NSF Collaboration in Mathematical Geosciences Research

FINAL REPORT

Ensemble Data Assimilation System Based on Control Theory

Expired 31 August 2008
CSU Account #: 5-33023
NSF #: 0327651

Milija Zupanski, PI

Cooperative Institute for Research in the Atmosphere
Colorado State University
Fort Collins, CO 80523-1375

Michael Navon, PI
NSF #: 0327818

Department of Scientific Computing
Florida State University
Tallahassee, FL 32306-4120

(Project web page at http://www.cira.colostate.edu/projects/ensemble/NSF_CMG.php)

Summary:

This research project has produced a novel methodology for ensemble data assimilation (EnsDA) based on control theory, named the Maximum Likelihood Ensemble Filter (MLEF). This methodology extends the applicability of current nonlinear filters to arbitrary nonlinear observation operators, with important implications to remote sensing and other nonlinear observations. In addition to the main goal of developing an ensemble data assimilation system based on control theory, at least two major new results were produced by this research: (1) non-Gaussian framework for data assimilation was first formulated within this project, and is now gaining a worldwide recognition, and (2) new non-differentiable unconstrained minimizations algorithms are formulated and tested within this project. In addition, the issue of insufficient number of degrees of freedom (DOF) was addressed during last year (2007-2008), resulting in an improvement of currently used error covariance localization techniques.

To date, this project produced *13* papers (*10* published, *2* submitted, *1* to be submitted, referenced below).

Two post-doctoral researchers are trained under this project; one at Colorado State University (CSU) and one at Florida State University (FSU), resulting in publishing 5 peer-reviewed manuscripts as first authors. Their work resulted in developing new mathematical formalism of non-Gaussian data assimilation, and in developing new methodology for ensemble reduction and inflation based on proper orthogonal decomposition. Both research subjects have been steadily gaining a worldwide interest. During the course of this project the postdoctoral researchers were exposed to state-of-the-art research and attended several international conferences/workshops. Dr. X. Xiong was also trained for one year under this project under the second PI at FSU. He contributed by investigating the particle filters in data assimilation. An additional researcher Dr. Mohamed Jardak was trained for a year by the second P.I. in collaboration with the first P.I. and his research work guided by both P.I.'s resulted in one submitted paper and a work in final stages of preparation.

Two related web pages developed at CSU and FSU are being frequently visited, totaling almost 40,000 hits in only few years, a large number for such focused research. International recognition of this research has been increasing by the day, with several universities and research laboratories in Europe and Asia now using the produced algorithm. In the United States, the MLEF has been used as a teaching tool at FSU, and as a research tool at federal agencies such as National Aeronautics and Space Administration, National Oceanic and Atmospheric Administration, and Department of Defense.

Broader impact of this research also includes new applications of the MLEF to carbon transport problem (Lokupitiya et al. 2008; Zupanski et al. 2007), a critical issue related to climate change. The uniqueness of the MLEF nonlinear capability is also attracting researchers addressing highly nonlinear problems, such as the impact of clouds on climate, and hydrologic applications to river-flow forecasting and data assimilation.

1. Goals and objectives

Our interdisciplinary collaborative research project has the following objectives:

1. Develop and test new Ensemble data assimilation (EnsDA) methodology based on control theory, with capability to assimilate nonlinear observations and employ a non-Gaussian Probability Density Function (PDF) assumption.

2. Evaluate the new methodology in assimilation of nonlinear observations. Pay special attention to Hessian preconditioning in highly nonlinear applications.

3. Make the new methodology available to universities and research institutions. Provide instructions to users.

4. Pending the results, develop new Hessian preconditioning / Optimization applicable to highly nonlinear data assimilation (D/A) applications.

2. Accomplishments

This research produced many important accomplishments over the 5-year period. A dedicated web page at http://www.cira.colostate.edu/projects/ensemble/NSF_CMG.php and the published papers we reference can also serve as a source of more detailed research results.

2.1. The MLEF algorithm

New ensemble data assimilation methodology has been developed. The new methodology / algorithm, named the Maximum Likelihood Ensemble Filter (MLEF). The MLEF algorithm is the first ensemble data assimilation (EnsDA) that explicitly includes control theory, by means of an iterative minimization. As such, it offers a novel solution to a nonlinear analysis problem in EnsDA.

Typically, EnsDA algorithms are developed using an analysis solution similar to the Extended Kalman Filter (EKF). As an improvement of EnsDA, matrices employing nonlinear operators are used in the place of linearized operators. Since the used analysis update is still of the same form as in the Kalman Filter, i.e. a linear matrix equation, there is a hidden assumption in EnsDA that nonlinear operators are not very nonlinear. In the MLEF, however, there is no such assumption. This is generally due to the use of control theory: instead of making assumptions to satisfy a pre-defined linear solution, the MLEF numerically solves an iterative minimization problem. As shown in Zupanski et al. (2008a), the MLEF equations can be derived without assuming linearity, and even without assuming differentiability of involved operators. Here we briefly explain main aspects of the MLEF in a commonly used Gaussian PDF framework.

The algorithmic structure of the MLEF is shown in Fig.1. One can see that the MLEF is built around the minimization algorithm, with separate modules corresponding to

prediction model, observation operator, minimization, and variable update. This allows a straightforward substitute of any of the modules, implying a relatively easy inclusion of other models, observations, and minimization algorithm.

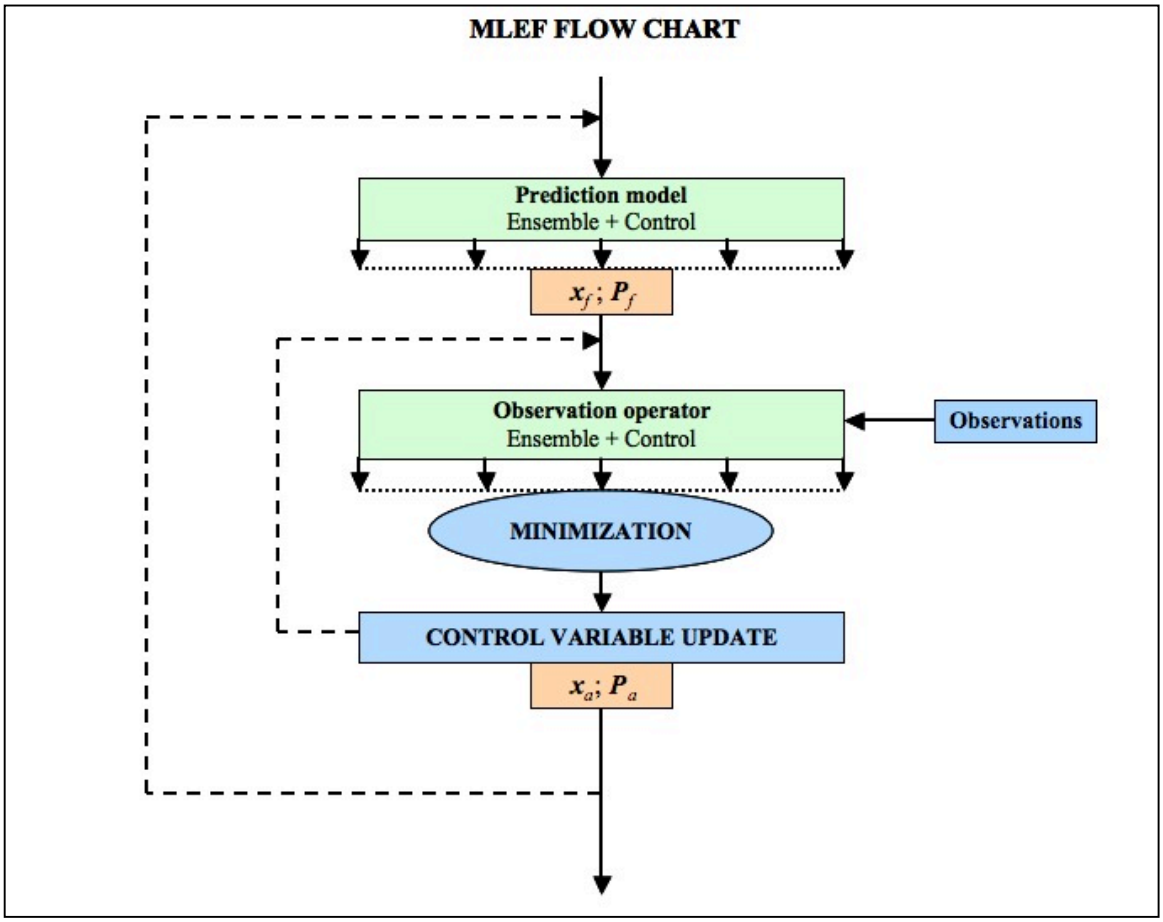

Fig.1. The MLEF algorithm flow chart. One can notice two loops: the inner loop corresponds to iterations of minimization, and the outer loop corresponds to data assimilation cycles. Dashed lines indicate that it is possible to have only one iteration at the time, or multiple, if desired.

