

# Implementation and testing of a simple data assimilation algorithm in the regional air pollution forecast model, DEOM

J. Frydendall<sup>1</sup>, J. Brandt<sup>2</sup>, and J. H. Christensen<sup>2</sup>

<sup>1</sup>DTU Informatic, Technical University of Denmark, 2800 Kgs. Lyngby, Denmark

<sup>2</sup>National Environmental Research Institute, Aarhus University, 4000 Roskilde, Denmark

Received: 16 January 2009 – Published in Atmos. Chem. Phys. Discuss.: 23 March 2009

Revised: 16 June 2009 – Accepted: 17 June 2009 – Published: 4 August 2009

**Abstract.** A simple data assimilation algorithm based on statistical interpolation has been developed and coupled to a long-range chemistry transport model, the Danish Eulerian Operational Model (DEOM), applied for air pollution forecasting at the National Environmental Research Institute (NERI), Denmark. In this paper, the algorithm and the results from experiments designed to find the optimal setup of the algorithm are described. The algorithm has been developed and optimized via eight different experiments where the results from different model setups have been tested against measurements from the EMEP (European Monitoring and Evaluation Programme) network covering a half-year period, April–September 1999. The best performing setup of the data assimilation algorithm for surface ozone concentrations has been found, including the combination of determining the covariances using the Hollingsworth method, varying the correlation length according to the number of adjacent observation stations and applying the assimilation routine at three successive hours during the morning. Improvements in the correlation coefficient in the range of 0.1 to 0.21 between the results from the reference and the optimal configuration of the data assimilation algorithm, were found. The data assimilation algorithm will in the future be used in the operational THOR integrated air pollution forecast system, which includes the DEOM.

throughout the world. The chemical weather can be seen as analogous to the meteorological weather. In particular, chemical weather emphasizes the strong influence of meteorological variability – and the chemical response to this variability – on air quality (Lawrence et al., 2005). In contrast to numerical weather forecasting, it is technically possible to carry out operational chemical weather forecasting without using data assimilation of the prognostic variables in the air pollution model. Without data assimilation of meteorological parameters during initialization, numerical weather forecast models would produce simulations that – even though the results would appear realistic – have nothing to do with the actual weather. A long-range chemistry-transport model (CTM) used for operational forecasting is driven by a numerical weather forecast model, and is bound by a emissions inventory as well as chemical lifetimes of the individual species. In this way the results from a chemical weather forecast will show good performance when compared against measurements. However, applying the data assimilation techniques which have been used by the weather forecasting community for since the late 1950s (Gandin, 1963), has the potential to make significant improvements in chemical weather forecasts and these techniques are now being introduced in air pollution models by various scientific communities.

At the National Center for Atmospheric Research, Atmospheric Chemistry Division, Boulder, USA an Optimum Interpolation routine (Lamarque et al., 1999) is being used to investigate CO in the troposphere. A group at the Data Assimilation Office, National Aeronautics & Space Administration (NASA) Goddard Space Flight Center USA, has used a Kalman Filter (M'énard, 2000; M'énard et al., 2000) to investigate chemical tracers. In Europe there are several groups working with chemical data assimilation: at the University of Cologne, Germany, a four-dimensional variational algorithm for atmospheric chemistry modelling has been developed and used in the EUROpean Air Pollution Dispersion (EURAD)

## 1 Introduction

Even though the field of chemical weather forecasting is still very much in the research and development phase, operational forecasting of the air pollution concentration is now being carried out on a routine basis in many countries



Correspondence to: J. Frydendall  
(jf@imm.dtu.dk)

model (Elbern et al., 1997; Elbern, 1999; Elbern et al., 2000). At the Delft University of Technology, Netherlands, a Kalman Filter has been developed (van Loon and Heemink, 1997) for atmospheric chemistry modelling. At the French meteorology laboratory, an Optimum Interpolation routine for ozone analysis has been developed (Blond et al., 2003; Blond and Vautard, 2004). At Norwegian Institute for Air Research, (NILU), a number of statistical interpolation methods are being employed for  $PM_{10}$  and applied in the Unified EMEP model (Denbyet al., 2008). In the Swedish Meteorological and Hydrological Institute (SMHI), an operational 2D-var method is under development for operation in the Multiscale Atmospheric Transport and Chemistry Model (MATCH) model (Denbyet al., 2008). At NERI, Denmark, besides of the work described in this paper, development and tests of a four-dimensional variational method have been made, see Zlatev and Brandt (2005, 2007).

Chemical transport models are valuable tools to understand the transport of chemical pollutants in the atmosphere. However, due to uncertainties e.g. caused by discretization of the governing equation, uncertainties in the simplified chemical reaction scheme or physical parameterizations, or erroneous emissions, the CTMs cannot truly represent the real world. On the other hand uncertainties in the measurements makes the comparison between the CTMs and the observations a non-trivial business. Data assimilation routines combines the information from the CTMs and the measurements by taking into account the model and observation uncertainties to make better representation of the air pollution fields.

A general problem in chemical data assimilation is, however, the lack of real-time data. In the meteorological community, a dense network for real-time meteorological measurements, both at the surface as well as radio soundings, was established many years ago. With respect to chemical data assimilation, the research groups typically have to collect the available sparse data sets on their own. However, more and more real-time surface observations are becoming available for assimilation, and even satellite measurements of e.g. the tropospheric column of  $NO_2$  can be obtained. Potentially, these data sets together provide relative high accuracy from the surface measurements combined with the greater spatial coverage from the satellite data. However, none of them provides an estimate of the vertical distribution of the chemical species.

An alternative to the direct use of data assimilation is post-processing approaches. In these postprocessing approaches, a moving training window is assigned to a fixed number of days where the model uncertainties are estimated from error residuals between model forecasts and observations. The model uncertainties estimates are used as bias corrections in the future forecast window. A nice example of such a postprocessing techniques is demonstrated in (Kang et al., 2008).

Data assimilation techniques applied in chemistry-transport models cannot only be used for operational forecasting of the chemical weather but also for generating ana-

lyzed fields covering a large time period of the different air pollution species e.g. for monitoring the air quality and assessing the impacts from air pollution. Examples could be integrated monitoring (using both models and measurements, see e.g. Hertel et al., 2007) of nitrogen species with respect to eutrophication in the marine and terrestrial ecosystems or integrated monitoring of ozone, nitrogen-oxides and particulate matter with respect to the impacts on human health.

## 2 The DEOM model

The long-range chemical transport model, the Danish Eulerian Operational Model (DEOM) (Brandt et al., 2000; Brandt et al., 2001a; Brandt et al., 2001b,c) has been developed at NERI for air quality forecasting. The model includes emissions, atmospheric transport and dispersion, chemical transformations and dry and wet depositions of 35 chemical species. The domain of the DEOM covers Europe and is constructed so that it is covered by the domain of the meteorological model, Eta, applied for operational weather forecasting at NERI and used as a driver for the DEOM. The Eta model is discretized on a staggered latitude/longitude system with shifted pole. The horizontal grid resolution is  $0.25^\circ \times 0.25^\circ$  corresponding to approximately  $39 \text{ km} \times 39 \text{ km}$  at  $60^\circ \text{ N}$ . The number of horizontal grid points is  $104 \times 175$  and the number of vertical layers is 32. The DEOM model is applied on a polar stereographic projection. The horizontal grid resolution is  $50 \text{ km} \times 50 \text{ km}$  at  $60^\circ \text{ N}$ . The number of grid points is  $96 \times 96$ . Three vertical layers are used in the DEOM model. The three layers are defined as a mixed layer (below the mixing height), a smog or reservoir layer between the mixing height and the advected mixing height from the previous day. The top layer is located between the advected mixing height and the free troposphere. The model has been a part of various inter-comparison studies and has shown comparable results with similar models, see e.g. Tilmes et al. (2002).

A splitting procedure, based on the ideas of McRae et al. (1982), is applied in the DEOM. The horizontal transport is discretized using an accurate space derivative algorithm. Time integration is performed with a predictor-corrector scheme with several correctors. For the horizontal dispersion, truncated Fourier series approximate the concentrations. Dry and wet depositions are computed directly using simple parameterizations. The chemical scheme used in the model is the CBM-IV scheme with 35 species. Chemistry is solved using the QSSA method (Hesstvedt et al., 1978).

The DEOM model is a part of the THOR integrated model system, Brandt et al. (2001a); Brandt et al. (2001b,c, 2005), capable of performing forecasting of meteorological and chemical weather for the general public as well as assessment and management for decision-makers in general. The system consists of several meteorological and air pollution models, developed at NERI over recent decades, and is capable of operating for different applications and at different

scales. Global meteorological data from NCEP are used as initial and boundary conditions for the numerical weather forecast model Eta. The weather data from this model are used to drive the air pollution models: the Danish Eulerian Operational Model (DEOM), the Urban Background Model (UBM), the Operational Street Pollution Model (OSPM) and others. Air pollution data from the DEOM is used as input to the UBM and the results from this model is used as input to the OSPM, see Brandt et al. (2001c).

