Ensemble-based data assimilation schemes for atmospheric chemistry models

Proefschrift

ter verkrijging van de graad van doctor
aan de Technische Universtiteit Delft,
op gezag van de Rector Magnificus prof.ir. K.C.A.M. Luyben,
voorzitter van het College voor Promoties,
in het openbaar te verdedigen op donderdag 22 april 2010 om 10:00 uur
door

Alina Lavinia BARBU

Diplôme d' études approfondies, Université de Technologie Compiègne, Frankrijk
geboren te Găeşti, Roemenië

Dit proefschrift is goedgekeurd door de promotoren:

Prof.dr.ir. A.W. Heemink
Prof.dr.ir. P.J.H. Builtjes

Samenstelling promotiecommissie:

Rector Magnificus,	voorzitter
Prof.dr.ir. A.W. Heemink,	Technische Universiteit Delft, promotor
Prof.dr.ir. P.J.H. Builtjes,	Free University of Berlin, Germany, promotor
Prof.dr.ir. C. Vuik,	Technische Universiteit Delft
Prof.dr.ir. H.W.J. Russchenberg,	Technische Universiteit Delft
Prof.dr. K. Ponnambalam,	University of Waterloo, Canada
Dr. P. Sakov,	NERSC Bergen, Norway
Dr. M. Schaap,	TNO Utrecht, The Netherlands

This research was carried out in the Delft Institute of Applied Mathematics, Delft University of Technology in cooperation with TNO Built Environment and Geosciences, Department of Air Quality and Climate, Utrecht, The Netherlands. Support of this work was provided by the Netherlands Institute for Space Research (SRON).

ISBN 978-90-9025297-1

Copyright © 2010 by Alina L. Barbu

All rights reserved. No part of the material protected by this copyright notice may be reproduced or utilized in any form or by any means, electronic or mechanical, including photocopying, recording, or by any information storage and retrieval system, without prior written permission from the author.

Printed in The Netherlands by Wöhrmann Print Service

To my sons, Andrei and Ştefan

Contents

1 Introduction and overview 1
- 1.1 Air pollution 1
  - 1.1.1 Sources of sulphur dioxide and nitrogen oxides 2
- 1.2 Atmospheric chemistry models 4
- 1.3 Measurements 5
- 1.4 Data assimilation 7
  - 1.4.1 Introduction 7
    - Bayesian perspective 8
  - 1.4.2 Filtering and Smoothing 10
  - 1.4.3 Kalman-like filtering 12
    - Extended Kalman filter 12
    - Ensemble-based filters 13
  - 1.4.4 Sources of errors in ensemble-based data assimilation . . 13
    - Gaussian assumption and linear updating 14
    - Use of localization 14
    - Use of covariance inflation 14
    - Sampling errors 14
    - Model errors and bias 15
  - 1.4.5 Measuring data assimilation performance 16
- 1.5 Motivation and overview 16

2 Ensemble filter and smoother 19
- 2.1 Introduction 19
- 2.2 Ensemble-based filters 21

2.2.1 The Ensemble Kalman Filter (EnKF) 21
2.2.2 The implementation of the EnKF 22
2.3 Smoother approach 24
2.3.1 Ensemble Kalman smoother (EnKS) 25
2.3.2 The FIFO-lag algorithm 25
2.4 Experimental setup 26
2.5 Assimilation experiments 28
2.5.1 Divergence problem 29
2.5.2 Smoother sensitivity to the ensemble size and lag length 34
2.5.3 Computational efficiency 36
2.6 Conclusions 36

3 Assimilation of sulphur dioxide and sulphate 39
3.1 Introduction 40
3.2 Deterministic LOTOS-EUROS model 41
3.3 Observation network for assimilation and validation 44
3.4 Model validation 45
3.5 Sequential data assimilation methodology 47
3.5.1 Stochastic state space representation 47
3.5.2 Observational operator 48
3.5.3 Ensemble Kalman filter and smoother 48
3.5.4 Localization procedure 50
3.6 Assimilation results 50
3.6.1 Experimental setup 50
3.6.2 Uncertainties in emissions 51
3.6.3 Uncertainties in emissions and reaction rate 54
3.7 Conclusions and discussion 60

4 Bias aware assimilation of OMI NO_2 tropospheric columns 65
4.1 Introduction 65
4.2 Satellite observations 67
4.3 The version v1.5 of the LOTOS-EUROS model 68
4.4 Model validation 68
4.5 State and bias estimation 71
4.6 Assimilation results and discussions 74
4.7 Conclusions 77

5 Square root filters 79
5.1 Introduction 80
5.2 Ensemble-based filters 81
5.2.1 The Ensemble Kalman Filter 82
5.2.2 Ensemble Square Root Filters 83

5.3 Reduced-rank square root filter 85
5.4 Smallest analysis increments 87
5.5 Experimental setup 88
5.5.1 2D pollution model 88
5.5.2 Lorenz 40-variables model 90
5.6 Results and discussions 90
5.6.1 2D pollution model 91
5.6.2 Lorenz 40-variables model 95
5.7 Conclusions 96

6 Summary, Conclusions and Outlook 99
6.1 Overview 99
6.2 Summary and conclusions 100
6.3 Outlook 101

Appendix 103
The LOTOS-EUROS model 103
Domain 103
Transport 104
Chemistry 104
Emissions 104
Removal processes 104

Bibliography 104

Samenvatting 117

Acknowledgements 121

Curriculum Vitae 123

CHAPTER 1

Introduction and overview

1.1 Air pollution

The perception of air pollution that we are most familiar with is represented by the smog that cause a reduction of the visibility. The air around industrial and densely populated areas is characterized by high concentrations of waste gases from fossil fuel combustion. These gases accumulate at such levels that they become visible by forming a colored combination of smoke and fog, simply referred as smog.

Under conditions of high temperature, the waste gases such as nitrogen oxides degrade in the atmosphere. One of the degradation product is the tropospheric ozone called summer smog, a highly reactive oxidant which damage natural ecosystem and is toxic to humans. Ground level ozone is formed when NOx and volatile organic compounds (VOCs) react in the presence of sunlight. Ozone can be transported by wind currents and cause health impacts far from original sources. Furthermore, ozone in the troposphere contributes to the enhanced greenhouse effect.

Atmospheric particulate matter is a complex mixture of anthropogenic and natural airborne particles. The main components are secondary inorganic aerosols (sulphate, nitrate and ammonium), combustion particles, primary and secondary biogenic aerosols, sea salt and earth crust materials. Winter smog is mainly related to the inorganic aerosol species and their precursor gases. Particulate matter (PM) or aerosol in ambient air has been also associated consistently with excess mortality and morbidity in human population, e.g. [12], [60]. Children, people with lung diseases such as asthma, and people who work or exercise outside are susceptible to adverse effects such as damage to lung tissue and reduction in lung function.

The various components of particulate matter in the atmosphere also have climate-forcing impacts, either contributing to or offsetting the warming effects of greenhouse gases. Aerosols affect climate by scattering and absorbing the incoming solar radiation and by their effect on the albedo and lifetimes of clouds [20] showed that the total radiative forcing by anthropogenic sulphate aerosols of the same order of magnitude, but opposite sign as the radiative forcing by man-made greenhouse gases. Due to the complex relation between the aerosol properties and the Earth's radiation balance, the highly inhomogeneous aerosol distribution in both space and time and the lack of knowledge of the PM properties on the global scale, the PM are considered one of the largest source of uncertainty in climate modeling.

Furthermore, secondary inorganic aerosol formation and transport has been studied for decades as they contribute to acidification of soils. Acid rain events occur in and downwind of regions with large amounts of man-made pollution. The main components of acid rain are ammonia (NH_3) as well as nitric acid (HNO_3) and sulphuric acid (H_2SO_4) produced by the atmospheric oxidation of nitrogen oxides and sulphur dioxide, respectively. These compounds react with other substances in the air to form acids which fall to earth as rain, fog, snow or dry particles that can be carried by wind for hundreds of miles. Vegetation systems is damaged and water sources become acidic and unsuitable for many fish. Acid rain also causes deterioration of cars, buildings and historical monuments.

Eutrophication is the phenomenon known as nutrient pollution. Increased nitrogen loading in water disturb the chemical balance of nutrients used by aquatic plants and animals. Additional nitrogen accelerates "eutrophication," which leads to oxygen depletion and reduces fish and shellfish populations.

Hence, for all these issues a thorough knowledge of the concentrations of pollutants as well as their sources and sinks is needed. Of all air pollution issues, this thesis is placed in the context of winter and summer smog with focus on their precursor gases, namely sulphur dioxide and nitrogen dioxide.

1.1.1 Sources of sulphur dioxide and nitrogen oxides

Sulphur dioxide is the chemical compound with the formula SO_2. The primary sources of sulphur dioxide are burning of fossil fuels and volcanic emissions. Sulphur dioxide has a rather short lifetime in air, turning into sulfate aerosol particles in about a day near the ground and in a month in the stratosphere. The sulfate can combine with water to create a haze of sulfuric acid aerosol. Figure 1.1 is an illustration of the emission sources (in megatons) for anthropogenic activities in Europe [35]. The major source is the energy generation which accounts for more than half of the total emissions. Secondary sources are industrial combustion and (international) shipping.

Nitrogen oxides (NOx) is an air pollutant harmful to humans and ecosys-

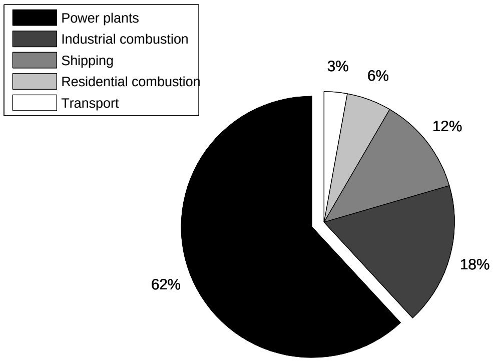

Figure 1.1: The relative importance of SOx emission sources for anthropogenic activities in Europe.

tems and plays a key role in tropospheric photochemistry as a precursor of ozone and by regulating the abundance of the OH radical. NOx emissions in the air are one of the largest sources of nitrogen pollution. In order to understand the role of nitrogen oxides in atmosphere, there is important to quantify its sources and sinks. The largest sources are over the continent and produce NOx at the surface layer mainly from fossil fuel combustion and biomass burning. Figure 1.2 illustrates the NOx sources and loads (in gigatones) in The Netherlands for 2006 as they have been reported in the EMEP database. Road transport is the most important source for nitrogen oxides. Four sources categories that contribute about equally to the other 60% of the total Dutch emission are industrial and non industrial combustion, power plants and other mobile sources. Compared to SOx, the NOx emissions are more diffuse as the sulphur emissions are dominated by large plants whereas the NOx emissions are dominated by individual small combustion bits, such as cars and other machinery. The large uncertainties (around 30-40%) in source estimates is a serious problem in quantifying the nitrogen dioxide and sulphur dioxide budgets and their impact on atmospheric chemistry. A good knowledge of the spatial distribution of these components is important for assessing the air quality over a given domain. Thereby, not only mean concentrations but also peak loads are relevant and air quality limits have been defined in Europe for both. Hence, also the variability or timing of the emissions throughout the year is important.

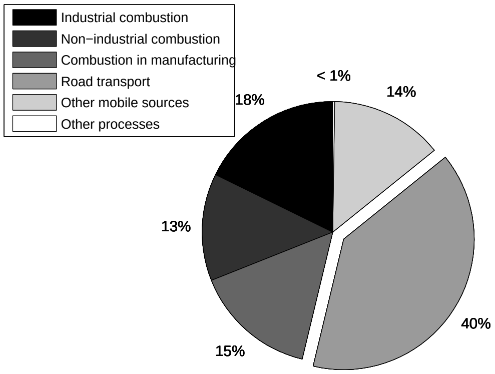

Figure 1.2: The relative importance of NOx emission sources for anthropogenic activities in The Netherlands in 2006.

1.2 Atmospheric chemistry models

Knowledge of the mechanism that governs atmospheric composition is summarized in models. During the last decades, different sources of information are incorporated into such models in order to describe more accurate the processes that influence atmosphere (weather, emissions and removals of gases, chemical reactions, transport processes, interactions with vegetation). These models are called chemistry transport models (CTMs). Since air pollution has non-local origin and pollutants may be transported within the atmosphere over hundreds of kilometers, the value of a CTM resides in its capability of predicting changes in the atmosphere from the scale of countries to continents or the whole globe. Atmospheric chemical modeling is highly important for understanding the impact of air pollutant emissions on the chemical composition of the atmosphere and the consequences on the natural environment and public health. Models are useful for predicting the effect of proposed changes in emission rates upon the level of air pollution. The atmospheric chemistry model that include predictions of the concentrations of ozone, toxic air pollutants, nitrogen compounds and atmospheric acids are important sources of information for public, industry and government policy.

Despite advances in computer technology, data collection and numerical modeling techniques, performance evaluations of chemical state of the atmosphere in air quality models demonstrate that their solution can contain important errors. The causes of these errors in the models come from various sources: uncertainties in deriving data input as emissions, approximations in the model formulation, bias due to the aims and limitations of the model, meteorological variations. Both man-made and natural emissions of air pollutants are highly uncertain due to their variability in time and space. For example, emissions from the combustion of fossil fuels as sulphur dioxide are larger in the winter than in the summer due to the increased burning of fossil fuels for domestic heating. Spatial variability of the emissions can be exemplified by the NOx emissions resulting from fuel combustion in the transport sector. Also natural emissions of nitrogen oxides by lightning are highly variable.

Taking into account that CTMs at continental scale are designed to simulate the fate of air pollution within boundary layer more accurate than the global models, several European chemistry transport models with the horizontal resolution of several tenths of kilometers have been developed such as EMEP [100], EURAD [54], CHIMERE [98] and others. To evaluate and predict the future trends in the concentrations of air pollution in the atmosphere, several simulation models have been developed in the Netherlands too. KNMI, RIVM and TNO have independently developed models to calculate the dispersion and chemical transformation of air pollutants in the lower troposphere over Europe. Two of these models are the TNO model LOTOS [15], [95] and the RIVM model EUROS [27], [106], [107]. LOTOS and EUROS were originally developed and used as photo-oxidant models by e.g. [15], [51]. During the last years attention was given to simulate the inorganic secondary aerosols SO_4, NH_4 and NO_3 by [97], [38] and carbonaceous aerosols [96]. The EUROS model also contains the possibility to perform simulations for persistent organic compounds [65]. The two models have a similar structure and comparable application areas. Hence, based on strategic and practical reasoning, RIVM/MNP and TNO agreed to collaborate on the development of a single chemistry transport model: LOTOS-EUROS. This unified model is used in this thesis.

1.3 Measurements

Scientific understanding of atmospheric chemistry is based upon experimental measurements performed in the laboratory, environmental chambers or in the field. Despite the fact that observational error distribution of atmospheric chemical system is often poorly known, significant advancements have been made in recent years in our ability to measure atmospheric chemistry.

Ground-based, air-bone and satellite information can be used in combination with an atmospheric chemistry model. Ground-based experiments can provide detailed information on gas phase concentrations, aerosol mass and size distribution and chemical composition. In situ measurements are considered the authoritative source for judgment of air quality. Several ground-based data networks are currently in operation to monitor air pollution concentrations, but these are limited in space and hence inadequate to provide a good coverage over the European domain. In general, in situ measurements provide accurate data. However, the representativity of the data is always an issue. In addition some measurement techniques may be prone to artifacts. For example, NO_2 measurements obtained with chemiluminescence monitors may be yield an overestimate as other components may be measured along and interpreted as NO_2. Moreover, the use of different equipment in different parts of a network may yield an inhomogeneous data set with respect to data quality.

The air quality data in Europe is organized through the EMEP network and national air quality agencies that provides the hourly and daily measurements near surface. The advantages of EMEP measurements are the common quality control standards applied as well as their site locations that makes the data relatively unaffected by local emissions. An important disadvantage is that the EMEP network provides a rather sparse distribution of stations for several species. On the other hand, local networks (urban, suburban) may have inhomogeneous data criteria selection or treatment of uncertainty.

Satellite measurements become in recent years a significant source of information for air pollution. Global observations are now available for a wide range of species including aerosols, tropospheric ozone and nitrogen dioxide. The advantage of satellite data is that these instruments yield in principle consistent data over a large domain and provide a full spatial coverage. They provide useful information on areas where other measurement are sparse or expensive, for example over sea. However, they are less precise than ground-based measurements and only supply data during cloud free conditions. This suggest that satellite data may be useful to improve the insight in the distributions of air pollutants when used complementary to ground-based observations. Satellite information of atmospheric composition need to be validated with independent measurements in order to be usable for air pollution monitoring. Validation, in this sense, means not only comparing numbers of a homogeneous quantity, but also allows for a correct interpretation of the satellite measurements. Examples of two types of data provided by in situ and satellite measurements, respectively are illustrated in Figure 1.3. The left panel represents the surface nitrogen dioxide hourly measured at Vredepeel station in The Netherlands (NL10 in the EMEP database), while the right panel shows the OMI NO_2 tropospheric column measured over a large region for the same day.

