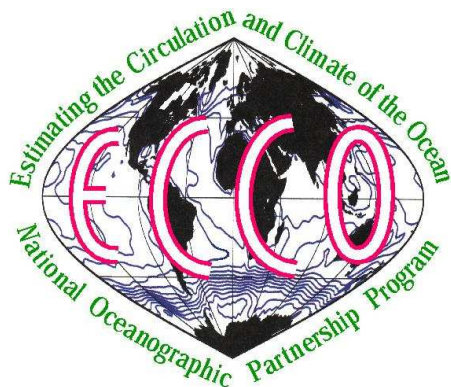


*The ECCO Report Series*¹

Budget Closures in Data Assimilation: Physical Consistency and Model Error Source Modeling

I. Fukumori²



Report Number 26

December 2003

¹ Estimating the Circulation and Climate of the Ocean (ECCO) is a consortium funded by a grant from the National Oceanographic Partnership Program (NOPP). Additional copies of this report are available at www.ecco-group.org, or contact D. Stammer (dstammer@ucsd.edu), I. Fukumori (if@pacific.jpl.nasa.gov), or J. Marshall (marshall@ocean.mit.edu).

² Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA 91109

Abstract

Data assimilation is a procedure that combines observations with models to improve descriptions of the modeled systems, such as the atmosphere and the ocean. Observations correct errors in models on the one hand, and models extrapolate data information in space and time on the other. Data assimilation is widely utilized in atmospheric sciences as well as in other disciplines of Earth Science, to help understand the workings of the respective systems. However, because of the data corrections, the temporal evolution described by these results are often physically inconsistent. For instance, heat and mass budgets cannot be closed, limiting the utility of the analyses. This article elucidates the nature of these inconsistencies and describes a solution to correct these problems. In particular, so-called smoothing algorithms can fix the deficiencies rendering model solutions physically consistent. Smoothing is half the data assimilation problem often ignored or forgotten in many applications. Smoothing is an inversion of the model that allows explicit estimation of model error sources. In comparison, sequential methods employed in most assimilation systems today are stationary inversions of the observations that correct the model state but not the sources of the error. This practice is historically rooted in data assimilation being developed in the context of weather forecasting as opposed to climate analyses that have different requirements. Identification of model error sources and their estimation by smoothing are fundamental to establishing a physically consistent estimate that is conducive to process studies of a temporally evolving dynamic system.

Capsule

Smoothing algorithms can fix problems in closing mass and heat budgets in data assimilation products.

Data Assimilation

Data assimilation is a procedure of combining observations with models, and is widely utilized in atmospheric sciences and related fields so as to improve descriptions of respective dynamic systems. However, the model-data combination often leaves imbalances in property budgets and other physical inconsistencies in the systems' estimated temporal evolution. This article discusses the nature of these problems and describes a solution to correct the deficiencies.

In data assimilation, observations correct errors in models on the one hand, and models extrapolate data information in space and time on the other. The combination can be regarded as least-squares averaging of observations and models. Consequently, results of assimilation are generally more complete and more accurate than those obtained by either observations or model simulations alone.

Data assimilation has long been employed in numerical weather forecasting as a means to optimize model initial conditions from which forecasts are initiated. In fact, data assimilation is often synonymous with model initialization using observations. Here, however, we consider data assimilation to mean a combination of models and observations in general.

Besides forecasting, data assimilated model results have also been utilized as a tool to analyze the nature of atmospheric circulation and to study the physical and dynamical processes that underlie the atmosphere's evolution. For instance, atmospheric re-analyses (e.g., Kalnay et al., 1996) provide coherent descriptions of the atmosphere that are utilized in diverse studies. The assimilations provide a summary description of the state of the atmosphere, freeing researchers from having to analyze individual observations in making inferences of the atmosphere and its changes.

More recently, data assimilation has also become an increasingly important element in oceanography (e.g., Stammer et al., 2002) as a tool to describe and to understand ocean circulation. The oceanographic interest is fueled in part by the recent increase in the amount of available observations and by a realization of the limitations of observing systems, vis-à-vis the complexity of ocean circulation, due to the sparse and incomplete nature of oceanographic measurements.

Physical (In)Consistency

Although data assimilation generally leads to improved accuracies, the utility of the assimilated estimates in help understanding dynamic systems is often limited, because of the estimates' physically inconsistent temporal evolution. This problem is described in Figure 1 that illustrates the temporal evolution of some element of a dynamic system (e.g., atmospheric surface pressure at some location) in a typical sequential data assimilation.

