NILU: OR 30/2007

NILU: OR 30/2007
REFERENCE: O-105077
DATE: APRIL 2008
ISBN: 978-82-425-1953-5 (print)
978-82-425-1954-2 (electronic)

Development and application of data assimilation in regional scale atmospheric chemistry models

Bruce Denby, Michael Kahnert, Jørgen Brandt, Jan Frydendall and Zahari Zlatev

Final report

for the Nordic Council of Ministers project

Development and application of data assimilation in regional scale atmospheric chemistry models

Project number 05FOX10

21 June 2007

Authors

Bruce Denby[1], Michael Kahnert[2], Jørgen Brandt[3], Jan Frydendall[3] and Zahari Zlatev[3]

[1]Norwegian Institute for Air Research (NILU), PO BOX 100, 2027 Kjeller, Norway

[2]Swedish Meteorological and Hydrological Institute (SMHI), 601 76 Norrköping, Sweden.

[3]National Environmental Research Institute (NERI), University of Aarhus, Denmark

Summary

This report presents the work and activities carried out during the Nordic Council of Ministers funded project 'Development and application of data assimilation in regional scale atmospheric chemical models' in the period 2005 - 2006. The project aim was to further develop and consolidate methodologies used in the Nordic countries for data assimilation in regional scale atmospheric chemical modelling. During the course of the two year project, work was carried out on a number of tasks including:

- Statistical interpolation of PM_{10} using the Unified EMEP model
- 2dvar analysis of ground observations using the MATCH model
- Statistical interpolation methods using the DEOM model
- Numerical testing of the 4dvar method
- Organization of an international workshop on data assimilation

The applications and methodologies treated within the project covered a wide range. Generally the data assimilation methods are applied for two separate applications. The first of these is for assessment purposes. By combining monitoring and modelling data the best possible spatial assessment can be achieved. This sort of application is evident in the first 2 tasks of the project in which model and monitoring data are combined offline to improve the spatial assessment of air quality as well as deposition rates for use in environmental impact and risk assessment. The second application is that of forecasting and is represented here by tasks 3 and 4. This also aims at establishing the best possible spatial assessment but this is carried out on a higher temporal scale, hourly, with the intention of providing the best possible initial conditions for the chemical forecast being made.

This document provides an overview of these research activities, carried out in relation to the Nordic Council of Ministers project, as well as a summary of the international workshop on data assimilation held as part of the project.

Data assimilation for atmospheric chemical models is under continuous development and is still a highly research oriented activity. It usefulness in providing the optimal combination of modelling and monitoring data from a number of sources is leading to improved assessment and forecasting of air quality in Europe. The development carried out, and shared knowledge obtained, during the course of this project has proven to be extremely useful for the further development of data assimilation methodologies in the Nordic countries.

Contents

1. Introduction

Data assimilation is a term referring to the various methods used to combine monitoring, or other forms of measured, data and model calculations. It describes a wide range of techniques, from the most simple post-modelling interpolation methods, to highly complex variational assimilation methods. Variational data assimilation techniques are often used in weather prediction models, and a number of European groups have already invested considerable efforts in applying this method to atmospheric chemical transport modelling. There are, however, also other common data assimilation methods, such as Kalman filters and ensemble methods, which can also be utilized.

This report presents the work and activities carried out during the NMR funded project 'Development and application of data assimilation in regional scale atmospheric chemical models' in the period 2005 - 2006. The project aim was to further develop and consolidate methodologies used in the Nordic countries for data assimilation in regional scale atmospheric chemical modelling. During the course of the two year project, work was carried out on a number of tasks including:

1. Statistical interpolation of PM_{10} using the Unified EMEP model
2. 2dvar analysis of ground observations using the MATCH model
3. Statistical interpolation methods using the DEOM model
4. Numerical testing of the 4dvar method
5. Organization of an international workshop on data assimilation

The applications and methodologies applied covered a wide range. Generally data assimilation methods are applied for two separate applications. The first of these is for assessment purposes. By combining monitoring and modelling data the best possible spatial assessment can be achieved. This sort of application is evident in the first 2 tasks of the project in which model and monitoring data are combined off line to improve the spatial assessment of air quality and deposition rates for use in risk assessment. The second application is that for forecasting. This also aims at establishing the best possible spatial assessment but this is carried out on a high temporal scale, hourly, with the intention of providing the best possible initial conditions for the chemical forecast being made. In addition to these major applications more fundamental work has been carried out on numerical methods for data assimilation.

The methodologies employed also vary. However each of the assimilation methods used (statistical interpolation, 2dvar and 4dvar) all have a common basis in variational methods. I.e. they try to minimise the estimated variance, or uncertainty, in the data to provide the best estimates.

The pollutants investigated also vary but ozone is a common pollutant that can be well treated with data assimilation methods. This is to a large extent the result of a large database for ozone monitoring. Other compounds, such as PM_{10} or base cations, are more limited in the available monitoring data.

The workshop on 'Data assimilation in regional scale atmospheric chemistry models' was held at NILU on 15 November, 2005, halfway through the project. More than 20 participants attended the one day workshop from the institutes involved and 8 scientific presentations were given. Two invited speakers, Hendrik Elbern and Arnold Heemink, who are experts in data assimilation methods also attended, providing valuable input and discussion. The report from the workshop has already been submitted as part of the first years reporting (Denby et al., 2006).

Data assimilation for atmospheric chemical models is under continuous development and is still a highly research oriented activity. It usefulness in providing the optimal combination of modelling

and monitoring data from a number of sources is leading to improved assessment and forecasting of air quality in Europe. The development carried out, and shared knowledge obtained, during the course of this project has proven to be extremely useful for the further development of data assimilation methodologies in the Nordic countries.

References

Denby, B., Brandt, J., Elbern, H., Frydendall, J., Heemink, A., Hvidberg, M., Kahnert, M., Tarrason, L., van Loon, M., Walker, S. E. and Zlatev, Z. (2006) Data assimilation in regional scale atmospheric chemical models. NMR Workshop at NILU, Kjeller Norway, 15 November 2005. Denby B. (ed.). Kjeller (NILU OR 43/2006).URL: www.nilu.no/data/inc/leverfil.cfm?id=22718&type=6

2. Statistical interpolation of PM_{10} using the Unified EMEP model: Application for air quality assessment in Europe and the Nordic countries

Contribution from NILU

2.1 Introduction

The aim of this work is to apply data assimilation methods for calculating the spatial distribution of PM_{10} exceedances, as defined in EU legislation (EC, 1999), on the European scale. Assessment of PM_{10} on this scale is required for estimating European wide health impacts and risk assessments suitable for policy development at the national and European level. Historically such assessments have been carried out with the use of ground based monitoring data only. This data is spatially inhomogeneous, providing information only at monitoring sites and as such cannot be used directly to calculate exceedances over the entire domain.

Chemical transport models (CTMs) may be used to calculate the spatial and temporal distribution of chemical compounds. For the case of PM_{10} regional scale CTMs tend to severely underestimate the measured concentrations (van Loon et al., 2004). As such CTMs by themselves are not useful tools for assessing limit value exceedances of PM_{10} in Europe, however they do provide important spatial information for the assessment. By combining such models with observations through the use of data assimilation techniques the quality of the assessment can be significantly improved.

There is a range of statistical interpolation techniques that can be used to assimilate, combine or 'fuse', data from different sources to create spatial concentration fields. These include traditional optimal interpolation methods (OI), Kalman filtering, kriging and residual kriging methods, regression and multiple regression techniques, e.g. Blonde et al. (2003); Horálek et al. (2005), Kassteele and Velders (2006), Fuentes and Raftery (2005) and Kasteele et al. (2006). The aim of all these methods is to combine various data sources, usually monitoring and modelling sources but also other information from remote sensing platforms or from emission or land use datasets, to provide the best estimate of the spatial distribution of a particular pollutant.

In this report a selection of basic data assimilation methods are applied, compared and assessed. These make use of statistical interpolation techniques, specifically linear regression and residual kriging. The methodologies are applied to the Unified EMEP model, together with the assimilation of ground based PM_{10} observations taken from the Airbase database, for the years 2003 and 2004.

This study presents an extension of the work conducted during the Air4EU FP6 project (www.air4eu.nl) into the application of assimilation methods on the European scale (Denby, 2007) and is in support of ongoing tasks by the European Topic Centre on Air and Climate Change to produce European wide assessment maps for health and ecosystem risk assessment applications (Horálek et al., 2007). This study gives special attention to the results of the assimilation techniques for the Nordic countries and discusses their applicability there.

2.2 Methodology

The region of Europe under focus is continental Europe, including Scandinavia. Daily mean concentrations of PM_{10} for the years 2003 and 2004 are calculated using input from the Unified EMEP model and from the Airbase database.

To compare the different assimilation methodologies cross-validation is used. This involves the removal of one station from the dataset and carrying out the assimilation process to determine the concentration at that point in space. This is done for all stations building up a model dataset that can be independently compared with measurements.

The result of the assimilations are analysed through assessment of the annual mean concentrations and the number of exceedances. Assessment of the methods is carried out using the statistical parameters root mean square error (RMSE), mean absolute error (MAE), coefficient of determination (R^2) and linear regression parameters of intercept and slope to indicate bias. Maps of the resulting exceedance fields are provided along with their assessed uncertainty maps. A number of individual Nordic stations are also used for assessment of the methodology.

2.2.1 Monitoring data

The monitoring data used is taken from the Airbase database (Airbase, 2005). Only monitoring sites described as rural background are used for the assimilation. A further selection of temporal coverage is made, including all monitoring sites with a temporal coverage > 25%.

In total 227 stations defined as rural background are registered in the Airbase database. Not all this stations are available. For 2003 the average number of stations available per day is 167, for 2004 this is 183.

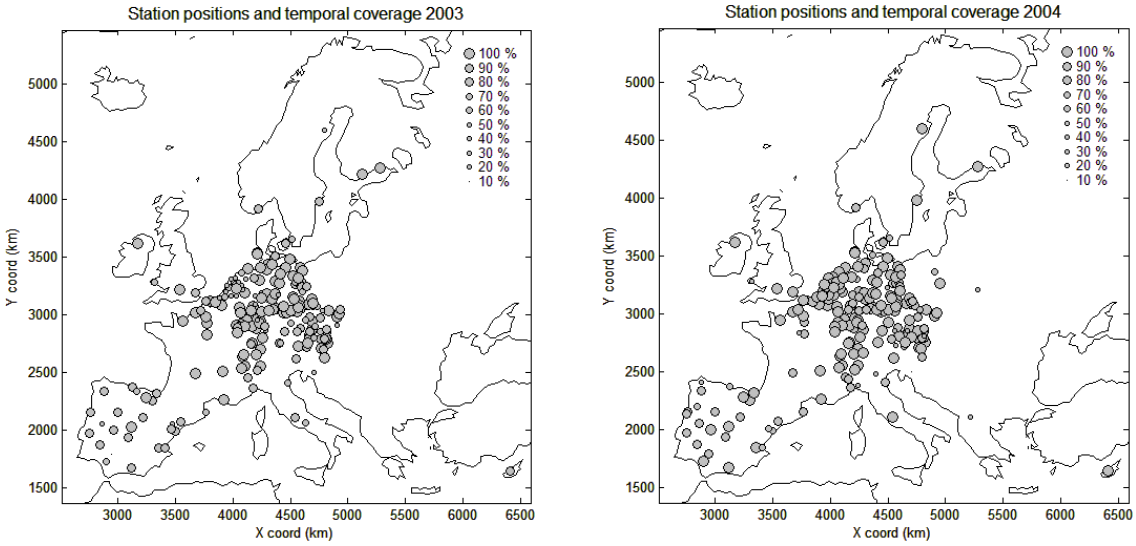


Figure 2.1. Positions and temporal coverage of the Airbase data used in the assimilations for the years 2003 (left) and 2004 (right). The size of the circles indicates the temporal coverage of the stations in %.

2.2.2 Unified EMEP model

The Unified EMEP model version rv2_1_2 (Simpson et al., 2003) has been used in the assimilation process. The model produces concentration fields of PM_{10} on a 50 x 50 km grid for all of Europe. In a recent intercomparison study for the years 1999 and 2001 (van Loon et al., 2004) the model, like many other European scale models, was shown to significantly underestimate PM_{10} concentrations with a relative bias of -45%. The underestimation is, most likely, caused by the large uncertainty in the modelling and measurement of secondary organic aerosols (SOA), and by the missing emission sources such as wind blown dust and resuspension. Despite this large bias the spatial distribution of PM_{10} concentrations represented by the EMEP model is expected to give valuable information to their spatial distribution.

For this study daily mean PM_{10} concentrations, as calculated by the Unified EMEP model, are used. When these concentrations are compared to observations bilinear interpolation is used from the 4 nearest model grids.

2.2.3 Statistical interpolation

In this study a number of statistical interpolation methods are employed on the daily mean PM_{10} model and observational datasets. These are linear regression, ordinary kriging, residual kriging of the linear regression and lognormal residual kriging with linear regression. These, and other, methods have been extensively tested in Horálek et al. (2006 and 2007) on the annual mean statistics. A more detailed study of their application on daily mean data can be found in Harálek et al. (2007), Chapter 6.

Linear regression

Linear regression of the daily mean model concentrations with the observed concentrations provides one of the simplest methods for correcting bias in the model. This can be written as

$$M'(x,y) = a + bM(x,y) \tag{2.1}$$

where the parameter *a* is the intercept, *b* the slope, *M* the model and *M'* the linear regression model. Linear regression minimises the mean square error of the regression model, in regard to the observations, to provide values for the parameters *a* and *b*. It is an effective method for removing the bias from the model but does not take into account the spatial variability of these parameters.

Linear regression was applied on a daily basis to the model. To avoid unrealistic results in the regression the following 2 conditions were applied.

1. If the intercept was found to be < 0 the regression was reanalysed by setting the intercept to 0
2. If the coefficient of determination was < 0.1 then the regression slope was set to 1 and the intercept calculated.

Though it has been shown (Horalék et al., 2006) that multiple regression with other spatially distributed data, such as topography, can improve the regression statistics for annual fields, only the CTM data was used in this comparison.

When linear regression is applied to the logarithm of the concentrations adjustment to the method, based on the correlation and intercept, is not carried out.

Ordinary kriging

Kriging is an often used interpolation method in the geosciences. It revolves around the assumption that there is a spatial correlation that can be described by a spatial variance function, the semi-variogram. This function can be used to interpolate to any point in space when observations are available. With kriging the value of the interpolated point is weighted with the observational values such that its variance is minimised. In other words, given the assumed nature of the spatial variance function the value given to the interpolated point is statistically the most likely one. Defining the semi-variogram function is thus critical to the method and should in principle be based on fits to the measured spatial variance. The most common function used to describe the semi-variogram is the spherical model, which is defined by the 3 parameters range, nugget and sill.

The weighting method is defined below where M_{OK} is the result of the ordinary kriging interpolation, λ_i is the weighting parameter and $O(x_i, y_i)$ is the observation *i* at position *x* and *y*.

$$M_{OK}(x,y) = \sum_{i=1}^{n} \lambda_i O(x_i, y_i) \tag{2.2}$$

Kriging works best when the spatial variability is much larger than the distance between measurement points. As such it has been applied, in air quality applications, to a large number of ozone interpolation applications where the density of measurement stations is high in comparison to the spatial variability.

