Ensemble-based Chemical Data Assimilation III: Filter Localization

Emil M. Constantinescu*, Adrian Sandu*,
Tianfeng Chai†, and Gregory R. Carmichael†

* Department of Computer Science, Virginia Polytechnic Institute and State University, Blacksburg, VA 24061. E-mail: {emconsta, sandu}@cs.vt.edu

† Center for Global and Regional Environmental Research, The University of Iowa, Iowa City, IA 52240. E-mail: {tchai, gcarmich}@cgrer.uiowa.edu

Abstract

Data assimilation is the process of integrating observational data and model predictions to obtain an optimal representation of the state of the atmosphere. As more chemical observations in the troposphere are becoming available, chemical data assimilation is expected to play an essential role in air quality forecasting, similar to the role it has in numerical weather prediction. Considerable progress has been made recently in the development of variational tools for chemical data assimilation. In this paper we implement and assess the performance of a localized "perturbed observations" ensemble Kalman filter (LEnKF). We analyze different settings of the ensemble localization, and investigate the joint assimilation of the state, emissions and boundary conditions. Results with a real model and real observations show that LEnKF is a promising approach for chemical data assimilation. The results also point to several issues on which future research is necessary.

1 Introduction

Data assimilation is the process by which measurements and model predictions are combined to obtain an accurate representation of the state of the modeled system. Data assimilation is recognized as essential in weather, climate analysis and forecast activities. In the first and second part of this study [Constantinescu et al., 2006b,c], we analyzed the potential and benefits of "perturbed observations" ensemble Kalman filter atmospheric chemical data assimilation applied in both ideal and real settings.

Variational techniques for data assimilation based on 4D-Var [Courtier et al., 1994, Rabier et al., 2000] are well-established in numerical weather prediction. 4D-Var framework represents the current state-of-the-art in meteorological [Courtier et al., 1994, Rabier et al., 2000] and chemical [Liao et al., 2005, Sandu et al., 2003, 2005, Sandu and Daescu, 2005] data assimilation. Ensemble Kalman filter (EnKF) data assimilation [Evensen, 1994, 2003, Burgers et al., 1998] has recently attracted considerable interest in numerical weather prediction. The cost of applying the Kalman filter [Kalman, 1960] to "large" models becomes tractable in the ensemble Kalman filter approach by using a Monte Carlo approximation to propagate the covariance.

In the first part of this study [Constantinescu et al., 2006b], we analyzed the performance of EnKF applied to chemical and transport models (CTM) in an idealized setting. A reference solution was considered to be the "truth" and was used both to build an initial unbiased ensemble and to generate artificial observations. The results indicate that EnKF is able to recover the reference solution with very good accuracy and to improve the forecast. Moreover, assimilation of the emission rates and lateral boundary conditions together with the model state proved beneficial for both analysis and forecast.

In the second part of this study [Constantinescu et al., 2006c], we considered the performance of EnKF applied to the same CTM (as in the second part) but now in a real setting. Real observations were used in the analysis for assimilation and in forecast for validation. The real setting posed a set of challenges and difficulties: filter divergence, ensemble spread was limited by the model positivity constraints, and the assimilation of the emission rates together with the state did not prove as beneficial as in the idealized case. Additionally, a comparison with 4D-Var [Courtier et al., 1994, Rabier et al., 2000, Liao et al., 2005, Sandu et al., 2003, 2005, Chai et al., 2006, Sandu and Daescu, 2005] was carried out in order to assess EnKF's effectiveness against a well proven technique. The results indicated that 4D-Var and EnKF produced similar quality results. EnKF, through background covariance inflation, may produce better results than 4D-Var near the observation sites, albeit spurious remote corrections are produced by EnKF and amplified by inflating the ensemble.

In this part of our study (the second), we investigate the "localization" of "perturbed observations" EnKF in order to avoid the spurious corrections noticed in the second part, and obtain an EnKF performance close to the one noticed in the ideal case (first part). The main contributions of this work are: (1) a discussion and experiments of the localized EnKF (LEnKF) in an operational-like setting using real data in order to avoid the development of spurious corrections, and (2) the assimilation of model parameters together with the states in order to improve the analysis and forecast results.

Houtekamer and Mitchell [Houtekamer and Mitchell, 1998] investigated a way to avoid EnKF ensemble covariance sub-sampling errors that generate spurious corrections through the use of an influence cutoff radius that removes the effect of remote observations. However, their results were not significantly improved probably due to the sharpness of the cutoff radius. Later, in [Houtekamer and Mitchell, 2001], the authors investigated the use of a correlation function that smoothly and monotonically decreases with distance, limiting the impact of remote observations, and thus avoiding spurious corrections. They showed that this technique is both effective by improving the accuracy of the results, and efficient by making the filter more parallelizable. A similar implementation and conclusions were also presented in [Hamill and Whitaker, 2001].

Recently Ott et. al. have investigated localization for ensemble Kalman filters (LEKF) [Ott et al., 2002] by projecting the ensemble on a local low dimensional subspace and applying the filter for that specific area. An implementation can be found in [Szunyogh et al., 2005]. Their results show great efficiency gains for the LEKF applied to the Lorenz model by reducing the ensemble size while keeping same error level as in larger sized EnKF results. This method will not be addressed in this paper.

The paper is structured as follows. Section 2.1 briefly review the EnKF and LEnKF methods. Section 3 describes the analysis setting and the filter localization approach used in the numerical experiments. Section 4.1 addresses a comparison between EnKF and LEnKF data assimilation applied to our atmospheric CTM. The assimilation of model parameters together with the model states is addressed in 4.2. A validation of the data assimilation results is carried out in Section 5. Finally, conclusions and future research directions are discussed in Section 6.

2 Data Assimilation

In this section we briefly review the EnKF approach to data assimilation, give an introduction to localized ensemble filtering, and discuss some techniques to avoid filter divergence.

Consider a nonlinear model $c_i = \mathcal{M}_{t_0 \to t_i}(c_0)$ that advances the state from the initial time

t_0 to future times t_i ($i \geq 1$). The model simulates the evolution of a real system (e.g., the polluted atmosphere). The model state c_i at t_i ($i \geq 0$) is an approximation of "true" state of the system c_i^t at t_i (more exactly, c_i^t is the system state projected onto the model space space).

Observations y_i of the real system are available at times t_i, and are corrupted by measurement and representativeness errors ε_i (assumed Gaussian with mean zero and covariance $\mathbb{R}_i$)

$$y_i = \mathcal{H}_i \left(c_i^t\right) + \varepsilon_i \ .$$

Here, $\mathcal{H}_i$ is an operator that maps the system/model state to observations.

The data assimilation problem is to find an optimal estimate of the state using the information from both the model (c_i, $i \geq 0$) and observations (y_i, $i \geq 0$).

2.1 The Localized Ensemble Kalman Filter (LEnKF)

The (ensemble) Kalman filter estimates the true state c_i^t at t_i using the information from the current best estimate c_i^f (the "forecast" or the background state) and the observations y_i. The optimal estimate c_i^a (the "analysis" state) is obtained as a linear combination of the forecast and observations that minimize the variance of the analysis (P^a)

$$c_i^a \ = \ c_i^f + P_i^f \, H_i^T \left(H_i \, P_i^f \, H_i^T + \mathbb{R}_i\right)^{-1} \left(y_i - \mathcal{H}_i(c_i^f)\right) = c_i^f + K_i \, \left(y_i - \mathcal{H}_i(c_i^f)\right) \ . \quad (1)$$

The forecast covariance P^f is estimated from an ensemble of runs (which produces an ensemble of E model states $c_i^f(e)$, $e = 1, \cdots, E$). The analysis formula (1) is applied to each member to obtain an analyzed ensemble. The working of the filter can be described in a compact notation as follows. The model advances the solution from t_{i-1} to t_i, then the filter formula is used to incorporate the observations at t_i:

$$c_i^f(e) = \mathcal{M}\big(c_{i-1}^a(e)\big) \ , \quad c_i^a(e) = \ c_i^f(e) + K_i \left(y_i - \mathcal{H}_i\big(c_i^f(e)\big)\right) \ , \quad e = 1, \cdots, E \ . \quad (2)$$

The results presented in this paper are obtained with the practical EnKF implementation discussed by Evensen [Evensen, 2003].

