Atmospheric Environment 60 (2012) 99–108

Contents lists available at SciVerse ScienceDirect

Atmospheric Environment

journal homepage: www.elsevier.com/locate/atmosenv

Data assimilation of surface air pollutants (O_3 and NO_2) in the regional-scale air quality model AURORA

Ujjwal Kumar*, Koen De Ridder, Wouter Lefebvre, Stijn Janssen

VITO – Flemish Institute for Technological Research, Mol, Belgium

HIGHLIGHTS

- Applied a bias-aware optimal interpolation in the air quality model AURORA in retrospective mode.
- Hollingsworth–Lönnberg method to estimate error covariance matrices.
- The validation was carried out by using "leave ten out approach".
- The method proves to be very promising especially in retrospective simulation.

ARTICLE INFO

Article history:
Received 16 March 2011
Received in revised form
31 May 2012
Accepted 1 June 2012

Keywords:
Data-assimilation
AURORA
Optimal interpolation
Hollingsworth–Lönnberg method

ABSTRACT

In the present work, a bias-aware optimal interpolation in conjunction with the Hollingsworth–Lönnberg method to estimate error covariance matrices was applied as data assimilation algorithm in the regional scale air quality model AURORA to assimilate ground level O_3 and NO_2 concentrations. The study was conducted over the domain Belgium including part of its neighbouring areas with grid resolution of 3×3 km^2. Data assimilation was carried out for the retrospective simulation in post-processing (offline) mode for a summer and a winter month. Observations were provided by the AIRBASE data archive. Since the air quality model AURORA is presumed to represent background conditions, only the background stations within the domain have been taken into account. The validation of the proposed method was carried out by leaving observations of ten monitoring stations out in one run of the data assimilation process and another ten stations out in the next run and so on (a "leave ten out approach"). The proposed method has been evaluated in both spatial as well temporal domain against various statistical indicators such as correlation coefficient (CORR), root mean square error (RMSE), index of agreement (IOA) and mean fractional bias (MFB). For both the O_3 and NO_2, the extensive validation results have clearly shown substantial improvement in the data assimilation results over AURORA free run in both the seasons. The results over 70 validation stations show that CORR increased from 0.4 to 0.8 for O_3, 0.3 to 0.6 for NO_2 while average RMSE reduced from 27.9 to 12.6 for O_3 and from 17.4 to 11.0 for NO_2 for the month of June. Similar improvements have been observed for the month of Dec as well. Spatial CORR, IOA for monthly means of both the O_3 and NO_2 concentrations were also increased considerably. The results clearly indicate that the applied bias aware optimal interpolation in conjunction with Hollingsworth–Lönnberg method is a very promising candidate for the statistical correction of regional scale air quality modelling results for the retrospective simulation.

1. Introduction

In the last two decades, various deterministic air quality models have come into being and are being routinely applied for operational forecasting/scenario studies of air pollution concentration in many countries throughout the world. An air quality model or chemical transport model (CTM) is usually driven by a numerical weather forecast model. However, it has been observed that there might arise considerable differences between modelled and measured air pollutant concentrations. The application of data assimilation techniques has the potential to reduce this gap, thus making significant improvements in air quality modelling results.

Data Assimilation (DA hereafter) in air quality models is a relatively recent research field. Constantinescu et al. (2007) applied EnKF as DA algorithm to the Sulphur Transport Eulerian

* Corresponding author. Current address: Chemistry & Climate Division, KNMI-Royal Netherlands Meteorological Institute, De-Bilt, The Netherlands.
E-mail addresses: ujjwalkumarin@yahoo.co.in, ujjwalkumarin@gmail.com (U. Kumar).

1352-2310/$ – see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.atmosenv.2012.06.005

Model (STEM) to simulate and assimilate various chemical species concentration over South East Asia. The concentration fields of both directly observed (O_3, NO_2) and unobserved species (HCHO, PAN) were considerably improved by EnKF data assimilation. Denby et al. (2009) compared two data-assimilation algorithms, statistical interpolation method residual kriging, and Ensemble Kalman filtering in the LOTOS-EUROS model. Although both methods improved the results, the statistical interpolation method performed significantly better than the EnKF. Wu et al. (2008) implemented and compared four DA algorithms, namely optimal interpolation, reduced-rank square root Kalman filter, ensemble Kalman filter, and four-dimensional variational assimilation, in the same benchmark settings for the Polyphemus model Polair3D (Boutahar et al., 2004) covering the domain of western Europe. The authors concluded that the optimal interpolation provided overall better performances, EnKF produced best forecasts during the end of prediction periods and the strongly constrained 4D-Var did a moderate job. Tombette et al. (2009) applied optimal interpolation in the aerosol model SIREAM (SIze-REsolved Aerosol Model), plugged to the chemistry-transport model Polair3D, covering the domain of European continent. The assimilation of PM_{10} observations significantly improved the one-day forecast of total particle mass (PM_{10} and $PM_{2.5}$). More recently, Frydendall et al. (2009) implemented optimal interpolation in the regional air pollution forecast model DEOM for ground level O_3 concentration over Europe. The optimal interpolation algorithm significantly improved the performance of the DEOM model when compared to the measurements.

In the present study, the regional scale air quality model AURORA was applied over Belgium and its neighbouring areas, covering a domain extending from approximately 49.9°N to 51.8°N, and 1.9°E to 7.1°E with a horizontal resolution of 3×3 km^2. A bias aware optimal interpolation was implemented in conjunction with the Hollingsworth–Lönnberg method (Hollingsworth and Lönnberg, 1986) in AURORA to assimilate ground level O_3 and NO_2 concentrations in order to improve the retrospective simulation results. The measured values of air pollutants concentrations were obtained from the AIRBASE data archive (http://air-climate.eionet.europa.eu/databases). Only the data from background monitoring stations were retained for the purpose of data-assimilation as, in principle, the AURORA model captures the background concentration of air pollutants.

The remainder of the paper is organized as follows. The air quality model AURORA and the data assimilation algorithm are described in Section 2. Section 3 presents the results of the free AURORA model run versus those obtained with AURORA with data assimilation, and Section 4 concludes the study.

2. Methodology

2.1. AURORA

AURORA (Air quality modelling in Urban Regions using an Optimal Resolution Approach) is a regional (limited area) Eulerian chemistry transport model developed at VITO (Lefebvre et al., 2011; De Ridder et al., 2008a; De Ridder et al., 2008b; Mensink et al., 2008; De Ridder et al., 2004; Lefebre et al., 2004; Mensink et al., 2002; Mensink et al., 2001). Advection is treated using the Walcek (2000) scheme, which is monotonic, exhibits a relatively limited numerical diffusion, and comes at a reasonable computational cost. Vertical diffusion is calculated with the Crank–Nicholson method (De Ridder and Mensink, 2002). Gaseous chemistry is treated by means of the carbon-bond IV scheme (Gery et al., 1989), which was enhanced to include the effect of biogenic isoprene emissions.

For the road traffic emissions, the MIMOSA4 model was used (Mensink et al., 2000; Vankerkom et al., 2009), which generates hourly output for different types of emissions, such as NO_2, PM_{10} and $PM_{2.5}$ for Flanders (see also Lefebvre et al., 2011). The non-traffic emissions of the different pollutants such as NO_x, PM_{10} and $PM_{2.5}$ for Flanders are based on the emission inventory compiled by the Flemish Environmental Agency. For the other regions outside the Flanders but still within the simulation domain, the EMAP-tool (Maes et al., 2009) was used to generate gridded emissions based on the EMEP data set.

By using the AURORA dispersion model we are aiming at producing air quality maps for Belgium and its neighbouring areas. Three different resolution steps were taken for the AURORA modelling. The nesting starts with a domain at 25×25 km^2 (71×71 points) resolution through a domain at 9×9 km^2 (71×71 points) to 3×3 km^2 resolution (121×71 points). The AURORA model run at a resolution of 25×25 km^2 was nested in the BelEUROS model (Deutsch et al., 2008a, 2008b). This model calculates air quality above the whole of Europe at a resolution of 60×60 km^2. Hourly BelEUROS data were interpolated to the AURORA model grid. All nesting was one-way nesting. The AURORA model outputs at 3×3 km^2 for ground level O_3 and NO_2 concentrations over the domain extending from approximately 49.9°N to 51.8°N and 1.9°E to 7.1°E have been used in the current study. Meteorological fields, required as input for AURORA, were simulated using the ARPS (Advanced Regional Prediction System) model, a non-hydrostatic mesoscale atmospheric model developed by the University of Oklahoma (Xue et al., 2000, 2001). More information on the AURORA model can be found in the European Model Database (http://air-climate.eionet.europa.eu/databases/MDS/index_html).

