BNL-98971-2012-JA

Atmos. Chem. Phys. Discuss., 12, 13515–13552, 2012
www.atmos-chem-phys-discuss.net/12/13515/2012/
doi:10.5194/acpd-12-13515-2012
© Author(s) 2012. CC Attribution 3.0 License.

Atmospheric Chemistry and Physics Discussions

This discussion paper is/has been under review for the journal Atmospheric Chemistry and Physics (ACP). Please refer to the corresponding final paper in ACP if available.

A three-dimensional variational data assimilation system for multiple aerosol species with WRF/Chem and an application to $PM_{2.5}$ prediction

Z. Li[1,2], Z. Zang[2], Q. B. Li[2,3], Y. Chao[2,3,5], D. Chen[2], Z. Ye[2], Y. Liu[4], and K.-N. Liou[2,3]

[1]Jet Propulsion Laboratory, California Institute of Technology, Pasadena, California, USA
[2]Joint Institute for Regional Earth System Science and Engineering, University of California, Los Angeles, California, USA
[3]Department of Atmospheric and Oceanic Sciences, University of California, Los Angeles, California, USA
[4]Brookhaven National Laboratory, Upton, New York, USA
[5]Remote Sensing Solutions, Inc., Pasadena, California, USA

Received: 8 April 2012 – Accepted: 15 May 2012 – Published: 31 May 2012

Correspondence to: Z. Li (zhijin.li@jpl.nasa.gov)

Published by Copernicus Publications on behalf of the European Geosciences Union.

Abstract

A three-dimensional variational data assimilation (3-DVAR) algorithm for aerosols in a WRF/Chem model is presented. The WRF/Chem model uses the MOSAIC (Model for Simulating Aerosol Interactions and Chemistry) scheme, which explicitly treats eight major species (elemental/black carbon, organic carbon, nitrate, sulfate, chloride, ammonium, sodium, and the sum of other inorganic, inert mineral and metal species) and represents size distributions using a sectional method with four size bins. The 3-DVAR scheme is formulated to take advantage of the MOSAIC scheme in providing comprehensive analyses of specie concentrations and size distributions. To treat the large number of state variables associated with the MOSAIC scheme, this 3-DVAR algorithm first determines the analysis increments of the total mass concentrations of the eight species, defined as the sum of the mass concentrations across all size bins, and then distributes the analysis increments over four size bins according to the background error variances. The number concentrations for each size bin are adjusted based on the ratios between the mass and number concentrations of the background state. This system has been applied to the analysis and prediction of $PM_{2.5}$ in the Los Angeles basin during the CalNex 2010 field experiment, with assimilation of surface $PM_{2.5}$ and speciated concentration observations. The results demonstrate that the data assimilation significantly reduces the errors in comparison with a down scaling simulation and improved forecasts of the concentrations of $PM_{2.5}$ as well as individual species for up to 24 h. Some implementation difficulties and limitations of the system are also discussed.

1 Introduction

Aerosols are airborne suspensions of minute particles and impose fundamental impacts on the earth's environment and climate and on human health. To understand the physical, chemical, radiative and dynamical processes associated with aerosols,

a variety of sophisticated atmospheric chemistry models have been developed and coupled with atmospheric models (Seinfeld and Pandis, 2006; Fast et al., 2006; Binkowski and Roselle, 2003). Parallel to the model development, the last decade has witnessed great progress in the technology for observing aerosols, ranging from in-situ speciated measurements to satellite- and surface-based remote sensing, leading to the establishment of a variety of observing networks (e.g., Diner et al., 2004).

The progress in both aerosol models and observing networks facilitates the development and implementation of aerosol data assimilation. Data assimilation is a methodology for the integration of all available observations into models to produce aerosol fields, which can be used to provide model initial conditions to improve forecasts, perform diagnostic analyses, and for other applications. The meteorological community has employed data assimilation for more than three decades to provide optimal initial conditions for numerical weather prediction models and develop reanalysis products for a wide spectrum of applications (Kalnay, 2003). In recent years, data assimilation has increasingly been applied to aerosol analysis.

Here we present an aerosol three-dimensional variational data assimilation (3-DVAR) scheme. This 3-DVAR scheme is developed for the WRF/Chem (Grell et al., 2005), with the comprehensive aerosol scheme known as the Model for Simulating Aerosol Interactions and Chemistry (MOSAIC) (Zaveri et al., 2008). MOSAIC was first implemented in WRF/Chem by Fast et al. (2006). Leveraging the generality of MOSAIC, this 3-DVAR system is used to estimate multi-species concentrations and their size distributions, and assimilate observations of not only total concentrations but also speciated concentrations.

The outline of this paper is as follows: In Sect. 2, some challenges faced by aerosol data assimilation and the strategies for meeting those challenges are described. In Sect. 3, a brief description of the MOSAIC scheme is given, and the analysis variables used in the 3-DVAR scheme are defined; Sect. 4 presents the 3-DVAR scheme and explains in detail the relationship between the observed and modeled variables. In Sect. 5, the method used to estimate the background error covariance is detailed, with

an emphasis on the vertical correlations. In Sect. 6, the 3-DVAR system is applied to the prediction of $PM_{2.5}$ in the Los Angeles basin during the CalNex (California Research at the Nexus of Air Quality and Climate Change) 2010 field experiment, and assessments of the performance of the data assimilation methodology and forecasts are presented. Finally, a summary and discussion are given in Sect. 7.

2 Challenges and strategies

Aerosol data assimilation faces some fundamental difficulties beyond those encountered in meteorological data assimilation. The difficulties arise primarily in treating a large number of state variables. A sophisticated model may explicitly treat more than a dozen species, which involve not only mass concentrations but also number concentrations. In particular, a large number of state variables are required to represent the wide range of aerosol size distributions, ranging from a few nanometers to around 100 μm in diameter. A modal method represents the size distributions by fitting the size distribution to a set of log-normal functions (Whitby, 1978). Four log-normal functions, known as the nucleation, Aitken, accumulation and coarse particle modes, are often used (Seinfeld and Pandis, 2006). Another method uses a set of bins of increasing size, and is referred to as a sectional or bin method (Gelbard et al., 1980; Jacobson, 1997). With either of these two methods, scores of state variables are then needed. A third method is to track the moments of the aerosol population (McGraw, 1997; Bauer et al., 2008). Although the moment method has been shown to be quite efficient, it still can lead to a large number of variables.

The large number of state variables poses multiple challenges in the practical implementation of data assimilation. Data assimilation is computationally demanding by nature. For generic formulations of different data assimilation schemes and their relationships, readers are referred to Daley (1991), Courtier et al. (1994), Ide et al. (1997), Cohn (1997), Ménard and Daley (1996) and Li and Navon (2001). Among the widely used data assimilation schemes, the three-dimensional variational data assimilation

(3-DVAR) scheme – the type used here – is the most computationally efficient. A 3-DVAR scheme iteratively minimizes a cost function that depends on error covariance matrices. Its demands on computational and memory resources increase rapidly as the number of state variables increases (see Sect. 4). A greater challenge is related to the limited number of observations. There are only a few hundred aerosol measuring surface stations in the United States, one of most dense networks in the world, and the measurements are limited to a few parameters and to the surface. Instrumented-aircraft measurements are even more limited in space and time. Satellite measurements provide global coverage, and the most common satellite observations are aerosol optical depths (AODs). The available observations are insufficient to constrain all the variables at spatial and temporal scales dictated by the inhomogeneity of aerosol emission sources and their relatively short atmospheric residence time.

Due to the afore-mentioned computational and observational requirements, it will be practically impossible for many years to establish a data assimilation system that can simultaneously and reliably estimate mass and number concentrations of all the major species at the size bins. We must therefore judiciously choose a limited number of variables to estimate, based on the aerosol treatment schemes, the desired accuracy for a given application, the geographical regions (urban, remote continental areas, oceans, etc.), the effectiveness of the use of observations available, and the computational feasibility.

Along this line, two types of schemes have been implemented. One scheme can be traced back to Collins et al. (2001), who assimilated AODs in a three-dimensional chemical transport model. This scheme uses AODs as the only data assimilation analysis variable and estimates their increments. Then the estimated AOD increments are transformed into species mass concentration increments, which are in turn added to the model forecast. Following Collins et al. (2001), a number of studies assimilated AODs in global and regional chemical transport models (Yu et al., 2003; Generoso et al., 2007; Adhikary et al., 2008; Zhang et al., 2008; Sandu and Chai, 2011). Another scheme first estimates the total aerosol mixing ratio increment and then distributes the

total increment to mass concentrations of individual species. This type of scheme was presented in Benedetti et al. (2008; 2009) and Mangold et al. (2011). In air quality oriented applications, the total concentration, which is often the concentrations of $PM_{2.5}$ (sizes smaller than 2.5 µm) and PM_{10} (size smaller than 10 µm), is used as the analysis variable (Denby et al., 2008; Tombeete et al., 2009; Pagowski et al., 2010).

