6-2016

An Ensemble Approach to Weak-Constraint Four-Dimensional Variational Data Assimilation

Jeremy A. Shaw
Portland State University, shaw@pdx.edu

Dacian Daescu
Portland State University, daescu@pdx.edu

Citation Details

Shaw, J., & Daescu, D. (2016). An Ensemble Approach to Weak-Constraint Four-Dimensional Variational Data Assimilation. Procedia Computer Science, 80, 496–506. http://doi.org/10.1016/j.procs.2016.05.329

Procedia Computer Science

Volume 80, 2016, Pages 496–506

ICCS 2016. The International Conference on Computational Science

An Ensemble Approach to Weak-Constraint Four-Dimensional Variational Data Assimilation

Jeremy A. Shaw[1] and Dacian N. Daescu[2]

[1] Portland State University, PO Box 751, Portland, OR, 97207, USA.
shaw@pdx.edu

[2] Portland State University, PO Box 751, Portland, OR, 97207, USA.
daescu@pdx.edu

Abstract

This article presents a framework for performing ensemble and hybrid data assimilation in a weak-constraint four-dimensional variational data assimilation system (w4D-Var). A practical approach is considered that relies on an ensemble of w4D-Var systems solved by the incremental algorithm to obtain flow-dependent estimates to the model error statistics. A proof-of-concept is presented in an idealized context using the Lorenz multi-scale model. A comparative analysis is performed between the weak- and strong-constraint ensemble-based methods. The importance of the weight coefficients assigned to the static and ensemble-based components of the error covariances is also investigated. Our preliminary numerical experiments indicate that an ensemble-based model error covariance specification may significantly improve the quality of the analysis.

Keywords: model error; weak constraint; variational data assimilation; ensemble methods; error covariance; error bias

1 Introduction

Four-dimensional variational data assimilation (4D-Var) provides an estimate to the state of a dynamical system through the minimization of a cost functional that measures the distance to a prior state (background) estimate and observations [20] over a time window $[t_0, t_N]$. The analysis fit to each information input component is determined by the specification of the error covariance matrices in the data assimilation system (DAS).

Unlike the extended Kalman filter, error covariances are typically not updated between 4D-Var assimilation cycles. A practical approach to improve the quality of the analysis is to include the "errors of the day" by using an ensemble-based estimate to the background error covariance matrix. Evensen [11] introduces this Monte Carlo alternative as the ensemble Kalman filter (EnKF) and it has since been implemented in various studies, e.g. [15], [17], [18]. Lorenc [21]

 Selection and peer-review under responsibility of the Scientific Programme Committee of ICCS 2016

doi:10.1016/j.procs.2016.05.329

and Fairbairn et al. [12] investigate the potential use of EnKF for numerical weather prediction (NWP) applications and its analysis performance, as compared with 4D-Var.

For large-scale dynamical systems, the number of ensemble forecasts is much smaller as compared to the dimension of the discrete state vector. Therefore, an ensemble-based representation to the background error covariance matrix is of low rank and corrupted by sampling errors. To alleviate these issues, several approaches have been considered for practical implementation including covariance localization [15, 16] and the formulation of hybrid methods that aim to synergistically combine the merits of variational and ensemble-based DA [1, 2].

Weak-constraint 4D-Var (w4D-Var) provides a theoretical framework to account for modeling errors in the analysis scheme. Trémolet [25] investigates some possible implementations of w4D-Var. In addition to the specification of the background error covariance ($\mathbf{B}$) matrix, the w4D-Var formulation requires information on the model error statistics and specification of the model error covariance. Up to now, the increased computational cost associated with w4D-Var has prevented its practical implementation. Various simplifications to reduce the computational burden have been considered, including writing the model error covariance as a scalar multiple of the background error covariance (see [8] for example) and modeling the model error [14, 26, 27]. Research to implement an ensemble data assimilation approach to model error covariance estimation in w4D-Var is at an incipient stage. Mitchell and Carrassi [24] use ensembles to account for model error, but in the context of the ensemble transform Kalman filter. Desroziers et al. [9] investigate a possible implementation of an ensemble 4D-Var using a four-dimensional ensemble covariance.

Ensemble data assimilation can estimate not only the model error covariance matrices, but also bias. Traditionally, an assumption is made that the errors in data assimilation are unbiased to simplify the computational cost or because the information about error biases is not available. Bias in data assimilation has been explored in the works by Dee [4], Dee and Da Silva [5], and Dee and Todling [6], where it is noted that errors in models and the data are often systematic rather than random. Attempts to correct for error bias have been made in the form of bias detection and correction methods and "bias-aware" data assimilation methods, including bias correction in variational data assimilation [7], but not in the context of w4D-Var. Bias-aware Kalman filters have been explored by Drécourt et al. [10].

This work investigates novel applications of ensemble and hybrid techniques to estimate the model error statistics in a w4D-Var DAS. The implementation of ensemble-based DA for w4D-Var is presented. A proof-of-concept and comparison of w4D-Var to the ensemble data assimilation and hybrid assimilation schemes is made in numerical experiments. The terminology and notation follow closely to that of Ide et al. [19] and Trémolet [25].

