NILU: OR 43/2006

NILU: OR 43/2006
REFERENCE: O-105077
DATE: JUNE 2006
ISBN: 82-425-1766-5

Data assimilation in regional scale atmospheric chemical models

NMR Workshop at NILU, Kjeller, Norway, 15 November 2005

Editor: Bruce Denby

norden
Nordic Council of Ministers

NILU: OR 43/2006
REFERENCE: O-105077
DATE: JUNE 2006
ISBN: 82-425-1766-5

Data assimilation in regional scale atmospheric chemical models

NMR Workshop at NILU, Kjeller, Norway, 15 November 2005

Editor: Bruce Denby
Contributing authors: Jørgen Brandt, Hendrik Elbern, Jan Frydendall, Arnold Heemink, Martin Hvidberg, Michael Kahnert, Leonor Tarrason, Maarten van Loon, Sam Erik Walker and Zahari Zlatev

Preface

This workshop is part of a Nordic Council of Ministers, Ocean and Air Group (NMR/HLG) funded project which supports the development and communication of scientific research between Nordic countries on the topic of 'data assimilation in regional scale atmospheric chemistry models'. The four institutes involved, NILU (Norway), met.no (Norway), DMU (Denmark) and SMHI (Sweden), all have active programmes in data assimilation. The intention of this project and workshop is to bring together these institutes to share knowledge and experience within a Nordic context and to further support development in this research area. In total 21 people attended the workshop, including invited experts in data assimilation from Europe. The workshop showed itself to be successful, being both informative and helpful to the participants. This report consolidates the presentations and discussions that took place during the workshop.

Contents

Data assimilation in regional scale atmospheric chemical models

NMR Workshop at NILU, Kjeller, Norway, 15 November 2005

1 Introduction – Bruce Denby

Data assimilation is a term referring to the various methods used to combine monitoring 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, that can also be utilized.

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.

The final part of the meeting was used to plan further activity of the project. It was decided that in 2006 a Nordic assimilation dataset should be compiled that will contain the relevant observational data for all the participating institutes including ground based, satellite and other remote sensed observations. This will lay the foundations for any further intercomparative assimilation studies and will allow the participating groups to cooperate closely on a single project. It will also

facilitate the development of knowledge in regard to the requirements of good observational data (for assimilation purposes) including the need for an understanding of information related to uncertainty, spatial and temporal representativeness and data capacity requirements. Compilation of the dataset will also aid in identifying gaps in the available data and will allow each institute to gain from the others expertise in their own particular field.

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

16:30 NMR project
Conclusions to be drawn from this workshop. Report from the workshop.
Next year? Intercomparison of methods?Another workshop?

17:00 End workshop

Participants list*

Consortium Institutes

NILU

Bruce Denby (co-ordinator)	*bde@nilu.no*
Yvan Orsolini	*orsolini@nilu.no*
Sam Erik walker	*sew@nilu.no*
Sverre Solberg	*sso@nilu.no*
Alena Bartonova	
Cathrine Lund Myhre	
Kjetil Tausend	
Agnes Dudek	
Tove Marit Svendby	
Caroline Forster	
Aasmund Fahre Vik	

Met.no

Maarten van Loon	*Maarten.v.Loon@met.no*
Leonor Tarrason	*Leonor.Tarrason@met.no*
Hanne Heiberg	

DMU

Jørgen Brandt	*jbr@dmu.dk*
Zahari Zlatev	*zz@dmu.dk*
Jan Frydendall	*jfr@dmu.dk*
Martin Hvidberg	*mhv@dmu.dk*

SMHI

Michael Kahnert	*michael.kahnert@smhi.se*

Invited speakers

Hendrik Elbern, EURAD	*he@eurad.Uni-Koeln.DE*
Arnold Heemink, TU Delft	*A.W.Heemink@ewi.tudelft.nl*

*Participants with email address included are actively involved in the assimilation project

Satellite data

2 The GEMS project and data assimilation with the Unified EMEP model – Leonor Tarrason

Maarten van Loon and Leonor Tarrason
Norwegian Meteorological Institute (met.no)
P.O. Box 43 Blindern 0313 Oslo, Norway
Email: maartenvl@met.no

The GEMS project

Large part of the data assimilation activities at met.no will be carried out in the EU funded Integrated Project GEMS (**G**lobal and regional **E**arth-system **M**onitoring using **S**atellite and in-situ data) within the Sixth Framework Programme. GEMS will create a new European operational system for operational global monitoring of atmospheric chemistry and dynamics and an operational system to produce improved medium-range and short-range air-chemistry forecasts, through much improved exploitation of satellite data. (see also the GEMS website: http://www.ecmwf.int/research/EU_projects/GEMS/).

The research teams involved will develop a global operational medium-range forecast / assimilation capability for dynamics and composition, exploiting all available satellite data.

The integrated forecast / assimilation capability will provide a powerful monitoring capability for greenhouse gases, reactive gases and aerosols. Sophisticated new inversion methods will be developed to infer surface fluxes of CO2 and other species through use of the surface flask data with the gridded atmospheric fields on transport and composition. The GEMS project will produce global retrospective analyses of the atmospheric dynamics and composition for the troposphere and stratosphere, and will be able to assess the impact of changes both on global and regional scale, examining extremes as well as means.

The global forecasts will provide key information on long-range transport of air pollutants to the regional forecast models, through the forecast boundary conditions used by the regional systems. The improved regional forecasts will be used by air-quality authorities at city level, in dozens of cities across Europe.

The contribution of met.no in this project is in the regional part, where the Unified EMEP model will be used as regional model, fed at the boundaries by global predicted fields. Also in the regional simulation data assimilation will be applied and hence one of the major tasks within GEMS is the development of data assimilation modules within the EMEP modelling system.

Data assimilation around the EMEP model

For the purpose of GEMS a Kalman Filter technique will be implemented around the EMEP model. The choice for this technique is based on existing experience with this kind of techniques within the EMEP team. Apart from this practical

argument, building a 4D-Var system would require the construction of the adjoint code of the EMEP model, which is a far from trivial task. The presentations and discussions at the workshop only confirmed this conclusion. In support of GEMS a project proposal has been submitted to the Norwegian Space Centre (NRS). In this proposal the emphasis is on treatment of satellite products by the remote sensing group at met.no. It is intended to directly assimilate observed radiances from space into the EMEP model. Usually, a derived product – aerosol optical depth (AOD) – is used for assimilation. Directly assimilating radiances will have the advantage that no assumptions need to be made on the composition and vertical distribution of the aerosols as is necessary for retrieving AOD values.

Discussion

Speaker	Comment
Zahari Zlatev	Communication between computers in real time must be difficult. Will you use grid computing in this project?
Leonor Tarrason	No this is not part of the project, just being able to communicate between the different databases is a priority in the project.
Hendrik Elbern	It may be interesting to start an initiative on this though! Grid computing is an interesting issue on a longer time perspective. It must be ensured however that, for routine applications, timely delivery is ensured.
Michael Kahnert	What is the main aim of the project? Is it to do an analysis (as a post-processing of CTM results) or to develop a forecasting capability (data assimilation)
Leonor Tarrason	The main goal of GEMS is to develop the operational capacity to forecast air quality in global and regional scale, using data assimilation techniques

3 Availability of Satellite Remote Sensing images of Atmospheric Species – Martin Hvidberg

Martin Hvidberg
National Environmental Research Institute (NERI)
Dept. of Atmospheric Environment (ATMI)
Roskilde, Denmark

Several people have inquired about an overview of "What atmospheric substances can be seen, measured, distinguished, by satellite remote sensing". This paper is an introductory presentation of what is available.
Remote sensing methods can be derived into categories in a number of ways. For the purpose of this work we are limiting ourselves to remote sensing of the Earth's atmosphere, and so excluding observations of the Earth's surface as well as observations of astronomical objects. In particular we are going to focus on the troposphere, since that is where we live. Though tropospheric conditions have influence an surface concentrations in general, and in modeling of these in particularly, we will not go into detail on this type of data even when it is provided by the satellite or sensor system. In addition we are going to limit the overview to passive remote sensing, specifically excluding active microwave systems and occultation GPS viewing systems.
The overview is based mainly on knowledge generated by the European network of excellence "ACCENT" and especially on work of the University of Bremen, Univ. Heidelberg, Univ. Toronto, Univ. Cambridge and KNMI in the Netherlands.

1. Viewing geometry

One of the main characteristics when selecting an atmospheric remote sensing product is the viewing geometry. Essentially there exist three systems: Nadir view, Occultation view and Limb view. Most satellite borne sensor systems uses one of these viewing geometries, but a few uses several in combination. The figure below illustrates these geometries and the main instruments using these geometries.

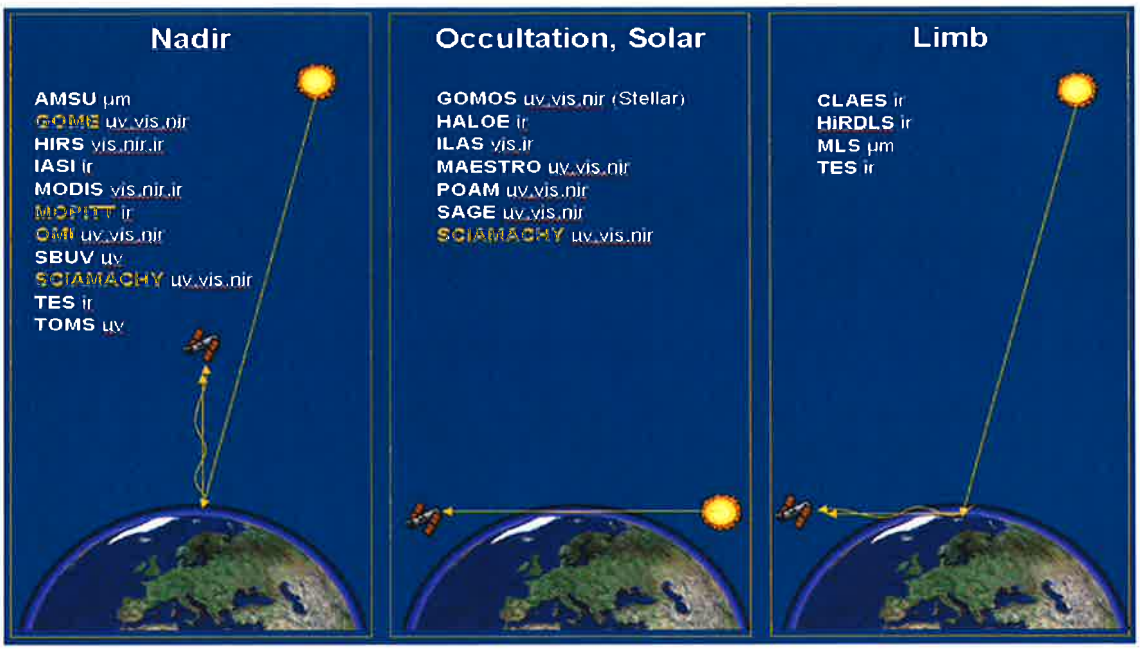


Figure 3.1: NERI – Martin Hvidberg

2. Wavelength region

Another important difference between sensor systems is the wavelengths at which they operate. There are three main areas used in atmospheric remote sensing, namely UltraViolet-VISible (UV_VIS), InfraRed (IR) and Microwave (μm).

3. Strategies

Various strategies are possible when designing a sensor system. It is less demanding to construct an instrument with the capability to measure "just" the total column than to build one with the ability to measure a vertical profile. The number of species looked for can also be limited by the selected design strategy. In the following section the most dominant combinations of viewing geometry, wavelength, etc. are introduced.

4. Nadir viewers

Nadir viewers utilize the backscattered electromagnetic radiation from the sun. Most frequently these techniques are used in the UV-VIS region. It can either be used in a dual wavelength reflectance ratio as technique, or in a multi wavelength technique.

Figure 3.2

For total column measurements
To measure a specific chemical species, measurements of incoming and outgoing radiation are made, to determine
total amount of that species. Two pairs of measurements are made. One measurement of incoming UV light and one measuring backscattered UV light, at a wavelength that is strongly absorbed by the chemical compound of interest.
A second pair of measurement of incoming and reflected radiation is made at a wavelength that is weakly absorbed
by the same species. The differences between the pairs of measurements at the two wavelengths are used to infer the amount present in the atmosphere of the given chemical compound.

For vertical profile measurements
Information on the vertical structure of the atmosphere can
be derived using the backscatter profiling technique in the
UV wavelength area. The atmosphere less absorbs light at
longer wavelengths than light at shorter wavelengths. Such longwave UV light is able to penetrate far into the atmosphere.

The backscattered radiation at specific UV wavelengths can only be scattered from above a particular height. Below this level, all the radiation is absorbed and there is no ackscattered radiance. This allows us to make a vertical measurement of a given species. Measurements at certain UV wavelengths are sensitive to specific portions
of the vertical profile.

Figure 3.3

5. Occultation viewers

Occultation view is looking at the sun, the moon or a star, through the atmosphere. This viewing geometry is utilizing the differences in atmospheric absorption spectrum. Looking at specific wavelengths that are known to be strongly absorbed by a particular chemical species it is possible to measure the presence of that given species, somewhere along the line of sight from the light source to the sensor.
Occultation techniques can be used within UV, VIS or IR wavelength areas.
Making this type of measurements while the satellite rises or sets behind the horizon enables the measurement of vertical profiles of the atmosphere.
The viewing geometry of occultation viewers severely limits the time and duration during which observations are possible. Both the satellite and the source of light have to be in the right place to make this technique possible.

Figure 3.4

6. Limb viewers

The limb viewing geometry again uses the scattered or emitted spectrum, rather than the absorption spectrum. The difference from the occultation techniques is that in Limb view the light source is not at the end of its line of sight. Limb viewers have, like occultation, a line of sight that is more or less at a tangent to the Earth.
The limb viewing technique is not limited to any specific wavelength of sunlight. Scattering techniques are used with UV, VIS and NIR and techniques based on emitted light are used with IR and Microwave.
This technique works best with ozone; however other trace gases like water vapor, nitrogen dioxide, and sulfur dioxide and aerosols are also measurable.
Compared to occultation, limb is less dependent on the position of both sun, Earth and satellite and can therefore collect data through considerably more hours every day.

Figure 3.5

Techniques utilizing emitted radiation, so called Limb emission or Limb viewing techniques, operating in the IR or Microwave areas do not even need the sun to be present. All objects with a temperature above absolute zero emit radiation, the wavelengths of this emitted radiation is characteristic for each atom and each molecule. Instruments based upon the limb emission technique infer amounts of gases from measurements radiation from different altitudes of the atmosphere.
In theory, the instrument could observe all the way to the surface, but below a certain altitude (under 10 km), clouds interfere with the emitted longwave radiation. The limb emission sensors are able to create a vertical profile of trace gas concentrations. The resulting vertical resolution is quite good, usually on the order of 3 kilometers.

Figure 3.6

7. Selected instruments

There exist close to 50 different sensor systems whose primary purpose is to monitor the Earth's atmosphere. Only a limited subset of these is relevant to the present research. Some are outdated technologically or maybe not even in orbit any more, and therefore only apply to long timeseries studies, e.g. TOMS. Some are not yet even released to the broader scientific community, e.g. OMI. In practice a short list of the sensor systems seems to be the once used by most research groups within atmospheric monitoring. The following, in no way exhaust the possibilities but are just a short presentation of a few, frequently used, instruments.

8. GOME

GOME and GOME-2 are Nadir viewing UV backscatter spectrometers, which means that they measure Earthshine spectra, that is: the sunlight, which is reflected back into space by molecules in the atmosphere and by the surface. The instrument also measures the solar spectrum directly. The ratio between the Earthshine and solar signal is a measure of the reflectivity of the Earth's atmosphere and surface.

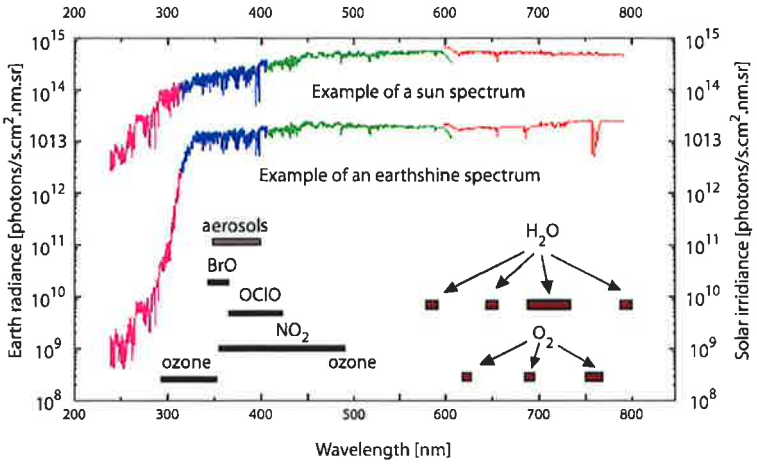
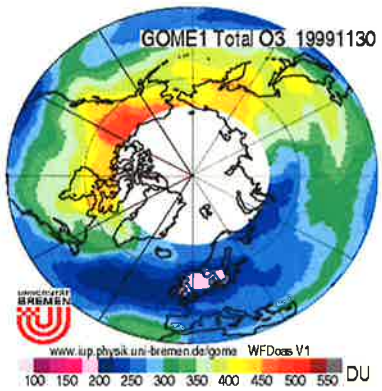


Figure 3.7. KNMI



GOME has been flying on ERS2 since 21 April 1995
Currently (Nov.'05) GOME-1 data are available from 28th June 1995 until 21st June 2003 (full global coverage). More recent data are currently being reprocessed and will be online very soon. The GOME-2 instrument is due to launch in 2006 aboard ESA's and EUMETSAT's first MetOp platform.

Figure 3.8: Mark Weber - Univ. Bremen

9. SCIAMACHY

SCIAMACHY was launched on March 2002, on Envisat. It is a spectrometer designed to measure sunlight, transmitted, reflected and scattered by the Earth's atmosphere or surface in the ultraviolet, visible and near infrared wavelength region
Data are available on request for 1. Jan 2003 till present from Univ. Heidelberg, Univ. Bremen, KNMI and through GMES-service.
The SCIAMACHY primary mission objective is to perform global measurements of trace gases in the troposphere and in the stratosphere.

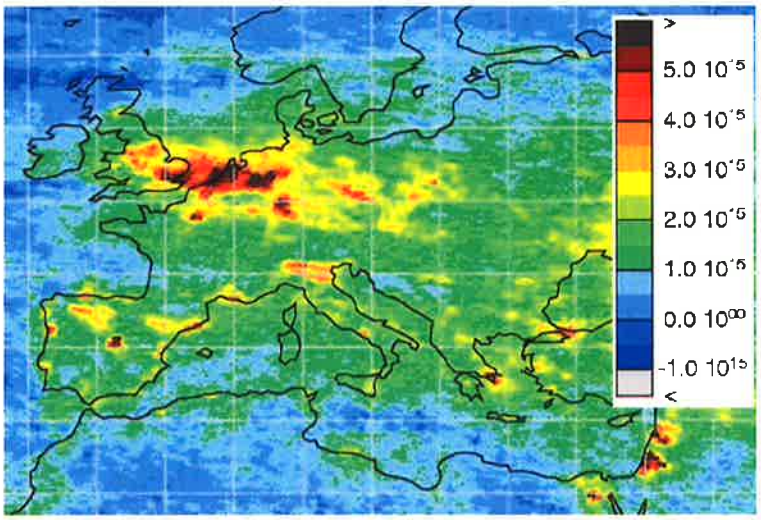


Figure 3.9: ***GOME column NO_2.*** *Tropospheric NO2 over Europe August 2002 . Units: Vertical Column [molec cm-2]. Andreas Richter - Univ. Bremen*

SCIAMACHY measurements yield the amounts and distribution of O_3, BrO, OClO, ClO, SO_2, CH_2O, NO_2, CO, CO_2, CH_4, H_2O, N_2O, pressure, Temperature, aerosol, radiation, cloud cover and cloud top height. A special feature of SCIAMACHY is the combined limb-nadir measurement mode, which enables the tropospheric column amounts of several trace gases to be determined. SCIAMACHY uses the same wavelengths in the UV-Vis as GOME-1 and -2 and has a spectral range extended into the infrared.

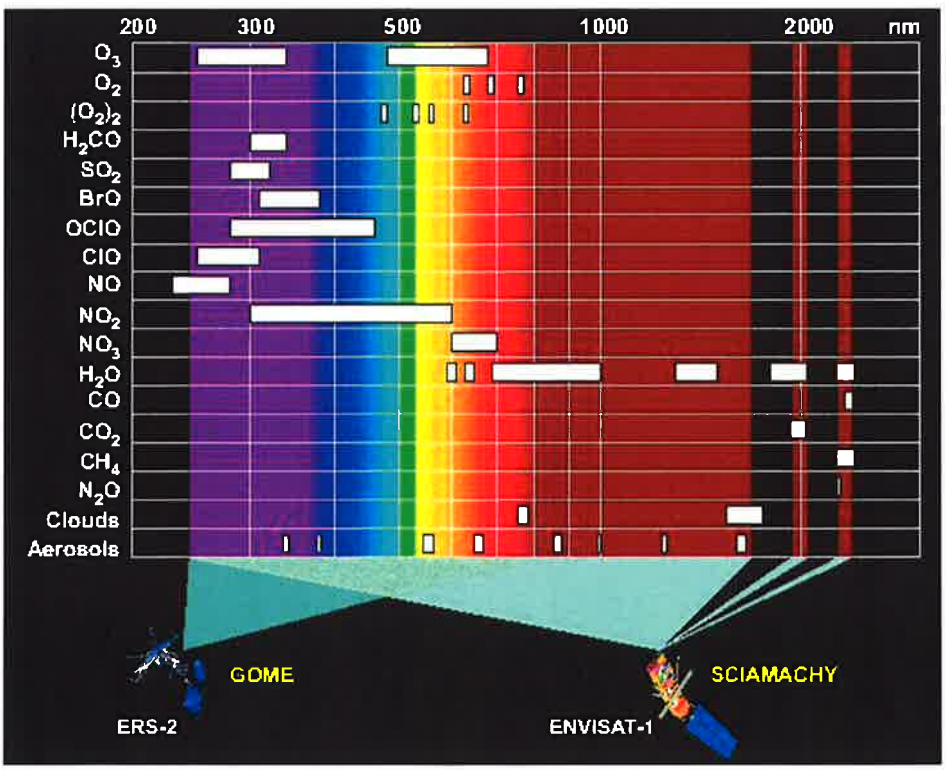


Figure 3.10: www-iup.physik.uni-bremen.de/sciamachy/

11. MOPITT

Measurements Of Pollution In The Troposphere (MOPITT) is a Canadian instrument launched in 1999 and consists of a IR Nadir looking sensor, aboard NASA's Terra satellite. It uses gas correlation spectroscopy to determine the abundance of carbon monoxide in the troposphere. The MOPITT sensor measures emitted and reflected radiance from the Earth in three spectral bands.
MOPITT data are available online as quick looks. Column total or concentration at 6 pre-defined heights.
The actual data on:"Derived CO levels", "gridded daily averages", "gridded monthly means", from 3. March 2000 till present are available upon request.

