QUARTERLY JOURNAL OF THE ROYAL METEOROLOGICAL SOCIETY
Q. J. R. Meteorol. Soc. **133**: 1229–1243 (2007)
Published online 9 July 2007 in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/qj.76

RMetS
ROYAL METEOROLOGICAL SOCIETY

Ensemble-based chemical data assimilation. I: General approach

Emil M. Constantinescu,[a] Adrian Sandu,[a]* Tianfeng Chai[b] and Gregory R. Carmichael[b]
[a] *Department of Computer Science, Virginia Polytechnic Institute and State University, Blacksburg, VA 24061, USA*
[b] *Center for Global and Regional Environmental Research, University of Iowa, Iowa City, IA 52240, USA*

ABSTRACT: Data assimilation is the process of integrating observational data and model predictions to obtain an optimal representation of the state of the atmosphere. As more chemical observations in the troposphere are becoming available, chemical data assimilation is expected to play an essential role in air-quality forecasting, similar to the role it has in numerical weather prediction. Considerable progress has been made recently in the development of variational tools for chemical data assimilation. In this paper we assess the performance of the ensemble Kalman filter (EnKF) and compare it with a state-of-the-art 4D-Var approach. We analyse 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. Copyright © 2007 Royal Meteorological Society

KEY WORDS data assimilation; ensemble Kalman filter; 4D-Var; chemical transport models

Received 29 March 2006; Revised 19 January 2007; Accepted 26 February 2007

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 and climate analysis and forecast activities, and is accomplished by means of a mature infrastructure. Both variational (Rabier *et al*., 2000) 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 (NWP).

Variational techniques for data assimilation are well established in NWP. Building on the early variational approach (Lorenc, 1986; Le Dimet and Talagrand, 1986; Talagrand and Courtier, 1987), the 4D-Var framework represents the current state of the art in meteorological (Courtier *et al*., 1994; Rabier *et al*., 2000) and chemical (Elbern *et al*., 2000; Elbern and Schmidt, 2001; Liao *et al*., 2005; Sandu *et al*., 2003; Sandu *et al*., 2005a; Sandu, 2006; Segers, 2002) data assimilation. Ensemble Kalman filter (EnKF) data assimilation (Evensen, 1994, 2003; Burgers *et al*., 1998) has recently attracted considerable interest in the context of NWP. With this approach, the cost of applying the Kalman filter (Kalman, 1960) to 'large' models becomes reasonable through the use of 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 in EnKF matching 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, although 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 EnKF to atmospheric chemical data assimilation. Considerable progress has been made recently in the development of variational tools for chemical data assimilation (Liao *et al*., 2005; Sandu *et al*., 2003; Sandu *et al*., 2005a; Sandu, 2006). However, compared with NWP and ocean applications, little work has been done to date on assimilating chemical observations using nonlinear ensemble filters. EnKF, the extended Kalman filter (van Loon and Heemink, 1997) and the 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 chemical-transport models (CTMs), and to improve the quality of

* Correspondence to: Adrian Sandu, Department of Computer Science, Virginia Polytechnic Institute and State University, Blacksburg, VA 24061, USA. E-mail: sandu@cs.vt.edu

the forecasts. A comparison between different flavours of the reduced Kalman filter is given in (Heemink and Segers, 2002). We study EnKF applied to a large number (66) of species rather than to only 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*., 2007a), we analysed 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 analysed how close this solution was 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 in 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:

- a comparison between 'perturbed-observations' EnKF and state-of-the-art 4D-Var in an operational-like setting using real data;
- an analysis of several methods for inflating the ensemble covariance and avoiding filter divergence;
- an examination of the localized EnKF in an operational-like setting; and
- a state-parameter inversion approach for our model.

Furthermore, the ensemble initialization is obtained through a novel approach based on autoregressive processes (Constantinescu *et al*., 2007b). A novel approach to localizing 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 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 inversion of the emissions and boundary conditions in a regional atmospheric CTM model.

This paper is structured as follows. Sections 2.1 and 2.2 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 presented and discussed in Section 4. Several strategies for inflating the ensemble covariance and avoiding filter divergence are addressed in Section 5. A validation of the data assimilation results is carried out in Section 6. Discussion and a summary are presented in Section 7.

2. Data assimilation

In this section we briefly review the 4D-Var and EnKF approaches to data assimilation. More details on 4D-Var can be found in (Courtier *et al*., 1994; Rabier *et al*., 2000), and on the EnKF in our previous study (Constantinescu *et al*., 2007a).

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 ($i \geq 1$). The model simulates the evolution of a real system (e.g. the polluted atmosphere). The model state $\mathbf{c}_i$ at t_i ($i \geq 0$) is an approximation to the 'true' state of the system $\mathbf{c}_i^{\mathrm{t}}$ at t_i. (More precisely, $\mathbf{c}_i^{\mathrm{t}}$ is the system state projected onto the model 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 $\mathbf{B}$. Observations $\mathbf{y}_i$ of the real system are available at times t_i, and are corrupted by measurement and representativity errors ε_i (assumed Gaussian with mean zero and covariance $\mathbf{R}_i$):

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

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

The data-assimilation problem is to find an optimal estimate of the state using the information both from the model ($\mathbf{c}_i$, $i \geq 0$) and from the observations ($\mathbf{y}_i$, $i \geq 0$).

2.1. 4D-Var

In 4D-Var (Courtier *et al*., 1994; Rabier *et al*., 2000), the best estimate of the initial state (conditioned by the observations $\mathbf{y}_0, \ldots, \mathbf{y}_n$) is obtained as the minimizer of the following cost function (which measures the model–observations misfit):

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

DOI: 10.1002/qj

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:

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

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.*, 2005a).

2.2. EnKF

The Kalman filter estimates the true state $\mathbf{c}_i^{\mathrm{t}}$ at t_i using the information from the current best estimate $\mathbf{c}_i^{\mathrm{f}}$ (the 'forecast' or the background state) and the observations $\mathbf{y}_i$. The optimal estimate $\mathbf{c}_i^{\mathrm{a}}$ (the 'analysis' state) is obtained as a linear combination of the forecast and observations that minimizes the variance of the analysis $\mathbf{P}^{\mathrm{a}}$:

$$\begin{aligned}\mathbf{c}_i^{\mathrm{a}} &= \mathbf{c}_i^{\mathrm{f}} \\ &\quad + \mathbf{P}_i^{\mathrm{f}}\mathbf{H}_i^{\mathrm{T}}\left(\mathbf{H}_i\mathbf{P}_i^{\mathrm{f}}\mathbf{H}_i^{\mathrm{T}} + \mathbf{R}_i\right)^{-1}\left(\mathbf{y}_i - \mathcal{H}_i\left(\mathbf{c}_i^{\mathrm{f}}\right)\right) \\ &= \mathbf{c}_i^{\mathrm{f}} + \mathbf{K}_i\left(\mathbf{y}_i - \mathcal{H}_i\left(\mathbf{c}_i^{\mathrm{f}}\right)\right). \end{aligned} \quad (3)$$

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

$$\left.\begin{aligned}\mathbf{c}_i^{\mathrm{f}}(e) &= \mathcal{M}\left(\mathbf{c}_{i-1}^{\mathrm{a}}(e)\right) \\ \mathbf{c}_i^{\mathrm{a}}(e) &= \mathbf{c}_i^{\mathrm{f}}(e) + \mathbf{K}_i\left(\mathbf{y}_i - \mathcal{H}_i\left(\mathbf{c}_i^{\mathrm{f}}(e)\right)\right)\end{aligned}\right\}, \quad (4)$$

for $e = 1, \ldots, E$. The results presented in this paper are obtained with the practical EnKF implementation discussed by Evensen (2003).

3. Experimental setting

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

3.1. 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.*, 2005a).