The aerosol data assimilation methods described above can be characterized as two-step schemes (Liu et al., 2011). The first step is to estimate the increments of lumped variables, such as AODs, $PM_{2.5}$ and PM_{10}; and the second step is to calculate the increments of individual species at specified sizes from the lumped variable increments. These two-step schemes are suboptimal. The optimal scheme would be to directly estimate all the prognostic variables in the forecast model. When a relatively simplified aerosol scheme is used, the number of state variables may be limited so that all the state variables can be estimated simultaneously (e.g., Liu et al., 2011; Sekiyama et al., 2011). We envision that for many years to come, two-step schemes will inevitably be used for most comprehensive and sophisticated aerosol schemes, which treat scores of variables to represent mass concentrations and number concentrations with multiple size distributions, but the number of lumped variables should gradually increase as the number of observations increases and computational technology advances.

3 Aerosol scheme and analysis variables

The design of an aerosol data assimilation scheme should be dependent on the aerosol scheme used. In particular, different aerosol schemes generally require different data assimilation analysis variables. Here we use the MOSAIC scheme in WRF/Chem.

MOSAIC is an actively developed aerosol module in WRF/Chem. It treats eight aerosol species, including elemental/black carbon (EC/BC), organic carbon (OC), nitrate (NO_3^-), sulfate (SO_4^{2-}), chloride (Cl^-), ammonium (NH_4^+), sodium (Na^+). Other unspecified inorganic species such as silica (SiO_2), other inert minerals, and trace

metals are lumped together as "other inorganic mass" (OIN). A sectional approach is adopted to represent aerosol size distributions. The size bins are defined by their lower and upper dry particle diameters. Each bin is assumed to be internally mixed so that all particles within a bin have the same chemical composition, while particles in different bins are externally mixed. The number of size bins, denoted as N_{bin} here, can be specified as appropriate for different applications. In MOSAIC, hence, the state variables consist of mass concentrations of as many as $8N_{\mathrm{bin}}$, along with the number concentrations of as many as N_{bin}.

While MOSAIC offers flexibility in specifying the number of size bins, four or eight bins are commonly used. Here 4 bins are used. Accordingly, the state variables consist of 32 mass concentrations and 4 number concentrations. Because of this large number of state variables, it is computationally impractical, if not impossible, to estimate all the variables simultaneously for the model resolution and domain size as we are concerned with. We thus choose instead to use a set of lumped variables as the data assimilation analysis variables and employ a two-step data assimilation scheme.

As in previous studies (Denby et al., 2008; Benedetti et al., 2009; Tombeete et al., 2009; Pagowski et al., 2010), we may use $PM_{2.5}$ or PM_{10} as the analysis variables. In this study, four bins are used, and they are located between 0.039–0.1 µm, 0.1–1.0 µm, 1.0–2.5 µm, and 2.5–10 µm. The total mass concentration of $PM_{2.5}$ or PM_{10} can be expressed as a lumped variable. Here we introduce more analysis variables and form lumped variables consisting of the total mass concentrations of the aforementioned eight species. Specifically, one data assimilation analysis variable is the total mass concentration of one aerosol specie, that is, the sum of the mass concentrations across the size bins used. We note that the data assimilation is designed to allow further lumping some of these eight variables for particular applications, as well as using the first two, three or all four bins.

Once the lumped variables are formed, the two-step scheme is used to obtain the analysis increment of these lumped variables by solving a 3-DVAR problem, and then to partition these increments into increments for the individual species in each of the size

bins. These partitioned increments are then added to the model forecast to produce the final analysis. Accordingly, the number of aerosols for each bin is adjusted.

4 Data assimilation scheme

Here we describe the basic framework and then address the partition of the increments of the lumped variables into the increments for the individual species in each of the size bins.

4.1 Basic formulation

We consider five analysis variables, x_{EC}, x_{OC}, $x_{\mathrm{NO_3}}$, $x_{\mathrm{SO_4}}$ and x_{OTR}, which are the total mass concentrations of EC, OC, NO_3^-, SO_4^{2-}, and OTR. The chloride, ammonium, sodium and "other organic aerosol" concentrations are lumped into one single variable OTR. Here we do not use eight species but only five species as analysis variables for simplicity and also because the speciated measurements that are assimilated later correspond to these five species. Following the notation suggested by Ide et al. (1997), we express these five vectors as x, that is, $x^{\mathrm{T}} = \left(x_{\mathrm{EC}}^{\mathrm{T}}, x_{\mathrm{OC}}^{\mathrm{T}}, x_{\mathrm{NO_3}}^{\mathrm{T}}, x_{\mathrm{SO_4}}^{\mathrm{T}}, x_{\mathrm{OTR}}^{\mathrm{T}}\right)$, where T stands for transpose. The incremental form of the 3-DVAR cost function is written as:

$$J(\boldsymbol{\delta x}) = \frac{1}{2}\boldsymbol{\delta x}^{\mathrm{T}}\mathbf{B}^{-1}\boldsymbol{\delta x} + \frac{1}{2}(\mathbf{H}\boldsymbol{\delta x} - \boldsymbol{d})^{\mathrm{T}}\mathbf{R}^{-1}(\mathbf{H}\boldsymbol{\delta x} - \boldsymbol{d}). \quad (1)$$

Here $\boldsymbol{\delta x}$ is the N-vector, known as the incremental state variable, which is defined as $\boldsymbol{\delta x} = x - x^{\mathrm{b}}$, where x^{b} is the forecast or background state generated by the MOSAIC scheme in WRF/Chem. $\mathbf{B}$ is the N × N-matrix, denoting the error covariance associated with x^{b}. The M-vector $\boldsymbol{d} = y - \mathbf{H}x^{\mathrm{b}}$ is known as the observation innovation, where $\boldsymbol{y}$ is an observation vector and the M × M-matrix $\mathbf{R}$ is the observation error covariance associated with the observation y. The M × N-matrix $\mathbf{H}$ is an observational operator that maps the state variable to the observation and is assumed to be linear here. The

minimization solution is the so-called analysis increment $\boldsymbol{\delta x}^{\mathrm{a}}$, and the final analysis is $x^{\mathrm{a}} = x^{\mathrm{b}} + \boldsymbol{\delta x}^{\mathrm{a}}$. This analysis is statistically optimal as a minimum error variance estimate (e.g., Jazwinski, 1970; Cohn, 1997) or a maximum likelihood (Bayesian) estimate if both forecast and observation errors have Gaussian distributions.

4.2 Construction of background error covariance

For a given set of observations, the performance of a 3-DVAR scheme is dictated by the specified background error covariance **B** and the observational error covariance **R** in (1). **R** can generally be specified in a straightforward way and will not be discussed in detail here. Statistically, an accurate estimate of **B** is required to render the analysis x^{a} the maximum likelihood estimate. More specifically, **B** plays the role of spreading out observational information contained in y to nearby model grid-points, smoothing out small scale noise, and enforcing basic dynamic balance constraints.

In practice, however, **B** is incorporated in a suboptimal way. The primary reason is that **B** is too large to handle numerically. For a high resolution model such as that used here, the number of model grid points is on the order of 10^6. The number of elements in **B** is therefore 10^{12} multiplied by the square of the number of analysis variables. With this size, **B** cannot be explicitly manipulated. A simplification or parameterization of **B** is required.

To pursue simplifications, we use the following factorization

$$\mathbf{B} = \mathbf{DCD}^{\mathrm{T}}, \tag{2}$$

where **D** is the root-mean-square error (RMSE) matrix, a diagonal matrix whose elements are RMSEs, and **C** is the correlation matrix. With this factorization, the RMSE and correlation matrices can be described and prescribed separately. Since **D** has an effective dimension the same as that of $\boldsymbol{\delta x}$, it is computationally treatable. The simplification applies primarily to **C**.

To simplify **C**, we use a three step procedure. The Cholesky factorization is first applied. Since **C** is symmetric and positive definite, the Cholesky factorization is

$$\mathbf{C} = \mathbf{C}^{\frac{1}{2}}(\mathbf{C}^{\frac{1}{2}})^{\mathrm{T}}, \tag{3}$$

where, the matrix $\mathbf{C}^{\frac{1}{2}}$ is a lower triangular matrix. This Cholesky factorization is used to ensure the symmetry and positive definiteness of **C** and reduce the computational cost. Using this Cholesky factorization, we can transform the analysis variable δx to δz through

$$\delta x = \mathbf{DC}^{\frac{1}{2}}\delta z. \tag{4}$$

Substituting Eq. (4) into Eq. (1), we obtain the desired form of Eq. (1) as

$$J(\delta z) = \frac{1}{2}\delta z^{\mathrm{T}}\delta z + \frac{1}{2}\left(\mathbf{HDC}^{\frac{1}{2}}\delta z - \boldsymbol{d}\right)^{\mathrm{T}}\mathbf{R}^{-1}\left(\mathbf{HDC}^{\frac{1}{2}}\delta z - \boldsymbol{d}\right). \tag{5}$$

The transformed cost function is generally better conditioned, and thus this transformation expedites the convergence when it is iteratively minimized.