2.1.a. Initial ensemble configuration

As in general filtering problem, on first needs an initial guess for the state vector, $\mathbf{x}^0 \in \mathbb{S}$, and its uncertainty given by the square-root error covariance matrix $\boldsymbol{P}_0^{1/2} : \mathbb{E} \to \mathbb{S}$, where $\mathbb{E} \subseteq \mathbb{S}$, $\mathbb{E}$ denoting the ensemble space and $\mathbb{S}$ the state space.

2.1.b. Prediction

The prediction step of the MLEF nonlinearly transports the state vector and its uncertainty in time. After defining a nonlinear prediction model $M : \mathbb{S} \to \mathbb{S}$, one obtains for the transport of the state vector

$$\boldsymbol{x}_t^f = M(\boldsymbol{x}_{t-1}^a) \tag{1.1}$$

where the subscripts t refers to time, and the superscripts f and a to the first guess (forecast) and analysis, respectively. For the transport of uncertainty, described in terms of a square-root forecast error covariance matrix, one obtains

$$\boldsymbol{P}_f^{1/2} = \left[\boldsymbol{p}_1^f \quad \boldsymbol{p}_2^f \quad \cdots \quad \boldsymbol{p}_{N_E}^f \right] \qquad \boldsymbol{p}_i^f = M(\boldsymbol{x}_{t-1}^a + \boldsymbol{p}_i^a) - M(\boldsymbol{x}_{t-1}^a) \tag{1.2}$$

where N_E is the dimension of $\mathbb{S}$ and $\{\mathbf{p}_i^a \in \mathbb{S}\,;(i=1,\ldots,N_E)\}$ are the columns of the square-root analysis error covariance from the previous cycle.

The equations (1.1)-(1.2) represent the prediction step of the MLEF. These equations indicate that ensemble perturbations in the MLEF are used as span-vectors of a dynamical uncertainty space. This implies that the state space dimension naturally limits the maximum number of ensembles in the MLEF, since ensemble perturbations refer to uncertainty of state space. In typical ensemble Kalman filters (EnKF) the desired number of ensembles is much larger (ideally the infinity), a consequence of using ensemble perturbations as a random sample.

2.1.c. Analysis

The analysis update step of the MLEF is based on maximizing the posterior PDF defined using the Bayes rule. This means that an assumption regarding the prior and conditional PDFs has to be made at this point. We now derive the MLEF analysis equations assuming Gaussian errors. In practice, instead of directly maximizing the posterior PDF, an equivalent solution is found by minimizing a negative logarithm of the posterior PDF. With Gaussian PDF assumption, this produces a cost function at time t in the form

$$J(\boldsymbol{x}) = \frac{1}{2}\left(\boldsymbol{x} - \boldsymbol{x}_t^f\right)^T \boldsymbol{P}_f^{-1}\left(\boldsymbol{x} - \boldsymbol{x}_t^f\right) + \frac{1}{2}\left[\boldsymbol{y}_t - h(\boldsymbol{x})\right]^T \boldsymbol{R}^{-1}\left[\boldsymbol{y}_t - h(\boldsymbol{x})\right] \tag{1.3}$$

where $\mathbf{R} : \mathbb{O} \to \mathbb{O}$ denotes the input observation error covariance, $\mathbb{O}$ is the observation space, and $h : \mathbb{S} \to \mathbb{O}$ is a nonlinear observation operator. In practice the inversion of the forecast error covariance in (1.3) is avoided by a change of variable,

$$\boldsymbol{x} - \boldsymbol{x}_t^f = \boldsymbol{G}\zeta \qquad \boldsymbol{G} = \boldsymbol{P}_f^{1/2}\left[\boldsymbol{I} + \left(\boldsymbol{Z}(\boldsymbol{x}_t^f)\right)^T \boldsymbol{Z}(\boldsymbol{x}_t^f)\right]^{-1/2} \tag{1.4}$$

where the new control variable ζ is defined in ensemble space $\mathbb{E}$. The matrix $\mathbf{Z} : \mathbb{S} \to \mathbb{O}$ is defined as

$$\boldsymbol{Z}(\mathbf{x}) = \left[z_1(\mathbf{x}) \quad z_2(\mathbf{x}) \quad \cdot \quad \cdot \quad z_{N_E}(\mathbf{x}) \right] \qquad z_i(\mathbf{x}) = \boldsymbol{R}^{-1/2}\left[H(\boldsymbol{x} + \boldsymbol{p}_i^f) - H(\boldsymbol{x})\right]. \tag{1.5}$$

This matrix is also referred to as the observation information matrix, being related to the Shannon definition of entropy and information measures [Rodgers (2001); Zupanski et al. (2007)]. Note that its value depends on the state vector $\boldsymbol{x}$, thus it has different values at the first guess and at the analysis.

The control variable ζ is iteratively updated using a minimization algorithm, eventually producing the optimal value. It can be shown that the change of variable (1.4) results in perfect preconditioning for quadratic cost function (1.3), implying a single iterative step in that situation. The matrix inversion and square root in (1.4) are accomplished using an

Eigenvalue Decomposition (EVD) algorithm, made applicable due to the $N_E \times N_E$ size of involved matrices. Note that in typical geosciences applications $N_S \gg N_E$, i.e. the matrix that is inverted in (1.4) is of considerably reduced size compared to the sparse $N_S \times N_S$ error covariance matrices in physical space, making manageable the computation. Finally, the analysis uncertainty in the MLEF is expressed through a square-root analysis error covariance defined as

$$\boldsymbol{P}_a^{1/2} = \left[\boldsymbol{p}_1^a \quad \boldsymbol{p}_2^a \quad \cdots \quad \boldsymbol{p}_{N_E}^a \right] \qquad \boldsymbol{p}_i^a = \left(\boldsymbol{P}_f^{1/2} \left[\boldsymbol{I} + \left(\boldsymbol{Z}(\boldsymbol{x}_t^a) \right)^T \boldsymbol{Z}(\boldsymbol{x}_t^a) \right]^{-1/2} \right)_i . \quad (1.6)$$

The equations (1.4)-(1.6) describe the analysis update step of the MLEF. The columns of the square-root analysis error covariance (1.6) are used as the initial forecast ensemble perturbations for the next assimilation cycle, and the algorithm continues to cycle.

2.2. Non-Gaussian framework

In order for non-Gaussian assumptions to be correctly applied, the D/A methodology has to be derived from scratch, i.e. from the Bayes rule. We follow this path in developing the framework for lognormal PDF. In papers by Fletcher and Zupanski (2006a, b) we explore a situation when only observation errors can be lognormal, still assuming Gaussian prior PDF. This is relevant for observations of cloud and aerosol variables, which are by definition positive-definite (i.e. concentrations, mixing ratios) and, as such, can be better represented by a lognormal PDF. In the paper by Fletcher and Zupanski (2008a) we extend this work to assuming that all errors can be of mixed normal-lognormal type, including the prior and the observation errors. This corresponds to a realistic situation with mixed Gaussian and non-Gaussian variables, such as temperature and specific humidity for example. Here we briefly explain the implications in the simplest problem, when only observation errors are lognormal, following Fletcher and Zupanski (2006a).

2.2.a. Cost function

From the very start, assuming lognormal observation errors requires a new definition of errors

$$\varepsilon = \frac{\boldsymbol{y}}{h(\mathbf{x})} \quad (1.7)$$

which also underlines a requirement for lognormal variables to be positive-definite, i.e. that $h(\boldsymbol{x}) > 0$. A multivariate lognormal PDF is

$$f(x) = \frac{1}{(2\pi)^{N/2} \left|\Sigma\right|^{1/2}} \left(\prod_{i=1}^{N} \frac{1}{x_i} \right) \exp\left\{ -\frac{1}{2} (\ln x - \mu)^T \Sigma^{-1} (\ln x - \mu) \right\} \quad (1.8)$$

with expectation μ m and covariance Σ. As a consequence of using (1.8) in Bayes rule, the cost function to be minimized is

$$J(\boldsymbol{x}) = \frac{1}{2}\left(\boldsymbol{x} - \boldsymbol{x}_t^f\right)^T \boldsymbol{P}_f^{-1}\left(\boldsymbol{x} - \boldsymbol{x}_t^f\right) + \frac{1}{2}\left[\ln \boldsymbol{y}_t - \ln h(\boldsymbol{x})\right]^T \boldsymbol{R}_L^{-1}\left[\ln \boldsymbol{y}_t - \ln h(\boldsymbol{x})\right] + \sum_{i=1}^{N}\left[\ln(\boldsymbol{y}_t)_i - \ln h_i(\boldsymbol{x})\right] \tag{1.9}$$

where $\mathbf{R}_L$ is the lognormal observation error covariance. The lognormal cost function (1.9) can be compared with Gaussian cost function (1.3).

