Localized Ensemble Kalman Dynamic Data Assimilation for Atmospheric Chemistry*

Adrian Sandu[1], Emil M. Constantinescu[1], Gregory R. Carmichael[2], Tianfeng Chai[2], John H. Seinfeld[3], and Dacian Dăescu[4]

[1] Department of Computer Science, Virginia Polytechnic Institute and State University, Blacksburg, VA 24061.
{asandu, emconsta}@cs.vt.edu
[2] Center for Global and Regional Environmental Research, The University of Iowa, Iowa City, 52242-1297.
{gcarmich,tchai}@cgrer.uiowa.edu
[3] Department of Chemical Engineering, California Institute of Technology, Pasadena, CA 91125.
seinfeld@caltech.edu
[4] Department of Mathematics and Statistics, Portland State University.
daescu@pdx.edu

Abstract. The task of providing an optimal analysis of the state of the atmosphere requires the development of dynamic data-driven systems (DDDAS) that efficiently integrate the observational data and the models. Data assimilation, the dynamic incorporation of additional data into an executing application, is an essential DDDAS concept with wide applicability. In this paper we discuss practical aspects of nonlinear ensemble Kalman data assimilation applied to atmospheric chemical transport models. We highlight the challenges encountered in this approach such as filter divergence and spurious corrections, and propose solutions to overcome them, such as background covariance inflation and filter localization. The predictability is further improved by including model parameters in the assimilation process. Results for a large scale simulation of air pollution in North-East United States illustrate the potential of nonlinear ensemble techniques to assimilate chemical observations.

1 Introduction

Our ability to anticipate and manage changes in atmospheric pollutant concentrations relies on an accurate representation of the chemical state of the atmosphere. As our fundamental understanding of atmospheric chemistry advances, novel data assimilation tools are needed to integrate observational data and models together to provide the best estimate of the evolving chemical state of the atmosphere. The ability to dynamically incorporate additional data into an executing application is a fundamental DDDAS concept (http://www.cise.nsf.gov/dddas.) We refer to this process as data assimilation. Data assimilation has proved vital for meteorological forecasting.

* This work was supported by the National Science Foundation through the award NSF ITR AP&IM 0205198 managed by Dr. Frederica Darema.

Y. Shi et al. (Eds.): ICCS 2007, Part I, LNCS 4487, pp. 1018–1025, 2007.
© Springer-Verlag Berlin Heidelberg 2007

In this paper we focus on the particular challenges that arise in the application of nonlinear ensemble filter data assimilation to atmospheric chemical transport models (CTMs). Atmospheric CTMs solve the mass-balance equations for concentrations of trace species to determine the fate of pollutants in the atmosphere [16]. The CTM operator, $\mathcal{M}$, will be denoted compactly as

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

where c represents the modeled species concentration, c^{in} the inflow numerical boundary conditions, u the wind fields, Q the surface emission rates, and the subscript denotes the time index. In our numerical experiments, we use the Sulfur Transport Eulerian Model (STEM) [16], a state-of-the-art atmospheric CTM.

Kalman filters [12] provide a stochastic approach to the data assimilation problem. The filtering theory is described in Jazwinski [10] and the applications to atmospheric modeling in [13]. The computational burden associated with the filtering process has prevented the implementation of the full Kalman filter for large-scale models. Ensemble Kalman filters (EnKF) [2,5] may be used to facilitate the practical implementation as shown by van Loon et al. [18]. There are two major difficulties that arise in EnKF data assimilation applied to CTMs: (1) CTMs have stiff components [15] that cause the filter to *diverge* [7] due to the lack of ensemble spread and (2) the ensemble size is typically small in order to be computationally tractable and this leads to filter *spurious corrections* due to sampling errors. Kalman filter data assimilation has been discussed for DDDAS in another context by Jun and Bernstein [11].

