation are reviewed below.

## 5.2 ECCO Near Real-Time Analysis System

The recursive nature of the Kalman filter and RTS smoother is particularly suitable for near real-time computation. However, the computational requirements of evaluating the state error covariance matrix  $\mathbf{P}$  make direct application of these methods impractical for most state-of-the-art ocean circulation models. Therefore, various methods have been put forth that approximate the derivation of  $\mathbf{P}$  so as to make Kalman filtering and RTS smoothing feasible. In ECCO, three approximations are concurrently employed:

- I. Time-asymptotic approximation (Fukumori *et al.* 1993),
- II. State reduction (Fukumori and Rizzoli, 1995),
- III. Partitioning (Fukumori, 2002).

The *time-asymptotic* approximation evaluates and employs a time-invariant representative limit of  $\mathbf{P}$ , thereby eliminating the computational cost associated with the continued model integration of the state error covariance matrix. Evaluation of this asymptotic limit is simplified by *partitioning* and *state reduction* where independent elements of  $\mathbf{P}$  are evaluated separately from one another (*partition*) and within each partition only the dominant modes of the error are estimated (*state reduction*). A reduced-state model is derived for each partition to evaluate the errors while the original fully nonlinear unapproximated model is used to integrate the state. The smaller dimensionality of each partitioned-reduced-state model reduces the computational cost of evaluating  $\mathbf{P}$ . Unlike global single-stage state reductions, the partitioning permits retaining many of the estimation problem's degrees of freedom without incurring excessive computational requirements.

The reader is referred to Chapter B and to references above for further discussion on theoretical aspects of these and other approximations. Here we review examples of implementing the approximations and their implications.

### 5.2.1 Identifying Process Noise

To first approximation, different model errors sources could be considered independent of one another. Then different process noise and their consequent model state errors could be evaluated separately in the context of a partitioned estimation (Fukumori, 2002).

Different sources of process noise cause different errors in the modeled state. For instance, the response of a model to changes in large-scale wind can be effectively described in terms of the gravest few vertical dynamic modes (e.g., Cane, 1984). In comparison, a model's response to changes in air-sea heat flux is to first approximation confined to the sea surface. The modeled process noise dictates the most effective state approximation (e.g., state reduction and partitioning), and, therefore, its identification is the first step in designing an assimilation system.

The ongoing ECCO near real-time assimilation estimates uncertainties of wind forcing and its resulting model state errors. This estimate should not be confused as one that considers all model errors are due to errors in wind, but it is an estimate of only a part of the errors, as discussed above, albeit one of the dominant ones. The model's controllability (ability to uniquely solve  $\mathbf{u}$  in Eq 1) limits aliasing of other model error sources to the particular process noise being estimated. The ECCO near real-time assimilation system described below is correspondingly designed to resolve the dominant response of the ocean to large-scale wind errors.

### 5.2.2 Regional Partitions

Due to large differences in temporal and spatial scales, wind-driven barotropic errors could be considered independent of baroclinic errors and thus estimated separately. Having sub-basin length scales, the baroclinic components are estimated individually among seven different basins across the globe (Figure 3). These regions include three separate tropical basins (Indian, Pacific, Atlantic) and four mid- and high-latitude basins (North Pacific, North Atlantic, South Atlantic and Indian, South Pacific). The regions overlap each other to minimize edge effects caused by the regional approximation; errors in overlapping areas are considered to be split among the different regions. The barotropic component, due to its large spatial scales, is estimated simultaneously over the entire model domain.

### 5.2.3 State Reduction

Within each partition, additional vertical and horizontal approximations are defined to further reduce the errors' dimension. Vertically, state errors are expanded in terms of vertical dynamic modes of velocity and vertical

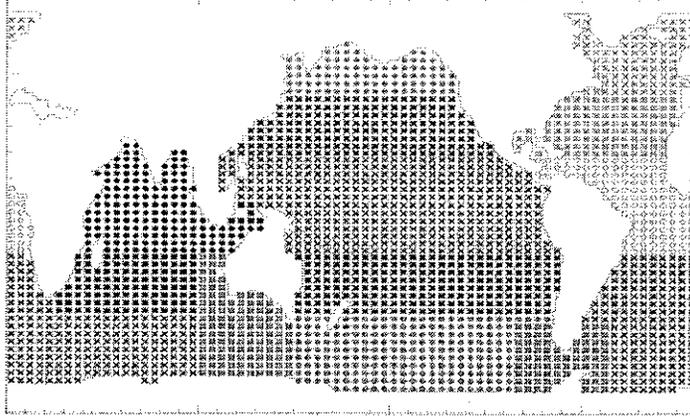


Figure 3. Coarse horizontal grid employed in ECCO partitioned reduced-state approximation. The different symbols denote different regional reduced-grid partitions used to estimate baroclinic errors of the model state.

displacement. For each baroclinic partition (Figure 3) the first five baroclinic modes are retained. Horizontally, large-scale errors are estimated by defining a coarse horizontal grid and an interpolation operator to map the coarse grid errors onto the model (fine) grid. The process noise (wind error) is reduced likewise, utilizing the same horizontal mapping operation.