To obtain a better understanding of the atmospheric composition measurement and modeling activities should be closely tied. Models are often built based on knowledge obtained from what has been observed and, on this basis,

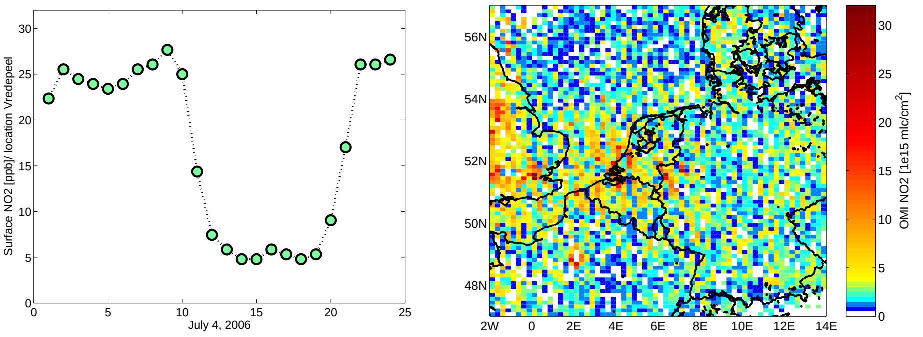

Figure 1.3: NO_2 measured at Vredepeel site (left) and NO_2 tropospheric column observed by OMI (right) at July, 4, 2006.

they are further validated. Reverse, models are used to simulate processes at sites or time for which no measurements are available. Besides, measurements seldom produce a complete picture of the processes, as sampling can only be done at limited locations and time. Moreover many measurements are expensive and several quantities cannot be measured at all. Consequently, it is highly beneficial when both sources of information are used talking into account the merits of both approaches. Combination of data and transport model can be applied for model validation and interpretation of different chemical component distributions, as well as for incorporation of data in the model which is main topic of the following chapters of this thesis.

1.4 Data assimilation

1.4.1 Introduction

Despite the complex chemical modeling in CTMs, these models still show the significant differences when compared with observations, in particular, being pronounced with air quality models. The causes of this bias are highly complex and only partly understood. On one hand, large uncertainties in emissions on spatial and temporal scales, formation routes and sinks of different species may cause model performances to be relatively poor. On the other hand, there are always errors and uncertainty associated with measurements.

Therefore, studying a physical system generally requires both a model for the time evolution of the system and observations. Incorporation of huge number of measurements in large scale models with acceptable computational time is far from obvious. The process of blending the results of a numerical model with the available measurements to obtain an accurate representation of the dynamical behavior of the modeled system is called data assimilation (DA). Data assimilation is recognized as essential in weather prediction, climate analysis and forecast activities in oceanography, hydrology and atmospheric chemistry [48].

Issues related to data assimilation involve choices of these three components: model and measurements, as well as the choice of how such information is combined. Data assimilation incorporates ideas from probability theory, statistics, control and system theory, optimization, estimation theory and other fields. Two largely used assimilation procedures in Geosciences are variational methods and filter techniques. Variational methods [102], [26] are based on the minimization of a penalty function which quantifies the difference between the model trajectory and observations over a period of interest, subject to weak or strong constraints. The minimization procedure requires an adjoint of the forward model. The 4D VAR framework represents the current state-of-art in meteorological data assimilation [26] and chemical data assimilation [33], [92]. While variational DA aims to combine a model with available observations over a time interval, sequential DA is an on-line method that updates the estimation of a state at each time when observations are available. Filtering consists of estimating the system state based on the observations up to this time. Starting with a number of pioneering applications in data assimilation [55] this approach has gain more and more popularity. Ensemble-based assimilation methods were originally developed as computationally feasible approximate solutions of the nonlinear filtering problem. Many studies aimed to discuss theoretically the relative merits of the two approaches to data assimilation and compare their advantages and disadvantages with various geophysical systems [81], [50], [69] including atmospheric chemistry applications [24].

Bayesian perspective

The Bayesian paradigm provides a coherent probabilistic approach for combining information, and thus is an appropriate framework for data assimilation and most effective when the uncertainty in both observations and model are accurately quantified [66]. The state-space approach of estimation theory has been originally presented by [49] in the context of atmospheric data assimilation. The starting point is an existing mathematical model governed by a set of partial differential equations that describes a physical process. In a air quality application, the equation system that describes the evolution of trace gas concentrations for several species in time is discretized according to:

$$x(k+1) \quad = \quad M(x(k)). \tag{1.1}$$

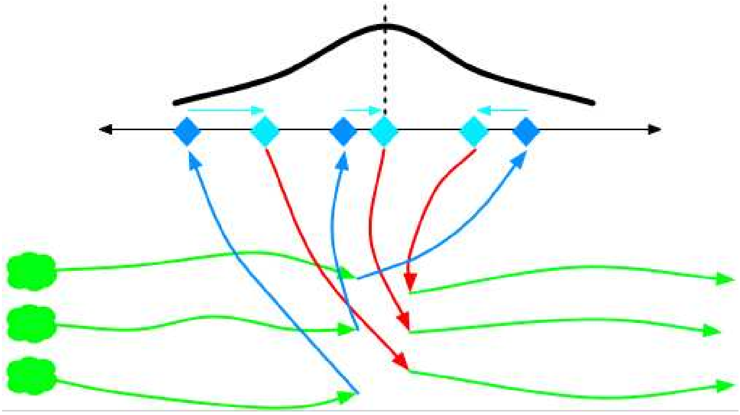

Figure 1.4: Schematic illustration of an ensemble-based technique in the Bayesian context.

where x is n-dimensional vector representing the state of the system at a given time. The elements of this vector are gas-phase concentrations. The state-space operator M describes the time evolution from the time k to k+1 of the state vector. A more realistic problem is represented by a physical system subject to unknown disturbances. For application of the data assimilation with a statistically based algorithm, a stochastic representation of the dynamical model should be written in a state-space form according to:

$$x(k+1) \quad = \quad M(x(k)) + w(k). \tag{1.2}$$

The random forcing term w is drawn from the normal distribution $N(0, Q)$ with Q the covariance matrix.

The state of the observational network is defined by the observation operator H that maps state variables x to observations y. We further assume that the measurements have white Gaussian errors v with covariance denoted by R:

$$y(k) = H(x(k)) + v(k), \quad v \sim N(0, R). \tag{1.3}$$

The full probability model or the joint probability distribution of all observable and unobservable quantities can be factored into two components:

$$P(x, y) = P(y|x)P(x). \tag{1.4}$$

The first factor is represented by the data distribution referred also as observation model. It is the distribution of the measurements, given the unobservable state.

$$P(y|x) \quad \longrightarrow \quad \textit{Data likelihood.} \tag{1.5}$$

If y corresponds to imperfect measurements of a chemical component, and x the true concentration of that species, the data likelihood quantifies the distribution of observation error in measuring concentration, reflecting possible biases as well as instrument errors. The second factor is the prior distribution that quantifies our a priori knowledge on the unobservable quantities of interest. For example, if x represents emissions, then one may base this prior distribution on historical information.

$$P(x) \quad \longrightarrow \quad Prior{=}Forecast \tag{1.6}$$

The posterior is the update of our prior knowledge about the state given the actual data:

$$P(x|y) \quad \longrightarrow \quad Posterior{=}Analysis \tag{1.7}$$

By applying Bayes' rule, we obtain the posterior distribution:

$$P(x|y) \propto P(y|x) P(x) \tag{1.8}$$

The joint state-space formulation is an application of the Bayesian update problem illustrated by the Figure 1.4. The modeled system advances in time until an analysis time; these integrations are represented by the green lines. When the distribution of the model state before the update, called the forecast distribution (formula 1.6), and the data likelihood are combined, the analysis distribution (formula 1.7) is provided (red lines). The process of applying an observational operator H to each sample of the prior state estimate and calculating the corresponding increments is illustrated by the blue lines. With this new model state, the system is propagated until the next assimilation time step (green lines).

1.4.2 Filtering and Smoothing

To design the best possible assimilation system, it is necessary to clearly define the goals of data assimilation (such as forecast initialization, monitoring the present situation, reanalysis), the physical characteristics of the processes involved, the properties of the observation network, and the limitations of the assimilation methods. Depending on the application it may be necessary to make forecasts or include information that is taken after the time of the estimate. If the estimate of the state at current time k is needed based on the measurements until time l, there are three different cases:

1. forecasting for $k > l$
2. filtering for $k = l$

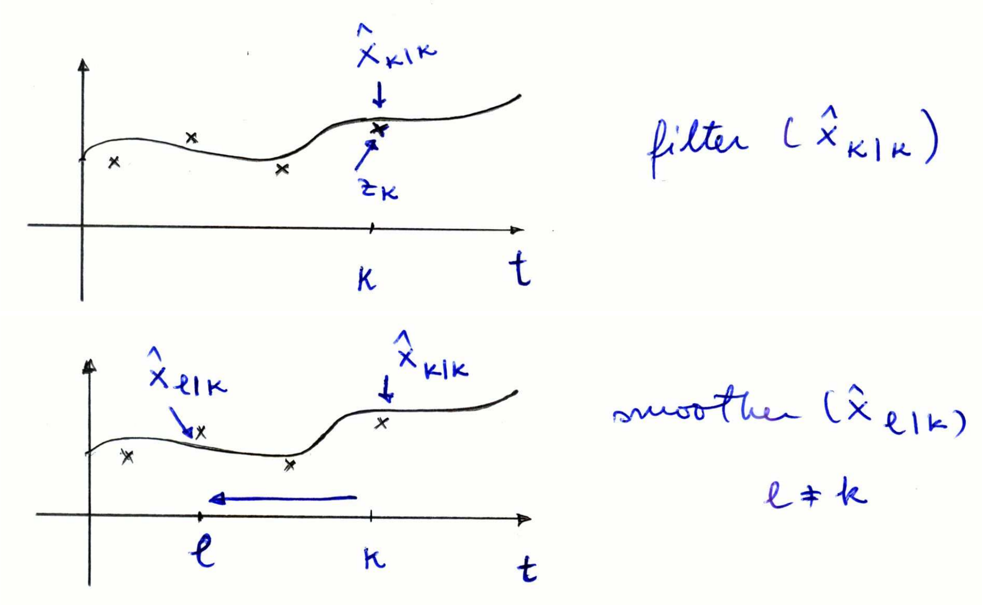

Figure 1.5: Schematic representation of the filtering (upper panel) and smoother procedures (lower panel).

3. smoothing for $k < l$

Given a stochastic model for dynamics and observations, the filter is able to compute the optimal estimate of the current state when all data from the past are available. The forecasting problem is related to the filtering in the sense that the filtering estimate is used as initial condition for the model to determine the forecast. In both techniques future measurements are not taken into account. For offline applications such as estimating time varying emissions, not including data after the analysis time is a disadvantage of the filter. In contrast to filtering, the smoother analysis results from retrospective assimilation of all observed data, both past and future measurements being incorporated into analysis. For an extended description of filter and smoother in terms of Bayesian statistics, [43] is given as reference.

The difference between filter and smoother is schematically illustrated in Figure 1.5.

Because more observations are used in producing a smoothed estimate than in producing an estimate at the current time, one expect the smoothed estimate to be more accurate and, at the same time, an increased complexity of the smoother techniques. To keep this procedure computationally feasible, the smoothed estimates may be obtained for a smaller part of the state vector only and for a limited set of retrospective data. The smoother can be derived by augmenting the state vector with the past values of the state. Then the new augmented model is used in combination with an filter in order to produce the smoothed estimates of the state.

For a linear model the 4D VAR method and smoother have been proved to be equivalent.

1.4.3 Kalman-like filtering

The classical Kalman filter provides an ideal framework for solving the sequential updating problem with linear model operators and Gaussian error distributions [68]. It can be derived from different perspectives: it can be identified as a recursive least squares problem or it can be designed to provide optimal solution to the Bayesian formulation ([66]) seeking either a maximum likelihood estimate, e.g. [80] or a minimum variance estimate of the state based on all observations available, e.g. [22]. The filter consists of two steps: forecast and analysis for both mean and covariance of a state estimate probability density function (PDF). These two moments fully characterize a Gaussian PDF.

Extended Kalman filter

The use of the standard Kalman filter in Geosciences is hampered by a number of factors. Firstly, due to the manipulation of the covariance matrix of the state, the cost of applying the Kalman filter to a system with large number of degrees of freedom becomes intractable [23]. For an atmospheric chemical model the dimension of the state can be more than 10^5. In order to derive a computationally feasible filter, simplifications have to be introduced. Secondly, the linearity and Gaussianity assumptions are two strong constraints imposed to a geophysical system. The forecast and analysis distributions cannot be obtained explicitly for non-Gaussian models and/ or nonlinear systems. Therefore variants were developed for nonlinear problems, missing observations, nonlinear updates and non-Gaussian distributions. An approach to handling nonlinear observations and evolution models is the Extended Kalman filter (EKF) that uses the full nonlinear model to propagate the state estimate, namely the PDF's mean, but uses a local linearization of the model to propagate the state's uncertainty, that is the PDF's covariance. The EKF has been largely studied in geophysical contexts, e.g. [49], [86]. The linearization of the error covariance evolution is often inadequate due to unbounded growth of the computed error variance. [85] showed poor performance of the EKF in application to high nonlinear systems as the Lorenz model when the data are inaccurate or sparse. In addition, to make the EKF computationally feasible for large scale applications, many reduced-state or low-rank filters have been proposed. Examples include the singular evolutive extended Kalman filter (SEEK) [89], the balanced truncation Kalman filter as has been proposed by [45].

Ensemble-based filters

Another class of low-rank filters rely on Monte Carlo integration of the Fokker-Plank-Kolmogorov equation governing the evolution of the PDF that describe the forecast statistics. The basic approach uses the Monte Carlo samples called ensembles or particles to approximate the forecast distribution with the full nonlinear forward model. One of the main advantages of this approach is that the tangent linear model is not required. General formulas for calculating the optimal nonlinear filter and smoother can be found in [105]. The performance of these algorithms depends to large extend on the number of replicates relative to the size of the state vector. Given the enormous number of state variables in a geophysical system, feasible implementation of the particle-based filters is still a problem. Therefore other approaches that rely on the Gaussian assumption of the PDF have been considered.

The Ensemble Kalman filter (EnKF) originally developed by [40] and [61] is one of such methods. Each particle in an ensemble is updated using the traditional Kalman gain calculated from the mean and the covariances of the prior ensemble. If the the data are randomly perturbed, the analysis ensemble is shown to have the proper statistics [17].

Another examples is given by the reduced-rank square root filter (RRSQRT) that can also be viewed as an ensemble methods [110]. In the RRSQRT filter formulation, the covariance matrix is expressed in a number of (orthogonal) ensembles or modes which are re-orthogonalized and truncated to a fixed number during each time step.

1.4.4 Sources of errors in ensemble-based data assimilation

The real settings posed a set of challenges related to each issue employed by data assimilation scheme, namely model, measurements and assimilation algorithms. Errors can be introduced at each level of the DA procedure. The application of DA to a CTM presents several challenges: filter divergence due to limited ensemble size, uncertainty in the model parameterization, additional constraints imposed by non-zero concentrations, unrealistic correlations, assimilation of strong related components, introduction of additional errors due to sampling procedures. However, several potential problems for the EnKF are worth to be mentioned.

Gaussian assumption and linear updating

Firstly, as all Kalman filtering schemes, it uses only the Gaussian part of the prior PDF and the updated ensemble preserves only two first moments of the posterior. Then the method can lead to unreliable or biased simulations when the statistics of model variable are strongly non-Gaussian. In addition, since the procedure assumes a linear relation between a state and data in calculating the Kalman gain, it is not suitable for cases in which the linearization of that relation is invalid. Particle filtering is a tool for solving these problems. This filter operates on a set of particles and their probabilities and not on the mean and covariance statistics and in the analysis step the correction is achieved by changing the weight associated to each particle.

Use of localization

Secondly, experience has shown that for a small number of ensembles, spurious long-distance correlations arise in the use of the ensemble-based methods due to the sub-sampling of the probability distribution. Thus, compensations are often employed. One of such compensation largely used in atmospheric data assimilation is the localization procedure. In the analysis step of algorithms localization reduces the impact of an observation on a state variable by a factor that is a function of a distance between the two. Different localization schemes have been proposed in order to avoid these spurious correlations. [62] have investigated the use of an influence cutoff radius that removes the impact of remote observations. The use of compactly supported correlation functions (see [47]) aims to provides a smoothed correction and monotonically decreasing with distance.

Use of covariance inflation

Several scenarios considered in experimental settings may lead to filter divergence and a decreasing ability towards the end of the assimilation window. A pragmatic method to prevent the collapse of the filter is covariance inflation [1] where the spread of the ensemble is artificially enlarged to make the filter more robust against model errors. Several procedures including additive inflation, multiplicative inflation and model-specific inflation (obtained through perturbing model parameters as emissions) have investigated in the atmospheric chemistry context by [25].