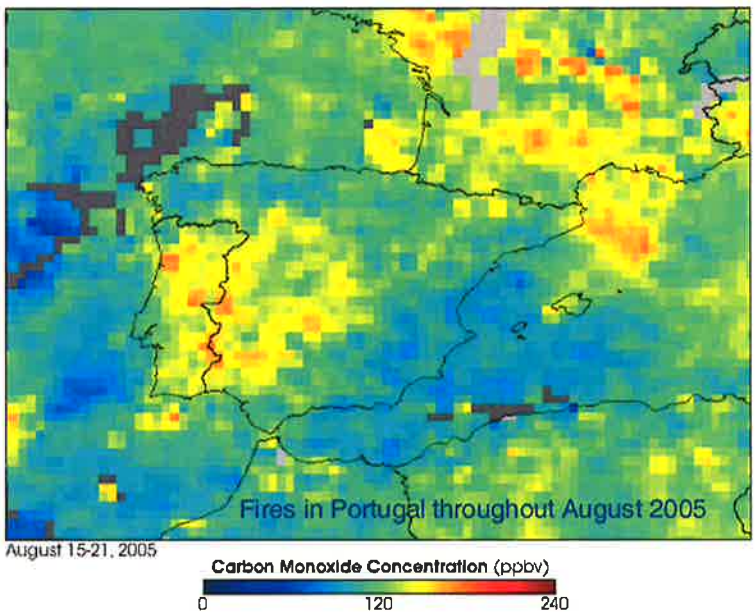


Figure 3.11: NASA - earthobservatory.nasa.gov/

12. A promising new satellite – Aura

On Thursday, July 15, 2004 Aura was launched at 6:01:59 a.m. local time from Vandenberg Air Force Base, aboard a Delta II rocket, later inserting the Aura satellite into a 705 kilometer orbit.
This completes the trilogy of satellites Terra, Aqua and Aura, the first series of NASA EOS satellites. While Terra monitors land, and Aqua monitors the Earth's water cycle, Aura will help understand the atmospheric system, global air quality, ozone recovery and climate change.
Each of Aura's four instruments, the Ozone Monitoring Instrument (OMI), the Tropospheric Emission Spectrometer (TES), the High Resolution Dynamics Limb Sounder (HIRDLS), the Microwave Limb Sounder (MLS) is designed to survey different aspects of Earth's atmosphere. Aura will survey the atmosphere throughout the troposphere and the lower stratosphere.

13. OMI

The OMI instrument employs hyper spectral imaging in a push-broom mode to observe solar backscatter radiation in the visible and ultraviolet. The hyper spectral capabilities will improve the accuracy and precision of the total ozone amounts. OMI will facilitate continuity in measurements from predecessors TOMS, SBUV, GOME, SCIAMACHY and GOMOS

Key air quality observations are O_3, NO_2, SO_2, BrO, OClO, and aerosol characteristics. The OMI instrument will distinguish between aerosol types, such as smoke, dust, and sulfates, and can measure cloud pressure and coverage, which provides data to derive tropospheric ozone concentrations.

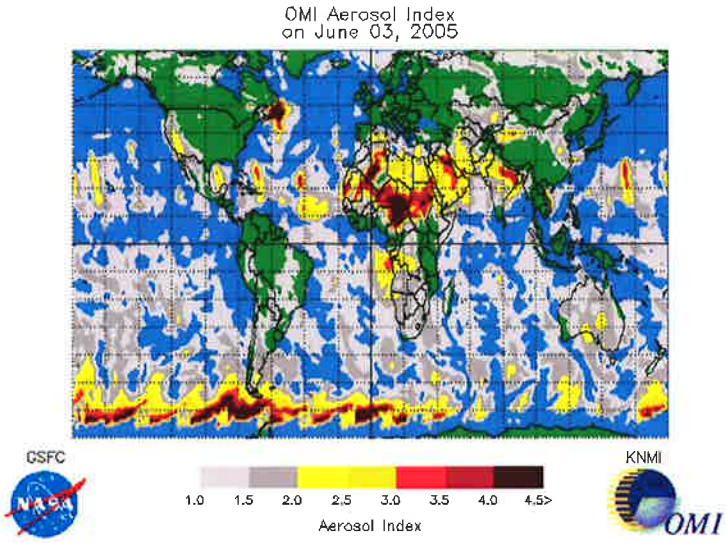


Figure 3.12: Ellen Brinksma, KNMI

14. TES

TES is a spectrometer that measures the infrared-light emitted by Earth's surface and by gases and particles in Earth's atmosphere. Spectrometers measure this radiation as a means of identifying the substances.
TES operates in a combination of limb and nadir mode. It generates three-dimensional profiles on a global scale of virtually all infrared-active species from Earth's surface to the lower stratosphere.

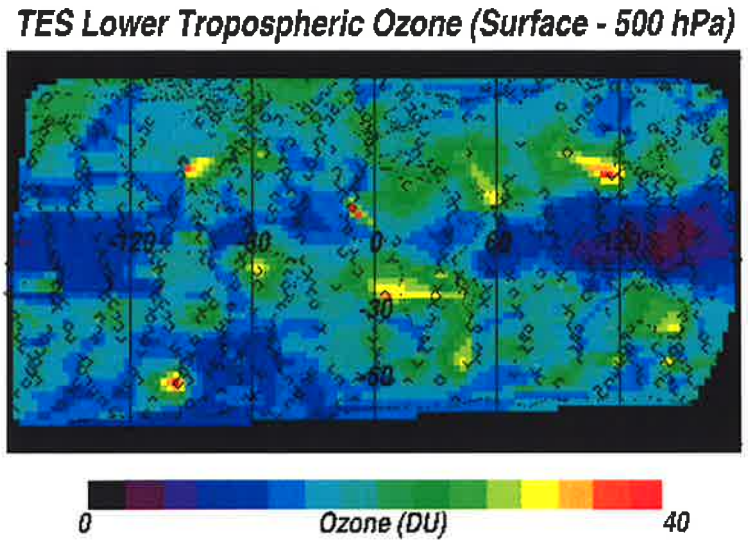


Figure 3.13: NASA, Jet Propulsion Lab. - tes.jpl.nasa.gov

15. HIRDLS

The High Resolution Dynamics Limb Sounder (HiRDLS) is a Scanning infrared limb sounder.
The HIRDLS instrument will obtain profiles over most of the globe, both day and night. Complete Earth coverage can be obtained in twelve hours.
It observes global distribution of temperature and concentrations of O_3, H_2O, CH_4, N_2O, NO_2, HNO_3, N_2O_5, CFC11, CFC12, ClONO2, and aerosols in the upper troposphere, stratosphere, and mesosphere

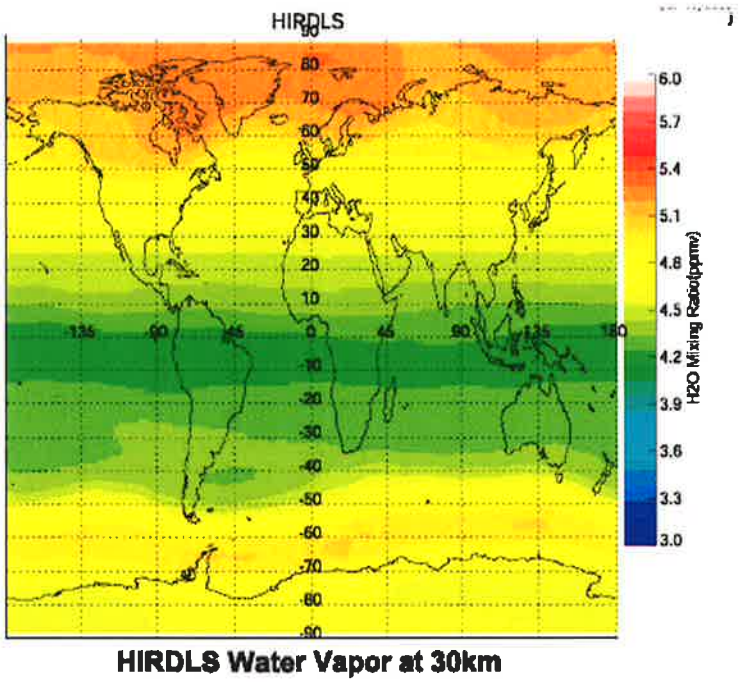


Figure 3.14: National Center for Atmospheric Research – www.eos.ucar.edu/hirdls/

16. MLS

Microwave Limb Sounder (MLS) is a passive microwave limb sounding radiometer / spectrometer. It measures thermal emission from the atmospheric limb.
The EOS MLS instrument will provide measurements of many chemical species involved in the destruction of stratospheric ozone. This instrument is a greatly enhanced version of the UARS MLS instrument, including use of latest technology to measure important species such as OH, BrO and many others which could not be measured by MLS at the time the UARS instrument was developed, as well as more precise measurements and measurements over a larger altitude range. The EOS Aura orbit will allow MLS measurements to be made to high latitudes every day on each orbit, whereas the UARS orbit required MLS high-latitude coverage to switch, approximately monthly, between the northern and southern hemispheres with critical periods being missed.

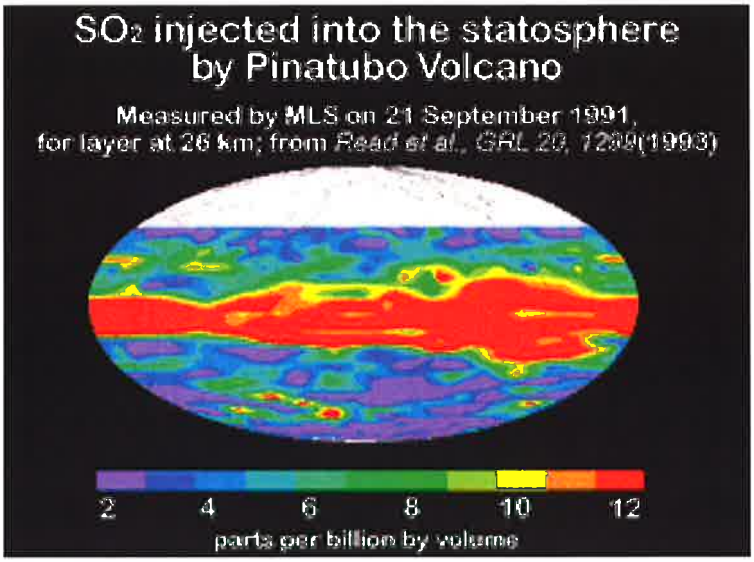


Figure 3.15: [Read et al., 1993] Dr. William Read. @mls.jpl.nasa.gov

17. Data Availability

Unfortunately after now 500 days in space, data from AURA are still to become available to the scientific community. Still there are high expectations for the usefulness and quality of these data when they become accessible to atmospheric scientists and others.

18. References:

aura.gsfc.nasa.gov/instruments/
envisat.esa.int/instruments/images/scia_heitran.html
www-iup.physik.uni-bremen.de/gome/wfdoas/
www-iup.physik.uni-bremen.de/sciamachy/
www.ccpo.odu.edu/SEES/ozone/oz_class.htm
www.esa.int/esaME/index.html
www.knmi.nl/gome_fd/doc/gomeintro.html

Acknowledgements:

Andreas Richter - Univ. Bremen, IUP, personal correspondence.
Ellen Brinksma - KNMI, personal correspondence.
Jim Drummond, Univ. Toronto personal correspondence.
Nicholas Savage – Univ. Cambridge, personal correspondence and proof reading.

Discussion

Speaker	Comment
Caroline Forster	Can you do the different types of measurements, e.g. Nadir, limb, with the same satellite?
Martin Hvidberg	Yes this is possible with some of the Satellites
Kjetil Tausend	How do you deal with the fact that the retrieval algorithm includes calibration with observational data? Is there a conflict of some form with data assimilation with the same data?

Michael Kahnert	I don't think there is a conflict as data assimilation can also be done in observational space.
General	A general discussion followed related to the representation of assimilated satellite data and land based data that is not independent from one another (Due to the fact that satellite data is not totally independent of the ground based data).

Variational methods

4 Development and implementation of a simple data assimilation algorithm – Jan Frydendall

Jan Frydendall and Jørgen Brandt
National Environmental Research Institute, Department of Atmospheric Environment
Frederiksborgvej 399, P.O. box 358, DK-4000 Roskilde, Denmark

Abstract

A simple algorithm for chemical data assimilation has been developed. The algorithm has been tested and implemented in the Eulerian chemical transport model, DEOM, used since 1999 for regional air pollution forecasting at NERI. DEOM is a part of the THOR integrated air pollution forecasting and management system (http://thor.dmu.dk). The data assimilation algorithm is shortly described and preliminary results from comparisons of model results with and without the data assimilation algoritm for a six months period in 1999 are shown.

The statistical interpolation algorithm

At NERI we wanted to get a deeper understanding of the data assimilation techniques. We wanted to understand what made these techniques work. Therefore, we did not start developing a very complicated data assimilation technique like the 3D/4D variational method or the extended Kalman filter. We have chosen the statistical interpolation technique for this study. It is fairly simple and yet it still gives a good insight to data assimilation.

1. The basics

Lets define the statistical interpolation scheme:

Statistical assumptions: The background error and the analysis error is defined as

$$\varepsilon_b = \mathbf{x}_b - \mathbf{x}_t \tag{1}$$

$$\varepsilon_a = \mathbf{x}_a - \mathbf{x}_t \tag{2}$$

and the observation error is defined

$$\varepsilon_o = \mathbf{y} - \mathbf{H}(\mathbf{x}_t) \tag{3}$$

From the background error we are now able to define the background covariance:

$$\mathbf{B} = \overline{(\varepsilon_b - \overline{\varepsilon}_b)(\varepsilon_b - \overline{\varepsilon}_b)^T} \tag{4}$$

the observation covariance:

$$\mathbf{R} = \overline{(\varepsilon_o - \overline{\varepsilon}_o)(\varepsilon_o - \overline{\varepsilon}_o)^T} \tag{5}$$

and the analysis covariance:

$$\mathbf{P}_a = \overline{(\varepsilon_a - \overline{\varepsilon}_a)(\varepsilon_a - \overline{\varepsilon}_a)^T} \tag{6}$$

- Linearized observation operator: the variation of the observation operator in the vicinity of the background state is linear. For any x close to $\mathbf{x}_b$, $\boldsymbol{H}(\mathbf{x})$ - $\boldsymbol{H}(\mathbf{x}_b) = \mathbf{H}(\mathbf{x} - \mathbf{x}_b)$, when $\mathbf{H}$ is a linear operator.
- Non-trivial errors: $\mathbf{B}$ and $\mathbf{R}$ are positive definite matrices.
- Unbiased errors: The expectation of the background and the observation errors is zero i.e.

$$\overline{\mathbf{x}_b - \mathbf{x}_t} = \overline{\mathbf{y} - \mathbf{H}(\mathbf{x}_b)} = 0 \tag{7}$$

- uncorrelated errors: observation and background errors are mutually uncorrelated i.e

$$\overline{(\mathbf{x}_b - \mathbf{x}_t)(\mathbf{y} - \mathbf{H}(\mathbf{x}_b))^T} = 0 \tag{8}$$

- If the background and the observation error p.d.f. are Gaussian, then $\mathbf{x}_a$ is also the maximum likehood estimator of $\mathbf{x}_t$.
- Linear analysis: we look for an analysis defined correction to the background which depend linearly on background observations departures.
- Optimum analysis: we look for an analysis state which is as close as possible to the true state in and root mean square sense. (i.e. it is a minimum variance estimate)

This leads to the statistical interpolations main equation:

$$\mathbf{x}_a = \mathbf{x}_b + \mathbf{K}(\mathbf{y} - \mathbf{H}(\mathbf{x}_b)) \tag{9}$$

$$\mathbf{K} = \mathbf{B}\mathbf{H}^T(\mathbf{H}\mathbf{B}\mathbf{H}^T + \mathbf{R})^{-1} \tag{10}$$

where the linear operator $\mathbf{K}$ is called the Kalman gain matrix of the analysis.

- In our setup we define the interpolation operator $\mathbf{H}$ as a linear interpolation between the grid locations and observations stations.
- The observation error covariance matrix as $\mathbf{R} = \sigma_o^2 \delta_i^j$ delta is the Kronecker delta and σ_o^2 is the error covariance of the observations.
- The background error covariance matrix is define as $\mathbf{B} = \sigma_b^2 \mathbf{f}(\mathbf{r})$ where σ_o^2 is the error covariance of the background and $f(r)$ is the correlation function.

The correlation function is define as

$$f(r) := \left(1 + \frac{|r|}{L}\right)\exp\left(-\frac{|r|}{L}\right) \tag{11}$$

r is the Euclidian distance in the model space and L is the correlation length. The function is depicted in the figure below

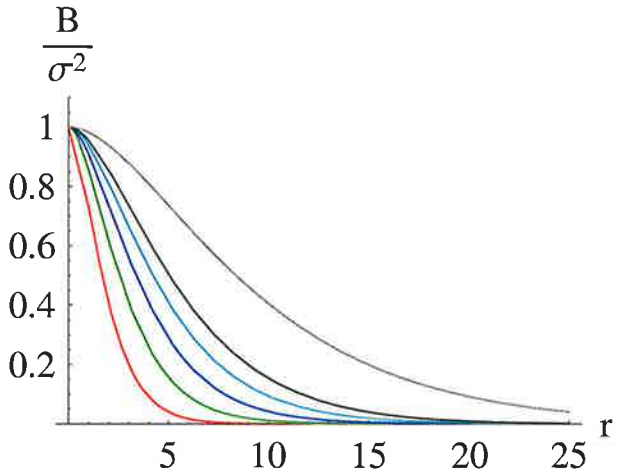


Figure 4.1: 1 Different correlations lengths are plotted. Red: L=1, Green: L=1.5, Blue: L=2, Azure: L=2.5, Black: L=3 and Grey: L=5. The greater the correlations length the slower is the descent of the of the correlations function

The error covariance determinations

In order to get fully specified error covariances matrices we have to do some analysis. The first and foremost method is the one described by (Hollingsworth and Lonnberg, 1986). It states that we have to make correlations between "observations minus background" separated by observation stations distances. We than have to fit a correlation model with the founded data. The results are depicted below:

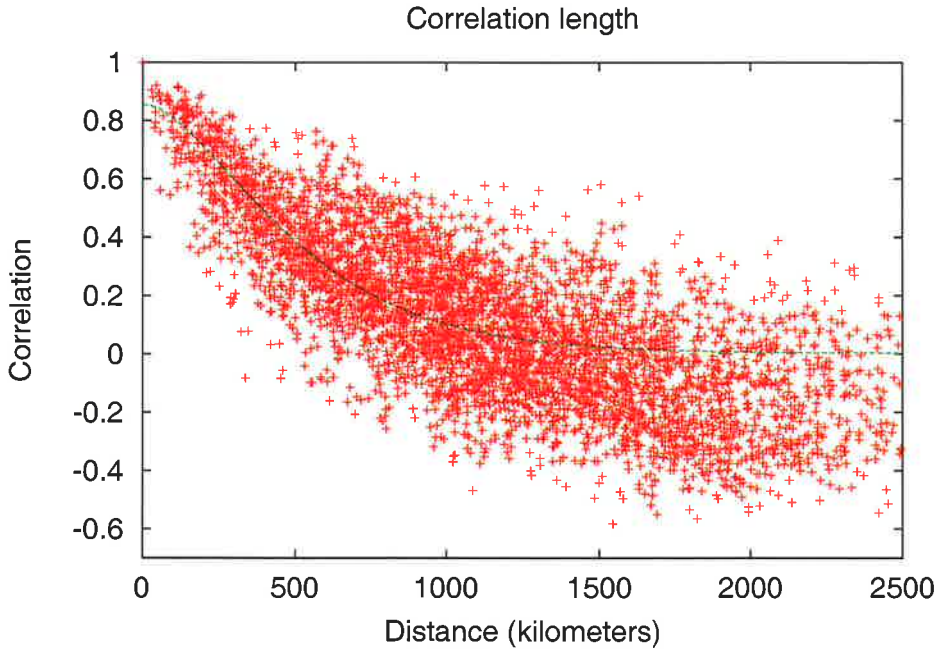


Figure 4.2: 2 The correlation function is fitted to the observation-minus-background correlation.

In order to get a good error covariance determination we had to use a large time series of six months to archive good results. This error covariance determination is very representative for a specific hour in the time series. Therefore we look at other ways to determine the error covariance for each run of the data assimilation algorithm. We looked at method derived by (Desroziers and Ivanov, 2001) where have found a iterative equation which would converge on the error covariance in a few steps if one could write the error covariance matrices as:

$$\mathbf{B} = \mathbf{e}_b\mathbf{B}$$
$$\mathbf{R} = \mathbf{e}_o\mathbf{R} \tag{12}$$

where e_b and e_o are scalars. Then the following algorithm should hole:

$$e_o(i+1) = 2\frac{J_o(\mathbf{x}_a(\mathbf{e}_o,\mathbf{e}_b)(i))}{Tr\left(\mathbf{I}-\mathbf{K}(\mathbf{e}_o,\mathbf{e}_b)(i)\mathbf{H}\right)} \tag{13}$$

$$e_b(i+1) = 2\frac{J^b(\mathbf{x}_a(\mathbf{e}_o,\mathbf{e}_b)(i))}{Tr\left(\mathbf{HK}(\mathbf{e}_o,\mathbf{e}_b)(i)\right)} \tag{14}$$

We did not have any success with the above algorithm yet, because the memory requirement is too large to handle for our computers. We currently are looking into ways of reducing the sizes of the error covariance matrices.