In the second part of this study we noted that EnKF's increments produced strong corrections far from the observation sites. This behavior is present to some extent in any Kalman filter implementation that employs a small ensemble of Monte Carlo simulations in order to approximate the background covariance (P^f). In its initial formulation, ensemble Kalman filter may suffer from spurious correlations caused by sub-sampling errors in the apriori (background) covariance estimates, when insufficient sampling points are used. This allows for observations to incorrectly impact remote model states, and possibly throwing the

model off balance or allowing it to reach unrealistic states. The filter *localization* introduces a restriction on the correction magnitude that an increment can have based on its remoteness.

One way to impose localization in EnKF is to use a decorrelation function with local support [Gaspari and Cohn, 1999] that monotonically decreases with distance for each increment [Hamill and Whitaker, 2001, Houtekamer and Mitchell, 2001], and apply it to the background covariance (P^f), making the background covariance distance decorrelated.

Consider a distance matrix, D with positive elements ($D_{k,l} \geq 0$, $D_{k,k} = 0$, $\forall k, l$), and a correlation function operator, ρ that maps distances in correction weights. Each element k, l in D represents the physical distance between the location of k and the location of l. Applying ρ to D produces a matrix of correction weights specific to the distances among the elements of D with $0 \leq \rho(D_{k,l}) \leq 1$ and $\rho(0) = 1$.

Following [Houtekamer and Mitchell, 2001], the EnKF relation (1) becomes

$$c_i^a = c_i^f + \left[\rho(D) \circ P_i^f\right] H_i^T \left(\left(H_i \left[\rho(D) \circ P_i^f\right] H_i^T\right) + \mathbb{R}_i \right)^{-1} \left(y_i - \mathcal{H}_i(c_i^f)\right), \quad (3)$$

where the operator '$\circ$' denotes the Schur product that applies the correction weights, $\rho(D)$, element wise to the background covariance, P^f (i.e. $\rho(D) \circ P^f \rightarrow \rho(D_{k,l}) \cdot P^f_{k,l}$,$\forall k, l$).

If H is linear the modified Kalman gain, $K^{\{\rho\}}$, can equivalently be expressed as

$$K^{\{\rho\}} = P_i^f H_i^T \left(\rho(D^y) \circ \left(H_i P_i^f H_i^T\right) + \mathbb{R}_i\right)^{-1}, \quad (4)$$

where D^y represents the distance among the observation sites (i.e. the location of y_i's).

Consequently, EnKF relation (1) becomes

$$c_i^a = c_i^f + \rho(D) \circ K_i^{\{\rho\}} \left(y_i - \mathcal{H}_i(c_i^f)\right). \quad (5)$$

In practice, the decorrelation function is first applied in the observation space ($\rho(D^y) \circ \left(H P^f H^T\right)$). Here, D^y is calculated as the distance among the observation sites and then applied to each entry in the $H P^f H^T$. If ρ is monotonic, the net effect on $H P^f H^T$ is the decrease of off-diagonal elements (cross-correlations), while keeping the diagonal elements (auto-correlations) unchanged. Hereafter, the modified Kalman gain, $K^{\{\rho\}}$, is computed.

Second, the decorrelation function $\rho(D)$ is Schur applied (elementwise) to the resulting (localized) Kalman gain, $K^{\{\rho\}}$, in order to compute each analysis increment. Here, the distance matrix D^c contains the distance from each state variable to each observation site. In this way, the contribution of each observation generated correction is damped according to the distance between the state location and the corresponding observation site.

The filter localization used in this paper is based on the approach described above. Further information about LEnKF can be found in [Houtekamer and Mitchell, 2001].

2.2 Preventing Filter Divergence

The "textbook application" of EnKF [Evensen, 2003] may lead to filter divergence. Filter divergence [Houtekamer and Mitchell, 1998, Hamill, 2004] is caused by progressive underestimation of the model error covariance magnitude during the integration – the filter becomes "too confident" in the model and "ignores" the observations in the analysis process. In part two of this study [Constantinescu et al., 2006c], we experienced such a behavior with a chemical atmospheric model due to the stiff chemistry and insufficient ensemble size. In that case, EnKF showed a decreasing ability to correct the ensemble state toward the observations at the end of the assimilation window. The cure was to artificially increase the covariance of the ensemble (effectively accounting for model errors) and therefore decrease the filter's confidence in the model results. In the second part [Constantinescu et al., 2006c], we investigated three ways to "inflate" the ensemble covariance in order to prevent filter divergence: additive inflation, multiplicative inflation, and model-specific inflation.

More details on preventing the filter divergence can be found in the second part of this study [Constantinescu et al., 2006c].

3 Experiment Setting

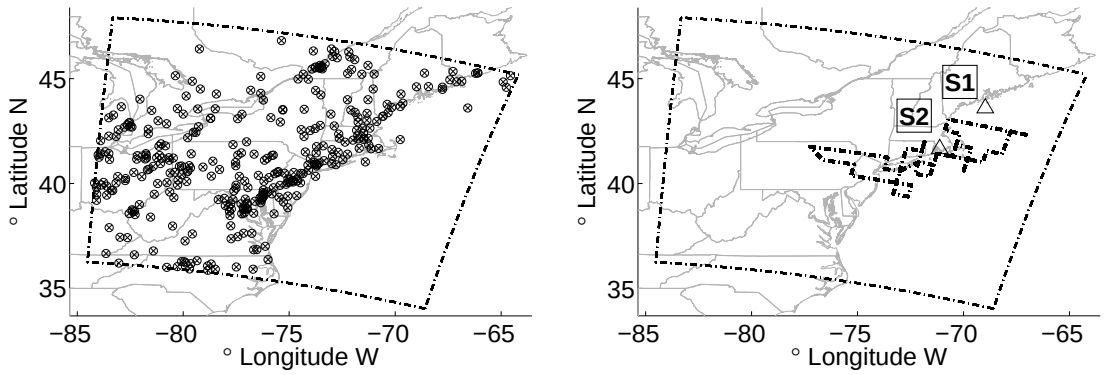
We now briefly review the framework of the analysis setting and present the approach taken to determine the decorrelation function for the filter localization. The numerical experiments presented in this paper use the same data assimilation setting as in the second part of this study [Constantinescu et al., 2006c].

Our data assimilation numerical experiments use the state-of-the-art atmospheric photochemistry and transport model STEM (Sulfur Transport Eulerian Model) [Carmichael et al., 2003] to solve the mass-balance equations for concentrations of trace species in order to determine the fate of pollutants in the atmosphere [Sandu et al., 2005], compactly written as

$$c_i = \mathcal{M}\left(c_{i-1}, u_{i-1}, c^{\text{in}}_{i-1}, Q_{i-1}\right) \ . \tag{6}$$

where c is the vector of concentrations (all species at all grid points), Q is the rate of surface emissions, u is the wind field, and c^{in} the Dirichlet boundary conditions. Subscripts denote time indices. The model also depends on other parameters (e.g., the turbulent diffusion, the air density) which are not explicitly represented here. The complete equations are described in the first part of our study [Constantinescu et al., 2006b] and in [Sandu et al., 2005].

The test case is a real-life simulation of air pollution in North–Eastern United States in July 2004 as shown in Figure 1 (the dash-dotted line delimits the computational domain).



(a) ICARTT ground stations (b) ozonesondes and P3 plane path

Figure 1: Ground measuring stations (a) in support of the ICARTT campaign (340 in total), and (b) two ozonesondes (S1, S2) and the flight path of a P3 plane that will be used for the numerical results/validation illustration.