Coupling models over different scales makes it possible to account for contributions from local, near-local as well as remote emission sources in order to describe the air quality at a specific location – e.g. in a street canyon or in a sub-urban area. The system provides high-resolution three-day forecasting of weather and air pollution, from regional scale over urban background scale and down to individual street canyons in cities – on both sides of the streets. The whole system is run operationally and automatically four times every day, initiated at 00:00 UTC, 06:00 UTC, 12:00 UTC and 18:00 UTC. The system is also applied in connection with the urban and rural monitoring programs in Denmark where the model results and measurements are used together via integrated monitoring to obtain the best available information level for the atmospheric environment and effects. It is planned that the data assimilation routine developed in this study is to be used as a basis for improvements in the air quality forecast at regional scale, which will also affect the results on urban scales.

### 3 The data assimilation algorithm

The data assimilation algorithm in this article is based on a Statistical Interpolation algorithm. The notation used is similar to the notation introduced by Ide et al. (1997). The observations  $y_o$  represent a measure of the real world. The data assimilation algorithm introduces this knowledge into the model and the combination of the model state  $x_b$  and observation state  $y_o$  is called the analysis state  $x_a$  which in theory should be a better representation of the real world than the background state or the observation state individually. The analysis state is obtained by weighting the model errors against the observation errors. This leads to the interpolation equation (Bouttier and Courtier, 1998):

$$\mathbf{x}_a = \mathbf{x}_b + \mathbf{K}(\mathbf{y}_o - \mathbf{H}(\mathbf{x}_b)) \quad (1)$$

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

where the linear operator  $\mathbf{K}$  is called the Kalman gain and is the weight matrix of the analysis.  $\mathbf{H}$  denotes the linear map between model space and observation space.

### 4 The background error covariance matrices

Three different background error covariance matrices  $\mathbf{B}=(\mathbf{B}_1, \mathbf{B}_2, \mathbf{B}_3)$  will be tested and compared to each other. It is assumed that the horizontal correlation is homogeneous and isotropic for the two first background error covariance matrices. For the last background error covariance matrix it is only assumed that the horizontal correlation is homogeneous. The first background error covariance matrix  $\mathbf{B}_1$  is the well known scaled Balgovind function (Balgovind et al., 1983),

$$C_0(\mathbf{r}) = \left(1 + \frac{|\mathbf{r}|}{L}\right) \exp\left(-\frac{|\mathbf{r}|}{L}\right) \quad (3)$$

where  $\mathbf{r}$  is Euclidean distance between the grid cell locations and  $L$  is the correlation length. For a thorough review on the properties of the Balgovind function and other correlation functions, see Gaspari and Cohn (1999).

The second background error covariance matrix  $\mathbf{B}_2$  is defined by Hoelzemann et al. (2001). The background error covariance matrix takes into account that adjacent observation stations can deteriorate the analysis field. The function is defined as follows: let  $\delta$  be the number of observation stations neighboring a model grid point, for which the radius of influence has to be estimated. The more adjacent the observation station is to the model grid point, the smaller the radius of influence. Taking 20% as the lower limit, the new  $L$  becomes

$$\tilde{L}(\delta) = \left(1 - \frac{\delta}{10}\right) L \quad (4)$$

where  $0 \leq \delta \leq 8$ . The new correlation matrix becomes

$$C_0(\mathbf{r}) = \left(1 + \frac{|\mathbf{r}|}{\tilde{L}}\right) \exp\left(-\frac{|\mathbf{r}|}{\tilde{L}}\right) \quad (5)$$

Finally the last background error covariance matrix  $\mathbf{B}_3$  takes into account that the observation spreading done by the background error covariance matrix should depend on the wind direction and the wind speed. With this approach the assumption on horizontal isotropic characteristic is abandoned in order to get a more realistic correlation function. The correlation length is decomposed into two correlation lengths: One that is parallel with the wind direction and one that is perpendicular to the wind direction, that is  $L \rightarrow L_{\parallel} + L_{\perp}$ . The isotropic correlation function can be interpreted as correlation circle in a 2-D system where the correlation length is the radius of the correlation circle. In the anisotropic case, the correlation circle will be transformed into a correlation ellipse with the major and minor axis given as a function of the correlation lengths in the wind directions.

Given the wind  $V=V(u, v)$ , we can calculate the rotations matrix and the transformation of  $(x, y)$ :

$$\begin{pmatrix} x' \\ y' \end{pmatrix} = \begin{pmatrix} x \cos \varphi + y \sin \varphi \\ -x \sin \varphi + y \cos \varphi \end{pmatrix} \quad (6)$$

where  $\varphi=v/u$ . Hence  $x \rightarrow \frac{x'}{L_{\parallel}}$  and  $y \rightarrow \frac{y'}{L_{\perp}}$ . The magnitude of the correlation lengths can be determined in several ways. In this case we let the magnitude be a function of the ratio between the wind components:

$$\begin{aligned} \frac{L_{\perp}}{L_{\parallel}} &= v/u \\ L_{\parallel} &= L \end{aligned} \quad (7)$$

Then the background correlation function becomes

$$C_0(r') = (1 + r') \exp(-r') \quad (8)$$

where  $r'=(x'^2 + y'^2)^{1/2}$ . The adjacent function (4) can, of course, be combined with (8).

#### 4.1 Covariance determination

An essential task in data assimilation is the estimation of the error parameters of the model and the observations. The approach chosen in this paper is called the Hollingsworth method (Hollingsworth and Lönnberg, 1986; Lönnberg and Hollingsworth, 1986; Daley, 1996). The idea is to look at the auto correlation function of the residuals between the model forecast and the observations. The sample correlations among all pairs of stations can be plotted as a function of separation distances, together with a curve representing a fitted auto correlation model, cf. Eq. (3). By extrapolation of the curve to the origin, the ratio between the observation and forecast error standard deviations can be determined. Another commonly used approach, which is used for estimating parameters in a very large state space model is based on the Ensemble Kalman Filter (Evensen, 1994; Burges et al., 1998), and this approach will be tested in the future.

In the determination of the background and observation error covariances, another problem becomes clear. There are on average only 90 observation stations operating for ozone in the EMEP network in a typical hour. However, it is not the same 90 stations are operating all the time – the location of the measured data is changing. On average the distribution of the stations is not centered around a specific area. If the stations were mainly located around a specific area this could mean that the interpolation operator  $\mathbf{H}$  would be sparse with only a few numbers different from zero grouped together. This would give problems when we want to create  $\mathbf{H}\mathbf{B}\mathbf{H}^T$  because the background error covariances matrix should be positive definite. This could result in a singular matrix and the data assimilation analysis would not be feasible. In order to avoid this problem we decide to let all the observation stations go in to  $\mathbf{H}$  and let missing measurements be controlled by the departures  $\mathbf{d}=\mathbf{y}_o-\mathbf{H}\mathbf{x}_a$ . The value zero is assigned to the missing measurements. In the final construction when  $\mathbf{d}$  is multiplied by the Kalman gain matrix  $\mathbf{K}$ , the zero value from the missing measurement would cancel the contribution to  $\mathbf{x}_a$ .

For estimating the background error covariance using the Hollingsworth method, a period of 6 months (April–September 1999) was used as a study period, and both measurements and model results were available for ozone. The departures at each observation station were calculated at 4 p.m. every day when the air pollution was well mixed. Furthermore, the maximum values of ozone are typically observed during the afternoon. From this departure the correlation with all the other departures was plotted as a function of their separation. The results can be seen in Fig. (1).

From Fig. (1) we want to fit the correlation function (3) with the data obtained from the six-months correlation study. In Fig. (1) the curve represents the correlation function. Now we are able to determine the background error covariances  $\sigma_b^2$  and the observation error covariances  $\sigma_o^2$ . The background error covariance can be determined from the interception with y-axes. From the interception we get the correlation  $q_b=\frac{\sigma_b^2}{\sigma_b^2+\sigma_o^2}$  and the observation error covariance can then be determined from the simple relation  $q_o=1-q_b$ . The final parameter that can be determined from the auto correlation function is the correlation length,  $L$ . As already discussed in the previous section, the correlation length is the distance at which two independent observation stations can be correlated in the model. Beyond that distance the stations will not be correlated in the model. In this study, we found the following parameters for surface ozone  $q_b=0.86$ ,  $q_o=0.14$  and  $L=270$  km.

