Atmos. Chem. Phys., 16, 3631–3649, 2016
www.atmos-chem-phys.net/16/3631/2016/
doi:10.5194/acp-16-3631-2016


Atmospheric Chemistry and Physics Open Access EGU

# Variational data assimilation for the optimized ozone initial state and the short-time forecasting

**Soon-Young Park**[1], **Dong-Hyeok Kim**[1], **Soon-Hwan Lee**[2], **and Hwa Woon Lee**[3]

[1]Institute of Environmental Studies, Pusan National University, Busan, Republic of Korea
[2]Department of Earth Science Education, Pusan National University, Busan, Republic of Korea
[3]Division of Earth Environmental System, Pusan National University, Busan, Republic of Korea

*Correspondence to:* Hwa Woon Lee (hwlee@pusan.ac.kr)

Received: 3 July 2015 – Published in Atmos. Chem. Phys. Discuss.: 20 October 2015
Revised: 21 January 2016 – Accepted: 17 February 2016 – Published: 17 March 2016

**Abstract.** In this study, we apply the four-dimensional variational (4D-Var) data assimilation to optimize initial ozone state and to improve the predictability of air quality. The numerical modeling systems used for simulations of atmospheric condition and chemical formation are the Weather Research and Forecasting (WRF) model and the Community Multiscale Air Quality (CMAQ) model. The study area covers the capital region of South Korea, where the surface measurement sites are relatively evenly distributed.

The 4D-Var code previously developed for the CMAQ model is modified to consider background error in matrix form, and various numerical tests are conducted. The results are evaluated with an idealized covariance function for the appropriateness of the modified codes. The background error is then constructed using the NMC method with long-term modeling results, and the characteristics of the spatial correlation scale related to local circulation are analyzed. The background error is applied in the 4D-Var research, and a surface observational assimilation is conducted to optimize the initial concentration of ozone. The statistical results for the 12-hour assimilation periods and the 120 observatory sites show a 49.4 % decrease in the root mean squared error (RMSE), and a 59.9 % increase in the index of agreement (IOA). The temporal variation of spatial distribution of the analysis increments indicates that the optimized initial state of ozone concentration is transported to inland areas by the clockwise-rotating local circulation during the assimilation windows.

To investigate the predictability of ozone concentration after the assimilation window, a short-time forecasting is carried out. The ratios of the RMSE (root mean squared error) with assimilation versus that without assimilation are 8 and 13 % for the +24 and +12 h, respectively. Such a significant improvement in the forecast accuracy is obtained solely by using the optimized initial state. The potential improvement in ozone prediction for both the daytime and nighttime with application of data assimilation is also presented.

## 1 Introduction

Data assimilation provides a consistent representation of the physical state such as the atmosphere by blending imperfect model predictions and noisy observations. As a technique that applies observational information to numerical models with the aim of increasing model predictability, data assimilation is actively used in numerical weather prediction (NWP) and ocean modeling studies (Daley, 1991; Courtier et al., 1998; Rabier et al., 2000; Kalnay, 2003; Navon, 2009; Evensen, 2007). With more chemical observations available in recent years, including the satellite data, data assimilation is expected to make more contributions to weather forecasting and further improve the predictability of air quality. When the data assimilation technique is used in an air quality model, it not only improves the initial concentration distribution of pollutants, but also optimizes the emissions. In addition to the boundary inflow concentration (Carmichael et al., 2008), emission is also one crucial factor in the numerical prediction of various air pollutants. Several data assimilation techniques have been developed. The four-dimensional variational (4D-Var) data assimilation requires an adjoint model for use in non-linear numerical models. This represents an

applied area in the use of adjoint sensitivity (Elbern and Schmidt, 2001; Penenko et al., 2002; Sandu et al., 2005; Hakami et al., 2007).

Research using the adjoint model in air quality models started in the mid-1990s. The adjoint models used in and before the year 2000 are well described in the review paper of Wang et al. (2001). Sandu and Chai (2011) and Carmichael et al. (2008) presented subsequent research, and described many areas in which the adjoint method has been applied. More recently, more comprehensive reviews including coupled chemistry meteorology models were well-addressed by Bocquet et al. (2015).

Elbern et al. (1997) were the first to assimilate tropospheric air quality data into the European air pollution dispersion model. They argued that back then the existing air quality data assimilation was limited solely to stratospheric ozone data from satellite observations, which is far less than enough for better air quality prediction. In their study, they performed data assimilation using both data generated by the model and various information from observations. The results indicated that when using the model-generated data, the predictability is improved not only for the chemical species directly related with those used in the data assimilation, but also for those not used in the data assimilation. In their following research, Elbern and Schmidt (2001) applied 4D-Var to cases of high summer ozone concentrations based on ground observations over Europe, and ozone sonde observations from other locations. The results of 6 h data assimilation showed improved predictability. In addition, they also examined the sensitivities of model simulation to data assimilation based on the radius of the influenced area when data assimilation was performed.

Chai et al. (2007) analyzed the effects of observations from various observation systems, such as ground, civil aviation, ship, ozone sonde, and lidar, on data assimilation. The ICARTT (International Consortium for Atmospheric Research on Transport and Transformation) data were obtained and used in the above research. In particular, they proposed a method to calculate background errors, which had not been addressed in detail in the previous research, and verified its performance in the interested modeling area. Boisgontier et al. (2008) assimilated tropospheric ozone concentrations in their regional ozone prediction study prior to the launch of the MetOp Satellite of European Organisation for the Exploitation of Meteorological Satellites (EUMETSAT) Polar System (EPS) in October 2006. Although the study performed data assimilation using the column ozone data ranging over 0–6 km in the troposphere, they expected that it would positively affect the accuracy of regional ozone prediction. The chemical data assimilation has been conducted using $NO_2$ and HCHO from the satellite, SCanning Imaging Absorption spectroMeter for Atmospheric CHartographY (SCHIAMACHY), together with air quality observations at the ground level (Zhang et al., 2008). The initial fields with assimilated observations were improved compared with that generated without data assimilation.

Gou and Sandu (2011) indicated that there differences might exist in the gradient results between discrete and continuous adjoint in the process of developing an adjoint model due to the high non-linearity in the advection equation of the air quality model. As a result, they argued that the discrete method is more accurate in the adjoint sensitivity study, and that the continuous method is faster in minimizing the cost function in the 4D-Var data assimilation. In their study of the background pollutants affecting ground ozone concentrations in western America during the summer, Huang et al. (2013) applied data assimilation not only to numerical simulations, but also to evaluation of the concentrations associated with transport. Based on analysis of the ground-observed ozone concentration, they suggested that the simulated surface $O_3$ error decreased by an average of 5 ppb and the reduction can be up to a maximum of 17 ppb with application of data assimilation. The estimated background $O_3$ that was transported from the eastern Pacific Ocean is about 3 ppb higher due to the application of data assimilation.

Most of the previous studies for chemical data assimilation have focused on a phenomena of meteorologically synoptic scale using satellite-based observation as well as ground-based data. The transport of air pollution forced by a local circulation such as land–sea breeze is poorly examined.

One of the important elements affecting results of data assimilation in the 4D-Var process is the background errors of the model (Talagrand and Courtier, 1987). Previous research has treated the background errors as scalar quantities with a Gaussian distribution, whereas there is a lack of research applying them in a matrix form and considering the three-dimensional covariance (Constantinescu et al., 2007; Singh et al., 2011; Sliver et al., 2013).

In this study, the region centered in the capital area of South Korea, where the ground observation sites are densely distributed, is selected for the study of data assimilation. The previously developed 4D-Var code has been modified to treat background errors in matrix forms, and various numerical tests have been conducted. The results are evaluated using an idealized covariance function. The realistic background errors are then obtained for the region around the capital of South Korea using long-term modeling results. Characteristics of the background errors generated in this study are analyzed. Also, the predictability of high ozone concentration was investigated by setting the initial ozone concentration as control variables in the cost function for the 4D-Var data assimilation.

## 2 Methods

### 2.1 4D-Var data assimilation

The variational method solves the data assimilation problem from an optimal control framework (Penenko and Obraztsov, 1976; Courtier and Talagrand, 1987; Le Dimet and Talagrand, 1986). We aim to find control variables that minimize the difference between the model predictions and observations. In the frame of strongly constrained 4D-Var data assimilation, the observational data at all times within the assimilation window are simultaneously considered. The control variables become the initial concentration distribution $\boldsymbol{c}_0$, and all results at future times are uniquely determined from this in the model.

In the maximum likelihood approach, the 4D-Var data assimilation gives the maximum a posteriori estimator of the true initial concentration distribution, which is obtained by minimizing the cost function:

$$\mathcal{J}(\boldsymbol{c}_0) = \frac{1}{2}\left(\boldsymbol{c}_0 - \boldsymbol{c}_0^b\right)^{\mathrm{T}} \mathbf{B}_0^{-1}\left(\boldsymbol{c}_0 - \boldsymbol{c}_0^b\right) + \frac{1}{2}\sum_{k=1}^{F}\left(\mathcal{H}(\boldsymbol{c}_k) - \boldsymbol{c}_k^{\mathrm{obs}}\right)^{\mathrm{T}} \mathbf{R}_k^{-1}\left(\mathcal{H}(\boldsymbol{c}_k) - \boldsymbol{c}_k^{\mathrm{obs}}\right). \quad (1)$$

Before data assimilation is performed, the current state that best estimates the true state is called a priori or background state $\boldsymbol{c}_0^b$. The random background errors are assumed to be unbiased and to have a normal distribution. $\mathbf{B}_0$ refers to the background error covariance (BEC). The observed value at time $k$ is $\boldsymbol{c}_k^{\mathrm{obs}}$. In general, the observational data are not accurately represented at the model grids. Additionally, in some cases, the observation instruments do not measure the meteorological variables directly (e.g., weather radar and satellite). Therefore, an observation operator $\mathcal{H}$ that converts a model space to an observation space is required. The observation error includes both measurement (instrument) error and representativeness error. The representativeness error occurs because of the error included in the observation operator itself and because the input data of $\mathcal{H}$ are not exactly the true state. Similar to the background error, the observation error is assumed to be unbiased and have a normal distribution. It is independent of other observation times, and usually is assumed to be spatially uncorrelated. Under this assumption, observation error covariance $\mathbf{R}_k$ becomes a diagonal matrix. In addition, the observation error and background error are assumed to be independent of each other. The interpretation of this equation is that the deviation of initial concentration $\boldsymbol{c}_0$ from the background field $\boldsymbol{c}_0^b$ is weighted by the inverse matrix of the background error covariance, whereas the differences between the model predictions $\mathcal{H}(\boldsymbol{c}_k)$ and observations $\boldsymbol{c}_k^{\mathrm{obs}}$ during assimilation windows are weighted by the inverse of error observation covariance matrix.

The 4D-Var analysis can be obtained by the initial concentration that minimizes Eq. (1) with respect to the model equation.

$$\boldsymbol{c}_0^a = \arg\min \mathcal{J}(\boldsymbol{c}_0) \ \text{ subject to } \ \boldsymbol{c}_t = \mathcal{M}_{t_0 \to t}(\boldsymbol{c}_0) \quad t = 1, \cdots, F \quad (2)$$

Here $\mathcal{M}$ represents the model solution operator and includes an atmospheric forcing, the emission rates, the chemical kinetics, and all the other parameters. Furthermore, the model provides analysis within the assimilation window using the optimal initial conditions: $\boldsymbol{c}_t^a = \mathcal{M}_{t_0 \to t}\left(\boldsymbol{c}_0^a\right)$. Formally, a gradient-based optimization procedure is used to obtain minimum value. Assuming a linear observation operator $\mathbf{H}_k = \mathcal{H}'(\boldsymbol{c}_t)$, the gradient of Eq. (2) with respect to $\boldsymbol{c}_0$ is

$$\nabla_{\boldsymbol{c}_0}\mathcal{J}(\boldsymbol{c}_0) = \mathbf{B}_0^{-1}\left(\boldsymbol{c}_0 - \boldsymbol{c}_0^b\right) + \sum_{k=1}^{F}\left(\frac{\partial \boldsymbol{c}_k}{\partial \boldsymbol{c}_0}\right)^{\mathrm{T}} \mathbf{H}_k^{\mathrm{T}} \mathbf{R}_k^{-1}(\mathbf{H}_k \boldsymbol{c}_k - \boldsymbol{c}_k^{\mathrm{obs}}). \quad (3)$$

In the gradient of 4D-Var cost function, $(\partial \boldsymbol{c}_k / \partial \boldsymbol{c}_0)^{\mathrm{T}}$ is a transposed derivative of future states with respect to the initial concentration. At this point, the adjoint model is used and through the solution of adjoint equation at $t_0$, the gradient of the cost function at the initial concentration is provided. The gradient for the 4D-Var's cost function can be effectively obtained by forcing the adjoint model with observation increments and calculating it backwards. When the forward and reverse adjoint models are performed, i.e., $\sum$ in the Eq. (3) is finished, it results in the problem of solving the following equation:

$$\nabla_{\boldsymbol{c}_0}\mathcal{J}(\boldsymbol{c}_0) = \mathbf{B}_0^{-1}\left(\boldsymbol{c}_0 - \boldsymbol{c}_0^b\right) + \boldsymbol{\lambda}_0 = 0. \quad (4)$$

$\boldsymbol{\lambda}_0$ is the sensitivity of the cost function (Eq. 1) defined for 4D-Var with respect to the initial concentration $\boldsymbol{c}_0$. Since $\mathbf{B}_0^{-1}$, $\boldsymbol{c}_0^b$, and $\boldsymbol{\lambda}_0$ values are known matrices and vectors, if the value of $\boldsymbol{c}_0$ that satisfies Eq. (4) is found, it becomes the analysis field $\boldsymbol{c}_0^a$. Solving the above equation is similar to solving a linear-algebraic problem such as $\mathbf{A}\boldsymbol{x} = \boldsymbol{b}$, and the solution can be obtained by various minimization algorithms (e.g., steepest descent, conjugate gradient and quasi-Newton methods).