The observations used for data assimilation are the ground-level ozone (O_3) measurements taken during the ICARTT (International Consortium for Atmospheric Research on Transport and Transformation) [ICARTT, Tang et al., 2006] campaign in summer 2004. Figure 1.a shows the location of the ground stations (340 in total) that measured ozone concentrations. The computational domain covers $1500 \times 1320 \times 20$ Km with a horizontal resolution of 60×60 Km and a variable vertical resolution (resulting in a 3-dimensional computational grid of $25 \times 22 \times 21$ points). The initial concentrations, meteorological fields, boundary values, and emission rates correspond to ICARTT conditions starting at 0 GMT of July 20^{th}, 2004. Our study also includes three validation measurements taken by two ozonesondes and a P3 plane (all shown in Figure 1.b).

All the simulations are started at the same time (0 GMT July 20^{th}) with a four hour initialization step (denoted as [-4,0] hours). This allows each of the ensemble members to reach quasi-steady-state before the assimilation window. The "best guess" of the state of the atmosphere at 0 GMT July 20^{th} is obtained from a longer simulation over the entire US performed in support of the ICARTT experiment [Tang et al., 2006]. This best guess is used to initialize the deterministic (non-assimilated) solution showed in the result sections.

The ensemble members are formed by adding a set of unbiased perturbations to the best guess at 0 GMT, and then evolving each member to 4 GMT July 20^{th}. The perturbation is constructed according to an autoregressive model, making it flow dependent. Additional information about its formation can be found in the first and in the second part of this paper [Constantinescu et al., 2006b,c]. Implementation considerations are given in [Constantinescu et al., 2006a].

The 24 hours assimilation window starts at 4 GMT July 20$^{\text{th}}$ and ends at 4 GMT July 21$^{\text{st}}$ (henceforth denoted as [0,24] hours). Observations are available at each integer hour in this window. The ozone O_3 observations used in this study are from the ICARTT ground stations (Figure 1).

EnKF adjusts the concentration fields of 66 "control" chemical species in each grid point of the domain every hour using (1). The ensemble size was chosen to be 50 members.

The 24 hours forecast window starts at 4 GMT July 21$^{\text{st}}$ and ends at 4 GMT July 22$^{\text{nd}}$ (the forecast windows will be denoted further as [24,48] hours). The model is initialized at 4 GMT July 22$^{\text{nd}}$ with the ensemble mean, and evolved in forecast mode for 24 hours.

The performance of each data assimilation experiment is measured by the R^2 correlation factor and root mean square (RMS) between the observation and the model solution (separate R^2 and RMS factors are computed in the assimilation and forecast windows).

A more detailed discussion can be found in part two of this paper [Constantinescu et al., 2006c].

3.1 Correlation Distance

The second part of this study [Constantinescu et al., 2006c] revealed that the small ensemble size leads to considerable sub-sampling errors in the estimated forecast covariances. As a consequence spurios long distance correlations lead to overcorrections of the state that deteriorate the analysis in areas remote from the observations. To alleviate these effects a decorrelation function ρ is used to restrict the magnitude of long-distance correlations. This "localization" process is based on the fact (verified experimentally) that the correlation between state variables decreases with the distance between their locations. The modified EnKF increment (obtained after localization) is calculated with (5).

In this study consider the decorrelation function ρ to be Gaussian:

$$\rho(D) \;=\; \exp\left[-\left(\frac{D}{L}\right)^2\right], \tag{7}$$

where D is the distance between two points and L is the decorrelation length (distance).

Since atmospheric flows are anisotropic (with different horizontal and vertical characteristic scales) we consider horizontal and vertical correlations separately. The decorrelation function becomes

$$\rho\left(D=\left[D^h D^v\right]\right) \;=\; \exp\left[-\left(\frac{D^h}{L^h}\right)^2-\left(\frac{D^v}{L^v}\right)^2\right], \tag{8}$$

where D^h, D^v are distances and L^h, L^v are the decorrelation lengths in the horizontal and vertical directions, respectively. In general the decorrelation distances change in space and

time, depending on the wind velocity, turbulent mixing, etc. In this study we only consider average decorrelation lengths in the horizontal and vertical directions. Better approximations are possible if time and position are considered for each filter application.

The NMC method [Parrish and Derber, 1992] is a popular technique used to estimate the background covariance in numerical weather prediction data assimilation. In NMC method, the differences between forecasts verifying at the same time are used to approximate the background error. The model error correlation coefficient c between two grid points (ℓ and k) is calculated as

$$c(l,k) = \frac{\langle \varepsilon_k \varepsilon_\ell \rangle}{\sqrt{\langle \varepsilon_k \varepsilon_k \rangle \cdot \langle \varepsilon_\ell \varepsilon_\ell \rangle}},$$

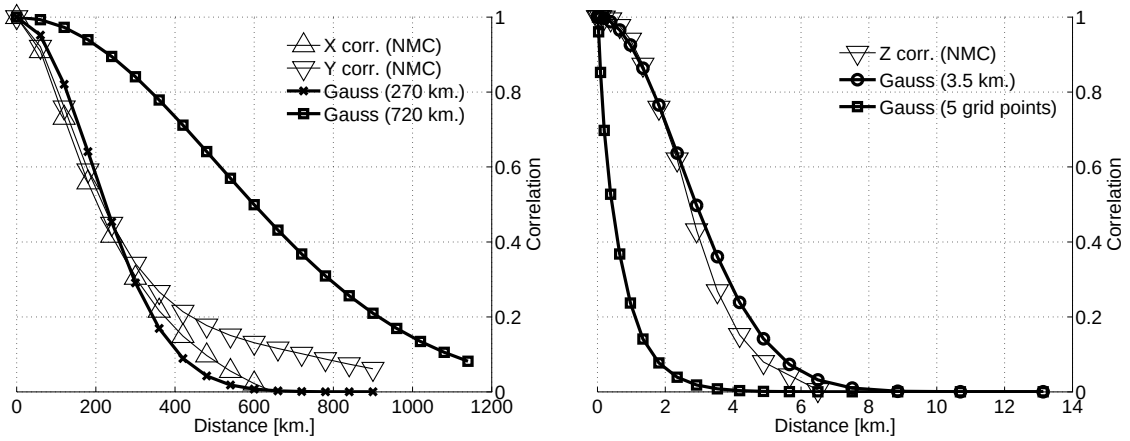
where '$\langle \cdot \rangle$' denotes the sample average and ε is the simulation deviation.

We consider three different forecasts initialized at one, two, and three days before the realization time. Deviation from the mean prediction is used as model "error" to construct the error covariance. Figure 2 shows the change of the correlation coefficients versus the horizotal and vertical distances. Also plotted are fitted Gaussian functions with horizontal and vertical correlation distances as $L^h = 270$ km and $L^v = 3.5$ km, respectively.

Additional experiments were performed using "intuitive" correlation distances of $L^h = 720$ km and $L^v = 5$ grid points. The results (not presented) showed that the NMC determined horizontal correlation improved the results, while the "intuitive" choice did not. However, the NMC averaged vertical correlation distance was too large to avoid spurious corrections at high altitudes (far from the observation sites), while the "intuitive" choice did avoid the vertical spurious corrections (as shown in the results section). This is likely due to the fact that the vertical grid has variable height, and the NMC height average resulted in a large correlation distance. For the rest of this study we consider the following correlation distances: $L^h = 270$ km and $L^v = 5$ grid points (shown in Figure 2.a,b together with the NMC estimated correlation distances).

4 Assimilation Results with Localized EnKF

In this section we discuss the results of data assimilation with the localized filter. First, we consider the state-only assimilation with different inflation techniques. Then, we show the results for joint state and parameter (emissions and lateral boundary conditions) assimilation.



(a) Horizontal correlation (b) Vertical correlation

Figure 2: Horizontal (a) and vertical (b) correlation-distance relationship. The correlation curves obtained through the NMC method (X,Y,Z corr.) are fitted with Gaussian functions with specific decorrelation distances (in parenthesis).

