PUBLISHED BY THE INTERNATIONAL METEOROLOGICAL INSTITUTE IN STOCKHOLM

Tellus (2009), 61A, 539–549

TELLUS

Data assimilation in a system with two scales—combining two initialization techniques

By JOAQUIM BALLABRERA-POY[1]*, EUGENIA KALNAY[2] and SHU-CHIH YANG[2,3], [1]*Institut de Ciències del Mar, CSIC, Passeig Marítim de la Barceloneta, 37-49, 08003 Barcelona, Spain;* [2]*AOSC, University of Maryland, College Park, MD, USA;* [3]*Global Modeling and Assimilation Office (GMAO), NASA/GSFC, Greenbelt, MD, USA*

(Manuscript received 14 April 2008; in final form 5 March 2009)

ABSTRACT

An ensemble Kalman filter (EnKF) is used to assimilate data onto a non-linear chaotic model, coupling two kinds of variables. The first kind of variables of the system is characterized as large amplitude, slow, large scale, distributed in eight equally spaced locations around a circle. The second kind of variables are small amplitude, fast, and short scale, distributed in 256 equally spaced locations. Synthetic observations are obtained from the model and the observational error is proportional to their respective amplitudes. The performance of the EnKF is affected by differences in the spatial correlation scales of the variables being assimilated. This method allows the simultaneous assimilation of all the variables. The ensemble filter also allows assimilating only the large-scale variables, letting the small-scale variables to freely evolve. Assimilation of the large-scale variables together with a few small-scale variables significantly degrades the filter. These results are explained by the spurious correlations that arise from the sampled ensemble covariances. An alternative approach is to combine two different initialization techniques for the slow and fast variables. Here, the fast variables are initialized by restraining the evolution of the ensemble members, using a Newtonian relaxation toward the observed fast variables. Then, the usual ensemble analysis is used to assimilate the large-scale observations.

1. Introduction

Current short-term (e.g. seasonal) climate predictions require the initialization of coupled models that exhibit a wide range of dynamic, and thermodynamic, phenomena interacting at various time- and space-scales (Meehl et al., 2001). The range of phenomena may even increase because of the growing awareness about the need to increase the resolution of coupled models to improve the skill of seasonal predictions (Saha et al., 2006). In such a context, assimilation of synoptic features or other small-scale features may become difficult, as they may evolve uncorrelated to observed parameters or features. Therefore, understanding the abilities of different data assimilation techniques to simultaneously reconstruct information at different scales is of importance for climate predictions.

Initialization of short-term climate or numerical weather prediction (NWP) models is currently performed using variational, either three- or four-dimensional, and/or via ensemble covariance estimations (see Kalnay et al., 2007 and references therein). The specific difficulties of ensemble methods to simultaneously reconstruct information at different scales are suggested by Lorenc (2003), based on the fact that finite-size ensembles can only fit a number of observations not larger than the number of ensemble members. Lorenc (2003) argues that, in the context of local analyses (usual strategy to deal with the spurious teleconnections that frequently taint sampled covariances), the size of the local domain must be large enough to facilitate the dynamic balance of large-scale fields such as temperature. As the analysis domain increases, fitting the observations could fail if: (1) small-scale fields, such as humidity, are thoroughly sampled and (2) the ensemble is not large enough. Failure to simultaneously fit all the observations will reduce the accuracy of the analysis and, then, the accuracy of forecasts (Lorenc, 2003). Such a concern differs from the usual initialization problem in NWP, which focuses on the identification of the balanced component (the slow manifold) in model initial states to filter out the detrimental development of fast gravity waves or other unbalanced components (the fast manifold).

The goal here is to asses, and increase if necessary, the accuracy of analyses when both large- and small-scale variables are simultaneously assimilated. More precisely, this work investigates the ability of an ensemble method to simultaneously identify the state of a strongly non-linear chaotic system with

*Corresponding author.
e-mail: joaquim@icm.csic.es
DOI: 10.1111/j.1600-0870.2009.00400.x

two separated spatio-temporal scales, when the small-scale signal is severely undersampled.

The ensemble Kalman Filter (EnKF) and the dynamic model are introduced are introduced in Section 2. The results of the data assimilation experiments are described in Section 3. The result of combining two different initialization methods (one for the large-scale variables and another for the short-scale ones) are shown in Section 4. Finally, Section 5 includes a summary and final remarks.

2. Methods

2.1. The ensemble Kalman Filter

Sequential data assimilation methods, based on the Kalman Filter (Kalman and Bucy, 1960), are being used in oceanography and meteorology either by closely following its original formulation but in coarse resolution models (e.g. Miller and Cane, 1989) or in reduced-dimensional subspaces in high resolution models (e.g. Cane et al., 1996; Evensen, 1994). Either way, transmission of the information from the innovation vector to the model space is given in a statistical manner:

$$\mathbf{x}_k^{\mathrm{a}} = \mathbf{x}_k^{\mathrm{b}} + \mathbf{P}_k^{\mathrm{b}}\mathbf{H}_k^T \left[\mathbf{H}_k\mathbf{P}_k^{\mathrm{b}}\mathbf{H}_k^T + \mathbf{R}_k\right]^{-1}\left(\mathbf{y}_k^{\mathrm{o}} - \mathbf{H}_k\mathbf{x}_k^{\mathrm{b}}\right). \tag{1}$$

See Kalnay (2003) and references therein for a detailed derivation and discussion of eq. (1). The convention used here is to write matrices as uppercase and vectors as lowercase. Moreover, $\mathbf{x}$ represents a n-dimensional vector (the state vector) defining the analysis space, $\mathbf{y}$ is the p-dimensional vector with the available observations; and $\mathbf{H}$ is the forward observation operator that maps the analysis space onto the observation space. The subscript k refers to the time step, the superscript b stands for the background (i.e. our knowledge of the state of the system in absence of observations) and superscript a stands for the analysis. The innovation vector, $\mathbf{d} = \mathbf{y}^{\mathrm{o}} - \mathbf{H}\mathbf{x}^{\mathrm{b}}$, accounts for the fraction of the information in the observations not being accounted for by the background field. The difference between the analysis and the background is the analysis update. Thus, eq. (1) estimates the analysis update as the weighted combination of the innovation vector, where the weights depend on the estimated background error covariance, $\mathbf{P}^{\mathrm{b}}$, and the estimated observational error covariance, $\mathbf{R}$.

The EnKF applied here closely follows the singular evolutive extended Kalman (SEEK) filter applied to a non-linear system (Ballabrera-Poy et al., 2001) and the local ensemble Kalman transform Kalman filter (LETKF) developed by Hunt et al. (2007). At each analysis step time t_k, the algorithm estimates the most probable state of the system through the average of an ensemble of r solutions of the model, $\mathbf{X}_k = (\mathbf{x}_k^1, \mathbf{x}_k^2, \ldots, \mathbf{x}_k^r) \in \mathbb{R}^{n\times r}$, whose spread is proportional to the degree of uncertainty associated with the current state of the system

$$\mathbf{P}_k^{\mathrm{b}} = \frac{\rho}{r-1}\left(\mathbf{X}_k - \bar{\mathbf{X}}_k\right)\left(\mathbf{X}_k - \bar{\mathbf{X}}_k\right)^T. \tag{2}$$