### 2.2 Background error covariance

Accurate error covariances for background and observation are important for the quality of data assimilation. A reasonable analysis may deteriorate because of misunderstanding of these covariances (Daescu, 2008). The background error covariance (BEC) is of utmost importance, as it weights the model error against the competing observation error, spreads information from observations to the adjacent area, and influences several parameters such as temperature and wind fields or chemical constituents (Elbern and Schmidt, 2001).

The adjoint code for CMAQ (CMAQ-ADJ) model was implemented from the project H98 (University of Houston,

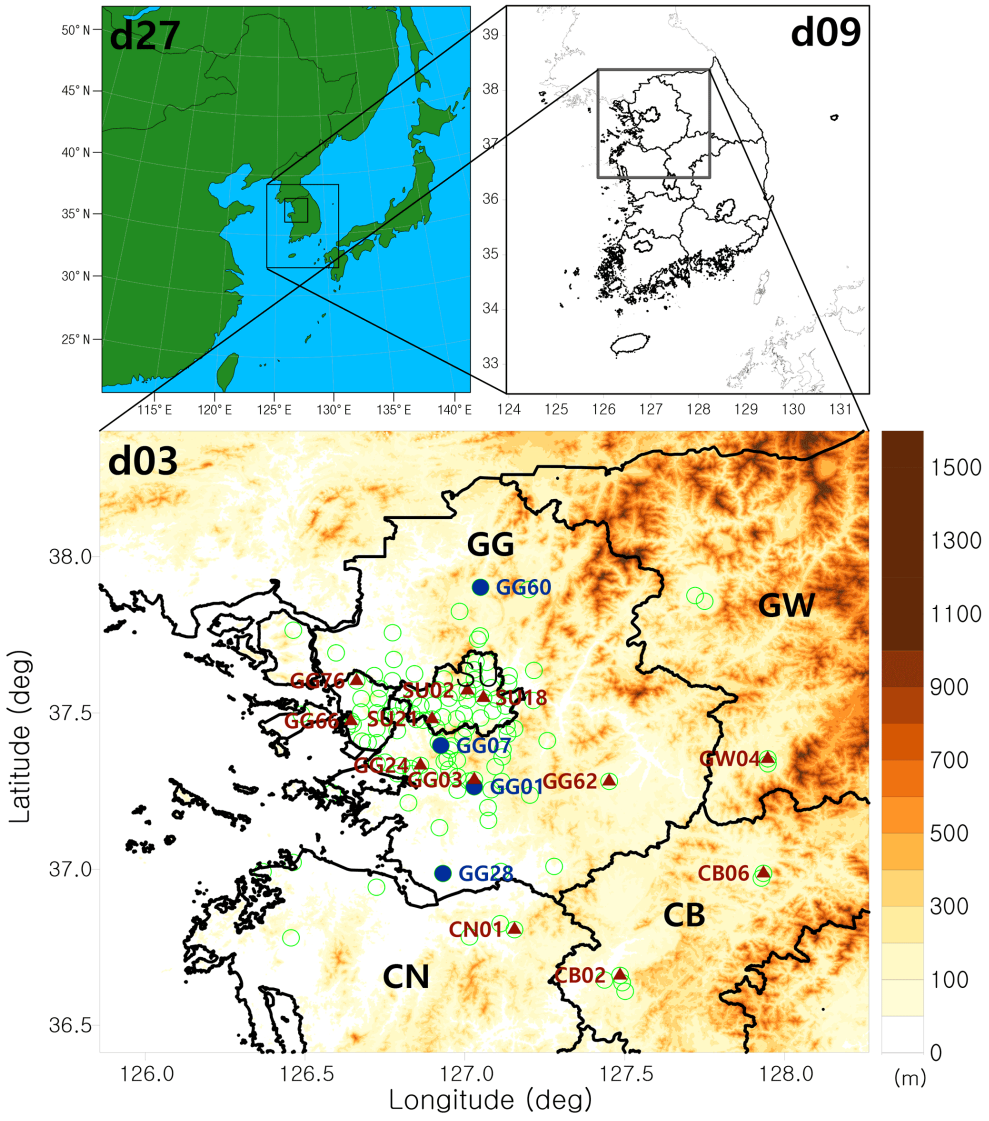


**Figure 1.** The model domains (d27, d09, and d03) for WRF (Weather Research and Forecasting). The domain size of CMAQ (Community Multiscale Air Quality) is mostly the same except that it has five grids fewer than WRF at lateral boundaries. The air quality monitoring sites at ground level are marked by green blank circles. Blue filled circles and red filled triangles indicate the selected locations for the idealized and realized background error covariance experiments, respectively. These experiments are conducted to investigate the diurnal variation of ozone during the assimilation window. Administrative district in the areas of Seoul, Gyeonggi-do, Gangwon-do, Chungcheongnam-do, and Chungcheongbuk-do is abbreviated to SU, GG, GW, CN, and CB, respectively, and also represented on the map.

2009) by Houston Advanced Research Center/Texas Environmental Research Consortium (HARC/TERC). The validation and several numerical tests of this code are well described in Hakami et al. (2007). Below is the defined cost function in CMAQ-ADJ to optimize initial condition, which refers to concentration at the initial time.

$$\mathcal{J}(\boldsymbol{c}_0) = \frac{1}{2\left(\sigma_0^b\right)^2}\left(\boldsymbol{c}_0 - \boldsymbol{c}_0^b\right)^{\mathrm{T}}\left(\boldsymbol{c}_0 - \boldsymbol{c}_0^b\right) + \frac{1}{2\left(\sigma_k^{\mathrm{obs}}\right)^2}\sum_{k=1}^{N}\left(\mathbf{H}_k\boldsymbol{c}_k - \boldsymbol{c}_k^{\mathrm{obs}}\right)^{T}\left(\mathbf{H}_k\boldsymbol{c}_k - \boldsymbol{c}_k^{\mathrm{obs}}\right) \quad (5)$$

This form only considers the model and observation errors as its variance, i.e., a constant value of $\left(\sigma_0^B\right)^2$ and $\left(\sigma_k^{\mathrm{obs}}\right)^2$ with Gaussian distribution.

**Table 1.** Configuration of WRF modeling system.

| WRF | d27 | d09 | d03 |
|---|---|---|---|
| Horizontal grid | 123 × 130 | 72 × 84 | 65 × 68 |
| Horizontal resolution | 27 km | 9 km | 3 km |
| Vertical layers | 33 layers (top: 50 hPa) | | |
| Physical options | WSM5 scheme | | |
| | Kain–Fritsch scheme | | |
| | Noah LSM | | |
| | Yonsei University PBL | | |
| | RRTM Longwave | | |
| | Dudhia Shortwave | | |
| Initial data | NCEP FNL data | | |
| Time period | 00:00 UTC 3 August–00:00 UTC 7 August 2008 | | |

**Table 2.** Configuration of CMAQ 4D-Var modeling system.

| CMAQ | | | d27 | d09 | d03 |
|---|---|---|---|---|---|
| Meteorological input | | | correspond to each WRF domain | | |
| Horizontal grid | | | 118 × 125 | 67 × 79 | 60 × 63 |
| Horizontal resolution | | | 27 km | 9 km | 3 km |
| Vertical layers | | | 15 layers (top: 20 km) | | |
| Other options | | | CB IV Chemical mechanism | | |
| | | | PPM advection | | |
| | | | Multiscale horizontal diffusion | | |
| | | | Eddy vertical diffusion | | |
| | | | RADM Cloud scheme | | |
| Emission data | | | INTEX-B | CAPSS | CAPSS |
| Time periods | Forward | | 00:00 UTC 3 August–00:00 UTC 7 August 2008 (4 days) | | |
| | 4D-Var | day time | 00:00 UTC 5 August–12:00 UTC 5 August 2008 (12 h, analysis) | | |
| | | | 12:00 UTC 5 August–12:00 UTC 6 August 2008 (24 h, forecast) | | |
| | | nighttime | 12:00 UTC 5 August–00:00 UTC 6 August 2008 (12 h, analysis) | | |
| | | | 00:00 UTC 6 August–00:00 UTC 7 August 2008 (24 h, forecast) | | |

If a BEC is to be correctly adopted, a cost function should be defined in the form of a matrix; this is denoted by the first term on the right-hand side in Eq. (1). The background part and its gradient of the cost function, written in Fortran codes, have been revised in this study to make the matrix operation possible. A numerical test is conducted to validate the suitability and effects of the revised codes.

The methods for obtaining the BEC of a numerical model are mainly divided into two types: an NMC method (Parrish and Derber, 1992) that defines the model error as the difference between the forecasting results at different initial times, and an ensemble method that uses a perturbed forecast. Recently, Kucukkaraca and Fisher (2006) introduced a technique for modeling a flow-dependent BEC. In Constantinescu et al. (2007), an autoregressive model was proposed for flow-dependent BEC in air quality data assimilation.

In this study, the BEC of the model is constructed by using the NMC method, which is the most intuitive and easily applied method.

## 3 Experimental design

If the observatory sites are distributed unevenly, results of data assimilation based on the variational theory will have low reliability, and it is difficult to minimize the cost function (Courtier and Talagrand, 1987). For this reason, the capital region of South Korea is selected for the present data assimilation study because measurement sites are relatively evenly distributed in this area. Figure 1 depicts the study area (d03), i.e., the capital region of South Korea along with the domain configuration for the other two nesting domains of coarse resolution. A total of 120 observatory sites are evenly distributed in the areas of Seoul (SU), Gyeonggi-do (GG), Gangwon-do (GW), Chungcheongnam-do (CN), and Chungcheongbuk-do (CB). The innermost domain, d03, is located in a geographical area with coasts to the west and the topography gradually rises towards the east. The Weather Research and Forecasting (WRF) model (Skamarock et al., 2008) is a mesoscale atmospheric model that has been widely used to simulate a local circulation pat-