2 Ensemble Data Assimilation

Weak-constraint 4D-Var provides a sequence of time-distributed analyses $\mathbf{x}_i^a \in \mathbb{R}^n$ that estimate the true state $\mathbf{x}_i^t$ of a dynamical system at time t_i of the data assimilation interval $[t_0, t_N]$. The nonlinear cost functional associated with w4D-Var is defined as

$$J(\mathbf{x}_0, \ldots, \mathbf{x}_N) = \frac{1}{2}[\mathbf{x}_0 - \mathbf{x}_0^b]^{\mathrm{T}}\mathbf{B}^{-1}[\mathbf{x}_0 - \mathbf{x}_0^b] + \frac{1}{2}\sum_{i=0}^{N}\left[\mathbf{y}_i - \mathbf{h}_i(\mathbf{x}_i)\right]^{\mathrm{T}}\mathbf{R}_i^{-1}\left[\mathbf{y}_i - \mathbf{h}_i(\mathbf{x}_i)\right] + \frac{1}{2}\sum_{i=1}^{N}[\boldsymbol{\eta}_i - \mathbf{q}_i]^{\mathrm{T}}\mathbf{Q}_i^{-1}[\boldsymbol{\eta}_i - \mathbf{q}_i] \quad (1)$$

where $\mathbf{x}_0^b \in \mathbb{R}^n$ is a prior (background) estimate of the true state at time t_0, $\mathbf{y}_i \in \mathbb{R}^{p_i}$ is the observation vector at time t_i, $\mathbf{h}_i$ denotes the observation operator that maps the state $\mathbf{x}_i$ into observation space, and $\boldsymbol{\eta}_i = \mathbf{x}_i - \mathcal{M}_i(\mathbf{x}_{i-1})$ represents the error in the forecast model $\mathcal{M}_i$ that advances the state from time t_{i-1} to time t_i. The vector $\mathbf{q}_i$ is defined as the statistical expectation of $\boldsymbol{\eta}_i$ and is referred to as model error bias. The statistical information on the background error $\boldsymbol{\varepsilon}^b = \mathbf{x}_0 - \mathbf{x}_0^b$, observational errors $\boldsymbol{\varepsilon}_i^o = \mathbf{y}_i - \mathbf{h}_i(\mathbf{x}_i)$, and model errors $\boldsymbol{\eta}_i$ is used to specify the positive definite matrices $\mathbf{B} \in \mathbb{R}^{n\times n}$, $\mathbf{R}_i \in \mathbb{R}^{p_i\times p_i}$, and $\mathbf{Q}_i \in \mathbb{R}^{n\times n}$, which represent the background, observation, and model error covariance matrices used in the data assimilation system.

In some applications, model error bias is not accounted for or is assumed to be zero, which eliminates the $\mathbf{q}_i$ term from the cost functional. In a strong-constraint 4D-Var system, a perfect model assumption $\mathbf{x}_i = \mathcal{M}_i(\mathbf{x}_{i-1})$ is used to completely eliminate the model error term from (1) and simplify the cost functional so that the only free variable is $\mathbf{x}_0$, the initial condition. By taking into account model error in w4D-Var, the control variable is sequence of states $\{\mathbf{x}_0, \ldots, \mathbf{x}_N\}$. Trémolet [25] describes other possible formulations of the control variable, including $\{\mathbf{x}_0, \boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_N\}$.

Ensemble data assimilation has been used in conjunction with variational methods in an attempt to capture the "errors of the day" and dynamically update the background error covariance. Ensembles can also be used to estimate the model error covariances $\mathbf{Q}_i$ and the model bias $\mathbf{q}_i$ by using ensembles for the analysis states $\mathbf{x}_{i,j}^a$. The steps needed to obtain ensemble estimates of model error are presented next.

2.1 Derivation of the Model Error Ensemble

When the true model error statistics are unknown, the data assimilation system specifications of the error bias $\mathbf{q}_i \neq \mathbf{q}_i^t$ and error covariance $\mathbf{Q}_i \neq \mathbf{Q}_i^t$ are made. The incremental algorithm, introduced by Courtier et al. [3], may be used to perform w4D-Var over a time window $[t_0, t_N]$. The states $\mathbf{x}_i^g$ at which the model and observation operators are linearized will utilize the assumed model error statistics, that is, let

$$\mathbf{x}_0^g = \mathbf{x}_0^b, \qquad \mathbf{x}_i^g = \mathcal{M}_i(\mathbf{x}_{i-1}^g) + \mathbf{q}_i, \qquad i = 1, \ldots, N. \tag{2}$$

The incremental method then produces the four-dimensional analysis

$$\mathbf{x}^a = \mathbf{x}^g + \mathbf{K}[\mathbf{y} - \mathbf{h}(\mathbf{x}^g)] \tag{3}$$

where $\mathbf{K}$ is the four-dimensional gain matrix analogous to the Kalman gain matrix.

