Computer Science Technical Report

TR-06-22 (926)

October 11, 2006

Emil M. Constantinescu, Tianfeng Chai,
Adrian Sandu, and Gregory R. Carmichael

"Autoregressive Models of Background Errors for Chemical Data Assimilation"

Computer Science Department
Virginia Polytechnic Institute and State University
Blacksburg, VA 24060
Phone: (540)-231-2193
Fax: (540)-231-6075
Email: sandu@cs.vt.edu
Web: http://www.eprints.cs.vt.edu

Autoregressive Models of Background Errors for Chemical Data Assimilation

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

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

The task of providing an optimal analysis of the state of the atmosphere requires the development of dynamic data-driven systems that efficiently integrate the observational data and the models. Data assimilation (DA) is the process of adjusting the states or parameters of a model in such a way that its outcome (prediction) is close, in some distance metric, to observed (real) states. It is widely accepted that a key ingredient of successful data assimilation is a realistic estimation of the background error distribution. This paper introduces a new method for estimating the background errors which are modeled using autoregressive processes. The proposed approach is computationally inexpensive and captures the error correlations along the flow lines.

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, physically consistent estimate of the evolving chemical state of the atmosphere.

The close integration of observational data is recognized as essential in weather and climate analysis and forecast activities, and this is accomplished by a mature experience and infrastructure in meteorological data assimilation [Daley, 1991; Kalnay, 2002; Courtier et al., 1998; Rabier et al., 2000]. Data assimilation is vital for meteorological forecasting and has started to be applied in chemical transport modeling [Elbern et al., 1997; Fisher and Lary, 1995; Van Loon et al., 2000; Menut et al., 2000].

In this paper we focus on data assimilation applied to atmospheric chemical transport models (CTMs). CTMs are designed to describe the fate and transport of atmospheric chemical constituents associated with the gas and aerosol phases. CTMs have become 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. They can be used in designing cost-effective emission control strategies for improved air quality, for the interpretation of observational data such as those obtained during intensive field campaigns, air-quality forecasting, and assessments into how we have altered the chemistry of the global environment.

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 $c(t^{\mathrm{B}})$ to the "final" state $c(t^{\mathrm{F}})$ (1). Perturbations (small errors) evolve according to the tangent linear model (2), and adjoint variables according to the adjoint model (3):

$$c(t^{\mathrm{F}}) = \mathcal{M}_{t^{\mathrm{B}}\to t^{\mathrm{F}}}\left(c(t^{\mathrm{B}})\right) \tag{1}$$

$$\delta c(t^{\mathrm{F}}) = M_{t^{\mathrm{B}}\to t^{\mathrm{F}}}\,\delta c(t^{\mathrm{B}}) \tag{2}$$

$$\lambda(t^{\mathrm{B}}) = M^{*}_{t^{\mathrm{F}}\to t^{\mathrm{B}}}\,\lambda(t^{\mathrm{F}})\,. \tag{3}$$

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 $\mathbb{B}$ (at t^{B}) to $\mathbb{P}$ (at t^{F}) according to

$$\mathbb{P} = M_{t^{\mathrm{B}}\to t^{\mathrm{F}}}\,\mathbb{B}\,M^{*}_{t^{\mathrm{F}}\to t^{\mathrm{B}}} + \mathbb{Q}\,, \tag{4}$$

where $\mathbb{Q}$ is the covariance of the model errors.

The background, or initial state of an atmospheric model, is not known exactly, and can be correctly represented only in a probabilistic framework that accounts for the uncertainty.

We can represent the background state c^{B} as the sum of an average (most likely) state $\bar{c}^{\mathrm{B}}$ plus an error (uncertainty) field δc^{B},

$$c^{\mathrm{B}} = \bar{c}^{\mathrm{B}} + \delta c^{\mathrm{B}} \, . \tag{5}$$

The error field is considered to be unbiased and with the background covariance $\mathbb{B}$,

$$\left\langle \delta c^{\mathrm{B}} \right\rangle = 0 \, , \qquad \left\langle \delta c^{\mathrm{B}} \left(\delta c^{\mathrm{B}} \right)^{T} \right\rangle = \mathbb{B} \, . \tag{6}$$

In ensemble forecasting, one of the major challenges is the specification of the initial ensemble. For a correct ensemble distribution, each member is drawn from the same 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. Molteni et al. [1996] and Barkmeijer et al. [1998, 1999] use the leading singular vectors (with respect to the energy norm) of the linear propagator to identify the directions in phase space associated with maximum perturbation growth during the early parts of the forecast period. Toth and Kalnay [1997] determine the directions of maximum error growth by "breeding" the perturbation vectors, i.e. letting the perturbations grow through the system evolution and periodic rescaling. Distance and flow information can also be used in ensemble initialization [Buehner, 2004; Zupanski, 2005].

The aim of this paper is to construct models of $\mathbb{B}$ which account for the spatial correlations of errors in atmospheric models in a "sensible" way, mimic the decay of the correlation with distance, and are computationally inexpensive and easy to implement. We focus on CTM applications and investigate the effectiveness of this new method on a CTM variational data assimilation problem. Constantinescu et al. [2006c] have already applied the approach described in this paper to an ensemble data assimilation problem. The contributions of this work include: (1) the introduction of a new method to generate autoregressive (AR) models for the background errors, (2) the application of these models to variational and ensemble data assimilation, and (3) the study of the effects of using the autoregressive models to solve a real chemical data assimilation problem.

The paper is organized as follows. Section 2 introduces the chemical transport models and discusses the correlation of errors. Section 3 develops the autoregressive error model approach and Section 4 describes the practical implementation. Section 5 illustrates the

use of the new background error model in a real, large scale data assimilation test, and Section 6 summarizes the results of this research.

2 Chemical Transport Models and State Errors

Chemical transport models solve the mass-balance equations for concentrations of trace species in order to determine the fate of pollutants in the atmosphere [Carmichael et al., 2003, 2006; Liao 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, ρ is the air density, and V_s^{dep} is the deposition velocity. The boundaries $\Gamma^{\{\text{in,out,ground}\}}$ represent the inflow, outflow, and ground boundaries, respectively. The evolution of c_s is described by the following equations

$$\begin{aligned}
\frac{\partial c_s}{\partial t} &= -u\nabla c_s + \frac{1}{\rho}\nabla(\rho K \nabla c_s) + \frac{1}{\rho} f_s(\rho c) + E_s, \quad t^0 \le t \le t^{\mathrm{B}}, \quad 1 \le s \le N_{\text{spec}} \\
c_s(t^0, x) &= c_s^0(x), \\
c_s(t, x) &= c_s^{\text{in}}(t, x) \quad \text{for} \quad x \in \Gamma^{\text{in}} \\
K\frac{\partial c_s}{\partial n} &= 0 \quad \text{for} \quad x \in \Gamma^{\text{out}} \\
K\frac{\partial c_s}{\partial n} &= V_s^{dep} c_s - Q_s \quad \text{for} \quad x \in \Gamma^{\text{ground}}
\end{aligned} \tag{7}$$

We refer to the equations (7) as the *forward model*.

A perturbation δc^0 of the initial conditions will result in perturbations $\delta c(t)$ of the concentration field at later times. The evolution of these perturbations is governed by the equations:

$$\begin{aligned}
\frac{\partial \delta c_s}{\partial t} &= -u\nabla \delta c_s + \frac{1}{\rho}\nabla(\rho K \nabla \delta c_s) + F_{s,*}(\rho c)\delta c + \phi_s\,, \quad t^0 \le t \le t^{\mathrm{B}}\,, \quad 1 \le s \le N_{\text{spec}} \\
\delta c_s(t^0, x) &= \delta c_s^0(x), \\
\delta c_s(t, x) &= \delta c_s^{\text{in}}(t, x) \quad \text{for} \quad x \in \Gamma^{\text{in}} \\
K\frac{\partial \delta c_s}{\partial n} &= 0 \quad \text{for} \quad x \in \Gamma^{\text{out}} \\
K\frac{\partial \delta c_s}{\partial n} &= V_s^{dep} \delta c_s - \delta Q_s \quad \text{for} \quad x \in \Gamma^{\text{ground}}
\end{aligned} \tag{8}$$

Equations (8) are referred to as the *tangent linear model* associated with the forward model

(7). Here $F = \partial f/\partial c$ denotes the Jacobian of the chemical rate function f, and $F_{s,*}$ is its s-th row. The stochastic forcing function ϕ_s describes the model errors.

Our approach to modeling the background errors is as follows. We consider a simulation that starts at t^0 (a distant time in the past) and ends at the background time t^{B}. During this interval errors (or uncertainties) in the conditions at time t^0 are evolved according to the TLM equations, and correlations between different components of the model error are established.

To better understand the relationship between the tangent linear model (8) which governs the evolution of perturbations and autoregressive models, we first discuss the one-dimensional advection diffusion equation, then the box chemical model.

2.1 Correlation of Errors

We now consider the correlation function between errors in two species, at two different locations, at the same time instant:

$$R_{s,q}(t,x,y) = \left\langle \, \delta c_s(t,x),\, \delta c_q(t,y) \, \right\rangle . \tag{9}$$

Here $\langle\cdot\rangle$ denotes the ensemble average.

