Background Error Covariance Estimation using Information from a Single Model Trajectory with Application to Ocean Data Assimilation

Christian L. Keppenne[1, 2] (christian.keppenne@nasa.gov)
Michele M. Rienecker[2] (michele.rienecker@nasa.gov)
Robin M. Kovach[1, 2] (robin.m.kovach@nasa.gov)
Guillaume Vernieres[1, 2] (guillaume.vernieres@nasa.gov)

[1]Science Systems and Applications Inc.
10210 Greenbelt Road, Suite 600
Lanham, Maryland 20706, USA

[2]Global Modeling and Assimilation Office
Code 610.1 , NASA Goddard Space Flight Center
Greenbelt, Maryland 20771, USA

Corresponding author:
Christian Keppenne
email: christian.keppenne@nasa.gov
telephone: 011-1-301-6145874
mail: Code 610.1. NASA Goddard Space Flight Center, Greenbelt, Maryland 20771, USA

Background Error Covariance Estimation using Information from a Single Model Trajectory with Application to Ocean Data Assimilation

Christian L. Keppenne[1,2], Michele M. Rienecker[2], Robin M. Kovach[1,2], and Guillaume Vernieres[1,2]

[1]Science Systems and Applications Inc.
10210 Greenbelt Road, Suite 600
Lanham, Maryland 20706, USA

[2]Global Modeling and Assimilation Office
Code 610.1 , NASA Goddard Space Flight Center
Greenbelt, Maryland 20771, USA

Abstract

An attractive property of ensemble data assimilation methods is that they provide flow dependent background error covariance estimates which can be used to update fields of observed variables as well as fields of unobserved model variables. Two methods to estimate background error covariances are introduced which share the above property with ensemble data assimilation methods but do not involve the integration of multiple model trajectories. Instead, all the necessary covariance information is obtained from a single model integration. The Space Adaptive Forecast error Estimation (SAFE) algorithm estimates error covariances from the spatial distribution of model variables within a single state vector. The Flow Adaptive error Statistics from a Time series (FAST) method constructs an ensemble sampled from a moving window along a model trajectory.

SAFE and FAST are applied to the assimilation of Argo temperature profiles into version 4.1 of the Modular Ocean Model (MOM4.1) coupled to the GEOS-5 atmospheric model and to the CICE sea ice model. The results are validated against unassimilated Argo salinity data. They show that SAFE and FAST are competitive with the ensemble optimal interpolation (EnOI) used by the Global Modeling and Assimilation Office (GMAO) to produce its ocean analysis. Because of their reduced cost, SAFE and FAST hold promise for high-resolution data assimilation applications.

Keywords:
Data assimilation; error covariance; Kalman filter; ensemble Kalman filter; ensemble optimal interpolation

1. Introduction

Following a seminal paper by Evensen (1994) introducing the ensemble Kalman filter (EnKF), ensemble data assimilation (EDA) methods have gained wide acceptance and usage in the geophysical sciences. While EDA methods differ in terms of the approach used to update or resample the ensemble of model states, they all require an ad hoc number of concurrent model integrations to estimate the distribution of background errors. This approach is essentially an $O(\underline{n})$ procedure, where n is the size of the model state vector. In contrast, the original Kalman (1960) filter algorithm propagates its background error covariance estimates by means of matrix multiplications of $O(\underline{n}^3)$. Hence, EDA methods are comparably economical from a numerical standpoint. Yet, their cost is significantly higher than that of conventional methods that do not involve ensemble model integrations. Thus, implementations of EDA methods must compromise between ensemble size and model resolution.

Because the analysis and error estimates depend on the state of each ensemble member, EDA methods are flow-adaptive. They also provide estimates of the cross-field covariance between observed and unobserved model fields that can be used to update unobserved system variables. For example, ocean sub-surface fields can be updated even if only surface observations are available.

The purpose of this paper is to introduce two data assimilation algorithms that share the abovementioned properties of EDA methods but, unlike EDA methods, rely on only a single model trajectory to estimate the necessary error-covariance information. As such, these methods obviate the requirement to compromise between ensemble size and model resolution. The Space Adaptive Forecast-error Estimation (SAFE) algorithm estimates error covariances from the spatial distribution of model variables in the neighborhood of every model grid cell in a single background state. Rather, the Flow Adaptive error Statistics from a Time series (FAST) algorithm estimates covariances from the recent distribution of high-pass filtered lagged instances of the model state vector sampled along the same trajectory. Because they do not require multiple integrations of the numerical model, SAFE and FAST are considerably less resource hungry than typical EDA methods and thus hold promise for high-resolution data assimilation applications.

The underlying assumption on which SAFE and FAST are based is that errors in the forecasts used in assimilation are primarily phase errors in space and/or time. For the ocean, this assumption makes sense as the dominant source of error can be related to errors in surface forcing, *i.e.,* the timing, intensity, or location of particular atmospheric synoptic events. Thus, the forecast (or background) errors can be related to the timing or intensity in the propagation or advection of oceanic anomalies.

The algorithms are outlined in Section 2 and compared to conventional assimilation techniques in Section 3 where they are applied to the assimilation of Argo temperature (T) profiles into the OGCM component of the NASA Global Modeling and Assimilation Office (GMAO) Goddard Earth Observing System (GEOS). Unassimilated Argo salinity (S) observations are used to validate the assimilation. Conclusions follow in Section 4.