An ensemble of analyses $\mathbf{x}_{i,j}^a$, where $i = 0, 1, \ldots, N$ and $j = 1, \ldots, N_e$, is used to produce a low-rank representation to the model error covariance. The setup is as follows.

- Prescribe the background error statistics $\mathbf{B}$, observation error covariances $\mathbf{R}_i$, and model error statistics $\mathbf{Q}_i$, $\mathbf{q}_i$ to be used for each w4D-Var problem, the same specification for each ensemble member.
- From the background state $\mathbf{x}_0^b$, form the background ensemble $\mathbf{x}_{0,j}^b = \mathbf{x}_0^b + \boldsymbol{\varepsilon}_j^b$, where $\boldsymbol{\varepsilon}_j^b$ is generated from the normal distribution $N(\mathbf{0}, \mathbf{B})$.
- Perturb the observation $\mathbf{y}_i$ to form an ensemble $\mathbf{y}_{i,j} = \mathbf{y}_i + \boldsymbol{\varepsilon}_{i,j}^o$, where the perturbation $\boldsymbol{\varepsilon}_{i,j}^o$ is normally distributed with mean zero and covariance $\mathbf{R}_i$, for $i = 0, 1, \ldots, N$ and $j = 1, \ldots, N_e$.

- For each member of the background ensemble $\mathbf{x}^b_{0,j}$, form the corresponding ensemble of guesses $\mathbf{x}^g_{i,j}$ according to (2) using the assumed model error bias $\mathbf{q}_i$.

A substitute for using the statistics of $\mathbf{B}$ to perturb the background is to use the approximate background error

$$\varepsilon = \|\mathbf{x}^a_0 - \mathbf{x}^b_0\|/n \tag{4}$$

as the standard deviation for the mean zero normally distributed perturbation. A multiplicative constant β can be included so that the standard deviation of the perturbations is $\beta\varepsilon$.

By performing w4D-Var using the incremental method with data $\mathbf{x}^g_{0,j}, \ldots, \mathbf{x}^g_{N,j}$ and observations $\mathbf{y}_{0,j}, \ldots, \mathbf{y}_{N,j}$, we get an ensemble of analysis states $\mathbf{x}^a_{i,j}$. The four-dimensional analysis ensemble $\mathbf{x}^a_j$ follows from (3)

$$\mathbf{x}^a_j = \mathbf{x}^g_j + \mathbf{K}_j[\mathbf{y}_j - \mathbf{h}(\mathbf{x}^g_j)] \tag{5}$$

where the gain matrix $\mathbf{K}_j$ may vary with the ensemble member j. In this framework, the best estimate of the true state is obtained as the ensemble average for each time

$$\bar{\mathbf{x}}^a_i = \frac{1}{N_e}\sum_{j=1}^{N_e} \mathbf{x}^a_{i,j}. \tag{6}$$

From (5) and (6), define ensemble estimates to model error

$$\boldsymbol{\eta}_{i,j} = \bar{\mathbf{x}}^a_i - \mathcal{M}_i(\mathbf{x}^a_{i-1,j}) \tag{7}$$

for $i = 1, \ldots, N$ and $j = 1, \ldots, N_e$. With the model error ensemble now available, the model error bias $\mathbf{q}^t_i$ is estimated by the ensemble mean

$$\mathbf{q}_{i,e} = \frac{1}{N_e}\sum_{j=1}^{N_e} \boldsymbol{\eta}_{i,j} \tag{8}$$

and the associated ensemble estimates to the model error covariance matrix are

$$\mathbf{Q}_{i,e} = \frac{1}{N_e - 1}\sum_{j=1}^{N_e} [\boldsymbol{\eta}_{i,j} - \mathbf{q}_{i,e}][\boldsymbol{\eta}_{i,j} - \mathbf{q}_{i,e}]^{\mathrm{T}}. \tag{9}$$

Now that estimates for the model error bias and model error covariance are available, $\mathbf{q}_{i,e}$ and $\mathbf{Q}_{i,e}$ may be used in a w4D-Var data assimilation system. This procedure is summarized in Algorithm 1. Possible ensemble-based assimilation schemes are described next.

2.2 Ensemble-based w4D-Var Schemes

Instead of prescribing the model error covariance matrices as static $\mathbf{Q}_i$ that do not change between assimilation cycles, one approach is to specify the model error bias and model error covariance matrices as $\mathbf{q}_{i,e}$ and $\mathbf{Q}_{i,e}$, respectively. This choice of specifying $\mathbf{q}_i = \mathbf{q}_{i,e}$ and $\mathbf{Q}_i = \mathbf{Q}_{i,e}$ utilizes the information from the "errors of the day" to improve the quality of the analysis. These specifications can be kept up-to-date in future time-steps by computing the ensemble estimates $\mathbf{q}_{i,e}$ and $\mathbf{Q}_{i,e}$ in each data assimilation cycle.