This paper addresses the following issues: (1) *Background covariance inflation* is investigated in order to avoid *filter divergence*, (2) *localization* is used to prevent spurious filter corrections caused by small ensembles, and (3) *parameters are assimilated together with the model states* in order to reduce the model errors and improve the forecast. The paper is organized as follows. Section 2 presents the ensemble Kalman data assimilation technique, Section 3 illustrates the use of the tools in a data assimilation test, and Section 4 summarizes our results.

2 Ensemble Kalman Filter Data Assimilation

Consider a nonlinear model $c_i = \mathcal{M}_{t_0 \to t_i}(c_0)$ that advances the state from the initial time t_0 to future times t_i $(i \geq 1)$. The model state c_i at t_i $(i \geq 0)$ is an approximation of "true" state of the system c^t_i at t_i (more exactly c^t_i is the system state projected onto the model space space). Observations y_i are available at times t_i and are corrupted by measurement and representativeness errors ε_i (assumed Gaussian with mean zero and covariance $\mathbb{R}_i$), $y_i = \mathcal{H}_i\left(c^t_i\right) + \varepsilon_i$. Here $\mathcal{H}_i$ is an operator that maps the model state to observations.

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

The (ensemble) Kalman filter estimates the true state c^t using the information from the current best estimate c^f (the "forecast" or the background state) and the observations y. The optimal estimate c^a (the "analysis" state) is obtained as

a linear combination of the forecast and observations that minimize the variance of the analysis (P^a)

$$c^a = c^f + P^f H^T \left(H P^f H^T + \mathbb{R}\right)^{-1} \left(y - \mathcal{H}(c^f)\right) = c^f + K\left(y - \mathcal{H}(c^f)\right). \quad (2)$$

The forecast covariance P^f is estimated from an ensemble of runs (which produces an ensemble of E model states $c^f(e)$, $e = 1, \cdots, E$). The analysis formula (2) is applied to each member to obtain an analyzed ensemble. 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. The filter can be described as

$$c_i^f(e) = \mathcal{M}\big(c_{i-1}^a(e)\big),\ c_i^a(e) = \ c_i^f(e) + K_i\left(y_i - \mathcal{H}_i\big(c_i^f(e)\big)\right). \quad (3)$$

The results presented in this paper are obtained with the practical EnKF implementation discussed by Evensen [5].

2.1 The Localization of EnKF (LEnKF)

The practical Kalman filter implementation employs a small ensemble of Monte Carlo simulations in order to approximate the background covariance (P^f). In its initial formulation, EnKF may suffer from spurious correlations caused by sub-sampling errors in the background covariance estimates. This allows for observations to incorrectly impact remote model states. The filter *localization* introduces a restriction on the correction magnitude based on its remoteness.

One way to impose localization in EnKF is to apply a decorrelation function ρ, that decreases with distance, to the background covariance. Following [8], the EnKF relation (2) with some simplifying assumptions becomes

$$c_i^a = c_i^f + \rho(D^c) \circ P_i^f\, H_i^T \left(\rho(D^y) \circ \left(H_i\, P_i^f\, H_i^T\right) + \mathbb{R}_i\right)^{-1} \left(y_i - \mathcal{H}_i(c_i^f)\right), \quad (4)$$

where $D^{\{c,y\}}$ are distance matrices with positive elements ($d_{i,j} \geq 0$), and $0 \leq \rho(d_{i,j}) \leq 1$, $\rho(0) = 1$, $\forall i, j$. The decorrelation function ρ is applied to the distance matrix and produces a decorrelation matrix (decreasing with the distance). The operation '$\circ$' denotes the Schur product that applies elementwise $\rho(D)$ to the projected covariance matrices $P^f\, H^T$ and $H\, P^f\, H^T$, respectively. Here, D^y is calculated as the distance among the observation sites, and D^c contains the distance from each state variable to each observation site.

