Q. J. R. Meteorol. Soc. (2006), **128**, pp. 1–999 doi: 10.1256/qj.yy.n

Ensemble-based Chemical Data Assimilation I: General Approach

By Emil M. Constantinescu[1], Adrian Sandu[1] *, Tianfeng Chai[2], and Gregory R. Carmichael[2]
[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, investigate several ways to avoid filter divergence, and investigate the assimilation of emissions. 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 (Liao et al., 2005; Sandu et al., 2003, 2005; Sandu, 2005) 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. (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 (Lorenc, 2003), Hamill (Hamill, 2004), and Kalnay (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 asses EnKF's practical merits.

The goal of this paper is to investigate the application of the ensemble Kalman filer (EnKF) to atmospheric chemical data assimilation. Considerable progress has been

* Corresponding author: 660 McBryde Hall, Virginia Tech, Blacksburg, VA 24061, E-mail: sandu@cs.vt.edu

made recently in the development of variational tools for chemical data assimilation (Liao et al., 2005; Sandu et al., 2003, 2005; Sandu, 2005). However, little work has been done to date to assimilate chemical observations using nonlinear ensemble filters.

In the a previus study (Constantinescu et al., 2006b), 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" without and with data assimilation. 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 state is beneficial for both the analysis and the forecast.

We now continue the analysis of EnKF for atmospheric chemical data assimilation and consider 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 observational data. The main contributions of this work are: (1) a discussion of several methods to inflate the ensemble covariance and avoid filter divergence, and (2) a comparison between "perturbed observations" EnKF and the state-of-the-art 4D-Var in an operational-like setting using real data.

The paper is structured as follows. Sections (*a*) and (*b*) briefly review the 4D-Var and EnKF methods, respectively. Section 3 describes the chemical transport model and the 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. The assimilation of emissions together the states is discussed in 7. Conclusions and future research directions are given in Section 8.

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 (Constantinescu et al., 2006b).

Consider a nonlinear model $\mathbf{c}_i = \mathcal{M}_{t_0 \to t_i}(\mathbf{c}_0)$ that advances the state from the initial time t_0 to future times t_i ($i \geqslant 1$). The model simulates the evolution of a real system (e.g., the polluted atmosphere). The model state $\mathbf{c}_i$ at t_i ($i \geqslant 0$) is an approximation of "true" state of the system $\mathbf{c}_i^t$ at t_i (more exactly $\mathbf{c}_i^t$ 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 $\mathbf{c}^{\mathrm{B}}$ (the "background" state, or the best initial guess) and its covariance matrix $\mathbb{B}$. Observations $\mathbf{y}_i$ 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 $\mathbb{R}_i$)

$$\mathbf{y}_i = \mathcal{H}_i\left(\mathbf{c}_i^t\right) + \varepsilon_i \ .$$

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 both the information from the model ($\mathbf{c}_i$, $i \geqslant 0$) and from the observations ($\mathbf{y}_i$, $i \geqslant 0$).

(*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 $\mathbf{y}_0 \cdots \mathbf{y}_n$) is obtained as the minimizer of the following cost function (which measures the model-observations misfit)

$$\mathcal{J}(c_0) = \frac{1}{2}\big(\mathbf{c}_0 - \mathbf{c}^{\mathrm{B}}\big)^T \mathbb{B}^{-1}\big(\mathbf{c}_0 - \mathbf{c}^{\mathrm{B}}\big) + \frac{1}{2}\sum_{i=0}^{n}\big(\mathbf{y}_i - \mathcal{H}_i(\mathbf{c}_i)\big)^T \mathbb{R}_i^{-1}\big(\mathbf{y}_i - \mathcal{H}_i(\mathbf{c}_i)\big). \tag{1}$$

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

$$\nabla_{\mathbf{c}_0}\mathcal{J} = \mathbb{B}^{-1}\ \big(\mathbf{c}_0 - \mathbf{c}^{\mathrm{B}}\big) + \sum_{i=0}^{n}\mathbf{M}^*_{t_i\to t_0}\mathbf{H}_i^T\ \mathbb{R}_i^{-1}\ \big(\mathbf{y}_i - \mathcal{H}_i(\mathbf{c}_i)\big), \tag{2}$$

where $\mathbf{M} = \mathcal{M}'$ is the tangent linear model associated with $\mathcal{M}$, $\mathbf{M}^*$ is the adjoint of $\mathbf{M}$, and $\mathbf{H} = \mathcal{H}'$ 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 $\mathbf{c}_i^t$ at t_i using the information from the current best estimate $\mathbf{c}_i^f$ (the "forecast" or the background state) and the observations $\mathbf{y}_i$. The optimal estimate $\mathbf{c}_i^a$ (the "analysis" state) is obtained as a linear combination of the forecast and observations that minimize the variance of the analysis ($\mathbf{P}^a$)

$$\mathbf{c}_i^a = \mathbf{c}_i^f + \mathbf{P}_i^f\mathbf{H}_i^T\left(\mathbf{H}_i\mathbf{P}_i^f\mathbf{H}_i^T + \mathbb{R}_i\right)^{-1}\left(\mathbf{y}_i - \mathcal{H}_i(\mathbf{c}_i^f)\right) = \mathbf{c}_i^f + \mathbf{K}_i\left(\mathbf{y}_i - \mathcal{H}_i(\mathbf{c}_i^f)\right). \tag{3}$$

The forecast covariance $\mathbf{P}^f$ is estimated from an ensemble of runs (which produces an ensemble of E model states $\mathbf{c}_i^f(e)$, $e = 1, \cdots, \mathrm{E}$). The analysis formula (3) 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:

$$\mathbf{c}_i^f(e) = \mathcal{M}\big(\mathbf{c}_{i-1}^a(e)\big),\ \mathbf{c}_i^a(e) = \mathbf{c}_i^f(e) + \mathbf{K}_i\Big(\mathbf{y}_i - \mathcal{H}_i\big(\mathbf{c}_i^f(e)\big)\Big),\ e = 1, \cdots, \mathrm{E}. \tag{4}$$

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 setting.

