Ensemble–Based Data Assimilation for Atmospheric Chemical Transport Models *

Adrian Sandu[1] Emil M. Constantinescu[1], Wenyuan Liao[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, liao}@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 (D^{3}AS) that efficiently integrate the observational data and the models. In this paper we discuss fundamental aspects of nonlinear ensemble data assimilation applied to atmospheric chemical transport models. We formulate autoregressive models for the background errors and show how these models are capable of capturing flow dependent correlations. Total energy singular vectors describe the directions of maximum errors growth and are used to initialize the ensembles. We highlight the challenges encountered in the computation of singular vectors in the presence of stiff chemistry and propose solutions to overcome them. Results for a large scale simulation of air pollution in East Asia illustrate the potential of nonlinear ensemble techniques to assimilate chemical observations.

Keywords: Dynamic data-driven applications and systems (D^{3}AS), data assimilation, background covariance, ensemble Kalman filter, total energy singular vectors, autoregressive processes.

1 Introduction

Our ability to anticipate and manage changes in atmospheric pollutant concentrations relies on an accurate representation of the chemical state of the

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

V.S. Sunderam et al. (Eds.): ICCS 2005, LNCS 3515, pp. 648–655, 2005.

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, physically consistent estimate of the evolving chemical state of the atmosphere. Data assimilation is vital for meteorological forecasting and has started to be applied in chemical transport modeling [7, 10, 20, 24].

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). The distinguishing feature of CTMs is the presence of nonlinear and stiff chemical interactions occurring at characteristic time scales that are typically much shorter than the transport time scales. CTMs propagate the model state forward in time from the "initial" state $x(t_0)$ to the "final" state $x(t_\mathrm{F})$ (1). Perturbations (small errors) evolve according to the tangent linear model (2) and adjoint variables according to the adjoint model (3):

$$x(t_\mathrm{F}) = \mathcal{M}_{t_0 \to t_\mathrm{F}} \ (x(t_0)) \tag{1}$$

$$\delta x(t_\mathrm{F}) = M_{t_0 \to t_\mathrm{F}} \ \delta x(t_0) \tag{2}$$

$$\lambda(t_0) = M^*_{t_\mathrm{F} \to t_0} \lambda(t_\mathrm{F}) \ . \tag{3}$$

Here $\mathcal{M}$, M, and M^* denote the solution operators of the CTM, the tangent linear, and the adjoint models respectively. The error covariance matrix evolves from $P(t_0)$ to $P(t_\mathrm{F})$ according to

$$P(t_\mathrm{F}) = M_{t_0 \to t_\mathrm{F}} \, P(t_0) \, M^*_{t_\mathrm{F} \to t_0} + Q \ , \tag{4}$$

where Q is the covariance of the model errors.

Kalman filter techniques [16] provide a stochastic approach to the data assimilation problem. The filtering theory is described in Jazwinski [15] and the applications to atmospheric modeling in [6, 19]. The computational burden associated to the filtering process has prevented the implementation of the full Kalman filter for large-scale transport-chemistry models. Ensemble Kalman filter techniques [8, 9, 13] may be used to facilitate the practical implementation as shown by van Loon et al. [24].

In the ensemble implementation of the Kalman filter [9] the statistics are represented by the ensemble mean and covariance. These statistics depend strongly on the background (initial) ensemble statistics $x(t_0)$ and $P(t_0)$. Since the probability density of the background state is not known exactly, it needs to be modeled. Previous efforts to develop flow dependent background covariance models are due to Riishojgaard [21], Hamill et al. [11], Houtekamer et. al. [14], and Buehner et. al. [1].

This paper brings the following new elements:

1. The background errors are modeled using autoregressive processes. Such models are computationally inexpensive and capture the error correlations along the flow lines.
2. Total energy singular vectors (TESVs) are calculated for chemically reactive flows. TESVs are the directions of maximum error growth over a finite time interval.

3. The above techniques are used to initialize the ensembles in a large scale data assimilation problem.

The paper is organized as follows. Section 2 presents the background error models and the calculation of TESVs. Section 3 illustrates the use of the tools in a large scale data assimilation test, and Section 4 summarizes the results of this research.

2 Construction of the Initial Ensemble

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 this section we consider autoregressive models for background errors and discuss the construction of TESVs. A more detailed discussion can be found in [5] and [18].