2.2.b. Hessian

Aside from the use of a logarithm, one can also notice an extra term that impacts transformation between Gaussian and lognormal variables. This term becomes relevant for gradient and for Hessian calculation, fundamentally impacting the D/A algorithm. The Hessian matrix becomes

$$\frac{\partial^2 J_L}{\partial x_i \partial x_j} = \left[\boldsymbol{P}_f^{-1} + \left(\frac{\partial h}{\partial x}\right)^T \left(\frac{\partial \ln \varepsilon^o}{\partial h}\right)\left(\boldsymbol{R}_L^{-1} + \boldsymbol{I}\right)\left(\frac{\partial \ln \varepsilon^o}{\partial h}\right)\left(\frac{\partial h}{\partial x}\right)\right]_{i,j} - \left[\boldsymbol{G}_i^T\left(\boldsymbol{R}_L^{-1}(\ln y - \ln h(x)) + \left(\frac{\partial \ln \varepsilon^o}{\partial h}\right)\boldsymbol{1}\right)\right]_j \tag{1.10}$$

where $\boldsymbol{G}_i = \frac{\partial}{\partial x_i}\left(\frac{\partial h}{\partial x}\right)$, and $\boldsymbol{1}$ is the identity vector. The form is similar to the Hessian matrix in Gaussian framework

$$\frac{\partial^2 J}{\partial x_i \partial x_j} = \left[\boldsymbol{P}_f^{-1} + \left(\frac{\partial h}{\partial x}\right)^T \boldsymbol{R}^{-1}\left(\frac{\partial h}{\partial x}\right)\right]_{i,j} - \left[\boldsymbol{G}_i^T \boldsymbol{R}^{-1}(y - h(x))\right]_j \tag{1.11}$$

with differences originating from the logarithm and the extra term in cost function (1.9). Important is to note that formal similarity between Hessian expressions (1.10) and (1.11) provides an algorithmic advantage in practical applications of lognormal system, by allowing a simple substitute of programs calculating the Hessian in Gaussian case, without altering other programs. The same is true for gradient calculation.

Our lognormal D/A framework development offers a new methodology for including general non-Gaussian PDFs in D/A. One needs to begin from the Bayes rule, followed by defining the cost function, and its gradient and Hessian matrix. In addition, it is important to understand whether the errors are additive or multiplicative. As a consequence of practical importance, our research indicate that it is possible to develop a general D/A algorithm for multiple choices of PDFs, by employing pre-defined programs for gradient and Hessian calculation.

2.3. Evaluating new EnsDA methodology

The MLEF has been applied to many different problems over the course of this research project.

2.3.a. Nonlinear observation operators

Nonlinearity of observations is an important issue for D/A, since majority of new information is coming from observations. This is especially true for remote sensing observations (satellite, radar), which typically include a highly nonlinear relation between the state and observed variable. The nonlinearity of observations has been one of the weakest points of EnsDA, since it typically employs the explicit linear solution of the Kalman filter (or Extended Kalman filter). Since in the MLEF the nonlinear observation issue is resolved via an iterative minimization, we examined the impact of nonlinear observations on Hessian preconditioning and minimization [Zupanski et al. (2008a)].

Using a 1-dimensional Burgers equation and synthetic observations, we define several nonlinear observation operators and compare the performance of the MLEF and the Fletcher-Reeves nonlinear conjugate-gradient minimization algorithm. The results for the cubic observation operator $h(x) = x^3$ are shown in Fig.2, indicating a clear superiority of the MLEF nonlinear minimization.

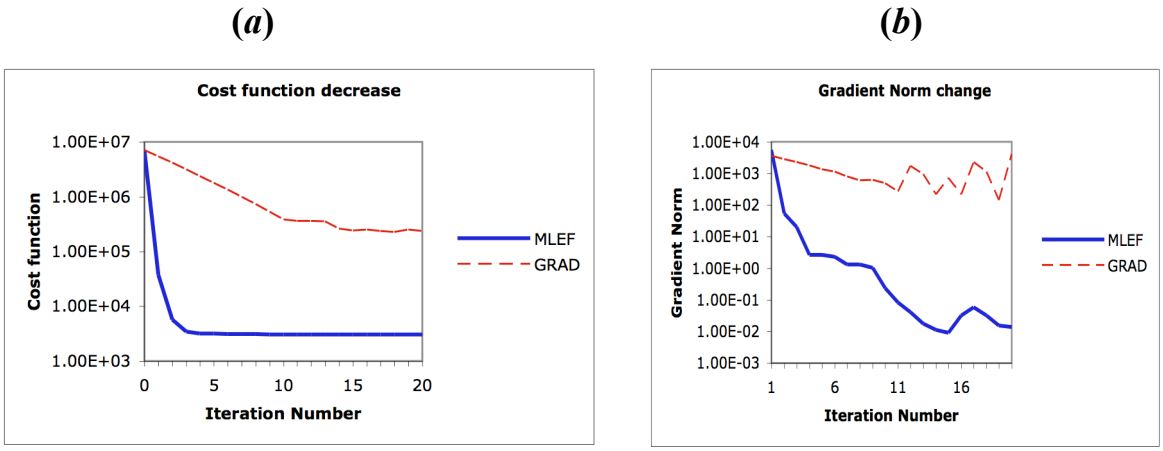

Fig.2. Minimization with the MLEF (full blue line) and with the Fletcher-Reeves nonlinear conjugate-gradient (dashed red line), in terms of the: (a) cost function, and (b) gradient norm. The results indicate much better performance of the MLEF, suggesting its applicability in nonlinear data assimilation.

2.3.b. Initial uncertainty

Specification of the initial uncertainty in EnsDA was addressed in the paper by Zupanski et al. (2006). 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, whilst the second is based on spatial smoothing of initially uncorrelated random perturbations. Data assimilation experiments are conducted using a global shallow-water model and synthetic 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 (Fig.3). The root-mean-square error rate of convergence of the data assimilation is improved, and the positive impact of initial correlations is noticeable throughout the data assimilation cycles. The sensitivity to the choice of the correlation length scale exists, although it is not very high. Given the insufficient information from observations and nonlinear dynamics, a common scenario in realistic geosciences applications, data assimilation can benefit from better-defined initial uncertainties.

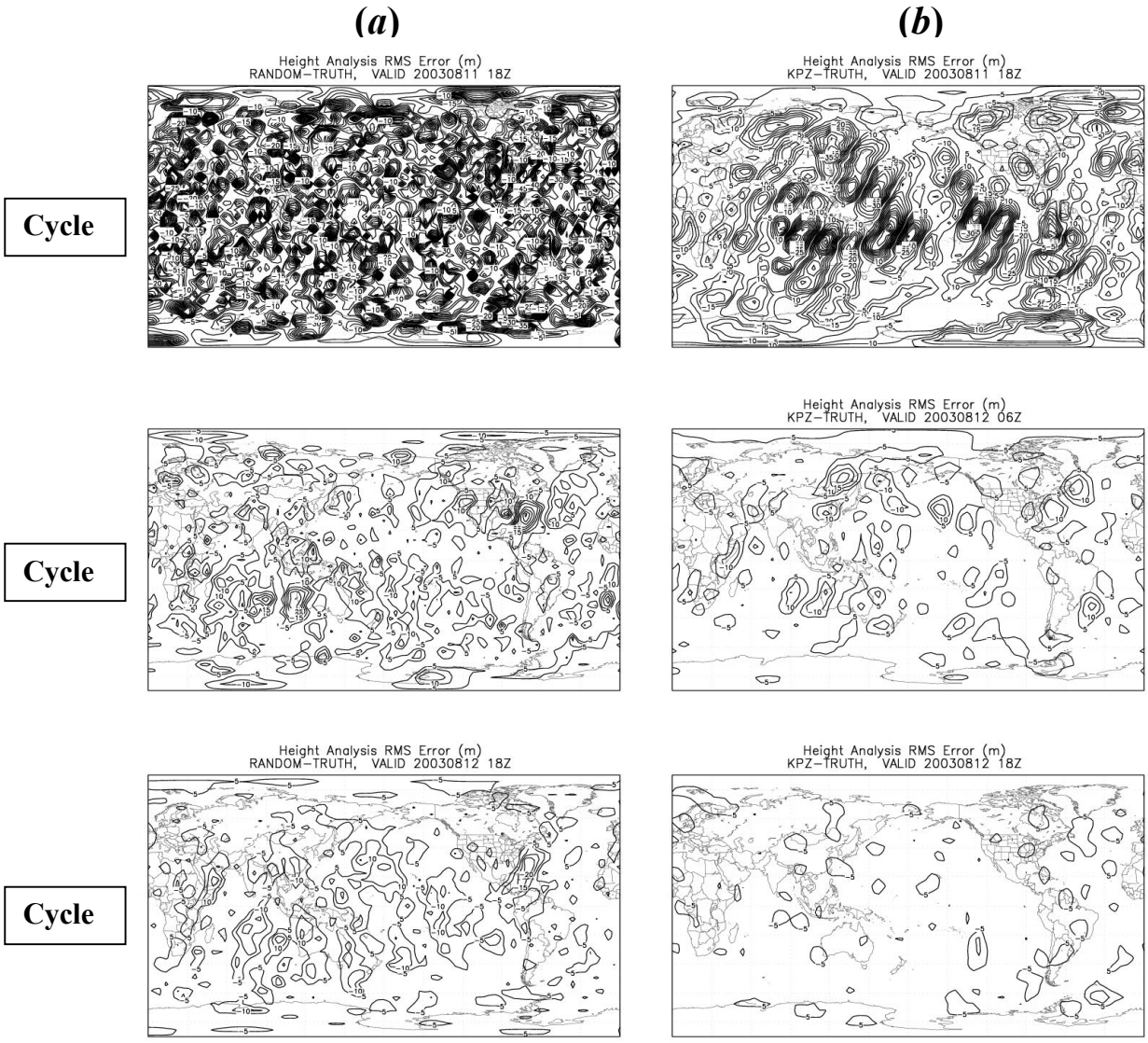

Fig.3. Height analysis error (m) in the analysis cycles 1, 3, and 5, obtained using: (*a*) uncorrelated-random, and (*b*) correlated-Kardar-Parisi-Zhang ensemble initiation methodology. The contour interval is 5 m. First thing to notice is that, no matter which ensemble initiation methodology is used, the MLEF efficiently reduces the noise level of the analysis error. Related to the choice of ensemble initiation methodology, it is easy to see that assigning correlations to initial perturbations improves results by creating smaller analysis error with less spatial noise.