In STEM, the evolution of N_{spec} species is described by the following equations:

$$\frac{\partial \mathbf{c}_s}{\partial t} = -\mathbf{u}\nabla\mathbf{c}_s + \frac{1}{\rho}\nabla\left(\rho\overline{\mathbf{K}}\nabla\mathbf{c}_s\right) + \frac{1}{\rho}\mathbf{c}_s(\rho c) + \mathbf{E}_s, \quad (5)$$

for $t^0 \le t \le t^{\mathrm{F}}$ and $1 \le s \le N_{\mathrm{spec}}$, and

$$\left.\begin{aligned}\mathbf{c}_s(t^0, x) &= \mathbf{c}_s^0(x) & \\ \mathbf{c}_s(t, x) &= \mathbf{c}_s^{\mathrm{in}}(t, x) & \text{for } x \in \Gamma^{\mathrm{in}} \\ \overline{\mathbf{K}}\frac{\partial \mathbf{c}_s}{\partial n} &= 0 & \text{for } x \in \Gamma^{\mathrm{out}} \\ \overline{\mathbf{K}}\frac{\partial \mathbf{c}_s}{\partial n} &= \mathbf{V}_s^{\mathrm{dep}}\mathbf{c}_s - \mathbf{q}_s & \text{for } x \in \Gamma^{\mathrm{ground}}\end{aligned}\right\},$$

where the concentration $\mathbf{c}_s$ for species s is dictated by the rate of surface emissions $\mathbf{q}_s$, the rate of elevated emissions $\mathbf{E}_s$, and the rate of chemical transformation $\mathbf{c}_s$, for this species. Here $\mathbf{u}$ denotes the wind-field vector, $\overline{\mathbf{K}}$ the turbulent-diffusivity tensor, ρ the air density, $\mathbf{c}^{\mathrm{in}}$ the Dirichlet boundary conditions, and $\mathbf{V}_s^{\mathrm{dep}}$ the deposition velocity.

The model can be written compactly as:

$$\mathbf{c}_i = \mathcal{M}\left(\mathbf{c}_{i-1}, \mathbf{u}_{i-1}, \mathbf{c}_{i-1}^{\mathrm{in}}, \mathbf{q}_{i-1}\right), \quad (6)$$

where $\mathbf{c}$ 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 and the air density) that are not explicitly represented here.

In regional air-quality simulations, the influence of the initial conditions fades over 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 over time. Moreover, stiff systems like chemistry are very stable: small perturbations are damped out quickly as the system tends to evolve towards a 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 multi-scale 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 three-dimensional fields. Correlations between chemical

DOI: 10.1002/qj

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 $\mathbf{c}_i$ be positive (negative concentrations are unphysical); therefore the error statistics associated with small concentrations are non-Gaussian. 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 constrain the solution of the optimization process to be positive. In the ensemble case, this problem can be alleviated by shifting the negative values to zero; however, biases may be introduced.

3.2. The case study

The test case is a real-life simulation of air pollution in the northeastern United States in July 2004, as shown in Figure 1 (where the dashed-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') campaign (ICARTT, 2004) in summer 2004 (5, 19 and 21 July). A detailed description of the ICARTT fields and data can be found in (Tang *et al*., 2007). 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 three-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 on 20 July 2004.

We have selected six stations throughout the domain in order to plot the time evolution of measured and modelled ozone concentrations and illustrate the effects of different data-assimilation scenarios. The selected stations are shown in Figure 1(b), and correspond to the following ICARTT codes:

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 (centre of the continental domain);
e 510590030 (in Washington, DC);
f 391514005 (inflow boundary).

Our study also includes three validation measurements taken by two ozone-sondes and a P3-B flight (all shown in Figure 1(b)).

3.3. Modelling 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 with zero mean (the model is unbiased) and covariance **B**. In EnKF the background covariance is used to generate the initial ensemble, while in 4D-Var the background covariance is used explicitly in the formulation of the cost function. A good approximation to the background error statistics is therefore essential for the success of both ensemble-based and variational data assimilation.

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

$$\mathbf{A} \cdot \delta \mathbf{c}^{\mathrm{B}} = \mathbf{S} \cdot \xi, \tag{7}$$

where

$$\mathbf{S} = \mathrm{diag}(\sigma_{i,j,k}),$$

represents the state covariances,

$$\xi(e) \sim \mathcal{N}(0, 1)^N$$

is a vector of N independent standard normal random variables, and **A** is a correlation-coefficient matrix. The AR background accounts for spatial correlations, distance

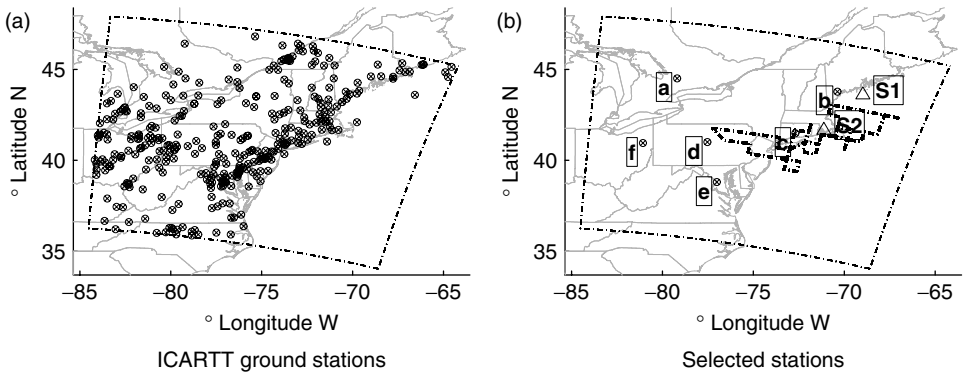

Figure 1. (a) Ground measuring stations supporting of the ICARTT campaign (340 in total). (b) Selected stations ('a' to 'f'), two ozone-sondes (S1 and S2), and the path of a P3-B flight, used to illustrate the numerical results and validation.

DOI: 10.1002/qj

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*., 2007b).

3.4. Analysis setting

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

All the simulations are started at the same time (0 GMT on 20 July) with a four-hour initialization step. This allows the background 4D-Var run and each of the ensemble members to reach a quasi-steady state before the assimilation window. We write the initialization window as '$[-4, 0]$ hours'. The 'best guess' of the state of the atmosphere at 0 GMT on 20 July is obtained from a longer simulation over the entire US, performed in support of the ICARTT experiment (Tang *et al*., 2007). This best guess is used to initialize the deterministic (unassimilated) solution shown in Section 6. The best guess evolved to 4 GMT on 20 July 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.

The 24-hour assimilation window starts at 4 GMT on 20 July and ends at 4 GMT on 21 July (written as '$[0, 24]$ hours'). Observations are available at each integer hour in this window (i.e. at $0, 1, \ldots, 24$ hours). The ozone observations used in this study are from the ICARTT ground stations (Figure 1). Not all the stations provide observations every hour (the number of hourly observations varies between 160 and 326 during the assimilation window). The observation error covariance (in both EnKF and 4D-Var), $\mathbf{R}$, is considered to be a diagonal matrix with a standard deviation of 0.25 ppbv from the measurement.

EnKF adjusts the concentration fields of 66 'control' chemical species in each grid point of the domain every hour using Equation (3). Two ensemble sizes are considered, with 50 and 200 members. Ensembles of 50 members are typical in NWP, 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 (Constantinescu *et al*., 2007a), a 50-member ensemble shows significant improvements against smaller ensembles. Furthermore, 50-member ensembles give numerical results that are computationally affordable given the computational resources available for the experiments. 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 on 20 July). The L-BFGS-B iterations are stopped when the cost function falls below 10^{-3} of its initial value ($\mathcal{J} = 10^{-3}\mathcal{J}_0$), or when the number of iterations exceeds 25, in order to maintain a computational workload comparable to that of the ensemble approach.