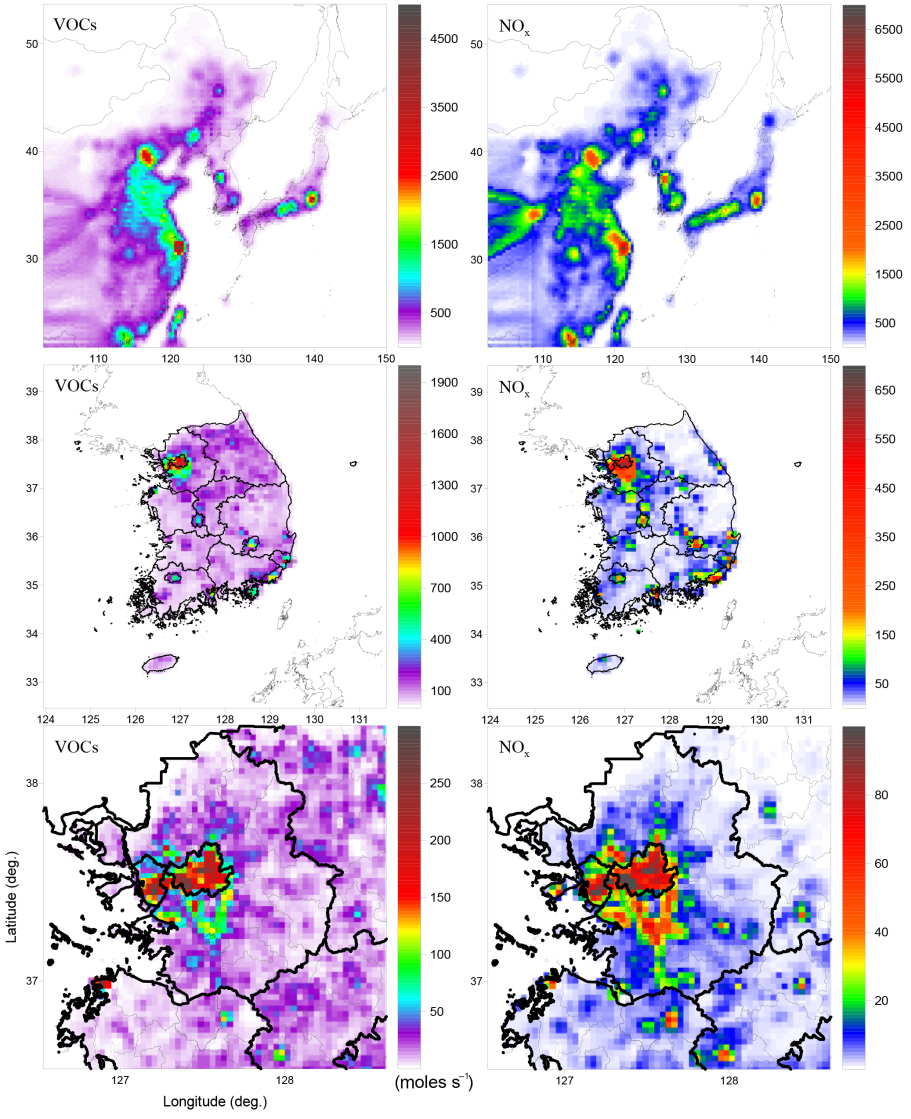


**Figure 2.** Horizontal distributions of emission rate for domain d27 (top), d09 (middle), and d03 (bottom). The left and right panels are for VOCs and $NO_x$ emission rates, respectively.

tern and provide the meteorological input data for the air quality model. The chemical formation and transportation of ozone is simulated by the Model-3 Community Multiscale Air Quality (CMAQ) model (Byun and Ching, 1999). This model simulates gas-phase chemistry using the carbon bond IV (CB-IV) photochemical mechanisms (Grey et al., 1989). To describe the chemical transformation, euler backward Iterative (EBI) (Hertel et al., 1993) solver is implemented. The advection is calculated by the piecewise-parabolic method (PPM) (Colella and Woodward, 1984), which is based on the finite volume subgrid definition of the advected scalar. The vertical diffusion in the planetary boundary layer is calculated following the approach in the Regional Acid Deposition Model, RADM (Chang et al., 1987), which is based on the similarity theory. Detailed settings used for the atmospheric and air quality model systems in the present study are presented in Tables 1 and 2, respectively. All time mentioned in this paper except those in Table 2 are local standard time (LST), which is 9 h earlier than the Coordinated Universal Time (UTC).

The experiment without assimilation was conducted as a forward run (FWD), which covers 4 days from 09:00 LST on 3 August to 09:00 LST on 7 August. In addition, data assimilation (4DV) was performed within the 12 h time window from 09:00 to 21:00 LST on 5 August. Figure 2 illustrates the spatial distribution of the total $NO_x$ and VOCs (volatile organic compounds) pollutants, which are out of the 24 emitted substances used in the CMAQ model. The domain d27 is located in the East Asian monsoon region, which includes most of China and Japan. The Intercontinental Chem-

ical Transport Experiment-Phase B (INTEX-B, Zhang et al., 2009) 2006 data were used as emissions; high emissions are mostly found over major cities of each country. The emissions applied to domains d09 and d03 are extracted from the CAPSS 2007 data (Lee et al., 2011).

The results of the WRF simulation for the synoptic pattern of surface pressure during the study period are presented in Fig. 3, along with the weather charts. The vector indicates surface wind, and the values of the contours are the concentrations of $O_3$. The model has successfully simulated the North Pacific high-pressure system, and adequately describes the local high-pressure system that developed in and around the East Sea on 4 August, as well as the high-pressure system that developed in and around the southwestern coastal region on 5 August. A clockwise synoptic flow developed because of the well-developed North Pacific high-pressure system. As a result, the long-distance transport from the pollution sources in China had little impact on the simulated pollutants.

Figure 4 shows the horizontal distributions of simulated ozone concentration and surface wind from 06:00 to 21:00 LST on 5 August at 3-hour intervals. At 06:00 LST, a southeasterly to easterly wind developed along the western coast, and the overall ozone concentration was low in this region. Accompanied with the increase in solar radiation after sunrise, the ozone concentration began to increase, and an onshore sea-breeze developed after 12:00 LST in the western coast. This sea breeze lasted from 18:00 to 21:00 LST. After sunset, the influence of the sea-breeze can be identified over areas where the ozone concentration decreased due to $NO_x$-titration. Afterwards, the dominant wind direction changed in a clockwise direction (figure omitted), and the local circulation did not extend far enough beyond the GG region.

## 4 Results

### 4.1 Effects of an idealized BEC

Two simple yet popular covariance models are Gaussian and Balgovind (Balgovind et al., 1983) functions expressed as

$$\omega(r) = \mathrm{EXP}\left(-\frac{r^2}{2L^2}\right), \text{ Gaussian} \quad (6)$$

$$\omega(r) = \left(1+\frac{r}{L}\right)\mathrm{EXP}\left(-\frac{r}{L}\right), \text{ Balgovind.} \quad (7)$$

To examine the appropriation of modified code, the Balgovind distribution expressed in Eq. (7) is selected for constructing the BEC that has the components of matrix form. Figure 5 shows the distribution patterns for Gaussian and Balgovind with respect to the distance between two grid points ($r$) and the characteristic length or radius of influence ($L$).

Table 3 summarizes a suite of numerical tests with and without data assimilation. In the tests with application of data assimilation, a matrix is constructed assuming that the BEC of the model has the form of a Balgovind function. The model domain is the innermost domain as illustrated in Fig. 1. The FWD test is conducted without data assimilation, and the other test is performed with data assimilation. The two types of tests are named as EXP_A and EXP_B, respectively.

EXP_A is a test that can be used to evaluate the characteristics of the BEC based on a single observation experiment. In this experiment, 100 ppb of $O_3$ was incorporated as an arbitrary value rather than actual observation data at the initial time at the center of the model domain. To emphatically show the background part of the cost function, the value 8.00, which is much larger than the basic value (0.08), is applied to $\sigma_k^{\mathrm{obs}}$ in Eq. (5). Using the function that sets the radius of influence to be 2, 5, and 10, the data assimilation characteristics for three BECs were examined.

In EXP_B, which is the second test, the effect of BEC used in 4D-Var is examined. Real observation data are used in EXP_B. The observation data include 12 h ozone concentration at 120 sites within the capital city regions. Two cases are investigated in the EXP_B (Table 3): the XBE case only considers variance that is not in a matrix form, and the OBE case uses the BEC in the matrix form that adopts the Balgovind function. In the XBE, two tests that take into consideration the different weighting between $\sigma_0^B$ and $\sigma_k^{\mathrm{obs}}$ are conducted separately. In XBE_r0.08, the observation data are assumed to be accurate and $\sigma_k^{\mathrm{obs}}$ is set to 0.08, which is the basic value for this model. For XBE_r8.00, $\sigma_k^{\mathrm{obs}}$ is set to 8.00, indicating that the results of the model are more important than the observation. For OBE_r8.00, $\sigma_k^{\mathrm{obs}}$ and $L$ are set to 8.00 and 5, respectively. The result of OBE_r0.08 is not analyzed because it is similar to the result of XBE_r0.08.

Among the results of the EXP_A, horizontal distributions of the analysis increment with respect to the radius of influence ($L$) are illustrated in Fig. 6. At the model grid point (29, 31), where arbitrary observation data were applied, all three tests showed an $O_3$ increment of about 50.0 ppb. The background concentration of $O_3$ at the grid was 40.1 ppb, but the value was up to about 90 ppb in the analysis when the synthetic observation of 100 ppb was applied. However, as the value of $L$ increased, the $O_3$ increment in the analysis occurs at more surrounding grids. Particularly, the analysis increments shown along the east–west cross section (Fig. 7) are distinguished on the 2-D graph according to the $L$ values. This result is attributed to the ideal function that is used, in which the error covariance information is expanded to the surrounding regions according to the $L$ values. These results indicate that the idealized BEC performs well in the revised codes, and proper analysis increments can be achieved when the spatial correlation is taken into account.

Figure 8 shows the daily changes in ozone concentration simulated by each experiment in the test EXP_B and from observations at selected sites. Exact locations of these

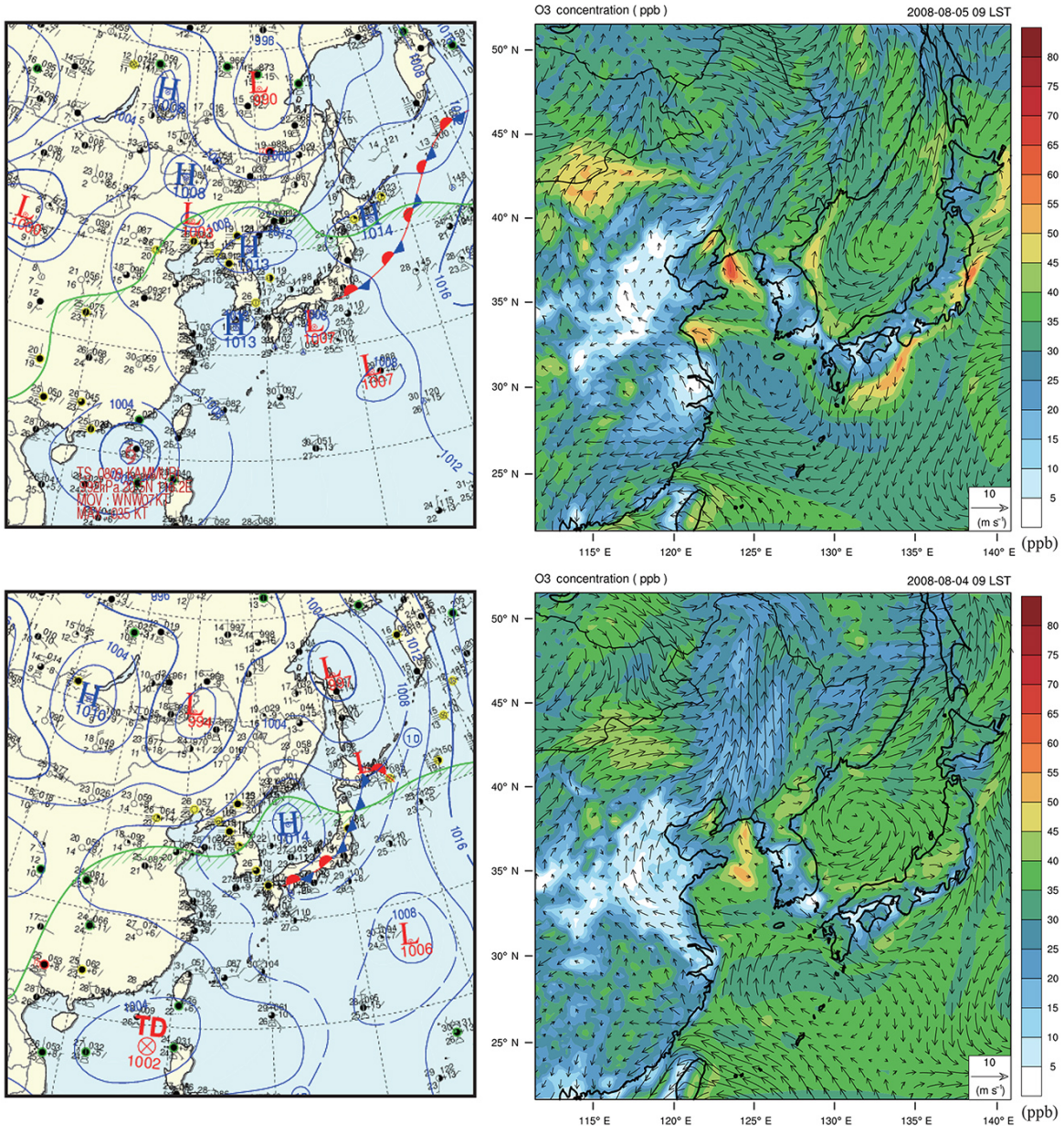


**Figure 3.** Synoptic weather charts (left) and simulated results (right) on 4 August (upper) and 5 August (lower). Filled contours and vectors represent ozone concentration and winds, respectively.

