An ensemble Kalman filter for short-term forecasting of tropospheric ozone concentrations

By K. EBEN[1]*, P. JURUŠ[1], J. RESLER[1], M. BELDA[1], E. PELIKÁN[1],
B. C. KRÜGER[2] and J. KEDER[3]

[1]*Institute of Computer Science, Academy of Sciences of the Czech Republic, Prague, Czech Republic*
[2]*Institute of Meteorology, University of Natural Resources and Applied Life Sciences (BOKU), Vienna, Austria*
[3]*Czech Hydrometeorological Institute, Prague, Czech Republic*

SUMMARY

An air-quality forecasting system based on the pair 'NWP model MM5–chemistry transport model CAMx' is proposed. A version of the ensemble Kalman Filter has been developed. The model-error covariance matrix is parametrized with the help of a covariance function and represented by an ensemble formed as a random selection from leading eigenvectors. The performance of the system is tested on the case of an ozone episode in June 2001. As a source of observations, the AirBase database has been used. Starting the forecast from analysed concentration fields improves the quality of forecast of the next day's ozone concentration maxima.

1. INTRODUCTION

In the Czech Hydrometeorological Institute (CHMI), a statistical model for forecasting daily maxima of ozone concentration has been in use since 2000. This system, developed jointly by the Institute of Computer Science (ICS) and CHMI, is based on data from the Czech air-quality monitoring system. Later on, a deterministic system was implemented in the ICS which is run in a quasi-operational regime (http://www.medard-online.cz). Due to uncertainties in the inputs for deterministic systems, it would be desirable to enhance the forecasting system with a data assimilating module. Such a module would provide a dynamic analysis of the ozone concentration field and improve forecasting skill.

Data assimilation in chemistry transport models (CTMs) has been recently studied in several papers. Some of them consider variational form using the adjoint model (e.g. Elbern *et al.* 2000; Wang *et al.* 2001) while others deal with sequential assimilation using various forms of ensemble Kalman filters (EnKF, e.g. van Loon and Heemink 1997; Segers *et al.* 2000). A comprehensive paper on Kalman filtering in air quality, covering many aspects of data assimilation, is Hanea *et al.* (2004).

The paper presents the results of a pilot study for a short-term ozone forecasting system. In order to be able to assess the influence of data assimilation, we have performed an experiment for a domain covering central and western Europe, for a period well covered with available measurements in the AirBase database. (AirBase is the public air-quality database system of the European Environmental Agency; http://air-climate.eionet.eu.int/)

2. THE FORECASTING MODEL PAIR

The numerical weather prediction (NWP) part of the system is formed by the MM5 model (Pennsylvania State University/NCAR, version 3.7.2, see www.mmm.ucar.edu/mm5). It has been configured for two domains; the coarse domain covers central and

* Corresponding author: Institute of Computer Science, Academy of Science of the Czech Republic, Pod vodarenskou vezi 2, 182 07 Prague, Czech Republic. e-mail:eben@cs.cas.cz

western Europe with 27 km resolution, and the nested domain covers the Czech Republic with 9 km resolution. There are 26 vertical levels which have arisen from the default 23 levels by subdividing some lower layers in order to get a better description of the planetary boundary layer. The NOAH land-surface model is used (Chen and Dudhia 2001). The CTM part of the system consists of the model CAMx (Comprehensive Air-quality Model with extensions, ENVIRON USA, version 4.20, http://www.camx.com) with two domains derived from the above-mentioned MM5 domains. CAMx is used with the SAPRC99 (Carter 2000) mechanism. The emission model is based on the EMEP* inventory (http://www.emep.int, with resolution about 50 km). For the countries which have taken part in the Panonia Ozone Project (POP), the emissions have been repositioned using the geographic distribution of the older POP inventory while keeping the new totals.

3. DESIGN OF THE EXPERIMENT

(*a*) *Configuration of the model and time periods of the experiment*

We have performed several experiments with different configurations of the CAMx model. It turns out that the choice of vertical diffusion scheme has a strong impact on the performance of CAMx. We have tested several schemes and have chosen the newest one (Byun *et al.* 1999) which gives the best accordance with measurements. The simulations took place for two time periods. The first period was used for station selection and tuning of parameters of the filters. The second period, when also an ozone episode took place, was used for evaluation of the filter performance.

