Ensemble-based Chemical Data Assimilation I: General Approach

By Emil M. Constantinescu1, Adrian Sandu1, *, Tianfeng Chai2, and Gregory R. Carmichael2

1 Department of Computer Science, Virginia Polytechnic Institute and State University, Blacksburg, VA 24061
2 Center for Global and Regional Environmental Research, The University of Iowa, Iowa City, IA 52240

(Received 25 March 2006; revised xx Xxxx 2006)

SUMMARY

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 assess the performance of the ensemble Kalman filter (EnKF) and compare it with a state of the art 4D-Var approach. We analyze different aspects that affect the assimilation process, and investigate several ways to avoid filter divergence. Results with a real model and real observations show that EnKF is a promising approach for chemical data assimilation. The results also point to several issues on which further research is necessary.

KEYWORDS: data assimilation, ensemble Kalman filter, 4D-Var, chemical transport models

1. INTRODUCTION

Data assimilation is the process by which model predictions utilize measurements to obtain an optimal representation of the state of the atmosphere. Data assimilation is recognized as essential in weather/climate analysis and forecast activities, and is accomplished by a mature experience/infrastructure. Both variational (Barkmeijer et al., 1999) and ensemble based (Molteni et al., 1996; Buizza et al., 2000) approaches to data assimilation are being successfully employed. 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.

Variational techniques for data assimilation are well-established in numerical weather prediction (NWP). Building on the early variational approach (Lorenc, 1986; Le Dimet and Talagrand, 1986; Talagrand and Courtier, 1987), the 4D-Var framework is the current state-of-the-art in meteorological (Courtier et al., 1994; Rabier et al., 2000) and chemical (Elbern et al., 2000a, 2001a; Liao et al., 2005; Sandu et al., 2003, 2005; Sandu, 2006; Segers, 2002) data assimilation. Ensemble Kalman filter 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.

Houtekamer et al. (2005) compare 3D-Var and EnKF in an operational (real) setting. Their results show difficulties for EnKF to match 3D-Var’s solution. To our knowledge, this is the first comparison between variational and ensemble data assimilation based on real data. Lorenc (2003), Hamill (2004), and Kalnay et al. (2005) discuss theoretically the relative merits of the two methods. They conclude that 4D-Var and EnKF have their own particular advantages and disadvantages, neither being a clear winner, albeit more research needs to be done at least to assess EnKF’s practical merits.

The goal of this paper is to investigate the application of the ensemble Kalman filter (EnKF) to atmospheric chemical data assimilation. Considerable progress has been made...
made recently in the development of variational tools for chemical data assimilation (Liao et al., 2005; Sandu et al., 2003, 2005; Sandu, 2006). However, compared with NWP, little work has been done to date to assimilate chemical observations using nonlinear ensemble filters. EnKF, extended Kalman filter (van Loon and Heemink, 1997) and reduced rank square root Kalman filter (Segers et al., 2000; Hanea et al., 2004) have been used in chemical data assimilation to recover ozone (van Loon et al., 2000), and various ways of accurately quantifying the uncertainty in sources have been investigated. This work shows that it is possible to successfully apply ensemble (Kalman) data assimilation approaches to atmospheric CTMs, and to improve the quality of the forecasts. A comparison among different flavors of reduced Kalman filters is given in (Heemink and Segers, 2002). We study EnKF applied to a large set of species (66) not only to one (ozone) as in previous studies, compare it to a variational technique, and investigate various ways to avoid filter divergence using real data.

In a previous study (Constantinescu et al., 2007), henceforth (Con07), we analyzed the performance of EnKF applied to chemical and transport models 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. One of the perturbed solutions was considered to be the “best guess”, and we analyzed how close this solution is to the “truth” with and without data assimilation. The results indicate that EnKF is able to recover the reference solution with very good accuracy and to improve the forecast.

Motivated by the encouraging results obtained in the idealized case, we continue the analysis of EnKF for atmospheric chemical data assimilation on a real scenario. The initial state is the best guess of the system, and we decrease the uncertainty by assimilating real observations. The “truth” is unknown and the assimilated solution is validated against independent observational data. The main contributions of this work are: (1) a comparison between “perturbed observations” EnKF and the state-of-the-art 4D-Var in an operational-like setting using real data, (2) an analysis of several methods to inflate the ensemble covariance and avoid filter divergence, (3) an examination of the localized EnKF (LEnKF) in an operational-like setting, and (4) a state-parameter inversion approach for our model. Furthermore, the ensemble initialization was obtained through a novel approach based on autoregressive processes (Constantinescu et al., 2006a). A novel approach to localize the inflation is proposed and is shown to considerably improve the filter performance.

This study is divided in two parts. In the first part (this paper) we begin by describing the differences between NWP and chemical data assimilation and, for the latter case, compare EnKF with a well established variational technique (4D-Var) using real data. Then, we investigate several ways to avoid filter divergence. We assess what is a “good” size of the ensemble for our application. In the second part of this study (the companion paper), we concentrate on the localized EnKF and show how the problems caused by inflating the ensemble covariance can be circumvented. Moreover, we show how the forecasting capabilities are improved via the inversion of the emissions and boundary conditions in a regional atmospheric CTM model.

The paper is structured as follows. Sections 2(a) and (b) briefly review the 4D-Var and EnKF methods, respectively. Section 3 describes chemical and transport models and the particular scenario used in this study, the ensemble initialization and 4D-Var background covariance formation, and the analysis setting. A comparison between 4D-Var and EnKF data assimilation applied to our atmospheric CTM is shown and discussed in Section 4. Several strategies to inflate the ensemble covariance and avoid filter
divergence are addressed in 5. A validation of the data assimilation results is carried out in 6. Discussion and a summary are given in Section 7.

