Ensemble-based Chemical Data Assimilation I: An Idealized Setting

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

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

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

Abstract

Data assimilation is the process of integrating observational data and model predictions to obtain an optimal representation of the state of the atmosphere. As more chemical observations in the troposphere are becoming available, chemical data assimilation is expected to play an essential role in air quality forecasting, similar to the role it has in numerical weather prediction. Considerable progress has been made recently in the development of variational tools for chemical data assimilation. In this paper we assess the performance of the ensemble Kalman filter (EnKF). Results in an idealized setting show that EnKF is promising for chemical data assimilation.

1 Introduction

Significant advancements have been made in recent years in our ability to measure and model the chemistry of the atmosphere. It is now possible to measure at surface sites and on mobile platforms many of the important primary and secondary atmospheric trace gases and aerosols. The spatial coverage is also expanding through growing capabilities to measure atmospheric constituents remotely using sensors mounted at the surface and in aircraft. From the modeling perspective chemical transport models (CTMs) have advanced to the point where they now specifically follow on the order of one hundred chemical species, interacting through chemical mechanisms involving hundreds of chemical reactions. However, while significant advances have occurred, atmospheric chemistry analyzes are hampered by the fact that chemical measurements and models are not closely integrated.

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. 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.

In this work we focus on data assimilation in chemical transport models (CTMs), which are designed to describe the fate and transport of atmospheric chemical constituents associated with the gas and aerosol phases. CTMs are an essential element in atmospheric chemistry studies, including important applications such as providing science-based input into best alternatives for reducing pollution levels in urban environments, designing cost-effective emission control strategies for improved air quality, for air-quality forecasting and assessments into how we have altered the chemistry of the global environment. Atmospheric chemical transport models pose specific challenges to data assimilation. The chemical interactions take place on a wide range of temporal scales (from $< 10^{-6}$ seconds to days). This makes the system numerically stiff. Moreover, the errors associated with misspecification of the initial conditions are often dominated by highly uncertain emission factors and uncertainty regarding the time-space distribution of anthropologically and naturally emitted pollutants. In regional models uncertainty in the specification of lateral boundary conditions considerably affects the solution. Therefore, to improve the analysis capabilities of CTMs, it is necessary to consider the estimation of emission parameters and lateral boundaries through data assimilation [Stewart, 1993, Menut, 2003].

In the variational approach (3D-Var, 4D-Var) the mismatch between model predictions and observations is quantified by a cost functional. Data assimilation is then formulated as an optimization problem where the model state and model parameters are adjusted to minimize this cost functional. Chemical data assimilation has advanced considerably in the past decade using the variational approach [Elbern and Schmidt, 1999, 2001, Elbern et al., 1997, 2000, 1999, Fisher and Lary, 1995, Menut et al., 2000, Sandu et al., 2003, 2005, Liao et al., 2005, Chai et al., 2006, Constantinescu et al., 2006a].

In this study we focus on the ensemble Kalman filter (EnKF) approach to chemical data assimilation, which has several highly attractive features. The computational model need not be modified, as there is no need for the tangent linear or adjoint models. The effects of non-linear dynamics are better captured than with the variational approaches (which are intrinsically linear). EnKF allows to easily account for model errors, and the calculations are almost ideally parallelizable. A detailed comparison of the relative merits of EnKF and 4D-Var in the context of Numerical Weather Prediction (NWP) can be found in [Lorenc, 2003, Kalnay et al., 2005].

EnKF has attracted considerable attention in meteorology. Houtekamer et. al. [Houtekamer and Mitchell, 2001, Houtekamer et al., 2005] have shown that significant gains can be obtained by applying ensemble Kalman filter (EnKF) to operational numerical weather prediction models. The sequential EnKF proposed in [Houtekamer and Mitchell, 2001] organizes observations into batches that are assimilated sequentially, thus increasing the computational efficiency. In [Mitchell and Houtekamer, 2002] the authors investigate three issues related to sequential EnKF, namely include ensemble size, balance, and model-error representation. Substantial imbalance in the analyzes can appear when the localization (the cutoff distance for correlations) is severe, but decreases as the localization is relaxed. Hunt et. al. developed 4D-EnKF [Hunt et al., 2004], a technique which allows observations to occur at times different than assimilation times. The linearized model dynamics is inferred from the ensemble, and the observational increments at intermediate times are propagated using the ensemble. Blond and Vautard [Blond and Vautard, 2004] used statistical interpolation to recover the surface ozone over Western Europe. They concluded that correcting only the initial conditions yields limited results, and other sources of uncertainty (like emissions or boundary conditions) need to be addressed in order to increase the prediction capability.

Ensemble Kalman filter has been used in chemical data assimilation to recover ozone and emissions [Van Loon et al., 2000, Heemink and Segers, 2002]. This work shows that it is possible to successfully apply the ensemble Kalman filter to an atmospheric CTM for data assimilation, and to improve the quality of the forecasts. The results also showed that although the data assimilation can significantly improve ozone estimates, it degrades the estimates of other important chemical species. A comparison among different flavors of reduced Kalman filters is given in [Heemink and Segers, 2002].

In this study we investigate the application of "perturbed observations" EnKF to chemical data assimilation. Here, we analyze the performance of EnKF data assimilation in an ideal setting, where a reference solution is considered the "truth" and is used to generate the initial ensemble, to obtain artificial observations, and to asses the quality of the results. The contributions of this work are: an analysis of EnKF on large scale chemical models, the use of model singular vectors and autoregressive background models to form the initial ensemble, study the effects of the ensemble size, emissions, and boundary conditions on chemical data assimilation.

The paper is structured in two parts. In the first part we analyze the performance of EnKF data assimilation in an ideal setting, where a reference solution is considered the "truth" and is used to generate the initial ensemble, to obtain artificial observations, and to asses the quality of the results. The second part of this study [Constantinescu et al., 2006b] continues the analysis in a real setting with real observations, discusses various strategies for covariance inflation, and compares the EnKF performance with a state-of-the art 4D-Var. In the third part of this study [Constantinescu et al., 2006c] we investigate the "localization" of EnKF.

This paper is structured as follows: In Section 2 we review the Kalman, ensemble Kalman, and chemical and transport models. Section 3 presents the construction of the initial ensemble. The analysis scheme is presented in Section 4. Our numerical results with EnKF data assimilation applied to a CTM are shown and discussed in Section 5. Conclusions and future research directions are given in Section 6.

Throughout this paper we use the notations from [Ide et al., 1997], where applicable.

2 Background

In this section we introduce the chemical transport models (Sec. 2.1) and review the theory of the ensemble Kalman filter (Sec. 2.2) used in our numerical experiments.

2.1 Chemical and Transport Models

Atmospheric chemistry and transport models 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]. Let c_s be the mole-fraction concentration of chemical species s, Q_s be the rate of surface emissions, E_s be the rate of elevated emissions, and f_s be the rate of chemical transformation for this species. Further, u is the wind field vector, K the turbulent diffusivity tensor, and ρ is the air density. The evolution of c_s is described by the following equations

$$
\begin{aligned}
\frac{\partial c_s}{\partial t} &= -u\nabla c_s + \frac{1}{\rho}\nabla(\rho K \nabla c_s) + \frac{1}{\rho} f_s(\rho c) + E_s, \quad t^0 \le t \le t^F, \quad 1 \le s \le N_{\rm spec}, \\
c_s(t^0, x) &= c_s^0(x), \\
c_s(t, x) &= c_s^{\rm in}(t, x) \quad \text{for} \quad x \in \Gamma^{\rm in}, \\
K\frac{\partial c_s}{\partial n} &= 0 \quad \text{for} \quad x \in \Gamma^{\rm out}, \\
K\frac{\partial c_s}{\partial n} &= V_s^{dep} c_s - Q_s \quad \text{for} \quad x \in \Gamma^{\rm ground}\ .
\end{aligned}
\tag{1}
$$

The model solution operator will be denoted compactly as

$$c_i = \mathcal{M}_{t_{i-1}\to t_i}\big(\, c_{i-1},\, u_{i-1},\, c_{i-1}^{\rm in},\, Q_{i-1}\,\big)\ . \tag{2}$$

where subscripts represent time, $c_i = c(t_i)$ etc.

A major difference between CTMs and NWP models is the presence of stiff chemical kinetic terms [Sandu et al., 1997] (represented as f_s in (1)). Stiff systems are very stable, and small perturbations of their state are rapidly damped out. Another difference between CTMs and NWP models is that the former does not solve the dynamic (momentum) equations. In practice CMTs are derived by prescribed meteorological fields (computed and analyzed off-line). In the future, however, it is expected that CTMs will be coupled with dynamic atmospheric models.

In our numerical experiments, we use the Sulphur Transport Eulerian Model (STEM) [Carmichael et al., 2003], a state-of-the-art chemical and transport atmospheric model. A further discussion of STEM's numerical methods and settings is presented in section 5.1.

2.2 The Ensemble Kalman Filter (EnKF)

Consider the discrete model (1) $\mathcal{M}_{t_{i-1}\to t_i} : \mathbb{R}^{\rm N} \to \mathbb{R}^{\rm N}$ that evolves the system's state vector $c \in \mathbb{R}^{\rm N}$ from time t_{i-1} to time t_i ($i \ge 1$). The model is an imperfect representation of a real system having the "true" state $c^t \in \mathbb{R}^{\rm N}$. The model predictions are not exact and therefore

$$c_i \;=\; \mathcal{M}_{t_{i-1}\to t_i}\big(c_{i-1}\big) + \eta_i\ , \tag{3}$$

where the random variable $\eta_i = c_i^t - c_i^f$ represents the *model error*. The model error is typically assumed to be Gaussian with mean zero (the model is unbiased) and covariance Q, $\eta_i \in \mathcal{N}(0, Q_i)$.

