Tellus (2006), 58A, 159–170
Printed in Singapore. All rights reserved

TELLUS

# Initiation of ensemble data assimilation

*By* M. ZUPANSKI[1]*, S. J. FLETCHER[1], I. M. NAVON[2], B. UZUNOGLU[3], R. P. HEIKES[4], D. A. RANDALL[4], T. D. RINGLER[4] and D. DAESCU[5], [1]*Cooperative Institute for Research in the Atmosphere, Colorado State University, Fort Collins, CO, USA;* [2]*Department of Mathematics and School of Computational Science and Information Technology, Florida State University, Tallahassee, FL, USA;* [3]*School of Computational Science and Information Technology, Florida State University, Tallahassee, FL, USA;* [4]*Department of Atmospheric Science, Colorado State University, Fort Collins, CO, USA;* [5]*Department of Mathematics and Statistics, Portland State University, Portland, OR, USA*

(Manuscript received 1 February 2005; in final form 31 October 2005)

ABSTRACT

The specification of the initial ensemble for ensemble data assimilation is addressed. The presented work examines the impact of ensemble initiation in the Maximum Likelihood Ensemble Filter (MLEF) framework, but is also applicable to other ensemble data assimilation algorithms. Two methods are considered: the first is based on the use of the Kardar-Parisi-Zhang (KPZ) equation to form sparse random perturbations, followed by spatial smoothing to enforce desired correlation structure, while the second is based on the spatial smoothing of initially uncorrelated random perturbations. Data assimilation experiments are conducted using a global shallow-water model and simulated observations. The two proposed methods are compared to the commonly used method of uncorrelated random perturbations. The results indicate that the impact of the initial correlations in ensemble data assimilation is beneficial. The root-mean-square error rate of convergence of the data assimilation is improved, and the positive impact of initial correlations is notable throughout the data assimilation cycles. The sensitivity to the choice of the correlation length scale exists, although it is not very high. The implied computational savings and improvement of the results may be important in future realistic applications of ensemble data assimilation.

## 1. Introduction

Ensemble data assimilation is a fast developing methodology designed to address the probabilistic aspect of prediction and analysis. It lends itself as a bridge between relatively independently developed data assimilation and ensemble forecasting methodologies. Beginning with the pioneering work of Evensen (1994), followed by Houtekamer and Mitchell (1998), there are now many ensemble data assimilation algorithms (Pham et al., 1998; Lermusiaux and Robinson, 1999; Brasseur et al., 1999; Keppenne, 2000; Bishop et al., 2001; Anderson, 2001, 2003; van Leeuwen, 2001; Whitaker and Hamill, 2002; Reichle et al., 2002a; Snyder and Zhang, 2003; Ott et al., 2004; Zupanski, 2005; Zupanski and Zupanski, 2006). Realistic applications with state-of-the-art models and real observations have also been intensively pursued in recent years (Houtekamer and Mitchell, 2001; Keppenne and Rienecker, 2002; Haugen and Evensen, 2002; Szunyogh et al., 2005; Houtekamer et al., 2005).

One of the important issues is the specification of the initial ensemble. Ideally, the initial ensemble perturbations, defined as a difference between the perturbed and the control initial conditions, should represent error statistics of the corresponding control model state. In practice, the error statistics are measured by the error covariance. It is known that the error covariance has a structure, in principle, defined by the model dynamics, and formally represented by correlations between model variables. In particular, one would like to create initial perturbations reflecting the structure of the error covariance.

Specification of the initial ensemble in ensemble data assimilation varies in the literature. It is generally recognized that the initial forecast error covariance should have a realistic correlation structure, with climatologically consistent perturbation magnitudes. For example, the Evensen (2003) approach is based on the use of a Fourier representation of the perturbations, effectively representing a prescribed correlation structure of the analysis error covariance. A random number generator is used to create random phase shifts. Houtekamer and Mitchell (1998) define a method which samples random perturbations with prescribed forecast error statistics (e.g. correlations). In their subsequent papers, the method is further refined (Mitchell and Houtekamer,

*Corresponding author.
e-mail: zupanskim@cira.colostate.edu

DOI: 10.1111/j.1600-0870.2006.00173.x

2000), and later generalized (Mitchell et al. 2002) to create geostrophically balanced perturbations within a primitive equations framework. Bishop et al. (2001) use a set of largest scale-orthogonal sine and cosine perturbations to initiate the ensemble data assimilation. Dowell et al. (2004) use prescribed ellipsoidal perturbations, with randomly chosen locations in the vicinity of the observation, with the idea that the correlations at the correct location should improve the algorithm performance.

Another generic group of methods for initiating ensemble data assimilation exploits the idea of using ensemble forecasting to develop balanced and correlated ensemble perturbations (Whitaker and Hamill, 2002; Hamill et al., 2003; Whitaker et al., 2004; Szunyogh et al., 2005; Anderson et al., 2005). This consists of forming uncorrelated random fields at some time in the past, sampled from a probability distribution with prescribed mean and standard deviation. Then, the ensemble forecasts with such defined initial perturbations are integrated until the ensemble perturbations develop realistic correlation structure. At that time, the ensemble forecast perturbations are used to form an initial forecast error covariance for the data assimilation scheme. The standard deviation of the initial uncorrelated random perturbations can be defined using the statistics of the forecast model, if known, or the forecast error used for the data assimilation, if available. In most cases, however, these estimates are not available, and one has to rely on the general estimates of the standard deviation.

There are a few potential difficulties with the last approach. If the prescribed initial standard deviation is too small, it may take a long time before realistic magnitudes of the perturbations are formed. Another possible problem is that the realistic forecast error correlations may take a long time to develop.

In this paper, we explore the possibility of improving the latter approach by considering correlated random initial fields. Two options are considered: one, to impose correlations directly on the uncorrelated random fields, and second, to first make spatially sparse random fields, and then impose correlations. The second method employs the Kardar-Parisi-Zhang (KPZ) equation to create spatially sparse random amplitude peaks. Both methods rely on imposing correlations on the initially uncorrelated random field. The proposed methods do not assume any particular form of perturbations (e.g. wave-like, or ellipsoidal), nor location for the random perturbations. This reflects the situation in realistic applications, where the optimal perturbations have an unknown, seemingly random structure and location.

In principle, all the above-mentioned ensemble initiation approaches that utilize correlated initial perturbations, including the methods presented here, should produce comparable results. The methods described in this study will be used in particular to evaluate the relevance of the initial correlations specification on ensemble data assimilation. The proposed methods are evaluated within the Maximum Likelihood Ensemble Filter (MLEF, Zupanski, 2005) framework. However, the methods are directly applicable to other ensemble data assimilation algorithms as well. The methodology will be explained in Section 2, experimental design will be presented in Section 3, results in Section 4, and conclusions will be drawn in Section 5.

## 2. Ensemble initiation methodology

### 2.1. The problem

The initiation of ensemble data assimilation is defined as the specification of perturbations which form the initial forecast error covariance at time $t_0$, before the first observation cycle. Since in this work, the MLEF algorithm (Zupanski, 2005) is used, let the square-root forecast error covariance be defined as

$$P_{\mathrm{f}}^{1/2} = \left(b_1^k, b_2^k, \ldots, b_S^k\right) \quad b_i^k = M\left(x_a^{k-1} + p_i^{k-1}\right) - M\left(x_a^{k-1}\right) \tag{1}$$

where $k$ is the time index representing the analysis cycle, $P_{\mathrm{f}}$ is the forecast error covariance at time $t_k$ (e.g. current analysis time), $S$ is the number of ensemble perturbations, $M$ is a non-linear forecast model integrated from time $t_{k-1}$ to time $t_k$, $b_i$ are the columns of the square-root forecast error covariance at time $t_k$, $\{p_i : i = 1, \ldots, S\}$ are the columns of the square-root analysis error covariance at time $t_{k-1}$, and $x_{\mathrm{a}}$ is the analysis at time $t_{k-1}$.