We considered a Gaussian distribution for the decorrelation function, ρ. Since our model generally has an anisotropic horizontal–vertical flow, we consider the two correlation components (and factors, δ) separately:

$$\rho\left(D^h, D^v\right) = \exp\left[-\left(D^h/\delta^h\right)^2 - \left(D^v/\delta^v\right)^2\right], \quad (5)$$

where D^h, D^v, δ^h, δ^v are the horizontal and vertical components. The horizontal correlation-distance relationship is determined through the NMC method [14]. The horizontal NMC determined correlations were fitted with a Gaussian distribution, $\delta^h = 270$ km. The vertical correlation was chosen as $\delta^v = 5$ grid points.

2.2 Preventing Filter Divergence

The "textbook application" of EnKF [5] may lead to filter divergence [7]: EnKF shows a decreasing ability to correct ensemble states toward the observations. This is due to an 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 solution is to increase the covariance of the ensemble and therefore decrease the filter's confidence in the model. The following are several ways to "inflate" the ensemble covariance.

The first method is the *additive inflation* [4], where the model errors are simulated by adding uncorrelated noise (denoted by η) to the model (η_-) or analysis (η_+) results. This increases the diagonal entries of the ensemble covariance. Since the correlation of the model errors is to a large extent unknown, white noise is typically chosen. With the notation (3), $c_i^f(e) = \mathcal{M}\big(c_{i-1}^a(e) + \eta_-(e)\big) + \eta_+(e)$. The second method is the *multiplicative inflation* [1], where each member's deviation from the ensemble mean is multiplied by a constant ($\gamma > 1$). This increases each entry of the ensemble covariance by that constant squared (γ^2). The ensemble can be inflated before ($\gamma_{\{-\}}$) or after ($\gamma_{\{+\}}$) filtering: $c_i^{\{f/a\}}(e) \leftarrow \big\langle c_i^{\{f/a\}} \big\rangle + \gamma_{\{-/+\}}$, where $\langle\cdot\rangle$ denotes the ensemble average.

A third possibility for covariance inflation is through perturbations applied to key model parameters, and we refer to it as *model-specific inflation*. This approach focuses on sources of uncertainty that are specific to each model (for instance in CTMs: boundary conditions, emissions, and meteorological fields). With the notation (3) and considering p as a set of model parameters, the model-specific inflation can be written as $c_i^f(e) = \mathcal{M}\big(c_{i-1}^a(e), \alpha_{i-1}(e)\, p_{i-1}\big)$, where $\alpha(e)$ are random perturbation factors of the model parameters.

2.3 Inflation Localization

The traditional approach to covariance inflation increases the spread of the ensemble equally throughout the computational domain. In the LEnKF framework, the corrections are restricted to a region that is rich in observations. These states are corrected and their variance is reduced, while the remote states (i.e., the states that are relatively far from the observations' locations) maintain their initial variation which is potentially reduced only by the model evolution. The spread of the ensemble at the remote states may be increased to unreasonably large values through successive inflation steps. And thus, the covariance inflation needs to be restricted in order to avoid the over-inflation of the remote states.

A sensible inflation restriction can be based on the localization operator, $\rho(D)$, which is applied in the same way as for the covariance localization. The localized multiplicative inflation factor, γ_ℓ, is given by

$$\gamma_\ell\,(i,j,k) = \max\left\{\rho\left(D^c\,(i,j,k)\right)\right\}(\gamma - 1) + 1\,, \tag{6}$$

where γ is the (non-localized) multiplicative inflation factor and i, j, k refer to the spatial coordinates. In this way, the localized inflation increases the ensemble spread only in the information-rich regions where filter divergence can occur.

3 Numerical Results

The test case is a real-life simulation of air pollution in North–Eastern U.S. in July 2004 as shown in Figure 1.a (the dash-dotted line delimits the domain). The observations used for data assimilation are the ground-level ozone (O_3) measurements taken during the ICARTT [9,17] campaign in 2004 (which also includes the initial concentrations, meteorological fields, boundary values, and emission rates). Figure 1.a shows the location of the ground stations (340 in total) that measured ozone concentrations and an ozonesonde (not used in the assimilation process). The computational domain covers $1500 \times 1320 \times 20$ Km with a horizontal resolution of 60×60 Km and a variable vertical resolution.

