The Local Ensemble Transform Kalman Filter and its implementation on the NCEP global model at the University of Maryland

Istvan Szunyogh (*), Elizabeth A. Satterfield (*), José A. Aravéquia (**), Elana J. Fertig (*), Gyorgyi Gyarmati (*), Eugenia Kalnay (*), Brian R. Hunt (*), Eric J. Kostelich (***), David D. Kuhl (*), Edward Ott (*), and James A. Yorke (*)

(*) University of Maryland, College Park, MD, United States of America
(**) Centro de Previsão de Tempo e Estudos Climáticos, Brazil
(***) Arizona State University, Tempe, AZ, United States of America
corresponding author e-mail: szunyogh@ipst.umd.edu

ABSTRACT

This paper describes the Local Ensemble Transform Kalman Filter (LETKF) data assimilation scheme and its implementation on the National Centers for Environmental Prediction (NCEP) Global Forecast System (GFS) model at the University of Maryland. Numerical results are shown for both simulated observations and observations of the real atmosphere. The role of flow-dependent information in data assimilation is discussed based on the results of the numerical experiments. Preliminary assimilation results with AMSU-A radiance observations are also presented.

1 Introduction

The informal *"Weather and Chaos"* project of the University of Maryland[1] was launched by Eugenia Kalnay and Jim Yorke in 2000. Their goal was to form an interdisciplinary group of numerical weather prediction and dynamical systems theory experts to conduct research that would lead to a better understanding of the behavior of high-dimensional spatio-temporally chaotic systems, such as the atmosphere, and to improved techniques for numerical weather prediction. A fundamental product of this collaborative research effort is the Local Ensemble Transform Kalman Filter (LETKF) data assimilation algorithm (Hunt et al., 2007, Ott et al. 2004). In this paper, we focus on results obtained with an implementation of the LETKF on a reduced (T62L28) resolution version of NCEP GFS model.

In section 2, we summarize the most important milestones in the development of the LETKF and describe its current implementation on the NCEP GFS model. In section 3, we assess the performance of this particular implementation, assimilating a set of conventional (non-radiance) observations of the atmosphere and discuss the implications of the results on the importance of capturing flow dependent aspects of the analysis and forecast error statistics. In section 4, we present preliminary assimilation results for the case when AMSU-A radiance observations are added to the observational data set. In section 5, we offer some concluding remarks.

[1] Further information on this project, including copies of our submitted, accepted and already published papers, can be found at http:/www.weatherchaos.umd.edu

2 The LETKF and its implementation on the NCEP GFS model

2.1 Ensemble based Kalman filtering

Ensemble based Kalman filters estimate the flow dependent background state $\overline{\mathbf{x}}^b(t)$ and the flow dependent background error covariance matrix $\mathbf{P}^b(t)$ based on an ensemble $\mathbf{x}^{b(i)}(t)$, $i = 1,2,\ldots,k$ of model forecast trajectories for the Δt long analysis time window $(t_a - \Delta t/2 \le t \le t_a + \Delta t/2)$:

$$\overline{\mathbf{x}}^b(t) = k^{-1}\sum_{i=1}^{k}\mathbf{x}^{b(i)}(t), \tag{1}$$

$$\mathbf{P}^b(t) = (k-1)^{-1}\sum_{i=1}^{k}\left(\mathbf{x}^{b(i)}(t) - \overline{\mathbf{x}}^b(t)\right)\left(\mathbf{x}^{b(i)}(t) - \overline{\mathbf{x}}^b(t)\right)^T. \tag{2}$$

An ensemble based Kalman filter calculates a k-member ensemble of analyses, $\mathbf{x}^{a(i)}$, $i = 1,2,\ldots,k$, based on the background information provided by Equations (1) and (2) and the observational information encapsulated in the observation operator $H(\mathbf{x})$, the observation error covariance matrix $\mathbf{R}$ and the vector of observations $\mathbf{y}^o$. The analysis ensemble provides an analysis mean

$$\overline{\mathbf{x}}^a = k^{-1}\sum_{i=1}^{k}\mathbf{x}^{a(i)}, \tag{3}$$

which estimates the atmospheric state at time t_a, and an estimate of the analysis error covariance matrix

$$\mathbf{P}^a = (k-1)^{-1}\sum_{i=1}^{k}(\mathbf{x}^{a(i)} - \overline{\mathbf{x}}^a)(\mathbf{x}^{a(i)} - \overline{\mathbf{x}}^a)^T, \tag{4}$$

which represents the uncertainty in the state estimate provided by the ensemble mean analysis. The analysis mean and error covariance matrix are determined by equations similar to those of a classical extended Kalman filter (e.g., Jazwinski 1970). For example, if $H(\mathbf{x})$ is a linear mapping represented by the matrix $\mathbf{H}$ and all observations occur at the analysis time, then one can write

$$\overline{\mathbf{x}}^a = \overline{\mathbf{x}}^b(t_a) + \mathbf{K}\left(\mathbf{y}^o - \mathbf{H}\overline{\mathbf{x}}^b(t_a)\right) \tag{5}$$

and

$$\mathbf{P}^a = (\mathbf{I} - \mathbf{K}\mathbf{H})\mathbf{P}^b(t_a), \tag{6}$$

where the Kalman gain matrix $\mathbf{K}$ is defined by

$$\mathbf{K} = \mathbf{P}^b(t_a)\mathbf{H}^T\left(\mathbf{H}\mathbf{P}^b(t_a)\mathbf{H}^T + \mathbf{R}\right)^{-1}. \tag{7}$$

(Notice that $\mathbf{P}^a$ does not depend on the observations $\mathbf{y}^o$.) The ensemble mean $\overline{\mathbf{x}}^a$ can be used as the initial condition of a deterministic forecast started at t_a, while the analysis ensemble $\mathbf{x}^{a(i)}$, $i = 1,2,\ldots,k$ provides the initial conditions for the model integrations that produce the background forecast trajectories for the next analysis time window. In this statistical model the estimated background and analysis error statistics are fully flow dependent, since

- the ensemble of background forecast trajectories provides flow-dependent information within the analysis time window and
- the analysis ensemble provides the information needed to propagate the flow-dependent information to the next analysis time window.

2.2 The challenge

Although ensemble-based Kalman Filter data assimilation schemes were proposed as early as 1994 (Evensen 1994; Burgers et al. 1998; Houtekamer and Mitchell 1998), when the development of the LETKF started in 2000, it was unclear whether an ensemble based Kalman filter could ever become a viable alternative to the variational techniques in an operational numerical weather prediction environment. There were three major concerns:

- The largest computationally affordable ensemble size would be too small compared to the degrees of freedom in the background uncertainties for a model that has hundreds of millions of state variables, i.e., the ensemble based estimate of the background error covariance matrix would be rank deficient.
- The linear algebraic calculations to obtain the analysis (ensemble) would be computationally too expensive.
- The severity of model errors would make a stable cycling of the Kalman filter impossible.