For simplicity, we consider the one-dimensional advection-diffusion-reaction of a single species c in an infinite spatial domain. Assume that u, ρ, K are constant in space and time, and that the chemical reaction is a simple decay equation, $f(\rho c) = -Lc$. The evolution of the concentration perturbation in time is governed by

$$\begin{aligned} &\frac{\partial \delta c}{\partial t} = -u\frac{\partial \delta c}{\partial x} + K\frac{\partial^2 \delta c}{\partial x^2} - L\delta c, \quad t^0 \le t \le t^{\mathrm{B}} \\ &t^0 = 0\,, \quad \delta c(0,x) = \delta c^0(x)\,. \end{aligned} \tag{10}$$

The general solution of the equation (10) is derived in Appendix A and has the form

$$\delta c(t,x) = \frac{e^{-Lt}}{\sqrt{\pi}}\,\frac{1}{2\sqrt{Kt}}\int_{\mathbb{R}} e^{-\left(\frac{x-z-ut}{2\sqrt{Kt}}\right)^2}\,\delta c^0(z)\,dz \tag{11}$$

Random initial perturbations Consider now that the initial perturbations δc^0 are a random process in space. Correlations develop due to the TLM dynamics, and the covariance function at time $t > 0$ as

$$\begin{aligned} R(t,x,y) &= \left\langle \, \delta c(t,x)\,,\, \delta c(t,y) \, \right\rangle \\ &= \frac{e^{-2Lt}}{4\pi Kt}\int_{\mathbb{R}}\int_{\mathbb{R}} e^{-\left(\frac{x-z-ut}{2\sqrt{Kt}}\right)^2-\left(\frac{y-w-ut}{2\sqrt{Kt}}\right)^2}\left\langle \, \delta c^0(z),\, \delta c^0(w) \, \right\rangle dz\,dw \end{aligned}$$

For the particular case where the initial random process has uniform variance and is totally uncorrelated

$$\left\langle \delta c^0(z), \delta c^0(w) \right\rangle = \sigma^2 \, \delta_{z-w}$$

the covariance function is

$$\begin{aligned} R(t,x,y) &= \sigma^2 \, \frac{e^{-2Lt}}{4\pi K t} \int_{\mathbb{R}} e^{-\left(\frac{x-z-ut}{2\sqrt{Kt}}\right)^2 - \left(\frac{y-z-ut}{2\sqrt{Kt}}\right)^2} \, dz \\ &= \sigma^2 \, \frac{e^{-2Lt}}{\sqrt{\pi}} \, \frac{e^{-\left(\frac{x-y}{\sqrt{8Kt}}\right)^2}}{\sqrt{8Kt}} \end{aligned}$$

For given x and y (with $x \neq y$) the covariance $R(t,x,y)$ as a function of time has a maximum value at

$$t_{\max} = \frac{\tau}{8} \left(\sqrt{1 + 4 \left(\frac{x-y}{D} \right)^2} - 1 \right)$$

where τ is the chemical lifetime of the species defined as the inverse of the destruction rate

$$\tau = \frac{1}{L}$$

and D is the "characteristic length"

$$D = \sqrt{\frac{K}{L}} = \sqrt{K\tau} \, .$$

The maximum value of the covariance of the solution at two locations $x \neq y$ is

$$\begin{aligned} R(t_{\max},x,y) &= \frac{\sigma^2}{\sqrt{\pi}\, D} \, \frac{e^{-\frac{1}{2}\sqrt{4\left(\frac{x-y}{D}\right)^2+1}}}{\sqrt{\sqrt{4\left(\frac{x-y}{D}\right)^2+1}-1}} \\ &\approx \frac{\sigma^2}{\sqrt{2\pi}\, D} \, \frac{e^{-\frac{|x-y|}{D}}}{\sqrt{\frac{|x-y|}{D}}} \quad \text{for} \quad |x-y| \gg D \end{aligned}$$

It is clear that the errors in initial conditions, when evolved through the tangent linear convection-diffusion-reaction equation, develop spatial correlations. The characteristic distance $D = \sqrt{K/L} = \sqrt{K\tau}$ is in fact the spatial correlation distance. It increases with increased diffusion strength and decreases with increased chemical destruction rate. Thus the developed spatial correlation distance is smaller for fast lived species and larger for long-lived species.

Note that in this simple example the spatial correlation at $t > 0$ between the solution at points x and y depends on the distance between the points $x - y$, the diffusion

coefficient K and the reaction rate L, but does not depend on the wind velocity u. However, it should be clear from the above derivation that if the initial condition is correlated ($\langle \delta c^0(z), \delta c^0(w) \rangle \neq 0$ for some $z \neq w$), or if it is decorrelated but the variance is space dependent ($\langle \delta c^0(z), \delta c^0(w) \rangle = \sigma^2(z)\, \delta_{z-w}$) then the correlation $R(t, x, y)$, $t > 0$, will depend on u as well.

Random forcing Consider now the simple model (10) started from a deterministic initial condition ($\delta c^0 = 0$) but excited by an additive white noise process $\zeta(t)$

$$\frac{\partial \delta c}{\partial t} = -u\frac{\partial \delta c}{\partial x} + K\frac{\partial^2 \delta c}{\partial x^2} - L\delta c + \zeta, \quad 0 \leq t \leq t^{\mathrm{B}} \tag{12}$$
$$\delta c(0, x) = 0\,, \quad \Big\langle\, \zeta \,\Big\rangle = 0\,, \quad \Big\langle\, \zeta(t_1, x_1)\, \zeta(t_2, x_2) \,\Big\rangle = \sigma^2\, \delta_{t_1 - t_2}\, \delta_{x_1 - x_2}\,.$$

The derivation presented in Appendix B reveals that the covariance function of the solution after long integration times $t \to \infty$ tends to the stationary value

$$R(t = \infty, x, y) = \frac{\tau\, \sigma^2}{\sqrt{2}D}\, e^{-\dfrac{\left|x - y\right|}{D}}$$

The spatial correlation distance in the stationary regime is $D = \sqrt{K/L} = \sqrt{K\tau}$.

The conclusion of this analysis is that the random perturbations in both the initial conditions and in the forcing, lead, through the dynamics of the tangent linear model, to random perturbations in the solution. The solution perturbations are correlated in space with a characteristic correlation distance $D = \sqrt{K/L}$. The correlation distance is influenced directly by the chemical reactions, with errors in fast lived species having a shorter correlation distance than long lived species, which is a sensible conclusion.

3 Autoregressive Models of Background Errors

3.1 One-dimensional Advection-Diffusion Equation

Consider the advection and diffusion of a single species c:

$$\frac{\partial c}{\partial t} = -u\frac{\partial c}{\partial x} + \frac{1}{\rho}\frac{\partial}{\partial x}\left(\rho K \frac{\partial c}{\partial x}\right), \quad t^0 \leq t \leq t^{\mathrm{B}} \tag{13}$$
$$c(t^0, x) = c^0(x)\,, \quad c(t, x) = c^{\mathrm{in}}(t, x) \quad \text{for} \quad x \in \Gamma^{\mathrm{in}}\,, \quad K\frac{\partial c}{\partial x} = 0 \quad \text{for} \quad x \in \Gamma^{\mathrm{out}}$$

We consider that the model started in the remote past from an uncertain initial condition $c^0 + \delta c^0$, and is evolving with known (deterministic) boundary conditions ($\delta c^{\mathrm{in}} = 0$).

The model state perturbations evolve in time according to the tangent linear model

$$\frac{\partial \delta c}{\partial t} = -u\frac{\partial \delta c}{\partial x} + \frac{1}{\rho}\frac{\partial}{\partial x}\left(\rho K \frac{\partial \delta c}{\partial x}\right) + \phi\,, \quad t^0 \le t \le t^{\mathrm{B}} \tag{14}$$
$$\delta c(t^0, x) = 0\,, \quad \delta c(t, x) = 0 \quad \text{for} \quad x \in \Gamma^{\text{in}}\,, \quad K\frac{\partial \delta c}{\partial x} = 0 \quad \text{for} \quad x \in \Gamma^{\text{out}}\,.$$

The system evolves subject to an external stochastic forcing function ϕ which represents the model errors. We want to estimate the cumulative effect of these perturbations at time t^{B}.

Consider a discretization of (14). The spatial discretization uses the first order upwind formula for advection and central finite differences for diffusion. A single implicit Euler step is taken from t^0 to t^{B}. The implicit Euler method is unconditionally stable. Moreover, this discretization is monotonic for any value of the step size.

The discrete tangent linear model reads

$$\begin{aligned}\delta c_j^{\mathrm{B}} = \delta c_j^0 &- \frac{u_j^+ \Delta t}{\Delta x}\left(\delta c_j^{\mathrm{B}} - \delta c_{j-1}^{\mathrm{B}}\right) - \frac{u_j^- \Delta t}{\Delta x}\left(\delta c_{j+1}^{\mathrm{B}} - \delta c_j^{\mathrm{B}}\right) \\ &+ \Delta t\,\frac{\left(\rho_{j+1}K_{j+1} + \rho_j K_j\right)\delta c_{j+1}^{\mathrm{B}} - 4\rho_j K_j\,\delta c_j^{\mathrm{B}} + \left(\rho_j K_j + \rho_{j-1}K_{j-1}\right)\delta c_{j-1}^{\mathrm{B}}}{2\rho_j \Delta x^2} + \Delta t\,\phi_j\end{aligned} \tag{15}$$
$$\delta c^0 = 0\,, \quad \delta c_j = 0 \quad \text{for} \quad x_j \in \Gamma^{\text{in}}\,, \quad K_j\frac{\partial \delta c}{\partial x} = 0 \quad \text{for} \quad x_j \in \Gamma^{\text{out}}\,.$$

Equation (14) has the form of an autoregressive model

$$\alpha_{j-1}\,\delta c_{j-1}^{\mathrm{B}} + \alpha_j\,\delta c_j^{\mathrm{B}} + \alpha_{j+1}\,\delta c_{j+1}^{\mathrm{B}} = \xi_j$$

where the random variable $\xi_j = \delta c_j^0 + \Delta t\,\phi_j$ is the noise added to the process. The values of the autoregressive coefficients are given by the discretization:

$$\begin{aligned}\alpha_{j-1} &= -\frac{\Delta t}{2\Delta x^2 \rho_j}\left(K_{j-1}\rho_{j-1} + \rho_j(K_j + 2\Delta x u_j^+)\right)\,, \\ \alpha_j &= 1 - \frac{\Delta t}{\Delta x^2}\left(\Delta x(u_j^- - u_j^+) - 2K_j\right)\,, \\ \alpha_{j+1} &= -\frac{\Delta t}{2\Delta x^2 \rho_j}\left(K_{j+1}\rho_{j+1} + \rho_j(K_j - 2\Delta x u_j^-)\right)\,.\end{aligned}$$

A general spatial discretization with a stencil of $2p+1$ points and solved in time by implicit Euler leads to the following AR model:

$$\sum_{k=-p}^{p} \alpha_{j+p}\,\delta c_{j+p}^{\mathrm{B}} = \xi_j \tag{16}$$

Monotonicity Consider the case where the initial perturbation is bounded, $\delta c_{\min} \leq \delta c^0 \leq \delta c_{\max}$, and the external forcing is null, $\phi = 0$. The monotonicity of the implicit Euler scheme coupled with the first order upwind advection and central diffusion implies that $\delta c_{\min} \leq \delta c^{\mathrm{B}} \leq \delta c_{\max}$. Therefore the magnitude of the initial perturbations is not increased, but correlations are developed.