We here assume that the background errors of different types of aerosols are not correlated. This is an ad hoc assumption and is used simply to circumvent the computational complexity of treating the cross-correlations between different types of aerosols. Using this assumption, **C** becomes a block diagonal matrix, with the main diagonal blocks being the correlation matrices of individual types of aerosols, and we have

$$\mathbf{C} = \begin{pmatrix} \mathbf{C}_{\mathrm{BC}} & & & & \\ & \mathbf{C}_{\mathrm{OC}} & & & \\ & & \mathbf{C}_{\mathrm{NO_3}} & & \\ & & & \mathbf{C}_{\mathrm{SO_4}} & \\ & & & & \mathbf{C}_{\mathrm{OTR}} \end{pmatrix}, \tag{6}$$

where $\mathbf{C}_{\mathrm{BC}}$, $\mathbf{C}_{\mathrm{OC}}$, $\mathbf{C}_{\mathrm{NO_3}}$, $\mathbf{C}_{\mathrm{SO_4}}$ and $\mathbf{C}_{\mathrm{OTR}}$ are the background error correlation matrices associated with the five types of aerosols.

Finally, a fundamental simplification of $\mathbf{C}$ can be achieved following Li et al. (2008a, b), in which a Kronecker product method is used to construct $\mathbf{C}$. Let $\mathbf{C}_S$ denote the background error correlation matrix of one specie, where S stands for EC, OC, NO_3^-, SO_4^{2-} or OTR. We can then approximately express $\mathbf{C}_S$ as

$$\mathbf{C}_S = \mathbf{C}_{Sx} \otimes \mathbf{C}_{Sy} \otimes \mathbf{C}_{Sz}, \tag{7}$$

where $\otimes$ denotes Kronecker product, which is also known as vector or tensor product (Graham, 1981). Here x, y, and z stand for the three coordinate directions in longitude, latitude and the vertical. $\mathbf{C}_{Sx}$ is thus an $n_x \times n_x$ correlation matrix in the x-direction. Accordingly, $\mathbf{C}_{Sy}$ is an $n_y \times n_y$ correlation matrix in the y-direction, and $\mathbf{C}_{Sz}$ an $n_z \times n_z$ correlation matrix in the z-direction. Here n_x, n_y, and n_z are the numbers of grid points in the x, y and z directions, respectively. The most desirable property of Eq. (7) is that these three one-dimensional correlation matrices are computationally treatable since n_x, n_y, and n_z are on the order of 10^2 to 10^3 in size in present day's atmospheric chemical models.

Another desirable property of Eq. (7) is that we have the Cholesky factorization

$$\mathbf{C}_S^{\frac{1}{2}} = \mathbf{C}_{Sx}^{\frac{1}{2}} \otimes \mathbf{C}_{Sy}^{\frac{1}{2}} \otimes \mathbf{C}_{Sz}^{\frac{1}{2}}. \tag{8}$$

Because of their treatable dimensions, $\mathbf{C}_{Sx}^{\frac{1}{2}}$, $\mathbf{C}_{Sy}^{\frac{1}{2}}$, and $\mathbf{C}_{Sz}^{\frac{1}{2}}$ are always pre-computed and saved, which renders this 3-DVAR scheme particularly efficient computationally. As showed in Li et al. (2008a, b), Eqs. (7) or (8) allows incorporating some forms of inhomogeneity and anisotropy in the correlations, while a particular desirable advantage of Eq. (7) is the capability of representing complex vertical correlations, which will be further discussed when the estimates of $\mathbf{C}_{Sx}$, $\mathbf{C}_{Sy}$, and $\mathbf{C}_{Sz}$ are described in Sect. 5.

4.3 Increments of lumped variables to species

The data assimilation scheme formulated above estimates five lumped variables formed from the mass concentrations of the aerosol species treated in MOSAIC, but

the ultimate analysis solution is concerned with the mass concentrations of those individual species. By definition, the increment of a lumped variable can be expressed as

$$\delta x_S = \sum_{l=1}^{L} \delta m_l, \tag{9}$$

where S again stands for EC, OC, NO_3^-, SO_4^{2-}, or OTR. Here δm_l is the mass concentration of one MOSAIC specie for a single size bin. Since EC, BC, NO_3^-, and SO_4^{2-} are MOSAIC species, the summation is across all the size bins, and L equals N_{bin}, the number of the size bins used. Since OTR includes four MOSAIC species, the summation is for the four species and across all the size bins, and L thus becomes $4N_{\text{bin}}$ for OTR. We calculate the analysis increment of mass concentration for one MOSAIC specie and one size bin as

$$\delta m_{Sl}^{\text{a}} = \frac{\sigma_{Sl}^2}{\sum_{i=1}^{L} \sigma_{Sl}^2} \delta x_S^{\text{a}}, \tag{10}$$

where σ_{Sl} are the root-mean-square (RMS) of the mass concentration background error for each specie and size bin. Equation (10) is used since it can be derived as a minimum error variance estimation with the constraint (Eq. 9) and under the assumption that δx_S^{a} has no error and the background errors associated with m_{Sl} are uncorrelated.

With δm_{Sl}^{a} at our disposal, we can obtain the final analysis mass concentrations for each MOSAIC species and size bin as

$$m_{Sl}^{\text{a}} = \begin{cases} m_{Sl}^{\text{b}} + \delta m_{Sl}^{\text{a}} & \text{if } \delta m_{Sl}^{\text{a}} \geq 0 \\ m_{Sl}^{\text{b}} + \delta m_{Sl}^{\text{a}} & \text{if } \delta m_{Sl}^{\text{a}} < 0; \quad m_{Sl}^{\text{b}} + \delta m_{Sl}^{\text{a}} \geq c_{rSl} \\ m_{Sl}^{\text{b}} & \text{if } \delta m_{Sl}^{\text{a}} < 0; \quad m_{Sl}^{\text{b}} + \delta m_{Sl}^{\text{a}} < c_{rSl} \end{cases}, \tag{11}$$

where c_{rSl} is a positive value. We use c_{rSl} to ensure that m^{a}_{Sl} is non-negative, and no reduction is made to the background state when the background concentration is lower than the observational errors. In practice, aerosol observational errors are often not well characterized. In the current application, we empirically specify $c_{rlS} = 0.1\sigma_{Sl}$.

After the mass concentration increments are added to the background state, the number concentration increments should be added accordingly. For simplicity, we assume that the ratio of the number concentration to the mass concentration for each size bin remains to the same as it was in the background state. Let's denote this ratio as γ. γ is calculated from the background state. The analysis of the number concentrations is

$$n^{a}_{k} = n^{b}_{k} + \gamma\left(m^{a}_{k} - m^{b}_{k}\right), \quad k = 1, \ldots, N_{\text{bin}} \tag{12}$$

where n^{b}_{k} is the background number concentration within one size bin, and m^{a}_{k} is the total mass concentration summed for all the species within one size bin and computed from m^{a}_{Sl} given by Eq. (11).

4.4 Observational operators

In present observing networks, in-situ observations primarily consist of total surface concentrations of particular matter, such as $PM_{1.0}$, $PM_{2.5}$ or PM_{10}, and speciated concentrations of EC, OC, NO^{-}_{3}, SO^{2-}_{4} and others. These two types of observations are assimilated in this study. The observational operators **H** represent the summation of the concentrations of the corresponding species across the size bins and a linear horizontal interpolation operator The total concentrations of particular matter are indirect observations, and the observational operators then represent the summation of the concentrations of all the species across the appropriate size bins. For example, the observational operator for the concentration of $PM_{2.5}$ represents the summation of the concentrations of all the species across the first three size bins (see Sect. 3).

The observational operator can be complicated when the observations are not the measured concentration at a given time, but the accumulated concentration for a period of time. A typical example is the observations obtained from the IMPROVE (Interagency Monitoring of Protected Visual Environments) network (http://vista.cira.colostate.edu/improve/). The observations are daily mean concentrations. To assimilate such temporal mean concentrations, we first compute the temporal mean observation innovation. The computed innovation is then assumed to be constant during the period of time and thus assimilated as the innovation at the given time.

For the daily mean concentrations from IMPROVE, for example, we compute the daily mean innovation for the period of time 12 h before and after the data assimilation time. The assumption used here is similar to that used in the First-Guess at the Appropriate Time (FGAT) method, which has been employed in several meteorological data assimilation systems (e.g., Simmons, 2000; Lorenc et al., 2000).

5 WRF/Chem configuration and background error covariance estimate

The WRF/Chem model is configured as a nested set of three spatial domains (Fig. 1). The large domain encompasses the Western United States and adjacent coastal region (36 km grid), the middle domain a smaller portion of the western states and the California coast (12 km grid), and the small domain the Los Angeles Basin (4 km grid). The nesting is two-way for both interior domains. Each domain has 40 vertical levels, with the vertical grid stretched to place the highest resolution in the lower troposphere. The analyses presented here will be primarily confined to the 4 km domain.