Algorithm 1 Computation of the ensemble estimates of model error.

procedure MODEL ERROR ENSEMBLE($\mathbf{B}$, $\mathbf{R}_i$, $\mathbf{Q}_i$, $\mathbf{q}_i$, N_e)
  for $j = 1, \ldots, N_e$ **do**
    $\mathbf{x}^b_{0,j} = \mathbf{x}^b_0 + \boldsymbol{\varepsilon}^b_j$ ▷ Perturb the background
    $\mathbf{x}^g_{0,j} = \mathbf{x}^b_{0,j}$ ▷ Set the guess states
    for $i = 1, \ldots, N$ **do**
      $\mathbf{x}^g_{i,j} = \mathcal{M}_i(\mathbf{x}^g_{i-1,j}) + \mathbf{q}_i$
    end for
    for $i = 0, 1, \ldots, N$ **do**
      $\mathbf{y}_{i,j} = \mathbf{y}_i + \boldsymbol{\varepsilon}^o_{i,j}$ ▷ Perturb the observations
    end for
    $(\mathbf{x}^a_{0,j}, \ldots, \mathbf{x}^a_{N,j}) = \text{w4DVar}(\mathbf{B}, \mathbf{R}_i, \mathbf{Q}_i, \mathbf{q}_i, \mathbf{x}^g_{i,j}, \mathbf{y}_{i,j})$ ▷ Analysis ensemble
  end for
  for $i = 0, 1, \ldots, N$ **do**
    Compute $\bar{\mathbf{x}}^a_i$ from equation (6)
    for $j = 1, \ldots, N_e$ **do**
      Compute $\boldsymbol{\eta}_{i,j}$ from equation (7)
    end for
  end for
  for $i = 1, \ldots, N$ **do**
    Compute $\mathbf{q}_{i,e}$ from equation (8)
    Compute $\mathbf{Q}_{i,e}$ from equation (9)
  end for
end procedure

The ensemble covariance matrices $\mathbf{Q}_{i,e}$ may have low rank due to a small ensemble size and additionally suffer from the presence of sampling error. To reduce this, one may replace $\mathbf{Q}_i$ in the data assimilation system by the Schur (elementwise) product of the ensemble covariance $\mathbf{Q}_{i,e}$ with a localization matrix

$$\mathbf{Q}_i = \mathbf{Q}_{i,e} \circ \mathbf{C}_i \tag{10}$$

where $\mathbf{C}_i$ is a properly selected correlation matrix. A popular correlation function to apply is the fifth-order rational function of compact support, given by equation (4.10) of Gaspari and Cohn [13].

Another option is to specify the model error covariance matrices as a linear combination of a static matrix $\mathbf{Q}_{i,c}$ and the ensemble covariance

$$\mathbf{Q}_i = \alpha_i \mathbf{Q}_{i,c} + (1 - \alpha_i)\mathbf{Q}_{i,e}. \tag{11}$$

A localization matrix $\mathbf{C}_i$ may applied to $\mathbf{Q}_{i,e}$ so that (11) is replaced by

$$\mathbf{Q}_i = \alpha_i \mathbf{Q}_{i,c} + (1 - \alpha_i)\mathbf{Q}_{i,e} \circ \mathbf{C}_i. \tag{12}$$

Similarly, the model error bias is specified as a linear combination of a static vector $\mathbf{q}_{i,c}$ and the ensemble average using the same parameter

$$\mathbf{q}_i = \alpha_i \mathbf{q}_{i,c} + (1 - \alpha_i)\mathbf{q}_{i,e} \tag{13}$$

where $0 \leq \alpha_i \leq 1$. This combination of two specifications of model error is referred to as hybrid data assimilation. For $\alpha_i = 1$, the specified model error will utilize the current static specification, or the status quo, while for $\alpha_i = 0$, it will be set to the ensemble model error statistics. Hybrid data assimilation is designed to combine the merits of both the static component $\{\mathbf{Q}_{i,c}, \mathbf{q}_{i,c}\}$ and the dynamic component $\{\mathbf{Q}_{i,e}, \mathbf{q}_{i,e}\}$ with a value of α_i satisfying $0 < \alpha_i < 1$ to improve the quality of the analysis more than the static and dynamic components can do alone.