2. Assimilation Algorithms

2.1 Preamble

Most sequential data assimilation algorithms are inspired by or derived from the Kalman filter (Kalman 1960) and involve the following steps,

$$\begin{aligned} \boldsymbol{x}_k^f &= \boldsymbol{M}(\boldsymbol{x}_{k-1}^a, \boldsymbol{f}_{k-1}), \quad (1a) \\ \boldsymbol{y}_k = \boldsymbol{H}_k(\boldsymbol{x}_k^t) + \boldsymbol{\varepsilon}_k, &\qquad E(\boldsymbol{\varepsilon}_k \boldsymbol{\varepsilon}_k^T) = \boldsymbol{R}_k, \quad (1b) \\ \boldsymbol{K}_k &= \boldsymbol{P}_k^f \boldsymbol{H}_k^T [\boldsymbol{H}_k \boldsymbol{P}_k^f \boldsymbol{H}_k^T + \boldsymbol{R}_k]^{-1}, \quad (1c) \\ \boldsymbol{x}_k^a &= \boldsymbol{x}_k^f + \boldsymbol{K}_k [\boldsymbol{y}_k - \boldsymbol{H}_k(\boldsymbol{x}_k^f)], \quad (1d) \end{aligned}$$

where the subscript k refers to the kth of a sequence of assimilations, $\boldsymbol{x}^f$ and $\boldsymbol{x}^a$ denote the model forecast and analyzed states, $\boldsymbol{M}$ is the model operator, and $\boldsymbol{f}_{k-1}$ represents the forcing between times t_{k-1} and t_k. The observations, $\boldsymbol{y}_k$, assimilated at time t_k are related to the true system state, $\boldsymbol{x}^t$, at time t_k by equation (1b) where $\boldsymbol{H}_k$ is the observation operator, E denotes the expectation operator and $\boldsymbol{\varepsilon}_k$, with covariance matrix $\boldsymbol{R}_k$, is the observation error vector. The Kalman gain matrix, $\boldsymbol{K}_k$, dictates how the observations and model forecast are weighted in the analysis computation (equation 1d). It depends on $\boldsymbol{H}_k$, $\boldsymbol{R}_k$ and the background error covariance matrix,

$$\boldsymbol{P}_k = E((\boldsymbol{x}^t - \boldsymbol{x}_k^f)(\boldsymbol{x}^t - \boldsymbol{x}_k^f)^T). \quad (2)$$

Of course, since $\boldsymbol{x}^t$ is unknown, $\boldsymbol{P}_k$ cannot be computed directly from equation (2) and must be estimated, either explicitly or implicitly, by some other means. In fact, the procedure used to estimate $\boldsymbol{P}_k$ can be used to classify data assimilation methods.

In most EDA methods, $\boldsymbol{P}_k$ is estimated from the statistical distribution of an ensemble of model forecasts,

$$\boldsymbol{x}_{i,k}^f = \boldsymbol{M}(\boldsymbol{x}_{i,k-1}^a, \boldsymbol{f}_{i,k-1}), \qquad i = 1, \ldots, n, \quad (3)$$

started from an ensemble of n analyzed model states at the previous analysis time, t_{k-1}. Following Houtekamer and Mitchell (2001), many EDA systems filter spurious long-range covariances resulting from finite ensemble sizes by (dropping the k subscript) decomposing $\boldsymbol{P}$ as

$$\boldsymbol{P} = \boldsymbol{P}_e \bullet \boldsymbol{C}, \quad (4)$$

where $\boldsymbol{P}_e$ represents the background covariances estimated from the ensemble of model states, $\boldsymbol{C}$ is a compactly supported correlation matrix and $\bullet$ denotes the Schur (i.e., element by element) product of two matrices.

In a class of methods known alternatively as ensemble optimal interpolation (EnOI: *e.g.,* Borovikov *et al.* 2005; Oke *et al.* 2005, 2010; Wan *et al.* 2010; Vernieres *et al.* 2012) or asymptotic ensemble filters, the time dependency is neglected and $\boldsymbol{P}$ is estimated from the statistics of one or more model run histories or from combinations of model histories. In many cases, EnOI methods are competitive with the flow-dependent EDA methods because they make up for the performance degradation due to neglecting the forecast-error evolution by estimating error statistics from a much larger ensemble.

Optimal interpolation (OI: Eliassen 1954) refers to an older class of data assimilation methods in which background error covariances are modeled with Gaussian functions or other analytically or empirically derived functions. Cross-field covariances are generally neglected in these methods and only the model field corresponding to the observed variable is updated.

2.2 Space Adaptive Forecast error Estimation (SAFE)

The SAFE algorithm attempts to combine the simplicity and cost effectiveness of OI with the large sample size of EnOI and the flow dependency of the EnKF. It estimates background error covariances by treating the state variables in neighboring grid cells surrounding every model grid point as if they were the state variables of other ensemble members at the same grid point. Because the size of the neighborhood determines the covariance amplitudes, rescaling is necessary. Note however that an error-covariance rescaling step is also implicitly present in many EDA methods where the background error covariance amplitude is determined by parameters of a covariance inflation procedure.

To facilitate the procedure in geophysical fluid models with complicated boundaries, the following algorithm is used. For simplicity of notation, we assume that the model state can be split according to

$$\boldsymbol{x} = [\boldsymbol{v}, \boldsymbol{w}], \quad \boldsymbol{P} = \begin{bmatrix} \boldsymbol{P}^{vv} & \boldsymbol{P}^{vw} \\ \boldsymbol{P}^{wv} & \boldsymbol{P}^{ww} \end{bmatrix}, \quad (5)$$

where $\boldsymbol{v}$ is an observed model field and $\boldsymbol{w}$ is unobserved. The generalization to more than two model fields is obvious. We also assume that all the data assimilated correspond to the same model quantity although the generalization to different observation types is also straightforward. In view of the above, the model update is split according to