(*a*) *The Model*

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).

The model can be written compactly as

$$\mathbf{c}_i = \mathcal{M}\left(\mathbf{c}_{i-1}, \mathbf{u}_{i-1}, \mathbf{c}^{\text{in}}_{i-1}, \mathbf{q}_{i-1}\right). \tag{5}$$

where $\mathbf{c}$ is the vector of concentrations (all species at all gridpoints), $\mathbf{q}$ is the rate of surface emissions, $\mathbf{u}$ is the wind field, and $\mathbf{c}^{\text{in}}$ 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 our previous study (Constantinescu et al., 2006b) and in (Sandu et al., 2005).

A strong constraint for this model is the requirement that $\mathbf{c}_s$ be positive. STEM numerical methods cannot produce or handle negative species concentrations, since they are not physically possible. This fact plays an important role in both variational and sequential data assimilation.

(*b*) *The Case Study*

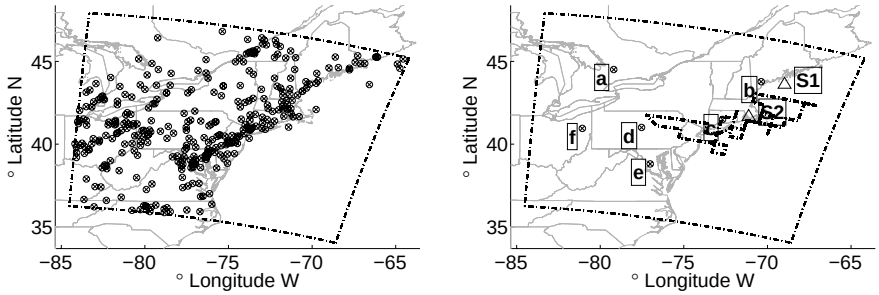
The test case is a real-life simulation of air pollution in North–Eastern United States in July 2004 as shown in Fig. 1 (the dash-dotted line delimits the computational domain). The observations used for data assimilation are the ground-level ozone (O_3) measurements taken during the ICARTT (International 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 60×60 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.

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 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.



(a) ICARTT ground stations (b) selected stations

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

In both EnKF and 4D-Var we consider background errors modeled by autoregressive (AR) processes of the form

$$\mathbf{A}\ \delta\mathbf{c}^{\mathrm{B}} = \mathbf{S}\ \xi\ , \quad \mathbf{S} = \mathrm{diag}(\sigma_{i,j,k})\ , \tag{6}$$

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

The ensemble initialization using AR perturbations was described in detail in the previous paper (Constantinescu et al., 2006b). The AR model is particularly advantageous in the 4D-Var context where the inverse of the AR background covariance

$$\mathbb{B}^{-1} = \mathbf{A}^T\ \mathbf{S}^{-2}\ \mathbf{A},$$

allows for an elegant formulation of the background term in the cost function (1)

$$\mathbf{z} = \mathbf{S}^{-1}\ \mathbf{A}\ \left(\mathbf{c} - \mathbf{c}^{\mathrm{B}}\right) \quad \Rightarrow \quad \frac{1}{2}\ \left(\mathbf{c}_0 - \mathbf{c}^{\mathrm{B}}\right)^T\ \mathbb{B}^{-1}\ \left(\mathbf{c}_0 - \mathbf{c}^{\mathrm{B}}\right) = \frac{1}{2}\ \mathbf{z}^T\ \mathbf{z}\ .$$

This formulation only requires one matrix-vector multiplication by the AR coefficient matrix $\mathbf{A}$, and one component-wise scaling (multiplication by the diagonal matrix $\mathbf{S}^{-1}$). A more detailed discussion can be found in (Constantinescu et al., 2006a; Liao et al., 2005).

(*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 20^{th}) 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 will denote the initialization window as [-4,0] hours. 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 results section. The best guess evolved to 4 GMT July 20^{th} 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 20^{th}.

The 24 hours assimilation window starts at 4 GMT July 20^{th} and ends at 4 GMT July 21^{st} (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).

EnKF adjusts the concentration fields of 66 "control" chemical species in each grid point of the domain every hour using (3). The ensemble size was chosen to be 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. The 200 member runs were 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 20^{th}). The L-BFGS iterations are stopped when the cost function is reduced to less than 10^{-3} of its initial value ($\mathcal{J} = 10^{-3}\mathcal{J}_0$), or when the number of iterations exceeds 25.

The 24 hours forecast window starts at 4 GMT July 21^{st} and ends at 4 GMT July 22^{nd} (the forecast windows will be denoted further as [24,48] hours). The model is initialized at 4 GMT July 22^{nd} 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.

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

$$\mathrm{R}^2(\mathbf{x},\mathbf{y}) = \frac{\left(n\sum_{i=1}^{n}\mathbf{x}_i\,\mathbf{y}_i - \sum_{i=1}^{n}\mathbf{x}_i\sum_{i=1}^{n}\mathbf{y}_i\right)^2}{\left(n\sum_{i=1}^{n}\mathbf{x}_i^2 - \left(\sum_{i=1}^{n}\mathbf{x}_i\right)^2\right)\left(n\sum_{i=1}^{n}\mathbf{y}_i^2 - \left(\sum_{i=1}^{n}\mathbf{y}_i\right)^2\right)}. \qquad (7)$$

In our experimental setting the deterministic (best guess) solution yields an R^2 of 0.24 in the analysis and 0.28 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 (Lorenc, 2003) and expanded by Kalnay (Kalnay 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). 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 "textbook application" of EnKF using 50 and 200 member ensembles. Table 1 shows the R^2 correlation results between the observations and model values for all state assimilated numerical experiments. For each scenario the ensemble size, setting information, and R^2 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 the 200 member ensemble are presented as EnKF experiment #11.