Sequential data assimilation progresses by first integrating the model forward in time, depicted by the red curve \widehat{ab} in Figure 1, from a point “a” at one instant (time t_1) to a point “b” at another time (time t_2). At the instant of point “b”, observations are available (point “o”), and the model simulated state “b” is corrected to be in closer agreement with these observations by assimilating the measurements. The correction is shown as the black line \overline{bc} and the corrected state is “c” which is usually not exactly in agreement with the observations “o” due in part to measurement errors. The procedure is then repeated from point “c” by integrating the model forward in time, carrying information obtained from the observations, until new measurements are available¹. The sequential correction \overline{bc} , in the terminology of estimation theory, is “filtering”, and its corrected solution is a filtered estimate.

Although the evolution \widehat{ab} is described by the modeled physics (e.g., advection, mixing, and external forcing), that of the data correction (\overline{bc}) is not. The data correction (\overline{bc}) corresponds to inaccuracies of the model state “b”. These inaccuracies are the consequence of errors in the model evolution \widehat{ab} , such as due to errors of the initial condition (point “a”) and errors in internal and external processes between “a” and “b”. However, the data correction itself is not explicitly described by such model error sources. Because of this ambiguity of what physical process data correction corresponds to, the modeled system's temporal evolution from the time of “a” to point “c” cannot be physically accounted for. For instance, budgets of mass and other properties

¹ In some systems, data information is assimilated at every model time-step by either interpolating observations in time or simply sampling the dynamic system as frequent as necessary. In such case, “a” and “b” describe the model state at consecutive time-steps and the discussion below applies equally to such systems.

cannot be closed between points “a” and “c”. The temporal evolution is thus physically inconsistent.²

The ambiguity of the data corrections limits the utility of data assimilated model solutions in understanding mechanisms that control the estimated temporal evolution and in deducing processes that underlie the general physical balance of the dynamic system. For instance, Figure 2 shows atmospheric surface pressure variability from an operational atmospheric analysis. The figure compares the equivalent of the model evolution between points “a” and “c” in Figure 1 (Figure 2a) with that of the data correction corresponding to the change from point “b” to “c” in the same figure (Figure 2b). Figure 2 shows that, on average, approximately 25% of the observed mass change in the atmosphere is not physically accounted for. Similarly, heat and other budgets cannot be closed and, consequently, identifying the cause of the changes in heat content and other properties is not straightforward in such analyses.

Smoothing

Resolving the assimilation’s physical inconsistency is to make sense of the data correction \overline{bc} in Figure 1. As stated, the data correction reflects inaccuracies in the model evolution \widehat{ab} . The solution, therefore, is to explicitly correct these error sources that \overline{bc} corresponds to. For instance, Trenberth et al. (1995) and others have discussed methods to ameliorate the mass imbalance in atmospheric analyses as in Figure 2 by choosing to adjust the velocity field during \widehat{ab} (e.g., advective divergence) while keeping other elements unchanged (e.g., surface pressure change itself). However, any particular choice in adjustment may violate other unconstrained budgets, such as that of heat, and degrade the overall accuracy of the model estimate.

² Most discussions of physical consistency in data assimilation have traditionally focused on the relationship among elements of the model state at a particular instant, such as geostrophic relationships between pressure and velocity and mapping data corrections to the slow manifold, etc. Such relationships among variables are typically sought so as to minimize “shocks” after models are initialized by the assimilated estimates. However, the consistency of temporal evolution that is considered here concerns the relationship between state estimates at different data-assimilated instances as opposed to that at a particular time.

The general solution to such adjustment lies in estimating the modeled system in its entirety so as to satisfy all relevant physical, dynamical, and observational constraints within their respective uncertainties. In particular, such estimation amounts to inverting the model evolution using the data correction \overline{bc} . Namely, the model’s temporal evolution algorithm can be regarded as a function of the initial condition (“a”), model physics and dynamics, and the various external forcings and boundary conditions. The model relates these independent quantities and processes to the end state “b” by an explicit mathematical relationship. Then, given an estimate of end state errors (\overline{bc}), this relationship can be inverted to correct inaccuracies in the independent variables that give rise to this model error, so that the model evolution is compatible with the data-corrected end state “c”. The result of such inversion is illustrated by the blue curve \widehat{dc} in Figure 1, that depicts adjustments to the model evolution and a corrected initial state “d”.

The inversion of the model evolution can be identified as “smoothing” in estimation theory. Smoothing is an inversion of the model in time that corrects model errors in the past using observations formally in the future. In comparison, filtering that was described previously (\overline{bc} in Figure 1) is an inversion of the observations, in particular, the theoretical relationship between the model state and the model equivalent of the observations. A simple mathematical exposition of these differences is provided in the Appendix. In particular, the common Kalman filtering and Rauch-Tung-Striebel smoothing can be identified as least-squares inversions of the observations and the model’s temporal evolution, respectively.