ARMA Models It is clear that the AR model (16) of the background error can be extended to an ARMA model (albeit at the expense of losing the direct relationship with the model (13)):

$$\sum_{k=-p}^{p} \alpha_{j+p}\, \delta c^{\mathrm{B}}_{j+p} = \sum_{k=-p}^{p} \beta_{j+p}\, \xi_{j+p}\ . \tag{17}$$

3.2 Box Model Chemistry

We now consider the following singular perturbation model for the chemical system:

$$\frac{d}{dt}\begin{bmatrix} c_{\mathrm{slow}} \\ c_{\mathrm{fast}} \end{bmatrix} = \begin{bmatrix} f\,(c_{\mathrm{slow}}, c_{\mathrm{fast}}) \\ \varepsilon^{-1}\, g\,(c_{\mathrm{slow}}, c_{\mathrm{fast}}) \end{bmatrix}, \quad \begin{bmatrix} c_{\mathrm{slow}}(t^0) \\ c_{\mathrm{fast}}(t^0) \end{bmatrix} = \begin{bmatrix} c^0_{\mathrm{slow}} \\ c^0_{\mathrm{fast}} \end{bmatrix}. \tag{18}$$

As a technical condition in the chemical system (18) we have the sub-Jacobian $\partial g/\partial c_{\mathrm{fast}}$ nonsingular, which implies that the limit DAE is of index-1 [Hairer et al., 1993].

The model (18) distinguishes between the fast and the slow species. The separation of scales is given by the parameter ε since c_{slow} evolves on $O(1)$ characteristic times while c_{fast} evolves on $O(\varepsilon)$ time scales. The smaller ε, the faster the evolution of c_{fast}, and in the limit $\varepsilon \to 0$ we have that

$$g\,(c_{\mathrm{slow}}, c_{\mathrm{fast}}) = 0\ .$$

This condition formally expresses the quasi-equilibrium of the system outside the initial transient. Since $\partial g/\partial c_{\mathrm{fast}}$ is nonsingular, the quasi-equilibrium relation allows to express the fast species as a function of the slow ones.

Small errors in the initial conditions propagate according to the tangent linear model

$$\frac{d}{dt}\begin{bmatrix} \delta c_{\mathrm{slow}} \\ \delta c_{\mathrm{fast}} \end{bmatrix} = \begin{bmatrix} \dfrac{\partial f}{\partial c_{\mathrm{slow}}} & \dfrac{\partial f}{\partial c_{\mathrm{fast}}} \\ \varepsilon^{-1}\dfrac{\partial g}{\partial c_{\mathrm{slow}}} & \varepsilon^{-1}\dfrac{\partial g}{\partial c_{\mathrm{fast}}} \end{bmatrix} \begin{bmatrix} \delta c_{\mathrm{slow}} \\ \delta c_{\mathrm{fast}} \end{bmatrix}, \quad \begin{bmatrix} \delta c_{\mathrm{slow}}(t^0) \\ \delta c_{\mathrm{fast}}(t^0) \end{bmatrix} = \begin{bmatrix} \delta c^0_{\mathrm{slow}} \\ \delta c^0_{\mathrm{fast}} \end{bmatrix}. \tag{19}$$

The quasi-equilibrium condition for the perturbations outside the initial transient is obtained by taking the limit $\varepsilon \to 0$

$$\frac{\partial g}{c_{\text{slow}}}\,\delta c_{\text{slow}} + \frac{\partial g}{c_{\text{fast}}}\,\delta c_{\text{fast}} = 0 \quad \Rightarrow \quad \delta c_{\text{fast}} = -\left(\frac{\partial g}{c_{\text{fast}}}\right)^{-1} \frac{\partial g}{c_{\text{slow}}}\,\delta c_{\text{slow}}\,. \tag{20}$$

This shows that due to quasi-equilibrium the errors in the fast components and slow components are strongly correlated (the errors in the fast components are determined by the errors in the slow ones).

One backward Euler step applied to (19) leads to the discrete model

$$\begin{bmatrix} I - \Delta t\,\dfrac{\partial f}{\partial c_{\text{slow}}} & -\Delta t\,\dfrac{\partial f}{\partial c_{\text{fast}}} \\ -\Delta t\,\varepsilon^{-1}\,\dfrac{\partial g}{\partial c_{\text{slow}}} & I - \Delta t\,\varepsilon^{-1}\,\dfrac{\partial g}{\partial c_{\text{fast}}} \end{bmatrix} \cdot \begin{bmatrix} \delta c^{\text{B}}_{\text{slow}} \\ \delta c^{\text{B}}_{\text{fast}} \end{bmatrix} = \begin{bmatrix} \delta c^{0}_{\text{slow}} \\ \delta c^{0}_{\text{fast}} \end{bmatrix}, \tag{21}$$

which again is an autoregressive model for the errors. Taking the limit $\varepsilon \to 0$ in (21) leads to equation (20) for the fast and slow components of δc^{B}. Consequently, the autoregressive model (21) captures the error correlations introduced by quasi-equilibrium in the stiff chemical system evolution.

4 Three-dimensional Multi-component AR Models

We now discuss the construction of three-dimensional autoregressive models for background errors. Consider a spatial domain $\mathcal{D}$ discretized using a structured grid of (N_X, N_Y, N_Z) gridpoints. We will denote by (i, j, k) the space gridpoint index. If N_S is the total number of different chemical species, then the dimension of the model state vector is $N = N_X \times N_Y \times N_Z \times N_S$.

The background state c^{B} is represented as the sum of the average state $\overline{c}^{\text{B}}$ plus an error (uncertainty) field δc^{B}. The error field has zero mean and background covariance $\mathbb{B}$ (5, 6).

Our basic assumption is that the background state errors form a multilateral autoregressive (AR) process [Hasselmann, 1976] of the form

$$\delta c^{\text{B}}_{i,j,k} + \alpha^{(\pm 1)}_{i,j,k}\,\delta c^{\text{B}}_{i\pm 1,j,k} + \beta^{(\pm 1)}_{i,j,k}\,\delta c^{\text{B}}_{i,j\pm 1,k} + \gamma^{(\pm 1)}_{i,j,k}\,\delta c^{\text{B}}_{i,j,k\pm 1} = \sigma_{i,j,k}\,\xi_{i,j,k}\,. \tag{22}$$

The model (22) captures bilateral correlations among neighboring grid points in the x, y and z directions (with the coefficients α, β, and γ respectively). Constant correlation coefficients α, β, γ imply fixed spatial directional correlations, whereas space-dependent coefficients allow to capture flow dependent correlations. The last term represents the

additional uncertainty at each grid point, with $\xi \in \mathcal{N}(0,1)$ normal random variables and σ the local error variance. The motivation behind multilateral AR models is the fact that (22) – with proper coefficients – can be regarded as a finite difference approximation of the tangent linear model of the advection-diffusion-reaction equation.

The AR process (22) can be represented compactly as

$$A\,\delta c^{\mathrm{B}} = \Sigma\,\xi\,, \quad \Sigma = \mathrm{diag}(\sigma_{i,j,k})\,, \quad \xi \in \mathcal{N}(0,1)\,. \tag{23}$$

The $N \times N$ background error covariance matrix is

$$\mathbb{B} = \left\langle \delta c^{\mathrm{B}} \left(\delta c^{\mathrm{B}}\right)^T \right\rangle = \left\langle A^{-1}\,\Sigma\,\xi \left(A^{-1}\,\Sigma\,\xi\right)^T \right\rangle = A^{-1} \left\langle \Sigma\,\xi\,\xi^T\,\Sigma \right\rangle A^{-T} = A^{-1}\,\Sigma^2\,A^{-T}\,, \tag{24}$$

and the corresponding correlation matrix is

$$\mathbb{C} = \mathrm{diag}(\mathbb{B})^{-1/2}\,\mathbb{B}\,\mathrm{diag}(\mathbb{B})^{-1/2}\,.$$

4.1 AR Models in Data Assimilation

We now discuss the application of the AR background covariance error model to ensemble Kalman filter [Houtekamer and Mitchell, 2001; Houtekamer et al., 2005; Constantinescu et al., 2006c,b] and to 4D-Var [Elbern and Schmidt, 1999, 2001; Elbern et al., 1999; Fisher and Lary, 1995; Menut et al., 2000; Chai et al., 2006a,b; Constantinescu et al., 2006a] data assimilation applications. A good review of the latest techniques in atmospheric chemical data assimilation can be found in [Carmichael et al., 2006]. It is well known that a representative background covariance matrix is essential for both techniques in order to achieve a good fit of the results.

In chemical atmospheric modeling data assimilation problems, the covariance matrix is usually approximated using the NMC method [Parrish and Derber, 1992]. In the NMC method, the differences between several forecasts verifying at the same time are used to approximate the background error Chai et al. [2006b]. The AR approach proves less expensive and more effective than the NMC approach.

4.1.1 Ensemble Data Assimilation: EnKF

In the ensemble Kalman filter data assimilation, the error is evolved in time through an ensemble of model runs. An important problem is the generation of the initial ensemble. Each member is formed by adding a different perturbation δc^{B} to the initial "best guess" (background) state. 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; Buehner, 2004; Zupanski, 2005]. In their formulation, the background representation relied on a certain mathematical model and/or a set of empirical assumptions. Here we introduce an analytic approach to this problem that can be adapted to a large class of models reducing the empirical assumptions to a minimum, if any.

Using the AR model (23), the background perturbation that defines the m-th member of the ensemble is obtained as

$$\delta c^{\mathrm{B}}_{[m]} = A^{-1}\,\Sigma\,\xi_{[m]}\,,$$

where $\xi_{[m]} \in (\mathcal{N}(0,1))^N$ is a vector of realizations of N independent normal random variables of mean 0 and standard deviation 1. The perturbation is generated by scaling the normal variables with the proper standard deviations, then solving a linear system with the AR coefficient matrix A.

The above described approach was successfully used by Constantinescu et al. [2006c,a,b] to initialize EnKF data assimilation experiments applied to a reactive transport problem.

4.1.2 Variational Data Assimilation: 4D-Var

In the 4D-Var data assimilation the best state estimate is obtained as the minimizer of the following cost function:

$$\mathcal{J} = \frac{1}{2}\left(c - c^{\mathrm{B}}\right)^T \mathbb{B}^{-1}\left(c - c^{\mathrm{B}}\right) + \frac{1}{2}\sum_{i=0}^{n}\left(\mathrm{obs}_i - H_i c_i\right)^T R_i^{-1}\left(\mathrm{obs}_i - H_i c_i\right) \tag{25}$$

Using the AR representation of the background covariance (24), we have

$$\mathbb{B}^{-1} = A^T\,\Sigma^{-2}\,A\,.$$

The 4D-Var cost function can be computed as

$$\begin{aligned} z &= \Sigma^{-1}\,A\left(c - c^{\mathrm{B}}\right), \\ \mathcal{J} &= \frac{1}{2}\,z^T z + \frac{1}{2}\sum_{i=0}^{n}\left(\mathrm{obs}_i - H_i c_i\right)^T R_i^{-1}\left(\mathrm{obs}_i - H_i c_i\right)\,. \end{aligned}$$

The AR model is particularly advantageous in the 4D-Var context where the evaluation of the background term in the cost function only requires one matrix-vector multiplication by the AR coefficient matrix A, and one component-wise scaling (multiplication by the diagonal matrix Σ^{-1}).