In this study ordinary kriging is applied, which is the type of kriging most often used. It separates itself from other kriging methods in that it requires the sum of the weights, λ_i, to be equal to 1. This leads to particular properties including that the interpolated field, far from the observations, approaches the mean of these observations. For a description of kriging methods one is referred to various books on the subject such as Webster and Oliver (2001) and Cressie (1993)

Ordinary kriging was applied to the daily mean observed concentrations. The kriging parameters of nugget, sill and range were determined in two ways, firstly by fitting the semi-variogram with a spherical variance function and secondly by optimising these parameters to obtain the minimum cross validation RMSE. The second of these methods is used throughout as it always provides the lowest RMSE values (Denby, 2007).

Residual kriging of the regression model
In this method the linear regression is used as a basis for the interpolation and the residuals, observed minus regression model, are spatially interpolated using ordinary kriging. This is similar to methods using kriging with external drift, e.g. Kassteele and Velders (2006), and has been extensively investigated in other studies, e.g. Horalek et al. (2006) and Denby (2007). Residual kriging of the regression model can be formulated, similarly to kriging, as:

$$M_{OK\_RES}(x,y) = \left[\sum_{i=1}^{n} \lambda_i \left(O(x_i,y_i) - M'(x_i,y_i)\right)\right] + M'(x,y) \qquad (2.3)$$

Log-normal residual kriging of the regression model
Log-normal residual kriging was also performed by transforming the concentrations using their natural logarithm. This has carried out since concentrations of PM_{10} are often log-normal in their frequency distribution and as such the log transformation should provide more normally distributed concentration values and normal distributions should work best for the statistical methods applied here. This has been shown to be the case in Horálek et al. (2005). For the log-normal residual kriging the logarithmic transformation has been made for both the linear regression and the residual kriging.

2.2.4 Assessment methodology

In general cross-validation is used to assess the quality of the methodology. This involves removing one of the observational sites from the data set and carrying out the interpolation at that point in space. The entire observational dataset is cycled through to produce a cross-validation dataset that can be compared statistically to the actual observations. Statistical parameters used are the root mean square error (RMSE), the mean absolute error (MAE), the coefficient of determination (R^2) and the regression slope and intercept.

2.3 Air quality mapping and assessment

The results of the cross-validation for the 2 yearly periods have been assessed for the methods outlined above and are presented in tables 2.1 and 2.2. It should firstly be noted that the EMEP model severely underestimates the PM_{10} concentrations. This can be seen in the large bias (regression intercept). Annual mean concentrations using all measurements are 19.9 μgm^{-3} and 23.5 μgm^{-3} for the years 2003 and 2004 respectively giving a model bias of around –50%.

The results of the assimilations show that normal and log-normal residual kriging provides the lowest RMSE and MAE of all the methods. There is however only a slight improvement, if any, with the log-normal method over the normal method in comparison for both years.

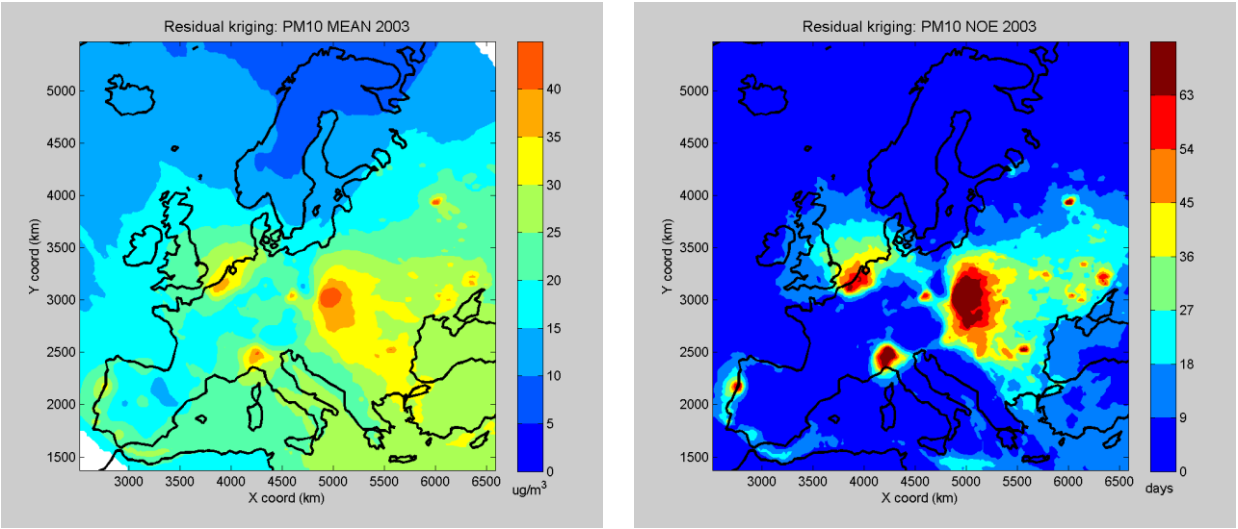
Table 2.1 Statistical analysis of the cross-validation results of the model and assimilation methods applied in the study for the years 2003 (top) and 2004 (bottom). The statistical analysis is based on the daily mean concentrations.

Cross-validation statistical parameter for 2003	**EMEP model**	**Linear regressio**	**Ordinary Kriging**	**Residual kriging**	**Log-normal residual kriging**
RMSE (μgm^{-3})	20.1	14.8	12.1	11.9	11.8
MAE (μgm^{-3})	14.0	9.21	7.3	7.1	6.9
Correlation (R^2)	0.27	0.36	0.57	0.59	0.60
Intercept (μgm^{-3})	13.5	1.2	0.7	0.9	1.8
Slope	0.89	0.95	0.97	0.96	1.0

Cross-validation statistical parameter for 2004	**Model**	**Linear regression**	**Ordinary Kriging**	**Residual kriging**	**Log-normal residual kriging**
RMSE (μgm^{-3})	17.1	12.6	10.1	9.9	9.9
MAE (μgm^{-3})	12.0	8.0	6.2	6.1	5.9
Correlation (R^2)	0.18	0.26	0.53	0.55	0.55
Intercept (μgm^{-3})	12.9	2.2	0.7	1.0	1.6
Slope	0.75	0.89	0.96	0.95	0.99

Maps are constructed at a resolution of 25 km. These show the regional background annual mean concentrations and NOE days for PM_{10}. The maps, for the 2 years assessed, are shown in figure 2.2 The year 2003 provided generally more exceedances of both the annual mean and the daily mean limit values. The area covered by daily mean exceedance is much larger than the annual mean. There are a number of regions that are in exceedance even on the regional level. These include the Nehterlands, the Poe valley in Italy and a large region in the industrialized regions of southern Poland. No exeedances on the regional scale are seen in the Nordic countries.

2003



2004

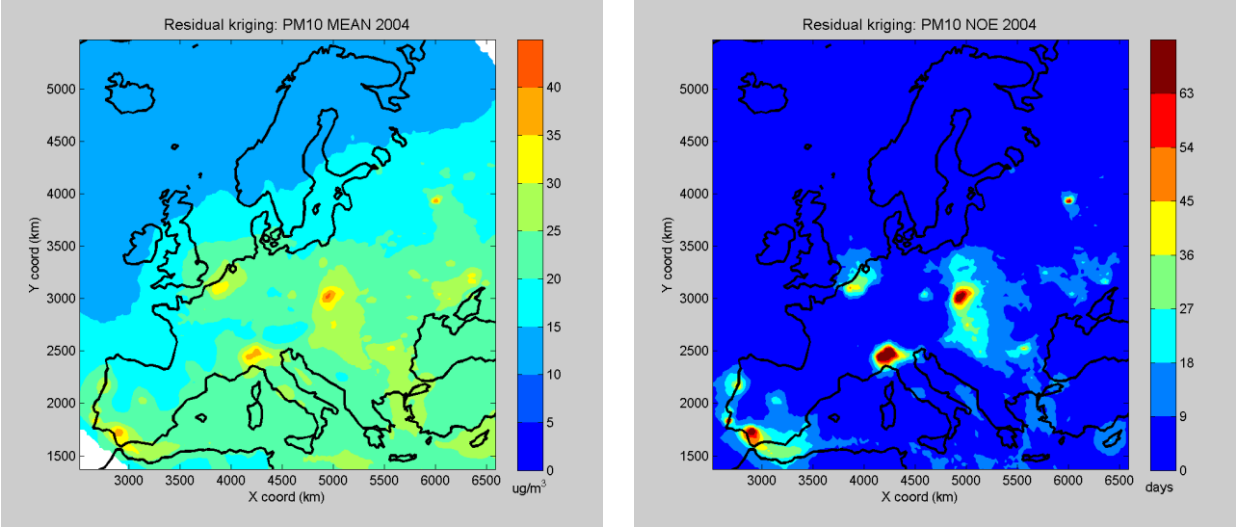


Figure 2.2. Maps showing the results of the residual kriging assimilation method for the 2 years 2003 (top) and 2004 (bottom). Shown are the annual mean PM_{10} concentrations (left), with a EU limit value of 40 µgm^{-3}, and the number of exceedance days (right). The limit value for daily mean exceedance is 50 µgm^{-3} and the maximum number of allowable exceedances of this value is 36 µgm^{-3}.

2.4 Uncertainty analysis and mapping

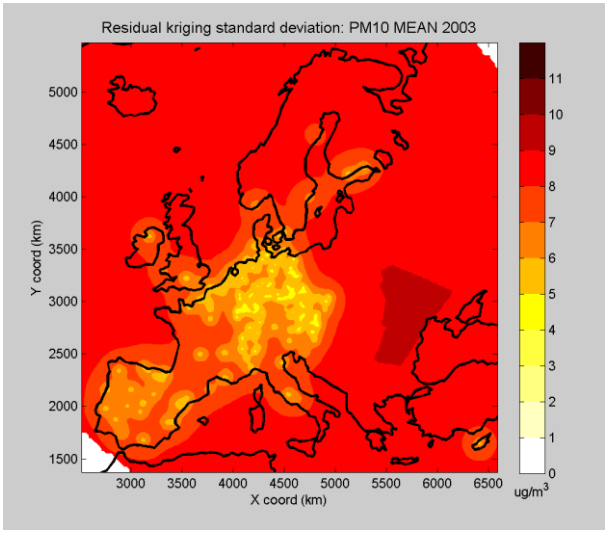
The maps presented in chapter 2.3 show the calculated spatial distribution of both annual mean concentrations and number of exceedances (NOE) of the daily mean limit of 50 µgm^{-3} for PM_{10}. These maps are constructed based on the daily mean residual kriging and regression fields. With each day an uncertainty, defined by the residual kriging variance, for each point in space can be determined. The variance is generally lowest close to stations and largest far from stations. However, the daily mean uncertainty does not tell us directly the annual mean or the NOE uncertainty. To calculate the uncertainty in these two parameters 2 approaches are used.

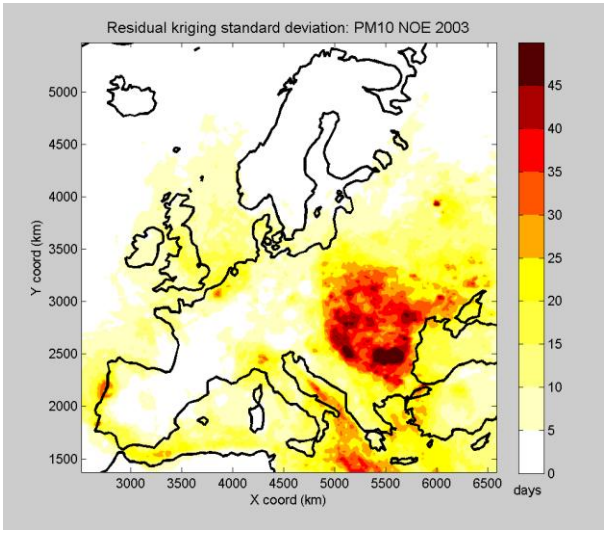
Annual means are calculated by taking the mean of all days of the year. It is not simply a matter of taking the mean of the variance for each day to obtain the annual average uncertainty field since

these fields are correlated to some extent with each other. To assess the annual mean uncertainty the temporal covariance matrix must be estimated that represents the correlation between the separate days of the year. Such an assessment, and the methodology for doing so, is described in (Denby, 2007). The temporal covariance was found to be extremely important for the annual mean uncertainty assessment, due to similar meteorological conditions that lead to correlated concentration distributions. Figure 2.3 (left) shows the calculated uncertainty fields, square root of the kriging variance, for the years 2003 and 2004. Both years show a similar uncertainty distribution with the absolute value being slightly lower for 2004 than 2003 due to generally lower concentrations in that year. Close to stations the annual mean uncertainty is around 4 – 5 μgm^{-3} whilst far from stations this uncertainty increases to around 8 – 9 μgm^{-3}. This is approximately the uncertainty expected from the linear regression model.

In an unbiased model the uncertainty in the NOE days can be derived directly from the probability of exceedance for each of the days. This approach will lead to quite low estimates of the uncertainty in NOE days. Representativeness bias will tend to dominate the NOE days uncertainty. This means that the uncertainty due to spatial representativeness of the annual mean concentrations determines the uncertainty in the NOE days. The uncertainty in the NOE can thus best be estimated by use of the annual mean standard deviation (SD) percentile band, based on the square root of the kriging variance. This is assessed by adding and subtracting the annual mean SD from the daily mean value ($\pm\sigma$), which reflects the model and representativeness bias. The uncertainty in NOE days can then be interpreted as being the maximum deviation, in number of days, from the $\pm\sigma$ calculations (Denby, 2007). Maps of this derived quantity are shown in figure 2.3 (right). Uncertainty in some areas far from observations can be very high (>50 days) showing that the estimates in these regions are actually very poor. In other areas closer to observations the estimated uncertainty in the NOE days is around 5 – 10 days.

2003





2004

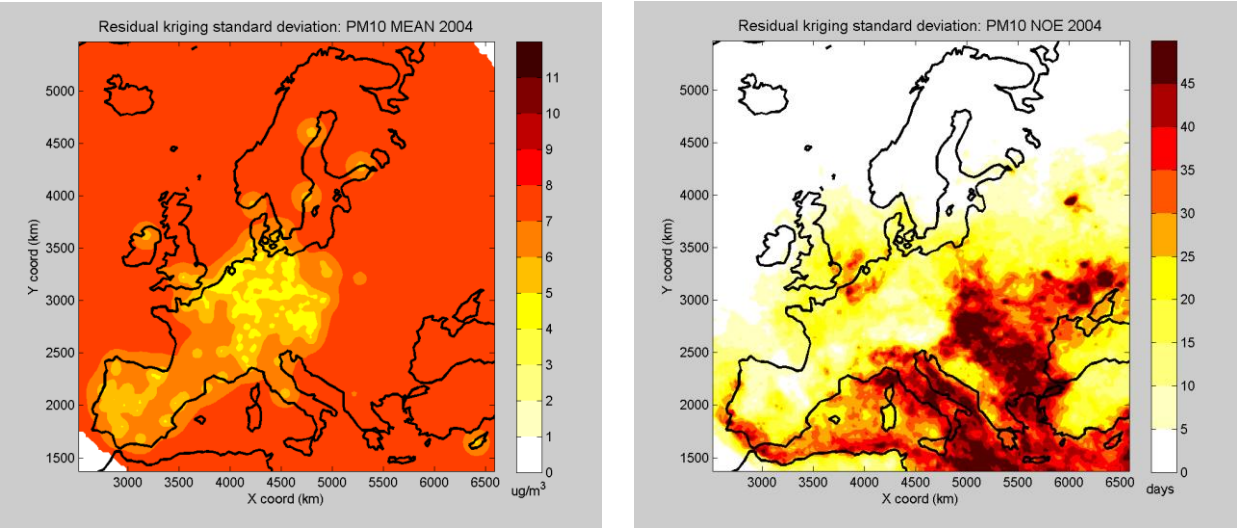


Figure 2.3. Maps showing the calculated uncertainty of the residual kriging assimilation method for the 2 years 2003 (top) and 2004 (bottom). Shown are the annual mean PM_{10} uncertainties (left) and the uncertainty in the number of exceedance days (right). Methods for calculating these uncertainties are described in the text.