2. Data Assimilation

In this section we briefly review the 4D-Var approach to data assimilation. More details on 4D-var can be found in (Courtier et al., 1994; Rabier et al., 2000), and on the ensemble Kalman filter in our previous study (Con07).

Consider a nonlinear model $c_i = M_{t_0 \rightarrow t_i}(c_0)$ 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 $c_t$ at $t_i$ (more exactly $c_i$ is the system state projected onto the model space space).

The initial model state is uncertain (and consequently, future states are also uncertain). For example, assuming a normal distribution of uncertainty, the initial state is characterized by its mean $c_B$ (the “background” state, or the best initial guess) and its covariance matrix $B$. Observations $y_i$ of the real system are available at times $t_i$ and are corrupted by measurement and representativeness errors $\varepsilon_i$ (assumed Gaussian with mean zero and covariance $R_i$)

$$y_i = \mathcal{H}_i(c_i) + \varepsilon_i.$$ 

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 both the information from the model ($c_i$, $i \geq 0$) and from the observations ($y_i$, $i \geq 0$).

(a) 4D-Var

In 4D-Var (Courtier et al., 1994; Rabier et al., 2000) the best estimate of the initial state (conditioned by the observations $y_0 \cdots y_n$) is obtained as the minimizer of the following cost function (which measures the model-observations misfit)

$$J(c_0) = \frac{1}{2}(c_0 - c^B)^T B^{-1} (c_0 - c^B) + \frac{1}{2} \sum_{i=0}^{n} (y_i - \mathcal{H}_i(c_i))^T R_i^{-1} (y_i - \mathcal{H}_i(c_i)). \quad (1)$$

A gradient-based minimization method is typically employed (in our experiments we used L-BFGS-B (Byrd et al., 1995)). The gradient of the cost function with respect to the initial state is obtained as

$$\nabla_{c_0} J = B^{-1} (c_0 - c^B) + \sum_{i=0}^{n} M^*_{t_i \rightarrow t_0} H^T_i R_i^{-1} (y_i - \mathcal{H}_i(c_i)), \quad (2)$$

where $M = \mathcal{M}'$ is the tangent linear model associated with $\mathcal{M}$, $M^*$ is the adjoint of $M$, and $H = \mathcal{H}'$ is the linearized observation operator. More information about variational data assimilation can be found in (Chai et al., 2006), and the adjoint derivation for the model we used in our numerical experiments can be found in (Sandu et al., 2005).

(b) The Ensemble Kalman Filter (EnKF)

The Kalman filter estimates the true state $c_t$ at $t_i$ using the information from the current best estimate $c_f$ (the “forecast” or the background state) and the observations $y_i$. The optimal estimate $c_a$ (the “analysis” state) is obtained as a linear combination of
the forecast and observations that minimize the variance of the analysis (\( P^a \))
\[
c_i^a = c_i^f + P_f^i H_i^T (H_i P_f^i H_i^T + R_i)^{-1} \left( y_i - \mathcal{H}_i(c_i^f) \right) = c_i^f + K_i \left( y_i - \mathcal{H}_i(c_i^f) \right),
\]
(3)
The forecast covariance \( P_f \) is estimated from an ensemble of runs (which produces an ensemble of \( E \) model states \( c_i^f(e), e = 1, \ldots, E \)). The analysis formula (3) is applied to each member to obtain an analyzed ensemble. The working of the filter applied to a perfect model 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 \( t_i \):
\[
c_i^f(e) = M\left( c_{i-1}^a(e) \right), \quad c_i^a(e) = c_i^f(e) + K_i \left( y_i - \mathcal{H}_i(c_i^f(e)) \right), \quad e = 1, \ldots, E.
\]
(4)
The results presented in this paper are obtained with the practical EnKF implementation discussed by Evensen (Evensen, 2003).

3. Experiment Setting

We now discuss the chemical and transport model used in our experiments, the particular scenario simulated, the ensemble initialization and 4D-Var background modeling, and the setting of the analysis.

(a) The Model

Our data assimilation numerical experiments use the state-of-the-art regional 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).

In STEM, the evolution of \( N_{\text{spec}} \) species is described by the following equations
\[
\frac{\partial c_s}{\partial t} = -u \nabla c_s + \frac{1}{\rho} \nabla (\rho K \nabla c_s) + \frac{1}{\rho} f_s (\rho c_s) + E_s, \quad t^0 \leq t \leq t^F, \quad 1 \leq s \leq N_{\text{spec}};
\]
\[
c_s(t^0, x) = c_{s0}(x),
\]
\[
c_s(t, x) = c_{s}^{\text{in}}(t, x) \quad \text{for} \quad x \in \Gamma^{\text{in}},
\]
\[
K \frac{\partial c_s}{\partial n} = 0 \quad \text{for} \quad x \in \Gamma^{\text{out}},
\]
\[
K \frac{\partial c_s}{\partial n} = V_s^{\text{dep}} c_s - q_s \quad \text{for} \quad x \in \Gamma^{\text{ground}},
\]
where the concentration \( c_s \) for species \( s \) is dictated by \( q_s \), the rate of surface emissions, \( E_s \) the rate of elevated emissions, and \( f_s \), the rate of chemical transformation for this species. Further, \( u \) denotes the wind field vector, \( K \) the turbulent diffusivity tensor, \( \rho \) the air density, \( c^{\text{in}} \) the Dirichlet boundary conditions, and \( V_s^{\text{dep}} \) the deposition velocity.