2.2. Data assimilation

Optimal interpolation (OI) was applied as data assimilation algorithm in post-processing (off-line) mode, using hourly simulated concentration fields together with measured values. If $\mathbf{z}$ represents the observations and $\mathbf{x}^{m}$ the model vector, the analysis vector $\mathbf{x}^{a}$ is obtained by weighing the model errors against the observation errors. This leads to the interpolation equation (Bouttier and Courtier, 2002; Kalnay, 2003):

$$\mathbf{x}^{a} = \mathbf{x}^{m} + \mathbf{K}(\mathbf{z} - \mathbf{H}\mathbf{x}^{m}) \quad (1)$$

where $\mathbf{K}$ is called the Kalman gain matrix, which depends on the background error covariance matrix $\mathbf{B}$ and the observational error covariance matrix $\mathbf{R}$. $\mathbf{H}$ is the observation operator that maps the model phase space onto the observation space. In this study, pollutant concentrations were modelled at 3 km resolution, and were directly compared with observations at the corresponding monitoring stations. Thus, in this case, $\mathbf{H}$ is simply a matrix consisting of 0's and 1's to match the correspondence between observations and modelled values. The Kalman gain $\mathbf{K}$ is obtained as follows in the least square sense (Bouttier and Courtier, 2002; Kalnay, 2003):

$$\mathbf{K} = \mathbf{B}\mathbf{H}^{\mathbf{T}}\left(\mathbf{H}\mathbf{B}\mathbf{H}^{\mathbf{T}} + \mathbf{R}\right)^{-1} \quad (2)$$

This is the standard bias-blind analysis equation when model and observations both are considered unbiased. However, biases can be present either in model or observations. For observations, the bias of the data normally arises from low representativeness of the observation stations. The background stations can normally be considered as unbiased. Hence, we select only background stations as representative stations. However, the model outputs are often biased. Therefore, we follow the bias-aware data-assimilation

method as proposed by Dee and da Silva (1998) and Dee and Todling (2000). The explanation for the bias aware data-assimilation method described below has been adapted from Dee and Todling (2000). Dee and da Silva (1998) showed how to produce unbiased analyses in a sequential data assimilation system when the forecast (model-output) is biased. The idea is to provide a running estimate of the bias and to correct the forecast (model-output) accordingly. The result is the replacement of (1) by the following two-step algorithm:

$$\widehat{\mathbf{b}}_{\mathbf{k}} = \widehat{\mathbf{b}}_{\mathbf{k-1}} - \mathbf{L}\left[\mathbf{z_k} - \mathbf{H}\left(\mathbf{x}_{\mathbf{k}}^{\boldsymbol{m}} - \widehat{\mathbf{b}}_{\mathbf{k-1}}\right)\right] \tag{3}$$

$$\mathbf{x}_{\mathbf{k}}^{\mathbf{a}} = \left(\mathbf{x}_{\mathbf{k}}^{\boldsymbol{m}} - \widehat{\mathbf{b}}_{\mathbf{k}}\right) + \mathbf{K}\left[\mathbf{z_k} - \mathbf{H}\left(\mathbf{x}_{\mathbf{k}}^{\boldsymbol{m}} - \widehat{\mathbf{b}}_{\mathbf{k}}\right)\right]. \tag{4}$$

The $n \times 1$ vector $\widehat{\mathbf{b}}_{\mathbf{k}}$ is the estimated forecast bias at time t_k. It is to be noted that (4) and (1) are identical when $\widehat{\mathbf{b}}_{\mathbf{k}} = 0$. The $n \times$ p matrix $\mathbf{L}$ defines the weighting coefficients for the bias update equation. Dee and Todling (2000) show that the optimal weights for the bias estimator are

$$\mathbf{L} = \mathbf{B^b H^T}\left[\mathbf{H B^b H^T} + \mathbf{H B H^T} + \mathbf{R}\right]^{-1} \tag{5}$$

where $\mathbf{B^b}$ is the error covariance matrix of the bias. For more details, please refer to Dee and Todling (2000). In this equation, we have used stationary covariance matrices.

2.3. Estimation of error covariance matrices

The specification of error covariance matrices $\mathbf{B}$ and $\mathbf{R}$ is an important step in data assimilation. In the current study, the background error covariance matrices were estimated using the Hollingsworth–Lönnberg method (Hollingsworth and Lönnberg, 1986; Lönnberg and Hollingsworth, 1986; Daley, 1996). The following explanation of the method has been adapted after Tilmes (2001). Let z_k and x_k be vector time series of observations and model simulated values, respectively, for an air pollutant at a station k out of M monitoring stations. One important assumption is that the data be free of biases, so that $\bar{z_k} = \bar{x_k}$, where the overbar denotes the average of a long time series at a station k. The model output x_k at a station k has been corrected for the bias using $\tilde{x_k} = x_k + b_k$ where $b_k = \bar{z_k} - \bar{x_k}$. Thereafter, the model errors have been calculated using the bias corrected model outputs, i.e., $e_{\text{model-error}} = z_k - \tilde{x}_k$. These model errors have been used to establish the model error covariance matrix $\mathbf{B}$.

Now the correlations R_{kl} between observation increments $(z_k - \tilde{x}_k)$ for each pair of sites k and l can be calculated. Plotting R_{kl} as a function of the distance r_{kl} between stations k and l on a scatter diagram yields the isotropic component of the correlation. The next step is to fit a curve $R(r)$ through these points. Provided homogeneity, the variance of the observation increments $\overline{(z_k - \tilde{x}_k)^2}$ should be independent of the site k and given that errors in the simulated fields and the observations are not mutually correlated, then

$$\frac{1}{M}\sum_{k=1}^{M}\overline{(z_k - \tilde{x}_k)^2} = E_O^2 + E_m^2 \tag{6}$$

where E_O^2 and E_m^2 denote the variances in the observations and the background (model) errors. The observation errors consist of the uncertainties in the measurements and the error of representativeness, i.e., the sub-grid variance. Now, $R_{kk} = 1$ by definition, but in general extrapolation of curve $R(r)$ to zero distance $R_z = \lim_{r-->0} R(r)$ does not yield unity. Further, as monitoring instruments installed at different monitoring stations are independent of each other, we can assume the observation errors between different monitoring stations are horizontally uncorrelated, but the background error might be correlated.

Thus, R_z is a measure for the horizontally correlated part of error:

$$R_z = \frac{E_m^2}{E_O^2 + E_m^2}. \tag{7}$$

E_O^2 and E_m^2 can be obtained from the Equations (3) and (4) after determining R_z and $1/M\sum_{k=1}^{M}\overline{(z_k - \tilde{x}_k)^2}$ from the data.

Based on the above mentioned assumptions, a very simple model for the error covariance matrices in the observations $\mathbf{R}$ and in the model field $\mathbf{B}$ can be derived:

$$\mathbf{R} = E_O^2 \cdot \mathbf{I}, \quad \mathbf{B} = E_m^2 \cdot \rho_m \tag{8}$$

where $\mathbf{I}$ is the identity matrix and the elements ρ_{kl} of the matrix ρ_m are defined by $\rho_{kl} = R(r_{kl})/R_z$, r_{kl} being the horizontal distance between stations k and l.

In addition to the background and observation error covariance matrices $\mathbf{B}$ and $\mathbf{R}$, the bias aware data-assimilation method also requires the specification of error covariance matrix $\mathbf{B^b}$ of the bias-estimates. The explanation for the $\mathbf{B^b}$ calculation below has been taken from Dee and Todling (2000). Dee and da Silva (1998) proposed the following model for the error covariances of the bias estimates:

$$\mathbf{B^b} = \gamma \mathbf{B} \tag{9}$$

with γ constant. This model assumes that the spatial correlations of the bias estimation errors are identical to those of the random components of the background errors, and that the two types of errors are balanced in the same way. To arrive at a means for determining an appropriate value for the parameter γ, we study the behaviour of the bias estimator at a single observation station that coincides with the model grid point. The bias gain (4) is then

$$\mathbf{L} = \lambda = \boldsymbol{\sigma_b^2}/(\boldsymbol{\sigma_b^2} + \boldsymbol{\sigma_m^2} + \boldsymbol{\sigma_o^2}) \tag{10}$$

with $\boldsymbol{\sigma_b^2}$, $\boldsymbol{\sigma_m^2}$, and $\boldsymbol{\sigma_o^2}$ the error standard deviations for the bias estimate, the model, and the observation, respectively, which we take to be stationary for the moment. In this case, bias Equation (3) becomes

$$\begin{aligned}\widehat{\mathbf{b}}_{\mathbf{k}} &= \widehat{\mathbf{b}}_{\mathbf{k-1}} - \lambda\left[\mathbf{z_k} - \left(\mathbf{x}_{\mathbf{k}}^{\mathbf{m}} - \widehat{\mathbf{b}}_{\mathbf{k-1}}\right)\right] \\ &= -\lambda\sum_{\mathbf{j}=0}^{\mathbf{k-1}}(1-\lambda)^{\mathbf{j}}\mathbf{v}_{\mathbf{k-j}}\end{aligned} \tag{11}$$

with $\mathbf{v_k} = \mathbf{z_k} - \mathbf{x}_{\mathbf{k}}^{\mathbf{m}}$, and $\widehat{\mathrm{b}}_0 = 0$.