In the last six years, several research groups have made significant progress with designing efficient implementations of ensemble-based Kalman filters, while model errors have been found a less serious obstacle than expected originally. We are aware of four successful attempts at assimilating observations of the real atmosphere with an implementations of an ensemble based Kalman filter (Houtekamer et al. 2005; Whitaker et al. 2004 and 2007; Szunyogh et al. 2007, Miyoshi and Sato 2007). Several successful implementations on limited area models, including the quasi-operational University of Washington Real-Time Ensemble Kalman Filter[2] project led by Greg Hakim, have also been reported.

The LETKF addresses the issues of potential rank deficiency and computational efficiency by a unique localization strategy that allows for an efficient implementation on a massively parallel computer. The two key features of this localization strategy are that

- the analysis is obtained independently for each model grid point, or more precisely, for each state variable, which makes processing the different grid points (state variables) in parallel on different processors possible and
- all observations that are thought to have useful information about the atmospheric state at a given grid point (or about a given state variable) are assimilated simultaneously.

2.3 Brief history of the LETKF

The idea of preparing an analysis grid point by grid point using flow-dependent information about the fastest growing background errors was first outlined by Kalnay and Toth (1994). This concept was turned into a mathematical algorithm, called the Local Ensemble Kalman Filter (LEKF), in a series of paper by Ott et al. (2002, 2004). These papers focused on investigating the conditions necessary to ensure that the analysis obtained grid point by grid point resulted in a smooth global analysis field. Here, we highlight only one of these conditions, which has important practical consequences: the local regions have to be sufficiently large to ensure that most of the observations assimilated at a given grid point are also assimilated at the neighboring grid points. Hunt et al. (2004) introduced a 4-dimensional formulation of the LEKF, replacing $\overline{\mathbf{x}}^b(t_a)$ and $\overline{\mathbf{P}}^b(t_a)$ with $\overline{\mathbf{x}}^b(t)$ and $\overline{\mathbf{P}}^b(t)$, $(t_a - \Delta t/2 \leq t \leq t_a + \Delta t/2)$. Finally, Hunt et al. (2007) introduced two important improvements into the LEKF:

[2]Real-time products from this project can be found at http://www.atmos.washington.edu/ enkf/.

- a computationally more efficient algorithm to carry out the linear algebraic calculations for each grid point and

- the flexibility to vary the size of the local regions for observations of different types; this flexibility is necessary for the simultaneous assimilation of conventional (in situ) measurements and satellite radiance observations.

The name of the algorithm was also changed from LEKF to LETKF to acknowledge tha similarity between the mathematical algorithms for a local region in the LETKF and for the global domain in the Ensemble Transform Kalman Filter (ETKF) of Bishop et al. (2001).

2.4 Square-root filters vs. perturbed observation schemes

The LETKF belongs to the family of *square-root filter algorithms* (Tippett et al., 2003). Square-root filters were introduced independently by Bishop et al. (2001), Anderson (2001), Whitaker and Hamill (2002) and Ott et al. (2002) . These schemes calculate $\mathbf{x}^{a(i)}$, $i = 1,2,\ldots,k$,

- first determining the analysis error covariance matrix $\mathbf{P}^a$ based on Equation (6)

- then generating $\mathbf{x}^{a(i)}$, $i = 1,2,\ldots,k$ so that Equation (4) is satisfied.

In contrast, perturbed-observation schemes (e.g., Burgers et al. 1998; Houtekamer and Mitchell 1998) do not explicitly determine $\mathbf{P}^a$. Instead, they prepare k independent analyses assimilating k sets of observations that are obtained by perturbing the observations $\mathbf{y}^o$ by k independently generated sets of observational noise. The observational noise is a random variable with zero mean and covariance matrix $\mathbf{R}$. As k increases, the random noise provides and increasingly more accurate representation of $\mathbf{R}$. For a small ensemble, however, the representation of the contribution of $\mathbf{R}$ to $\mathbf{P}^a$ is affected by sampling errors, a problem that does not affect the square-root filters. Thus, a squre-root filter provides a more accurate estimate of $\mathbf{P}^a$, which, in practice, allows for the use of a smaller ensemble. This effect is best illustrated by Whitaker and Hamill (2002).

Square-root filters are sometimes called *deterministic filters*, while the perturbed-observation schemes are sometimes called *probabilistic filters*. This terminology can be misleading, since both types of schemes treat the error components as probabilistic variables and use the model to deterministically propagate the error statistics.

2.5 Simultaneous vs. serial assimilation of the observations

Obtaining the analysis grid point by grid point, as we do in the LETKF, has been considered computationally less efficient than assimilating the observations serially (Evensen 2006). Serial schemes update the state estimate at all grid points where the given observation can provide some information about the state. However, when implemented on a parallel computer, the LETKF is significantly more efficient than a serial scheme. For instance, Whitaker et al. (2007) found that, for the model and observational data set considered in their paper and in Szunyogh et al. (2007), the serial scheme was nearly an order of magnitude more expensive than the LETKF in terms of computational wall clock time. For completeness, it should be added, that this advantage may diminish, as also demonstrated by Whitaker et al. (2007), when there is significant redundancy between observations. In that case, the serial scheme allows for an efficient data thinning procedure that eliminates a large portion of the observations without compromising the accuracy of the analyses.

2.6 Implementation of the LETKF on the NCEP GFS model

In what follows, we explain how the LETKF algorithm of Hunt et al (2007) is implemented on the NCEP GFS. The analysis calculations are carried out in model grid space. Since the NCEP GFS model is a spectral-transform model, performing the grid space calculations requires that the members of the background ensemble first be transformed from spectral to grid space. Once the analysis step is completed, the analysis ensemble is transformed back from grid to spectral space. The components of the state vector $\mathbf{x}$ in our description of the LETKF algorithm are the model grid point variables.

Following the convention at NCEP, we prepare analyses with six-hour frequency at 0000UTC, 0600UTC, 1200UTC, and 1800UTC, using observations from a 6-hour window centered at the analysis time. The background trajectories are obtained by integrating each ensemble member for 9 hours starting from the analysis ensemble of the previous analysis time. That is, the members of the global background ensembles are all 3- to 9-hour lead time forecast trajectories. To suppress gravity wave oscillations, the background trajectories are initialized with a digital filter algorithm (Lynch and Huang 1992) employing a 3-hour half window.

The data assimilation procedure we follow for conventional (non-radiance) observations is summarized by the nine steps given below. The modifications to the algorithm that are needed for the assimilation of satellite radiance observations, in addition to the assimilation of conventional observations, are described in Section 4.