2.3.c. Hurricane Katrina with Weather Research and Forecasting (WRF) model

The Maximum Likelihood Ensemble Filter (MLEF) is used with the Weather Research and Forecasting model to assimilate real observations in the hurricane Katrina case [Zupanski et al. (2008b)]. In our research we focused on the issue of insufficient number of degrees of freedom in high-dimensional realistic applications of ensemble data assimilation, such as of hurricane Katrina development and landfall. A modified local domains approach is developed and tested with the MLEF algorithm. The modification addresses the calculation of the observation information matrix by creating a correlation between local domains. This approach avoids the need to interpolate/smooth local analyses directly, and thus has a potential to produce dynamically more consistent local analyses. The results with the local MLEF show that error covariance localization has a clear positive impact (Fig.4). Even with 96 ensembles, the global (i.e. non-localized) MLEF cannot achieve the performance of the local MLEF with 32 ensembles.

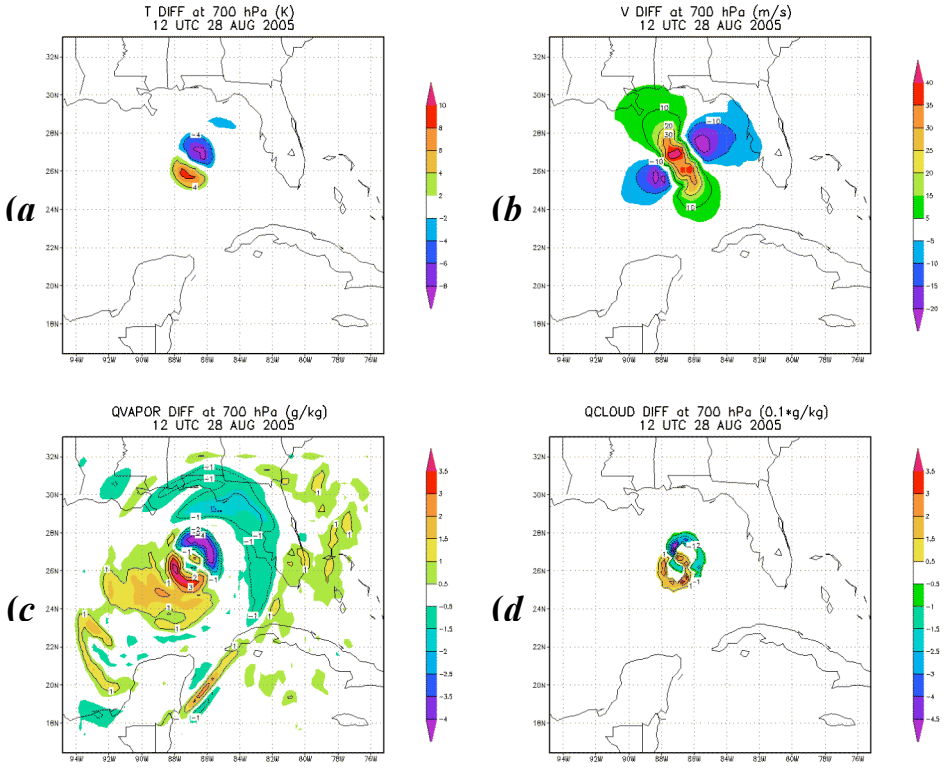

Fig.4. Differences between the 6-hour forecast from the LMLEF and the NOOBS forecast at 700 hPa, valid at 1200 UTC on August 28, for: (a) temperature (K), (b) north-south wind (ms^{-1}), (c) specific humidity (gkg^{-1}), and (d) cloud water ($10^{-1}\,gkg^{-1}$). These results indicate the positive impact of error covariance localization within the MLEF. The results also show that, even with only 32 ensembles, the local MLEF is able to focus on the hurricane producing only dynamically defined local impact of data assimilation, without a noise in the forecast. It is also interesting to note that dynamics/physics of the hurricane defines a naturally larger correlations for specific humidity (panel (c)), while it maintains shorter correlations for other variables.

2.3.d. Sensitivity of EnsDA to ensemble size in shallow-water dynamics

We examined how the ensemble size impacts EnsDA performance, in several studies using the MLEF with the CSU's 2-dimensional shallow water equations model on the sphere. Our results have implications for complex, high-dimensional model applications, indicating that for a well-developed dynamical system the required number of ensembles may not be that large (i.e. much smaller than state dimension). We also develop a new general methodology for controlling the ensemble size, which could also be used to monitor the quality of EnsDA.

In the paper by Fletcher and Zupanski (2008b) we look for the optimal number of ensemble members, with respect to the root mean square (rms) error of the three state variables in the shallow water equations model. We vary the size of the observations sets for three different Rossby-Haurwitz waves in the shallow water equations model. These waves generate different types of dynamics from fast, shallow motions to fast, tall waves with vortices, to a flow similar to geostrophic balance. We show that, for faster flows, we require less ensemble members than for slow balanced flows. We also present an explanation for this behavior in the form of hybrid Lyapunov-bred vectors. These results have implications for complex, high-dimensional applications, indicating that for a well-developed dynamical system the required number of ensembles may not be that large (i.e. much smaller than state dimension).

In another study [Uzunoglu et al. (2007)] we address the question of whether it is possible to consistently reduce the number of ensemble members at a later stage in the assimilation cycles. An extension to this is given this reduction is it possible to reintroduce the ensemble members at a later time if the order of accuracy is decreasing, is also considered. To address these questions we present an adaptive methodology for reducing and inflating ensemble size by projecting the ensemble on to a limited number of its leading Empirical Orthogonal Functions (EOFs) through a Proper Orthogonal Decomposition (POD). An adaptive methodology for reducing and inflating ensemble size was successfully applied for two different test cases with the shallow water equations model that typically resulted in a reduction of up to a factor of half in the number of ensemble members required for successful implementation.

2.4. Additional accomplishments

These goals were not originally set, but were nevertheless achieved under this project.

2.4.a. Non-Differentiable Minimization

Based on promising results obtained by the MLEF for nonlinear operators, we extended our research to non-differentiable minimization. We were able to define two non-differentiable minimization algorithms as extensions of the standard gradient-based minimization algorithms [Zupanski et al. (2008a)]: the generalized conjugate-gradient and the generalized quasi-Newton non-differentiable minimization algorithms.

Preliminary examination indicates a superior performance of non-differentiable minimization compared to gradient-based minimization methods. Given the difficulties of currently available non-differentiable algorithms in dealing with high-dimensional systems found in geosciences, our ensemble based non-differentiable minimization can have an advantage due to its high efficiency.

2.4.b. Error covariance localization

In an effort to improve the MLEF performance in high-dimensional, realistic geosciences applications, we developed an algorithm to increase the degrees of freedom (DOF) in the analysis stage of the filter [Zupanski et al. (2008b)]. This was achieved by an error covariance localization algorithm, which limits the distant correlations, yet still controls the amount of noise typically introduced by the localization procedure. The introduced error covariance localization is a modification of the local-domains approach of Ott et al. (2004), such that the interface between local analyses is smooth and it does not produce spurious noise. At the same time, the computational efficiency of the original method is conserved. The approach taken was to apply the modification to the block observation weight matrix, rather than to the analysis directly. As an illustration, for $K+1 \le i \le L-K$ the i-th diagonal block can be written as

$$\boldsymbol{W}_{new}^{i} = \frac{\gamma_K^2 \boldsymbol{W}^{i-K} + \cdots + \gamma_1^2 \boldsymbol{W}^{i-1} + \gamma_o^2 \boldsymbol{W}^{i} + \gamma_1^2 \boldsymbol{W}^{i+1} + \cdots + \gamma_K^2 \boldsymbol{W}^{i+K}}{\gamma_0^2 + 2\gamma_1^2 + 2\gamma_2^2 \cdots + 2\gamma_K^2}$$

In the above formula L is the number of local domains, and K is the number of neighboring domains used in modification. This new blocks include the correlation between the original local domains, extending to the K-th neighbor. The calculation of new block diagonal matrices can be interpreted as an interpolation from neighboring local domains. The use of the MLEF with the atmospheric WRF model in simulation of hurricane Katrina indicates the relevance of error covariance localization in high-dimensional geosciences applications. The developed error covariance localization has been successfully tested in several other applications, and is now an optional component of the MLEF algorithm.