We use version 3.3 of WRF/Chem. In WRF/Chem, the chemistry (both aerosol and gas-phase) and meteorological components are fully coupled (Grell et al., 2005; Fast et al., 2006). The parameterizations of physical processes used that are most relevant to aerosols are summarized as follows: the Monin-Obukhov surface layer scheme, the Yonsei University (YSU) boundary-layer scheme, the Morrison 2-moment microphysics scheme, the Noah land surface model and the Dudhia radiation scheme for longwave

and shortwave interactions with clear-air and clouds (Skamarock et al., 2005 and references therein). For the chemical processes, the MOSAIC (4 bin) aerosol scheme is used. The Carbon Bond Mechanism (CBM-Z) scheme is used for the Gas-phase chemistry processes. The emissions were derived from National Emission Inventory 2005 (NEI'05) for both aerosols and trace gases (Mckeen et al., 2002).

In Sect. 4.2, we developed an approximate expression for the background error covariance. According to Eqs. (2) and (7), we need to estimate the RMSE matrix **D** and one-dimensional correlation matrices for each analysis variable. To estimate these matrices, we follow a methodology used in meteorology. An overview of methods to diagnose background error statistics for application in Numerical Weather Prediction (NWP) is provided in Bannister (2008). The main approaches are based on either statistics of observations minus model differences at observation locations, or on model fields generated on the model grid that can be used statistically as a proxy of the background error; this second method is known as the NMC method (Parrish and Derber, 1992). The observation innovation method cannot be used in aerosol data assimilation because of the lack of three-dimensional speciated observations, and hence the NMC method is used here.

The NMC method has been used for estimating the aerosol concentration background error covariance (Benedetti and Fisher, 2007; Kahnert, 2008). In the ECMWF 4DVAR system (Benedetti and Fisher, 2007), the differences between 48 h and 24 h forecasts of aerosol mixing ratios for sea salt, desert dust and continental particulates are assumed to be a proxy of the background error. While the NMC method has not yet been fully justified in the context of aerosol data assimilation, the ECMWF 4DVAR system indicates that it can be useful at least in some circumstances. We estimate the aforementioned error covariance matrices following Benedetti and Fisher (2007).

We generate one month of the 48 h and 24 h forecast differences from 15 May 2010 to 14 June 2010. The forecast is initialized using the North American Regional Reanalysis (NARR) (Mesinger et al., 2006). The interior boundary conditions and sea surface temperatures are updated at each initialization, with the lateral boundary conditions

updated continuously throughout the forecast. Note that the NARR fields do not include any aerosol variables. The initial conditions for the aerosol species are simply from the forecast without being updated. Hence, the forecast difference arises from the difference in the meteorological fields, which in turn give rise to differences in vertical and horizontal transports, dry and wet depositions, and chemical processes that are sensitive to temperature, moisture and cloud water content. It is assumed that these differences are representative of the short-term forecast errors in transport-related aerosol processes, and can be used to calculate background error statistics.

We directly estimate the RMSE matrix **D**. The domain average RMSEs for five species are shown in Fig. 2. The RMSEs differ among the species. The largest RMSE is associated with NO_3^-, while the smallest RMSE is associated with OC. The vertical distributions of the RMSE for all the species display a relatively rapid decrease with height.

The fine structures of the RMSE vertical distribution are related to boundary layer heights. In Fig. 3, the boundary layers primarily consist of marine layers with depths of less than 400 m, and inland boundary layers with depths of around 1000 m. There is a noticeable increase in the SO_4^{2-} RMSEs at the boundary layer height.

The three-dimensional correlation matrix $\mathbf{C}_S$ has been approximately factorized and expressed as the Kronecker product of three one-dimensional correlation matrices $\mathbf{C}_{Sx}$, $\mathbf{C}_{Sy}$ and $\mathbf{C}_{Sz}$ for each species in Eq. (7), and these three one-dimensional matrices need to be estimated. Although the horizontal correlation matrices $\mathbf{C}_{Sx}$ and $\mathbf{C}_{Sy}$ can be directly computed using the proxy of the background error (Li et al., 2008b), we here use Gaussian functions to represent then correlation functions for constructing $\mathbf{C}_{Sx}$ and $\mathbf{C}_{Sy}$, and also assume that the correlations are isotropic. With these two assumptions, a correlation function between two points r_1 and r_2 in the horizontal can be expressed as $c_S(r_1, r_2) = e^{-(r_2-r_1)^2/2L_S^2}$, where L_S is the horizontal correlation scale. The correlation length scale L_S becomes the only parameter that need to be estimated. This correlation length scale is estimated using the proxy of the background error. The correlation c_S reduces to the value of e^{-1} when the distance of two points r_1 and r_2 is

measured L_S. This distance averaged over the model domain is used as the estimate of L_S. The estimated L_S are 36, 32, 20, 52, 48 km for the five species, EC, OC, NO_3^-, SO_4^{2-}, and OTR, respectively.

The estimated correlation length scales are significantly different among distinct species. The largest scale is associated with SO_4^{2-}, which indicates that the background error has relatively large scales horizontally, and the influence of an SO_4^{2-} observation could spread farther than other species. In contrast, the smallest correlation length scale is associated with NO_3^-, and it is about 2/5 of the scale associated with SO_4^{2-}. Such differences among the correlation length scales suggest a need to use multi-species concentrations as the analysis variables.

The vertical correlation matrices $\mathbf{C}_{Sz}$ are computed directly from the proxy data. This can be done since $\mathbf{C}_{Sz}$ is only an $n_z \times n_z$ matrix. A directly-computed $\mathbf{C}_{Sz}$ helps represent the complicated vertical structures of the vertical correlations due to the discontinuity-like transition of the vertical distributions between the boundary layer and the free atmosphere above. Such structures are difficult to represent using analytical functions. The computed vertical correlation matrices of $\mathbf{C}_{Sz}$ are displayed in Fig. 4. The vertical correlation for OC is not shown, because it is similar to that for EC. A salient and common feature of these vertical correlations is their strong relation to the boundary layer heights. Consistent with the discontinuity-like transition in the background vertical RMSE around the top of the boundary layer, the vertical correlations display a jump at the top of boundary layer. Another feature is worth mentioning is that the vertical correlation for NO_3^- shows a relatively small vertical scale.

6 Applications to the prediction of $PM_{2.5}$

The Greater Los Angeles area continues to be the most polluted metropolitan region in the U.S., and prediction of the spatial and temporal distributions of $PM_{2.5}$ in the region remains a challenge. Here the 3-DVAR system is used to assimilate $PM_{2.5}$

measurements along with some speciated concentration measurements, and then predictions are performed. This data assimilation and prediction experiment has been carried out for a period of one month from 12:00 UTC, 15 May to 12:00 UTC, 14 June 2010. This period of time was chosen because the CalNex 2010 field experiment took place at this time (http://www.esrl.noaa.gov/csd/calnex/). CalNex 2010 was a major climate and air quality study in California conducted by the National Oceanic and Atmospheric Administration (NOAA) and the California Air Resources Board (ARB), and thus more observations are available for assimilation and evaluation.

6.1 Implementation of data assimilation and forecast

To assess the data assimilation analyses and subsequent predictions, we compare the results from the experiments with and without data assimilation. The experiments are carried out as follows. The forecast is initialized at 00:00 UTC daily using the NARR meteorological fields, and then the forecast is run out to 36 h, that is, till 12:00 UTC on the following day. The first 12 h of the forecast are discarded as a model spin-up for the meteorological fields. This spin-up allows the WRF/Chem to produce not only proper clouds and precipitation but also fine structures in the wind fields associated with the higher resolution of the model and surface topography. The chemical fields, including both the gaseous and aerosol species, are updated daily at 12:00 UTC after the 12 h spin-up of the meteorological fields, using the forecast from the previous day. For convenience, we refer to these results as the control analyses or forecasts.

The aerosol data assimilation is carried out every 6 h. A time window of 6 h is used to resolve diurnal variations. Specifically, the data assimilation is carried out at 12:00 UTC, after the above-mentioned spin-up, and the 6 h aerosol forecast initialized by the prior aerosol data assimilation analysis valid at 06:00 UTC is used as the background state. Then the aerosol analysis obtained is used as the initial condition at 12:00 UTC for the subsequent 6 h forecast, which generates the background state for the aerosol data assimilation at 18:00 UTC. In this way, the data assimilation and forecast cycle can continuously advance in time. It is worth noting that the aerosol feedback to the

radiation and cloud processes is turned off in the WRF/Chem and thus the meteorological fields are exactly the same as those in the control experiment.

The aerosol observations assimilated have two types as mentioned in Sect. 4.4. One type is mass concentrations of $PM_{2.5}$ from Air Resources Board (ARB) of California Environmental Protection Agency (available at http://www.arb.ca.gov/aqmis2/aqdselect.php), which consist of hourly concentration measurements. There are a total of 42 stations in the model domain (Fig. 1). The hourly observations are assimilated, but only at those hours when the data assimilation is carried out. The other type are the daily mean speciated mass concentrations from the IMPROVE network, and the daily mean innovations (see Sect. 4.4) are assimilated at 12:00 UTC. For those observations assimilated, a simple quality control is applied. First, observations with negative values are rejected; second, the differences between the observations and 6 h forecasts (O-F) are examined, and those observations with the O-F values greater than 120 $g\,m^{-3}$ are rejected. The observational error is specified as half of the background root mean square errors.