3 Numerical Experiments

The performance of ensemble and hybrid formulations of w4D-Var assimilation described in subsection 2.2 is investigated in comparative numerical experiments performed with the multi-scale model of Lorenz [22]

$$\frac{dx_k}{dt} = x_{k-1}(x_{k+1} - x_{k-2}) - x_k - \frac{hc}{b}\sum_{j=1}^{J} y_{jk} + F \tag{14a}$$

$$\frac{dy_{jk}}{dt} = cby_{j+1,k}(y_{j-1,k} - y_{j+2,k}) - cy_{jk} + \frac{hc}{b}x_k \tag{14b}$$

where $k = 1, \ldots, K$ and $j = 1, \ldots, J$. The y_{jk} variables vary at a smaller scale than the x_k variables and are arranged as $y_{11}, y_{21}, \ldots, y_{J1}, y_{12}, \ldots, y_{J2}, \ldots, y_{JK}$. They also extend cyclically so that $y_{J+1,1} = y_{11}$. We will refer to the model given by (14) by LZ96. In this experiment, the "true" state of the dynamical system is represented by the integration of (14) by the fourth-order Runge-Kutta method with $b = c = 10$, $h = 1$, $K = 40$, $J = 10$ and $F = 8$. By ignoring the effects of the y_{jk} variables, the Lorenz 40-variable model [23]

$$\frac{dx_k}{dt} = x_{k-1}(x_{k+1} - x_{k-2}) - x_k + F \tag{15}$$

will only approximate the true state evolution and model error is now introduced by the unrepresented small-scale dynamics. Thus, for the data assimilation process, the true state $\mathbf{x}_i^t$ at time t_i will be the x-values produced from the integration of (14), whereas the forecast model $\mathcal{M}_i$ will be the integration of (15) using a constant step-size $\Delta t = 0.05$, which identifies to a 6-hour time period. The integration of (14) requires a smaller time step to preserve numerical stability, so a 6-hour forecast is achieved through ten smaller time-steps with $\Delta t = 0.005$.

A data assimilation window consists of the current time t_0 and three time-steps, representing an assimilation window $[t_0, t_3]$. Observational data are generated from the true state with the observational error taken from the distribution $N(0, (\sigma^o)^2)$ with the standard deviation specified as $\sigma^o = 0.55$. The observation operator satisfies $\mathbf{h}_i(\mathbf{x}_i) = \mathbf{x}_i$ for $i = 1, 2, 3$.

An analysis will be produced from w4D-Var after setting up the background error covariance $\mathbf{B}$ by running the extended Kalman filter using the true model error statistics for 700 time-steps with $\mathbf{B}$ initialized to the identity matrix. The background $\mathbf{x}_0^b$ for each step of the extended Kalman filter is taken to be a forecast of the previous analysis perturbed by random noise. After the spin-up cycle is complete, $\mathbf{B}$ will then remain static for w4D-Var assimilation.

A comparative analysis is done to investigate the performance of the ensemble and hybrid assimilation methods to gauge their benefits. Three w4D-Var schemes (henceforth referred to as Control, Weak Ensemble, and Weak Hybrid) are run concurrently in order to properly compare and contrast the results. For each assimilation system, the background and observation error covariances are specified as described above, whereas the model error statistics are set as follows.

- (Control) Mis-specified model error covariances specified as $\mathbf{Q} = 2\,\mathrm{diag}(\mathbf{Q}^t)$ and model bias $\mathbf{q} = \mathbf{0}$ is considered as the status-quo and serves as the basis for comparing against the other schemes.

- (Weak Ensemble) Use equations (8) and (10) from an ensemble size of 20. The background was perturbed using (4) and multiplicative factor $\beta = 10$, which was used to make sure the ensemble had sufficient spread. The localization matrix is obtained using the fifth-order rational function of Gaspari and Cohn [13] with correlation length 8.

- (Weak Hybrid) Use equations (12) and (13) with $\alpha = 0.5$. The static components are set to $\mathbf{Q}_c = 2\,\mathrm{diag}(\mathbf{Q}^t)$ and $\mathbf{q}_c = \mathbf{0}$ and the ensemble covariance with localization is the same one computed for the weak ensemble scheme.

Additionally, three strong-constraint 4D-Var schemes are considered. With the same observation error covariances as the w4D-Var systems, the background error covariance matrix is set as follows.

- (Strong 4D-Var) Set $\mathbf{B}$ as the matrix from the spin-up cycle.

- (Strong Ensemble) Set $\mathbf{B}$ as an ensemble estimate obtained from an ensemble of 20 strong-constraint 4D-Var assimilation systems. Further details are in the next paragraph.

- (Strong Hybrid) Use the hybrid $\mathbf{B} = \alpha\mathbf{B}_c + (1-\alpha)\mathbf{B}_e$, where the static component $\mathbf{B}_c$ is the covariance produced from the spin-up cycle and $\mathbf{B}_e$ is the background ensemble error covariance with localization computed for the strong ensemble scheme. The parameter α is also set to 0.5 like the weak hybrid scheme.

The ensemble-based schemes for the strong-constraint 4D-Var utilize background perturbations computed using (4) with multiplicative factor $\beta = 5$. In this case, the analysis $\mathbf{x}_0^a$ in (4) is obtained from a strong-constraint 4D-Var run. The observations $\mathbf{y}_i$ are perturbed in the same manner as described in subsection 2.1. Each analysis ensemble member $\mathbf{x}_{0,j}^a$ is forecast to the beginning of the next assimilation cycle and $\mathbf{B}_e$ is defined to be the sample covariance from the ensemble of these forecasts.