In the numerical results section of this paper we show a comparison between data assimilation experiments applied to the same reactive flow problem using different types of background covariances: diagonal, NMC, and AR.

4.2 Implementation Aspects

Our approach is to construct the AR model (22) 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.

The dimension of the background covariance matrix is $N \times N$ with $N \sim 10^6 - 10^8$ for realistic chemical transport models. Operator and dimensional splitting allow the representation of A and $\mathbb{B}$ as products of small, sparse matrices, thus reducing dramatically the costs associated with matrix-vector multiplications and linear system solutions, as well as the total storage requirements.

Specifically, let us consider a three dimensional atmospheric model solved by splitting the horizontal transport from the vertical transport and the chemistry. The concentration of species s in gridpoint (i, j, k) at t^n is denoted $c_{i,j,k,s}$. The vector of concentrations of species s in a horizontal plane is denoted by $c_{1:N_X,1:N_Y,k,s}$, and in a column by $c_{i,j,1:N_Z,s}$. The vector of concentrations of all species in the gridpoint is denoted by $c_{i,j,k,1:N_S}$. The vector of all concentrations is $c = c_{1:N_X,1:N_Y,1:N_Z,1:N_S}$

Using monotonic space discretizations and the backward Euler time integration method, the solution of the horizontal transport over a time step Δt is obtained as

$$H_{k,s}(\Delta t)\, \delta c^{n+1}_{1:N_X,1:N_Y,k,s} = \delta c^{n}_{1:N_X,1:N_Y,k,s} \;, \quad \forall\, k, s \;, \tag{26}$$

and the solution of the vertical transport as

$$V_{i,j,s}(\Delta t)\, \delta c^{n+1}_{i,j,1:N_Z,s} = \delta_{i,j,1:N_Z,s} c^{n} \;, \quad \forall\, i, j, s \;. \tag{27}$$

Over the entire domain we will write formally the horizontal discretization operator as

$$\left(I_{(1\le k\le N_Z)\times(1\le s\le N_S)} \otimes H_{k,s}(\Delta t)\right) \delta c^{n+1} = \delta c^{n} \,,$$

and the vertical discretization operator as

$$\left(I_{(1\le i\le N_X)\times(1\le j\le N_Y)\times(1\le s\le N_S)} \otimes V_{i,j,s}(\Delta t)\right) \delta c^{n+1} = \delta c^{n} \,.$$

where the Kronecker product operator $\otimes$, denotes the fact that the operation is repeated for each species, horizontal slice, or column.

Similarly, the solution is changed during one timestep due to chemical processes. In absence of transport processes (which are accounted for separately), the chemical interactions at each grid point are independent of other gridpoints, and are represented by a system of ODE

$$c'_{i,j,k} = f\left(t, c_{i,j,k}\right) \;, \quad \forall\, i, j, k \,.$$

Errors are propagated through the tangent linear chemical model, which is also an independent set of ODE at each grid point

$$\delta c'_{i,j,k,1:N_S} = F\left(t, c_{i,j,k,1:N_S}\right)\, \delta c_{i,j,k,1:N_S}\,, \quad F(t,c) = \frac{\partial f(t,c)}{\partial c}\,, \quad \forall\, i,j,k\,.$$

This error equation at each grid point is discretized in time by the backward Euler method to obtain

$$C_{i,j,k}(\Delta t)\, \delta c^{n+1}_{i,j,k,1:N_S}\Big(I - \Delta t\, F(t^n, c^n_{i,j,k,1:N_S})\Big)\, \delta c^{n+1}_{i,j,k,1:N_S} = \delta c^n_{i,j,k}\,,$$

or, over the entire domain

$$\Big(I_{(1\le i\le N_X)\times(1\le j\le N_Y)\times(1\le k\le N_Z)} \otimes C_{i,j,k}(\Delta t)\Big)\, \delta c^{n+1} = \delta c^n\,. \tag{28}$$

The autoregressive model obtained through operator splitting is of the form:

$$\begin{aligned} A = &\Big(I_{(1\le i\le N_X)\times(1\le j\le N_Y)\times(1\le k\le N_Z)} \otimes C_{i,j,k}(\Delta t)\Big) \\ &\cdot \Big(I_{(1\le i\le N_X)\times(1\le j\le N_Y)\times(1\le s\le N_S)} \otimes V_{i,j,s}(\Delta t)\Big) \\ &\cdot \Big(I_{(1\le k\le N_Z)\times(1\le s\le N_S)} \otimes H_{k,s}(\Delta t)\Big) \end{aligned} \tag{29}$$

A symmetric operator split version is also possible.

With dimensional splitting, the storage of A requires $N_X N_Y (N_Z \times N_Z)$ for the vertical operator, and $N_Z(N_X N_Y \times N_X N_Y)$ for the horizontal operator, yielding a reduction in storage of $\frac{N_X N_Y + N_Z}{N_X N_Y N_Z}$ times compared to the full storage of A. Furthermore, each of the operators are sparse matrices to a certain degree. Inverting A can be computed by inverting $V(\cdot)$ and $H(\cdot,\cdot)$ independently.

4.3 Chemical Lifetime and Correlation Distances

In CTMs, different species can have widely different "chemical lifespans". Short-lived species (e.g., OH) take part in fast chemical reactions and their abundance varies quickly with time. Our theoretical analysis in Section 2.1 has shown that the spatial correlation distance is limited by the characteristic lifetime of the chemical species. Specifically, for the chemical species s with a chemical lifetime τ_s an integration of length $O(\tau_s)$ is necessary for the spatial correlations to develop; but longer integration times will lead to spurious spatial correlations to develop. Our practical experience has revealed that fast species like NO_2 need correlation lengths smaller than O_3, while slow species like HCHO need longer correlation lengths. The reason for choosing variable correlation lengths for different species has been explained analytically in the previous sections. Another sensible reason

is that fast reacting species vanish in relative short amounts of time, and thus they cannot give correlations past a certain "destruction" or damping time. Slower reaction species persist a longer amount of time, hence the correlation distance needs to be longer in time.

For a correct representation of the spatial correlations being limited by the chemical lifetimes, we take the following approach. For each chemical species s the transport part of the autoregressive model (29) is constructed by applying m_s consecutive implicit Euler steps with step size Δt such that $m_s \Delta t \approx \tau_s$, the chemical lifetime of species s. Similarly, $p_{i,j,k}$ chemical steps are applied to allow chemical correlations in grid (i, j, k) to fully develop during the time interval $n_{i,j,k}\Delta t$.

The AR coefficient matrix reads

$$\begin{aligned} A = &\left(I_{(1\le i\le N_X)\times(1\le j\le N_Y)\times(1\le k\le N_Z)} \otimes C_{i,j,k}^{n_{i,j,k}}(\Delta t)\right) \\ &\cdot \left(I_{(1\le i\le N_X)\times(1\le j\le N_Y)\times(1\le s\le N_S)} \otimes V_{i,j,s}^{m_s}(\Delta t)\right) \\ &\cdot \left(I_{(1\le k\le N_Z)\times(1\le s\le N_S)} \otimes H_{k,s}^{m_s}(\Delta t)\right) \end{aligned} \tag{30}$$

In our experiments, the resulting background covariance matrix turns out to be well conditioned, easy to compute, and with acceptable storage requirements.

4.4 Construction of Spatial Operators

The individual spatial operators $V_{i,j,s}(\Delta t)$ and $H_{k,s}(\Delta t)$ depend on the meteorological data:

$$V_{i,j,s} = V_{i,j,s}\left(w^n_{i,j,1:N_Z}, Kv^n_{i,j,1:N_Z}, \Delta t\right)$$

and

$$H_{k,s} = H_{k,s}\left(u^n_{1:N_X,1:N_Y,k}, Kh^n_{1:N_X,1:N_Y,k}, \Delta t\right),$$

where u^n, v^n, w^n the latitudinal, longitudinal, and vertical components of the wind field, and Kh^n and Kv^n are the horizontal and vertical turbulent diffusion coefficients respectively. In the regular finite difference approach, the operators are constructed using the meteorological field values at the current time t^n. In order to capture correlation patterns developed over a longer time interval, the transport AR operators are constructed from time averaged meteorological data. The averaging interval can be for example 12 hours before the background time, t^{B}:

$$V_{i,j,s} = V_{i,j,s}\left(\frac{1}{N}\sum_{n=1}^{N} w^n_{i,j,1:N_Z}, \frac{1}{N}\sum_{n=1}^{N} Kv^n_{i,j,1:N_Z}, \Delta t\right)$$

and

$$H_{k,s} = H_{k,s}\left(\frac{1}{N}\sum_{n=1}^{N} u^n_{1:N_X,1:N_Y,k}, \frac{1}{N}\sum_{n=1}^{N} Kh^n_{1:N_X,1:N_Y,k}, \Delta t\right).$$

5 Numerical Results

In this section we show some preliminary numerical results of our AR model of the error covariance in the context of 4D-Var data assimilation. We construct an error covariance matrix for a data assimilation atmospheric chemical and transport application, and we test it against other ways of modeling the background error, namely diagonal (D) and NMC. NMC is a popular technique used in these type of applications, while D is a simple and accessible approach to modeling the background errors that does not consider error cross correlations among the solution components yielding a diagonal background matrix. The investigation is carried out in a variational data assimilation framework using 4D-Var. Constantinescu et al. [2006c] employed the AR model described in this work to initialize an ensemble data assimilation experiment applied to a problem similar to the one described in this paper.

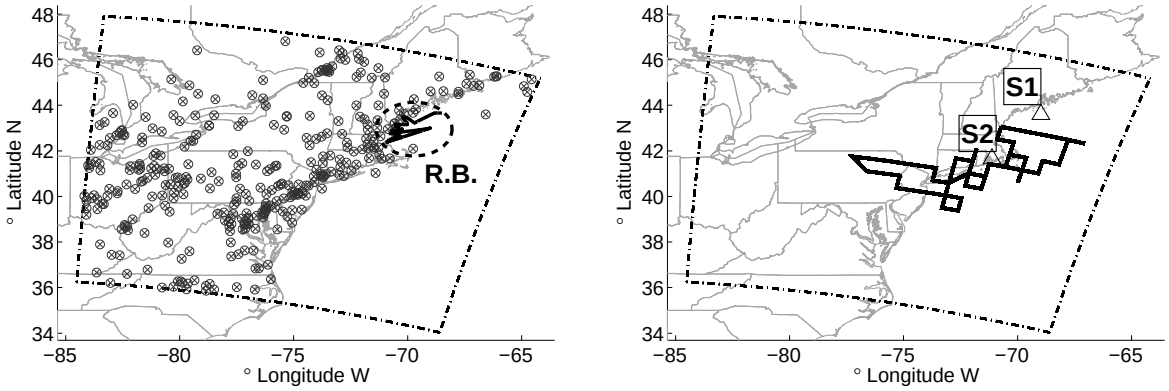
5.1 The Test Problem

Variational data assimilation experiments are carried out on a real-life scenario of air pollution in North–Eastern United States in July 2004 as shown in Fig. 1 (the dash-dotted line delimits the computational domain). We analyze the convergence of the optimization algorithm and the fit of the assimilation results in the presence of different background error models.