The coarse grid is defined as a 5°-by-3° and 6°-by-6° (zonal and meridional resolution) grid for baroclinic and barotropic partitions, respectively. Objective mapping (Bretherton *et al.*, 1976) is employed as the coarse-to-fine grid interpolation operator, which can also be identified as a least-squares operator in itself (Eq 3). The interpolation assumes no underlying error and a Gaussian covariance function using the coarse grid dimensions as the correlation distance. To prevent spurious correlation across land (e.g., Pacific Ocean to Atlantic Ocean across the Isthmus of Panama), distances between model grid points used to define the mapping operation are computed around the model's land points.

The reduced state error thus consists of dynamic mode amplitudes  $\mathbf{a}_u$ ,  $\mathbf{a}_v$ ,  $\mathbf{a}_\eta$  defined on a coarse grid for zonal and meridional velocity and vertical displacement, respectively. The approximated control  $\mathbf{a}_\tau$  is the magnitude of wind error defined on the same coarse horizontal grid. These approximated errors are related to those of the model state and model forcing by,

$$\delta \mathbf{u} = \mathbf{D}_{vel} \mathbf{O} \mathbf{a}_u, \quad \delta \mathbf{v} = \mathbf{D}_{vel} \mathbf{O} \mathbf{a}_v, \quad \delta \boldsymbol{\eta} = \mathbf{D}_\eta \mathbf{O} \mathbf{a}_\eta, \quad \delta \boldsymbol{\tau} = \mathbf{O} \mathbf{a}_\tau \quad (27)$$

where  $\delta \mathbf{u}$ ,  $\delta \mathbf{v}$ ,  $\delta \boldsymbol{\eta}$ , and  $\delta \boldsymbol{\tau}$  are errors of model zonal and meridional velocity, vertical displacement, and wind stress, respectively. The  $\mathbf{D}$ s consist of structures of vertical dynamic modes of respective variables that project the errors vertically to the model grid.  $\mathbf{O}$  denotes the horizontal mapping

operator from the coarse grid to the model grid. Errors of other state variables are diagnostically derived using estimates in Eq (27). For instance, errors of temperature  $\mathbf{T}$  and salinity  $\mathbf{S}$  are derived from those of displacement by,

$$\delta\mathbf{T} = \frac{\partial\mathbf{T}}{\partial z} \delta\boldsymbol{\eta}, \quad \delta\mathbf{S} = \frac{\partial\mathbf{S}}{\partial z} \delta\boldsymbol{\eta} \quad (28)$$

and errors of sea level can be defined as a function of  $\delta\boldsymbol{\eta}$  and density (temperature and salinity). Errors from different partitions are summed together to form the overall model error estimate.

The total dimension of each partitioned-reduced-state is summarized in Table 1. The largest partition is the tropical Pacific cell consisting of nearly 12000 elements. In comparison, the total dimension of the model state (horizontal velocity, temperature, and salinity on the model grid) is 8 million.

*Table 1.* The reduced-state dimension of seven baroclinic partitions and the global barotropic partition. Each baroclinic partition employs the five gravest baroclinic modes. Each partition has three variables; zonal and meridional velocity and vertical displacement.

Partition	Grid Points	Dimension
Tropical Indian	308	4620
Tropical Pacific	787	11805
Tropical Atlantic	350	5250
South Pacific	633	9495
South Atlantic & Indian	664	9960
North Pacific	271	4065
North Atlantic	198	2970
Global Barotropic	963	2889

#### 5.2.4 Derivation of State Error Covariance Matrix

A time-invariant state error covariance matrix is derived for each separate partition by computing the asymptotic limit of the respective Riccati equation. (The Riccati equation describes the temporal evolution of the state error covariance matrix when integrating the model and assimilating observations. See ChapterB.) The computation employs a representative approximation of the assimilation problem in which time-invariant system matrices  $\mathbf{A}$ ,  $\mathbf{G}$ ,  $\mathbf{H}$ ,  $\mathbf{Q}$ , and  $\mathbf{R}$ , (Eqs 1, 5 and 10) are derived and used. The so-called “doubling algorithm” provides an effective means to integrate the corresponding Riccati equation in increasing time-steps of powers of two (i.e., doubling) (Fukumori *et al.* 1993).