4.1 State-Only Data Assimilation

The localization of EnKF is needed in order to prevent the filter from making spurious corrections based on remote observations [Mitchell and Houtekamer, 1999, Houtekamer and Mitchell, 2001, Hamill and Whitaker, 2001, Hamill, 2004]. A data assimilation validation test done in the second part [Constantinescu et al., 2006c] with independent observations (not used in the analysis process) showed that EnKF without localization produces a good analysis near the location of the observations, but spurious corrections corrupted the solution far from them. Moreover, ensemble covariance inflation further improved the performance of the analysis near the observation sites, but also amplified the spurious corrections.

In this section we investigate the performance of localized EnKF and the effect of ensemble covariance inflation on the solution. The state only EnKF and LEnKF data assimilation R^2 and RMS results presented in this study are shown in Table 1. The EnKF scenarios were extensively discussed in the second part of our study [Constantinescu et al., 2006c]. LEnKF results are obtained using the same settings for the filter as in EnKF, but now with localization. Significant improvements are noticed for LEnKF data assimilation in the analysis than the assimilation with the non-localized EnKF. The R^2 values show a much better fit (in the analysis): EnKF's results are around 0.60 while LEnKF's score above 0.80. Moreover, the "textbook" LEnKF application has a better performance than any of the EnKF scenarios, pointing out to a possible large amount of spurious correlations developed even in the dense observation regions. The forecast performance shows some improvement for

LEnKF but not significant, indicating large errors in the model since the analyzed solution shows a good fit.

Figure 3 shows the ground level ozone field concentration at 14 EDT in the assimilation window (July 21st) measured by the ICARTT stations and predicted by the best guess, a 4D-Var assimilation for the same scenario (described in the second part of our study [Constantinescu et al., 2006c]), LEnKF "textbook" application, and LEnKF with the three inflation methods.

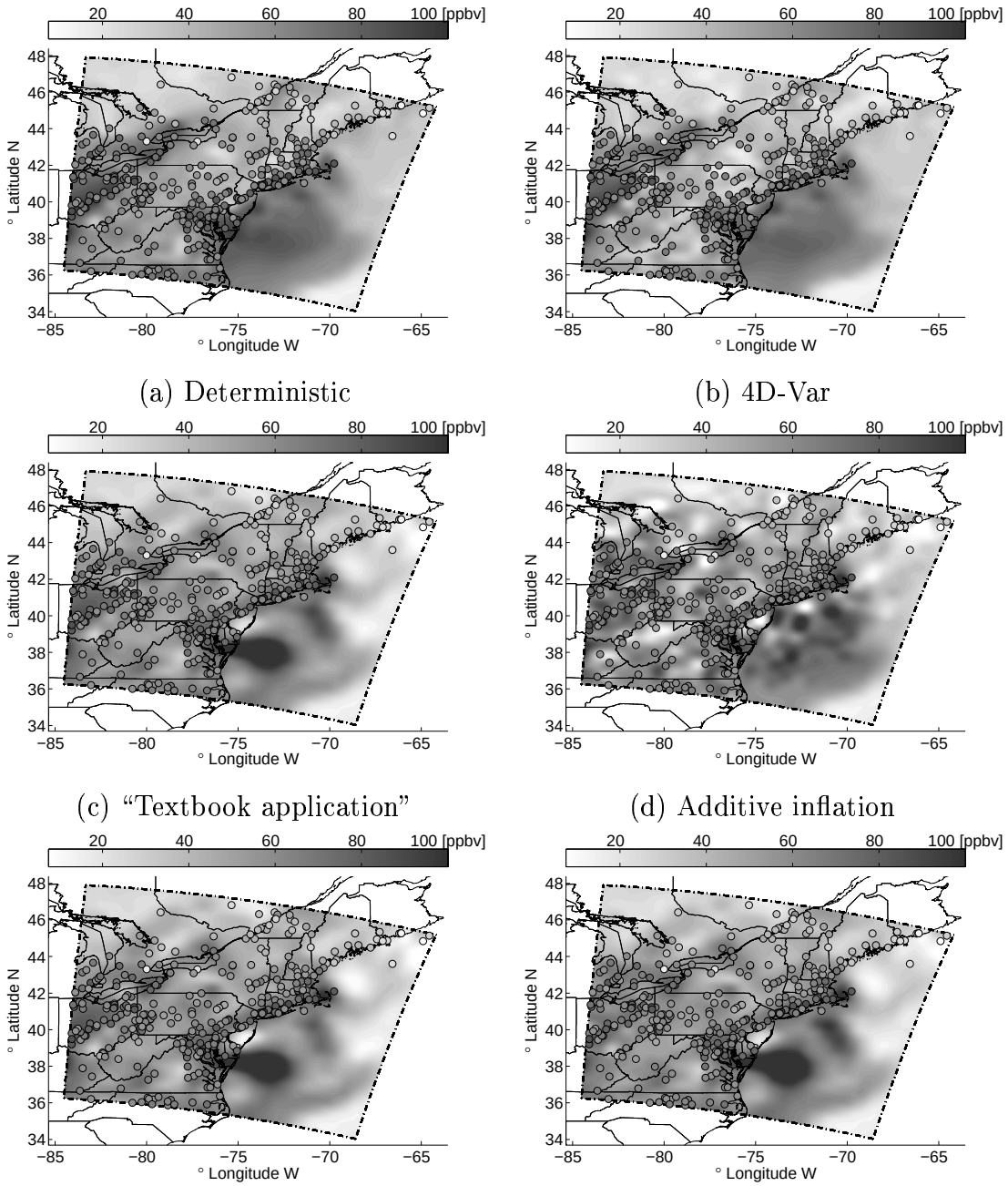
LEnKF textbook application does not seem to suffer from filter divergence as it displayed in the second part [Constantinescu et al., 2006c], possibly because the corrections are done at lower levels (close to the observation sites - decreasing with altitude) and the spread of the ensemble is maintained by the initial field variation coming away from the observation sites (a flow of "uncertainty"). This process can prevent the filter from diverging, but this effect should only be temporary.

LEnKF with additive inflation shows a very good performance in the analysis window. However, a qualitative inspection of the field (Figure 3.d, R^2=0.92) shows "punctual " corrections. This is expected and is in agreement with the theory: The filter is constrained to make "strong" corrections at the observation sites but it disregards the neighboring information due to the amplification of the diagonal background correlation factors. This effect can be quantitatively noticed in the forecast window (Table 1) where this approach gives the lowest R^2 value.

LEnKF with multiplicative (Figure 3.f, R^2=0.82) and model specific inflation (Figure 3.e, R^2=0.88) show similar qualitative (the field is smooth) and quantitative results. However, model-specific inflation is more grounded in our intuition by its relation to the main sources of uncertainty in CTMs which are treated explicitly.

4.2 Joint State and Parameter Data Assimilation

In regional chemistry and transport modeling, the influence of the initial conditions is rapidly diminishing with time, and the concentration fields are "driven" by emissions (denoted as EM) and by lateral boundary conditions (denoted as BC). Since both emissions and lateral boundaries are generally poorly known, it is of considerable interest to improve their values using information from observations through data assimilation. In this setting we have to solve a state-parameter assimilation problem [Derber, 1989, Annan et al., 2005, Evensen, 2005]. Our study of EnKF in an idealized setting [Constantinescu et al., 2006b] has revealed that the combined state-parameter assimilation has the potential to further improve both the analysis and the forecast. In the second part of the study [Constantinescu et al., 2006c],



(a) Deterministic

(b) 4D-Var

(c) "Textbook application"

(d) Additive inflation

(e) Model-specific inflation

(f) Multiplicative inflation

Figure 3: Ground level ozone field concentration at 14 EDT (July 21st) in the assimilation window measured by the ICARTT stations (shown in color coded filled circles) and predicted by: (a) the best guess, (b) 4D-Var, (c) and LEnKF "textbook" application, and LEnKF with the three inflation techniques (d) additive, (e) model-specific, and (f) multiplicative.