Since the initiation of ensemble data assimilation is the process of defining the perturbation vectors $\{b_i : i = 1, \ldots, S\}$ at time $t_0$, we would like to utilize the formulation (1) in order to produce balanced perturbations, i.e. perturbations constrained by model dynamics. The use of (1) implies that the problem of defining the initial $\{b_i : i = 1, \ldots, S\}$ at time $t_0$ is substituted by the problem of defining the perturbations $\{p_i : i = 1, \ldots, S\}$ at some previous time. Let $t_0 - \tau$ denote the time when the initial perturbations $\{p_i : i = 1, \ldots, S\}$ are defined, where $\tau$ is a prescribed time interval. Typically, for global models, $\tau$ is a time interval ranging between 6 and 24 h. Therefore, a typical method for the initiation of ensemble data assimilation consists of: (1) specification of the perturbations $\{p_i : i = 1, \ldots, S\}$ at time $t_0 - \tau$, and (2) ensemble forecasting from $t_0 - \tau$ to $t_0$, used to define the perturbations $\{b_i : i = 1, \ldots, S\}$ at time $t_0$.

There are two important practical aspects of initiating an ensemble forecast. One is that the initial ensemble perturbations $\{p_i : i = 1, \ldots, S\}$ need to have an inherent randomness, reflecting the fact that the magnitude and location of unstable initial perturbations are not known a priori. Second issue is that a correlation structure in the initial perturbations is desired, since the outer product of perturbation vectors forms an error covariance. Matching these two requirements may not be simple, due to an additional restriction imposed by the ensemble size.

### 2.2. The ensemble initiation methods

Let $Z = \{z_1, z_2, \ldots, z_S\}$ be uncorrelated normal random variables belonging to a normal $N(0, \sigma_z^2)$ distribution, where

$\{z_i : i = 1, \ldots, S\}$ are the perturbation vectors defined in model space. The covariance is defined as $\mathrm{Cov}(Z) = E(ZZ^T) = \sigma_z^2 I$, where $E$ denotes mathematical expectation, and the superscript T denotes the transpose. A new random variable $P$ can be obtained by applying a change of variable $P = \boldsymbol{F}Z$, where $\boldsymbol{F}$ is a nonsingular linear operator. The covariance of the transformed variable $P$ is

$$\mathrm{Cov}(P) = E(PP^{\mathrm{T}}) = E(\boldsymbol{F}ZZ^{\mathrm{T}}F^{\mathrm{T}}) = \boldsymbol{F}E(ZZ^{\mathrm{T}})\boldsymbol{F}^{\mathrm{T}} = \sigma_z^2 \boldsymbol{F}\boldsymbol{F}^{\mathrm{T}}. \quad (2)$$

If the operator $\boldsymbol{F}$ is normalized, the matrix $\boldsymbol{F}\boldsymbol{F}^{\mathrm{T}}$ defines the correlation matrix for $P$. Furthermore, if $\boldsymbol{F}$ itself is defined as a correlation matrix with characteristic length $L$, the characteristic length of $\boldsymbol{F}\boldsymbol{F}^{\mathrm{T}}$ is $L/\sqrt{2}$ (e.g. Gaspari and Cohn, 1999).

Using the relation (2) and the approximation $b_i = M(x_a + p_i) - M(x_a) \approx \mathbf{M}p_i$, where $\boldsymbol{M}$ is the Jacobian of the nonlinear model $M$, the forecast error covariance at time $t_0$ is

$$P_f = E(\boldsymbol{M}PP^{\mathrm{T}}\boldsymbol{M}^{\mathrm{T}}) = \boldsymbol{M}E(PP^{\mathrm{T}})\boldsymbol{M}^{\mathrm{T}} = \sigma_z^2(\boldsymbol{M}\boldsymbol{F})(\boldsymbol{M}\boldsymbol{F})^{\mathrm{T}}. \quad (3)$$

Note that the above approximation is used only to illustrate the covariance structure. In real situations, the nonlinear difference is used instead. We now proceed with defining three variants of the described ensemble initiation methodology, by focusing on the specification of the initial ensemble $\{p_i : i = 1, \ldots, S\}$ at time $t_0 - \tau$. The ensemble forecasting from $t_0 - \tau$ to $t_0$ is employed in all three variants, as the means for creating dynamically balanced perturbations at time $t_0$.

*2.2.1. The uncorrelated-random method.* The uncorrelated-random method is a simple technique used to create initially uncorrelated random perturbations, and, with some modifications, the method is most often used to specify the initial ensemble. In practice, a standard uncorrelated normal random variable with zero mean and unit variance at time $t_0 - \tau$ is created first. This can be done using the Box–Muller method (Box and Muller, 1958), which transforms an independent random variable uniformly distributed between 0 and 1 into a normal $N(0, 1)$ variable. Using a prescribed standard deviation $\sigma_z$, one can create an uncorrelated random variable $Z \sim N(0, \sigma_z^2)$. Then, using the linear transformation $\boldsymbol{F} = \boldsymbol{I}$, the actual initial perturbation used in (1) is $P = Z$, i.e. $\{p_i = z_i; i = 1, \ldots, S\}$. This is followed by an ensemble integration from $t_0 - \tau$ to $t_0$. If $\tau$ is adequately chosen, the forecast error covariance at time $t_0$ will have the correlations developed and balanced by the forecast model equations, as implied by eq. (1).

*2.2.2. The correlated-random method.* The correlated-random method is a straightforward extension of the uncorrelated-random method. The difference comes from the definition of the initial random variable $P$. In principle, the correlations at time $t_0$ will be developed according to eq. (1), but this may take a long time. In order to improve the forecast error covariance at time $t_0$, a change of variable $P = \boldsymbol{F}Z$ is introduced at time $t_0 - \tau$, which creates correlated random perturbations. In our applications, the matrix $\boldsymbol{F}$ is a block-diagonal Toeplitz matrix, with the elements calculated using the space-limited compactly supported function (4.4) from Gaspari and Cohn (1999). Each block corresponds to a particular model variable, and possibly to a particular model vertical level, as explained in Zupanski et al. (2005, Section 4a). Since the shallow-water model is two-dimensional, this implies using the blocks given by eqs. (20) and (21) from Zupanski et al. (2005).

*2.2.3. The correlated Kardar-Parisi-Zhang method.* Since we anticipate smoothing of the initially uncorrelated random perturbations, the correlated-random method can produce smeared perturbations, without any particular spatial pattern. A typical forecast error covariance, however, would likely show dominant spatial patterns in the area of dynamical instability. Thus, it may be beneficial to first create random perturbations with spatially sparse (e.g. distant), large amplitude peaks, and then to apply smoothing. If the distance between the peaks corresponds to the correlation length scale, the smoothed field will have a desired appearance with few dominant spatial patterns. In order to create the uncorrelated random patterns the KPZ equation is used (Kardar et al., 1986; Beccaria and Curci, 1994; Newman and Bray, 1996; Maunuksela et al., 1999; Marinari et al., 2002) in one-dimensional form

$$\frac{\partial h}{\partial t} = \frac{\partial^2 h}{\partial x^2} + \left(\frac{\partial h}{\partial x}\right)^2 + \xi(x, t) \quad (4)$$

where $h$ is the perturbation and $\xi$ is a random forcing, in our application white Gaussian random noise with unit variance. This equation is generally used to explain the dynamics of interfaces moving through random media. It can be described as a dynamic renormalization procedure used in statistical turbulence theory (Verma, 2000). In another application related to dynamic localization of Lyapunov vectors, the Lyapunov vectors can be viewed as exponentials of the roughened interface, and thus can be represented by the KPZ equation (Pikovsky and Politi, 1998). In principle, any other differential equation with random forcing could be used instead of the KPZ equation. Our motivation for adopting the KPZ equation is that the KPZ model integration produces spatially sparse uncorrelated random perturbations.