To better compare the performance of the control system to the other five assimilation schemes, the ratio of errors between the five other schemes and the control is considered. For example, if the weak ensemble scheme outperforms the control scheme, then the ratio of the weak ensemble errors to the control errors will be less than 1. A ratio of 1 means the schemes have the same performance and larger than 1 if the weak ensemble scheme performs worse than the control.

Figure 1 shows the ratio of the monthly (30-day) average analysis errors for each method to the control, that is, $\varepsilon^a(\langle\text{method}\rangle)/\varepsilon^a(\text{control})$, where the analysis error is the difference between the LZ96 x-values and the w4D-Var analysis. It is noticed that the strong hybrid and weak ensemble schemes have similar performance, whose ratios fluctuate near 1. The weak ensemble scheme does not seem to perform much better than the control scheme, however the weak hybrid errors show an improvement over the entire assimilation period. At month 7, an improvement of about 7% is achieved. The fact that the ensemble scheme ratios are sometimes slightly larger than 1 can be attributed to two important components: the factor β that controls the background ensemble spread and the ensemble size. The choice to set $\beta = 10$ for w4D-Var ensemble schemes was made to ensure that $\mathbf{Q}_e$ did not suffer from being orders of magnitude smaller than $\mathbf{Q}^t$. Due to the banded structure of the localization matrix, the model error correlations are not fully accounted for in the ensemble and hybrid methods. Still, enough

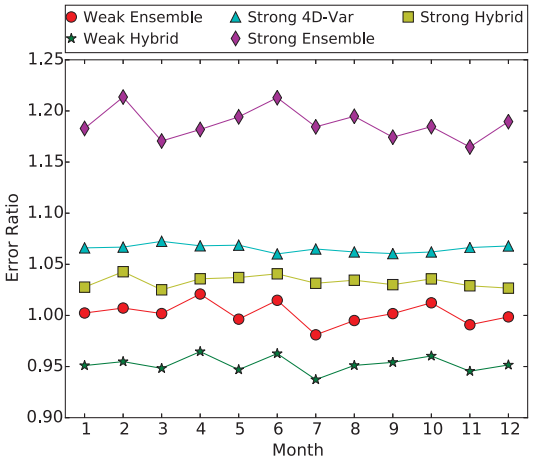

Figure 1: The ratio of the global monthly analysis error to the control w4D-Var experiment.

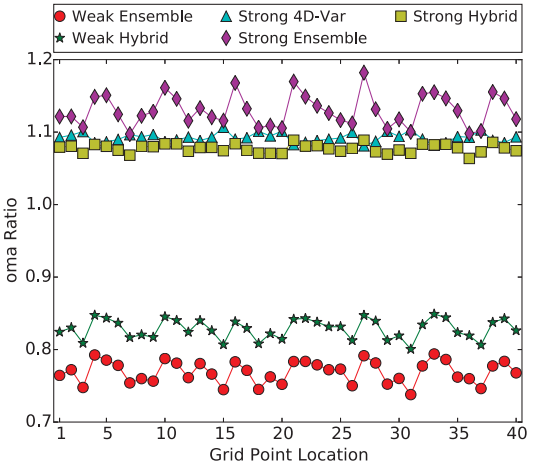

Figure 2: The ratio of the average oma differences to the control.

of the correlation structure was recovered from the ensembles to reduce the monthly hybrid w4D-Var error averages.

Figure 2 shows the ratio of the three-year averaged observed-minus-analysis (oma) $\|\mathbf{y}_i - \mathbf{h}_i(\mathbf{x}_i^a)\|$ difference for each grid point to the control. It shows that the analyses for the weak ensemble scheme better fit with the observations than the hybrid scheme, even though Figure 1 shows the hybrid scheme had a lower average analysis error. Since the background, observational, and model error components of w4D-Var are weighted by their corresponding inverse covariance matrices in the cost functional, the analysis fit to the observations is affected by the magnitudes of the error covariance matrices. In particular, $\mathbf{Q}_e$ having a larger magnitude than the hybrid model error covariance reduces the weight of model error in the analysis and increases the relative weights of the background and observations. Recalling that $\mathbf{B}$ and $\mathbf{R}$ remain unchanged between the two schemes, it can be inferred that $\mathbf{Q}_e$ has a larger magnitude and that the hybrid specification better represents the true model error statistics. Some evidence to support this conjecture is shown in Figure 3, which compares the prescribed model error variance to the ensemble and hybrid model error variances, obtained from the three-year average covariance matrices.