The model can be written compactly as
\[
c_i = M\left( c_{i-1}, u_{i-1}, c_{i-1}^{\text{in}}, q_{i-1} \right).
\]
(6)
where \( c \) is the vector of concentrations (all species at all grid points). 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.
In regional air quality simulations, the influence of the initial conditions fades in time, and the concentration fields become largely driven by emission and removal processes and by lateral boundary conditions. Consequently, for ensemble simulations, the initial spread of the ensemble is diminished in time. Moreover, stiff systems like chemistry are very stable - small perturbations are damped out quickly in time as the system tends to evolve toward quasi-steady-state. After a short time the chemical evolution collapses onto a low dimensional manifold in state space. With prescribed meteorological fields, the stiff effects are important for the entire dynamics of the system.

An additional difficulty arises from the truly multiscale character of CTM dynamics. Many chemical and aerosol processes take place at very short temporal and spatial scales. Observations of chemical and particulate concentrations are strongly influenced by the local variability, yet we use them to constrain large 3-D fields. Correlations between chemical species (due to chemical interactions) and between chemical and dynamic variables (due to transport processes) need to be correctly represented by small ensembles.

A strong constraint for this model is the requirement that $c_i$ be positive (negative concentrations are non-physical). STEM numerical methods also require positive concentrations. This fact plays an important role in both variational and sequential data assimilation since the assimilation process may produce negative concentration values. In the variational approach, it is typically easy to bound the solution of the optimization process to be positive. In the ensemble case, this problem can be alleviated through shifting the negative values to zero; however, biases may be introduced.

\subsection*{(b) The Case Study}

The test case is a real-life simulation of air pollution in North–Eastern United States in July 2004 as shown in Fig. 1 (the dash-dotted line delimits the computational domain). 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) campaign in summer 2004 (July 5, 19 and 21). A detailed description of the ICARTT fields and data can be found in (Tang et al., 2006). 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 $60 \times 60$ 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 20th, 2004.

We selected a number of six stations throughout the domain to plot the time evolution of measured and modeled ozone concentrations and illustrate the effect of different data assimilation scenarios. The selected stations are shown in Fig. 1.b and correspond to the following ICARTT IDs: '#a' - 00065001 (close to the Great Lakes), '#b' - 230310038 (coastal station, close to Portland, ME), '#c' - 90070007 (coastal station, close to New York, NY), '#d' - 420270100 (center of the continental domain), '#e' - 510590030 (in Washington DC), '#f' - 391514005 (inflow boundary). Our study also includes three validation measurements taken by two ozonesondes and a P3-B flight (all shown in Fig. 1.b).
(c) **Modeling the Background Errors**

Our current knowledge of the state of the atmosphere (at the beginning of the simulation) is represented by the “background” field and its error. In practice, little is known about the background error; it is typically assumed to be Gaussian and with zero mean (the model is unbiased) and covariance $\mathbb{B}$. In EnKF the background covariance is used to generate the initial ensemble, while in 4D-Var the background covariance is used explicitly in formulation of the cost function. A good approximation of the background error statistics is therefore essential for the success of both ensemble and variational data assimilation.

In both EnKF and 4D-Var (Constantinescu et al., 2007; Sandu et al., 2005) we consider background errors modeled by autoregressive (AR) processes of the form

$$
\mathbf{A} \mathbf{\delta} \mathbf{e}^{\mathbb{B}} = \mathbf{S} \mathbf{\xi}, \quad \mathbf{S} = \text{diag}(\sigma_{1,j,k}),
$$

where $\mathbf{\xi}(\mathbf{e}) \in (\mathcal{N}(0, 1))^N$ (of mean 0 and standard deviation 1) is a vector of $N$ independent normal random variables, $\mathbf{A}$ is a correlation coefficient matrix, and $\sigma$ represents the state covariances. The AR background accounts for spatial correlations, distance decay, and chemical lifetime. For more details on the construction and application of the AR background model the reader is referred to (Constantinescu et al., 2006a).

(d) **Analysis Setting**

This section discusses the setting of both 4D-Var and EnKF data assimilation experiments.

All the simulations are started at the same time (0 GMT July 20th) with a four hour initialization step. This allows the background 4D-Var run and each of the ensemble members to reach quasi-steady-state before the assimilation window. We shall denote the initialization window as [-4,0] hours. The “best guess” of the state of the atmosphere at 0 GMT July 20th 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 results section. The best guess evolved to 4 GMT July 20th represents the background state in 4D-Var. The ensemble members are formed by adding a set of unbiased perturbations to the best guess at 0 GMT, then evolving each member to 4 GMT July 20th.

The 24 hours assimilation window starts at 4 GMT July 20th and ends at 4 GMT July 21st (henceforth denoted as [0,24] hours). Observations are available at each integer hour.
in this window (i.e., at 0, 1, . . ., 24 hours). The ozone $O_3$ observations used in this study are from the ICARTT ground stations (Fig. 1). Not all the stations provide observations each hour (the number of hourly observations varies between 160 and 326 during the assimilation window). The observation error covariance (both in EnKF and 4D-Var), $R$, was considered a diagonal matrix with $(0.25\text{ppbv})^2$ standard deviation form the respective measurement.

EnKF adjusts the concentration fields of 66 “control” chemical species in each grid point of the domain every hour using (3). Two ensemble sizes were considered with 50 and 200 members. Ensembles of 50 members are typical in numerical weather prediction and they are thought to provide a good balance between accuracy and computational efficiency; however, the ensemble size is application-specific and is given by the principal directions of the error growth. In our idealized experiment (Con07), a 50 member ensemble showed significant improvements against smaller ensembles. On the other hand, 50 members ensembles provide a convenient computational cost for our numerical results pending the computational resources available at that time. The 200 member runs have been performed mainly for comparison purposes.