The correlation factor between model and observations in the assimilation window is $\mathrm{R}^2 = 0.24$ for the non-assimilated run. It grows to $\mathrm{R}^2 = 0.40$ for the solution assimilated with the 50 member ensemble and to $\mathrm{R}^2 = 0.49$ 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 ($\mathrm{R}^2 = 0.52$). 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).

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

ID	Method	Details	R^2 analysis	R^2 forecast
-	Deterministic	Best guess solution, no assimilation	0.24	0.28
-	4D-Var	50 iterations w/ AR background	0.52	0.29
1	EnKF(50)	"textbook application"	0.38	0.30
2	EnKF(50)	additive inflation: $\mathcal{N}(0,\ 6$ppb) white noise added *before* filtering if O_3 >5ppb	0.60	0.30
3	EnKF(50)	additive inflation: $\mathcal{N}\ (0,\ 6$ppb) white noise added *after* filtering if O_3 >5ppb	0.71	0.30
4	EnKF(50)	multiplicative inflation: $\gamma_- \leqslant 4, \gamma_+ = 1$	0.61	0.30
5	EnKF(50)	multiplicative inflation: $\gamma_- = 1, \gamma_+ \leqslant 4$	0.61	0.29
6	EnKF(50)	multiplicative inflation: $\gamma_- \leqslant 4, \gamma_+ \leqslant 4$	0.62	0.32
7	EnKF(50)	multiplicative inflation: $\gamma_- \leqslant 10, \gamma_+ \leqslant 8$	0.63	0.31
8	EnKF(50)	model-specific inflation: 10% emissions, 10% boundaries, 3% wind	0.58	0.32
9	EnKF(50)	model-specific inflation: 10% emissions, 10% boundaries, 10% wind	0.59	0.30
10	EnKF(50)	combined inflation: $\gamma_- \leqslant 10$, $\gamma_+ \leqslant 4$, 10% emissions, 10% boundaries, 5% wind	0.72	0.33
11	EnKF(200)	"textbook application"	0.49	0.30
12	EnKF(200)	multiplicative inflation: $\gamma_- \leqslant 4, \gamma_+ \leqslant 2$	0.82	0.28
13	EnKF(200)	multiplicative inflation: $\gamma_- \leqslant 10, \gamma_+ \leqslant 8$	0.85	0.23

TABLE 1. The R^2 measure of model-observations match in the assimilation and forecast windows for the EnKF (with different ensemble sizes) and 4D-Var data assimilation.

We will now look at several ways to prevent the filter divergence by inflating the ensemble covariance.

5. PREVENTING FILTER DIVERGENCE

The previous section shows that the "textbook application" of EnKF (Evensen, 2003) 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). 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).