Sampling errors

Thirdly, although it was shown that the EnKF results in an unbiased estimate of the error covariance for large ensemble sizes, this procedure also introduces additional sampling errors. To eliminate these sampling errors a number of deterministic filters that do not require random number realizations in the analysis step have been proposed. Ensemble Square Root Filters (ESRFs) are all deterministic filters that achieve the proper EKF analysis error covariance statistics by updating the ensemble mean and then linearly transforming the ensemble members with the use of rotations, translations and rescaling in various directions.

Model errors and bias

In addition, it has been recognized that the lack of complete information about the statistics of model (and observation) errors may impact significantly the assimilation. The adjoint method that provide a tool to tune the model to available data involve an assumption of perfect model structure. This assumption is too restrictive for atmospheric applications. Sequential procedures require quantification of the uncertainty both in the observations and background state. There are two main factors that create the background uncertainty: inaccurate initial conditions from the previous analysis and deficiencies in the model that may play a major role in forecasting activities, e.g. [87]. A central problem is the discrepancy between the model dynamics and the actual dynamics that is generally termed as model error. Several solutions have been proposed to estimate and/or account for the model errors such as bias by [28], white noise process by [57] and first-order Markov process by [116].

A fundamental assumption of the standard Kalman filter is that the model and observations are unbiased. Bias in observation reflects the instrument inaccuracies, representativeness errors or, for remote sensing observations, errors in the retrieval algorithm. After quality control, the measurements are supposed to be largely corrected. By contrast, model forecast are hardly ever unbiased due to the inaccurate physical parameterization, discretization, erroneous boundary conditions, forcing errors, etc. Hence, the forecast error may contain a random and a systematic component.

A typical characteristics of Kalman filter-based techniques is that the forecast skills shows that the model drift back to a biased state after the analysis is performed. This suggests that state updating alone is not an adequate solution to improve model results persistently. For off-line bias correction scheme, typically a time function of the bias is estimated in advance from the model and observation or analysis climatology.

Different approaches for on-line estimation of forecast bias in the presence of bias-free observing system have been proposed. A common practice was to augment the original model state with several uncertain parameters that are designed as bias terms, or more general, as model error terms [66]. [46] gives the theoretical framework for treating separate-bias problem. The estimation of the bias is decoupled from the computation of bias-blind estimate of the state.

The forecast bias represents the expectation of the forecast error that is defined as the difference between the true state of the system and the forecasted state estimate. In general it is difficult to derive a bias evolution model since the forecast bias is dependent on the state and parameters of the atmospheric model and correlated in time.

In spite of these challenges, ensemble-based methods have attractive features that can be exploit for successful DA with atmospheric chemistry models.

1.4.5 Measuring data assimilation performance

An important aspect in data assimilation is quantifying the performance. Various methods are employed to asses the capability of a data assimilation system. There are two main direction of measuring the performance of a specific method or comparing different data assimilation techniques. One approach is the so called twin experiment where the data used in the assimilation is produced with a model by perturbing several parameters, initial conditions or the forcing. This approach has been used in this thesis for comparing between different types of filter and smoother algorithms. The advantage of performing a twin experiment is that the the estimate provided by the DA scheme can be compared with the true values of the model variables. A drawback is that it is difficult to estimate the performance of the algorithm in the context of real settings. Therefore a second method of verifying the performance is to incorporate real data into the model. Additional validation schemes should be used since the performance of the entire DA system is a result of different causes. In this thesis it is shown that the performance of assimilation may dramatically depends on how the parameters of the model and information to be integrated in the system are chosen.

1.5 Motivation and overview

It has been proved that sequential data assimilation based on low-rank Kalman filter approach is suitable for air pollution problems. Applications to the tropospheric ozone have been developed and used successfully with the LOTOS model by [99] as well as with the EUROS model by [53]. This thesis extends the previous work by investigating the application of ensemble-based data assimilation methods to the unified atmospheric chemistry model LOTOS-EUROS. It describes the potential and benefits of ensemble filtering and smoothing techniques for atmospheric chemical data assimilation applied in both ideal and real settings. Central in all projects is the use of different type of data for assimilation purpose. This study has been carried out with focus on four different topics that can be formulated as the following scientific questions:

- How can we use the retrospective data assimilation with an atmospheric chemistry model?
- How accurate can the sulphur dioxide and sulphate concentrations be estimated using a single component setup compared to the combined assimilation procedure?
- How can the OMI satellite product contribute to a better understanding of LOTOS-EUROS capabilities in predicting the tropospheric ozone?
- Can we select an optimal algorithm from the data assimilation perspective?

Following the structure imposed by these topics, the content of this thesis can be split into four parts, each of them corresponding to a paper. The first part described in Chapter 2 concerns the concepts of filtering and smoothing, with the emphasis on the comparison of different schemes from accuracy and efficiency points of view. It describes the use of retrospective data to estimate the concentrations and emissions in a twin experiment.

The next part of this thesis described in Chapter 3 aims to answer to the second question by investigating the application of an ensemble Kalman filter procedure to the LOTOS-EUROS model. The focus is on an important issues related to the use of different sets of data. It describes the process of going from a single to a multi-component data assimilation for two strongly dependent species, sulphur dioxide and sulphate in an experiment conducted over whole Europe for one year period of simulation. In addition, a stochastic environment is built around the model to provide insight of the model parameters such as emissions and reaction rate. In the end the filtering and smoothing procedures should provide an accurate and comprehensive analysis of the atmospheric field and parameters.

In Chapter 4 the forecast bias estimation problem is treated within an application focused on the tropospheric nitrogen dioxide. The model simulations and OMI satellite information are compared and combined in a bias-based data assimilation experiment. Model bias is due to the presence of random and persistent error in the model forecast caused by incorrect physical parameterization and limited chemistry scheme. Therefore, the identification and correction of forecast model bias and sources of uncertainty are important components of data assimilation system that may improve the model capabilities in prediction of pollution.

Chapter 5 discusses the last question. It starts with a theoretical background of low-rank filters analyzing their capabilities. This is followed by an improved method for a specific class of low-rank algorithms which addresses a crucial problem for data assimilation, namely the model error.

Finally the summary, conclusions and outlook of this thesis are presented in the Chapter 6.

CHAPTER 2

Ensemble filter and smoother

Abstract: Large uncertainties in emissions, formation routes and sinks of aerosol particles cause that model performances for particulate matter and its components are relatively poor. Presently, our focus is on the component for which enough information is available: sulphate.

In this chapter data assimilation schemes based on ensemble filtering and ensemble smoothing techniques are used to combine the results of a simplified chemistry transport model with measurement information to estimate emission parameters. For this purpose a number of smoothing algorithms: the ensemble Kalman smoother, a fixed-lag ensemble smoother and the smoothing implementation as proposed by [90] have been applied. The problem of filter and smoother divergence is also discussed in this chapter.

2.1 Introduction

Particulate matter (PM) in ambient air has been associated consistently with excess mortality and morbidity in human population (e.g. [2], [8]). Atmospheric particulate matter is a complex mixture of anthropogenic and natural airborne particles. The main components are secondary inorganic aerosols (sulphate, nitrate and ammonium), combustion particles (OC and EC), sea salt and earth crust materials. The various components of the particulate matter in the atmosphere also have climate-forcing impacts, either contributing to or offsetting the warming effects of greenhouse gases e.g. [12], [60]. Furthermore, secondary inorganic aerosol formation and transport has been studied for decades as they contribute to acidification of soils. Hence, for all these

---

This chapter is a slightly revised version of [3].

issues a thorough knowledge of the concentrations of particles as well as their sources and sinks is urgently needed.

During the last decades, models were developed to describe the fate of (particulate) pollutants over Europe. However, large uncertainties in emissions, formation routes and sinks of particles cause model performances for PM and its components to be relatively poor [108]. The use of these models to assess the state of the atmosphere can be strengthened using data assimilation. Data assimilation schemes combine the results of a numerical atmospheric chemistry model with measurement information to obtain an optimal reconstruction of the dynamic behavior of the system. Such methods have been applied for ozone [108], [32], [52] but not for components of the particulate matter.

We are developing such a data assimilation system to be used, in the end, to estimate parameters such as conversion rates and emission strengths for PM and its precursors. Application of data assimilation to the components of particulate matter is presently hampered by a number of factors. Except for sulphate, observations are sparse and often very uncertain [97]). Consequently, a first attempt should be directed at sulphate. Furthermore, the problems at hand are associated with long time scales and therefore, modeling studies have a time window of one or more years. A data assimilation experiment for these time windows using a large chemistry transport model is a formidable task. Different data assimilation schemes have been used in air pollution applications. Ensemble-based assimilation is easy to implement, suitable for real-time estimation of concentrations and allows a very general statistical description. The ensemble Kalman filter, one of the Monte Carlo sequential methods only data prior to the time of analysis are used and cannot reconstruct emissions or concentrations at previous time. For an accurate estimation of emissions ensemble-based smoothers are required. The general behavior and the efficiency of three implementations of the ensemble smoother approach in a twin experiment has been studied.

The chapter is organized as follows. The Ensemble Kalman filter together with an implementation of this algorithm are given in section 2.2. The ensemble smoother techniques are discussed in section 2.3, as well as the implementation of three algorithms, followed in section 3 by the description of the twin experiment. The model is described in the next section. The results of data assimilation calculations are presented and discussed in section 5, i.e., the general behavior of a smoother assimilation, the performance of the algorithms and their efficiency. The last section summarizes the concluding remarks.

2.2 Ensemble-based filters

A sequential data assimilation procedure has been developed based on the Kalman filter (KF) technique. In general, KF computes probability density functions (pdf's) of the true state given:

1. a transition model to propagate the state in time, and
2. observations with associated representation error

For application of the filter algorithms to a dynamical model, a stochastic representation should be written in a state-space form according to:

$$x(k+1) \quad = \quad M(x(k)) + w(k), \quad w \sim N(0, Q). \tag{2.1}$$

The state-space operator M describes the time evolution from the time k to k+1 of the state vector x. The random forcing term w is drawn from the normal distribution $N(0, Q)$ with Q the covariance matrix. The state of the observational network is defined by the observation operator H that maps state variables x to observations y. We further assume that the measurements have white Gaussian errors v with covariance denoted by R:

$$y(k) = H(x(k)) + v(k), \quad v \sim N(0, R). \tag{2.2}$$

If the initial pdf, the model uncertainty, and the representation error are expressed as normal random vectors, and if, in addition the transition model is linear, the KF provides the complete solution of the estimation problem. In practice, this algorithm cannot be implemented for large scale applications. Since the size of the state vector is usually very large (at least 10^4 elements), the storage and propagation of the covariance matrix become impossible. Another reason is that the the CTMs are nonlinear, for example because chemical reactions are included.

2.2.1 The Ensemble Kalman Filter (EnKF)

A common used alternative for the KF is the Ensemble Kalman Filter (EnKF). this algorithm has been successfully used for variety of many applications, including air pollution models. As a stochastic method, the EnKF is based on the representation of the probability density of the state estimate in an ensemble of N states, $\xi_1, \xi_2, \ldots \xi_N$. Each ensemble member is assumed to be a single sample out of a distribution of the true state. Whenever necessary, statistical moments are approximated with sample statistics. In the first step of the algorithm an ensemble of N states $\xi^a(0)$ is generated to represent the uncertainty in $x(0)$. In the second step, the *forecast*, the stochastic model propagates the distribution of the true state from the time k to $k+1$:

$$\begin{aligned}\xi_j^f(k+1) &= M(\xi_j^a(k)) + w_j(k), && (2.3)\\ x^f &= \frac{1}{N}\sum_{j=1}^{N}\xi_j^f. && (2.4)\end{aligned}$$

The ensemble covariance matrix of forecast errors P^f is assumed to be carried at time k by the ensemble of perturbations denoted by L^f.

$$\begin{aligned}L^f &= \left[\xi_1^f - x^f, \xi_2^f - x^f, \ldots, \xi_N^f - x^f\right], && (2.5)\\ P^f &= \frac{1}{N-1}L^f\left(L^f\right)^T. && (2.6)\end{aligned}$$

When the measurements become available, the mean and the covariance are replaced with equivalent ones in the *analysis step* using the ensemble Kalman gain:

$$K = P^f H^T\left(HP^fH^T + R\right)^{-1}, \tag{2.7}$$

The stochastic analysis step defines the standard EnKF with perturbed observations. The analysis ensemble involves the update of each ensemble members the and their ensemble covariance matrix.

$$\begin{aligned}\xi_j^a &= \xi_j^f + K\left[y - H\xi_j^f + v_j\right], && (2.8)\\ P^a &= (I - KH)\,P^f, && (2.9)\end{aligned}$$

where the ensemble of state vectors is generated by the realizations w_j and v_j of the white noise processes w and d, respectively.

The advantages of this algorithm are that P^f is positive definite and that the tangent linear model is not required anymore because the ensembles are propagated through the model using the original operator. Also, in the final implementation of the algorithm, $P^f(k)$ need not to be computed. As a result, the computational effort required for the EnKF is approximately N model simulations. The errors in the state are of the statistical nature and decrease slowly with the number of replicates.

2.2.2 The implementation of the EnKF

In this section we briefly review the practical implementation of the EnKF. More details can be found in [41]. The matrix A holding the ensemble members

$\xi_i \in \mathcal{R}^n$ is defined by:
$A = (\xi_1, \xi_2, \ldots, \xi_N) \in \mathcal{R}^{n\times N}$, N being the number of the ensemble members and n is the size of the state vector. Then the ensemble perturbation matrix is $L = A - \overline{A} = A\,(I - \mathbf{1}_N)$, where the ensemble mean is stored in each column of $\overline{A} = A\mathbf{1}_N$.

The EnKF uses an ensemble of forecasts to estimate background-error covariances. [61] showed that in order to maintain sufficient spread in the ensemble and prevent filter divergence, the observations should be treated as random variables. The concept of using perturbed sets of observations to update each ensemble member is introduced. The perturbed observations consist of the actual observations and random noise.

We consider the vector of measurements $d \in \mathcal{R}^m$, where m is the number of the measurements and define N vectors of perturbed observations as $v_j = v + \epsilon_j$ for every $j = 1, \ldots, N$, which can be stored in the columns of a matrix: $V = (v_1, v_2, \ldots, v_N) \in \mathcal{R}^{m\times N}$, while the ensemble of perturbations is stored in the matrix $\Upsilon = (\epsilon_1, \epsilon_2, \ldots, \epsilon_N) \in \mathcal{R}^{m\times N}$.
Now we are able to construct the ensemble representation of the observations error covariance matrix:

$$R_{ens} = \frac{\Upsilon\Upsilon^T}{N-1}.$$

The filter analysis at current time l in terms of the ensemble covariance matrix P^f is formulated as:

$$\begin{aligned} A_l^a &= A_l^f + P_l^f H^T \left(H P_l^f H^T + R_{ens}\right)^{-1} (V - HA_l)\,, \\ A_l^a &= A_l^f + L_l L_l^T H^T \left(H L_l^T H^T + \Upsilon\Upsilon^T\right)^{-1} (V - HL_l)\,. \end{aligned} \tag{2.10}$$

It is assumed that the ensemble perturbations and observation errors are uncorrelated, i.e $HL\Upsilon \equiv 0$. Then the following decomposition is hold:
$HLL^TH^T + \Upsilon\Upsilon^T = (HL + \Upsilon)\,(HL + \Upsilon)^T$, and the problem can be reformulated by using several new matrices X^1, X^2, X^3 defined as following:

$$\begin{aligned} A^a &= A^f + LL^TH^T((HL+\Upsilon)\,(HL+\Upsilon)^T)^{-1}(V - HA), \\ A^a &= A^f + L\,(HL)^T X^1, \\ A^a &= A^f + LX^2, \\ A^a &= A^f + A^f\,(\mathbf{I} - \mathbf{1}_N)\,X^2, \\ A^a &= A^f\left(\mathbf{I} + X^2\right), \\ A^a &= A^f X^3, \end{aligned} \tag{2.11}$$

where we have used $\mathbf{1}_N X^2 \equiv 0$. This final form of the analysis ensemble is obtained by transforming the predicted ensemble with the matrix X^3 in a nonlinear way.

2.3 Smoother approach

Given a stochastic model for dynamics and observations, the Kalman filter is able to compute the optimal estimate of the current state when all data from the past are available, but future measurements are not taken into account. For offline applications such as estimating time varying emissions, not considering data after the analysis time is the disadvantage of the filter. In contrast to filtering, the smoother analysis results from retrospective assimilation of all observed data, both past and future measurement being incorporated into analysis.

It is useful to notice that smoothing is essential to oceanographic, meteorological and hydrological investigations. Many applications use data from the past. Such retrospective data analysis can be formulated in terms of smoothing techniques. There are three different classes: fixed point smoothing, which requires estimates of system state at only a single time instance, fixed interval smoothing which requires estimates at multiples times distributed throughout an interval and fixed lag smoothing which requires estimates in a lag window W prior to the most recent measurement. Of the three types of smoothers, fixed lag algorithm is most appropriate for the reconstruction of emissions in our application since only the most recent emissions significantly impact the concentration field. The fixed lag smoother can be derived in the context of ensemble-based filter by augmenting the state vector with the past values of the state. Then the new augmented model is used in combination with an ensemble filter in order to produce the smoothed estimates. It is worth to be mentioned that at the end of the interval the smoother estimate is identical to that produced by the filter, given the same observation network and the same initial statistics of the state.