$$\boldsymbol{v}^a = \boldsymbol{v}^f + \underbrace{\boldsymbol{P}^{vv}\boldsymbol{H}^T\left[\boldsymbol{H}\boldsymbol{P}^{vv}\boldsymbol{H}^T + \boldsymbol{R}\right]^{-1}\left[\boldsymbol{y} - \boldsymbol{H}(\boldsymbol{v}^f)\right]}_{\Delta\boldsymbol{v}}. \qquad (6a)$$

$$\boldsymbol{w}^a = \boldsymbol{w}^f + \boldsymbol{P}^{wv}\boldsymbol{H}^T\left[\boldsymbol{H}\boldsymbol{P}^{vv}\boldsymbol{H}^T + \boldsymbol{R}\right]^{-1}\left[\boldsymbol{y} - \boldsymbol{H}(\boldsymbol{v}^f)\right], \qquad (6b)$$

$$= \boldsymbol{w}^f + \boldsymbol{P}^{wv}\left(\boldsymbol{P}^{vv}\right)^{-1}\Delta\boldsymbol{v}. \qquad (6c)$$

The application of equation (6c) is further simplified by assuming that the w background error in grid cell (i, j, k) is predominantly related to the v error in grid cell (i, j, k) and negligibly related to the v errors in other grid cells, thus neglecting the off diagonal elements of $\boldsymbol{P}^{vv}$ in (6c). Instead the unobserved model field is updated according to

$$\boldsymbol{w}^{a}_{ijk} = \boldsymbol{w}^{f}_{ijk} + \frac{P^{wv}_{ijk}}{P^{vv}_{ijk}} \Delta \boldsymbol{v}_{ijk}, \quad i=1,\cdots,I, \quad j=1,\cdots,J, \quad k=1,\cdots,K, \qquad (6d)$$

where I, J and K denote the number of grid cells along the x, y, and z space dimensions, respectively. Heuristically, these simplifications are related to the assumption that if a and b are correlated and b and c are correlated, then a and c are correlated.

The first step is to estimate the background error variance of the observed field (the procedure is the same regardless whether this field is prognostic or diagnostic) with

$$\boldsymbol{\sigma}^2_{vv} = diag(\boldsymbol{P}^{vv}) = \Theta([\boldsymbol{v} - \Theta(\boldsymbol{v})]^2), \qquad (7)$$

where Θ is a local 3D averaging operator. For our implementation, repetitive application of a gridpoint (spatial) Laplacian smoother was found to be effective. The results of Section 3 (Figure 1) indicate that the size of the regions over which the averaging is applied is of little consequence.

The variance estimate is rescaled such that

$$\left\| diag(\boldsymbol{H}\boldsymbol{P}^{vv}\boldsymbol{H}^{t}) \right\| = \gamma^2 \left\| diag(\boldsymbol{R}) \right\|, \quad (8)$$

where the double vertical bar stands for an L2 vector norm. The parameter γ is prescribed. It is a scalar representing the global mean (asymptotic) target ratio of background error variances to data error variances and its role is similar to that of multiplicative covariance inflation parameters used in many EDA applications. Note that this formulation assumes a steady state regime where the average global mean error variance increase between successive assimilations equals the mean error variance decrease resulting from each assimilation step.

After estimating the background error variances, the update of equation (6a) is applied. This step corresponds to an OI analysis with the model background error variances calculated with equation (7). Let π_{12} represent the covariance of the v background errors at locations 1 and 2. It is estimated with

$$\pi_{12} = \sigma_{v1}\sigma_{v2}\rho_{12}, \quad (9a)$$
$$\rho_{12} = c_0(max(\tfrac{1}{L_v}|v_1 - v_2|, \tfrac{1}{L_x}|x_1 - x_2| + \tfrac{1}{L_y}|y_1 - y_2| + \tfrac{1}{L_z}|z_1 - z_2|)), \quad (9b)$$

where σ_{v1} and σ_{v2} are estimated with equation (7), the Ls are length scales in units of the variable v and in the three space dimensions and c_0 is the popular function given by equation (4.10) of Gaspari and Cohn (1999), or any other compactly supported correlation function. Alternatively, Euclidian distance can be used in the right hand side of equation (9b) at the expense of a slightly higher operation count. The *max* function selects the largest of its arguments.

Equation (9b) ensures that π_{12} is 0 if either v_1 differs significantly from v_2 or if locations 1 and 2 are very distant from each other. The intent is that in the majority of cases,

$$\pi_{12} = \sigma_{v1}\sigma_{v2}c_0((|v_1 - v_2|)/L_v),$$

and the modulation of the background error covariances with the c_0 function enforces error covariance localization in a state-dependent manner. The formulation with the *max* function is pertinent to the ocean where strong gradients often coincide with zero correlation surfaces. In other applications, one can replace equation (9b) with

$$\pi_{12} = \sigma_{v1}\sigma_{v2}c_0\left(\left(\left((v_1 - v_2)/L_v\right)^2 + \left((x_1 - x_2)/L_x\right)^2 + \left((y_1 - y_2)/L_y\right)^2 + \left((z_1 - z_2)/L_z\right)^2\right)^{\frac{1}{2}}\right).$$

The local cross-field covariances of the v and w errors in every grid cell are estimated with

$$\sigma^2_{vw} = \Theta([v - \Theta(v)][w - \Theta(w)]). \qquad (10)$$

They are used to update the fields of unobserved variables according to equation (6b-d).