(*b*) *Bias considerations and selection of measurement stations*

Unless we use some kind of a bias-aware filter (Dee and da Silva 1998), a basic requirement for Kalman filtration is that the model has to be unbiased. In reality, however, both the model and the observations are biased. The 'true' value actually corresponds to some unobservable average quantity. The difference between the observed value and its model counterpart should be split into a systematic part (bias) and random error in both model and observations. Such a task is obviously hard to accomplish. The best quantity we have at our disposition is the mean innovation, i.e. the mean difference between the modelled and observed value. If the mean innovation is far from zero, it could be due to the model bias as well as due to lack of representativeness of the station with respect to the corresponding grid square. The most common sort of bias in ozone modelling is the orographic bias. In Hanea *et al.* (2004), any station with altitude over 500 m was rejected from data assimilation. Since our resolution is finer and our region of interest has higher altitude, we accepted stations at altitudes under 900 m.

Another sort of model bias can arise from inadequate boundary conditions, especially if clean boundary conditions are employed. To reduce this effect, we have performed a simulation for the same time period on a large grid with the same centre as our grid but with 50 cells added on each side. From this simulation, boundary conditions have been extracted and provided for our experiments.

From a technical viewpoint, each background station which has the mean innovation close to zero is eligible for data assimilation. The process of station selection was done as follows:

(i) We confined ourselves to background stations, as is usual in air-quality modelling studies.

(ii) For the first simulation period, a free run of CAMx was performed. Stations with an absolute value of mean innovation greater than a threshold were blacklisted. This threshold was $20\,\mu g\,m^{-3}$ for O_3 and $15\,\mu g\,m^{-3}$ for NO and NO_2.

(iii) Cluster analysis of station locations based on grid coordinates was performed in order to get an even spread of selected stations. Each cluster seed was selected for assimilation. This ensured one station per grid point.

Due to daily cyclical behaviour of ozone concentrations, further reduction of bias would probably require its parametrization which is beyond the scope of this paper. From now on we shall assume the model to be unbiased.

4. THE PROPOSED DATA ASSIMILATION SYSTEM

We perform data assimilation for the coarse domain of the CTM. Uncertainty in meteorological fields, emissions and other inputs and parameters is covered by the model error.

(a) Choice of the state vector

There are 56 species entering the CAMx model with different importance for photochemical reactions. The only measurements routinely available are those of ozone and nitrogen oxides, since others are rarely measured at background stations. Two extreme choices of state vector are

(i) all values of concentrations of the modelled species in the CAMx model, i.e. concentrations of 56 species on the 3D grid. This represents over three million values.

(ii) values of species which have some measurements available, i.e. in our case O_3, NO, NO_2, on the 2D grid of the ground layer.

A reasonable compromise should take into the state vector those species which have a strong or moderate relationship with measured species. If such species are neglected, the analysis update does not take place for them and the model tends to restore the original values. On the other hand, if unrelated variables are included in the state vector, spurious correlations may arise and the computational load increases. We performed an experiment with all 56 species on a 2D grid of the ground layer included in the state vector and monitored the ensemble spread of single species. Then we dropped those species which gave insignificant spread. As a result we formed our state vector of 25 variables[*] in the 2D grid of the ground layer. Sample plots of two analyses with different choices of state vector are in Figs. 1(a) and (b).

(b) Parametrization of the model error and its representation by an ensemble

Let us recall the forecast step of the extended Kalman filter equations (for the notation and a discussion of model error in NWP models see e.g. Hamill and Whitaker (2005)):

$$\mathbf{x}_k^{f} = \mathcal{M}(\mathbf{x}_{k-1}^{a}) \tag{1}$$

$$\mathbf{P}_k^{f} = \mathbf{M}\mathbf{P}_{k-1}^{a}\mathbf{M}^{T} + \mathbf{Q}_k, \tag{2}$$

[*] O_3, NO, NO_2, CRES, PAN2, HCHO, CCHO, RCHO, ISOPRENE, HNO_3, HO_2H, CO_2H, CCO-OOH, ACET, MEK, COOH, ROOH, CO, ETHENE, ALK1, ALK3, ALK4, ALK5, SO_2, SULF. For the notation see Carter (2000), Table A-1.