Observations $y \in \mathbb{R}^{\rm P}$ of the true state $c^t \in \mathbb{R}^{\rm N}$ are available at discrete times t_i, $i \ge 0$

$$y_i \;=\; \mathcal{H}_i\big(c_i^t\big) + \varepsilon_i, \tag{4}$$

where the random variable ε_i represents the *observation error*. The observation operator $\mathcal{H} : \mathbb{R}^{\mathrm{N}} \to \mathbb{R}^{\mathrm{P}}$ maps the state space into the observation space. Let $\langle \cdot \rangle$ denote the statistical average. The observation error is typically assumed to be Gaussian with mean zero and covariance R, $\varepsilon_i \in \mathcal{N}(0, R_i)$.

The Kalman filter [Kalman, 1960, Fisher, 2002] gives an optimal estimate of the true state c^t using the model approximate solution (the *forecast*) $c^f \in \mathbb{R}^{\mathrm{N}}$, and the *observations* $y \in \mathbb{R}^{\mathrm{P}}$. This optimal estimate of the state is called the *analysis* $c^a \in \mathbb{R}^{\mathrm{N}}$. The analysis is obtained as a linear combination of the forecast and observations that minimize the variance of the analysis

$$c_i^a = c_i^f + K_i\, d_i \;, \quad d_i = y_i - \mathcal{H}_i\left(c_i^f\right) \;, \tag{5}$$

where K_i is the Kalman gain matrix and d_i the innovation vector. Assuming that the model and observation errors are uncorrelated the Kalman gain is given by

$$K_i = P_i^f H_i^T \left(H_i P_i^f H_i^T + R_i\right)^{-1} \;, \tag{6}$$

where $H = \mathcal{H}'$ is the linearized observation operator and $P_i^f = \langle \eta_i \eta_i^T \rangle$ is the *forecast error covariance*. We denote by $\langle \cdot \rangle$ the statistical average.

The filter works as follows. The best estimate of the state at t_{i-1} is the analysis c_{i-1}^a. This state is propagated to t_i using the model (3) to obtain the model forecast c_i^f. The filter (5)–(6) is then applied to combine the forecast state and the observations and obtain the analysis c_i^a

$$c_i^f = \mathcal{M}_{t_{i-1}\to t_i}\left(c_{i-1}^a\right) + \eta_i \;, \quad c_i^a = c_i^f + K_i\left(y_i - \mathcal{H}_i\left(c_i^f\right)\right) \;.$$

The forecast covariance matrix P_i^f is evolved from the previous step

$$P_i^f = M_{t_{i-1}\to t_i}\, P_{i-1}^a\, M_{t_i \to t_{i-1}}^* + Q_i \;, \tag{7}$$

where $M = \mathcal{M}'$ is the tangent linear model of (3) and M^* the adjoint of M. The analysis covariance matrix P_i^a is given by the filter as

$$P_i^a = P_i^f - K_i H_i P_i^f \;. \tag{8}$$

The Kalman filter is not practical for large systems, because of the prohibitive computational cost needed to invert the large matrix in (6) and to propagate the covariance matrix in time (7). Approximations are needed to make the Kalman computationally feasible. One such approximation is provided by the ensemble Kalman filter (EnKF).

First proposed by Evensen [Evensen, 1994], and then later clarified by Burgers [Burgers et al., 1998], the ensemble Kalman filter [Fisher, 2002] uses a Monte-Carlo approach to propagate covariances. An ensemble of E states (labeled $e = 1, \cdots, E$) is used to sample the probability distribution of the background error. Each member is advanced in time and analyzed separately to produce an ensemble of analyzed states

$$c_i^f(e) = \mathcal{M}_{t_{i-1}\to t_i}\left(c^a_{i-1(e)}\right) + \eta_i(e)\ , \quad c_i^a(e) = c_i^f(e) + K_i\left(y_i(e) - \mathcal{H}_i\left(c_i^f(e)\right)\right)\ , \quad e = 1, \cdots, E\ .$$

The forecast and the analysis covariances are estimated from the statistical samples

$$P_i^{\{a,f\}} \approx \frac{1}{\mathrm{E}-1} \sum_{e=1}^{\mathrm{E}} \left(c_i^{\{a,f\}} - \langle c_i^{\{a,f\}}\rangle_{\mathrm{E}}\right) \left(c_i^{\{a,f\}} - \langle c_i^{\{a,f\}}\rangle_{\mathrm{E}}\right)^T\ . \tag{9}$$

where $\langle\cdot\rangle_{\mathrm{E}}$ represents the ensemble average.

The ensemble Kalman filter raises several issues. First the rank of estimated covariance matrix is usually several orders of magnitude smaller than the dimension of the matrix. Two methods have been used to fix the rank-deficiency problem: splitting the analysis increment into two parts and increasing the rank of estimated covariance [Houtekamer and Mitchell, 2001]. Next, the random errors in the statistically estimated covariance decrease slowly, only by the square-root of the ensemble size. Furthermore, the subspace spanned by random vectors for expressing the forecast error is not optimal. In spite of the problems, ensemble Kalman filter has many attractive features.

Evensen [Evensen, 1992, 1993] discussed the implementation of the extended Kalman filter for data assimilation in a multilayer quasi-geostrophic model. In [Evensen, 1994] Evensen proposed to replace the error covariance equation in the extended filter by a Monte-Carlo solution to the "full" Kolmogorov equation. The error statistics needed in extended Kalman filter can be calculated directly from the ensemble. The numerical results presented in this study are based on the practical EnKF implementation presented by Evensen in [Evensen, 2003].

3 The Initial Ensemble

One of the challenges with ensemble forecasting is the specification of the initial ensemble. For a correct ensemble, each member is drawn from the same probability distribution function (pdf) that produced the true system state, and is impossible to distinguish between ensemble members and truth. Hansen [Hansen, 2002] argues that the initial ensemble should sample the (local) system attractor. A good approximation of the background error statistics, and a correct initialization of the ensemble are essential for the success of ensemble data assimilation.

In the ECMWF ensemble prediction system [Molteni et al., 1996] the ensemble perturbations are generated from the leading singular vectors of the linearized propagator. These vectors identify the directions in phase space associated with maximum perturbation growth.

In this section we consider the autoregressive models for background errors and discuss the construction of model singular vectors. A more detailed discussion can be found in [Constantinescu et al., 2006a, Liao et al., 2005].

3.1 Flow-Dependent Models of Background Error

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; a typical assumption is that it has a Gaussian distribution with zero mean (the model is unbiased) and covariance $\mathbb{B}$. In EnKF the background covariance is used to generate the initial ensemble. A good approximation of the background error statistics is therefore essential for the success of data assimilation.

The initial state of each member e, $e = 1, \cdots, E$ is formed by adding a different perturbation $\delta c^{\mathrm{B}}(e)$ to the initial "best guess" (background) state

$$c_0(e) = c^{\mathrm{B}} + \delta c^{\mathrm{B}}(e) \ , \quad e = 1, \cdots, E \ .$$

The ensemble of perturbations should correctly sample the probability distribution of background errors. Building the initial ensemble based on the distance and flow dependence has been discussed in [Riishojgaard, 1998, Hamill and Whitaker, 2001, Buehner, 2004].

In this study the background covariance is modeled by autoregressive (AR) processes [Constantinescu et al., 2006a] of the form

$$\delta c^{\mathrm{B}}_{i,j,k} + \alpha^{(\pm 1)}_{i,j,k}\, \delta c^{\mathrm{B}}_{i\pm 1,j,k} + \beta^{(\pm 1)}_{i,j,k}\, \delta c^{\mathrm{B}}_{i,j\pm 1,k} + \gamma^{(\pm 1)}_{i,j,k}\, \delta c^{\mathrm{B}}_{i,j,k\pm 1} = \sigma_{i,j,k}\, \xi_{i,j,k}, \tag{10}$$

where α, β, and γ are the autoregressive coefficients, subscripts refer to the spatial coordinates, and σ represents the error variance. The AR process can be represented compactly as

$$A\, \delta c^{\mathrm{B}} = S\, \xi \ , \quad S = \mathrm{diag}(\sigma_{i,j,k}) \ , \tag{11}$$

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 perturbation that defines the initial state of the e-th member of the ensemble is

$$\delta c^{\mathrm{B}}_{\mathrm{AR}}(e) = A^{-1}\, S\, \xi(e) \ , \quad e = 1, \cdots, E \ .$$

where $\xi(e) \in \left(\mathcal{N}(0,1)\right)^N$ is a vector of N independent normal random variables of mean 0 and standard deviation 1. This perturbation is generated by scaling the normal variables ξ with the proper standard deviations, then solving a linear system with the AR coefficient matrix A. The background covariance matrix is $\mathbb{B} = A^{-1}\, S^2\, A^{-T}$. The AR model (10) is constructed using the coefficients A of a discretization of the advection–diffusion–reaction operator. A computationally efficient approach is to obtain A via operator splitting of the chemistry and transport, followed by dimensional splitting of the three-dimensional advection-diffusion equation. This model of the background covariance accounts for spatial correlations, distance decay, and chemical lifetime [Constantinescu et al., 2006a].

3.2 Model Singular Vectors

Model singular vectors are the directions of the most rapidly growing perturbations over a finite time interval. We measure the magnitude of the perturbations in the concentration fields using L^2 weighted norms. The ratio between perturbation energies at the final (t^{F}) and initial time (t^0) offers a measure of error growth:

$$\sigma^2 = \frac{\|\delta x(t^{\mathrm{F}})\|_F^2}{\|\delta x(t^0)\|_G^2} = \frac{\langle \delta x(t^0), M^*_{t^{\mathrm{F}}\to t^0} F M_{t^0\to t^{\mathrm{F}}} \delta x(t^0)\rangle}{\langle \delta x(t^0), G \delta x(t^0)\rangle} \tag{12}$$

Here G is a positive definite and F a positive semidefinite matrix. In (12) we use the fact that perturbations evolve in time according to the dynamics of the TLM. Model singular vectors are defined as the directions of maximal error growth, i.e. the vectors $s_k(t^0)$ that maximize the ratio σ^2 in equation (12). These directions are the solutions of the following generalized eigenvalue problem:

$$M^*_{t^{\mathrm{F}}\to t^0}\, F\, M_{t^0\to t^{\mathrm{F}}}\, s_k(t^0) = \sigma_k^2\, G\, s_k(t^0) \tag{13}$$

The left side of (13) involves one integration with the tangent linear model followed by one integration with the adjoint model.