Note that there is an important dependence between the sparseness of the random patterns and the imposed correlation length scale: an average distance between the uncorrelated random amplitude peaks (i.e. the sparseness of the random patterns) should correspond to the correlation length scale. This requirement assures that the non-zero perturbations are defined over the full integration domain. The sparseness of random patterns depends on the length of time integration of the KPZ equation: the longer the integration, the sparser the patterns. The empirical relation used in this algorithm is time $= \alpha \times L$, where time refers to the integration time of the KPZ equation, $L$ is a prescribed length scale, and $\alpha$ is an empirical parameter. In our application $\alpha = 0.2$ is chosen, based on the trial-and-error results. In choosing

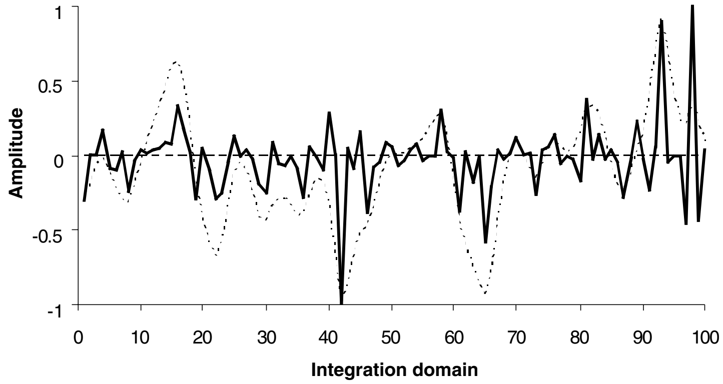

*Fig. 1.* The KPZ equation forecast perturbation vector at the end of 200 nondimensional time steps: (*a*) without imposed correlations (solid line), and (*b*) correlations imposed with the length scale of 5 nondimensional units (dotted line). This is a forecast snap-shot for the one-dimensional integration domain of 100 grid points.

the length scale $L$, one should be guided by general experience, if available. Then, the trial-and-error could be used within those bounds. For example, if the variable is atmospheric pressure, it is generally known that the correlation length scale is of the order of 1000 km, thus one can set the bounds to be 1000 and 3000 km. If the trial-and-error experiments suggest length scales of 10–100 km, however, they should not be used, and a lower bound of 1000 km should be used instead. In other words, a common sense should be used in combining trial-and-error and previous experience.

A typical result of the one-dimensional integration of eq. (4) is shown in Fig. 1. One can note uncorrelated random perturbations at each grid point, with only a few dominating peaks, thus indicating a spatially localized pattern. The sparseness of these patterns is what we are seeking for. The impact of the correlation imposed on the KPZ random perturbation is seen as a smooth dotted line in Fig. 1. The created smooth line represents a correlated random perturbation that would be used as an initial ensemble perturbation.

When applying the KPZ equation in discrete form, a simple centered finite differencing is used for the spatial derivatives, and the one-level forward scheme for the time integration (e.g. Haltiner and Williams, 1980). The particular algorithmic steps relevant to the creation of sparse random perturbations are as follows (1) given the correlation length scale $L$, define the number of time steps for the KPZ equation integration according to the empirical formula, (2) along each of the forecast model coordinates, integrate the one-dimensional KPZ equation, followed by imposing correlations along that coordinate. A sequential application of the one-dimensional KPZ equation in the direction of one model coordinates is used to define the perturbation in a two-dimensional domain. For example, in the case of a longitude–latitude grid the one-dimensional KPZ equation integration along the latitudinal rings is performed first, independently for each ring. Then, using the obtained values as initial conditions, the one-dimensional KPZ equation is integrated along each of the longitudinal rings. This procedure is not unique (since it depends on the order of integration), nor optimal, but the important consequence is that the resulting perturbation is sparse and localized in space. If needed, the perturbation $h$ is occasionally renormalized by imposing an upper limit $|h| \leq N\sigma$, where $\sigma$ is the standard deviation ($\sigma = 1$ and $N = 3$ in this case).

### 2.3. Algorithmic details

All described ensemble initiation methods can be presented in the following sequence of calculations, assuming that the initial time of data assimilation is $t_0$:

(1) define the time interval $\tau$, and specify the initial conditions at time $t_0 - \tau$.

(2) at time $t_0 - \tau$, prescribe the standard deviation $\sigma_z$ and the correlation length scale $L$,

(3) at time $t_0 - \tau$, create the initial ensemble perturbations using one of the methods (e.g. uncorrelated-random, correlated-random, correlated-KPZ),

(4) perform ensemble forecasting from $t_0 - \tau$to $t_0$, and

(5) at time $t_0$ use the ensemble forecast perturbations to form the (square-root) forecast error covariance.

As suggested earlier, a potential advantage of this ensemble initiation method is that it is algorithmically simple, yet it includes the nonlinear forecast model as a balance constraint for ensemble perturbations. There are, however, no other balance constraints used in the algorithm, and no specific evidence of the gravity waves in the solution is noted. A possible reason may be that, in the MLEF, the square-root forecast error covariance columns are defined as a difference between two short-range nonlinear forecasts. Since a global model integration typically acts as a gravity-wave filter (e.g. Haltiner and Williams, 1980, Chapter 11-5), and since the analysis increment is generally defined as a linear combination of these forecast differences, the filtering impact of the short-range forecasts is magnified over many assimilation cycles, eventually making the impact of gravity waves small, or negligible.

Since the forecast model is a component of an ensemble data assimilation algorithm, the use of various forecast models is essentially transparent from the user's point of view. A potential drawback is that the ensemble forecasting in step (4) may be computationally demanding in some cases.

## 3. Experimental design

In this section, few basic experiments are defined in order to illustrate the impact of the initial ensemble on ensemble data assimilation. As mentioned earlier, the MLEF methodology (Zupanski, 2005) is used in all experiments.

### 3.1. Model

A finite-difference shallow-water model developed at Colorado State University is used in this study (Heikes and Randall, 1995a,b), with the improved numerical scheme which better conserves the potential enstrophy and energy (Ringler and Randall, 2002). This is a global model, constructed on a twisted icosahedral grid. The grid consists of hexagons and pentagons, effectively reducing the pole problem. The prognostic variables are height, velocity potential, and stream function. The time integration scheme is the third-order Adams-Bashforth scheme (Durran, 1991). The model has been successfully tested on the suite of seven test cases described by Williamson et al. (1992) (e.g. Heikes and Randall, 1995a). The number of grid cells used in this study is 2562, which corresponds to a model resolution of approximately 4.5° of longitude–latitude. The height points are defined at the center of each cell, while the wind components are defined at the two opposite cell corners. This results in approximately two times more wind points than height points. Both components of the wind (e.g. east-west and north-south) are defined at each wind point. Overall, the total number of prognostic variables is 12 800.

### 3.2. Observations

The observations are created by adding random perturbations from a normal distribution $N(0, \boldsymbol{R})$ to a model forecast, which we refer to as the truth. This implies a perfect model assumption, since the same model is used in the assimilation. Although the model equations formally predict the velocity potential and the stream function, more conventional wind observations are created, and later assimilated. The observation error covariance matrix, $\boldsymbol{R}$, is assumed to be diagonal, i.e. no correlation between observations is assumed. The observation error chosen for the height is 5 m, and for the wind is 0.5 ms$^{-1}$. There are 1025 observations defined in each analysis cycle, uniformly distributed around the globe. These observations consist of 513 height observations and 512 wind observations. Since the two wind components (east-west and north-south) are colocated, there are 256 observation points for each wind component. The observations are assimilated every 6 h.

### 3.3. Experiments

The initial conditions are defined by the fifth test case from Williamson et al. (1992), which corresponds to a geostrophically balanced zonal flow over an isolated conical mountain. The initial zonal flow is 20 ms$^{-1}$, and the mountain is centered at 30°N, 90°W, with a height of 2000 m. This setup is characterized by the excitation of Rossby and gravity waves, with notable nonlinearity occurring in the vicinity of the mountain.