*2.4.c. Gaussian resampling for particle filte*r

We proposed in this work [Xiong et al. (2006)] an *a posteriori* Gaussian resampling (GR) method for the particle filter (PFGR) that aims to increase the stability of the particle filter PF and maintain the ensemble spread, while allowing for a potential generalization to higher-dimensional models. In our work we presented simulation results of a numerical test of the method using the Lorenz model comparing PFGR and EnKF. This method was adopted in recent work of Salman (2008a, 2008b) for Lagrangian data assimilation. Data assimilation results are shown in Fig.5. The PFGR yields satisfactory results when tested in the framework of a low dimension Lorenz model. The most computationally expensive part involves the singular decomposition of a matrix with dimensions of the ensemble sample size. The estimated prior and posterior PDF obtained

by the kernel density technique are shown in Fig.6, indicating a similar posterior PDF of the EnKF and the PFGR.

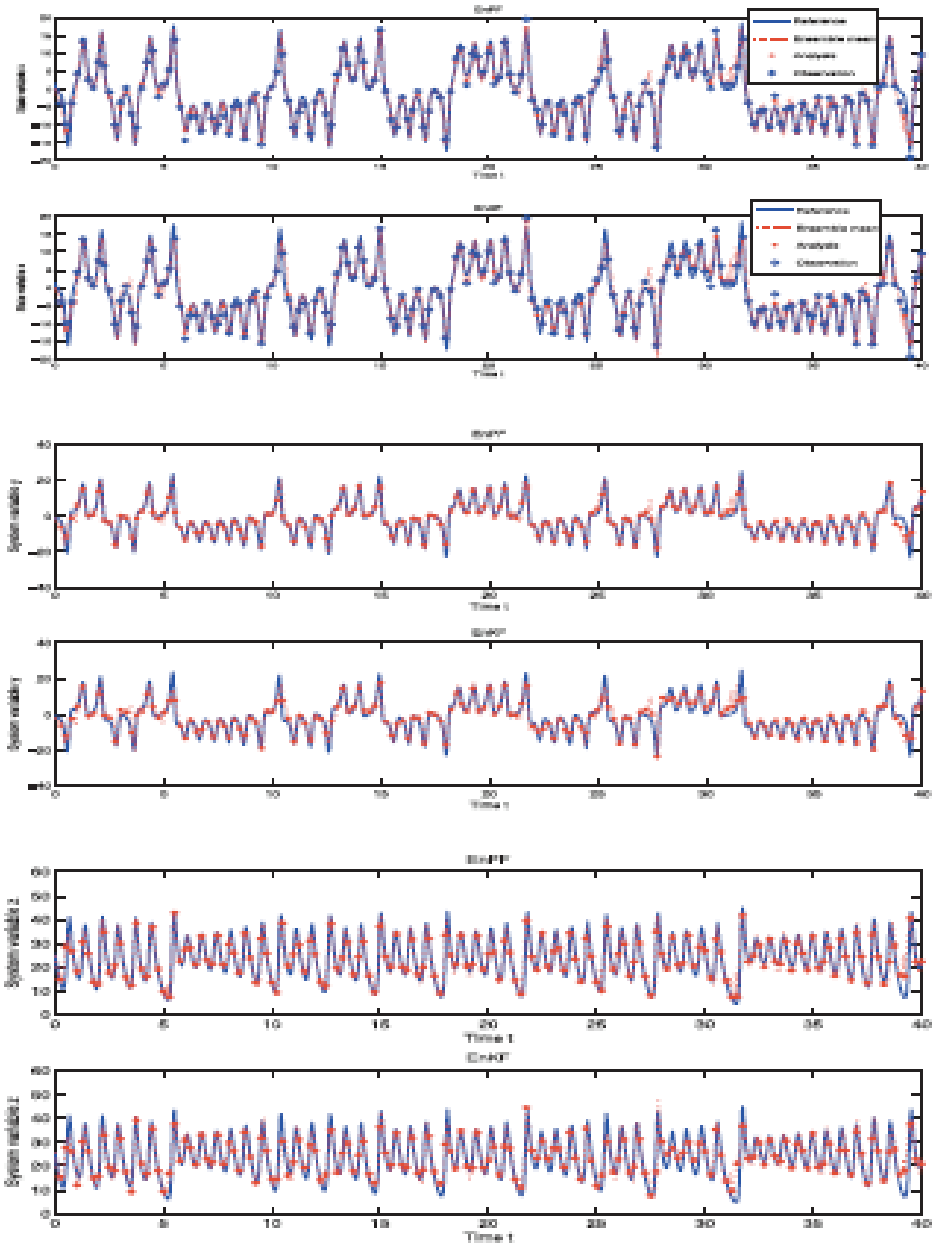

Fig.5. Results of data assimilation experiments with EnKF and PFGR. System variables x, y, z, reference solution, observations and ensemble mean prediction. 160 observations within 40 s run time. Zero model error variance. Measurements on x only with observation variance 2.0. Ensemble size 1000. The performance of two filters is comparable, producing similar number of spikes (mispredictions).

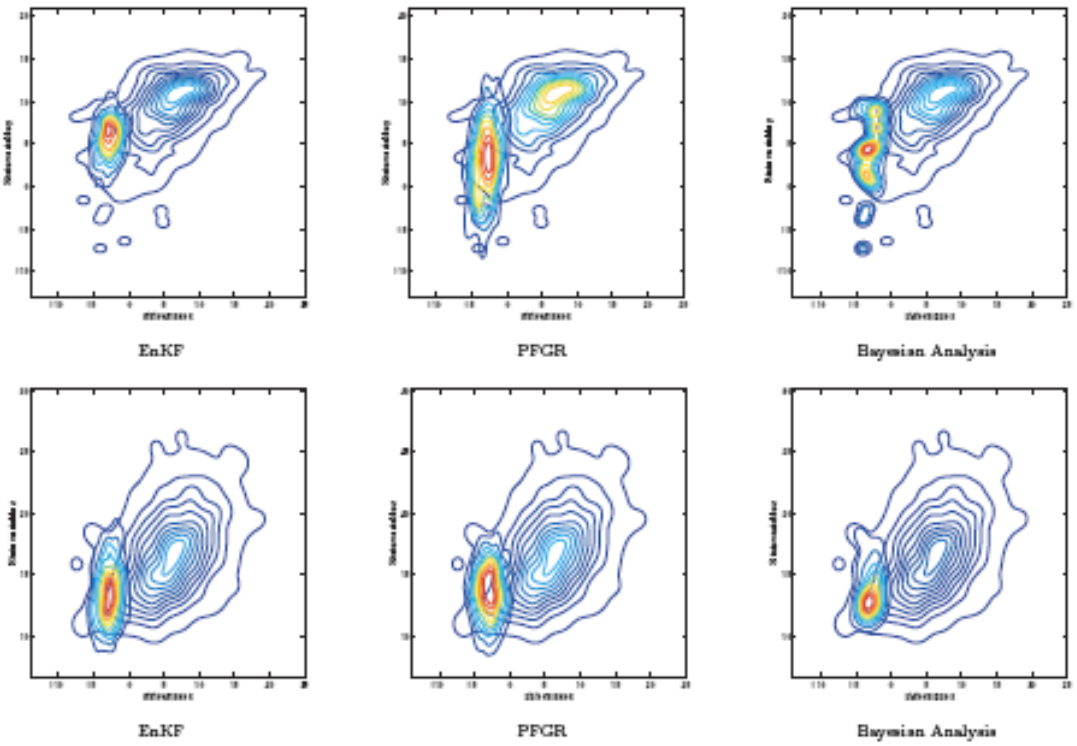

Fig.6. Kernel density estimate of the prior and posterior probability density function integrated over z direction (upper row) and y direction (lower row). Observation value x=-3.884. EnKF, PFGR and Bayesian Analysis. The prior and posterior ensemble data obtained from the same runs as of Fig.5 at t=34.5 s. The probability profiles of the EnKF and the PFGR posterior ensemble show similarity.

2.4.d. Comparison of sequential data assimilation methods for the Kuramoto-Sivashinsky equation

The Kuramoto-Sivashinsky equation plays an important role as a low-dimensional prototype for complicated fluid dynamics systems having been studied due to its chaotic pattern forming behavior. Up to now, efforts to carry out data assimilation with this 1-d model were restricted to variational adjoint methods domain and only Chorin and Krause (2004) tested it using a sequential Bayesian filter approach. In this work we compare three sequential data assimilation methods namely the Kalman filter (EnKF) with covariance localization and inflation approach, the sequential Monte-Carlo particle filter approach (PF) and the Maximum Likelihood Ensemble Filter methods (MLEF). This comparison is to the best of our knowledge novel.