The 24-hour forecast window starts at 4 GMT on 21 July and ends at 4 GMT on 22 July (written as '$[24, 48]$ hours'). The model is initialized at 4 GMT on 22 July, 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 presented by the positivity of chemical concentration fields, a constraint inherent to chemical-transport modelling. 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 value between the observations and the model solution (separate R^2 and RMS values are computed in the assimilation and forecast windows). The R^2 correlation and RMS value of two series $\mathbf{x}$ and $\mathbf{y}$ of length n are

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

and

$$\mathrm{RMS}(\mathrm{x,y}) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\mathbf{x}_i - \mathbf{y}_i\right)^2},$$

respectively.

In our experimental setting, the deterministic (best guess) solution yields an R^2 (RMS) of 0.24 (22.1 ppbv) in the analysis and 0.28 (23.5 ppbv) in the forecast window. 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 NWP has been given by Lorenc (2003) and expanded by Kalnay *et al*. (2005). Hamill (2004) undertakes a theoretical analysis of the two approaches. 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 concerning 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

DOI: 10.1002/qj

task in the presence of stiff chemistry (Sandu *et al*., 2005a). 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*., 2007b). 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 affect the computational time needed to advance the model state for each ensemble member. It is not yet clear how the computational cost of EnKF compares with that of 4D-Var (for similar levels of performance). In this context, we consider perfect models for this comparison. To the best of our knowledge, comprehensive tests of EnKF versus 4D-Var have not yet been carried out.

The EnKF forecast can be performed 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 case, 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 perform a 'noiseless' application of EnKF using 50- and 200-member ensembles. Table I shows the R^2 and RMS 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 the 200-member ensemble 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 ppbv) for the non-assimilated run. It grows to $R^2 = 0.40$ (RMS= 23.5 ppbv) for the solution assimilated with the 50-member ensemble, and to $R^2 = 0.49$ (RMS= 16.29 ppbv) 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 ppbv). None of the methods, however, is able considerably to improve the model–observations correlation in the forecast window.

To further understand the behaviour of the filter, we look at the time evolution of ozone concentrations at

Table I. The R^2 and RMS(ppbv) measures of model–observations match in the assimilation and forecast windows for the EnKF (with different ensemble sizes) and 4D-Var data assimilation.

ID	Method	Details	R^2 (RMS) analysis	R^2 (RMS) forecast
–	Deterministic	Best-guess solution, no assimilation	0.24 (22.1)	0.28 (23.5)
–	4D-Var	50 iterations with AR background	0.52 (16.05)	0.29 (22.4)
1	EnKF(50)	'noiseless' application	0.38 (18.18)	0.30 (23.15)
2	EnKF(50)	additive inflation: $\mathcal{N}(0, (6\,\text{ppbv})^2)$ white noise added *before* filtering if $O_3 > 5$ ppbv	0.60 (15.17)	0.30 (23.22)
3	EnKF(50)	additive inflation: $\mathcal{N}(0, (6\,\text{ppbv})^2)$ white noise added *after* filtering if $O_3 > 5$ ppbv	0.71 (11.95)	0.30 (23.18)
4	EnKF(50)	multiplicative inflation: $\gamma_- \leq 4$, $\gamma_+ = 1$	0.61 (14.27)	0.30 (23.13)
5	EnKF(50)	multiplicative inflation: $\gamma_- = 1$, $\gamma_+ \leq 4$	0.61 (14.27)	0.29 (24.31)
6	EnKF(50)	multiplicative inflation: $\gamma_- \leq 4$, $\gamma_+ \leq 4$	0.62 (14.2)	0.32 (24.33)
7	EnKF(50)	multiplicative inflation: $\gamma_- \leq 10$, $\gamma_+ \leq 8$	0.63 (14.16)	0.31 (24.24)
8	EnKF(50)	model-specific inflation: 10% emissions, 10% boundaries, 3% wind	0.58 (14.45)	0.32 (22.63)
9	EnKF(50)	model-specific inflation: 10% emissions, 10% boundaries, 10% wind	0.59 (14.14)	0.30 (23.21)
10	EnKF(50)	combined inflation: $\gamma_- \leq 10$, $\gamma_+ \leq 4$, 10% emissions, 10% boundaries, 5% wind	0.72 (11.51)	0.33 (24.04)
10a	EnKF(50)	multiplicative variance adjustment ($1 \leq \gamma_v \leq 3.9$) and model-specific inflation: 30% emissions, 30% boundaries, 20% wind	0.67 (12.7)	0.19 (62.06)
11	EnKF(200)	'noiseless' application	0.49 (16.29)	0.30 (23.74)
12	EnKF(200)	multiplicative inflation: $\gamma_- \leq 4$, $\gamma_+ \leq 2$	0.82 (9.36)	0.28 (37.63)
13	EnKF(200)	multiplicative inflation: $\gamma_- \leq 10$, $\gamma_+ \leq 8$	0.85 (8.57)	0.23 (39.64)

DOI: 10.1002/qj

the selected ground stations. Figure 2 shows the time series of ozone observations, and the unassimilated EnKF #1 and 4D-Var solutions. After the first 12 hours, the EnKF solution comes very close to the unassimilated one, and 'ignores' further observations. Clearly the filter diverges. Without an effective influence from 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 Figure 5. Increasing the ensemble size to 200 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 to be expected, since we consider a perfect model for the purpose of the comparison between 4D-Var and EnKF.