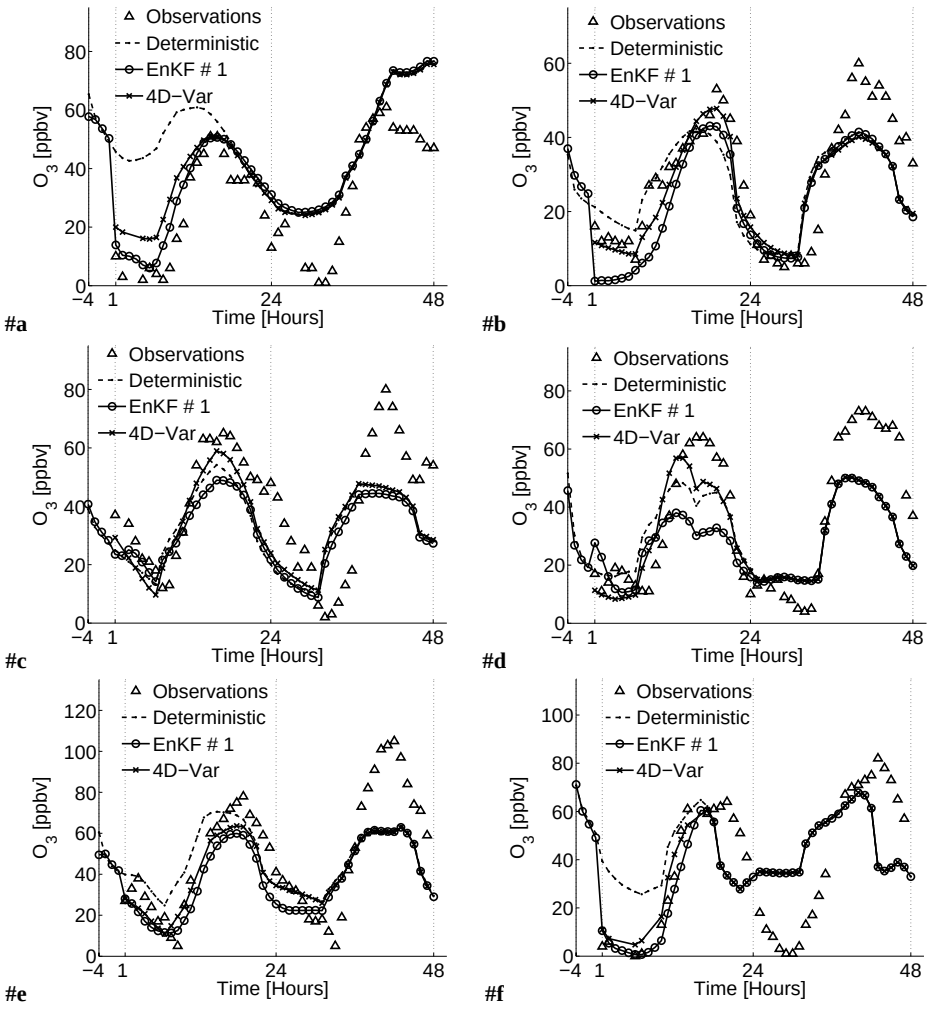


Figure 2. Ozone concentrations measured at the selected stations and predicted by EnKF#1 (50 members, "textbook application") and 4D-Var (50 iterations). The overall measure shows comparable results, and in the case of EnKF it shows clearly that the filter divergences after some time.

(*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 $\eta \in \mathcal{N}(0,\, \mathcal{Q})$ of mean zero and covariance matrix $\mathcal{Q}$.

The most intuitive way is to add noise to the forecast solution. The net result is to increment the forecast covariance by $\mathcal{Q}$. With the notation (4)

$$\mathbf{c}_i^f\,(e) = \mathcal{M}\left(\mathbf{c}_{i-1}^a(e)\right) + \eta(e)\;, \quad e = 1, \cdots, \mathrm{E}\;, \quad \Rightarrow \quad \mathbf{P}_{\mathrm{i}}^{\mathrm{f}} \leftarrow \mathbf{P}_{\mathrm{i}}^{\mathrm{f}} + \mathcal{Q}\;.$$

In the ideal situation $\mathcal{Q}$ should reflect the correlation of the model errors. Since these are very much unknown one typically chooses white noise, i.e. the covariance matrix $\mathcal{Q}$ is diagonal (η 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 $\mathbf{c}_i^f$). 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)

$$\mathbf{c}_i^f(e) = \mathcal{M}\left(\mathbf{c}_{i-1}^a(e) + \eta(e)\right) , \quad e = 1, \cdots, \mathrm{E} .$$

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 textbook 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 γ^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

$$\mathbf{c}_i^f(e) \leftarrow \left\langle \mathbf{c}_i^f \right\rangle + \gamma_- \left(\mathbf{c}_i^f(e) - \left\langle \mathbf{c}_i^f \right\rangle\right) , \quad e = 1, \cdots, \mathrm{E} \quad \Rightarrow \quad \mathbf{P}^{\mathrm{f}} \leftarrow \gamma_-^2 \, \mathbf{P}^{\mathrm{f}} ,$$

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

$$\mathbf{c}_i^a(e) \leftarrow \left\langle \mathbf{c}_i^a \right\rangle + \gamma_+ \left(\mathbf{c}_i^a(e) - \left\langle \mathbf{c}_i^a \right\rangle\right) , \quad e = 1, \cdots, \mathrm{E} \quad \Rightarrow \quad \mathbf{P}^{\mathrm{a}} \leftarrow \gamma_+^2 \, \mathbf{P}^{\mathrm{a}} .$$