2.5 Assessment of the assimilation at Nordic sites

A number of Nordic stations are included in the assimilation and the effectiveness of the assimilation can be checked at these sites using the cross-validation technique. Generally Nordic stations are more isolated than other European sites and so the residual kriging method is not expected to provide as significant improvement as with other continental sites. In figure 2.4 the positions of 9 Nordic sites are shown. In regard to the positions of the stations it should be noted that the two Danish stations DK0048A and DK0005R are placed just 1km apart. In 2003 data from the Danish stations DK0048A was missing and in 2004 data from the Finish station FI0032A was also missing.

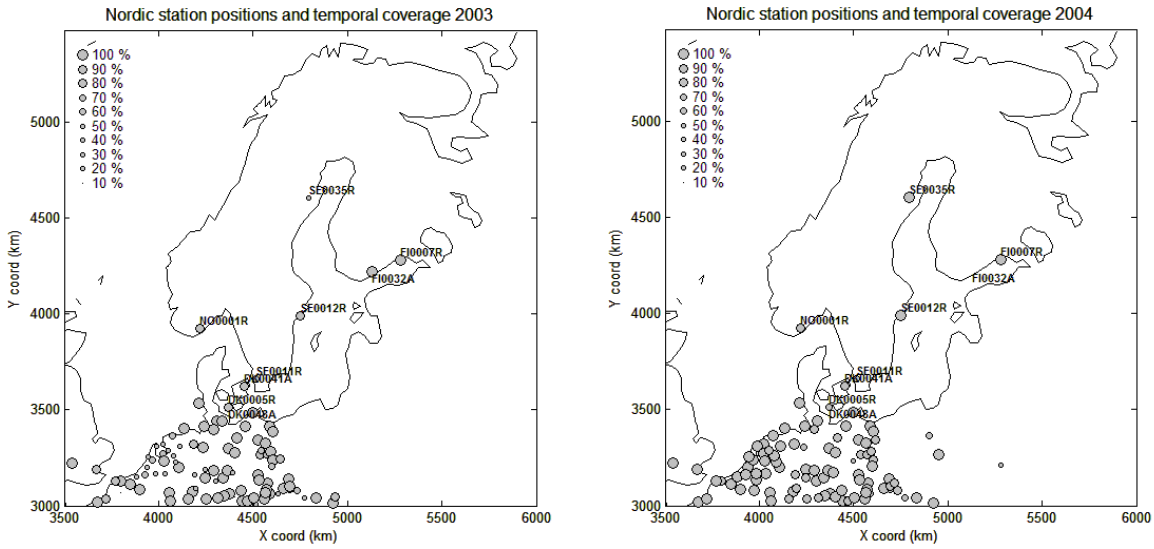


Figure 2.4. Positions and temporal coverage of the Nordic stations used in the assessment.

In figures 2.5 - 2.7 The annual mean, coefficient of determination and daily mean RMSE are compared for all 9 Nordic stations. The following conclusions can be drawn:

1. The annual mean bias is reduced with the assimilation technique for all stations except the Norwegian station (NO0001R) in both years
2. The daily mean RMSE is reduced with the assimilation technique for all stations except the Norwegian station (NO0001R) in both years
3. The coefficient of determination is improved with the assimilation technique for all stations except the Swedish station (SE0035R) in both years

These results indicate the effectiveness of the assimilation technique even in the more isolated areas where observations are few. The major exception in the 9 stations studied is the Norwegian station at Birkenes (NO0001R). The reason for this may lie in its physical proximity to continental polluted regions, such as Germany and the Netherlands, whilst the air mass passing over the station has its origins far from the continent. Other stations that have similarly low concentration levels, such as the Swedish station SE0035R, are physically further removed from the continental stations and will be less influenced by these and more by the intervening stations. This indicates one of the short comings of the methodology. However, it is of interest to note that the uncertainty calculations, figure 2.3 indicates an uncertainty of around 8 μgm^{-3} at these more remote Nordic stations. In this regard the Birkenes station, with an average bias of 8.4 μgm^{-3} is not an unlikely result and simply reflects existing uncertainties in the assimilation method.

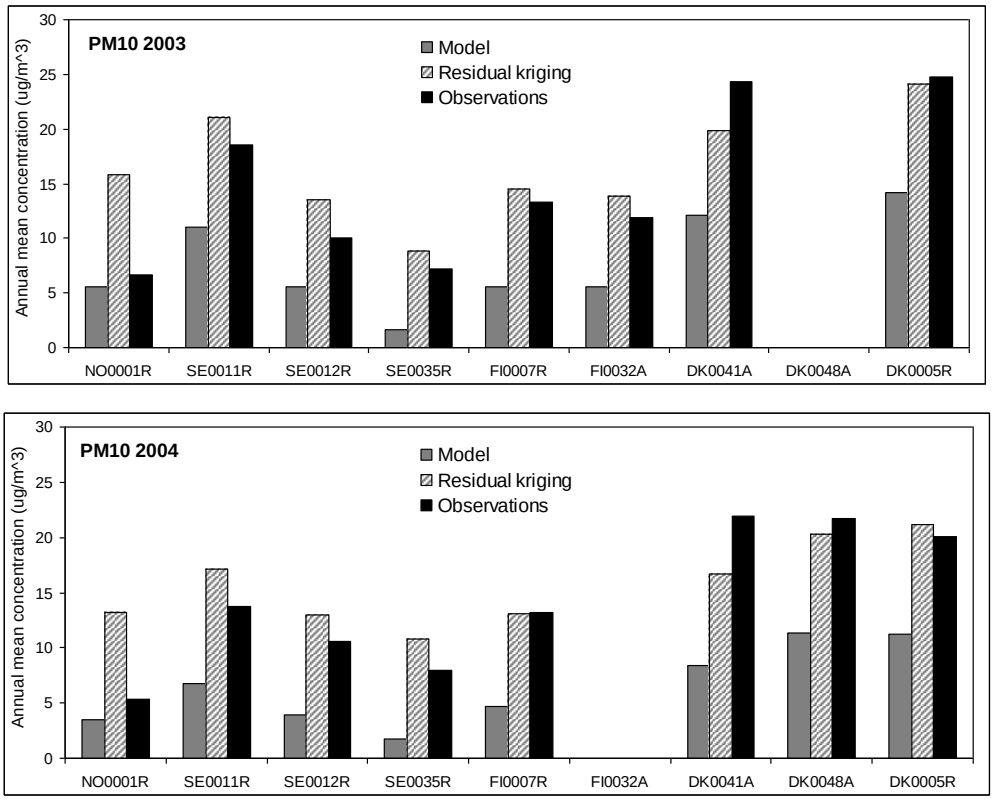


Figure 2.5. Annual mean concentrations at the 9 Nordic stations for the years 2003 (top) and 2004 (bottom). Shown are the results from the Unified EMEP model and the residual kriging assimilation technique as well as the observed values.

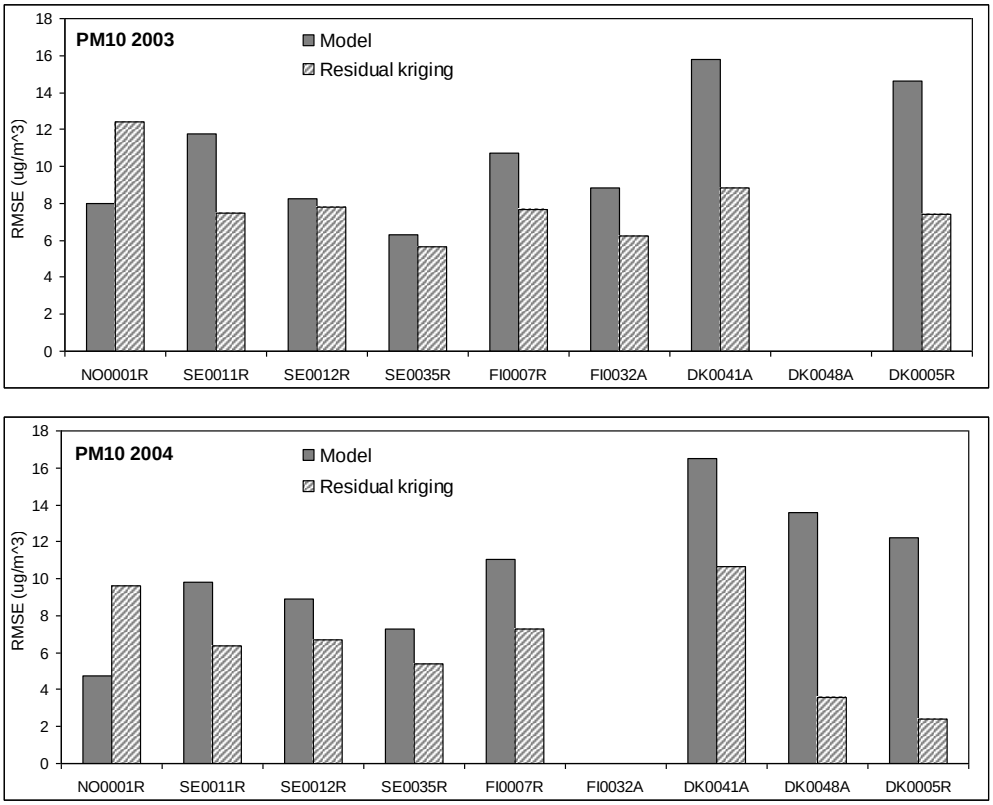


Figure 2.6. Daily mean RMSE at the 9 Nordic stations for the years 2003 (top) and 2004 (bottom). Shown are the results from the Unified EMEP model and the residual kriging assimilation technique.

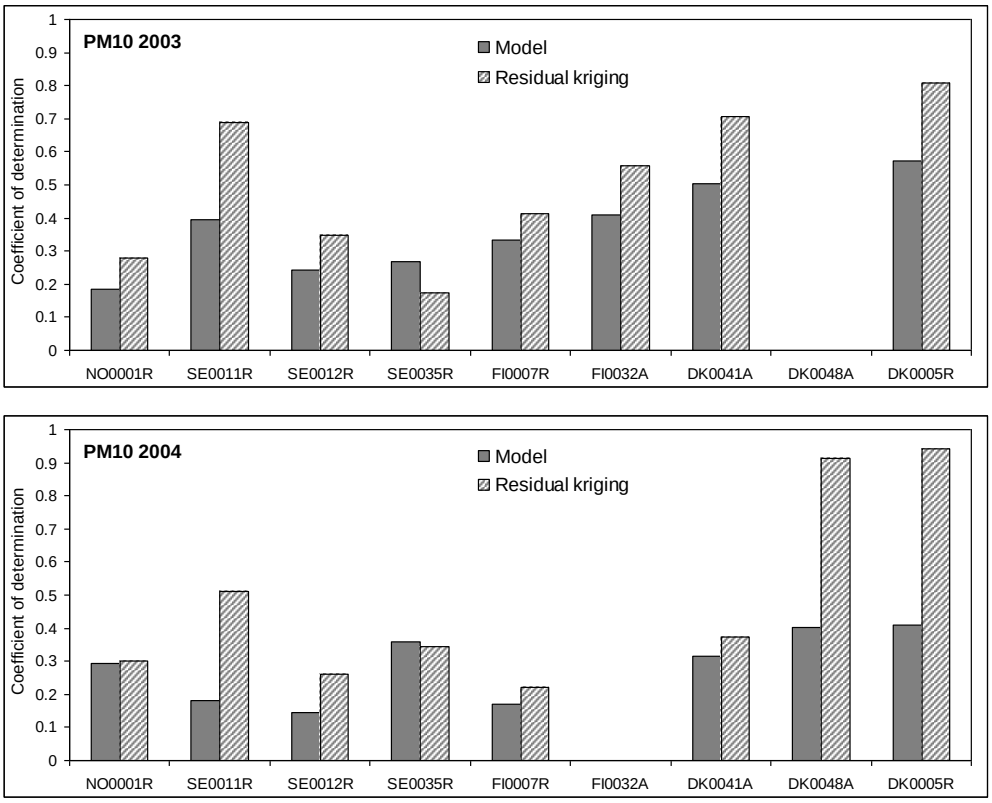


Figure 2.7. Daily mean coefficient of determination (R^2) at the 9 Nordic stations for the years 2003 (top) and 2004 (bottom). Shown are the results from the Unified EMEP model and the residual kriging assimilation technique.

2.6 Conclusions and future work

The assimilation technique applied in this study, residual kriging of the regression model, has been shown in this and a number of other studies to be effective in improving the spatially distributed concentrations of PM_{10}. The technique is straight forward and can be applied using currently available GIS software, though the work carried out here was scripted in Matlab. For the case of PM_{10} where model results are generally poor the need to assimilate these with observations is clear and this methodology is capable of providing much improved results in regard to the CTM themselves. However, as has been indicated by this particular study, it does not improve results at all sites.

In this regard the assimilation technique also allows the spatial uncertainty to be determined. This is an important step that enables a more critical and objective assessment of the model and assimilation results. It also allows a clearer understanding of the quality of the maps for policy or decision making purposes. The calculation and display of uncertainty maps is one of the future research activities that will lead from this study.

The methodology described here is currently being compared with a more sophisticated data assimilation method, that being ensemble Kalman filtering, applied to the LOTOS-EUROS model. This will enable a clearer assessment of the effectiveness of this technique compared to more advanced data assimilation methodologies.

Acknowledgments

This work has been carried out in conjunction with the FP6 research project Air4EU (www.air4eu.nl) and the European Topic Centre for Air and Climate Change (ETC/ACC) spatial mapping activities for the European Environmental Agency (EEA).

References

AirBase (2005) The European Air Quality Measurements Database. EEA-ETC/ACC.

http://air-climate.eionet.europa.eu/databases/airbase/index_html

Denby, B. (2007) Basic data assimilation: application to the European scale. Air4EU case study report (Air4EU – CS D7.1.13)

http://www.air4eu.nl/reports_products.html

Cressie, N. (1993). Statistics for Spatial Data (Rev. Ed.).: New York, Wiley.

Blond, N., Bel, L. and Vautard, R. (2003) Three-dimensional ozone data analysis with an air quality model over the Paris area, *J. Geophys. Res., 108,* 4744.

EC (1999) First daughter directive, Council directive 1999/30/EC, relating to limit values for sulphur dioxide, nitrogen dioxide and oxides of nitrogen, particulate matter, and lead in ambient, Official Journal of the European Communities, L 163, 29.06.1999, 41-60.

http://eur-lex.europa.eu/LexUriServ/site/en/oj/1999/l_163/l_16319990629en00410060.pdf

Fuentes, M. and Raftery, A. E. (2005) Model evaluation and spatial interpolation by bayesian combination of observations with outputs from numerical models. *Biometrics, 66,* 36–45.