Method & Details	R^2 analysis (RMS)	R^2 forecast (RMS)
Deterministic solution, no assimilation	0.24(22.1)	0.28(23.5)
EnKF, "textbook application"	0.38(18.2)	0.30(23.2)
EnKF, model-specific inflation: 10% emissions, 10% boundaries, 3% wind	0.58(14.4)	0.32(22.6)
EnKF, multiplicative inflation: $\gamma_- \leq 4$, $\gamma_+ \leq 4$	0.62(14.2)	0.32(24.3)
EnKF, additive inflation: $\mathcal{N}(0, 6\text{ppb})$ white noise added *before* filtering if O_3 >5ppb	0.60(15.2)	0.30(23.2)
LEnKF, "textbook application"	0.81(9.79)	0.32(22.9)
LEnKF, model-specific inflation: 10% emissions, 10% boundaries, 3% wind	0.88(7.59)	0.32(22.5)
LEnKF, multiplicative inflation: $\gamma_- \leq 4$, $\gamma_+ \leq 4$	0.82(9.52)	0.32(22.8)
LEnKF, additive inflation: $\mathcal{N}(0, 6\text{ppb})$ white noise added *before* filtering if O_3 >5ppb	0.92(6.16)	0.31(22.7)

Table 1: The R^2 and RMS measures of model-observations match in the assimilation and forecast windows for EnKF and LEnKF (with different inflation "types") data assimilation.

where real observations were available, the assimilation of emissions showed no improvement of either the forecast or the analysis. This behavior is likely due to the small ensemble size spurious correlations which are developed between emissions and ozone concentrations. The filter tries to compensate the model-observations mismatch by over-adjusting the emission rates. In this section we discuss the same state-parameter assimilation setting for LEnKF.

In the numerical experiments we follow the approach discussed in the first and second part of this study [Constantinescu et al., 2006b,c]. The emission rates and lateral boundary conditions are multiplied by specific correction coefficients. These correction coefficients are appended to the model state (more exactly, to the vector of control variables). The EnKF data assimilation is then carried out with the extended model state. With the notation (6)

$$\begin{bmatrix} c_i^f \\ \alpha_i^{\mathrm{EM}} \\ \alpha_i^{\mathrm{BC}} \end{bmatrix} = \begin{bmatrix} \mathcal{M}_{t_{i-1}\to t_i}\left(c_{i-1}^a, u_{i-1}, \left(1+\alpha_{i-1}^{\mathrm{BC}}\right) c_{i-1}^{\mathrm{in}}, \left(1+\alpha_{i-1}^{\mathrm{EM}}\right) Q_{i-1}\right) \\ \alpha_{i-1}^{\mathrm{EM}} \\ \alpha_{i-1}^{\mathrm{BC}} \end{bmatrix}.$$

For the parameters α, we consider a different correction parameter for each species and ground-level grid point. In practice one may consider a coarser resolution, e.g., one correction

Method & Details	R^2 analysis(RMS)		R^2 forecast (RMS)			
	ST	ST+ EM+ BC	ST	ST+ EM	ST+ BC	ST+ EM+ BC
LEnKF, "textbook application"	0.81 (9.79)	0.88 (7.75)	0.34 (22.0)	0.34 (21.9)	0.41 (20.5)	0.42 (20.3)
LEnKF, multiplicative inflation: $\gamma_- \leq 2$, $\gamma_+ \leq 2$	-	0.89 (7.50)	0.34 (22.0)	0.34 (22.0)	0.40 (20.5)	0.40 (20.5)
LEnKF, multiplicative inflation: $\gamma_- \leq 4$, $\gamma_+ \leq 4$	0.82 (9.52)	0.89 (7.48)	0.34 (22.0)	0.34 (22.0)	0.41 (20.6)	0.41 (20.5)

Table 2: Model-observations agreement (R^2 and RMS) for the EnKF data assimilation of only the state and of the joint state (ST), emissions (EM) and lateral boundary conditions (BC) parameters. Visible improvements in both the analysis and the forecast are obtained by adjusting the emissions and lateral boundary conditions.

factor per species per geographic area. The initial ensemble of correction factors is an independent set of normal variables $\alpha_0 \in \mathcal{N}(0, 0.3)$. As in the case of the model states localization, the Kalman increments for the model parameters are localized in the same way: The distance between the observations and each parameter "location" is calculated, and the same correlation function, ρ, is applied to each of the corresponding Kalman increment.

Table 2 shows the model-observations agreement (R^2 and RMS) after LEnKF data assimilation for the state only and for the joint state, emission, and lateral boundary conditions correction coefficients. The forecast results include four scenarios. The state only (ST) forecast considers the solution provided by the analysis without accounting for the emission and boundary conditions correction coefficients (during the forecast window). The rest of the forecast results consider the emission (EM) and boundary condition (BC) coefficients during the forecast window in turn and together. The results show improvement in both the forecast and the analysis windows when emissions and lateral boundary conditions are used in the assimilation process. While in the EnKF setting the assimilation of emission rates was not beneficial, in the LEnKF case under consideration, the addition of the emissions and boundary conditions to the assimilation process improves the assimilated solution. The joint state and parameters forecast results are consistent when both EM and BC correction factors are used, however, the importance of boundary conditions is clearly emphasized by our forecast results.

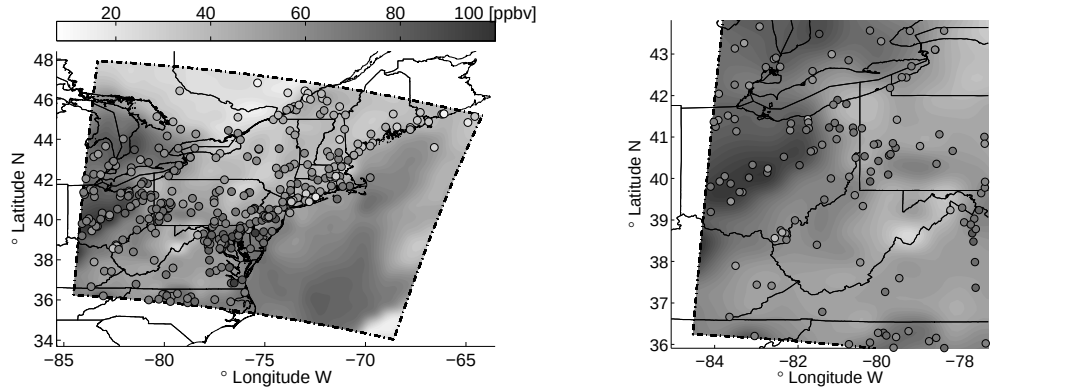
Figure 4 shows the ground level ozone field concentration at 14 EDT in the forecast window (July 22^{nd}) measured by the ICARTT stations and predicted by the best guess, the state-only LEnKF and the joint state and parameters LEnKF with multiplicative inflation. It is clear that assimilating parameters improves the results especially on the inflow boundary. Figures 4.a,b ("best guess") and 4.c,d (state-only LEnKF) show a smooth but qualitatively poor forecast result near the Western boundary. However, Figures 4.e,f (LEnKF, joint state and parameters) show an improvement in the prediction fit near the inflow boundary. Moreover, the quantitative results (R^2 and RMS) show an overall improvement in the analyzed solution, albeit no clear qualitative difference can be distinguished in the rest of the domain.

More research is needed to understand the use of EnKF data assimilation to correct for emissions in chemical transport models in an optimal way.