For the state vector, we determine recursive equations for the estimate for all k and some fixed lag W:

$$x_{k-W|k} = E\left[x_{k-W}|y(0), y(1), \ldots, y(k)\right]$$

and the associated error covariance. By using the augmentation techniques the state vector, model and the observation operator become, respectively:

$$X(k) = \left[x(k), x(k-1), \ldots, x(k-W)\right], \tag{2.12}$$

$$X(k) = \begin{bmatrix} M(k-1) & 0 & \ldots & 0 & 0 \\ I & 0 & \ldots & 0 & 0 \\ 0 & I & \ldots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \ldots & I & 0 \end{bmatrix} X(k-1) + \begin{bmatrix} G(k) \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} w(k), \tag{2.13}$$

$$y(k) = \begin{bmatrix} H(k) & 0 & \dots & 0 & 0 \end{bmatrix} \begin{bmatrix} x(k) \\ x(k-1) \\ \vdots \\ x(k-W) \end{bmatrix} + v(k).$$

In other words, the one-step prediction estimate of the entire state of the augmented state vector Equation 2.12 contains the smoothed estimates of the state for lag length up to W. The Kalman filter applied to this augmented model (Equation 2.13) lead to the equations involving a state estimates, augmented Kalman gain matrix and augmented covariance matrix, respectively.

2.3.1 Ensemble Kalman smoother (EnKS)

The Ensemble Kalman Smoother (EnKS), described in [43], updates the ensemble at prior times every time when new measurements are available. The updates exploit the space-time correlations between the model forecast at measurement locations and the model state at a prior time. Thus, every time a new set of measurements becomes available the ensemble at the current and all prior times can be updated. For convenience, we write the formula only for one new available observation from the future data. The EnKS analysis for the prior time l and $k > l$ can be found analogously to the analysis given by Equation 2.10:

$$A_l^a = A_l^f + L_l L_k^T H^T \left(H L_k L_k^T H^T + \Upsilon \Upsilon^T \right)^{-1} (V - H A_k).$$

Considering the definition of X_3 obtained from Equation 2.11 the matrix of coefficients at time k is used on the updated ensemble at the prior time l. Therefore, we compute the smoothed estimate at the time l in the past using data from the future. In the fixed lag smoothing approach the state at time k is updated with observations in a fixed time window $(k, k+W]$, where W is the lag length. Equivalently, the fixed lag smoothing updates the state in $[k-W, k)$ if the observation is available at time k. The equation produces the following formula as in the EnKF analysis:

$$A_k^a = A_k^f \prod_{j=k+1}^{k+W} X_j^3. \tag{2.14}$$

2.3.2 The FIFO-lag algorithm

The faster fixed lag algorithm developed by [90] is computational improvement to the previous ensemble smoothing technique. This algorithm is called

FIFO because the smoother is implemented via the first-in-first-out queue. The idea is that the new information is added to the front of the queue and old information is removed from the back of the queue. Introducing a new matrix $X^4 = \prod_{j=k+1}^{k+W} X_j^3$, from Equation 2.14 is derived:

$$A_k^a = A_k^f X_k^4. \tag{2.15}$$

In the fixed lag smoother X_k^4 and A^a do not need to be computed separately. To define a forward recursion, X^3 is initialized to the identity at all unobserved model steps. The matrix X^4 is initialized at $k = 0$ as in the following formula:

$$X_0^4 = \prod_{j=1}^{W} X_j^3. \tag{2.16}$$

The following recursion defines fixed-lag smoothing and is done on a forward pass:

$$X_k^4 = \prod_{j=k+1}^{k+W} X_j^3 = \left(X_k^3\right)^{-1} X_{k-1}^4 X_{k+W}^3. \tag{2.17}$$

2.4 Experimental setup

Our data assimilation experiments use a reduced chemical transport model. In reality, sulphur dioxide is emitted and transported away from the source. During transport process sulphur dioxide SO_2 is converted into sulphate SO_4. In our model we consider the 2D advection diffusion equation for the transport of SO_2 and SO_4:

$$\frac{\partial c}{\partial t} + u\frac{\partial c}{\partial x} + v\frac{\partial c}{\partial y} = \nu\frac{\partial^2 c}{\partial x^2} + \nu\frac{\partial^2 c}{\partial y^2} + S,$$

with the square domain $[0, D] \times [0, D]$ and zero initial conditions. Here, c is the concentration, $[u, v]$ is the velocity field, ν represents the dispersion coefficient, and S is the source term. A backward Lagrangian scheme is used to discretize this equation on the 30×30 grid. For all experiments the velocity field (see Figure 2.1) is assumed to be known and constant in time. During transport two pathways convert SO_2 into sulphate SO_4: gas phase reaction with the OH radical and heterogeneous reactions in clouds and fogs. The gas phase reaction is taken from the standard CBM-IV chemistry mechanism [114]. The reaction rates involved in the heterogeneous routes are very uncertain. Here, we approximate these complicated routes by a linear first order reaction. This

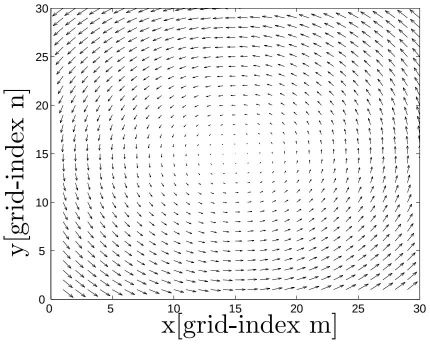

Figure 2.1: The velocity field

approach is an option in the LOTOS-EUROS model [95]. Many more authors, e.g., [103] have used this approach to account for missing reaction pathways and/or a lack of modeled oxidant levels to oxidize the SO_2 [70]. Other models, e.g. [2] use explicit cloud chemistry but neglect the pH-dependency, which also yields a linear system of reactions. These approaches are commonly used in Europe. Moreover, [44], have shown the linear behavior of the SO_2-SO_4 system in Europe after the mid-nineties. This linear behavior has also been observed for the European Regional CTMs for long term use and policy support in the EURODELTA study [94].

We consider the following chemistry which describes the conversion of SO_2 to SO_4:

$$SO_2 + OH \xrightarrow{k_1} SO_4,$$

$$SO_2 \xrightarrow{k_2} SO_4,$$

where the rates $k_1 = 1.5[\frac{1}{ppbmin}]$ and $k_2 = 8.3e-5[\frac{1}{min}]$ are considered to be constant. The second reaction is used to represent cloud chemistry and other heterogeneous oxidation pathways. A loss term $k_3 = 0.1[\frac{1}{hours}]$ is considered to represent a constant deposition rate for SO_4 component. The change in concentration fields can be described as follows:

$$\begin{aligned} \frac{\partial SO_2}{\partial t} &= -(k_1 OH + k_2) SO_2, \\ \frac{\partial SO_4}{\partial t} &= (k_1 OH + k_2) SO_2 - k_3. \end{aligned} \tag{2.18}$$

OH concentrations were taken from the LOTOS-EUROS model [16] for a warm and cold summer day. The unit for concentration is ppb. Physical meaning for one time step is 24 hours. The OH concentrations show a daytime maximum and are very small during the night. We use the splitting operator to separate

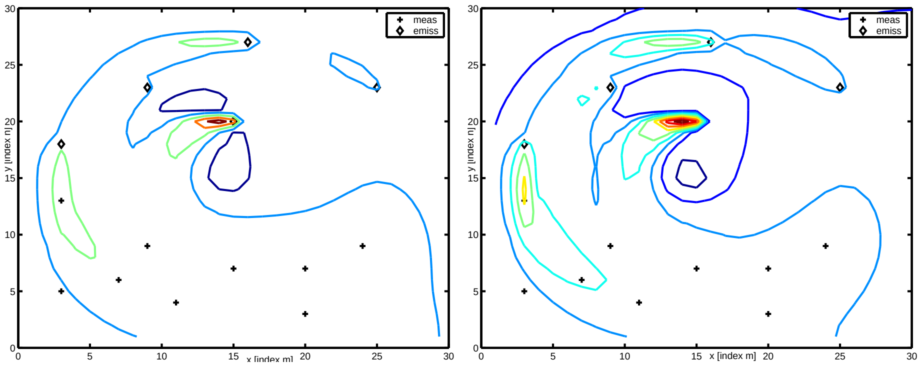

Figure 2.2: True and modeled sulphate concentrations after $k = 100$ time steps using 80 ensembles. The diamonds represent the emission points and the crosses indicate the measurements.

the processes, chemistry and transport, and to solve them separately, but in the same time step. In our model these two processes are supposed to be uncontaminated by random noise.

2.5 Assimilation experiments

The focus in the experiments is on the improvement of the accuracy of the simulated sulphur concentrations and emissions using the smoothing procedure and on the improvement of the efficiency of the smoothing algorithms. In this section several aspects of the smoother assimilation are discussed. To compare the algorithms in twin experiments, a reference simulation was designed first. The reference solution was generated by inserting constant SO_2 emissions at five grid cells. The increase of concentration per time step for these location was $\{0.2, 0.5, 0.4, 0.2, 0.25\}$, respectively. Observations are generated in nine locations of the domain, see Figure 2.2. The final product is SO_4. We use only sulphate measurements for our data assimilation experiments. Therefore, the observations were generated from simulated true concentrations of SO_4 as:

$$y(k) = H(k)c_{SO_4}(k) + v(k),$$

where H is the observations operator and v is the observational noise process $v \sim N(0, R)$. The emissions are treated as the model input by defining our model according to:

$$\begin{bmatrix} c(k+1) \\ e(k+1) \end{bmatrix} = \begin{bmatrix} M_c & M_e \\ 0 & I \end{bmatrix} \begin{bmatrix} c(k) \\ e(k) \end{bmatrix} + \begin{bmatrix} 0 \\ I \end{bmatrix} w(k),$$

where c contains the sulphur and sulphate concentrations, e represents the SO_2 emission vector and w is white noise $w \sim N(0, Q)$. The augmented state vector $x = [c_{SO_2}, c_{SO_4}, e_{SO_2}]$ consists of 1805 compounds: 30×30 concentrations of SO_2, 30×30 concentrations of SO_4, and 5 points of pollutant emissions. The representation of the state vector becomes:

$$x(k+1) \quad = \quad M(k)x(k) + w(k).$$

A number of simulations has been performed using different smoothing implementations. The assimilation results depend on the model and parameters involved in the processes, e.g., the number of observations, the accuracy of the algorithms and the noise specification. For evaluating the smoothing algorithms we compute the root-mean square (rms) error:

$$\text{rms} = \sqrt{\frac{1}{\alpha^2 T} \sum_{m,n,k} \left[c_{m,n}(k) - \hat{c}_{m,n}(k)\right]^2},$$

where $c_{m,n}(k)$ are the true generated concentrations, $\hat{c}_{m,n}(k)$ are the computed estimates, α is the number of grid points in one direction, and T is the time steps number.

2.5.1 Divergence problem

A divergence problem might occur when the covariance matrix of the error covariance matrix is too small [66]. In this case the filter gain will be small too and as a result, the observations will have little impact on the state estimate. The difference between an actual observation and its predicted value is called the innovation. In our application, the divergence of the filter is illustrated by the comparison of the realizations of the innovations with their theoretical statistics.

Experiment 1: Deterministic emissions in the truth and filter model.
In the first experiment, the simulated reality was considered to be deterministic, i.e. the emissions are treated as constants. The process noise is assumed to be zero in the filter model therefore, the Kalman gain converges to zero. Studying the realizations of the innovations, we can detect filter divergence. The deterministic case is illustrated in Figures 2.3. It is shown that the 'deterministic' filter cannot estimate the reference solution properly. In addition, the standard deviation is very small; the filter solution is convergent, but to the wrong estimate. The results illustrated in Figure 2.3 and Figure 2.4 are obtained by using 30 and 80 ensemble replicates, respectively. By increasing the ensemble size the estimates are only improved in the beginning of assimilation window.

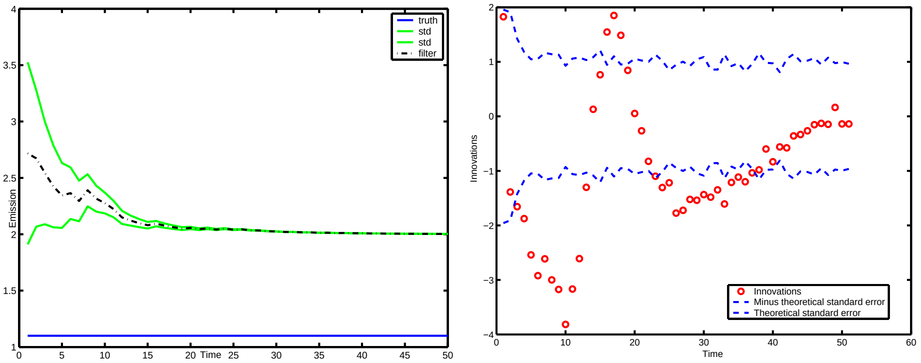

Figure 2.3: Results of the deterministic case in experiment 1 using 30 ensemble members; (top) The time series evolution of the averaged emission; the reference solution (continuous line), the filter estimate (dash-dot line) and the estimated standard deviations (full gray lines). (bottom) The innovations of the filter (circles) and theoretical standard error (dot line).

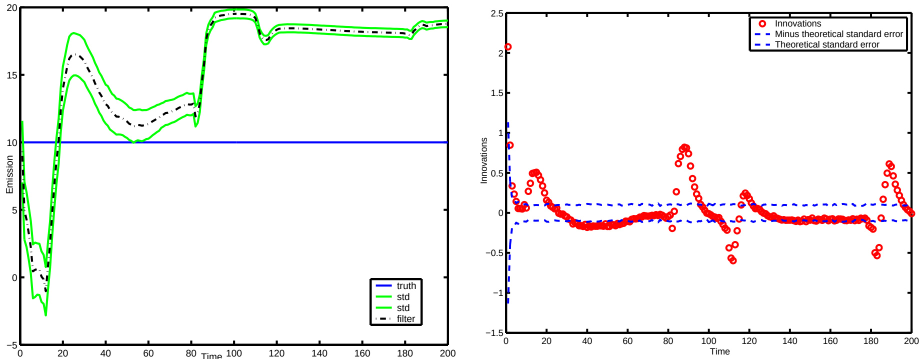

Figure 2.4: Results of the deterministic case using 80 ensembles members; (top) The time series evolution of the averaged emission; the reference solution (continuous line), the filter estimate (dash-dot line) and the estimated standard deviations (full gray lines). (bottom) The innovations of the filter (circles) and theoretical standard error (dashed line).

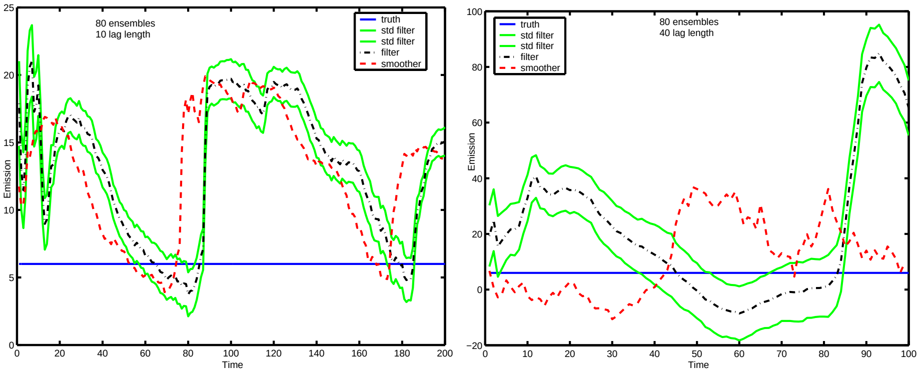

Figure 2.5: Results of the stochastic filter model in experiment 2; (top) Averaged emission using 80 replicates and 10, 40 lag length, respectively; the reference solution (continuous line), the filter estimate (dash-dot line) and the estimated standard deviations (full gray lines); (bottom) The innovations of the filter (circles) and theoretical standard error (dashed line).

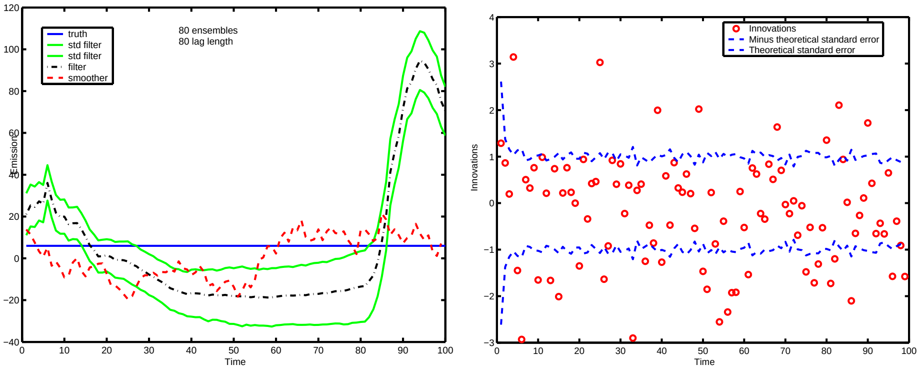

Figure 2.6: Results of the stochastic filter model in experiment 2; (top) Averaged emission using 80 replicates and 80 lag length; the reference solution (continuous line), the filter estimate (dash-dot line) and the estimated standard deviations (full gray lines); (bottom) The innovations of the filter (circles) and theoretical standard error (dashed line).