We compare in detail the relative performance of above filters for both linear and nonlinear observation operators [Jardak et al. (2008a)]. The results of these sequential data assimilation tests are discussed and conclusions are drawn as to the suitability of these data assimilation methods in the presence of linear and nonlinear observation

operators. The results with the MLEF are shown in Fig.7 for linear and in Fig.8 for nonlinear observation operators.

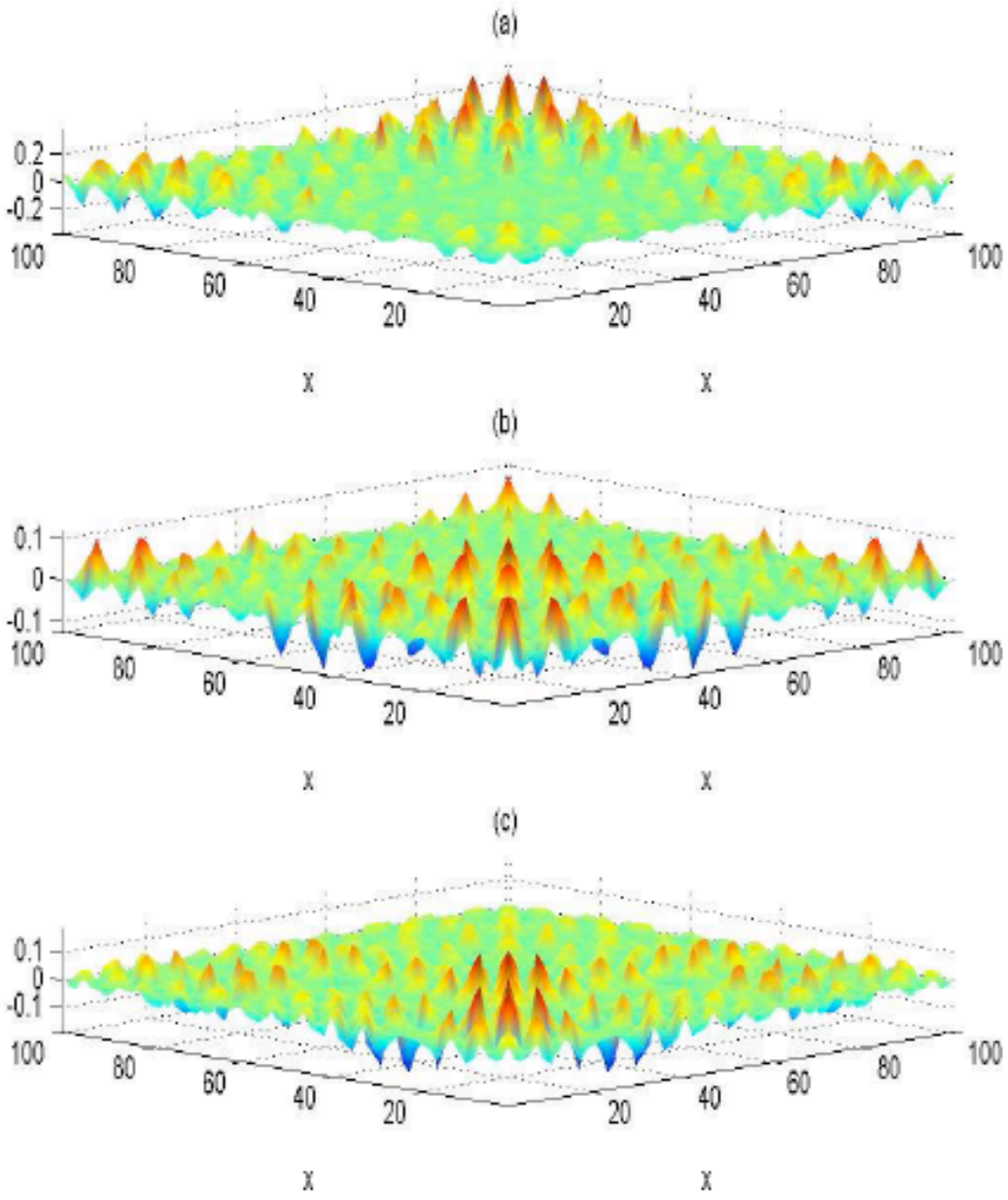

Fig.7. Analysis error covariance using 100 ensemble members. (a) after t = 20 time units , (b) after t = 100 time units and (c) after t = 250 time units, linear observation operator case, MLEF method

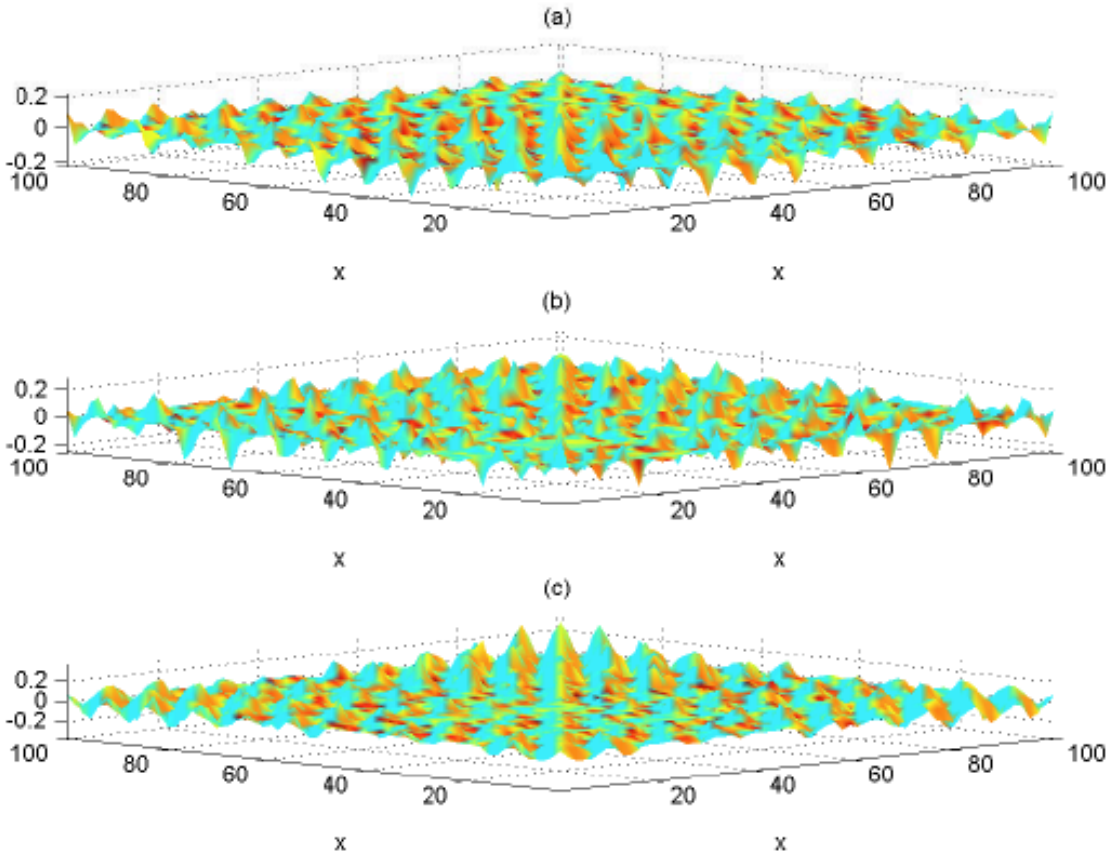

Fig.8. Analysis error covariance using 100 ensemble members. (a) after t = 20 time units , (b) after t = 100 time units and (c) after t = 250 time units, nonlinear observation operator case, MLEF method.

2.4.e. Ensemble Data Assimilation methods for the shallow water equations with linear and nonlinear observation operators

Many problems in the geosciences require estimation of the state of a system that changes over time using a sequence of noisy measurements made on the system. Data assimilation is the process of fusing observational data and model predictions to obtain an optimal representation of the state of the atmosphere.

A new comparison of three frequently used sequential data assimilation methods illuminating their strengths and weaknesses in the presence of linear and nonlinear observation operators is presented [Jardak et al. (2008b)]. The ensemble Kalman filter (EnKF), the particle filter (PF) and the Maximum Likelihood Ensemble Filter (MLEF) methods were implemented in a spectral shallow water equations model in spherical geometry using the Rossby-Haurwitz wave no 4 as initial conditions. Different error metrics illustrate their relative performances and serve as indicator for real life applications. Results for the EnKF are illustrated in Fig.9 (geopotential) and Fig.10 (velocity).