The estimated covariances from the analysis will vary over the seasons and over the local regions i.e. Southern and Northern Europe. The ideal correlation function should be adapted to the fit the local regions and be varied over different seasons. However, in this implementation the basic correlation function will only be tested to determine the effects of the error covariance in the data assimilation routine. Experiments with finding proper correlation functions have been carried out by (Houtekamer and Herschel, 2001), (Hamill et al., 2001) for the EnKF. Finding better error covariances is a investigation in itself and is beyond the scope of this paper.

## 5 The data assimilation experiments

In these experiments, the data assimilation algorithm is implemented into the DEOM and the effects of applying the algorithm with different configurations are tested against measurements. First, the DEOM was run for the test period from April to September 1999, to make a reference analysis. The summer period is chosen because there are more ozone episodes in the summer months, which is mainly due to warmer temperatures and much higher global radiation in these periods compared to winter periods. In the following tests the model results are compared to measurements and the improvements relative to the reference run without the data assimilation are examined. Improvements in both the

correlation and bias should be expected, since the discrepancies between the observations and the model results have been used to adjust the model results with a weight function. The test period was chosen because it was a well documented period with several ozone episodes and a relatively large temporal spatial coverage of the measurements from the EMEP network.

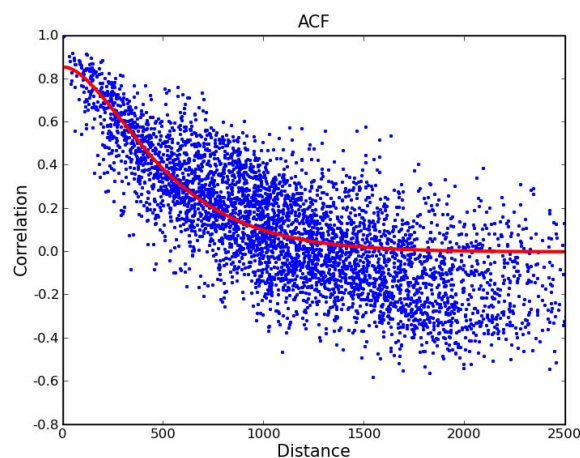
In this study the tests will be concentrated on the daily maximum values of ozone concentrations. The DEOM model usually performs well with respect to predicting the daily maximum values, which means that the background field from the DEOM model will be less erroneous, compared to other parameters. In this study, it is believed that the data assimilation will decrease the bias and increase the correlation and hence decrease the normalized mean square error, when compared to the measurements.

The measurement data from the EMEP ozone network includes 207 observation stations within the DEOM model grid. All the tests will be conducted over the entire period of 6 months. The data assimilation routine is activated once every day at 12:00 UTC, unless otherwise stated in the description of the tests. The analyzed model fields are compared to the same observation stations that are used in the data assimilation analysis, but at a different time. The comparison is made for the daily maximum ozone concentration, which usually takes place 4–6 h (at 16:00 UTC–18:00 UTC) later than when the assimilation procedure was conducted. This gives a separation in time between the assimilation time and the actual comparison time of 4–6 h.

Another way of evaluating the assimilation process could be to use only half of the observation stations in the data assimilation and use the other half as control/validation stations. This approach should give some information about the spatial separation that arises from the missing observation stations and the stations that are included in the analysis. When the analysis is compared to the observation stations that were excluded in the analysis, the improvement in the analysis field should be seen. However, the number of measurement stations is relatively small, and as mentioned above, the time separation between the observations used for assimilation and the observations used for validation for the daily ozone maximum should be large enough to avoid problems, since the ozone concentrations are transported and chemically produced in the model domain between the time of assimilation and time of validation.

Nine different model runs were performed with the data assimilation algorithm implemented in the DEOM, to carry out the eight experiments, besides a reference run. The model runs are:

1. Reference: the reference run of the DEOM model without the data assimilation routine activated.
2. Experiment 1: the assimilation algorithm conducted with correlation function (3) using equal weights i.e.



**Fig. 1.** The correlation function (3) is fitted to the departures correlation as a function of the distance between them in km.

$\sigma_b^2=1$ ,  $\sigma_o^2=1$  and  $L=3$  grid units (in this case corresponding to 150 km).

3. Experiment 2: run with optimal weights found by the Hollingsworth method.
4. Experiment 3: as experiment 2 with the assimilation routine activated three times a day, on 10:00 UTC, 11:00 UTC and 12:00 UTC.
5. Experiment 4: run with the anisotropic correlation function (8) with determined weights.
6. Experiment 5: as experiment 2 with the correlation function taking into account the density of observations by (4).
7. Experiment 6: combination of experiments 4 and 5 with both the anisotropic and the density of observations correlation function. The assimilation routine is activated once per day at 12:00 UTC.
8. Experiment 7: as experiment 6 with the assimilation routine activated three times a day, at 10:00 UTC, 11:00 UTC and 12:00 UTC.
9. Experiment 8: run with the correlation function with optimal weights as in experiment 2, adjusted with the formula as in experiment 5 and with the assimilation routine activated three times a day, on 10:00 UTC, 11:00 UTC and 12:00 UTC as in experiment 3.

For all the experiments described above, the model results of the daily maximum value of ozone was validated against measurements from EMEP and examined in the following three different ways (corresponding to average over space, no averaging and average over time, respectively):

1. Time series of the daily maximum value as mean over all stations, where all the observations and calculated values are averaged over space for every day and plotted as function of time.
2. Scatter plots of the daily maximum value including the observations and calculated daily maximum values for all times and locations.
3. Scatter plots of the mean of the daily maximum value at each station, where the observations and the calculated daily maximum values for all stations are averaged over the time period.

### 5.1 Statistical results from the experiments

In the following subsections, the DEOM model results from all the experiments combined with the different ways of averaging compared to measurements are given. The model results were compared to measurements, and statistics were calculated for every experiment. The statistics are the correlation coefficient, the student's *t*-test for significance of the correlation coefficient, the fractional bias, and the normalized mean square error.

The statistics for the whole period April–September 1999 for the daily maximum values of ozone from all experiments 1–9 and for the different averaging methods described above are presented in Tables 1, 2, and 3.

The great number of statistics from the different assimilation scenarios made a direct comparison difficult. Therefore a ranking system was used to determine the best performing configuration of the data assimilation setup. In the ranking system ranks were assigned as the number 1 for the experiment with the best statistic, 2 for the second best, and so on up to 8. If two statistics had the same value they were assigned the same rank, and the successive rank was skipped. Only the corresponding statistics were compared with each other. In the end the best performing assimilation setup could be determined from the ranking with the lowest total value. Results from the ranking can be seen in Table 4. The ranking was performed for each month, April to September, and one ranking for the entire period.

From Table 4 it is clear that the assimilation experiment 8 is the best performing including the combination of determining the covariances using the Hollingsworth method, varying the correlation length according to the number of adjacent observation stations and applying the assimilation routine at three successive hours. It can be seen that the correlation coefficient is improved by 0.21 and the students *t* test has gone up by 50.7. The fractional bias and normalized mean square error have decreased by  $1.8 \times 10^{-3}$  and  $1.7 \times 10^{-2}$ , respectively. Having a variable correlation length increases the correlation for stations that are adjacent. It can be seen from statistics from individual stations (not shown here) that the performance improved for these kind of stations.

**Table 1.** Statistics calculated for the reference model run and experiments 1–8 compared to measured data from EMEP, covering the period April–September 1999, based on the time series of the daily maximum value as mean over all stations, where all the observations and calculated values are averaged over space for every day and plotted as function of time.

Model run	Correlation coefficient	Students <i>t</i> test	Fractional bias	NMSE
Reference	0.86	22.9	5.2E-03	3.1E-03
Experiment 1	0.94	36.6	3.8E-03	1.4E-03
Experiment 2	0.94	38.6	1.8E-03	1.3E-03
Experiment 3	0.96	47.6	8.4E-03	1.0E-03
Experiment 4	0.93	34.7	9.8E-03	1.6E-03
Experiment 5	0.95	39.2	5.1E-03	1.3E-03
Experiment 6	0.94	36.7	7.3E-03	1.5E-03
Experiment 7	0.95	41.8	1.1E-02	1.2E-03
Experiment 8	0.96	45.1	3.4E-03	1.0E-03

**Table 2.** Statistics calculated for the reference model run and experiments 1–8 compared to measured data from EMEP, covering the period April–September 1999, based on the Scatter plots of the daily maximum value including the observations and calculated daily maximum values for all times and locations.

Model run	Correlation coefficient	Students <i>t</i> test	Fractional bias	NMSE
Reference	0.62	101.5	5.2E-03	4.5E-02
Experiment 1	0.72	131.4	3.8E-03	3.3E-02
Experiment 2	0.73	137.8	1.8E-03	3.2E-02
Experiment 3	0.76	149.8	8.5E-03	2.9E-02
Experiment 4	0.74	143.1	5.1E-03	3.0E-02
Experiment 5	0.71	129.8	9.7E-03	3.4E-02
Experiment 6	0.73	138.1	7.2E-03	3.1E-02
Experiment 7	0.75	143.7	1.1E-02	3.0E-02
Experiment 8	0.76	152.2	3.4E-03	2.8E-02

**Table 3.** Statistics calculated for the reference model run and experiments 1–8 compared to measured data from EMEP, covering the period April–September 1999, based on the Scatter plots of the mean of the daily maximum value at each station, where the observations and the calculated daily maximum values for all stations are averaged in time.