4D-Var adjusts the initial concentrations of the 66 control chemical species at each grid point at the beginning of the assimilation window (4 GMT July 20th). The L-BFGS iterations are stopped when the cost function is reduced to less than $10^{-3}$ of its initial value ($J = 10^{-3} J_0$), or when the number of iterations exceeds 25.

The 24 hours forecast window starts at 4 GMT July 21st and ends at 4 GMT July 22nd (the forecast windows will be denoted further as [24,48] hours). The model is initialized at 4 GMT July 22nd with the evolved optimal solution in case of 4D-var, and with the ensemble mean in case of EnKF, and evolved in forecast mode for 24 hours.

An important challenge is raised by the positivity of chemical concentration fields, a constraint inherent to chemical transport modeling. In 4D-Var positivity can be imposed as a bound constraint in the optimization procedure (and is easily accommodated by L-BFGS-B (Byrd et al., 1995)). In EnKF it is difficult to impose the positivity constraint and the analysis (3) may result in negative concentrations. The simple strategy of setting all negative concentrations to zero introduces bias in the analysis. A limit on the ensemble variance can alleviate this problem.

The performance of each data assimilation experiment is measured by the $R^2$ correlation and RMS factors between the observations and the model solution (separate $R^2$ and RMS factors are computed in the assimilation and in the forecast windows). The $R^2$ correlation and RMS factor of two series $x$ and $y$ of length $n$ are

\[
R^2(x, y) = \frac{\left(n \sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i \right)^2}{\left(n \sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2 \right) \left(n \sum_{i=1}^{n} y_i^2 - (\sum_{i=1}^{n} y_i)^2 \right)},
\]

\[
RMS(x, y) = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (x_i - y_i)^2}.
\]

In our experimental setting the deterministic (best guess) solution yields an $R^2$ (RMS) of 0.24 (22.1) in the analysis and 0.28 (23.5) in the forecast windows. We aim to improve these results by using the two assimilation methods (EnKF and 4D-Var).

4. COMPARISON BETWEEN ENKF AND 4D-VAR

An excellent comparison of the relative merits of EnKF and 4D-Var in the context of numerical weather prediction was given by Lorenc (2003) and expanded by Kalnay
E.M. Constantinescu et al. (2005). Hamill makes a theoretical analysis of the two approaches in (Hamill, 2004). A direct comparison of operational systems involving 3D-Var and EnKF can be found in (Houtekamer et al., 2005), where promising results for ensemble filtering are shown.

Similar arguments for the relative merits of EnKF and 4D-Var can be considered in the context of CTMs. EnKF is simple to implement, while 4D-var requires the construction of adjoint models, a non-trivial task in the presence of stiff chemistry (Sandu et al., 2005). EnKF allows for a simple integration of model errors, whereas 4D-Var assumes a perfect model. The ensemble propagates the forecast covariance and a very good estimate of the background covariance is readily available at the beginning of the next assimilation cycle. On the other hand the 4D-Var optimal solution is consistent with model dynamics throughout the assimilation window. 4D-Var naturally incorporates asynchronous observations while for EnKF asynchronous observations require a more involved framework (Hunt et al., 2004). A consistent derivation of the initial ensemble in EnKF is difficult (Constantinescu et al., 2006a). Moreover, in the presence of stiff chemistry, it is likely that each application of the filter will throw the model state off the quasi-steady-state; consequently, after each assimilation cycle a new stiff transient will be introduced, and this may considerably impact the computational time needed to advance the model state for each ensemble member. It is not clear at this time how does the computational cost of EnKF compare with that of 4D-Var (in order to obtain similar performance); in this context, we considered perfect models for this comparison. To the best of our knowledge comprehensive tests of EnKF versus 4D-Var have not been carried out so far.

The EnKF forecast can be done by evolving each individual member (ensemble forecast) or by performing a single model integration initialized with the best estimate (the ensemble average at the end of the assimilation window). In the latter situation the forecast costs of 4D-Var and EnKF are the same. On the other hand the ensemble forecast provides an estimate of uncertainty in model predictions over the forecast window. In the results presented in this paper the forecasts after EnKF assimilation are computed using a single model integration.

We first performed a “noiseless application” of EnKF using 50 and 200 member ensembles. Table 1 shows the $R^2$ and RMS results between the observations and model values for all state assimilated numerical experiments. For each scenario the ensemble size, setting information, and $R^2$ (RMS) for the analysis and forecast windows are presented. The results with the 50 member ensemble are presented as EnKF experiment #1, and the results with 200 are presented as EnKF experiment #11.

The correlation factor between model and observations in the assimilation window is $R^2 = 0.24$ (RMS = 22.1) for the non-assimilated run. It grows to $R^2 = 0.40$ (RMS = 23.5) for the solution assimilated with the 50 member ensemble and to $R^2 = 0.49$ (RMS = 16.29) for the solution assimilated with the 200 member ensemble. The large ensemble solution comes close to the correlation factor of the 4D-Var assimilated solution ($R^2 = 0.52$, RMS = 16.05). None of the methods, however, is able to considerably improve the model-observations correlation in the forecast window.