Additional observations of speciated concentrations are available during the CalNex 2010 experiment (available at http://www.esrl.noaa.gov/csd/calnex/) from one station located on the California Institute of Technology campus (see Fig. 1). These observations consist of the concentrations of EC, OC, NO_3^-, SO_4^{2-}, and others. They are not assimilated but are used as independent observations to assess the skill of the analyses.

6.2 Assessment of data assimilation analyses

We first compare the analysis $PM_{2.5}$ against the observations that are assimilated. To quantify the difference, we analyze biases, correlations and root-mean-square differences between data assimilation analyses and observations. For simplicity, we refer to root-mean-square differences as root-mean-squared errors (RMSEs), although they are not RMSEs in a strict sense because the observation errors can be significant.

These biases, correlations and RMSEs will also be used later in analyzing the forecasts.

Figure 5 presents a scatter plot of the $PM_{2.5}$ concentrations from the control and data assimilation analysis, respectively, against the observations for a period of one month. Observations from each of the 42 stations and at all four times of the day with data assimilation, that is, at 00:00, 06:00, 12:00 and 18:00 UTC, are used. The control $PM_{2.5}$ results display a significant underestimation. The observed mean concentration of $PM_{2.5}$ is 21.5 $\mu g\,m^{-3}$, while that of the control analysis is 14.9 µm, a bias of −7.6 $\mu g\,m^{-3}$ and thus about 30% lower than the observed. The correlation is 0.51 and the RMSE is 11.0 $\mu g\,m^{-3}$, compared with a standard deviation of 26.8 $\mu g\,m^{-3}$ in the observed $PM_{2.5.}$ In the data assimilation analysis, the bias is greatly reduced, to as small as −1.0 $\mu g\,m^{-3}$. The correlation between the analysis and observed $PM_{2.5}$ is as high as 0.87, while the RMSE decreases to 4.2 $\mu g\,m^{-3}$. These results show that this 3-DVAR scheme can effectively assimilate the $PM_{2.5}$ observations.

We have introduced multi-species concentrations as analysis variables, aiming to enhance the capability of the scheme in reproducing species concentrations. The concentrations of four major species – EC, OC, NO_3^-, and SO_4^{2-} are evaluated here, using the speciated observations obtained at the super observing station (Fig. 1). We note that these speciated observations are not assimilated and thus are independent data.

Figure 6 shows the scatter plots of analysis species concentrations against the observations at the super observing station. The concentrations of all four species are improved in both their correlation and RMSE. Relative to the control results, the correlation increases from 0.1 for EC to 0.2 for NO_3^- and the RMSEs are reduced by 10 % for OC to 50 % for NO^{3-}. The RMSE increases for EC, but this is due to the bias. The biases in the NO_3^- and SO_4^{2-} concentrations are appreciably reduced, but an increase in bias occurs in the EC and OC concentrations. The increase of these biases is worth discussing. The EC and OC concentrations show a positive bias in both the control and analysis results, and this bias is opposite to that of the $PM_{2.5}$ bias in sign,

while the NO_3^- and EC, SO_4^{2-} concentrations show a negative bias. Note that the assimilation of IMPROVE observations have relatively little impact on the analysis at this location, because there is no IMPROVE station in the area surrounding this location. Thus the assimilation of $PM_{2.5}$ gives rise to increases in the biases in the EC and OC concentrations. This increase in bias actually indicates an inherit difficulty when only $PM_{2.5}$ observations are assimilated, that is, the individual species concentration bias can deteriorate when the $PM_{2.5}$ and species concentration biases have opposite signs.

6.3 Assessment of forecasts

One of the ultimate goals of developing this 3-DVAR system is to improve our capability for predicting aerosol concentrations. Here we evaluate forecast skill in $PM_{2.5}$ concentrations out to 24 h. Figure 7 presents the correlations and RMSEs as a function of forecast lead time. The 24 h forecast is initialized with the analysis at 12:00 UTC, which corresponds to 05:00 LT. The 0h forecast is the data assimilation analysis at 12:00 UTC. The correlations and RMSEs show that the forecast has consistent skill. Comparing the correlations and RMSEs with those from the control forecast without aerosol data assimilation, a positive impact on skill can be seen all the way out to 24 h.

Examining all forecast durations from 6 h to 24 h, one characteristic seen is that there is not a monotonic decrease in skill as the forecast duration increases. Actually, during the period from 6 h to 24 h, the forecast correlation increases. The RMSE shows a maximum value for the 6 h forecast and is basically flat from 12 h to 24 h. Relative to the control forecast, both the correlation and RMSEs display no significant variations from 12 h to 24 h. Another characteristic is the large increase in correlation and decrease in RMSEs during the first 6 h. These two characteristics of the forecast skill are not often encountered in meteorological data assimilation, and may suggest an inherit challenge in aerosol data assimilation or the need for the incorporation of dynamic balance among aerosol components and between aerosol and gaseous components.

We have seen (Fig. 1) that the $PM_{2.5}$ observations are heterogeneously distributed. A related question is whether the forecast skill is also spatially heterogeneous in association with this heterogeneous distribution of observations. We have analyzed the spatial distribution of both the correlations (not shown) and RMSEs for the 24 h forecast. The distributions of RMSE are shown in Fig. 8. From Fig. 8a and b, we see that the RMSE of the 24 h forecast shows a significant decrease in areas where the RMSE is relatively larger in the control forecast without data assimilation, including most of the Los Angeles area. To show this decrease in the RMSE more clearly, the difference between the control forecast and the 24 h forecast with data assimilation is shown in Fig. 8c. A decrease in RMSE is seen at most of the observing locations, but there are four locations where the RMSEs increase. We note that these four locations are located along the coast.

The forecast of individual species is more challenging than that of the total $PM_{2.5}$ concentration, but the results are encouraging. Figure 9 presents the 24 h forecast correlations and RMSEs at the super observing station. Most encouraging is that, in terms of both correlations and RMSE, the error reduction in the analysis NO_3^- concentration (Fig. 9) persists up to 24 h. The correlations of the OC and SO_4^{2-} concentrations are also larger than that of the control forecast up to 24 h, and the correlation of the EC concentrations is larger than that of the control forecast up to 18h. Except the NO_3^- forecast, the RMSEs of the individual concentrations are also larger than that of the control forecast for a relatively short period of time, and this is because a significant part of the RMSEs is due to the bias.

7 Conclusions and discussions

A 3-DVAR data assimilation system for the MOSAIC aerosol scheme in WRF/Chem has been developed and presented. MOSAIC provides a comprehensive representation of aerosol species and size distributions that result from a variety of pollutant

emissions. This 3-DVAR scheme is formulated and implemented in an attempt to take full advantage of the MOSAIC scheme.

The 3-DVAR system developed can be characterized as a two-step scheme. The total concentrations of individual species, defined as the sum of the concentrations across all size bins, are used as analysis variables, and thus the analysis variables consist of the eight concentrations of the MOSAIC aerosol species. In this study, four size bins is used. The first step is to determine analysis increments for these eight concentrations following a 3-DVAR methodology (Sect. 4.1). The second step is to distribute the analysis increments over four size bins. The distribution is inversely related to the background error variances. The number concentrations are adjusted based on the ratio of the mass and number concentration in the background state.

This system was applied to the analysis and prediction of $PM_{2.5}$ in the Los Angeles basin during CalNex 2010. Surface $PM_{2.5}$ and speciated concentration observations were assimilated. To evaluate the performance of the scheme, we carried out control forecasts, which were initialized with the North America Regional Reanalysis (NARR) and used both aerosol and trace gas emissions derived from National Emission Inventory 2005 (NEI'05). The comparison of the control forecasts and the forecasts with aerosol data assimilation against observations demonstrated that the data assimilation generated analyses with significantly reduced errors, which improved the subsequent forecasts of $PM_{2.5}$ up to 24 h. We also evaluated the performance of the forecasts of elemental carbon, organic carbon, nitrate, and sulfate concentrations. The data assimilation significantly improved the forecast of sulfate concentrations up to 24 h, while the improvement extended up to 12–18 h for the other three species.

We have emphasized the use of multi-species concentrations as analysis variables. Because of this, this 3-DVAR has the capability of simultaneously assimilating total and speciated concentrations observations, such as total $PM_{2.5}$ concentrations and speciated concentrations from the IMPROVE network. The use of multi-species concentrations is also desirable for representations of the background error covariance. In Sect. 5, we showed that both the horizontal and vertical scales of the background

errors correlations are significantly different, and this difference needs to be accounted for. We carried out an experiment (not presented), in which the background error correlation of each species was replaced by that estimated for $PM_{2.5}$ using the 48 h minus 24 h forecast described in Sect. 5, and showed that this degraded the performance.

In this study, only surface observations were assimilated, but this 3-DVAR scheme has been developed to assimilate additional observations from aircraft and satellites. With more observations assimilated, this data assimilation system is expected not only to further improve forecasts but also to be of use in other applications. For example, the output can be used to interpolate limited observations for the evaluation of numerical models of aerosol-related processes such as aerosol-cloud interactions.