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

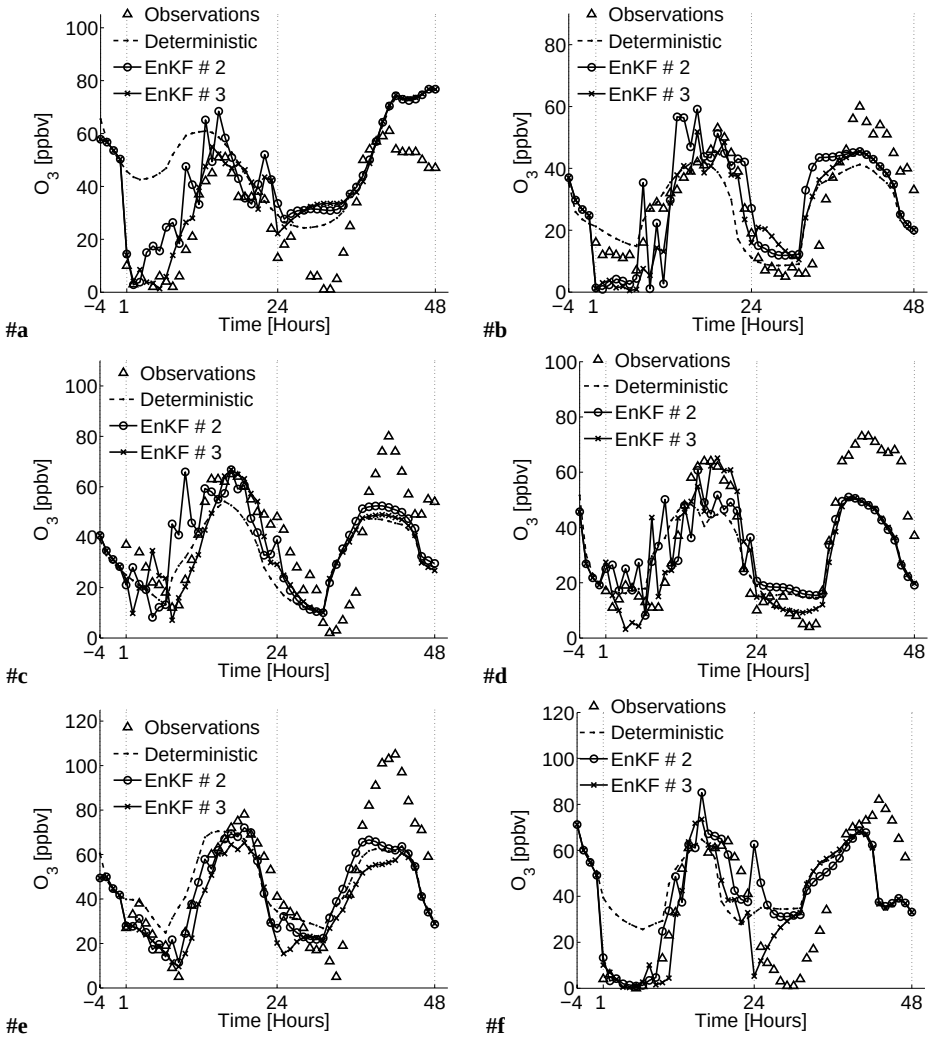


Figure 3. Ozone concentrations measured at the selected stations and predicted by EnKF #2 (± 6ppb white noise added before each filtering step if O_3 $>$5ppb) and #3 (± 6ppb white noise added after each filtering step if O_3 $>$5ppb). EnKF #2 shows oscillations, while EnKF #3 produces good quality results.

inflation factors are usually estimated by trial and error. Typical values found in the meteorological literature (, Anderson 2001) are small ($1.01 \leqslant \gamma \leqslant 1.2$). 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 (γ_-) and aposteriori (γ_+) inflation factors. We estimate the variance of the observed species (O_3), 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 $\gamma_\pm$ to prevent over-inflation.

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 agreement of model predictions and data than the textbook EnKF

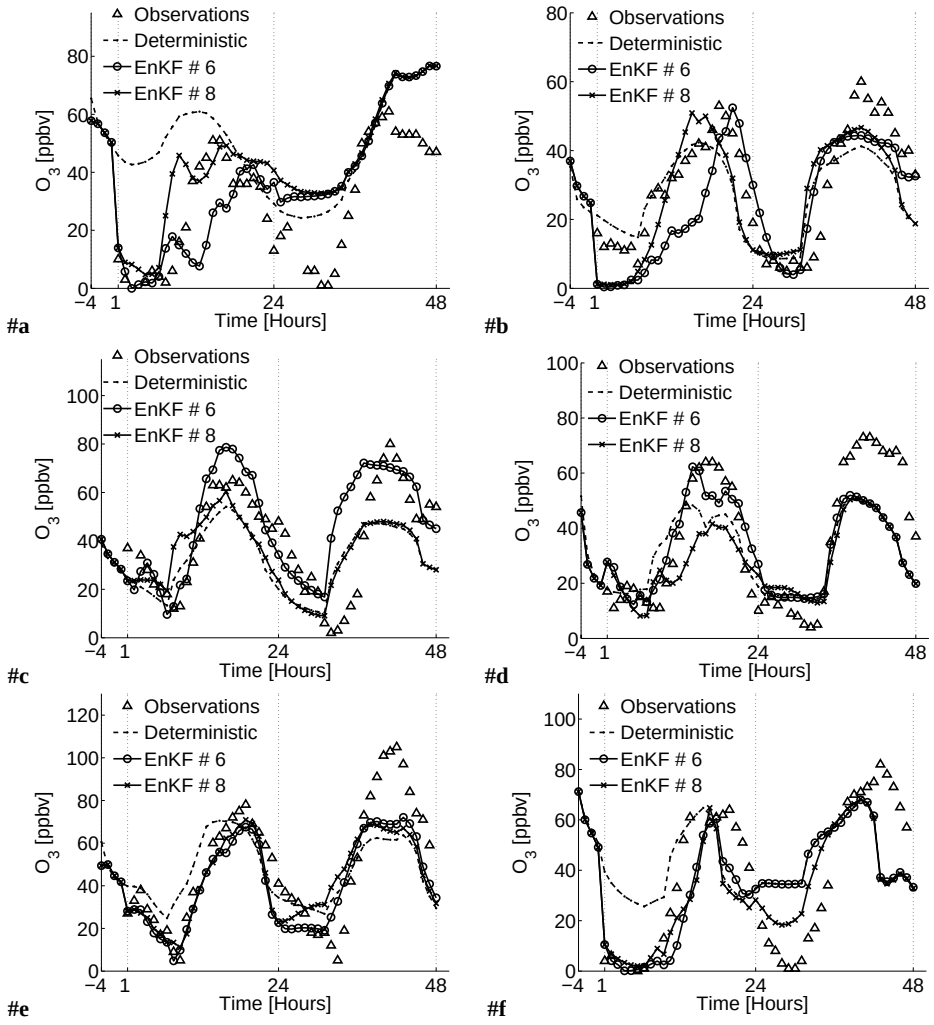


Figure 4. Ozone concentrations measured at the selected stations and predicted by EnKF #6 (multiplicative inflation with $\gamma_- \leqslant 4$ and $\gamma_+ \leqslant 4$) and EnKF #8 (10% Em + Lat, 3% Wind). EnKF #6 uses multiplicative inflation, while EnKF #8 uses model parameter inflation, they both produce same quality results overall.

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) inflations. The best R^2 agreement (in analysis and forecast) is obtained with moderate bounds for the inflation factors ($\gamma_- \leqslant 4$ and $\gamma_+ \leqslant 4$ in EnKF #6). The effects of covariance over-inflation can be noticed for EnKF #7 ($\gamma_- \leqslant 10$ and $\gamma_+ \leqslant 8$), where the forecast R^2 is degraded, albeit the analysis R^2 proves to be the largest.