where $\mathcal{M}$ is the (nonlinear) model operator, $\mathbf{M}$ its tangent linear operator and $\mathbf{Q}_k$ stands for the model-error term. A typical behaviour of ensembles in NWP models is characterized by growing of the ensemble spread during the integration of the model (Ehrendorfer 1997). This is, however, not the case in ensembles of CTM states. The substantial portion of forcing in the model causes a relatively quick convergence to a common mean for any ensemble of perturbed model states. We have performed experiments in which we switched off the model error term in the middle of the simulation and put $\mathbf{Q}_k = 0$. During the subsequent 24 hours of simulation, the ensemble spread (measured by the r.m.s. error from the ensemble mean) declined approximately to 20% of its original value for O_3 and to 10% for NO_2. For this reason it is essential to model properly the error matrix $\mathbf{Q}$.

In addition to the input errors mentioned above, model error encompasses the error of insufficient or missing physics, errors arising in model orography, physiographic data, etc. Some uncertainties, especially those related to emissions, are hard to quantify. A detailed study of sensitivity of a CTM with respect to emissions is Mallet and Sportisse (2005).

The complexity and lack of information on model errors has led us to somewhat arbitrary parametrization of the matrix $\mathbf{Q}$, related to optimum interpolation. In optimum interpolation, the overall error is constant in time and it is determined geographically. The corresponding covariance matrix is generated with the help of a spatial covariance function, isotropic or anisotropic. Different choices for such a function are described in Gaspari and Cohn (1999). Let x and y be some grid points in the horizontal, and $d(x, y)$ be their Euclidean distance. We start with the isotropic homogeneous correlation function

$$B_0(x, y) = \{1 + d(x, y)\}\, e^{-d(x,y)}. \tag{3}$$

This function has also been used in Hoelzemann (2000), who reports a relatively successful performance of the optimum interpolation data assimilation scheme of chemical constituents in CTM EURAD (EURopean Air Pollution Dispersion, www.eurad.uni-koeln.de/). Various generalizations are proposed there. We use this function (called Balgovind function in Hoelzemann 2000) as a template for the covariance function generating the model-error covariance matrix $\mathbf{Q}$. First we construct an anisotropic generalization of the function (3).

We put

$$B(x, y) = e(x, y)\{1 + r(x, y)\}\, e^{-r(x,y)}, \tag{4}$$

where

$$r(x, y) = \frac{d(x, y)}{d_r\{1 + r_w w(x, y)\}},$$

$$e(x, y) = e_{\mathrm{b}}\{1 - e_{\mathrm{lu}}|c_{\mathrm{lu}}(x) - c_{\mathrm{lu}}(y)|\},$$

$w(x, y)$ is a projection of the wind vector onto the direction connecting x and y and $c_{\mathrm{lu}}(x)$ is a scoring function. The scoring function describes the type of grid point according to the type of diurnal cycle of ozone concentrations (roughly urban versus rural) determined by land use; e_{b}, e_{lu}, r_w are suitable constants. The idea of land-use-based generalization also occurs in Hoelzemann (2000).

We suppose that the model errors in different species are uncorrelated so that the matrix $\mathbf{Q}$ has a form of

$$\mathbf{Q} = \begin{pmatrix} \sigma_1^2 \mathbf{Q}_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma_m^2 \mathbf{Q}_m \end{pmatrix},$$

where σ_i^2 is the error variance of a single species, which has to be specified and $\mathbf{Q}_i$ is some template matrix derived from the function (4) with parameters specific for i.

While it is clear that the *values* of some species may show strong correlations, it is hard to infer whether these correlations may be present in model errors. Moreover, the correlations of values depend on weather conditions; sometimes they show daily cycles, etc. The behaviour of model error relations is unclear in this case, therefore the only viable choice so far is to adopt the assumption of zero correlations.

Similarly, for the species not measured we assume a perfect model, i.e. zero variance of model error. Although there is no methodological reason for doing so, adding random noise to these variables without correcting them with a subsequent analysis might lead to an uncontrolled increase of variance.