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

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

5. Preventing filter divergence

The preceding 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

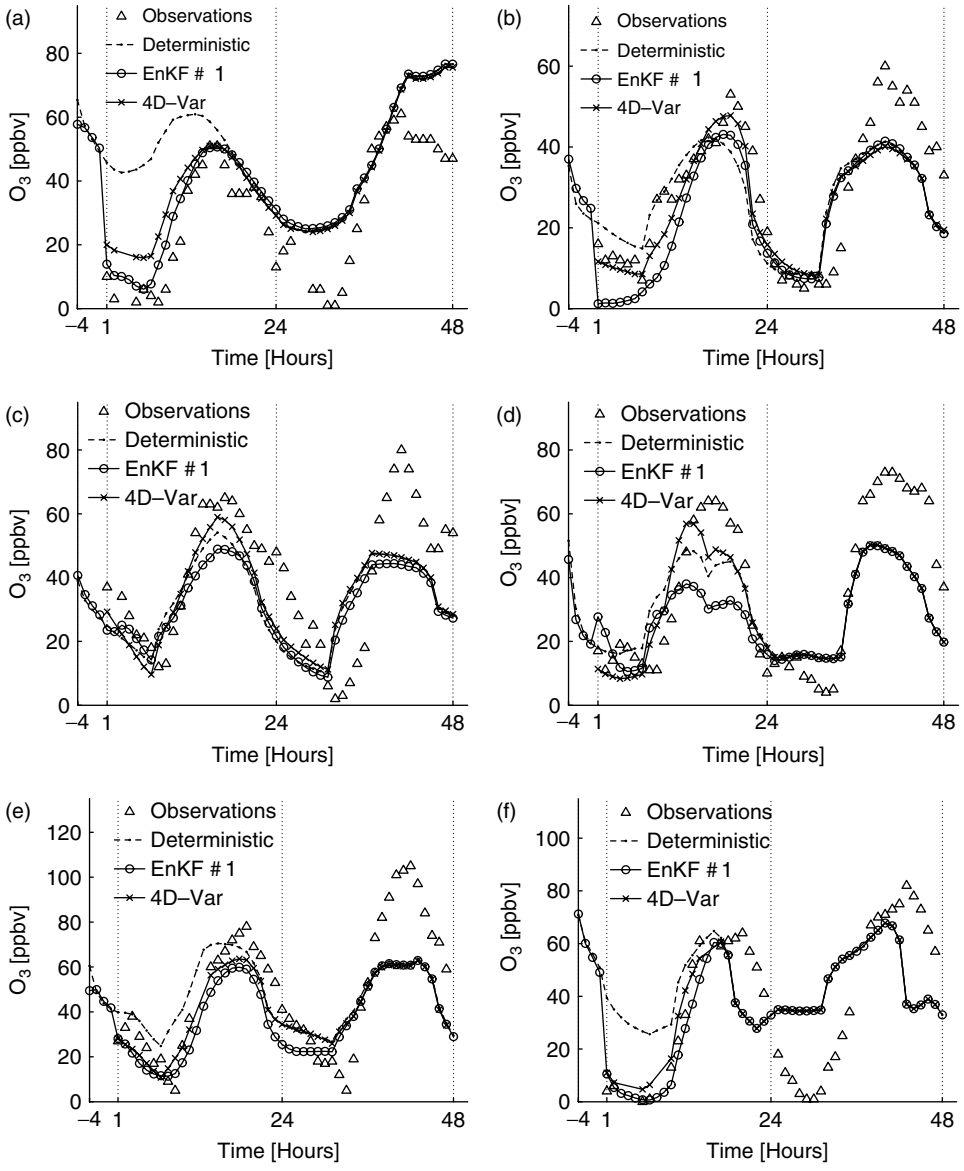

Figure 2. Ozone concentrations measured at the selected stations ('a' to 'f') and predicted by EnKF #1 (50 members, 'noiseless' application) and 4D-Var (50 iterations). The overall measure shows comparable results, and in the case of EnKF it shows clearly that after some time the filter diverges.

DOI: 10.1002/qj

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 thereby 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 *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 matrix. The second method is *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 the square of that constant. Finally, we discuss covariance inflation obtained through perturbing key model parameters: we call this *model-specific inflation*. We note that a better approach can be obtained by constructing multi-model ensembles (McKeen *et al*., 2005).

5.1. Additive inflation

The additive inflation process (Corazza *et al*., 2002) consists in 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 $\eta \sim \mathcal{N}(0, \mathcal{Q})$ 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 of Equation (4),

$$\mathbf{c}_i^{\mathrm{f}}(e) = \mathcal{M}\big(\mathbf{c}_{i-1}^{\mathrm{a}}(e)\big) + \eta_i(e)$$

for $e = 1, \ldots, E$, and

$$\mathbf{P}_i^{\mathrm{f}} \to \mathbf{P}_i^{\mathrm{f}} + \mathcal{Q}.$$

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

Another way is to add the noise immediately 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. To 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 of Equation (4),

$$\mathbf{c}_i^{\mathrm{f}}(e) = \mathcal{M}\big(\mathbf{c}_{i-1}^{\mathrm{a}}(e) + \eta(e)\big)$$

for $e = 1, \ldots, E$. The experiment EnKF #3 presented in Table I adds white noise to the ozone after each assimilation step. The noise has a standard deviation of 6 ppbv, and is added only if the ozone concentration is larger than 5 ppbv, 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. This is reflected in our results (Figure 3). Both experiments 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 number of unstable modes (model states) mean that EnKF #2 performs poorly in the forecast, while EnKF #3 performs better than the noiseless EnKF #1. In our case, EnKF #2 develops oscillations in the solution, as can be seen in Figure 3.

5.2. 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 (diagonal and off-diagonal entries of the covariance matrix are scaled by the same amount).

One can inflate the forecast ensemble before filtering,

$$\mathbf{c}_i^{\mathrm{f}}(e) \to \langle \mathbf{c}_i^{\mathrm{f}} \rangle + \gamma_-\big(\mathbf{c}_i^{\mathrm{f}}(e) - \langle \mathbf{c}_i^{\mathrm{f}} \rangle\big)$$

for $e = 1, \ldots, E$,

$$\mathbf{P}_i^{\mathrm{f}} \to \gamma_-^2 \mathbf{P}_i^{\mathrm{f}}$$

(where $\langle \cdot \rangle$ denotes an ensemble average); or one can inflate the analyzed ensemble after filtering,

$$\mathbf{c}_i^{\mathrm{a}}(e) \to \langle \mathbf{c}_i^{\mathrm{a}} \rangle + \gamma_+\big(\mathbf{c}_i^{\mathrm{a}}(e) - \langle \mathbf{c}_i^{\mathrm{a}} \rangle\big)$$

for $e = 1, \ldots, E$,

$$\mathbf{P}_i^{\mathrm{a}} \to \gamma_+^2 \mathbf{P}_i^{\mathrm{a}}.$$

The inflation of the analysis covariance prepares an ensemble of larger spread for the integration over the

DOI: 10.1002/qj

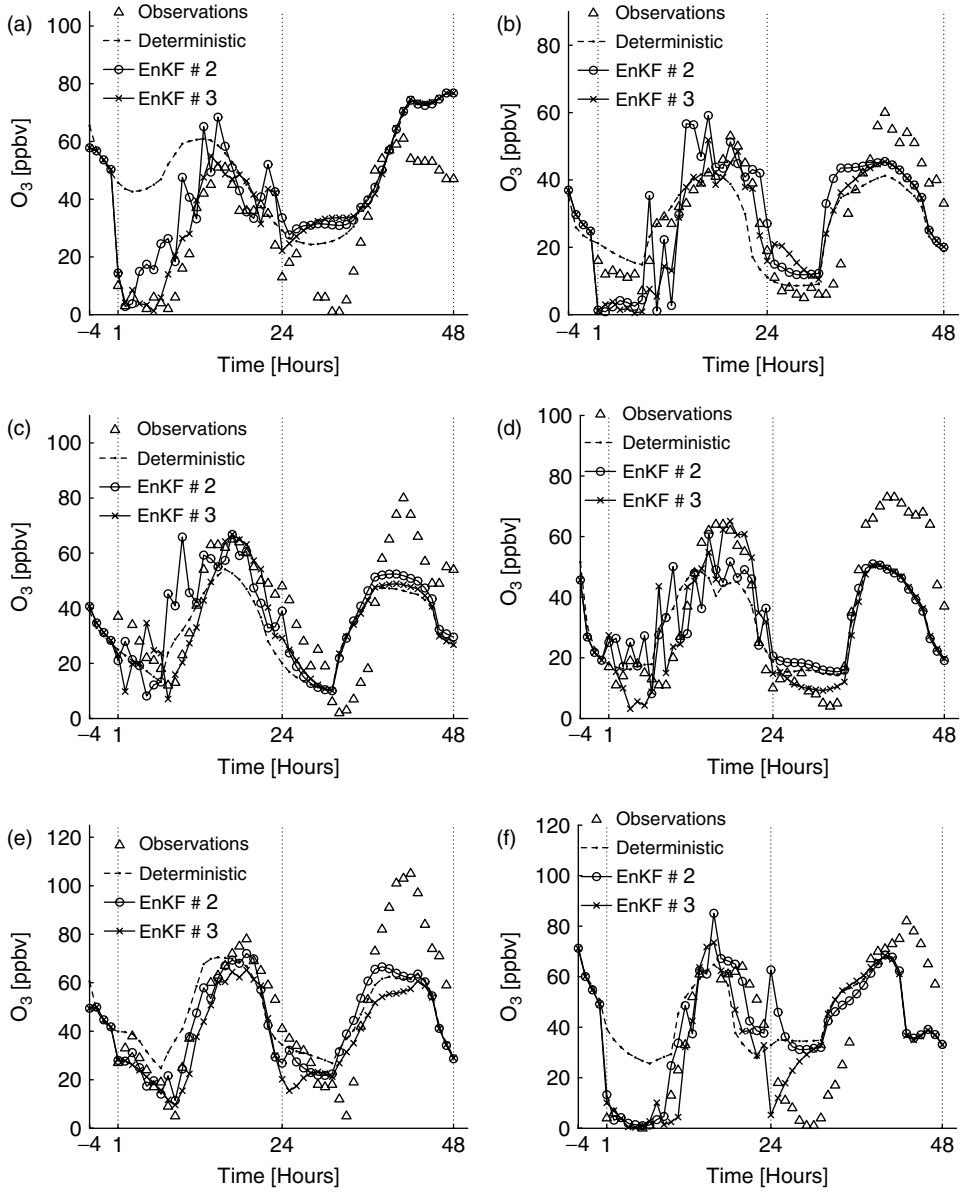

Figure 3. Ozone concentrations measured at the selected stations ('a' to 'f') and predicted by EnKF #2 (±6 ppbv white noise added before each filtering step if $O_3 > 5$ ppbv) and #3 (±6 ppbv white noise added after each filtering step if $O_3 > 5$ ppbv). EnKF #2 shows oscillations, while EnKF #3 produces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