2.1 Modeling the Background Errors

The background state x^{B} is represented as the sum of the average state $\overline{x}^{\mathrm{B}}$ plus an error (uncertainty) field δx^{B}, $x^{\mathrm{B}} = \overline{x}^{\mathrm{B}} + \delta x^{\mathrm{B}}$. The error field has zero mean $\langle \delta x^{\mathrm{B}} \rangle = 0$, and background covariance $B = \langle \delta x^{\mathrm{B}} \left(\delta x^{\mathrm{B}}\right)^T \rangle$. Our basic assumption is that the background state errors form a multilateral autoregressive (AR) process [12] of the form

$$\delta x^{\mathrm{B}}_{i,j,k} = \alpha_{\pm 1} \delta x^{\mathrm{B}}_{i\pm 1,j,k} + \beta_{\pm 1} \delta x^{\mathrm{B}}_{i,j\pm 1,k} + \gamma_{\pm 1} \delta x^{\mathrm{B}}_{i,j,k\pm 1} + \sigma_{i,j,k}\, \xi_{i,j,k} \ . \qquad (5)$$

Here (i, j, k) are gridpoint indices on a 3 dimensional (structured) grid. The model (5) captures the correlations among neighboring grid points, with α, β ,γ representing the correlation coefficients in the x, y and z directions respectively. The last term represents the additional uncertainty at each grid point, with $\xi \in \mathcal{N}(0, 1)$ normal random variables and σ local error variances. The motivation behind multilateral AR models is the fact that (5), – with proper coefficients – can be regarded as a finite difference approximation of the advection-diffusion equation.

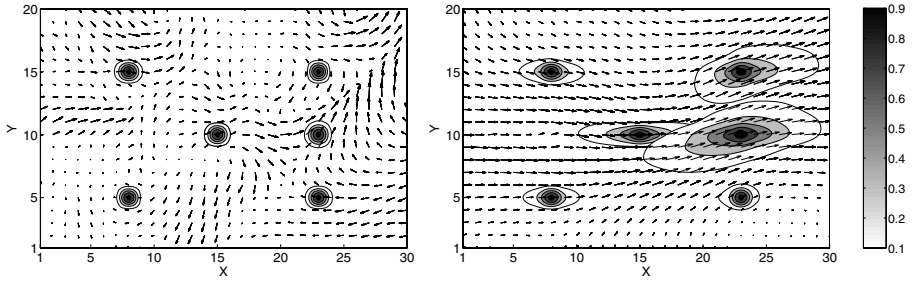
The AR process (5) can be represented compactly as

$$A\, \delta x = \xi \ . \qquad (6)$$

Note that A is a very sparse matrix. The background error covariance matrix is $B = A^{-1} A^{-T}$, and the correlation matrix is $D = \mathrm{diag}(B)^{-1/2} B \mathrm{diag}(B)^{-1/2}$.

Constant correlation coefficients α, β, γ imply fixed spatial directional correlation whereas variable coefficients may be used to capture flow dependent correlations. In this paper we use the scaled wind speeds u, v, and w as correlation coefficients. For example, the correlation coefficients in the x direction are given by $\alpha_{i,j,k} = u_{i,j,k} / \max_{i,j,k}(\sqrt{u^2_{i,j,k} + v^2_{i,j,k} + w^2_{i,j,k}})$. This approach leads to very well conditioned covariance matrices B.

To illustrate the autoregressive models we consider the wind fields over East Asia on 0 GMT, March 1st, 2001, corresponding to the Trace-P field campaign [3]. An autoregressive model (5) of background errors is constructed using flow dependent coefficients (scaled wind velocities). Top views of the spatial correlations of the resulting uncertainty fields are shown in Figure 1 for several gridpoints located on the ground layer (a) and on the top layer (b). The correlations match the shape and magnitude of the wind field. Note that the wind speed near the ground is smaller than at the top and this is reflected by the correlations.