In all experiments a 1000 ensemble members are used. Such high number is not necessarily needed, but it helps in relaxing the restrictions due to a limited ensemble size, thus allowing a more focused examination of the initial ensemble specification. The observations are defined at model grid points, implying the linearity of the observation operator. The initial conditions of the forecast run used to define the observations are defined at $t_0$. The assimilation is performed over 15 days, which corresponds to 60 data assimilation cycles of 6-h intervals. The initial conditions for the experimental run are defined the same way as in the forecast run used to create the observations (e.g. test case 5), however initiated 6 h earlier in order to create erroneous initial conditions. Thus, the prescribed time interval $\tau$ is 6 h, and the initial ensemble is defined at $t_0 - 6$ h. With this setup, the forecast error covariance at time $t_0$ has a standard deviation about two times larger than the observation error (i.e. approximately 10 m for height and 1 ms$^{-1}$ for winds), which is considered to be realistic.

The experiments are separated into two groups: (1) the three ensemble initiation methods, i.e. the uncorrelated-random, correlated-random, and correlated-KPZ, are compared, and (2) the sensitivity of the analysis to the correlation length scale is evaluated.

The results are compared using the root-mean-square (RMS) analysis error, defined as a difference between the analysis and the truth (e.g. the forecast used to create observations), valid at the time of the analysis. In addition, the $\chi^2$ test (e.g. Menard et al., 2000), and the rank histogram of normalized innovation vectors (e.g. observation minus guess) (Reichle et al., 2002b; Zupanski, 2005), are also used.

## 4. Results

### 4.1. Impact of the ensemble initiation methodology

Three data assimilation results are presented in this subsection, which differ only in the method used to create the initial ensemble perturbations at time $t_0 - \tau$: (1) the uncorrelated-random method, (2) the correlated-random method, and (3) the *correlated-KPZ* method. For the correlated-random and the correlated-KPZ methods the correlation length scale of 4000 km is used for all the three variables, the height and the two wind components.

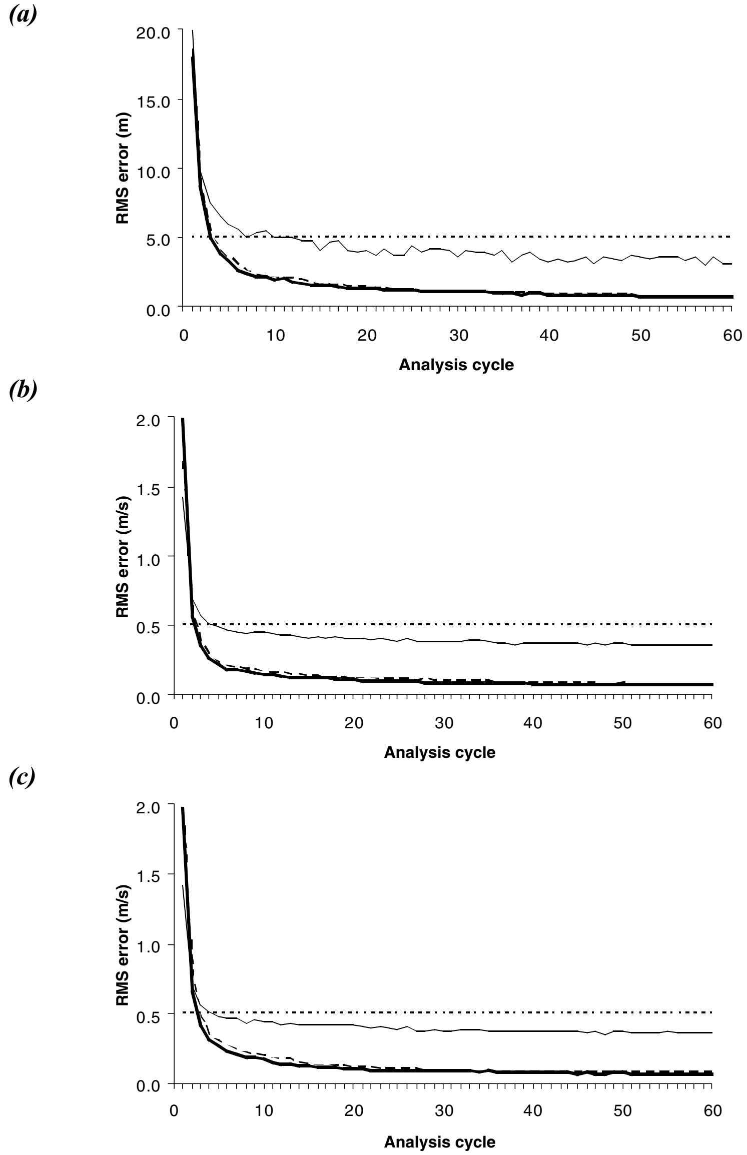

*Fig. 2*. Analysis RMS error for: (*a*) height (m), (*b*) east-west wind component (ms$^{-1}$), and (*c*) north-south wind component (ms$^{-1}$). The results are obtained using: (1) uncorrelated-random method (thin solid line), (2) correlated-random method (dashed line), and (3) correlated-KPZ method (thick solid line). The observation error standard deviation is indicated by a dotted line.

The performance of the three ensemble initiation methods is shown in Fig. 2, as the analysis RMS errors for the height and wind components. For reference, the observation errors are plotted as well. First thing to note is that the analysis RMS error is smaller than the observation error in all experiments, indicating a successful performance of the algorithms for all the three ensemble initiation methods. One can also note a significantly improved performance of the algorithms which use a correlated initial ensemble, i.e. the correlated-random and the correlated-KPZ methods. This suggests that initial correlations are important for the ensemble initiation. Between themselves, however, the correlated-random and the correlated-KPZ results do not differ very much. A closer look (not shown here) indicates that the correlated-KPZ method does produce consistently smaller RMS errors than the correlated-random method, however not at the significant level. This indicates that the sparseness of the perturbations is not an important feature in the initial ensemble. The uncorrelated random correlation method eventually produces analysis RMS errors smaller than the observation errors, but the RMS errors remain relatively large. The correlated-random and the correlated-KPZ methods, on the other hand, both achieve good convergence in less than 1 d (i.e. 3–4 cycles) and overall much smaller analysis RMS errors. The height RMS error reduces to 0.8 m after 60 analysis cycles, while the wind RMS errors are reduced to less than 0.1 ms$^{-1}$. It is interesting to note that there is no visible trend of the RMS errors in the experiments to become closer as the assimilation proceeds. Although one would expect the RMS errors from all experiments to become equal eventually, the RMS error difference between the experiments remains approximately constant.

This apparent paradox will be further explored in the subsequent section.

In terms of the normalized innovation vector statistics, the correlated-random and the correlated-KPZ methods again show an improved performance over the uncorrelated-random method. Since the results of the correlated-random and the correlated-KPZ methods are very similar, only the results of the correlated-KPZ method are shown (e.g. Figs. 3 and 4). The optimal value for the $\chi^2$ test is 1. The results in Fig. 3 indicate large deviations from the optimal value in the uncorrelated random perturbation experiment, eventually settling in the 1.2–1.3 range. On the other hand, the results of the correlated-KPZ experiment show much better values, closer to 1.

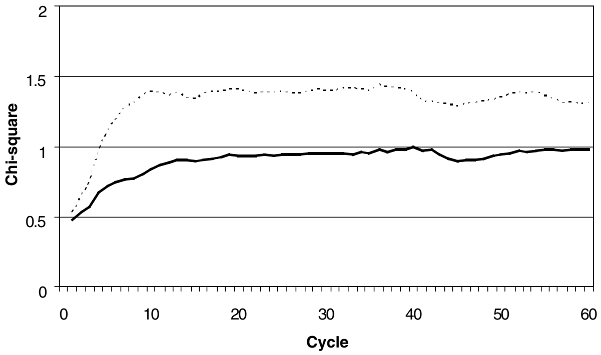

*Fig. 3.* Innovation vector statistics illustrated by the $\chi^2$ test. The results are shown for: (1) uncorrelated-random method (dotted line) and (2) correlated-KPZ method (solid line).