1. The nonlinear observation operator H is applied to each ensemble background trajectory, $\mathbf{x}^{b(i)}(t)$, $i = 1,2,\ldots,k$, to obtain the *global background observation ensemble* $\{\mathbf{y}^{b(i)}\}$, $i = 1,\ldots,k$. The ensemble average of the background ensemble is calculated in observational space by calculating the ensemble average $\overline{\mathbf{y}}^b$ of $\{\mathbf{y}^{b(i)}\}$. Then a *global background observation ensemble perturbations matrix* $\mathbf{Y}^b$ is constructed by taking its columns to be the vectors obtained by subtracting $\overline{\mathbf{y}}^b$ from each ensemble member $\mathbf{y}^{b(i)}$. Here, H is a four-dimensional observational operator that interpolates the background ensemble members to the time and location of the observations. The time interpolation is carried out by outputting the background ensemble trajectories with a 1-hour frequency and applying a linear interpolation to obtain the background state at the observation time. In the two horizontal spatial dimensions, H is a simple bilinear interpolation. Since the vertical coordinate in the NCEP GFS model is sigma (defined as the ratio of the pressure to the surface pressure), while the vertical coordinate of the observations is pressure, the vertical interpolation is somewhat more complicated than the temporal and the horizontal interpolations: for each ensemble member, we first calculate the pressure at the sigma levels, multiplying sigma by the background surface pressure of the given ensemble member; the 28 sigma levels define 28 sigma layers, where the lowest layer is defined by the surface (where sigma is 1) and the lowest sigma level; we find the sigma layer that contains the observation; finally we linearly interpolate using the logarithm of the pressure values at the bottom and top of the sigma layer. Since the logarithm of the surface pressure is part of the state vector, H is a nonlinear function for all observations.

2. The *global background ensemble perturbation matrix* $\mathbf{X}^b$ is constructed by first taking the ensemble mean $\overline{\mathbf{x}}^b(t_a)$ of $\{\mathbf{x}^{b(i)}(t_a)\}$ at the analysis time and then taking the i-th column of the matrix to be $\mathbf{x}^{b(i)}(t_a) - \overline{\mathbf{x}}^b(t_a)$.

3. All data needed to obtain the analysis ensemble at a given grid point are selected. Below, $\overline{\mathbf{y}}^b_{[\ell]}$ and $\mathbf{Y}^b_{[\ell]}$ refer to the rows of $\overline{\mathbf{y}}^b$ and $\mathbf{Y}^b$, which represent *local background* information at the location of the observations that were chosen for the calculation of the *local analysis ensemble*, $\{\mathbf{x}^{a(i)}_{[\ell]}\}$, at the given grid point, likewise, $\mathbf{y}^o_{[\ell]}$ is the *local vector of observations*, and $\mathbf{R}_{[\ell]}$ is the *local observation error covariance matrix* for the chosen set of observations. (Hereafter, the subscript $[\ell]$ refers to a local region associated with an arbitrary grid point.)

4. The matrix $\mathbf{C}_{[\ell]} = (\mathbf{Y}^b_{[\ell]})^T \mathbf{R}^{-1}_{[\ell]}$ is computed. We define the entries of $\mathbf{R}_{[\ell]}$ by the values provided by NCEP in the operational observational data files. In addition, we assume that the observational errors

are uncorrelated. This assumption makes $\mathbf{R}_{[\ell]}$ diagonal, thus the calculation of $\mathbf{C}_{[\ell]}$ is computationally inexpensive. [Hunt et al. (2007) suggest solving $\mathbf{R}_{[\ell]}\mathbf{C}_{[\ell]}^T = \mathbf{Y}_{[\ell]}^b$ for $\mathbf{C}_{[\ell]}$ when $\mathbf{R}_{[\ell]}$ is non-diagonal.] In some of the experiments, we reduce the weight of observations that are located further than a given distance (e.g., 500 km in this paper) from the analysis grid point. We achieve this by multiplying the entries of $\mathbf{R}_{[\ell]}^{-1}$ by a factor $\mu(r) \leq 1$, which is a monotonically decreasing function of the distance, r, between the location of the analysis grid point and the location of the observation.

5. The eigensystem of $\left[(k-1)\mathbf{I}/\rho + \mathbf{C}_{[\ell]}\mathbf{Y}_{[\ell]}^b\right]$ is determined. The eigenvectors and eigenvalues are then used to calculate the matrix $\tilde{\mathbf{P}}_{[\ell]}^a = \left[(k-1)\mathbf{I}/\rho + \mathbf{C}_{[\ell]}\mathbf{Y}_{[\ell]}^b\right]^{-1}$. Here $\rho \geq 1$ is a multiplicative covariance inflation factor. In our implementation, ρ is a smoothly varying three-dimensional scalar field. (See Section 3.1 for more details.)

6. The matrix $\mathbf{W}_{[\ell]}^a = [(k-1)\tilde{\mathbf{P}}_{[\ell]}^a]^{1/2}$ is computed using the eigenvalues and eigenvectors calculated in the previous step.

7. The vector $\bar{\mathbf{w}}_{[\ell]}^a = \tilde{\mathbf{P}}_{[\ell]}^a \mathbf{C}_{[\ell]}(\mathbf{y}_{[\ell]}^o - \bar{\mathbf{y}}_{[\ell]}^b)$ is computed and then added to each row of $\mathbf{W}_{[\ell]}^a$. The columns of the resulting matrix are the *weight vectors* $\{\mathbf{w}_{[\ell]}^{a(i)}\}$.

8. The analysis ensemble members $\{\mathbf{x}_{[\ell]}^{a(i)}\}$ at the analysis grid point are obtained from $\mathbf{x}_{[\ell]}^{a(i)} = \mathbf{X}_{[\ell]}^b \mathbf{w}_{[\ell]}^{a(i)} + \bar{\mathbf{x}}_{[\ell]}^b$; here $\mathbf{X}_{[\ell]}^b$ and $\bar{\mathbf{x}}_{[\ell]}^b$ represent the rows of $\mathbf{X}^b$ and $\bar{\mathbf{x}}^b(t_a)$ corresponding to the analysis grid point.

9. After completing step (iii)-(viii) for each grid point, the results of step (viii) are collected to form the global analysis ensemble $\{\mathbf{x}^{a(i)}\}$.

A detailed discussion of the parallel computer implementation of the algorithm and the assessment of one particular code implementation on a Beowulf cluster are presented in Szunyogh et al. (2007).

3 Results with conventional observations

Our implementation of the LETKF on the NCEP GFS model has been validated in three steps, gradually adding realistic features to the observing network:

- A time series of "true" states was generated by an integration of the GFS model at T62L28 resolution starting from an operational NCEP analysis truncated to T62L28 resolution. Simulated vertical soundings were generated by adding random observational noise, with a presumed covariance matrix $\mathbf{R}$, to the "true" grid point values of the surface pressure, virtual temperature, and two horizontal components of the wind vector. Sensitivity to the observational coverage was investigated by rejecting observations at a given percentage of randomly selected locations. A detailed assessment of the results from these experiments was published in Szunyogh et al. (2005). Here, we show results only for the case when observations are retained at 10% of the grid points.

- Using a time series of "true" states simulated observations were generated at the locations of the observations of the real atmosphere. To obtain the location and the type of the observations we used, a data base that included all observations that were operationally assimilated at NCEP between 1 January 2004 0000UTC and 29 February 2004 1800UTC, with the exception of satellite radiance observations, but including all satellite-derived wind observations. We also excluded all types of surface observations except for the surface pressure. Some results from these experiments were published in Szunyogh et al. (2007).