Model run	Correlation coefficient	Students <i>t</i> test	Fractional bias	NMSE
Reference	0.67	8.7	4.6E-03	9.1E-03
Experiment 1	0.74	10.6	4.4E-03	7.6E-03
Experiment 2	0.76	11.1	2.4E-03	7.3E-03
Experiment 3	0.78	12.0	9.0E-03	7.0E-03
Experiment 4	0.75	10.9	9.0E-03	7.5E-03
Experiment 5	0.78	11.9	4.5E-03	6.7E-03
Experiment 6	0.77	11.6	6.6E-03	7.0E-03
Experiment 7	0.80	12.8	1.1E-02	6.4E-03
Experiment 8	0.81	13.2	3.0E-03	3.0E-03

**Table 4.** The outcome from the ranking process of the combined ranks of correlation coefficient, fractional bias and the normalized mean square error including the results for daily maximum values averaged in time, not averaged and averaged in space for every month during the period April–September 1999 as well as for the whole period (all). The global rank with the value 1 is the best performing configuration of the model and the value 9 refers to the worst performing model configuration.

Model run	Apr	May	Jun	Jul	Aug	Sep	Apr–Sep	Total	Global rank
Reference	55	44	69	59	67	62	75	431	7
Experiment 1	53	48	29	52	33	28	50	293	5
Experiment 2	67	63	44	59	49	35	35	352	6
Experiment 3	30	40	27	41	37	9	36	220	2
Experiment 4	90	90	90	85	67	90	72	584	9
Experiment 5	38	28	45	28	29	36	32	236	3
Experiment 6	79	73	81	81	70	78	52	514	8
Experiment 7	29	24	38	22	49	38	46	246	4
Experiment 8	9	10	17	11	17	21	12	97	1

In all the experiments where hourly successive assimilation was conducted, the model performance is improved. This is clear because more information from the observations is used to correct the background field. This suggests that doing sequential assimilation like from the Ensemble Kalman filter or 4-D variational assimilation would enhance the model performance significantly by updating the model at every observations time.

From the ranking table it can be seen that the decomposition of the correlation length into two lengths determined from the wind directions performed worst of all scenarios. This could be due to the way we determined the size of the correlation ellipse, where the size of the perpendicular correlation length was determined from the wind speed ratio  $v/u$ . The wind ratio could make the ellipse too narrow so that observation spreading could be too small in some areas. Also experiments 6 and 7 did not perform well, which can be explained from the results from the anisotropic error covariance matrix, which destroys the signal from the observations stations to the model in these experiments too. It should be noted that experiment 3 with the determined error covariances performs much better than the experiment 2 with equal weights. Determining the weights is the most logical way to bring information from the model error and the observation error into the assimilation routine. As stated earlier the covariances was determined from a long time period, which might not be optimal for all time periods, where the weights are less representative.

## 5.2 Direct comparison of the reference model run and the best performing configuration

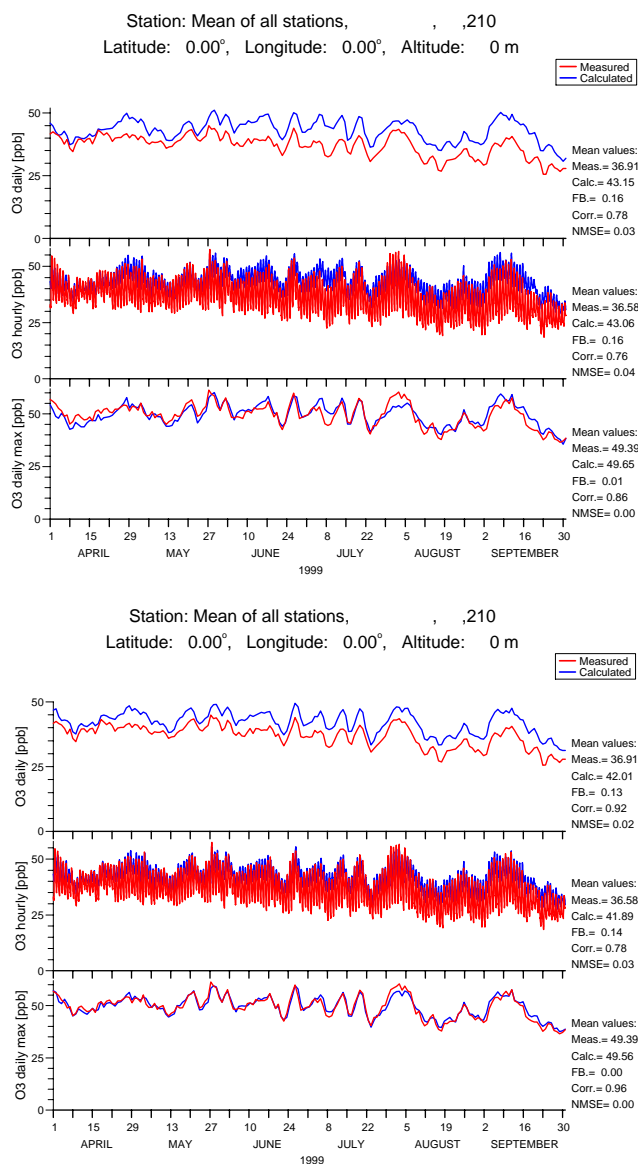
In this subsection the visualization results from the reference model run and the best performing model results from experiment 8 are shown.

In Fig. 2 the time series of the observations and the model calculations as mean over all stations from the EMEP network are shown. The figure includes times series of daily mean, hourly values and daily maximum values. From the daily mean and the daily maximum values it becomes clear that the assimilation routine pulls the model calculations toward the measurements and thereby decreases the fractional bias and increases the correlation between the observed and modeled time series.

In Figs. 3 and 4 the frequency distributions of the three statistics, calculated at the individual measurement stations for the period April–September 1999, are compared to the reference for the daily mean and daily maximum values, respectively. The figures show that the assimilation routine significantly increases the correlation for a number of stations, which can be seen in the way the histogram shifts to the right compared to the reference. The fractional bias and the normalized mean square error are relatively small in both figures for most of the measurement stations. A small change in the fractional bias, which has a tendency to be centered more around zero, can be observed from the figures. Furthermore, a shift towards smaller values of the NMSE is seen.

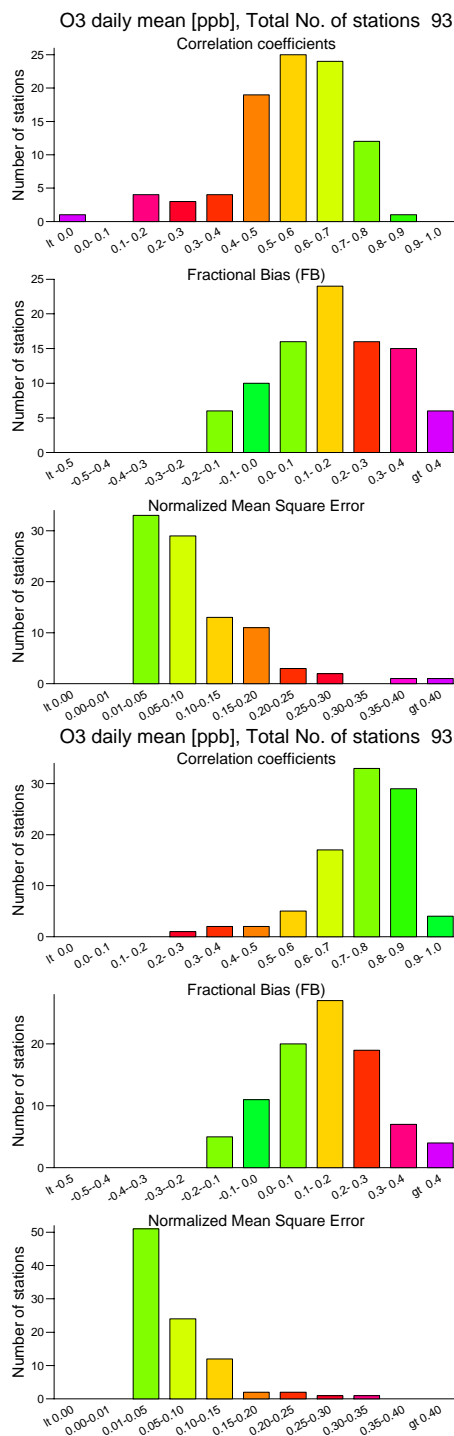
In the scatter plot in Fig. 5 showing the daily mean ozone values, we can see that the assimilation routine again improves the model outcome in the way the scatter plot gets more trimmed around the 1:1 line and the correlation coefficient increases from 0.49 to 0.68. The same is true for the scatter plots shown in Fig. 6 including all the daily maximum values from all the measurement stations as well as the corresponding model results.