The ensemble representing the time- and flow-dependent model error is generated in the following way:

(i) We generate the template matrix with the covariance function (4), the dimension corresponding to a single species and ground layer only. Then, a sparse approximation of this covariance matrix is taken by zeroing covariances under a given threshold. The threshold has to be selected carefully, since an overestimated threshold can cause loss of positive definiteness of the matrix.

(ii) Using the ARPACK software for sparse matrices, we generate an ensemble of the first 500 principal eigenvectors of the approximated matrix. The vectors have the form of $\mathbf{V}\Sigma^{1/2}$, where $\mathbf{Q}_i = \mathbf{V}\Sigma\mathbf{V}^{\mathrm{T}}$.

(iii) We select those species for which measurements are available or which can be correlated reasonably to the measurements. In our case we select O_3, NO, NO_2 only.

(iv) Let N be the ensemble size. For each of the selected species, a uniform random sample of N members is taken from the 500 eigenvectors. These N members are scaled by a scalar as a compensation for loss of spread due to selecting only N from 500 eigenvectors. In the case study below, N was equal to 60. They define perturbations which represent the covariance matrix of model errors for the given species.

(v) The eigenvectors are multiplied by σ_i^2, concatenated and padded with zeros so that the resulting vector has the dimension of the whole state vector. The set of all these vectors represents the matrix $\mathbf{Q}$, where all variances σ_i^2 are zero except those corresponding to the selected species.

(c)   *The analysis step in the ground layer*

Let $\mathbf{A}_b = (\mathbf{x}_1^b, \mathbf{x}_2^b, \ldots, \mathbf{x}_N^b)$ be the background ensemble propagated from the previous model step, and $\mathbf{A}_q = (\mathbf{x}_1^q, \mathbf{x}_2^q, \ldots, \mathbf{x}_N^q)$ be the ensemble representing model error as constructed above. If $\mathbf{A}_b'$ and $\mathbf{A}_q'$ are the centred counterparts of $\mathbf{A}_b$ and $\mathbf{A}_q$ respectively, and $\mathbf{A}'^f = \mathbf{A}_b'^f + \mathbf{A}_q'^f$, we can estimate $\mathbf{P}_k^f$ as

$$\mathbf{P}_k^f \approx \frac{\mathbf{A}'^f \mathbf{A}'^{f\mathrm{T}}}{N-1}.$$

The data assimilating part of the system is a version of an ensemble square-root filter, in the setting without perturbed observations and with covariance localization (Hamill *et al.* 2001; Whitaker and Hamill 2002). Let $\mathbf{H}$ be the observation operator, let $\mathbf{R}$ be the observation covariance matrix and ρ a localization function. The mean update equation is

$$\bar{\mathbf{x}}^a = \bar{\mathbf{x}}^f + \mathbf{K}(\mathbf{y} - \mathbf{H}\bar{\mathbf{x}}^f),$$

where

$$\mathbf{K} = \rho \circ (\mathbf{P}^{\mathrm{f}}\mathbf{H}^{\mathrm{T}})\mathbf{C}^{-1},$$
$$\mathbf{C} = \rho \circ (\mathbf{H}\mathbf{P}^{\mathrm{f}}\mathbf{H}^{\mathrm{T}}) + \mathbf{R},$$

and the symbol $\circ$ denotes the matrix Schur product. The localization function ρ is chosen according to formula (4.10) of Gaspari and Cohn (1999), as a piecewise rational function of fifth order with compact support.

The perturbations update equation takes the form

$$\mathbf{A}'^{\mathrm{a}} = (\mathbf{I} - \widetilde{\mathbf{K}}\mathbf{H})\mathbf{A}'^{\mathrm{f}},$$

where

$$\widetilde{\mathbf{K}} = \rho \circ (\mathbf{P}^{\mathrm{f}}\mathbf{H}^{\mathrm{T}})(\mathbf{C}^{-\frac{1}{2}})^{\mathrm{T}}(\mathbf{C}^{\frac{1}{2}} + \mathbf{R}^{\frac{1}{2}})^{-1}.$$

A sample plot of localized correlations of species is in Figs. 2(a) and (b). They have arisen as a combined result of the anisotropic covariance function in the parametrization of the matrix $\mathbf{Q}$ and the development of the ensemble in time.