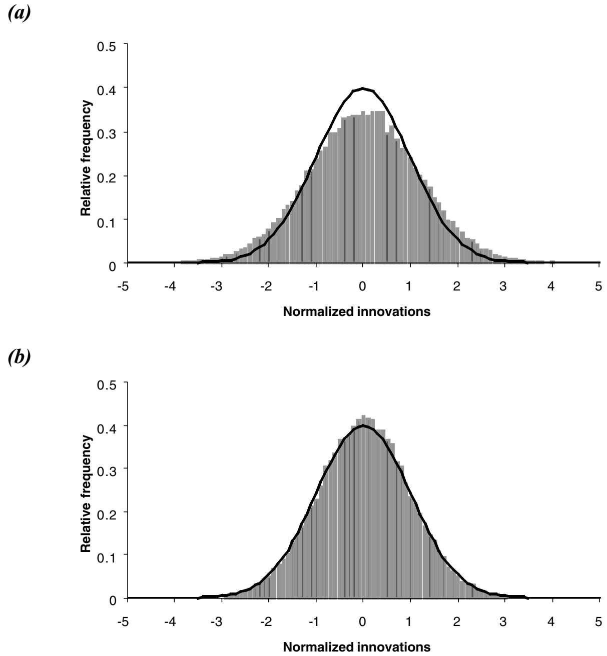

*Fig. 4.* Innovation vector statistics illustrated by a rank histogram of normalized innovations. The results are obtained using: (*a*) uncorrelated-random method, and (*b*) correlated-KPZ method. The solid line represents the $N(0, 1)$ normal distribution.

The rank histogram (Fig. 4) shows a comparison between the $N$(0,1) normal distribution (zero mean and unit variance), and the histogram of normalized innovations $(R + HP_fH^T)^{-1/2}(y - H(x_a))$. The observation operator is denoted $H$, and $\boldsymbol{y}$ is the observation vector. Details of the inverse square-root matrix calculation can be found in Zupanski (2005). Although small deviations from the $N(0,\ 1)$ distribution can be expected due to the impact of weak nonlinearity of the shallow-water model, the rank histogram (Fig. 4a) suggests an underestimation of the error covariance in the uncorrelated random perturbation method. To see this, note that if the forecast error covariance $P_{\mathrm{f}}$ is underestimated, the normalized innovations will be slightly larger, thus the innovation vector realizations away from 0 will be more abundant. Since the rank histogram is normalized by dividing the number of normalized innovations within a bin by the total number of innovation vectors, the histogram will indicate a larger spread, and thus a smaller maximum. Even in the uncorrelated random perturbation experiment, however, there are no significant outliers, meaning that ensembles are adequately covering the necessary range of perturbations (i.e. ensemble spread is adequate). Overall, the innovation vector statistics indicates a stable performance of the MLEF algorithm in both ensemble initiation experiments.

A comparison of the impact of the uncorrelated-random and the correlated-KPZ ensemble initiation methods on the height analysis increment (i.e. analysis minus truth) is shown in Fig. 5, during first several data assimilation cycles. Note that all plots in Fig. 5 have the same contour interval of 5 m. This is used in order to better illustrate a dramatic reduction of the analysis error. It is clear that the correlated-KPZ experiment produces much smoother analysis increments, eventually resulting in superior performance. By cycle 5, the height analysis increments in correlated-KPZ experiment (Fig. 5b) are generally smaller than 5 m, with only a few small areas with about 10 m. On the other hand, the analysis increments in the uncorrelated random initial perturbation experiment (Fig. 5a) are quite noisy, especially in the first cycle. Nevertheless, it appears that the analysis increment noise is dramatically reduced in both experiments, suggesting a robustness of the data assimilation algorithm.

### 4.2. *Impact of decorrelation length*

Given the results of the previous subsection, it is interesting to learn how sensitive the correlated-KPZ results are to the choice of the length scale. The analysis RMS error results obtained using the uncorrelated-random (i.e. 0 km decorrelation length), and the correlated-KPZ method with decorrelation lengths of 1000, 4000, and 7000 km, are shown in Fig. 6. The errors indicate that, in this example, larger decorrelation lengths generally produce better analysis. Most improvement is noted when the decorrelation length is increased in the range between 0 and 4000 km.

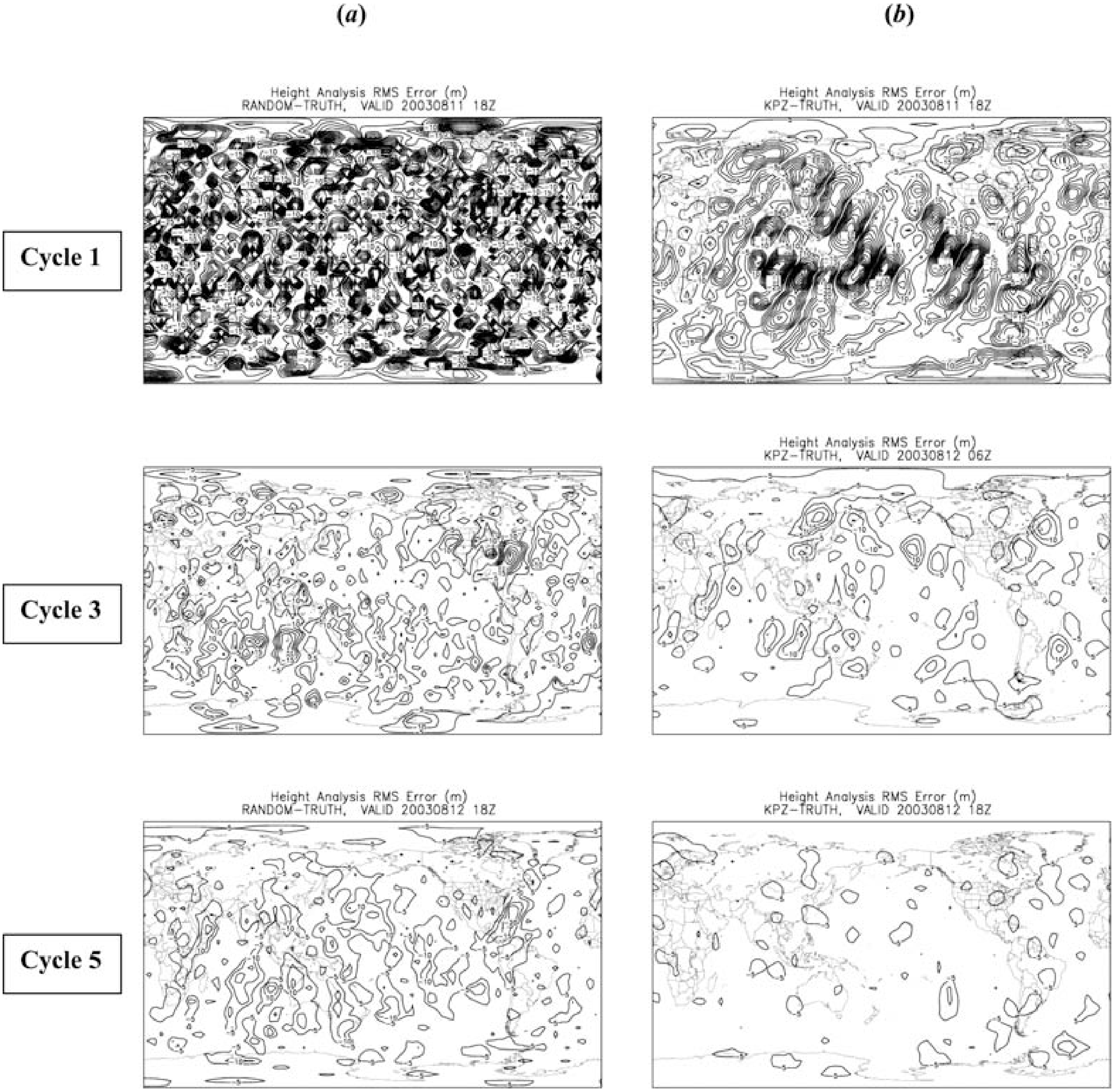
(a)
(b)
Height Analysis RMS Error (m)
RANDOM-TRUTH, VALID 20030811 18Z
Height Analysis RMS Error (m)
KPZ-TRUTH, VALID 20030811 18Z
Cycle 1
Height Analysis RMS Error (m)
KPZ-TRUTH, VALID 20030812 06Z
Cycle 3
Height Analysis RMS Error (m)
RANDOM-TRUTH, VALID 20030812 18Z
Height Analysis RMS Error (m)
KPZ-TRUTH, VALID 20030812 18Z
Cycle 5