Despite the promising performance of this 3-DVAR system, a few fundamental assumptions used warrant further examination. The most fundamental assumption is that the background and observational errors are Gaussian. However, the evidence shows that aerosol concentrations tend to be non-Gaussian (Seinfeld and Pandis, 2006). While the background errors may not necessarily follow the aerosol concentration distributions, there is a possibility that the background errors are non-Gaussian. In addition, aerosol concentrations are strictly nonnegative quantities, therefore errors in these quantities cannot be strictly Gaussian distributed, although they may be approximately so since the Gaussian density assigns positive probability to negative values of these quantities (Cohn, 1997). A systematic analysis is needed to characterize the background error distribution. Another assumption is that the background errors are uncorrelated between different types of species and between distinct size bins. This is an ad hoc assumption made simply to render the problem computationally manageable. The correlations between different bins can be significant, because one species often arises from same sources and transfers across the bins. The consequences of this assumption should be quantified, and relaxation of this assumption should be pursued. In the present implementation, gaseous components are not involved. However, aerosol particles in the atmosphere contain a variety of volatile compounds (ammonium, nitrate, chloride, volatile organic compounds) that can exist either in the particulate or

gas phase, and these two phases are often in thermodynamic equilibrium. Assimilation of gaseous components may be required to further improve the forecast of aerosol species concentrations.

Acknowledgements. The research described in this publication was carried out, in part, at the Jet Propulsion Laboratory (JPL), California Institute of Technology, under a contract with the National Aeronautics and Space Administration (NASA). This research was supported in part by the U.S. Department of Energy Earth System Modeling (ESM) program via the FASTER project www.bnl.gov/esm/, the JPL Director's Research and Development Fund (DRDF), and NASA grants NNX09AF07G and NNX08AF64G from the Atmospheric Chemistry Modeling and Analysis Program (ACMAP).

References

Adhikary, B., Kulkarni, S., D'allura, A., Tang, Y., Chai, T., Leung, L. R., Qian, Y., Chung, C. E., Ramanathan, V., and Carmichael, G. R.: A regional scale chemical transport modeling of Asian aerosols with data assimilation of AOD observations using optimal interpolation technique, Atmos. Environ., 42, 8600–8615, doi:10.1016/j.atmosenv.2008.08.031, 2008.

Bannister, R. N.: A review of forecast error covariance statistics in atmospheric variational data assimilation. I: Characteristics and measurements of forecast error covariances, Q. J. Roy. Meteor. Soc., 134, 1951–1970, 2008.

Bauer, S. E., Wright, D. L., Koch, D., Lewis, E. R., McGraw, R., Chang, L.-S., Schwartz, S. E., and Ruedy, R.: MATRIX (Multiconfiguration Aerosol TRacker of mIXing state): an aerosol microphysical module for global atmospheric models, Atmos. Chem. Phys., 8, 6003–6035, doi:10.5194/acp-8-6003-2008, 2008.

Benedetti, A. and Fisher, M.: Background error statistics for aerosols, Q. J. Roy. Meteor. Soc., 133, 391–405, 2007.

Benedetti, A. and Janiskova, M.: Assimilation of MODIS cloud optical depths in the ECMWF model, Mon. Weather Rev., 136, 1727–1746, doi:10.1175/2007MWR2240.1, 2008.

Benedetti, A., Morcrette, J.-J., Boucher , O., Dethof , A., Engelen, R. J., Fisher, M., Flentje, H., Huneeus, N., Jones, L., Kaiser, J. W., Kinne, S., Mangold, A., Razinger, M., Simmons, A. J., and Suttie, M.: Aerosol analysis and forecast in the European Centre for Medium-Range

Weather Forecasts Integrated Forecast System: 2. Data assimilation, J. Geophys. Res., 114, D13205, doi:10.1029/2008JD011115, 2009.

Binkowski, F. S. and Roselle, S. J.: Models-3 Community Multiscale Air Quality (CMAQ) model aerosol component, 1. Model description, J. Geophys. Res., 108 (D6), 4183, doi:10.1029/2001JD001409, 2003.

Cohn, S. E. : Estimation theory for data assimilation problems: basic conceptual framework and some open questions, J. Meteorol. Soc. Jpn., 75, 1B, 257–288, 1997.

Collins, W. D., Rasch, P. J., Eaton, B. E., Khattatov, B. V., Lamarque, J.-F., and Zender, C. S.: Simulating aerosols using a chemical transport model with assimilation of satellite aerosol retrievals: methodology for INDOEX, J. Geophys. Res., 106, 7313–7336, doi:10.1029/2000JD900507, 2001.

Courtier, P., Thepaut, J.-N., and Hollingsworth, A.: A strategy for operational implementation of 4-D-Var, using an incremental approach, Q. J. Roy. Meteor. Soc., 120, 1367–1388, 1994.

Daley, R.: Atmospheric Data Assimilation, Cambridge Atmospheric and Space Science Series, Cambridge University Press, Cambridge, UK, 457 pp., 1991.

Denby, B., Schaap, M., Segers, A., Builtjes, P., and Horálek, J.: Comparison of two data assimilation methods for assessing PM_{10} exceedances on the European scale, Atmos. Environ., 42, 7122–7134, 2008.

Diner, D. J., Ackerman, T. P., Anderson, T. L., Bosenberg, J., Braverman, A. J., Charlson, R. J., Collins, W. D., Davies, R., Holben , B. N., Hostetler, C. A., Kahn, R. A., Martonchik , J. V., Menzies, R. T., Miller, M. A., Ogren, J. A., Penner, J. E., Rasch, P. J., Schwartz, S. E., Seinfeld, J. H., Stephens, G. L., Torres, O., Travis, L. D., Wielicki, B. A., and Yu, B.: PARAGON: an integrated approach for characterizing aerosol climate impacts and environmental interactions, B. Am. Meteor. Soc., 85, 1491–1501, doi:10.1175/BAMS-85-10-1491, 2004.

Fast, J. D., Gustafson Jr., W. I., Easter, R. C., Zaveri, R. A., Barnard, J. C., Chapman, E. G., and Grell, G. A.: Evolution of ozone, particulates, and aerosol direct forcing in an urban area using a new fully-coupled meteorology-chemistry-aerosol model, J. Geophys. Res., 111, D21305, doi:10.1029/2005JD006721, 2006.

Gelbard, F., Tambour, Y., and Seinfeld, J. H.: Sectional representations for simulating aerosol dynamics, J. Colloid Interf. Sci., 76, 541–556, 1980.

Generoso, S., Bréon, F.-M., Chevallier, F., Balkanski, Y., Schulz, M., and Bey, I.: Assimilation of POLDER aerosol optical thickiness into the LMDz–INCA model: implications for the Arctic aerosol burden, J. Geophys. Res., 112, D02311, doi:10.1029/2005JD006954, 2007.

Graham, A.: Kronecker Product and Matrix Calculus with Applications, Ellis Horwood Ltd., Chichester, England, 130 pp., 1981.
Grell, G. A., Peckham, S. E., Schmitz, R., McKeen, S. A., Frost, G., Skamarock, W. C., and Eder, B.: Fully coupled "online" chemistry within the WRF model, Atmos. Environ., 39, 6957–6976, doi:10.1016/j.atmosenv.2005.04.027, 2005.
Ide, K., Courtier, P., Ghil, M., and Lorenc, A. C.: Unified notation for data assimilation, J. Meteorol. Soc. Jpn., 75 (1B), 71–79, 1997.
Jacobson, M. Z.: Development and application of a new air pollution modeling system-II. Aerosol module structure and design, Atmos. Environ., 31, 131–144, 1997.
Jazwinski, A. H.: Stochastic Processes and Filtering Theory, Academic Press, New York, 376 pp., 1970.
Kalnay, E.: Atmospheric modeling, data assimilation and predictability, Cambridge Univ. Press, New York, 341 pp., 2003.
Kahnert, M.: Variational data analysis of aerosol species in a regional CTM: background error covariance constraint and aerosol optical observation operators, Tellus B, 60, 753–770, doi:10.1111/j.1600-0889.2008.00377.x, 2008.
Li, Z., Chao, Y., McWilliams, J. C., and Ide, K.: A three-dimensional variational data assimilation scheme for the Regional Ocean Modeling System, J. Atmos. Ocean. Tech., 25, 2074–2090, 2008a.
Li, Z., Chao, Y., McWilliams, J. C., and Ide, K.: A three-dimensional variational data assimilation scheme for the Regional Ocean Modeling System: implementation and basic experiments, J. Geophys. Res., 113, C05002, doi:10.1029/2006JC004042, 2008b.
Li, Z. and Navon, I. M.: Optimality of variational data assimilation and its relationship with the Kalman filter and smoother, Q. J. Roy. Meteor. Soc., 127, 661–683, 2001.
Liu, D. C. and Nocedal, J.: On the limited memory BFGS method for large scale optimization, Math. Program., 45, 503–528. doi:10.1007/BF01589116, 1989.
Liu, Z., Liu, Q., Lin, H.-C., Schwartz, C. S., Lee, Y.-H., Wang, T: Three-dimensional variational assimilation of MODIS aerosol optical depth: implementation and application to a dust storm over East Asia, J. Geophys. Res., 116, D23206. doi:10.1029/2011JD016159, 2011.
Lorenc, A. C., Ballard, S. P., Bell, R. S., Ingleby, N. B., Andrews, P. L. F., Barker, D. M., Bray, J. R., Clayton, A. M., Dalby, T., Li, D., Payne, T. J. and Saunders, F. W.: The Met. Office global 3-dimensional variational data assimilation scheme, Q. J. Roy. Meteor. Soc., 126, 2991–3012, 2000.