(*d*) *The propagation of analysis increments in the vertical*

Since we assimilate observations in the ground layer only, we perform the 2D analysis for the ground layer. We want higher-layer concentrations to be consistent with the increments in the ground layer. If no action is taken, the increment would be quickly vertically diffused away. To prevent this, we impose artificial vertical diffusion with fixed boundary conditions. This diffusion will propagate our ground-level analysis increment in the vertical direction and reduce the shock invoked in the model. We employ the CAMx vertical diffusion equation

$$\frac{\partial c}{\partial t} = \frac{\partial}{\partial z}\left(K_{\mathrm{v}}\varrho\frac{\partial(c/\varrho)}{\partial z}\right), \tag{5}$$

where c here stands for the increment of concentration, z is the vertical dimension, K_{v} denotes the vertical diffusion coefficient and ϱ is air density. We can consider the increment separately since the equation is linear. We fix the bottom concentration equal to the analysis increment in the ground layer and the top concentration equal to zero. Artificial diffusion is performed for a specified time period (1 hour in our case).

5. A CASE-STUDY

Our ultimate target is a practically viable forecasting system for tropospheric ozone concentrations. Such a system would perform ensemble filtering hourly, in the manner suggested above. A forecast for the next day would be computed each day in the afternoon, starting from the last available analysis. The case-study reflects this conception.

We generated an ensemble of 60 members. The time period used for selection of stations and tuning of parameters of the filter was 1–20 June 2001; the period used for evaluation, when also an ozone episode occurred, was 22–29 June 2001.

Data from stations selected from the AirBase database according to section 3(b) have been assimilated. This gave about 450 measurements of O_3, 300 measurements of NO_2 and 200 measurements of NO each hour. Starting from the EnKF analysis at 1700*

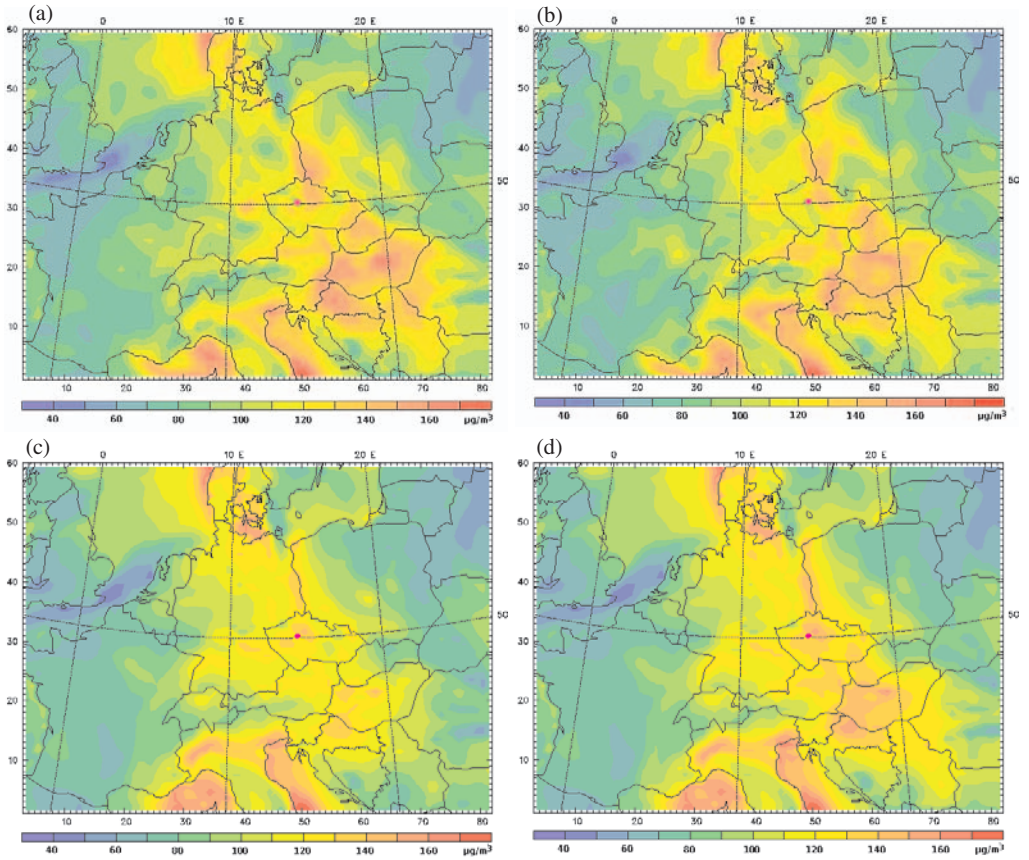* Central European Time (CET = UTC + 1 hour) is used throughout.