In Figs. 7 and 8 scatter plots are given, including mean values for all measurement stations for the daily mean and daily maximum values, respectively. In these figures average values are made over time, whereas in Fig. 2, the averaging is carried over space. In Fig. 7, the correlation coefficient increases from 0.37 to 0.58 and the bias decreases a little. For

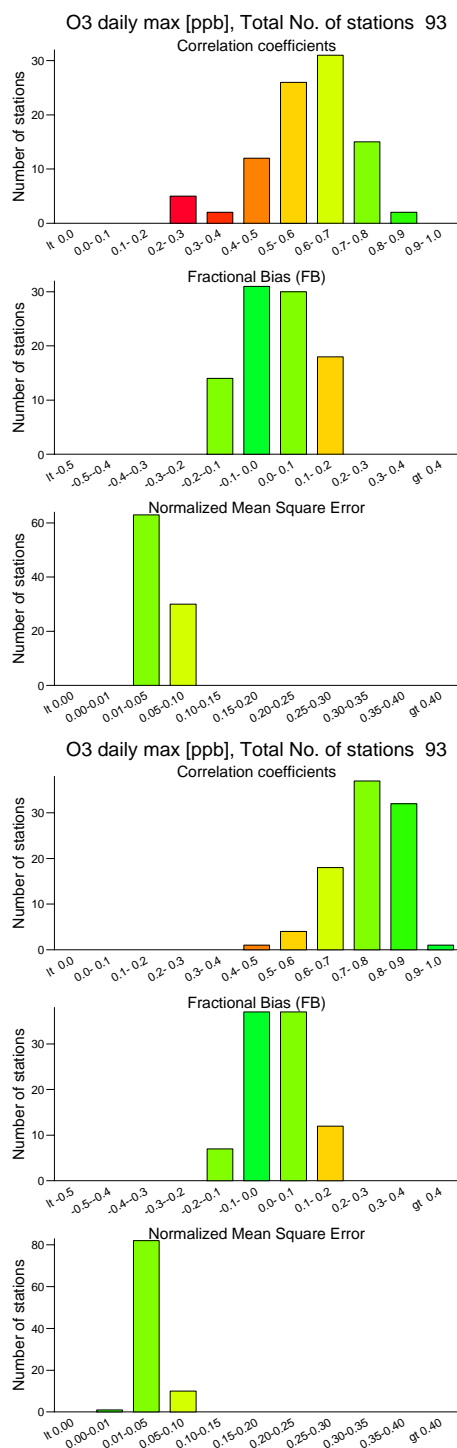


**Fig. 2.** Time series, covering the period April–September 1999, of measured and modelled values taken as a mean over all the measurement stations in the EMEP network of daily mean, hourly and daily maximum values of O<sub>3</sub>. The upper figure shows the results from the reference run without applying the data assimilation technique. The lower figure shows the results from a model run including the data assimilation as the configuration in experiment 8.

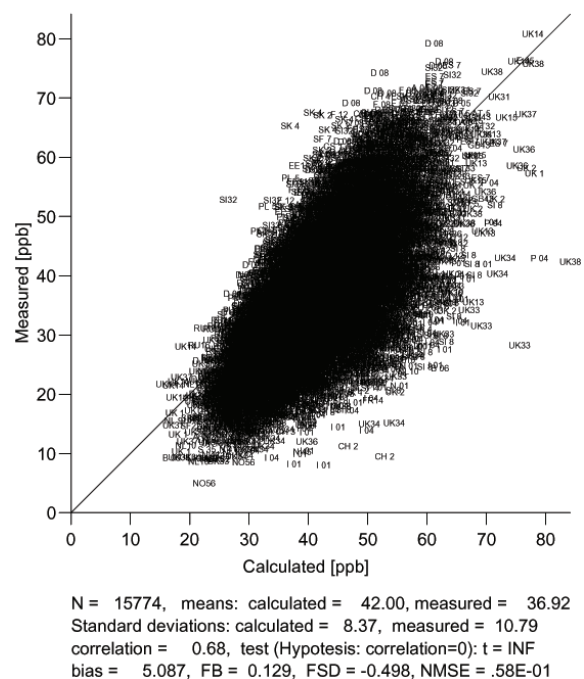
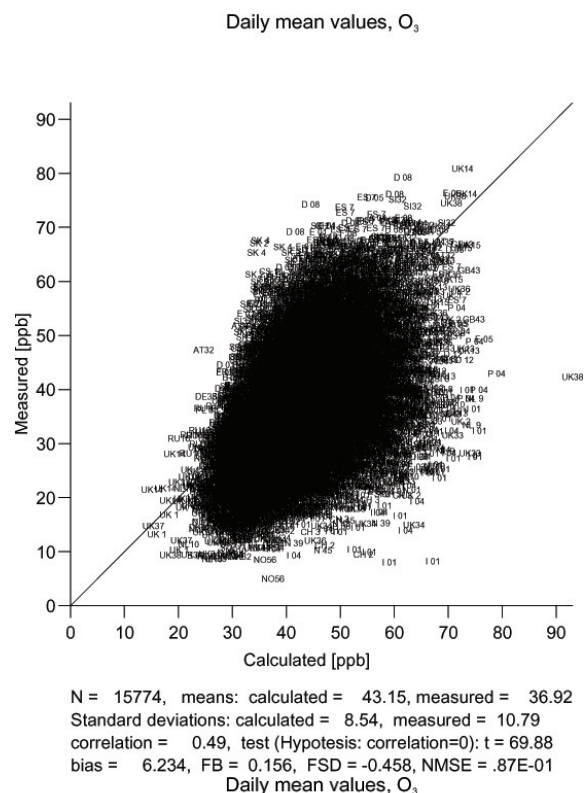
the daily maximum values displayed in Fig. 8, an increase in the correlation coefficient from 0.67 to 0.81 is seen. Also here the bias and the normalized mean square error decrease, as expected. In both figures the increase in the correlation coefficient is significant, which can be seen in the increase of the student's *t*-test parameter. An increase in the *t*-test parameter of more than 2.632 means that the increase is significant within a significance level of 1%.