Filtered estimates carry data information forward in time as the model is temporally integrated. In comparison, the sequential smoother (estimation of \widehat{dc}) corrects these prior filtered estimates (“a”) and processes in between, using formally future observations (“o”); i.e., smoothed estimates (“d”) are based on both past and formally future observations. The use of additional observations result in smoothed estimates generally being more accurate than corresponding filtered results.

Smoothing is also an inversion across time whereas filtering is a static inversion of instantaneous fields. In fact, given “d”, the model at earlier times can be corrected similarly by extending the smoothing further back in time so that the evolution from these earlier times is also consistent with “d”, and in turn, “c”. Such estimate is illustrated by the extension of the blue curve \widehat{dc} to earlier instances ($t < t_1$). From its construction, note that such smoothed estimates are dependent on when the smoothing

was initiated (i.e., the end time). For instance, the blue dashed curve and blue circles depict a smoothed estimate initiated from some future instant beyond “c” ($t > t_2$).

In addition to the state, smoothing explicitly corrects other sources of model inaccuracies that dictate the model evolution, such as external forcings, model parameterization, etc. Because of these corrections, the temporal evolution of the smoothed estimate is physically consistent. For instance, the temporal evolution of the smoothed estimate is literally “smooth” as depicted in Figure 1. Unlike the filtered solution denoted by the red curves and black lines in Figure 1, the smoothed evolution along the blue solid curve and the blue dashed curve can be accounted for by explicit physical processes embodied in the model. Budgets of heat, mass, and other properties can be closed in correspondence to the model’s underlying physical principles.

Most discussions of smoothing focus on the improved accuracy of its results over those of filtering. However, the assimilated results’ consistent temporal evolution is a singular virtue of smoothing. In comparison, no matter how much observations are utilized to improve the accuracy of filtered estimates, filtered solutions, by construction, generally cannot be made to satisfy the model physics.

Note, however, that smoothing does not concern forecasting. Being an inversion of the model backwards in time, smoothing corrects model errors in the past but does not alter the terminal end state or any future estimate. This is depicted in Figure 1 by the end state “c” of the smoothed evolution \widehat{dc} being identical with the filtered estimate “c”. Forecasting that is initiated at time “c” using only observations up until that instant is not affected by smoothing \widehat{dc} that corrects the model in the past. This independence explains why smoothing has not been an issue in numerical weather forecasting. However, smoothing is necessary for climate analyses and/or other applications that require physically consistent descriptions of the dynamic system’s temporal evolution. In particular, closed budgets and other physical consistencies are conducive to analyzing the workings of dynamic systems.

Model Error Source Modeling

In the past, smoothing has largely been ignored in Earth Science, due in part to an historical emphasis on estimating the state, particularly in the context of initializing models for forecasting. Some progress has recently been made in the application of smoothing. For example, the Consortium for “Estimating the Circulation and Climate of the Ocean” (ECCO) has established a series of smoothed ocean data assimilation

analyses (e.g., Stammer et al., 2002. See also <http://www.ecco-group.org>). Satellite observations and in situ measurements are assimilated into a near global ocean general circulation model so as to study ocean circulation and its effect on other elements of the Earth system.

A key to achieving a physically consistent estimate is in explicit modeling and estimation (smoothing) of model error sources. This modeling concerns not so much their statistical properties but their explicit mathematical formulation. Model error sources do not solely consist of errors in initial condition but also inaccuracies in other elements of the model algorithm, including the model physics, model parameters, external forcings, etc. In Figure 1, the smoothed estimate \widehat{dc} consists not only of adjusting the initial condition “d”, but also of other elements that are independent of “d” affecting the model trajectory to “c”.

In ECCO, model error sources estimated to date include errors in prescribed atmospheric forcings (wind, heat flux, fresh-water flux), inaccuracies in model parameter values (e.g., mixing coefficients), and errors in the initial condition. Because their effect on model evolution are distinct (controllability; see, for example, Gelb, 1974), errors from different sources can generally be distinguished from each other, given adequate observations of the state’s temporal evolution; i.e., errors are not unduly misinterpreted.

However, there are other error sources in the ECCO model that have not yet been accounted for, such as errors caused by model numerics (finite differencing errors), inaccuracies due to inadequate resolution such as key passages in the model ocean, errors in water mass formation, etc. Owing to the use of comprehensive observations and smoothing algorithms, the present ECCO estimates are optimal and physically consistent, but because of these remaining undetermined errors, the estimates are yet incomplete. How much and how many types of model errors are corrected are as important, if not more, as how much observations are assimilated.

Data assimilated products are only accurate to the extent of what has been corrected by the assimilation. Understanding inherent limitations of these estimates, especially the physical inconsistency of filtered estimates vis-à-vis the consistency of smoothed estimates, is important when utilizing their respective products. Modeling and estimating model error sources as complete as possible is arguably one of the most important issues in advancing the fidelity and utility of data assimilated model products.