To further understand the behavior of the filter we look at the time evolution of ozone concentrations at the selected ground stations. Figures 2.a-f show the time series of ozone observations, and the non-assimilated, EnKF #1, and 4D-Var solutions. After the first 12 hours the EnKF solution comes very close to the non-assimilated one and “ignores” further observations. Clearly the filter diverges. Without an effective influence of the new observations the solution is driven by emissions and (lateral) boundary conditions. Another result in support of the filter divergence is EnKF #11, shown
in Fig. 5. Increasing the ensemble size to 200 members doubles the accuracy of the estimated covariances. The analysis is improved by a small factor in the beginning of the assimilation window when the ensemble variance is large enough, but after 12 hours the filter diverges as well (and fails to bring any improvement in the second half of the assimilation window or in the forecast). These results are expected since we consider a perfect model for the purpose of the 4D-V ar - EnKF comparison.

A conclusion of this numerical experiment is that both EnKF and 4D-V ar methods perform well in the beginning of the assimilation window.

Now we look at several ways to prevent the filter divergence by inflating the ensemble covariance.

5. Preventing Filter Divergence

The previous section shows that the “noiseless application” of EnKF (Evensen, 2003) (perfect model assumption) to our particular scenario leads to filter divergence: EnKF shows a decreasing ability to correct the ensemble state toward the observations at the end of the assimilation window. 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. The cure is 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 this section we investigate several ways to “inflate” the ensemble covariance in order to prevent filter divergence. The first method is the additive inflation (Corazza et al., 2002), where we simulate model errors by adding uncorrelated noise to model results. This increases the diagonal entries of the ensemble covariance. The second
method is the *multiplicative inflation* (Anderson, 2001), where each member’s deviation from the ensemble mean is multiplied by a constant. An “online” estimation of the inflation constant is possible (Kalnay et al., 2005; Miyoshi, 2005). This increases each entry of the ensemble covariance by that constant squared. Finally we discuss the covariance inflation obtained through perturbing key model parameters, and we call it *model-specific inflation*. We note that a better approach can be obtained by constructing multi-model ensembles (McKeen et al., 2005).

(a) Additive Inflation

The additive inflation process (Corazza et al., 2002) consists of adding random noise to the model solution; the noise can be thought of as a representation of the unknown model error. With the assumption that the model error is unbiased we add white noise \( \eta \in \mathcal{N}(0, Q) \) of mean zero and covariance matrix \( Q \).
The most intuitive way is to add noise to the forecast solution. The net result is to increment the forecast covariance by $Q$. With the notation (4)

$$c_f^i(e) = M(c_{i-1}^o(e) + \eta(e)),$$  
$e = 1, \ldots, E, \quad \Rightarrow \quad P_f^i \leftarrow P_f^i + Q.$

In the ideal situation $Q$ should reflect the correlation of the model errors. Since these are very much unknown one typically chooses white noise; i.e., the covariance matrix $Q$ is diagonal ($\eta$ is a vector of independent random variables). The experiment EnKF #2 presented in Table 1 is an application of the filter with additive inflation, with white noise added before assimilation (to $c_f^i$). An independent random perturbation drawn from a normal distribution with mean zero and standard deviation of 6 ppb is added to ozone in each grid point. Note that the perturbations can be negative and large and the perturbed ozone concentration can become negative; in this case the concentrations are set to zero. In order to avoid excessive biases induced by the truncations a perturbation is added in a grid point only if the ozone concentration is larger than 5 ppb.

Another way is to add the noise right after each assimilation step. This noise is evolved through the model (from $t_{i-1}$ to $t_i$) and the resulting perturbation in the forecast state will present appropriate correlations. The forecast covariance is thus added a covariance matrix that captures at least some of the off-diagonal elements of the model error covariance. With the notation (4)

$$c_f^i(e) = M(c_{i-1}^o(e) + \eta(e)),$$  
$e = 1, \ldots, E.$

The experiment EnKF #3 presented in Table 1 adds white noise to ozone after each assimilation step. The noise has a standard deviation of 6 ppb and is added only if the ozone concentration is larger than 5 ppb to minimize biases resulting from truncation.

Adding white noise before the assimilation has a negative impact on the off diagonal elements of the background covariance by diminishing their relative weight, while adding perturbations after the assimilation (and before the integration) allows correlations to redevelop, and this is reflected in our results (Fig. 3). They both perform well in the analysis where the increased variation of the background allows the filter to better account for the observations. The lack of off diagonal correlation and the amount of unstable modes (model states) render the EnKF #2 to perform poorly in the forecast, while EnKF #3 performs better than the noiseless EnKF (#1). In our case, EnKF #2 developed oscillations in the solution, as they can be noticed in the Figs. 3.a-f.

(b) Multiplicative Inflation

The multiplicative approach to covariance inflation (Anderson, 2001) is to enlarge the spread of the ensemble about its mean by a scalar factor $\gamma > 1$. The result is an increase of the ensemble covariance by $\gamma^2$ while the ensemble mean remains unchanged. The filter trust in the model is thus degraded while the correlations developed through the ensemble evolution are preserved (both diagonal and off-diagonal entries of the covariance matrix are scaled by the same amount).

One can inflate the forecast ensemble before filtering

$$c_f^i(e) \leftarrow \langle c_f^i(e) \rangle + \gamma_-(c_f^i(e) - \langle c_f^i(e) \rangle), \quad e = 1, \ldots, E, \quad \Rightarrow \quad P_f \leftarrow \gamma_2^2 P_f,$$

(where $\langle \cdot \rangle$ denotes the ensemble average) or the analyzed ensemble after filtering:

$$c_a^i(e) \leftarrow \langle c_a^i(e) \rangle + \gamma_+(c_a^i(e) - \langle c_a^i(e) \rangle), \quad e = 1, \ldots, E, \quad \Rightarrow \quad P_a \leftarrow \gamma_2^2 P_a.$$

The inflation of the analysis covariance prepares an ensemble of larger spread for the integration over the next time interval. Note that the multiplicative covariance inflation
procedure changes the concentrations and may lead to negative concentration values. One needs to set these negative concentrations to zero, which may change the ensemble mean (and bias the estimate).