Some preliminary results

We have implemented the statistical interpolation algorithm into the DEOM model (Brandt et al., 2001). The DEOM model is a three-layer Eulerian atmospheric chemistry transport model, which is designed to predict ozone values as well as many other chemical species. The data assimilation algorithm is set to correct the background field at 10:00 UTC, 11:00 UTC and 1200 UTC every day in a period from April to September in 1999 and focus is on the daily maximum ozone values occurring in the late afternoon at the same days. As DEOM is a 3 layer model, where the lowest model layer is defined by the mixing height, the model can not describe the nocturnal values of ozone in the surface layer. Model results are compared to measurement data from all available station in the EMEP network. The correlation length is L = 270 km and covariances are those found by the Hollingsworth method. In the following figures (except figure 3), the left figures includes results obtained from the reference model run (without data assimilation) and the right figures includes results obtained from the analysis model run (including data assimilation).

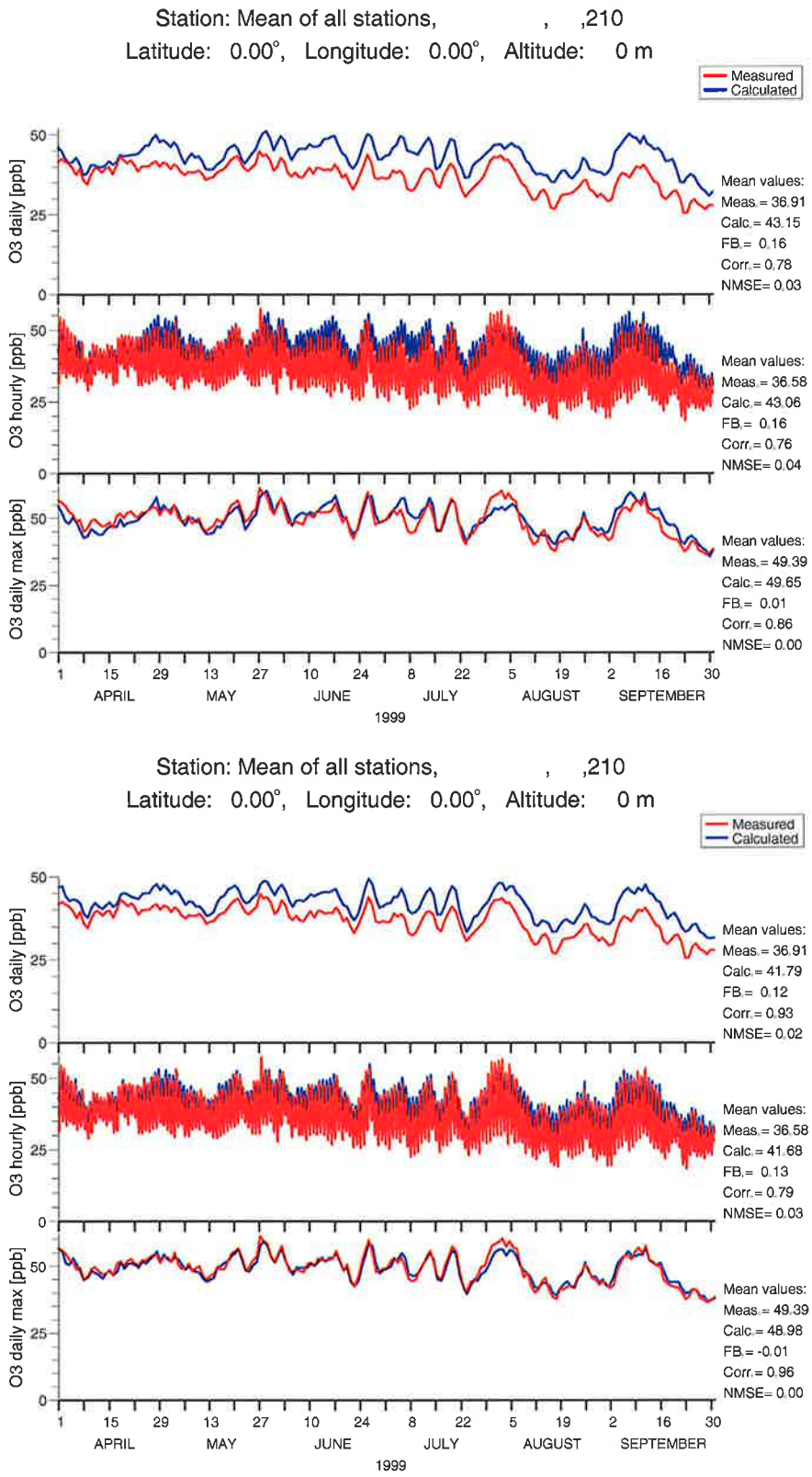


Figure 4.3:Comparison of calculated and measured daily mean, hourly and daily maximum ozone values taken as a mean over all measurement stations. One clearly sees the effect of the data assimilation process by looking at the top (without data assimilation) and bottom (with data assimilation) time series. The correlation coefficient increase and the bias and NMSE decrease

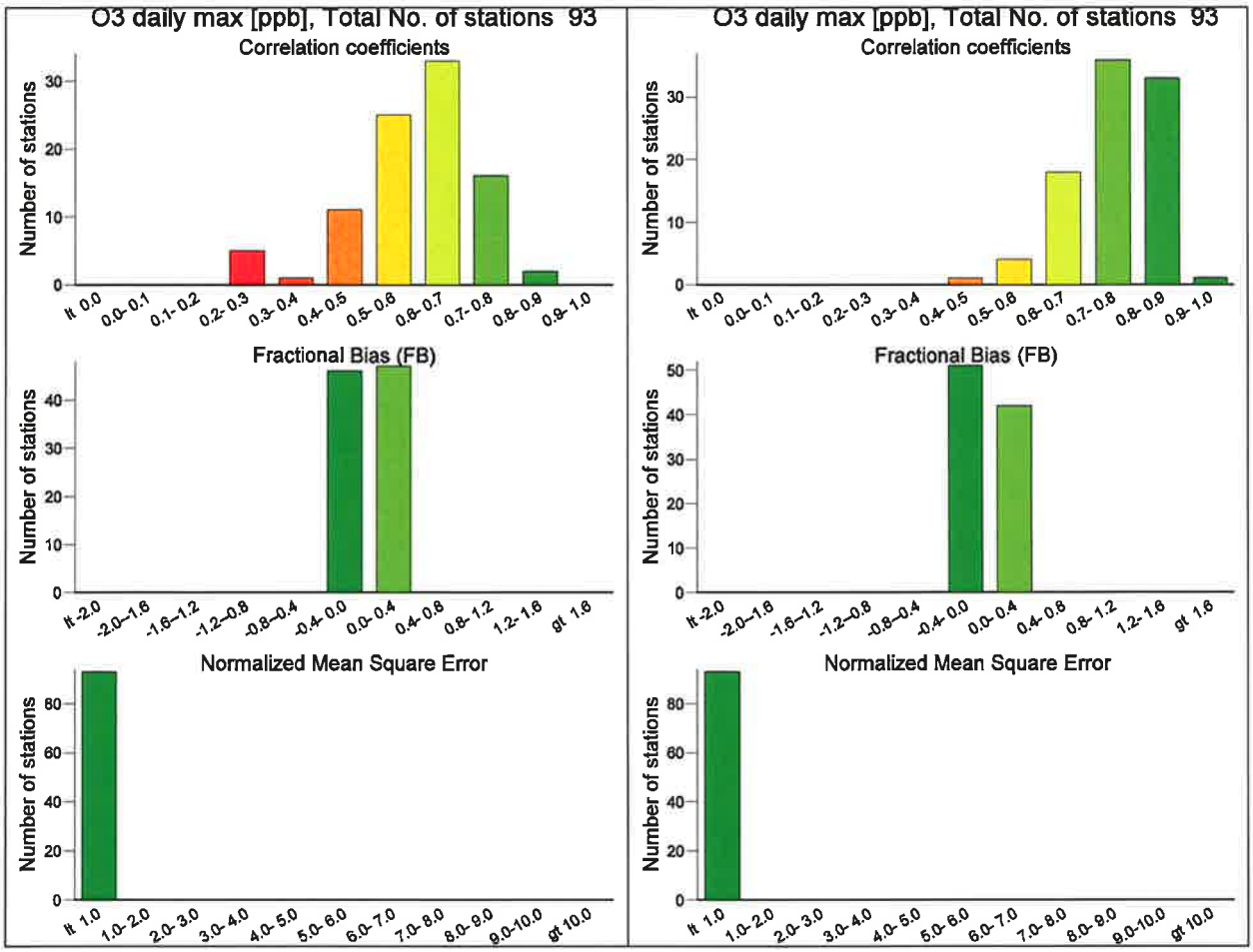


Figure 4.4 :4 Frequency distributions of the correlation coefficient, the fractional bias and the Normalised Mean Square Error estimated from calculated and measured data at EMEP measurement stations. The distribution of the correlation coefficient is generally shifted to the right by 0.2.

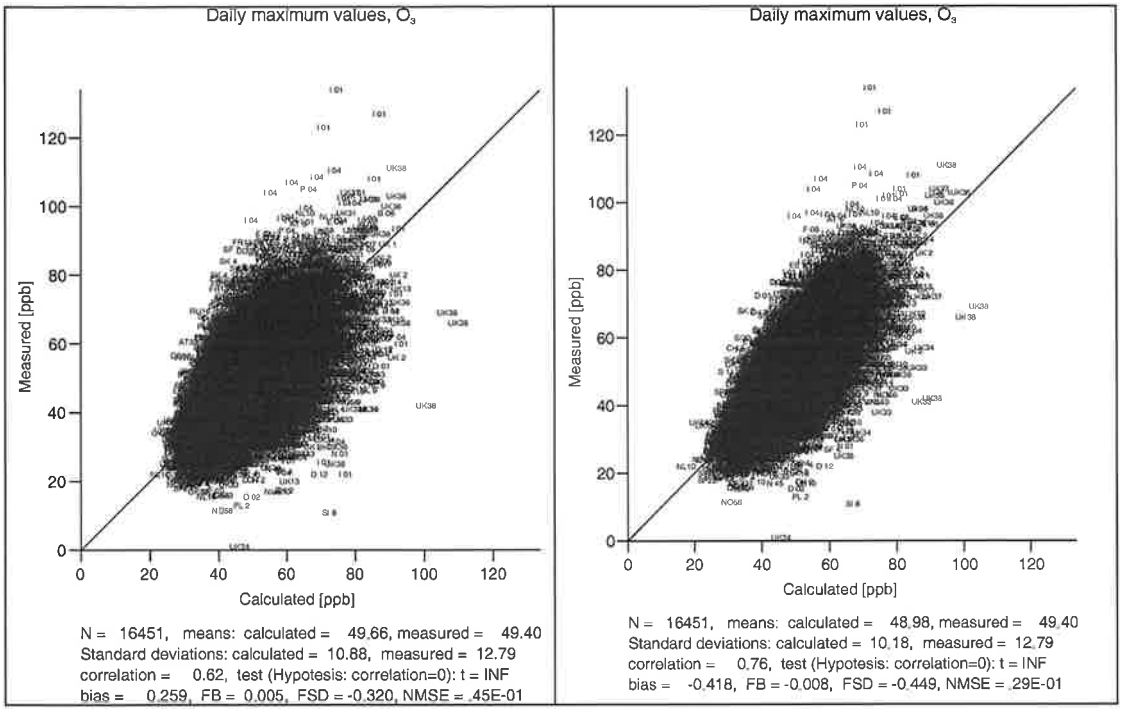


Figure 4.5: 5 Comparison of calculated and measured daily maximum values of ozone for all measurement stations for the period Apr.-Sep. 1999. The daily maximum values of ozone. The correlation coefficient increases from 0.62 to 0.76 when the data assimilation is applied

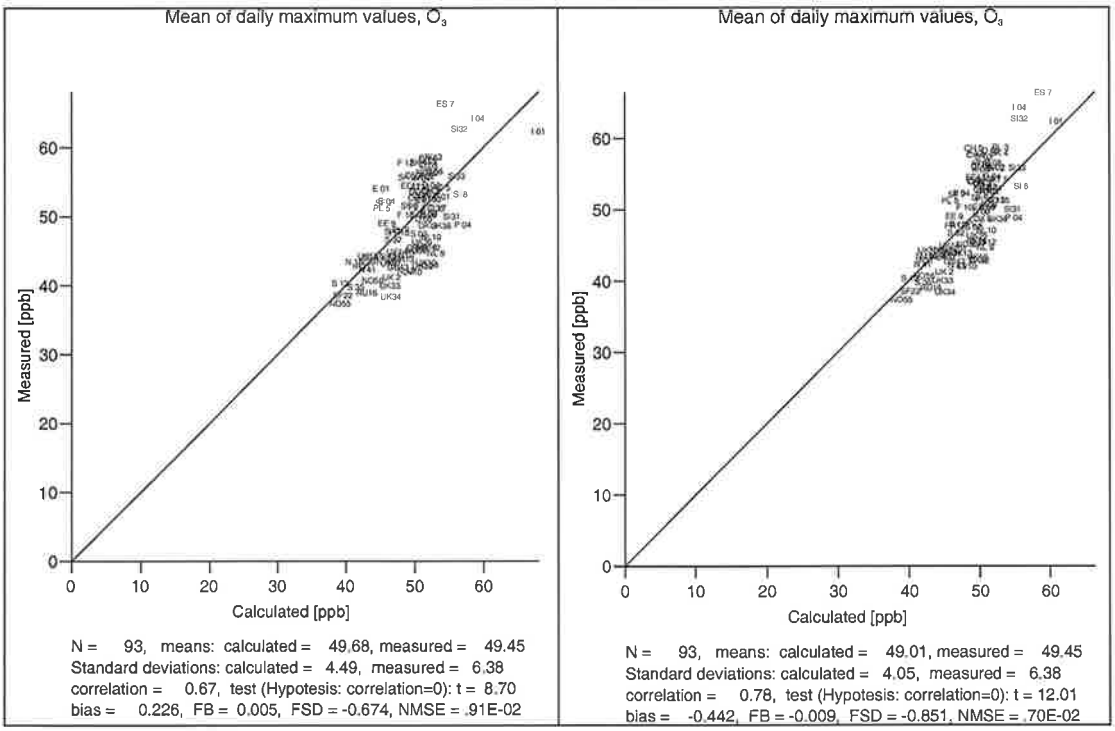


Figure 4.6: 6 Comparison of calculated and measured mean values of the daily maximum values of ozone at each measurement station for the period Apr.-Sep. 1999. The correlation coefficient increases from 0.67 to 0.78 when the data assimilation is applied.

REFERENCES

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

Desroziers, G. and Ivanov, S., 2001, Diagnosis and adaptive tuning of observation-error parameters in a variational assimilation: Quarterly Journal of the Royal Meteorological Society, **127,** 1433-1452.

Hollingsworth, A. and Lonnberg, P., 1986, 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.

Discussion

Speaker	Comment
Arnold Heemink	How do you choose your projection operators?
Jan Frydendall	This has not been decided yet
Michael Kahnert	We use reduced eigenvalue decomposition and select at least so many eigenvalues that the smallest eigenvalue is smaller than 10% of the largest one
Maarten van Loon	Suggest you look at Johannes Flemings work as he has done very similar studies.

5 Applying Variational Data Assimilation in connection with an Atmospheric Chemical Scheme – Zahari Zlatev

Zahari Zlatev and Jørgen Brandt
National Environmental Research Institute
Frederiksborgvej 399, P. O. Box 358
DK-4000 Roskilde, Denmark

Abstract

The chemical schemes are among the most difficult components in large-scale environmental models. Therefore, these components should be treated efficiently in the efforts to make the environmental models able to produce reliable results when these are used in different important for the modern society comprehensive studies. The requirement for efficient treatment of the chemical schemes is increased when data assimilation is to be applied in conjunction with the model under consideration. Some of the problems, which are to be resolved when data assimilation is used together with a particular chemical scheme, are discussed in this note. Several experiments were carried out in an attempt to investigate the minimal requirements that are to be imposed on the availability of observations in order to ensure successful implementation of data assimilation. Results from these experiments are presented and discussed.

Key words: Environmental model, Chemical schemes, Data assimilation, Numerical methods.

1. The atmospheric chemical scheme

The major properties of the particular atmospheric chemical scheme, which is used in this note, can shortly be described as follows:

- The chemical scheme contains 56 species.
- Among the 56 chemical species, which are involved in the selected chemical scheme, are:
  - sulphur pollutants,
  - nitrogen pollutants,
  - ozone,
  - ammonia-ammonium,
  - several radicals,
  - isoprene and
  - many hydrocarbons.
- The chemical scheme can be described mathematically by a systems of ordinary differential equations (ODEs): $\mathbf{dc/dt = f(t,c)}$, where $\mathbf{t}$ is the

time-variable and $\bar{c}$ is a vector the components of which are the 56 chemical species involved in the chemical scheme.

- It is very difficult to handle efficiently this system of ODEs because it is:
  - stiff,
  - badly scaled,
  - there are temporal variations in a wide range,
  - many species (these participating in the photochemical reactions) contain sharp gradients in the periods around sun-rises and sun-sets.

These properties of the chemical systems of ODEs are illustrated in Fig. 1, where the temporal variation of one of the chemical species, isoprene, is given for the period starting at 6:00 in the morning and finishing at 24:00 in the next day. The sharp gradients at sun-rises and sun-sets are clearly seen in Fig. 1.

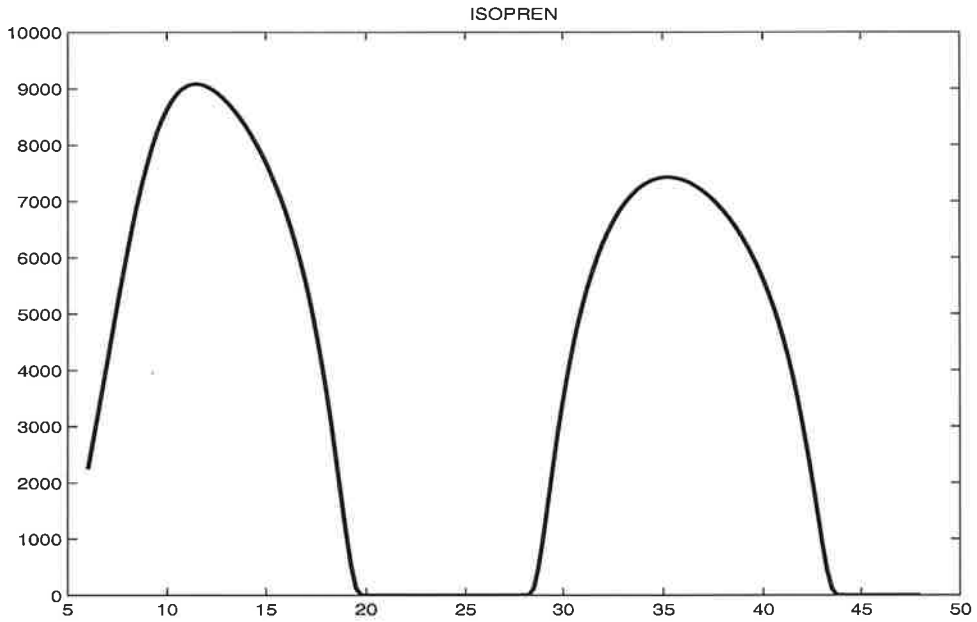


Figure 1

Temporal variation of the isoprene concentrations (measured in number of molecules per cubic centimetre) in the period from 6:00 to 24:00 on the next days (from 6 to 48 on the horizontal axis.

2. The data assimilation algorithm

The data assimilation algorithm is shortly described in this section (this algorithm is fully described in Zlatev and Brandt, 2005). Any data assimilation algorithm tries to minimize a functional of the form:

$$(1) \qquad J\{\bar{c}_0\} = \frac{1}{2} \sum_{p=0}^{P} \left\langle W(t_p)\left(\bar{c}_p - \bar{c}_p^{obs}\right),\ \bar{c}_p - \bar{c}_p^{obs} \right\rangle ,$$

where $\mathbf{J}\{\bar{\mathbf{c}}_0\}$ depends on the initial value $\bar{\mathbf{c}}_0$ of the vector of the concentrations, $\mathbf{W}(\mathbf{t}_p)$ is a matrix containing some weights and $\langle\ ,\ \rangle$ is an inner product in an appropriately defined Hilbert space (it will be assumed that the usual vector space is used, i.e. that $\bar{\mathbf{c}} \in \Re^s$ where $\mathbf{s}$ is the number of chemical species involved in the model; in this particular study this number is 56, but in the treatment of a full environmental model this number can be very large, because a discretized large-scale environmental model contains very often many millions of components). It is seen that $\mathbf{J}\{\bar{\mathbf{c}}_0\}$ depends on both the weights and the differences between calculated by the model concentrations $\bar{\mathbf{c}}_p$ and observations $\bar{\mathbf{c}}_p^{obs}$ at the time-levels $\{0, 1, \ldots, \mathbf{P}\}$ at which observations are available. $\mathbf{W}(\mathbf{t}_p)$ will be assumed to be the identity matrix $\mathbf{I}$ in this study, but in general weights are to be defined in some way.

The task in this note is to find an improved initial field $\bar{\mathbf{c}}_0$, which minimizes $\mathbf{J}\{\bar{\mathbf{c}}_0\}$, but it should be emphasized that data assimilation can also be used for other purposes (as, for example to improve the emissions). Some optimization algorithm must be used to minimize $\mathbf{J}\{\bar{\mathbf{c}}_0\}$. Most of the optimization algorithms are based on the application of the gradient of $\mathbf{J}\{\bar{\mathbf{c}}_0\}$. The adjoint equation has to be defined and used in the calculation of the gradient of $\mathbf{J}\{\bar{\mathbf{c}}_0\}$.