Model matrices  $\mathbf{A}$ ,  $\mathbf{G}$ , and  $\mathbf{H}$  are derived using a coarse grain Green's function of the corresponding partitioned reduced-state model and is

computed by combining the state approximation and the original unapproximated model. For instance, a general state and control perturbation (error),  $\delta \mathbf{x}$  and  $\delta \mathbf{u}$ , can be written as,

$$\begin{aligned}\delta \mathbf{x} &= \mathbf{B} \delta \mathbf{x}' + \mathbf{N} \mathbf{n} \\ \delta \mathbf{u} &= \tilde{\mathbf{B}} \delta \mathbf{u}' + \tilde{\mathbf{N}} \mathbf{m}\end{aligned}\quad (29)$$

where  $\mathbf{B}$  and  $\mathbf{N}$  define the range and null space of a particular partitioned reduced state (control) approximation described in Sections 5.2.2 and 5.2.3, respectively, and  $\delta \mathbf{x}'$  and  $\mathbf{n}$  are their amplitudes.  $\tilde{\mathbf{B}}$ ,  $\tilde{\mathbf{N}}$ ,  $\delta \mathbf{u}'$  and  $\mathbf{m}$  are corresponding counterparts for the control.

Then, given a general (nonlinear) model, Eq (19), the perturbations satisfy,

$$\delta \mathbf{x}_{t+1} = A(\tilde{\mathbf{x}} + \delta \mathbf{x}, \tilde{\mathbf{u}} + \delta \mathbf{u}) - A(\tilde{\mathbf{x}}, \tilde{\mathbf{u}}) \quad (30)$$

where  $\tilde{\mathbf{x}}$  and  $\tilde{\mathbf{u}}$  are a representative state and control, respectively. (Time-means are used.) Substituting Eq (29) into (30) and multiplying both sides of the equation with the pseudo inverse of  $\mathbf{B}$ , denoted  $\mathbf{B}^*$ , and noting the orthogonality between  $\mathbf{B}$  and  $\mathbf{N}$ , we have,

$$\delta \mathbf{x}'_{t+1} = \mathbf{B}^* \left( A(\tilde{\mathbf{x}} + \mathbf{B} \delta \mathbf{x}' + \mathbf{N} \mathbf{n}, \tilde{\mathbf{u}} + \tilde{\mathbf{B}} \delta \mathbf{u}' + \tilde{\mathbf{N}} \mathbf{m}) - A(\tilde{\mathbf{x}}, \tilde{\mathbf{u}}) \right) \quad (31)$$

The approximation's dependence on the null space ( $\mathbf{n}$  and  $\mathbf{m}$ ) is a source of error in defining a reduced-state model. However, because of their orthogonality, this dependency could be ignored if range and null space perturbations remain within their respective domain through the model integration, as  $\mathbf{B}^*$  in Eq (31) will nullify any resulting null space perturbation. For example, to first approximation, a particular dynamic mode remains the same mode and large-scale perturbations remain large-scale. Then Eq (31) could be approximated in closed form in the reduced-space as,

$$\delta \mathbf{x}'_{t+1} = \mathbf{B}^* \left( A(\tilde{\mathbf{x}} + \mathbf{B} \delta \mathbf{x}', \tilde{\mathbf{u}} + \tilde{\mathbf{B}} \delta \mathbf{u}') - A(\tilde{\mathbf{x}}, \tilde{\mathbf{u}}) \right) \quad (32)$$

defining the partitioned reduced-state model.

Corresponding partitioned reduced-state matrices  $\mathbf{A}'$  and  $\mathbf{G}'$  that linearize Eq (32) around the representative state and control ( $\tilde{\mathbf{x}}$  and  $\tilde{\mathbf{u}}$ ),

$$\mathbf{B}^* \left( A(\tilde{\mathbf{x}} + \mathbf{B} \delta \mathbf{x}', \tilde{\mathbf{u}} + \tilde{\mathbf{B}} \delta \mathbf{u}') - A(\tilde{\mathbf{x}}, \tilde{\mathbf{u}}) \right) \approx \mathbf{A}' \delta \mathbf{x}' + \mathbf{G}' \delta \mathbf{u}' \quad (33)$$

are derived as coarse grain Green's functions. (The prime denotes the individual partitioned reduced-state equivalent.) Namely, an arbitrary column of the two matrices,  $(\mathbf{A}')_i$  and  $(\mathbf{G}')_i$ , can be numerically derived as,

$$\begin{aligned}(\mathbf{A}')_i &= \mathbf{A}' \mathbf{e}_i + \mathbf{G}' \mathbf{0} = \mathbf{B}^* \left( A(\tilde{\mathbf{x}} + \mathbf{B} \mathbf{e}_i, \tilde{\mathbf{u}}) - A(\tilde{\mathbf{x}}, \tilde{\mathbf{u}}) \right) \\ (\mathbf{G}')_i &= \mathbf{A}' \mathbf{0} + \mathbf{G}' \mathbf{e}_i = \mathbf{B}^* \left( A(\tilde{\mathbf{x}}, \tilde{\mathbf{u}} + \tilde{\mathbf{B}} \mathbf{e}_i) - A(\tilde{\mathbf{x}}, \tilde{\mathbf{u}}) \right)\end{aligned}\quad (34)$$

where  $\mathbf{e}_i$  is the corresponding column of the identity matrix of appropriate dimension and  $\mathbf{0}$  is a vector of zeroes.