- The observations of the real atmosphere, which were used to obtain the locations and types of the observations in the second set of simulated observations experiments, were assimilated. The most important results from these experiments were published in Szunyogh et al. (2007).

3.1 Results with observations of the real atmosphere

One objective of our numerical experiments was to compare the performance of the LETKF with that of a state-of-the-art operational data assimilation system. The state-of-the-art reference data assimilation system used in the Szunyogh et al. (2007) was the Spectral Statistical Interpolation (SSI) of NCEP (Parrish and Derber 1992), one particular implementation of 3D-Var, which was used in operation at NCEP until April 2007. NCEP generously provided analyses that were generated with the SSI using the version of the model (at T62L28 resolution) and the observational data set, which we used in our LETKF experiments. In what follows, we refer to this data set as the *NCEP benchmark*. NCEP also made available their operational high-resolution (T254L64) analyses for the two-month period of the study. We used these operational analyses for verification purposes after truncating them to the T62L28 resolution. The most important differences between the NCEP operational and benchmark systems were in their resolution and the use of satellite radiances in the operational system. [The same NCEP benchmark and operational data sets are also used in Whitaker et al. (2007).]

The parameters of the LETKF were set to the following values:

- The ensemble had $k = 60$ members.
- In the horizontal direction, observations were considered from an 800-km radius neighborhood of the grid point at which the state was to be estimated.
- The observations had equal weight [$\mu = 1$, see step (iv) of the LETKF algorithm] within a 500-km radius of the grid point, beyond which the weight of the observations, μ decreased linearly to zero at 800 km.
- In the vertical direction, observations were considered from a layer around the grid point, and this layer had a depth of 0.35 scale height between model levels 1 and 15 (below sigma = 0.372), above which the depth gradually increased to 2 scale heights at the top of the atmosphere (defined by sigma = 0.003). Here, the scale height was defined as the vertical distance in which the surface pressure decreased by a factor of *e*, that is, the unit-scale-height deep vertical layer was defined by $\ln(\sigma_1/\sigma_2) = 1$, where σ_1 was sigma at the bottom of the layer and σ_2 was sigma at the top of the layer.
- The covariance inflation coefficient ρ tapered from 1.25 at the surface to 1.2 at the top of the model atmosphere in the SH extratropics and from 1.35 to 1.25 in the NH extratropics, while ρ changed smoothly throughout the tropics (between 25° S and 25° N) from the values of the SH extratropics to the values of the NH extratropics.

For further implementation details see Szunyogh et al. (2007). Here, we show only one representative result: the comparison of 48-hour forecasts started from the LETKF and the NCEP benchmark analyses. (Figure 3.1). In the SH extratropics the LETKF forecasts are more similar to the operational analyses throughout the entire depth of the model atmosphere. In the NH extratropics, the performance of the two data assimilation systems is similar: the LETKF provides more accurate predictions of all variables at altitudes higher than the 100 hPa pressure level, while below that level, the benchmark shows slightly better agreement with the operational model. Although the differences between the forecast errors are small at most levels, these differences are statistically significant except for the geopotential height below 300 hPa and the wind between 200 and 100 hPa.

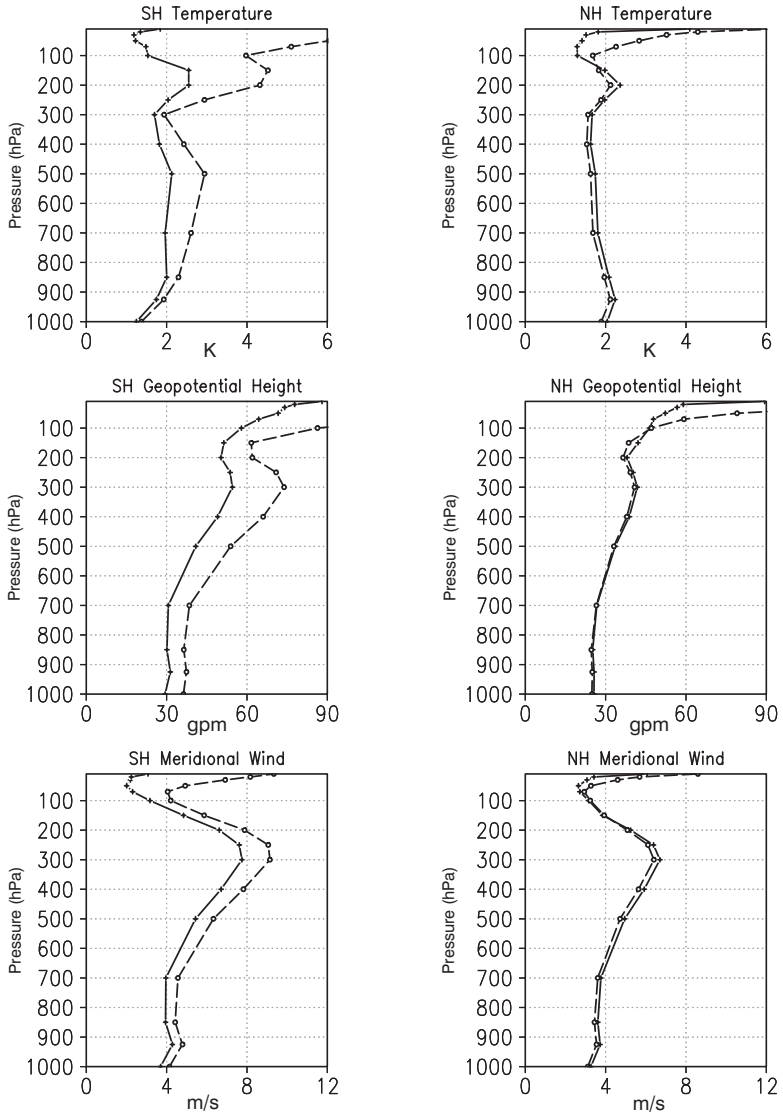

Figure 1: Root-mean-square error of the 48-hour LETKF forecasts (crosses connected by solid lines) and the 48-hour NCEP benchmark forecasts (open circles connected by dashes). The error is estimated using the operational NCEP analyses as proxy for truth. Left panels show the results for the SH extratropics, while right panels show the results for the NH extratropics. The averages are taken over all model grid points south of $20^\circ N$ *in the SH and over all model grid points north of* $20^\circ N$ *in the NH, and over all forecasts started between 11 January 2004 0000UTC and 27 February 2004 1800UTC. (The last two days of February are not included in the verification, because the verifying data set was not available beyond the end of February.) Figure is from Szunyogh et al. 2007.*

3.2 The role of flow-dependent information

In an ensemble based Kalman filter, the background ensemble perturbations provide a representation of the background uncertainty. *The analysis can reduce the error in the background when good quality observation are available to make the correction in the space captured by the background ensemble.* The role of the analysis ensemble $\mathbf{x}^{a(i)}$ is to represent the space of uncertainties taking into account the corrections in the state estimate achieved by the assimilation of the observations.