An algorithm for performing data assimilation for any model is given in Fig. 2 (again, more details about this algorithm can be found in Zlatev and Brandt, 2005). Several remarks are needed in connection with this algorithm:

- The Jacobian matrix of the right-hand-side vector $\mathbf{f}(\mathbf{t}, \bar{\mathbf{c}})$ has to be calculated and used to form the adjoint equation.
- The algorithm consists of forward calculations (which are carried out in the first inner loop in the red box in Fig. 2) and backward calculations (these are performed in the second inner loop in the red box in Fig. 2).
- It is assumed, in the algorithm presented in Fig. 2, that every time when a time-point in which observations are available is reached one proceeds with backward calculations. This will be inefficient when the number of observations $\mathbf{P}$ ($\mathbf{P\_STEP}$ in the algorithm from Fig. 2) is large. It is possible to carry out the backward calculations only once (after performing the forward computations over the whole time-interval; see again Zlatev and Brandt, 2005).

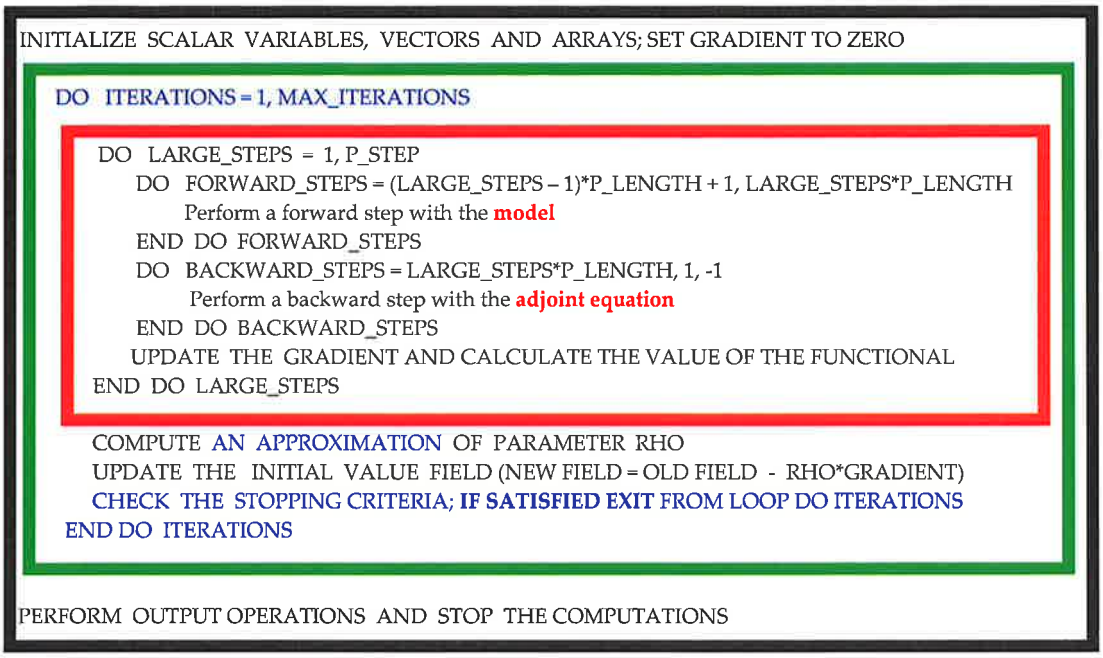


Figure 2

An algorithm for performing variational data assimilation.

- The values of vector $\bar{\mathbf{c}}$ found during the forward calculations are unfortunately also needed during the backward calculations. Thus, these values have either to be stored (during the forward calculations) or recomputed (during the backward calculations). This is causing great problems when the model treated is large.
- Five different numerical methods have been used in the experiments. Only results obtained by using the well-known Backward Euler Method will be used in the following part of this note.

3. Organization of the experiments

The experiments were carried out under the following assumptions:

- An assimilation window of length 6 hours, starting at 6:00 in the morning and finishing at noon (12:00), is always used.
- It is assumed that observations are available at the starting point and at the end of each hour (i.e. at seven time-points).
- Perturbations of the initial solution of all species by 50% errors are always applied.
- The length of the forecast window is 42 hours.
- The following actions were always successively performed:
    - The initial values of the concentrations are perturbed in the beginning of computations.
    - Data assimilation is then applied to improve the initial values (values of a reference solution, calculated by a very small time-stepsize, , being applied as "observations").

- Finally, a forecast is calculated over the full length of the time-interval (42 hours) by using the improved by the data assimilation algorithm initial values.

- A component-wise relative error estimation is used to calculate both the global error and the error made in calculation of the chemical species number . This means that the following formulae (where is the number of time-steps, while is the number of chemical species) were used:

$$\text{(2)} \qquad \text{ERROR} = \max_{n=1,2,\ldots,N;\, i=1,2,\ldots,q} \left(\frac{\left| c_{n,i}^{\text{model}} - c_{n,i}^{\text{ref}} \right|}{c_{n,i}^{\text{ref}}} \right)$$

$$\text{(3)} \qquad \text{ERROR}_i = \max_{n=1,2,\ldots,N} \left(\frac{\left| c_{n,i}^{\text{model}} - c_{n,i}^{\text{ref}} \right|}{c_{n,i}^{\text{ref}}} \right)$$

- The following notation is used in Table 1 –Table 5:
  - **Error0_P** – Error in the perturbed initial solution.
  - **ErrorF_P** – Error which is obtained when the calculations are performed by using the perturbed initial solution (i.e. without using data assimilation to improve the initial solution).
  - **Error0_I** – Error in the improved (by performing data assimilation) initial solution.
  - **ErrorF_I** – Error which is obtained when the calculations are performed by using the improved initial solution.

4. Experiments

Four important cases were studied:

- **Observations of all 56 chemical species are available.** Results are presented in Table 1 (errors obtained when all species are taken into account) and Table 2 (errors obtained when only the ozone concentrations are taken into account). The errors caused by the numerical method are dominating in this case. Therefore, decreasing the time-stepsize (which leads to an increase of the number of steps) results in decreasing of the errors.
- **Only ozone observations are available.** Results are presented in Table 3. The accuracy is not improved, because ozone is a secondary pollutant, which participate in reactions with many other species (the accuracy of which cannot be improved when only ozone "observations" are available). The errors caused by reactions of ozone with perturbed and not improved (because of the lack of "observations") chemical species is dominating. Therefore increasing the accuracy of the numerical method by decreasing the time-stepsize has practically no effect on the accuracy of the ozone concentrations.
- **Observations of a primary pollutant (sulphur di-oxide) are available.** Results are presented in Table 4. The data assimilation algorithm leads to an

improvement of the quality of the sulphur di-oxide concentrations, because this compound does not participate in many chemical reactions and, thus, the influence of the perturbed and not improved when only sulphur di-oxide "observations" are available chemical species is limited. Increasing the number of steps (i.e. decreasing the time-stepsize) results in improvements, but the errors do not decrease in such a regular way as in the case where "observations of all chemical species are available; compare the results in Table 4 with the results in Table 1 and Table 2.

Steps	Error0_P	ErrorF_P	Error0_I	ErrorF_I
1008	0.47	0.48	2.0E-3	3.2E-1
2016	0.49	0.50	1.0E-3	1.5E-1
4032	0.47	0.47	5.0E-5	7.4E-2
8064	0.48	0.49	2.5E-4	3.6E-2
16128	0.46	0.48	1.3E-4	1.8E-2
32256	0.49	0.50	6.3E-5	9.0E-3

Table 1

Global errors, i.e. errors calculated by formula (2), which are obtained when "observations" of **all** chemical species are available. The behaviour of the errors in the improved solutions (the last two columns) is nearly perfect (doubling the number of steps, which means halving the stepsize Δt, leads to halving the error; this should be expected because the Backward Euler Method is of order one).

Observations of a group of pollutants are available (ozone + nitrogen di-oxide). The results shown in Table 3 indicate that it is necessary to have observations of some of the species, which react with ozone, in order to improve the accuracy of the ozone concentrations by applying data assimilation. One of the important species, nitrogen di-oxide, has been chosen. Results obtained by using "observations" of ozone and nitrogen di-oxide are shown in Table 5. The errors caused by the perturbed and not improved (because of lack of observations) species is dominant also when observation of nitrogen di-oxide are available together with ozone observations. This is demonstrated by the fact that improving the accuracy of the numerical method by increasing the number of steps (decreasing the time-stepsize) has no effect on the accuracy. However, adding "observations" of nitrogen di-oxide has a positive effect on the accuracy of the ozone concentrations achieved when data assimilation is used (the accuracy is improved; mainly by a factor of approximately ten).

Steps	Error0_P	ErrorF_P	Error0_I	ErrorF_I
1008	0.47	0.48	2.0E-3	2.4E-3
2016	0.49	0.50	1.0E-3	1.2E-3
4032	0.47	0.47	5.0E-5	6.0E-4
8064	0.48	0.49	2.5E-4	3.0E-4
16128	0.46	0.48	1.3E-4	1.5E-4
32256	0.49	0.50	6.3E-5	7.5E-5

Table 2

Errors in the calculation of the ozone compound, i.e. errors calculated by formula (3) applied for ozone, which are obtained when "observations" of **all** chemical species are available. The behaviour of the errors in the improved solutions (the last two columns) is nearly perfect (doubling the number of steps, which means halving the stepsize Δt, leads to halving the error; this should be expected because the Backward Euler Method is of order one).

Steps	Error0_P	ErrorF_P	Error0_I	ErrorF_I
1008	0.48	0.50	0.25	0.25
2016	0.47	0.50	0.36	0.36
4032	0.48	0.50	0.32	0.33
8064	0.46	0.50	0.49	0.49
16128	0.46	0.50	0.34	0.35
32256	0.47	0.50	0.54	0.54

Table 3

Errors in the calculation of the ozone concentrations, i.e. errors calculated by formula (3) applied for ozone, which are obtained when **only ozone** "observations" are available. The application of data assimilation has negligible effect in the efforts to improve the solution. Increasing the accuracy of the numerical method by increasing the number of steps does not lead to improvements of the accuracy.

Steps	Error0_P	ErrorF_P	Error0_I	ErrorF_I
1008	0.46	0.57	2.6E-2	2.6E-2
2016	0.48	0.50	2.6E-2	2.6E-2
4032	0.49	0.50	6.8E-3	6.8E-3
8064	0.49	0.50	5.5E-2	5.5E-2
16128	0.47	0.50	2.3E-4	2.3E-4
32256	0.50	0.50	5.3E-4	5.3E-5

Table 4

Errors in the calculation of the sulphur di-oxide concentrations, i.e. errors calculated by formula (3) applied for sulphur di-oxide, which are obtained when **only sulphur di-oxode** "observations" are available. The application of data assimilation leads to improved accuracy. Increasing the accuracy of the numerical method by increasing the number of steps results in improvements of the accuracy but not in such a regular way as in Table 1 and Table 2.

Steps	Error0_P	ErrorF_P	Error0_I	ErrorF_I
1008	0.48	0.50	0.046	0.046
2016	0.47	0.50	0.046	0.046
4032	0.48	0.50	0.046	0.046
8064	0.46	0.50	0.019	0.019
16128	0.46	0.50	0.045	0.045
32256	0.47	0.50	0.018	0.018

Table 5

Errors in the calculation of the ozone concentrations, i.e. errors calculated by formula (3) applied for sulphur di-oxide, which are obtained when **both ozone and nitrogen di-oxide** "observations" are available. The application of data assimilation leads to improved accuracy (by a factor of approximately 10). Increasing the accuracy of the numerical method by increasing the number of steps does not lead to improvements of the accuracy.

5. Conclusions

The experiments, part of which were discussed in the previous section indicate that the following conclusions can be drawn:

- There are no problems when observations from all chemical species involved in the atmospheric chemical scheme are available. However, it is not realistic to expect that this will be the case in real situations.
- If observations from only one primary (and not very active chemically) pollutant (such as sulphur di-oxide) are available, then one could also achieve good results.
- If observations from only one secondary and very active chemically pollutant (such as ozone) are available, then the data assimilation algorithm is not giving considerable improvements of the accuracy of the pollutant under consideration.
- If one takes some group of pollutants, which react with each other, then improvements of the results could be achieved. This has been demonstrated by adding nitrogen di-oxide observations to the ozone observations (see the results in Table 5).

References

Zlatev, Z. and Brandt, J., 2005, Testing the Accuracy of a Data Assimilation Algorithm, International Journal of Computational Science and Engineering, to appear.

Discussion

Speaker	Comment
Michael Kahnert	Is the perturbation you introduce realistic? In regard to the chemical balance?
Leonor Tarrason	It makes sense that SO2 has less error because it is not reactive.
Sam Erik Walker	How are the observational error defined?

6 Application of 2-dimensional variational data analysis in MATCH - Michael Kahnert

Michael Kahnert
Swedish Meteorological and Hydrological Institute (SMHI)
Folkborgsv. 1, 601 76 NORRKÖPING, Sweden

A two-dimensional variational data analysis (2dvar) algorithm has been implemented into the Multiple-Scale Atmospheric Transport and Chemistry Modelling System (MATCH). The algorithm is designed to analyse ground-based chemical observations of gas and particle species. It is currently run in single-variate mode. Numerical problems associated with the inversion of the background error covariance matrix are alleviated by performing reduced eigenvalue decomposition. Variational quality control is implemented to allow for an automatised discarding of suspicious data.

The 2dvar algorithm has been validated against an older optimum interpolation (OI) algorithm. It is applied to interpolate measurements from Swedish and Norwegian background stations in order to determine the total (Swedish + long-range transport) air concentrations of ozone, and concentrations of SOx, NOx, and NHx in air and precipitation. Interpolated ozone concentrations are used as input to the MATCH-Sweden model, which is operationally applied in the Swedish National Environmental Monitoring Programme to map the Swedish contributions to dry and wet deposition of SOx, NOx, and NHx. Air and precipitation measurements of oxidised sulphur and of oxidised and reduced nitrogen are further interpolated with the data analysis algorithm to determine the long-range transport contribution to the total deposition rates. First tests have been performed to analyse ozone measurements in conjunction with a background ozone field computed with the MATCH-photochemistry model. It is planned for the future to use MATCH-photochemistry computations as a background field also for analysing SOx, NOx, and NHx.

In view of the possible use of remote sensing data in conjunction with a chemical data analysis algorithm, the errors related to computing optical properties based on MATCH results were assessed. To this end, the MATCH-photochemistry model was used to compute secondary inorganic aerosol mass (SIA) and primary particulate matter (PPM). Further, the MATCH-sea salt model was used to compute sea salt mass. The validation period was 2002. From the computational results the aerosol optical depth (AOD) and the backscattering coefficient (β_{sca}) where computed. AOD results were compared with sun photometer measurements from the AERONET station on Gotland, and β_{sca} results where compared to lidar measurements from the EARLINET station in Hamburg. The AOD computations underestimated the observations at all wavelengths, which can be attributed to the missing organic and black carbon mass in the computations. The β_{sca} results showed a generally good agreement with the measurements, except in the boundary layer, where the measurements often showed higher values, probably due to local pollution sources, which were not resolved by the model. The results of this study indicate that remote sensing observations could provide valuable additional data for chemical data analysis. However, to use these data in a

meaningful way in the MATCH-2dvar algorithm, it would be necessary to modify the system such that it could be run in multi-variate mode, and one would have to perform a comprehensive study of the model error statistics to produce a more accurate estimate of the background error covariance matrix.

Discussion

Speaker	Comment
Michael Kahnert	What is the advantage of 2D var over optimal interpolation if they are basically the same? How to discriminate between local effects and regional features in the observations? How to model background error covariances? Answer: e.g. ensemble-method We analyse the long-range transport part instead of the total concentrations of air pollutants, since LRT concentrations are more regionally representative. Should one run the model with local emissions and subtract the result from the total concentrations, or is it better to run the model without local concentrations and interpret the result as the LRT part?
Michael Kahnert	How large are the errors in modelling optical properties of aerosols based on CTM results in view of assimilating remote sensing data?
General	A conclusion was that the main error source in the current model version is that organic carbon and soot are not accounted for in the model. The uncertainties related to modelling aerosol optical depth and backscattering coefficient are - at least in comparison - small.
Jørgen Brandt	It is better to run the model with all emissions and again without the local emissions to separate the total, LRT, and local contributions.
Michael Kahnert	It will be difficult to run our Europe-model with an equally high grid resolution as our Sweden-model
Jørgen Brandt	One should add the capability of running with nested resolution to the model

7 Implementation and performance experiences with chemical 4Dvar assimilation – Hendrik Elbern

H. Elbern and A. Strunk
Rhenish Institute for Environmental Research at the University of Cologne (RIU)
and
Helmholtz virt. Inst. for Inverse Modelling of Atmospheric Chemical Composition (IMACCO)

The slides presented by Henrik Elbern are included in the appendix to this report. The contents of the talk included the following topics and conclusions.

1. General problem appraisal
2. BLUEs as a partial answer
    a. Chemical 4Dvar
    b. Operation parameter choice
    c. 4D-variational implementation issues
    d. Enhanced spatial resolution
3. Summary
    a. Chemical weather forecasts are a multiple scale problem
    b. Chemical data assimilation rests on sparse and heterogeneous observations, with variable error characteristics (incl. error of representativity)
    c. Initial value optimisation is insufficient, as at least emission rates are less known and more important
    d. The ability for inversion is therefore required to optimize emission rates
    e. Much is to be done for optimising multivariate covariance matrices
4. Conclusion
    a. Fine grid resolution is required for air quality modelling, however
    b. The finer the grid the more critical is meteorological modelling
    c. Covariance matrices and performance statistics are essential; operational application are essential
    d. Tangent linear approximations and Gaussian error assumptions often violated
    e. Sequential assimilation algorithms appear to be less suited for treatment of temporal error correlations
    f. Ensemble ideas appear useful in one or another way

Discussion

Speaker	Comment
Zahari Zlatev	How do you deal with the increasing resolution? With gradients, meteorology and emissions? How do you deal with the linear tangent approximations when they are not valid.
Hendrik Elbern	Increasing resolution is firstly a computational problem. As regards the degrees of freedom additionally incurred , the radius of influence is the proper device to expand the information gained by the measurement over the affected model domain. On the time scale considered in air quality, meteorology must be firstly optimized by meteorological data assimilation, with little benefits to be expected from chemistry observations. Emissions are taken as first guess, and then subject to inversion based optimisation. The tangent linear approximation is only valid if the first guess model run is in the proper chemical scenario. For example, the model run should be in a VOC a NOx restricted domain from the outset.
Arnold Heemink	What is your experience with using the discretized adjoint of the continuous forward model in stead of using the adjoint of the discrete forward model?
Hendrik Elbern	I think I can leave it to the mathematicians to work out the technique to be adopted because I consider it to give less of an error than other sources in real case applications. Our experience is based on a box model basis with this issue and the differences found were not significant.
Arnold Heemink	How many model computations for a gradient computation?
Hendrik Elbern	Requires 3 computations.
Leonor Tarrason	Can you break the emissions down to sectors for assimilation?
Hendrik Elbern	The emissions are broken down in terms of each grid point, (also aloft to account for effective stack heights), and the 19 emitted species. The shape of the diurnal profile of the temporal evolution is taken from the emission model.
Yvan Orsolini	Stratospheric assimilation. Do you have to worry about chemical balance.
Hendrik Elbern	No. This is exactly what 4D var does!
Michael Kahnert	How would you include deposition in this. Will assimilation improve deposition velocities?
Hendrik Elbern	This is a possibility and is planned but it is not implemented as yet

Ensemble methods

8 An introduction to Sequential Importance Resampling – Sam Erik Walker

Sam Erik Walker
Norwegian Institute for Air Research (NILU)
PO BOX 100, Kjeller, Norway

1 Background

Assessing air quality can potentially be improved by combining models and observations using different approaches. For regional scale air quality assessments, this means primarily to combine a given regional scale air pollution model with available air quality observations using different methods of data assimilation.

The SIR (Sequential Importance Re-sampling) method (*Van Leeuwen, 2003*) is a relatively new data assimilation method, based on a completely general Bayesian statistical framework (*Box and Tiao, 1992; Berger, 1985*). The method makes no assumptions of linearity in the model equations, nor that the model or observation errors should be Gaussian. This is in contrast with most other well-known methods of data assimilation, such as the optimal interpolation (OI) method (*Gandin, 1963*), the 3D-Var or 4D-Var methods (*Lorenc, 1986); Lewis and Derber, 1985*), or different variants of the Kalman filter (*Kalman, 1960*), which assumes that the model evolution is linear and that the involved errors are Gaussian. The SIR-method could therefore be interesting to apply in connection with regional scale air pollution models since such models will generally be non-linear with non-Gaussian errors when one includes photochemistry and/or aerosol-chemistry operators.

The method is also known as a Sequential Monte Carlo (SMC) method, Markov Chain Monte Carlo (MCMC) method, or Particle Filter (PF). The recent book by (*Doucet et al., 2001*) provides a good overview and insight into this particular class of statistically based data assimilation methods. Another good source of information is provided by the web site: http://www-sigproc.eng.cam.ac.uk/smc/index.html.

The SIR-method, as well as other methods in this class, has already been applied with good results in different scientific fields such as control theory, tracking, perception etc, where non-linearity in the model plays a role. More recently the SIR-method has been applied in oceanography (*Van Leeuwen, 2003*), and in ecosystem population modelling (*Losa et al., 2005*). No applications, however, seem yet to exist for regional scale air pollution modelling.

2 Bayesian statistics

As mentioned the SIR-method is based on Bayesian statistics, which in turn is based on Bayes' theorem. For a model where the model state is defined as a vector **x** containing generally n components, and observations are defined as a vector **y** containing generally m components, Bayes' theorem can (in our context) be written:

$$P(\mathbf{x}^t|\mathbf{x}^f,\mathbf{y}^o) = k \cdot \pi(\mathbf{x}^t|\mathbf{x}^f) \cdot L(\mathbf{x}^t|\mathbf{y}^o) \quad (2.1)$$

where $\mathbf{x}^f$ denotes the forecasted model state (before observations have been used), and $\mathbf{x}^t$ denotes the true model state, which we want to determine as accurately as possible, based on the model forecasted state $\mathbf{x}^f$ and the set of observations $\mathbf{y}^o$. For a regional scale model the model state could e.g., be defined as the one-dimensional vector of all 3D grid cell concentrations, while the observation vector can be defined as the one-dimensional vector of all (m) (simultaneous) regional air quality measurements.