Horálek, J., KurFürst, P., Denby, B., De Smet, P., De Leeuw, F., Brabec, M. and Fiala, J. (2005) Interpolation and assimilation methods for European scale air quality assessment and mapping, Part II: Development and testing new methodologies (ETC/ACC Technical Paper 2005/8).

http://air-climate.eionet.europa.eu/reports/ETCACC_TechnPaper_2005_8_Spatial_AQ_Dev_Test_Part_II

Horálek, J., KurFürst, P., Denby, B., De Smet, P., De Leeuw, F., Swart, R., van Noije T. and Fiala, J. (2007) Spatial mapping of air quality for European scale assessment (ETC/ACC Technical Paper 2006/6).

Kassteele, van de J., Koelemeijer, R. B. A., Dekkers, A. L. M., Schaap, M., Homan, C. D. and Stein, A. (2006) Statistical mapping of PM_{10} concentrations over Western Europe using secondary information from dispersion modelling and MODIS satellite observations. *Stoch. Environ. Res. Risk Assess, Vol 21*, 183–194

Kassteele, van de J. and G.J.M Velders (2006). Uncertainty assessment of local NO2 concentrations derived from error-in-variable external drift kriging and its relationship to the 2010 air quality standard. *Atmos. Environ.; 40,* 2583-2595.

Simpson, D., Fagerli, H., Jonson, J. E., Tsyro, S., Wind, P. and Tuovinen, J.-P. (2003) Transboundary acidification, eutrophication and ground level ozone in Europe. Unified EMEP Model Description. Oslo, Norwegian Meteorological Institute (EMEP/MSC-W Status Report 1/2003. Part I).

http://www.emep.int/index_model.html

Van Loon,M, Roemer, M., and Builtjes, P. (2004) Model intercomparison in the framework of the review of the unified EMEP model. (TNO-Rep. R 2004/282).

Webster, R. and Oliver, M.A. (2001) Geostatistics for Environmental Scientists. Chichester, John Wiley & Sons.

3. Operational 2dvar analysis of ground observations in the Swedish National Environmental Surveillance Programme

Contribution from SMHI

3.1 Introduction

The Multiple-Scale Atmospheric Chemistry and Transport Model (MATCH) [Robertson et al., 1999] is used in a wide range of different applications ranging from emergency alert to photo-chemistry assessments on various scales. Within the scope of the national environmental surveillance programme assessments of both air quality and critical loads are conducted on an annual basis. In this context it is of specific interest to discriminate between the long-range regional background and the local contribution to the air quality situation as well as to map dry and wet deposition of acidifying and eutrophying chemical species over Sweden.

To assess the domestic contribution to both air quality and the environmental impact of air pollution, MATCH is applied on a high-resolution grid (11 km) with detailed national emission inventories. A quasi-stationary Sulphur-Nitrogen chemistry scheme is employed. The chemistry is only weakly non-linear, which justifies the use of a chemistry scheme that accounts for only a subset of the atmospheric contents. The ozone field is an input parameter to the S-N chemistry scheme.

Variational data analysis has only recently been incorporated into the operational process. The background field is computed with the MATCH photochemistry version run on a European scale with European emission inventories and with low resolution (44 km). The analysis system is employed for three main purposes. (i) The ozone field is determined by analysing hourly ground observations in conjunction with MATCH photochemistry results. The analysis result is used as input to the MATCH-Sweden model. (ii) MATCH-Sweden results are subtracted from MATCH-photochemistry results and from observations. The differences, which are interpreted as the long-range transport (LRT) contributions, are analysed. By adding the MATCH-Sweden results to the analysis results one obtains a high-resolution total field of NO_x, SO_x, and NH_x concentrations in air and precipitation, which has been corrected by the observations. (iii) MATCH-seasalt modelling results are analysed by means of ground observations of Na^+, Ca^{2+}, Mg^{2+}, and K^+. The LRT and total fields of secondary inorganic compounds, as well as analysis results for base cations, serve as input to the MATCH deposition module, which computes critical loads on a high-resolution spatial scale over Sweden.

3.2 Methodology

The basis of the variational data analysis approach shall be sketched out briefly. Let us assume we have observations of some quantity y at k different observation points $\mathbf{R}_1,\ldots,\mathbf{R}_k$. The field of observations shall be denoted by the vector $\mathbf{y}=(y(\mathbf{R}_1),\ldots,y(\mathbf{R}_k))^T$. In addition there are modelling results $x^{(b)}$ available of some quantity x at the model domain's grid points $\mathbf{r}_1,\ldots,\mathbf{r}_m$, where the field of observations (the so-called background field) shall be denoted by the vector $\mathbf{x}^{(b)}=(x^{(b)}(\mathbf{r}_1),\ldots,x^{(b)}(\mathbf{r}_m))^T$. In practice, y and x may represent the same quantity. For instance, one may have an observation field $\mathbf{y}$ of ozone concentrations and a corresponding field $\mathbf{x}^{(b)}$ of computed ozone concentrations. However, y and x may also represent different but related quantities. For instance, y may represent satellite radiance observations from which the ozone concentrations can be retrieved. Let $\mathbf{h}$ denote an operator that allows for interpolating the background field to the observation points, such that $\mathbf{h}(\mathbf{x}^{(b)})$ can be directly compared to $\mathbf{y}$. The variational approach starts by defining a cost functional

$$J[\mathbf{x}] = \frac{1}{2}(\mathbf{x}-\mathbf{x}^{(b)})^T \cdot \mathbf{B}^{-1} \cdot (\mathbf{x}-\mathbf{x}^{(b)}) + \frac{1}{2}(\mathbf{y}-\mathbf{h}(\mathbf{x}))^T \cdot \mathbf{O}^{-1} \cdot (\mathbf{y}-\mathbf{h}(\mathbf{x})),$$

where **B** and **O** denote the background and observational error covariance matrices, respectively. The cost functional *J* is a measure for how much a field **x** simultaneously deviates from both the background field $\mathbf{x}^{(b)}$ and the observations **y**. The smaller the error of an observation or a background value the more the corresponding deviation contributes to the cost functional. The optimum "combination" of observations and background information is found by minimising the cost functional. The corresponding field $\mathbf{x}^{(a)}$ that minimises *J* is called the field of analysed values. It satisfies the condition $\nabla J(\mathbf{x}^{(a)}) = \mathbf{0}$. In practice one determines $\mathbf{x}^{(a)}$ iteratively. One starts with some initial guess $\mathbf{x}^{(0)}$, computes $J[\mathbf{x}^{(0)}]$ and $\nabla J[\mathbf{x}^{(0)}]$, and uses the gradient information in a standard descent algorithm to obtain an improved field $\mathbf{x}^{(1)}$ that lies closer to the minimum of *J*. This procedure is repeated iteratively to successively find improved fields $\mathbf{x}^{(i)}$ with each iteration step *i*, until the algorithm has converged with sufficient accuracy towards the analysed field $\mathbf{x}^{(a)}$.

Observations and model results are combined in the variational method in a systematic, well-controlled and reproducible fashion. For instance, less reliable observations are assigned a larger error variance and will automatically receive a lower weight in the determination of the analysed field $\mathbf{x}^{(a)}$. The error variances are the diagonal elements of the error covariance matrix. The assignment of the observational error variances can be fully automatised by making use of a data flagging system, as well as by taking the impact of meteorological conditions on the reliability of observations into account. Data flagging instead of data pre-processing allows for commenting a data file in a well-defined way without manipulating the actual data, which enhances the reproducibility of results. The background error *covariances* determine how far the information contained in an observation is propagated out into the area surrounding the observation point.

When applying the variational data analysis technique to air pollution monitoring problems one usually deals with a two dimensional field of observations near the ground, which needs to be combined with model results. The corresponding implementation of the variational approach is therefore referred to as the two-dimensional variational (2dvar) data analysis method.

3.3 2dvar post-processing of MATCH results

3.3.1 Ozone

The concentration of ozone at the surface can vary on a much smaller spatial scale than, for instance, sulphate or nitrate concentrations. The existing network of ozone monitoring stations in Sweden is insufficient for resolving these spatial variations. It is therefore essential that the information available from measurements is supplemented by model results. The MATCH model is capable of accounting for various effects, such as spatially varying ozone deposition rates due to physiographic variations, and thus to provide the necessary supplementary information to ozone monitoring data. The merits of combining ozone observations and model results by use of data analysis are illustrated in Figure 3.1.

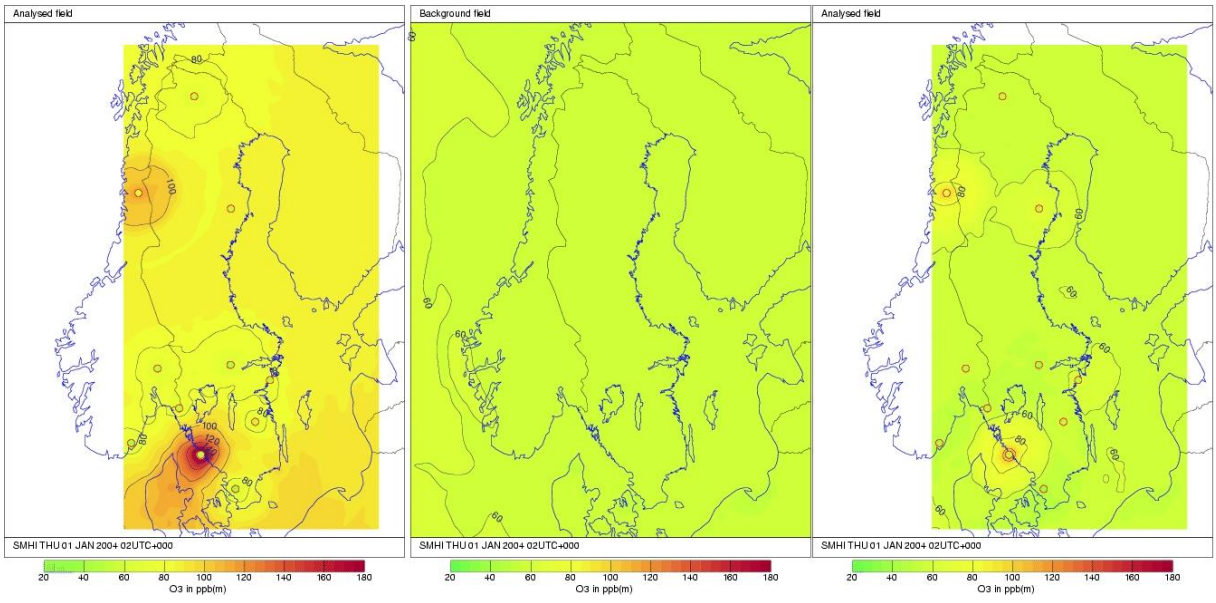


Figure 3.1. O_3 concentration field obtained from an interpolation of observations (left), the MATCH photochemistry model (middle), and a 2dvar analysis of the MATCH results and observations (right).

A mere interpolation of observations (left panel) has two main drawbacks. Firstly one can obtain unrealistic results from regionally unrepresentative observations. Secondly the results in data void regions are highly unreliable, since they are based on an interpolation of observations from distant stations. By combining model results (middle panel) with observations by use of the 2dvar data analysis technique one obtains an analysed ozone concentration field (right panel) in which the model result is corrected by the observations. However, the influence of unrepresentative observations is considerably reduced as compared to a pure interpolation of measurements. Observation error variances and variational quality control parameters can be tuned to reduce the influence of or reject unreliable observations. Further, in data void regions the analysis relies mostly on the modelled background field, which yields considerably more reliable results.

3.3.2 Sulphur and nitrogen compounds

Figure 3.2 shows an example for SO_2. The left panel shows SO_2 air concentrations computed on an 11x11 km^2 grid using Swedish emissions with the MATCH-Sweden model, employing the S-N chemistry scheme, and using the analysed ozone concentrations from the previous example as an input field. The middle panel shows a clip-out of the SO_2 concentration field computed on a 44x44 km^2 grid covering all of Europe with the MATCH-photochemistry model using European emission inventories. The Swedish contribution is subtracted from both the photochemistry results and the SO_2 observations, the difference is interpreted as the LRT-contribution (thus neglecting non-linear chemistry effects), and the LRT field is analysed. The main idea behind analysing the LRT instead of the total concentration field is that the LRT field can be assumed to vary smoothly on a larger spatial scale than the total field, which contains local, small-scale contributions. Subsequently, the high-resolution Swedish contribution is again added to the LRT analysis result. The resulting high-resolution analysis of the total field is presented in the right panel. By comparing the middle and the right panel one observes both the corrections to the background field by the observations and the modifications of the low-resolution background field due to high-resolution information about the Swedish contribution.

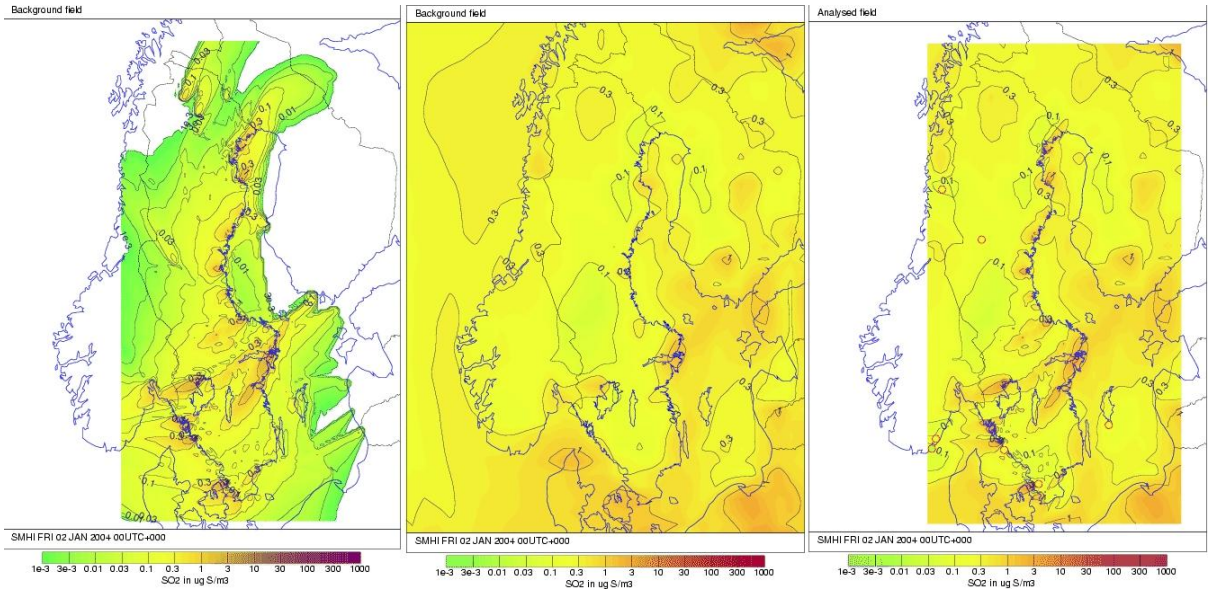


Figure 3.2. SO_2 concentration field computed with the high-resolution MATCH-Sweden model (left), the low-resolution MATCH-Europe photochemistry model (middle), and the high-resolution 2dvar-analysis (right).