Acknowledgment

The author is grateful to Carl Wunsch and Tong Lee for helpful comments on an early version of this manuscript. 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 research was carried out in part at the Jet Propulsion Laboratory (JPL), California Institute of Technology, under contract with the National Aeronautics and Space Administration.

Appendix: Sequential Filtering and Smoothing as Least-Square Inversions of a Model

The mathematical structures of the filter and smoother help elucidate the nature of these respective assimilation methods. In particular, sequential filtering and smoothing algorithms can be identified as inversions of separate elements of the model that solve different parts of the data assimilation problem. We provide a simple mathematical review to illustrate these differences. The description also demonstrates the physical consistency of smoothed estimates and illustrates the significance of model error source modeling.

Mathematically, data assimilation is an inverse problem whereby the state of a dynamic system, \mathbf{x} , is estimated given a set of observations, \mathbf{y} , and a model; e.g.,

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

where \dots denote similar equations at different instances, t . 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. Vector \mathbf{u} denotes the model's controls and includes inhomogeneous terms of the model (e.g., boundary condition and forcing) and a representation of sources of model error (process noise). For simplicity, we assume a linear model in this discussion. The problem above and the solution 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 correspond to instances that observations are available.³

³ Model controls typically vary between times t and $t + 1$. Thus, while $\mathbf{x}_{(t)}$ denotes the model state at time t , $\mathbf{u}_{(t)}$ can be recognized as a concatenation of model controls during times $t \leq T < t + 1$.

The least-squares solution provides a general solution to linear inverse problems such as Eq (1). The least-squares solution, $\hat{\mathbf{a}}$ (the $\hat{\cdot}$ denotes an estimate), for a general linear inverse problem,

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

given \mathbf{b} is

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

(e.g., Lawson and Hanson, 1974). \mathbf{a}_0 is a prior estimate of \mathbf{a} , and \mathbf{R}_{aa} and \mathbf{R}_{bb} are prior error covariance matrices of \mathbf{a}_0 and \mathbf{b} , respectively, that are independent of each other. (The latter also includes uncertainties in the representativeness of Eq (2).)

Sequential filtering and smoothing algorithms can be recognized as least-squares estimations of the form Eq (3). For instance, the statistically optimal Kalman filter⁴ (e.g., Gelb, 1974) corrects model forecasts $\hat{\mathbf{x}}_{(t,-)}$ with observations to yield an analysis (i.e., corrected estimate) $\hat{\mathbf{x}}_{(t)}$ such that,

$$\hat{\mathbf{x}}_{(t)} = \hat{\mathbf{x}}_{(t,-)} + \mathbf{P}_{(t,-)}\mathbf{H}^T(\mathbf{H}\mathbf{P}_{(t,-)}\mathbf{H}^T + \mathbf{R}_{(t)})^{-1}(\mathbf{y}_{(t)} - \mathbf{H}\hat{\mathbf{x}}_{(t,-)}) \quad (4)$$

The minus sign in the argument denotes estimates prior to assimilation at the particular instant. $\mathbf{P}_{(t,-)}$ and $\mathbf{R}_{(t)}$ are error covariance matrices of $\hat{\mathbf{x}}_{(t,-)}$ and $\mathbf{y}_{(t)}$, respectively, where the latter also includes representation errors (Cohn, 1997) of the observation equations in (1) (upper half of 1). The model forecast $\hat{\mathbf{x}}_{(t,-)}$ is in turn estimated by integrating the model from the analysis at a previous instant $\hat{\mathbf{x}}_{(t-1)}$; viz.,

$$\hat{\mathbf{x}}_{(t,-)} = \mathbf{A}\hat{\mathbf{x}}_{(t-1)} + \mathbf{G}\mathbf{u}_{(t-1)} \quad (5)$$

The filtering algorithm is schematically described in Figure 1 that illustrates the temporal evolution of some element of a dynamic system. The state estimates $\hat{\mathbf{x}}_{(t-1)}$,

⁴ Other common sequential assimilation methods such as direct insertion, nudging, optimal interpolation, and 3dVAR are statistically suboptimal variants of the Kalman filter. These other assimilation schemes utilize ad hoc weights or prescribed state error covariance matrices (\mathbf{P} in Eq 4) instead of solutions of the Riccati equation used in Kalman filtering. The latter evaluates model error evolution taking into account the nature of the model error source and the relative dynamical and statistical effects of model physics and observations. Explicit modeling of model error sources is essential for achieving physical consistency later by smoothing as described below by Eq 9.