An important decision in the multiplicative covariance inflation is the choice of the inflation factors $\gamma_{\pm}$. Small inflation factors do not prevent filter divergence. Large values lead to overconfidence in measurements, may amplify spurious correlations, and may lead to large biases after the negative concentrations are set to zero. The inflation factors are usually estimated by trial and error. Typical values found in the meteorological literature (Anderson, 2001) are small ($1.01 \leq \gamma \leq 1.2$). Values in this range did not bring any noticeable improvement to the analysis in our tests. Therefore we have implemented an adaptive scheme to determine the magnitude of the apriori ($\gamma_{-}$) and aposteriori ($\gamma_{+}$) inflation factors. We estimate the variance of the observed species

Figure 3. Ozone concentrations measured at the selected stations and predicted by EnKF #2 (±6ppb white noise added before each filtering step if $O_3 > 5$ppb) and #3 (±6ppb white noise added after each filtering step if $O_3 > 5$ppb). EnKF #2 shows oscillations, while EnKF #3 produces good quality results.

$\gamma_{-}$ and $\gamma_{+}$ are estimated by trial and error. Typical values found in the meteorological literature (Anderson, 2001) are small ($1.01 \leq \gamma \leq 1.2$). Values in this range did not bring any noticeable improvement to the analysis in our tests. Therefore we have implemented an adaptive scheme to determine the magnitude of the apriori ($\gamma_{-}$) and aposteriori ($\gamma_{+}$) inflation factors. We estimate the variance of the observed species.
(O₃), and balanced it against the observation variance, while allowing the ensemble variance to have “reasonable” values (> 1%). Upper bounds are imposed on the choice of γ⁺ to prevent over-inflation.

Another choice for the apriori inflation factor can be based on Kalman filtering theory which requires that the ensemble and innovation spreads be of similar magnitude (Evensen, 2003):

\[
c^f = c^t + \eta, \quad y = Hc^t + \varepsilon, \quad d = y - Hc^f, \\
\langle dd^T \rangle \approx \langle H\eta\eta^T H^T \rangle + \langle \varepsilon\varepsilon^T \rangle = HP^FH^T + \mathbb{R}.
\]

Here η are the model errors. In order to balance the ensemble and innovation spreads, a multiplicative inflation factor can be approximated by:

\[
\gamma_- = \gamma_v = \max \left( \frac{\text{trace} \left( \langle dd^T \rangle - \mathbb{R} \right)}{\text{trace} \left( HP^FH^T \right)}, 1 \right),
\]

where the trace of the covariance matrix is used to approximate the average covariance, and \(\langle \cdot \rangle\) represents the ensemble average. Henceforth, the specific choice of \(\gamma_v\) in (9) will be referred to as adaptive multiplicative adjustment.

Multiplicative inflation results are shown in Table 1 (experiments EnKF #4 to EnKF #7). For each example we present the upper bound of the inflation factors (the lower bound is always 1). In all examples (EnKF #4 to EnKF #7) the multiplicative inflation leads to a better \(R^2\) (RMS) and agreement of model predictions and data than the noiseless EnKF application (EnKF #1). The combination of apriori and aposteriori inflation (EnKF #6) leads to a better agreement with data than either apriori only (EnKF #4) or aposteriori only (EnKF #5) inflation. The best \(R^2\) (RMS) agreement (in analysis and forecast) is obtained with moderate bounds for the inflation factors (\(\gamma_- \leq 4\) and \(\gamma_+ \leq 4\) in EnKF #6). The effects of covariance over-inflation can be noticed for EnKF #7 (\(\gamma_- \leq 10\) and \(\gamma_+ \leq 8\) and EnKF #10a), where the forecast \(R^2\) (RMS) is degraded, albeit the analysis \(R^2\) (RMS) proves to be the largest (smallest).

Figure 4 presents the time series of ozone concentrations at the six selected ground stations. The assimilated ozone series follow the observations much closer than the non-assimilated ones in the analysis window. However, the improvements in the forecast capabilities are modest.

The effects of over-inflating the covariance can be seen in experiments EnKF #12 and EnKF #13 which use 200 member ensembles. The results of EnKF #13 presented in Fig. 5 show that the assimilated results become oscillatory (effects also noticed for EnKF #7, but not shown in this study). The \(R^2\) and RMS agreement between the assimilated solutions and the data is remarkable in the assimilation window, but the forecast skill is deteriorated when compared to the non-assimilated solution. By decreasing too much the confidence in the model the solution is over-constrained by the observations and reflects less and less the model dynamics.

The adaptive multiplicative adjustment defined by (9) and used in scenario EnKF #10a yields a poor fit, especially in the forecast. The poor forecast performance is probably due to spurious remote corrections that get amplified by the additional multiplicative inflation. We shall consider correcting this aspect via localization in the second part of this study.

(c) Model-Specific Inflation

While the additive and multiplicative covariance inflation algorithms are general, we now focus on the sources of uncertainty that are specific to CTMs: boundary
conditions, emissions, and meteorological fields. We account for these uncertainties by perturbing the model parameters (i.e., creating an ensemble of model parameters that mimics the appropriate distribution of parameter space uncertainty). Each ensemble member then runs with a different set of model parameter values. This leads naturally to an increased spread of the ensemble of states, i.e., to covariance inflation. With the notation (6)

$$c_i^f(e) = M(c_{i-1}^a(e), \alpha_{i-1}^U(e) u_{i-1}, \alpha_{i-1}^{BC}(e) c_{i-1}^{in}, \alpha_{i-1}^{EM}(e) q_{i-1}), \quad e = 1, \cdots, E,$$

where $\alpha_{U,BC,EM}(e) \in \mathcal{N}(1, \sigma_{U,BC,EM})$ are random perturbation factors of the model parameters.