This kind of data analysis is carried out operationally for daily averaged concentrations of SO_2, NO_2, NH_x, sulphate, and nitrate in air, as well as for monthly averaged concentrations of sulphate, nitrate, and NH_x in precipitation.

3.3.3 Base cations

In the previous examples the background error covariance matrix was assumed to be homogeneous and isotropic, i.e. invariant under translations and rotations. For the analysis of base cations we drop this assumption. To this end we introduce empirical coast-class correction factors [Lövblad et al., 2004] based on measurements of the variation of base cation concentrations at varying distance from the coast (see Figure 3.3, left panel). By constructing a matrix with these coast-class correction factors on the matrix diagonal and zeros in the off-diagonal elements one obtains a similarity transformation that introduces inhomogeneity and anisotropy in the background error covariance matrix such that the information from the observations is propagated out to the sea over a larger radius than over land. Since the inversion of a diagonal matrix is trivial this similarity transformation does not deteriorate the conditioning of the background error covariance matrix.

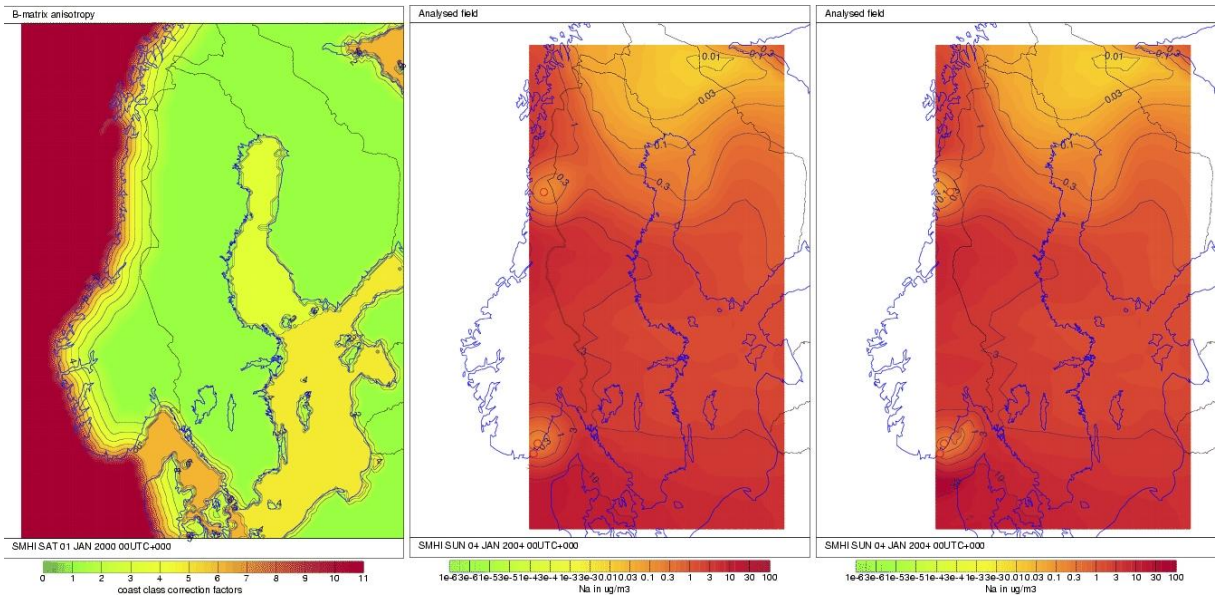


*Figure 3.3. Empirical coast-class correction factors (left), Na-analysis with an isotropic, homogeneous **B**-matrix (middle), and with anisotropy and inhomogeneity introduced into the **B**-matrix by the coast-class correction (right).*

The effect of this anisotropy can be seen in the sodium analysis shown in the middle and right panels in Figure 3.3, which are based on employing an isotropic and an anisotropic background error covariance matrix, respectively. In the isotropic case (middle) the observations induce corrections to the background field that are propagated isotropically from the observation sites into the surrounding region. By contrast, in the anisotropic case (right) the propagation of the correction to the background field is inhibited land-inwards and enhanced over the sea. Thus we efficiently account for the empirical fact that the representativeness of base cation observations decreases from coastal sites land inwards.

This kind of data analysis is carried out operationally for daily averaged concentrations in air and monthly averaged concentrations in precipitation of Na^+, Ca^{2+}, Mg^{2+}, and K^+. The background field is computed by the MATCH seasalt model.

The analysis results for sulphur and nitrogen compounds as well as for base cations are used as input to the MATCH deposition module for computing critical loads. The results of the deposition computations are published on SMHI's websites (www.smhi.se, -> miljö, -> atmosfärskemi).

3.3.4 Aerosol optical depth

As a pilot study for assimilating remote sensing information we combined the results from the MATCH photochemistry, seasalt, and primary particulate matter (PPM) models with computations of aerosol optical properties. As a result we obtain from the MATCH results profiles of differential extinction optical depth and aerosol backscattering coefficient, as well as column-accumulated extinction optical depth. The results can be compared with sunphotometer and lidar observations.

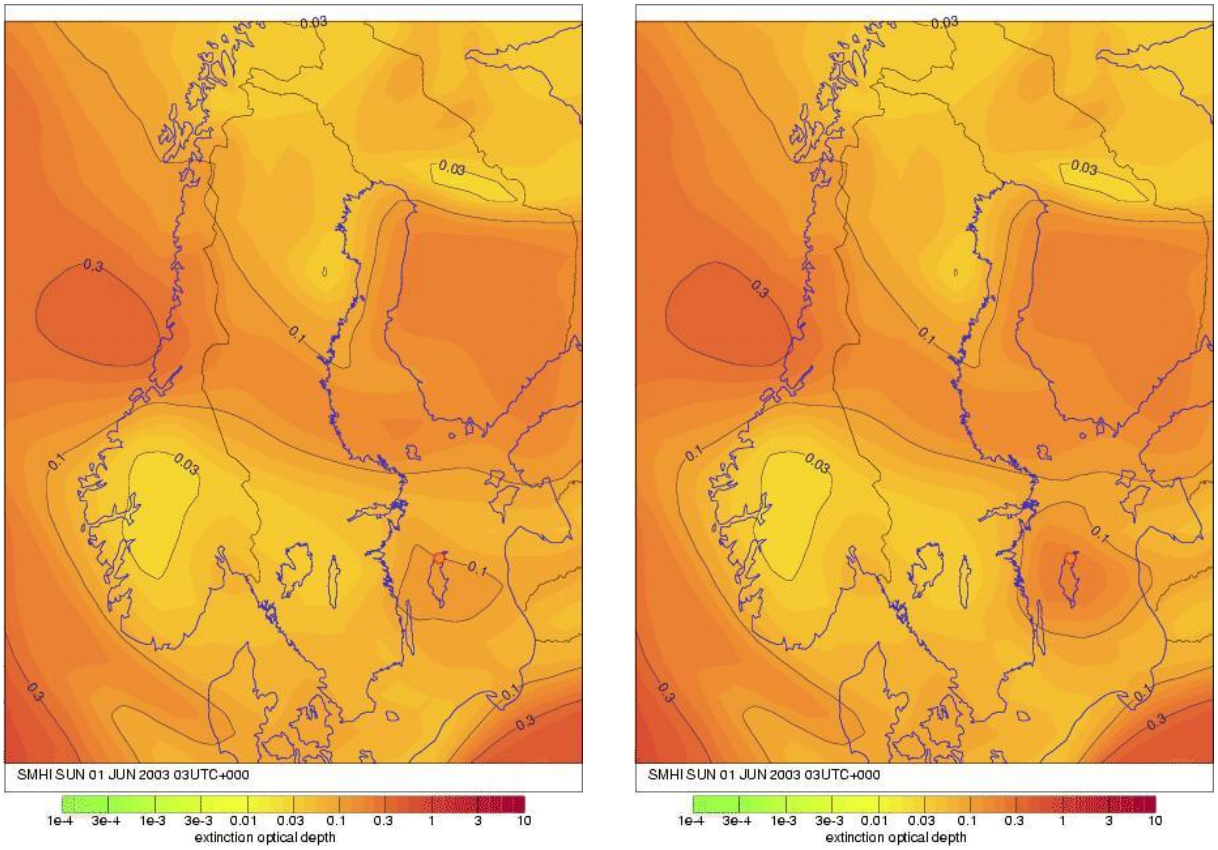


Figure 3.4. Aerosol optical depth computed from the MATCH photochemistry, sea salt, and PPM models (left), and data analysis result (right). Only one observation from the AERONET station on Gotland had been available for the analysis.

In Figure 3.4 a data analysis of sunphotometer observations of accumulated extinction optical depth (OD) is shown. Figure 3.5 shows a time series of OD observations, model results, and analysis results for the AERONET station on Gotland. The study illustrates the suitability of the model to be used for computation of optical properties, and the capability of data assimilation techniques to improve the model results. However, the current uni-variate version of the data analysis module is not capable of producing corrected concentration fields of particulate pollutants from the data analysis of optical parameters. To this end a multi-variate extention of the code will have to be implemented.

One can further see in Figure 3.5 that the currently used 2dvar data analysis methodology only yields a correction of the background at those times at which observations become available. This is because the 2dvar analysis method performs a post-processing of model results. A full data assimilation system would combine the data analysis process with the model integration over time, thus propagating the information from the observations not just out into a spatial region around the observation site, but also spreading the information in time.

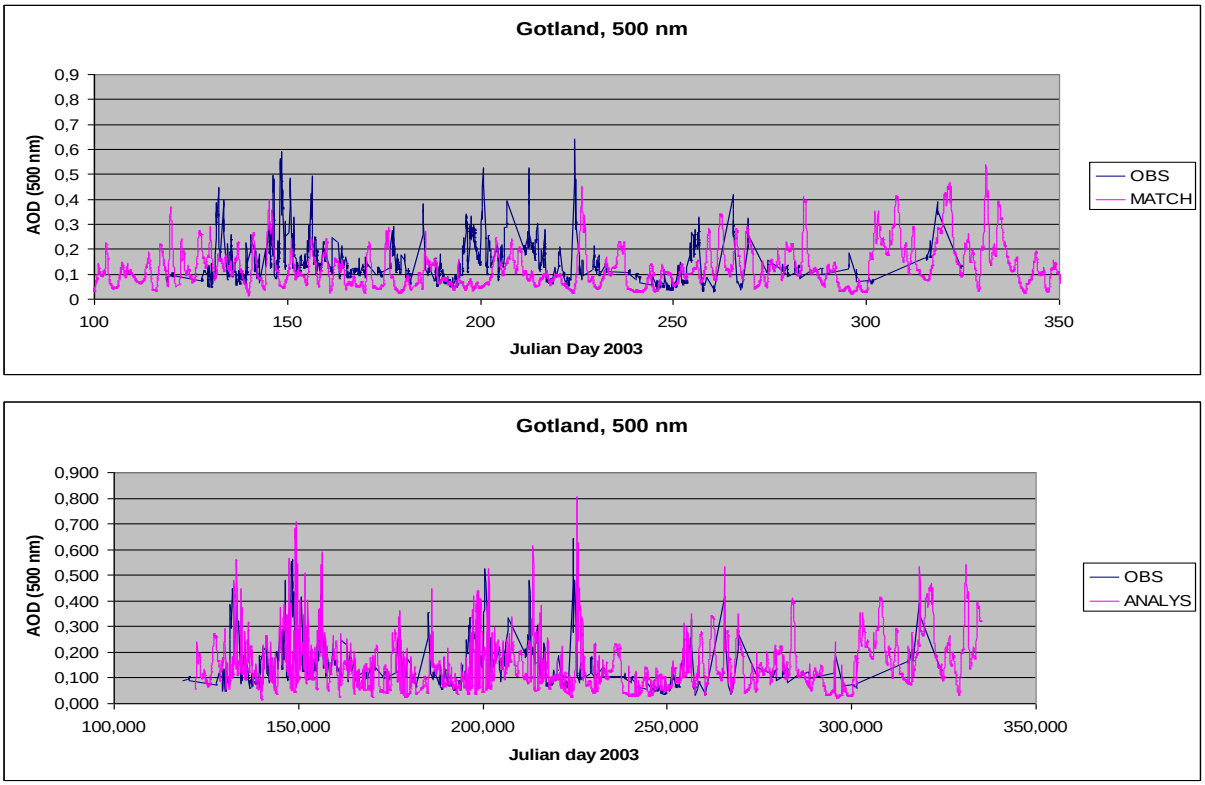


Figure 3.5: Upper panel: comparison of observed (blue) and modelled (pink) aerosol extinction optical depth at 500 nm on Gotland. Lower panel: Corresponding comparison of observations (blue) and data analysis results (pink). Note that the straight lines in the observation time series indicate data void regions. Thus the analysis relies on the background field during these periods.

3.4 Future outlook

The current operational analysis system has a number of weak points. We run a uni-variate analysis with a background error covariance matrix modelled by

$$B_{i,j} = A \cdot \exp\left[(x_i - x_j)^2 / 2\sigma^2\right],$$

or, for base cations, by

$$\mathbf{B}_{bc} = \mathbf{C} \cdot \mathbf{B} \cdot \mathbf{C}^T,$$

where **C** is a diagonal matrix containing the coast class correction factors, and where x_i denote the elements of the state vector. This approach yields analysed fields that are, in general, not consistent with the governing equations. Also, the method is not capable of exploiting information from one observed species to obtain corrections for chemically correlated species. It would be desirable to extend the current method to a multi-variate analysis that employs a background error covariance matrix based on the model's statistical properties. This would (i) yield a more "correct" analysis result, (ii) ensure consistency with the governing equations, which would prepare the analysis code to be extended to a full 4dvar version, and (iii) allow us to use observations available with high spatial and temporal resolution (such as ozone) to correct the background field for correlated species (such as NO_x) for which fewer observations are available.

As already suggested, it would be an asset if the information from available observations would not only be propagated out in space but also in time. This would require a four-dimensional data assimilation system, such as 4dvar or an ensemble Kalman filter.

References

Robertson, L., J. Langner, and M. Engardt, 1999, An Euerian limited-area atmospheric transport model. *J. Appl. Meteor.*, **38**, 190-210.

Lövblad, G., C. Persson, , T. Klein, T. Ruoho-Airola, M. Hovmand, L. Tarrason, K. Tørseth, T. Larssen, F. Moldan, and L. Rapp, 2004, The deposition of base cations in the Nordic countries. IVL Report B1583, IVL, Gothenburg.

4. Statistical interpolation with the DEOM model for ozone forecasting

Contribution from NERI

Abstract

In this chapter experiments with a statistical Interpolation algorithm applied in the DEOM model at NERI is described. The algorithm has been developed and optimized via nine different experiments where different model setups have been tested. The results from each experiment have been validated against measurements from the EMEP network. The best performing setup of the data assimilation algorithm has been found and will be used in the further application of the technique at NERI. The results will eventually be used in the THOR operational air pollution forecast system. The data assimilation technique used in this chapter is described in more detail in Frydendall, (2006); Frydendall and Brandt (2006) and Frydendall et al., (2007)

4.1 Introduction

Data assimilation has been introduced in air pollution models with good results over the last few years. Various scientific communities have developed and used different data assimilation techniques. A research group at Atmospheric Chemistry Division, National Center for Atmospheric Research, Boulder, USA used an Optimum Interpolation routine (Lamarque et al., 1999) to investigate CO in the troposphere. At the Data Assimilation Office, NASA Goddard Space Flight Center USA, a group has used a Kalman Filter to investigate chemical tracers (Ménard and Chang, 2000; Ménard et al., 2000). At the University of Cologne, Germany, a 4-DVar routine for atmospheric chemistry modelling have been developed (Elbern et al., 1997; Elbern and Schmidt 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 analyses has been developed (Blond et al., 2003; Blond and Vautard, 2003). At NERI, Denmark, a four dimensional variational method is under development (See chapter 5) and Zlatev and Brandt (2006).