$\hat{\mathbf{x}}_{(t,-)}$, and $\hat{\mathbf{x}}_{(t)}$ respectively correspond to states represented by points “a”, “b”, and “c” in Figure 1. The two relationships, Eqs (5) and (4) describe the red and black segments \widehat{ab} and \overline{bc} , respectively, in the same figure.

Comparison of Eq (4) with Eq (3) shows that the Kalman filter is nothing but a least-squares inversion of the observation equation in Eq (1)⁵. At the same time, this correspondence also indicates that the lower half of Eq (1) is not solved (inverted) by the Kalman filter algorithm; i.e., Kalman filters only solve half the assimilation problem. In fact, substituting Eq (5) in (4) gives,

$$\hat{\mathbf{x}}_{(t)} = \mathbf{A}\hat{\mathbf{x}}_{(t-1)} + \mathbf{G}\mathbf{u}_{(t-1)} + \mathbf{K}(\mathbf{y}_{(t)} - \hat{\mathbf{x}}_{(t,-)}) \quad (6)$$

where, for brevity, we define $\mathbf{K} \equiv \mathbf{P}_{(t,-)}\mathbf{H}^T(\mathbf{H}\mathbf{P}_{(t,-)}\mathbf{H}^T + \mathbf{R}_{(t)})^{-1}$ (Kalman filter). Eq (6) is different from the lower half of (1) due to the third term on the right-hand-side of Eq (6). This third term corresponds to the data correction \overline{bc} in Figure 1, and is not explicitly part of the model physics (a function of \mathbf{A} or \mathbf{G}) that describes the model’s temporal evolution.

The physical inconsistency in (6) can be resolved by recognizing that the lower half of the inverse problem (1) is not yet solved. Namely, given the updated estimate $\hat{\mathbf{x}}_{(t)}$ by (4), the lower half of (1) can be identified as defining another inverse problem for (re-)estimating $\mathbf{x}_{(t-1)}$ and $\mathbf{u}_{(t-1)}$. Namely,

$$\hat{\mathbf{x}}_{(t)} = (\mathbf{A} \quad \mathbf{G}) \begin{pmatrix} \mathbf{x}_{(t-1)} \\ \mathbf{u}_{(t-1)} \end{pmatrix} \quad (7)$$

which is another inverse problem of form (2) where the coefficient matrix is $(\mathbf{A} \quad \mathbf{G})$ and the unknown vector is $(\mathbf{x}_{(t-1)}^T \mathbf{u}_{(t-1)}^T)^T$.

Eq (7) can also be solved by least-squares. In particular, an exact solution can be sought that satisfy mass conservation, etc, because model error sources \mathbf{u} are explicitly included in the formulation. This amounts to setting $\mathbf{R}_{bb} = 0$ in Eq (2). The filtered estimate $\hat{\mathbf{x}}_{(t-1)}$ and the a priori control $\mathbf{u}_{(t-1)}$ provide the prior solutions to (7), and their error covariance matrix defines the equivalent of \mathbf{R}_{aa} ,

$$\begin{pmatrix} \mathbf{P}_{(t-1)} & 0 \\ 0 & \mathbf{Q}_{(t-1)} \end{pmatrix}. \quad (8)$$

⁵ This is slightly an oversimplification, as the Kalman filter algorithm also describes the time-evolution and the derivation of the model state error covariance matrix, \mathbf{P} in Eq 4, in addition to the state estimate discussed here.

$\mathbf{Q}_{(t-1)}$ is the prior error covariance matrix of $\mathbf{u}_{(t-1)}$ (process noise). As in the standard formulation of the Kalman filter, we assume errors in $\mathbf{u}_{(t-1)}$ to be uncorrelated in time and, therefore, uncorrelated with errors in the estimate $\hat{\mathbf{x}}_{(t-1)}$, resulting in the off-diagonal blocks in (8) being zero. (Note that \mathbf{P} in (4) and (8) are functions of \mathbf{Q} as described by the Kalman filter algorithm.)

Then substituting (8) in (3) along with other elements equivalent to those of (2) we have,

$$\begin{aligned} \begin{pmatrix} \hat{\mathbf{x}}_{(t-1,+)} \\ \hat{\mathbf{u}}_{(t-1,+)} \end{pmatrix} &= \begin{pmatrix} \hat{\mathbf{x}}_{(t-1)} \\ \mathbf{u}_{(t-1)} \end{pmatrix} + \begin{pmatrix} \mathbf{P}_{(t-1)}\mathbf{A}^T (\mathbf{A}\mathbf{P}_{(t-1)}\mathbf{A}^T + \mathbf{G}\mathbf{Q}_{(t-1)}\mathbf{G}^T)^{-1} \\ \mathbf{Q}_{(t-1)}\mathbf{G}^T (\mathbf{G}\mathbf{Q}_{(t-1)}\mathbf{G}^T + \mathbf{A}\mathbf{P}_{(t-1)}\mathbf{A}^T)^{-1} \end{pmatrix} \\ &\quad \times (\hat{\mathbf{x}}_{(t)} - \mathbf{A}\hat{\mathbf{x}}_{(t-1)} - \mathbf{G}\mathbf{u}_{(t-1)}) \end{aligned} \quad (9)$$