This approach is well grounded in our intuition - the main sources of uncertainty in CTMs are treated explicitly. The uncertainties are propagated through the tangent linear model of the forward model, and hence, the state subspace spanned by the ensemble is
consistent with model dynamics and the state errors are correlated according to model dynamics. This approach was also used to construct the background error covariance (Constantinescu et al., 2006a). The violations of the positivity constraint arising from the additive and multiplicative inflation procedures are avoided.

The numerical results for model-specific covariance inflation are presented in Table 1 (examples EnKF #8 and #9). In both examples the boundary conditions and emissions are added normal random perturbations with a standard deviation equal to 10% of their nominal values, i.e., $\alpha^{BC,EM} \sim N(1, 0.1)$. The perturbation of the wind fields is 3% in example EnKF #8 ($\alpha^U \sim N(1, 0.03)$) and 10% in example EnKF #9 ($\alpha^U \sim N(1, 0.1)$).

The model-observations agreement of the results of experiment EnKF #8 are similar to those of experiment EnKF #6 (which uses multiplicative inflation), but the model-specific inflation is easier to implement and its solution is in better agreement with model dynamics. A comparison between the results of EnKF #6 and #8 at the selected

Figure 5. Ozone concentrations measured at the selected stations and predicted by EnKF #11 (no inflation) and #13 (multiplicative inflation $\gamma_- \leq 10, \gamma_+ \leq 8$) both with 200 members. EnKF #11 has no inflation and diverges, while EnKF #13 is overinflated and becomes unstable, at least in the forecast window.
ground stations is shown in Fig. 4 and confirms that model-specific inflation lead to similar performance as the multiplicative inflation.

Further inflation through the wind fields leads to a degradation of both the analysis and forecast results, as seen in example EnKF #9 in Table 1. The experiment EnKF #10 represents a hybrid strategy where both model-specific and multiplicative inflation yield good results.

Note that a more sophisticated method to account for model errors (and consequently inflate the ensemble covariance and avoid filter divergence) is to use a multi-model ensemble (McKeen et al., 2005). Another approach to prevent filter divergence is to prevent the ensemble inbreeding (Houtekamer and Mitchell, 2001) by breaking the filter into two parts that cross act on the other’s input. These approaches are not discussed in this paper.

### 6. Validation of the Filter and Assimilation Results

The filter validation procedure involves the comparison between the ensemble and innovation spreads (Evensen, 2003). After inflation, the model errors are better predicted. However, even when ensemble covariance inflation is used, the model error is under-predicted by a factor ranging from three to ten in the case of model-specific inflation, and five for the multiplicative inflation with $\gamma \leq 4$. We show (in example EnKF#10a) that excessive inflation leads to a degradation of the results, especially in forecast mode, and possible biases. In Table 2 we list the ensemble and innovation covariance estimates. These results come in support of the inflation techniques described in this paper: the disagreement between the ensemble and innovation spreads is reduced through covariance inflation. However, even in scenario EnKF#10a where the agreement is forced through adaptive multiplicative inflation, the filter and the model (mainly through the chemical processes) damp the ensemble variation to nearly half of the innovation spread. This leads to forced apriori multiplicative inflation that also amplifies the spurious corrections and possibly introduces biases.

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. The rest of 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 (Fig. 1.b). The ozonesondes were launched at 14 GMT (S1) and at 22 GMT (S2) July 20th. The NOAA

<table>
<thead>
<tr>
<th>Time [GMT]</th>
<th>EnKF #1</th>
<th>EnKF #6</th>
<th>EnKF #8</th>
<th>EnKF #10a</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\text{HPH}^T$ $\text{dd}^T - \text{R}$</td>
<td>$\text{HPH}^T$ $\text{dd}^T - \text{R}$</td>
<td>$\text{HPH}^T$ $\text{dd}^T - \text{R}$</td>
<td>$\text{HPH}^T$ $\text{dd}^T - \text{R}$</td>
</tr>
<tr>
<td>2</td>
<td>0.58</td>
<td>13.36</td>
<td>2.26</td>
<td>29.04</td>
</tr>
<tr>
<td>3</td>
<td>0.36</td>
<td>14.26</td>
<td>2.32</td>
<td>29.04</td>
</tr>
<tr>
<td>4</td>
<td>0.26</td>
<td>15.87</td>
<td>2.27</td>
<td>29.04</td>
</tr>
<tr>
<td>5</td>
<td>0.22</td>
<td>16.58</td>
<td>2.45</td>
<td>13.36</td>
</tr>
<tr>
<td>6</td>
<td>0.17</td>
<td>18.13</td>
<td>2.09</td>
<td>13.36</td>
</tr>
<tr>
<td>7</td>
<td>0.15</td>
<td>19.07</td>
<td>2.20</td>
<td>13.38</td>
</tr>
<tr>
<td>8</td>
<td>0.18</td>
<td>24.22</td>
<td>3.55</td>
<td>13.36</td>
</tr>
<tr>
<td>9</td>
<td>0.16</td>
<td>25.48</td>
<td>2.20</td>
<td>14.26</td>
</tr>
<tr>
<td>10</td>
<td>0.14</td>
<td>23.29</td>
<td>2.11</td>
<td>12.88</td>
</tr>
<tr>
<td>11</td>
<td>0.14</td>
<td>21.59</td>
<td>2.75</td>
<td>13.48</td>
</tr>
</tbody>
</table>

**TABLE 2.** Trace in [ppbv] scaled by the number of observations of the covariance matrices from the validation relation (8). The covariance estimation is shown for noisel ess filter application (EnKF #1), multiplicative inflation (EnKF #6), and model-specific inflation (EnKF #8, #10a) for the first ten hours of the assimilation window.
P3-B plane was flown between 14–22 GMT along the trajectory shown in Fig. 1.b at different altitudes (corresponding to grid vertical levels 3–16 in our model).