The eigenvalue problem (13) is solved by software packages like ARPACK [Maschhoff and Sorensen] using Lanczos iterations. The symmetry of the matrix $M^*\, F\, M$ required by Lanczos imposes to use the discrete adjoint M^* of the tangent linear operator M in (13). The computation of discrete adjoints for stiff systems is a nontrivial task [Sandu et al., 2003]. In addition, computational errors (which can destroy symmetry) have to be small. A more detailed discussion can be found in [Liao et al., 2005].

An initial random perturbation can be constructed in the space of the model singular vectors as follows

$$\delta c^{\mathrm{B}}_{\mathrm{SV}} = \sum_k \alpha_k\, \xi_k\, s_k(t^0)$$

where $\xi_k \in \mathcal{N}(0,1)$ are normal random variables and α_k are appropriate scaling coefficients. Adding an initial perturbation in the space spanned by dominant singular vectors ensures that the ensemble spans the directions of maximal error growth.

4 The Analysis Scheme

In this paper we follow closely the classical implementation of "perturbed observations" EnKF as described in [Evensen, 2003].

The initial state of each ensemble member is obtained by adding to the background both an autoregressive perturbation (which captures flow-dependent error correlations, see Section 3.1) and a perturbation in the space of dominant model singular vectors (which samples the directions of maximal error growth, see Section 3.2)

$$c_0(e) = c^{\mathrm{B}} + \delta c^{\mathrm{B}}_{\mathrm{AR}}(e) + \delta c^{\mathrm{B}}_{\mathrm{SV}}(e) \ , \quad e = 1, \cdots, E \ . \tag{14}$$

Emissions and lateral boundary conditions are major sources of uncertainty in regional atmospheric CTMs. After some simulation time the solution is driven less by the initial conditions and more by emissions and boundary conditions. EnKF can be extended to include the emission and lateral boundary condition in the assimilation process (and solve the state-parameter estimation problem [Derber, 1989, Annan et al., 2005, Evensen, 2005]). Correction coefficients α^{EM} and α^{BC} are used to adjust the (prescribed) emission rates and lateral boundary conditions, respectively, in each grid point. The correction coefficients can be viewed as model parameters, and are padded to the controlled state variables to form and extended state vector

$$\begin{bmatrix} c_i \\ \alpha_i^{\mathrm{EM}} \\ \alpha_i^{\mathrm{BC}} \end{bmatrix} = \begin{bmatrix} \mathcal{M}_{t_{i-1}\to t_i}\left(c_{i-1},\, u_{i-1},\, c^{\mathrm{in}}_{i-1},\, Q_{i-1} \right) \\ \alpha_{i-1}^{\mathrm{EM}} \\ \alpha_{i-1}^{\mathrm{BC}} \end{bmatrix} \ . \tag{15}$$

An uncorrelated unbiased perturbation is used for the initial emission and lateral boundary conditions. The ensemble propagates and the filter corrects the extended state vector. The corrected emissions and boundary values are then used during the forecast.

A correct estimation of model errors is important in data assimilation in order to quantify the correct level of "trust" in the model forecast. A direct approach to accounting for model errors is to add noise to the ensemble of model forecasts. In this study we have taken a different approach, namely we have randomly perturbed the emissions and boundary conditions for each member run.

5 Numerical Results

In this section we experimentally investigate the performance and the feasibility of EnKF data assimilation in the context of photochemical and transport models. For this purpose we consider an idealized setting in which the “truth” is a reference solution computed with the model, and artificial observations are generated by perturbing the “true” (reference) values.

5.1 The Test Problem

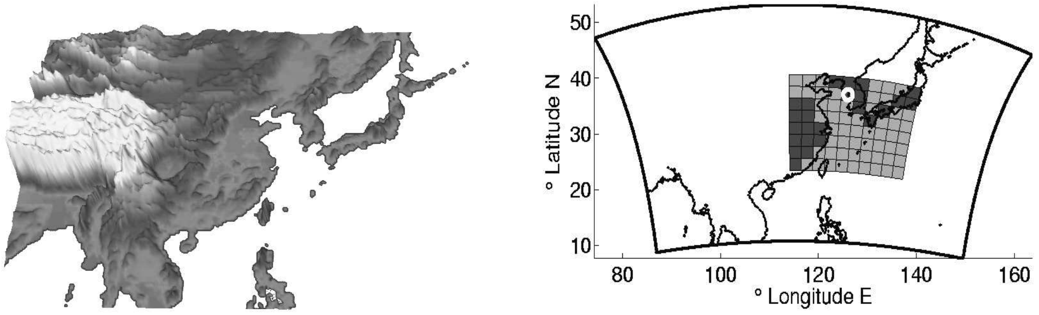
The test problem is a simulation of air pollution in South-East Asia using the STEM model and TraceP [Carmichael et al., 2003] conditions.

The chemical reaction and transport equation (1) is solved using an operator splitting approach. STEM uses linear finite difference discretization of the transport terms. Horizontal transport is solved using a directional x and y split approach, and a third order 1D upwind finite difference formula [Sandu et al., 2005]. The diffusion terms are discretized using second order central differences. The advection inflow boundary uses a first order upwind scheme, which makes the order of whole scheme quadratic for the interior points. The vertical advection scheme by first order upwind finite difference and the diffusion term is discretized by the second order central differences [Sandu et al., 2005]. Atmospheric chemical kinetics result in stiff ODE equations that use a stable numerical integration that preserve linear invariants.

The gas phase mechanism is SAPRC-99 [Carter, 2000] which accounts for 93 chemical species (88 variable and 5 constant) involved in 235 chemical reactions. The chemistry time integration is done by Rosenbrock 2 numerical integrator [Sandu and Daescu, 2005], implemented using the kinetic preprocessor (KPP) [Damian et al., 2002].

The numerical experiment is a real-life simulation of air pollution in South-East Asia in support of the TraceP field experiment (NASA TRAnsport and Chemical Evolution over the Pacific) [Carmichael et al., 2003]. The meteorological fields, boundary values, and emission rates correspond to TraceP starting at 0 GMT of March 1st to 0 GMT March 3rd, 2001. The simulated region (shown in Figure 1.a) covers $7200 \times 4800 \times 20$ Km, and is covered by a 3-dimensional computational grid with $30 \times 20 \times 18$ points; the grid has 240×240 Km horizontal resolution and varying vertical height.

The following numerical experiments consider a 24 hour assimilation window (0 GMT of March 1st to 0 GMT March 2nd, 2001) followed by a 24 hours forecast window (0 GMT of March 2nd to 0 GMT March 3rd, 2001) in order to assess the performance of the analysis scheme.



(a) South-East Asia (b) Observations and verification area

Figure 1: a) The simulated physical domain (East Asia); b) The computational domain and the location of the ground observations (dark), the column observations (light o), and the ground projection of the parallelipipedic verification area (light).

5.2 Analysis Setting

An idealized ensemble is constructed by adding perturbations to the "true" (reference) solution c^t. The idealized ensemble together with artificial observations, Hc^t, allow us to study performance of EnKF applied to chemical transport models in isolation from other issues like data and model errors.

A parallelipipedic verification area is defined above Korea (Figure 1.b). We are interested to improve the estimates of the concentration fields within the verification area. We will assess the quality of the assimilated fields for ozone (O_3), nitrogen dioxide (NO_2), as well as for species that are not observed directly: formaldehyde (HCHO), peroxyacyl nitrate PAN, and carbon monoxide CO. The verification region is chosen away from the model boundaries in order to avoid the boundary artifacts in the assimilation process.

The analysis setting used in the numerical experiments has the following characteristics:

- *Reference solution.* The reference solution is started at 0 GMT of March 1^{st} and ends at 0 GMT March 3^{rd}, 2001 (48 hours) with the TraceP initial concentrations.
- *Assimilated solution.* We follow the assimilation results for one particular ensemble member, based on the principle that the ensemble members cannot be statistically distinguished between them and they equally well represent the truth. Note that in the idealized setting used here the ensemble is unbiased (is constructed about the "truth") and remains essentially unbiased throughout the simulation. Thus the ensemble mean is essentially indistinguishable from the reference solution.
- *Observations.* Artificial observations are obtained from the reference run for ozone (O_3) and one of its chemical precursors, nitrogen dioxide (NO_2) on the ground level in Korea, Japan, and part of China, and along a vertical column above Korea. In total

there are 24 observed grid points on the ground, and 17 observed gridpoints along the column. The location (grid coordinates) of the observations is presented in Figure 1.b.

- *Assimilation window.* The assimilation window starts at 0 GMT March 1st, and ends at 0 GMT March 2nd (denoted from now on as the interval $[0, 24]$ hours). Observations are available at 6, 12, 18, and 24 hours.
- *Forecast:* The forecast window starts at 0 GMT March 2nd, and ends at 0 GMT March 3rd (denoted as the interval $[24, 48]$ hours).
- *States.* The control states are the concentrations of 66 different species, including the observed ones.
- *Parameters.* The correction factors applied to the emission rates and lateral boundary conditions are considered model parameters, and are assimilated in the state-parameter estimation experiments.
- *Model singular vectors.* Model singular vectors are computed for the assimilation window with respect to the verification region at the final time. The dominant 40 model singular vectors were used to initialize the ensemble.