sites are marked in Fig. 1. At the site GG01, the observed (black solid line) concentration of ozone, which is higher than 100 ppb, was not simulated in the FWD (blue solid line). In XBE_r0.08 (green solid line), although the BEC is not applied, the simulated $O_3$ concentration is close to the observation in almost all the time slots. Comparing results of the two experiments that applied 8.00 for $\sigma_k^{\text{obs}}$, the effect of BEC can be determined. In the case of XBE_r8.00 (red dotted line), the simulated changes in $O_3$ concentration are similar to that simulated by the FWD because the weighted value in the FWD is high. When the BEC is taken into consideration for the same $\sigma_k^{\text{obs}}$(OBE_r8.00), the result is similar to that of XBE_r0.08. This result demonstrates the effect of the spreading analysis increment to its surrounding region where the observation sites are densely distributed. Although the weight of the observations is not set very high, improvements in the field analysis by spatial correlation are still achieved. At the GG07 site, this trend is quite significant with the OBE_r8.00 test, giving a result similar to that of XBE_r0.08. At the GG60 position, the model results are significantly improved, but the nighttime ozone is still overestimated. However, at the GG28 site, which is located at a

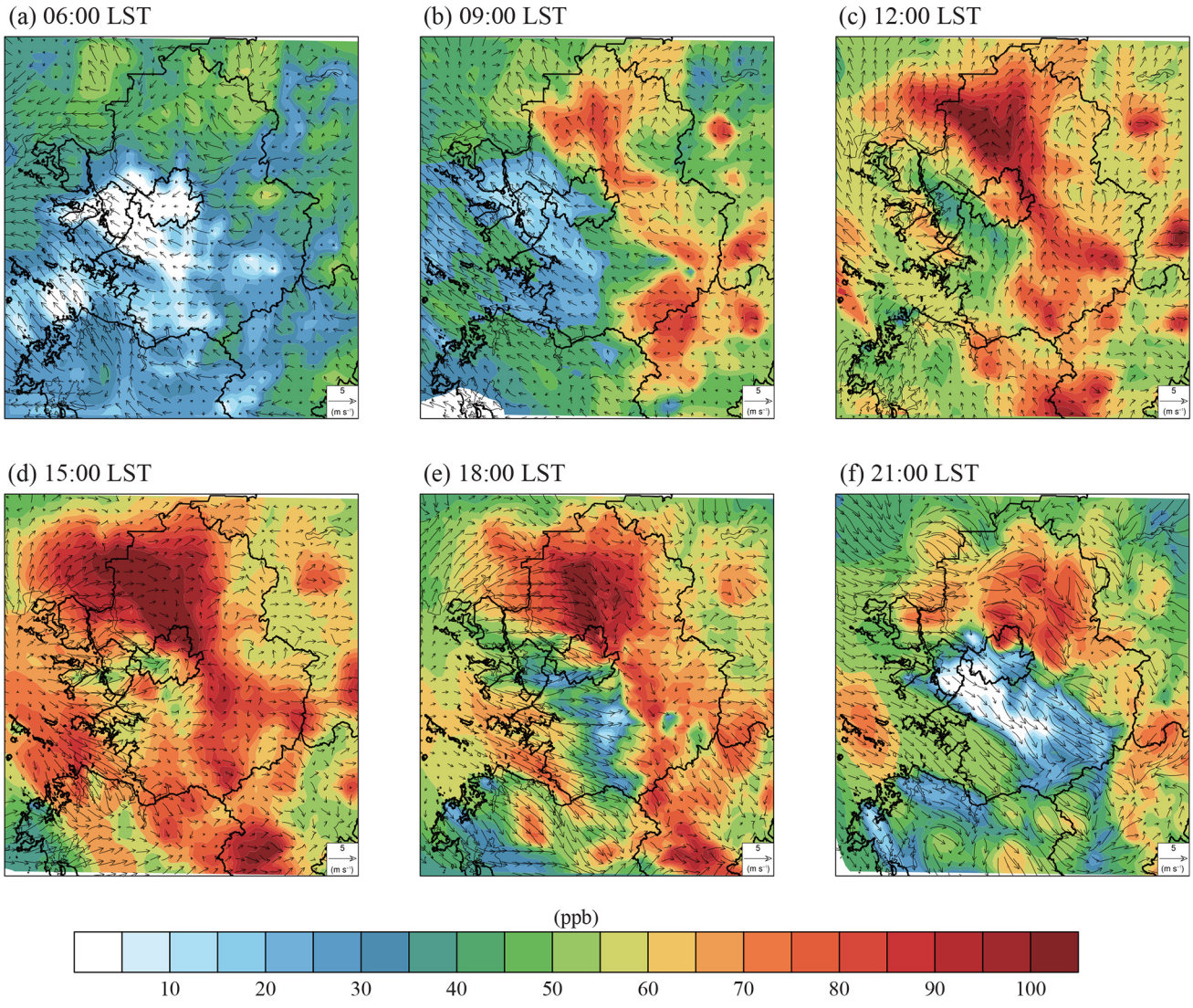


**Figure 4.** Diurnal variations of horizontal distribution of ozone (contour) and wind (vector) at 3-hour interval starting from 06:00 LST on 5 August.

**Table 3.** Experimental design for the idealized background error covariance test. The FWD case is conducted and the results are compared with that of the 4D-Var run.

| Assimilation | Case | | Observation data | Radius of Influence | $\sigma_0^B$ | $\sigma_k^{\text{obs}}$ |
|---|---|---|---|---|---|---|
| Forward run | FWD | | n/a | n/a | n/a | n/a |
| 4D-Var | EXP_A | L02 | 100 ppb at | $L = 02$ | BEC | 8.00 |
| run | (single | L05 | (29, 31) | $L = 05$ | BEC | 8.00 |
| | obs.) | L10 | | $L = 10$ | BEC | 8.00 |
| | EXP_B | XBE_r0.08 | 12 h $O_3$ | n/a | 1.00 | 0.08 |
| | | XBE_r8.00 | at all 120 | n/a | 1.00 | 8.00 |
| | | OBE_r8.00 | sites | $L$ =05 | BEC | 8.00 |

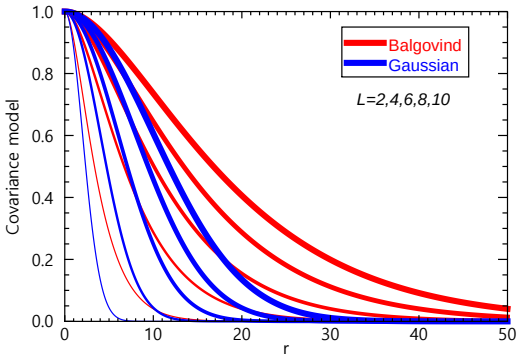


**Figure 5.** Covariance distribution for Gaussian (blue) and Balgovind (red) functions with respect to the distance ($r$) and the values of radius of influence ($L$).

region where observation sites are sparsely distributed, the BEC effect is barely observed. The results of XBE_r8.00 are similar to those of OBE_r8.00, except after 18:00 LST. This indicates that the effect of BEC, which considers the spatial correlation, can be distinct mainly over regions where the observation sites are densely distributed.

### 4.2 Development of realistic BEC

The BEC is obtained using the NMC (National Meteorological Center, now National Centers for Environmental Prediction) approach (Parrish and Derber, 1992), which is based on a real simulation for the realistic 4D-Var data assimilation study.

Figure 9 describes the method to define the model error. The error statistics for the CMAQ model is defined by the differences between +48 and +24 h forecast:

$$\boldsymbol{\epsilon}^i = \boldsymbol{c}^i_{+48\,\mathrm{h}} - \boldsymbol{c}^i_{+24\,\mathrm{h}}. \tag{8}$$

The BEC matrix has 2 800 526 400 components for a 3-D model with a number of grids $Nx \times Ny \times Nz = 60 \times 63 \times 14 = 52\,920$. To avoid storing the error covariance matrix explicitly, we assume $\mathbf{B}$ can be written as

$$\mathbf{B} = \mathbf{X} \otimes \mathbf{Y} \otimes \mathbf{Z} \otimes \mathbf{C}, \quad \text{(Chai et al., 2007)} \tag{9}$$

where $\mathbf{X} = [Nx \times Nx]$, $\mathbf{Y} = [Ny \times Ny]$, and $\mathbf{Z} = [Nz \times Nz]$ representing the error correlation in the three directions. $\mathbf{C}$ is the error covariance matrix at a single grid point that refers to the error variances and correlation between different species. In this study, $\mathbf{C}$ is considered to be constant, which means there is no correlation between the species.

It seems to be error-prone to invert ill-conditioned matrices. Based on the property of Kronecker product, $\mathbf{B}^{-1}$ can be expressed as

$$\mathbf{B}^{-1} = (\mathbf{X} \otimes \mathbf{Y} \otimes \mathbf{Z})^{-1} = \mathbf{X}^{-1} \otimes \mathbf{Y}^{-1} \otimes \mathbf{Z}^{-1}. \tag{10}$$

Singular value decomposition (SVD) is applied to $\mathbf{B}$ matrix. For example, a general $m \times n$ matrix $\mathbf{A}$ can be written as

$$\mathbf{A} = \mathbf{U}\,\mathbf{\Sigma}\,\mathbf{V}^{\mathrm{T}}. \tag{11}$$

For the symmetric matrices, such as $\mathbf{X}$, $\mathbf{Y}$, $\mathbf{Z}$

$$\mathbf{A} = \mathbf{U}\,\mathbf{\Sigma}\,\mathbf{U}^{\mathrm{T}}. \tag{12}$$

Then the inverse of $\mathbf{A}$ is easily calculated:

$$\mathbf{A}^{-1} = \mathbf{U}\,\mathbf{\Sigma}^{-1}\mathbf{U}^{\mathrm{T}}. \tag{13}$$

The accuracy of inverted BEC through these process has been confirmed by an algebraic calculation such as $\mathbf{B}^{-1}\mathbf{B} = \mathbf{I}$ and by comparing the vector $\boldsymbol{x}$ between $\mathbf{B}\boldsymbol{x} = \boldsymbol{y}$ and $\boldsymbol{x} = \mathbf{B}^{-1}\boldsymbol{y}$.

The error correlations between the vertical layers of the model are given in Fig. 10. Moving further away from a pertinent layer, the error correlation decreases. Judging from the diagonalized structure of errors, the correlation was found to be roughly a function of the physical distance between the layers. Examining the vertical error correlations for the magnitude of the boundary layer, high correlations can be found up to the fourth layer for the correlations in the vicinity of ground surface. This result indicates that an improvement in the model simulation can be achieved in the neighboring layers by performing DA using the observation data of upper layers that are located from the surface to the boundary layer.

In Fig. 11, the error correlations are plotted as a function of distance between two layers. When the distribution of correlations versus distance is fitted to a simple function, $e^{-\frac{\Delta z^{1.2}}{l_z^{1.2}}}$, the vertical length scale is $l_z = 300$ m. Although some high values deviate from this function, generally low correlation coefficients agree well with this function. The correlation coefficients versus the horizontal distance are illustrated in Fig. 12. On average, for both the north-south and east-west directions, $l_h$ is identified to be 10 km, and a function $e^{-\frac{\Delta z^{1.0}}{l_z^{1.0}}}$ fits well with the results. Particularly, the correlation coefficient for the east-west direction is somewhat higher than that for the south-north direction. This is partly attributed to the effect of middle latitude synoptic westerly and partly due to the land-sea breeze that occurs frequently in August in the capital city region, which produces circulation in the east-west direction.