In reality, a finite size ensemble cannot fully capture the space of uncertainties and observations are not always available to make the corrections in the well-captured directions of uncertainties. We have found the experiments with simulated observations extremely useful in exploring the flow dependence of the performance of the ensemble in capturing the background and analysis uncertainties. In this case the "true" state is known and the difference between the analysis or the forecast and the true state can be projected on the ensemble perturbations. Since in the LETKF the ensemble based covariance and the observations are considered only from a local volume around the grid points, the projection can be calculated for each grid point by calculating the projection based on only those state vector components that are associated with grid points from the given local region. To obtain a quantitative diagnostic quantity, we normalize the square of the projection of the error on the ensemble with the square of the magnitude of the difference between the forecast and the "true state" . In the ideal case, when the error is fully captured by the ensemble, this quantity, called the *explained variance* in our papers, is equal to 1. When the ensemble completely misses the error, the explained variance is 0.

Another diagnostic quantity that we have found extremely useful in explaining the behavior of the LETKF is the *E-dimension* (Ensemble dimension) which was originally introduced by Patil et al. (2001) and was described in detail in Oczkowski et al. (2005). The E-dimension measures the steepness of the eigenvalue spectrum of the ensemble based estimate of the analysis or forecast error covariance matrix. Numerically, the formula for the E-dimension is

$$E = \frac{\left(\sum_{i=1}^{k} \sqrt{\lambda_i}\right)^2}{\sum_{i=1}^{k} \lambda_i}, \tag{8}$$

where λ_i, $i = 1,\ldots,k$, are the eigenvalues of the covariance matrix. Similarly to the explained variance, the E-dimension can be calculated for each grid point based on information from the related local region. The E-dimension is a real number between 1 and k. The lower the E-dimension, the steeper the eigenvalue spectrum of the covariance matrix. The heuristic meaning of the E-dimension is that in the case of low E-dimension, a few orthogonal error patterns dominate the field of uncertainty captured by the ensemble, and in the case when the E-dimension is high, many equally important error structures contribute to the ensemble based estimate of the uncertainty. [For further illustration of the properties of the E-dimension on simple examples see Oczkowski et al. 2005)].

The relationship between error growth, explained variance and E-dimension is illustrated by Figure 3.2. Two panels are shown for each of the three experiments: the top panels are for the randomly distributed simulated observations, the middle panels are for the realistically distributed simulated observations, and the bottom panels are for the observations of the real atmosphere. The left panels are for analysis time, while the right panels are for 120-hour forecast lead time.There is a simple relationship between the E-dimension and the explained variance: *the lower the E-dimension, the more likely that the explained variance is high.* This is a property that is independent of the experimental design and the forecast lead time.

Interestingly, the distribution of the E-dimension with the explained variance at analysis time is more similar for the two experiments, in which the observations are realistically distributed (two lower left panels). This indicates that the relationship between E-dimension and explained variance at analysis time is more affected by the spatial distribution of the observations than by the model errors. What makes this result even more interesting is the fact that the model errors have a large effect on the magnitude of the analysis and forecast errors: the errors for the real atmosphere are about twice as large as for the perfect model scenario.

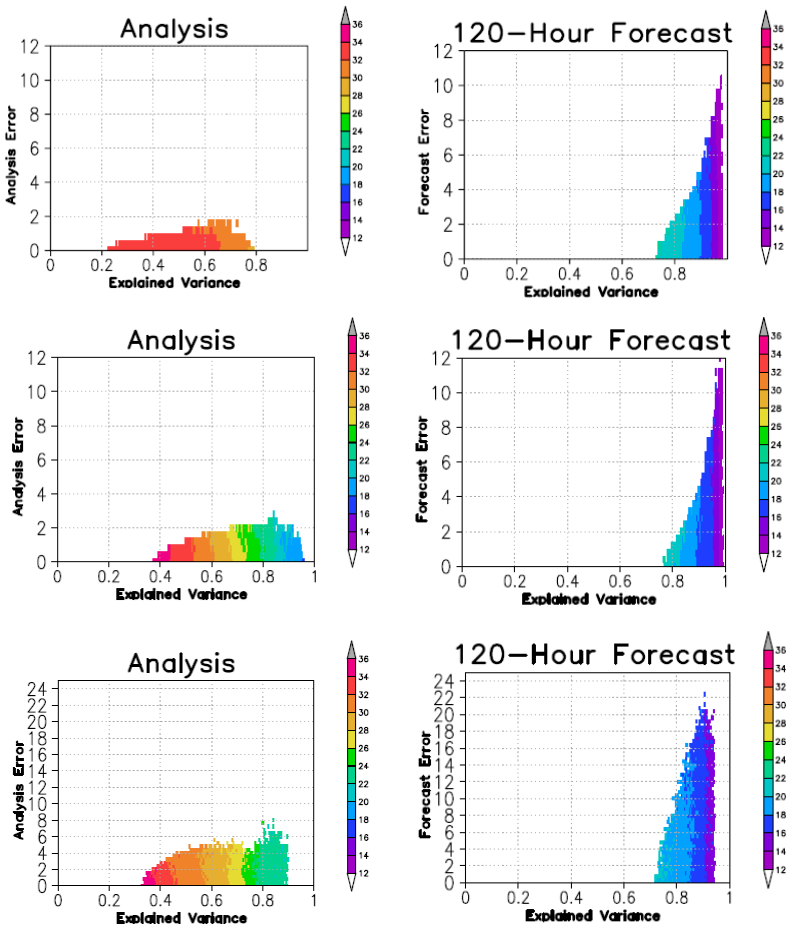

Figure 2: Joint distribution of the analysis/forecast error and the explained variance for the Northern Hemisphere extra-tropics. The bin increments are 0.005 for the explained variance and 0.4 for the forecast error. Color shades indicate the mean E-dimension for each non-empty bin. Shown are the distributions for the experiments that assimilate randomly distributed simulated observations (top panels), simulated observations at the locations of the observations of the real atmosphere (middle panel) and observations of the real atmosphere (bottom panel). Left panels show the distribution at analysis time, while the right-hand-side panels show the distribution at 120-hour forecast lead time. Note the different scale for the analysis/forecast errors in the bottom panels. The statistics were collected for forecasts stared over the periods of 15 January, 2000, 0000 UTC- 15 February, 2000, 1200 UTC (top panels), 11 January, 2004, 0000 UTC- 29 February, 2004, 1200 UTC (middle panels) and 11 January, 2004, 0000 UTC- 15 February, 2004, 0000 UTC (bottom panels).

The larger the forecast error the more certain that the E-dimension is low. Also, the lower the E-dimension the more certain that the explained variance is high (right-hand-side panels). In other words, the lower the predictability the more certain that the predictability of predictability is high. In practical terms, *when the forecast error is large we can expect the ensemble to do a good job in capturing the space of uncertainties.* This general rule applies to both the perfect model and the real atmosphere. The only major difference between the two cases is that for the real atmosphere the explained variance saturates at around 0.9 instead of 1 and that the lower bound for the explained variance increases less rapidly with the forecast error.