The plus in the argument on the left hand side of (9) denotes estimates that utilize formally future observations (as contained in $\hat{\mathbf{x}}_{(t)}$ by 4) as opposed to filtered estimates (Eq 4) and model forecasts (Eq 5). The smoother’s assimilation of additional observations results in improved accuracy in its state estimate over that of the filtered solution.

Previous filtered estimates $\hat{\mathbf{x}}_{(t-2)}$ and $\hat{\mathbf{u}}_{(t-2)}$ can also be improved by inverting the equivalent of Eq (7) using the estimate $\hat{\mathbf{x}}_{(t-1,+)}$ of Eq (9). By induction, other filtered estimates at earlier times can be improved by such inversion recursively back in time.

Equation (9) and the recursive equation it defines can be recognized as the Rauch-Tung-Striebel (RTS) fixed-interval smoother, and the improved solutions are its smoothed estimates (e.g., Bryson and Ho, 1975). Note the correspondence between (9) and (3) that illustrates that smoother estimations of $\hat{\mathbf{x}}_{(t-1,+)}$ and $\hat{\mathbf{u}}_{(t-1,+)}$ are inversions of the model operators \mathbf{A} and \mathbf{G} , respectively, that describe the temporal evolution of the model.

The temporal evolution of such smoothed estimate is depicted by the blue solid curve in Figure 1. Filtered state estimates $\hat{\mathbf{x}}_{(t-1)}$ and $\hat{\mathbf{x}}_{(t)}$ and smoothed estimate $\hat{\mathbf{x}}_{(t-1,+)}$, respectively, correspond to points “a”, “c”, and “d”.⁶

⁶ Smoothed state estimates intervening “c” and “d” can be obtained by solving the equivalent of Eq (9) in smaller time increments. Alternatively, the same can be obtained by the equivalent of Eq (10), integrating the model forward in time using smaller time increments and intervening smoothed control estimates.

By construction, unlike the Kalman filter estimate, the smoothed solution given by (9), and the recursively smoothed estimates at earlier instances, satisfy the model equations in (1) (lower half of (1)). In particular,

$$\hat{\mathbf{x}}_{(t',+)} = \mathbf{A}\hat{\mathbf{x}}_{(t'-1,+)} + \mathbf{G}\hat{\mathbf{u}}_{(t'-1,+)} \quad (10)$$

for $t' < t$. This is also obvious by substituting the equivalent of Eq (9) at time $t' - 1$ on the right hand side of (10), noting $\hat{\mathbf{x}}_{(t)}$ on the right-hand-side of (9) corresponds to $\hat{\mathbf{x}}_{(t',+)}$ for $t' < t$.

Note that for Eq (10) to be *physically consistent*, the estimated error source $\hat{\mathbf{u}}$ and its model \mathbf{G} must be chosen sensibly in Eq (1). For instance, errors corresponding to external forcing imply certain properties for \mathbf{G} (and, therefore, \mathbf{Q} in Eq 8) and those associated with mixing (e.g., depth of ocean surface mixed layer) require others. In particular, assuming uncorrelated white noise in state space (i.e., $\mathbf{G}\mathbf{u} = \delta$, where individual elements of δ is uncorrelated white noise), such as for temperature in the interior of the ocean, would imply internal sources and sinks that is not physically sensible. Sensible modeling of model errors is important for achieving a physically consistent smoothed estimate.

Although smoothed solutions can satisfy model equations (10), smoothing should not be confused with so-called “strong constraint” estimation (Sasaki, 1970) that assumes that models have no errors except in initial condition. 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 (difference between $\hat{\mathbf{u}}$ and \mathbf{u} in Eq 9), as opposed to leaving them unknown, is what allows for the temporal evolution of the solution to be physically consistent.