The performance of a hybrid data assimilation system is closely determined by the weight assigned to the static and ensemble-based components of the error covariances. This aspect is investigated by running the hybrid data assimilation scheme for different α for $\mathbf{Q} = \alpha\mathbf{Q}_c + (1-\alpha)\mathbf{Q}_e \circ \mathbf{C}$, where $\mathbf{Q}_c$ is the static component of $\mathbf{Q}$ specified as the control error covariance. For $\alpha = 0$, the system runs in ensemble mode while for $\alpha = 1$, the system runs as the control, the status quo. The weight α varies from 0 to 1 in steps of $\Delta\alpha = 0.025$ and the ratio of the time- and space-averaged analysis errors over a three year period to the control verses the choice of α is shown in Figure 4. The ensemble size for generating $\mathbf{Q}_e$ is 20, as before. The results show that the error corresponding to pure ensemble mode provides an improvement over the control data assimilation system by about 2.5%. Further reduction in the analysis error is achieved due to the specification of the hybrid covariance with $0 < \alpha \leq 0.95$. In particular, the hybrid covariance matrix corresponding to approximately $\alpha = 0.625$ provides the greatest reduction in the analysis error, about 7.5% improvement over the control.

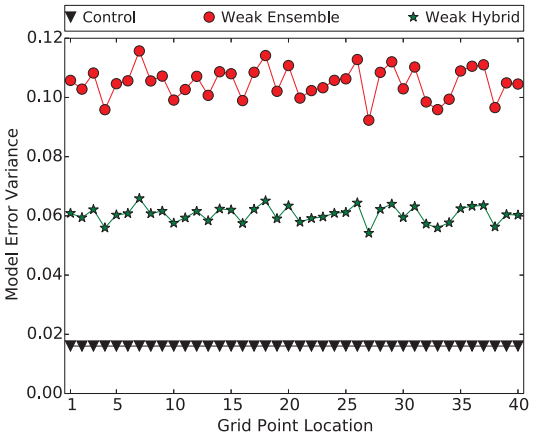

Figure 3: A comparison of the prescribed model error variance to the time-averaged ensemble and hybrid model error variances.

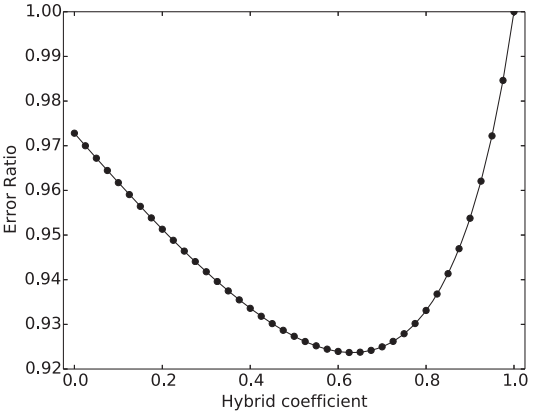

Figure 4: The ratio of the time- and space-averaged analysis errors to the control verses the hybrid scalar weight α for the model error covariances.

4 Conclusion

This article provides a framework for performing ensemble and hybrid data assimilation in a w4D-Var setting. A practical approach is considered that relies on an ensemble of w4D-Var systems solved by the incremental algorithm to obtain an ensemble of analysis sequences, the best estimates of the true state from which an ensemble of model error estimates are formed. These model error ensembles provide insight to the true nature of the model error covariance matrices. Model error bias has traditionally been assumed to be zero, and it may be possible to incorporate information about model error bias to improve the quality of the analysis with future research in this area.

In some situations, such as the case when the number of ensemble members is small, the ensemble covariance matrices will have low rank and may not be a completely reliable representation of the true model error statistics. The weighted combination of a static matrix, a diagonal matrix for example, and the ensemble covariance can prove to be an improvement over the ensemble matrices alone. A further improvement is to remove the random noise within the ensemble covariance by using a localization matrix.

The results of our numerical experiments provide a proof-of-concept for using ensembles in a w4D-Var setting. Specifying the model error covariances as the ensemble covariances with localization can improve the analysis error. Further improvement can be made in a hybrid setting with a good choice of the scalar weights.

Further research is needed to improve upon ensemble and data assimilation in w4D-Var. For a data assimilation window $[t_0, t_N]$ of $N+1$ times, N model error ensembles of size N_e are formed and are used to estimate each model error covariance $\mathbf{Q}_1, \ldots, \mathbf{Q}_N$. When the dimension of the state space is large, this can be computationally expensive, so it would be desirable to have a small ensemble size and still obtain a good estimate of the model error statistics. To do this becomes a question of how to optimally perturb the background and observations when forming the initial ensemble. Another improvement can be made by determining how to specify the hybrid scalar weights to get the best improvement in the quality of the analysis.

The numerical results in this study assumed an idealized observing system in which all

states in the dynamical system are observed. In practical applications, the performance of the data assimilation system is closely determined by the observing system configuration and further research is needed to investigate the performance of both ensemble and hybrid w4D-Var assimilation schemes.

Acknowledgment

This work was supported by the NASA Modeling, Analysis, and Prediction Program under award NNX13AN94G.

References

[1] H. Cheng, M. Jardak, M. Alexe, and A. Sandu. A hybrid approach to estimating error covariances in variational data assimilation. *Tellus A*, pages 1–15, dec 2011.