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

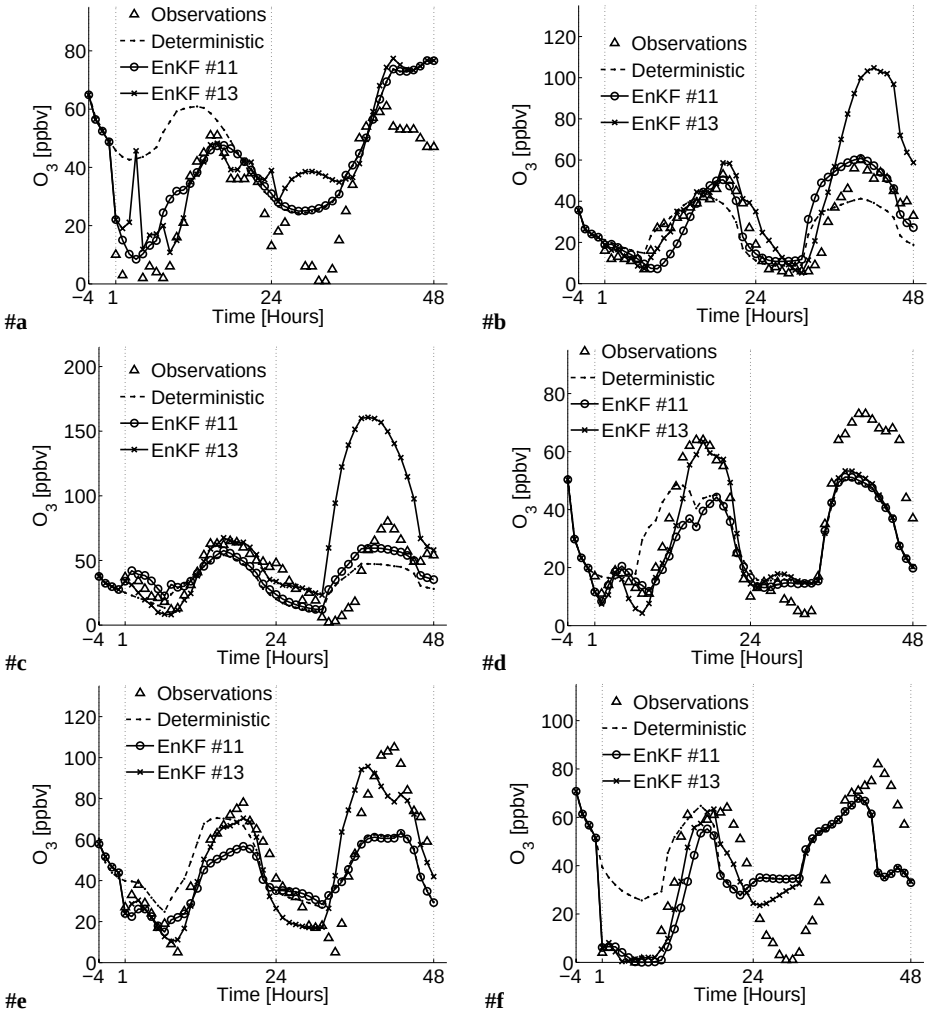


Figure 5. Ozone concentrations measured at the selected stations and predicted by EnKF #11 (no inflation) and #13 (multiplicative inflation 1:10;1:8) both with 200 members. EnKF #11 has no inflation and diverges, while EnKF #13 is overinflated and becomes unstable, at least in the forecast window.

EnKF #7, but not shown in this study). The R^2 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 overconstrained by the observations and reflects less and less the model dynamics.

(*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 (5)

$$\mathbf{c}_i^f(e) = \mathcal{M}\left(\mathbf{c}_{i-1}^a(e),\, \alpha_{i-1}^{\mathrm{U}}(e)\,\mathbf{u}_{i-1},\, \alpha_{i-1}^{\mathrm{BC}}(e)\,\mathbf{c}_{i-1}^{\mathrm{in}},\, \alpha_{i-1}^{\mathrm{EM}}(e)\,\mathbf{q}_{i-1}\right), \quad e = 1, \cdots, \mathrm{E},$$

where $\alpha^{\mathrm{U,BC,EM}}(e) \in \mathcal{N}(1,\, \sigma^{\mathrm{U,BC,EM}})$ 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. Moreover the state subspace spanned by the ensemble is consistent with model dynamics and the state errors are correlated according to model dynamics. 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., $\alpha^{\mathrm{BC,EM}} \in \mathcal{N}(1,\, 0.1)$. The perturbation of the wind fields is 3% in example EnKF #8 ($\alpha^{\mathrm{U}} \in \mathcal{N}(1,\, 0.03)$) and 10% in example EnKF #9 ($\alpha^{\mathrm{U}} \in \mathcal{N}(1,\, 0.1)$).

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 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 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 (Fig. 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 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,