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, and this 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 do not bring any noticeable improvement to the analysis in our tests. Therefore we have implemented an adaptive scheme to determine the magnitude of the a priori (γ_-) and a posteriori (γ_+) inflation factors. We estimate the variance of the observed species (O_3), and balance it against the observation variance, while allowing the ensemble variance to have 'reasonable' values (greater than 1%). Upper bounds are imposed on the choice of $\gamma_{\pm}$ to prevent over-inflation.

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

$$\langle \mathbf{d}\mathbf{d}^{\mathrm{T}} \rangle \approx \langle \mathbf{H}\eta\eta^{\mathrm{T}}\mathbf{H}^{\mathrm{T}} \rangle + \langle \varepsilon\varepsilon^{\mathrm{T}} \rangle = \mathbf{H}\mathbf{P}^{\mathrm{F}}\mathbf{H}^{\mathrm{T}} + \mathbf{R}, \qquad (8)$$

where

$$\mathbf{c}^{\mathrm{f}} = \mathbf{c}^{\mathrm{t}} + \eta,$$

DOI: 10.1002/qj

$$\mathbf{y} = \mathbf{H}\mathbf{c}^{\mathrm{t}} + \varepsilon,$$

and

$$\mathbf{d} = \mathbf{y} - \mathbf{H}\mathbf{c}^{\mathrm{f}}.$$

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

$$\gamma_- = \gamma_{\mathrm{v}} = \max\left\{\frac{\mathrm{tr}\left(\langle \mathbf{d}\mathbf{d}^{\mathrm{T}}\rangle - \mathbf{R}\right)}{\mathrm{tr}\left(\mathbf{H}\mathbf{P}^{\mathrm{F}}\mathbf{H}^{\mathrm{T}}\right)}, 1\right\}, \tag{9}$$

where the trace of the covariance matrix is used to approximate the average covariance, and $\langle\cdot\rangle$ denotes an ensemble average. Henceforth, the specific choice of γ_{v} in Equation (9) will be referred to as 'adaptive multiplicative adjustment'.

Multiplicative inflation results are shown in Table I (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), multiplicative inflation leads to a better R^2 (RMS) and agreement of model predictions and data than the 'noiseless' application (EnKF #1). The combination of a priori and a posteriori inflation (EnKF #6) leads to a better agreement with data than either a-priori-only (EnKF #4) or a-posteriori-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, although the analysis R^2 proves to be very large and RMS very small.

Figure 4 presents the time series of ozone concentrations at the six selected ground stations. The assimilated ozone series follow the observations much more closely than the unassimilated 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 Figure 5 show that the assimilated results become oscillatory (an effect 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 compared to the unassimilated solution. When the confidence in the model is decreased by too much, the solution becomes over-constrained by the observations and reflects the model dynamics less and less well.

The adaptive multiplicative adjustment defined by Equation (9) and used in experiment 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.

5.3. 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 of Equation (6), $\mathbf{c}_i^{\mathrm{f}}(e)$ can be written as

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

for $e = 1, \ldots, E$, where

$$\alpha^{\mathrm{U,BC,EM}}(e) \sim \mathcal{N}\left(1, (\sigma^{\mathrm{U,BC,EM}})^2\right)$$

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 the model dynamics and the state errors are correlated according to model dynamics. This approach is also used to construct the background error covariance (Constantinescu *et al.*, 2007b). 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 I (experiments EnKF #8 and #9). In both experiments, to the boundary conditions and emissions are added normal random perturbations with standard deviations equal to 10% of their nominal values, i.e. $\alpha^{\mathrm{BC,EM}} \sim \mathcal{N}(1, 0.1^2)$. The perturbation of the wind fields is 3% in experiment EnKF #8 ($\alpha^{\mathrm{U}} \sim \mathcal{N}(1, 0.03^2)$) and 10% in example EnKF #9 ($\alpha^{\mathrm{U}} \sim \mathcal{N}(1, 0.1^2)$).

The model–observations agreement of experiment EnKF #8 is similar to that of experiment EnKF #6 (which uses multiplicative inflation), but model-specific inflation is easier to implement and its solution is in better agreement with the model dynamics. Figure 4 compares the results of EnKF #6 and #8 at the selected ground stations, and confirms that model-specific inflation leads to performance similar to that of multiplicative inflation.

Further inflation through the wind fields leads to a degradation of both the analysis and the forecast results, as seen in experiment EnKF #9 in Table I. 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