Dee and Todling (2000) further showed that, in the absence of model bias correction, the time series of observed-minus-model residual typically have coloured spectra. A value for the parameter λ is determined such that the spectra of the bias-corrected observed-minus-model residuals become as flat as possible, in some well defined sense. In the time-frequency domain, (11) corresponds to

$$\beta_n = R_n(\lambda)\nu_n \tag{12}$$

Where β_n, ν_n are the Fourier coefficients for wave-number $n > 0$ of the time series $\widehat{\mathbf{b}}_{\mathbf{k}}$, ν_k, respectively. The response function R_n is

$$R_n(\lambda) = -\lambda\left[1 - (1-\lambda)e^{2\pi i \Delta t/n}\right], \tag{13}$$

A practical method for estimating λ (Dee and Todling, 2000) is to compute the average normalized power spectrum P_n of the

residuals for a set of stations, and then to find λ that minimizes the functional

$$f(\lambda) = \sum_n n^2\{|[1 + R_n(\lambda)]P_n| - 1\}^2 \quad (14)$$

A value of λ can thus be computed. Having determined λ and with $\boldsymbol{\sigma_m^2}$ and $\boldsymbol{\sigma_o^2}$ given, (9)–(10) imply

$$\gamma = \frac{\lambda}{1-\lambda}\frac{\boldsymbol{\sigma_m^2} + \boldsymbol{\sigma_o^2}}{\boldsymbol{\sigma_m^2}} \quad (15)$$

which, in conjunction with (9), completes the specification of the bias estimation error covariance model.

3. Results and discussion

Data assimilation was carried on hourly basis for the two air pollutants O_3, NO_2 for two different months (June and December 2007) representing two different seasons summer and winter. Fig. 1 shows the map of the domain and the location of the background AIRBASE observation stations used in the study. The data from 73 background observation stations within the domain have been used for both O_3 and NO_2. All processes of data-assimilation were carried out in post-processing offline mode. For the purpose of validation, the data from 10 monitoring stations were left out in one run of data assimilation. In the next run, another 10 stations were left out and so on (leave ten out approach). Thus, the validation has been made possible for the 70 available observation stations by carrying out 7 runs.

3.1. Data assimilation results

First step in the OI data-assimilation is to infer and establish the error covariance matrices. To evaluate the background and observational error covariance matrices **B** and **R**, it is important to investigate the relation between correlation coefficient of observation increments for a pair of observation stations vs. distance between those paired observation stations. This was done both for the O_3 and NO_2 for the months of June-2007 and Dec-2007. Fig. 2(a) and Fig. 2(b) show this relation for the hourly O_3 concentrations for the month of June-07 and Dec-2007, respectively. Fig. 2(c) and Fig. 2(d) represent the same relation for the hourly NO_2 concentrations for the months of Jun-2007 and Dec-2007, respectively. The fitted exponential curves for these relations [as shown in the Fig. 1] were used to establish both the background (**B**) and observational (**R**) error covariance matrices as described in Section 2.3. Once **B** and **R** determined, the Kalman gain matrix can easily be determined according to Equation (2) and finally the analysis field is obtained using Equation (1).

To estimate the covariance matrix $\mathbf{B^b}$ of the bias estimates, the residuals (observed minus model) time series have been analysed at various observation stations. The functional $f(\lambda)$ of Equation (14) has been optimized by tuning the values of λ and the value of γ (Equation (15)) has been calculated. We get a low value of $\gamma \sim 0.05$ and this has been used to calculate the $\mathbf{B^b}$ and **L** (gain matrix of bias Equation (3)).

Using the "leave ten out approach", the validation results for most of the observation stations have been obtained. The validation results for each observation station were evaluated against various statistical indicators such as CORR (correlation coefficient), RMSE (root mean square error), IOA (index of agreement) and MFB (mean fractional bias), as described below. Let O_i represents an observation data at a time instant i and P_i represents the predicted data at the same time instant i, i.e., P_i is the value obtained through either (i) AURORA free run without data-assimilation (AURORA-FREE) or, (ii) AURORA run with data assimilation (AURORA-DA). Let N be the number of observations at a particular observation station. The various statistical indicators are defined as below (Wilks, 2006; Borrego et al., 2008):

$$\mathrm{CORR} = \frac{\frac{1}{N-1}\sum_{i=1}^{N}(O_i - \overline{O})(P_i - \overline{P})}{\left(\frac{1}{N-1}\sum_{i=1}^{N}(O_i - \overline{O})^2\right)^{1/2}\left(\frac{1}{N-1}\sum_{i=1}^{N}(P_i - \overline{P})^2\right)^{1/2}}$$

$$\mathrm{RMSE} = \left(\frac{1}{N}\sum_{i=1}^{N}(P_i - O_i)^2\right)^{1/2}$$

$$\mathrm{IOA} = 1 - \frac{\sum_{i=1}^{N}(P_i - O_i)^2}{\sum_{i=1}^{N}(|P_i - \overline{O}| + |O_i - \overline{O}|)^2}$$

$$\mathrm{MFB} = \frac{1}{N}\sum_{i=1}^{N}\frac{(P_i - O_i)}{(P_i + O_i)/2}$$

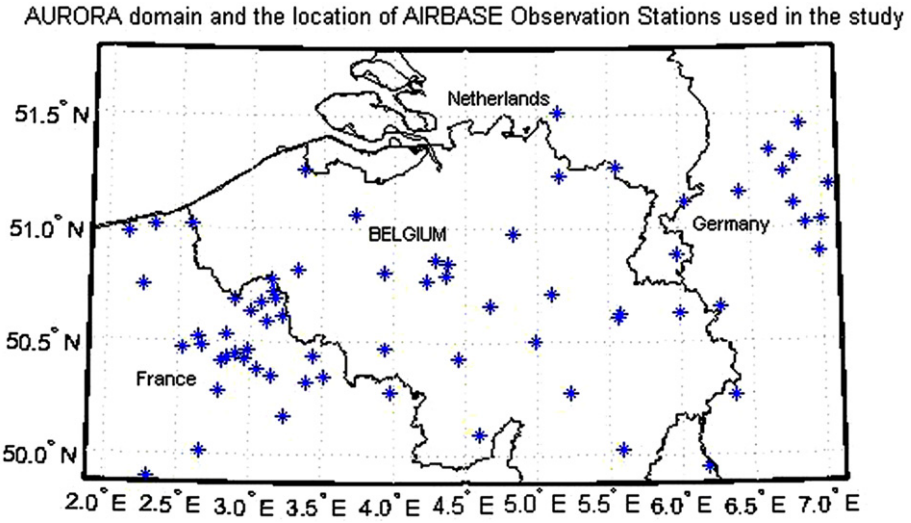

Fig. 1. Map of the AURORA domain. The marked blue stars on the map show the location of the background AIRBASE observation stations used in the study. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