Model implementation of Eq (34), and in particular, the pseudo inverse  $\mathbf{B}^*$ , requires some consideration. Since vertical displacement is not a variable in most models and inverting Eq (28) can be difficult where stratification is weak, vertical velocity is integrated in time in Eq (34) to diagnose  $\delta\eta$  (cf. Section 5.2.3). Because of their orthogonality, implementing the pseudo inverse of the vertical transformation (the  $\mathbf{D}$ s in Eq 27 that make up  $\mathbf{B}$ ) is trivial. However, the pseudo inverse of the horizontal operator  $\mathbf{O}$  is not. The objective mapping operator is relatively sparse, and, therefore,  $\mathbf{O}$  in Eq (27) is implemented as a sparse matrix multiplication retaining only significant elements of the matrix. However, the pseudo inverse of  $\mathbf{O}$  tends to be a fairly large and dense matrix. An effective means of implementing the inversion of  $\mathbf{O}$  is as,

$$\mathbf{O}^* = (\mathbf{O}^T \mathbf{O})^{-1} \mathbf{O}^T \quad (35)$$

Matrix  $(\mathbf{O}^T \mathbf{O})^{-1}$  is a relatively small matrix that can be precomputed and stored. The left multiplication by  $\mathbf{O}$  transpose can be achieved algorithmically given the sparse matrix  $\mathbf{O}$  that is already available. (A multiplication by  $\mathbf{O}^T$  is an adjoint of  $\mathbf{O}$ .)

The partitioned reduced-state observation matrix  $\mathbf{H}'$  can be numerically derived similarly to those in Eq (34):

$$(\mathbf{H}')_i = \mathbf{H}' \mathbf{e}_i = H(\bar{\mathbf{x}} + \mathbf{B} \mathbf{e}_i) - H(\bar{\mathbf{x}}) \quad (36)$$

where  $H$  is a function defining the model equivalent of the observations.

The time-asymptotic approximation employs a time-invariant system in which not only the model ( $\mathbf{A}'$  and  $\mathbf{G}'$ ) but the observation matrix  $\mathbf{H}'$  and the data and model error covariance matrices  $\mathbf{R}$  and  $\mathbf{Q}'$  are stationary. (Only the operators  $\mathbf{A}'$ ,  $\mathbf{G}'$ , and  $\mathbf{H}'$  and the statistics  $\mathbf{R}$  and  $\mathbf{Q}'$  are assumed stationary, not the state, control, or observation.) However, since in practice what is observed ( $H$ ) varies in time, a representative set of observations is assumed to be available regularly in deriving the state error covariance matrix. For instance, to simulate the coverage and accuracy of satellite altimeter data, a three-day assimilation cycle is assumed during which all satellite altimeter data within the 10-day repeat period is available but with 3-times the assumed data error.

The resulting system matrices are used to integrate the corresponding Riccati Equation to its asymptotic limit utilizing the doubling algorithm.

### 5.2.5 Implementation

Although the derivation of  $\mathbf{P}$  assumed a 3-day assimilation interval, actual assimilation is performed every 6-hours (model time-step is 1 hour), assimilating all available observations within 3-hours of the assimilation

instant. No observation is utilized more than once as dictated by standard estimation theory. The 6-hour assimilation interval is a compromise between computational requirements associated with applying the Kalman filter more frequently and the resolution of high frequency variability of the ocean (e.g., wind-driven barotropic motion).

For computational efficiency, the assimilation employs an alternate form of the Kalman gain matrix from the common formulation of Eq (5),

$$\mathbf{P}_i^f \mathbf{H}^T (\mathbf{H} \mathbf{P}_i^f \mathbf{H}^T + \mathbf{R}_i)^{-1} = \mathbf{P}_i^a \mathbf{H}^T \mathbf{R}_i^{-1} \quad (37)$$

The alternate form on the right hand side of Eq (37) employs the analysis error covariance instead of the forecast error covariance, and has fewer computational steps than the left hand side, when the respective state error covariance matrices are given.

In the partitioned reduced-state formulation, the filter (data) increment (the difference between analyzed state and forecast state, i.e., the third term of Eq 7),  $\Delta \hat{\mathbf{x}}_i^a$ , can be written as a sum of the increments in different partitions,

$$\Delta \hat{\mathbf{x}}_i^a = \sum_j \mathbf{B}_i \Delta \hat{\mathbf{x}}_{i,j}^{a'} \quad (38)$$

where,

$$\Delta \hat{\mathbf{x}}_{i,j}^{a'} = \mathbf{P}_i^{a'} \mathbf{H}_{i,j}'^T \mathbf{R}_i^{-1} (\mathbf{y}_i - H(\hat{\mathbf{x}}_i^f)) \quad (39)$$

is the filter increment of an individual partitioned reduced-state (subscript  $i$ ). In Eq (39), the reduced state observation matrix  $\mathbf{H}'$  (Eq 36) can be used. Alternatively, Eq (39) could be implemented as,

$$\Delta \hat{\mathbf{x}}_{i,j}^{a'} = \mathbf{P}_i^{a'} \mathbf{B}_i^T \mathbf{H}_i^T \mathbf{R}_i^{-1} (\mathbf{y}_i - H(\hat{\mathbf{x}}_i^f)) \quad (40)$$

using the adjoint of the model observation operator  $H$  as the left multiplication  $\mathbf{H}^T$ . Eq (40) involves less approximation and is of particular convenience when the observation operator is an implicit function of the state and its adjoint is available.