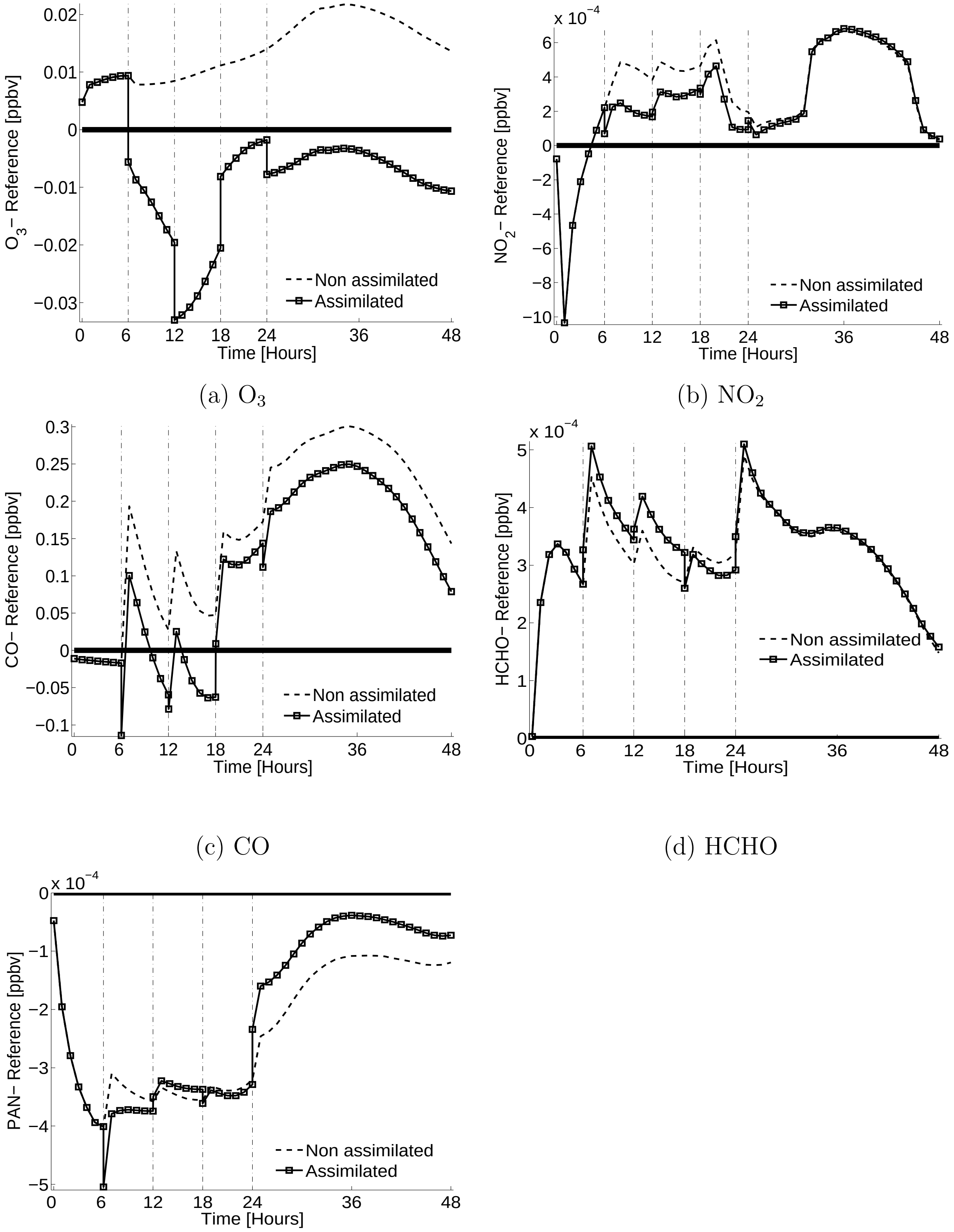
5.3 Ensemble Bias

In the numerical results we present the concentrations of several chemical species (O_3, NO_2, CO, HCHO, and PAN) averaged over the verification area. The concentration units are parts per billion (volumetric) – ppbv. Among the selected species only O_3 and NO_2 are directly observed; CO, HCHO, and PAN are adjusted by assimilating the observations of ozone and nitrogen dioxide.

Figure 2 shows the absolute ensemble bias for the selected species during the assimilation and forecast windows. The ensemble has a very small bias and this bias does not increase over time.

5.4 Ensemble Size

The ensemble size determines the accuracy to which the forecast error covariance is approximated. A small ensemble size leads to under-prediction of the forecast error [Houtekamer and Mitchell, 1998, Mitchell and Houtekamer, 1999, 2002], and ultimately may lead to filter divergence. Filter divergence [Houtekamer and Mitchell, 1998, Hamill, 2004] is caused by progressive underestimation of the model error covariance and coerces filter to neglect the observations in the analysis process. A large ensemble is expensive (the cost increases linearly with the ensemble size while the accuracy of the covariance estimate improves by



(a) O_3

(b) NO_2

(c) CO

(d) HCHO

(e) PAN

Figure 2: Ensemble bias (ensemble average minus the reference solution) for 48 hours of simulation. The first 24 hours are the assimilation window, the next 24 hours are the forecast. The ensemble remains essentially unbiased.

its square root). An important question is how large should the ensemble be, and how to determine its size.

The appropriate ensemble size depends on the application and model. We performed several simulations with ensembles of 10, 22, and 50 members. The results are presented in Figure 3. The reference and the analysis concentration fields of O_3 (directly observed) and CO (not observed) are averaged over the verification area. Smaller ensembles (Figure 3.a,d) have smaller spreads and under-represent model errors. Figures 3.c and 3.e show that the large ensemble (50 members) provides analysis solutions that are very close to the reference for both for the observed and not observed species. In the next experiments we will consider 50-member ensembles.

5.5 Ensemble Convergence

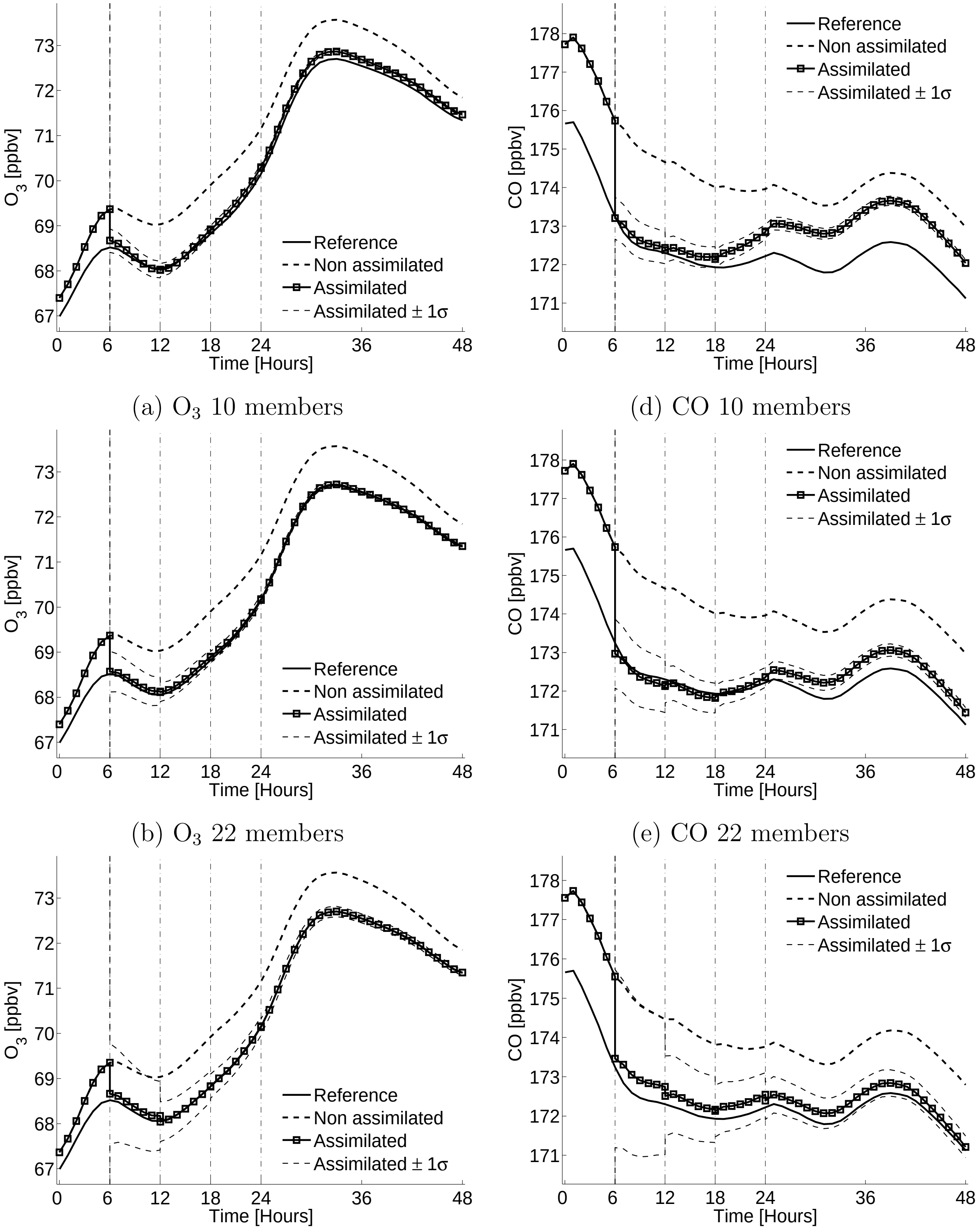
Figure 4 shows the convergence of the assimilated solution for several chemical species (O_3, NO_2, CO, HCHO, and PAN). Only O_3 and NO_2 are directly observed. Note that the ensemble spread is decreasing slowly during the ensemble evolution in time. As expected, sharp reductions in the ensemble spread are seen at the assimilation times. Both directly observed and unobserved species are assimilated correctly. Short lived species like NO_2 (in Figure 4.b) do not show considerable difference between the assimilated and non assimilated solution.