2.3 Flow Adaptive error Statistics from a Time series (FAST)

Unlike SAFE which uses the spatial distribution of model variables to estimate error covariances, FAST computes the analysis increment at time t_k from n previous instances of the model state vector sampled from the recent history of the current model run,

$$X_k = \{x_{k-j} - \overline{x}_{(k)}, \quad j = 0, \ldots, n-1\}, \quad (11a)$$

$$\overline{x}_{(k)} = \frac{1}{n}\sum_{j=0}^{n-1} x_{k-j}, \quad (11b)$$

where $x_k = x(t_k), x_{k-1} = x(t_{k-1} = t_k - \tau)$, etc., for a given time lag τ. Arguably, τ should be such that x_{k-1} differs significantly from x_k while it still contains information that is useful at t_k. For simplicity, τ is set to the assimilation interval in this study.

While one could attempt to compute the analysis from X_k without further preprocessing as though it were made of the current state of each member of an ensemble of model trajectories, the resulting error covariance estimates would be dominated by the instances furthest away from the center of the time window since $\overline{x}_{(k)}$ is the simple moving average of length n estimated at time $t_{k-n/2}$. To prevent this from occurring and improve the assimilation performance, the lagged state instances are first high-pass filtered and then resampled to remove the remaining sequential ordering information.

The high-pass filtering takes the form

$$\boldsymbol{X}'_k = \left\{\boldsymbol{x}_{k-j} - \boldsymbol{x}^0_{k-j}, \quad j = 0,\ldots,n-1\right\}, \quad (12)$$

where the sequence of $\boldsymbol{x}^0_k$ is an exponential moving average (EMA) of the model state history,

$$\boldsymbol{x}^0_k = \alpha\,\boldsymbol{x}_k + (1-\alpha)\,\boldsymbol{x}^0_{k-1}, \quad (13)$$

where $0 \le \alpha \le 1$. A good choice to filter out time scales longer than half the sampling time window is $\alpha = 4/(n+2)$. The case with $\alpha = 0.5$ is essentially equivalent to forming the ensemble of first order differences over the time window.

The resampling,

$$\boldsymbol{X}''_k = \left\{\boldsymbol{x}''_{k-j} = \sum_{i=0}^{n-1} \beta_{ij}\boldsymbol{x}'_{k-i}, \quad j = 0,\ldots,n-1\right\}, \quad (14a)$$
$$\boldsymbol{X}'''_k = \left\{\boldsymbol{x}''_{k-j} - \bar{\boldsymbol{x}}'', \quad j = 0,\ldots,n-1\right\}, \quad (14b)$$

uses weights, β_{ij}, drawn from a uniform random distribution.

FAST makes the same calculations to estimate background error covariances and compute assimilation increments with the ensemble of deviations from equation (14b) as the EnKF makes with its ensemble of model states at time t_k (*e.g.,* equation 2b-f of Keppenne *et al.* 2008). One notable difference is that FAST calculates only one increment. Because a single model integration is involved, the ensemble size (n) can be increased at a very minimal cost

2.4 GEOS-5 Modeling and Ocean Data Assimilation System

2.4.1 GEOS-5 atmosphere-ocean general circulation model

The SAFE and FAST algorithms are tested in Section 3 in the context of assimilating Argo temperature data into the GFDL MOM4.1 ocean model coupled to the NASA GEOS-5 AGCM and to the Los Alamos CICE ice model (all of which comprise the GEOS-5 AOGCM). The model configuration is the same as that used for the GMAO ocean analysis (Vernieres *et al.* 2012). In summary, the OGCM is run with a geopotential vertical coordinate on a ½° grid with a gradual meridional refinement to ¼° at the Equator and with 40 vertical levels. The grid is Cartesian south of 60°N and tripolar northward thereof. The AGCM grid is 1° × 1.25° with 72 levels. The CICE model is run on the same horizontal grid as the OGCM. The AGCM is constrained by replaying the Modern-Era Retrospective analysis for Research and Applications (MERRA: Rienecker *et al.* 2011) while the ocean observations are assimilated. The replay procedure replaces the AGCM state with the state of the analysis every six hours.

2.4.2 GEOS integrated ocean data assimilation system (iODAS)

The components of the GEOS-5 AOGCM are connected to each other and to the GEOS integrated ocean data assimilation system (iODAS) with the Earth System Modeling Framework (ESMF). Besides SAFE and FAST, an EnOI utilizing a steady state ensemble of forecast-error estimates is used in Section 3 as a comparison benchmark. The parallel implementation of iODAS follows Keppenne and Rienecker (2003).

SAFE. FAST and EnOI background error covariances are localized according to equation (4) where the element of $\boldsymbol{C}$ corresponding to the *i*th and *j*th model state variables at space-time locations (x_i, y_i, z_i, t_i) and (x_j, y_j, z_j, t_j), is given by

$$c_{ij} = c_0(\max(\tfrac{1}{L_r}|r_i - r_j|, \tfrac{1}{L_x}|x_i - x_j| + \tfrac{1}{L_y}|y_i - y_j| + \tfrac{1}{L_z}|z_i - z_j|))\, c_0(\tfrac{1}{L_t}|t_i - t_j|), \quad (15)$$

where r_i and r_j are the adaptive localization variable at locations *i* and *j* and the *r* field is the observed variable. Note the similarity with equation (9b), except for the appearance of the temporal term, $c_0(\tfrac{1}{L_t}|t_i - t_j|)$. The latter results from differences between the measurement times and the analysis time. The application of equation (15) to modulate the background error covariances enforces a state-dependent error-covariance localization, even when the raw covariances are time-independent, as is the case with EnOI.