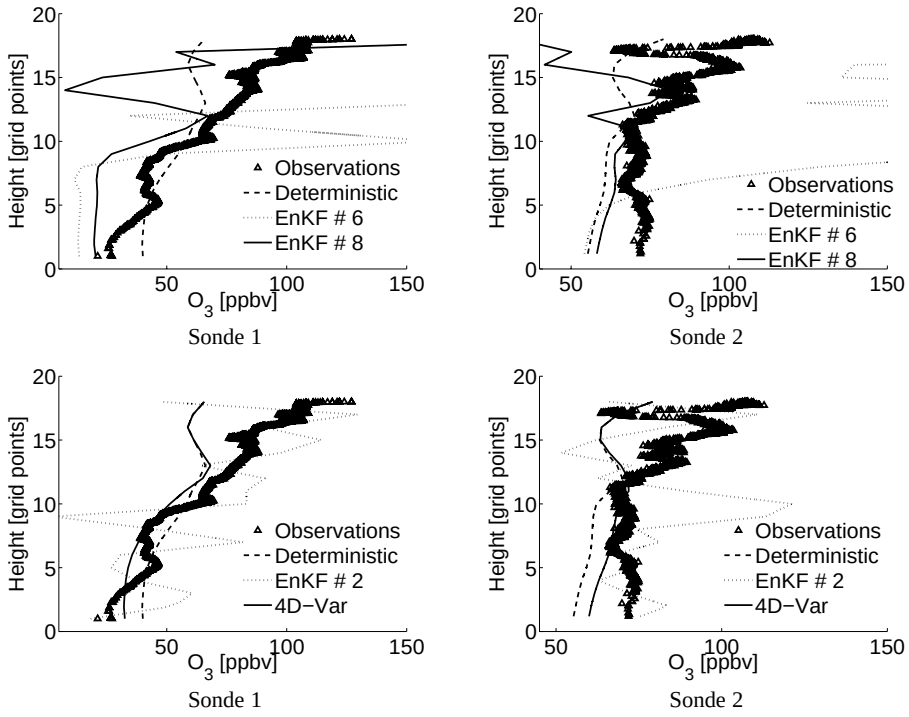


Figure 6. Ozone concentrations measured by ozonesondes and predicted by the model after data assimilation with 4D-Var, EnKF #2 (additive inflation), EnKF #6 (multiplicative inflation with $\gamma_- \leqslant 4$ and $\gamma_+ \leqslant 4$), and EnKF #8 (model-specific inflation with 10% perturbations on the emissions and boundary conditions and 3% perturbations of the wind).

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., 2002). 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. ASSIMILATION OF EMISSIONS

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 and by lateral boundary conditions. Since both emissions and lateral boundaries are in general 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 this section we discuss the state-parameter assimilation in the case where real observations are available.

In the numerical experiments we follow the approach discussed in our previous study (Constantinescu et al., 2006b). The emission rates 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 (5)

$$\begin{bmatrix} \mathbf{c}_i^f \\ \alpha_i^{\mathrm{EM}} \end{bmatrix} = \begin{bmatrix} \mathcal{M}_{t_{i-1}\to t_i}\left(\mathbf{c}_{i-1}^a, \mathbf{u}_{i-1}, \mathbf{c}_{i-1}^{\mathrm{in}}, \left(1+\alpha_{i-1}^{\mathrm{EM}}\right)\ \mathbf{q}_{i-1}\right) \\ \alpha_{i-1}^{\mathrm{EM}} \end{bmatrix}.$$

We considered two cases for the parameters α. In the first, one scalar correction parameter is considered per chemical species for all ground-level gridpoints (coarse resolution). In the second a different correction parameter is considered for each species and each ground-level gridpoint (finest resolution). In practice one may consider intermediate resolutions, e.g., one correction factor per species per geographic area. The initial ensemble of correction factors is an independent set of normal variables $\alpha_0 \in \mathcal{N}(0, 0.3)$, that produces a perturbation of $\pm 100\%$ of the initial value.

Table 2 shows the model-observations agreement (R^2) after EnKF data assimilation for only the state and for the combined state and emission correction coefficients. We only consider multiplicative inflation scenarios. Table 2 presents the results obtained with one correction factor per gridpoint; the results with one correction factor for all gridpoints are similar and are not presented. The results show no improvement of either the forecast or the analysis when emissions are assimilated. It is likely that due to the small ensemble size spurious (stronger-than-real) correlations are being developed between emissions and ozone concentrations. The filter tries to compensate the model-observations mismatch by over-adjusting the emission rates. This de facto introduces model errors which have a negative impact on the quality of the analysis and forecast. While in the idealized setting the assimilation of emission was beneficial, in the real case under consideration it degrades the assimilated solution. Considerably more research is needed to understand the use of EnKF data assimilation to correct for emissions in chemical transport models.

Method	Details	R^2 analysis		R^2 forecast	
		State	Em.	State	Em.
Deterministic	–	0.24	0.24	0.28	0.28
EnKF(50)	"textbook application"	0.40	0.38	0.30	0.31
EnKF(50)	multiplicative inflation: $\gamma_- \leqslant 4$, $\gamma_+ \leqslant 4$	0.62	0.62	0.32	0.31
EnKF(50)	multiplicative inflation: $\gamma_- \leqslant 10$, $\gamma_+ \leqslant 8$	0.63	0.65	0.35	0.26

TABLE 2. Model-observations agreement for the EnKF data assimilation of only the state and of the combined state and emissions parameters (Em). No visible improvements in either the analysis or the forecast are obtained by adjusting the emissions.

8. CONCLUSIONS AND FUTURE WORK

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 (Constantinescu et al., 2006b) 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. 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 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., 2002).

Since the solution of a regional CTM is largely influenced by uncertain lateral boundary conditions and by 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 has visibly improved the quality of the analysis. In the real case under consideration assimilation of emissions does not improve the analysis and degrades the forecast solution. The correct assimilation of states and emissions using EnKF is a challenging problem, and considerably more research is needed to fully understand it.

The numerical experiments in the idealized setting (Constantinescu et al., 2006b) 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.

Several research directions emerge from the analysis carried out in this paper. The localized versions of the ensemble filter need to be studied to alleviate spurious corrections of the chemical fields far away from the observation sites. Considerable work is required to use ensemble data assimilation to reduce uncertainties in emission inventories. Finally, to fully understand the ensemble data assimilation one needs to fully understand the capability of the ensemble to capture correlations due to chemical interactions. Specifically, one needs to consider observations of several different chemical species, and to asses the impact of these observations to the correction of other chemical fields. In this paper we have considered the "perturbed observations" version of EnKF. The performance of the "square root" EnKF variants will need to be assessed. On the longer term we would like to develop hybrid methods that combine the advantages of the 4D-Var and EnKF data assimilation approaches. We plan to pursue these questions in the near future.