The overline indicates the average over the ensemble. The constant ρ is known as forgetting factor or covariance inflation. The algorithm does not explicitly evaluate the error covariance matrix, but extensively exploits its definition:

$$\begin{aligned}\mathbf{H}\mathbf{P}_k^{\mathrm{b}}\mathbf{H}^T &\approx \frac{\rho}{r-1}\left(\mathbf{H}\mathbf{X}_k - \overline{\mathbf{H}\mathbf{X}}_k\right)\left(\mathbf{H}\mathbf{X}_k - \overline{\mathbf{H}\mathbf{X}}_k\right)^T,\\ \mathbf{P}_k^{\mathrm{b}}\mathbf{H}^T &\approx \frac{\rho}{r-1}\left(\mathbf{X}_k - \overline{\mathbf{X}}_k\right)\left(\mathbf{H}\mathbf{X}_k - \overline{\mathbf{H}\mathbf{X}}_k\right)^T.\end{aligned} \tag{3}$$

Equation (3) assume that the forward observation operator, $\mathbf{H}$, and the average operator permute (which is indeed the case when the forward observation operator is linear). Making use of eq. (2) and the Sherman–Morrison–Woodbury formula, eq. (1) may be written as

$$\begin{aligned}\mathbf{x}_k^{\mathrm{a}} &= \mathbf{x}_k^{\mathrm{b}} + \mathbf{E}_k\Lambda^{\mathrm{a}}\left(\mathbf{H}\mathbf{E}_k\right)^T\mathbf{R}_k^{-1}\left(\mathbf{y}_k^{\mathrm{o}} - \mathbf{H}_k\mathbf{x}_k^{\mathrm{b}}\right),\\ \Lambda^{\mathrm{a}} &= \left[\frac{r-1}{\rho}\mathbf{I} + \left(\mathbf{H}\mathbf{E}_k\right)^T\mathbf{R}_k^{-1}\left(\mathbf{H}\mathbf{E}_k\right)\right]^{-1},\end{aligned} \tag{4}$$

where $\mathbf{x}_k^{\mathrm{b}} = \bar{\mathbf{X}}_k$ and $\mathbf{E}_k = \mathbf{X}_k - \bar{\mathbf{X}}_k$. Notice that large covariance inflation coefficients reduce the size of the term being added to the diagonal, this effect being more noticeable for small ensembles. Thus, although the first term inside the bracket does increase the diagonal weight of a matrix (reducing the condition number of the matrix to be inverted), large covariance inflations reduce the conditionality of the matrix inversion in eq. (4). On the other hand, and following Hunt et al. (2007), the matrix inversion of eq. (4) is done through the eigenvalue factorization of the right-hand side of the second equation:

$$\begin{aligned}\Lambda^{\mathrm{a}} &= \Lambda^{\mathrm{b}^{-1}} = \Pi^{\mathrm{a}}\,\Pi^{\mathrm{a}^T},\\ \Pi^{\mathrm{a}} &= \Pi^{\mathrm{b}}\mathbf{W}^{-1/2},\\ \Lambda^{\mathrm{b}} &= \frac{r-1}{\rho}\mathbf{I} + \left(\mathbf{H}\,\mathbf{E}_k\right)^T\mathbf{R}_k^{-1}\left(\mathbf{H}\,\mathbf{E}_k\right) = \Pi^{\mathrm{b}}\,\mathbf{W}\,\Pi^{\mathrm{b}^T}.\end{aligned} \tag{5}$$

Matrices Π^{b} and $\mathbf{W}$ contain the eigenvectors and eigenvalues of Λ^{b}, respectively. The square form of eq. (5) guarantees that Λ^{a} has indeed the properties of a covariance matrix. The covariance of the analysis error is

$$\mathbf{P}_k^{\mathrm{a}} = \mathbf{E}_k\Lambda^{\mathrm{a}}\mathbf{E}_k^T = \mathbf{E}_k\Pi^{\mathrm{a}}\left(\mathbf{E}_k\Pi^{\mathrm{a}}\right)^T, \tag{6}$$

which can be written as

$$\mathbf{P}_k^{\mathrm{a}} = \frac{1}{r-1}\mathbf{E}_k^{\mathrm{a}}\mathbf{E}_k^{\mathrm{a}^T} \tag{7}$$

by defining the directions of the analysis error as

$$\mathbf{E}_k^{\mathrm{a}} = (r-1)^{1/2}\mathbf{E}_k\Pi^{\mathrm{a}}. \tag{8}$$

As noted by Hunt et al. (2007), eq. (8) guarantees that the ensemble defining the analysis error is unbiased by respect the ensemble mean. Finally, the new ensemble is given by

$$\mathbf{X}_k^{\mathrm{a}} = \mathbf{x}_k^{\mathrm{a}} + \mathbf{E}_k^{\mathrm{a}}, \tag{9a}$$

where $\mathbf{x}_k^{\mathrm{a}}$ is given by eq. (4). The ensemble of analysis states given by (9a) is being used as initial conditions of the model to dynamically propagate the members of the ensemble up to

the next analysis time, t_{k+1}, where eqs. (2)–(9a) will provide the new estimation of the state of the system and its expected error. As in Corazza et al. (2007), random initial perturbations may also be applied after every analysis step to allow the ensemble to explore the error growth on a space larger than the one given by the r-directions of the analysis space:

$$\mathbf{X}_k^{\mathrm{a}} = \mathbf{x}_k^{\mathrm{a}} + \mathbf{E}_k^{\mathrm{a}} + \sigma\ \mathbf{s} \bullet \eta, \tag{9b}$$

In this equation, $\mathbf{s} \bullet \eta$ is the Schur product between $\eta \in N(0, 1)$, a normal random variable of zero mean and unit variance, and $\mathbf{s}$, the vector containing the appropriate scaling for each component of the state vector (defined in Section 3.2). The amplitude of the random noise is determined by the coefficient σ. Equation (9b) is, thus, used instead of eq. (9a).

To emphasize the importance of the background error covariance matrix (either in its full version or estimated from a finite-size ensemble), eq. (1) may be rewritten, dropping the time subscript, as

$$\begin{aligned} \mathbf{x}^{\mathrm{a}} &= \mathbf{x}^{\mathrm{b}} + \mathbf{P}^{\mathrm{b}}\,\mathbf{x}^{\mathrm{d}}, \\ \mathbf{x}^{\mathrm{d}} &= \mathbf{H}^T \left[\mathbf{H}\,\mathbf{P}^{\mathrm{b}}\,\mathbf{H}^T + \mathbf{R}\right]^{-1} \mathbf{d}. \end{aligned} \tag{1b}$$

Here, $\mathbf{x}^{\mathrm{d}}$ is the n-dimensional vector, constructed by mapping the weighted innovation vector from the observation locations back to the analysis space, via the transpose of the forward observation operator. The components of the vector $\mathbf{x}^{\mathrm{d}}$ are all zeros except for those components that directly map onto the observation space. The key role of the $\mathbf{P}^{\mathrm{b}}$ is clearly understood by considering the case in which a single observation (corresponding to the Kth component of the state vector, x_K) is being assimilated. In this case, all the components of the vector $\mathbf{x}^{\mathrm{d}}$ are zero, except for its Kth component, which is equal to $(y^{\mathrm{o}} - x_K^{\mathrm{b}})/(P_{KK}^{\mathrm{b}} + R)$. Equation (1b) evidences that the spatial distribution of the analysis update is obtained by statistically projecting the information from the component K to the rest of the space via the covariances between the background error at K with the background error of all the other points of the system. If the point K were statistically uncorrelated with the rest of the system, no components other than K would be corrected by eq. (1).