3. Application

To validate SAFE and FAST, we ran four AOGCM experiments assimilating T profiles from the broad-scale global array of temperature/salinity profiling floats (Argo: Gould *et al.* 2004). In the SAFE, FAST and EnOI runs, both T and S ocean model fields are updated. As in the GMAO production ocean analysis (Vernieres *et al.* 2012), the EnOI background error covariances are computed from the leading 20 EOFs (with the corresponding ensemble mean removed) of an ensemble of 186 short-term forecast-error estimates from coupled GEOS-5 forecasts. TOI is an univariate OI run in which the background error variance corresponds to the cumulative variance of the EOFs used in the EnOI run. Note tha the TOI run is included for completeness, even though it is known that assimilation that does not update salinity carefully can give a poor analysis (e.g., Sun et al. 2007). Besides the assimilated Argo T data, unassimilated Argo S data are used for validation. A control run without data assimilation was also run.

The runs cover a two-year period starting January 1, 2010. The ocean initial conditions are the same for all runs and come from the GMAO ocean analysis (Vernieres *et al.* 2012). The GEOS-5 replay procedure constrains the atmosphere to MERRA over the period of the runs. Every five days, data from a 5-day time window centered about the analysis time are processed. The operator $\boldsymbol{H}$ is a 4-dimensional interpolation operator to the time and location of the observations. The observational error model is vertically Gaussian to reflect correlated errors in each ARGO profile and the absence of error correlations between distinct profiles. The observational error variance varies as a function of depth according to the magnitude of the vertical gradient. Details are provided in Vernieres *et al.* (2012). The assimilation increments are applied incrementally over a five-day period, as in the incremental analysis update procedure of Bloom *et al.* (1996), but without rewinding the model clock (Keppenne *et al.* 2008).

The SAFE run estimates its background error covariances from equations (7) and (10) where the Θ operator consists of 10 diffusion steps. To improve the performance in the low latitudes, SAFE error covariances are explicitly disabled when they involve a grid cell within the 10°N-10°S latitude band and another grid cell outside of it. This step is exclusively applied in the SAFE run to prevent the state variables at grid cells outside the waveguide from participating in the estimation of the background error variance (Θ operator) at grid cells inside the waveguide. The FAST run applies equations (11-14) with a five-day lag, n=20 and α=0.18. Only 20 lags are used to facilitate comparison with EnOI, since the latter uses a static ensemble of 20 leading EOFs. The error-covariance localization scales (Ls in equation 15) are the same in all runs and are identical to those used in Vernieres *et al.* (2012).

Figure 1 shows that varying the size of the neighborhood used in the SAFE background error estimation (number of Θ smoothing iterations in equation 7) has little effect on the performance of the SAFE algorithm. It shows the evolution of global RMS OMF reduction from the corresponding RMS OMF from the control run without data assimilation, such that negative numbers indicate that the analysis is closer to the data than the control. Fig. 1a corresponds to the assimilated T data and Fig. 1b to the unassimilated S data. The three cases shown correspond to 5 (red), 10 (blue) and 20 (green) iterations of a Laplacian filter. While the case with 20 iterations produces a somewhat larger RMS S OMF reduction, the differences from the other two cases are small.

Figure 2 illustrates how high-pass filtering and resampling the sequence of background states from which FAST estimates error covariances affects the assimilation performance. It shows the global RMS OMF reduction from the control for both T and S in five cases. The green lines correspond to the full FAST methodology (equations 11-14) with n=20 and α=0.18 (period 10 EMA). The four other cases shown correspond to (1) the deviations from their ensemble mean of the most recent 20 unfiltered background states sampled every five days (magenta), (2 and 3) the deviations from their ensemble mean of the most recent 20 first order time differences (cyan) and second-order time differences (blue) of background states sampled every five days and (4) the EnOI run (red). Clearly, computing covariances from unfiltered background states, a procedure which corresponds to using signal covariances, results in the poorest performance for S data even though it draws the model state closest to the T data. The performance obtained with the dynamic ensembles of most recent first and second order time differences is close to that obtained with the static ensemble of leading EOFs. FAST with 50-day high-pass filtering (period-10 EMA removal from a time series with L = 5 days) performs best for S and achieves a good compromise for T. Presumably, the 50-day filtering retains pertinent information and avoids aliasing to the lower frequencies but it is possible that better results could be obtained with a different high-pass period.

To illustrate the SAFE and FAST error covariance models, Figure 3 shows time sequences of zonal vertical cross sections at the Equator through the SAFE (Fig. 3 a-d) and FAST (Fig. 3 e-h) background error standard deviation estimates for the model's ocean temperature. The succession is shown with a 3-month interval. The FAST and SAFE sections are qualitatively similar. Yet, the SAFE estimates are noticeably smoother because the number of grid cells participating in the SAFE spatial averaging is larger than the number of lagged state instances used in the FAST computations. Also note the general resemblance to the corresponding section through the time-independent background error standard deviation field used by EnOI and TOI (Fig. 3i). The differences between the equatorial sections are largest in the Indian and Atlantic Ocean.

The processing time of each run with data assimilation expressed in terms of the time taken by the control run on 30 Intel Altix Sandy Bridge nodes (360 2.8 GHz cores) is shown in Figure 4.

TOI takes 70% longer than the control run while FAST and EnOI both take about twice as long as TOI and SAFE takes nearly 50% longer than TOI. For comparison, the best case scenario for a 20-member ensemble run in which ensemble members are run sequentially is also shown. Running ensemble members in parallel, while possible with the GEOS iODAS would require many more compute nodes.

Figure 5 illustrates the background error covariance models used in each run by showing marginal T and S assimilation increments corresponding to the impact of a unit T innovation at (0ºN, 140ºW, 180m) at the end of the runs (January 1, 2012). The top row of panels (a), (e) and (i) shows zonal sections through the corresponding marginal T increments in the SAFE (left), FAST (middle) and EnOI (right) runs. The 2nd row of panels (b), (f) and (j) shows corresponding meridional T sections. Panels (c), (g) and (k) (3rd row) and the bottom row of panels (d), (h) and (l) show zonal and meridional sections through the corresponding marginal S increments.