*Fig. 5.* Height analysis increment (m) in the analysis cycles 1, 3, and 5 obtained using: (*a*) uncorrelated-random, and (*b*) correlated-KPZ methodology. The contour interval is 5 m. The continental outlines are indicated only for perspective.

When increasing the length from 4000 to 7000 km, there is only a marginal improvement in first few cycles, the difference becoming negligible eventually. One can speculate that, in this case, the true analysis error covariance has characteristic length scales between 4000 and 7000 km.

As noted earlier, the RMS errors from experiments with different correlation length scales do not seem to converge to a common value, indicating a longlasting impact of the specification of the initial ensemble. Recent results from chaos theory provide a possible explanation for this behavior. Lorenzo et al. (2003) examined the system of coupled Lorenz chaotic cells, and found that the specification of the correlation length of perturbations has a nontrivial impact on the system predictability. In applications to one-dimensional coupled chaotic oscillators, Lopez et al. (2004) found that correlations of spatiotemporal perturbations developed in the system contain important information about the sub-leading Lyapunov exponents, and consequently impact the overall perturbation growth. This subject was further pursued and it was found (2005, Cristina Primo, private communication) that, for a dynamical systems with weak chaos, a

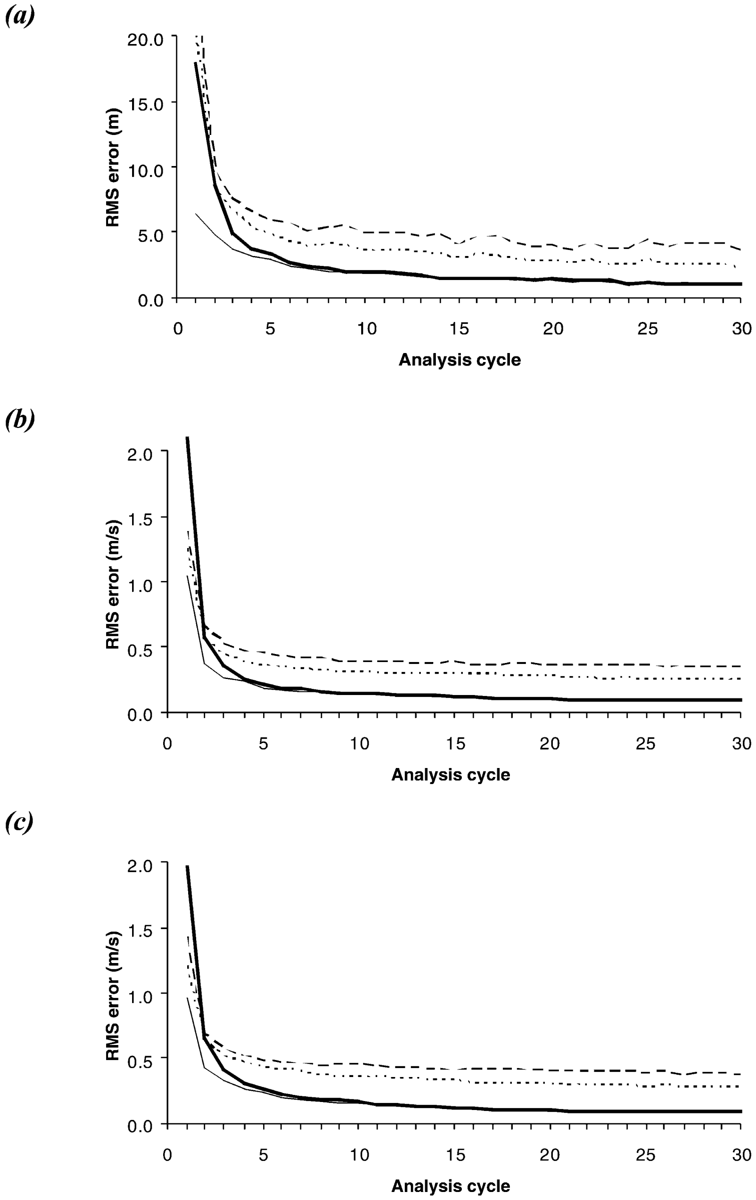

*Fig. 6.* Sensitvity of the analysis RMS errors to correlation length scale in the correlated-KPZ method, for: (a) height (m), (b) east-west wind component (ms$^{-1}$), and (c) north-south wind component (ms$^{-1}$). Shown results are for decorrelation lengths of: (1) 0 km (e.g. uncorrelated – dashed line), (2) 1000 km (dotted line), (3) 4000 km (thick solid line), and (4) 7000 km (thin solid line). The results with correlation length of 4000 km are same as shown in Fig. 2 for the correlated-KPZ results.

specification of the spatial correlations of the initial perturbation can be felt for long time. In the case of strong chaotic behavior of the system, the errors saturate quickly, and the impact of the initial spatial correlations is lost. The weakly chaotic results are in general agreement with the behavior noted in Figs. 2 and 6. Note that, in the case of a global shallow-water model with an isolated mountain, examined here, the system is effectively nonchaotic. The RMS error of the perturbations (forecast minus truth) is not changing significantly in time, indicating a lack of instability and no notable error growth after a couple of days. Thus, one would expect that the specification of the initial ensemble has a longlasting impact in this case.

On the other hand, if the ensemble amplitude is changed (e.g. SD $\sigma_z$ in eqs. (2) and (3)), while keeping the initial correlations unchanged, the RMS error converges to a common asymptotic value. In Fig. 7 the impact of the magnitude of the initial SD is shown in the correlated-random experiment, with initially specified correlation length scale $L = 4000$ km in all experiments. The SD varies from 1 to 10 m for height, and from 0.1 to 1 ms$^{-1}$ for winds, i.e. by one order of magnitude. Yet, in all of the experiments the analysis RMS errors quickly converge to a common value. This is also in agreement with other ensemble data assimilation results (e.g. Whitaker et al., 2004; Zhang et al., 2004; Szunyogh et al., 2005).