### 4.3 Validation time results

4D-Var experiments are performed in this study, using actual observations with the distribution of the initial concentration of $O_3$ as the control variable. The observed hourly $O_3$ concentrations at 120 sites located within the domain d03 are used. In formula (1), $\boldsymbol{c}_0$ of ozone is considered as the control variable, and the BEC established in 4.2 is applied as the model error ($\mathbf{B}_0^{-1}$). The representativeness error is not considered, because the observatory sites are manually placed

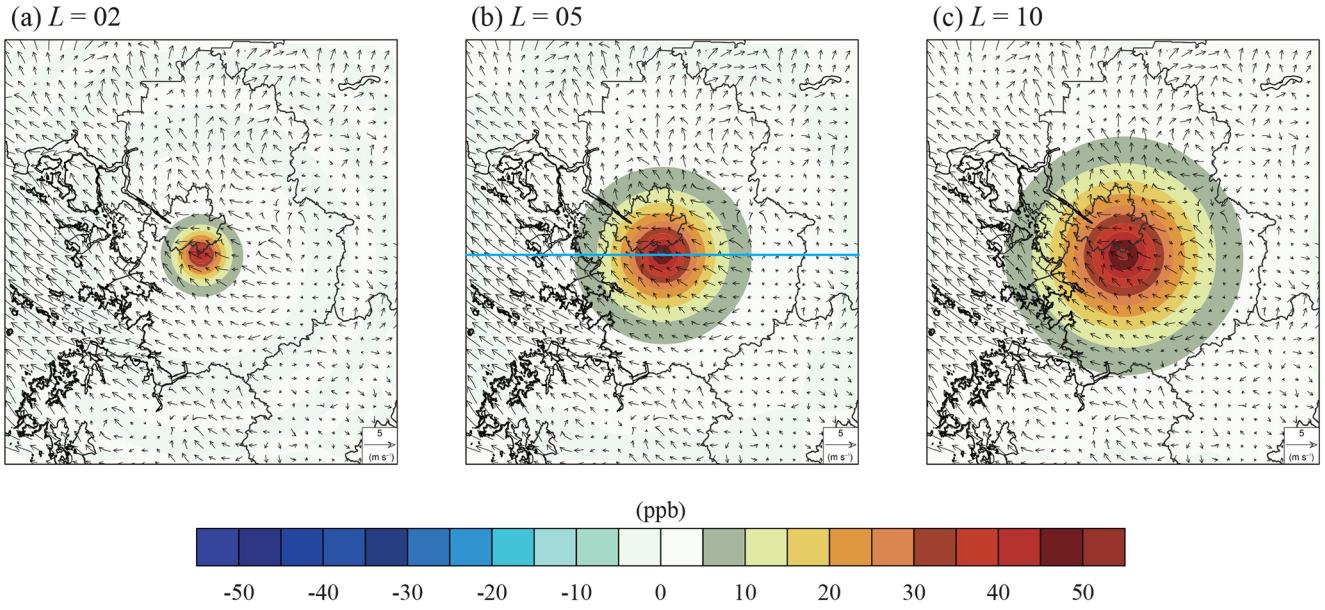


**Figure 6.** Horizontal distribution of analysis increments at surface resulted from the single observation experiment (EXP_A) with respect to radius of influence ($L$). Blue line in panel **(b)** stands for the location where the cross-sectional values of analysis increments are examined.

**Table 4.** Statistics of the model results.

| Description | Variable | Statistic definition* |
|---|---|---|
| Mean obs. | $\overline{O}$ | $(1/N)\sum_{i=1}^{N} O_i$ |
| Mean model | $\overline{M}$ | $(1/N)\sum_{i=1}^{N} M_i$ |
| Mean bias | MB | $(1/N)\sum_{i=1}^{N}(M_i - O_i)$ |
| Normalized mean bias | NMB (%) | $(1/N)\sum_{i=1}^{N}(M_i - O_i)/\overline{O} \times 100$ |
| root mean square error | RMSE | $\sqrt{(1/N)\sum_{i=1}^{N}(M_i - O_i)^2}$ |
| index of agreement | IOA | $1 - \frac{\sum_{i=1}^{N}(M_i - O_i)^2}{\sum_{i=1}^{N}(\lvert M_i - \overline{O}\rvert + \lvert O_i - \overline{O}\rvert)^2}$ |

* $M$ = modeled, $O$ = observed.

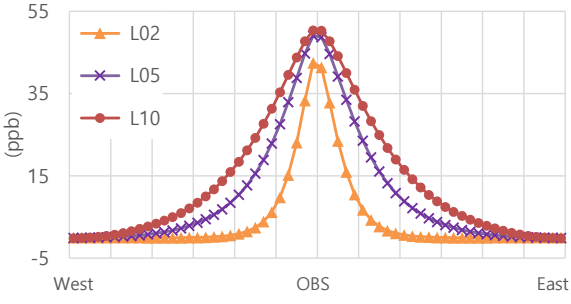


**Figure 7.** Cross-section of analysis increments along the blue line in Fig. 6b as the radius of influence ($L$) values are increase.

on grids close to the measurement sites. The observation error $\mathbf{R}_k^{-1}$ is a diagonal matrix that has same diagonal components, which is 1 % of the observed concentration.

The observation results of the diurnal variation of $O_3$ at several sites during the 12-hour time window are shown in Fig. 13, along with results of the FWD and 4DV experiments. The sites are selected in accordance with the administrative districts as shown in Fig. 1. The daytime high concentrations of $O_3$ above 100 ppb are not well simulated in the FWD, whereas they are captured in the 4DV experiment. At almost all the sites the high values of $O_3$ concentration simulated by the 4DV experiment are found to be close to the observational values. Looking at the results of the FWD, it is found that the ozone concentration at GW04 and CB06 is above 80 ppb at 09:00 LST, while the 4DV significantly reduces the

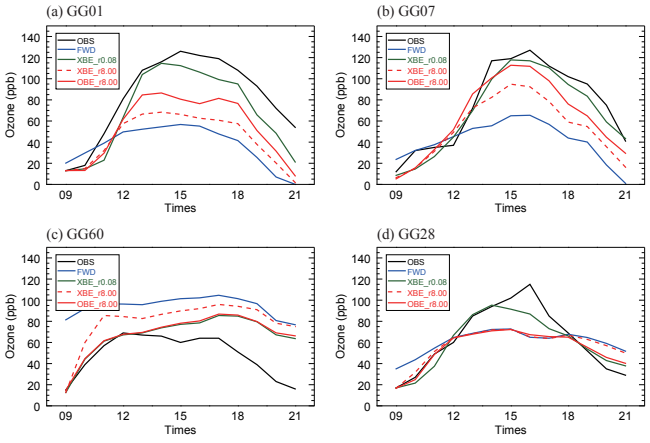


**Figure 8.** Diurnal variations of surface ozone from the results of EXP_B at **(a)** GG01, **(b)** GG07, **(c)** GG60, and **(d)** GG28. Black and blue solid lines indicate observation (OBS) and results of forward run (FWD), respectively. XBE_r0.08 (green solid), XBE_r8.00 (red dashed), and OBE_r8.00 (red solid) represent 4D-Var run results with and without considering the background error in matrix form where the observation error ($\sigma_k^{\text{obs}}$) is 0.08 and 8.00.

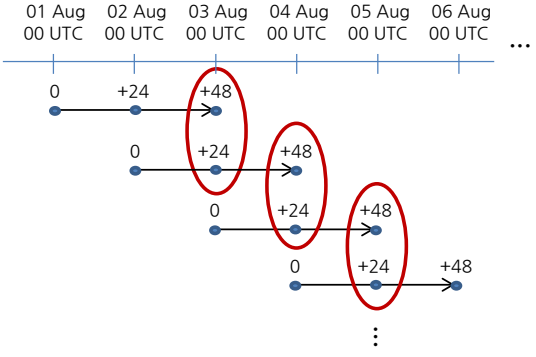


**Figure 9.** Schematic illustration for the NMC approach to obtain the background error covariance (BEC) matrix.

errors in the initial condition. However, 4DV cannot properly simulate the high concentrations of $O_3$ in the early afternoon at some sites, for example at the site GG76, and the high concentration of $O_3$ at SU21 remains underestimated. These problems are caused by uncertainties in ozone precursors that exist in both the initial conditions and in the emissions. This can probably be solved by changing the control variables and optimizing the amounts of emissions and by improving initial concentrations of the pollutants. In addition, the accuracy of the simulation for the ozone concentration in Incheon areas is directly affected by the pollutants coming from the Yellow Sea. Hence it is necessary to optimize the boundary data.

The root mean square error (RMSE) and index of agreement (IOA) of simulated results at each iteration step of 4D-Var using observation data from all sites were calculated, and the results are shown in Fig. 14 (the definitions of statistical variables used in this research are listed in Table 4). Results

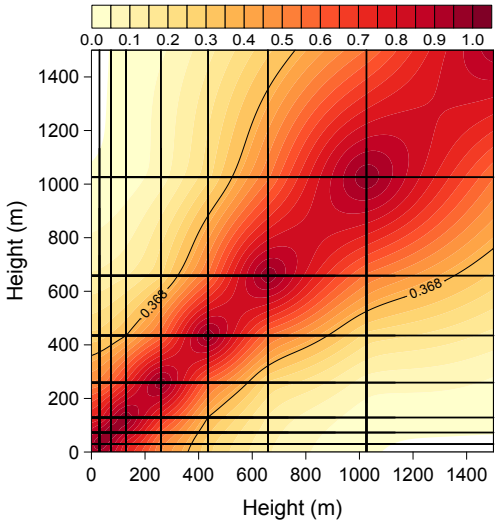


**Figure 10.** Model error correlation coefficients between vertical levels. The physical height of each level is indicated by the non-uniform grid line only in the layer below 1553 m, which is the eight layer of CMAQ.

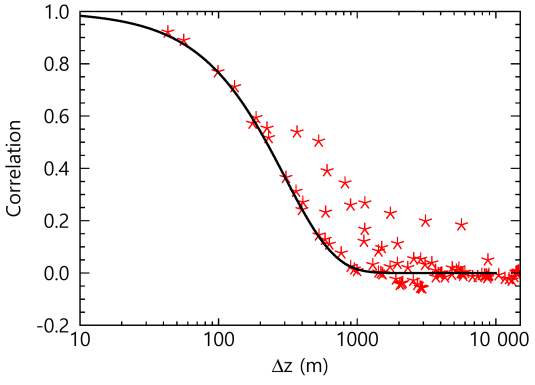


**Figure 11.** Model error correlation coefficients between two layers, as a function of $\Delta z$ (the distance between two levels). The fitted line is $R = e^{-\frac{\Delta z^{1.2}}{l_z^{1.2}}}$, where $l_z = 300$ m.

at the starting point, i.e., iteration = 1, are the statistical results of the FWD results, where RMSE and IOA are 35.1 ppb and 0.576, respectively. After approximately 20 iterations, RMSE decreases to 20 ppb or less, and IOA increases to 0.9 or more. Thereafter, there are little changes in these statistical variables, implying that the results of 4DV have converged.

Figure 15 gives the diurnal variations of the two statistical variables. As the statistical results are derived from 120 observatory sites over a fixed period of time, they actually represent the errors and general agreement in spatial distribution of $O_3$ concentration. The FWD results show a decrease in RMSE and an increase in IOA until 11:00 LST, but a rapid increase in RMSE and a decrease in IOA occur after 11:00 LST. This is caused by the inaccurate simulation of

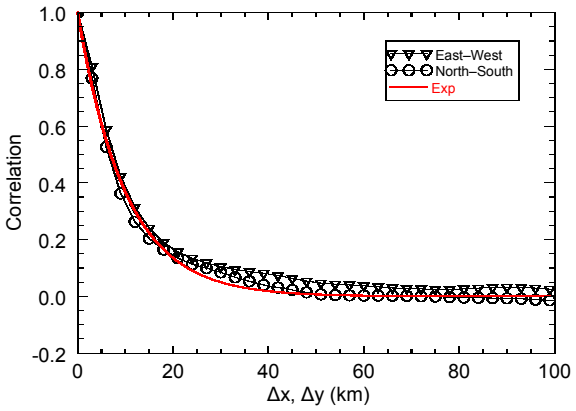


**Figure 12.** Model error correlation coefficients as a function of horizontal distance $\Delta x$ or $\Delta y$, which corresponds to east-west (revert triangles) and north–south (blank circles) direction, respectively. They can be fitted to $R = e^{-\frac{\Delta h^{1.0}}{l_h^{1.0}}}$, where $l_h = 10$ km.

high ozone concentrations during the daytime. The value of RMSE then decreases again after 16:00 LST, but large errors of $O_3$ concentration up to 30 ppb or more are still evident. In contrast, in the 4DV results, the RMSE and IOA for the initial concentration of $O_3$ are 2.9 ppb and 0.954 respectively, suggesting that the errors in the initial state are significantly reduced. Afterwards, IOA continues to decrease and reach the value of 0.543 at 21:00 LST, but this value is still higher than that in the FWD result (i.e., 0.363). The value of RMSE increases at the beginning 2 h and is close to the FWD result, but it never becomes larger than 20 ppb thereafter. In particular, the RMSE shows the maximum decrease of 27.4 ppb at 16:00 LST, which means that the accuracy of the simulation for high daytime ozone concentration has been substantially improved.