5.6 Improvements in Forecast Capability

We now investigate the impact that EnKF data assimilation has on the forecast capability of the model. The estimation of state only and the combined estimation of parameters and state are discussed. The numerical results present the error fields, i.e., the differences between the perturbed (assimilated or non-assimilated) fields and the reference solution.

A comparison between the errors in the assimilated and in the non assimilated solutions are shown in Figure 5 (O_3), Figure 6 (NO_2), Figure 7 (CO), Figure 8 (HCHO), and Figure 9 (PAN). The two-dimensional plots are obtained by averaging the errors across all vertical layers. The errors are shown at the end of the assimilation window (24h) and at the end of the forecast window (48h). Data assimilation considerably improves the estimates of chemical species, both directly observed (Figures 5, 6) and unobserved (Figures 7, 8, 9). The filter is capable of correctly accounting for the inter-species correlations formed during the model (chemistry) integration.

Boundary conditions play an important role in determining the concentration fields in regional models. Since we use unperturbed numerical boundary conditions, a very small



(a) O_3 10 members

(d) CO 10 members

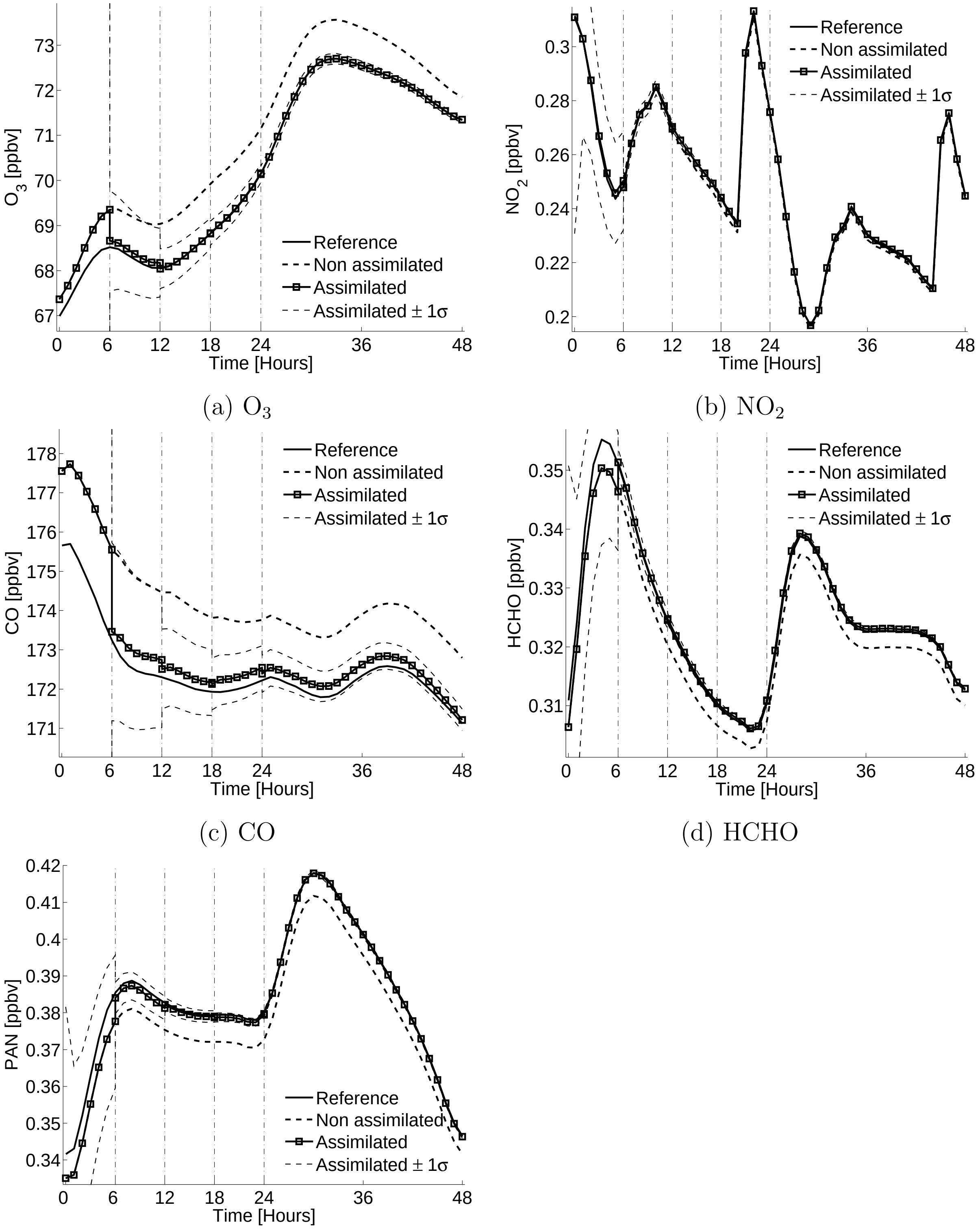
(b) O_3 22 members

(e) CO 22 members

(c) O_3 50 members

(f) CO 50 members

Figure 3: Assimilation with different ensemble sizes. The convergence of the 50-member ensemble is considered sufficient for both the observed and not observed species.



(a) O_3

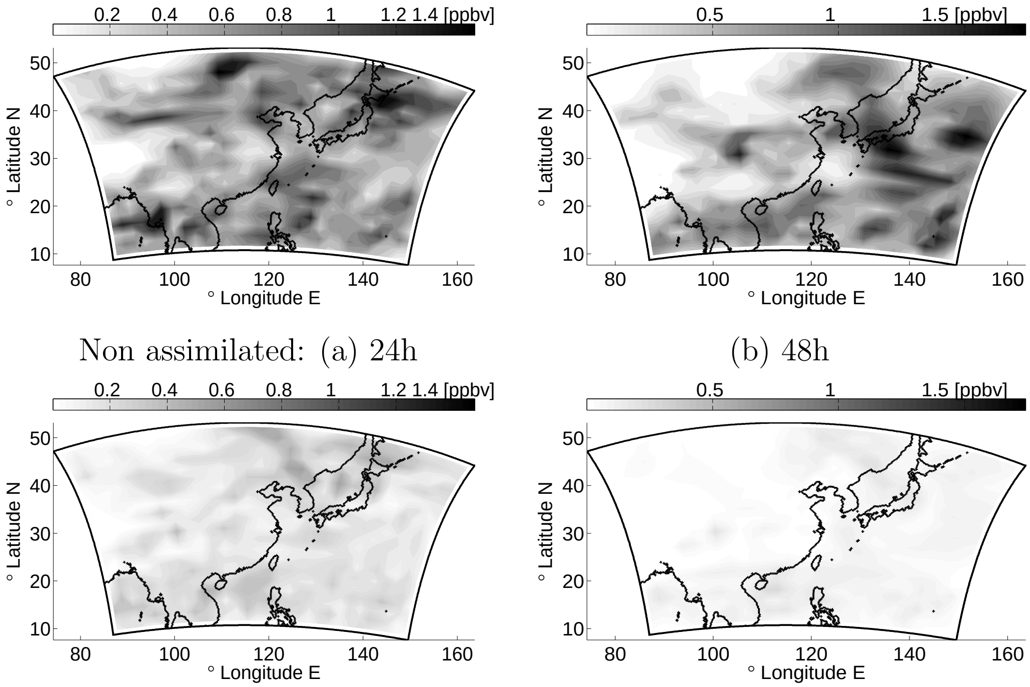
(b) NO_2

(c) CO

(d) HCHO

(e) PAN

Figure 4: Assimilated solution averaged on the verification region – 24 hours assimilation, 24 hours forecast. The ensemble converges.



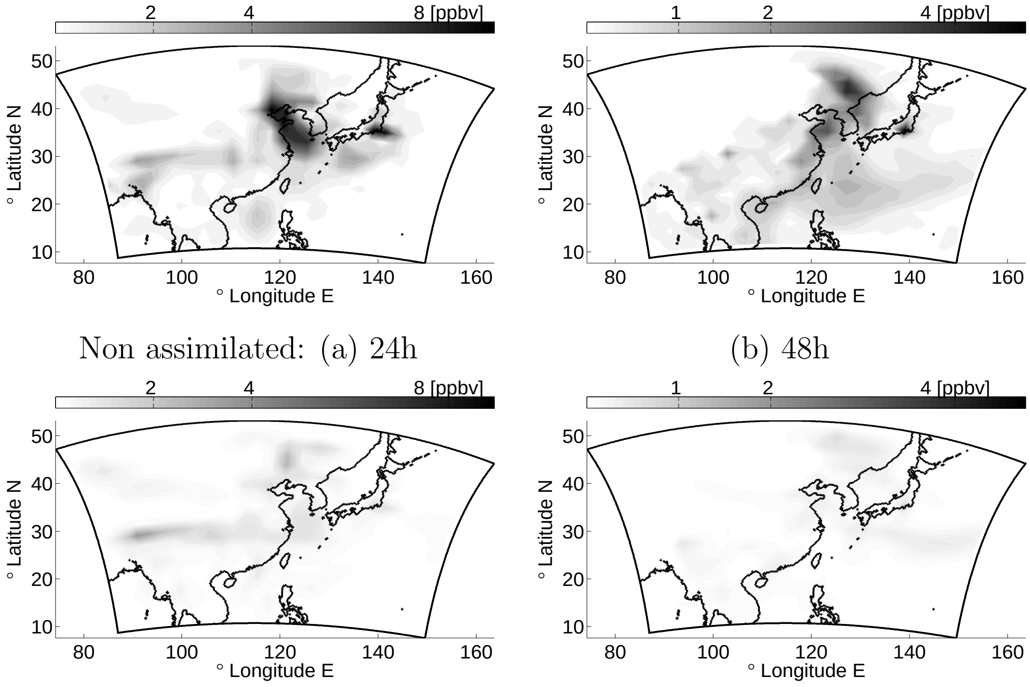
Non assimilated: (a) 24h (b) 48h

Assimilated: (c) 24h (d) 48h

Figure 5: Comparison of errors in O_3 non assimilated and assimilated fields (vertically averaged). The O_3 estimate is considerably improved by data assimilation.

error is noticed near the inflow boundary – East, North-East side of the domain – as the ensemble members and the reference solution are all determined by the same inflow boundary values.