The simulations are started at 0 GMT July 20^{th} with a four hour initialization step ([-4,0] hours). The "best guess" of the state of the atmosphere at 0 GMT July 20^{th} is used to initialize the deterministic solution. The ensemble members are formed by adding a set of unbiased perturbations to the best guess, and then evolving each member to 4 GMT July 20^{th}. The perturbation is formed according to an AR model [3] making it flow dependent. The 24 hours assimilation window starts at 4 GMT July 20^{th} (denoted by [1,24] hours). Observations are available at each integer hour in this window, i.e., at 1, 2, ..., 24 hours (Figure 1.a). EnKF adjusts the concentration fields of 66 "control" chemical species in each grid point of the domain every hour using (2). The ensemble size was chosen to be 50 members (a typical size in NWP). A 24 hour forecast window is also considered to start at 4 GMT July 21^{st} (denoted by [24,48] hours).

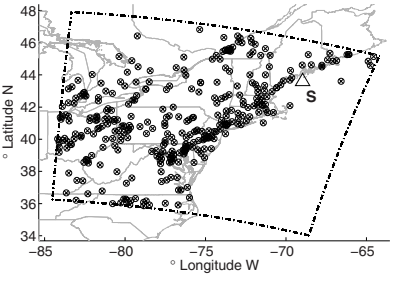


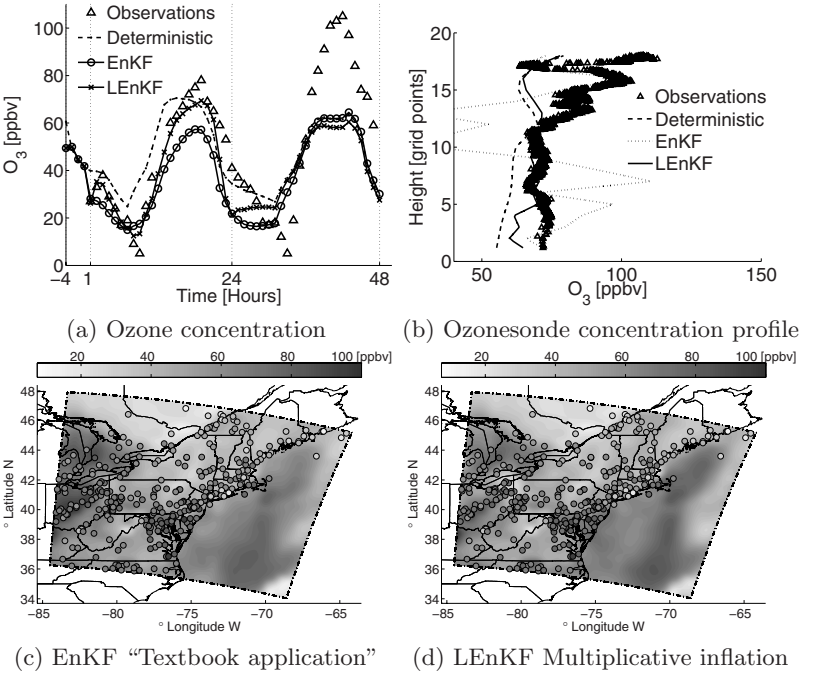
Fig. 1. Ground measuring stations in support of the ICARTT campaign (340 in total) and the ozonesonde (S) launch location

The performance of each data assimilation experiment is measured by the R^2 correlation factor (correlation^2) between the observation and the model solution. The R^2 correlation results between the observations and model values for all the numerical experiments are shown in Table 1. The deterministic (best guess) solution yields an R^2 of 0.24 in the analysis and 0.28 in the forecast windows. In Table 1 we also show the results for a 4D-Var experiment.