Table 5 shows the statistical results based on simulations with the 12-hour assimilation periods and from the 120 observatory sites. The simulation result of the 4DV experiment is 61.4 ppb, which is close to the average concentration of observed ozone of 63.6 ppb. A 49.4 % decrease in RMSE and a 59.9 % increase in IOA in the results of the 4DV (i.e., the difference between FWD and 4DV) demonstrate the great improvement caused by data assimilation. Mean bias, normalized by the average observed concentration (MMB), was −21.2 % in FWD, and −3.4 % in 4DV. This result of NMB implies that the tendency to underestimate daytime ozone is mitigated by application of data assimilation.

To compare the spatial distribution of the simulated $O_3$ with that of the observed concentrations, the 4DV results are presented in Fig. 16. The concentrations of $O_3$ at observatory sites are indicated with colored circles using the same color scales as the contours. At 09:00 LST, 4DV shows a homogeneous distribution, with concentrations of $O_3$ in and around Seoul to be almost zero. However, in eastern GG, GW, and CB, where the observatory sites are sparsely distributed, the concentration of $O_3$ decreases to zero only near the observatory sites. For the high concentration of ozone, i.e., 100 ppb or higher, which appears at 15:00 LST, the FWD results are approximately 50–60 ppb in Seoul (Fig. 4), and the 4DV results are consistent with the observed concentrations. However, at 18:00 LST, the difference between FWD and 4DV results grows more remarkable. Low ozone concentration appears even in central Seoul and in southeastern GG in the FWD concentration simulation at 21:00 LST, which is attributed to excessive $NO_x$ titration. However, for the 4DV results, the distribution of $O_3$ concentration in Seoul areas shows a pattern similar to that of the observations.

**Table 5.** Statistics for the observed (OBS) and simulated (FWD and 4DV) results. The FWD indicates the simulation without data assimilation. 4DV results are obtained by assimilating all observed surface $O_3$ with realized background error covariance matrix during 12 h time windows.

| Statistics | FWD | 4DV | OBS |
|---|---|---|---|
| Mean (ppb) | 50.1 | 61.4 | 63.6 |
| RMSE (ppb) | 35.1 | 17.8 | – |
| IOA | 0.576 | 0.921 | – |
| MB (ppb) | −13.5 | −2.1 | – |
| NMB (%) | −21.2 | −3.4 | – |

Figure 17 shows the difference between results of FWD and 4DV (4DV results minus that of the FWD). These differences can be regarded as analysis increments and their effects during assimilation windows. At 09:00 LST, the analysis increments are negative in most of the area, but are positive over some of the western coast area and the CN area, which is affected by the clockwise circulation of the sea-breeze. These analysis increments, which are also evident in the result of the reanalysis of initial conditions, are transported to inland areas by the local circulation. As a result, the differences between the FWD and 4DV experiments become larger, and the areas of positive values become larger too, encompassing the SU and GG areas. This process makes it possible to simulate the high concentration of daytime ozone.

### 4.4 Predictability of ozone

The direct comparison with the observation data used during the assimilation window has a limit in the verification of results. Forecasts of FWD and 4DV with different initial conditions after the time window (Table 2) are performed in this part. Figure 18a depicts the temporal variation of ozone concentration, which is obtained by averaging the results of all the observatory sites and those of corresponding model grids during the 12-hour assimilation period and the 12-hour forecast. During the period for validation, the FWD overestimates $O_3$ in the morning and underestimates it after 12:00 LST while the 4DV shows a tendency that almost conforms to that of the observations. The forecast is initial-

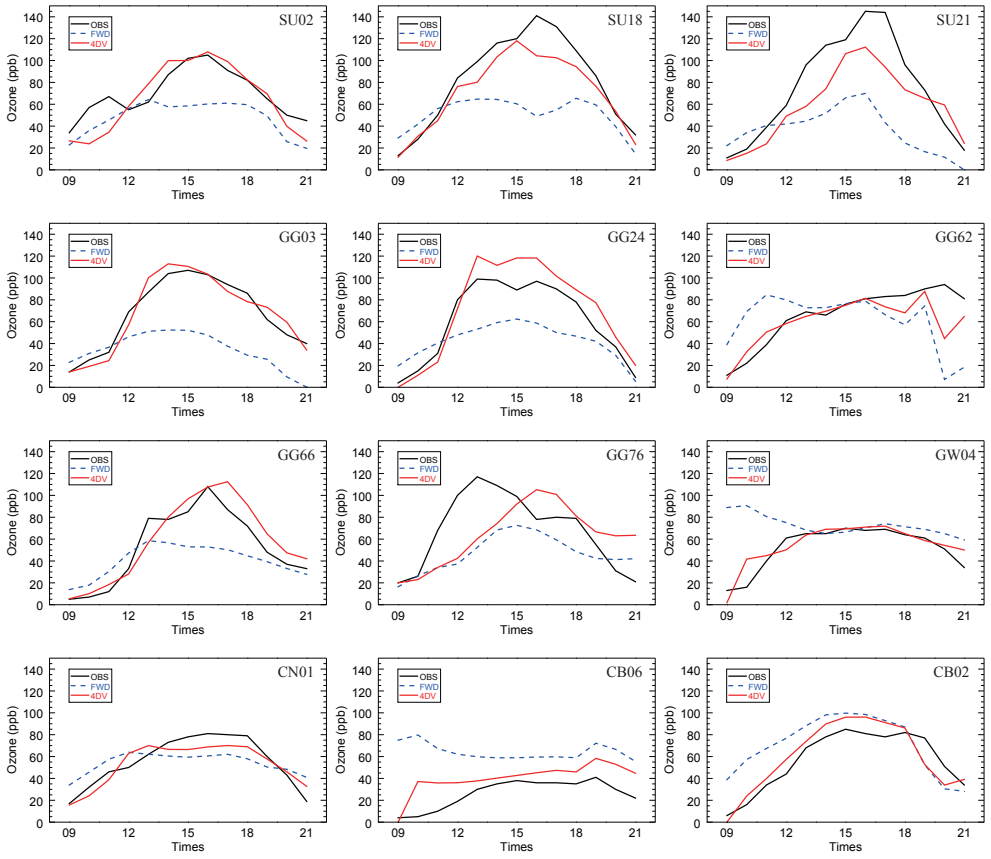


**Figure 13.** Time variations of surface ozone concentration at selected sites whose specific locations are marked by red filled triangles in Fig. 1 during daytime on 5 August. Black solid lines are observed results, and blue bashed and red solid lines indicate simulated results from the FWD and 4DV, respectively.

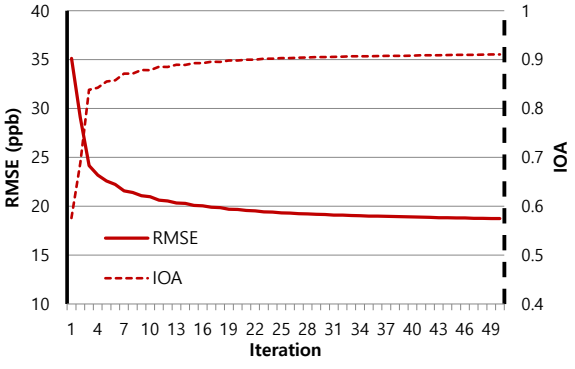


**Figure 14.** Decreasing root mean square error (RMSE, solid line) and increasing index of agreement (IOA, dashed line) with respect to each iteration step. The RMSE and IOA are calculated by comparing 4D-Var data assimilation (4D-Var) results during time window with observed $O_3$ concentration.

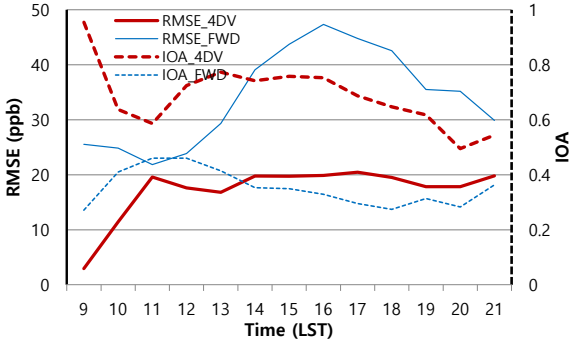


**Figure 15.** Diurnal variations of statistical results of IOA (dashed) and RMSE (solid) during the assimilation time window. The results with assimilation (4DV) are indicated by red and thick lines, and those without assimilation (FWD) are the blue and thin lines.

ized at 21:00 LST, on 5 August, and run for 24 h. The results of the first 12 h are plotted in the figure. Both experiments show a tendency to forecast high levels of nighttime ozone. However, while the FWD shows a rising tendency after 21:00 LST, the 4DV gives a declining ozone tendency and therefore provides a better forecast than the FWD. Figure 19b indicates the reduced forecast errors in the results of the 4DV, along with the time variations of statistical vari-

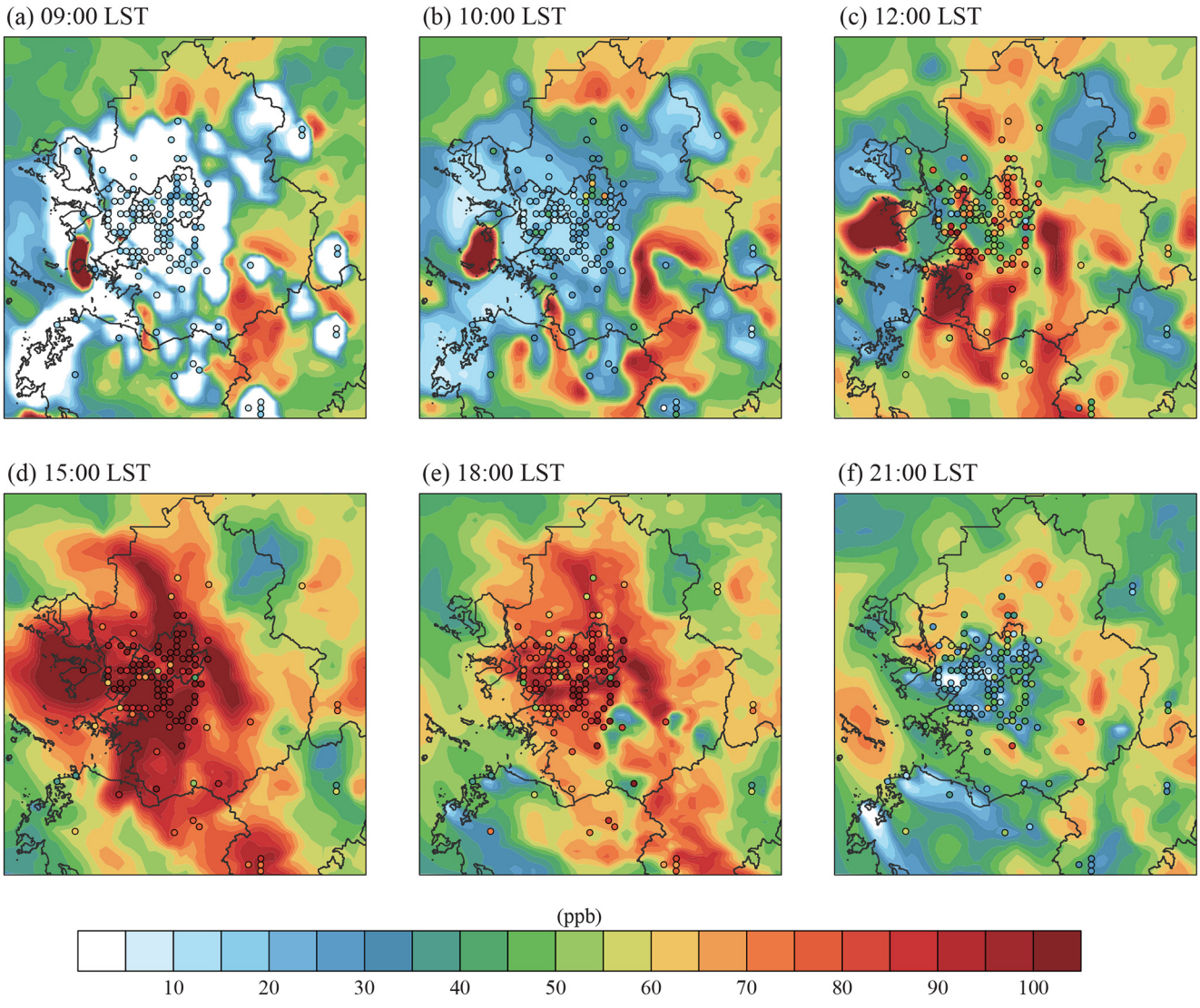


**Figure 16.** Horizontal distributions of surface ozone and its time variations. The plotted time is valid at **(a)** 09:00, **(b)** 10:00, **(c)** 12:00, **(d)** 15:00, **(e)** 18:00, and **(f)** 21:00 LST on 5 August. Contour value stands for simulated results of 4DV experiment and the filled circles with the same color scale as the contours indicate observed values.

ables, for the forecast period. At 21:00 LST, the 4DV error is only 19.8 ppb, much smaller than that of the FWD. This is attributed to the initial condition that is 10.0 ppb less than that of FWD. After 21:00 LST, the effect of improved initial condition diminishes gradually, although the RMSE in the 4DV results is still smaller than that in the FWD results. To quantitatively evaluate the overall improved predictability, the ratio of the reduced RMSE in the 4DV to that in the FWD experiments is calculated. Results indicate that the ratio is 8 % for the +24 h, and 13 % for the +12 h. This improvement in the forecast accuracy is achieved solely by using the assimilated initial condition, and more improvements are therefore expected by further optimizing the number of parameters such as emissions and boundary conditions.