2.2. The model

The equations of the model are

$$\begin{aligned} \frac{\mathrm{d}X_i}{\mathrm{d}t} &= X_{i-1}\,(X_{i+1} - X_{i-2}) - X_i + F + G_i, i = 1, \ldots, I, \\ \frac{\mathrm{d}Y_{j,i}}{\mathrm{d}t} &= cbY_{j+1,i}\left(Y_{j-1,i} - Y_{j+2,i}\right) \\ &\quad - cY_{j+1,i} + H_i, j = 1, \ldots, J, \\ G_i &= -h\tfrac{c}{b}\sum_{j=1}^{J} Y_{j,i}, \\ H_i &= h\tfrac{c}{b}X_i. \end{aligned} \tag{10}$$

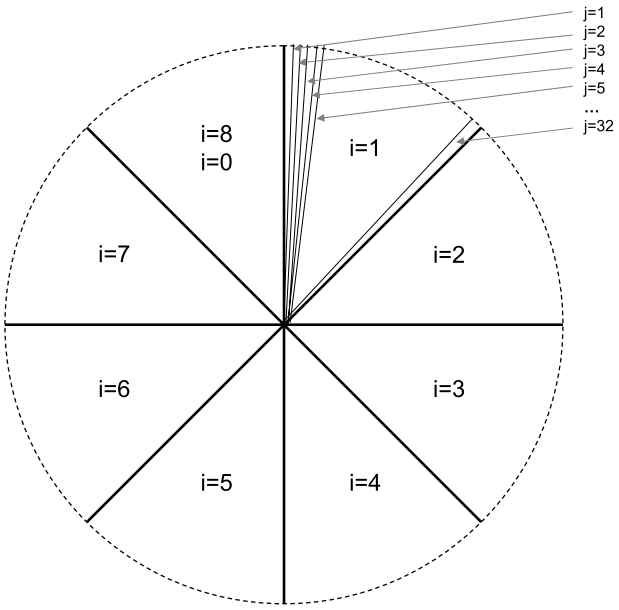

Fig. 1. Spatial representation of the model. The system is given as eight periodic cells diving a circle. Each cell has an associated X variable representing a climate average. Each one of these cells includes 32 gridpoints, over which a small-scale variable Y is defined.

The dynamic equations for X are based on the model of Lorenz and Emanuel (1998), which has been profusely used for testing methodological approaches in data assimilation (see, among others, Lorenz and Emanuel, 1998; Anderson, 2001; Whitaker and Hamill, 2002; Ott et al., 2004; Descamps and Talagrand, 2007). The dynamic equations for X have an additional term, G_i, that links the time evolution of X_i with the average value of $Y_{j,i}$, $j = 1, \ldots, J$. The 'spatial' distribution of these variables is shown in Fig. 1. The X variables are cyclic, and each 'X-cell' contains 32 Y-variables; so the total number of model variables is $264 = 8 + 8 \times 32$. Equations for Y are formally the same than for X, with the only differences being the side of the advection term and the nature of the forcing term. In both sets of equations, the non-linear term simulates advection while conserving the energy of X and Y. As stated by Lorenz and Emanuel (1998), these equations may be looked at as the evolution of an unspecified scalar meteorological variable, as vorticity or temperature, around a latitude circle with no latitude or vertical dependence. The values used in this work are $I = 8$, $J = 32$, $b = 10$, $c = 10$, $h = 1$ and $F = 18$. The value of $c = 10$ indicates that Y decays 10 times faster than X. The relative size of both kinds of variables is controlled by the amount of energy drawn from the slow variables to the fast variables, that is, the value of the coefficient h.

The time units were defined by Lorenz and Emanuel (1998) to be such that 0.05 units correspond to 6 h. The time step of the model is 0.01 units, equivalent to 20 time steps per day. A fourth-order Runge–Kutta scheme is used to advance the time

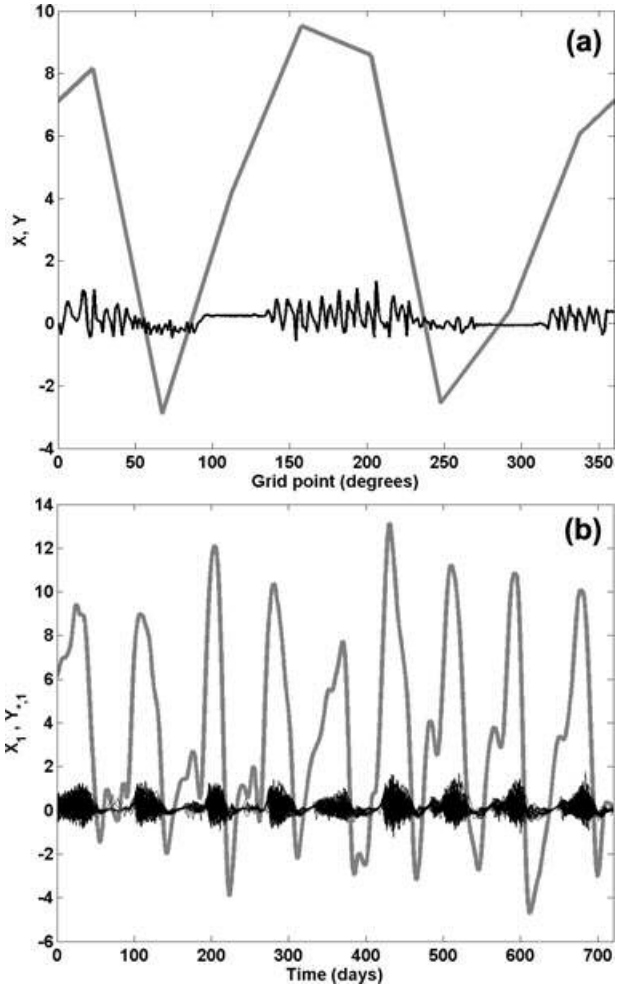

Fig. 2. (a) Snapshot of the spatial distribution of a solution at an arbitrary time. (b) Time evolution of the variables of the first cell.

steps of the model. The system is initialized from random initial conditions and spun up for 10 yr (of 360 d). Then, a reference simulation is run for an additional 10-yr period, from which the statistical properties of the model are evaluated. On the other hand, the first year of the reference solution is sampled to build a set of observations to be assimilated on the model.

A snapshot of the model is shown in Fig. 2a, and the time evolution of the variables in the first segment (i.e. X_1 and $Y_{*,1}$) is shown in Fig. 2b. In both cases, the thick line corresponds to X and the thin lines correspond to Y. The amplitude of the slow variable is one order of magnitude larger than the amplitude of the fast variables (the average of the standard deviation of variables X is 4.5, whereas the average standard deviation of Y is 0.29). These plots show that the amplitude of the envelope of the fast variables is modulated by the amplitude of the slow one. The average autocorrelation for both types of variables is shown in Fig. 3. The decorrelation time is simply estimated as the e-folding timescale, that is, the scale for which the correlation becomes smaller than 0.37. Thus, the decorrelation time of variables X is of about 20 d. Variables Y have a decorrelation time of about 2 d. The results shown in Fig. 3 allows concluding that although some variables Y may have a close value at some given time (gridpoints 100–140 in Fig. 2b), their time evolution is such that they rapidly become uncorrelated from each other.

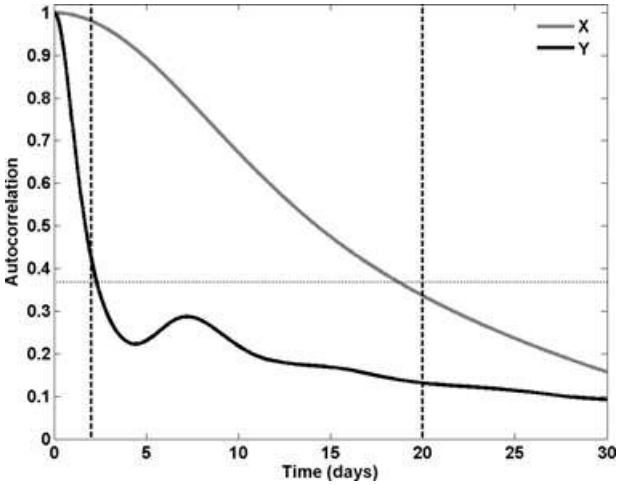

Fig. 3. Average of the autocorrelation of the variables of the model (the horizontal dotted line represents the e-folding value).