5.1.1 The Model

The numerical tests use the state-of-the-art regional atmospheric chemical transport model STEM [Carmichael et al., 2003]. The chemical reaction and transport equation (7) is solved using an operator splitting approach. STEM uses linear finite difference discretization of the transport terms. The advection terms are solved using a third order 1D upwind finite difference formula [Sandu et al., 2005]. The diffusion terms are discretized using second order central differences. The order of whole scheme is quadratic for the interior points. 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), and involves in 235 chemical reactions. The chemistry time integration is done by Rosenbrock 2 numerical integrator [Sandu et al., 2003], and is implemented using the kinetic preprocessor (KPP) [Damian et al., 2002].

The computational domain covers $1500 \times 1320 \times 20$ Km with a horizontal resolution of



(a) Ground stations and R. Brown location (b) Ozonesondes and P3 airplane path

Figure 1: Computational domain and (a) AIRNow ground measuring stations in support of the ICARTT campaign (340 in total) and Ronald H. Brown (R.B.) vessel location, and (b) two ozonesondes (S1, S2) and the flight path of a P3 airplane.

60×60 Km and a variable vertical resolution (resulting in a 3-dimensional computational grid of $25 \times 22 \times 21$ points). The initial concentrations, meteorological fields, boundary values, and emission rates correspond to ICARTT (International Consortium for Atmospheric Research on Transport and Transformation) [ICARTT] conditions starting at 12 GMT of July 20^{th}, 2004.

5.1.2 Analysis Setting

Now we briefly describe the analysis setting of the 4D-Var data assimilation experiments.

The simulations are started at 8 EDT July 20^{th}. We consider a 12-hour assimilation window that starts at 8 EDT July 20^{th} and ends at 20 EDT July 20^{th} during which model predictions are fitted with the observations in order to decrease the cost function (25).

The "best guess" of the state of the atmosphere is obtained from a longer simulation over the entire US performed in support of the ICARTT experiment [Tang et al., 2006]. This best guess is used to initialize the deterministic (non-assimilated) solution shown in the results section. The best guess evolved to 8 EDT July 20^{th} represents the background state in 4D-Var.

The observations comprise of ground-level (AIRNow [EPA, 2004]), airplane (P3 [NOAA, 2004a] and others), and ozonesonde O_3 measurements taken during the ICARTT campaign in summer 2004 [ICARTT]. Figure 1.a shows the location of the ground stations (340 in total) that measured ozone concentrations. Not all the stations provide observations each hour (the number of hourly observations varies between 160 and 326 during the assimilation window). A detailed description of the ICARTT fields and data can be found in

[Chai et al., 2006b; Tang et al., 2006].

An independent set of measurements are used to validate the data assimilation results using different background models. These measurements are collected by a NOAA vessel called Ronald H. Brown [NOAA, 2004b]. The location of the Ronald H. Brown ship is shown in Fig. 1.b.

4D-Var adjusts the initial concentrations of the ozone at each grid point at the beginning of the assimilation window (8 EDT July 20^{th}). The optimization algorithm used to minimize the cost function is L-BFGS-B [Byrd et al., 1995]. The optimization process is manually stopped after a certain number of iterations which depends on the decrease of the cost function in the particular data assimilation scenarios under consideration. More information about the optimization setup can be found in [Chai et al., 2006b].

The AR model was constructed using the averaged 12 hour wind fields prior to the assimilation window: 20 EDT July 19^{th} - 8 EDT July 20^{th}. The inverse of the $\mathbb{B}$ matrix for the NMC model used in our numerical experiments was obtained using a truncated SVD [Gwak and Masada, 2004]. This approach inverts only the contributions corresponding to the largest singular values, and thus circumvents the errors coming from inverting the NMC matrix which can be ill conditioned and reduces the cost function computational effort. The NMC model used in this paper for the numerical experiments is described in [Chai et al., 2006b].

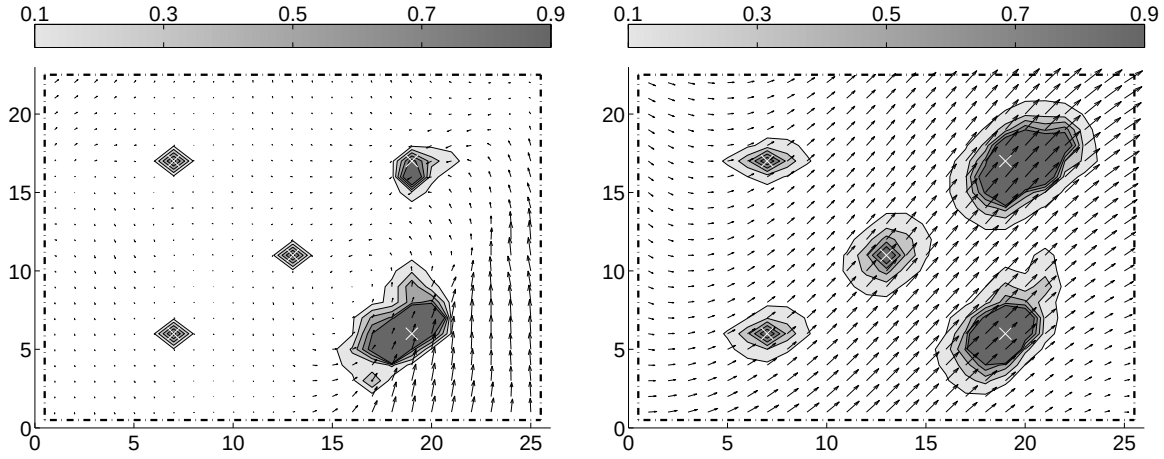
The performance of each data assimilation experiment is measured both by RMS and R^2 correlation factor between observations and model predictions. The RMS and R^2 correlation factor of two series X and Y of length n are

$$\text{RMS}(X,Y) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(X-Y)^2}$$

$$\text{R}^2(X,Y) = \frac{\left(n\sum_{i=1}^{n}X_iY_i - \sum_{i=1}^{n}X_i\sum_{i=1}^{n}Y_i\right)^2}{\left(n\sum_{i=1}^{n}X_i^2 - \left(\sum_{i=1}^{n}X_i\right)^2\right)\left(n\sum_{i=1}^{n}Y_i^2 - \left(\sum_{i=1}^{n}Y_i\right)^2\right)}.$$

5.2 AR Models Capture Flow Dependent Correlations

To illustrate the correlations generated by autoregressive models we consider the wind fields over North–Eastern United States (see Fig. 1.a), on July 20^{th}, 2004, corresponding to the ICARTT field campaign. An autoregressive model (22) of background errors is constructed using flow dependent coefficients. Top views of the spatial correlations of the resulting uncertainty fields are shown in Figure (2) for several grid points located on the



(a) Ground level correlations (b) Top level correlations

Figure 2: Horizontal background error correlations captured by the AR model: (a) ground and (b) top levels. Shown are five points (marked with white "x" symbols) using the ICARTT wind fields on July 20th, 2001.

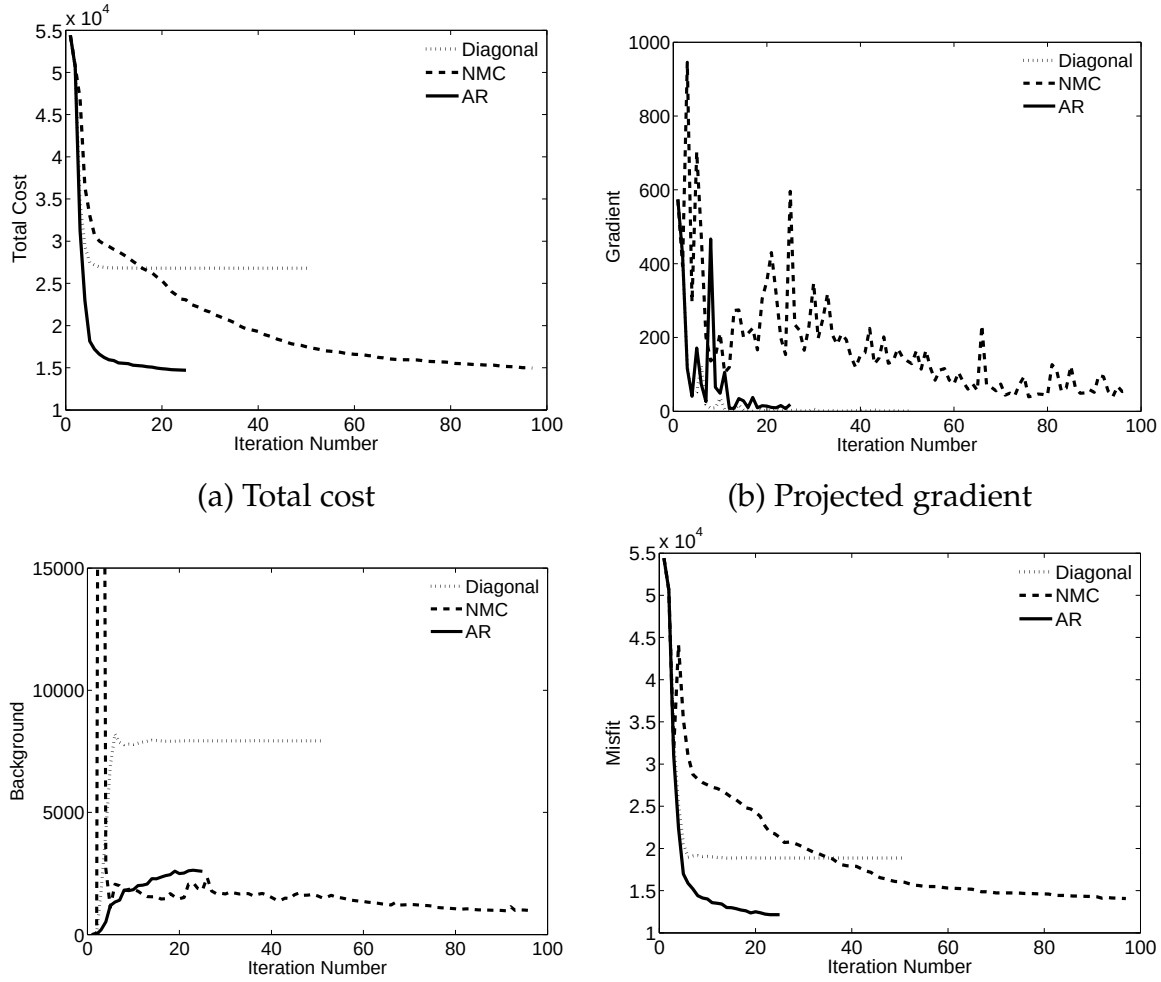
ground layer (a) and on the top layer (b). The correlations match the shape and magnitude of the wind field. Note that the wind speed near the ground is smaller than at the top and this is reflected by the correlations. Moreover, from the numerical point of view, the covariance matrix is well conditioned: cond($\mathbb{B}$)=640.