A general problem in chemical data assimilation is that usually only ground based observation are available to assimilate. This is of course a problem since the vertical distribution of the chemical species is not included in the measurements. Furthermore, many of the chemical species are only measured as daily mean values and not as hourly values as be would be preferable. New types of data on chemical species are now available from satellite observations. However, this makes data assimilation more challenging since satellite observations are often filled with spatial and temporal holes (P.F.Levelt et al., 1998; Khattatovet al., 2000; 2001).

4.2 The Chemical Transport Model DEOM

The Danish Eulerian Operational Model (DEOM) (Brandt et al., 2001a; 2001b, 2001c; 2003) was developed at NERI for air pollution modelling, forecasting and assessment. The domain of DEOM is covering Europe and constructed so that it is covered by the domain of the numerical weather forecast model Eta also applied at NERI for 3-days forecasting. The Eta model is discretized on a staggered latitude/longitude system with shifted pole. The horizontal grid resolution is $0.25^o \times 0.25^o$ corresponding to approximately 39 km × 39 km at 60^oN. The number of horizontal grid points is 104 × 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 km × 50 km at 60^oN. The number of grid points is 96 × 96. Three vertical layers are used in the DEOM model. The three layers are defined as a mixed layer (below the mixing height), a reservoir layer between the mixing height and the advected mixing height from the previous day and a top layer representing the free troposphere. The model has previously been included in inter-comparison exercises showing good results (see e.g. Tilmes et al., 2002).

The DEOM model calculates transport, dispersion, deposition and chemistry of 35 species. A splitting procedure, based on the ideas of McRae et al. (1982), is applied. 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; 2001b; 2001c; 2003) along with other models. One of the main goals with the THOR system is to carry out air pollution forecasts at all scales - from the northern hemisphere scale, over European scale and the urban background scale down to the street level at both sides of the streets. The data assimilation routine developed in this study will go into the THOR system between the measured data and the regional chemical transport model DEOM, in this case covering the European domain.

4.3 The Assimilation experiments

In these experiments, the data assimilation algorithm is implemented into DEOM and the effect of applying the algorithm is tested against measurements. However, before the implementation DEOM was run for the testing period of April to September, 1999, to make a reference analysis. The testing 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.

The tests will in this study be concentrated on the daily maximum values of ozone concentrations. The DEOM model is traditionally good in predicting the daily maximum values, which means that the background field from the DEOM model will be less erroneous, compared to other parameters. 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 month for month over the entire period. The data assimilation routine is activated once every day at 12 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 hours (at 16 UTC - 18 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 hours.

Nine different tests will be performed with the data assimilation algorithm implemented in DEOM, the nine tests are:

(1) The reference run of the DEOM model without the data assimilation routine.

(2) The assimilation algorithm will be conducted with the correlation function given by (Balgovin et al., 1983; Daley, 1996)

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

Using equal weights i.e. σ_b^2 =1 , σ_o^2 = 1 and L = 3

(3) Run with optimal weights found by the Hollingsworth method (see, e.g. Daley, 1996)

(4) As (3) with the assimilation routine activated three times a day, on 10 UTC, 11 UTC and 12 UTC.

(5) Run with an anisotropic correlation function depending on the wind direction with determined weights.

(6) As (3) with the correlation function taking into account the density of observations by

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

where δ is the number of stations (between 1 and 8) (Hoelzemann et al., 2001).

(7) As (5) + (6) with one assimilation per day at 12 UTC.

(8) As (7) with the assimilation routine activated three times a day, on 10 UTC, 11 UTC and 12 UTC.

(9) Run with the correlation function in (3) with optimal weights (4) adjusted with the formula in (6) and with the assimilation routine activated three times a day, on 10 UTC, 11 UTC and 12 UTC.

4.4 The assimilation initialisation and test setup

It was decided to use six months from April to September in 1999 as our test period. The results from these runs are described in detail in Frydendall, (2006); Frydendall et al., (2007). In the tests the model results are compared to measurements and the improvements relative to the reference run without the data assimilation are examined. It should be clear that improvements in the correlation and bias should be found, since the discrepancy between the observations and the model results have been added to the model with a weight function. This procedure will automatically increase the correlation and the bias with the observations. 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 include in the analysis. When the analysis is compared to the observation stations that were exclude in the analysis, the improvement in the analysis field should be seen. However, 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 chemical produced in the model domain.

4.5 The tests and their ranks

The DEOM model was run for all nine tests. The model results were compared to measurements and statistics were calculated for every test. 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 daily maximum value of ozone was examined in the following different ways:

1) The daily maximum value as mean of all stations, where all the observation stations and calculated values are averaged over space for every day and plotted as function of time.
2) The daily maximum value including the observations and calculated daily maximum values for all times and locations
3) 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.
4) Daily mean values including all days and stations, and
5) Daily mean values for all stations, where the observations and the calculated daily maximum values for all stations are averaged in time.

The great number of statistics from the different assimilation scenarios made the comparison a big task. Therefore a ranking system was used to determine the best performing model setup. In the ranking system the ranks were assigned the number 1 to the best statistic, 2 to the second best, and so on up 10. If two statistics had the same value it got the same rank, and the successive rank was skipped. Only the corresponding statistics were compared with each other. In the end, the best assimilation setup from the rank could be determined. Details of all the tests including the statistics can be found in Frydendall, (2006).

Overall, it is found that the last assimilation test 9) is the best including the combination of varying the correlation length according to the number adjacent observations station and the assimilation routine applied at three successive hours. Having a variable correlation length increases the correlation for stations that are adjacent. Furthermore, the weights have been determined using the weights found by the Hollingsworth method.

The tests 7) and 8) showed a less good performance. This is due to the anisotropic error covariance matrix depending on the wind direction, which is destroying the signal from the observations stations to the model.

To summarize the results, the results from the reference and from the best performing model results (test 9) are shown in table 4.1. In the table the improvement achieved by the data assimilation is evident. It can be seen that the correlation coefficient e.g. is improved by 0.27 for the daily mean values including all days and stations. All in all, the data assimilation improves all statistical parameters for the forecasted daily maximum ozone concentrations.

Table 4.1. The summarized results for ozone from the reference run without data assimilation and test (9) are shown for the entire period April – September 1999.

		Correlation coefficient	**Students t-test**	**Fractional bias**	**NMSE**
Reference	Daily maximum day by day a	0.8	22.8	5.2E-0	3.1E-0
Test (9)	averaged over stations	0.9	45.0	3.4E-0	1.0E-0
Reference	Daily maximum including all day	0.6	101.4	5.2E-0	4.5E-0
Test (9)	and stations	0.7	152.1	3.4E-0	2.8E-0
Reference	Mean of daily maximum for eac	0.6	8.7	4.6E-0	9.1E-0
Test (9)	station	0.8	13.2	3.0E-0	3.0E-0
Reference	Daily mean values including a	0.4	69.8	0.15	8.7E-0
Test (9)	days and stations	0.6	Na	0.12	5.8E-0
Reference	Daily mean values averaged ov	0.3	3.7	0.15	4.8E-0
Test (9)	time for all stations	0.5	6.7	0.12	3.5E-0

4.6 Analysis of two ozone episodes

In this section, two ozone episodes that occurred on September 7, 1999 and September 12, 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 setup described in test (9) is used. The results are presented in figure 4.1, where the reference run is shown in the left column and the analyzed fields in the right column. Both runs are continuous, started on September 1st, with initial data from a previous run for the month before. In the model run using the data assimilation, the data is assimilated each day at 10 UTC, 11 UTC and 12 UTC. For both episodes there are some differences between the reference and the analyzed fields especially for September 12th, where the ozone concentrations in the Mediterranean area are decreased. In this area the assimilation routine has pulled the concentration level down. Also in central Europe ozone concentration levels are lower compared to the reference. In the Scandinavian region the ozone concentration levels also have decreased in both episodes. In general we see that the DEOM model had an overestimation of ozone concentration for these two days in September. The data assimilation routine corrected the calculations to match the measurements

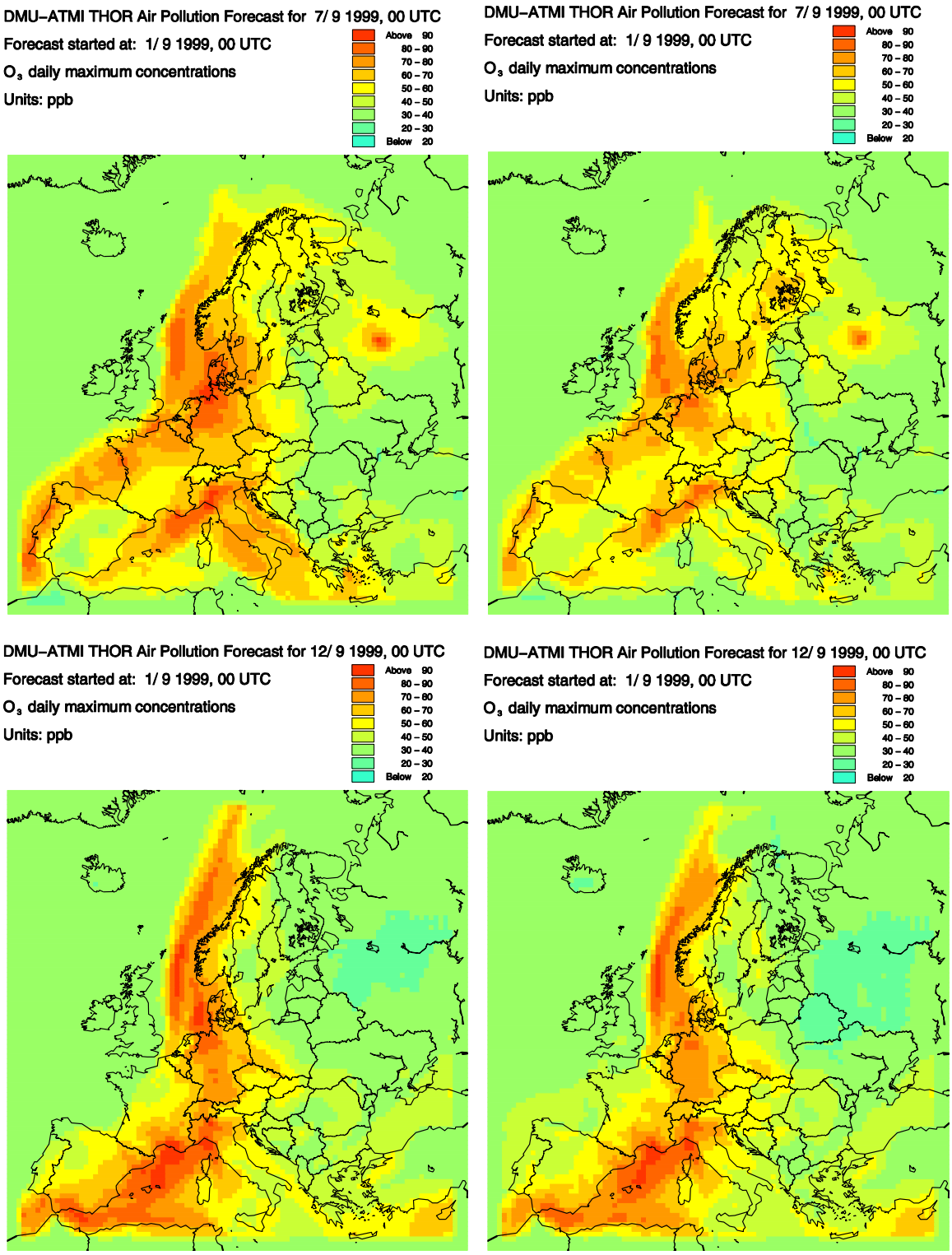


Figure 4.1. Daily maximum ozone concentrations calculated using DEOM for two different days during an ozone episode in Europe in September 1999. The left figures are the reference run without using data assimilation. The right figures are the corresponding result including the data assimilation technique based on the setup used in 9). The upper figures show the situation on September 7, 1999 and the lower figure show the situation on September 12, 1999. Especially around the first episode (top figures) the effect of data assimilation is seen, while during the second episode, the ozone field is less changed by the algorithm.

4.7 Conclusions

The first results with the data assimilation routine based on Statistical Interpolation have been conducted with the DEOM model. Nine different experiments including different setup of the data assimilation algorithm were defined and tested against measurements and ranked according to the performance. The results from the experiments have shown that the data assimilation routine together with a CTM are beneficial for obtaining better performance of the short term ozone forecasts using the CTM model. Improvement in the correlation coefficients in the range of 0.1 to 0.27 between the reference and scenario (9) were seen. Additionally, there were significant reductions of bias and NMSE.

References

Balgovind, R., Dalcher, A., Ghil, M., and Kalnay, E.: A Stochastic- Dynamic Model for the Spatial Structure of Forecast Error Statistics, Monthly Weather Review, 111, 701–722, 1983.

Blond, N. and R.Vautard: Three-dimensional ozone analyses and their use for short-term ozone forecasts, Journal of Geophysical Research, 109, 2003.

Blond, N., L.Bel, and R.Vautard: Three-dimensional ozone data analysis with an air quality model over the Paris area, Journal of Geophysical Research, 108, 4744, 2003.

Bouttier, F. and Courtier., P.: Data assimilation concepts and methods., Tech. rep., ECMWF training course lecture notes, 1998.

Brandt, J., J. H. Christensen, L. M. Frohn, F. Palmgren, R. Berkowicz and Z. Zlatev, "Operational air pollution forecasts from European to local scale". *Atmospheric Environment*, Vol. 35, Sup. No. 1, pp. S91-S98, 2001a

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, Physics and Chemistry of the Earth,Part B: Hydrology,Oceans and Atmosphere, 26, 781–786, 2001b.

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, Physics and Chemistry of the Earth,Part B: Hydrology,Oceans and Atmosphere, 26, 825–830, 2001c.

Brandt, J., J. H. Christensen, L. M. Frohn and R Berkowicz, 2003: “Air pollution forecasting from regional to urban street scale – implementation and validation for two cities in Denmark”. *Physics and Chemistry of the Earth,* Vol. 28, pp. 335-344, 2003.

Daley, R. Atmospheric data assimilation. 1996. Cambridge University Press. Cambridge Atmospheric and Space Science series. Ref Type: Serial (Book,Monograph)

Elbern, H. and Schmidt, H.: A four-dimensional variational chemistry data assimilation scheme for Eulerian chemistry transport modeling, Journal of Geophysical Research-Atmospheres, 104, 18 583–18 598, 1999.

Elbern, H., Schmidt, H., and Ebel, A.: Variational data assimilation for tropospheric chemistry modeling, Journal of Geophysical Research-Atmospheres, 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, Environmental Modelling & Software, 15, 539–548, 2000.