The spatial cross-correlations (X_1, X_*) and (X_1, $Y_{*,*}$)are displayed in Figs. 4a and b. The thick grey line indicates (X_1, X_*), and the thin black line corresponds to (X_1, Y_*). Again, e-fold lines are superposed to identify non-significant correlations. Thus, Fig. 4a indicates that X_1 displays long-range correlations while being uncorrelated with all the Y variables outside its own segment. The correlations ($Y_{1,1}$, X_*) and ($Y_{1,1}$,$Y_{*,*}$) are shown in Fig. 4b. The thick grey lines indicate correlation against X variables, and the think black lines indicate correlation against the Y variables. The variable $Y_{1,1}$ shows no correlation with any other variable other than the X from its own segment.

Therefore, the system described by eqs. (10) couples two variables with different spatial and temporal scales. The large-amplitude variables are slow and display long-range correlations, whereas the small-amplitude variables decorrelate rapidly and are uncorrelated from each other (either if they belong to the same segment or if they belong to different segments).

3. Experiments

3.1. Set up and observational error

The experimental set up is based on twin experiments in which the trajectory to reconstruct comes from the model itself. The period of the data assimilation experiments is 1 yr. The first year of the reference simulation described in Section 2.2 is defined as the truth. Observations are obtained by sampling the truth and adding a random observational error. Observations are

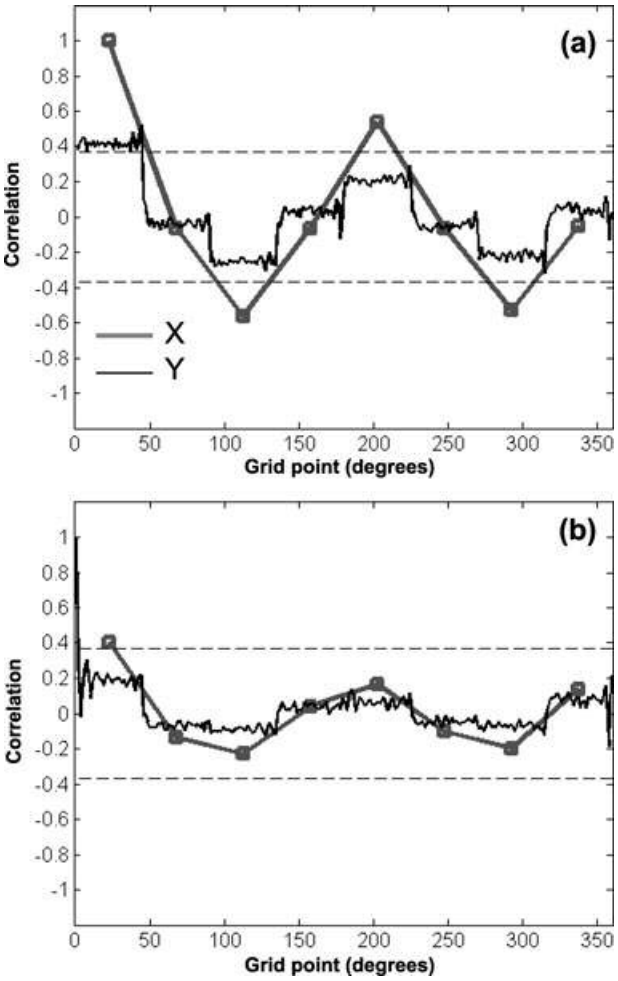

Fig. 4. (a) Spatial correlation of X_1 with all the variables of the system. (b) Spatial correlation of $Y_{1,1}$ with all the variables of the system. Thick lines indicate correlation with X variables and thin lines are correlations with Y variables.

assimilated every 6 h, that is, every five time steps. The spatial distribution of observations in each assimilation experiment will be specified later. A false initial condition is used to initialize the data assimilation experiment. For each data assimilation experiment, comparison is made with the experiment where no observations are assimilated. Assessment of the ability of the assimilation to bring the system toward the truth is given separately for each kind of variable (slow and fast).

The amplitude of the observational noise is taken proportional to the amplitude of the signal, using the same ratio for both kinds of variables. However, different approaches may be used to estimate the range of each kind of variable. Let X and Y be the matrices obtained by pooling together all X_{ji} and Y_{ji} variables, respectively. If the amplitude of the signal were defined by its range (i.e. difference between the maximum and the minimum), the amplitudes of X and Y would be 20.67 and 2.62, indicating that slow variables are almost eight times larger than the fast ones. However, the standard deviation of X and Y are 4.54 and 0.29, respectively, whose ratio is twice (15.6) the one calculated from their ranges. This disparity indicates that the range strongly overestimates the size of the fast variables, which would result as an artificially large error for them. As an alternative, the interquartile range (IQR, i.e. the difference between the 75 and the 25 percentiles) is used here to define the amplitude of the signal. In the case under consideration, the IQR values are 7.21 and 0.31. Finally, the amplitude of the noise is taken, approximately, as the 15% of the IQR, being 1.0 for X and 0.05 for Y.

3.2. Filter initialization

The EnKF requires an initial ensemble, whose ensemble average represents the first guess and whose spread parametrizes the expected error of the first guess. For the experiments below, the algorithm is initialized using the mean state of the 10-yr reference simulation and the variance around the mean during the same period. Notice that variance of the system gives a measure of the faithfulness of the mean in representing the true state of the system. For example, regions where the variance is near zero indicates regions where all the solutions of the model are very close to the same value and the long-term mean is a good guess about the state of the system at that region. On the contrary, regions with a large variance are regions where the value of the model has a large spread and the mean is a less accurate representation of the state of the system.

The initial ensemble is constructed from the multivariate empirical orthogonal functions (MEOFs) of the daily averaged outputs of the model for the 10 yr of the reference simulation. The multivariate vector $\mathbf{z}_k = (s_x X_k,\, s_y Y_k) \in \mathbb{R}^{264}$ is constructed by gathering all the 264 model variables and scaling them by the inverse of their standard deviation (4.54 for all the slow variables and 0.292 for all the fast ones). This multidimensional vector is written as $\mathbf{z}_k = \mathbf{s} \bullet \mathbf{x}_k$, where $\mathbf{x}_k = (X_k, Y_k)$. The first 250 MEOFs are retained and rescaled to reintroduce the variance lost by this truncation. The initial ensemble, with r members, is given by

$$\mathbf{X}_{ij} = \bar{\mathbf{x}}_i + \mathbf{s}_i^{-1} \sum_{k=1}^{250} \sqrt{D_k}\mathbf{S}_{ik}\eta_{kj}, \quad i = 1, \ldots, 264,\ j = 1, \ldots, r, \tag{11}$$

where $\bar{\mathbf{z}}$ is the long-term mean of the multivariate vector, $\mathbf{D}$ is the diagonal matrix with the eigenvalues, $\mathbf{S}$ is the rectangular matrix with the eigenvectors and $\eta_{kj} \in N(0, 1)$; so, $\eta\eta^T/(r-1) \approx \mathbf{I}$.