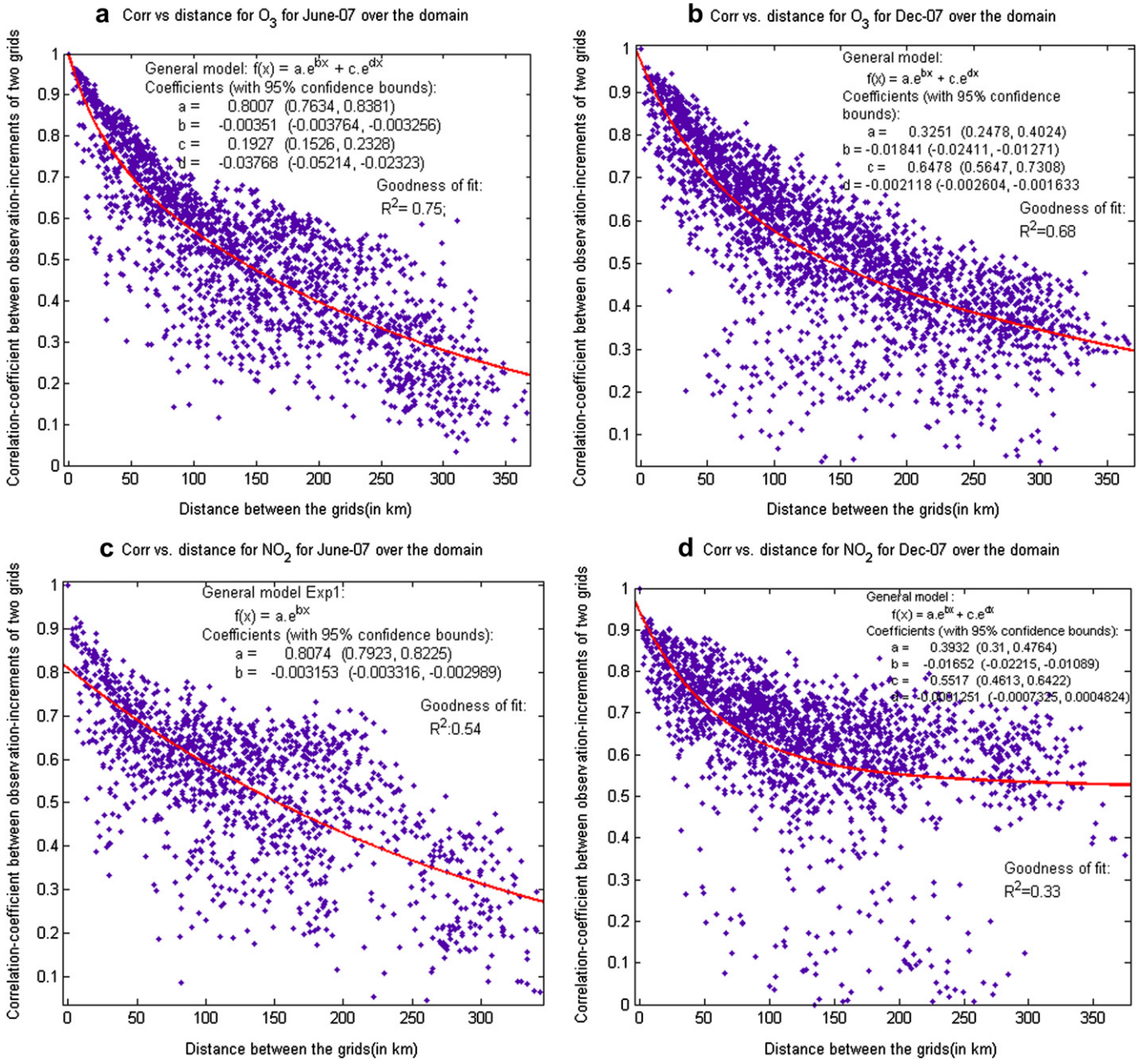

Fig. 2. Best fit exponential curve in the least square sense and the scatter plots for the correlation coefficient of observation increments between a pair of observation stations vs. distance between those pair of observation stations for the hourly (a) O_3 concentrations for the month of Jun-2007, (b) O_3 concentrations for the month of Dec-2007, (c) NO_2 concentrations for the month of Jun-2007, (b) NO_2 concentrations for the month of Dec-2007.

Fig. 3 presents the O_3 concentrations validation results for the 10 observation stations. It is pertinent to mention here that the observations data from the shown 10 observation stations (named in Fig. 3) have not been used for the data assimilation while presenting the results. The validation results for the above mentioned statistical indicators have been shown for the months June-2007 [Fig. 3(a,b,c,d)] and Dec-2007 [Fig. 3(e,f,g,h)]. For both the months June-2007 and Dec-2007, Fig. 3 clearly reveals that AURORA-DA results have significantly improved over AURORA-FREE for each statistical indicator at the 10 observation stations. CORR and IOA have significantly increased (by a factor of 1.2–2) for AURORA-DA while RMSE and MFB have substantially reduced (by a factor of about 0.5) for the data assimilated results (AURORA-DA). Though for sake of brevity and clarity, only 10 validation stations have been shown in the figure, the results from all the possible 70 validation stations from 7 runs of data-assimilation show that RMSE decreased and CORR improved for all the 70 validation stations for both the month of June and Dec. For NO_2, the similar validation results for Jun-2007 and Dec-2007 have been presented in Fig. 4. For both the seasons (summer June-2007, winter Dec-2007), AURORA-DA has performed significantly well over AURORA-FREE except at station BETN064 in Fig. 4 (f). For AURORA-DA, CORR and IOA have significantly increased (by a factor of 1.5–2), MFB has substantially reduced, RMSE has also considerably decreased (by a factor of about 0.5). Fig. 4(f) shows that except at station BETN064, RMSE has decreased at all the other 9 stations. In fact, for the month of June-2007, out of 70 validation stations, except at 4 stations where RMSE were already low ($\leq$10 $\mu g\ m^{-3}$), RMSE has decreased by a factor of half at all the other

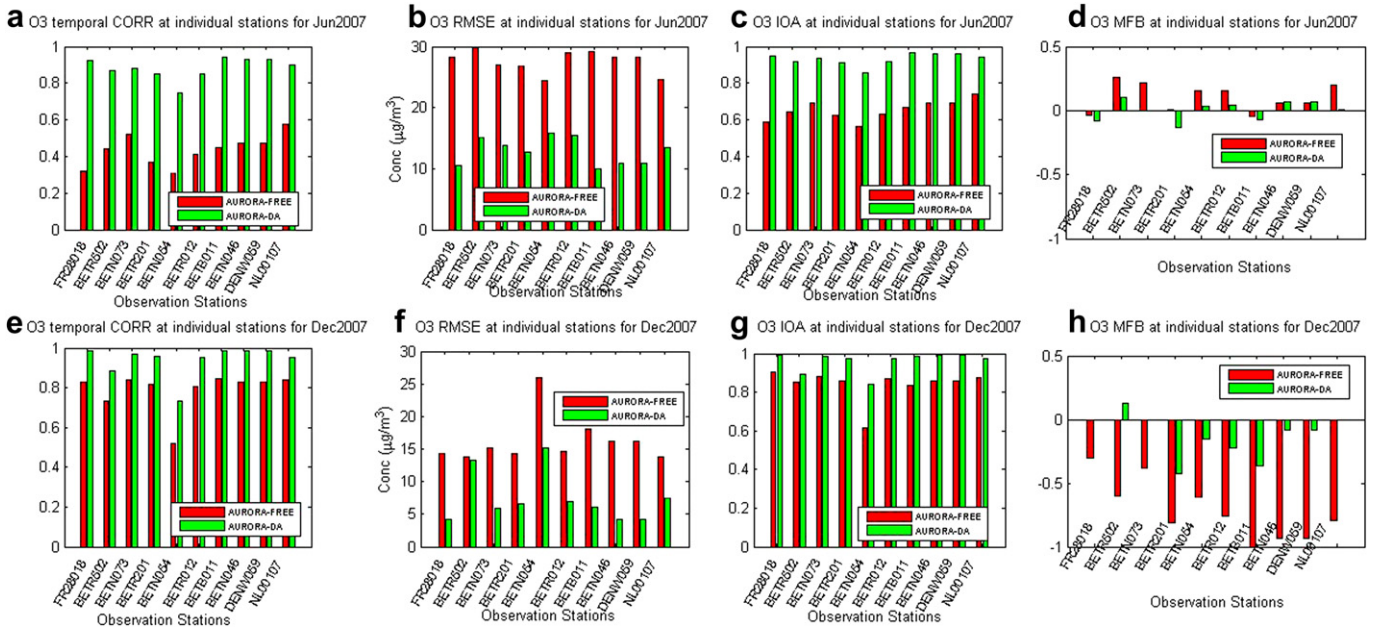

Fig. 3. O_3 validation results for ten individual observation stations, (a) CORR for Jun-2007, (b) RMSE for Jun-2007, (c) IOA for Jun-2007, (d) MFB for Jun-2007, (e) CORR for Dec-2007, (b) RMSE for Dec-2007, (c) IOA for Dec-2007, (d) MFB for Dec-2007.