We next study the effect of assimilating the emissions and the lateral boundary conditions together with the model states. Specifically, we append the model state a vector of correction factors for the emissions and the lateral boundary conditions. One scalar correction factor is added for each gridpoint and chemical species. A comparison between the errors in the state-only assimilated solutions and the errors in the combined state-parameter assimilated solutions are shown in Figure 10 (O_3), Figure 11 (NO_2), Figure 12 (CO), Figure 13 (HCHO), and Figure 14 (PAN). The errors fields are shown at ground level (first model layer) at the end of the assimilation window (24h) and at the end of the forecast window (48h). For all the chemical species the combined state-parameter estimation leads to improvements in analysis accuracy over the state-only estimation.



Non assimilated: (a) 24h (b) 48h

Assimilated: (c) 24h (d) 48h

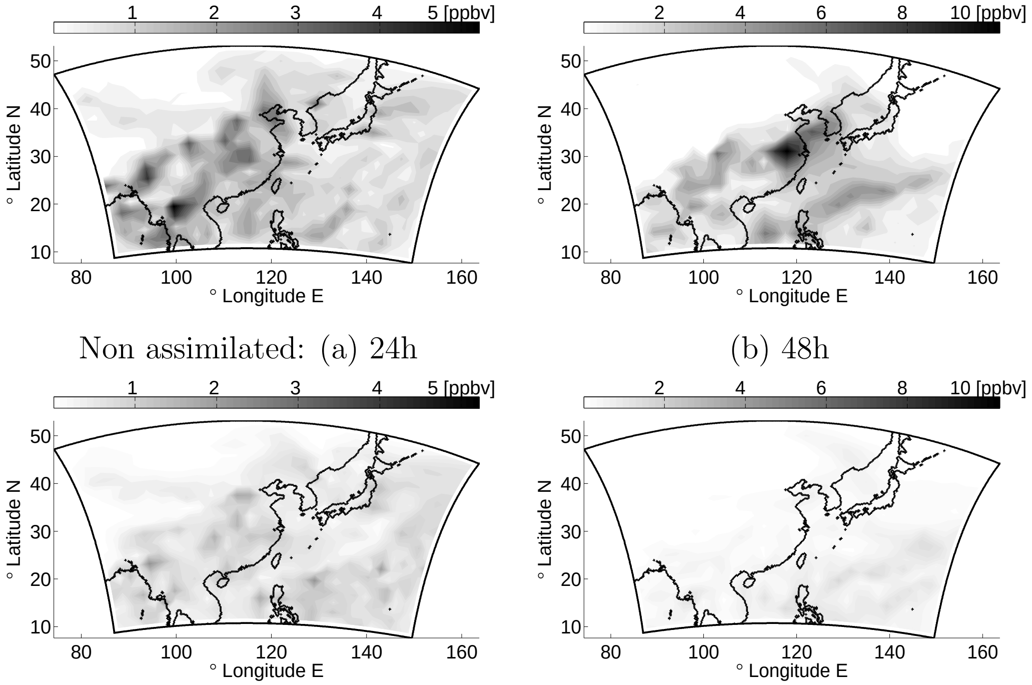
Figure 6: Comparison of errors in NO_2 non assimilated and assimilated fields (vertically averaged). The NO_2 estimate is considerably improved by data assimilation.

6 Conclusions and Future Work

In this paper we investigate the application of the ensemble Kalman filter technique to chemical data assimilation in atmospheric photochemical and transport (atmospheric) models.

To focus on the basic algorithmic issues the analysis is carried out in an idealized setting. A reference solution is considered to be the "true" state of the atmosphere and is used to generate artificial observations and to assess the quality of the analysis. Our analysis focuses on a verification region above Korea, chosen away from the boundaries in order to avoid the interference of boundary effects with the filter performance. An idealized ensemble is constructed by adding unbiased perturbations to the reference solution. Initial perturbations are constructed by the superposition of two processes. An autoregressive model of the background errors that account for flow-dependent correlations developed before the starting time of the assimilation. The second set of perturbations is along the dominant singular vectors computed with respect to the verification region above Korea. These perturbations undergo a maximum growth in 24 hours of evolution (among all directions in state space at the initial time).

In our experiment the ensemble bias remains insignificant at least for 48 hours. This



Non assimilated: (a) 24h

(b) 48h

Assimilated: (c) 24h

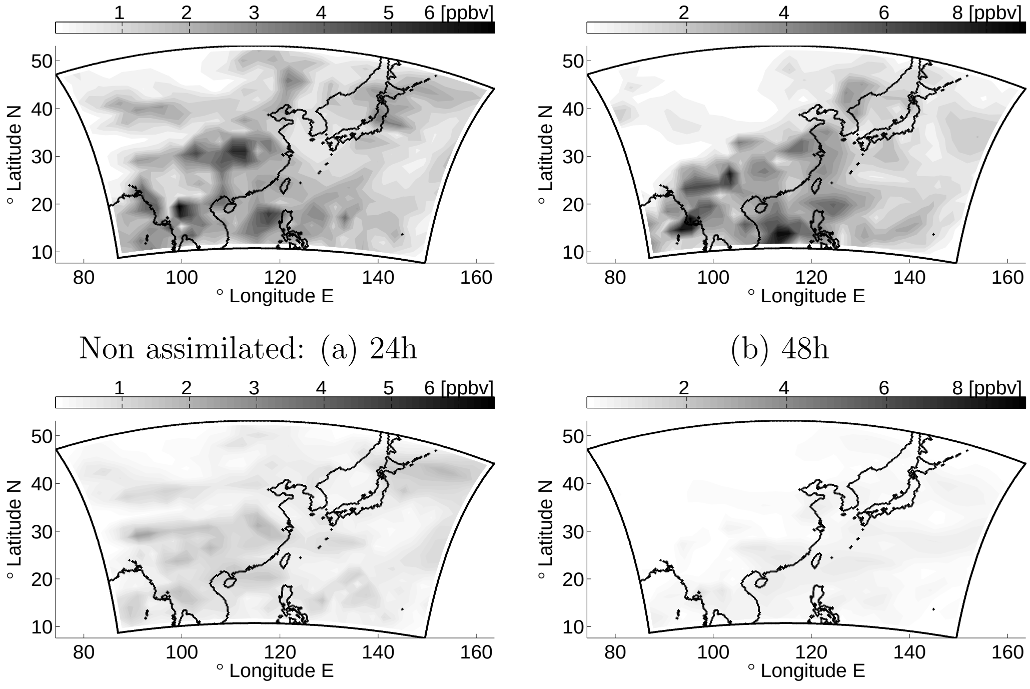
(d) 48h

Figure 7: Comparison of errors in HCHO non assimilated and assimilated fields (vertically averaged). The HCHO estimate is considerably improved by data assimilation.

characteristic greatly helps the EnKF data assimilation. The ensemble bias can become an issue in real/operational circumstances where the addition of perturbations may lead to negative concentrations; setting these perturbed concentrations to zero may result in biased estimates.

In the numerical experiments carried out here the ensemble spread is always positive, and there was no need for covariance inflation. The ensemble spread slowly decreases with time even without assimilation. The chemical kinetic system is stiff and therefore very stable – small perturbations are damped out quickly in time. Without simulating the atmospheric dynamics (meteorological fields are prescribed) this stiff effects are important. The decrease of the ensemble spread in time is different than what is typically observed in data assimilation with numerical weather prediction models. The shrinking spread may pose the danger of filter divergence if the spread becomes too small. Different approaches to covariance inflation will be discussed in the second part of this study. As atmospheric models are slowly evolving toward solving chemistry and dynamics together, future studies should consider ensemble data assimilation with integrated numerical weather prediction and chemistry models.

Ensemble size is an important parameter to represent correctly the distribution of error probabilities. Small ensembles underestimate the forecast errors, while large ensembles are



Non assimilated: (a) 24h (b) 48h

Assimilated: (c) 24h (d) 48h

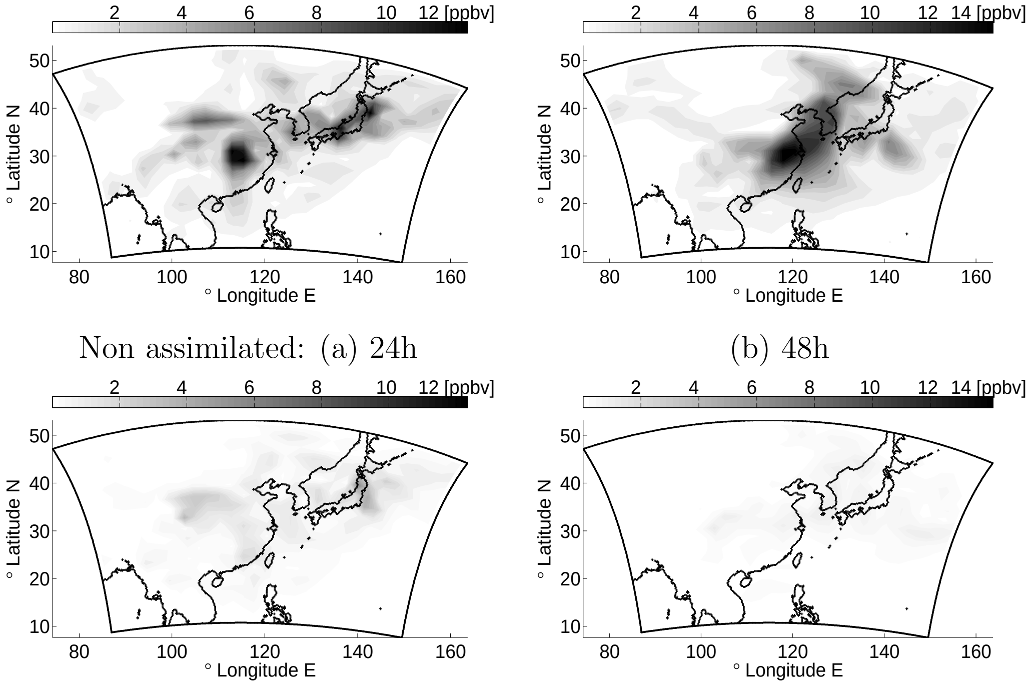
Figure 8: Comparison of errors in CO non assimilated and assimilated fields (vertically averaged). The CO estimate is considerably improved by data assimilation.

costly. In our idealized experiment 50 members proved to be a good choice, requiring no covariance inflation.