DOI: 10.1002/qj

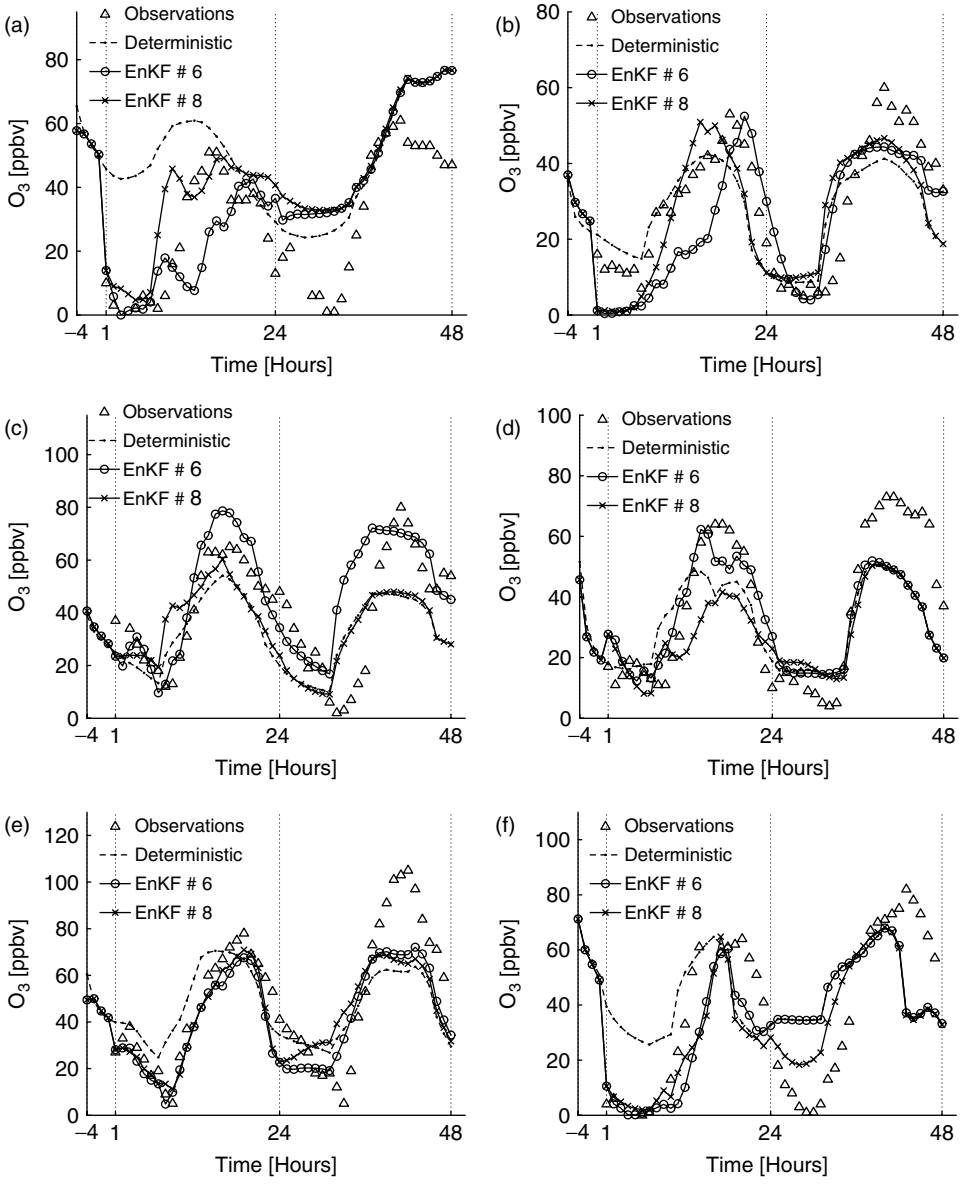

Figure 4. Ozone concentrations measured at the selected stations ('a' to 'f') and predicted by EnKF #6 (multiplicative inflation with $\gamma_- \leq 4$ and $\gamma_+ \leq 4$) and EnKF #8 (10% emissions and boundaries, 3% wind). EnKF #6 uses multiplicative inflation, while EnKF #8 uses model-parameter inflation; they produce results of the same overall quality.

multi-model ensemble (McKeen *et al*., 2005). Another approach to preventing filter divergence is to prevent the ensemble from inbreeding (Houtekamer and Mitchell, 2001) by breaking the filter into two parts each of which acts 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 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 3 to 10 in the case of model-specific inflation, and 5 for multiplicative inflation with $\gamma_\pm \leq 4$. We have shown (in experiment EnKF #10a) that excessive inflation leads to a degradation of the results, especially in forecast mode, and possible biases. In Table II we list the ensemble and innovation covariance estimates. These results 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) dampen the ensemble variation to about half of the innovation spread. This leads to forced a priori multiplicative inflation that also amplifies the spurious corrections and possibly introduces biases.

The data assimilation experiments presented in this paper use only ground ozone observations. While the

DOI: 10.1002/qj

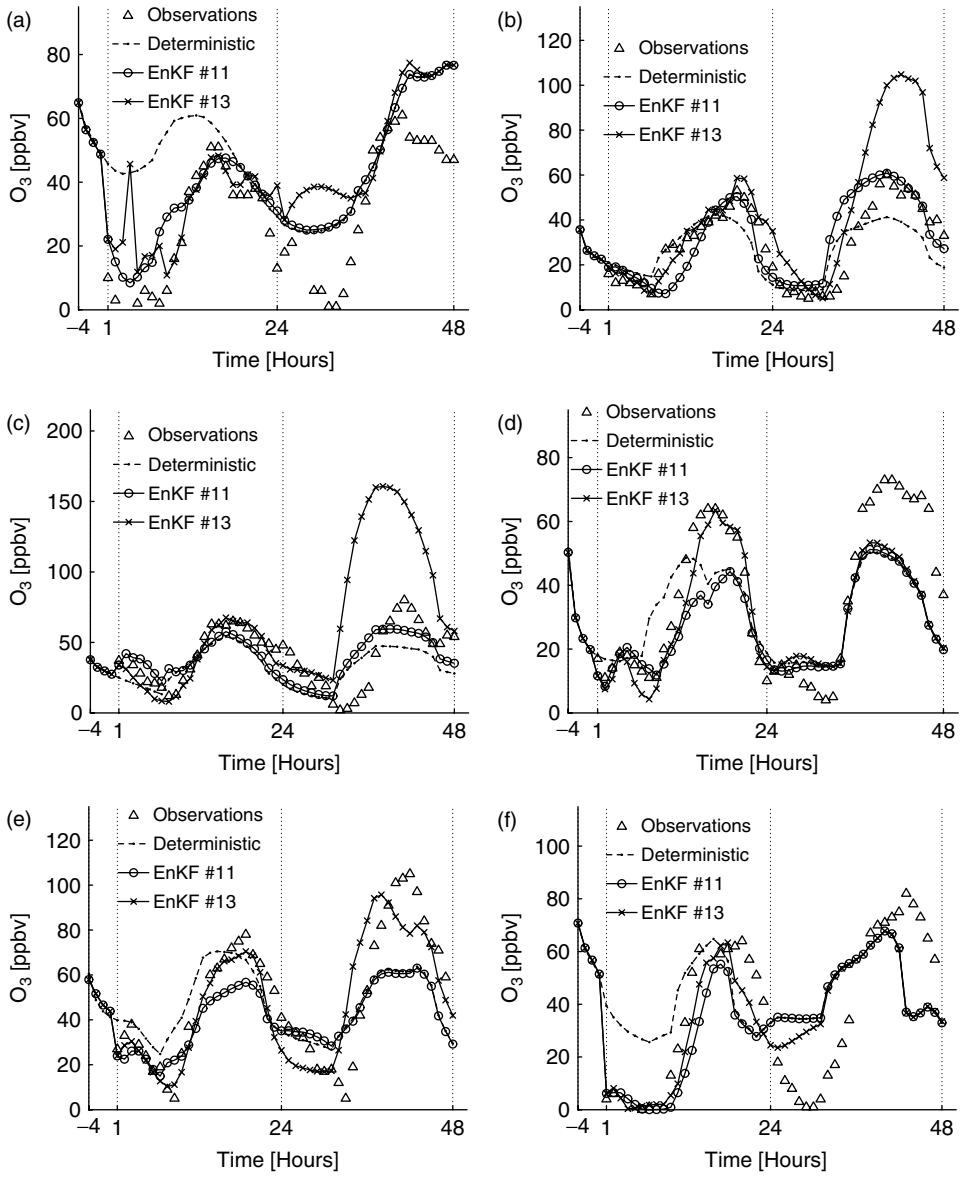

Figure 5. Ozone concentrations measured at the selected stations ('a' to 'f') and predicted by EnKF #11 (no inflation) and #13 (multiplicative inflation, $\gamma_- \leq 10$, $\gamma_+ \leq 8$), both with 200 members. EnKF #11 has no inflation and diverges, while EnKF #13 is over-inflated and becomes unstable, at least in the forecast window.

ground stations provide a rich dataset, 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 datasets were obtained by the two ozone-sondes S1 and S2 and during the P3-B flight (Figure 1(b)). The ozone-sondes were launched at 14 GMT (S1) and 22 GMT (S2) on 20 July. The NOAA P3-B plane was flown between 14 GMT and 22 GMT along the trajectory shown in Figure 1(b) at different altitudes (corresponding to vertical grid levels 3–16 in our model).

Figure 6 shows the vertical profiles of the ozone concentrations measured by the two ozone-sondes (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 (at 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, the vertical ozone profile is oscillatory, with the peaks taking unreasonable values. The vertical profiles obtained with 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 unassimilated solution, and show no oscillations.

The oscillatory behaviour of the EnKF solutions at higher levels is probably due to spurious correlations

DOI: 10.1002/qj

Table II. Trace (in ppbv) scaled by the number of observations of the covariance matrices from the validation relation (8). The covariance estimation is shown for the noiseless-filter application (EnKF #1), multiplicative inflation (EnKF #6), and model-specific inflation (EnKF #8 and #10a) for the first 10 hours of the assimilation window.