Figure 1. Maps of ozone concentration ($\mu g\ m^{-3}$) at 1500 on 28 June 2001 peaking episode. (a) EnKF analysis, 25 species in the state vector; (b) EnKF analysis, 3 species in the state vector; (c) free model run; and (d) forecast starting from the analysis at 1700 on 27 June.
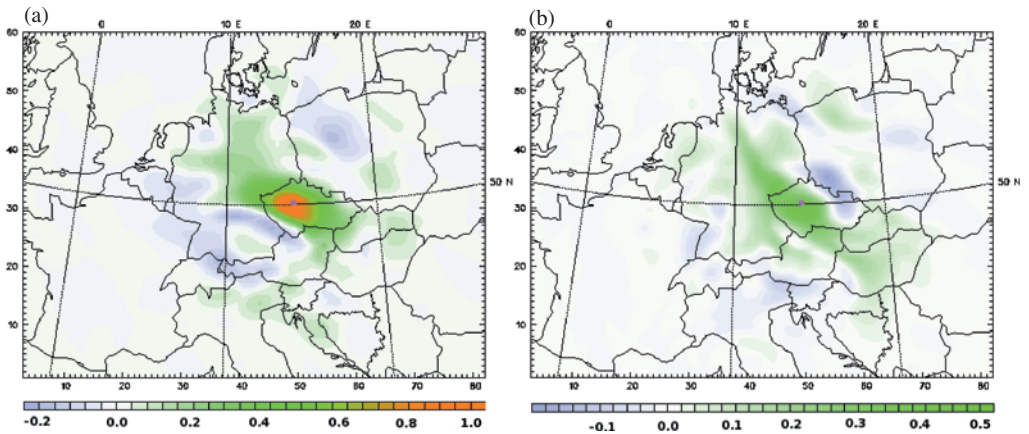

Figure 2. Maps of localized sample correlations of a species located in one grid point (Prague) with the rest of the grid: (a) $O_3 - O_3$ at 0800 on 26 June 2001, and (b) $NO_2 - CO$ at 2300 on 27 June 2001.

TABLE 1.   MEDIANS OF ABSOLUTE ERRORS WITH RESPECT TO OBSERVATIONS

Day	All values			$O_3 > 130 \ \mu g \ m^{-3}$		
	Free run	Analysis	Forecast	Free run	Analysis	Forecast
22	20.1	18.4	19.0	41.1	30.6	39.0
23	17.6	15.6	17.0	31.6	24.5	28.7
24	15.9	15.1	15.1	24.0	17.2	20.9
25	21.9	17.3	19.8	30.0	20.6	26.1
26	31.2	22.0	27.2	42.0	21.3	35.2
27	24.0	21.8	23.2	37.4	20.5	30.1
28	16.4	16.8	16.5	27.6	8.8	16.0
29	16.1	14.6	16.4	28.1	20.2	26.6

From the Figures and the Table, it can be seen that the analysis shifts all the error distribution towards zero. A significant improvement is obtained if we confine ourselves to values of elevated concentrations, where also improvement of the forecast is evident. The forecast improvement suggest that the analysis positively influences modelling of the night reservoir of ozone and nitrogen oxides and their spatial distribution.

Figures 1(a) and (b) show two versions of the analysis, one with the state vector formed from O_3, NO and NO_2, the other one with the state vector formed of the 25 species as described in section 4(a). The two versions of the analysis show local differences. We have run forecasts starting from both versions with very similar results.