The above results suggest that the impact of the initial correlation length does depend on the strength of the chaotic behavior of a dynamical system. In application to realistic weather and climate models, as well as to simpler chaotic systems, this

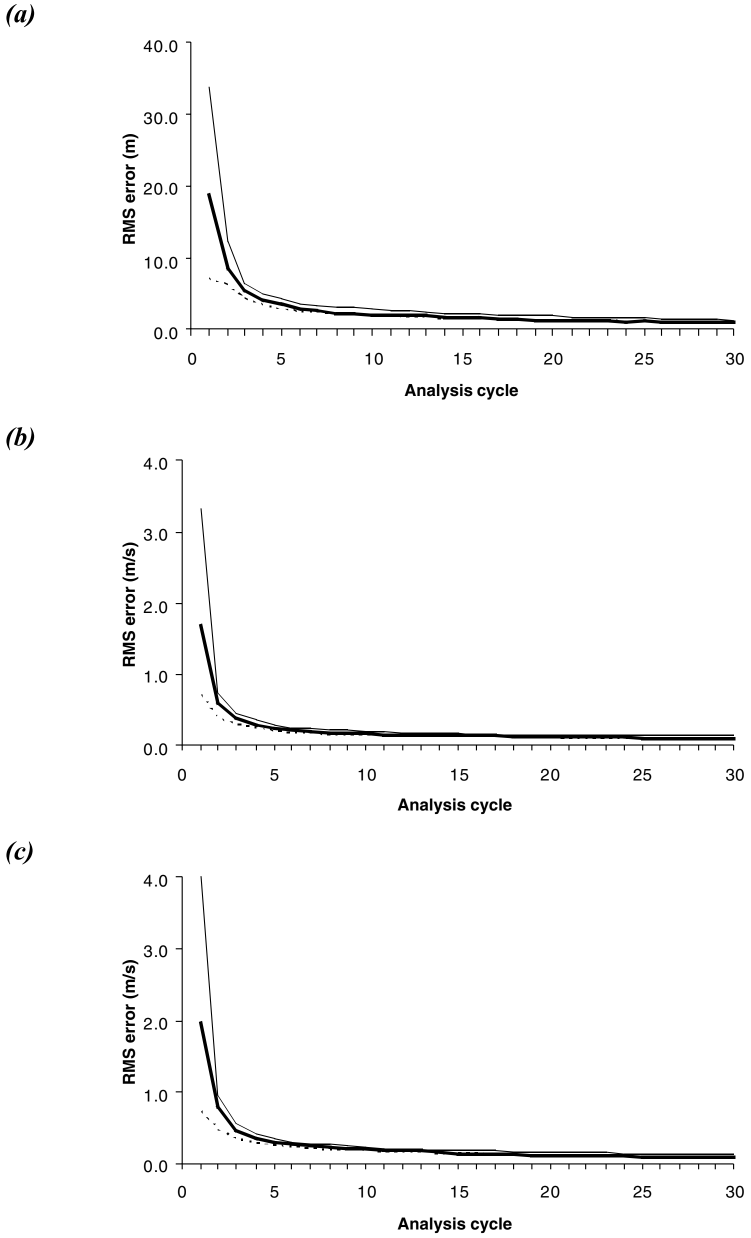

*Fig. 7.* Sensitvity of the analysis RMS errors to the initial standard deviation (amplitude) in the correlated-random method, for: (a) height (m), (b) east-west wind component ($ms^{-1}$), and (c) north-south wind component ($ms^{-1}$). Shown results are for the initial height SD of: (1) 1 m (dotted line), (2) 5 m (thick solid line), and (3) 10 m (thin solid line), and for the initial wind SD of: (1) 0.1 $ms^{-1}$ (dotted line), (2) 0.5 $ms^{-1}$ (thick solid line), and (3) 1.0 $ms^{-1}$ (thin solid line).

would imply a reduced sensitivity to initial correlations. However, poorly estimated initial error covariance, combined with a limited size of the ensemble and an inadequate observational coverage, could all contribute to ensemble data assimilation divergence, even before realistic correlations could be developed. Therefore, specification of the initial ensemble should be an important component of ensemble data assimilation.

The beneficial impact of initial correlations also suggests that an improved ensemble data assimilation algorithm performance can be expected if the forecast error covariance structure is well-known. This may be the case in operational numerical weather prediction (NWP) centers, or for forecast models with a longer history of applications. If the correlation length scale is poorly known, however, one would expect a lesser benefit of correlated initial perturbations. This could happen when many control variables are involved, some with unreliable correlation length scales, such as the physics related variables (e.g. clouds and precipitation). However, the comparison between the uncorrelated random results and the correlated-KPZ results for 1000 km length scale (Fig. 6) suggests that even in those situations the use of the correlated-KPZ method (or the correlated-random method as implied by Fig. 2) may be superior to the uncorrelated random method. This is a subject worth exploring in future realistic applications of the method.

## 5. Summary and Conclusions

Three methods for the specification of the initial ensemble for ensemble data assimilation are presented in this paper: the uncorrelated-random, the correlated-random, and the correlated-KPZ method. The correlated-random and the correlated-KPZ algorithms consist of two steps: (1) creating uncorrelated random perturbations (spatially sparse in the case of the KPZ equation), and (2*)* imposing correlations of a chosen length scale on the perturbations. The expectation is that the initial error covariance is more realistic, thus enabling the ensemble data assimilation to perform better, and achieve faster convergence in terms of the analysis RMS error. The assimilation algorithm employed in the study is the MLEF, applied to the assimilation of simulated observations using a global shallow-water model with zonal flow over an isolated mountain. The presented ensemble initiation methods are directly applicable to other ensemble data assimilation algorithms.

Results indicate a superior performance of the ensemble data assimilation scheme if the initial correlations are specified. This is confirmed by the analysis RMS scores, as well as by the innovation vector statistics. Sensitivity of the correlated ensemble initiation methods to the input correlation length parameter exists, but it is relatively small if a good estimate of the correlation length scale is known.

Overall results indicate that the specification of the initial ensemble may be an important component of an ensemble data assimilation algorithm. It was found that initially specified correlations can have a longlasting impact, if the system is weakly chaotic. In the near future complex atmospheric models will be used, which are inherently more chaotic than the examined shallow-water model with zonal flow over an isolated mountain. This allows further examination of the impact of the initial correlations, extended to a stronger chaotic regime. A possible difficulty in future realistic applications is that some variables (e.g. pressure, temperature, and winds) may have better known correlation statistics than microphysical variables (e.g. clouds and precipitation). Then, an important question is how to achieve an improvement in the context of multiple control variables with poorly known correlation statistics.

A possible improvement of the correlated-KPZ method may be achieved by using the formulation of the KPZ equation with spatially correlated noise (e.g. Janssen et al. 1999), such that the two algorithmic steps collapse into just one step. This is more appealing from the mathematical, as well as from the practical, point of view, since the method would become algorithmically simpler. At present, however, taking into account the algorithmic complexity of the correlated-KPZ method, the correlated-random method appear to be a more practical choice for ensemble initiation than the correlated-KPZ method. Hopefully, a focused research on the ensemble initiation methods will eventually lead to more robust ensemble data assimilation algorithms, important for future realistic applications.

## 6. Acknowledgments

We would like to thank Dusanka Zupanski for many helpful discussions and careful reading of the manuscript. We would also like to thank two anonymous reviewers for their careful reviews which considerably improved the manuscript. Our gratitude is also extended to the National Center for Atmospheric Research, which is sponsored by the National Science Foundation, for the computing time used in this research. This work was supported by the National Science Foundation Collaboration in Mathematical Geosciences grant 0327651.

### References

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

Anderson, J. L. 2003. A local least squares framework for ensemble filtering. *Mon. Weather Rev.* **131**, 634–642.

Anderson, J. L., Wyman, B., Zhang, S. and Hoar, T. 2005. Assimilation of surface pressure observations using an ensemble filter in an idealized global atmospheric prediction system. *J. Atmos. Sci.* **62**, 2925–2938.

Beccaria, M. and Curci, G. 1994. Numerical simulation of the Kardar-Parisi-Zhang equation. *Phys. Rev. E* **50**, 4560–4563.

Bishop, C., Etherton, B. J. and Majumdar, S. J. 2001. Adaptive sampling with the ensemble transform Kalman filter. Part I: theoretical aspects. *Mon. Weather Rev.* **129**, 420–436.

Box, G. E. P. and Muller, M. E., 1958. A note on the generation of random normal deviates. *Ann. Math. Stat.* **29**, 610–611.

Brasseur, P., Ballabrera, J. and Verron, J. 1999. Assimilation of altimetric data in the mid-latitude oceans using the SEEK filter with an eddy-resolving primitive equation model. *J. Marine Syst.* **22**, 269–294.