Another diagnostic is the spread-skill relationship: the correlation between the spatio-temporally changing fields of the root-mean-square error in the ensemble mean forecast and the ensemble spread. It may seem that for a well-designed ensemble prediction system this correlation should always be high. The reality, however, is more complex than that (e.g., Barker 1991; Houtekamer 1993; Whitaker and Loughe 1998, Kuhl et al. 2007). In particular, when the error field is dominated by random errors the correlation is expected to be nearly zero. The top panel of Figure 3.2 depicts such a case for the analysis time in the two extratropics: 10% coverage of randomly placed observations can efficiently remove errors with well-defined structures, the correlation at analysis time is nearly zero. As the errors grow with time, and well-defined structures develop, the correlation also grows rapidly reaching a maximum value of about $0.4-0.5$ at around 60-hour forecast lead time. When the simulated observations are placed at the locations of the conventional observations of the atmosphere, the initial correlation in the extratropics becomes higher, especially in the Southern Hemisphere (middle panel of Figure 3.2) . This indicates that the observational coverage, especially in the Southern Hemisphere, is not sufficient to remove all background errors correctly captured by the ensemble. (This result also explains the presence of low-dimensional regions and the associated high explained variance at analysis time in the middle-left panel of Figure 3.2.) The picture does not change much when the simulated observations are replaced with observations of the real atmosphere: the correlation at analysis time is 0.4 in the Southern Hemisphere and 0.32 in the Northern Hemisphere (bottom panel of Figure 3.2). The main difference between the figure for the perfect model scenario and the figure for the real atmosphere is the lower correlation for the case of the real atmosphere at the longer forecast times. This is most likely due to the effects of model errors.

The correlation in the tropics behaves differently from the correlation in the extratropics. For the randomly distributed simulated observations the initial correlation is relatively high (about 0.31). This indicates that a 10% observational coverage is not sufficient to remove all background errors that are correctly identified by the ensemble. When the observations are taken at realistic locations (middle and bottom panel), at analysis time the correlation is the lowest in the tropics, which is primarily due to the larger correlations in the extratropics in this case.

4 Preliminary assimilation results with satellite radiance observations

We have recently implemented the CRTM radiative transfer model of JCSDA on our implementation of the LETKF on the NCEP GFS model. We illustrate our capability to assimilate satellite radiance observations by assimilating observations taken by the AMSU-A sensor of the Aqua satellite. The application of the LETKF to satellite radiance observations and to conventional observations is different in the following important details:

- In step 1 of the LETKF algorithm $H(\mathbf{x})$ is replaced by $H(\mathbf{x})+b(\mathbf{x})$, where $H(\mathbf{x})$ is calculated by the CRTM and $b(\mathbf{x})$ is a state dependent bias component. In our current implementation, we assume that the bias can be efficiently approximated by $b(\mathbf{x}) = \alpha T_s + \beta s + \gamma$, where T_s is the skin temperature, s is the scan angle, and α, β and γ are global constants, i.e., they are the same for all observations of a given satellite channel. The constants α, β and γ for the different satellite channels are estimated as part of the LETKF algorithm.
- The estimates of the parameters α, β and γ for the different satellite channels, which are needed to calculate $b(\mathbf{x})$ are obtained by the method of *state augmentation*, i.e., by applying the LETKF to an

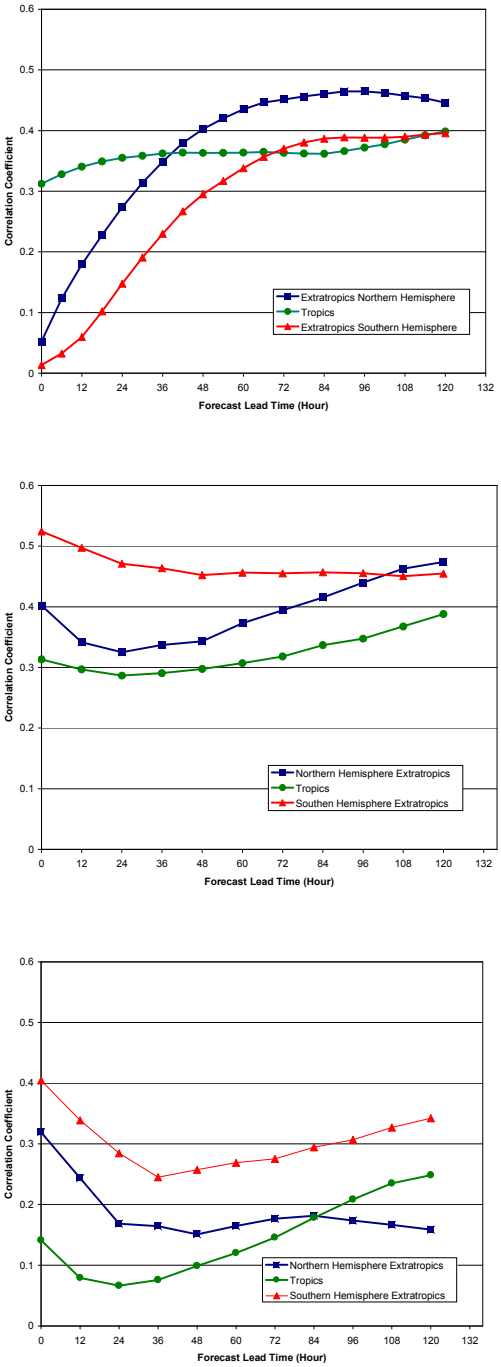

Figure 3: Correlation between the root-mean-square error of the ensemble mean forecast and the ensemble spread. The analyses are obtained by assimilating simulated observation at random locations (top panel) simulated observations at the locations of the conventional observations of the real atmosphere (middle panel) and conventional observations of the real atmosphere (bottom panel). Results are shown for three different geographical area: Northern Hemisphere extratropics (blue), Southern Hemisphere extratropics (red) and tropics (green). The average is taken over all forecasts started between 11 January 2004 0000UTC and 25 February 2004 1800UTC.

augmented state vector $\mathbf{z}$ instead of $\mathbf{x}$. The components of $\mathbf{z}$ are composed of the components of $\mathbf{x}$ and the parameters α, β and γ for the different satellite channels.

- An estimate of α, β and γ for each satellite channel is obtained for each model grid point at which satellite radiance observations are assimilated. The global estimate of these parameters for the different satellite channels is obtained by a weighted average of the estimates for the different grid points. (See Ferig et al. 2007b for a detailed description and an evaluation of the performance of this bias correction strategy.)

- Different localization strategies are applied to the conventional and the satellite radiance observations in step 3 of the localization algorithm:

  – While the calculation of $H(\mathbf{x})$ for the conventional observation is typically based on background information from a strongly localized region around the observation, the calculation of $H(\mathbf{x})$ with the CRTM is based on the background atmospheric state from the entire vertical column of the model atmosphere. (We recall that in the LETKF, H is applied to the global atmospheric state $\mathbf{x}$ and that the localization is done afterwards in observation space.)