Mangold, A., De Backer, H., De Paepe, B., Dewitte, S., Chiapello, I., Derimian, Y., Kacenelenbogen, M., Léon, J.-F., Huneeus, N., Schulz, M., Ceburnis, D., O'Dowd, C., Flentje, H., Kinne, S., Benedetti, A., Morcrette, J.-J., and Boucher, O.: Aerosol analysis and forecast in the European Centre for Medium-Range Weather Forecasts Integrated Forecast System: 3. Evaluation by means of case studies, J. Geophys. Res., 116, D03302, doi:10.1029/2010JD014864, 2011.
McGraw, R.: Description of atmospheric aerosol dynamics by the quadrature method of moments, Aerosol Sci. Tech., 27, 255–265, 1997.
McKeen, S. A., Wotawa, G., Parrish, D. D., Holloway, J. S., Buhr, M. P., Hubler, G., Fehsenfeld, F. C., and Meagher, J. F.: Ozone production from Canadian wildfires during June and July of 1995, J. Geophys. Res., 107 (D14), 4192, doi:10.1029/2001JD000697, 2002.
Ménard, R. and Daley, R.: The application of Kalman smoother theory to the estimation of 4-DVAR error statistics, Tellus A, 48, 221–237, 1996.
Mesinger, F., DiMego, G., Kalnay, E., Mitchell, K., Shafran, P. C., Ebisuzaki, W., Jovic, D., Woollen, J., Rogers, E., Berbery, E. H., Ek, M. B., Fan, Y., Grumbine, R., Higgins, W., Li, H., Lin, Y., Manikin, G., Parrish, D., and Shi, W.: North American regional reanalysis, B. Am. Meteor. Soc., 87, 343–360, doi:10.1175/BAMS-87-3-343, 2006.
Parrish, D. F. and Derber, J. C.: The National Meteorological Center's spectral statistical interpolation analysis system, Mon. Weather Rev., 120, 1747–1763, 1992.
Pagowski, M., Grell, G. A., McKeen, S. A., Peckham, S. E., and Devenyi, D.: Three-dimensional variational data assimilation of ozone and fine particulate matter observations: some results using the weather research and forecasting – chemistry model and grid-point statistical interpolation, Q. J. Roy. Meteor. Soc., 136, 2013–2024, 2010.
Sandu, A. and Chai, T. F.: Chemical data assimilation an overview, Atmosphere, 2, 426–463, 2011.
Seinfeld, J. H. and Pandis, S. N.: Atmospheric Chemistry and Physics: From Air Pollution to Climate Change, 2nd edn., J. Wiley, New York, 2006.
Sekiyama, T. T., Tanaka, T. Y., Shimizu, A., and Miyoshi, T.: Data assimilation of CALIPSO aerosol observations, Atmos. Chem. Phys., 10, 39–49, doi:10.5194/acp-10-39-2010, 2010.
Simmons, A.: Assimilation of satellite data for numerical weather prediction: basic importance, concepts and issues, ECMWF seminar proceedings, Exploitation of the new generation of satellite instruments for numerical weather prediction, 4–8 September 2000, Reading, UK, 21–46, 2000.

Skamarock, W. C., Klemp, J. B., Dudhia, J., Gill, D. O., Barker, D. M., Duda, M. G., Huang, X., Wang, W., and Powers, J. G.: A description of the advanced research WRF version 3, NCAR Tech. Note, NCAR/TN-475 + STR, 8 pp., Natl. Cent. for Atmos. Res., Boulder, Colorado, available at: http://www.mmm.ucar.edu/wrf/users/docs/arw_v3.pdf, 2008.

Tombette, M., Mallet, V., and Sportisse, B.: PM_{10} data assimilation over Europe with the optimal interpolation method, Atmos. Chem. Phys., 9, 57–70, doi:10.5194/acp-9-57-2009, 2009.

Whitby, K. T.: The physical characteristics of sulfate aerosols, Atmos. Environ., 12, 135–159, 1978.

Yu, H., Dickinson, R. E., Chin, M., Kaufman, Y. J., Geogdzhayev, B., and Mishchenko, M. I.: Annual cycle of global distributions of aerosol optical depth from integration of MODIS retrievals and GOCART model simulations, J. Geophys. Res., 108 (D3), 4128, doi:10.1029/2002JD002717, 2003.

Zaveri, R. A., Easter, R. C., Fast, J. D., and Peters, L. K.: Model for Simulating Aerosol Interactions and Chemistry (MOSAIC), J. Geophys. Res., 113, D13204, doi:10.1029/2007JD008782, 2008.

Zhang, J., Reid, J. S., Westphal, D., Baker, N., and Hyer, E.: A system for operational aerosol optical depth data assimilation over global oceans, J. Geophys. Res., 113, D10208, doi:10.1029/2007JD009065, 2008.

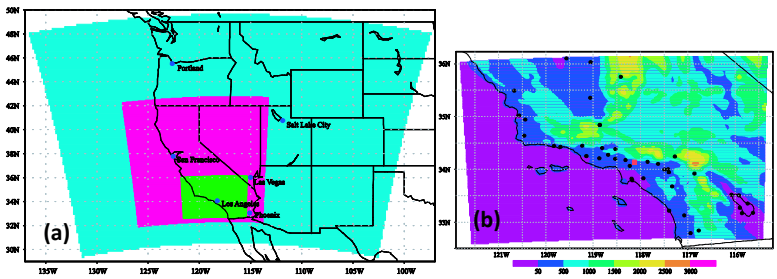

Fig. 1. Model domains, topography and location of observations. The three nested model domains are shown in the left panel. The right panel gives the topography in the small domain along with the locations of the 42 $PM_{2.5}$ monitoring stations (black dot) operated by the California Air Resources Board (ARB) and one super observing station (red square), which was located on the campus of the California Institute of Technology and collected speciated aerosol concentrations during the CalNex 2010 field experiment.

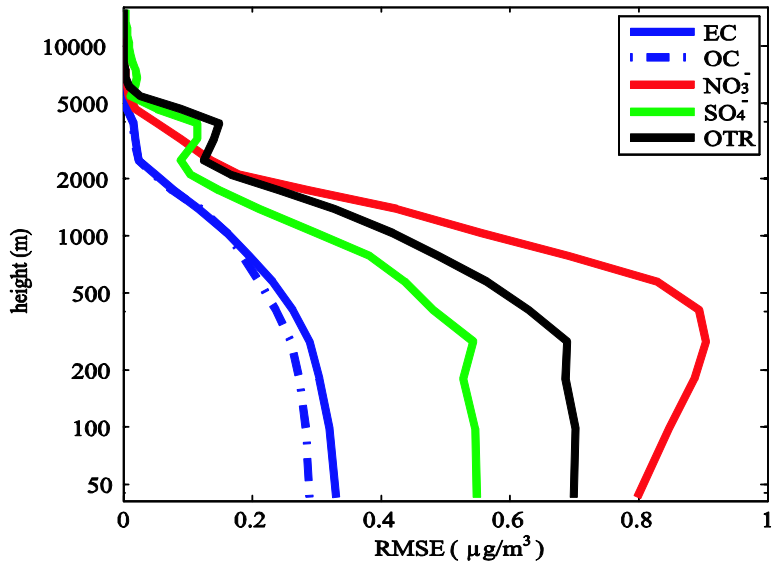

Fig. 2. Vertical distribution of the root-mean-square of the background errors in mass concentration for five species, estimated using the differences between 24 and 48 h forecasts valid at the same time.

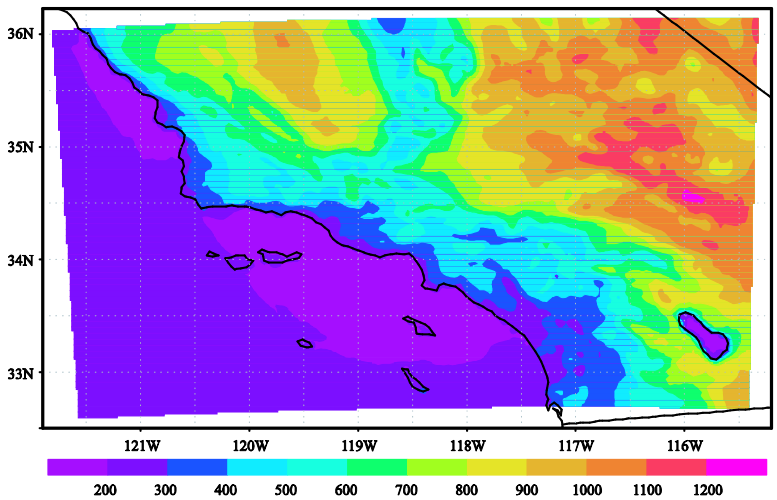

Fig. 3. Boundary layer heights (m) at 11:00 UTC, 18 October 2011 generated by WRF/Chem, initialized at 00:00 UTC, 16 October 2011 with the North America Regional Reanalysis (NARR).