3.3. Measuring the performance of the assimilation

One of the advantages of twin experiments is that they confer a perfect knowledge about the state to be reconstructed, simplifying the measure of the performance of the assimilation. The results of the data assimilation will be given separately for the slow variables and the fast ones. For each kind of variables, the

distance between the analysis will be given in terms of the root mean square (rms) of the difference between the true value and the reconstructed one:

$$\text{rms}_X = \left[\frac{1}{I}\sum_{i=1}^{I}\left(X_i - X_i^t\right)^2\right]^{1/2},$$
$$\text{rms}_Y = \left[\frac{1}{I\cdot J}\sum_{\substack{j=1\\ i=1}}^{J,I}\left(Y_{ji} - Y_{ji}^t\right)^2\right]^{1/2} \quad (12)$$

Although the rms varies with time, the results reported below correspond to the average value over the last half of the experiment (i.e. the average over the last 6 months) in the expectation that the life of any initial transient is shorter than 6 months. The resulting rms values are compared with the rms of a free-evolving simulation starting from the same initial condition and no data assimilation. Comparison against the free-evolving simulation allows to account for those situations where the model might converge, by itself, towards the true solution.

3.4. Assimilation of all observations (EN1)

During a first set of experiments observations at all gridpoints are assimilated (i.e. $\mathbf{H} = \mathbf{I}$) every 6 h. These experiments are used to study the sensitivity of the assimilation to the various choices for the ensemble size, r; the covariance inflation, ρ, and the amplitude of the added random fluctuations, σ. The results are evaluated in terms of the rms analysis error, normalized by the rms error from no assimilation. Figure 5 indicates that the ensemble size and the amplitude of the random perturbation have a larger impact on the rms of the reconstructed states than the amplitude of the covariance inflation. Figure 5a shows the 6-month average rms as a function of the ensemble size. For small ensembles, large-scale variables have a smaller rms than the short-scale variables. Ensembles of size 50 allow a similar degree of convergence for both kinds of signals. Larger ensembles provide little reduction of the error. Figure 5b illustrates the low sensitivity of the rms error to the value of the covariance inflation ($r = 50$). The amplitude of the noise has a large impact on the performance of the filter (Fig. 5c). The largest errors are obtained when $\sigma = 0$ and diminish as the amplitude of the noise increases, until a minimum is reached (around the value of 0.08). Note that the error obtained using $\sigma = 1$ (i.e. an order of magnitude larger than the optimal value) is about 60% smaller than when the ensemble is integrated without adding any noise ($\sigma = 0$).

Figure 6 shows the rms for the slow (6a) and fast (6b) variables, when all data are assimilated using $r = 50$, $\rho = 1.0$, $\sigma = 0.10$ in an experiment called Ensemble 1 (EN1). The figures compare the analysis error from the assimilation with the error of the free-evolving simulation (FREE experiment) starting from the same initial field. Also superposed is the observational error (i.e. 1.0 and 0.05 for X and Y). Table 1 displays the

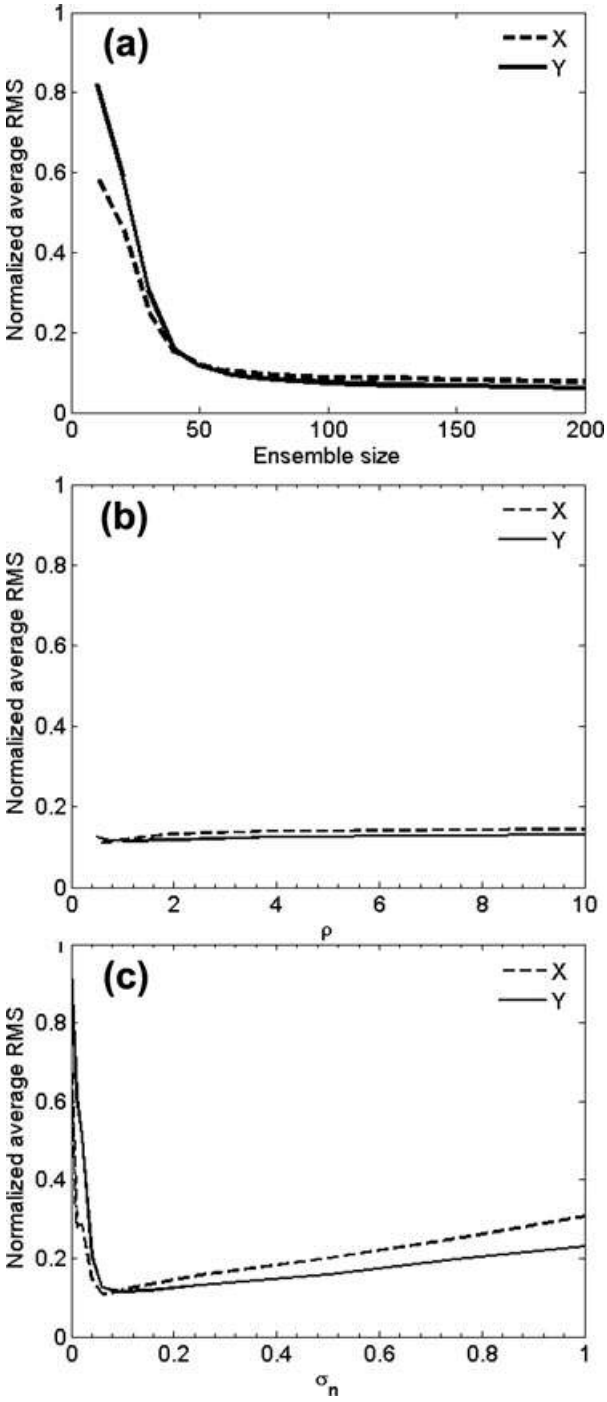

Fig. 5. Six-month average of the rms error of the assimilation as a function of: (a) ensemble size; (b) covariance inflation and (c) the size of the random perturbations. Results are normalized by the rms of the error of the free simulation.

6-month average of the analysis error and the error of the FREE experiment. The analysis error, for either slow or fast variables, is smaller than the corresponding observational error, indicating that the assimilation algorithm allows constraining both kinds of

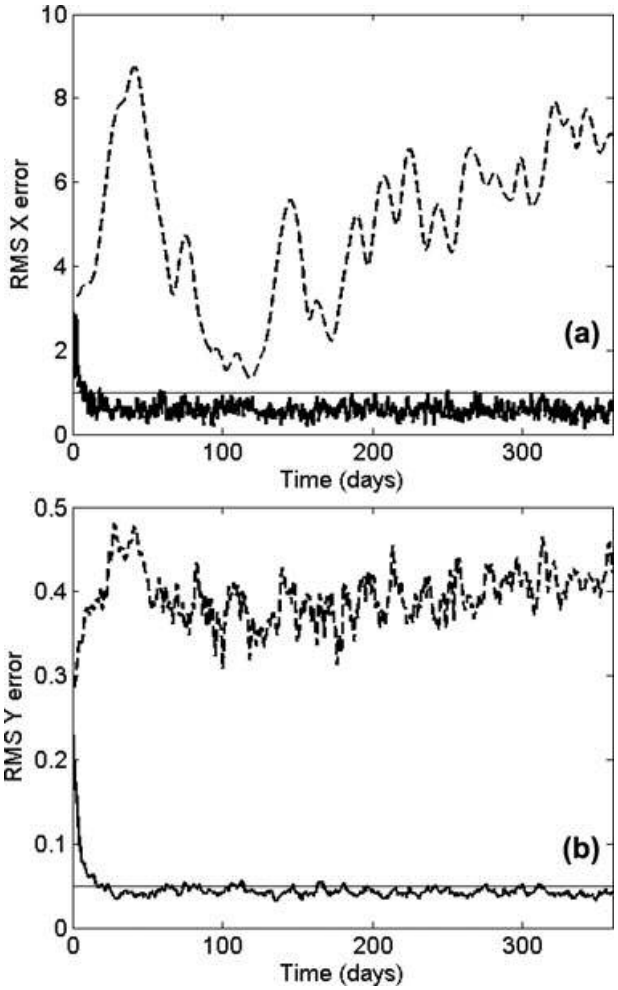

Fig. 6. The rms error after each analysis step for the X (a) and Y (b) variables. All variables are assimilated. The thin straight line indicates the amplitude of the corresponding observational error. Dashed lines indicate the error of the free simulation.