According to Eq. (2.1) the posterior density P of the true state $\mathbf{x}^t$ given both the forecasted state $\mathbf{x}^f$ and observations $\mathbf{y}^o$ can be calculated as the product of the prior density π and the likelihood function L. The prior density π summarises (in a probabilistic sense) our prior beliefs about the true model state $\mathbf{x}^t$ given the forecasted model state $\mathbf{x}^f$, but before any observations $\mathbf{y}^o$ are used, while the posterior density P summarises our beliefs about the true model state $\mathbf{x}^t$ after observations are used. The link between them is the likelihood function L. This function is defined by $L(\mathbf{x}|\mathbf{y}) = p(\mathbf{y}|\mathbf{x})$ where p is the probability density of the observations **y** given that the true model state is **x**. A likelihood function L could for example be defined by assuming that the observation errors are multi-dimensional Gaussian:

$$L(\mathbf{x}^t|\mathbf{y}^o) = \frac{1}{(2\pi)^{n/2}|\mathbf{R}|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{y}^o - H(\mathbf{x}^t))^T \mathbf{R}^{-1}(\mathbf{y}^o - H(\mathbf{x}^t))\right) \quad (2.2)$$

In Eq. (2.2) **R** represents the observations error covariance matrix (usually diagonal with observation error variances along the diagonal, while H represents an observation operator linking model states with expected observations, i.e., mapping any model state vector **x** in the n-dimensional model space into a vector of expected observations H(**x**) in the m-dimensional observation space. The values of the m-vector H(**x**) can be viewed as a set of expected observations, given that **x** represents the true model state. In the likelihood function L, **y** is kept constant equal to the actual set of observations $\mathbf{y}^o$ while **x** is allowed to vary. For example by using the Gaussian function for L in Eq. (2.2) the most likely true model states are those model states **x** for which H(**x**) is close to $\mathbf{y}^o$. This is then weighted against the prior density π in order to form the posterior density P. The value k in Eq. (2.2) is simply a constant so that P becomes a proper density, i.e., that P integrates to 1 over the model space.

In the SIR-method the user has complete freedom in specifying the prior density π and the likelihood function L. This is in contrast to other data assimilation method like Optimal Interpolation (OI), 3D-Var or 4D-Var, or Kalman filters, which can

also be viewed as Bayesian statistical based methods, but where these functions are always assumed to be Gaussian. The freedom to be able to specify them arbitrarily is generally an important advantage of the SIR-method over the other methods.

3 The SIR-method

The SIR-method operates with an ensemble of model states $\{\mathbf{x}^{(i)}, i = 1,\ldots,N\}$, where N denotes the number of ensemble members (the ensemble size). The number of ensemble members is specified by the user and kept constant in the method for all time steps. This is similar to other ensemble based data assimilation methods like the ensemble Kalman filter method (EnKF) (*Evensen, 1994; 2003; 2004*), and the reduced rank Kalman filter methods (*Verlaan and Heemink, 1997; Heemink et al., 2001; Segers, 2002*).

In the SIR-method all ensemble members are considered to be equally likely, i.e., they will all have the same (discrete) probability 1/N. It is the positions $\mathbf{x}^{(i)}$ of the ensemble members, i.e., their spatial densities, in the model state space which forms the basis for the approximation of the involved Bayesian prior and posterior probability density functions (PDFs). Based on available observations, a re-sampling step is included in the method, where each ensemble member will either be kept and possibly multiplied (made into several identical or almost identical copies) or removed from the ensemble, based on the calculated likelihood-function values. The method is illustrated in the Figs. 1-3 below.

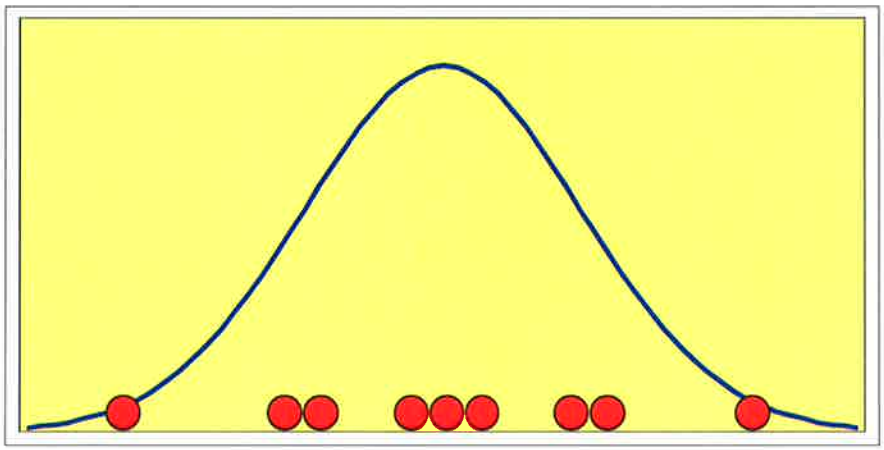


Figure 1: A Gaussian prior PDF (blue curve) represented by a discrete approximation in the form of an ensemble of N = 9 red points with equal probability 1/9.

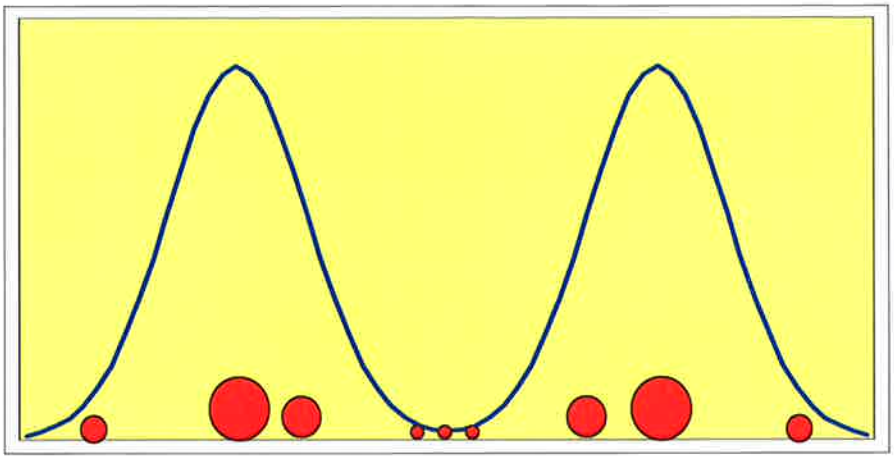


Figure 2: A non-Gaussian posterior PDF (blue curve) obtained by multiplying the prior PDF shown in Fig. 1 with a likelihood function. The same ensemble is used, but now the ensemble members have different probabilities (represented by the new sizes of the red points)

In the figures the red points represents a set of ensemble members, which forms a discrete approximation of the probability densities plotted. Fig. 1 shows a (Gaussian) prior PDF, which is updated by a likelihood function to form the posterior PDF as shown in Fig. 2. The size of the red points represents the probabilities of the different ensemble members before and after multiplying with the likelihood function. A re-sampling step is then performed in order to form a new ensemble where each ensemble member again has equal weighting or (discrete) probability 1/N. This is illustrated in Fig. 3 (below) where the larger points are replaced with 2 or more new points, while the smallest points have been removed from the ensemble. This ensures that the posterior density is again well represented with a new ensemble based on equal weighting.

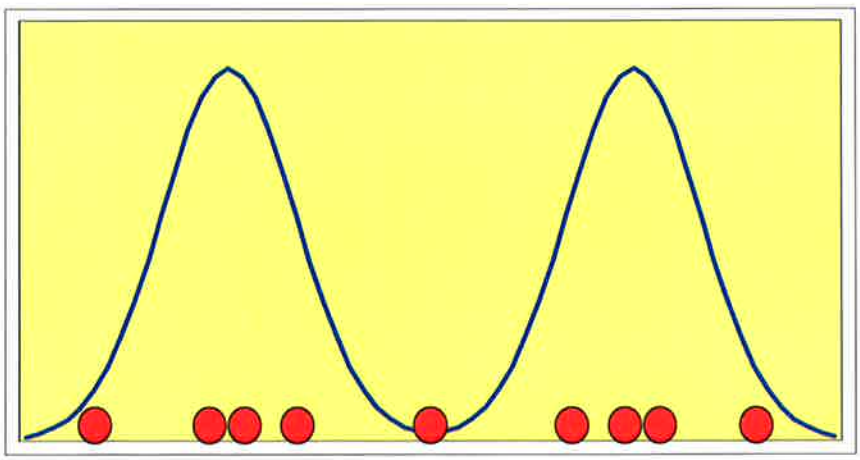


Figure 3: After re-sampling the non-Gaussian posterior PDF (blue curve) is represented by a new ensemble of red points again with equal probabilities (1/9).

We will now turn to a more formal description of the method.

Algorithmic steps of the SIR-method

Assume that we have en ensemble of N finally analysed or assimilated model states from a previous time step k-1:

$$\left\{ \mathbf{x}_{k-1}^{a,(1)}, \mathbf{x}_{k-1}^{a,(2)}, ..., \mathbf{x}_{k-1}^{a,(N)} \right\} \tag{3.1}$$

where the ensemble members have the same posterior (discrete) probability $P_i = 1/N$ for $i = 1,\ldots,N$. For each ensemble member, a new forecasted model state for the next time step k can be produced, by running the model to the next time step k:

$$\mathbf{x}_k^{f,(i)} = M_{k-1}(\mathbf{x}_{k-1}^{a,(i)}, \boldsymbol{\theta}_k, \boldsymbol{\eta}_k^{(i)}) \text{ for } i = 1,...,N \tag{3.2}$$

In Eq. (3.2) $\boldsymbol{\theta}_k$ represents a set of additional model parameters, and $\boldsymbol{\eta}_k$ represents a vector of stochastic variables describing model errors or noise. The parameter vector $\boldsymbol{\theta}_k$ represents parameters of the model not included in the state vector $\mathbf{x}_k$. The model noise vector $\boldsymbol{\eta}_k$ is important in the SIR-method, as it will be used to generate a natural spread of the model calculated (forecasted) ensemble members at time step k. The noise vector can be viewed as representing natural dynamical model errors, which will be added as random forcing in the model equations.

Based on the new forecasted ensemble at time step k, an approximate minimum variance estimate of the true model state $\mathbf{x}_k^t$ at time step k before using available observations is then:

$$\overline{\mathbf{x}}_k^f = \sum_{i=1}^{N} w_i \mathbf{x}_k^{f,(i)} \tag{3.3}$$

where $w_i = 1/N$ for $i = 1,...,N$ denotes the ensemble weights. Note that Eq. (3.3) with the weighting chosen is an approximation of the expectance value associated with the new prior PDF at time step k, since each ensemble member has the same prior probability $\pi_i = 1/N$ for $i = 1,...,N$ inherited from the same equal (1/N) probabilities for the ensemble members at the previous time step $k - 1$.

An estimate of the uncertainty of the estimated true model state can further be calculated using the following expression for the second-order moments (variance-covariance matrix) of the prior PDF based on the same ensemble:

$$\mathbf{P}_{e,k}^f = \frac{1}{N-1} \sum_{i=1}^{N} \left(\mathbf{x}_k^{f,(i)} - \overline{\mathbf{x}}_k^f \right) \left(\mathbf{x}_k^{f,(i)} - \overline{\mathbf{x}}_k^f \right)^T \tag{3.4}$$

Using available observations $\mathbf{y}_k^o$ at time step k, updated weights $\hat{w}_i$ (probabilities) will be calculated using a likelihood function L as follows:

$$w_k^{f,(i)} = \frac{1}{N}; \quad w_k^{a,(i)} = \frac{1}{N} \cdot L(\mathbf{x}_k^{f,(i)} | \mathbf{y}_k^o) \text{ for } i = 1,...,N \tag{3.5}$$

where the superscript ^{a} denotes assimilated values (after observations have been used), and where L denotes the likelihood function of $\mathbf{x}$ given the observations $\mathbf{y}$. As mentioned earlier, in Bayesian statistics the likelihood function is generally defined by the conditional probability density function $p(\mathbf{y}|\mathbf{x})$ of observations $\mathbf{y}$ given that the true model state is $\mathbf{x}$, and then reversing the arguments, making L a function of $\mathbf{x}$ given $\mathbf{y}$, i.e., $L(\mathbf{x}|\mathbf{y}) = p(\mathbf{y}|\mathbf{x})$.

A likelihood function L could for example be defined by assuming that the observations are multidimensional Gaussian given the model state $\mathbf{x}$:

$$L(\mathbf{x}_k^{f,(i)}|\mathbf{y}^o) = \frac{1}{(2\pi)^{n/2}|\mathbf{R}|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{y}^o - H(\mathbf{x}_k^{f,(i)}))^T \mathbf{R}^{-1}(\mathbf{y}^o - H(\mathbf{x}_k^{f,(i)}))\right) \tag{3.6}$$

As described earlier, $\mathbf{R}$ here represents the observations error variance-covariance matrix, while H represents an observation operator mapping conceived true model states $\mathbf{x}$ into expected observations $H(\mathbf{x})$ in the observation space.

The next step of the SIR-method is then to perform a re-sampling of the ensemble using the new assimilated probabilities or weights $w_k^{a,(i)}$. This is illustrated in Fig. 4 (below).

The re-sampling is done by sampling a new set of N ensemble members, with replacement, using the old ensemble with the probability distribution given by the assimilated probabilities $w_k^{a,(i)}$ defined by Eq. (3.5). Old ensemble members that correspond well with the observations (high weights) will thus be kept and possibly multiplied (several copies might be made), while those corresponding poorly with the observations (low weights) might be removed. After the re-sampling step, all ensemble members will again have equal weights or probabilities $w_k^{a,(i)} = 1/N$.

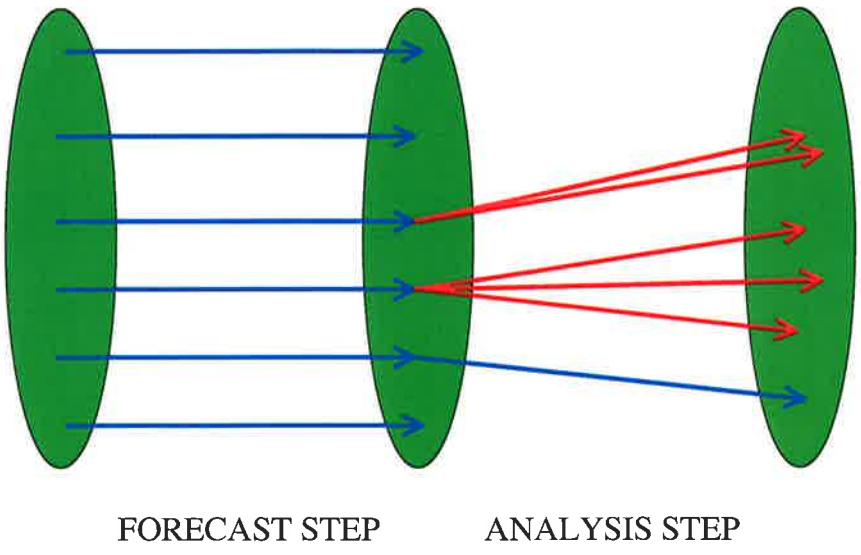


Figure 4: Re-sampling of ensemble members in the SIR-method.

Based on the new assimilated ensemble at time step k, an approximate minimum variance estimate of the true model state $\mathbf{x}_k^t$ at time step k after using available observations can again be written:

$$\mathbf{x}_k^a = \sum_{i=1}^{N} w_i \mathbf{x}_k^{a,(i)} \tag{3.7}$$

where $w_i = 1/N$ for i = 1,...,N. An estimate of the uncertainty of this estimate is again provided by the following formula:

$$\mathbf{P}_{e,k}^a = \frac{1}{N-1} \sum_{i=1}^{N} \left(\mathbf{x}_k^{a,(i)} - \mathbf{x}_k^a \right) \left(\mathbf{x}_k^{a,(i)} - \mathbf{x}_k^a \right)^T \tag{3.8}$$

In the observation space, an approximate minimum variance estimate of the true observations can be calculated using the following expression:

$$\mathbf{C}_k^a = \frac{1}{N} \sum_{i=1}^{N} H(\mathbf{x}_k^{a,(i)}) \tag{3.9}$$

with associated uncertainty (variance-covariance matrix) calculated by:

$$\mathbf{Q}_{e,k}^a = \frac{1}{N-1} \sum_{i=1}^{N} \left(H(\mathbf{x}_k^{a,(i)}) - \mathbf{C}_k^a \right) \left(H(\mathbf{x}_k^{a,(i)}) - \mathbf{C}_k^a \right)^T \tag{3.10}$$

Eqs. (3.2) - (3.10) represents the algorithmic steps of the SIR-method from one time step to the next, and can be repeated for all involved time periods k = 1,...,K.

A challenge in the method is to create a good initial ensemble $\left\{ \mathbf{x}_0^{a,(1)}, ..., \mathbf{x}_0^{a,(N)} \right\}$ at time step k = 0. This is usually done by starting with a simple initial (e.g., global background) model state, and then spinning up the model using some iterations with Eq. (3.2) applying "sensible" perturbations of the involved model state $\mathbf{x}_k$ and parameters $\boldsymbol{\theta}_k$ using the random noise vector $\boldsymbol{\eta}_k$.

This completes the description of the algorithmic steps of the SIR-method.

Discussion

Note from the above description that the SIR-method handles each ensemble member separately except for the re-sampling step (which is of low complexity). The method is therefore highly suited for parallel processing. It is also conceptually very easy to implement since it avoids the use of any minimization procedure or matrix inversion. For large (regional scale) models the computational time should be roughly proportional to N runs of the model on a single processor machine, but if several processors are available the model forecasts could be run in parallel.

Due to the simple update equations the method can in principle handle any kind of non-linearities in the model M itself or in the observation operator H. A regional

scale model will typically be non-linear due to the inclusion of photochemistry and/or aerosol chemistry in the model equations. The observation operator H could be non-linear if e.g., line- or point source sub-grid models are included in this operator in order to compare with observations.

A nice feature of the method is that it can also easily provide estimates of uncertainties in the model state variables, e.g., in the calculated regional scale model concentrations. It is also possible to calculate probabilities of exceedence of limit values.

The ensemble size N needed in practical applications with the method generally depends on the regional scale model itself, the number of model state variables, and the number and placement of observations. It is difficult to specify in advance exactly how large N must be. A trial and error procedure must usually be exercised in order to find the optimal number of ensemble members. Probably it must be at least in the interval 25-100. If N is chosen too small, the method may suffer from convergence problems in high dimensions since it depends on a Monte Carlo random draw approach. It is probably most easy to use in situations where the posterior PDF has a single maximum.

A smoother version of the method also exists – called Guided SIR – that uses a range of observations over a specific time window to calculate the likelihood function (*Van Leeuwen, 2003*). This may enhance the applicability of the method for regional scale models since the ensemble size can then probably be made smaller.

4 Application on a simple 1D atmospheric transport model with photochemistry

The SIR-method has recently been tested on a one-dimensional atmospheric advection-diffusion model with photochemistry (*Walker et al., 2005*). Simulated experiments, defining a set of true input parameters, and resulting model concentration, were performed to see if the method could handle systematic (bias) and unsystematic (random) errors in the input data, and still be able to produce assimilated values close to the true state. The effects on the performance of using different observations likelihood functions, such as Gaussian or Lorentz (Student's t) distributions were also tested.

The 1D model tested was:

$$\frac{\partial \mathbf{c}}{\partial t} = -u\frac{\partial \mathbf{c}}{\partial x} + \frac{\partial}{\partial x}\left(k_x \frac{\partial \mathbf{c}}{\partial x}\right) + \mathbf{R} + \mathbf{q} \qquad (4.1)$$

where **c** is a space (x) and time (t) varying concentration vector (μg/m) containing the species NO_2, NO and O_3, u is the wind speed and k_x a turbulent eddy diffusivity coefficient. **R** denotes the non-linear fast reaction NO_2-NO-O_3 photochemistry operator, and **q** represents emissions of the same three species. Boundary and initial conditions were given by $\mathbf{c}(x,t) = \mathbf{c}_B$ for $x = 0$ and $x = n\Delta x$, and for $t = 0$, where $\mathbf{c}_B$ denotes a set of background concentrations of the three species. The physical domain $[0, n\Delta x]$ was divided into n grid cells each with length Δx. For the tests performed here $n = 50$ and $\Delta x = 1000$ m. Eq. (4.1) was

then discretized and solved on an hourly basis using hourly input data of u, k_x, **q** and $\mathbf{c}_B$, and separate operators for advection, diffusion and photochemistry (*Bott, 1989; Slørdal et al, 2003*).

Experimental set-up

The model was run for 2 weeks (336 hours). Realistic hourly values of wind speed (u) and temperature difference (ΔT_{10m-2m}) were taken from a meteorological station close to Oslo, Norway. The station is placed in a relatively flat and homogenous area (z_0 = 0.1 m). A meteorological pre-processor was used to calculate horizontal turbulence intensities v_s, and diffusion coefficients k_x as $0.1 \cdot \Delta x \cdot \sigma_v$ (*Slørdal et al., 2003*). Expected values of emissions (**q**) and background concentrations ($\mathbf{c}_B$) were set equal to 10^{-3} , $9 \cdot 10^{-3}$ and 0 µg/m, and 10, 0 and 50 µg/m respectively for each of the three species, constant for all hours.

The model state vector **x** was defined as the concentration grid vector **c**. In order to create the initial ensemble and to update the ensemble from one time step to the next, actual input parameters u, k_x, **q** and $\mathbf{c}_B$ to the model was drawn randomly using lognormal distributions. The hourly observed values were used as mean values in these distributions, and the standard deviations were assumed to be 40% of these values. The values were set equal for all grid cells.

True values $()^t$ of the above parameters were defined using the expectance values and an assumed bias factor f_b = 1.2 (20% bias) as follows:

$$u^t = E(u) \cdot f_b \,,\; k_x^t = E(k_x) \cdot f_b \text{ and } \mathbf{q}^t = E(\mathbf{q}) / f_b \,.$$

True background values were always assumed to be unbiased i.e. $\mathbf{c}_B{}^t = E(\mathbf{c}_B)$. Pseudo-observations of NO_2 were assumed to be Gaussian or Lorentz-distributed around the true model concentrations using a standard deviation equal to 5% of the true value for each hour. We assumed no observations of NO or O_3.