66 validation stations. For the month of December, only 2 stations have been observed to show a little increase in RMSE (where AURORA-FREE RMSE were already low ($\leq$10 μg m^{-3}) while for the rest of the 68 validation stations, RMSE has decreased by a factor of half. Overall, the validation results at each individual observation stations clearly show that there is appreciable improvement in the results of AURORA-DA over AURORA-FREE for the O_3 as well as NO_2 concentrations for both the seasons.

In order to gain more insight into the validation results, the time series plots for O_3 and NO_2 concentrations have been depicted in Fig. 5 for one observation station. The data of this observation station has been completely excluded from the data assimilation, thus, the entire time-series has been generated using the AURORA-DA only. The daily mean and daily 8-h maximum O_3 concentration at an observation station BETN073 for the month of June-2007 and Dec-2007 have been shown in Fig. 5 (a,b) and Fig. 5(c,d), respectively. While defining the daily mean and daily 8-h maximum O_3 concentration, AIRBASE guidelines have been followed (http://air-climate.eionet.europa.eu/databases/airbase/aggregation_statistics.html, accessed in Nov-2010). An inspection of Fig. 5(a,b,c,d) clearly reveals that the daily mean as well as the daily 8-h maximum O_3 concentrations from AURORA-DA (green line) follows the actual

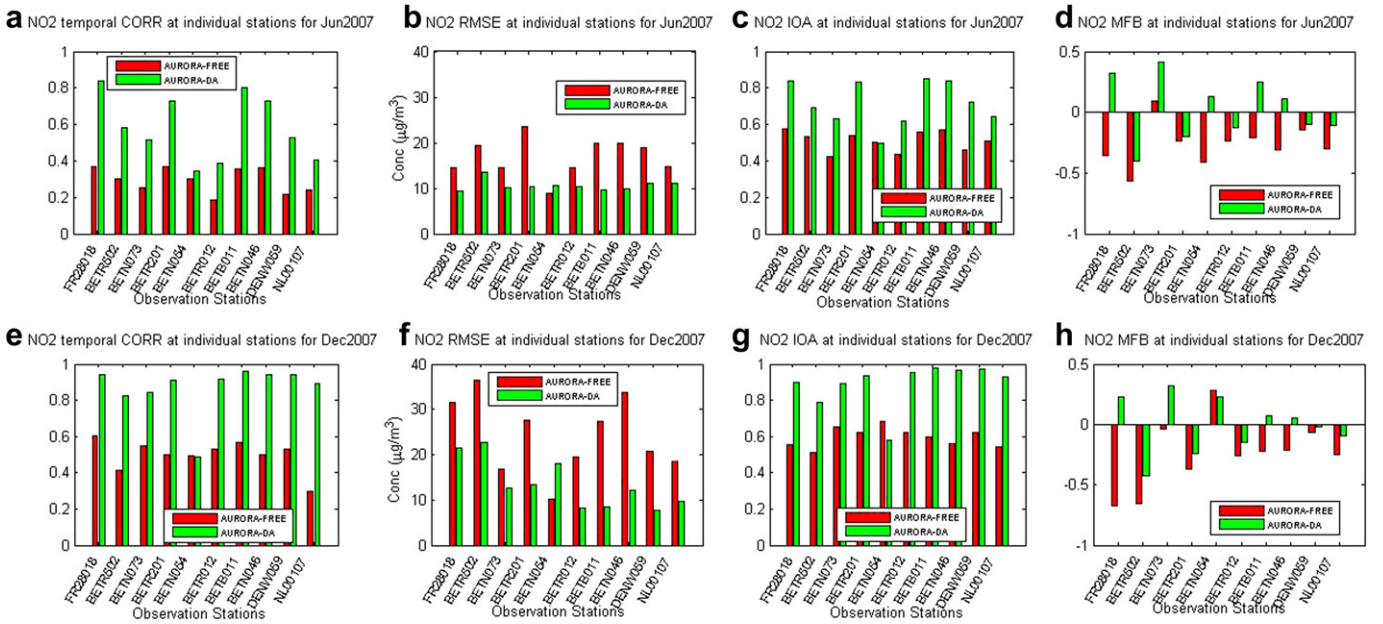

Fig. 4. NO_2 validation results for ten individual observation stations, (a) CORR for Jun-2007, (b) RMSE for Jun-2007, (c) IOA for Jun-2007, (d) MFB for Jun-2007, (e) CORR for Dec-2007, (b) RMSE for Dec-2007, (c) IOA for Dec-2007, (d) MFB for Dec-2007.

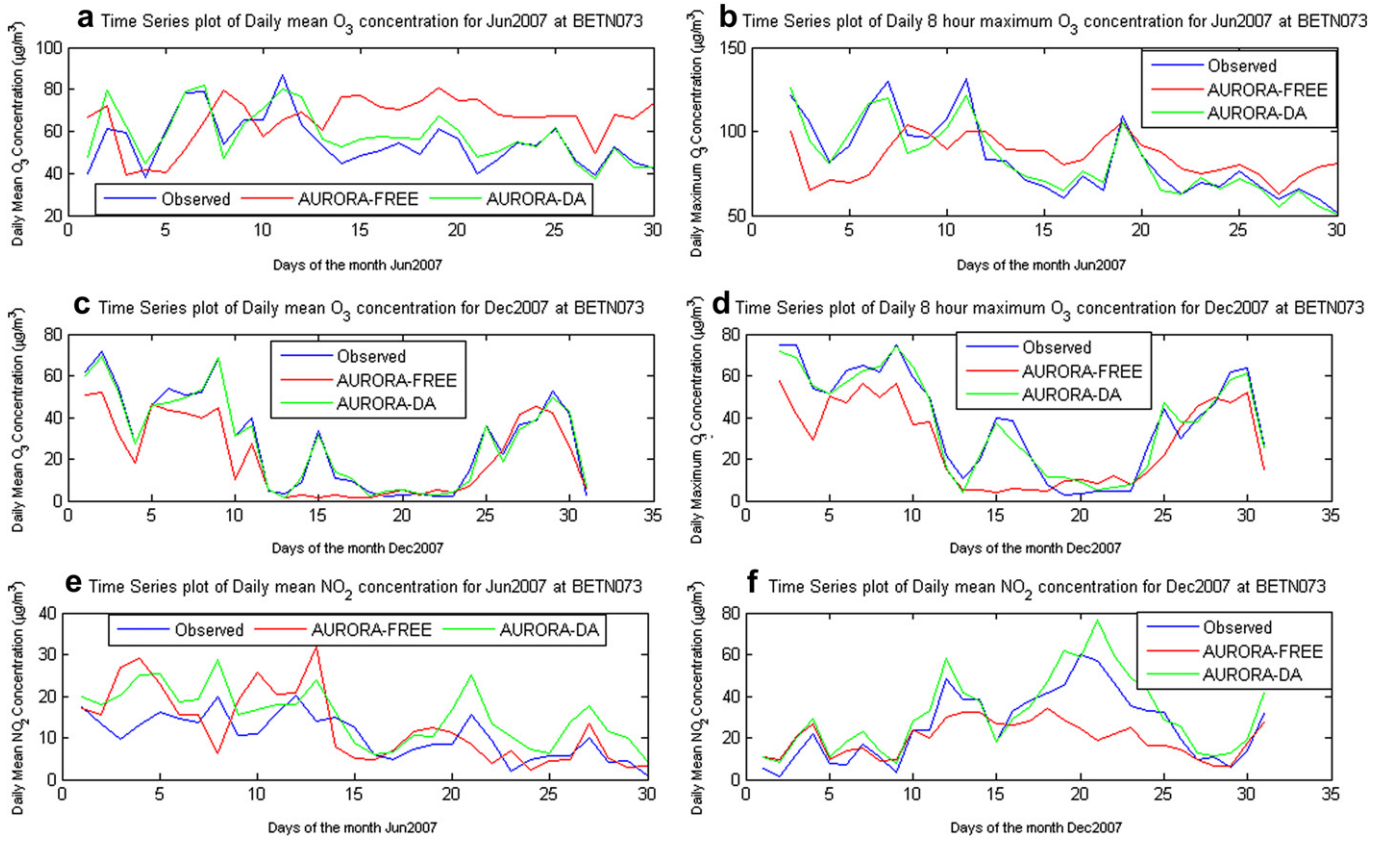

Fig. 5. Time-series plots at a validation station for the observed values, AURORA-FREE and AURORA-DA for (a) daily mean O_3 concentrations for Jun-2007, (b) daily 8-h maximum O_3 concentrations for Jun-2007, (c) daily mean O_3 concentrations for Dec-2007, (b) daily 8-h maximum O_3 concentrations for Dec-2007, (e) daily mean NO_2 concentrations for Jun-2007, (a) daily mean NO_2 concentrations for Dec-2007.