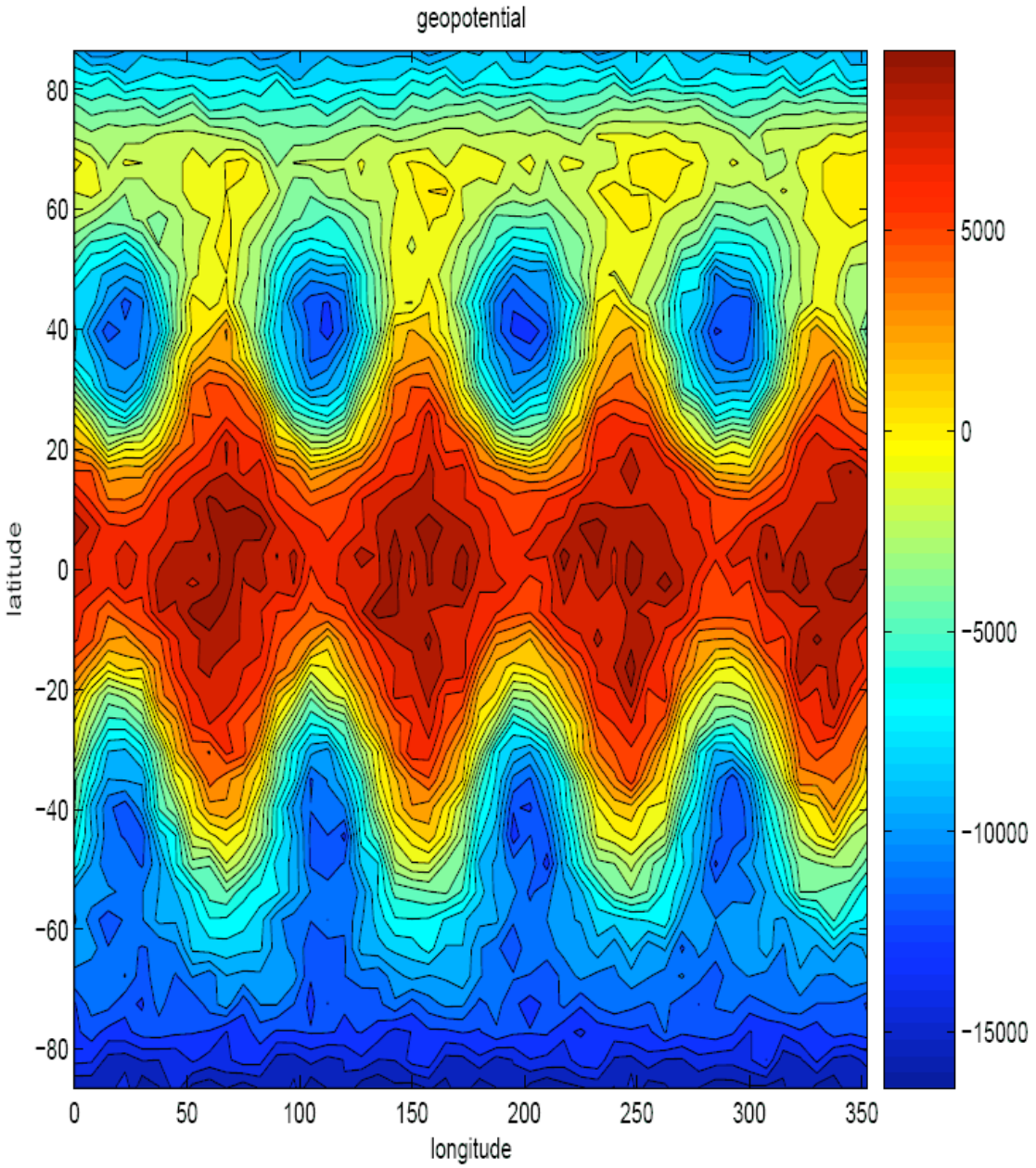

Fig.9. Geopotential field. Linear observation operator after 12 days of integration, 216 observations, 300 ensembles, 30 cycles , 2% perturbation. Rossby Haurwitz test case.(EnKF)

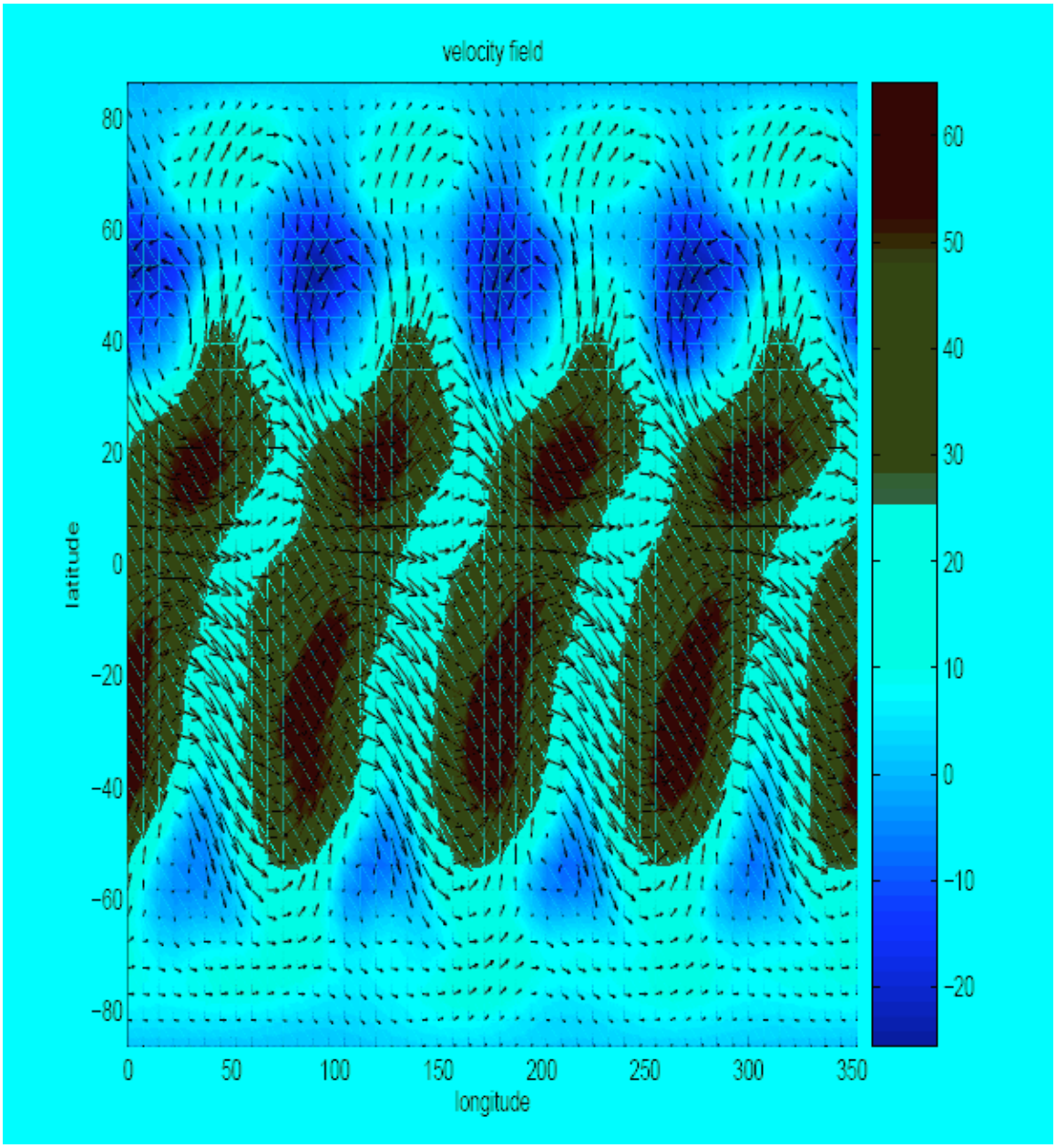

Fig.10. Velocity field. Linear observation operator after 12 days of integration, 216 observations, 300 ensembles, 30 cycles , 2% perturbation. Rossby Haurwitz test case.(EnKF)

3. Education and Training

Two postdoctoral researchers sponsored under this project (Dr. Bahri Uzunoglu at FSU and Dr. Steven Fletcher at CSU) were exposed to the cutting-edge scientific research, resulted in a successful education and training. Both postdoctoral scientists conducted a novel research that led to new directions of their personal scientific development, and

they both published papers with this work. In addition, the MLEF algorithm has been used as an educational tool and a class project at several universities in the US, Europe, and Asia.

Dr. Bahri Uzunoglu developed new methodology for controlling ensemble size by relying on POD theory [Uzunoglu et al. (2007)]. His results have been steadily gaining international recognition, as more scientists are beginning to use his ideas, not only in EnsDA but also in designing reduced-order variational methods. Dr. Steven Fletcher focused on non-Gaussian PDF assumptions in D/A, resulting in several papers describing a lognormal PDF framework for D/A [Fletcher and Zupanski (2006a,b; 2008a)]. His non-Gaussian work has been internationally recognized, impacting not only EnsDA but also variational D/A research.