In fact, smoothed state estimates alone (upper half of Eq 9) do not generally satisfy model equations;

$$\hat{\mathbf{x}}_{(t',+)} \neq \mathbf{A}\hat{\mathbf{x}}_{(t'-1,+)} + \mathbf{G}\mathbf{u}_{(t'-1)} \quad (11)$$

However, given smoothed control estimates ($\hat{\mathbf{u}}_{(t,+)}$ for all t) and the smoothed initial condition ($\hat{\mathbf{x}}_{(0,+)}$ where $t = 0$ is initial time), smoothed state estimates at other instances could be derived sequentially in time ($t > 0$) by simply using the forward model (10). Thus, it could be argued that smoothed control estimates $\hat{\mathbf{u}}_{(t,+)}$ are more

fundamental than smoothed state estimates $\hat{\mathbf{x}}_{(t,+)}$, because the latter can be derived from the former using the model, but not vice versa.⁷

Although the discussion above has focused on sequential methods of 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 4dVAR⁸ is equivalent to the RTS smoother (Eq 9). The adjoint and 4dVAR estimation directly solve for the smoothed solution without deriving the intermediate filter estimates.

⁷ The terminology, “state estimation”, is often used synonymously with “data assimilation”. However, model controls are generally not directly part of the model state. Smoothing’s model error source estimation ($\hat{\mathbf{u}}$ in 9) could be better identified as “control estimation”.

⁸ The present generation of 4dVAR implemented at the European Centre for Medium-Range Weather Forecasts (ECMWF) (Rabier et al., 2000) is a fixed-lag smoother that does not allow for model error sources (controls) except for errors in the initial condition.

References

- Bryson, A. E., Jr. and Y.-C. Ho, 1975. "Applied Optimal Control", Revised Printing. Hemisphere, New York, 481pp.
- Cohn, S. E., 1997. An introduction to estimation theory, *Journal of the Meteorological Society of Japan*, **75**, 257–288.
- Gelb, A., 1974. "Applied Optimal Estimation", M.I.T. Press, Cambridge, MA 374 pp.
- Kalnay, E., and coauthors, 1996. The NCEP/NCAR 40-year reanalysis project, *Bulletin of the American Meteorological Society*, **77**, 437–471.
- Lawson, C. L., and R. J. Hanson, 1974. "Solving Least-Squares Problems", Prentice-Hall, Englewood Cliffs, NJ, 340pp.
- Rabier, F., H. Järvinen, E. Klinker, J.-F. Mahfouf, and A. Simmons, 2000. The ECMWF operational implementation of four-dimensional variational assimilation. I: Experimental results with simplified physics, *Quarterly Journal of the Royal Meteorological Society*, **126A**, 1143–1170.
- Sasaki, Y., 1970. Some basic formalisms in numerical variational analysis, *Monthly Weather Review*, **98**, 875–883.
- Stammer, D., C. Wunsch, I. Fukumori, and J. Marshall, 2002: State Estimation in Modern Oceanographic Research, EOS, Transactions, American Geophysical Union, 83(27), 289&294-295.
- Trenberth, K. E., J. W. Hurrell, and A. Solomon, 1995. Conservation of mass in three dimensions in global analyses, *Journal of Climate*, **8**, 692–708.

Figure Legends

Figure 1: Schematic of a state element’s temporal evolution in a typical sequential data assimilation. Abscissa is time and ordinate is the state’s value. The red and black symbols and curves illustrate filtered estimation. The estimation progresses by integrating the model from an initial condition (red cross denoted “a”) to another (black cross denoted “b”), when a set of observations (triangle denoted “o”) becomes available and is used to correct the model state (red cross denoted “c”). The procedure is then repeated until the next set of data are available. The blue curves and symbols describe smoothed estimates that utilize all observations within a given period, to correct not only the model state but also sources of model error. The blue solid curve employs data up until time t_2 . The blue dashed curve illustrates a smoothed estimate using data beyond the temporal interval that is shown.

Figure 2: Average variation of atmospheric surface pressure; a) changes between analyses over 6-hour periods, b) 6-hourly data corrections. Variations are standard deviations based on the operational analyses and forecasts of atmospheric surface pressure by the National Centers for Environmental Prediction (NCEP) from 29 May to 6 December 2001. Units are in mbar. Global area weighted averages are 6 mbar and 1.5 mbar for (a) and (b), respectively.