Due to inaccuracies in the marine geoid estimate, the analysis assimilates altimetric sea level anomaly relative to its temporal mean instead of absolute sea surface height. For each partitioned reduced state, Eq (39) is computed by,

$$\Delta \hat{\mathbf{x}}_{i,j}^{a'} = \mathbf{P}_i^{a'} \mathbf{H}_{i,j}'^T \mathbf{R}_i^{-1} ((\mathbf{y}_i - \bar{\mathbf{y}}) - (H(\hat{\mathbf{x}}_i^f) - \bar{\mathbf{m}})) \quad (41)$$

where  $\bar{\mathbf{y}}$  and  $\bar{\mathbf{m}}$  are time-mean altimetric sea level and its model simulation equivalent, respectively. This particular formulation corrects the model sea level variability without altering the model time-mean within the linearized time-asymptotic approximation. Such approximation is further sensible considering that errors in the time-mean state (bias) are due to time-correlated errors for linear models. Standard Kalman filtering and smoothing

formulations assume temporally uncorrelated process noise and such correlated model errors require modification to the canonical estimation procedure. Thus, assimilation of other observations (e.g., temperature profiles) is similarly restricted to their temporal anomalies. Time-invariant process noise can alternatively be estimated separately from such temporally uncorrelated errors.

For computational efficiency, Eq (41) is carried out from the right as a series of left multiplications (or operations) of the innovation vector (i.e., data-model difference) and its products; i.e., no matrix-matrix multiplication is performed to compute the coefficient matrix in Eq (41). Contributions from different partitions are summed together (Eq 38) to correct the entire model forecast. The unapproximated fully nonlinear model is then integrated in time using the resulting analysis with all diagnostic variables updated consistently with these data increments.

In terms of the partitioned reduced-state formulation, the smoother increment (difference between analysis and smoother; second term on the right hand side of Eq 11) can also be written as a sum of smoother increments in the partitioned reduced state;

$$\begin{pmatrix} \Delta \hat{\mathbf{x}}_t^s \\ \Delta \hat{\mathbf{u}}_t^s \end{pmatrix} = \sum_i \begin{pmatrix} \mathbf{B}_i \Delta \hat{\mathbf{x}}_{t,j}^{rs} \\ \tilde{\mathbf{B}}_i \Delta \hat{\mathbf{u}}_{t,j}^{rs} \end{pmatrix} \quad (42)$$

where,

$$\begin{pmatrix} \Delta \hat{\mathbf{x}}_{t,j}^{rs} \\ \Delta \hat{\mathbf{u}}_{t,j}^{rs} \end{pmatrix} = \begin{pmatrix} \mathbf{P}_i^{ra} \mathbf{A}_i^T \mathbf{P}_i^{rf-1} \\ \mathbf{Q}_i^r \mathbf{G}_i^T \mathbf{P}_i^{rf-1} \end{pmatrix} (\Delta \hat{\mathbf{x}}_{t+1,i}^{rs} + \Delta \hat{\mathbf{x}}_{t+1,i}^{ra}) \quad (43)$$

are the smoother increments of state and control of a particular partitioned reduced state. The partitioned form of the smoother increment recursion, Eq (43), uses the second form of the smoother gain in Eq (13) and the definitions of  $\Delta \hat{\mathbf{x}}^s$  and  $\Delta \hat{\mathbf{x}}^a$  to rearrange the last term in Eq (11). Unlike the filter, elements of the approximate smoother gain matrix in Eq (43) are time-invariant and thus, for computational efficiency, the gain matrix can be explicitly derived and used in deriving the smoother increments. Smoother increments of different partitions are summed together to correct the entire model state and control (Eq 42). However, because of the approximations, the resulting smoothed state and smoothed controls do not exactly satisfy the model equations and are, strictly speaking, physically inconsistent. Instead, estimates of a smoothed state that is fully consistent with the control are derived by re-integrating the model in time using the smoothed control estimates.