(a) Ground Level (b) Top Level

Fig. 1. Background error correlations for the Trace-P wind fields on March 1, 2001

2.2 Total Energy Singular Vectors

Total energy singular vectors (TESVs) 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 ("energy") norms. The ratio between perturbation energies at the final (t_{F}) and initial time (t_0) offers a measure of error growth:

$$\sigma^2 = \frac{\|\delta x(t_{\mathrm{F}})\|_B^2}{\|\delta x(t_0)\|_A^2} = \frac{\langle \delta x(t_0), M^*_{t_{\mathrm{F}} \to t_0} B M_{t_0 \to t_{\mathrm{F}}} \delta x(t_0) \rangle}{\langle \delta x(t_0), A \delta x(t_0) \rangle} \tag{7}$$

Here A is a positive definite and B a positive semidefinite matrix. In (7) we use the fact that perturbations evolve in time according to the dynamics of the tangent linear model (2). TESVs 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 (7). These directions are the solutions of the following generalized eigenvalue problem:

$$M^*_{t_{\mathrm{F}} \to t_0} \, B \, M_{t_0 \to t_{\mathrm{F}}} \, s_k(t_0) = \sigma_k^2 \, A \, s_k(t_0) \tag{8}$$

The left side of (8) involves one integration with the tangent linear model followed by one integration with the adjoint model.

The eigenvalue problem (8) is solved by software packages like ARPACK [17] using Lanczos iterations. The symmetry of the matrix $M^* \, B \, M$ required by

Lanczos imposes to use the discrete adjoint M^* of the tangent linear operator M in (8). The computation of discrete adjoints for stiff systems is a nontrivial task [22]. In addition, computational errors (which can destroy symmetry) have to be small. A considerable loss of symmetry during the stiff transient is observed in practice [18]. This is due to the fact that the initial perturbations are away from the slow (non-stiff) manifold. To correct this we apply the tangent linear model on the initial perturbation for a short time, which is equivalent to "projecting" the initial perturbation onto the slow evolution manifold. In order to preserve operator symmetry, another projection using the adjoint model needs to be performed at the end of the adjoint integration. Consequently the matrix-vector products are computed as $w = \Pi^* M^*_{t_F \to t_0} B M_{t_0 \to t_F} \Pi x$, where Π and Π^* denote the projection operations performed with the tangent linear and the adjoint models respectively.

3 Numerical Results

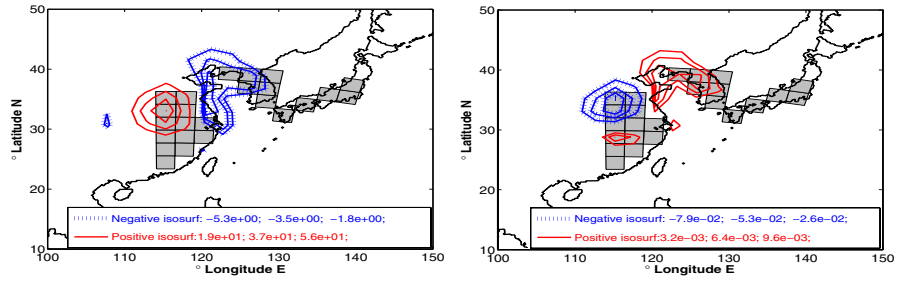
The numerical tests use the state-of-the-art regional atmospheric chemical transport model STEM [3]. The simulation covers a region of 7200 Km × 4800 Km in East Asia and uses a $30 \times 20 \times 18$ computational grid with a horizontal resolution of 240 Km × 240 Km. The chemical mechanism is a variant of SAPRC-99 [4] and accounts for 93 different chemical species. The simulated conditions during March 1$^{\text{st}}$, 2001, correspond to the Trace-P [3]) field experiment.

We consider artificial observations generated every 6 hours by a reference simulation starting at 0 GMT, March 1$^{\text{st}}$, 2001. The observations are ground level ozone (O_3) concentrations at 24 gridpoints over Japan, Korea, and East China. These grid points are referred to as the "target area" (the gray area in Figure 2).

For the calculation of TESVs the final perturbation energy measures the perturbations of (O_3) and nitrogen dioxide (NO_2) in the target area at the final time. The perturbation norm at the initial time accounts for perturbations in all chemical species, scaled by typical concentration values [18]. The O_3 and NO_2 sections of the dominant TESV are shown in Figure 2. We notice that dominant TESV is localized near the target area, and that it is strongly correlated with the adjoint variable corresponding to a similar target function.

The data assimilation process uses an ensemble with 50 members. The ensemble is run for 6 hours in forecast mode, then is analyzed using the artificial observations in the ensemble Kalman framework [9]. The assimilated ensemble is then advanced in time for another 6 hours, then analyzed again, etc. until the end of the 24 hours simulation interval.