5.3 Comparison Between AR, Diagonal, and NMC Background Error Results

In this section we analyze the data assimilation results using the variational framework (4D-Var) described in previous sections. Here we consider the analysis scenario described in sec. 5.1.

Figure 3 shows the optimization parameters when using different background operators: D, NMC, and AR. The total cost represents the cost functional described by (25). The first term in (25) is referred to as the "background", while the second term is called "misfit". The background contribution in (25) constrains the optimization solution from "departing" from the best guess solution according to the background error model. The misfit acts in the opposite direction by trying to fit the solution with the observations. Each iteration in Fig. 3 represents at least one forward and one adjoint model time integration which is the most costly part of the data assimilation procedure.

Based on the results shown in Fig. 3, we conclude that when using the D background model, the optimization solution quickly converges to a solution which does not fit very



(a) Total cost

(b) Projected gradient

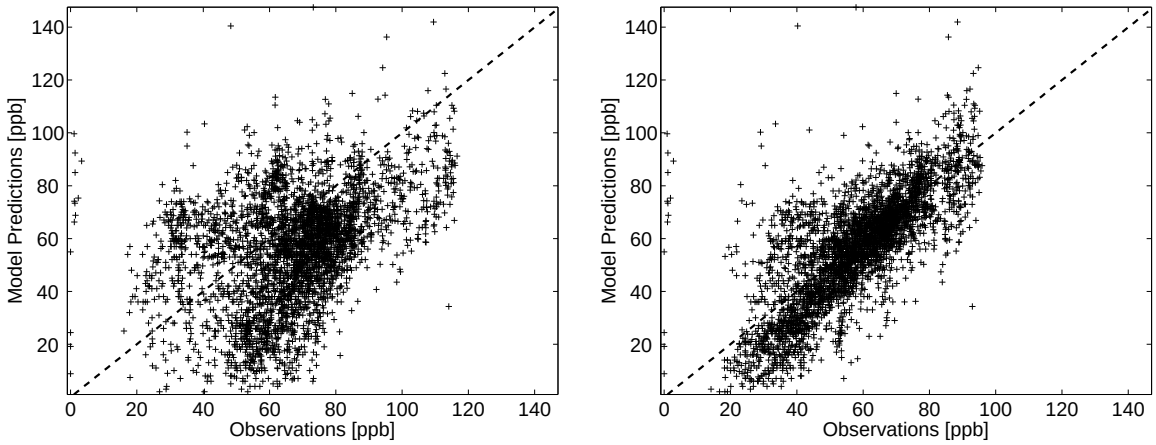
(c) Background contribution

(d) Misfit

Figure 3: Evolution of the (a) total cost function, (b) projected gradient, and the contributions of the (c) background and (d) misfit parts of the cost function when using the D (50 iterations), NMC (100 iterations), and AR (25 iterations) background operators.

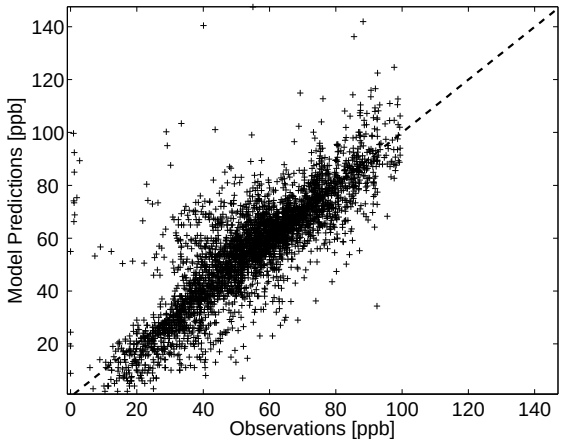
well with the observations, while both AR and NMC converge to solutions that better fit the observations. However, the cost function using the AR operator converges to a solution in 25 iterations, which is much faster than when using the NMC operator (which is typically four times slower than the AR convergence); moreover, the AR solution has a slightly better fit. The use of a diagonal background (D) model clearly impairs the optimization process by misrepresenting (ignoring) the correlations among background error components.

Figure 4 shows scatter plots of observations against model predictions during the analysis window of the unoptimized solution and the optimized solution using the D, NMC, and AR background operators. Below each figure we show the RMS and R^2 measures

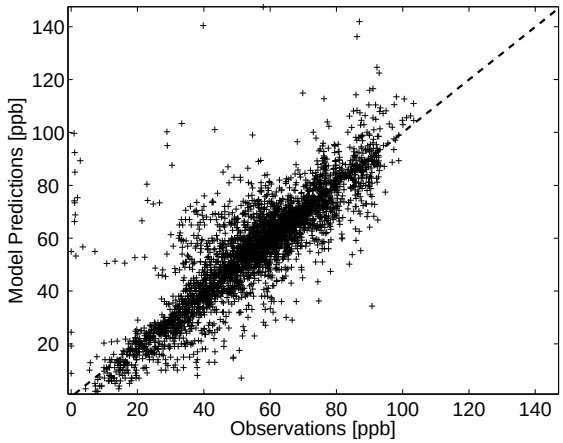


(a) Best Guess (RMS=24.46; R^2=0.15)

(b) D (RMS=14.03; R^2=0.55)



(c) NMC (RMS=11.92; R^2=0.68)



(d) AR (RMS=11.05; R^2=0.72)

Figure 4: Scatter plots of all observations vs. model predictions used during the analysis window for the best guess and the optimized initial conditions using D, NMC, and AR background operators. For each scenario we show the RMS and R^2 measures.

for the corresponding time window and scenario. In our experimental setting, both RMS and R^2 measures show the best fit for the optimized solution using the AR background (RMS=11.05; R^2=0.72). The results using the NMC model (RMS=11.92; R^2=0.68), qualitatively, are very close to the ones using AR; however, as described above, the number of iterations required by the NMC is significantly greater than for the AR model. The solution using the D model (RMS=14.03; R^2=0.55) shows an improvement when compared against the unoptimized solution (RMS=24.46; R^2=0.15); however, the convergence is limited to an unfit solution (the total cost is about 2.7e+04, while AR and NMC amount to about 1.e+04). A summary of all the results used in the assimilation process is shown in Table 1. Here we show the combined and ground (AIRNow), airplane measurements,

Scenario	All (3596)		AIRNow (2075)		Airplane (1486)		Ozonesonde (35)	
	RMS	R^2	RMS	R^2	RMS	R^2	RMS	R^2
Best guess	24.46	0.15	26.30	0.04	21.73	0.08	20.02	0.09
Diagonal	14.03	0.55	13.98	0.43	14.16	0.40	10.92	0.89
NMC	11.92	0.68	10.90	0.62	13.32	0.50	5.51	0.95
AR	11.05	0.72	9.65	0.70	12.84	0.53	4.60	0.96

Table 1: RMS [ppb] and R^2 fit measures of the best guess and optimized solution using D, NMC and AR background operators. We show the RMS and R^2 for *all*, *AIRNow*, *airplane*, and *ozonesonde* observations used during the assimilation process (in parenthesis we show their corresponding number).

and ozonesonde results in separate columns.

The evolution of the ozone concentration measured by the P3 airplane (see Fig. 1.b) and predicted by the model using the the best guess, and the optimized solution using the three types of background error models under investigation is shown in Figures 5. Here we note that the solutions that use AR and NMC fit equally well the measurements.

In Figure 6 we show the evolution of the ozone concentration measured by the Ronald H. Brown vessel (see Fig. 1.a) and predicted by the model using the unoptimized solution and the optimized (D, NMC, and AR) solutions. These measurements come from an independent source and are not used during the assimilation process. Here we consider the analysis and a 66 hour forecast window. The results are summarized in Table 2. These results support the conclusions presented above: 1) AR solution is closest to the observations (during analysis), 2) NMC also shows a good fit, while D falls behind. The overall validation results (see tab. 2) show a nearly equally good fit of the optimized solutions using AR and NMC background models. We note that during our experiments we also considered an AR model constructed with wind fields that corresponded to another period of time. The results, not shown in this study, were unsatisfactory, reinforcing the idea that the background errors depend on the particular flow and time.

6 Conclusions and Future Work

Data assimilation is the process to integrate observations and models in order to obtain simulation results that closer match reality. The information from observations can be used to adjust the initial conditions and other model parameters like emissions and boundary

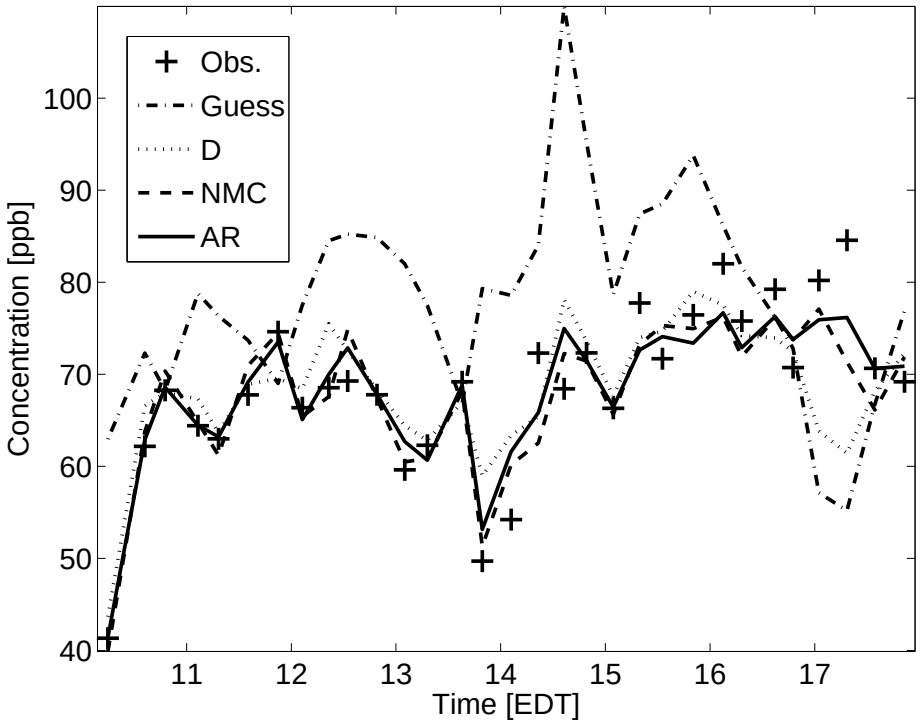


Figure 5: Time series of the observed ozone concentration, best guess, and optimized solution using D, NMC, and AR background operators for the P3 flight campaign.

conditions. It is widely accepted that the quality of the assimilation results depends on how well the errors in the initial conditions (the "background errors") are represented. The background errors are typically not well known and they need to be modeled. In this paper we construct autoregressive models of the background errors and apply them in the context of chemical transport models. The proposed approach is general and can be used in other applications, e.g., numerical weather prediction models.