[2] A. M. Clayton, A. C. Lorenc, and D. M. Barker. Operational implementation of a hybrid ensemble/4D-Var global data assimilation system at the Met Office. *Quarterly Journal of the Royal Meteorological Society*, 139(675):1445–1461, jul 2013.

[3] P. Courtier, J. N. Thépaut, and A. Hollingsworth. A strategy for operational implementation of 4DVar, using an incremental approach. *Quarterly Journal of the Royal Meteorological Society*, pages 1367–1387, 1994.

[4] D. P. Dee. Bias and data assimilation. *Quarterly Journal of the Royal Meteorological Society*, 131(613):3323–3343, oct 2005.

[5] D. P. Dee and A. M. Da Silva. Data assimilation in the presence of forecast bias. *Quarterly Journal of the Royal Meteorological Society*, pages 269–295, 1998.

[6] D. P. Dee and R. Todling. Data assimilation in the presence of forecast bias: The GEOS moisture analysis. *Monthly Weather Review*, 128:3268–3282, 2000.

[7] D. P. Dee and S. Uppala. Variational bias correction of satellite radiance data in the ERA-Interim reanalysis. *Quarterly Journal of the Royal Meteorological Society*, 135(644):1830–1841, 2009.

[8] J. C. Derber. A variational continuous assimilation technique. *Monthly Weather Review*, 1989.

[9] G. Desroziers, J.-T. Camino, and L. Berre. 4DEnVar: link with weak-constraint 4D-Var and different possible implementations. *Quarterly Journal of the Royal Meteorological Society*, 140(October):2097–2110, 2014.

[10] J. P. Drécourt, H. Madsen, and D. Rosbjerg. Bias aware Kalman filters: Comparison and improvements. *Advances in Water Resources*, 29:707–718, 2006.

[11] G. Evensen. Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics. *Journal of Geophysical Research: Oceans*, 99(C5):10143–10162, 1994.

[12] D. Fairbairn, S. R. Pring, A. C. Lorenc, and I. Roulstone. A comparison of 4DVar with ensemble data assimilation methods. *Quarterly Journal of the Royal Meteorological Society*, may 2013.

[13] G. Gaspari and S. E. Cohn. Construction of correlation functions in two and three dimensions. *Quarterly Journal of the Royal Meteorological Society*, 125(April 1998):723–757, 1999.

[14] A. Griffith and N. Nichols. Adjoint methods in data assimilation for estimating model error. *Flow, Turbulence and Combustion*, 65:469–488, 2000.

[15] T. M. Hamill, J. S. Whitaker, and C. Snyder. Distance-dependent filtering of background error covariance estimates in an ensemble Kalman filter. *Monthly Weather Review*, 129:2776–2790, 2001.

[16] P. L. Houtekamer and H. L. Mitchell. A sequential ensemble Kalman filter for atmospheric data assimilation. *Monthly Weather Review*, pages 123–137, 2001.

[17] P. L. Houtekamer and H. L. Mitchell. Ensemble Kalman filtering. *Quarterly Journal of the Royal Meteorological Society*, 131:3269–3289, 2005.

[18] P. L. Houtekamer, H. L. Mitchell, G. Pellerin, M. Buehner, M. Charron, L. Spacek, and B. Hansen. Atmospheric Data Assimilation with an Ensemble Kalman Filter: Results with Real Observations. *Monthly Weather Review*, 133(3):604–620, 2005.

[19] K. Ide, P. Courtier, M. Ghil, and A. C. Lorenc. Unified Notation for Data Assimilation: Operational, Sequential and Variational. *Journal of the Meteorological Society of Japan*, 75(1B):181–189, 1997.

[20] F. Le Dimet and O. Talagrand. Variational algorithms for analysis and assimilation of meteorological observations: theoretical aspects. *Tellus A*, 1986.

[21] A. C. Lorenc. The potential of the ensemble Kalman filter for NWP a comparison with 4D-Var. *Quarterly Journal of the Royal Meteorological Society*, 129(595):3183–3203, oct 2003.

[22] E. N. Lorenz. Predictability: A problem partly solved. *Proc. Seminar on predictability*, 1996.

[23] E. N. Lorenz and K. A. Emanuel. Optimal Sites for Supplementary Weather Observations: Simulation with a Small Model. *Journal of the Atmospheric Sciences*, 55(3):399–414, feb 1998.

[24] L. Mitchell and A. Carrassi. Accounting for model error due to unresolved scales within ensemble Kalman filtering. *Quarterly Journal of the Royal Meteorological Society*, 2014.

[25] Y. Trémolet. Accounting for an imperfect model in 4D-Var. *Quarterly Journal of the Royal Meteorological Society*, 132(621):2483–2504, oct 2006.

[26] D. Zupanski. A general weak constraint applicable to operational 4DVAR data assimilation systems. *Monthly Weather Review*, pages 2274–2292, 1997.

[27] D. Zupanski and M. Zupanski. Model Error Estimation Employing an Ensemble Data Assimilation Approach. *Monthly Weather Review*, 134(1994):1337–1354, 2006.