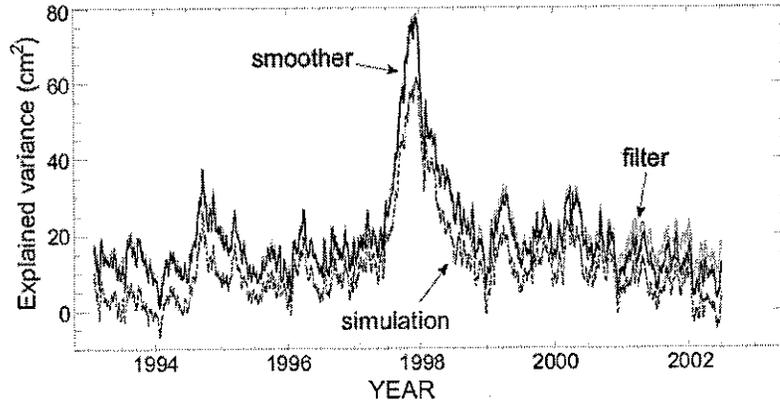


Figure 4. Model explained observed altimetric sea level anomaly variance; simulation (broken curve), Kalman filter (gray curve), smoother (solid curve). The explained variance is defined as the difference between data variance and model-data residual variance. Note the smoother results being nearly indistinguishable from the filter's (except near the end, 2001–2002) whereas the simulation's explained variance is substantially less than these throughout the experiment. Results are from the ECCO near real-time assimilation.

### 5.2.6 Assessment

The assimilation is assessed by examining its self-consistency and by comparisons with independent observations. Being a least-squares estimate, the estimates' errors are non-increasing functions of the amount and accuracy of the observations that are assimilated. An estimates' systematic degradation would indicate the assimilation's inaccurate assumption and/or errors in the implementation. Some examples of such assessment are briefly described below.

One of the useful and readily available measures for assessing assimilation is the difference between observations and their model equivalent, in particular, the innovation sequence (i.e., difference between observations and a filter's forecast). For instance, Figure 4 compares the amount of data variance (sea level) explained by the different model estimates. Explained variance is defined as,

$$\langle \mathbf{y}\mathbf{y}^T \rangle - \langle (\mathbf{y} - H(\mathbf{x}))(\mathbf{y} - H(\mathbf{x}))^T \rangle \quad (44)$$

The second term, the residual variance, is the variance of what the model cannot explain, and thus the difference with the data variance (first term) is a measure of what the model resolves. As the forecast does not yet utilize the particular observations, the innovation sequence also provides a measure of skill with respect to independent observations.

Figure 4 illustrates that the approximate Kalman filter explains significantly more data variance than does the model simulation without data

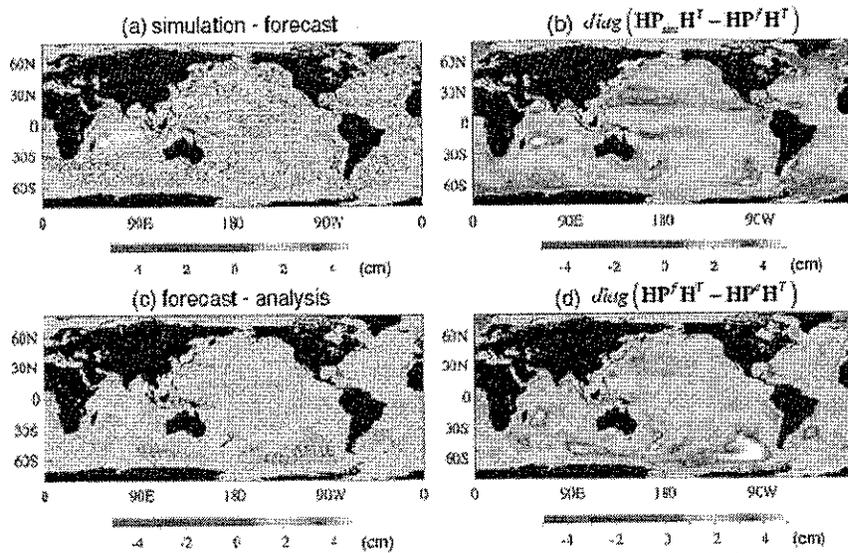


Figure 5. An assessment of model-data residuals with respect to their theoretical expectations. The panels show reductions in root-mean-square sea level residuals by assimilation of satellite altimeter data with a global ocean general circulation model. Panels (a) and (b) are differences between simulation and forecast and its theoretical expectation based on estimated errors, respectively; a positive value indicates an improvement by the latter model. Panels (c) and (d) are the same except between forecast and analysis. Note the first order consistency between (a) and (b) and between (c) and (d). Gray areas in (a) and (c) denote regions with no observations. Results correspond to assimilation using calibrated prior error estimates of Figure 2. (From Fukumori, *et al.*, 1999.)

constraints. Moreover, the figure shows that the smoothed estimate (smoothed-wind-driven model simulation) explains nearly as much variance as does the Kalman filter (model forecast), thus demonstrating the fidelity of the approximate smoother.

The assimilation's self-consistency can be assessed by comparing model-data differences with their theoretical expectations, i.e., formal error estimates computed and utilized by the Kalman filter algorithm. Figure 5 illustrates an example of such comparison. Differences between different model residuals are comparable to their respective theoretical expectations in both overall amplitude and spatial distribution. The absolute magnitudes of these estimates are statistically consistent, as the model-data difference of the simulation is also comparable with its theoretical expectation (Figure 2).