Time (GMT)	EnKF #1		EnKF #6		EnKF #8		EnKF #10a	
	$\mathbf{HPH}^{\mathrm{T}}$	$\mathbf{dd}^{\mathrm{T}} - \mathbf{R}$	$\mathbf{HPH}^{\mathrm{T}}$	$\mathbf{dd}^{\mathrm{T}} - \mathbf{R}$	$\mathbf{HPH}^{\mathrm{T}}$	$\mathbf{dd}^{\mathrm{T}} - \mathbf{R}$	$\mathbf{HPH}^{\mathrm{T}}$	$\mathbf{dd}^{\mathrm{T}} - \mathbf{R}$
2	0.58	13.36	2.26	29.04	1.29	13.36	13.25	13.38
3	0.36	14.26	2.32	29.04	1.01	12.88	13.30	13.48
4	0.26	15.87	2.27	29.04	1.26	14.12	14.47	14.59
5	0.22	16.58	2.45	13.36	1.20	13.12	13.17	13.31
6	0.17	18.13	2.09	13.36	1.13	12.96	13.06	13.19
7	0.15	19.07	2.20	13.38	1.17	12.39	12.30	12.49
8	0.18	24.22	3.55	13.36	1.15	17.93	18.61	18.69
9	0.16	25.48	2.20	14.26	1.28	18.95	19.12	19.19
10	0.14	23.29	2.11	12.88	1.43	16.08	18.96	19.04
11	0.14	21.59	2.75	13.48	1.37	14.82	22.71	22.75

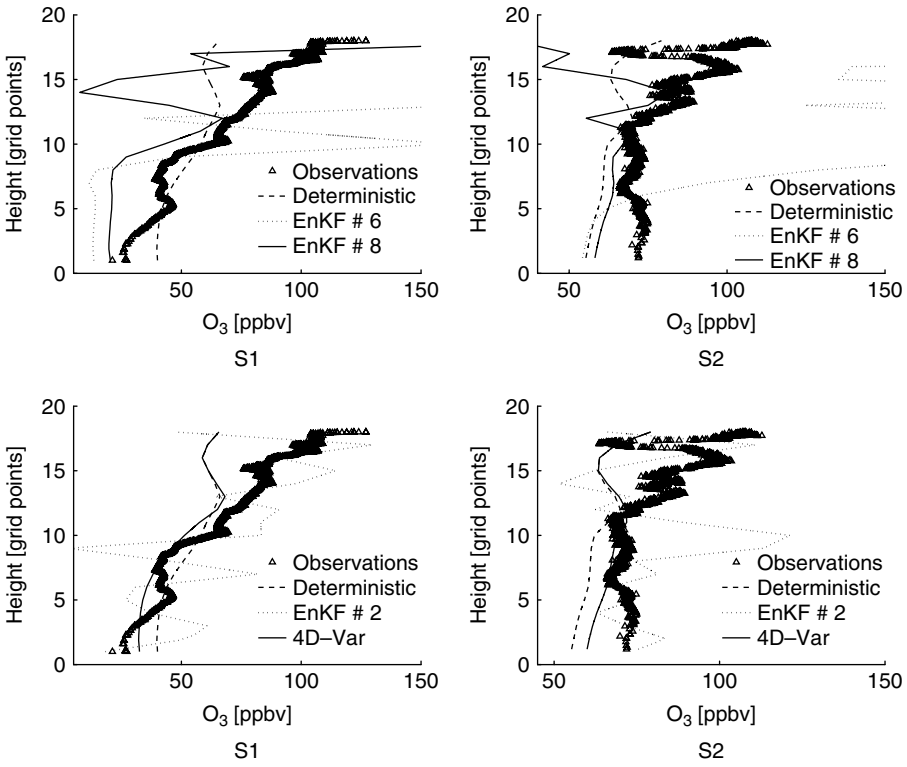

Figure 6. Ozone concentrations measured by ozone-sondes S1 and S2 and predicted by the model after data assimilation with 4D-Var, EnKF #2 (additive inflation), EnKF #6 (multiplicative inflation with $\gamma_- \leq 4$ and $\gamma_+ \leq 4$), and EnKF #8 (model-specific inflation with 10% perturbations of the emissions and boundary conditions and 3% perturbations of the wind).

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 ground level. The spurious correlations are due to the limited size of the ensemble. They are strongest for the multiplicative-inflation experiment (where all the correlations, including the spurious ones, are increased every cycle), and very mild for model-specific inflation (which better captures the real correlations). To alleviate the spurious correlations inherent to limited-size ensembles, one should consider techniques to explicitly localize the correlations (Houtekamer and Mitchell, 2001; Ott *et al*., 2004). This approach forces the corrective influence that each observation site exerts on the concentration field, making it 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 flight, not shown in this paper, allow us to draw conclusions that closely parallel those from the ozone-sondes. EnKF data assimilation with additive and multiplicative covariance inflation (and no localization) does not perform very well in the upper levels of the atmosphere

DOI: 10.1002/qj

because of the over-corrections required by spurious correlations. The solution obtained with model-specific inflation, and the 4D-Var solution, follow the observations well, although no visible improvement is obtained compared to the unassimilated concentrations. Clearly, to fully constrain the ozone field one needs to include measurements of the vertical ozone profiles in the assimilation.

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 (Constantinescu *et al*., 2007a) 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 show 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 over time, as the fields are largely determined by emissions and by lateral boundary conditions. Consequently, the initial spread of the ensemble is diminished over time. Moreover, stiff systems (like the chemistry components in the STEM model) are stable: small perturbations are damped out quickly, since fast transients are quickly 'attracted' to a (slow) low-dimensional manifold. Without simulating the atmospheric dynamics (meteorological fields are prescribed), these stiff effects are important.

In order to prevent filter divergence, the spread of the ensemble needs to be explicitly increased. We have 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 a very good agreement between model predictions and data in the assimilation window, but amplifies spurious correlations inherent to small ensembles, and causes the concentration fields away from the observation sites to deteriorate greatly. Model-specific covariance inflation is obtained by perturbing the meteorological fields, emissions, and lateral boundary conditions. The agreement of model predictions and observations is similar to that obtained with multiplicative inflation. However, model-specific covariance inflation does not over-amplify spurious correlations, and seems to be the best choice for chemical and transport modelling.

Experimental results show that 4D-Var and EnKF (without over-constraining the solution) produce results of similar quality. By inflating the covariance we can better constrain the EnKF solution near the ground level, and obtain a very good match between model predictions and observations in the assimilation window. This, however, causes the analysis quality at high levels (away from the observations) to deteriorate sharply. 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 will be considered in the second part of this study.

The numerical experiments in the idealized setting (Constantinescu *et al*., 2007a) used vertically-distributed observations. The numerical experiments with real data (presented in this paper) use 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 for properly constraining three-dimensional concentration fields. In the experiments presented here 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 behaviour of EnKF, one should consider the correction of other chemical fields as well.

Acknowledgements

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

References

Anderson JL. 2001. An ensemble adjustment Kalman filter for data assimilation. *Mon. Weather Rev.* **129**: 2884–2903.

Buizza R, Barkmeijer J, Palmer TN, Richardson DS. 2000. Current status and future developments of the ECMWF ensemble prediction system. *Meteorol. Appl.* **7**: 163–175.

Burgers G, van Leeuwen PJ, Evensen G. 1998. Analysis scheme in the ensemble Kalman filter. *Mon. Weather Rev.* **126**: 1719–1724.

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

Carmichael GR, Tang, Kurata, Uno, Streets, Woo, Huang, Yienger, Lefer, Shetter, Blake, Atlas, Fried, Apel, Eisele, Cantrell, Avery, Barrick, Sachse, Brune, Sandholm, Kondo, Singh, Talbot, Bandy, Thorton, Clarke, Heikes. 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)**: 10 649–10 671.