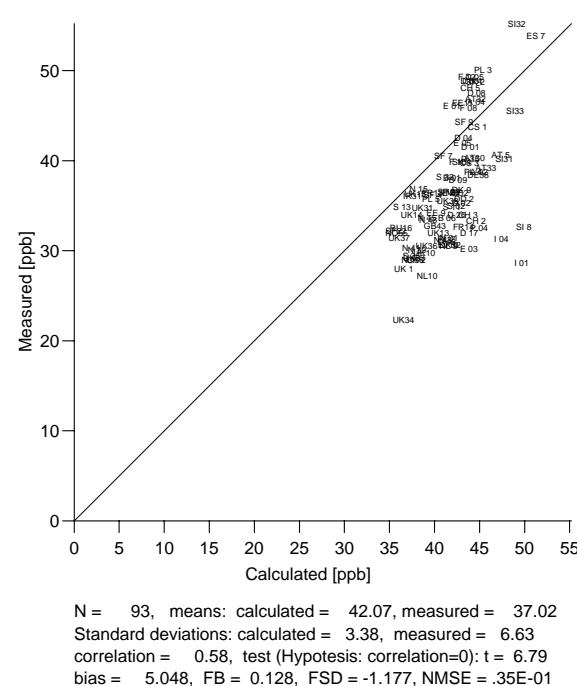
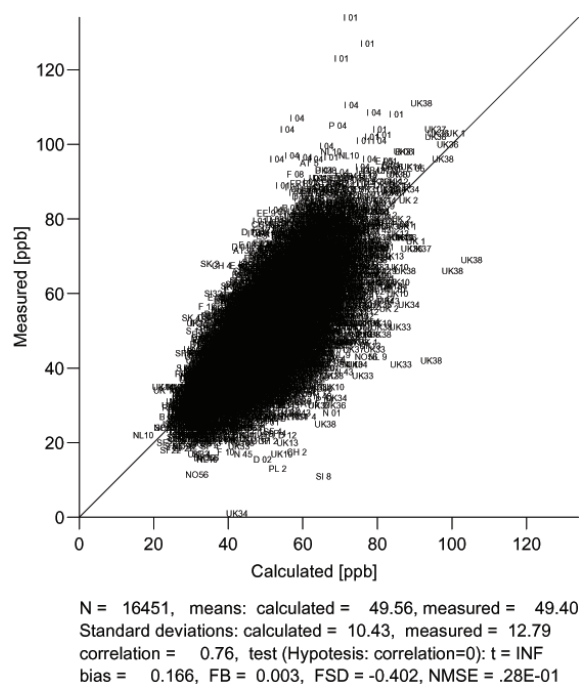
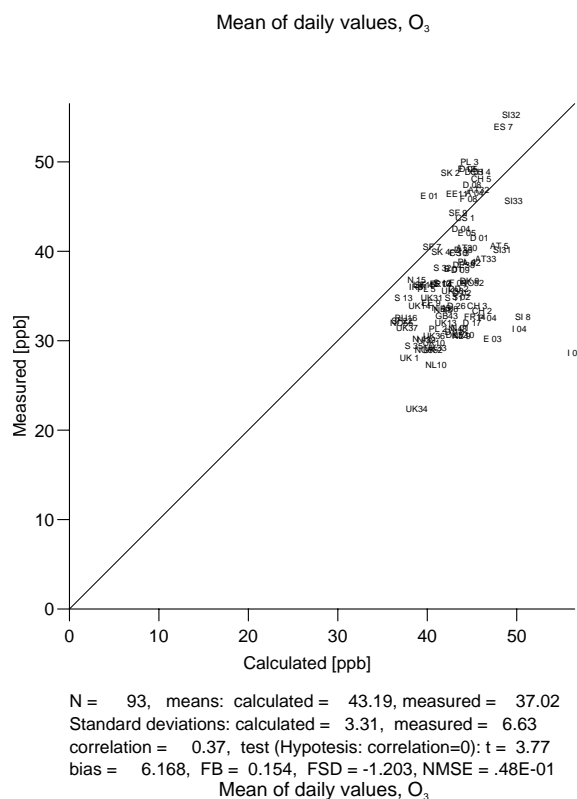
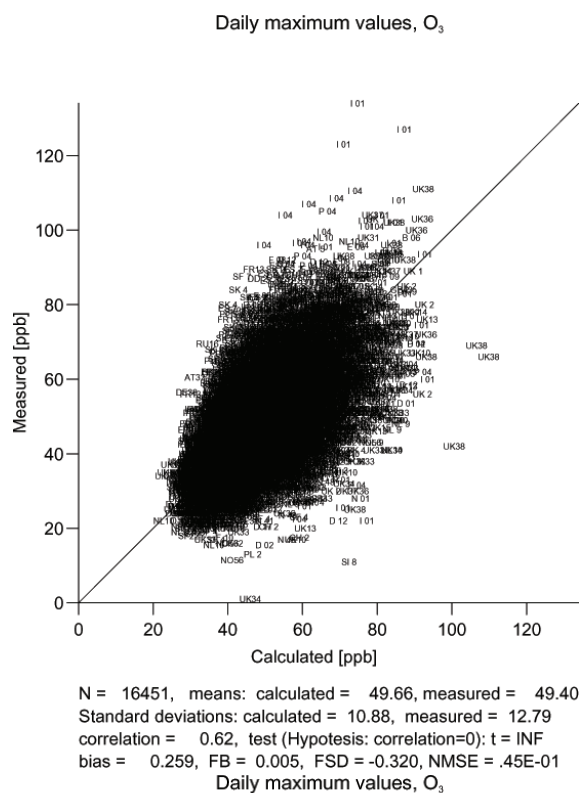
**Fig. 3.** Frequency distributions for the correlation coefficient, the fractional bias and the normalized mean square error, including the statistics from comparisons between measurements and model results for the daily mean values at each measurement stations within the EMEP network. The testing period is April–September 1999. The upper figure shows the results from the reference run without applying the data assimilation technique. The lower figure shows the results from a model run including the data assimilation as the configuration in experiment 8.



**Fig. 4.** Frequency distributions for the correlation coefficient, the fractional bias and the normalized mean square error, including the statistics from comparisons between measurements and model results for the daily maximum values at each measurement stations within the EMEP network. The testing period is April–September 1999. The upper figure shows the results from the reference run without applying the data assimilation technique. The lower figure shows the results from a model run including the data assimilation as the configuration in experiment 8.

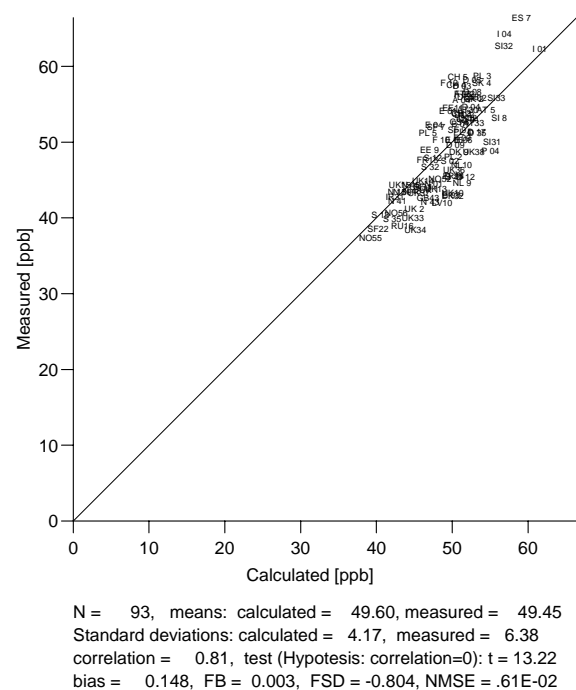
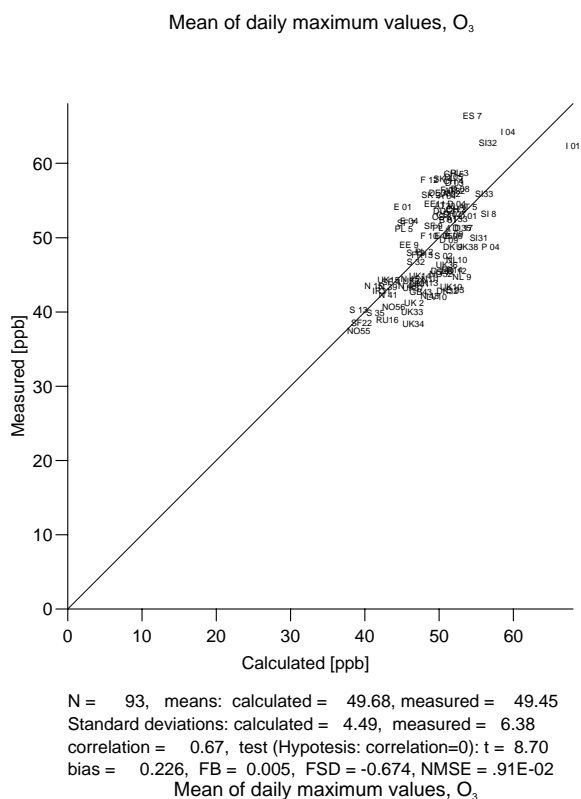


**Fig. 5.** Scatter plot showing a comparison between modelled and measured values of the daily mean values of O<sub>3</sub> including all values during the period April–September 1999, at all stations in the EMEP network. The upper figure shows the results from the reference run without applying the data assimilation technique. The lower figure shows the results from a model run including the data assimilation as the configuration in experiment 8.



**Fig. 6.** Scatter plot showing a comparison between modelled and measured values of the daily maximum values of O<sub>3</sub> including all values during the period April–September 1999, at all stations in the EMEP network. The upper figure shows the results from the reference run without applying the data assimilation technique. The lower figure shows the results from a model run including the data assimilation as the configuration in experiment 8.

**Fig. 7.** Scatter plot showing a comparison between modelled and measured values of the daily mean values of O<sub>3</sub> taken as a mean over the period April–September 1999, at all stations in the EMEP network. The upper figure shows the results from the reference run without applying the data assimilation technique. The lower figure shows the results from a model run including the data assimilation as the configuration in experiment 8.



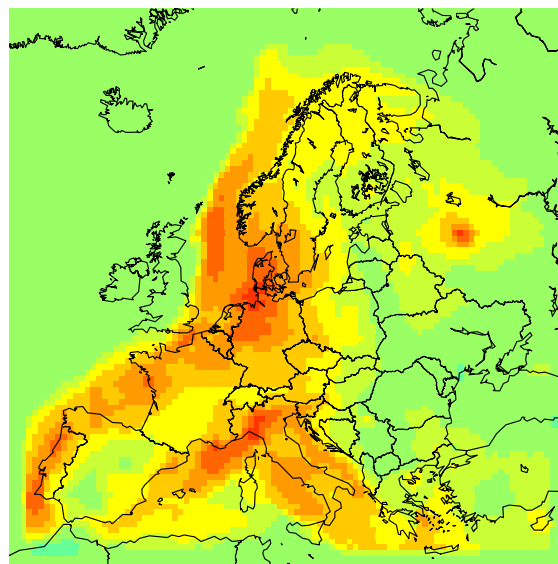
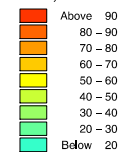
**Fig. 8.** Scatter plot showing a comparison between modelled and measured values of the daily maximum values of O<sub>3</sub> taken as a mean over the period April–September 1999, at all stations in the EMEP network. The upper figure shows the results from the reference run without applying the data assimilation technique. The lower figure shows the results from a model run including the data assimilation as the configuration in experiment 8.