Different initial perturbations are considered as follows. The first simulation ("D") uses an uncorrelated background. The initial perturbation is of the form $\delta x^B = 30\%\ x^B \cdot \xi$, where $\xi \in \mathcal{N}(0,1)$ and x^B is the initial concentration vector. The second simulation ("AR") uses a flow dependent AR model for background errors. The initial perturbation is $\delta x^B = A^{-1}\ \left(30\%\, x^B \cdot \xi\right)$, as described in section 2. The third simulation ("AR+SV") adds perturbations along the largest 40

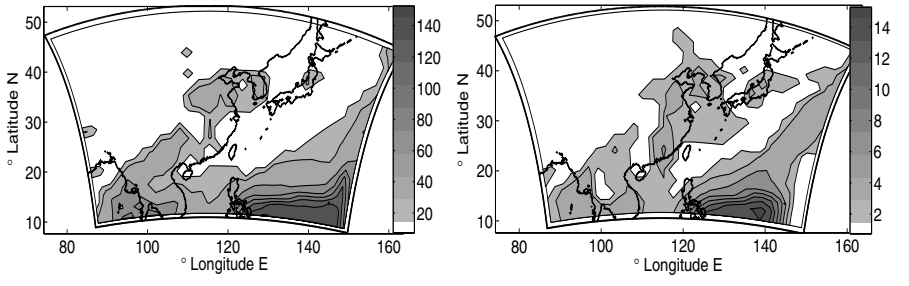


(a) O_3 Section (b) NO_2 Section

Fig. 2. The dominant TESV (for ground level O_3 concentration in the gray area) after 24 hours of evolution

TESVs to the flow dependent perturbations given by the autoregressive model. The TESV perturbations undergo the maximum growth over a 24 hour interval. Reducing uncertainty along these directions impacts the overall accuracy improvements obtained through data assimilation.

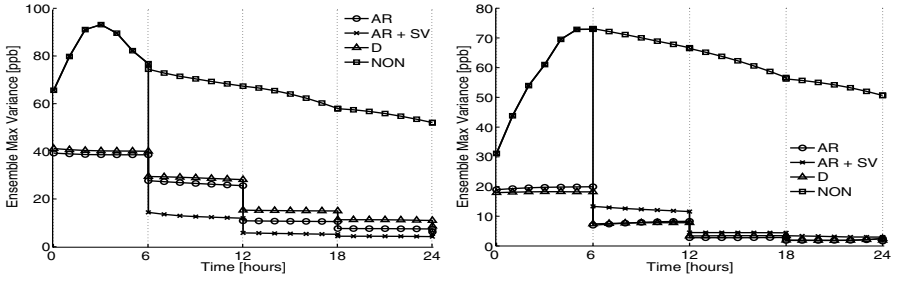
Figure 3 shows the ensemble standard deviation at ground level at the initial and final times using *AR+SV* background perturbations. Data assimilation leads to a large decrease in the ensemble standard deviation after 24 hours.



(a) Initial Time (b) Final Time (w/ Data Assimilation)

Fig. 3. Ensemble standard deviation at ground level with *AR+SV* background perturbations

The 24 hours time evolution of the ensemble O_3 standard deviation over the entire domain is shown in Figure 4(a), and over the target area in Figure 4(b). Different initial perturbations are considered with a diagonal correlation (*D*), an autoregressive correlation (*AR*), and the superposition of autoregressive and TESV perturbations (*AR+SV*). *NON* denotes the non-assimilated ensemble (initialized with *AR+SV*). The first analysis (at 6 hours) has the highest impact on the quality of the solution. Different ensembles perform differently under data assimilation. The *AR* initialized ensemble gives slightly better solutions than the *D* initialized one. The *AR+SV* ensemble performs best over the target area and very well over the entire domain.



(a) Entire Domain (b) Target Area

Fig. 4. Time evolution of the ensemble standard deviation for different initial perturbations: diagonal correlation (D), autoregressive correlation (AR), and the superposition of autoregressive and TESV perturbations ($AR+SV$). NON denotes the non-assimilated ensemble (initialized with $AR+SV$). All other ensembles are analyzed every 6 hours using O_3 ground level observations in the target area

4 Conclusions

This paper discusses some of the challenges associated with the application of nonlinear ensemble filtering data assimilation to atmospheric chemical transport models. The distinguishing feature of these models is the presence of nonlinear and stiff chemical interactions occurring at very short characteristic time scales.