Figure 6 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 assimilating data with 4D-Var, EnKF #2 (additive inflation), EnKF #6 (multiplicative inflation), and EnKF #8 (model-specific inflation).

The EnKF solutions are very close to observations near the observation sites (on or close to the ground level where the solution is constrained). At higher altitudes, however, the assimilated ozone fields are very different for different assimilation methods. For additive (EnKF #2) and multiplicative (EnKF #6) inflation tests the vertical ozone profile is oscillatory, with the peaks taking unreasonable values. The vertical profiles obtained with the model-specific inflation (EnKF #8) have reasonable values. The 4D-Var profiles are close to observations and close to the EnKF solution near the observation sites. At high altitudes the 4D-var profiles come closer to the non-assimilated solution and show no oscillations.

The oscillatory behavior of the EnKF solutions at higher levels is likely due to spurious correlations between these levels and the ground. Spurious long-range correlations imply that the ensemble is strongly correcting the ozone in the upper levels in response to model-observations mismatch at the ground level. The spurious correlations are due to the limited size of the ensemble. They are the strongest for the multiplicative inflation experiment (where all the correlations, including the spurious ones, are increased every cycle) and very mild for the model-specific inflation (which better captures the real correlations). To alleviate the spurious correlations inherent with limited size ensembles one should consider techniques to explicitly localize the correlations (Houtekamer and Mitchell, 2001; Ott et al., 2004). This approach forces the correction that each observation site exerts on the concentration field to decrease with the distance from the observation site. Limiting the spatial influence in EnKF will be considered in future work.

The ozone concentrations measured during the P3-B plane, not shown in this paper, allow us to draw conclusions that closely parallel those of the ozonesondes. EnKF data assimilation with additive and 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 solution obtained with the model-specific inflation as well as the 4D-Var solution 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 in the assimilation measurements of the vertical ozone profiles as well.

7. DISCUSSION

This paper presents a comparison between “perturbed observations” EnKF and state-of-the-art variational data assimilation (4D-Var) applied to the assimilation of real observations into an atmospheric photochemical and transport model. Our previous study (Con07) considered an idealized setting for data assimilation and showed a very promising performance of EnKF. The experiments discussed in this paper reveal the difficulties and challenges of assimilating real data.

Experiments showed that the filter diverges quickly (after about 12 hours of assimilation) with both 50 and 200 member ensembles under the perfect model assumptions. 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 the chemistry components in STEM model) are stable - small perturbations are damped out quickly in time since fast transients are quickly “attracted” to a (slow) low dimensional manifold. Without simulating the atmospheric dynamics (meteorological fields are prescribed) this stiff effects are important.

In order to prevent filter divergence the spread of the ensemble needs to be explicitly increased. We investigated three different approaches to ensemble covariance inflation: additive, multiplicative, and model-specific. Additive inflation reduces the relative magnitude of the off-diagonal correlations and limits the potential of the subsequent analysis. Multiplicative inflation allows for a very good agreement of model predictions and data in the assimilation window, but amplifies spurious correlations inherent with small-sized ensembles, and greatly deteriorates the concentration fields away from the observations sites. Model-specific covariance inflation was obtained by perturbing the meteorological fields, emissions, and lateral boundary conditions. The agreement of model predictions and observations is similar to the one obtained with the multiplicative inflation. However, model-specific covariance inflation does not overly-amplify spurious correlations and seems to be the best choice for chemical and transport modeling.

Experimental results show that 4D-Var and EnKF (without over-constraining the solution) produce similar quality results. By inflating the covariance we were able
to better constrain the EnKF solution near the ground level, and obtain a very good match of model predictions and observations in the assimilation window. This, however, sharply deteriorates the analysis quality at high levels (away from the observations). In our validation results, 4D-Var does not produce spurious corrections far from the observation sites. In 4D-Var, as expected, the analysis effects are smaller away from the observation sites. To obtain similar results with EnKF one needs to consider limiting the correlation distances explicitly, using ideas similar to the localized EnKF (Ott et al., 2004). This approach is considered in the second part of this study.

The numerical experiments in the idealized setting (Con07) used vertically distributed observations. The numerical experiments with real data (this paper) used only ground level observations, and the validation results show that the improvements in the vertical profiles are small (even with 4D-Var). It is likely that information on the vertical distribution of the concentration fields is very important 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 the discussion should encompass the correction of other chemical fields as well.

ACKNOWLEDGEMENT

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.

ACKNOWLEDGEMENTS

REFERENCES


M. Corazza, E. Kalnay, and D. Patil. 2002 Use of the breeding technique to estimate the shape of the analysis "errors of the day". *J. Geophys. Res.*, 10, 233–243


ICARTT. 2004 ICARTT home page:http://www.al.noaa.gov/ICARTT


E. Kalnay, H. Li, T. Miyoshi, S.C. Yang, and J. Ballabrera-Poy. 2005 4D-Var or ensemble Kalman filter? *Submitted to Tellus*


S. McKeen, et. al. 2005 Assessment of an ensemble of seven real-time ozone forecasts over eastern North America during the summer of 2004. *J. Geophys. Res. - Atmos.*, 110(D21307), 16