variables with the similar accuracy level, although it seems as if the slow variables are better reconstructed. Figure 6 shows that no trends are present in the time evolution of the error, suggesting the stability of the implemented EnKF. As a summary, when all the observations are assimilated, no differences appear in the behaviour of the reconstructed variables due to their differences in the length of the decorrelation scale or amplitude. In the context of the problem pointed out by Lorenc (2003), it indicates that a 50-member ensemble (small compared with the size of the system) may still simultaneously fit large- and small-scale fields when small-scale fields are exhaustively sampled.

3.5. Assimilation of X observations (EN2)

The next experiment of ensemble assimilation, called EN2, only assimilates X observations (i.e. $p = 8$). The rms of the error

Table 1. The rms of the error of the free run (FREE) and the five assimilation experiments described in the text

EXP	rms_X	rms_Y
FREE	6.18	0.41
EN1	0.60	0.05
EN2	0.47	0.29
EN3	0.65	0.39
CI1	0.48	0.02
CI3	0.48	0.27

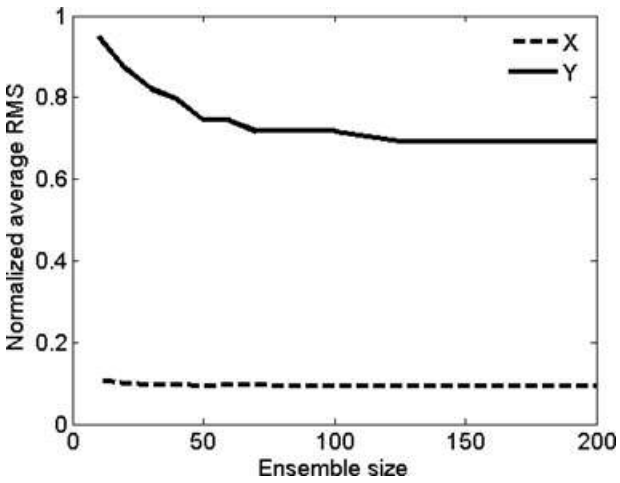

Fig. 7. Six-month averaged rms error after each analysis step as a function of the ensemble size. Only X variables are assimilated.

(Table 1) indicates that the reconstruction of the fast signal has an error 60 times larger than when all observations are assimilated. Although this does not come as a surprise, it is interesting to note that the lack of convergence of the fast signal does not deteriorate the reconstruction of the slow signal. On the contrary, the analysis error of the X variable is smaller when X observations are assimilated alone. In the same vein, Fig. 7 shows that an ensemble with only 10 members is already able to make the slow signal converge. This fact contrasts with the sharp increase of the error shown in Fig. 5a, when the size of the ensemble is reduced. These results can be explained in terms of overfitting, as the number of observations to be fit is reduced whereas the number of degrees of freedom of the filter is kept constant.

3.6. Assimilation of all X but oversampled Y observations (EN3)

The next batch of experiments, called EN3, assimilates all the large-scale variables (eight observations) together with a few observations of the high frequency variables (16 evenly distributed observations), i.e. $p = 24$. The assimilation experiment with $r = 50$, $\rho = 1.0$, $\sigma = 0.10$ fails because one of the members of the

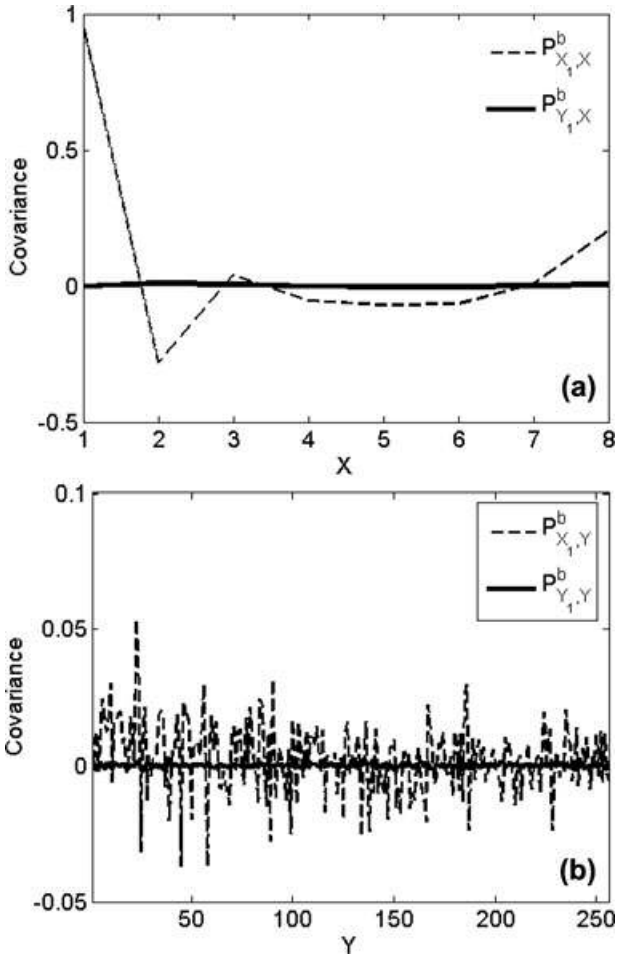

Fig. 8. Background covariance between X_1 (dashed) and Y_1 (continuous) and all the X variables (a) and Y variables (b). Obtained from the ensemble members at day 360 from experiment EN3.

ensemble becomes unstable and dies. The assimilation fails for all the experiments in which $\sigma < 0.12$. The filter becomes stable again, once the amplitude of the noise is larger than 0.13. Defining strategies to ensure the stability of the filter by replacing unstable members is beyond the scope of this work, and results are shown for those combinations of parameters for which the filter is stable. Experiment EN3 shown in Table 1 corresponds to $\sigma = 0.15$.

A surprising feature of the values in Table 1 is that the rms of the error is significantly worse for both X and Y than the errors obtained from EN2 (no assimilation of Y observations). Such an error increase cannot be explained by the arguments of Lorenc (2003), because the 50-member ensemble was previously able to fit sets of 264 and 8 observations in experiments EN1 and EN2, respectively. Moreover, although increasing the ensemble size reduces the error (Fig. 4a indicates a monotonous decrease of the error as the ensemble size increases), the performance of EN3 (even when 200 members are used) is always lower than the performance of EN2.