Figure 2.a shows the O_3 concentration measured at a Washington DC station and predicted by the EnKF and LEnKF with model-specific inflation. Figure 2.b shows the ozone concentration profile measured by the ozonesonde for the EnKF and LEnKF with additive inflation. Two effects are clear for the "textbook" EnKF. The filter diverges after about 12 hours (2.a), and spurious corrections are made at higher altitudes (2.b), as the distance from the observation (ground) sites increases. The vertical profile in Figure 2.b shows great improvement in the analyzed solution of LEnKF. The results in Table 1 confirm the benefits of localization by dramatically improving the analysis and forecast fit.

Table 1. The R^2 measure of model-observations match in the assimilation and forecast windows for EnKF, 4D-Var, and LEnKF. (Multiplicative inflation: $\gamma_- \leq 4$, $\gamma_+ \leq 4$; Model-specific inflation: 10% emissions, 10% boundaries, 3% wind).

Method & Details	R^2 analysis	R^2 forecast
Deterministic solution, no assimilation	0.24	0.28
EnKF, "textbook application"	0.38	0.30
4D-Var - 50 iterations	0.52	0.29
LEnKF, model-specific inflation	0.88	0.32
LEnKF, multiplicative inflation	0.82	0.32
LEnKF, additive inflation	0.92	0.31
LEnKF with parameter assimilation, and multiplicative localized inflation	0.89	0.41



(a) Ozone concentration (b) Ozonesonde concentration profile

(c) EnKF "Textbook application" (d) LEnKF Multiplicative inflation

Fig. 2. Ozone concentration (a) measured at a Washington DC station (ICARTT ID: 510590030) and predicted by EnKF ("textbook") and LEnKF with model-specific inflation, and (b) measured by the ozonesonde for EnKF and LEnKF. Ground level ozone concentration field (c,d) at 14 EDT in the forecast window measured by the ICARTT stations (shown in color coded filled circles) and predicted EnKF and LEnKF.

3.1 Joint State and Parameter Assimilation

In regional CTMs the influence of the initial conditions is rapidly diminishing with time, and the concentration fields are "driven" by emissions and by lateral boundary conditions. Since both of them are generally poorly known, it is of considerable interest to improve their values using information from observations. In this setting we have to solve a joint state-parameter assimilation problem [6].

The emission rates and lateral boundary conditions are multiplied by specific correction coefficients, α. These correction coefficients are appended to the model state. The LEnKF data assimilation is then carried out with the augmented model state. With the notation (1), LEnKF is applied to

$$\left[c_i^f \;\; \alpha_i^{\{1,2\}} \right]^T = \left[\mathcal{M}_{t_{i-1} \to t_i} \left(c_{i-1}^a, u_{i-1}, \alpha_{i-1}^{\{1\}} c_{i-1}^{\text{in}}, \alpha_{i-1}^{\{2\}} Q_{i-1} \right) \;\; \alpha_{i-1} \right]^T .$$

For α, we consider a different correction for each species and each gridpoint. The initial ensemble of correction factors is an independent set of normal variables and the localization is done in the same way as in the state-only case.

The R^2 after LEnKF data assimilation for combined state and emission correction coefficients (presented in Table 1) show improvements in both the forecast and the analysis windows. Figures 2.(c,d) show the ground level ozone field concentration at 14 EDT in the forecast window measured by the ICARTT stations, EnKF with state corrections and LEnKF with joint state-parameter corrections. In the LEnKF case under consideration the addition of the correction parameters to the assimilation process improves the assimilated solution (especially on the inflow boundary (West)).

4 Conclusions

This paper discusses some of the challenges associated with the application of nonlinear ensemble filtering data assimilation to atmospheric CTMs. Three aspects are analyzed in this study: *filter divergence* - CTMs tend to dampen perturbations; *spurious corrections* - small ensemble size cause wrong increments, and *model parametrization errors* - without correcting model errors in the analysis, correcting the state only does not help in improving the forecast accuracy.