The differences apparent in Figure 5 result primarily from differences in covariance modeling approach (static ensemble in EnOI, time-lagged ensemble in FAST, spatial covariance in SAFE). However, differences also arise from differences in the state adaptive error-localization of equation (15) since the differences between the respective background states have increased over time (particularly evident in Figure 6). The amplitude differences between the SAFE, FAST and EnOI marginal gains reflect differences in the background error estimates at the observation location. In this example, there is more correspondence between the shapes of the marginal T and S increments from the EnOI (panels (i) and (k) and panels (j) and (l)) than those from SAFE or FAST. The amplitude of the T marginal increment is also largest in the EnOI run. Yet, the amplitude of the S marginal increment is relatively small in the EnOI run, reflecting lower covariance between the T and S error estimates at this particular observation location.

To further illustrate how the SAFE, FAST and EnOI error-covariance models differ, Figure 6 shows the time evolution (sampled every three months) of zonal sections through the marginal S increment corresponding to a unit T innovation at the same Equatorial location considered in Figure 5. Not surprisingly since the EnOI estimates background covariances from a static ensemble, its marginal S gain at this location displays the least temporal variation. The latter result from how the background T field (*r* in equation (15)) changes with time. Conversely, the FAST marginal S gain varies the most with time as one could have expected because the corresponding background error covariances are high pass filtered by design and represent errors/uncertainties at periods shorter than 50 days in this case. Clearly, the FAST covariances are influenced by tropical instability waves which mostly occur between July and November and have wavelengths of 1000-2000 km and periods of 20-40 days (e.g., Willett et al., 2006). While the SAFE background error covariance calculations also depend on the background fields, the resulting covariances only capture variability in space, not in time.

Figure 7 quantifies the improvement (negative values) or worsening (positive values) over the control by showing to what extent the RMS OMF statistics differ from the corresponding statistics from the control run. RMS OMF differences are shown in each panel for the SAFE (blue), FAST (red), EnOI (green) and TOI (magenta) runs. Figure 7a corresponds to the assimilated Argo T data, while Figures 7b and 7c correspond to the unassimilated Argo S data above and below 300 meters. While the four data assimilation methods perform similarly for T, FAST stands out for its better performance in terms of S, especially in the upper ocean (Fig. 7b). On the other hand, the underperformance of TOI, which degrades the model salt field compared to the control run, is especially striking in the thermocline (Fig. 7c).

The global RMS observation minus forecast (OMF) differences corresponding to the T data are comparable in the four runs with T data assimilation (SAFE: 0.76 ºC, FAST: 0.88 ºC, EnOI: 0.76 ºC, TOI: 0.87 ºC), as they each explain approximately the same fraction of the T innovation variance of the control run (1.27^2 $ºC^2$). This result is as expected given that each run sets $\gamma=1$ in equation (8) to facilitate the comparison. Figure 8 further illustrates the respective performance of each run with T assimilation. The difference of the RMS OMF (horizontally and over time) in the data assimilation runs from that in the control is shown as a function of depth for 2011 (blue: SAFE, red: FAST, green: EnOI, magenta: TOI). Negative numbers mean that the data assimilation brings the (5-day lead) forecast state closer to the data than the control and should be the norm if the data are unbiased. Figure 8a corresponds to the assimilated T data and Figure 8b to the unassimilated S data. For T, the level of improvement over the control is similar for all runs and is largest near a depth of 100 meters. For S, the results are markedly different. TOI is worse than the control over the entire water column and while SAFE, FAST and EnOI all improve upon the control over the entire column, FAST produces the largest improvement over the entire depth range.

The horizontal distributions of the differences in RMS S OMF from those of the control during 2011 for each of the SAFE, FAST, EnOI and TOI runs are shown in Figure 9 for the upper 300 meters and in Figure 10 for depths greater than 300 meters. In the upper ocean, SAFE, EnOI and TOI all show significant degradations from the control in the Western Equatorial South Pacific (red areas in Figs. 9a, 9c, and 9d). FAST does better in the same area and performs best overall (Fig. 7b). Since the upper ocean salt content is heavily influenced by precipitation and evaporation and the corresponding fluxes are constrained to the MERRA forcing in all runs, including the control, it is not surprising that the analyses (which all assimilate T only) do not outperform the control at the surface and in the mixed layer. Positive impacts on the model salinity from the T data assimilation are most likely to manifest themselves further away from the surface. Accordingly, the positive impact of the S field correction in the SAFE, FAST and EnOI runs is more apparent below 300 meters, especially in the Northern Atlantic, Gulf Stream and Kuroshio areas and in the area of the West Australian and Leeuwin currents in the Southeast Indian Ocean. While FAST performs best overall, it under-performs the control in the Indian sector of the Southern Ocean. Since the comparison is restricted to 2011, these regional comments are not definitive.

4. Outlook

When EDA schemes are applied to complex numerical models, the ensemble size is always a limiting factor or the object of compromise. The methodologies introduced here are designed to possess the main advantages of EDA methods, namely the ability to update state variables even if unobserved (or not directly assimilated) and to adaptively estimate the spatial distribution of background errors, without incurring the cost of ensemble integrations.

While SAFE is nearly as economical as conventional OI, our results hint that it is somewhat less effective as FAST or EnOI in updating fields of unobserved variables. The better performance of FAST in this respect may stem in part from its error covariance model ability to capture sub-seasonal variability and in part from the fact that it does not rely on the type of heuristic assumption made with SAFE between equations (6c) and (6d).