Results

The results are shown in Figs. 5 and 6 (below) for ensemble sizes N = 25 and N = 100 respectively. In both figures hourly concentrations of NO_2 from grid cell number 27 (of the 50 cells) are plotted. Only the tests performed with the Lorentz (Student's t) distribution are shown here. It was found that this gave somewhat more stable and consistent improvements than using Gaussian distributions.

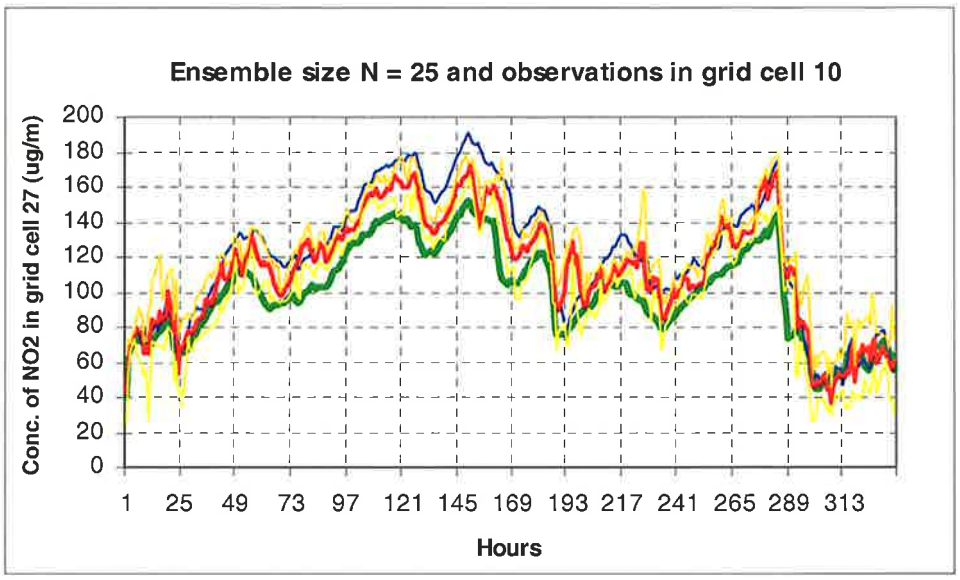


Figure 5: Results of the SIR-method using an ensemble size N = 25 and observations in grid cell 10 only. Concentration values of NO_2 from grid cell 27 (of 50) are plotted. Unit: µg/m.

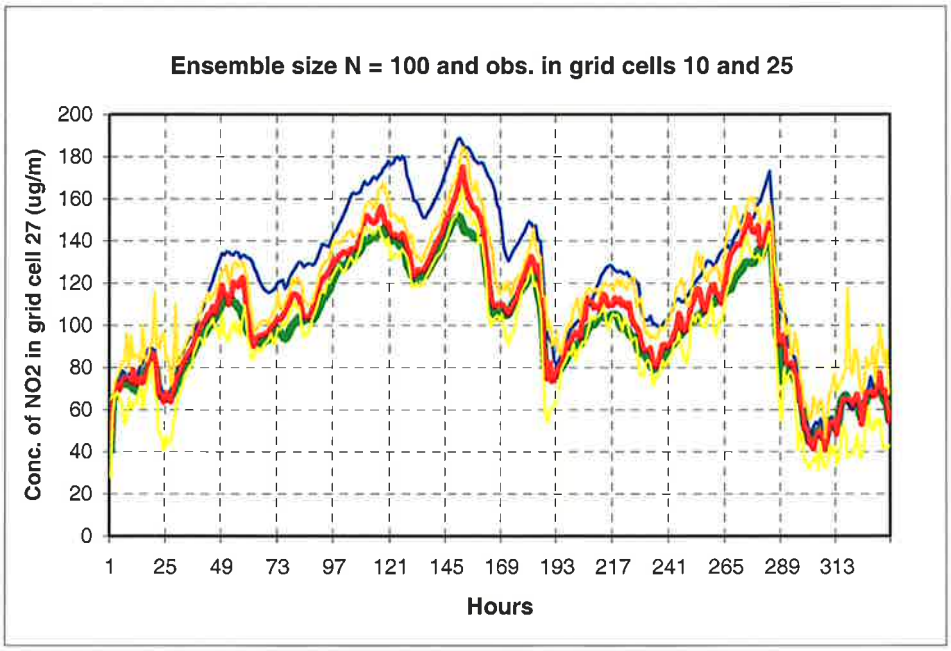


Figure 6: Results of the SIR-method using an ensemble size N = 100 and observations in grid cell 10 and 25. Concentration values of NO_2 from grid cell 27 (of 50) are plotted. Unit: µg/m.

From the figures we see that the assimilated concentrations (red curve) lie consistently closer to the true concentrations (green curve) than the unassimilated concentrations (blue curve), although the improvement varies with time. This shows that the SIR-method works reasonably well on our test problem. The yellow and orange curves in the figure represents respectively the 2.5 and 97.5 percentiles of the assimilated (posterior) concentration distributions based on the

ensemble members. Increasing the ensemble size N from 25 to 100 and the number of observations from 1 (grid cell 10) to 2 (grid cells 10 and 25) improves the results. Increasing N further does not lead to any great improvements, since the model error statistics seems to be well represented with 100 ensemble members. Increasing the number of observations leads to some improvements in the results, but moderately after two observations have been introduced. This is probably due to the 1D structure of the model, and the fact that the emissions are distributed homogeneously in all grid cells. Most of the information about the true state seems to be contained in a few observations of NO_2.

In Fig. 7, the probability of exceeding 100 µg/m (as an example), and in Fig. 8, the number of unique ensemble members is shown as a function of time (hours) for the run with N = 25 and observations of NO_2 in grid cell 10. As can be seen from the figure we avoided ensemble collapse (i.e., very few unique members in the ensemble) during the run with N = 25. This was also true for the run with N = 100.

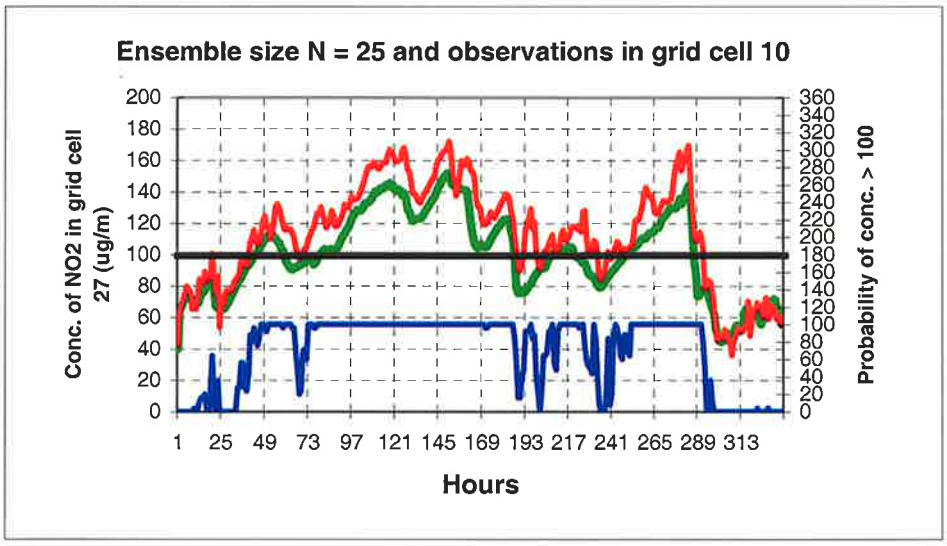


Figure 7: Same plot as in Fig. 1, but in addition the blue curve plotted at the bottom shows the calculated probability that the true concentration exceeds 100 µg/m.

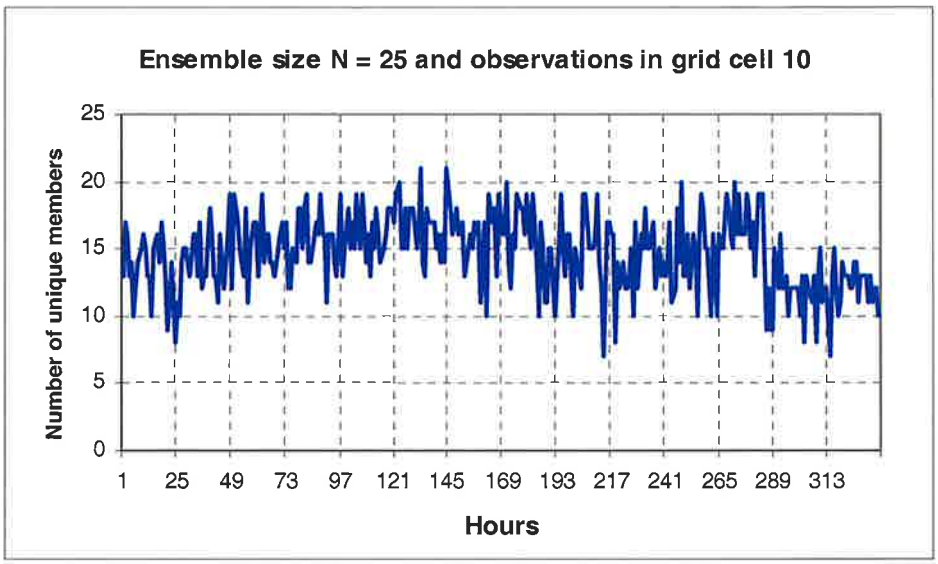


Figure 8: Number of unique ensemble members.

Thus, to summarize, the SIR-method seems to work well on the 1D advection-diffusion photochemistry model tested here, reducing both model bias and uncertainties if observations of NO_2 are available. Likelihood functions based on Lorentz (Student's t) distribution seems to generally give the best results.

5 Concluding remarks

To summarize our presentation of the SIR-method, the method has the following properties on the plus side:

- It is easy to implement, no minimization procedure or matrix inversion needed
- It is flexible to model parameter or physics stochastic errors
- It does not intrude on or change the model physics. Only changes the probabilities of different model states
- It is especially suited for non-linear models with non-Gaussian PDFs
- The discrete PDFs will converge towards true PDFs if the ensemble size goes to infinity
- It is easy parallelizable if more processors are available
- It is easy to calculate assimilated model uncertainties even when there are no observations

On the minus side:

- It may need a large ensemble size N since it is a Monte-Carlo method in high dimensions
- It may be difficult to create a good initial ensemble representing the prior PDF
- It may be difficult to track the true model state if the PDFs involved have many local maxima (multi-modal distribution)

It remains to be seen though, if the issues of non-linearity and non-Gaussian prior and posterior PDFs, are sufficient to warrant the use of this method for regional scale air pollution models, over other more traditional methods of data assimilation, like optimal interpolation, the variational methods or Kalman filter methods.

References

Berger, J.O., Statistical Decision Theory and Bayesian Analysis, *Springer Verlag, 1985*

Bott, A. (1989) a positive definite advection scheme obtained by non-linear renormalization of the advective fluxes. *Mon. Weather Rev. 117, 1006-1015 and 2633-2636.*

Box, G.E.P and G.C. Tiao (1992) Bayesian Inference in Statistical Analysis, Wiley Classics Library Ed., New York.

Doucet A., de Freitas N., Gordon N., editors, Sequential Monte Carlo methods in practice, *Springer Verlag, New York, 2001.*

Evensen, G. (2004) Sampling strategies and square root analysis schemes for the EnKF, *Ocean Dynamics*, **54**, 539-560.

Evensen, G. (2003) The ensemble Kalman filter: Theoretical formulation and practical implementation, *Ocean Dynamics*, **53**, 343-367.

Evensen, G. (1994) Sequential data assimilation with a non-linear quasi-geostrophic model using Monte-Carlo methods to forecast error statistics. *J. Geophys. Res.*, **99 (C5)**, 10143-10162.

Gandin, L.S. (1963) Objective analysis of meteorological fields, *Gidrometeorologicheskoe Izvdatel'stvo*, Leningrad, USSR. English translation by Israeli Program for Scientific Translations, Jerusalem, 1965.

Heemink, A.W., M. Verlaan, and A.J. Segers (2001) Variance reduced ensemble Kalman filtering. *J. Mon. Wea. Rev.* **129**, 1718-1728.

Kalman, R. (1960) A new approach to linear filtering and prediction problems. *Trans. ASME, Ser. D, J. Basic Eng.* **82**, 35-45.

Lewis, J. and J. Derber (1985) The use of adjoint equations to solve a variational adjustment problem with advective constraint. *Tellus* **37A**, 309-322.

Lorenc, A. (1986) Analysis methods for numerical weather prediction. *Quart. J. Roy. Meteor. Soc.* **112**, 1177-1194.

Losa, S. et al., Sequential Importance Re-sampling Filtering in Ecosystem Modelling, *European Geosciences Union General Assembly, Vienna, 24-29 April 2005*

Segers, A. (2002) Data assimilation in atmospheric chemistry models using Kalman filtering, Ph.D. thesis, Delft University.

Slørdal, L.H., Walker, S.E., Solberg, S. (2003) The urban air dispersion model EPISODE applied in AirQUIS$_{2003}$. Technical description. *Kjeller, Norwegian Institute for Air Research (NILU TR 12/2003).*

Van Leeuwen, P. J. (2003) A variance minimizing filter for large scale applications. *J. Mon. Wea. Rev.* **131**, 2071-2084.

Verlaan, M. and A.W. Heemink (1997) Tidal flow forecasting using reduced rank square root filters. *Stochastic Hydrology and Hydraulics* **11**, 346-368.

Walker, S.-E. (2005) Application of sequential importance re-sampling for 1D atmospheric chemical data assimilation. Poster presented at the *IAMAS 2005 conference* in Beijing, China, 2-11 August 2005.

Discussion

Speaker	Comment
Sam Erik Walker	Is this methodology to complex to apply in reality?
Hendrik Elbern	This approach could identify and solve some of the difficult problems

9 Data assimilation in atmospheric chemistry models using ensemble methods – Arnold Heemink

Arnold Heemink
Delft University of Technology

The slides presented by Arnold Heemink are included in the appendix to this report. The contents of the talk included the following major topics and conclusions.

1. Introduction and motivation for the use of data assimilation and ensemble methods including real life applications of data assimilation
2. A description of ensemble Kalman filter algorithms for large scale systems including
    a. Linear dynamics F(k) and constant parameters: State estimation using Kalman filtering
    b. Ensemble Kalman filters (EnKF)
    c. Reduced Rank square root filtering (RRSQRT)
    d. Reduced-rank Kalman filters (RRKF)
    e. Complementary Orthogonal sub space Filter For Efficient Ensembles (COFFEE)
3. A presentation of methods for model reduced variational data assimilation
4. Conclusions
    a. Kalman filtering of many large scale non linear numerical models is now feasible using Ensemble Kalman filtering
    b. For the estimation of constant parameters the variational methods are superior
    c. The adjoint implementation may be avoided using model reduction

Discussion

Speaker	Comment
Zahari Zlatev	THe production of adjoint codes is not so large a problem now days as converters are available
Arnold Heemink	This is true if you are programming now, you can write the code so that it is suitable for the creation of adjoint models. However when you have a code that 20 PhD students have contributed to then this is not so easy
Arnold Heemink	Ensemble Kalman filter probability distributions are always positive definite. It is notable that the ensemble filter type is more robust than the reduced rank filters.
Arnold Heemink	Adjoint methods are better than ensemble methods when assimilating non-stochastic parameters
Arnold Heemink	Ensemble kalman filter is difficult to beat

10 NMR project discussion

The final part of the meeting was used to plan further activity of the project. It was decided that in 2006 a Nordic assimilation dataset should be compiled that will contain the relevant observational data for all the participating institutes including ground based, satellite and other remote sensed observations. This will lay the foundations for any further intercomparative assimilation studies and will allow the participating groups to cooperate closely on a single project. It will also facilitate the development of knowledge in regard to the requirements of good observational data (for assimilation purposes) including the need for an understanding of information related to uncertainty, spatial and temporal representativeness and data capacity requirements. Compilation of the dataset will also aid in identifying gaps in the available data and will allow each institute to gain from the others expertise in their own particular field.

Appendix A

Slides from the presentations by Hendrik Elbern and Aarnold Heemink

Implementation and Performance Experiences with 4D-var Chemical Data Assimilation

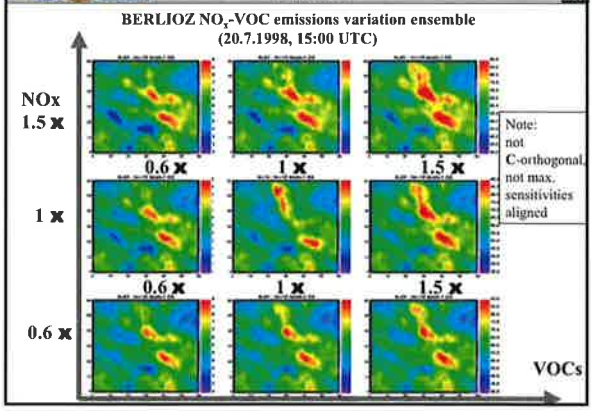
H. Elbern, A. Strunk
Rhenish Institute for Environmental Research at the University of Cologne (RIU)
and
Helmholtz virt. Inst. for Inverse Modelling of Atmospheric Chemical Composition (IMACCO)

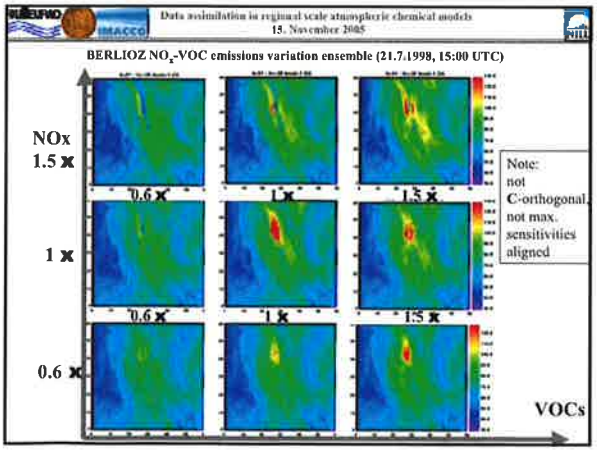
Contents

- General problem appraisal
- BLUEs as a partial answer
  - chemical 4Dvar
- Conclusions

1. General problem appraisal

- Is data assimilation useful?
- →Finally the "users" decide.
- →Users are Environmental agencies, who have to care for the "person in the street".
- → Use kerb site stations, really?
- → Consequences for the implementation and computational demands?

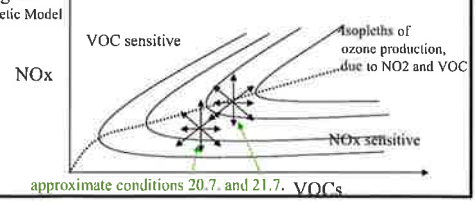




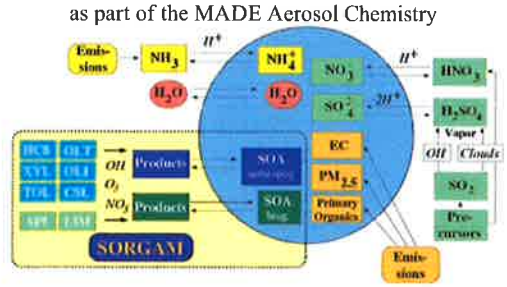
$E_{mpirical}$ K_{inetic} M_{odel} $A_{pproach}$ scheme

- Nitrogene oxides and numerous hydrocarbons act highly nonlinearly as precursors of ozone.
- Chemical conditions are either controlled by NOx or VOC deficit, delineating the "chemical regime".
- Both 4D-var and Kalman filter should start with the proper chemical regime.