The fidelity of the assimilated analyses permits diverse studies and applications of not only ocean circulation (e.g., Fukumori *et al.*, 2004) but also of ocean biogeochemical processes (e.g., McKinley, 2002) and geodetic investigations (e.g., Dickey *et al.*, 2002.) For instance, Figure 6 illustrates such an application and an assessment of the data assimilated model estimate.

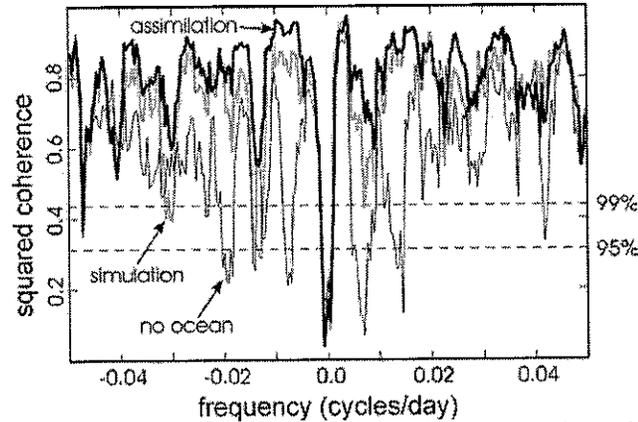


Figure 6. Coherence between observed and modeled excitation of Earth's wobble (polar motion); NCEP atmosphere reanalysis (thin black; "no ocean"), ECCO simulation plus NCEP atmosphere (gray; "simulation"), ECCO assimilation plus NCEP atmosphere (thick black; "assimilation"). Also shown are the 95% and 99% confidence levels. (Gross, 2003, personal communication. See Gross *et al.*, 2003, for related results.)

The figure shows coherence between observed excitation of Earth's polar motion (the wobble of Earth's rotation axis relative to the terrestrial frame) and that estimated by atmospheric (National Centers of Environmental Prediction (NCEP) Reanalysis, Kalnay *et al.*, 1996) and oceanic models. While changes in atmospheric circulation (thin black curve) account for most of the observed polar motion, adding the ocean estimate (gray curve) significantly improves the coherence at almost all frequencies. Moreover, the ocean assimilation (thick black curve) further improves the coherence illustrating the impact of ocean data assimilation in improving the estimate of ocean circulation. Satellite navigation employs estimates of polar motion and thus would benefit from forecasts as well as near real-time ocean analysis systems.

## 6. SUMMARY

Data assimilation concerns correcting models using observations. Although the concept is straightforward, there are various subtleties involved in both what data assimilation solves and how the computation is carried out. A careful understanding of these issues is helpful in assimilating observations, in utilizing their results, and in further improving their estimates.

Data assimilation can be considered a process of fitting models to observations. A solution that is consistent with both observations and model

physics is sought. However, given that all models are in one way or another approximations of the real world (ocean), there are some, sometimes many, aspects of the observations that are real but inconsistent with the models. These aspects that models cannot inherently simulate (representation errors) therefore cannot be part of the assimilated solution and must be properly accounted for. Forcing models to agree with such measurements can lead to increased inaccuracies and inconsistencies. An assessment of what models do and do not simulate is important in carrying out the assimilation, and an understanding of what the assimilated estimates resolve is fundamental to utilizing the results.

Mathematically, data assimilation is an inverse problem. The temporally evolving state of the model and sources of model error are estimated by inverting model equations that consist of those relating the model state to the observations and those describing the model's temporal evolution.

The Kalman filter and other common filtering methods are inversions of the model equivalent of the observations but not of the model evolution, and, therefore, do not completely solve the assimilation problem. Smoothers additionally invert the model evolution completing the estimation, providing estimates of both model state and model error sources.

While state estimation is often used synonymously with data assimilation, it is in fact the estimation of the model error sources (process noise) that is most fundamental. Given smoothed model error corrections, and apart from corrections to the initial condition, the smoothed state can be derived by integrating the model in time, but not vice versa.

Because of model errors, data assimilated state estimates by themselves are not physically consistent, in the sense that the estimated states' temporal evolution cannot be physically accounted for. For instance, budgets of heat and other properties cannot be closed in terms of explicit physical processes. The smoother's explicit estimation of model error sources resolves the physical inconsistency, rendering the assimilated solution amenable to various process studies and applications.

Although methods of data assimilation are well known, their implementation is often hampered by the models' large dimension and their complex nonlinearities. Many approximations have been put forth that render their implementation feasible and practical. The near real-time assimilation system of the Consortium for "Estimating the Circulation and Climate of the Ocean" (ECCO) employs a hierarchy of such approximations to maximize utilization of observations.

The fidelity and scope of these and other analyses lend themselves to various studies in ocean circulation and their application. However, existing products are in certain respects yet incomplete. The present near real-time ECCO estimates utilize a simplification by only estimating errors resulting from uncertainties in wind forcing. Other ECCO estimates also estimate

errors in diabatic forcing and uncertainties in some of the model parameters. However, there are many other model error sources that have not yet been addressed. Expanding the estimated suite of process noise remains a central task in further improving ECCO and other assimilation estimates.