5 Validation of the Assimilation Results

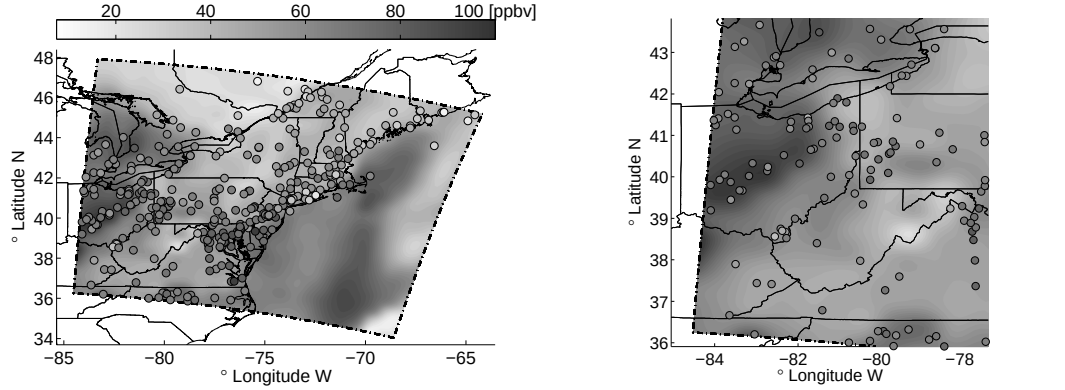
The data assimilation experiments in this paper use only ground ozone observations. While the ground stations provide a rich data set, the concentration fields are not constrained at any of the upper levels. Moreover, no chemical species except ozone is constrained. This section presents a validation of the assimilation results against three independent vertically distributed observations. These data sets were obtained by the two ozonesondes, S1 and S2, and during the P3-B flight (Figure 1.b). The ozonesondes were launched at 14 GMT (S1) and at 22 GMT (S2) July 20^{th}. The NOAA P3-B plane was flown between 14–22 GMT July 20^{th} along the trajectory shown in Figure 1.b at different altitudes (corresponding to grid vertical levels 3–16 in our model).

Figure 5 represents the vertical profile of the ozone concentrations measured by the two ozonesondes (S1 and S2) together with the concentrations predicted by the model after the EnKF and LEnKF data assimilation with additive inflation, multiplicative inflation, and model-specific inflation. The EnKF solutions near the observation sites (on or close to the ground level) where the solution is constrained show a good fit, and the vertically developed correlations improve the solution in that vicinity. At higher altitudes, however, the ozonesondes show an oscillatory behavior of the ozone profile, especially for the additive and multiplicative inflation settings. Model-specific inflation shows a less oscillatory behavior. However, the LEnKF profiles show no oscillations. LEnKF solution gives a fit as good as EnKF does close to the observation sites, and comes closer to the non-assimilated solution at higher altitudes, where there is no information about the truth solution, and thus the model prediction prevails.



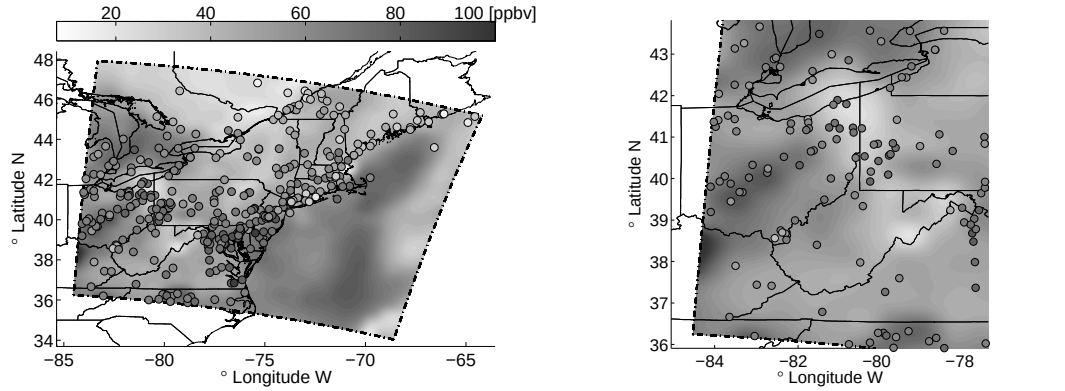
(a) Deterministic (b) Western boundary zoom of (a)

LEnKF: State only



(c) Multiplicative inflation (d) Western boundary zoom of (c)

LEnKF: State + Emissions + Boundary Conditions



(e) Multiplicative inflation (f) Western boundary zoom of (e)

Figure 4: Ground level ozone field concentration at 14 EDT (July 22$^{\text{nd}}$) in the forecast window measured by the ICARTT stations (shown in color coded filled circles) and predicted by: (a,b) the best guess, (c,d) LEnKF with state-only corrections, and (e,f) LEnKF with joint state and parameter corrections. Assimilating parameters improves the results especially on the West - inflow boundary (shown on the right column).

The LEnKF approach forces the correction that each observation exerts on the concentration field to decrease with the distance from the observation site, and thus limits the spatial influence.

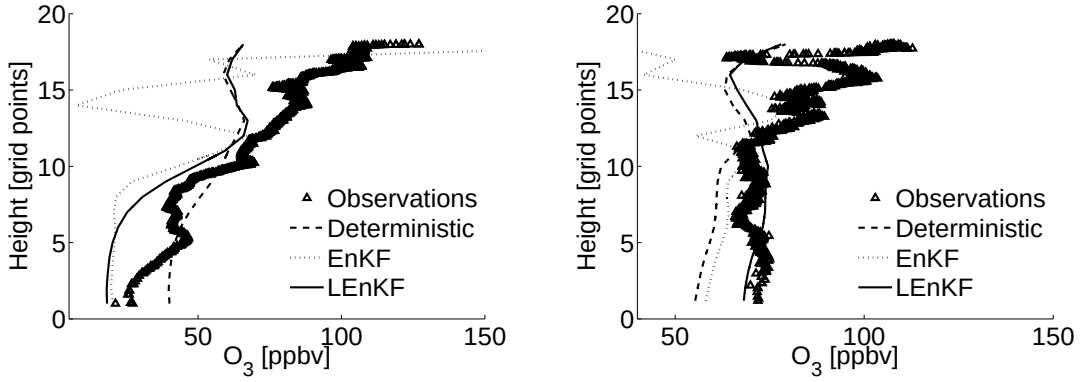
Figure 6 represents the ozone concentrations measured during the P3-B plane flight together with the concentrations predicted after data assimilation with LEnKF and EnKF with multiplicative inflation (Figure 6.a), and model-specific inflation (Figure 6.b). The conclusions closely parallel those of the ozonesondes. EnKF data assimilation with multiplicative covariance inflation (and no localization) is not performing very well in the upper levels of the atmosphere due to over-corrections required by spurious correlations. The LEnKF solution and EnKF solution obtained with the model-specific inflation follow the observations well, although no visible improvement is obtained when compared to the non-assimilated concentrations. Clearly, to fully constrain the ozone field one needs to include measurements of the vertical ozone profiles in the assimilation as well.

6 Conclusions and Future Work

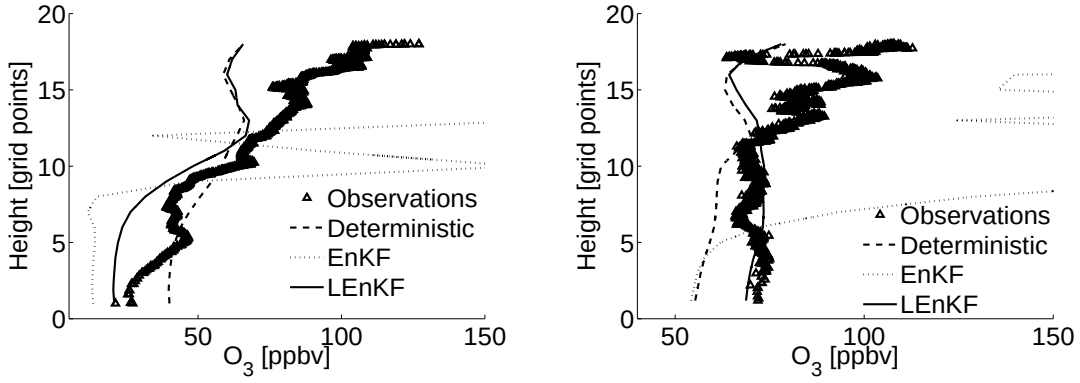
This paper discusses the application of "perturbed observations" EnKF to chemical data assimilation into atmospheric chemistry and transport models. The first part of this study [Constantinescu et al., 2006b] considers an idealized setting for data assimilation and shows a very promising performance of EnKF. The second part of this study [Constantinescu et al., 2006c] reveals the difficulties and challenges of assimilating real data, and discusses different approaches to covariance inflation in order to prevent filter divergence. The third part (this paper) discusses the covariance localization in chemical data assimilation.