DMU–ATMI THOR Air Pollution Forecast for 7/9 1999, 00 UTC

Forecast started at: 1/9 1999, 00 UTC

O<sub>3</sub> daily maximum concentrations

Units: ppb

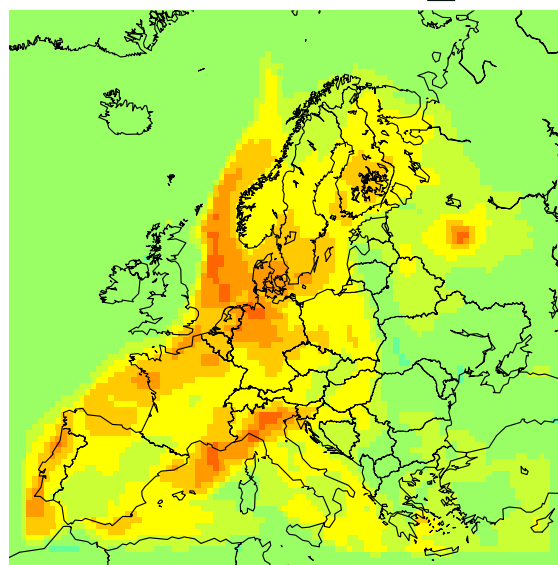
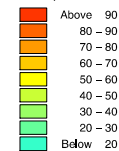


DMU–ATMI THOR Air Pollution Forecast for 7/9 1999, 00 UTC

Forecast started at: 1/9 1999, 00 UTC

O<sub>3</sub> daily maximum concentrations

Units: ppb



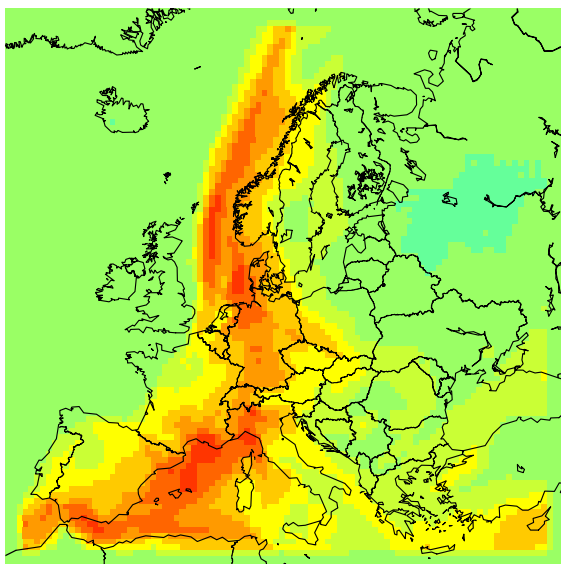
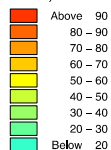
**Fig. 9.** Daily maximum ozone concentrations calculated using DEOM during an ozone episode in Europe on 7 September 1999. The upper figure shows the results from the reference run without using data assimilation. The lower figure shows the corresponding result including the data assimilation of surface O<sub>3</sub> based on the configuration in experiment 8.

DMU-ATMI THOR Air Pollution Forecast for 12/ 9 1999, 00 UTC

Forecast started at: 1/ 9 1999, 00 UTC

O<sub>3</sub> daily maximum concentrations

Units: ppb

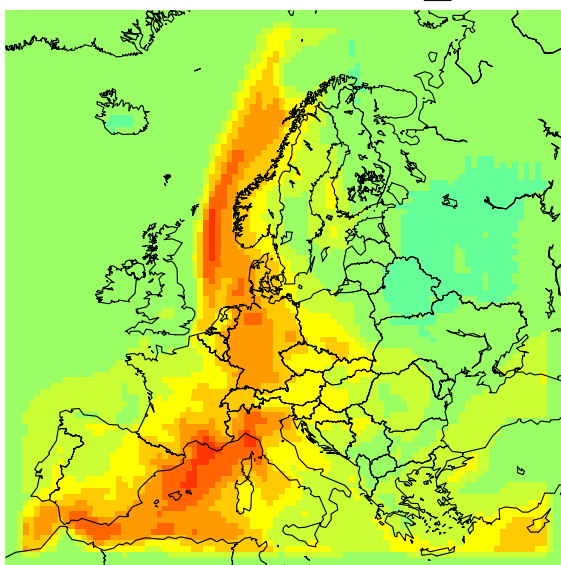
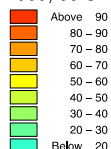


DMU-ATMI THOR Air Pollution Forecast for 12/ 9 1999, 00 UTC

Forecast started at: 1/ 9 1999, 00 UTC

O<sub>3</sub> daily maximum concentrations

Units: ppb



**Fig. 10.** Daily maximum ozone concentrations calculated using DEOM during an ozone episode in Europe on 12 September 1999. The upper figure shows the results from the reference run without using data assimilation. The lower figure shows the corresponding result including the data assimilation of surface O<sub>3</sub> based on the configuration in experiment 8.

### 5.3 Analysis of two ozone episodes

In this section, two ozone episodes that occurred on 7 and 12 September 1999 will be examined. The effect of using the data assimilation algorithm is compared to the reference run where no data assimilation is applied. The model configuration described in experiment 8 is used. The results are presented in Figs. 9 and 10, respectively, where the reference run is shown in top figures and the analyzed fields in the lower figures. Both model runs are carried out continuously, starting on 1 September, with initial data from a previous run for the month before. In the model run using the data assimilation technique, the data is assimilated each day at 10:00 UTC, 11:00 UTC and 12:00 UTC.

For both episodes there are some differences between the reference and the analyzed fields. This is the case especially for 7 September see Fig. 9, where the ozone concentrations in the Mediterranean area are decreased considerable. In this area the assimilation algorithm has pulled the general concentration level down. Also in Central Europe and in the Scandinavia region, ozone concentrations are lower compared to the reference. For 12 September see Fig. 10, the differences between the reference and the assimilated results are smaller, however, corrections are seen for smaller areas, especially in the area east of Spain and south west of Denmark.

In general, we see that the DEOM model overestimated the ozone concentration for these two days in September 1999. The overall ozone concentrations are corrected towards the observations and thereby improve the prediction capability of the DEOM model.

## 6 Conclusions

This study reports the first results of a data assimilation routine that has been developed based on Statistical Interpolation for the DEOM model. Eight different experiments including different configurations of the data assimilation algorithm were defined and tested against measurements from the EMEP network for the period April–September 1999. In order to find the optimal configuration of the algorithm, the model results from the different experiments were ranked according to the performance.

The Statistical Interpolation algorithm significantly improved the performance of the DEOM model when compared to the measurements. The Statistical Interpolation algorithm generally improved the correlation coefficient with 0.10 and the fractional bias with  $2 \times 10^{-3}$  and normalized mean square error with  $2 \times 10^{-2}$  for the overall ozone daily maximum concentrations.

The best performing setup of the data assimilation algorithm was found to be the configuration in experiment 8, including the combination of determining the covariances using the Hollingsworth method, varying the correlation

length according to the number of adjacent observation stations and applying the assimilation routine at three successive hours during the morning at 10:00 UTC, 11:00 UTC and 12:00 UTC. The results from the experiments have shown that the data assimilation routine together with a CTM is beneficial for obtaining better performance of the short-term ozone forecasts using the CTM model. Improvements in the correlation coefficients in the range of 0.1 to 0.21 between the reference and configuration in experiment 8 were seen. Additionally, there were significant reductions in bias and NMSE.

Two ozone episodes that occurred on 7 and 12 September 1999 were examined in order to make visual testing of the behavior of the algorithm for artificial behavior. It was concluded from this experiment that the data assimilation routine did not introduced any sharp gradients into the model which could lead to artificial model solutions.

It was expected that the data assimilation routine should have some effect on e.g. the NO<sub>2</sub> concentrations when altering the ozone concentrations. In experiment 8 there was no clear indication that the NO<sub>2</sub> concentration was effected significantly (not shown here). In the next step the NO<sub>2</sub> measurements could also be assimilated into the DEOM model. However, the measurement of NO<sub>2</sub> is only given as daily mean values. This means that the measurements cannot be used directly as was the case for the ozone measurements, where the hourly values were more representative for the model time step. Methods for correct representation of the daily measurements in the DEOM model using data assimilation can probably be developed by e.g. assimilating daily fields into daily mean values from the model, and then using the fraction between the two to adjust the NO<sub>2</sub> concentration at higher time resolution. This requires, however, a number of new tests and is beyond the scope of this paper. A next step of using the algorithm will be operational data assimilation of NO<sub>2</sub> data from satellite measurements.