6.   CONCLUSIONS AND OUTLOOK

The results of the pilot experiment show the benefit of ensemble Kalman filtering for both analysis and forecast of tropospheric ozone concentration fields. The performance improvement is evident in regions with elevated ozone concentrations, whereas overall statistics show a less pronounced result. Further performance improvements could be expected if a reasonable parametrization of the station and model bias were to be employed. This would enable a more efficient assimilation of urban background stations and would develop a better base for forecasting urban air quality. Model output statistics could be used for this purpose. Optimal tuning of ensemble size, variances in the model and observation-error matrices and other building elements of the system are the matter of future work.

REFERENCES

Byun, D. W. and Ching, J. K. S. (Eds.)   1999   'Science algorithms of the EPA Models-3 Community Multiscale Air Quality (CMAQ) Modeling System.' EPA-600/R-99/030, Office of Research and Development, US Environmental Protection Agency, Washington DC, USA

Carter, W. P. L.   2000   'Documentation of the SAPRC-99 Chemical Mechanism for VOC Reactivity Assessment.' Final Report to California Air Resources Board Contract No. 92-329 and 95-308. http://pah.cert.ucr.edu/~carter/absts.htm

Chen, F. and Dudhia, J.   2001   Coupling an advanced land-surface/hydrology model with the Penn State–NCAR MM5 modeling system. Part I: Model implementation and sensitivity. *Mon. Weather Rev.*, **129**, 569–585

Dee, D. P. and da Silva, A. M. 1998 Data assimilation in the presence of forecast bias. *Q. J. R. Meteorol. Soc.*, **124,** 269–295

Ehrendorfer, M. 1997 Predicting the uncertainty of numerical weather forecasts: a review. *Meteorol. Zeitschrift*, **6,** 147–183

Elbern, H., Schmidt, H., Talagrand, O. and Ebel, A. 2000 4D-variational data assimilation with an adjoint air quality model for emission analysis. *Environ. Modell. Software,* **15,** 539–548

Gaspari, G. and Cohn, S.E. 1999 Construction of correlation functions in two and three dimensions. *Q. J. R. Meteorol. Soc.*, **125,** 723–757

Hamill, T. M. and Whitaker, J. S. 2005 Accounting for the error due to unresolved scales in ensemble data assimilation: a comparison of different approaches. *Mon. Weather Rev.*, **133,** 3132–3147

Hamill, T. M., Whitaker, J. S. and Snyder, C. 2001 Distance-dependent filtering of background error covariance estimates in an Ensemble Kalman Filter. *Mon. Weather Rev.*, **129,** 2776–2790

Hanea, R. G., Velders, G. J. M. and Heemink, A. 2004 Data assimilation of ground-level ozone in Europe with a Kalman filter and chemistry transport model. *J. Geophys. Res.*, **109,** D10302, doi: 10.1029/2003JD004283

Hoelzemann, J. 2000 'Inhomogeneous and anisotropic assimilation of ozone observations for chemical transport modelling'. Dipl. thesis, EURAD-group, Institute for Geophysics and Meteorology, University of Cologne, Germany

Mallet, V. and Sportisse, B. 2005 A comprehensive study of ozone sensitivity with respect to emissions over Europe with a chemistry-transport model. *J. Geophys. Res.*, **110,** D22302, doi: 10.1029/2005JD006234

Segers, A., Heemink, A., Verlaan, M. and van Loon, M. 2000 A modified RRSQRT-filter for assimilating data in atmospheric chemistry models. *Environ. Mod. Software,* **15,** 663–671

van Loon, M. and Heemink, A. W. 1997 'Kalman filtering for nonlinear atmospheric chemistry models: First experiment'. Report MAS/R9711, Cent. voor Wisknude en Inf., Amsterdam, the Netherlands

Wang, K. Y., Lary, D. J., Shallcross, D. E., Hall, S. M. and Pyle, J. A. 2001 A review on the use of the adjoint method in four-dimensional atmospheric-chemistry data assimilation. *Q. J. R. Meteorol. Soc.*, **127,** 2181–2204

Whitaker, J. S. and Hamill, T. M. 2002 Ensemble data assimilation without perturbed observations. *Mon. Weather Rev.*, **130,** 1913–1924