Experiments showed that the textbook application of EnKF diverges quickly after about 12 hours of assimilation with small ensembles. In regional air quality simulations the influence of the initial conditions fades in time, as the fields are largely determined by emissions and by lateral boundary conditions. Consequently, the initial spread of the ensemble is diminished in time. Moreover, stiff systems (like chemistry) are stable – small perturbations are damped out quickly in time. Without simulating the atmospheric dynamics (meteorological fields are prescribed) these stiff effects are important. In order to prevent filter divergence the spread of the ensemble needs to be explicitly increased. We consider three different approaches to ensemble covariance inflation: additive, multiplicative, and model-specific. They increase the model solution fit to the observations; however, they also amplify spurious correlations inherent to small-sized ensembles, and greatly deteriorate the analyzed concentration fields away from the observations sites.

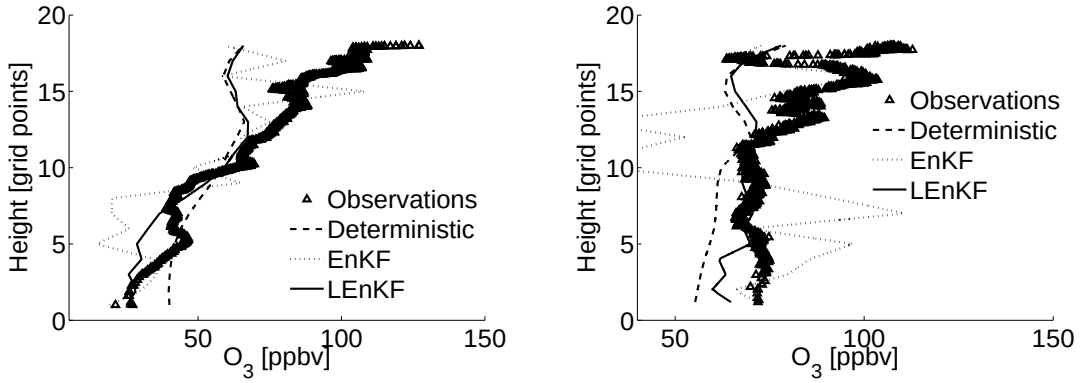
The spurious corrections of the chemical fields located far from the observation sites are



Model-specific inflation: (a) Sonde # 1 (b)Sonde # 2

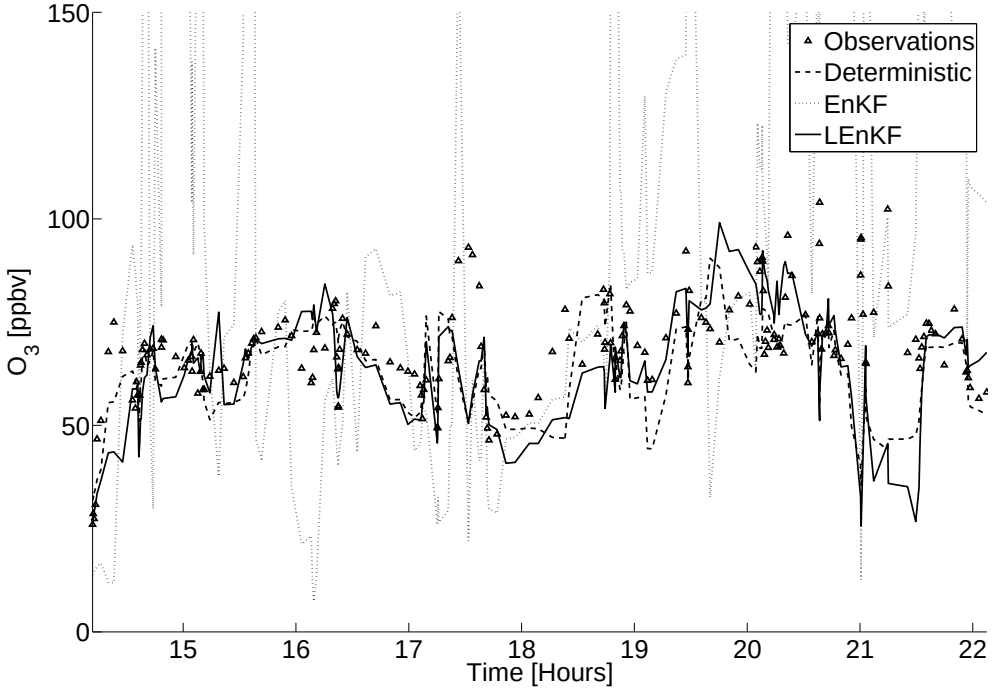


Multiplicative inflation: (c) Sonde # 1 (d)Sonde # 2

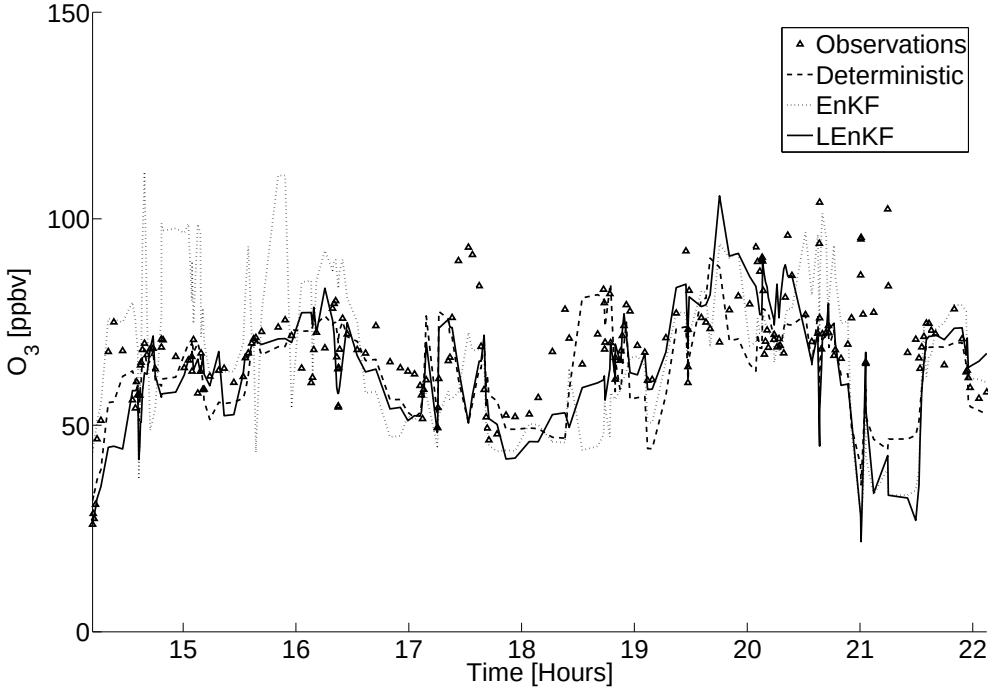


Additive inflation: (e) Sonde # 1 (f)Sonde # 2

Figure 5: Ozone concentrations measured by ozonesondes and predicted by the model after data assimilation with EnKF and LEnKF with the three types of inflation: (a,b) model-specific, (c,d) multiplicative, and (e,f) additive inflation.



(a) Multiplicative inflation



(b) Model-specific inflation

Figure 6: Ozone concentrations measured during the P3-B plane flight, and ozone concentrations predicted after data assimilation with EnKF and LEnKF with (a) multiplicative inflation, and (b) with model-specific inflation. EnKF with multiplicative inflation shows unreasonable oscillations at the higher levels due to over-corrections resulting from spurious correlations. LEnKF and EnKF with model-specific inflation solutions are in better agreement with the observations.

alleviated by using a localized version of the ensemble filter. A decorrelation function is used in order to restrict the filter corrections to the information rich regions, and to dampen the spurios remote correlations developed due to the small ensemble size. The correlation distances are estimated empirically using the NMC approach. Results show that the localized EnKF considerably improves the assimilation results with small (practical) ensemble data assimilation.