EKMA diagram
(Empirical Kinetic Model Approach)

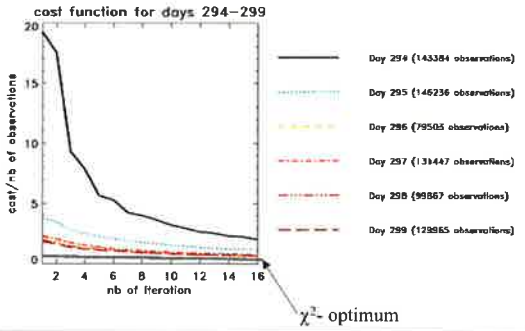


Example: chemical complexity:
The EURAD Secondary ORGanic Aerosol Model (SORGAM)
as part of the MADE Aerosol Chemistry



Digression to stratospheric chem. DA (I)

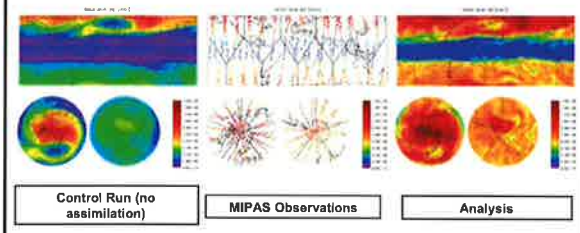
Results: Cost-functions for Oct. 21 – Oct. 26



χ^2- optimum

Digression to stratospheric chem. DA (II)

Results for HNO_3 at 28 hPa (~24 km), Nov. 4, 2003 12:00 UTC



Control Run (no assimilation) | MIPAS Observations | Analysis

Model error characteristics
(non-random walk model errors)

- model formulation: deficiencies in
  - completeness w.r.t. VOC, aerosols
  - boundary and surface layer resolution
  - horizontal resolution
- surface boundary conditions: uncertainties in
  - emissions: anthrop., bio (incl. wildfires); strength, injection heights
  - deposition (dry, wet), sedimentation
- meteo: synoptics: errors in
  - phase/amplitude,
  - PBL,
  - rain

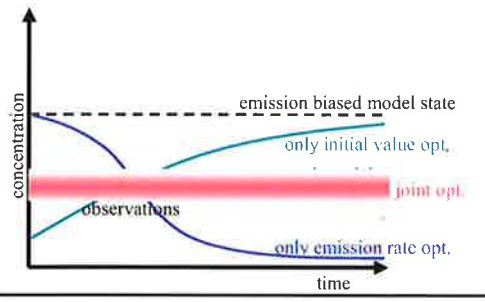
Issues in terms of air quality
data assimilation design not addressed by BLUEs

- chemical scenario mismatches:
  - tangent linear approximations poor
  - 2nd order statistics insufficient (covariances, Gaussian error characteristics)
- phase/feature mismatches
- on/off processes
- temporal error characteristics: bias+temporal correlation + white noise

2. BLUEs as a partial answer

- chemical 4Dvar
- addressing:
  - operation parameter choice
  - 4D-variational implementation issues
  - enhanced spatial resolution

Hypothesis:
initial state and emission rates are least known



In the troposphere for **emission rates** the product (*paucity of knowledge*importance*)
is high

Emission Rate Optimization

minimize cost function

$$J(\mathbf{x}(t_0),\mathbf{e}) = \frac{1}{2}(\mathbf{x}^b(t_0)-\mathbf{x}(t_0))^T\mathbf{B}_0^{-1}(\mathbf{x}^b(t_0)-\mathbf{x}(t_0)) + \frac{1}{2}\int_{t_0}^{t_n}(\mathbf{e}_b(t)-\mathbf{e}(t))^T\mathbf{K}^{-1}(\mathbf{e}_b(t)-\mathbf{e}(t))dt + \frac{1}{2}\int_{t_0}^{t_n}\{\mathbf{y}^o(t)-H[\mathbf{x}(t)]\}^T\mathbf{R}^{-1}(\mathbf{y}^o(t)-H[\mathbf{x}(t)])dt$$

deviations from background initial state
deviations from a priori emission rates
model deviations from observations

$\mathbf{x}^b(t_0)$ background state at $t=0$
$\mathbf{x}(t)$ model state at time t
$\mathbf{e}_b(t_0)$ background emission rate at $t=0$
$\mathbf{e}(t)$ emission rate field at time t
$\mathbf{K}$ emission rate error covariance matrix
$H[\ ]$ forward interpolator
$\mathbf{y}^o(t)$ observation at time t
$\mathbf{B}_0$ background error covariance matrix

3. Advanced Methods 4D-var

Transport-diffusion-reaction equation and its adjoint

Tendency Equations

direct chemistry transport equation

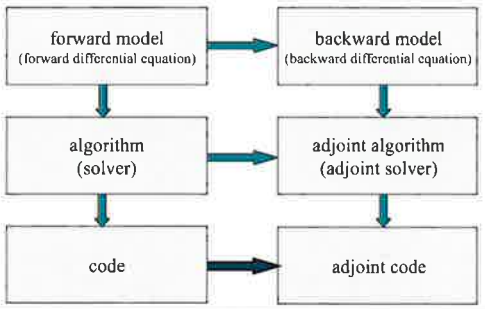
$$\frac{\partial c_i}{\partial t} + \nabla\cdot(\mathbf{v}c_i) - \nabla\cdot(\rho\mathbf{K}\nabla\frac{c_i}{\rho}) - \sum_{r=1}^{R} s_i(r)\left[k(r)\prod_{j=1}^{U} c_j^{s_j(r)-1}\right] = E_i + D_i$$

c_i	concentration of species i	c_i^*	adjoint of concentration of species i
$\mathbf{v}$	wind velocity	s	stoichiometric coefficient
$k(r)$	reaction rate of reaction r	$\mathbf{K}$	diffusion coefficient
U	number of species in the mechanism	R	number of reactions in the mechanism
E_i	emission rate of species i (source)	D_i	deposition rate of species i (sink)

adjoint chemistry transport equation

$$-\frac{\partial c_i^*}{\partial t} - \mathbf{v}\nabla c_i^* - \frac{1}{\rho}\nabla\cdot(\rho\mathbf{K}\nabla c_i^*) + \sum_{r=1}^{R}\left[k(r)\frac{\partial c_i}{\partial c_i}\right]\prod_{j=1}^{U} c_j^{s_j(r)-1}\sum_{i=1}^{U}(s_i(r)-s_i(r-1))c_i^* = 0$$

The 4D variational method:
Key development: construction of the adjoint code



Adjoint integration "backward in time"
(slide from lecture 1)

How to make the parameters of resolvents i $M(t_{i-1},t_i)$ available in *reverse* order??

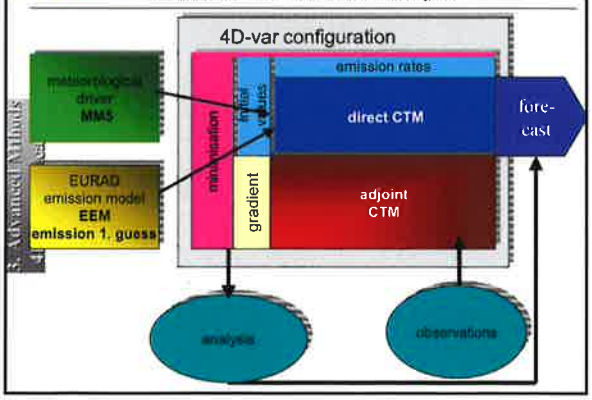
direct model $\frac{d\mathbf{x}}{dt} = \mathcal{M}(\mathbf{x}) + \mathbf{e}(t)$, $\frac{d\mathbf{x}^b}{dt} = \mathcal{M}(\mathbf{x}^b) + \mathbf{e}_b(t)$ (1)
tangent linear model $\delta\mathbf{x}(t_i) = \mathbf{M}(t_i,t_{i-1})\delta\mathbf{x}(t_{i-1}) = \prod_{i=1}^{n}\mathbf{M}(t_i,t_{i-1})\delta\mathbf{x}(t_0)$ (2)
adjoint model $-\frac{d\mathbf{x}^*(t)}{dt} = \mathcal{M}^T(\mathbf{x}^*(t)) + \mathbf{R}^{-1}(\mathbf{y}^o(t) - H[\mathbf{x}(t)])$ (3)

gradient of the cost function

$$\nabla_{\mathbf{x}(t_0),\mathbf{e}}J = -\mathbf{B}_0^{-1}(\mathbf{x}^b(t_0)-\mathbf{x}(t_0)) - \mathbf{K}^{-1}(\mathbf{e}^b(t)-\mathbf{e}(t)) - \sum_{i=1}^{n}\left[\prod_{i=1}^{n}\mathbf{M}^T(t_{i-1},t_i)\right]\mathbf{R}^{-1}(\mathbf{y}^o(t_n) - H[\mathbf{x}(t_n)])$$

Find minimum of $J(\mathbf{x}(t_0),\mathbf{e})$ with $\nabla_{\mathbf{x}(t_0),\mathbf{e}}J$ by use of a minimization routine

Mesoscale EURAD 4D-var data assimilation system

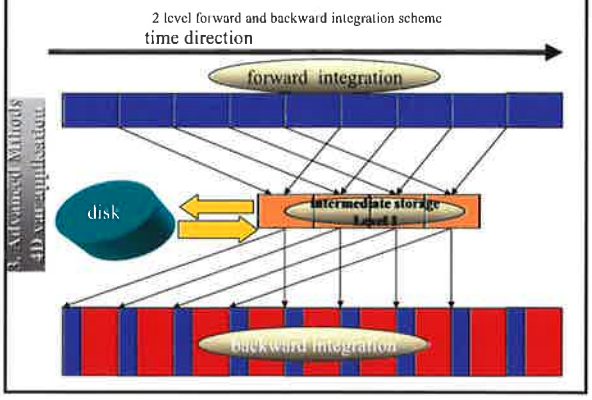


Computational complexity estimate of the variational algorithm

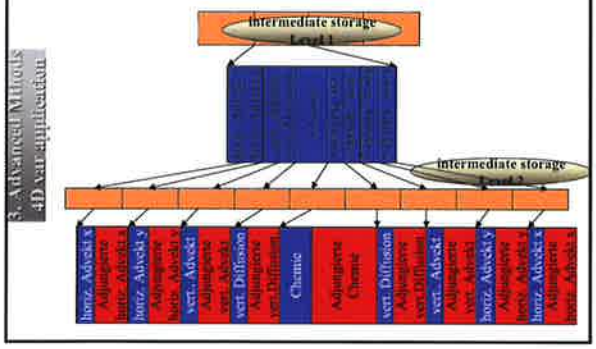
Nx*Ny*Nz	spatial dimensions $O(10^4\text{-}10^5)$
N_c	# constituents O(100)
N_T	# time steps of assimilation window O(10-100)
N_o	# operators O(10)
const	intermediate results $O(10^6)$

Storage strategy	# forward runs/iteration	storage	Complexity [T_{fwd}]
total storage	1	const*N_x*N_y*N_z*N_c*N_T*N_o $O(10^{12}\text{-}10^{13})$	3
operatorwise 1 level	2	N_x*N_y*N_z*N_c*N_T*N_o $O(10^8\text{-}10^9)$	4
dynamic stepwise 2 levels	3	N_x*N_y*N_z*N_c*N_T $O(10^7\text{-}10^8)$	5

2 level forward and backward integration scheme



storage sequence: level 2, operator split
(each time step)



Incremental Formulation

- Analysis State: $x^a = x^b + \delta x^a$
  $u^a = u^b + \delta u^a$
- New „State" Variables: $v = \mathbf{B}^{-1/2}\delta x$
  $w = \mathbf{K}^{-1/2}\delta u$
- Cost Function:
  $$J(v,w) = \frac{1}{2}v^T v + \frac{1}{2}w^T w + \frac{1}{2}\left[\mathbf{H}\delta x_i - d_i\right]^T \mathbf{R}^{-1}\left[\mathbf{H}\delta x_i - d_i\right]$$
- Gradient:
  $$\nabla_v J = \nabla_v J_{1V} + \nabla_v J_O = v + \mathbf{B}^{T/2}\nabla_{\delta v} J_O$$
  $$\nabla_w J = \nabla_w J_{EF} + \nabla_w J_O = w + \mathbf{K}^{T/2}\nabla_{\delta w} J_O$$

Background Error Covariance Matrix B

- must be provided as an operator (size is of order 10^{13})
- we would like to have an operator which can easily be factorised by $B^{1/2}B^{T/2}$
- *a choice under testing:*
  - generalized diffusion equation serves for a valid operator generating a positive definite covariance operator
  - diffusion equation is self adjoint
  - $B^{1/2}$ and $B^{T/2}$ by applying the diffusion operator half the diffusion time

$$\mathbf{B} = \Sigma\,\mathbf{C}\,\Sigma$$
$$\mathbf{C} = \mathbf{C}^{1/2}\mathbf{C}^{T/2}$$
$$\mathbf{C}^{1/2} = \Lambda\mathbf{L}_v^{1/2}\mathbf{L}_h^{1/2}$$
$$\mathbf{C}^{T/2} = \mathbf{L}_h^{T/2}\mathbf{L}_v^{T/2}\Lambda$$

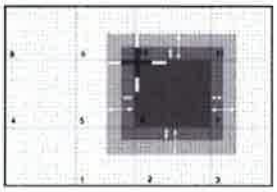
Calculation Speed → Parallel Implementation

- three complete forward runs + one single adjoint run
- an adjoint piece of code needs about twice the time

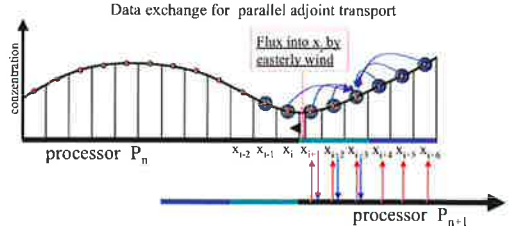
→ 20 iterations need (3 + 2) • 20 = 100 times of a forward run

→ parallel implementation needed

- grid partitioning: „domain decomposition method"

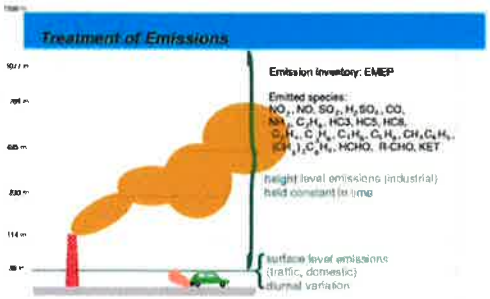


Data exchange for parallel adjoint transport



- Grid points contributing to the calculation of the easterly flux into x_i (4th order pol.)
- Grid points contributing to the calculation of the adjoint of grid point at x_{i-3}

Treatment of the inverse problem to infer emission rates

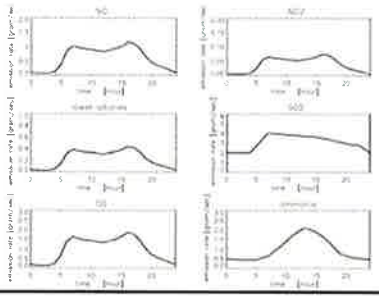


Normalised diurnal cycle of anthropogenic surface emissions *f(t)*

emission(t)=f(t;location,species,day) * v(location,species)

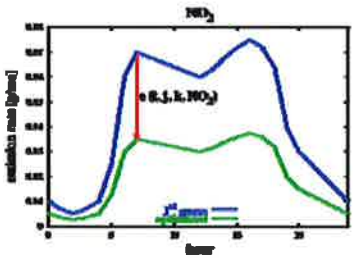
day in (working day, Saturday, Sunday) *v optimization parameter*

3c Advanced Methods 4D var application

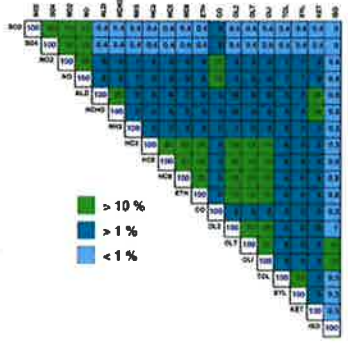


Optimisation of emission rates

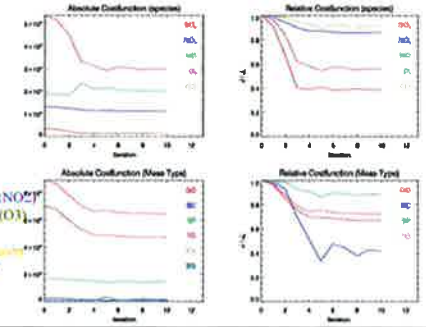
- amplitude optimisation for each emitted species and grid cell

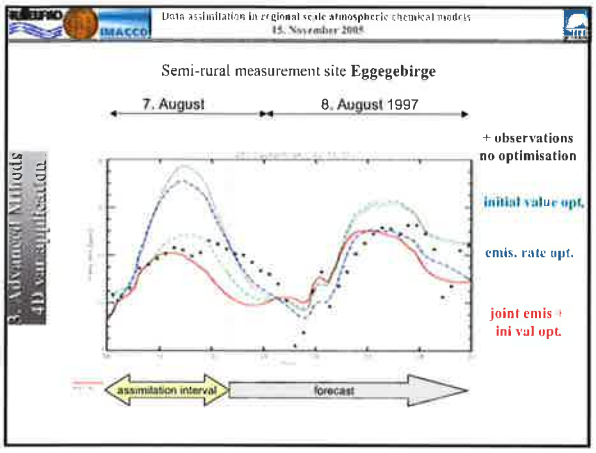


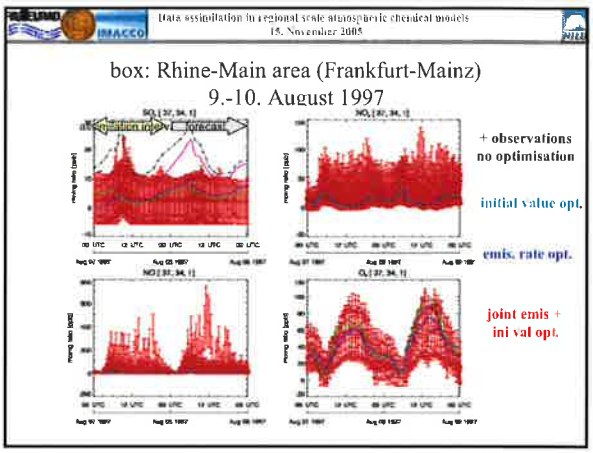
Background emission rate covariance matrix

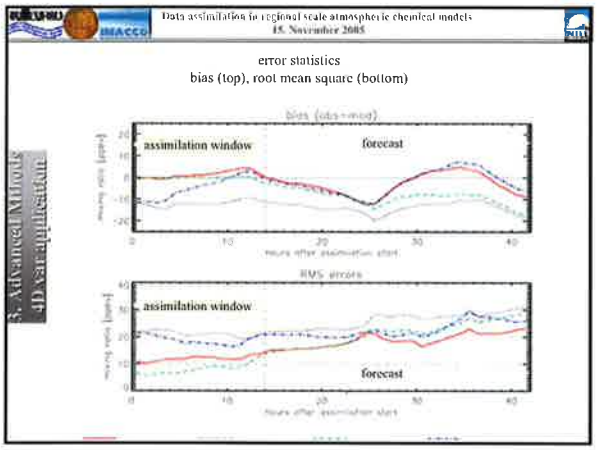


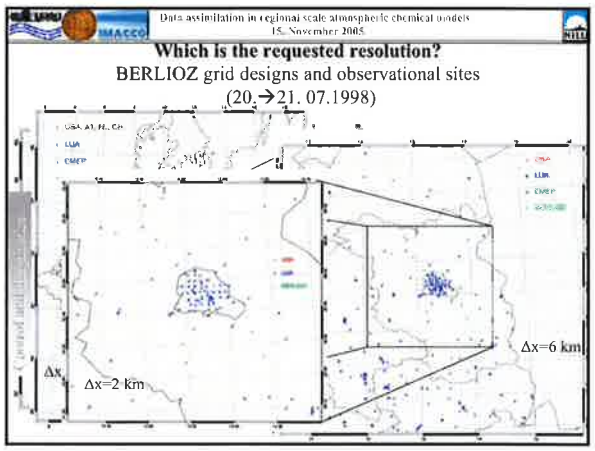
Reduction of the partial cost functions
relative to observation type (coarse grid, 54 km)

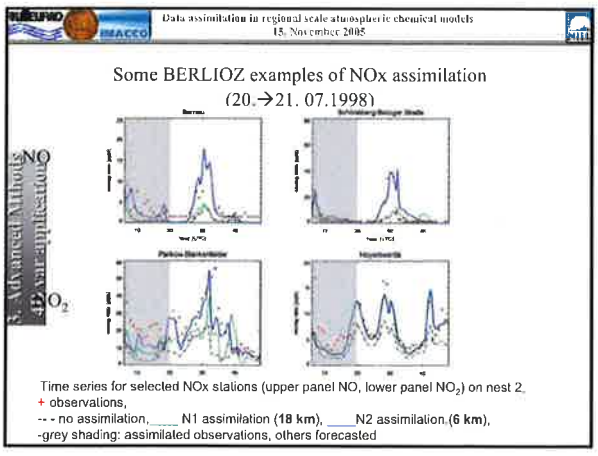


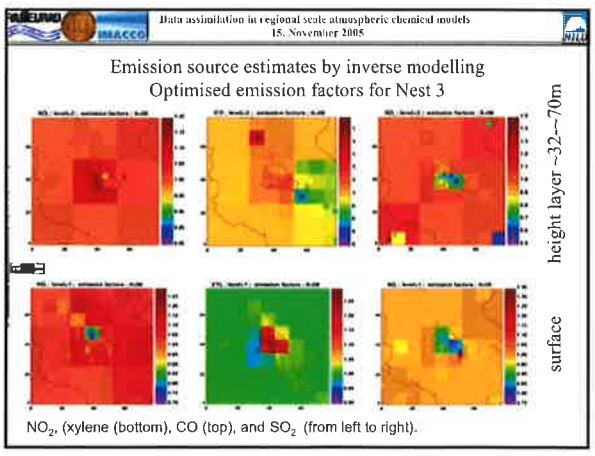

Data assimilation in regional scale atmospheric chemical models
15. November 2005
Semi-rural measurement site Eggegebirge
7. August
8. August 1997
+ observations
no optimisation
initial value opt.
emis. rate opt.
joint emis + ini val opt.
3. Advanced Methods 4D var application
assimilation interval
forecast

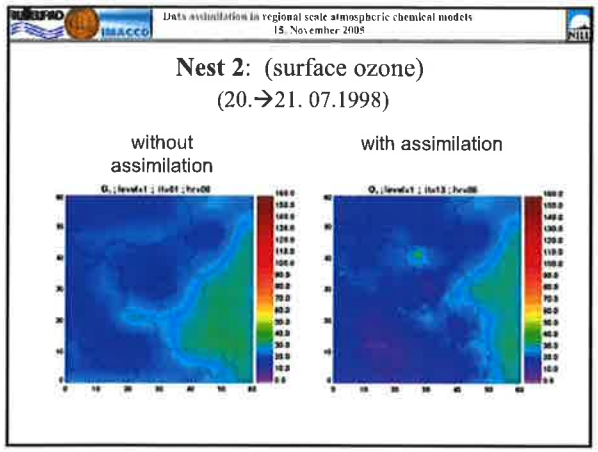

Data assimilation in regional scale atmospheric chemical models
15. November 2005
box: Rhine-Main area (Frankfurt-Mainz)
9.-10. August 1997
+ observations
no optimisation
initial value opt.
emis. rate opt.
joint emis + ini val opt.