Of course, nothing precludes one from using FAST or SAFE to boost the ensemble size of an EDA scheme. SAFE background error estimates can be combined with those obtained with a dynamical ensemble as is usually done with OI covariances in hybrid EDA schemes. Several FAST trajectories can be run concurrently and the resulting time lagged ensembles combined into a single ensemble. Another area where SAFE and FAST seem to hold promise is in complex production systems where running an EDA scheme would require that the ensemble size or model resolution be severely limited, and in high-resolution data assimilation applications where numerical cost is critical. To illustrate this, we increased the MOM and CICE horizontal resolution to a 0.1° global tripolar grid with gradual meridional refinement to 0.05° and the GEOS-5 AGCM resolution to 0.25° × 0.3125°, while keeping the number of verticals levels unchanged (MOM/CICE: 40, AGCM: 72). We then started running the high resolution CGCM on 960 2.8 GHz Altix Sandy Bridge cores with a 5-minute time step replaying the MERRA reanalysis in its AGCM component and initializing its OGCM component with a horizontally constant hydrostatic equilibrium condition. Each day, a multi-scale (bi-scale) ocean analysis took place. First, T, S and current fields from the 0.5° GMAO ocean analysis (Vernieres *et al.* 2012) were assimilated into the 0.1° global OGCM using SAFE and updating only the fields of observed variables. The covariance localization scales were the same as those used to produce the ocean analysis in this step. Following the assimilation of the 0.5° production analysis, the 0.1° temperature analysis was refined by using SAFE to assimilate daily 0.25° Reynolds (2007) SSTs, shortening the horizontal localization scales to one fifth of the production analysis values. SAFE was used because FAST would have required the availability of past background states. One could choose to continue the analysis with FAST after the initial spin up.

Figures 11 and 12 illustrate the rapid convergence of the ocean surface conditions from the multi-scale ocean analysis to the Reynolds data. They show details of the SST field on August 27, 2007, 27 days into the run. In each of Figures 11 and 12, panel (a) correspond to the 0.1° analysis, panel (b) shows the 0.25° Reynolds SST data and panel (c) shows the corresponding detail from the 0.5° production analysis. Had one wanted to produce such a fine analysis with EDA, the computational resource requirement would have been overwhelming (about 1 hour of wall clock time per simulation day per ensemble member on 960 cores).

5. Acknowledgement

This work is supported by NASA's Modeling Analysis and Prediction Program under WBS 802678.02.17.01.25. The infrastructure for the runs is provided by the NASA Center for Climate Simulation (NCCS). Yuri Vikhliaev, Max Suarez and Bin Zhao helped configure the GEOS-5 modeling system.

6. Bibliography

Bloom, S.C., L.L. Takacs, A.M. DaSilva, and D. Ledvina, 1996: Data assimilation using incremental analysis updates. *Mon. Wea. Rev.*, **124**, 1256-1271.

Borovikov, A., M.M. Rienecker, C.L. Keppenne, and G.C. Johnson, 2005: Multivariate error covariance estimates by Monte-Carlo simulation for assimilation studies in the Pacific Ocean. *Mon. Wea. Rev.*, **133**, 2310-2334.

Eliassen A., 1954: Provisional report on calculation of spatial covariance and autocorrelation of the pressure field. Report 5. Videnskaps Akademiet Institut for Vaer Og Klimaforskning, Oslo, Norway, 12pp.

Evensen, G., 1994: Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics. *J. Geophys. Res.*, **99 (C5)**, 10,143-10,162.

Gaspari, G., and S.E. Cohn, 1999: Construction of correlation functions in two and three dimensions. *Quart. J. Roy. Meteor. Soc.*, **125B (554)**, 723-757.

Gould, J., D. Roemmich, S. Wijffels, H. Freeland, M. Ignaszewsky, X. Jianping, S. Pouliquen, Y. Desaubies, U. Send, K. Radhakrishnan, K. Takeuchi, K. Kim, M. Danchenkov, P. Sutton, B. King, B. Owens and S. Riser, 2004: Argo Profiling Floats Bring New Era of In Situ Ocean Observations, *EOS, Trans. AGU*, **85 (19)**, 179, 190-191.

Houtekamer, P.L., and H.L. Mitchell, 2001: A sequential ensemble Kalman filter for atmospheric data assimilation. *Mon. Wea. Rev.*, **129**, 123-137.

Kalman, R., 1960: A new approach to linear filtering and prediction problems. *J. Basic Eng.*, **D82**, 35-45.

Keppenne, C.L., and M.M. Rienecker, 2003: Assimilation of temperature into an isopycnal ocean general circulation model using a parallel ensemble Kalman filter. *J. Mar. Sys.*, **40-41**, 363-380.

Keppenne, C.L., M.M. Rienecker, J.P. Jacob and R.M. Kovach, 2008: Error covariance modeling in the GMAO ocean ensemble Kalman filter, *Mon. Wea. Rev.*, **136**, 2964-2982.

Oke, P.R., A. Schiller, D.A. Griffin and G.B. Brassington, 2005: Ensemble data assimilation for an eddy resolving ocean model, *Q.J. Roy. Met. Soc.*, **131**, 3301-3311.

Oke, P.R.; G.B. Brassington; D.A. Griffin and A. Schiller, 2010: Ocean data assimilation: a case for ensemble optimal interpolation, *Aust. Meteorolog. & Oceanogr. J.*, **59**, 67-76.

Reynolds R.W., T.M. Smith, C. Liu, D.B. Chelton, K.S. Casey, and M.G. Schlax, 2007: Daily high-resolution blended analyses for sea surface temperature. *J. Climate*, **20**, 5473-5496.