Experiment 2: Deterministic emissions in the truth model and stochastic emissions in the filter model.
In the second experiment we propose to remedy the divergence by adding a fictitious noise to the emissions in the filter model.

$$e_{k+1} = e_k + w_k. \tag{2.19}$$

The important question is how much noise to add. On the one hand, small uncertainties in emissions may be too modest in preventing divergence. On the other hand, decreasing the confidence in the model leads to very fluctuating estimates. In Figures 2.5 and 2.6 the behavior of the filter and smoother is shown for the estimation of the constant emission parameter. We have used 80 ensemble replicates and the smoother has been used with 10, 40 and 80 lag lengths. The estimate is improving due to the increasing amount of information from the past data. In this context the best approximation of the reference constant emission is obtained using 80 lag length. The innovation plot shows a better agreement between the innovations and their theoretical statistics than in the case of experiment 1 (Figure 2.3). Using 80 ensembles, the smoother with the window of 30-40 length has no effective influence comparing to the filter assimilation. Increasing the lag length up to 80 the smoother provides better estimates.

Experiment 3: Stochastic emissions in the truth and filter model.
In the third experiment we changed the model by considering the reference emissions as fluctuating too. The performance of the smoother comparing to the filter in the reconstruction of sulphur emissions and sulphur itself are shown in Figure 8. The smoother solution provides a better fit to the reference trajectory. The large peaks provided by the EnKF solution are reduced using the smoother. We also have investigated the sensitivity and efficiency of the smoother implementations using this third case in the next sections of this chapter.

In Figure 2.7 (plot bar) we notice that the best improvement is obtained with moderate lag lengths (between 5 and 20). After this window size the smoother becomes less accurate, but it still provides better estimation comparing to the filter.
Concerning the divergence of the smoother, we have been confronted with two problems. The first one is related to the filter diverge problem. The smoother fails also when the filter does. The second problem concerning the smoother itself is when the smoother could not provide a more accurate estimates of the emission parameters comparing to the filter results due to the accumulation of the numerical errors of the algorithm. After some integration time, the smoother estimate becomes even less accurate comparing to the filter solution.

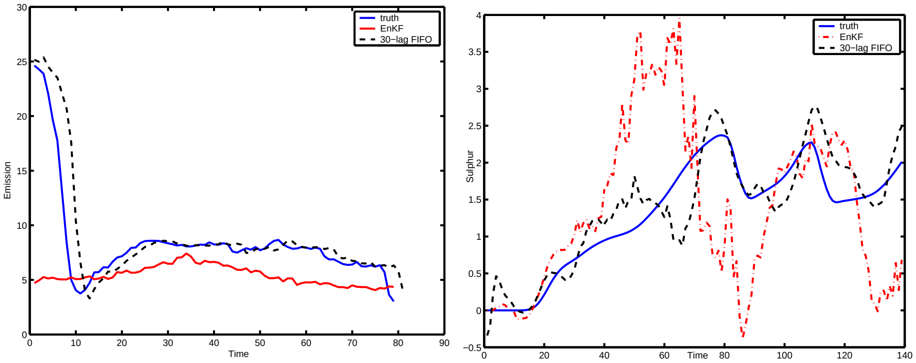

Figure 2.7: Time evolution of the emission and SO_2 concentration for the stochastic case, experiment 3 using the EnKF and FIFO with 30 lag length; full line is the reference solution, the filter is the gray dash-dot line and the smoother is the dotted line.

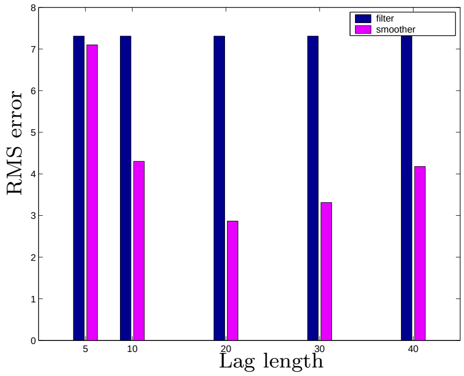

Figure 2.8: Root-mean-square estimation errors depicted using the EnKF and EnKS for different lag lengths; the filter result is represented by black bars and the smoother solution by the gray bars.

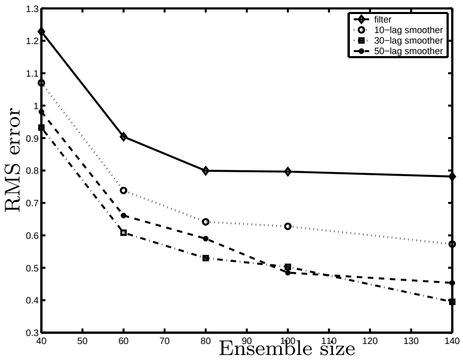

Figure 2.9: Root-mean-square estimation error of the assimilation using the ensemble fixed lag smoother as a function of the ensemble size; the filter is depicted by the full line and the smoothers with 10, 30, 50 lag length are represented by small dotted, dotted and dash-dotted lines, respectively.

2.5.2 Smoother sensitivity to the ensemble size and lag length

A parameter that is important for the assimilation in all ensemble-based algorithms is the number of the ensembles used in the assimilation. Simulations with different ensemble size were performed to study the sensitivity of the model to this parameter (see Figure 2.9). The accuracy of the estimates increases with the number of replicates, but the computational time becomes larger. The errors are calculated for the 80 last steps of the simulation and averaged over 20 runs, so the effect of the initial condition is not very important in the calculation of the errors. The accuracy of the filter does not improve significantly when the number of ensembles is larger than 80 as it is shown in Figure 2.9. However, by using the lagged smoother with 30 lag length the rms error is still decreasing when the ensemble size is increasing after 80 members. Figure 2.10 depicts the rms estimation errors provided by the filter and smoother for several lag lengths and ensembles. A number of 9 measurements has been used in this case. The lag length, chosen to stabilize the error, depends on the number of observations considered. A number of the simulations has been performed using different set of measurements. Figure 2.11 illustrates the effect of data density by depicting the rms error of the fixed lag smoother using three different sets of the measurement points. As one would expect, the errors decrease with the number of the measurements. Using the larger number of observations causes a rapid saturation of the smoother solution.

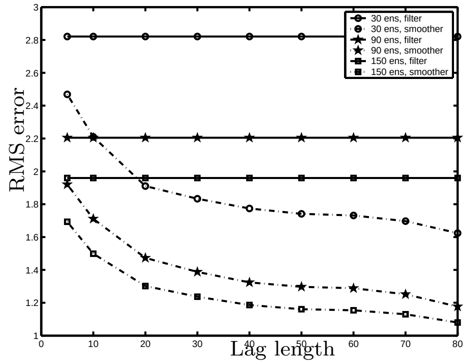

Figure 2.10: Root-mean-square estimation error of the assimilation using the ensemble fixed-lag smoother for several lag lengths and ensemble sizes.

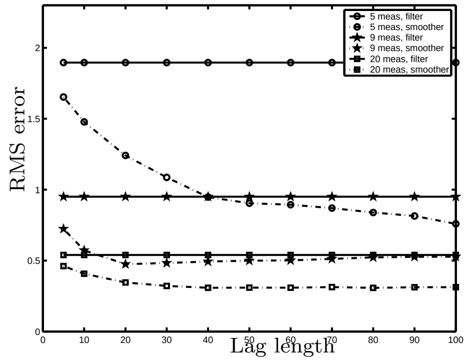

Figure 2.11: Root-mean-square estimation error of the assimilation using three different sizes of the measurement sets: 5, 9, and 20, respectively.

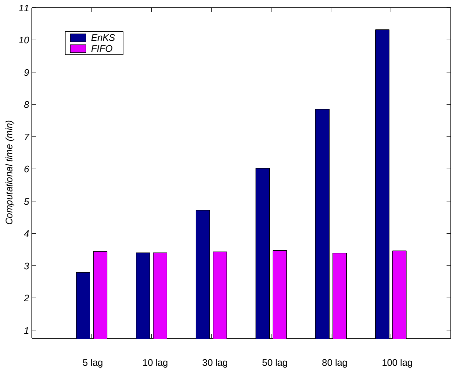

Figure 2.12: The comparison of the computational aspect between the EnKS and FIFO for several lag lengths.

2.5.3 Computational efficiency

By using the EnKS the computational effort required is a function of the lag length. The incremental cost of this algorithm over the EnKF shows a linear dependence on the lag length. The FIFO lag algorithm is independent of the lag length and more efficient than EnKS after a certain lag length. Comparing the smoothers with 80-lag length for example, the computational time for the FIFO lag algorithm is 50% smaller than for the EnKS (see Figure 2.12). However, the computational effort is less when the observations are sparse in time.

2.6 Conclusions

In this study we have used a simplified version of a chemistry transport model to perform experiments with a number of smoother algorithms. We have discussed that the linearity of the SO_2-SO_4 system used here is common to most models used for long term studies. Refraining ourselves to bulk aerosol approaches we argue that the results presented here are valid for the SO_2-SO_4 system as well as all primary aerosol components, such as sea salt and carbonaceous (combustion) particles. These primary components behave linearly as well in a bulk approach. For components with a nonlinear behavior (nitrate, SOA) we assess that the ensemble-based data assimilation scheme is able to take into account the nonlinearity of the system.

We have focused our work on the behavior of the smoother techniques: the algorithmic aspects of various data assimilation schemes and the accuracy and efficiency of the smoother algorithms. In this chapter three algorithms for emission parameter estimation using the ensemble smoother have been presented: the fixed lag smoothing, ensemble Kalman smoother and FIFO lag algorithm. The algorithms have been compared with each other. The simulation experiments used in this study showed the feasibility of the ensemble smoothing schemes to reconstruct the pollutant emissions. The smoother techniques are able to improve the filter estimates and provide more accurate results in terms of the rms error.

Although the techniques presented are found to be efficient, an application to real life still imposes large computational difficulties. In this study it is shown that the most efficient algorithm is the FIFO implementation, which produces the same estimates with less computational effort. The FIFO lag algorithm is faster than the EnKS, especially for large lag lengths. This helps to make the ensemble smoothing a practical option for the retrospective analysis of large air pollution data sets.

This chapter also studied the difficulties of data assimilation under the perfect-model assumption. Experiments showed that the filter (and the smoother) may diverge with small and also with large ensemble size. For preventing filter and smoother divergence, our approach is based on a stochastic model for the emission parameters in order to keep the covariance matrix sufficiently large. By using very noisy observations, the analysis from the smoother scheme is less accurate. When a larger set of observations with the same perturbations and the same noise specifications in emissions is used, the filter (and smoother) is over-inflated.

Since the estimate of a chemistry transport model is highly influenced by uncertain emissions it is important to adjust these parameters using retrospective data assimilation. In the next chapter we will apply our results to the LOTOS-EUROS model for the year 2003 in order to asses the impact of the real data on the reconstruction of the emitted pollutants.

CHAPTER 3

Assimilation of sulphur dioxide and sulphate

Abstract Fine particulate matter (PM) is relevant for human health and its components are associated with climate effects. The performance of chemistry transport models for PM, its components and precursor gases is relatively poor. The use of these models to assess the state of the atmosphere can be strengthened using data assimilation. This study focuses on simultaneous assimilation of sulphate and its precursor gas sulphur dioxide into the regional chemistry transport model LOTOS-EUROS using an ensemble Kalman filter. The process of going from a single component setup for SO_2 or SO_4 to an experiment in which both components are assimilated simultaneously is illustrated. In these experiments, solely emissions, or a combination of emissions and the conversion rates between SO_2 and SO_4 were considered uncertain. In general, the use of sequential data assimilation for the estimation of the sulphur dioxide and sulphate distribution over Europe is shown to be beneficial. However, the single component experiments gave contradicting results in direction in which the emissions are adjusted by the filter showing the limitations of such applications. The estimates of the pollutant concentrations in a multi-component assimilation have found to be more realistic. We discuss the behavior of the assimilation system for this application. The model uncertainty definition is shown to be a critical parameter. The increased complexity associated with the simultaneous assimilation of strongly related species requires a very careful specification of the experiment, which will be the main challenge in the future data assimilation applications.

This chapter is a slightly revised version of [5].

3.1 Introduction

Many studies have found relations between the aerosol concentrations in the atmosphere and mortality and respiratory problems ([12], [60]).

The composition of aerosol is determined by their origin and the physical and chemical processes, which they undergo in the atmosphere. A distinction can be made in primary and secondary aerosols. Primary aerosols are emitted in the atmosphere, whereas secondary aerosols are product of chemical reactions. Over Europe the primary emissions is largely due to anthropogenic activities. Natural emissions are biomass burning as forest fires and volcanoes.

During the last decades, models were developed to describe the fate of (particulate) pollutants over Europe. Despite advances in computer technology, data collection and numerical modeling techniques, performance evaluations of chemical state of the atmosphere in air quality models demonstrate that their solution can contain important errors. The causes of these errors are highly complex and only partly understood. Large uncertainties in emissions on spatial and temporal scales, formation routes and sinks of particles cause model performances for PM and its components to be relatively poor [109]. In order to decrease modeling errors due to the imperfect knowledge on initial or boundary conditions, the emission rate and chemical processes, data assimilation techniques have been proved to be very effective.

Variational methods and Kalman filter techniques have been applied in air pollution modeling for ozone, e.g. [109], [32], but seldom for components of particulate matter. Most of the applications are single-species experiments in which one component or almost independent species are assimilated into a model system. These experiments are often dedicated to the estimation of emissions, the driving force in air pollution. Examples include studies for estimation of air pollution emissions of NOx, VOC, CO and SO_2, e.g. [56], [72], [52] and emissions of greenhouse gases such as methane, e.g. [63], [115] and carbon dioxide [36]. However, it is recognized that information on multiple species is required to better constrain the models due to the tight coupling in the chemical processes in the atmosphere [18].

A study by [33] uses as methodology 4-dimensional variational multi-component assimilation to assess the potential and limits of estimating pollutant precursor sources. A multi-component sequential data assimilation scheme in which the species to be incorporated are strongly dependent is not often performed. A study by [83] showed that in case the model uncertainty is low, chemically consistent results can be obtained by assimilating multiple components. However, in case the model uncertainties are large, an experiment may pose additional challenges for the implementation of a data assimilation scheme. For example, [21] proved that it is necessary to impose additional constraints on the correlations to ensure the consistency between assimilated and non-assimilated fields. In case of large model uncertainties, these findings may indicate that the results obtained in a single component setup may be different from a multi-component experiment.

Application of data assimilation to the components of particulate matter is presently hampered by a number of factors. The most important issues are sparse and often uncertain observations for PM and its components and the tendency of models to underestimate PM concentrations [30]. In this chapter we focus on sulphur dioxide (SO_2) and particulate sulphate (SO_4), for which a reasonably large number of observations are available in Europe [58]. The key question is how accurate can the sulphur dioxide and sulphate concentrations be estimated using a single component setup compared to the combined assimilation procedure? We use an ensemble Kalman filtering technique to answer this question.

These chapter is organized as follows. In the second section the LOTOS-EUROS model used in this study is introduced. Section 3 concerns the description of the observations from the the European Monitoring and Evaluation Programme (EMEP) database used in our assimilation and validation experiments, followed in section 4 by the comparison between model results and measured data for the year 2003. Sequential data assimilation methodology that encompasses the definition of the stochastic environment and ensemble-based filter algorithm is presented in section 5. Section 6 contains the assimilation results concerning sulphur dioxide and sulphate aerosol for different experimental setups. Finally, the last section summarizes the concluding remarks of our research.

3.2 Deterministic LOTOS-EUROS model

We use model version $v1.3.$ as in [95]. This and earlier model versions have been applied in numerous studies related to (inorganic) PM, e.g. [30], [101], [95]. Below we describe the model features relevant for this chapter. The model is described according to:

$$c(k+1) = M_{LE}(c(k)). \tag{3.1}$$

Here, the state space operator of the LOTOS-EUROS model is denoted by M_{LE}. This operator computes the concentration vector c, which contains all considered components for each grid cells, at time k+1 hour given the concentration at time k. The time interval between two consecutive steps is one hour in which the model performs three time steps of 20 minutes.

The model domain used in this chapter is bound at 35° and 70° North and 10° West and 40° East covering Europe from the western part of Russia to the Atlantic Ocean and from the Mediterranean Sea to Scandinavia (see Figure

3.1). In this chapter the model is driven by meteorological data produced at the Free University of Berlin employing a diagnostic meteorological analysis system [71].