Experiments showed that the filter diverges quickly. The influence of the initial conditions fades in time as the fields are largely determined by emissions and by lateral boundary conditions. Consequently, the initial spread of the ensemble is diminished in time. Moreover, stiff systems (like chemistry) are stable - small perturbations are damped out quickly in time. In order to prevent filter divergence, the spread of the ensemble needs to be explicitly increased. We investigated three approaches to ensemble covariance inflation among which model-specific inflation is the most intuitive. The "localization" of EnKF is needed in order to avoid the spurious corrections noticed in the "textbook" application. The correlation distances are approximated using the NMC method. Furthermore, covariance localization prevents over-inflation of the states that are

remote from observation. LEnKF increased both the accuracy of the analysis and forecast at the observation sites and at distant locations (from the observations).

Since the solution of a regional CTM is largely influenced by uncertain lateral boundary conditions and by uncertain emissions it is of great importance to adjust these parameters through data assimilation. The assimilation of emissions and boundary conditions visibly improves the quality of the analysis.

References

1. J.L. Anderson. An ensemble adjustment Kalman filter for data assimilation. *Mon. Wea. Rev*, 129:2884–2903, 2001.
2. G. Burgers, P.J. van Leeuwen, and G. Evensen. Analysis scheme in the ensemble Kalman Filter . *Mon. Wea. Rev*, 126:1719–1724, 1998.
3. E.M. Constantinescu, T. Chai, A. Sandu, and G.R. Carmichael. Autoregressive models of background errors for chemical data assimilation. *To appear in J. Geophys. Res.*, 2006.
4. M. Corazza, E. Kalnay, and D. Patil. Use of the breeding technique to estimate the shape of the analysis "errors of the day". *Nonl. Pr. Geophys.*, 10:233–243, 2002.
5. G. Evensen. The ensemble Kalman filter: theoretical formulation and practical implementation. *Ocean Dynamics*, 53, 2003.
6. G. Evensen. The combined parameter and state estimation problem. *Submitted to Ocean Dynamics*, 2005.
7. P.L. Houtekamer and H.L. Mitchell. Data assimilation using an ensemble Kalman filter technique . *Mon. Wea. Rev*, 126:796–811, 1998.
8. P.L. Houtekamer and H.L. Mitchell. A sequential ensemble Kalman filter for atmospheric data assimilation . *Mon. Wea. Rev*, 129:123–137, 2001.
9. ICARTT. ICARTT home page:http://www.al.noaa.gov/ICARTT .
10. A.H. Jazwinski. *Stochastic Processes and Filtering Theory*. Academic Press, 1970.
11. B.-E. Jun and D.S. Bernstein Least-correlation estimates for errors-in-variables models . *Int'l J. Adaptive Control and Signal Processing*, 20(7):337–351, 2006.
12. R.E. Kalman. A new approach to linear filtering and prediction problems. *Trans. ASME, Ser. D: J. Basic Eng.*, 83:95–108, 1960.
13. R. Menard, S.E. Cohn, L.-P. Chang, and P.M. Lyster. Stratospheric assimilation of chemical tracer observations using a Kalman filter. Part I: Formulation. *Mon. Wea. Rev*, 128:2654–2671, 2000.
14. D.F. Parrish and J.C. Derber. The national meteorological center's spectral statistical-interpolation analysis system. *Mon. Wea. Rev*, (120):1747–1763, 1992.
15. A. Sandu, J.G. Blom, E. Spee, J.G. Verwer, F.A. Potra, and G.R. Carmichael. Benchmarking stiff ODE solvers for atmospheric chemistry equations II - Rosenbrock solvers. *Atm. Env.*, 31:3,459–3,472, 1997.
16. A. Sandu, D. Daescu, G.R. Carmichael, and T. Chai. Adjoint sensitivity analysis of regional air quality models. *J. of Comp. Phy.*, 204:222–252, 2005.
17. Y. Tang et al. The influence of lateral and top boundary conditions on regional air quality prediction: a multi-scale study coupling regional and global chemical transport models. *Submitted to J. Geophys. Res.*, 2006.
18. M. van Loon, P.J.H. Builtjes, and A.J. Segers. Data assimilation of ozone in the atmospheric transport chemistry model LOTOS. *Env. Model. and Soft.*, 15:603–609, 2000.