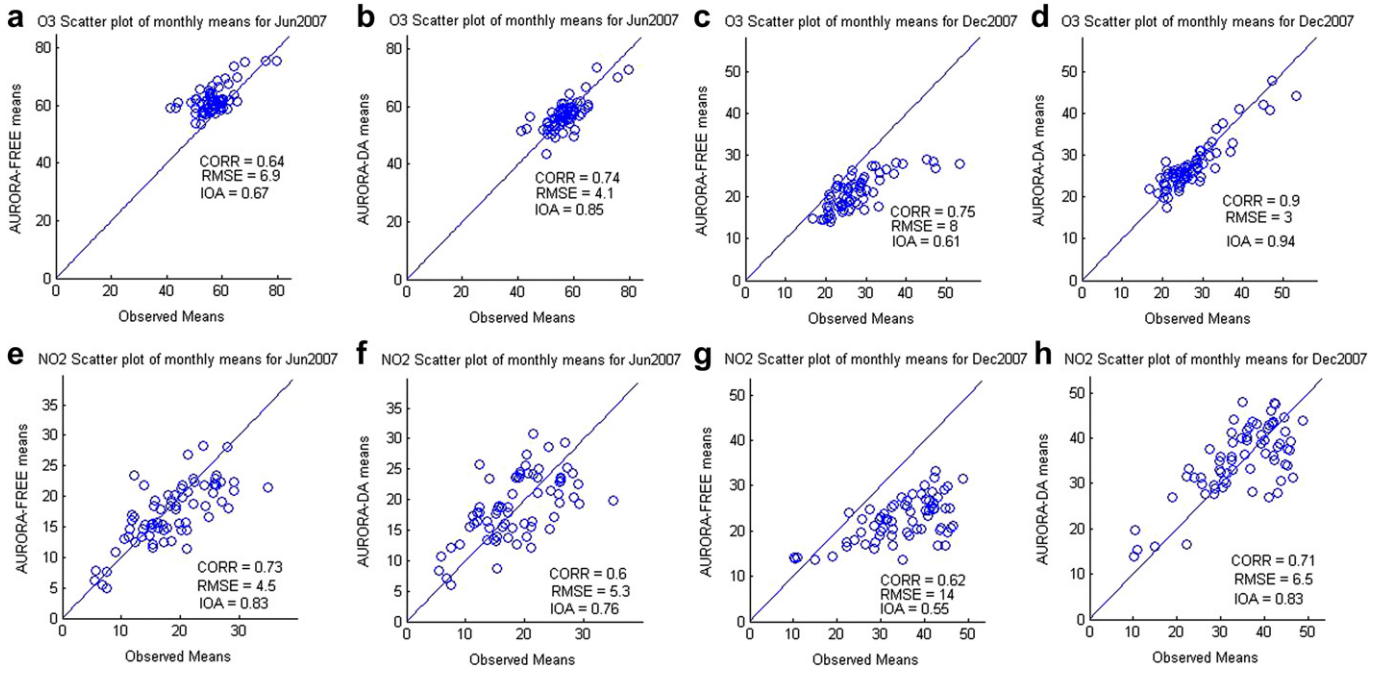

Fig. 6. Scatter plots between the monthly means of observed and (a) AURORA-FREE O_3 concentrations for Jun-2007, (b) AURORA-DA O_3 concentrations for Jun-2007, (c) AURORA-FREE O_3 concentrations for Dec-2007, (d) AURORA-DA O_3 concentrations for Dec-2007, (e) AURORA-FREE NO_2 concentrations for Jun-2007, (f) AURORA-DA NO_2 concentrations for Jun-2007, (g) AURORA-FREE NO_2 concentrations for Dec-2007, (h) AURORA-DA NO_2 concentrations for Jun-2007. CORR, IOA and RMSE have also been shown associated with each scatter plot.

observations (blue-line) better than the AURORA-FREE (red line) in both summer (Jun-2007) and winter (Dec-2007). Fig. 5(e,f) shows the daily mean NO_2 concentrations at the observation station BETN073 for Jun-2007 and Dec-2007. For both months, AURORA-DA (green line) follows the actual observations (blue line) more closely than that of AURORA-FREE (red line).

The monthly means for each of the 70 validation stations were calculated for AURORA-DA and were compared with the actual observations. The monthly means from AURORA-FREE runs were also calculated and compared with the actual observations. These validation results are shown as scatter plots in Fig. 6. A comparison of Fig. 6(a) and Fig. 6(b) clearly reveals that for the month of Jun-2007, AURORA-DA has performed significantly better than AURORA-FREE for O_3 concentrations as CORR has improved from 0.64 to 0.74, RMSE has decreased from 6.9 to 4.1 and IOA has increased from 0.68 to 0.87. Similar better performance of AURORA-DA for O_3 concentrations is noted for the month of Dec-2007 [Fig. 6(c) and Fig. 6(d)]. For NO_2, AURORA-DA has yielded almost the same results as that of AURORA-FREE for the month of Jun-2007 [Fig. 6(e) and Fig. 6(f)]. However, the AURORA-DA performed considerably better in the winter (Dec-2007) [Fig. 6(g) and Fig. 6(h)]. Thus, it has clearly been seen that AURORA-DA has improved the results on the spatial scale as well.

Next we present the maps of the monthly mean of daily mean and daily 8-h maximum O_3 concentrations over the study domain with and without data-assimilation. Fig. 7(a) and Fig. 7(b) present the monthly mean (Jun-2007) O_3 concentrations for the free run (AURORA-FREE) and AURORA with data assimilation (AURORA-DA), respectively. Fig. 7(c) and Fig. 7(d) represent the monthly mean of daily maximum O_3 concentrations for the month of Jun-2007 for the AURORA-FREE and AURORA-DA, respectively. The comparison shows how the data-assimilation corrects for the overestimated O_3 concentration in and around Belgium especially in the urban areas. Fig. 7(e) and Fig. 7(f) present the monthly mean of daily mean O_3 concentrations for Dec-2007 for AURORA-FREE and AURORA-DA, respectively while Fig. 7(g) and Fig. 7(h) depict the monthly mean of daily maximum O_3 concentrations for Dec-2007 in case of AURORA-FREE and AURORA-DA, respectively. The comparison of AURORA-FREE [Fig. 7(e) and Fig. 7(g)] and AURORA-DA results [Fig. 7(f) and Fig. 7(h)] clearly reveal that the data-assimilation has improved the underestimation of O_3 concentrations by AURORA-FREE in both the cases. Fig. 8(a) and Fig. 8(b) show the monthly mean NO_2 concentrations for AURORA-FREE and AURORA-DA, respectively. A comparison of Fig. 8(a) and 8(b) shows that the underestimation of NO_2 is corrected by the data-assimilation. Fig. 8(c) and 8(d) presents the monthly mean NO_2 concentrations