Dowell, D. C., Zhang, F., Wicker, L. J., Snyder, C. and Crook, N. A. 2004. Wind and temperature retrievals in the 17 May 1981 Arcadia, Oklahoma, supercell: ensemble Kalman filter experiments. *Mon. Weather Rev.* **132**, 1982–2005.

Durran, D. R. 1991. The third-order Adams-Bashforth method: an attractive alternative to leapfrog time differencing. *Mon. Weather Rev.* **119**, 702–720.

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.

Evensen, G. 2003. The ensemble Kalman filter: theoretical formulation and practical implementation. *Ocean Dyn.* **53**, 343–367.

Gaspari, G. and Cohn, S. E. 1999. Construction of correlation functions in two and three dimensions. *Quart. J. Roy. Meteor. Soc.* **125**, 723–757.

Haltiner, G. J. and Williams, R. T. 1980. *Numerical Prediction and Dynamic Meteorology*. 2nd Edition. Wiley & Sons, 477 pp.

Hamill, T. M., Snyder, C. and Whitaker, J. S. 2003. Ensemble forecasts and the properties of flow-dependent analysis-error covariance singular vectors. *Mon. Weather Rev.* **131**, 1741–1758.

Haugen, V. E. J. and Evensen, G. 2002. Assimilation of SLA and SST data into an OGCM for the Indian Ocean. *Ocean Dyn.* **52**, 133–151.

Heikes, R. and Randall, D. A. 1995a. Numerical integration of the shallow-water equations on a twisted icosahedral grid. Part I: basic design and results of tests. *Mon. Weather Rev.* **123**, 1862–1880.

Heikes, R. and Randall, D. A. 1995b. Numerical integration of the shallow-water equations on a twisted icosahedral grid. Part II: a detailed description of the grid and an analysis of numerical accuracy. *Mon. Weather Rev.* **123**, 1881–1887.

Houtekamer, P. L. and Mitchell, H. L. 1998. Data assimilation using an ensemble Kalman filter technique. *Mon. Weather Rev.* **126**, 796–811.

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

Houtekamer, P. L., Mitchell, H. L., Pellerin, G., Buehner, M., Charron, M., and co-authors. 2005. Atmospheric data assimilation with an ensemble Kalman filter: results with real observations. *Mon. Weather Rev.* **133**, 604–620.

Janssen, H. K., Tauber, U. C. and Frey, E. 1999. Exact results for the Kardar-Parisi-Zhang equation with spatially correlated noise. *Eur. Phys. J. B* **9**, 491–511.

Kardar, M., Parisi, G. and Zhang, Y. C. 1986. Dynamic scaling of growing interfaces. *Phys. Rev. Lett.* **56**, 889–892.

Keppenne, C. L. 2000. Data assimilation into a primitive-equation model with a parallel ensemble Kalman filter. *Mon. Weather Rev.* **128**, 1971–1981.

Keppenne, C. L. and Rienecker, M. M. 2002. Initial testing of massively-parallel ensemble Kalman filter with the Poseidon isopycnal ocean general circulation model. *Mon. Weather Rev.* **130**, 2951–2965.

Lermusiaux, P. F. J. and Robinson, A. R. 1999. Data assimilation via error subspace statistical estimation. Part I: theory and schemes. *Mon. Weather Rev.* **127**, 1385–1407.

Lopez, J. M., Primo, C., Rodriguez, M. A. and Szendro, I. G. 2004. Scaling properties of growing noninfinitesimal perturbations in space-time chaos. *Phys. Rev. E* **70**, 056224(5).

Lorenzo, M. N., Santos, M. A. and Perez-Munuzuri, V. 2003. Spatiotemporal stochastic forcing effects in an ensemble consisting of arrays of diffusively coupled Lorenz cells. *Chaos* **13**, 913–920.

Marinari, E., Pagnani, A., Parisi, G. and Racz, Z. 2002. Width distributions and the upper critical dimension of Karadar-Parisi-Zhang interfaces. *Phys. Rev. E* **65**, 026136.

Maunuksela, J., Myllys, M., Timonen, J., Alava, M. J. and Ala-Nissila, T. 1999. Kardar-Parisi-Zhang scaling in kinetic roughening of fire fronts. *Physica A* **266**, 372–376.

Menard, R., Cohn, S. E., Chang, L.-P. and Lyster, P. M. 2000. Assimilation of stratospheric chemical tracer observations using a Kalman filter. Part I: formulation. *Mon. Weather Rev.* **128**, 2654–2671.

Mitchell, H. L. and Houtekamer, P. L., 2000. An adaptive ensemble Kalman filter. *Mon. Weather Rev.* **128**, 416–433.

Mitchell, H. L., Houtekamer, P. L. and Pellerin, G. 2002. Ensemble size, balance, and model-error representation in an ensemble Kalman filter. *Mon. Weather Rev.* **130**, 2791–2808.

Newman, T. J. and Bray, A. J. 1996. Strong-coupling behaviour in discrete Kardar-Parisi-Zhang equations. *J. Phys. A.* **29**, 7917–7928.

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**(4), 273–277.

Pham, D. T., Verron, J. and Roubaud, M. C. 1998. A singular evolutive extended Kalman filter for data assimilation in oceanography. *J. Marine Syst.* **16**, 323–340.

Pikovsky, A. and Politi, A. 1998. Dynamic localization of Lyapunov vectors in spacetime chaos. *Nonlinearity* **11**, 1049–1062.

Reichle, R. H., McLaughlin, D. B. and Entekhabi, D. 2002a. Hydrologic data assimilation with the ensemble Kalman filter. *Mon. Weather Rev.* **130**, 103–114.

Reichle, R. H., Walker, J. P., Koster, R. D. and Houser, P. R. 2002b. Extended versus ensemble Kalman filtering for land data assimilation. *J. Hydrometorol.* **3**, 728–740.

Ringler, T. D. and Randall, D. A. 2002. A potential enstrophy and energy conserving numerical scheme for solution of the shallow-water equations on a geodesic grid. *Mon. Weather Rev.* **130**, 1397–1410.

Snyder, C. and Zhang, F. 2003. Assimilation of simulated Doppler radar observations with an ensemble Kalman filter. *Mon. Weather Rev.* **131**, 1663–1677.

Szunyogh, I., Kostelich, E. J., Gyarmati, G., Patil, D. J., Hunt, B. R., and co-authors 2005. Assessing a local ensemble Kalman filter: perfect model experiments with the NCEP global model". *Tellus* **57A**, 528–545.

van Leeuwen, P. J. 2001. An ensemble smoother with error estimates. *Mon. Weather Rev.* **129**, 709–728.

Verma, M. K. 2000. Intermittency exponents and energy spectrum of the Burgers and KPZ equations with correlated noise. *Physica A* **277**, 359–388.

Whitaker, J. S., Compo, G. P., Wei, X. and Hamill, T. M. 2004. Reanalysis without radiosondes using ensemble data assimilation. *Mon. Weather Rev.* **132**, 1190–1200.

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

Williamson, D. L., Drake, J. B., Hack, J. J., Jakob, R. and Swartztrauber, P. N. 1992. A standard test set for numerical approximations to the shallow-water equations in spherical geometry. *J. Comput. Phys.* **102**, 221–224.

Zhang, F., Snyder, C. and Sun, J. 2004. Impacts of initial estimate and observation availability on convective-scale data assimilation with an ensemble Kalman filter. *Mon. Weather Rev.* **132**, 1238–1253.

Zupanski, M. 2005. Maximum likelihood ensemble Filter: theoretical aspects. *Mon. Weather Rev.* **133**, 1710–1726.

Zupanski, D. and Zupanski, M. 2006. Model error estimation employing ensemble data assimilation approach. *Mon. Weather Rev.*, in press.

Zupanski, M., Zupanski, D., Vukicevic, T., Eis, K. and Vonder Haar, T. 2005. CIRA/CSU four-dimensional variational data assimilation system. *Mon. Weather Rev.* **133**, 829–843.