*Acknowledgements.* This paper was partly supported by the project “GMES Service Extensions for Denmark: EO Improved Air Quality Forecast”, funded by the European Space Agency under the umbrella of PROMOTE, and partly by the Danish Research School ITMAN.

Edited by: S. Galmarini

## 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(4), 701–722, 1983.
- Blond, N. and Vautard, R.: Three-dimensional ozone analyses and their use for short-term ozone forecasts, *J. Geophys. Res.*, 109(17), D17303, doi:10.1029/2004JD004515, 2004.
- Blond, N., Bel, L., and Vautard, R.: Three-dimensional ozone data analysis with an air quality model over the Paris area, *J. Geophys. Res.*, 108(D23), 4744, doi:10.1029/2003JD003679, 2003.
- Bouttier, F. and Courtier, P.: Data assimilation concepts and methods, Tech. rep., ECMWF training course lecture notes, 59 pp., 1998.
- Brandt, J., Christensen, J., Frohn, L., Berkowicz, R., and Palmgren, F.: The DMU-ATMI THOR Air Pollution Forecast System, System Description, 2000.
- Brandt, J., Christensen, J. H., Frohn, L. M., and Berkowicz, R.: Operational air pollution forecasts from regional scale to urban street scale. Part 1: system description, *Phys. Chem. Earth Pt. B*, 26, 781–786, 2001a.
- Brandt, J., Christensen, J. H., Frohn, L. M., and Berkowicz, R.: Operational air pollution forecasts from regional scale to urban street scale. Part 2: performance evaluation, *Phys. Chem. Earth Pt. B*, 26, 825–830, 2001b.
- Brandt, J., Christensen, J. H., Frohn, L. M., Palmgren, F., Berkowicz, R., and Zlatev, Z.: Operational air pollution forecasts from European to local scale, *Atmos. Environ.*, 35, 91–98, 2001c.
- Brandt, J., Christensen, J., Frohn, L. M., Berkowicz, R., Skjøth, C. A., Geels, C., Hansen, K. M., Frydendall, J., Hedegaard, G. B., Hertel, O., Jensen, S. S., Hvidberg, M., Ketzler, M., Olesen, H. R., Løfstrøm, P., and Zlatev, Z.: THOR – an operational and integrated model system for air pollution forecasting and management from global to local scale, in: Proceedings from the First ACCENT Symposium, Urbino, Italy, 12th–16th September 2005, 6 pp., ACCENT, 2005.
- Burges, G., van Leeuwen, P. J., and Evensen, G.: Analysis Scheme in the Ensemble Kalman Filter, *American Meteorological Society, Mon. Weather Rev.*, 126(6), 1719–1724, 1998.
- Daley, R.: *Atmospheric Data Analysis*, Cambridge University Press, 1996.
- Denby, B., Kahnert, M., Brandt, J., Frydendall, J., and Zlatev, Z.: Development and application of data assimilation in regional scale atmospheric chemistry models, Tech. Rep. NILU O 30/2007, Ref: O-105077, 47 pp., April 2008.
- Elbern, H. and Schmidt, H.: A four-dimensional variational chemistry data assimilation scheme for Eulerian chemistry transport modeling, *J. Geophys. Res.-Atmos.*, 104, 18 583–18 598, 1999.
- Elbern, H., Schmidt, H., and Ebel, A.: Variational data assimilation for tropospheric chemistry modeling, *J. Geophys. Res.-Atmos.*, 102, 15 967–15 985, 1997.
- Elbern, H., Schmidt, H., Talagrand, O., and Ebel, A.: 4D-variational data assimilation with an adjoint air quality model for emission analysis, *Environ. Modell. Softw.*, 15, 539–548, 2000.
- Evensen, G.: Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics, *J. Geophys. Res.*, 99, 10 143–10 162, 1994.
- Gandin, L. S.: *Objective Analysis of Meteorological Fields*, Program Scientific Translations, 1963.
- Gaspari, G. and Cohn, S. E.: Construction of correlation functions in two and three dimensions, *Q. J. Roy. Meteor. Soc.*, 125, 723–757, 1999.
- Hamill, T. M., Whitaker, J. S., and Snyder, C.: Distance-Dependent Filtering of Background Error Covariance Estimates in an Ensemble Kalman Filter, *Mon. Weather Rev.*, 129(11), 2776–2790, 2001.
- Hertel, O., Ellermann, T., Palmgren, F., Berkowicz, R., Løfstrøm,

- P., Frohn, L. M., Geels, C., Skjøth, C. A., Brandt, J., Christensen, J., Kemp, K., and Ketzel, M.: Integrated Air Quality Monitoring – Combined use of measurements and models in monitoring programmes, *Environ. Chem.*, 4, 65–74, 2007.
- Hesstvedt, E., Hov, Ø., and Isaksen, I. S. A.: Quasi-Steady-State Approximations in Air-Pollution Modeling – Comparison of Two Numerical Schemes for Oxidant Prediction, *Int. J. Chem. Kinet.*, 10, 971–994, 1978.
- Hoelzemann, J. J., Elbern, H., and Ebel, A.: PSAS and 4D-var data assimilation for chemical state analysis by urban and rural observation sites, *Phys. Chem. Earth Pt. B*, 26, 807–812, 2001.
- Hollingsworth, A. and Lönnberg, P.: The Statistical Structure of Short-Range Forecast Errors As Determined from Radiosonde Data. Part I: The Wind-Field, *Tellus A*, 38, 111–136, 1986.
- Houtekamer, P. L. and Herschel, M. L.: Data Assimilation Using an Ensemble Kalman Filter Technique, *Mon. Weather Rev.*, 126(3), 796–811, 1998.
- Ide, K., Courtier, P., Ghil, M., and Lorenc, A. C.: Unified Notation for Data Assimilation: Operational, Sequential and Variational, Special Issue on Data Assimilation in Meteorology and Oceanography: Theory and Practice, *J. Meteorol. Soc. Jpn.*, 75, 181–189, 1997.
- Kang, D., Mathur, R., Rao, T. S., and Yu, S.: Bias adjustment techniques for improving ozone air quality forecasts, *J. Geophys. Res.*, 113, D23308, doi:10.1029/2008JD010151, 2008.
- Lamarque, J. F., Khatatov, B. V., Gille, J. C., and Brasseur, G. P.: Assimilation of Measurement of Air Pollution from Space (MAPS) CO in a global three-dimensional model, *J. Geophys. Res.-Atmos.*, 104, 26 209–26 218, 1999.
- Lawrence, M. G., Hov, Ø., Beekmann, M., Brandt, J., Elbern, H., Eskes, H., Feichter, H., and Takigawa, M.: The Chemical Weather, *Environ. Chem.*, 2, 6–8, 2005.
- Lönnberg, P. and Hollingsworth, A.: 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, 1986.
- McRae, G. J., Goodin, W. R., and Seinfeld, J. H.: Numerical-Solution of the Atmospheric Diffusion Equation for Chemically Reacting Flows, *J. Comput. Phys.*, 45, 1–42, 1982.
- M'énard, R. and Chang, L. P.: Assimilation of Stratospheric Chemical Tracer Observations Using a Kalman Filter. Part II:  $\chi^2$ -Validated Results and Analysis of Variance and Correlation Dynamics, *Mon. Weather Rev.*, 128, 2672–2686, 2000.
- M'énard, R., Cohn, S. E., Chang, L. P., and Lyster, P. M.: Assimilation of Stratospheric Chemical Tracer Observations Using a Kalman Filter. Part I: Formulation, *Mon. Weather Rev.*, 128, 2654–2671, 2000.
- Tilmes, S., Brandt, J., Flatoy, F., Bergstrom, R., Flemming, J., Langer, J., Christensen, J. H., Frohn, L. M., Hov, O., Jacobsen, I., Reimer, E., Sten, R., and Zimmermann, J.: Comparison of five eulerian air pollution forecasting systems for the summer of 1999 using the german ozone monitoring data, *J. Atmos. Chem.*, 42, 91–121, 2002.
- van Loon, M. and Heemink, A. W.: Kalman filtering for nonlinear atmospheric chemistry models: first experiences, *Tech. Rep. MAS-R9711*, Centrum voor Wiskunde en Informatica, 1997.
- Zlatev, Z. and Brandt, J.: Testing Variational Data Assimilation Modules, *Springer-Verlag*, 395–402, 2005.
- Zlatev, Z. and Brandt, J.: Testing the accuracy of a data assimilation algorithm, *International Journal of Computational Science and Engineering*, 3, 305–313, 2007.