Significant emissions of sulphur oxides (SO_x) occur in most parts of Europe. The main sources of sulphur oxides are combustion of sulphur containing fuels. When the fuels are burned, sulphur is oxidized to sulphur dioxide. Within this study, emissions of SO_x (SO_2 and SO_4) have been taken from the TNO emission database, which is available at a resolution 0.25° by 0.125 ° for the year 2000 [111]. The amount of primary sulphate emissions is estimated to be 3% from the total SO_x emissions. The country total emissions in this data base are equal to those in the CAFE 2000 baseline emissions. Table 3.1 shows the European emissions per major source category. The most important source group of emitting SO_x is the power generation (11 megatons). International shipping is a significant source too (2.16 megatons) compared to the total land based emissions (15.85 megatons). Area sources are injected into the lowest layer, whereas the emissions from point sources are injected according to stack height. The annual emission totals are translated to hourly emissions in LOTOS-EUROS using prescribed temporal factors as in [16].

`Sources`	`EU 27`	`Non EU`
Power generation	7.62	3.44
Residential combustion	0.69	0.30
Industrial combustion	1.45	0.93
Industrial processes	0.67	0.16
Fossil fuel extraction	0	0.003
Road transport	0.14	0.06
Other mobile sources	0.26	0.05
Waste treatment	0.003	0.0
Agriculture	0.002	0.02
Total land emissions	10.86	4.99
Shipping emissions	2.16	
Total emissions	**18.01**	

Table 3.1: Total emissions of SO_x (megatons) for anthropogenic activities in Europe.

In this study we focus on the sulphur cycle. There are two distinct pathways to oxidize SO_2 to SO_4 in the atmosphere. The first is the oxidation of SO_2 by the hydroxyl radical (OH):

$$(R_1)\ SO_2 + OH \xrightarrow{k_1} SO_4, \quad k_1 = 1.5\ [1/\text{min.}1/\text{ppb}].$$

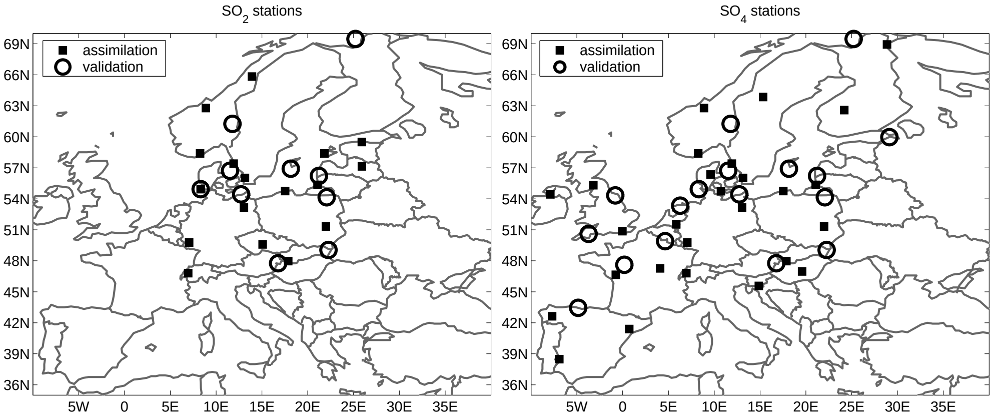

Figure 3.1: Geographical location of the stations for sulphur dioxide (left) and sulphate (right); squares and circles represent assimilation and validation stations, respectively. Note that two German stations, one for assimilation and the other for validation are situated close to each other and therefore, represented by the overlapping marks.

The second pathway is the oxidation of SO_2 in the aqueous phase (clouds, fog, etc) by ozone and hydrogen peroxide. Therefore, in addition to the gas phase reaction of OH with SO_2, the oxidation pathways in clouds is represented by a simple first order reaction:

$$(R_2) \qquad \mathrm{SO_2} \xrightarrow{k_2} \mathrm{SO_4}.$$

The reaction rate is a function of cloud cover and relative humidity:

$$k_2 = \bar{k_2} * (1 + 2 * c) * \gamma, \tag{3.2}$$

$$\gamma = \max\{1, 1 + 0.1 * (\mathrm{RH} - 90)\}, \tag{3.3}$$

where $\bar{k_2} = 8.3 * 10^{-5}$ [1/min.1/ppb], $c \in [0,1]$ is cloud cover and γ is a function of relative humidity RH (%) described by the latter expression. This parameterization enhances the oxidation rate under cool and humid conditions. This simple parameterization performs well compared to models which include the aqueous phase chemistry explicitly (see [70], [95]). The change in concentration fields according to reactions R_1 and R_2 can be described as follows:

$$\frac{\partial[\mathrm{SO_2}]}{\partial t} = -(k_1 * [\mathrm{OH}] + k_2) * [\mathrm{SO_2}], \tag{3.4}$$

$$\frac{\partial[\mathrm{SO_4}]}{\partial t} = (k_1 * [\mathrm{OH}] + k_2) * [\mathrm{SO_2}]. \tag{3.5}$$

The brackets applied to a chemical component means the concentration of that species. We denote by $\bar{r}$ the reaction rate:

$$\bar{r} = k_1 * [\,\mathrm{OH}] + k_2. \tag{3.6}$$

Since sulphur dioxide has a negligible effect on OH formation, the simulations of sulphur cycle can run independently using predefined OH concentrations. LOTOS-EUROS contains this option and we have used this approach for the assimilation experiments as it decreases the computational costs substantially.

3.3 Observation network for assimilation and validation

The LOTOS-EUROS model is designed to represent regional background concentrations. Therefore, measurements of SO_2 and SO_4 have been gathered as daily averages from EMEP database [35] which provides high quality data for background concentrations. The coverage of the data network is dense in central and northern regions, but sparse for large parts of Europe, especially for sulphur dioxide [58]. Sites with less than 30 days of measurements have been removed. Mountain stations (altitude higher than 700 m) have been also excluded for the model to observations comparison due to the fact that the model is not able to represent well the elevated sites. The number of remaining stations measuring SO_2 only, SO_2 and SO_4, or SO_4 only is shown in the Table 3.2. Consequently, a total number of 27 stations providing data for sulphur dioxide and 44 for sulphate distributed over Europe (see Figure 3.1) is used in this study.

For validation purposes it is useful to split the set of observations into two parts. The first set of data will be incorporated in the assimilation process to obtain the optimal estimate of the state. The second data set will not be used for the assimilation, but only to verify the results. A selection of stations, chosen to represent different regions equally was made for each of these purposes. In total, the assimilation set contains 17 sites for SO_2 from which 6 stations measure only SO_2 and 27 for SO_4 with 16 stations providing sulphate data only. The validation set contains 10 sites for SO_2 (all of them measuring both species) and 17 for SO_4 (see Table 3.1).

In following sections the LOTOS-EUROS model output is compared with EMEP data using four statistical measures. The analysis is based on pairs of modeled (M) and measured observations (O) at a number of S stations over

Species	SO_2	SO_2 & SO_4	SO_4	Total
Stations	6	21	23	50
Assimilation stations	6	11	16	33
Validation stations	0	10	7	17

Table 3.2: The number of observation locations where SO_2 only, both SO_2 and SO_4 and SO_4 only, respectively, are measured from which a number of assimilation and validation stations has been chosen.

D days. The ratio of the model results over the observed data is defined as:

$$\text{ratio} = \frac{\sum_{s=1}^{S} \sum_{d=1}^{D} M_{s,d}}{\sum_{s=1}^{S} \sum_{d=1}^{D} O_{s,d}}. \tag{3.7}$$

The residual is the sum of the absolute difference between the model and observed results for all stations and whole year.

$$\text{residual} = \frac{1}{S} \sum_{s=1}^{S} \frac{1}{D} \sum_{d=1}^{D} |M_{s,d} - O_{s,d}|. \tag{3.8}$$

Another statistical parameter is the root mean square (rms) error defined as:

$$\text{rms} = \frac{1}{S} \sum_{s=1}^{S} \sqrt{\frac{1}{D} \sum_{d=1}^{D} \left(M_{s,d} - O_{s,d}\right)^2}. \tag{3.9}$$

The average correlation coefficient is defined using the correlation in time at the individual stations as:

$$\text{corr}_s = \frac{\sum_{d=1}^{D} \left(O_{s,d} - \bar{O}_s\right)\left(M_{s,d} - \bar{M}_s\right)}{\sigma_{s,O} * \sigma_{s,M}} \tag{3.10}$$

$$\text{corr} = \frac{1}{S} \sum_{s=1}^{S} corr_s. \tag{3.11}$$

Here $\bar{O}_s$ and $\bar{M}_s$ are the observed and modeled means at a station s respectively with corresponding standard deviations σ.

3.4 Model validation

In this section we describe the simulated sulphur dioxide and sulphate distributions for the year 2003 as well as an evaluation against measurements. The

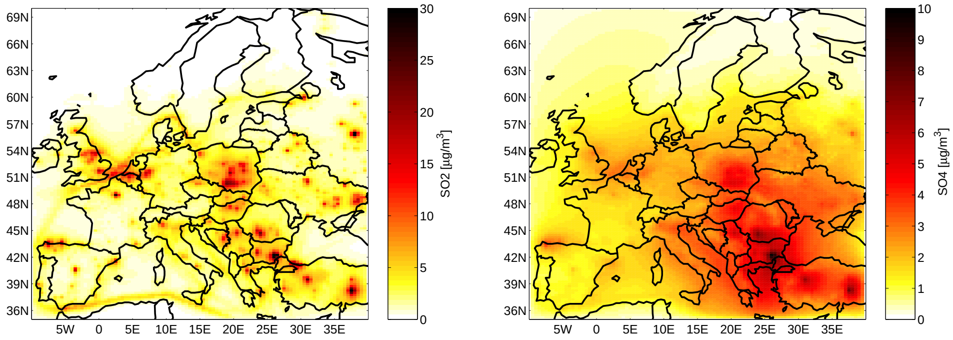

Figure 3.2: Annual mean modeled distribution (μg/m^3) of sulphur dioxide (left) and sulphate (right) over Europe.

annual average modeled concentrations of sulphur dioxide and sulphate are illustrated in Figure 3.2. The concentrations of sulphur dioxide are high around the densely populated and heavily industrialized regions in the Balkans, central Europe, northern Spain, England and the Benelux countries. Furthermore, the ship tracks can be recognized. Low values are found in Scandinavia and southern France. A sulphate concentration band of 2 to 5 μgm^{-3} is found over western Europe to the Balkans with maximum concentrations in southeastern Europe. Highest sulphate concentration in this region range between 5 and 10 μgm^{-3}.

Table 2 (in section 6) shows the statistical parameters introduced in section 3 (Equations 3.7- 3.11) for a comparison between the model and the observations. The model tends to overestimate EMEP SO_2 concentrations especially in winter. This is a well known problem in most models. The annual mean value of SO_2 given by the model at all stations is 2.3 μgm^{-3} compared to a measured value of 1.6 μgm^{-3}, which gives an average ratio of 1.4. A large part of the bias is due to high modeled concentrations at several coastal stations (SE14, SE11, DK08) and in Poland (PL02) (see Figure 3.3). Hence, the high mean overestimation is largely determined by a few extreme values. However, a general overestimation is observed for almost all locations (see Figure 3.6). For most stations the correlation coefficient lies in the range between 0.40 and 0.60. Lower correlation coefficients (less than 0.30) have been found at sites in Norway, where concentrations are very low.

The SO_4 modeled annual average is 1.8 μgm^{-3}, compared to 2.5 μgm^{-3} for observed concentrations. Hence, for SO_4 the model underestimates the measured concentrations with 28 % on average. We find the modeled data to be consistently lower than the observed data (see Figure 3.6). The correlation coefficient is on average 0.47 for sulphate. The temporal behavior of the model is illustrated for a German station (DE07) with a correlation coefficient of 0.61 in Figure 3.3. The model underestimates especially the peak concentrations. The spring period with very high observed concentrations was subject of a model inter-comparison exercise in which none of the models could explain the observed concentrations [101]. For sulphate, poor correlation coefficients have been found for a number of stations situated in Spain, France and Russia. High correlations (larger than 0.7) were obtained for sites in in Sweden and Denmark.

In summary, except few stations with a large overestimation, the model is able to reproduce the SO_2 observations with a slight overestimation, while for SO_4 the model underestimates the observed concentrations systematically. These results suggest that the general model underestimation of sulphate may be caused by an underestimation of conversion rate between the species.

3.5 Sequential data assimilation methodology

3.5.1 Stochastic state space representation

For application of the filter algorithm to the LOTOS-EUROS model, a stochastic representation should be defined for the model error. The knowledge of uncertainties is crucial for a successful data assimilation. Using a stochastic model for several uncertain parameters, an assimilation scheme is able to produce an optimal estimate of the state and parameters given the observations.

Consider a parameter p that has been modeled uncertain using a colored noise process λ_p which has the following equation in scalar form:

$$\begin{aligned} \lambda_p(k+1) = \alpha\lambda_p(k) + \sigma\sqrt{1-\alpha^2}w(k), \\ w(k) \sim N(0,1). \end{aligned} \tag{3.12}$$

The coefficient $\alpha \in [0,1]$ represents the time correlation parameter. Setting α to zero, we obtain a white noise sequence with zero mean and variance σ. If α is one, the random process is reduced to a constant value. The temporal covariance $E(\lambda_p(k+l)\lambda_p(k))$ is equal to α^l. λ_p is a stationary Gaussian process with an exponential covariance function using the parameterization $\alpha = \exp(-1/\tau)$ for a given time correlation length τ.

Following [56], a stochastic model state is formed by augmenting the state vector (Equation 2.1) with the noise process λ_p as:

$$\begin{aligned}\begin{bmatrix} c(k+1) \\ \lambda_p(k+1) \end{bmatrix} &= \begin{bmatrix} M_{LE}(c(k), \lambda_p(k)) \\ \alpha\lambda_p(k) \end{bmatrix} \\ &+ \begin{bmatrix} 0 \\ \sigma\sqrt{1-\alpha^2} \end{bmatrix} w(k), \end{aligned} \tag{3.13}$$

For each element of the vector of the parameters, the associated noise process to this element can have different values of the time correlation length and/or standard deviation.

The new augmented state vector is denoted by:

$$x(k+1) = M(x(k)) + Gw(k). \tag{3.14}$$

The nonlinear operator M describes the time evolution from k to k+1 of the augmented state vector x which contains the concentrations and parameter vector, and Gw is a forcing term.

3.5.2 Observational operator

The state of the observational network is defined by the observation operator H that maps state vector x to observation space y. Since y contains daily average concentrations, the state vector has to be augmented with daily average values of the observed components. The daily average fields in the state are updated every time step of one hour given the new instantaneous concentration c. The operator H then simply selects the grid cells in the daily average fields in x that correspond to observation locations.

The measurements have white Gaussian errors v with covariance denoted by R:

$$y(k) = H(x(k)) + v(k), \quad v \sim N(0, R). \tag{3.15}$$

This error accounts for the uncertainty in the actual observation at a specific station as well as for its representativeness error.

3.5.3 Ensemble Kalman filter and smoother

A data assimilation system has been developed around the LOTOS-EUROS model based on the Ensemble Kalman Filter (EnKF) technique. In its Monte Carlo formulation, the pdf of the state is not expressed in terms of a mean and covariance only, but is described by an ensemble of model state. The spread between the ensembles replicates should describe the uncertainty in the value of the state.

In the first step of this algorithm an ensemble of N states $\xi^a(0)$ is generated to represent the uncertainty in $x(0)$. In the second step, the *forecast step*, the stochastic model propagates the ensemble members from the time $k-1$ to k:

$$\xi_j^f(k) = M(\xi_j^a(k-1)) + w_j(k-1), \tag{3.16}$$

where w_j represent the realizations of a white noise process w. The model state is represented by the ensemble mean x^f:

$$x^f(k) = \frac{1}{N}\sum_{j=1}^{N} \xi_j^f(k). \tag{3.17}$$

The forecast error covariance matrix P^f is assumed to be carried by the ensemble of perturbations L^f:

$$L^f(k) = \left[\xi_1^f(k) - x^f(k), \dots, \xi_q^f(k) - x^f(k)\right], \tag{3.18}$$

$$P^f(k) = \frac{1}{N-1} L^f(k) L^f(k)^\top. \tag{3.19}$$

When the measurements become available, the ensemble replicates are updated in the *analysis step* using the Kalman gain:

$$\begin{gathered} K(k) = P^f(k)H(k)^\top * \\ \left[H(k)P^f(k)H(k)^\top + R(k)\right]^{-1}, \end{gathered} \tag{3.20}$$

$$\begin{gathered} \xi_j^a(k) = \xi_j^f(k) + K(k) * \\ \left[y(k) - H(k)\xi_j^f(k) + v_j(k)\right], \end{gathered} \tag{3.21}$$

where v_j represent the realizations of the white noise processes v.

Advantages of the ensemble formulation is that the dynamical model is not restricted to linearity and, the implementation could be rather simple. The number of required ensemble members depends on the complexity of the pdf to be captured, which is usually determined by the nonlinearity of the model and the description of the involved uncertainties. In practice, an ensemble with 10-100 members is acceptable to keep computations feasible.