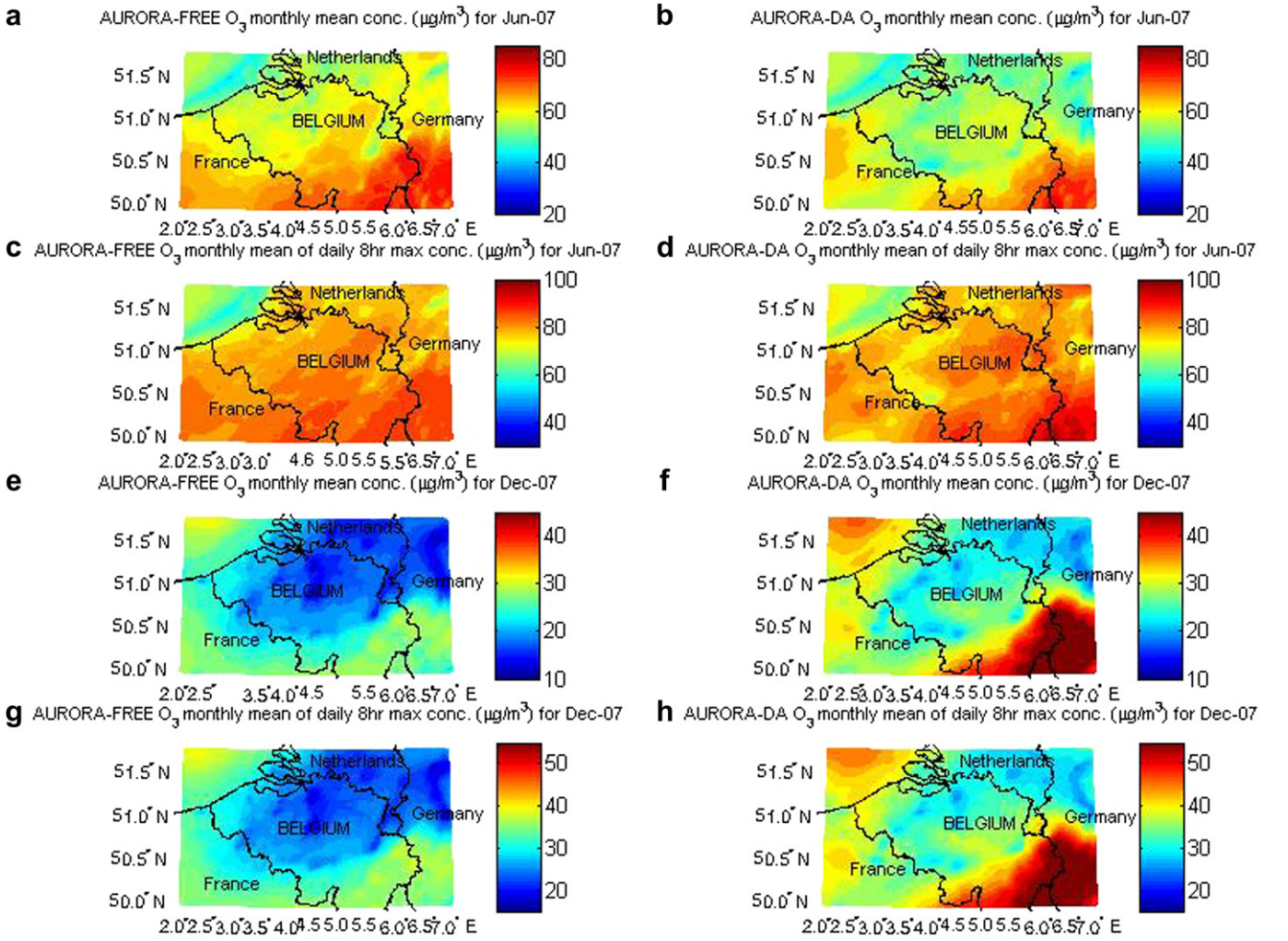

Fig. 7. Maps for the monthly mean concentrations (in μg m^{-3}) of (a) AURORA-FREE O_3 for Jun-2007, (a) AURORA-DA O_3 for Jun-2007, (c) AURORA-FREE daily maximum O_3 for Jun-2007, (d) AURORA-DA daily maximum O_3 for Jun-2007, (e) AURORA-FREE O_3 for Dec-2007, (f) AURORA-DA O_3 for Dec-2007, (g) AURORA-FREE daily maximum O_3 for Dec-2007, (h) AURORA-DA daily maximum O_3 for Dec-2007.

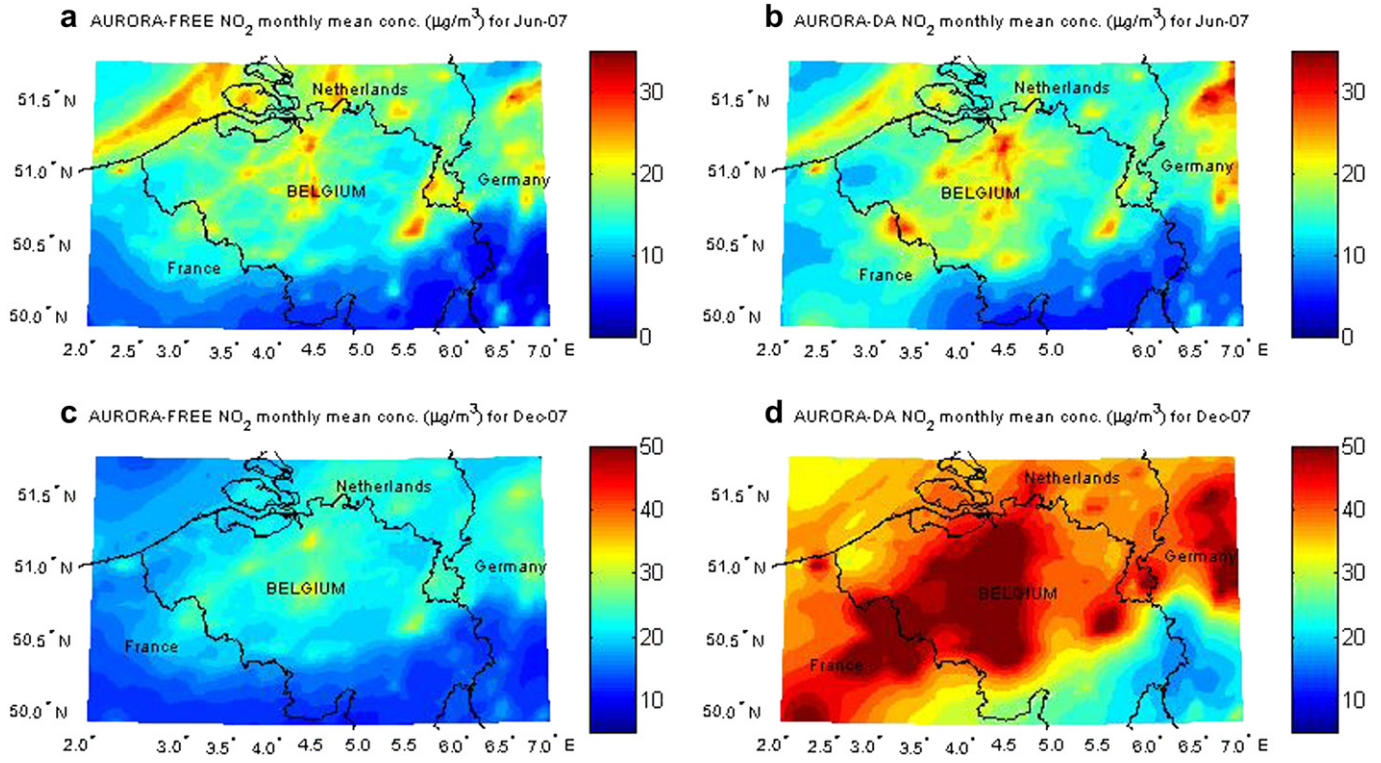

Fig. 8. Maps for the monthly mean concentrations (in μg m^{-3}) of (a) AURORA-FREE NO_2 for Jun-2007, (a) AURORA-DA NO_2 for Jun-2007, (c) AURORA-FREE NO_2 for Jun-2007, (d) AURORA-DA NO_2 for Dec-2007.

for Dec-2007 for AURORA-FREE and AURORA-DA, respectively. The comparison of Fig. 8(c) and Fig. 8(d) clearly reveals how the underestimation of NO_2 concentrations have been corrected by data-assimilation in the month of Dec-2007 also.

4. Conclusion

In the present study, a bias-aware optimal Interpolation in conjunction with the Hollingsworth–Lönnberg method to estimate background error covariance has been applied to the outputs of the regional air quality model AURORA in retrospective mode. Cross-validation was used to evaluate the accuracy of the assimilated concentration fields. The results clearly reveal that this simple scheme works well in the context of air quality modelling, substantially improving the retrospective simulation results. Also, the method is computationally very cheap compared to advanced ensemble or variational techniques, and can be directly applied in post-processing mode. Earlier studies such as Wu et al. (2008) and Denby et al. (2009) also found somewhat better results for O_3 and PM_{10} when using the simpler schemes such as optimal interpolation (in conjunction with Balgovind function (Balgovind et al., 1983)) and statistical interpolation (residual kriging), as compared to advanced ensemble techniques such as EnKF. However, the focus of the present study is on the retrospective simulation of small regional scale air quality modelling and the study has clearly shown the great potential of the applied scheme (i.e., a bias aware optimal interpolation with Hollingsworth–Lönnberg method) in data-assimilation of air quality models on such scale.

Acknowledgement

This work was carried out with support of the European Commission (LIFE+ project ATMOSYS and FP7 project PASODOBLE).

References

Balgovind, R., Dalcher, A., Ghil, M., Kalnay, E., 1983. A stochastic-dynamic model for the spatial structure of forecast error statistics. Monthly Weather Review 111 (4), 701–722.

Borrego, C., Monteiro, A., Ferreira, J., Miranda, A.I., Costa, A.M., Carvalho, A.C., Lopes, M., 2008. Procedures for estimation of modelling uncertainty in air quality assessment. Environment International 34, 613–620.