In this paper we discuss the construction of background error models using multilateral autoregressive (AR) processes. The AR coefficients are given by a monotonic discretization of the tangent linear model; thus they capture the error correlations resulting from the error propagation through the model. The resulting AR models are computationally inexpensive and represent well the error correlations along the flow lines. Correlations between errors in different chemical species (arising due to stiff chemical interactions) are also captured by the AR model. The AR model can be extended to an ARMA model.

The full background covariance matrix requires an $O(n^2)$ storage where $n \sim 10^7$ is the number of state and parameter variables. In the discussed AR framework an operator splitting approach (applied to the tangent linear model) allows easy computation of the product of the covariance matrix (or its inverse) times a vector. This removes the

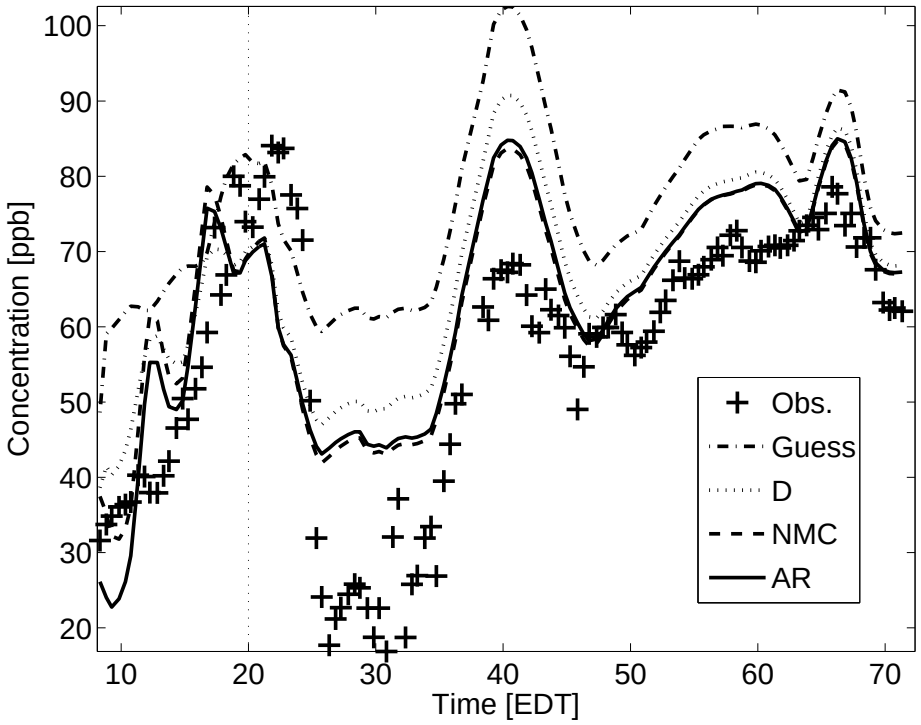


Figure 6: Time series of the observed ozone concentration, best guess, and optimized solution using D, NMC, and AR background operators for the Ronald H. Brown platform validation measurements during the analysis (8-20 EDT) and forecast windows (20-72 EDT).

need to store the entire covariance matrix and greatly reduces the computational costs. Moreover, the resulting background covariance matrix is full rank and well conditioned. Sparse matrix representations of individual operators further reduce the storage and the computational costs. In contrast, the NMC method requires several long integrations with the full model and is computationally expensive. The resulting covariance matrices are typically ill conditioned.

AR models can be applied in both variational and ensemble data assimilation settings. In this paper we consider the variational approach, where the autoregressive model is used to compute the background term of the cost function. In ensemble data assimilation, AR models can be used to generate the initial ensemble. The inversion of the inverse covariance matrix is obtained easily via operator splitting. This method has already been applied successfully in an EnKF data assimilation study using real data [Constantinescu et al., 2006c].

Several data assimilation experiments have been carried out with the STEM chemical transport model using observed data from a simulation of air pollution in Eastern U.S. A

Scenario	Analysis (71)		Forecast (306)	
	RMS	R^2	RMS	R^2
Best guess	19.77	0.72	23.42	0.49
Diagonal	11.41	0.68	15.98	0.59
NMC	12.22	0.59	12.80	0.68
AR	10.64	0.66	13.29	0.66

Table 2: RMS [ppb] and R^2 fit measures of the best guess and optimized solution using D, NMC and AR background operators for the independent validation results, Ronald H. Brown, that are not used during the assimilation process. We show the *analysis*, and *forecast* measures separately (in parenthesis we show their corresponding number).

comparison of data assimilation results has been carried out using the AR model of the background errors, a model obtained through the NMC approach and a diagonal model. In our setting, the AR model results are superior to the other methods tested (measured by the RMS and the R^2 fit indicators). Of particular importance is the fact that the AR model is significantly better than the diagonal model. The optimization using the AR model takes considerably fewer iterations than the one using the NMC model. The results of the diagonal model are inferior when compared with AR or NMC. An independent set of observations was used for validation. The optimized solution using the AR and NMC models were shown to improve the analysis and the forecast.

We note that the results we present in this paper using the AR and NMC background models can be further improved. The NMC background operator was obtained using some simplifications [Chai et al., 2006b] from the original NMC method [Parrish and Derber, 1992], while in the AR approach we used only the decorrelation distances provided by the chemistry, and thus further tuning is possible. Furthermore, a combination of AR and NMC background contributions can be explored (e.g., the use the decorrelation distance provided by the NMC when fine tuning the AR model).

Future work will consider the use of the AR background models in an operational setting. This is possible since AR needs averages of known meteorological data (from the period preceding the initial time). Such AR models can be constructed using current data for the next assimilation window.

A Solution of Linear, One-Dimensional Convection-Diffusion-Reaction Equation

For simplicity we consider the one-dimensional advection-diffusion-reaction of a single species c in an infinite spatial domain. Assume that u, ρ, K are constant in space and time, and that the chemical reaction is a simple decay equation, $f(\rho c) = -Lc$. The evolution of the concentration perturbation in time is governed by

$$\begin{aligned} \frac{\partial \delta c}{\partial t} &= -u\frac{\partial \delta c}{\partial x} + K\frac{\partial^2 \delta c}{\partial x^2} - L\delta c, \quad t^0 \le t \le t^{\mathrm{B}}, \\ t^0 &= 0, \quad \delta c(0,x) = \delta c^0(x). \end{aligned}$$

The Fourier transform (in space) of the solution in (10) is

$$\widehat{\delta c}(t,\omega) = \mathcal{F}\Big(\delta c(t,x)\Big) = \frac{1}{\sqrt{2\pi}}\int_{\mathbb{R}} e^{-i\omega x}\,\delta c(t,x)\,dx\,.$$

For a Dirac delta space distribution

$$\mathcal{F}\Big(\delta_{x-z}\Big) = \frac{1}{\sqrt{2\pi}} e^{-i\omega z}\,.$$

Taking the Fourier transform of (10) leads to

$$\begin{aligned} \frac{d\widehat{\delta c}}{d} &= \Big(-i\omega u - K\omega^2 - L\Big)\,\widehat{\delta c}\,, \quad \widehat{\delta c}(0,\omega) = \widehat{\delta c}^{\,0}(\omega) \\ \widehat{\delta c}(t,\omega) &= e^{\Big(-i\omega u - K\omega^2 - L\Big)t}\,\widehat{\delta c}^{\,0}(\omega) \\ \delta c(t,x) &= \mathcal{F}^{-1}\Big(\widehat{\delta c}(t,\omega)\Big) = \frac{e^{-Lt}}{\sqrt{2\pi}}\int_{\mathbb{R}} e^{i\omega(x-ut)}\,e^{-K\omega^2 t}\,\widehat{\delta c}^{\,0}(\omega)\,d\omega\,. \end{aligned}$$

If the initial condition is a Dirac delta function $\delta c(0,x) = \delta_{x-z}$ then the solution at later times $t > 0$ is

$$\begin{aligned} \mathfrak{c}_z(t,x) &= \frac{e^{-Lt}}{\sqrt{2\pi}}\int_{\mathbb{R}} e^{i\omega(x-ut)}\,e^{-K\omega^2 t}\,\frac{e^{-i\omega z}}{\sqrt{2\pi}}\,d\omega \\ &= \frac{e^{-Lt}}{\sqrt{2\pi}}\,\frac{1}{\sqrt{2\pi}}\int_{\mathbb{R}} e^{i\omega(x-z-ut)}\,e^{-K\omega^2 t}\,d\omega \\ &= \frac{e^{-Lt}}{\sqrt{2\pi}}\,\mathcal{F}^{-1}\Big(e^{-K\omega^2 t}\Big)\Big|_{x-z-ut} \\ &= \frac{e^{-Lt}}{\sqrt{\pi}}\,\frac{e^{-\left(\frac{x-z-ut}{2\sqrt{Kt}}\right)^2}}{2\sqrt{Kt}} \end{aligned}$$

By the principle of superposition the solution corresponding to a general initial condition $\delta c(0,z) = \delta c^0(z)$ is obtained by summing up elementary solutions that correspond to delta initial conditions:

$$\begin{aligned}\delta c(t,x) &= \int_{\mathbb{R}} \delta \mathfrak{c}_z(t,x)\, \delta c^0(z)\, dz \\ &= \frac{e^{-Lt}}{\sqrt{\pi}} \frac{1}{2\sqrt{Kt}} \int_{\mathbb{R}} e^{-\left(\frac{x-z-ut}{2\sqrt{Kt}}\right)^2} \delta c^0(z)\, dz\end{aligned}$$

B Solution of Linear, One-dimensional Convection–Diffusion–Reaction Equation with Random Forcing