3.5.4 Localization procedure

Due to finite size of the ensemble set, spurious correlations between elements in the state vector arise. Some other unrealistic correlations through Erisman the domain may be introduced by the noise processes. Such undesired correlations at large distances can be eliminated by coupling the ensemble Kalman filter with the localization of covariances (see [62]). The localization is achieved using a Schur product of the covariance matrices of the background error and a correlation function with local support. The Kalman gain K is calculated according to:

$$K = \left(f \circ P^f\right) H^\top \left[H \left(f \circ P^f\right) H^\top + R\right]^{-1}. \tag{3.22}$$

The Schur product is defined by $f \circ P^f$ and represents the element-wise multiplication of a correlation matrix with the covariance matrix P^f. The correlation matrix is obtained by applying the correlation function f to the Euclidean distance between two points. This function has a Gaussian form and depends on one parameter. Only the measurements within a distance depending on this bounding parameter are used to update the model state. The correlations decrease to zero at a finite radius determined by this length scale parameter which was set to 100 km. Experiments showed that the use of larger radius imposed the use of larger ensemble size to avoid the degradation of the filter performances.

3.6 Assimilation results

3.6.1 Experimental setup

The ensemble Kalman filter is applied to different stochastic versions of LOTOS-EUROS. The application runs over the European domain and is directed to strongly dependent species, namely sulphur dioxide and sulphate. The study comprises assimilation procedure in two different setups. In the first set of experiments (section 6.2.) the targeted parameter is represented by the noisy emissions. To this an uncertain chemical reaction rate is added in the second set of experiments (section 6.3). Daily ground-based measurements derived from the EMEP database for the year 2003 have been used in the analysis step of the EnKF algorithm that is performed at midnight. It is assumed that the uncertainty in the measurements is defined using a fixed standard deviation of 10%. The assimilation of SO_2 or SO_4 only, as well as simultaneous incorporation of both species in the model, is addressed in the next sections.

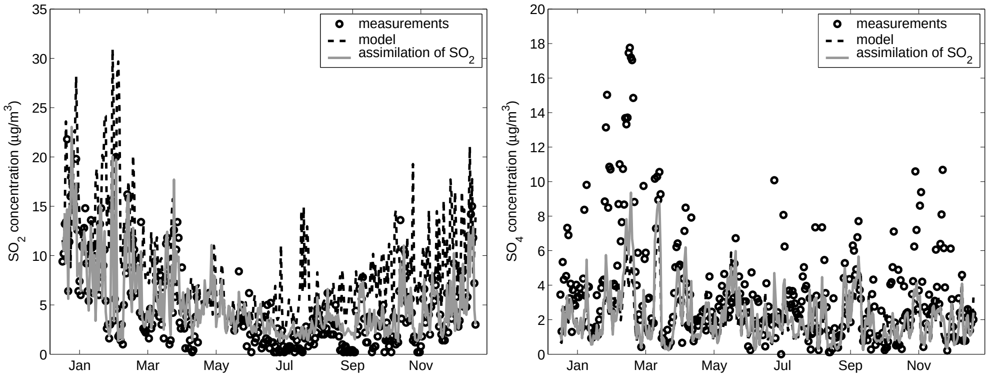

Figure 3.3: Annual time series of daily average sulphur dioxide concentrations at PL02 station (left) and sulphate concentrations at DE07 station (right) obtained by using uncertain emissions and assimilating SO_2 concentration data only. Note that the model and assimilation results in the right panel are almost identical.

The behavior of our data assimilation system is studied using EnKF with 15 ensemble members. Computational costs are dominated by the propagation of the ensemble members and, thus, increase with the size of ensemble set. Experiments showed that using an ensemble of 30 members the performance of the algorithm did not increase significantly to justify the additional computational cost. The small ensemble size imposes the use of a localization distance of 100 km (section 5.3.) that has been found to give satisfactory results.

3.6.2 Uncertainties in emissions

The assimilation system has been applied to a number of cases to illustrate the process of going from a single to a multi-component assimilation experiment. The first two experiments mimic typical single component experiments in which we assimilate only SO_2 or SO_4, respectively. The third experiment explores the results of incorporating both components into the analysis. As emissions are one of the major sources of uncertainty in air quality modeling, we follow other single component studies by using the perturbation factor on the emission parameter to compensate the model errors.

The stochastic model has been built by adding uncertainties to the deterministic SO_2 emissions denoted by $\bar{e}(k)$ in a grid cell:

$$e(k) = \max\left(0, \bar{e}(k)\left(1 + \lambda_e(k)\right)\right). \tag{3.23}$$

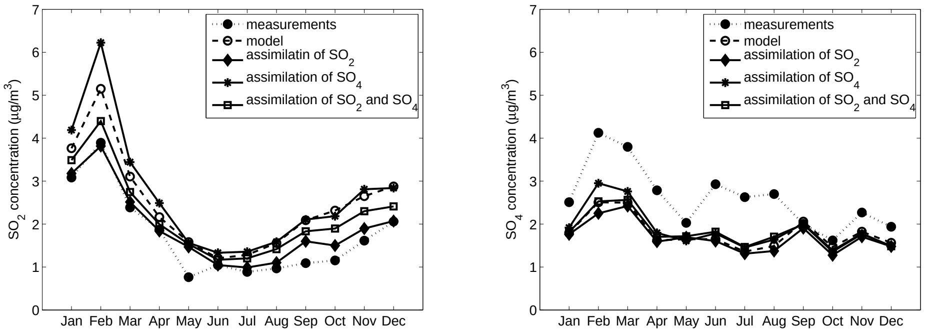

Figure 3.4: Comparison of the measured seasonal variation of the sulphur dioxide (left) and sulphate (right) concentrations with those obtained by the model and the three assimilation experiments with uncertain emissions only. The data represents the averaged concentration over all assimilation stations.

The standard deviation of the colored noise process λ_e set to 40%. A standard deviation of 30% for sulphur dioxide emissions was considered in [67]. We have increased the uncertainty since additional errors can be introduced by the way how the emissions are broken down into hourly emission estimates. The time correlation length is set to three days, about the half lifetime of the sulphur dioxide.

To illustrate how the assimilation procedure performs, two assimilation stations have been selected: PL02 (Jarczew) and DE07 (Neuglobsow). The first station has been chosen to illustrate the model tendency to overestimate the sulphur dioxide and the second one has been selected to study the behavior of SO_4 concentrations that is representative for central and northern Europe. Figure 3.3 shows the time series for sulphur dioxide provided by the assimilation of SO_2 measurements only. At PL02 site (left panel) where a high deviation of the simulated SO_2 concentrations from the measured data was noticed, the time series showed a substantial improvement. The assimilation reduces the bias between the model and the measurements by half on average. At DE07 station (right panel) the model underestimates sulphate over the whole year but especially during February-March. The filter is not able to lower the residue between the model and measurements for this period. Moreover, due to the decrease in sulphur dioxide emissions, the sulphate concentrations are lowered slightly.

The results of the assimilation process are summarized in Figure 3.4 which illustrates three experiments either assimilating only SO_2, only SO_4 or both

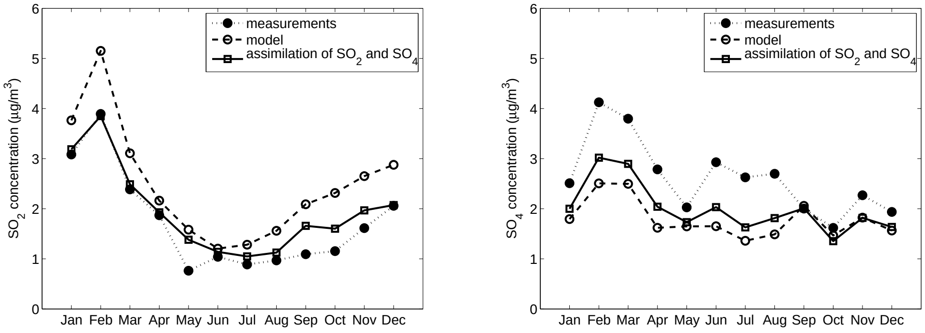

Figure 3.5: Monthly mean sulphur dioxide (left) and sulphate (right) concentrations averaged over assimilation stations obtained by using uncertainties in emissions and reaction rate and assimilating SO_2 and SO_4 simultaneously.

components. In this figure we compare the impact of these experiments as function of season for all assimilation stations. The left panel shows that the model exhibits a rather constant bias of 0.7 μgm^{-3} throughout the year. By assimilation of only SO_2 measurements the ensemble Kalman filter is able to compensate this bias almost completely by reducing the SO_2 emissions around the assimilation stations. Consequently, also the sulphate concentrations are slightly reduced as the lower SO_2 concentrations also cause a lower conversion of SO_2 to SO_4. In the second experiment only sulphate observations are assimilated. Though the filter is not very effective in increasing the sulphate concentrations, the SO_2 emissions are increased to convert more SO_2 to SO_4 and decrease the bias in sulphate levels. Hence, the SO_2 concentrations in the system rise and the overestimation of SO_2 is increased. The same features are also observed at the validation stations. As expected from the validation of the model, the filter needs to decrease or increase the emissions to optimally estimate SO_2 or SO_4 concentrations, respectively.

Theoretically, by using more information to update the state of the system the performance of the filter should increase. Therefore our third experiment concerns the combined assimilation of SO_2 and SO_4 measurements. In this case the filter balances the influence of the opposing measurements. This yields an improved SO_2 estimate in which the bias is reduced by half and the absolute values fall roughly in the middle between the model and the first experiment with only SO_2 assimilation. For sulphate the combined assimilation yields values close to the model results.

In these experiments the degree of freedom in the stochastic system is only associated with emissions. The results of our first multi-component experiment shows that only considering the stochastic emission forcing, the model error is misspecified and impacts the data assimilation analysis considerably. Therefore, we expand our experiments with an additional source of uncertainty.

3.6.3 Uncertainties in emissions and reaction rate

The overestimation of sulphur dioxide in combination with the underestimation of sulphate may indicate that the conversion rate between these species is too slow. Imprecise knowledge about heterogeneous reaction pathways and rates in clouds and fogs [70] or uncertainties in the estimated OH concentrations make the reaction rate uncertain. Hence, for the next experiment, model uncertainty has been assumed for the emissions (as in the previous experiments) as well as the reaction rate given in the Equation 3.6 :

$$r(k) = \max\left(0, \bar{r}(k)\left(1 + \lambda_r(k)\right)\right). \tag{3.24}$$

The standard deviation of the colored noise process λ_r added to the deterministic reaction rate $\bar{r}$ was set to 50%. This choice is justified by the poor knowledge of the conversion of SO_2 to SO_4 related to the uncertainties listed above. The time correlation parameter was set to three days as for emissions.

The results of this assimilation experiment are summarized in the Figure 3.5. There we compare the assimilated monthly averaged concentrations with the measured and simulated data. From January to April, the positive bias in the modeled SO_2 concentrations is reduced by the assimilation for both analyzed as well as verification sites. For the same period the negative bias in the SO_4 concentrations is decreased by half. The period May-August is characterized by the ability of the system to capture the low SO_2 concentrations. There remains a negative bias of sulphate, feature which is detected over the validation stations also. In the last period of the year (September-December) we have noticed a reduction of SO_2 by half on average. Being quite well estimated by the model the sulphate component hardly benefits from the assimilation. All these features has been detected over the validation stations too (results not shown here).

The benefit of using two stochastic parameters in the experiment is significant for the performance of the assimilation, especially for the first half of the year (compare to Figure 3.4). The use of the uncertain reaction rate has a positive impact on both components. For both SO_2 and SO_4 the assimilated concentrations are similar or closer to the concentrations obtained in the single component assimilation of SO_2 and SO_4, respectively.

Figure 3.6 illustrates a more elaborate comparison of the annual averaged measured data with either the modeled (open marks) or the assimilated values

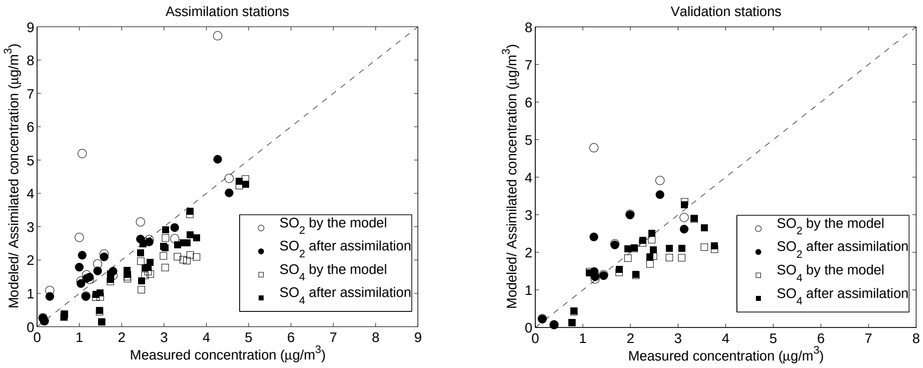

Figure 3.6: Comparison of the annual mean concentrations of sulphur dioxide and sulphate from the model and assimilation of both components against measurements for the assimilation (left) and validation (right) stations.

(filled marks). The results are provided for the two sets of stations: assimilation and validation. The large overestimation of SO_2 concentrations observed at several sites is significantly reduced. In general, the SO_2 concentrations are slightly lowered by the assimilation. For sulphate the filtering procedure shows a reduction of the discrepancy for all stations. As for most data assimilation experiments the reduction in the bias is smaller for the validation stations than for the stations which contribute to the assimilation process. More, due to the localization effect the improvement around the assimilation sites extends only to limited areas.

Table 3.3 summarizes the statistical performance of the model and assimilation in the experiment with both sources of uncertainty. The annual mean for sulphur dioxide calculated for all 27 stations was reduced from 2.3 to 1.9 μgm^{-3}. Since sulphate was underestimated, the assimilation increases the modeled mean from 1.8 to 2.0 μgm^{-3}. The residual and rms error decrease for both components, but more for sulphur dioxide than for sulphate. Also the correlation increases significantly to values above 0.6 for both components. Although the average sulphate concentration does not change much, the increased mean correlation and decreased average residual show that the temporal behavior for sulphate did improve significantly. In Table 3.4 the quantification in percents of the multi-component assimilation impact is given for the two different setups. An overall increased performance of multi-component scheme with two uncertain model parameters over the assimilation experiment with only one uncertain parameter is noticed for sulphate component. For example, the improvement of the rms error increases from 11 to 20%.

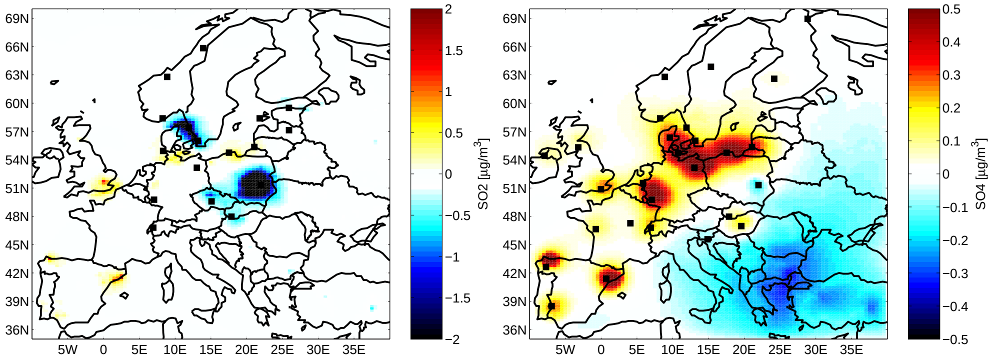

Figure 3.7: The difference between assimilated and simulated annual mean distributions for sulphur dioxide (left) and sulphate(right) over Europe. The assimilation stations are indicated with square marks.

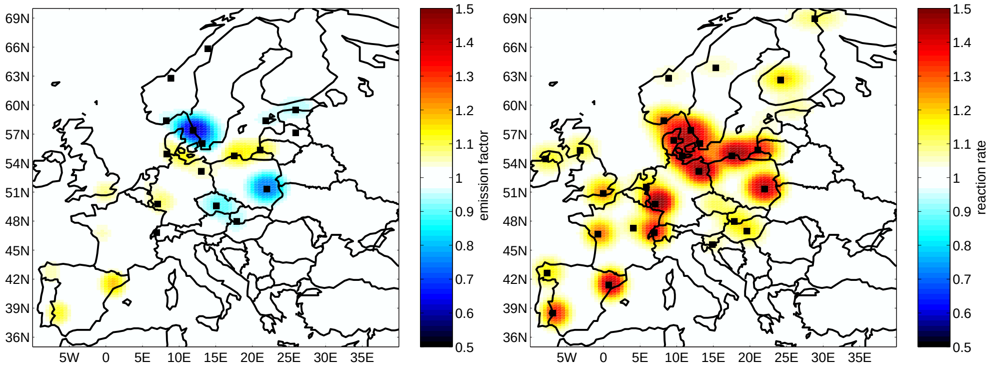

Figure 3.8: Annual mean distributions of the corrections factors for emissions (left) and reaction rate (right) obtained by the simultaneous assimilation of both sulphur compounds. The assimilation stations are indicated with square marks.