For the approximate Kalman filter and RTS smoother, such extension requires an explicit modeling of the process noise that is physically sensible (identification of operator  $\mathbf{G}$  in Eq 1) and in identifying an effective approximation (partition and state reduction operators and basis set  $\mathbf{B}$  and  $\tilde{\mathbf{B}}$  in Eq 32) that would resolve the corresponding errors in the model state. An effective basis set not only has a small dimension but must also form a closed dynamic system (Eq 31 approximated as Eq 32). Understanding the nature of the modeled system is imperative to such design.

## ACKNOWLEDGMENTS

This study is a contribution of the Consortium for Estimating the Circulation and Climate of the Ocean (ECCO) funded by the National Oceanographic Partnership Program. This work was carried out in part at the Jet Propulsion Laboratory (JPL), California Institute of Technology, under contract with the National Aeronautics and Space Administration.

## REFERENCES

- Bretherton, F. P., R. E. Davis, and C. B. Fandry, 1976: A technique for objective analysis and design of oceanographic experiments applied to MODE-73, *Deep-Sea Res.*, **23**, 559-582.
- Cane, M. A., 1984: Modeling sea level during El Niño, *J. Phys. Oceanogr.*, **14**, 1864-1874.
- Cohn, S. E., 1997: An introduction to estimation theory, *J. Met. Soc. Japan*, **75**, 257-288.
- Dickey, J. O., S. L. Marcus, O. de Viron, and I. Fukumori, 2002. Recent Earth oblateness variations: Unraveling climate and postglacial rebound effects, *Science*, **298**, 1975-1977.
- Fu, L.-L., I. Fukumori and R. N. Miller, 1993: Fitting dynamic models to the Geosat sea level observations in the Tropical Pacific Ocean. Part II: A linear, wind-driven model, *J. Phys. Oceanogr.*, **23**, 2162-2181.
- Fukumori, I., J. Benveniste, C. Wunsch, and D. B. Haidvogel, 1993: Assimilation of sea surface topography into an ocean circulation model using a steady-state smoother, *J. Phys. Oceanogr.*, **23**, 1831-1855.
- Fukumori, I., and P. Malanotte-Rizzoli, 1995: An approximate Kalman filter for ocean data assimilation; an example with an idealized Gulf Stream model, *J. Geophys. Res.*, **100**, 6777-6793.
- Fukumori, I., R. Raghunath, L. Fu, and Y. Chao, 1999: Assimilation of TOPEX/POSEIDON data into a global ocean circulation model: How good are the results?, *J. Geophys. Res.*, **104**, 25,647-25,665.
- Fukumori, I., 2002. A partitioned Kalman filter and smoother, *Mon. Weather Rev.*, **130**, 1370-1383.

- Fukumori, I., T. Lee, B. Cheng, and D. Menemenlis, 2004. The origin, pathway, and destination of Niño3 water estimated by a simulated passive tracer and its adjoint, *J. Phys. Oceanogr.*, **34**, 582-604.
- Gross, R. S., I. Fukumori, and D. Menemenlis, 2003. Atmospheric and oceanic excitation of the Earth's wobbles during 1980-2000, *J. Geophys. Res.*, **108** (B8), 2370, doi:10.1029/2002JB002143.
- Kalnay, E., and coauthors, 1996. The NCEP/NCAR 40-year reanalysis project, *Bull. Amer. Meteorol. Soc.*, **77**, 437-471.
- Marshall, J. C., A. Adcroft, C. Hill, L. Perelman, and C. Heisey, 1997: A finite-volume, incompressible Navier Stokes model for studies of the ocean on parallel computers, *J. Geophys. Res.*, **102**, 5753-5766.
- McKinley, G. A., 2002. Interannual variability of air-sea fluxes of carbon dioxide and oxygen, Ph.D. thesis, Massachusetts Institute of Technology.
- Menemenlis, D., I. Fukumori, and T. Lee, 2004. Using Green functions to calibrate an ocean general circulation model, *Mon. Weather Rev.*, (in press).
- Paduan, J.D., and P.P. Niiler, 1993: Structure of velocity and temperature in the northeast Pacific as measured with Lagrangian drifters in Fall 1987, *J. Phys. Oceanogr.*, **23**, 585-600.
- Sasaki, Y., 1970: Some basic formalisms in numerical variational analysis, *Mon. Weather Rev.*, **98**, 875-883.
- Stammer, D., C. Wunsch, I. Fukumori, and J. Marshall, 2002: State Estimation in Modern Oceanographic Research, EOS, Trans. Amer. Geophys. Union, 83(27), 289&294-295.
- Wunsch, C., 1996: *"The Ocean Circulation Inverse Problem"*, Cambridge Univ. Press, New York, NY, 442pp.