  – A satellite radiance observation is assimilated at all model levels at which the *weighting function* in the radiative transfer model is larger than a given percentage of the maximum of the weighting function. [A detailed theoretical and experimental justification of this approach is provided by Fertig et al. (2007a)]. The results show below are obtained with a cut-off value of 60%.

Here, we show only one particular result to demonstrate the positive effects of the AMSU-A radiance observations on the accuracy of the analysis (Figure **??**). These observations have a large positive impact on the short term (48-hour) forecast accuracy in the Southern Hemisphere and a mixed impact elsewhere. Since the observation are primarily associated with the virtual temperature state variable of the atmosphere, the large improvement in the meridional component of the wind in the Southern Hemisphere indicates, that the ensemble-based estimate of the covariance between the temperature and wind background errors is sufficiently accurate to lead to a major improvement of the wind analysis.

5 Concluding remarks

Our results suggest the flow-dependent background error covariance information improves the accuracy of the analysis, the short-term deterministic forecasts and the short-term ensemble forecasts in, and downstream of, the regions where the observational coverage is relatively sparse. We have also demonstrated that the LETKF algorithm is an efficient practical approach to utilize the flow-dependent information:

- The LETKF can be efficiently implemented on a parallel computer architecture.

- At the reduced model resolution and for the reduced observational data set used in our numerical experiments, the LETKF provides more accurate analyses and forecasts than the SSI in regions of sparse observational coverage and comparably accurate analyses and forecasts in regions of dense observational coverage.

- Preliminary results with AMSU-A observations suggest that the bias in the satellite radiance observations can be efficiently estimated and corrected with the LETKF data assimilation scheme.

We believe that the main advantage of the LETKF over other formulations of the ensemble based Kalman filter is its computational efficiency, since theoretical considerations and the results of numerical experiments suggest that other formulations of the ensemble based Kalman filter should be similarly accurate. We also note that the

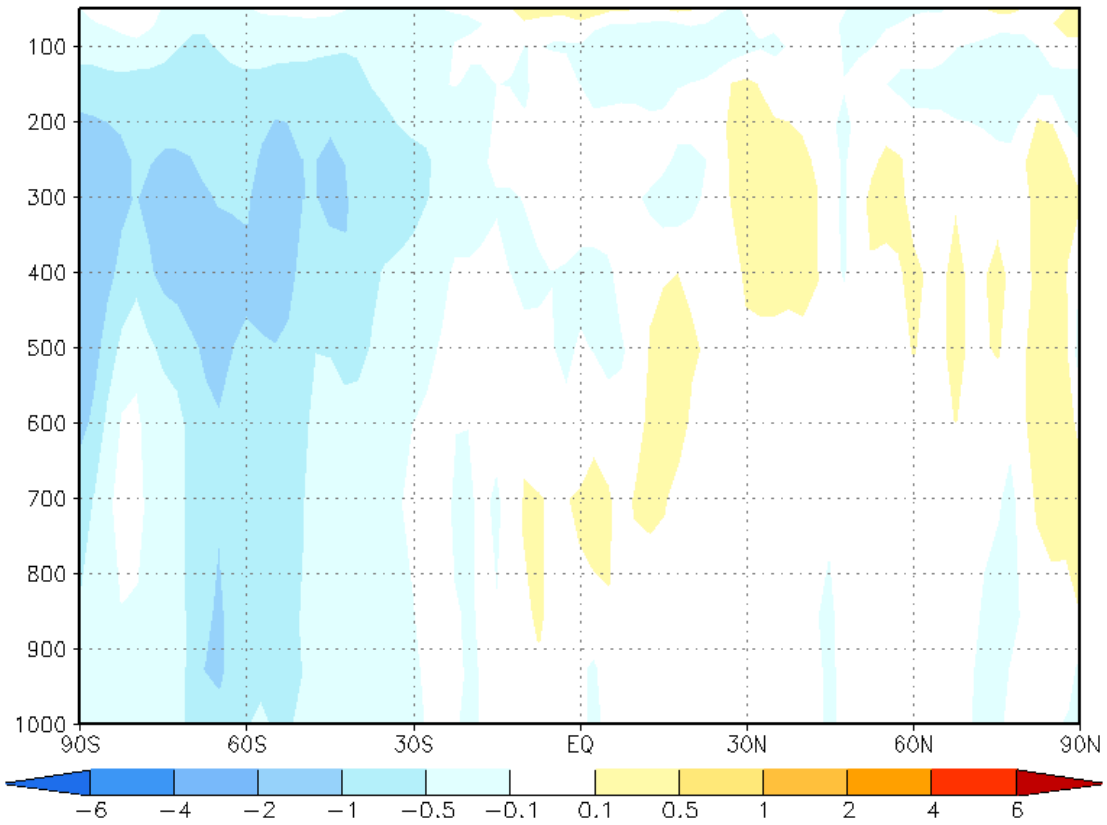

Figure 4: Height-latitude cross section of the impact of the assimilation of AMSU-A radiance observations on the quality of 48-hour forecasts of the meridional wind. Shown is the difference between the zonal mean root-mean-square errors in the forecasts with AMSU-A and conventional observations and the forecasts with only conventional observations . Negative values indicate locations where the AMSU-A observations improve the forecasts.The average is taken over all forecasts started between 16 January 2004 0000UTC and 27 February 2004 1800UTC. The color shades show the difference in ms^{-1}.

current version of the LETKF algorithm can also be viewed as one particular implementation of a general framework for an efficient implementation of the ensemble based Kalman filter. For instance, the equation to update the state estimate (to obtain the analysis) could be replaced by a variational scheme; one could also add balance constraints to filter gravity waves, either by modifying the current update equation or within the variational framework.

While the results so far are encouraging, there are many important issues left to explore:

- Our results and the results of others working with ensemble based Kalman filters, most importantly those of Whitaker et al. (2007), show little advantage of using flow-dependent information in regions of dense observational coverage. Since these results were obtained at model resolutions that are much lower than the current operational model resolutions, the question arises whether we can expect to find more advantages of using flow-dependent information about the estimation errors when the resolution is increased to the resolution of the current and future operational models.

- Closely related to the previous issue, the results suggests that increasing the observational density may have a similarly positive effect to that obtained by using flow dependent information about the error estimates. Considering the rapid increase of the number of remotely sensed observations, how much improvement can we expect in the future from incorporating flow dependent information into the data assimilation process?

- One potential advantage of an ensemble based data assimilation system is that it provides initial conditions for an ensemble prediction system at no extra cost and in a way that is consistent with the data assimilation system. Little attention has been paid, so far, to the evaluation of the medium- and extended-range ensemble predictions that can be produced with the initial conditions generated by the ensemble based data assimilation systems.