The concentration fields of both directly observed an unobserved species are considerably improved by EnKF data assimilation. Improvements are assessed by directly comparing the analyzed fields with the reference solution. Moreover, data assimilation has improved the forecast for at least 24 hours after assimilation. Improvements in the chemical species that are not directly observed shows that the ensemble is capable of correctly representing inter-species error correlations, established through the chemical interactions.

Additional improvements are possible by assimilating for state, emissions, and lateral boundary conditions simultaneously. Emission rates and lateral boundary conditions correction factors integration in the assimilation process is immediate and a straight forward process.

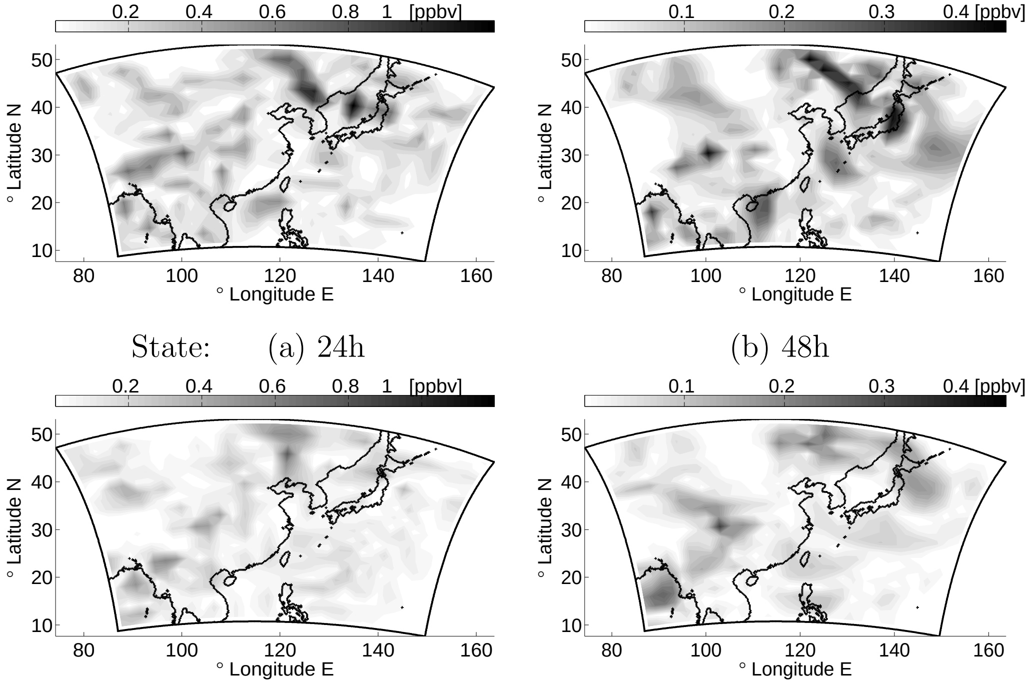
The EnKF assimilation scheme is very simple to implement with no changes to the original model code. Although no assessment on the computational cost was carried out, the scheme is well suited for parallel computation. The cost scales linearly with the size of the ensemble, pending that with the growing number of ensemble members the filter computational expense can be neglected.



Non assimilated: (a) 24h (b) 48h

Assimilated: (c) 24h (d) 48h

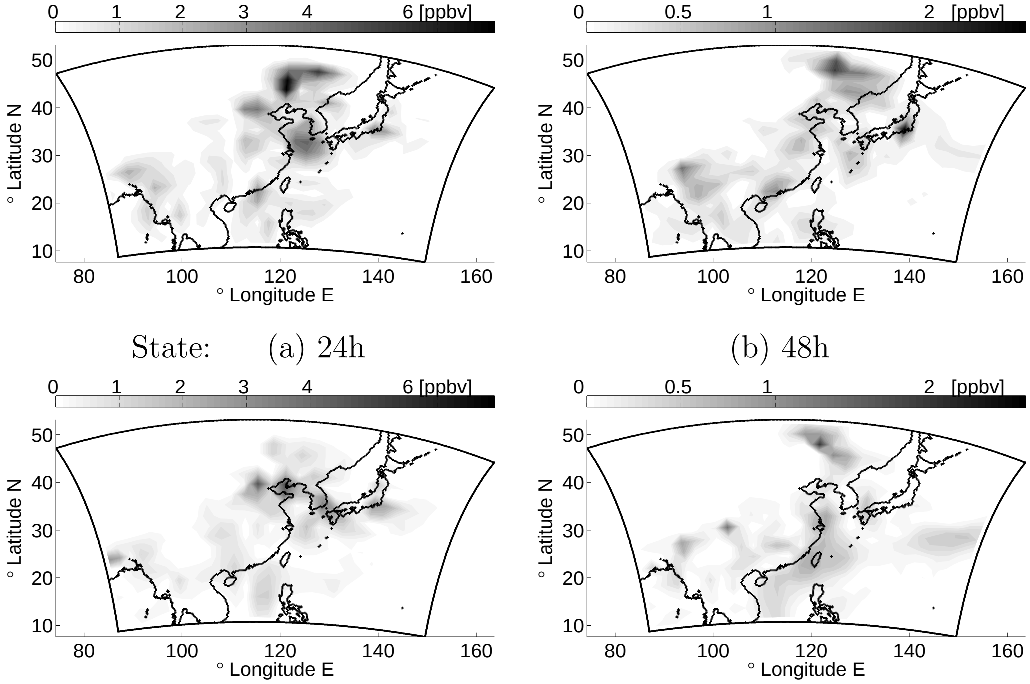
Figure 9: Comparison of errors in PAN non assimilated and assimilated fields (vertically averaged). The PANestimate is considerably improved by data assimilation.



State: (a) 24h (b) 48h

State + emissions: (c) 24h (d) 48h

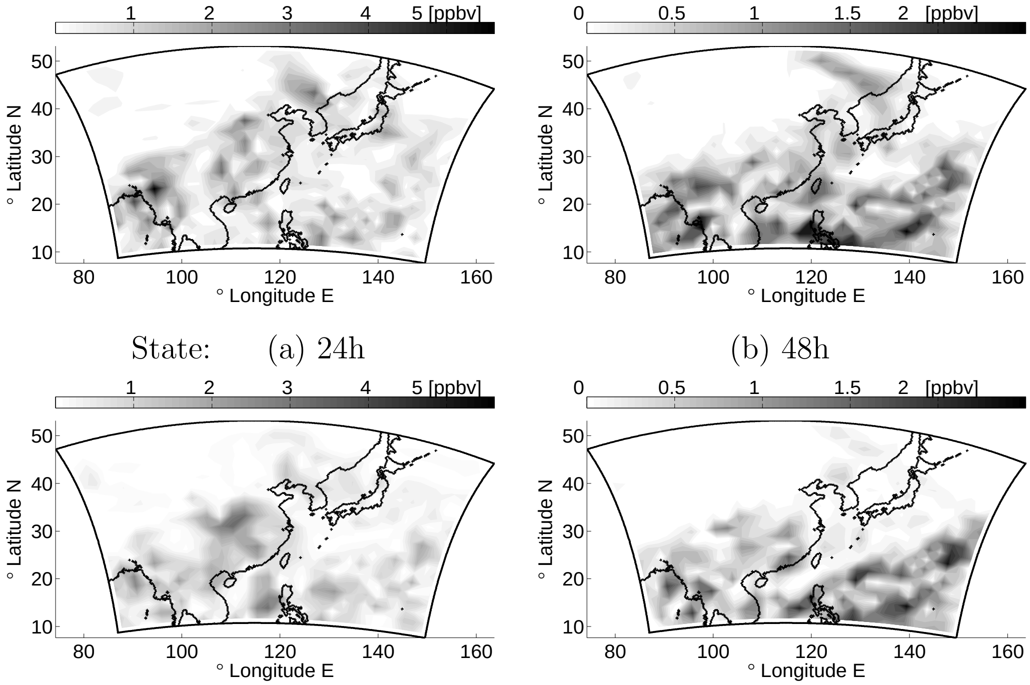
Figure 10: Ground O_3 state and state + emissions assimilated error levels