Figure 8a shows the background error covariance between slow variables and both X_1 and $Y_{1,1}$ estimated from the ensemble spread at the end of the assimilation experiment (day 360). Similarly, Fig. 8b shows the respective covariances of X_1 and $Y_{1,1}$ with the fast variables. There are differences between Figs. 4 and 8. Whereas the correlations shown in Fig. 4 are estimated using 72 000 outputs of the model, correlations in Fig. 8 are estimated with only 50. On the other hand, Fig. 4 measures the degree of association between the temporal evolution of the variables of the model. By neglecting correlations between the dashed lines, the conclusion to be drawn from Fig. 4 is that short-scale variables evolve independent of each other. In contrast, Fig. 8 gives a measure of the correlation of the background error at different locations. However, because of the results shown in Fig. 4, no correlation between the background errors in the fast variables should exist. Accordingly, it must be noticed that the ordinate-axis scale in Fig. 8b is one order of magnitude smaller than in Fig. 8a. Thus, the correlation from Fig. 8 indicates that observing $Y_{1,1}$ would have a small contribution on both the large- and small-scale variables (compared with the impact that X_1 would have had) except in the case of large innovation. If the model deviates from the observations for a fast-scale variable, the non-zero spurious correlation will propagate it to the rest of the system.

A retrospective analysis of the results of EN1, EN2 and EN3 is that the simultaneous assimilation using the EnKF of slow and fast variables always hurts the slow variables, suggesting that the cumulated effect of the spurious covariances from the fast variables not only fails to improve the performance of the assimilation but even increases the risk of filter failure.

4. Alternative initialization (CI)

The results of the previous section show that assimilation of uncorrelated variables is prone to the errors associated with spurious covariances and may degrade the results of the filter. This could be prevented by damping the amplitude of the spurious covariances. However, setting these covariances to zero means that no information of the fast variables would propagate to the rest of the system, and these observations would just correct the value at the observation location. However, although variables may be statistically independent one from the others, they are dynamically linked. For example, Fig. 9 shows the temporal evolution, during an assimilation cycle, of an anomaly of a fast variable. After 6 h, the anomaly has dynamically propagated to the neighbouring locations. This indicates that although sequential assimilation algorithms (either global or local) may fail to assimilate these data, other methods based on the dynamic propagation of the signal might be an alternative. One candidate is the adjoint variational approach that, in its simplest implementation, intends to identify the initial condition for which

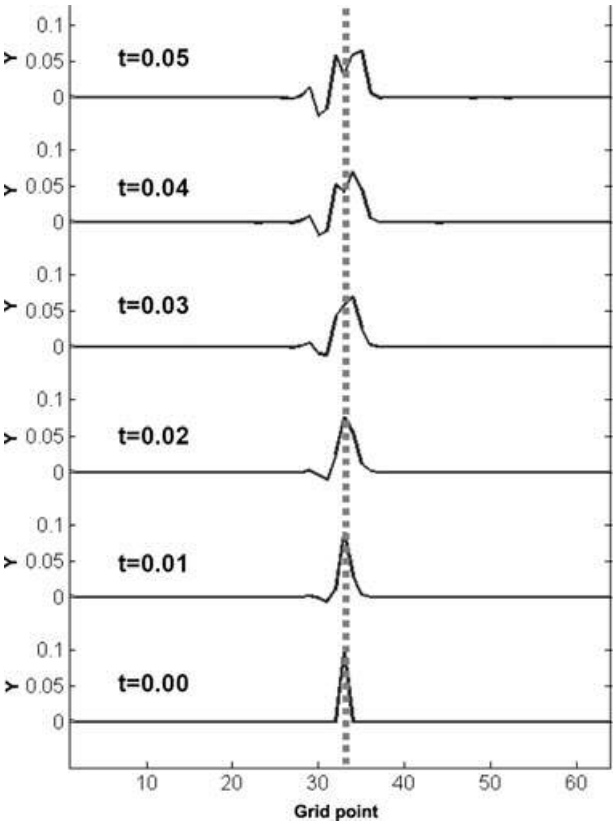

Fig. 9. Time evolution, during one assimilation cycle, of an *Y* anomaly initially located at gridpoint 33 (the dotted line indicates its initial position).

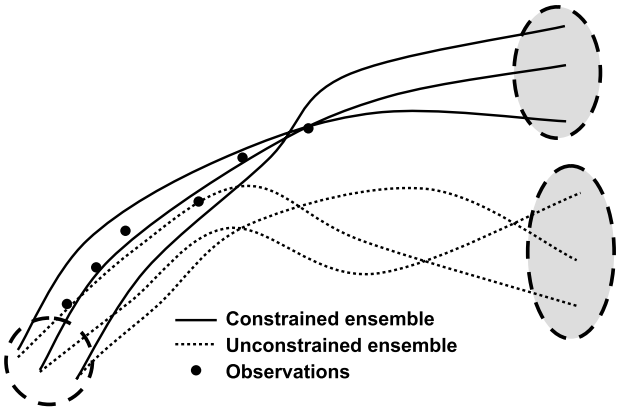

Fig. 10. Diagram illustrating differences between constrained and unconstrained trajectories of an ensemble during one assimilation period.

the system evolves the closest to the observations. Here, a much simpler technique is investigated, based on a combination of two initialization techniques: the EnKF and Newtonian Relaxation or Nudging (Anthes, 1974). In Nudging, assimilation is done by introducing, in the dynamic equations, an additional term, pulling the model solutions toward the observed values.

In the combined initialization experiments, here called CI, the ensemble is being used to initialize the large-scale variables, as in experiment EN2. As noted before, the reduced number of observations allows the filter to closely fit the data, providing an error smaller than the one obtained when all observations (both *X* and *Y*) are being simultaneously assimilated. The initialization of the fast components is done by constraining their evolution using the nudging technique to pull the fast components toward their observed values (Fig. 10). In this setting, implementation of the Newtonian Relaxation is straightforward, as the decorrelation scale of the fast variables (1.5 d) is much larger than the period between analyses (6 h). The experiment with CI with 24 observations is called CI3 (as it uses the same set of observations as the EnKF experiment EN3). The results are shown in Table 1. The performance of the CI has been found to be robust for a large range of values of the nudging coefficient (which is the same for all *Y* observations). Values for the nudging coefficient have been explored by increasing them from zero, up to a value for which the model explodes. The smallest value of the coefficient providing a significant reduction of the error of the fast variables is chosen as the nudging coefficient. In CI3, the amplitude of the coefficient is 100, which corresponds to a half-day relaxation timescale. The results shown in Table 1 indicate that CI3 does not dramatically reduce the error from EN2. However, notice that coupled initialization has been able to slightly reduce the overall error of fast variables, without any degradation of the slow variables reconstruction. Thus, they are significantly better than the results from EN3. Moreover, when this approach is applied to assimilate all the variables (called CI1 as it represents the CI of the same observations used in the ensemble assimilation EN1), the error is 0.48 and 0.023 for *X* and *Y*, respectively (Fig. 11). The CI reduces, by half, the error of the fast variables reported by EN1, indicating that constraining the dynamic evolution of the ensemble is doing a better job than trying to assimilate the fast variables with a sequential filter.