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

J.L. Anderson. 2001 An Ensemble Adjustment Kalman Filter for Data Assimilation. *Mon. Weather Rev.*, **129,** 2884–2903

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

J. Barkmeijer, R. Buizza, and T.N. Palmer 1999 3D-Var Hessian singular vectors and their potential use in the ECMWF Ensemble Prediction System. *Q. J. R. Meteorol. Soc.*, **125,** 2333–2351

R. Buizza, J. Barkmeijer, T.N. Palmer, and D.S. Richardson. 2000 Current status and future developments of the ECMWF ensemble prediction system. *Meteorol. Appl.*, **7,** 163–175

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

R. Byrd, P. Lu, and J. Nocedal. 1995 A limited memory algorithm for bound constrained optimization. *SIAM J Sci. Stat. Comp.*, **16 (5),** 1190–1208

G.R. Carmichael, et. al. 2003 Regional-scale Chemical Transport Modeling in Support of the Analysis of Observations obtained During the Trace-P Experiment. *J. Geophys. Res.*, **108(D21 8823),** 10649–10671

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

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

E.M. Constantinescu, A. Sandu, T. Chai, and G.R. Carmichael. 2006b Ensemble-based chemical data assimilation: An idealized setting. *SUBMITTED to Atmos. Environ.*

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

P. Courtier, J.-N. Thepaut, and A. Hollingsworth. 1994 A strategy for operational implementation of 4D-Var, using an incremental approach. *Q. J. R. Meteorol. Soc.*, **120,** 1367–1387

J. Derber. 1989 A variational continuous assimilation scheme. *Mon. Weather Rev.*, **117,** 2437–2446

G. Evensen. 1994 Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forcast error statistics . *J. Geophys. Res.*, **99(C5),** 10143–10162

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

G. Evensen. 2005 The combined parameter and state estimation problem. *SUBMITTED to Ocean Dyn.*

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

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

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

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

P.L. Houtekamer, H.L. Mitchell, G. Pellerin, M. Buehner, M. Charron, L. Spacek, and B. Hansen. 2005 Atmospheric data assimilation with the ensemble Kalman filter: Results with real observations. *Mon. Weather Rev.*, **133(3),** 604–620

B.R. Hunt, E. Kalnay, E. Kostelich, E. Ott, D. Patil, T. Sauer, I. Szunyogh, J. Yorke, and A. Zimin.	2004	Four-dimensional ensemble Kalman filtering. *Tellus A,* **56,** 273–277
ICARTT.	2004	ICARTT home page:http://www.al.noaa.gov/ICARTT
R.E. Kalman.	1960	A new approach to linear filtering and prediction problems . *Transaction of the ASME - J. of Basic Eng.,* **82,** 35–45
E. Kalnay, H. Li, T. Miyoshi, S.C. Yang, and J. Ballabrera-Poy.	2005	4D-Var or ensemble Kalman filter. *PHYSICA D,* SUBMITTED
F.X. Le Dimet and O. Talagrand.	1986	Variational algorithms for analysis and assimilation of meteorological observations. *Tellus,* **38 A,** 97–110
W. Liao, A. Sandu, G.R. Carmichael, and T. Chai.	2005	Total energy singular vector analysis with atmospheric chemical transport models. *SUBMITTED*
A.C. Lorenc.	1986	Analysis methods for numerical weather prediction. *Q. J. R. Meteorol. Soc.,* **112,** 1177–1194
A.C. Lorenc.	2003	The potential of the ensemble Kalman filter for NWP – a comparison with 4D-Var. *Q. J. R. Meteorol. Soc.,* **129(595),** 3183–3203
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
F. Molteni, R. Buizza, T.N. Palmer, and T. Petroliagis.	1996	The new ECMWF ensemble prediction system: methodology and validation. *Q. J. R. Meteorol. Soc.,* **122,** 73–119
E. Ott, B. R. Hunt, I. Szunyogh, A. V. Zimin, E. J. Kostelich, M. Corazza, E. Kalnay, D. J. Patil, and J. A. Yorke.	2002	A local ensemble Kalman filter for atmospheric data assimilation. *ArXiv Physics e-prints*
F. Rabier, H. Jarvinen, E. Klinker, J.F. Mahfouf, and A. Simmons.	2000	The ECMWF operational implementation of four-dimensional variational assimilation. I: Experimental results with simplified physics. *Q. J. R. Meteorol. Soc.,* **126,** 1148–1170
A. Sandu.	2005	Discrete Runge-Kutta adjoints. *International Conference on Computational Sciences (ICCS-2006), Reading, UK, ACCEPTED*
A. Sandu, D. Daescu, and G.R. Carmichael.	2003	Direct and adjoint sensitivity analysis of chemical kinetic systems with kpp: I – theory and software tools. *Atmos. Environ.,* **37,** 5083–5096
A. Sandu, D. Daescu, G.R. Carmichael, and T. Chai.	2005	Adjoint sensitivity analysis of regional air quality models. *J. Comput. Phys.,* **204,** 222–252
O. Talagrand and P. Courtier.	1987	Variational assimilation of meteorological observations with the adjoint vorticity equation. Part I: Theory. *Q. J. R. Meteorol. Soc.,* **113,** 1311–1328
Y. Tang, et. al.	2006	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 J. Geophys. Res.*