The above result shows a forecast for the nighttime ozone with application of the daytime data assimilation. However, high concentrations of ozone that have harmful effects to human health are often found during daytime. Therefore, the effects of the assimilation over a time window in the nighttime upon the forecast accuracy of daytime ozone concentration are also carried out. The period for validation of data assimilation is set to be 12 h, from 12:00 UTC on 5 August to 00:00 UTC on 6 August (Table 2). The +12 h forecast period for 4DV in Fig. 18a corresponds to that of the FWD during the validation period in Fig. 18b. In the results with assimilation of nighttime ozone, the estimated ozone concentration approaches that of the observation, and the variation tendency conforms to the observation. In the ensuing fore-

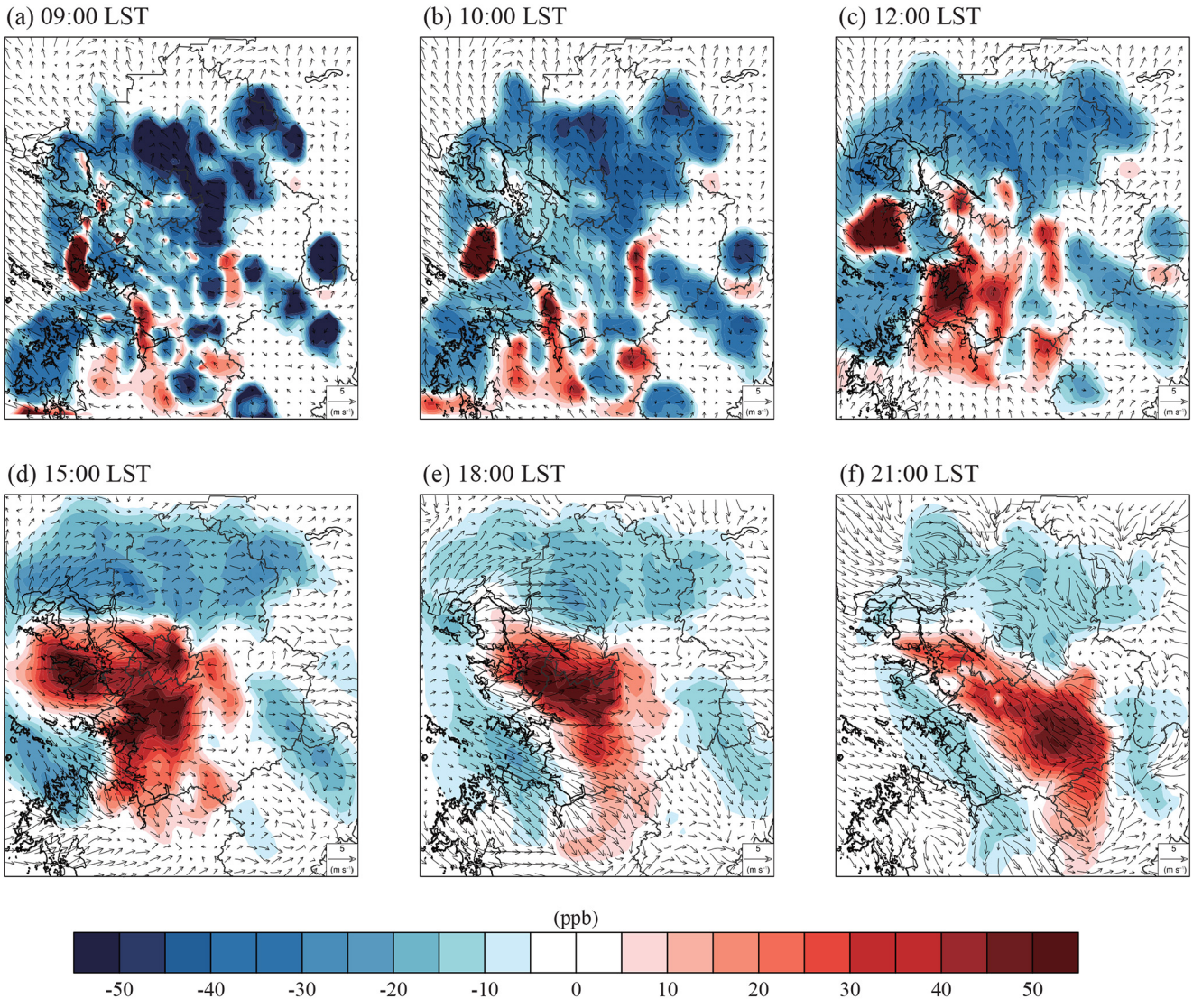


**Figure 17.** The same as Fig. 16 except that the contour value is analysis increments **(a)** and its impact on daytime ozone.

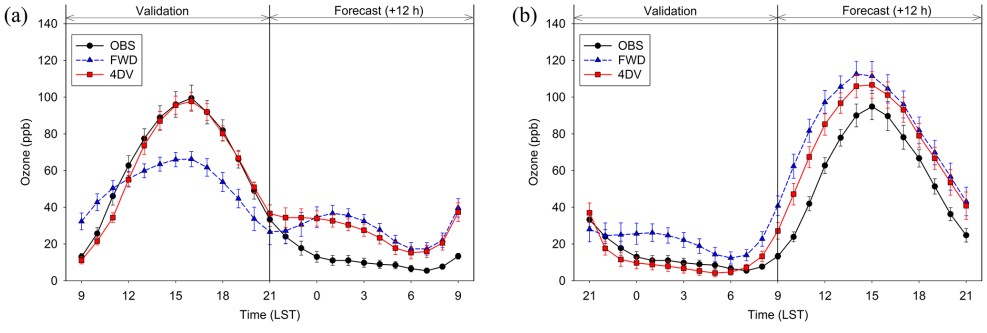


**Figure 18.** Time variations of observed and forecast ozone concentration after **(a)** daytime and **(b)** nighttime assimilation. All 120 sites data are averaged and its 3 standard errors also displayed with vertical bars. Triangle over blue dashed line, circle over red solid line, and dot over black solid line stand for forward run (FWD), 4D-Var run (4DV), and observation (OBS) results, respectively.

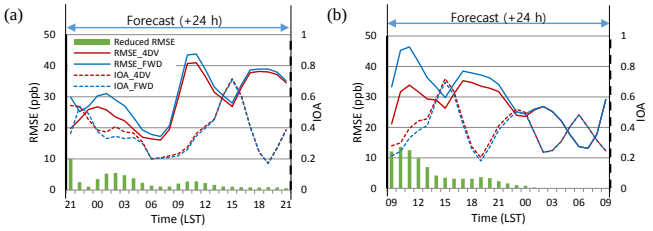


**Figure 19.** Time variations of RMSE (solid lines) and IOA (dashed lines) for 24 h forecast after **(a)** daytime and **(b)** nighttime assimilation. Red and blue lines indicate the statistical results for 4D-Var run (4DV) and forward run (FWD), respectively. Hourly reduced RMSE values are also marked along the axis of abscissas.

cast period, both of the experiments show a diurnal variation in the simulated ozone, but the FWD results demonstrate deviations from the observation, which are caused by the overestimated initial concentration at 09:00 LST. In the morning, the maximum reduced RMSE (Fig. 19b) is 13.6 ppb, and all the reductions of RMSEs are more than 10.0 ppb. After 09:00 LST, the value of the reduced RMSE decreases. The improvement in forecast accuracy, obtained by calculating the ratio of reduced errors, is 11 % for +24 h, and 17 % for +12 h, indicating that the improvement achieved by the nighttime assimilation is higher than that by the daytime assimilation. However, the effects of the improved initial condition by 4D-Var in the daytime ozone forecast cannot last for more than 12 h.

Optimized ozone after data assimilation did not show a significant change in the other chemical components (not shown here). Ozone is a secondary produced pollutant, and has no direct emission sources. Other components, especially the precursors of ozone, are mostly dependent on its emission information. Our next study will be optimizing the initial condition for $NO_x$ and VOCs to improve the predictability of $O_3$. If the multivariate background error covariance is well established, this optimization will be achieved although the control variable is different from the observed variables.

## 5 Conclusions

In this study, we presented an approach that uses an adjoint model in data assimilation. To incorporate observation data in a numerical model, the 4D-Var that is designed to improve predictability of ozone concentration is conducted by optimization of the initial values. The model systems used in the present study include WRF, CMAQ and CMAQ-ADJ.

The previously developed adjoin code for 4D-Var considers the background error of the model in the cost function as a constant. In this study, the code is revised to reflect the information of errors belonging to the actual subject areas. Verification of the revised code is conducted. Two numerical experiments are first performed by defining an ideal matrix with the assumption that the background error has a Balgovind function distribution. The results are verified. It is found that synthetic observation information is effectively spread over the neighboring areas.

In order to define the realistic model error, the NMC method that is widely used in meteorological DA is adopted in this study. The background error covariance is constructed based on the 29 differences between 48 h forecasts and 24 h forecasts, which are taken as the model error. The forecasts are performed over August, with daily initialization and a forecast period of 48 h. The vertical correlation of the model results is constructed as a diagonal and symmetric matrix; the length scale in the correlation analysis of vertical distance is about 300 m, and the scale of length in the averaged east-west and south-north correlation is about 10 km (the east-west correlation is higher than the north-south correlation).

The generated background error of the model simulation is applied in the 4D-Var research, and the surface observation is incorporated by DA to optimize the initial concentration of ozone. As a result of DA in a 12 h time window during the daytime of 5 August, the 4DV experiment shows a diurnal variation of $O_3$ concentration that conforms well to the observation, while the experiment without DA (FWD) either overestimates or underestimates the $O_3$ concentration. In the statistical result, the RMSE decreases by about 49.4 %, and the IOA increases by 59.9 %, suggesting that the initial conditions of ozone concentration are successfully improved by application of DA. The analysis increments, which are the extents of improvement of the initial conditions, spread along the route of the sea breeze that blows in from Incheon during the daytime and blows out during the evening, causing an improvement in the statistical results for the calculation area over 12 h. In addition, a potential improvement for the ozone predictability is presented using the optimized initial condition after the time window. In particular, a larger improvement in the predictability of daytime ozone concentration is expected if DA is performed over the nighttime than in the daytime.

Data assimilation has been playing an essential role in air quality modeling study. For this reason, the following studies need to be conducted for further operational applications of data assimilation.

In addition to ground data, other observations such as the data from ozone sonde, airplanes, and satellites, need to be exploited.

In the case of long-range transport, the inflow boundary condition needs to be optimized by considering it as a control variable in 4D-Var data assimilation.

Instead of using the averaged values of BEC data (which is used in the present research) to easily obtain the inverse matrix, the error correlation with different length scales at each grid should be considered. For this purpose, the preconditioning procedure, which modifies the form of the cost function, should be applied.