Boutahar, J., Lacour, S., Mallet, V., Quélo, D., Roustan, Y., Sportisse, B., 2004. Development and validation of a fully modular platform for numerical modelling of air pollution: POLAIR. International Journal of Environment and Pollution 22, 17–28.

Bouttier, F., Courtier, P., 2002. Data assimilation concepts and methods. Meteorological Training Course Lecture Series. ECMWF. http://www.ecmwf.int/newsevents/training/lecture_notes/pdf_files/ASSIM/Ass_cons.pdf.

Constantinescu, E.M., Sandu, A., Chai, T., Carmichael, G.R., 2007. Assessment of ensemble-based chemical data assimilation in an idealized setting. Atmospheric Environment 41, 18–36.

Daley, R., 1996. Atmospheric Data Analysis. Cambridge University Press.

Dee, D.P., da Silva, A.M., 1998. Data assimilation in the presence of forecast bias. Quarterly Journal of the Royal Meteorological Society 124, 269–295.

Dee, D.P., Todling, R., 2000. Data assimilation in the presence of forecast bias: the GEOS Moisture analysis. Monthly Weather Review 128, 3268–3282.

De Ridder, Mensink, C., 2002. Improved algorithms for advection and vertical diffusion in AURORA. In: Borrego, Schayes (Eds.), Air Pollution Modeling and Its Application XV. Kluwer Academic/Plenum Publishers, New York, pp. 395–401.

De Ridder, K., Lefebre, F., Bañuelos, A., Pérez-Lacorzana, J.M., Dufek, J., Adamec, V., Damsgaard, O., Thierry, A., Bruse, M., Bürger, M., Weber, C., Hirsch, J., 2004. An integrated methodology to assess the benefits of urban green space. The Science of the Total Environment 334-335, 489–497.

De Ridder, K., Lefebre, F., Adriaensen, S., Arnold, U., Beckroege, W., Bronner, C., Damsgaard, O., Dostal, I., Dufek, J., Hirsch, J., IntPanis, L., Kotek, Z., Ramadier, T., Thierry, A., Vermoote, S., Wania, A., Weber, C., 2008a. Simulating the impact of urban sprawl on air quality and population exposure in the German Ruhr area. Part I: reproducing the base state. Atmospheric Environment 42, 7059–7069.

De Ridder, K., Lefebre, F., Adriaensen, S., Arnold, U., Beckroege, W., Bronner, C., Damsgaard, O., Dostal, I., Dufek, J., Hirsch, J., IntPanis, L., Kotek, Z., Ramadier, T., Thierry, A., Vermoote, S., Wania, A., Weber, C., 2008b. Simulating the impact of urban sprawl on air quality and population exposure in the German Ruhr area. Part II: development and evaluation of an urban growth scenario. Atmospheric Environment 42, 7070–7077.

Denby, B., Schaap, M., Segers, A., Builtjes, P., Horalek, J., 2009. Comparison of two data assimilation methods for assessing PM10 exceedances on the European scale. Atmospheric Environment 42, 7122–7134.

Deutsch, F., Janssen, L., Vankerkom, J., Lefebre, F., Mensink, C., Fierens, F., Dumont, G., Roekens, E., 2008b. Modeling changes of aerosol compositions over Belgium and Europe. International Journal of Environment and Pollution 32, 162–173.

Deutsch, F., Mensink, C., Vankerkom, J., Janssen, L., 2008a. Application and validation of a comprehensive model for PM_{10} and $PM_{2.5}$ concentrations in Belgium and Europe. Applied Mathematical Modeling 32, 1501–1510.

Frydendall, J., Brandt, J., Christensen, J.H., 2009. Implementation and testing of a simple data assimilation algorithm in the regional air pollution forecast model, DEOM. Atmospheric Chemistry and Physics 9, 5475–5488.

Gery, M.W., Whitten, G.Z., Killius, J.P., Dodge, M.C., 1989. A photochemical kinetics mechanism for urban and regional scale computer modeling. Journal of Geophysical Research 94, 925–956.

Hollingsworth, A., Lönnberg, P., 1986. The statistical structure of short-range forecast errors as determined from radiosonde data. Part I: the wind field. Tellus 38A, 111–136.

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

Lefebre, F., De Ridder, K., Lewyckyj, N., Janssen, L., Cornelis, J., Geyskens, F., Mensink, C., 2004. In: Borrego, C., Incecik, S. (Eds.), Air Pollution Modeling and Its Applications XVI. Kluwer Academic/Plenum Publishers, New York, pp. 511–519.

Lefebvre, W., Fierens, F., Trimpeneers, E., Janssen, S., Van de Vel, K., Deutsch, F., Viaene, P., Vankerkom, J., Dumont, G., Vanpoucke, C., Mensink, C., Peelaerts, W., Vliegen, J., 2011. Modeling the effects of a speed limit reduction on traffic-related elemental carbon (EC) concentrations and population exposure to EC. Atmospheric Environment 45, 197–207. http://dx.doi.org/10.1016/j.atmosenv.2010.09.026.

Lonnberg, P., Hollingsworth, A., 1986. The statistical structure of short-range forecast errors as determined from radiosonde data. Part II: the covariance of height and wind errors. Tellus A 38, 137–161.

Maes, J., Vliegen, J., Van de Vel, K., Janssen, S., Deutsch, F., De Ridder, K., Mensink, C., 2009. Spatial surrogates for the disaggregation of CORINAIR emission inventories. Atmospheric Environment 43, 1246–1254.

Mensink, C., De Ridder, K., Lewyckyj, N., Delobbe, L., Janssen, L., Van Haver, Ph, 2001. Computational aspects of air quality modelling in urban regions using an optimal resolution approach. In: Margenov, S., Wasniewski, J., Yamalov, P. (Eds.), 2001. Large-scale Scientific Computing, Lecture Notes in Computer Science, vol. 2179, pp. 299–308.

Mensink, C., De Ridder, K., Deutsch, F., Lefebre, F., Van de Vel, K., 2008. Examples of scale interactions in local, urban, and regional air quality modelling. Atmospheric Research 89, 351–357.

Mensink, C., De Ridder, K., Lewyckyj, N., Lefebre, F., Janssen, L., Cornelis, J., Adriaensen, S., Ruts, M., 2002. AURORA: an air quality model for urban regions using an optimal resolution approach. In: Ninth International Conference on the Modelling, Monitoring and Management of Environmental Problems, ENVIROSOFT 2002; Bergen; 6 May 2002 through 8 May 2002; Code 63549.

Mensink, C., De Vlieger, I., Nys, J., 2000. An urban transport emission model for the Antwerp area. Atmospheric Environment 34, 4595–4602.

Tilmes, S., 2001. Quantitative estimation of surface ozone observation and forecast errors. Physics and Chemistry of the Earth, Part B: Hydrology, Oceans and Atmosphere 26 (10), 759–762.

Tombette, M., Mallet, V., Sportisse1, B., 2009. PM10 data assimilation over Europe with the optimal interpolation method. Atmospheric Chemistry and Physics 9, 57–70.

Vankerkom, J., De Vlieger, I., Schrooten, L., Vliegen, J., Styns, K., 2009. Beleidsondersteunend onderzoek: Aanpassingen aan het emissiemodel voor wegtransport MIMOSA. Studie uitgevoerd in opdracht van VMM – MIRA, 2009/TEM/R/084.

Walcek, C.J., 2000. Minor flux adjustment near mixing ratio extremes for simplified yet highly accurate monotonic calculation of tracer advection. Journal of Geophysical Research 105, 9335–9348.

Wilks, D.S., 2006. Statistical Methods in Atmospheric Sciences. Elsevier Inc., pp. 49–58.

Wu, L., Mallet, V., Bocquet, M., Sportisse, B., 2008. A comparison study of data assimilation algorithms for ozone forecasts. Journal of Geophysical Research 113, D20310. http://dx.doi.org/10.1029/2008JD009991.

Xue, M., Droegemeier, K.K., Wong, V., 2000. The Advanced Regional Prediction System (ARPS) – a multiscale non-hydrostatic atmospheric simulation and prediction tool. Part I: model dynamics and verification. Meteorology and Atmospheric Physics 75, 161–193.

Xue, M., Droegemeier, K.K., Wong, V., Shapiro, A., Brewster, K., Carr, F., Weber, D., Liu, Y., Wang, D.-H., 2001. The Advanced Regional Prediction System (ARPS) – A multiscale non-hydrostatic atmospheric simulation and prediction tool. Part II: model physics and applications. Meteor. Atmos. Physics 76, 134–165.