Chai T, Carmichael GR, Sandu A, Tang Y, Daescu DN. 2006. Chemical data assimilation of transport and chemical evolution over the Pacific (TRACE-P) aircraft measurements. *J. Geophys. Res.* **111**(D02301): doi:10.1029/2005JD005883.

Constantinescu EM, Sandu A, Chai T, Carmichael GR. 2007a. Assessment of ensemble-based chemical data assimilation in an idealized setting. *Atmos. Environ.* **41**(1): 18–36.

Constantinescu EM, Chai T, Sandu A, Carmichael GR. 2007b. Autoregressive models of background errors for chemical data assimilation. *J. Geophys. Res.* (to appear) DOI: 2006JD008103.

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

Courtier P, Thepaut J-N, Hollingsworth A. 1994. A strategy for operational implementation of 4D-Var, using an incremental approach. *Q. J. R. Meteorol. Soc.* **120**: 1367–1387.
Elbern H, Schmidt H. 2001. Ozone episode analysis by 4D-Var chemistry data assimilation. *J. Geophys. Res.* **106**: 3569–3590.
Elbern H, Schmidt H, Talagrand O, Ebel A. 2000. 4D-variational data assimilation with an adjoint air quality model for emission analysis. *Environ. Modell. Softw.* **15**: 539–548.
Evensen G. 1994. Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics. *J. Geophys. Res.* **99**(C5): 10 143–10 162.
Evensen G. 2003. The ensemble Kalman filter: theoretical formulation and practical implementation. *Ocean Dynam.* **53**(4): 343–367 .
Hamill TM. 2004. 'Ensemble-based atmospheric data assimilation'. Technical report, University of Colorado and NOAA-CIRES Climate Diagnostics Center, Boulder, Colorado, USA.
Hanea RG, Velders GJM, Heemink A. 2004. Data assimilation of ground-level ozone in Europe with a Kalman filter and chemistry transport model. *J. Geophys. Res.* **109**(D10-302): 1–19.
Heemink AW, Segers AJ. 2002. Modeling and prediction of environmental data in space and time using Kalman filtering. *Stoch. Env. Res. Risk A.* **16**(3): 225–240.
Houtekamer PL, Mitchell HL. 1998. Data assimilation using an ensemble Kalman filter technique. *Mon. Weather Rev.* **126**: 796–811.
Houtekamer PL, Mitchell HL. 2001. A sequential ensemble Kalman filter for atmospheric data assimilation. *Mon. Weather Rev.* **129**: 123–137.
Houtekamer PL, Mitchell HL, Pellerin G, Buehner M, Charron M, Spacek L, Hansen B. 2005. Atmospheric data assimilation with the ensemble Kalman filter: Results with real observations. *Mon. Weather Rev.* **133**(3): 604–620.
Hunt BR, Kalnay E, Kostelich E, Ott E, Patil D, Sauer T, Szunyogh I, Yorke J, Zimin A. 2004. Four-dimensional ensemble Kalman filtering. *Tellus A* **56**: 273–277.
ICARTT. 2004. http://www.al.noaa.gov/ICARTT.
Kalman RE. 1960. A new approach to linear filtering and prediction problems. *J. Basic Eng. – T. ASME* **82**: 35–45.
Kalnay E, Li H, Miyoshi T, Yang SC, Ballabrera-Poy J. 2005. 4D-Var or ensemble Kalman filter? *Tellus* (submitted).
Le Dimet FX, Talagrand O. 1986. Variational algorithms for analysis and assimilation of meteorological observations. *Tellus* **38A**: 97–110.
Liao W, Sandu A, Carmichael GR, Chai T. 2005. Singular vector analysis for atmospheric chemical transport models. *Mon. Weather Rev.* **134**: 2443–2465.
Lorenc AC. 1986. Analysis methods for numerical weather prediction. *Q. J. R. Meteorol. Soc.* **112**: 1177–1194.
Lorenc AC. 2003. The potential of the ensemble Kalman filter for NWP – a comparison with 4D-Var. *Q. J. R. Meteorol. Soc.* **129**(595): 3183–3203.
McKeen S, Wilczak, Grell, Djalalova, Peckham, Hsie, Gong, Bouchet, Menard, Moffet, McHenry, McQueen, Tang, Carmichael, Pagowski, Chan, Dye, Frost, Lee, Mathur. 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.
Miyoshi T. 2005. *Ensemble Kalman Filter Experiments with a Primitive-Equation Global Model*. Ph.D. thesis, University of Maryland.
Molteni F, Buizza R, Palmer TN, Petroliagis T. 1996. The new ECMWF ensemble prediction system: methodology and validation. *Q. J. R. Meteorol. Soc.* **122**: 73–119.
Ott E, Hunt BR, Szunyogh I, Zimin AV, Kostelich EJ, Corazza M, Kalnay E, Patil DJ, Yorke JA. 2004. A local ensemble Kalman filter for atmospheric data assimilation. *Tellus A* **56**(5): 415–428.
Rabier F, Jarvinen H, Klinker E, Mahfouf JF, Simmons A. 2000. The ECMWF operational implementation of four-dimensional variational assimilation. I: Experimental results with simplified physics. *Q. J. R. Meteorol. Soc.* **126**: 1148–1170.
Sandu A. 2006. 'On the properties of Runge–Kutta discrete adjoints'. *International Conference on Computational Sciences (ICCS-2006), Reading, UK (LCNS 3994):* 550–557.
Sandu A, Daescu DN, Carmichael GR. 2003. Direct and adjoint sensitivity analysis of chemical kinetic systems with KPP: I – theory and software tools. *Atmos. Environ.* **37**: 5083–5096.
Sandu A, Daescu DN, Carmichael GR, Chai T. 2005a. Adjoint sensitivity analysis of regional air quality models. *J. Comput. Phys.* **204**: 222–252.
Sandu A, Constantinescu EM, Liao W, Carmichael GR, Chai T, Seinfeld JH, Daescu DN. 2005b.. 'Ensemble filter data assimilation for atmospheric chemical and transport models'. *International Conference on Computational Sciences* (ICCS-2005) (LNCS 3515) 648–656.
Segers AJ. 2002. *Data Assimilation in Atmospheric Chemistry Models Using Kalman Filtering*. Ph.D. thesis, TU Delft.
Segers AJ, Heemink AW, Verlaan M, van Loon M. 2000. Modified RRSQRT-filter for assimilating data in atmospheric chemistry models. *Environ. Modell. Softw.* **15**: 663–671.
Talagrand O, Courtier P. 1987. Variational assimilation of meteorological observations with the adjoint vorticity equation. Part I: Theory. *Q. J. R. Meteorol. Soc.* **113**: 1311–1328.
Tang Y, Carmichael GR, Thongboonchoo N, Chai T, Horowitz LW, Pierce RB, Al-Saadi JA, Pfister G, Vukovich JM, Avery MA, Sachse GW, Ryerson TB, Holloway JS, Atlas EL, Flocke FM, Weber RJ, Huey LG, Dibb JE, Streets DG, Brune WH. 2007. Influence of lateral and top boundary conditions on regional air quality prediction: a multiscale study coupling regional and global chemical transport models. *J. Geophys. Res.* **112**: No. D10, D10S18 doi: 10.1029/2006JD007515.
van Loon M, Heemink AW. 1997. Kalman filtering for nonlinear atmospheric chemistry models: first experiences. *Modelling, Analysis and Simulation* **MAS-R9711**: CWI, Amsterdam, March.
van Loon M, Builtjes PJH, Segers AJ. 2000. Data assimilation of ozone in the atmospheric transport chemistry model LOTOS. *Environ. Modell. Softw.* **15**: 603–609.

DOI: 10.1002/qj