Discussion Paper | Discussion Paper | Discussion Paper | Discussion Paper |

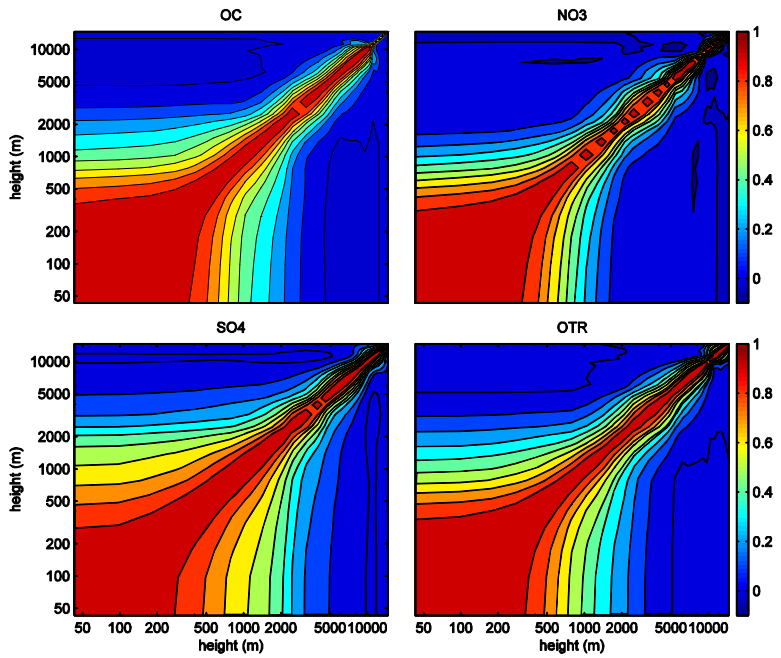

Fig. 4. Vertical correlations of the background errors for EC, NO_3^-, SO_4^{2-}, and OTR. These correlations are computed using differences between 24 and 48 h forecasts valid at the same time. A localization was applied with a vertical scale of 3500 m. The contour interval is 0.1.

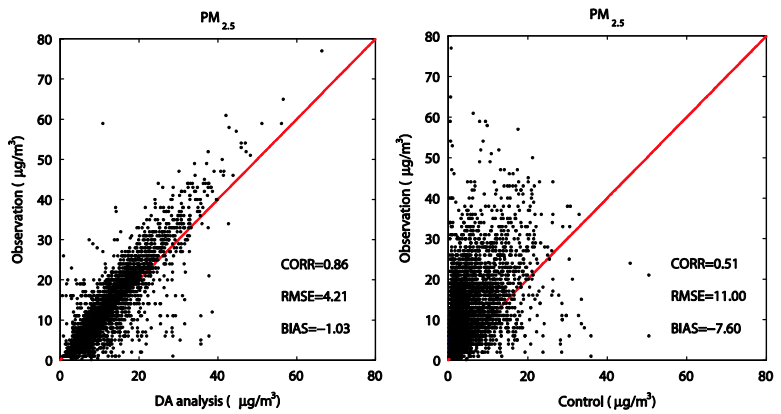

Fig. 5. Scatter plots of the $PM_{2.5}$ mass concentrations against observations in the analysis with **(a)** and without **(b)** data assimilation. The observations are assimilated, and consist of the 00:00, 06:00, 12:00 and 18:00 UTC observations from the 42 ARB monitoring stations during the period 12:00 UTC, 5 May to 12:00 UTC, 14 June 2010.

Discussion Paper | Discussion Paper | Discussion Paper | Discussion Paper |

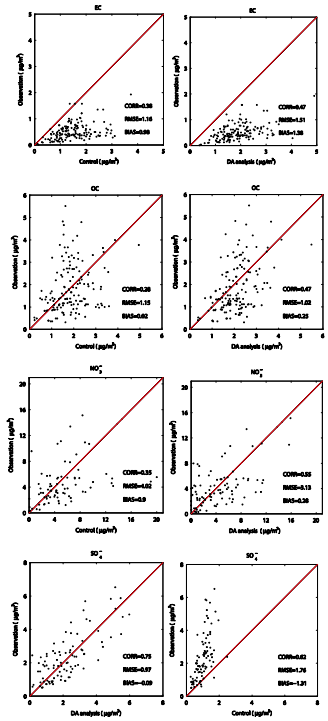

Fig. 6. Scatter plots of model mass concentrations against observations with (left column) and without (right column) data assimilation for the species of EC, OC, NO_3^-, and SO_4^{2-}, respectively. These observations are not assimilated and are thus independent data. These observations are obtained hourly from the super monitoring station during the CalNex 2010 field experiment.

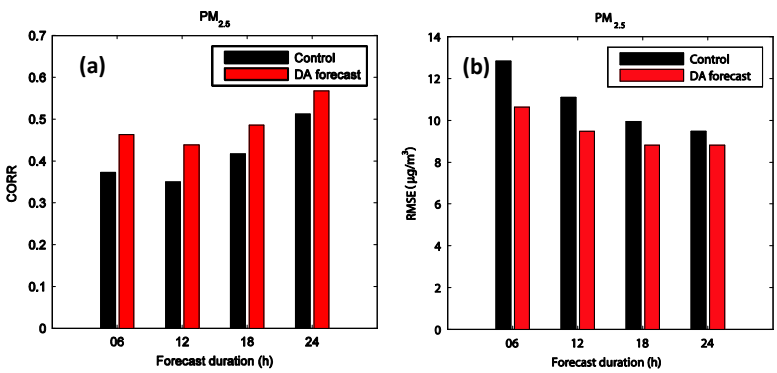

Fig. 7. Correlations **(a)** and root-mean square errors (RMSEs in $\mu g\,m^{-3}$) **(b)** of the total $PM_{2.5}$ concentration forecasts against observations as a function of forecast duration. The forecast is initialized at 12:00 UTC. Both correlations and RMSEs are calculated against the observations from 42 stations during the period 12:00 UTC, 5 May to 12:00 UTC, 14 June 2010. The black bars are for the control forecast without data assimilation, and the red bars for the forecast with data assimilation.

Discussion Paper | Discussion Paper | Discussion Paper | Discussion Paper |

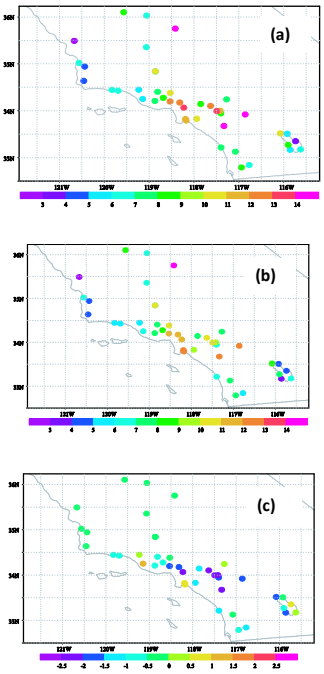

Fig. 8. Spatial distribution of RMSEs of the 24 h forecasts. **(a)** shows the RMSEs of the control forecast, and **(b)** the RMSEs of the 24 h forecast with data assimilation. The RMSE difference between the control and 24 h forecast with data assimilation is shown in the panel **(c)**, in which a negative value indicates a decrease of RMSE arising from data assimilation. The RMSEs are in $\mu g\,m^{-3}$.

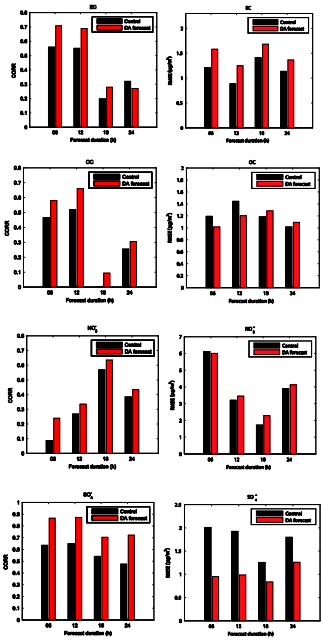

Fig. 9. Correlations (left) and root-mean square errors (RMSEs in $\mu g\,m^{-3}$) (right) of the specie concentration forecasts against observations as a function of forecast duration. The forecast is initialized at 12:00 UTC. Both correlations and RMSEs are calculated against the observations from 42 stations during the period 12:00 UTC, 5 May to 12:00 UTC, 14 June 2010. The black bars for the control forecast without data assimilation, and the red bars for the forecast with data assimilation.