State: (a) 24h (b) 48h

State + emissions: (c) 24h (d) 48h

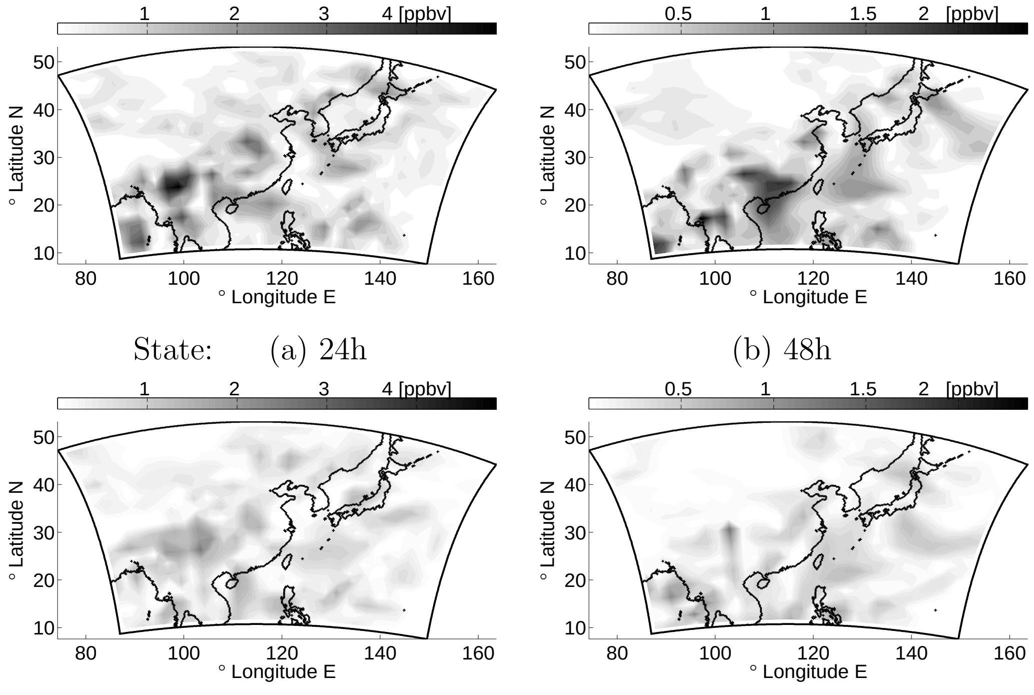
Figure 11: Ground NO_2 state and state + emissions assimilated error levels



State: (a) 24h (b) 48h

State + emissions: (c) 24h (d) 48h

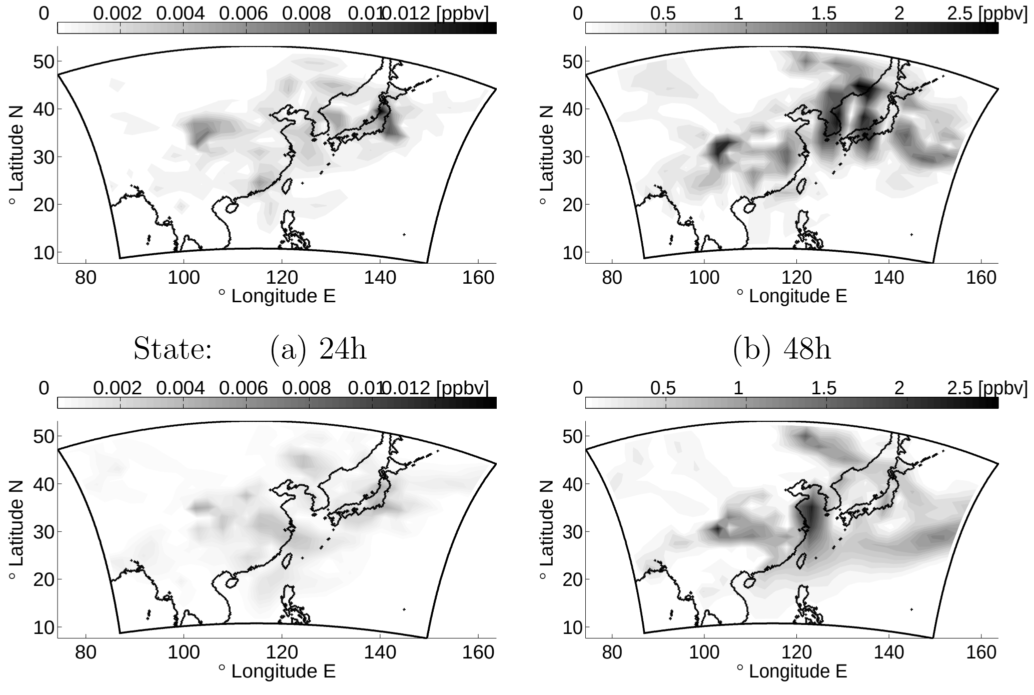
Figure 12: Ground HCHO state and state + emissions assimilated error levels



State: (a) 24h (b) 48h

State + emissions: (c) 24h (d) 48h

Figure 13: Ground CO state and state + emissions assimilated error levels



State: (a) 24h (b) 48h

State + emissions: (c) 24h (d) 48h

Figure 14: Ground PAN state and state + emissions assimilated error levels

Acknowledgments

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

References

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

N. Blond and R. Vautard. Three-dimensional ozone analyses and their use for short-term ozone forecasts. *Journal of Geophysical Research*, 109(D18):17303–+, 2004.

M. Buehner. Ensemble-derived stationary and flow-dependent background error covariances: Evaluation in a quasi-operational NWP setting. *Quarterly Journal of the Royal Meteorological Society*, 131:1013–1044, 2004.

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

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

W.P.L. Carter. Implementation of the SAPRC-99 Chemical Mechanism into the Models-3 Framework. Technical report, United States Environmental Protection Agency, January 2000.

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

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

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

E.M. Constantinescu, A. Sandu, T. Chai, and G.R. Carmichael. Ensemble-based Chemical Data Assimilation III: Filter Localization. *In preparation*, 2006c.

V. Damian, A. Sandu, M. Damian, F. Potra, and G.R. Carmichael. The kinetic preprocessor KPP - a software environment for solving chemical kinetics. *Computers and Chemical Engineering*, 26:1567–1579, 2002.

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

H. Elbern and H. Schmidt. A 4D-Var chemistry data assimilation scheme for Eulerian chemistry transport modeling. *Journal of Geophysical Research*, 104(5):18,583–18,598, 1999.

H. Elbern and H. Schmidt. Ozone episode analysis by 4D-Var chemistry data assimilation. *Journal of Geophysical Research*, 106(D4):3569–3590, 2001.

H. Elbern, H. Schmidt, and A. Ebel. Variational data assimilation for tropospheric chemistry modeling. *Journal of Geophysical Research*, 102(D13):15,967–15,985, 1997.

H. Elbern, H. Schmidt, and A. Ebel. Implementation of a parallel 4D-Var chemistry data assimilation scheme. *Environmental Management and Health*, 10:236–244, 1999.

H. Elbern, H. Schmidt, O. Talagrand, and A. Ebel. 4D-variational data assimilation with an adjoint air quality model for emission analysis. *Environmental Modeling and Software*, 15:539–548, 2000.

G. Evensen. Using the extended Kalman filter with a multi-layer quasi-geostrophic ocean model. *Journal of Geophysical Research*, 97(C11):17905–17924, 1992.

G. Evensen. Open boundary conditions for the extended Kalman filter with a quasi-geostrophic model. *Journal of Geophysical Research*, 98(C19):16529–16546, 1993.

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

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

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

M. Fisher. Assimilation Techniques(5):Approximate Kalman Filters and Singular Vectors. *European Centre for Medium-Range Weather Forecasts*, 2002.

M. Fisher and D.J. Lary. Lagrangian four-dimensional variational data assimilation of chemical species. *Quarterly Journal of the Royal Meteorological Society*, 121:1681–1704, 1995.

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

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

J.A. Hansen. Accounting for model error in ensemble-based state estimation and forecasting. *Monthly Weather Review*, 130:2373–2391, 2002.

A.W. Heemink and A.J. Segers. Modeling and prediction of environmental data in space and time using Kalman filtering. *Stochastic Environmental Research and Risk Assessment(SERRA)*, 16(3):225–240, 2002.

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

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

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

B.R. Hunt, E. Kalnay, E. Kostelich, E. Ott, D. Patil, T. Sauer, I. Szunyogh, J. Yorke, and A. Zimin. Four-dimensional ensemble Kalman filtering. *Tellus A*, 56:273–277, 2004.

K. Ide, Courtier P., Ghil M., and Lorenc A.C. Unified notation for data assimilation: Operational sequential and variational. *Journal of the Meteorological Society of Japan*, 75(1B): 181–189, 1997.

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

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

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

A.C. Lorenc. The potential of the ensemble Kalman filter for NWP – a comparison with 4D-Var. *Quarterly Journal of the Royal Meteorological Society*, 129(595):3183–3203, 2003.

K. Maschhoff and D. Sorensen. Parallel arpack home page, . URL `http://www.caam.rice.edu/∼kristyn/parpack_home.html`.

L. Menut. Adjoint modeling for atmospheric pollution process sensitivity at regional scale. *Journal of Geophysical Research*, 108(D17), 2003.

L. Menut, R. Vautard, M. Beekmann, and C Honor. Sensitivity of photochemical pollution using the adjoint of a simplified chemistry-transport model. *Journal of Geophysical Research - Atmospheres*, 105-D12(15):15,379–15,402, 2000.

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

H.L. Mitchell and P.L. Houtekamer. Ensemble size, balance, and model-error representation in an ensemble Kalman filter. *Monthly Weather Review*, 130:2791–2808, 2002.

F. Molteni, R. Buizza, T.N. Palmer, and T. Petroliagis. The new ECMWF ensemble prediction system: methodology and validation. *Quarterly Journal of the Royal Meteorological Society*, 122:73–119, 1996.

L.P. Riishojgaard. A direct way of specifying flow-dependent background error correlations for meteorological analysis systems. *Tellus A*, 50(1):42–42, 1998.

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

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

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

A. Sandu, Blom J.G., Spee E., Verwer J.G., Potra F.A., and Carmichael G.R. Benchmarking stiff ODE solvers for atmospheric chemistry equations II - Rosenbrock Solvers. *Atmospheric Environment*, 31:3,459–3,472, 1997.

R.W. Stewart. Multiple steady states in atmospheric chemistry. *Journal of Geophysical Research*, 98(.17):20601–20612, November 1993.

M. Van Loon, P.J.H. Builtjes, and A.J. Segers. Data assimilation of ozone in the atmospheric transport chemistry model LOTOS. *Environmental Modeling and Software*, 15:603–609, 2000.