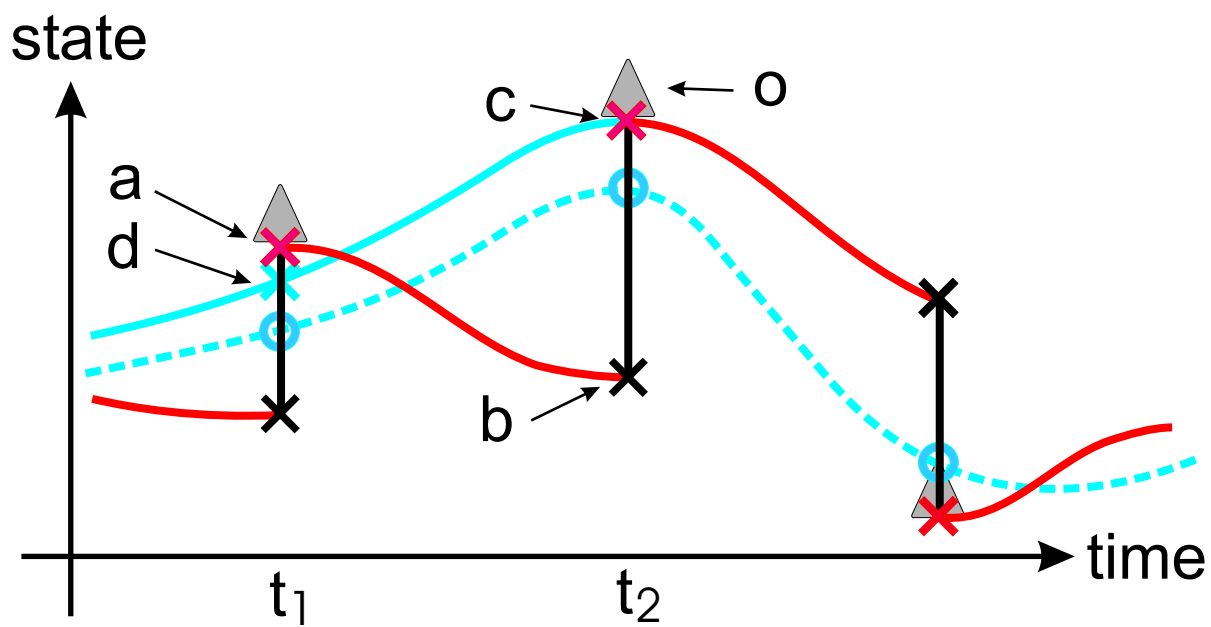
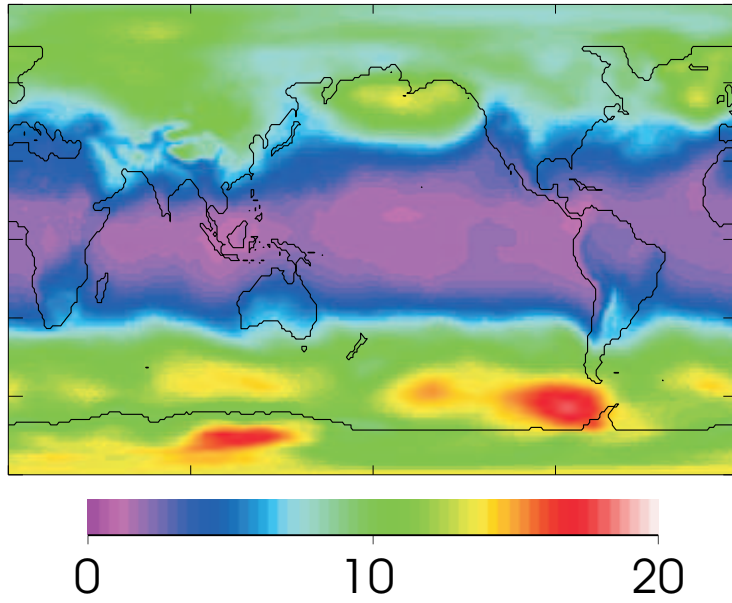


Fig 1

(a) Temporal Variability



(b) Data Corrections

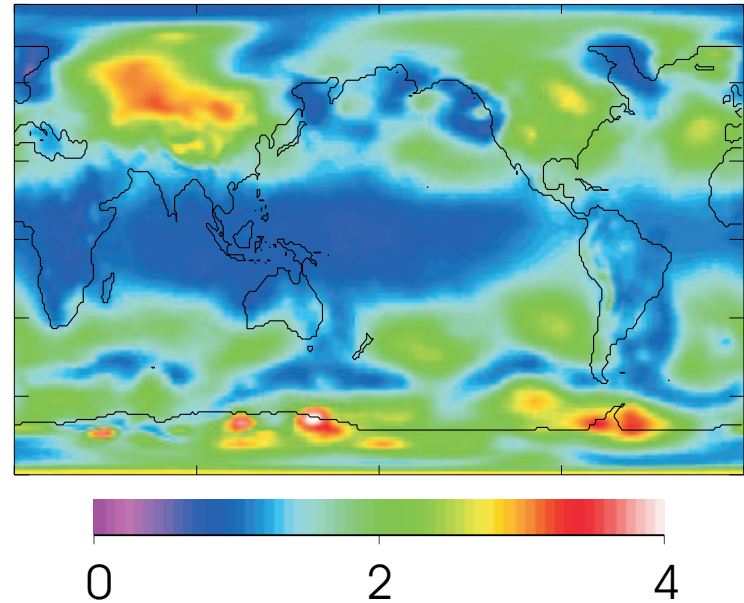


Fig 2