Frydendall, J., 2006 “Development, implementation and testing of a simple chemical data assimilation algorithm”. Niels Bohr Institute, University of Copenhagen, pp. 98.

Frydendall, J. and J. Brandt, 2006: “Development and implementation of a simple data assimilation algorithm”., pp. 8. Paper from NMR Workshop on Data Assimilation held at NILU, December, 2005.

Frydendall, J., J. Brandt and J. H. Christensen, 2007: “Implementation and testing of a data assimilation algorithm in the operational air pollution forecast system, THOR”. In preparation for *Atmospheric Chemistry and Physics*, 2007

Gaspari, G. and Cohn, S. E.: Construction of correlation functions in two and three dimensions, Quarterly Journal of the Royal Meteorological Society, 125, 723–757, 1999.

Hesstvedt, E., Hov, O., and Isaksen, I. S. A.: Quasi-Steady-State Approximations in Air-Pollution Modeling - Comparison of Two Numerical Schemes for Oxidant Prediction, International Journal of Chemical Kinetics, 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, Physics and Chemistry of the Earth Part B-Hydrology Oceans and Atmosphere, 26, 807–812, 2001.

Hollingsworth, A. and Lonnberg, P.: The Statistical Structure of Short-Range Forecast Errors As Determined from Radiosonde Data .1. the Wind-Field, Tellus Series A-Dynamic Meteorology and Oceanography, 38, 111–136, 1986.

Ide, K., Courtier, P., Ghil, M., and Lorenc, A. C.: unified Notation for Data Assimilation: operational, Sequential and Variational, J. Met. Soc. Japan, Special Issue on Data Assimilation in Meteorology and Oceanography: Theory and Practice, 181–189, 1997.

Khattatov, B., Jean-Francois, L., Lawrence, V. L., Richard, M., Pieternel, L., Xuexi, T., Guy, P. B., and John, C. G.: Assimilation of satellite observations of long-lived chemical species in global Atmos. Chem. Phys., 0000, 0001–14, 2006 www.atmos-chem-phys.org/acp/0000/0001/

Khattatov, B., L.Lyjak, and J.Gille: On applications of photochemical models to the design of measurement strategies, Geophysical Research Letters, 28, 2377–2380, 2001.

Lamarque, J. F., Khattatov, 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, Journal of Geophysical Research-Atmospheres, 104, 26 209–26 218, 1999.

McRae, G. J., Goodin, W. R., and Seinfeld, J. H.: Numerical-Solution of the Atmospheric Diffusion Equation for Chemically Reacting Flows, Journal of Computational Physics, 45, 1–42, 1982.

M´enard, R. and Chang, L. P.: Assimilation of Stratospheric Chemical Tracer Observations Using a Kalman Filter. Part II: chi2- Validated Results and Analysis of Variance and Correlation Dynamics, Monthly Weather Review, 128, 2672–2686, 2000.

M´enard, R., Cohn, S. E., Chang, L. P., and Lyster, P. M.: Assimilation of Stratospheric Chemical Tracer Observations Using a Kalman Filter. Part I: Formulation, Monthly Weather Review, 128, 2654–2671, 2000.

P.F.Levelt, Khattatov, B., J.C.Gille, G.P.Brasseur, X.X.Tie, and J.W.Waters: Assimilation of MLS ozone measurements in the global three-dimensional chemistry transport model ROSE, Geophysical Research Letters, 25, 4493–4496, 1998.

Tilmes, S., Brandt, J., Flatoy, F., Bergstrom, R., Flemming, J., Langer, J., Christensen, J. H., Frohn, L. M., Hov, O., Jacobsen, I., Reimer, E., Stren, R., and Zimmermann, J.: Comparison of five Eulerian air pollution forecasting systems for the summer of 1999 using the german ozone monitoring data, Journal of Atmospheric Chemistry, 42, 91–121, 2002.

van Loon, M. and A.W.Heemink: Kalman filtering for nonlinear atmospheric chemistry models: first experiences, Tech. Rep. MASR9711, Centrum voor Wiskunde en Informatica, 1997.

Zlatev, Z. & Brandt, J. 2006: Testing Variational Data Assimilation Modules. In: Lirkov, I., Margenov, S. & Wasniewski, J. (Eds.): Large-Scale Scientific Computing. 5th International Conference, LSSC 2005, Sozopol, Bulgaria, June 6-10, 2005. Springer. - Lecture Notes in Computer Science 3743: 395-402.

5. Using a set of test-examples to study some properties of variational data assimilation algorithms

Contribution from NERI

Abstract

Variational data assimilation algorithms can successfully be used in different fields of science and engineering. An attempt to utilize available sets of observations in the efforts to improve (i) the models used to study different phenomena and/or (ii) the model results can systematically be carried out when data assimilation algorithms are used.

The main idea, on which the variational data assimilation algorithms are based, is pretty general. A functional is formed by using a weighted inner product of differences of model results and measurements. The value of this functional is to be minimized. Forward and backward computations are carried out by using the model under consideration and its adjoint equations (both the model and its adjoint are defined by systems of differential equations). The major difficulty is caused by the huge increase of both the computational load (normally by a factor more than 100) and the storage needed. This is why it might be appropriate to apply some splitting procedure in the efforts to reduce the computational work.

Five test-examples have been created. Different numerical aspects of the data assimilation algorithms and the interplay between the major computational parts of any data assimilation algorithm (numerical algorithms for solving differential equations, splitting procedures and optimization algorithms) have been studied by using these tests. The presentation will include results from testing carried out in the study.

5.1 Basic ideas

Assume that observations are available at time-points $t_p, p \in \{1,2,...,P\}$. These observations can be taken into account in an attempt to improve in some sense the results obtained by a given model. This can be done by minimizing the value of the following functional (see [7]):

$$\text{(1)} \qquad J(c_0) = \frac{1}{2}\sum_{p=0}^{P} \langle W_p (\bar{c}_p - \bar{c}_p^{obs}), (\bar{c}_p - \bar{c}_p^{obs}) \rangle$$

where (a) the functional $J(c_0)$ is depending on the initial value c_0 of the vector of the concentrations at time t_0 (because the model results c_p depend on c_0), (b) $W(p)$ is a matrix containing some weights (it will be assumed here that $W(p)$ is the identity matrix, but some weights have to be used in all practical problems) and (c) $\langle , \rangle$ is an inner product in an appropriately defined Hilbert space (it will be assumed in this paper that the usual vector space is used, i.e. it is assumed that $\bar{c} \in \Re^s$ where s is the number of chemical species which are involved in the model).

An optimization algorithm has to be used in order to minimize the functional $J(c_0)$. Most of the optimization algorithms are based on the application of the gradient of $J(c_0)$. The adjoint equation of the model under consideration has to be derived and used in the calculation of the gradient of the functional $J(c_0)$. Most of the scientific and engineering models are described mathematically by systems of differential equations. Therefore the adjoint equations are also described by

systems of differential equations. This short analysis shows clearly that a data assimilation algorithm is a very complicated numerical procedure. The time and storage requirements are the major difficulty. Such a procedure consists of (i) a good optimization algorithm and (ii) good numerical algorithms for solving differential equations. In order to reduce the time and storage requirements it is also necessary (iii) to apply some good splitting technique.

5.2 Need for a good set of test-examples

The final aim is to apply the data assimilation technique to large-scale air pollution models for studying the transport of harmful air pollutants over Europe ([11], [12]). The ideas discussed here are very general and can successfully be applied in connection of many other models which lead (after some kind of semi-descretization) to stiff systems of ordinary differential equations (ODEs). Before applying a data assimilation algorithm to a given model it is necessary to check carefully (a) the correctness of its modules and (b) the efficiency of the numerical algorithms applied in the different modules. This can successfully be done only if good test-examples are available. The chemical part of an environmental model is normally the most time consuming part (and the most difficult one because it introduces stiffness in the model). This is why it is especially important to test carefully the correctness and the efficiency of the chemical part. The chemical part of an environmental model can be represented as a stiff system of ODEs:

$$\mathbf{(2)} \qquad \frac{d\bar{c}}{dt} = f(t,\bar{c}), \qquad \bar{c} \in \Re^s,$$

where vector $\bar{c}$ contains s components and function f is in general nonlinear. Five test-examples were devised (see [12]). We start with a very simple linear system. Then the complexity is gradually increased. The second test-example is a non-linear but autonomous system. The third test-example is a non-linear and non-autonomous system with a Jacobian matrix which does not depend explicitly on t. The fourth test-example is a non-linear and non-autonomous system with a Jacobian matrix which depends explicitly on t. The last test-example is a chemical scheme with 56 chemical species, which is really used in many environmental models ([11], [12]). It is described by a non-linear and non-autonomous system of ODEs. Both the right-hand-side function and the Jacobian matrix depend on t. It is not possible to express the dependence on t analytically, because some chemical rates depend on some quantities (as, for example, the temperature) which are dependent on the time variable. Analytical solution is not available, but a reference solution has been calculated with a time-stepsize $\Delta t = 10^{-5}$. The values of this solution were saved at the end of every period of 15 min. The so-found reference solution is used to check the accuracy achieved in different runs.

The first four examples are taken from the book of Lambert ([5]), while the fifth example is, as mentioned above, similar to the schemes used in the EMEP models (see [9], [11] and [12]).

5.3 Calculating the gradient of the functional

It is convenient to explain the basic ideas that are used when the gradient of $J(\bar{c}_0)$ is calculated by the following very simple example. Assume that observations are available only at five time-points: t_0, t_1, t_2, t_3 and t_4. The gradient of the functional can be calculated in the following way. Assume that some tool, *model*, by which the values of the unknown vectors $\bar{c}(t_0)$, $\bar{c}(t_1)$, $\bar{c}(t_2)$, $\bar{c}(t_3)$ and $\bar{c}(t_4)$ can be calculated, is available. The tool, the *model*, can be, for example, some air pollution model, but in some simpler cases *model* can simply refer to some solver for systems of PDEs or ODEs. Under this assumption, the calculations have to be performed, **for the particular example with** $P = 4$, in five consecutive steps.

- **Step 1.** Use the *model* to calculate $\bar{\mathbf{c}}_1$ (performing integration, in a forward mode, from time-point $\mathbf{t}_0$ to time-point $\mathbf{t}_1$). Calculate the adjoint variable $\bar{\mathbf{q}}_1 = \bar{\mathbf{c}}_1 - \bar{\mathbf{c}}_1^{\mathbf{obs}}$. Form the adjoint equation (corresponding to the *model* used in the forward mode; adjoint equations will be discussed in Section 5). Perform backward integration (by applying the adjoint equation) from time-point $\mathbf{t}_1$ to time-point $\mathbf{t}_0$ to calculate the vector $\bar{\mathbf{q}}_0^1$, where the lower index shows that $\bar{\mathbf{q}}_0^1$ is calculated at time-point $\mathbf{t}_0$, while the upper index shows that $\bar{\mathbf{q}}_0^1$ is obtained by using $\bar{\mathbf{q}}_1 = \bar{\mathbf{c}}_1 - \bar{\mathbf{c}}_1^{\mathbf{obs}}$ as an initial vector in the backward integration.

- **Step 2 to Step 4.** Perform the same type of calculations, as those in Step 1 to obtain $\bar{\mathbf{q}}_0^2$, $\bar{\mathbf{q}}_0^3$ and $\bar{\mathbf{q}}_0^4$. More precisely, the following operations are to be carried out for $\mathbf{p} = \mathbf{2, 3, 4}$:

  (a) use the forward mode to proceed from time-point $\mathbf{t}_{\mathbf{p-1}}$ to time-point $\mathbf{t}_\mathbf{p}$,

  (b) form the adjoint variable $\bar{\mathbf{q}}_\mathbf{p} = \bar{\mathbf{c}}_\mathbf{p} - \bar{\mathbf{c}}_\mathbf{p}^{\mathbf{obs}}$,

  (c) use the adjoint equation in a backward mode from time-point $\mathbf{t}_\mathbf{p}$ to time-point $\mathbf{t}_0$ to calculate $\bar{\mathbf{q}}_0^\mathbf{p}$.

- **Step 5.** The sum of the vectors $\bar{\mathbf{q}}_0^1$, $\bar{\mathbf{q}}_0^2$, $\bar{\mathbf{q}}_0^3$, $\bar{\mathbf{q}}_0^4$ obtained in Step 1 to Step 4 and vector $\bar{\mathbf{q}}_0^0 = \bar{\mathbf{c}}_0 - \bar{\mathbf{c}}_0^0$ gives an approximation to the required gradient of the functional $\mathbf{J}\{\mathbf{u}_0\}$.

It is clear that the above procedure can easily be extended for any number $\mathbf{P}$ of time-points at which observations are available.

The gradient of the functional $\mathbf{J}\{\mathbf{u}_0\}$ is calculated by performing one forward step from time-point $\mathbf{t}_0$ to time-point $\mathbf{t}_\mathbf{P}$ and $\mathbf{P}$ backward steps from time-points $\mathbf{t}_\mathbf{p}, \mathbf{p} \in \{\mathbf{2,...,P}\}$, to time-point $\mathbf{t}_0$. This explains the main idea, on which the data assimilation algorithms are based, in a very clear way, but it is expensive when $\mathbf{P}$ is large. In fact, the computational work can be reduced, performing only once the backward calculations (see, for example, [1] or [7]).

5.4 Solving the system of ODEs

Six numerical algorithms for solving stiff systems of ODEs have been used in the experiments. The algorithms selected by us are listed below:

- the Backward Euler Algorithm,
- the Implicit Mid-point Rule,
- a Second-order Modified Diagonally Implicit Runge-Kutta Algorithm,
- a Fifth-order Three-stage Fully Implicit Runge-Kutta Algorithm,
- a Second-order Two-stage Rosenbrock Algorithm,
- the Trapezoidal Rule.

The Implicit Mid-point Rule and the Trapezoidal Rule are **A-stable** algorithms. All the other algorithms are **L-stable**. More details about the selected numerical algorithms and their properties can be found in [3], [4], [5] and [10].

5.5 Solving the adjoint equations

It is necessary to distinguish between linear models and non-linear models when the adjoint equations are formed and treated numerically. Assume that the model is linear and, furthermore, that the model is written in the following general form:

$$\mathbf{(3)} \qquad \frac{d\bar{c}}{dt} = A\bar{c}.$$

Denote by $\bar{q}$ the adjoint variable. Then the adjoint equation can be written as

$$\mathbf{(4)} \qquad \frac{d\bar{q}}{dt} = -A^{*}\bar{q},$$

where A^{*} is the conjugate operator of A. If the problem is discretized by using some numerical algorithm, then operator A will be represented by a matrix which is normally also denoted by A. If the adjoint equation is discretized, then the transposed matrix A^{T} will appear in the discretized version of (4).

Consider now a non-linear model:

$$\mathbf{(5)} \qquad \frac{d\bar{c}}{dt} = B(\bar{c}).$$

The adjoint equation of the model presented in (5) can be written as

$$\mathbf{(6)} \qquad \frac{d\bar{q}}{dt} = -\left[B'(\bar{c})\right]^{*}\bar{q},$$

where $B'(\bar{c})$ is obtained by differentiation of B. In the discrete case, we will have the transposed matrix of the Jacobian of B in (6).