The numerical experiments in the idealized setting [Constantinescu et al., 2006b] used vertically distributed observations. The numerical experiments with real data presented in this paper used ground level observations only; the validation results show that the improvements in the vertical profiles are small. It is likely that the information on the vertical distribution of the concentration fields is very important in order to properly constrain three-dimensional concentration fields. In the current experiments, we have used only ozone observations to adjust the concentration fields of 66 different chemical species. The only assimilation results presented are for the ozone fields; to further understand the behavior of EnKF future studies should analyze the corrections of other chemical fields as well.

Since the solution of regional CTMs is largely influenced by uncertain lateral boundary conditions and uncertain emissions, it is of great importance to adjust these parameters through data assimilation. In the idealized setting [Constantinescu et al., 2006b] the assimilation of emissions and boundary conditions considerably improves the quality of the analysis. In the real case [Constantinescu et al., 2006c] the assimilation of emissions did not improve the analysis and degraded the forecast solution. The joint assimilation of states, emissions, and lateral boundary conditions with localized EnKF improves the analysis and the forecast solutions. Joint parameter and state data assimilation in this context is a challenging problem, and considerably more research is needed to fully resolve it.

More work is required to completely understand the use of ensemble data assimilation to reduce uncertainties in emission inventories and in boundary conditions. One challenge arises from the long integration times needed to develop meaningful correlations between the emission rates or boundary conditions and the concentration fields. Another challenge is posed by large spurious correlations which lead the filter to correct the emission rates and boundary conditions in order to compensate for other sources of error.

Finally, to fully understand the ensemble chemical data assimilation, one needs to explore the capability of small ensembles to capture correlations due to chemical interactions. The challenge is due to very short temporal and spatial scales on which chemical interactions take place. One needs to consider how observations of several different chemical species improve the analysis of other chemical fields.

In this paper we considered the "perturbed observations" version of EnKF. The per-

formance of the "square root" EnKF variants will need to be assessed in the context of chemical data assimilation. In the future we plan to develop hybrid methods that combine the advantages of the 4D-Var and EnKF data assimilation approaches.

Acknowledgments

This work was supported by the National Science Foundation through the awards NSF CAREER ACI–0413872, NSF ITR AP&IM 0205198, NSF CCF 0515170, by NOAA, and by the Houston Advanced Research Center through the award H59/2005.

References

J.D. Annan, J.C. Hargreaves, N.R. Edwards, and R. Marsh. Parameter estimation in an intermediate complexity earth system model using an ensemble Kalman filter. *Ocean Modelling*, 8:135–154, 2005.

G. Burgers, P.J. van Leeuwen, and G. Evensen. Analysis scheme in the ensemble Kalman Filter . *Monthly Weather Review*, 126:1719–1724, 1998.

G.R. Carmichael, Y. Tang, G. Kurata, I. Uno, D. Streets, J.H. Woo, H. Huang, J. Yienger, B. Lefer, R. Shetter, D. Blake, E. Atlas, A. Fried, E. Apel, F. Eisele, C. Cantrell, M. Avery, J. Barrick, G. Sachse, W. Brune, S. Sandholm, Y. Kondo, H. Singh, R. Talbot, A. Bandy, D. Thorton, A. Clarke, and B. Heikes. Regional-scale Chemical Transport Modeling in Support of the Analysis of Observations obtained During the Trace-P Experiment. *Journal of Geophysical Research*, 108(D21 8823):10649–10671, 2003.

T. Chai, G.R. Carmichael, A. Sandu, Y. Tang, and D.N. Daescu. Chemical data assimilation of Transport and Chemical Evolution over the Pacific (TRACE-P) aircraft measurements. *Journal of Geophysical Research*, 111(D02301):10.1029/2005JD005883, 2006.

E.M. Constantinescu, A. Sandu, T. Chai, and G.R. Carmichael. Autoregressive models of background errors for chemical data assimilation. *In preparation*, 2006a.

E.M. Constantinescu, A. Sandu, T. Chai, and G.R. Carmichael. Ensemble-based chemical data assimilation I: An idealized setting. *In preparation*, 2006b.

E.M. Constantinescu, A. Sandu, T. Chai, and G.R. Carmichael. Ensemble-based chemical data assimilation II: Real observations. *In preparation*, 2006c.

P. Courtier, J.-N. Thepaut, and A. Hollingsworth. A strategy for operational implementation of 4D-Var, using an incremental approach. *Quarterly Journal of the Royal Meteorological Society*, 120:1367–1387, 1994.

J. Derber. A variational continuous assimilation scheme. *Monthly Weather Review*, 117: 2437–2446, 1989.

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

G. Evensen. The Ensemble Kalman Filter: theoretical formulation and practical implementation. *Ocean Dynamics*, 53, 2003.

G. Evensen. The combined parameter and state estimation problem. *Ocean Dynamics*, SUBMITTED, 2005.

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

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

T.M. Hamill. Ensemble-based atmospheric data assimilation. Technical report, University of Colorado and NOAA-CIRES Climate Diagnostics Center, Boulder, Colorado, USA, 2004.

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

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

ICARTT. ICARTT home page:http://www.al.noaa.gov/ICARTT . URL `http://www.al.noaa.gov/ICARTT`.

R.E. Kalman. A new approach to linear filtering and prediction problems . *Transaction of the ASME- Journal of Basic Engineering*, 82:35–45, 1960.

W. Liao, A. Sandu, G.R. Carmichael, and T. Chai. Total energy singular vector analysis with atmospheric chemical transport models. *SUBMITTED*, 2005.

H.L. Mitchell and P.L. Houtekamer. An adaptive ensemble Kalman filter. *Monthly Weather Review*, 128:416–433, 1999.

E. Ott, B. R. Hunt, I. Szunyogh, A. V. Zimin, E. J. Kostelich, M. Corazza, E. Kalnay, D. J. Patil, and J. A. Yorke. A local ensemble Kalman filter for atmospheric data assimilation. *ArXiv Physics e-prints*, 2002.

D.F. Parrish and J.C. Derber. The national meteorological center's spectral statistical-interpolation analysis system. *Monthly Weather Review*, (120):1747–1763, 1992.

F. Rabier, H. Jarvinen, E. Klinker, J.F. Mahfouf, and A. Simmons. The ECMWF operational implementation of four-dimensional variational assimilation. I: Experimental results with simplified physics. *Quarterly Journal of the Royal Meteorological Society*, 126:1148–1170, 2000.

A. Sandu and D. Daescu. Discrete adjoints for stiff odes. *In preparation*, 2005.

A. Sandu, D. Daescu, and G.R. Carmichael. Direct and adjoint sensitivity analysis of chemical kinetic systems with kpp: I – theory and software tools. *Atmospheric Environment*, 37:5,083–5,096, 2003.

A. Sandu, D. Daescu, G.R. Carmichael, and T. Chai. Adjoint sensitivity analysis of regional air quality models. *Journal of Computational Physics*, 204:222–252, 2005.

S. Szunyogh, E.J. Kostelich, G. Gyarmati, D.J. Patil, B.R. Hunt, E. Kalnay, E. OTT, and J.A. Yorke. Assessing a local ensemble Kalman filter: perfect model experiments with the National Centers for Environmental Prediction global model. *Tellus A*, 57(4):528–545, 2005.

Y. Tang, G.R. Carmichael, N. Thongboonchoo, T. Chai, L.W. Horowitz, R.B. Pierce, J.A. Al-Saadi, G. Pfister, J.M. Vukovich, M.A. Avery, G.W. Sachse, T.B. Ryerson, J.S. Holloway, E.L. Atlas, F.M. Flocke, R.J. Weber, L.G. Huey, J.E. Dibb, D.G. Streets, and W.H. Brune. The influence of lateral and top boundary conditions on regional air quality prediction: a multi-Scale study coupling regional and global chemical transport models. *SUBMITTED to Journal of Geophysical Research*, 2006.