When considering the error covariance used in the modeling study, it is possible to conduct DA research using observation variables that are different to the control variables.

The study proposes a method to improve predictability by applying DA technology to air quality forecasts. Results of the present study provide helpful information to policy makers in charge of emission regulation. With more information related to a variety of air pollutants becoming available in the future, for example data from the geostationary orbit environmental satellite that is planned to operate in 2018 (Lee et al., 2010) and other observation systems, it is necessary to handle vast amount of observation data for better chemical weather forecasting (Carmichael et al., 2008). This study can be considered to be a preliminary research in this aspect.

*Acknowledgements.* We would like to thank the CSL research group at Virginia Tech for providing the CMAQ-ADJ code used in this study. This work was conducted under the framework of Research and Development Program of the Korea Institute of Energy Research (grant number: GP2014-0030).

Edited by: A. Baklanov

## References

Balgovind, R., Dalcher, A., Ghil, M., and Kalnay, E.: A Stochastic-Dynamic Model for the Spatial Structure of Forecast Error Statistics, Mon. Weather Rev., 111, 701–722, doi:10.1175/1520-0493(1983)111<0701:Asdmft>2.0.Co;2, 1983.

Bocquet, M., Elbern, H., Eskes, H., Hirtl, M., Žabkar, R., Carmichael, G. R., Flemming, J., Inness, A., Pagowski, M., Pérez Camaño, J. L., Saide, P. E., San Jose, R., Sofiev, M., Vira, J., Baklanov, A., Carnevale, C., Grell, G., and Seigneur, C.: Data assimilation in atmospheric chemistry models: current status and future prospects for coupled chemistry meteorology models, Atmos. Chem. Phys., 15, 5325–5358, doi:10.5194/acp-15-5325-2015, 2015.

Boisgontier, H., Mallet, V., Berroir, J. P., Bocquet, M., Herlin, I., and Sportisse, B.: Satellite data assimilation for air quality forecast, Simul. Model. Pract. Th., 16, 1541–1545, doi:10.1016/j.simpat.2008.01.008, 2008.

Byun, D. W. and Ching, J. K. S.: Science algorithms of the EPA models-3 Community Multiscale Air Quality (CMAQ) modeling system, EPA/600/R-99/030, US EPA, Research Triangle Park, USA, 1999.

Carmichael, G. R., Sandu, A., Chai, T., Daescu, D. N., Constantinescu, E. M., and Tang, Y.: Predicting air quality: Improvements through advanced methods to integrate models and measurements, J. Comput. Phys., 227, 3540–3571, doi:10.1016/j.jcp.2007.02.024, 2008.

Chai, T., Carmichael, G. R., Tang, Y., Sandu, A., Hardesty, M., Pilewskie, P., Whitlow, S., Browell, E. V., Avery, M. A., Nédélec, P., Merrill, J. T., Thompson, A. M., and Williams, E.: Four-dimensional data assimilation experiments with International Consortium for Atmospheric Research on Transport and Transformation ozone measurements, J. Geophys. Res., 112, D12S15, doi:10.1029/2006jd007763, 2007.

Chang, J. S., Brost, R. A., Isaksen, I. S. A., Madronich, S., Middleton, P., Stockwell, W. R., and Walcek, C. J.: A three-dimensional Eulerian acid deposition model: Physical concepts and formulation, J. Geophys. Res.-Atmos, 92, 14681–14700, doi:10.1029/JD092iD12p14681, 1987.

Colella, P., and Woodward, P. R.: The Piecewise Parabolic Method (PPM) for gas-dynamical simulations, J.Comp. Phys., 54, 174–201, doi:10.1016/0021-9991(84)90143-8, 1984.

Constantinescu, E. M., Chai, T., Sandu, A., and Carmichael, G. R.: Autoregressive models of background errors for chemical data assimilation, J. Geophys. Res., 112, D12309, doi:10.1029/2006jd008103, 2007.

Courtier, P. and Talagrand, O.: Variational Assimilation of Meteorological Observations With the Adjoint Vorticity Equation. Ii: Numerical Results, Q. J. Roy. Meteor. Soc., 113, 1329–1347, doi:10.1002/qj.49711347813, 1987.

Courtier, P., Andersson, E., Heckley, W., Pailleux, J., Vasiljevic, D., Hamrud, M., Hollingsworth, A., Rabier, E., and Fisher, M.: The ECMWF implementation of three-dimensional variational assimilation (3D-Var). I: Formulation, Q. J. Roy. Meteor. Soc., 124, 1783–1807, doi:10.1002/qj.49712455002, 1998.

Daescu, D. N.: On the Sensitivity Equations of Four-Dimensional Variational (4D-Var) Data Assimilation, Mon. Weather Rev., 136, 3050–3065, doi:10.1175/2007mwr2382.1, 2008.

Daley, R.: Atmospheric Data Analysis, Cambridge University Press, Cambridge, UK, 1991.

Elbern, H. and Schmidt, H.: Ozone episode analysis by four-dimensional variational chemistry data assimilation, J. Geophys. Res., 106, 3569–3590, doi:10.1029/2000jd900448, 2001.

Elbern, H., Schmidt, H., and Ebel, A.: Variational data assimilation for troospheric chemistry modeling, J. Geophys. Res., 102, 15967–15985, 1997.

Evensen, G.: Data Assimilation: The Ensemble Kalman Filter, Springer Berlin Heidelberg, Germany, 2009.

Gery, M. W., Whitten, G. Z., Killus, J. P., and Dodge, M. C.: A photochemical kinetics mechanism for urban and regional scale computer modeling, J. Geophys. Res.-Atmos., 94, 12925–12956, doi:10.1029/JD094iD10p12925, 1989.

Gou, T. and Sandu, A.: Continuous versus discrete advection adjoints in chemical data assimilation with CMAQ, Atmos. Environ., 45, 4868–4881, doi:10.1016/j.atmosenv.2011.06.015, 2011.

Hakami, A., Henze, D. K., Seinfeld, J. H., Singh, K., Sandu, A., Kim, S., Byun, D., and Li, Q.: The Adjoint of CMAQ, Environ. Sci. Technol., 41, 7807–7817, doi:10.1021/es070944p, 2007.

Hertel, O., Berkowicz, R., Christensen, J., and Hov, Ø.: Test of two numerical schemes for use in atmospheric transport-chemistry models, Atmos. Environ. A, 27, 2591–2611, doi:10.1016/0960-1686(93)90032-T, 1993.

Huang, M., Carmichael, G. R., Chai, T., Pierce, R. B., Oltmans, S. J., Jaffe, D. A., Bowman, K. W., Kaduwela, A., Cai, C., Spak, S. N., Weinheimer, A. J., Huey, L. G., and Diskin, G. S.: Impacts of transported background pollutants on summertime western US air quality: model evaluation, sensitivity analysis and data assimilation, Atmos. Chem. Phys., 13, 359–391, doi:10.5194/acp-13-359-2013, 2013.

Kalnay, E.: Atmospheric Modeling, Data Assimilation and Predictability, Cambridge University Press, Cambridge, UK, 2003.

Kucukkaraca, E. and Fisher, M.: Use of Analysis Ensembles in Estimating Flow-dependent Background Error Variances, European Centre for Medium-Range Weather Forecasts, ECMWF technical memorandum, Reading, UK, 429, 2006.

Le Dimet, F. X. and Talagrand, O.: Variational algorithms for analysis and assimilation of meteorological observations, Tellus, 38A, 97–110, 1986.

Lee, S., Hong, Y., Song, C.-K, Lee, J., Choi, W.-J., Kim D., Moon, K.-J., and Kim, J.: Plan of Korean Geostationary Environment Satellite over Asia-Pasific region, EGU General Assembly, Vienna, Austria, April 2010, EGU2010-7595-1, 2010.

Lee, D.-G., Lee, Y.-M., Jang, K.-W., Yoo, C., Kang, K.-H., Lee, J.-H., Jung, S.-W., Park, J.-M., Lee, S.-B., Han, J.-S., Hong, J.-H., and Lee, S.-J.: Korean National Emissions Inventory System and 2007 Air Pollutant Emissions, Asian J. Atmos. Environ., 5, 278–291, doi:10.5572/ajae.2011.5.4.278, 2011.

Navon, I.: Data Assimilation for Numerical Weather Prediction: A Review, in: Data Assimilation for Atmospheric, Oceanic and Hydrologic Applications, edited by: Park, S. and Xu, L., Springer Berlin Heidelberg, Germany, 21–65, 2009.

Parrish, D. F. and Derber, J. C.: The National Meteorological Center's Spectral Statistical-Interpolation Analysis System, Mon. Weather Rev., 120, 1747–1763, doi:10.1175/1520-0493(1992)120<1747:TNMCSS>2.0.CO;2, 1992.

Penenko, V., Baklanov, A., and Tsvetova, E.: Methods of sensitivity theory and inverse modeling for estimation of source parameters, Future Gener. Comp. Sy., 18, 661–671, doi:10.1016/S0167-739X(02)00031-6, 2002.

Penenko, V. V. and Obraztsov, N. N.: A variational initialization method for the fields of meteorological elements, Soviet Meteor. Hydrol., 11, 1–11, 1976.

Rabier, F., Jarvinen, H., Klinker, E., Mahfouf, J. F., and Simmons, A.: The ECMWF operational implementation of four-dimensional variational assimilation. I: Experimental results with simplified physics, Q. J. Roy. Meteor. Soc., 126, 1143–1170, doi:10.1256/Smsqj.56414, 2000.

Sandu, A. and Chai, T.: Chemical Data Assimilation – An Overview, Atmosphere, 2, 426–463, doi:10.3390/atmos2030426, 2011.

Sandu, A., Daescu, D. N., Carmichael, G. R., and Chai, T.: Adjoint sensitivity analysis of regional air quality models, J. Comput. Phys., 204, 222–252, doi:10.1016/j.jcp.2004.10.011, 2005.

Silver, J. D., Brandt, J., Hvidberg, M., Frydendall, J., and Christensen, J. H.: Assimilation of OMI $NO_2$ retrievals into the limited-area chemistry-transport model DEHM (V2009.0) with a 3-D OI algorithm, Geosci. Model Dev., 6, 1–16, doi:10.5194/gmd-6-1-2013, 2013.

Singh, K., Jardak, M., Sandu, A., Bowman, K., Lee, M., and Jones, D.: Construction of non-diagonal background error covariance matrices for global chemical data assimilation, Geosci. Model Dev., 4, 299–316, doi:10.5194/gmd-4-299-2011, 2011.

Skamarock, W. C., Klemp, J. B., Dudhia, J., Gill, D. O., Barker, D. M., Duda, M. G., Huang, X.-Y., Wang, W., and Powers, J. G.: A Description of the Advanced Research WRF Version 3, National Center for Atmospheric Research, Boulder, Colorado, USA, 2008.

Talagrand, O. and Courtier, P.: Variational Assimilation of Meteorological Observations With the Adjoint Vorticity Equation. I: Theory, Q. J. Roy. Meteor. Soc., 113, 1311–1328, doi:10.1002/qj.49711347812, 1987.

University of Houston: Air Quality Modeling of TexAQS-II Episodes with Data Assimilation, TERC Project H98, Final Report, Houston Advanced Research Center (HARC), Houston, USA, 2009.

Wang, K. Y., Lary, D. J., Shallcross, D. E., Hall, S. M., and Pyle, J. A.: A review on the use of the adjoint method in four-dimensional atmospheric-chemistry data assimilation, Q. J. Roy. Meteor. Soc., 127, 2181–2204, doi:10.1256/Smsqj.57615, 2001.

Zhang, L., Constantinescu, E. M., Sandu, A., Tang, Y., Chai, T., Carmichael, G. R., Byun, D., and Olaguer, E.: An adjoint sensitivity analysis and 4D-Var data assimilation study of Texas air quality, Atmos. Environ., 42, 5787–5804, doi:10.1016/j.atmosenv.2008.03.048, 2008.

Zhang, Q., Streets, D. G., Carmichael, G. R., He, K. B., Huo, H., Kannari, A., Klimont, Z., Park, I. S., Reddy, S., Fu, J. S., Chen, D., Duan, L., Lei, Y., Wang, L. T., and Yao, Z. L.: Asian emissions in 2006 for the NASA INTEX-B mission, Atmos. Chem. Phys., 9, 5131–5153, doi:10.5194/acp-9-5131-2009, 2009.