A correct initialization of the ensemble is necessary for a successful application of nonlinear filtering data assimilation. We propose to model background errors using multilateral autoregressive processes. Such models are computationally inexpensive and capture well the error correlations along the flow lines. Total energy singular vectors are calculated for chemically reactive flows. A dual projection technique (with the tangent linear and with the adjoint models) is proposed to keep the linearized solutions on the slow manifold and preserve the symmetry of the chemistry tangent linear – adjoint operators.

The data assimilation test problem considered here is based on a large scale simulation of air pollution in East Asia in March 2001. The ensembles are initialized using autoregressive models of background errors and total energy singular vectors. The superposition of these two types of initial perturbations leads to an ensemble that performs very well both over the target area and over the entire computational domain.

References

1. M. Buehner. Ensemble-derived stationary and flow-dependent background error covariances: Evaluation in a quasi-operational NWP setting. *Q.J.R.M.S.*, accepted, 2004.
2. G.R. Carmichael. STEM – A second generation atmospheric chemical and transport model. *URL: http:// www.cgrer.uiowa.edu*, 2003.

3. G.R. Carmichael et. al. Regional-scale chemical transport modeling in support of the analysis of observations obtained during the trace-p experiment. *J. Geophys. Res.*, pages 10649–10671, 2004.
4. 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.
5. E.M. Constantinescu, A. Sandu, G.R. Carmichael, and T. Chai. Autoregressive models of background errors for chemical data assimilation. In preparation, 2005.
6. R. Daley. *Atmospheric Data Analysis*. Cambridge University Press, 1991.
7. H. Elbern, H. Schmidt, and A. Ebel. Variational data assimilation for tropospheric chemistry modeling. *J. Geophys. Res.*, 102(D13):15,967–15,985, 1997.
8. G. Evensen. 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, 1994.
9. G. Evensen. The ensemble Kalman filter: theoretical formulation and practical implementation. *Ocean Dyn.*, 53, 2003.
10. M. Fisher and D.J. Lary. Lagrangian four-dimensional variational data assimilation of chemical species. *Q.J.R.M.S.*, 121:1681–1704, 1995.
11. T.M. Hamill and J.S. Whitaker. Distance-dependent filtering of background error covariance estimates in an ensemble Kalman filter. *Mon. Wea. Rev.*, 129:2776–2790, 2001.
12. K.F. Hasselmann. Stochastic climate models. Part I. Theory. *Tellus*, 28:473–484, 1976.
13. P.L. Houtekamer and H.L. Mitchell. Data assimilation using an Ensemble Kalman Filter Technique. *Mon. Wea. Rev.*, 126(3):796–811, 1998.
14. 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. *Mon. Wea. Rev.*, accepted, 2003.
15. A.H. Jazwinski. *Stochastic Processes and Filtering Theory*. Academic Press, 1970.
16. R.E. Kalman. A new approach to linear filtering and prediction problems. *Trans. ASME, Ser. D: J. Basic Eng.*, 83:95–108, 1960.
17. Lehoucq, R., K. Maschhoff, D. Sorensen, C. Yang. ARPACK Software (Parallel and Serial), http://www.caam.rice.edu/software/ARPACK.
18. W. Liao, A. Sandu, G.R. Carmichael, and T. Chai. Total energy singular vector analysis of atmospheric chemical transport models. Submitted, 2005.
19. 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.
20. L. Menut, R. Vautard, M. Beekmann, and C. Honoré. Sensitivity of photochemical pollution using the adjoint of a simplified chemistry-transport model. *J. Geophys. Res.*, 105-D12(15):15,379–15,402, 2000.
21. L.P. Riishojgaard. A direct way of specifying flow-dependent background error correlations for meteorological analysis systems. *Tellus A*, 50(1):42–42, 1998.
22. A. Sandu, D. Daescu, and G.R. Carmichael. Direct and adjoint sensitivity analysis of chemical kinetic systems with KPP: I – theory and software tools. *Atm. Env.*, 37:5,083–5,096, 2003.
23. A. Sandu, Dacian N. Daescu, Gregory R. Carmichael, and Tianfeng Chai. Adjoint sensitivity analysis of regional air quality models. *J. Comp. Phys.*, accepted, 2004.
24. M. van Loon, P.J.H. Builtjes, and A.J. Segers. Data assimilation of ozone in the atmospheric transport chemistry model LOTOS. *Env. Model. Soft.*, 15:703–709, 2000.