Rienecker, M.M., M.J. Suarez, R. Gelaro, R. Todling, J. Bacmeister, E. Liu, M.G. Bosilovich, S.D. Schubert, L. Takacs, G.-K. Kim, S. Bloom, J. Chen, D. Collins, A. Conaty, A. da Silva, et al., 2011. MERRA - NASA's Modern-Era Retrospective Analysis for Research and Applications. *J. Climate*, **24**, 3624-3648.

Vernieres, G., C.L. Keppenne, M.M. Rienecker, J.P. Jacob and R.N. Kovach, 2012: The GEOS-ODAS description and evaluation. Technical Report Series on Global Modeling and Data Assimilation, NASA/TM-2012-104606.

Wan, L., L. Bertino and J. Zhu, 2010: Assimilating Altimetry Data into a HYCOM Model of the Pacific: Ensemble Optimal Interpolation versus Ensemble Kalman Filter, *J. Atmos. Ocean. Tech.*, **27 (4)**, 753-765.

Willett, C.S., R.R. Leben, and M. Lavin, 2006: Eddies and tropical instability waves in the eastern tropical Pacific: A review. *Prog. Oceanogr.*, **69**, 218-238.

Figure captions

Figure 1. Reduction of SAFE RMS OMF over the corresponding RMS OMF from the control run without data assimilation for (a) assimilated Argo T and (b) unassimilated Argo S data. The three cases shown correspond to SAFE runs in which the background error covariance estimation involves 5 (red), 10 (blue) and 20 (green) steps of a diffusive (Laplacian) filter. Negative (vs. positive) values correspond to improvements (vs. worsening) over the control.

Figure 2. Reduction of RMS OMF over the corresponding RMS OMF from the control run without data assimilation for (a) assimilated Argo T and (b) unassimilated Argo S data in runs assimilating the Argo T data every five days and in which the background error covariances are estimated with either EnOI using a static ensemble of 20 leading error EOFs (EnOI: red), a lagged ensemble of the 20 most recent unfiltered background states (0 order: magenta), an ensemble of the 20 most recent first-order time differences (1^{st} order: cyan), an ensemble of the 20 most recent second-order time differences (2^{nd} order: blue), or FAST with 20 lags and 50-day high pass filtering (FAST: green). Negative (vs. positive) values correspond to improvements (vs. worsening) over the control.

Figure 3. Temperature background error standard deviation estimates along the Equator in the SAFE, FAST and EnOI runs of Section 3 and corresponding from top to bottom to March 31, 2011 (a: SAFE, e: FAST), June 30, 2011 (b: SAFE, f: FAST), September 30, 2011 (c: SAFE, g: FAST) and December 31, 2011 (d: SAFE, h: FAST) Panel (i) shows the time independent background error standard deviation estimate used by both the EnOI and TOI runs. The color scale shown to the right of panel (i) is applicable for all panels.

Figure 4. Processing time per month of model simulation expressed in units of the corresponding processing time from the control run. Note the logarithmic scale. The EnKF case corresponds to a best case scenario for a 20-member EnKF run in which ensemble members are run sequentially.

Figure 5. Zonal and meridional sections through the marginal contribution to the T and S assimilation increments in PSU corresponding to a unit T innovation at (0ºN, 140ºW, 180m) in the SAFE (a-d), FAST (e-h) and EnOI (i-l) runs on January 1, 2012. Zonal (meridional) sections are labeled W-E (S-N). (a), (e), (i) correspond to T zonal sections, (b), (f), (j) to T meridional sections, (c), (g), (k) to S zonal sections and (d), (h), (l) to S meridional sections. The top color bar applies to all the panels in the top two rows. The bottom color bar applies to the bottom two rows.

Figure 6. Zonal sections through the marginal contribution to the S assimilation increment in PSU corresponding to a unit T innovation at (0ºN, 140ºW, 180m) in the SAFE (a-e), FAST (f-j) and EnOI (k-o) runs on (from top to bottom) January 1, 2010, April 1, 2010, July 1, 2010, October 1, 2010 and January 1, 2011. The color bar to the right applies to all the panels.

Figure 7. (a) RMS OMF difference with RMS OMF from the control run without data assimilation for (a) assimilated Argo T data, (b) unassimilated Argo S data in the upper 300

meters and (c) unassimilated Argo S data below 3000 meters. RMS OMF differences quantify the improvement (negative values) or worsening (positive values) over the control and are shown in each panel for the SAFE (blue), FAST (red), EnOI (green) and TOI (magenta) runs.

Figure 8. Global average of RMS OMF over the control as a function of depth for (a) assimilated T data and (b) unassimilated S data in the second year (2011) of the SAFE (blue), FAST (red), EnOI (green) and TOI (magenta) runs. Negative (positive) numbers indicate a reduction (increase) in RMS OMS statistics over the control run.

Figure 9. Horizontal distribution of RMS OMF differences for the unassimilated S data during 2011 with the corresponding RMS OMF from the control run. The data are binned over 0-300-meter deep by 1º zonal by 1º meridional boxes. Negative values identify areas where the analysis is closer to the Argo observations than the corresponding state from the control run and vice versa. The four panels correspond to the SAFE (a), FAST (b), EnOI (C) and TOI (d) runs.

Figure 10. Same as Figure 9 for the Argo S observations below 300 meters.

Figure 11. Eastern equatorial pacific detail of SST field on August 27, 2007 in (a) the high-resolution 0.1° multi-scale global ocean analysis, (b) the 0.25° Reynolds SST data set assimilated in the second step of each daily multi-scale assimilation and (c) the 0.5° GMAO ocean analysis assimilated in the first-step of the multi-scale procedure.

Figure 12. Same as Figure 11 for the western north Pacific east of Japan.