Dr. X. Xiong was also trained for one year under this project under the second PI at FSU. He contributed by investigating the particle filters in data assimilation [Xiong et al. (2006)].

Dr Mohamed Jardak was trained by both PI's for the last year, and will now relocate to work with Prof Adrian Sandu of Virginia Tech on projects related to data assimilation. His work resulted in a comparison of MLEF with both EnKF and various versions of particle filters in the presence of both linear and nonlinear observation operators for both the Kuramoto-Sivashinsky model [Jardak et al. (2008a)] as well as for a spectral model of the global shallow water equations [Jardak et al. (2008b)].

Over the 4-year course of this project the results were regularly presented at international and cross-disciplinary conferences/workshops, by postdoctoral researchers and by PIs. Overall, there were 14 conferences/workshops that included presentations based on the results from this research. We regularly acknowledge the NSF support in those presentations.

4. Dissemination: making new EnsDA methodology available

One of the goals of this research project was to provide the developed EnsDA algorithm to a broad spectrum of users, especially for educational purposes. This was accomplished through personal contacts at Conferences/Workshops and through web pages at Colorado State University (http://www.cira.colostate.edu/projects/ensemble/NSF_CMG.php) and Florida State University (https://people.scs.fsu.edu/~navon/). The MLEF algorithm documentation is included on the Colorado State University web page.

The MLEF algorithm has been used as an educational tool and a class project at several universities and other educational institutions in the US (Florida State University, Colorado State University), Europe (International Centre for Theoretical Physics, Trieste, Italy; University of Belgrade, Belgrade, Serbia), and Asia (Ewha Womans University, Seoul, South Korea).

Results of the research conducted under this NSF project have been published in best international journals, totaling 13 papers. The publications are listed in the reference list is given below.

Participation at inter-disciplinary conferences/workshops was also one of the paths of dissemination of the research results. The postdoctoral researchers and the PIs presented the work related to this NSF research project at many scientific meetings throughout the world, including the meetings of *European Geophysical Union* (EGU), *American Geophysical Union* (AGU), *World Meteorological Organization* (WMO), *International Union of Geodesy and Geophysics* (IUGG), *American Meteorological Society* (AMS), and *Society for Industrial and Applied Mathematics* (SIAM).

In March 2006 we organized the Workshop on Predictability, Observations and Uncertainties in Geosciences [Zupanski and Navon (2007)], held at FSU and co-sponsored by the NSF CMG and the School of Computational Science at FSU. (13-15 March 2006) that also contributed in disseminating our research [Zupanski and Navon (2007)]. The idea of the workshop was to learn about and to discuss common mathematical concepts behind uncertainty estimation in a broad application spectrum, thus including over 50 participants from hydrology, weather, climate, geology, nuclear sciences, and mathematics, and among them the late Prof. Ed Lorenz of Massachusetts Institute of Technology, the "father" of chaos theory. The workshop promoted including students and young researchers from mathematics and engineering to work on geosciences problems.

5. Broader impact

This research has achieved international recognition, and has been considered as a method of choice in many areas of geosciences. This includes meteorological applications of the MLEF to clouds and precipitation (NASA Global Precipitation Mission), climate applications in Multiscale Modeling Framework (Center for Multiscale Modeling of Atmospheric Processes –CMMAP), hydrology applications to river-flow D/A (NOAA Office of Hydrology Development), as well as to carbon D/A.

One of the most relevant impacts of this research has been to the carbon transport problem [Zupanski et al. (2007); Lokupitiya et al. (2008)]. Carbon transport is relevant for prediction of climate change, as well as for identifying regional and global sources and sinks. The MLEF was especially impressive in predicting uncertainties of sources and sinks, as well as being flexible to include parameter and model bias estimation.

There is another indication of the broader impact of this research, measured by the number of visits to our web sites. To date, there are more than 36,000 visits!

6. References

***Peer-reviewed publications supported by this NSF project*:**

Fletcher, S.J., and M. Zupanski, 2008b: A study of ensemble size and shallow water dynamics with the Maximum Likelihood Ensemble Filter. *Tellus*, **60A**, 348-360.

Fletcher, S. J., and M. Zupanski, 2008a: Implications and Impacts of Transforming Lognormal Variables into Normal Variables in VAR. *Met. Zeit.*, **16**, 755-765.

Fletcher, S. J., and M. Zupanski, 2006b: A Hybrid Normal and Lognormal Distribution for Data Assimilation. *Atmos. Sci. Lett.,* **7,** 43-46.

Fletcher, S. J., and M. Zupanski, 2006a: A Data Assimilation Method for Log-normally Distributed Observational Errors. *Q. J. Roy. Meteorol. Soc.*, **132**, 2505-2519.

Jardak, M., I. M. Navon and M. Zupanski, 2008a: Comparison of sequential data assimilation methods for the Kuramoto-Sivashinsky equation. *International Journal for Numerical Methods in Fluids., (submitted)*.

Jardak, M., I. M. Navon and M. Zupanski, 2008b: Ensemble Data Assimilation for the shallow water equations. (To *be submitted to JGR Atmospheres)*.

Uzunoglu, B., S.J. Fletcher, I. M. Navon, and M. Zupanski, 2007: Adaptive Ensemble Size Reduction and Inflation. *Q. J. R. Meteorol. Soc.,* **133**, 1281-1294.

Xiong, X., I.M. Navon and B. Uzunoglu, 2006: A Note on the Particle Filter with Posterior Gaussian Resampling. *Tellus A,* **58A,** 456-460.

Zupanski, M., D. Zupanski, S. J. Fletcher, M. DeMaria, and B. Dumais, 2008b: Ensemble data assimilation with the Weather Research and Forecasting model: The hurricane Katrina case. *J. Geophys. Res*., (*submitted*).

Zupanski, M., 2008: Theoretical and Practical Issues of Ensemble Data Assimilation in Weather and Climate. Chapter in the book titled "*Data Assimilation for Atmospheric, Oceanic, and Hydrologic Applications"*, S. K. Park, Editor, Springer, (*in print*).

Zupanski, M., I. M. Navon, and D. Zupanski, 2008a: The Maximum Likelihood Ensemble Filter as a non-differentiable minimization algorithm. *Q. J. R. Meteorol. Soc*., **134**, 1039-1050.

Zupanski, M., and I.M. Navon, 2007: Predictability, Observations, and Uncertainties in Geosciences. *Bull. Amer. Meteor. Soc.*, **88**, 1431-1433.

Zupanski, M., S.J. Fletcher, I.M. Navon, B. Uzunoglu, R.P. Heikes, D.A. Randall, T.D. Ringler, and D. Daescu, 2006: Initiation of Ensemble Data Assimilation. *Tellus*, **58A**, 159-170.

***Peer-reviewed publications resulting from this NSF project* (*broader impact*):**

Carrio, G.G., W.R. Cotton, D. Zupanski, and M. Zupanski, 2008: Development of an aerosol retrieval Method: Description and preliminary tests. *J. Appl. Meteor. Climate*., doi: 10.1175/2008JAMC1729.1.

Hyunhee K., S. K. Park, D. Zupanski, and M. Zupanski, 2008: Application of the Maximum Likelihood Ensemble Filter to Data Assimilation for the Typhoon Megi. *Mon. Wea. Rev.*, (*submitted*).

Lokupitiya, R.S., D. Zupanski, A.S. Denning, S.R. Kawa, K.R. Gurney, M. Zupanski, W. Peters, 2008: Estimation of global CO_2 fluxes at regional scale using the maximum likelihood ensemble filter. *J. Geophys. Res.*, **113**, D20110, doi:10.1029/2007JD009679.

Zupanski, D., A.Y. Hou, S.Q. Zhang, M. Zupanski, C.D. Kummerow, and S. H. Cheung, 2007: Application of information theory in ensemble data assimilation. *Q. J. R. Meteorol. Soc.*, **133**, 1533-1545.

Zupanski, D., A. S. Denning, M. Uliasz, M. Zupanski, A. E. Schuh, P. J. Rayner, W. Peters and K. D. Corbin, 2007: Carbon flux bias estimation employing Maximum Likelihood Ensemble Filter (MLEF). *J. Geophys. Res.*, **112**, D17107, doi:10.1029/2006JK008371.

Other references

Chorin, A.J., and P. Krause, 2004: Dimensional reduction for a Bayesian filter. *PNAS,* **101**, 15013–15017.

Ott, E., B. R. Hunt, I. Szunyogh, A. V. Zimin, E. J. Kostelich, M. Corazza, E. Kalnay, and D. J. Patil, 2004: A local ensemble Kalman filter for atmospheric data assimilation. *Tellus,* **56A,** 415–428.

Salman, H., 2008a: A hybrid grid/particle filter for Lagrangian data assimilation. I: Formulating the passive scalar approximation. *Q. J. R. Meteorol. Soc.*, **134**, 1539-1550.

Salman, H., 2008b: A hybrid grid/particle filter for Lagrangian data assimilation. II: Appliocation to a model vortex flow. *Q. J. R. Meteorol. Soc.*, **134**, 1551-1565.