- Current data assimilation systems cannot take into account directly the effects of model errors. Estimation and correction of the effects of model errors is, in principle, possible within both the variational framework (weak-constraint formulation) and the ensemble based Kalman filter framework (e.g., Baek et al. 2006). It is unclear which framework provides a more efficient way to estimate, and to correct for, the effect of model errors.

Acknowledgements

The current version of the LETKF data assimilation system is the result of six years of theoretical work and code development. Initial funding for this project was provided by the W. M. Keck Foundation and the James S. McDonnell Foundation through a 21st Century Research Award. Our efforts have also been funded by the Army Research Office, by the NASA AIRS program, by the Office of Naval Research (Physics), and by the National Science Foundation (Grants #0104087 and PHYS 0098632). The implementation of the LETKF on the NCEP GFS has been carried out as part of a project for the comparison of different ensemble-based Kalman filter algorithms. This project has been organized by Zoltan Toth of NCEP and was supported by a NOAA THORPEX grant. The NCEP benchmark analysis and forecast was prepared by Yocheng Song of NCEP and the operational observational data files were made us available by Richard Wobus. We have greatly benefited from discussions with members of the other teams participating in the comparison project. We are especially thankful to Jeff Whitaker of the Earth System Research Laboratory of NOAA for sharing the experience he accumulated with other implementations of ensemble-based Kalman filters on the NCEP GFS. E. J. K. gratefully acknowledges partial support from the NSF Interdisciplinary Grants in the Mathematical Sciences Program under grant number DMS-0408102.

References

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

Baek, S.-J., B. Hunt, E. Kalnay, E. Ott, and I. Szunyogh,2006: Local ensemble Kalman filtering in the presence of model bias.*Tellus,* **58A,** 293-306.

Barker, T. W., 1991: The relationship between spread and forecast error in extended-range forecasts. *J. Climate,* **4,** 733-742.

Bishop, C. H., B. J. Etherton, and S. Majumdar, 2001:Adaptive sampling with the Ensemble Transform Kalman Filter.Part I: Theoretical aspects. *Mon. Wea. Rev.*, **129,**420-436.

Burgers, G., P. J. van Leeuwen, and G. Evensen, 1998: Analysis scheme in the ensemble Kalman filter. *Mon. Wea. Rev.*, **126,** 1719-1724.

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.

Evensen, G., 2007: *Data assimilation: The ensemble Kalman filter*. Springer, Berlin.

Fertig, E. J., B. R. Hunt, E. Ott, and I. Szunyogh, 2007a: Assimilating nonlocal observations with a local ensemble Kalman filter. *Tellus*, to appear.
Preprint available at http://www.weatherchaos.umd.edu/publications.php.

Fertig, E. J., and coauthors, 2007b: Observation bias correction with an ensemble Kalman filter. *Tellus*, under review.
Preprint available at http://www.weatherchaos.umd.edu/publications.php.

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

Houtekamer, P. L., 1993: Global and local skill forecasts. *Mon. Wea. Rev.*, **121,** 1834-1846.

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

Houtekamer, P. L., H. L. Mitchell, G. Pellerin, M. Buehner, M. Charron, L. Spacek, and B. Hansen, 2005: Atmospheric data assimilation with an ensemble Kalman filter: Results with real observations. *Mon. Wea. Rev.*, **133,** 604-620.

Hamill, T. M., J. S. Whitaker, and C. Snyder, 2001: Distance-depemdent filtering of background error covariance estimates in an ensemble Kalman filter. *Mon. Wea. Rev.*, **129,** 2776-2790.

Hunt, B. R., E. Kalnay, E. J. Kostelich, E. Ott, D. J. Patil, T. Sauer, I. Szunyogh, J. A. Yorke, and A. V. Zimin, 2004: Four-dimensional ensemble Kalman filtering. *Tellus,* **56A,** 273-277.

Hunt, B. R., E. J. Kostelich, and I. Szunyogh, 2007: Efficient data assimilation for spatiotemporal chaos: a Local Ensemble Transform Kalman Filter. *Physica D,* **230,** 112-126.

Kalnay, E., and Z. Toth, 1994: Removing growing errors in the analysis cycle. *Preprints of the 10th conference on Numerical Weather Prediction, Amer. Meteorol. Soc.*, 1994, pp212-215, Boston, MA.

Jazwinski, A. H., 1970: *Stochastic processes and filtering theory.* Academic Press, San Diego, CA.

Kuhl, D., and coauthors, 2007: Assessing predictability with a local ensemble Kalman filter. *J. Atmos. Sci.*, **64**, 1116-1140.

Lynch, P., and P. M. Huang, 1992: Initialization of the HIRLAM model using a digital filter. *Mon. Wea. Rev.*, **120**, 1019-1034.

Miyoshi, T. and Y. Sato, 2007: Assimilating Satellite Radiances with a Local Ensemble Transform Kalman Filter (LETKF) Applied to the JMA Global Model (GSM). *SOLA*, **3**, 37-40.
Available online at http://www.jstage.jst.go.jp/article/sola/3/0/3_37/_article

Oczkowski, M., I. Szunyogh, and D. J. Patil, 2005:Mechanisms for the development of locally low dimensional atmosphericdynamics. *J. Atmos. Sci.*, **65**, 1135–1156.

Ott, E. and coauthors, 2002: Exploiting local low dimensionality of the atmospheric dynamics for efficient Kalman filtering. *Physics/0203058*.

Ott, E. and coauthors, 2004: A Local Ensemble Kalman Filterfor atmospheric data assimilation. *Tellus*, **56A**,415–428.

Patil, D. J. and coauthors, 2001: Local low dimensionalityof atmospheric dynamics, *Phys. Rev. Lett.*, **86**, 5878–5881.

Parrish, D., and J. Derber, 1992: The National Meteorological Center's spectral statistical interpolation analysis system. *Mon. Wea. Rev.*, **120**, 1747-1763.

Szunyogh, and coauthors, 2005: Assessing a localensemble Kalman filter: Perfect model experiments with the NCEP globalmodel. *Tellus*, **57A**, 528-545.

Szunyogh, and coauthors, 2007: A local ensemble transform Kalman filter data assimilation system for the NCEP global model. *Tellus*, **59A**, to appear.
Preprint available at http://www.weatherchaos.umd.edu/publications.php.

Tippett, M. K., J. L. Anderson, C. H. Bishop, T. M. Hamill, J. S. Whitaker, 2003: Ensemble square-root filters, Mon. Wea. Rev., 131, 1485-1490.

Whitaker, J. S., and A. F. Loughe, 1998: The relationship between ensemble spread and ensemble mean skill. *Mon. Wea. Rev.*, **126**, 3292-3302.

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

Whitaker, J. S., G. P. Compo, X. Wei, and T. H. Hamill, 2004: Reanalysis without radiosondes using ensemble data assimilation. *Mon. Wea. Rev.*, **132**, 1190-1200.

Whitaker, J. S., T. M. Hamill, X. Wei, Y. Song, and Z. Toth, 2006: Ensemble data assimilation with the NCEP Global Forecast System. *Mon. Wea. Rev.*, to appear.
Preprint available by request from the authors.