It is seen that the adjoint equations are always linear; compare (4) and (6). However, the right-hand-side in the linear case does not depend on the *model* variable $\bar{c}$. In the non-linear case this is not true. The right-hand-side of (6) depends on $\bar{c}$. This fact has serious implications: the values of $\bar{c}$ calculated during the forward mode (when the *model* is treated) are to be saved and used when the adjoint equation is handled (when the backward mode is carried out).

If the chemical scheme (2) is considered, then (6) can be rewritten as

$$\mathbf{(7)} \qquad \frac{d\bar{q}}{dt} = -\left[\frac{\partial f(t,\bar{c})}{\partial \bar{c}}\right]^{T}\bar{q}.$$

It is clear now that the numerical algorithms from the previous section can easily be adapted for the adjoint equation (7) of the chemical scheme (2). For example, the application of the Backward Euler Algorithm in connection with adjoint equation (7) leads to the following formula for the backward computations:

$$\textbf{(8)} \qquad \bar{q}_n = \bar{q}_{n+1} - \Delta t \left[- \frac{\partial f(t_n, \bar{c}_n)}{\partial \bar{c}_n} \right]^T \bar{q}_n .$$

The fact that the adjoint equation is used in the backward mode is taken into account when (8) is derived.

5.6 Application of splitting procedures

The application of data assimilation algorithms leads to very time-consuming problems (the computer time may be increased by a factor up to 100 and even more). Therefore splitting, which is commonly used during the treatment of large-scale environmental models, is even more needed when these are used together with data assimilation techniques. The test-examples, which are discussed in Section 2, were treated both without splitting and with by four splitting procedures: (i) sequential splitting, (ii) symmetric splitting, (iii) weighted sequential splitting and (iv) weighted symmetric splitting. Much more details about different splitting procedures can be found in [12].

The splitting of each of the first four test-examples is not very critical. Let us consider as an example the splitting applied in connection with the second test-example. The operator on the right-hand-side of this example is $f_1 = y_2$, $f_2 = y_2(y_2 - 1)/y_1$ (where y_1 and y_2 are the components of vector $\bar{c}$. It is split into two operators: (a) $f_1^{(1)} = 0$, $f_2^{(1)} = -y_2^{(1)} / y_1^{(1)}$ and (b) $f_1^{(2)} = y_2^{(2)}$, $f_2^{(2)} = (y_2^{(2)})^2 / y_1^{(2)}$. The sum of these two operators is equal, component-wise, to the original operator in the right-hand-side of the second test-example (i.e. $f_1^{(1)} + f_1^{(2)} = f_1$ and $f_2^{(1)} + f_2^{(2)} = f_2$).

It is not very obvious how to split the fifth test-example. We grouped in the first sub-model the species which react with ozone. The remaining species formed the second sub-model.

At each time-step during the forward mode the splitting was carried out as usual (see, for example, [11] and [12]). At each time-step during the backward mode the splitting operators are applied in reverse order (compared with the order applied in the corresponding forward time-step).

5.7 Minimizing the functional

The problem of minimizing the functional (1) is an **unconstrained** optimization problem. Therefore, the subroutine E04DGF from the NAG Library, which performs unconstrained optimization, has been used in the beginning ([12]). However, we realized very quickly that **it is better to impose some constraints.** There are often physical reasons for doing this (in the chemical scheme, for example, the concentrations of the chemical species should be kept non-negative). Therefore, the next choice was subroutine E04KDF also from the NAG Library. This is a rather flexible subroutine. It requires simple bounds for the variables of the functional. It is quite reasonable to assume that such bounds could always be derived in real-life problems (by using the physical properties of the studied processes).

The problem with subroutine E04KDF is not the determination of the bounds for the variables, but rather the necessity to scale the model, which is very often a rather difficult task. Unfortunately, such a requirement is, to our knowledge, common for all optimization algorithms.

5.8 Numerical results

Only the ability of the data assimilation algorithms **to improve the initial values** of the solution was tested numerically. This is important for forecasting high pollution levels. However, the data assimilation algorithms can also be used for many other purposes (see, [1], [2], [6], [8], [12]).

A perturbation parameter α was introduced. The values of the initial solution were always perturbed by using ten different values of α (introducing relative errors of $5\%, 10\%, \ldots, 50\%$ in the initial values). Data assimilation is used to improve the initial values. The improved initial values are then used to calculate the solution over an increased time-interval. The analytical solution (the reference solution for the fifth test-example) is used to evaluate the relative error, component-wise, at the end of each time-step (each period of 15 min. for the fifth test-example). The max-norm of the vector or relative errors found over the whole time-interval is taken and used in the comparisons of the results from the different runs.

Each test-example has been run with the six numerical algorithms and the five splitting procedures (including here also the case where no splitting is used). Furthermore, for the first four test-examples we start with a time-stepsize $\Delta t = 0.25$ and carry out successively 18 additional runs (every time reducing the time-stepsize by a factor of two). For the fifth test-example we start with a time-stepsize $\Delta t = 150$ and carry out successively 10 additional runs (reducing again the time-stepsize by a factor of two every time when a new run is started).

The results from many runs ([12]) show that (i) reducing the time-stepsize leads to a reduction of the error according to the order of the combined algorithm (numerical algorithm + splitting procedure), (ii) if the time-stepsize is sufficiently small then the error obtained with the data assimilation algorithm is practically the same as the error obtained by using exact initial values without data assimilation (which means that the results are optimal in some sense), (iii) the numerical algorithms that are only A-stable (the Implicit Mid-Point Rule and the Trapezoidal Rule; see [3] and [5]) have difficulties for large time-stepsizes when the stiff chemical scheme is to be handle and (iv) if no splitting is used, then it might be more efficient in some cases to use high-order algorithms (the Fifth-order Three-stage Fully Implicit Runge-Kutta Algorithm performed better, for all five test-examples, than the other algorithms when no splitting was used).

It should be emphasized here that the stability problems, which were mentioned in (iii), disappear when some splitting procedures are used ([12]). Since the chemical scheme is a rather general and sufficiently large problem, this fact indicates that the splitting procedures may have some stabilizing effect when stiff systems of ODEs are to be handled.

Some results, which were obtained in the efforts to improve the initial value and the accuracy of the ozone component in the scheme with 56 chemical species, are presented in Table 5.1. The notation can be explained in the following way: (a) $\mathbf{ERROR\_0\_P}$ is giving the relative error in the perturbed initial condition, (b) $\mathbf{ERROR\_0\_I}$ is giving the relative error in the improved initial condition, (c) $\mathbf{ERROR\_F\_P}$ is giving the global relative error obtained by using the perturbed initial condition, (d) $\mathbf{ERROR\_F\_I}$ is giving the global relative error obtained by using the improved initial condition.

Table 5.1. Numerical results obtained when the chemical scheme with 56 compounds is run. The Backward Euler Algorithm is used without splitting. The initial value of the ozone concentration is perturbed by a factor $\alpha = 0.5$.

Steps	ERROR_0_P	ERROR_F_P	ERROR_0_I	ERROR_F_I
1008	0.47	0.48	2.0E-03	2.4E-03
2016	0.49	0.50	1.0E-03	1.2E-03
4032	0.47	0.47	5.0E-04	6.1E-4
8064	0.48	0.48	2.5E-04	3.2E-04
16128	0.46	0.48	1.3E-04	1.7E-04
32256	0.49	0.48	6.3E-05	8.8E-05

It is clearly seen that reducing the stepsize (i.e. multiplying the number of time-steps by a factor of two) leads to a reduction of both the initial guess and the global error by a factor of two. This is precisely the expected behaviour (because the Backward Euler Algorithm is of order one). Much more numerical results might be found in [12]. **Recently an extended series of tests with a two-dimensional advection-diffusion-chemistry model has been run. Paper, describing these results, is in preparation.**

5.9 Conclusions

The results from several thousand runs indicate that the data assimilation modules are able to improve the initial values of the solution if (a) the numerical algorithms used are sufficiently accurate and (b) the initial perturbations are not very large.

On the other hand, the results indicate also that both the computing time and the storage needed are increased by a factor which is very often greater than 100. Therefore, it is necessary (i) to continue the search for faster but still sufficiently accurate numerical algorithms, (ii) to apply faster computers, (iii) to exploit efficiently the cache memory and (iv) to parallelize the codes.

References

1. Elbern, H., Schmidt, H.: A four-dimensional variational chemistry data assimilation scheme for Eulerian chemistry transport modelling. Journal of Geophysical Research, **104** (1999) 18583-18598.
2. Elbern, H., Schmidt, H., Talagrand, O., Ebel, A.: 4D-variational data assimilation with an adjoint air quality model for emission analysis. Environmental Modelling & Software, **15** (2000) 539-548.
3. Hairer, E., Wanner, G.: Solving Ordinary Differential Equations, II: Stiff and Differential-algebraic Problems. Springer, Berlin-Heidelberg-New York-London, 1991.
4. Hundsdorfer, W., Verwer, J. G.: Numerical solution of time-dependent advection-diffusion-reaction equations. Springer, Berlin-Heidelberg-New York-London, 2003.
5. Lambert, J. D.: Numerical Algorithms for Ordinary Differential Equations. Wiley, Chichester-New York-Brisbane-Toronto-Singapore, 1991.
6. Le Dimet, F.-X., Navon, I. M., Daescu, D. N.: Second order information in data assimilation. Monthly Weather Review, **130** (2002) 629-648.
7. Lewis, J. M., Derber, J. C.: The use of adjoint equations to solve a variational adjustment problem with advective constraints. Tellus, **37A** (1985) 309-322.
8. Sandu, A., Daescu, D. N., Carmichael, G. R., Chai, T.: Adjoint sensitivity analysis of regional air quality models. Journal of Computational Physics, 2005; to appear.
9. Simpson, D., Fagerli, H., Jonson, J. E., Tsyro, S. G., Wind, P., Tuovinen, J-P.: Transboundary Acidification, Eutrophication and Ground Level Ozone in Europe, Part I. Unified EMEP Model Description. EMEP/MSC-W Status Report 1/2003. Norwegian Meteorological Institute, Oslo, Norway, 2003.
10. Zlatev, Z.: Modified diagonally implicit Runge-Kutta algorithms. SIAM Journal on Scientific and Statistical Computing, **2** (1981) 321-334.
11. Zlatev, Z.: Computer treatment of large air pollution models. Kluwer Academic Publishers, Dordrecht-Boston-London, 1995.

12. Zlatev, Z. and Dimov, I.: Computational and numerical challenges in environmental modeling. Elsevier, Amsterdam-Boston-Heidelberg-London-New York-Oxford-Paris-San Diego-San Francisco-Singapore-Sydney-Tokyo, 2006.

6. International workshop on data assimilation with regional scale atmospheric chemical transport models

The workshop on 'Data assimilation in regional scale atmospheric chemistry models' was held at NILU on 15 November, 2005, halfway through the project. More than 20 participants attended the one day workshop from the institutes involved and 8 scientific presentations were given. Two invited speakers, Hendrik Elbern and Arnold Heemink, who are experts in data assimilation methods also attended, providing valuable input and discussion. The report from the workshop has already been submitted as part of the first years reporting (Denby et al., 2006).

6.1 Summary

This workshop was organized as the first activity of the NMR funded project on 'Data assimilation in regional scale atmospheric chemical models'. The aim of the workshop was to establish links between the participating institutes (NILU, DMU, SMHI, met.no) and plan and co-ordinate future activities. Presentations by all the institutes were given to establish the methodologies currently employed, the level of expertise and the future research intentions of the participating institutions. In addition two invited speakers attended the workshop, Henrik Elbern and Arnold Heemink, who are acknowledged experts in the field of data assimilation in chemical transport modelling. Their attendance was vital to help place the work in a European perspective and for their critical appraisal and first hand knowledge of the techniques currently employed.

The workshop was held at NILU on 15 November 2005. 8 separate presentations were given with a large amount of time devoted to discussion (see attached agenda). In total up to 21 people participated in the workshop. 11 from the participating institutes who are directly involved with the project, 2 invited speakers and a number of interested parties from both NILU and met.no. A list of participants is also included. Discussions ranged from the very technical to the philosophical with a number of recommendations for methodologies and problem solving being discussed.

The presentations from the project participants have been consolidated for this report, which will be used as reference for further development and cooperation. The presentations from the invited speakers have been summarized, with the presented slides contained in an appendix. At the end of each presentation is a table containing some of the discussion points brought up during the meeting.

6.2 Workshop agenda

09:30	Introduction and welcome *Bruce Denby, NILU*
Satellite data	
09:45	The GEMS project *Leonor Tarrason, Met.no*
10:00	Availability of Satellite Remote Sensing images of Atmospheric Species *Martin Hvidberg, DMU*
Variational methods	
10:30	Development and implementation of a simple data assimilation algorithm *Jan Frydendall, DMU*
11:00	Applying variational data assimilation for an atmospheric chemical scheme *Zahari Zlatev, DMU*
11:30	COFFEE BREAK
11:45	Application of 2-dimensional variational data analysis in MATCH *Michael Kahnert, SMHI*
12:15	Implementation and performance experiences with chemical 4Dvar assimilation *Hendrik Elbern, EURAD, Cologne*

12:45 LUNCH
13:30 Discussion

Ensemble methods

14:30 An introduction to Sequential Importance Resampling
Sam Erik Walker, NILU
15:00 Data assimilation in atmospheric chemistry models using ensemble methods
Arnold Heemink, TU Delft
15:30 Discussion

Project discussion

16:30 NMR project: Conclusions to be drawn from this workshop and future work
All
17:00 End workshop

References

Denby, B., Brandt, J., Elbern, H., Frydendall, J., Heemink, A., Hvidberg, M., Kahnert, M., Tarrason, L., van Loon, M., Walker, S. E. and Zlatev, Z. (2006) Data assimilation in regional scale atmospheric chemical models. NMR Workshop at NILU, Kjeller Norway, 15 November 2005. Denby B. (ed.). Kjeller (NILU OR 43/2006).URL: www.nilu.no/data/inc/leverfil.cfm?id=22718&type=6

Norwegian Institute for Air Research (NILU)

P.O. Box 100, N-2027 Kjeller, Norway

An institute in CIENS and in the Environmental Research Alliance of Norway

REPORT SERIES	REPORT NO. OR 30/2007	ISBN 978-82-425-1953-5 (print) 978-82-425-1954-2 (electronic ISSN 0807-7207	
SCIENTIFIC REPORT			
DATE	SIGN.	NO. OF PAGES 47	PRICE NOK 150,-

TITLE	PROJECT LEADER
Development and application of data assimilation in regional scale atmospheric chemistry models	Bruce Denby
	NILU PROJECT NO. O-105077
AUTHOR(S) Bruce Denby, Michael Kahnert, Jørgen Brandt, Jan Frydendall and Zahari Zlatev	CLASSIFICATION * A
	CONTRACT REF. Project number 05FOX10

REPORT PREPARED FOR
Nordic Council of Ministers

ABSTRACT
This report presents the work and activities carried out during the Nordic Council of Ministers funded project 'Development and application of data assimilation in regional scale atmospheric chemical models' in the period

2005 - 2006. The project aim was to further develop and consolidate methodologies used in the Nordic countries for data assimilation in regional scale atmospheric chemical modelling.

NORWEGIAN TITLE

KEYWORDS		
Data assimilation	Atmospheric chemistry models	Nordic

ABSTRACT (in Norwegian)
[Skriv abstract på norsk]

* Classification
- A Unclassified (can be ordered from NILU)
- B Restricted distribution
- C Classified (not to be distributed)