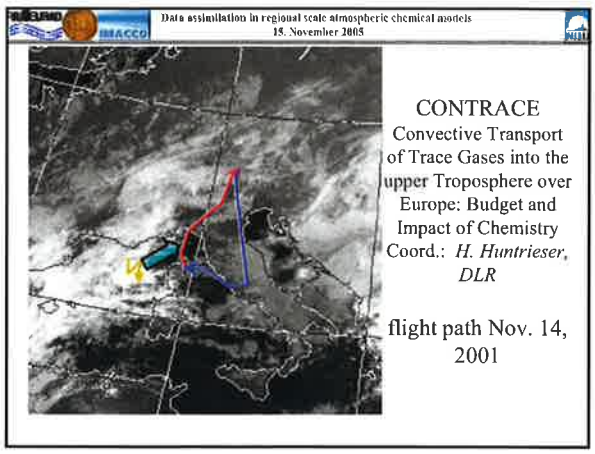

Data assimilation in regional scale atmospheric chemical models
15. November 2005
error statistics
bias (top), root mean square (bottom)
assimilation window
forecast
3. Advanced Methods 4D var application

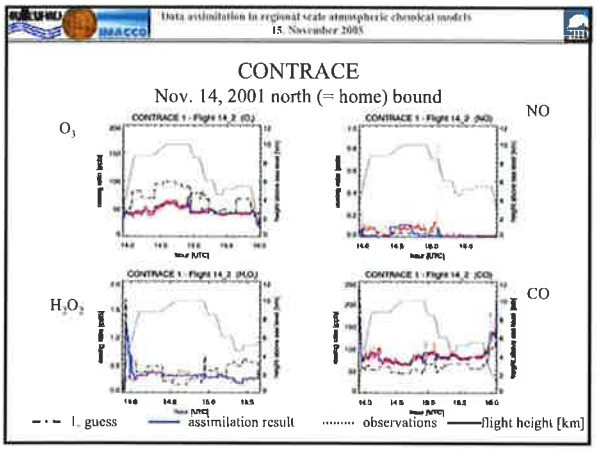

Data assimilation in regional scale atmospheric chemical models
15. November 2005
Which is the requested resolution?
BERLIOZ grid designs and observational sites
(20.→21. 07.1998)
Δx=6 km
Δx=2 km
Δx
3. Advanced Methods 4D var application

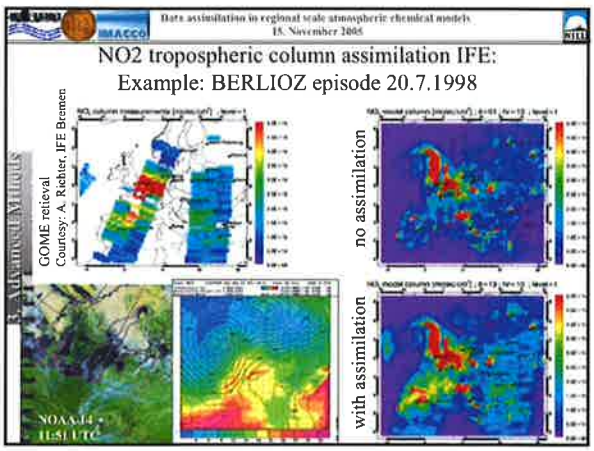

Data assimilation in regional scale atmospheric chemical models
15. November 2005
Some BERLIOZ examples of NOx assimilation
(20.→21. 07.1998)
NO
NO_2
3. Advanced Methods 4D var application
Time series for selected NOx stations (upper panel NO, lower panel NO_2) on nest 2.
+ observations,
-- - no assimilation, ____ N1 assimilation (18 km), ____ N2 assimilation (6 km),
-grey shading: assimilated observations, others forecasted

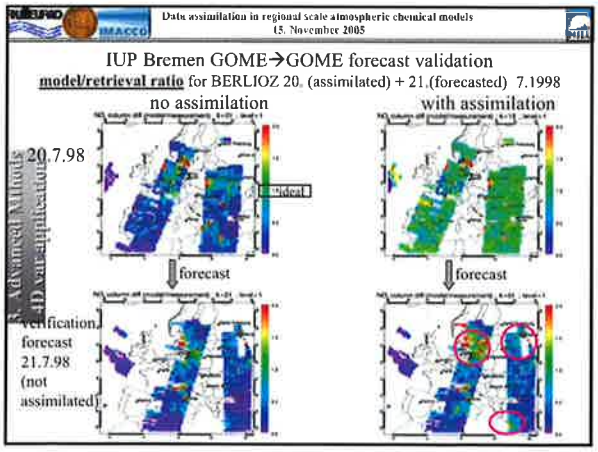

Data assimilation in regional scale atmospheric chemical models
15. November 2005
Emission source estimates by inverse modelling
Optimised emission factors for Nest 3
height layer ~32–70m
surface
NO_2, (xylene (bottom), CO (top), and SO_2 (from left to right).

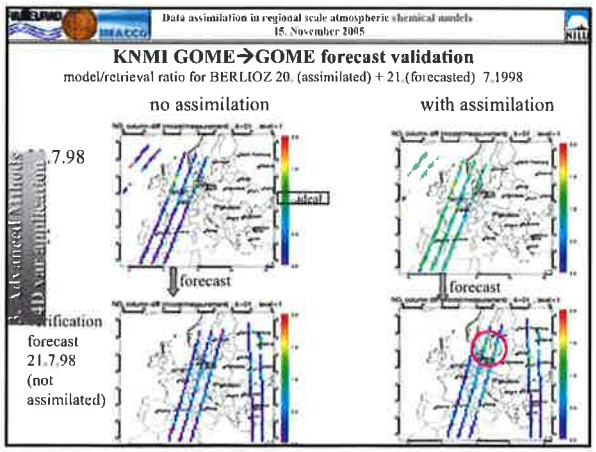

Data assimilation in regional scale atmospheric chemical models
15. November 2005
Nest 2: (surface ozone)
(20.→21. 07.1998)
without assimilation
with assimilation


Data assimilation in regional scale atmospheric chemical models
15. November 2005
CONTRACE
Convective Transport of Trace Gases into the upper Troposphere over Europe: Budget and Impact of Chemistry
Coord.: H. Huntrieser, DLR
flight path Nov. 14, 2001


Data assimilation in regional scale atmospheric chemical models
15. November 2005
CONTRACE
Nov. 14, 2001 north (= home) bound
O3
NO
H2O2
CO
1. guess
assimilation result
observations
flight height [km]

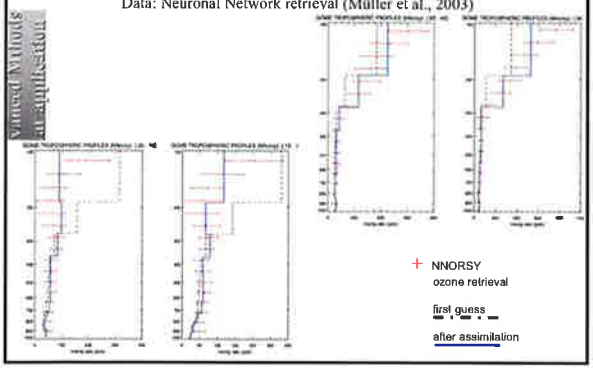

Data assimilation in regional scale atmospheric chemical models
15. November 2005
NO2 tropospheric column assimilation IFE:
Example: BERLIOZ episode 20.7.1998
GOME retieval
Courtesy: A. Richter, IFE Bremen
no assimilation
with assimilation
NOAA-14 ~ 11:51 UTC


Data assimilation in regional scale atmospheric chemical models
15. November 2005
IUP Bremen GOME→GOME forecast validation
model/retrieval ratio for BERLIOZ 20. (assimilated) + 21.(forecasted) 7.1998
no assimilation
with assimilation
20.7.98
ideal
forecast
verification forecast 21.7.98 (not assimilated)


Data assimilation in regional scale atmospheric chemical models
15. November 2005
KNMI GOME→GOME forecast validation
model/retrieval ratio for BERLIOZ 20. (assimilated) + 21.(forecasted) 7.1998
no assimilation
with assimilation
ideal
forecast
verification forecast 21.7.98 (not assimilated)

GOME ozone profile assimilation BERLIOZ (20.7.1998)

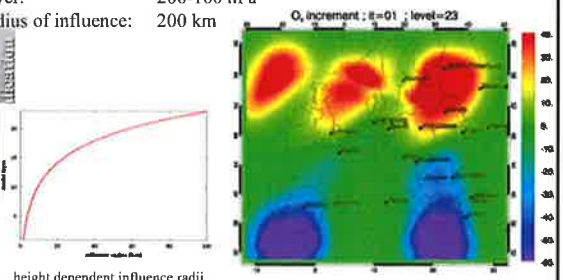
Data: Neuronal Network retrieval (Müller et al., 2003)



Ozone observation increment 20.07.1998

layer: 200-100 hPa

radius of influence: 200 km

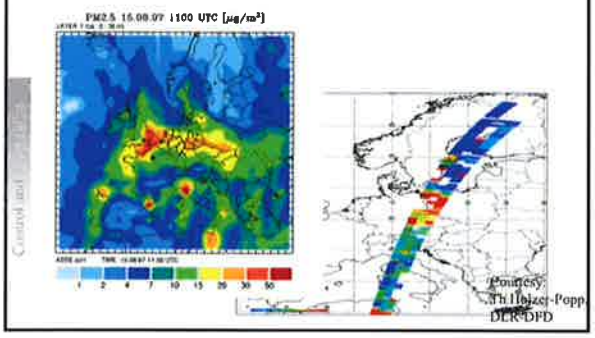


height dependent influence radii

An examplefor aerosol data assimilation

PM2.5 [µg/m³]

Note: colour code scaling different!



Summary

- chemical weather forecasts are a multiple scale problem
- chemical data assimilation rests on sparse and heterogeneous observations, with variable error characteristics (incl. error of representativity)
- initial value optimisation is insufficient, as at least emission rates are less known and more important
- the ability for inversion is therefore required to optimize emission rates
- much is to be done for optimising multivariate covariance matrices

Preliminary conclusions

- fine grid resolution is required for air quality modelling, however
- the finer the grid the more critical is meteorological modelling
- covariance matrices and performance statistics are essential → operational application are essential
- tangent linear approximations and Gaussian error assumptions often violated
- sequential assimilation algorithms appear to be less suited for treatment of temporal error correlations
- ensemble ideas appear useful in one or another way

DATA ASSIMILATION IN ATMOSPHERIC CHEMISTRY MODELS USING ENSEMBLE METHODS

Arnold Heemink

Delft University of Technology

Joint work with Martin Verlaan, Remus Hanea, Alina Barbu and Peter Vermeulen

12/14/2005 1

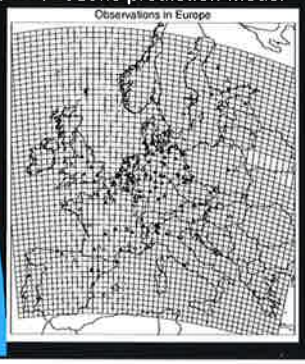
Overview

- Introduction and motivation: Some real life applications of data assimilation
- Ensemble Kalman filter algorithms for large scale systems
- Model reduced variational data assimilation

12/14/2005 2

Some real life applications of data assimilation

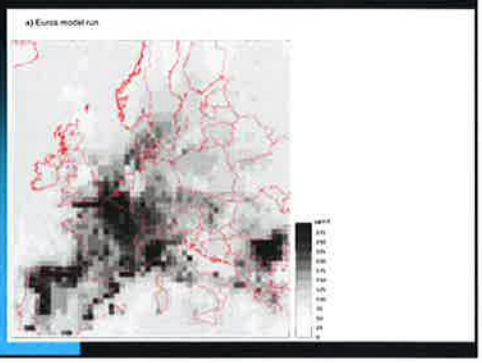
Grid of Ozone prediction model



12/14/2005 3

Some real life applications of data assimilation

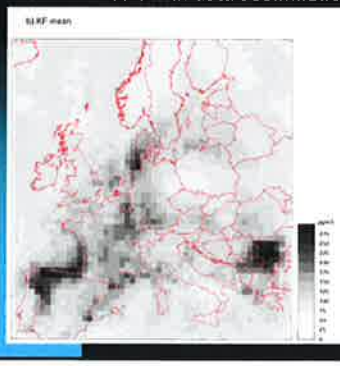
Model result without data assimilation



12/14/2005

Some real life applications of data assimilation

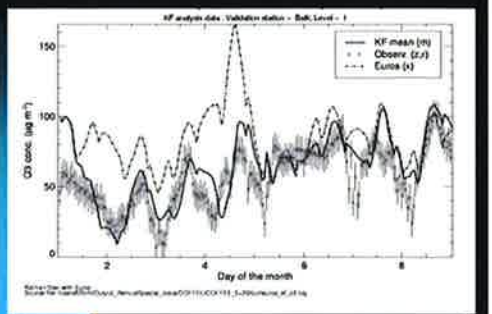
Model result with data assimilation



12/14/2005

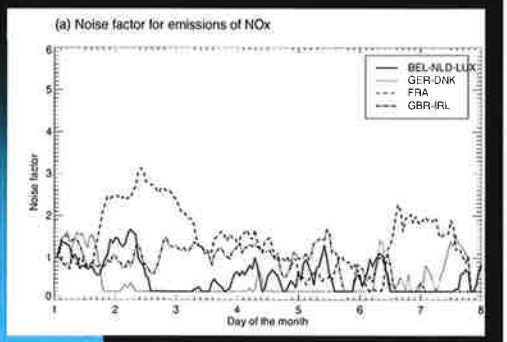
Some real life applications of data assimilation

Ozone concentration in validation station



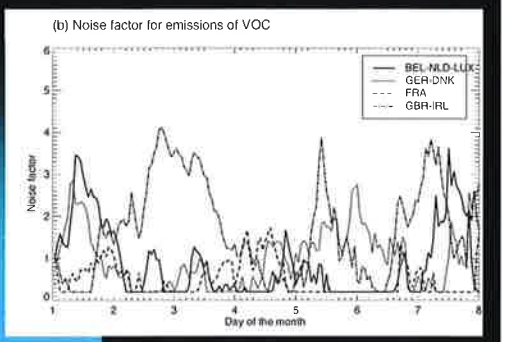
12/14/2005

Some real life applications of data assimilation



12/14/2005 7

Some real life applications of data assimilation



12/14/2005 8

State space model

The (non linear) physics:

$$X_{k+1} = f(X_k, p, k) + G(k)W_k$$

where X is the state, p is vector of uncertain parameters, f represents the (numerical) model, G is a noise input matrix and W is zero mean system noise with covariance Q

The measurements:

$$Z_k = M(k)X_k + V_k$$

where M is the measurement matrix and V is zero mean measurement noise with covariance R

12/14/2005 9

Formulation of the weak constraint data assimilation problem

It is desired to combine the data with the stochastic model in order to obtain an optimal estimate of the state and parameters of the system.

We define the criterion (MAP estimate):

$$J(p, X_k) = \sum_{k=1}^{K} \|Z_k - M(k)X_k\|_R^2 + \sum_{k=0}^{K-1} \|X_{k-1} - f(X_k, p, k)\|_{(GQG^T)^{-1}}^2 + \alpha \|p - p_0\|_{C^{-1}}^2$$

12/14/2005 10

Linear dynamics F(k) and constant parameters: State estimation using Kalman filtering

A recursive algorithm for k=1,2,... to determine :

- X_k^f Optimal estimate of the state at time k using measurements up to and including k-1
- P_k^f Covariance matrix of the estimation error
- X_k^a Optimal estimate of the state at time k using measurements up to and including k
- P_k^a Covariance matrix of the estimation error

12/14/2005 11

Ensemble Kalman filter (EnKF)

To represent the probability density of the state estimate N ensemble members are chosen randomly:

$$\hat{x} = \frac{1}{N} \sum \xi_i$$

$$S = [\ldots \frac{\xi_i - \hat{x}}{\sqrt{N-1}} \ldots]$$

$$P = SS^T$$

12/14/2005 12

A model reduction approach to data assimilation

Consider the q dimensional sub space:

$P=[p_1 \; p_2 \; \ldots \; p_q]$

And project the original model onto this sub space:

$$r_{k+1} = [P^T \frac{\partial(f)_k}{\partial(X)_k} P] r_k$$

$$Z_k = [M(k)P] r_k + v_k$$

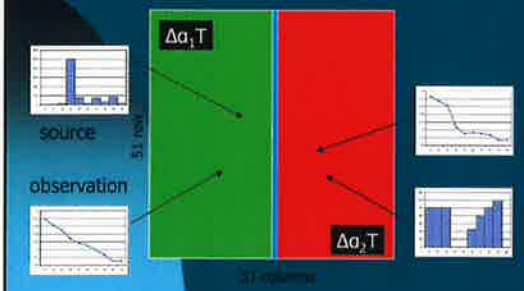
We now have a q dimensional, explicit (approximate) system description including its adjoint!

The sub space can be determined by computing the EOF (Empirical Orthogonal Functions) of an ensemble of model simulations

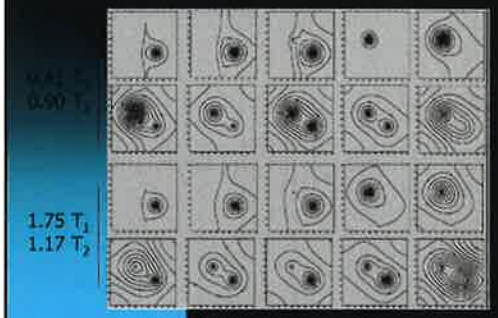
A model reduction approach to data assimilation

- Generate an ensemble of N model simulations
- Determine the q dominant EOF's: sub space P
- Project original model onto P (this requires q additional model simulations). The adjoint is now available too.
- Perform the optimization in reduced space
- Repeat the process from the start if necessary

Test model: Ground water flow (Diffusion equation)

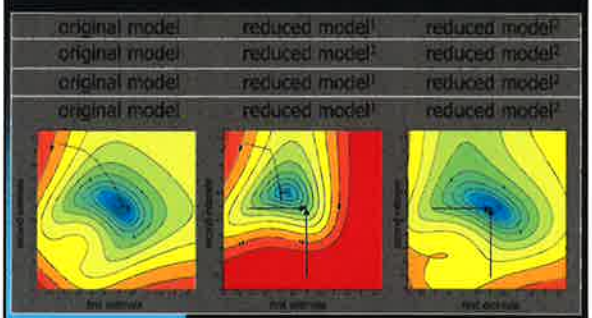


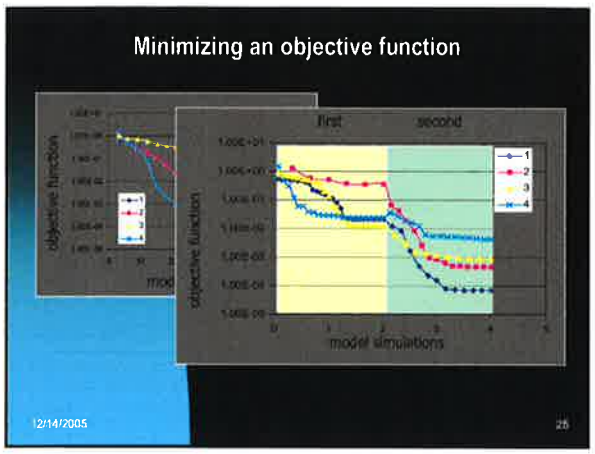
Model Results – snapshots water level

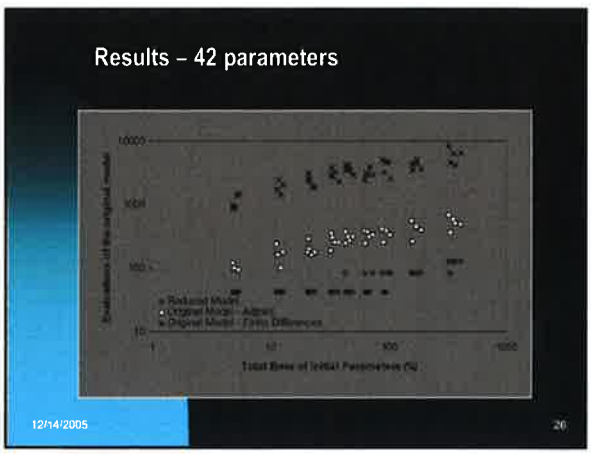


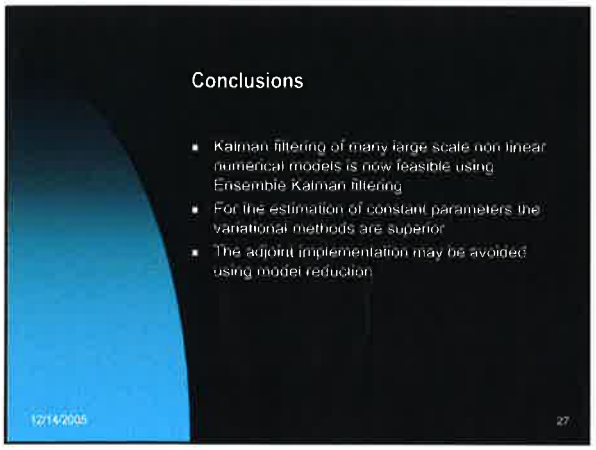
EOF - patterns

Minimizing an objective function




Minimizing an objective function
first
second
objective function
model simulations
1
2
3
4
12/14/2005
25


Results – 42 parameters
12/14/2005
26


Conclusions
Kalman filtering of many large scale non linear numerical models is now feasible using Ensemble Kalman filtering
For the estimation of constant parameters the variational methods are superior
The adjoint implementation may be avoided using model reduction
12/14/2005
27

Norwegian Institute for Air Research (NILU)
P.O. Box 100, N-2027 Kjeller, Norway

REPORT SERIES	REPORT NO. OR 43/2006	ISBN 82-425-1766-5	
SCIENTIFIC REPORT		ISSN 0807-7207	
DATE 22.09.06	SIGN. Kjetil Torkfeld	NO. OF PAGES 72	PRICE NOK 150,-

TITLE	PROJECT LEADER
Data assimilation in regional scale atmospheric chemical models NMR Workshop at NILU, Kjeller, Norway, 15 November 2005	Bruce Denby NILU PROJECT NO. O-105077
AUTHOR(S) Editor: Bruce Denby Contributing authors: Jørgen Brandt , Hendrik Elbern, Jan Frydendall , Arnold Heemink , Martin Hvidberg, Michael Kahnert, Leonor Tarrason, Maarten van Loon, Sam Erik Walker and Zahari Zlatev	CLASSIFICATION * A CONTRACT REF. 05FOX10

REPORT PREPARED FOR
NMR/LHG
Vandenheden
Miljøstyrelsen
Strandgade
291401
København K

ABSTRACT
This workshop is part of a Nordic Council of Ministers (NMR) funded project which supports the development and communication of scientific research between Nordic countries on the topic of 'data assimilation in regional scale atmospheric chemistry models'. The four institutes involved, NILU (Norway), met.no (Norway), DMU (Denmark) and SMHI (Sweden), all have active programmes in data assimilation. The intention of this project and workshop is to bring together these institutes to share knowledge and experience within a Nordic context and to further support development in this research area. In total 21 people attended the workshop, including invited experts in data assimilation from Europe. The workshop showed itself to be successful, being both informative and helpful to the participants. This report consolidates the presentations and discussions that took place during the workshop.

NORWEGIAN TITLE
Data assimilation in regional scale atmospheric chemical models

KEYWORDS		
Data assimilation	Atmospheric chemistry models	Regional

ABSTRACT (in Norwegian)

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