5. Summary

Assimilation of variables whose evolution is uncorrelated with the rest of variables of the system is found to degrade the performance of an EnKF. This result has been obtained by assimilating synthetic data in a non-linear coupled model with two different kinds of variables. The variables in the model may represent processes with separate spatial-/timescales, for example, synoptic waves and convection. In this coupled model, the evolution of the small-scale variables is not correlated either with the other small-scale variables or with the large-scale variables from which they draw their energy. The detrimental effect of the assimilation of small-scale variables with the EnKF has been observed either when all of them are assimilated or when they are oversampled,

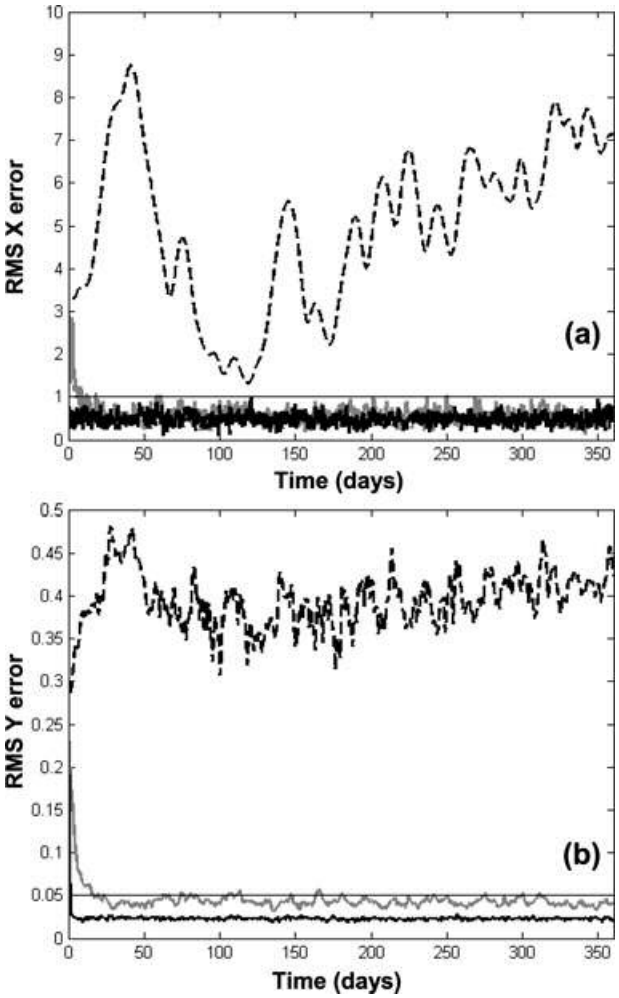

Fig. 11. As *Fig. 6*. All *X* variables are assimilated using an ensemble filter and all *Y* variables are used to constrain the evolution of the ensemble members. Grey lines show the error when all variables are being assimilated with EN1 (Fig. 6).

and little improvement is gained by increasing the size of the ensemble.

This absence of correlation translates to 'locally' diagonal covariance matrices, which do not support the statistical propagation of the information from the observed locations to the rest of the model variables. The spurious sample covariance matrix using a finite ensemble sample is the prime suspect to explain the decrease of the performance of the EnKF when assimilating a strongly undersampled set of these uncorrelated variables. The scenario investigated in this manuscript differs from the scenario drawn by Lorenc (2003), about the limits of EnKFs to assimilate a rich set of observations of small-scale variables. These scenarios are of interest for any data assimilation system (operational or process oriented) dealing with multivariate sets of observations, as they point out the importance of a proper characterization of the statistical properties of each observed variable before accepting it in their data assimilation system.

As a conclusion, this work suggests that in a system with multiple scales, different initialization methods should be combined: one method for the slow component and another method for the fast one. Ensemble techniques are suited to reconstruct the large scale or both scales when all the system is observed (as long as the ensemble size is large enough as pointed out by Lorenc, 2003) but may not be suited to simultaneous reconstruct large and small signals from an undersampled small-scale signal. As an example of the combination of two initialization methods, fast variables have been assimilated using Newtonian relaxation, that is, a technique that exploit the dynamic constraints of the model. The time evolution of the members of the ensemble has been modified by introducing a relaxation term to constrain the evolution of the system by pulling the fast variables to their observed values. The EnKF is then used to assimilate only the correlated variables. This approach is simple to implement, it does not increase the numerical cost of the ensemble simulations; and it has been found to be stable for a large range of relaxation coefficients. When all the observations are assimilated, it provides an analysis error smaller than the error obtained when all the observations are simultaneously assimilated using the original EnKF. When only few uncorrelated variables are assimilated, the new approach does not show the harmful effect of including the uncorrelated variables discussed in Section 3.6. Moreover, the errors of the overall fast variables is reduced from the error obtained when only large-scale variables are being assimilated (Section 3.5), indicating that some information from the short-scale variables is being now indeed propagated to the other short-scale variables. Although encouraging, the usefulness of this CI approach should be validated with other coupled models and even with the same model but with different coupling strength.

6. Acknowledgments

This study is supported by the Spanish National Science Program under contracts ESP2005–06823-C05 and ESP2007–65667-C04.

References

Anderson, J. L. 2001. An ensemble adjustment filter for data assimilation. *Mon. Wea. Rev.* **129**, 2884–2903.

Anthes, R. A. 1974. Data Assimilation and initialization of hurricane prediction models. *J. Atmos. Sci.* **31**, 702–719.

Ballabrera-Poy, J., Brasseur, P. and Verron, J. 2001. Dynamical evolution of the error statistics with the SEEK filter to assimilate altimetric data in eddy-resolving ocean models. *Q. J. R. Meteorol. Soc.* **127**, 233–253.

Cane, M. A., Kaplan, A., Miller, R. N., Tang, B., Hackert, E. C. and coauthors. 1996. Mapping tropical Pacific sea level: data assimilation

via a reduced state space Kalman filter. *J. Geophys. Res.* **101**, 22599–22617.

Corazza, M., Kalnay, E. and Yang, S.-C. 2007. An implementation of the Local Ensemble Kalman Filter in a quasi geostrophic model and comparison with 3D-Var. *Nonlin. Proc. Geophys.* **14**, 89–101.

Deschamps, L. and Talagrand, O. 2007. On some aspects of the definition of initial conditions for ensemble prediction. *Mon. Wea. Rev.* **135**, 3260–3272.

Evensen, G. 1994. Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics. *J. Geophys. Res.* **99**, 10143–10162.

Hunt, B. R., Kostelich, E. J. and Szunyogh, I. 2007. Efficient data assimilation for spatiotemporal chaos: a local ensemble transform Kalman filter. *Physica D* **230**, 112–126.

Kalman, R. E. and Bucy, R. S. 1960. A new approach to linear filtering and prediction problems. *J. Basic Eng.* (*ASME*) **82D**, 35–45.

Kalnay, E. 2003. *Atmospheric Modeling, Data Assimilation and Predictability*. Cambridge University Press, Cambridge, 364 pp.

Kalnay, E., Li, H., Miyoshi, T., Yang, S.-C. and Ballabrera-Poy, J. 2007. 4-D-Var or ensemble Kalman filter? *Tellus* **59A**, 758–773.

Lorenc, A.C. 2003. The potential of the ensemble Kalman filter for NWP—a comparison with 4D-Var. *Q. J. R. Meteorol. Soc.* **129**, 3183–3203.

Lorenz, E. N. and Emanuel, K. A. 1998. Optimal sites for supplementary weather observations: simulations with a small model. *J. Atmos. Sci.* **55**, 399–414.

Meehl, G. A., Lukas, R., Kiladis, G. N., Weickmann, K. M., Matthews, A. J. and co-authors. 2001. A conceptual framework for time and space scale interactions in the climate system. *Clim. Dyn.* **17**, 753–775.

Miller, R.N. and Cane, M. A 1989. A Kalman Filter analysis of sea level height in the tropical Pacific. *J. Phys. Oceanogr.* **19**, 773–790.

Ott, E., Hunt, B. R., Szunyogh, I., Zimin, A. V., Kostelich, E. J. and co-authors. 2004. A local ensemble Kalman filter for atmospheric data assimilation. *Tellus* **56A**, 415–428.

Saha, S., Nadiga, S., Thiaw, C., Wang, J., Wang, W. and co-authors. 2006. The NCEP Climate Forecast System. *J. Clim.* **19**, 3483–3517.

Whitaker, J. S. and Hamill, T. M. 2002. Ensemble data assimilation without perturbed observations. *Mon. Wea. Rev.* **130**, 1913–1924.