Consider now the simple model (10) started from a deterministic initial condition ($\delta c^0 = 0$) but excited by an additive white noise ζ

$$\begin{aligned}&\frac{\partial \delta c}{\partial t} = -u\frac{\partial \delta c}{\partial x} + K\frac{\partial^2 \delta c}{\partial x^2} - L\delta c + \zeta, \quad 0 \le t \le t^{\mathrm{B}} \\ &\delta c(0,x) = 0\,, \quad \left\langle \zeta \right\rangle = 0\,, \quad \left\langle \zeta(t_1,x_1)\,\zeta(t_2,x_2) \right\rangle = \sigma^2\, \delta_{t_1-t_2}\, \delta_{x_1-x_2}\,.\end{aligned}$$

The solution can be obtained via Duhamel's principle from the solution (11) of the non-forced (homogeneous) equation (12) as

$$\delta c(t,x) = \int_0^t \frac{e^{-L(t-\theta)}}{2\sqrt{\pi K(t-\theta)}} \int_{\mathbb{R}} e^{-\left(\frac{x-z-u(t-\theta)}{2\sqrt{K(t-\theta)}}\right)^2} \zeta(\theta,z)\, dz\, d\theta$$

The covariance function

$$
\begin{aligned}
R(t,x,y) = & \int_0^t \int_0^t \frac{e^{-L(2t-\theta-\mu)}}{4\pi K\sqrt{(t-\theta)(t-\mu)}} \\
& \int_{\mathbb{R}} \int_{\mathbb{R}} e^{-\left(\frac{x-z-u(t-\theta)}{2\sqrt{K(t-\theta)}}\right)^2 - \left(\frac{y-w-u(t-\mu)}{2\sqrt{K(t-\mu)}}\right)^2} \left\langle \zeta(\theta,z), \zeta(\mu,w) \right\rangle dz\, dw\, d\theta\, d\mu \\
= & \sigma^2 \int_0^t \int_0^t \frac{e^{-L(2t-\theta-\mu)}}{4\pi K\sqrt{(t-\theta)(t-\mu)}} \\
& \int_{\mathbb{R}} \int_{\mathbb{R}} e^{-\left(\frac{x-z-u(t-\theta)}{2\sqrt{K(t-\theta)}}\right)^2 - \left(\frac{y-w-u(t-\mu)}{2\sqrt{K(t-\mu)}}\right)^2} \delta_{\theta-\mu}\, \delta_{z-w}\, dz\, dw\, d\theta\, d\mu \\
= & \sigma^2 \int_0^t \int_0^t \frac{e^{-L(2t-\theta-\mu)}}{4\pi K\sqrt{(t-\theta)(t-\mu)}} \\
& \left(\int_{\mathbb{R}} \int_{\mathbb{R}} e^{-\left(\frac{x-z-u(t-\theta)}{2\sqrt{K(t-\theta)}}\right)^2 - \left(\frac{y-w-u(t-\mu)}{2\sqrt{K(t-\mu)}}\right)^2} \delta_{z-w}\, dz\, dw \right) \delta_{\theta-\mu}\, d\theta\, d\mu \\
= & \sigma^2 \int_0^t \int_0^t \frac{e^{-L(2t-\theta-\mu)}}{4\pi K\sqrt{(t-\theta)(t-\mu)}} \\
& \left(\int_{\mathbb{R}} e^{-\left(\frac{x-z-u(t-\theta)}{2\sqrt{K(t-\theta)}}\right)^2 - \left(\frac{y-z-u(t-\mu)}{2\sqrt{K(t-\mu)}}\right)^2} dz \right) \delta_{\theta-\mu}\, d\theta\, d\mu \\
= & \sigma^2 \int_0^t \int_0^t \frac{e^{-L(2t-\theta-\mu)}}{4\pi K\sqrt{(t-\theta)(t-\mu)}} \left(\sqrt{\frac{4\pi K(t-\theta)(t-\mu)}{2t-\theta-\mu}}\, e^{-\frac{(x-y-u(\mu-\theta))^2}{4K(2t-\theta-\mu)}} \right) \delta_{\theta-\mu}\, d\theta\, d\mu \\
= & \frac{\sigma^2}{\sqrt{4\pi K}} \int_0^t \int_0^t \frac{e^{-L(2t-\theta-\mu)}}{\sqrt{2t-\theta-\mu}}\, e^{-\frac{(x-y-u(\mu-\theta))^2}{4K(2t-\theta-\mu)}}\, \delta_{\theta-\mu}\, d\theta\, d\mu \\
= & \frac{\sigma^2}{\sqrt{\pi}} \int_0^t e^{-2L(t-\theta)}\, \frac{e^{-\left(\frac{x-y}{\sqrt{8K(t-\theta)}}\right)^2}}{\sqrt{8K(t-\theta)}} d\theta
\end{aligned}
$$

In the limit $t \to \infty$ we obtain that

$$
\begin{aligned}
R(\infty,x,y) &= \frac{\sigma^2}{\sqrt{\pi}} \int_0^\infty e^{-2L\theta}\, \frac{e^{-\left(\frac{x-y}{\sqrt{8K\theta}}\right)^2}}{\sqrt{8K\theta}} d\theta \\
&= \frac{\sigma^2}{\sqrt{2KL}}\, e^{-\frac{|x-y|}{\sqrt{K/L}}} \\
&= \frac{\tau\,\sigma^2}{\sqrt{2}D}\, e^{-\frac{|x-y|}{D}}
\end{aligned}
$$

where

$$\tau = \frac{1}{L} \quad \text{and} \quad D\sqrt{K/L}\,.$$

Acknowledgments

This work was supported by the National Science Foundation through the awards NSF CAREER ACI–0413872, NSF ITR AP&IM 0205198, and by NOAA.

References

J. Barkmeijer, M. van Gijzen, and F. Bouttier. Singular vectors and estimates of the analysis error covariance metric. *Quarterly Journal of the Royal Meteorological Society*, 124(549): 1695–1713, 1998.

J. Barkmeijer, R. Buizza, and T.N. Palmer. 3D-Var hessian singular vectors and their potential use in the ECMWF ensemble prediction system. *Quarterly Journal of the Royal Meteorological Society*, 125:2333–2351, 1999.

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.

R. Byrd, P. Lu, and J. Nocedal. A limited memory algorithm for bound constrained optimization. *SIAM Journal of Scientific and Statistical Computing*, 16(5):1190–1208, 1995.

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.

G.R. Carmichael, A. Sandu, T. Chai, D.N. Daescu, E.M. Constantinescu, and Y. Tang. Predicting air-quality: Current status and future directions. *submitted to Journal of Computational Physics*, 2006.

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, 2006a.

T. Chai, G.R. Carmichael, Y. Tang, A. Sandu, M. Hardesty, P. Pilewskie, S. Whitlow, E.V. Browell, M.A. Avery, V. Thouret, P. Nedelc, J.T. Merrill, and A. Thompson. Four dimensional data assimilation experiments with ICARTT (International Consortium for Atmospheric Research on Transport and Transformation) Ozone Measurements. *submitted to Journal of Geophysical Research*, 2006b.

E.M. Constantinescu, A. Sandu, T. Chai, and G.R. Carmichael. Ensemble-based chemical data assimilation II: Real observations. Technical Report TR-06-07, Computer Science, Virginia Tech, April 2006a.

E.M. Constantinescu, A. Sandu, T. Chai, and G.R. Carmichael. Ensemble-based chemical data assimilation III: Filter localization. Technical Report TR-06-08, Computer Science, Virginia Tech, April 2006b.

E.M. Constantinescu, A. Sandu, T. Chai, and G.R. Carmichael. Assessment of ensemble-based chemical data assimilation in an idealized setting. *To appear in Atmospheric Environment*, 2006c.

P. Courtier, E. Andersson, W. Heckley, J. Pailleux, D. Vasiljevic, M. Hamrud, A. Hollingsworth, F. Rabier, and M. Fisher. The ECMWF implementation of three-dimensional variational assimilation (3D-Var). I:Formulation. *Quarterly Journal of the Royal Meteorological Society*, 124(550):1783–1807, 1998.

R. Daley. *Atmospheric Data Analysis*. Cambridge University Press, 1991.

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.

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.

EPA. AIRNow, 2004. URL `http://airnow.gov/`.

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.

K.-W. Gwak and G.Y. Masada. Regularization embedded nonlinear control designs for input-constrained and ill-conditioned thermal system. *Journal of Dynamic Systems, Measurement, and Control*, 126(3):574–582, 2004. doi: 10.1115/1.1789973.

E. Hairer, S.P. Norsett, and G. Wanner. *Solving Ordinary Differential Equations II: Stiff and Differential-Algebraic Problems*, chapter VII. Springer, 1993.

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

K.F. Hasselmann. Stochastic climate models. Part I. Theory. *Tellus*, 28:473–484, 1976.

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.

ICARTT. ICARTT home page. URL `http://www.al.noaa.gov/ICARTT`.

E. Kalnay. *Atmospheric Modeling, Data Assimilation and Predictability*. Cambridge University Press, 2002.

W. Liao, A. Sandu, G.R. Carmichael, and T. Chai. Singular vector analysis for atmospheric chemical transport models. Submitted to Monthly Weather Review, 2005.

L. Menut, R. Vautard, M. Beekmann, and Honoré C. 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.

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.

NOAA. P3, 2004a. URL http://www.al.noaa.gov/ICARTT/fieldoperations/fomp.shtml.

NOAA. New England Air Quality Study - Intercontinental Transport and Chemical Transformation, 2004b. URL http://saga.pmel.noaa.gov/Field/NEAQS-ITCT/.

D.F. Parrish and J.C. Derber. The national meteorological center's spectral statistical-interpolation analysis system. *Monthly Weather Review*, (120):1747–1763, 1992.

F. Rabier, H. Jarvinen, E. Klinker, J.F. Mahfouf, and A. Simmons. The ECMWF operational implementation of four-dimensional variational assimilation. I: Experimental results with simplified physics. *Quarterly Journal of the Royal Meteorological Society*, 126:1148–1170, 2000.

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, 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.

Y. Tang, G.R. Carmichael, N. Thongboonchoo, T. Chai, L.W. Horowitz, R.B. Pierce, J.A. Al-Saadi, G. Pfister, J.M. Vukovich, M.A. Avery, G.W. Sachse, T.B. Ryerson, J.S. Holloway, E.L. Atlas, F.M. Flocke, R.J. Weber, L.G. Huey, J.E. Dibb, D.G. Streets, and W.H. Brune. 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 Journal of Geophysical Research*, 2006.

Z. Toth and E. Kalnay. Ensemble forecasting at NCEP and the breeding method. *Monthly Weather Review*, 125:3297–3319, 1997.

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.

M. Zupanski. A method for initialization of ensemble data assimilation. *submitted to Tellus*, 2005.