Figure 3.7 shows the difference between annual averaged assimilated and modeled concentrations for sulphur dioxide (left panel) and sulphate (right panel), respectively. The corresponding correction factors for the emission strength and reaction rate that lead to these concentration changes are depicted in the Figure 3.8. The spatial distribution shows that the annual averaged SO_2 concentrations are lowered by the filter in central Europe and over the sea area between Denmark and Sweden. In these regions the overestimation of sulphur dioxide levels is compensated by an emission reduction of about 25%. The emission forcing for other zones is larger than the deterministic values by about 10%. Similar improvements on the analyzed correction factors for SO_2 emission rates have been obtained by [34] in a 4-dimensional variational experiment with a low resolution model. Also the general features of emission correction distribution (reduction in central European area and amplification in Iberian peninsula) are similar with those obtained in our study (see Figure 3.8, left panel). The reaction rate is increased everywhere (see Figure 3.8, right panel) showing that the filter uses both sources of uncertainty to compensate the discrepancies. At the measurement locations in western Europe the model underestimates the sulphate concentrations and therefore the filter tends to lower the bias by increasing the emissions and conversion rate of SO_2 into SO_4. Also by adding an extra source of uncertainty in the reaction rate, the filter become less confident in the model and consequently, the sulphate measurements have more impact on the analysis. As result more sulphate is produced, whereas the net effect for SO_2 is slightly negative to neutral. If the stochastic model is accurate these results suggest that the reaction rate in the model may be underestimated.

One should be aware that the availability of measurement data has an impact on the results. In parts of Spain the sulphur dioxide emissions and concentrations are increased. The reason is that there are no Spanish SO_2 measurements in the EMEP database for 2003. Hence, the results in Spain resemble those of the single component experiment with only sulphate measurement data. We would also like to draw attention to the lowered sulphate mass in south-eastern Europe. The results should be carefully interpreted due to the lack of any measurements. In this area small difference between model and ensemble mean become evident. It also illustrates that the updates by the assimilation scheme impact areas outside the radius of influence of a particular station through advection, e.g. North Sea. Nonetheless, we stress that for these applications it is important that monitoring data provide complete sets of related components.

The augmented state vector with the past parameters lead to the ensemble smoother instead of ensemble filter. In contrast to filtering, the smoother analysis results from a retrospective assimilation of all observed data, both

	Model		**Uncertain emissions and reaction rate**	
	S0$_2$	**S0$_4$**	**S0$_2$**	**S0$_4$**
observed mean	1.6	2.5	1.6	2.5
calculated mean	2.3	1.8	1.9	2.0
ratio	1.4	0.7	1.1	0.8
residual	1.5	1.4	0.9	1.1
rms	2.1	2.1	1.5	1.7
corr	0.48	0.47	0.62	0.65

Table 3.3: Statistical comparison of modeled and assimilated concentrations of SO_2 and SO_4 in the experiment using uncertain emissions and reaction rate parameters. All statistics have been averaged over all considered stations.

Improvement	**Uncertain emissions**		**Uncertain emissions and reaction rate**	
%	**S0$_2$**	**S0$_4$**	**S0$_2$**	**S0$_4$**
ratio	14	2	17	10
residual	33	12	40	21
rms	28	11	28	20
corr	22	22	22	27

Table 3.4: Statistical improvement (given in %) of assimilation over the model simulation in the experiment with either uncertain emissions or uncertain emissions and reaction rate, respectively, for for both species.

past and future measurement being incorporated into analysis. Increasing the amount of data available for each time step may reduce the analysis errors. For the emission estimation purpose the use of the most recent observations is suitable. Therefore in our study we use the fixed lag smoother approach (see Chapter 2 for the theoretical background). The results that have been obtained by using a fixed lag smoother with a window of three days are shown in the Figure 3.9. The correction achieved for the emission factors using the fixed lag ensemble smoother is around 15% showing that the sulphur dioxide emissions have daily persistent behavior. The regions that are affected by the use of smoother are similar when compared with the filter (Figure 3.8, left side). Note that the adjustments of the smoother on the emission estimates points on the same direction as filter suggesting that the influence of the emissions is amplified in time. Therefore the convergence of the smoother algorithm is expected to be reached by using more ensemble members. The same behavior is noticed on the reaction rate with amplified corrections in the central part of the European domain as well as in the Iberian peninsula.

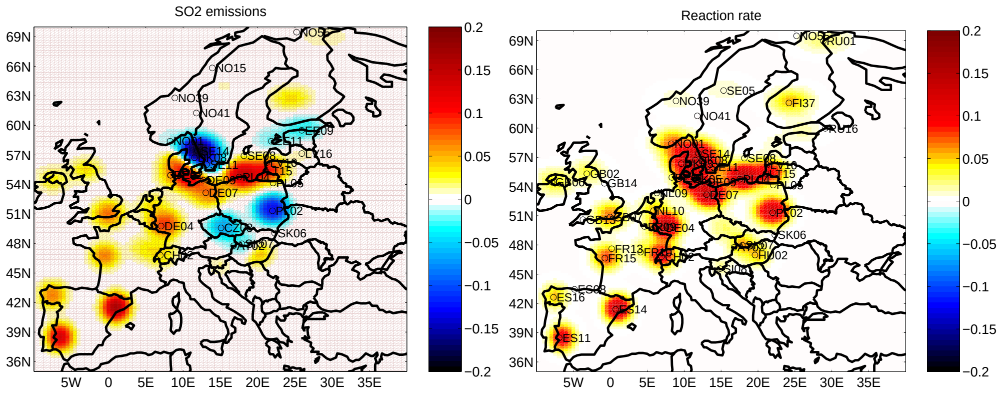

Figure 3.9: The difference between 3 lag length smoother and filter results for the emission (left) and reaction rate (right) factors by using 100 km localization distance.

Additional experiments have been performed by using a localization distance set to 350 km. Figure 3.10 illustrates the difference between the smoother and filter for the two parameters considered in this study when 15 ensembles are used. The left side figure suggests that the influence of the assimilation when an ensemble smoother is used with a larger distance overrules the effects of the assimilation. It may be an indication that several stations with very high values of sulphur dioxide measurements should be screened before performing an assimilation experiment. Also, our conclusion concerning the overestimation of shipping emissions in the North Sea is confirmed. The higher and overall reduction of emissions results in a lower sulphate concentrations with strong impact on the estimation of the reaction rate (3.10, right side).

This sensitivity experiment suggests that a smoother algorithm that localizes the analysis in time combined with a larger localization distance may produce unrealistic results if the ensemble size is not large enough to capture the characteristics of the system. The quality of smoothed estimates is determined by how accurate the smoother computes the covariance matrix those structure is imposed by the model uncertainty and model physics. Figure 3.11 shows the difference between the smoother and filter results for the emission (left) and reaction rate (right) factors by using 350 km localization distance and 45 ensemble members. The smoother tends to lower the bias by using both sources of uncertainty. The general features of emissions distributions (reduction in central Europe and amplification in the western and south-western parts of the domain) are similar with those obtained in Figure 3.9. When compared with Figure 3.10, the impact on the reduction of emissions is decreased in the central Europe. The reaction rate is significantly increased that results in producing more sulphate and reducing the bias.

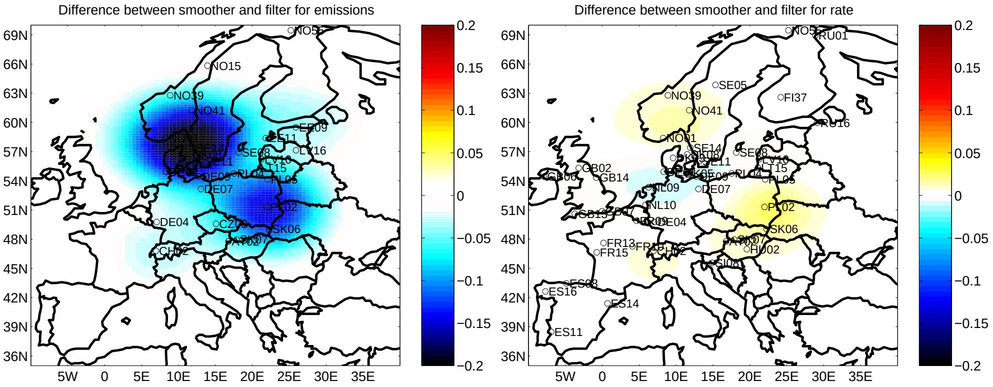

Figure 3.10: The difference between 3 lag length smoother and filter results for the emission (left) and reaction rate (right) factors by using 350 km localization distance and 15 ensembles.

3.7 Conclusions and discussion

In this chapter a sequential data assimilation scheme has been applied to a sulphur cycle version of the LOTOS-EUROS model using ground-based

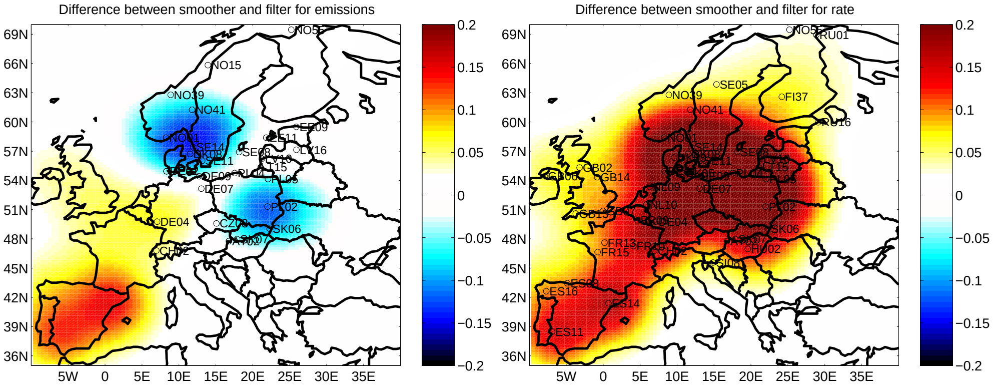

Figure 3.11: The difference between 3 lag length smoother and filter results for the emission (left) and reaction rate (right) factors by using 350 km localization distance and 45 ensembles.

observations derived from the EMEP database for 2003. Our goals were to construct a stochastic environment around the model for estimating the concentrations of two closely related chemical components, sulphur dioxide and sulphate, and to gain insight into the behavior of the assimilation system for a multi-component setup in contrast to a single component experiment.

Comparison of the deterministic model with measured data showed that the annual average modeled sulphate concentration is systematically underestimated. For sulphur dioxide the model generally overestimates the annual mean concentrations slightly, but large overestimations were observed for stations in the south-eastern part of Poland and near the coast in Denmark and Sweden.

Extensive simulations with the ensemble Kalman filter show that two issues are crucial for the assimilation performance: the available observation data, and the choice of stochastic parameters. Concerning the first, the number of measurement sites and the density of measurement network play an important role. Due to inhomogeneous density of data corroborated with the localization effect imposed by the ensemble-based filter, the influence of assimilation on model state and parameters is limited to certain areas. Hence, by including a more extensive set of monitoring data the spatial impact of the assimilation can be improved. Future studies dedicated to provide accurate concentration maps over the whole domain should encompass a larger number of observation sites. A more extensive data set could be obtained using data from national networks in combination with rigorous quality assurance to ensure the use of a data set representative for the model resolution. Although a larger data set would improve the results with respect to the estimated concentration fields in a quantitative sense, our conclusions about the use of a multi-component strategy are not expected to change. This is illustrated by the fact that the improvement in the performance by the multi-component strategy is observed at both assimilation and validation stations.

Concerning the second issue, the assimilation of multiple species is advised in case of large model uncertainties which can be attributed to several causes. For example, the single component experiments with assimilation of either SO_2 or SO_4 give contradicting directions in which the emissions are adjusted. The multi-component analysis forced us to redefine the model error. We could demonstrate that with a more accurate description of the model error using two noisy parameters (emissions and reaction rate) instead of one (emissions) the multi-component assimilation performs better. The experiments shows that the filter technique is able to correct the model error, if the uncertain model parameters are specified correctly.

To acquire a best possible estimate of air pollutant emissions many studies, including ours, focus on the estimation of the emission strength. We found that the direction in which the emissions are adjusted depends strongly on the setup of our assimilation experiments. We feel that the emission estimates presented here are influenced by the fact that they are used to compensate for other systematic errors and uncertainties in the model description. Consequently, even for the relatively simple system of SO_2 and SO_4, our modeling capabilities should be strengthened before reliable parameter estimations can be obtained by using data assimilation. On the other hand, specific features of our results may be interpretable. For example, the system significantly adjusts the emissions around the stations in the northern coastal regions. There the impact of the local shipping emissions is very large and our results indicate that these emissions could be too high. The corrections provided by the smoother over the filter lay between 14 and 20%.

Additional experiments have been performed using different specifications for the definition of the stochastic model to investigate the impact on the analyzed concentration fields. These experiments used a different magnitude of the noise in the emissions and the reaction rate, a larger time correlation or a larger influence zone for the localization scheme. In all these experiments no important improvements have been obtained compared to the results presented above. This suggests that the remaining biases should be attributed to other sources of uncertainty than those in the configuration of our (stochastic) model.

The underestimation of sulphate fields may be explained partially by a too slow conversion rate of the sulphur dioxide into its product. The modeling of the conversion rate as noisy process helps the filter to reduce the bias because it provides a more accurate description of the model error and enlarges the ensemble spread which allows the sulphate measurements to have more impact. However, the bias in sulphate concentrations is not lowered very effectively. Uncertainties in the removal processes could play a role as well. For example, our dry deposition scheme provides rather high removal rates for particles [37]. Based on the result of our study we are investigating the sensitivity to the choice of deposition scheme. Also, non-inventory sources of sulphur dioxide may regionally explain a part of the underestimation of sulphate concentrations. During the summer of 2003 huge forest areas were burnt at the Iberian peninsula [59], which may have significantly contributed to the underestimation of the sulphate levels there.

In short, we have shown that one should move from single component applications of data assimilation to multi-component applications. The increased complexity associated with this move requires a very careful specification of the multi-component experiment, which will be the main challenge for the future.

CHAPTER 4

Bias aware assimilation of OMI NO_2 tropospheric columns

Abstract: This study aims to investigate the impact of nitrogen dioxide tropospheric columns provided by the Ozone Monitoring Instrument on the chemical transport model LOTOS-EUROS. A large discrepancy between the simulated NO_2 columns and those derived from satellite measurements has been found. The study comprises assimilation procedure that takes into account a bias correction scheme. By applying an ensemble Kalman filter to a stochastic version of the model, the bias is shown to be reduced significantly. The impact of assimilating OMI NO_2 data is evaluating by the response of the modeled ozone concentrations.

4.1 Introduction

Interest on nitrogen oxides (NOx) in the atmosphere is related to the influence of the NO_2 compound on several gases of high importance as ozone (O_3) and hydroxyl (OH). Nitrogen oxide sources are mainly anthropogenic with the automobile, domestic heating and electric power production being the most important sources. NOx, which is the sum of nitric oxide (NO) and nitrogen dioxide (NO_2), is primarily emitted in the form of NO. Oxidation by ozone quickly forms NO_2 which is converted back to NO by photolytic decay. A photochemical equilibrium between NO and NO_2 is reached within minutes. At some distance from the immediate source, the bulk of planetary boundary layer NOx is constituted of NO_2. A good knowledge of the spatial distribution of NO_2 is important for assessing the air quality over a given domain.

This chapter is a slightly revised version of [4].

Since several years, satellite missions dedicated to the tropospheric sounding have been developed and are now operational (IASI, GOME2 aboard Metop in 2007; TES, OMI aboard EOS-AURA in 2004, SCIAMACHY on Envisat in 2002). Satellite observations are starting to be used for the understanding and monitoring of air quality at large scale and need to be exploited.

A number of studies focused on comparison between satellite and modeled NO_2 columns found that the forecast bias is a commune characteristic of different models. As NO_2 has a relatively short lifetime, the description of its chemical production and loss and the distribution and magnitude of sources are of significance. Negative or positive bias is explained by several causes as overestimation of dry deposition, treatment of vertical mixing and limited chemistry scheme [93]. In a paper by [8] the cloud free SCIAMACHY observations were compared with surface measurements and simulations over Western Europe performed with the regional air-quality model CHIMERE showing that the model underestimates surface NO_2 concentrations for urban and suburban stations. A comparison study by [76] has concluded that the overestimation of the tropospheric NO_2 columns retrieved from GOME with a factor of 2 to 3 in a global circulation model (GCM) is most likely due to the missing NOx sink process.

In order to apply a data assimilation scheme, the systematic bias should not be neglected since it has been shown by [28] that a biased forecast causes always a biased update. The bias problem is not characteristic to the atmospheric data assimilation only, but also to ocean and land data assimilation. Many authors have addressed this issue with a focus on bias estimation and correction, e.g. [29] for assimilation of humidity data in the Goddard Earth Observing System.

Many studies involving CTM evaluations indicate the presence of persistent bias between observations and the simulated fields of chemical species, but only few of them take into account a bias aware data assimilation scheme. A study by [74] applied the methodology of separate bias estimation for an unbiased assimilation of carbon monoxide retrievals. Also [5] treated the problem of a bias for sulphate and its precursor gas in a multi-component data assimilation procedure by using a stochastic model to decrease the errors in the modeling system.

This study is based on the combination of satellite NO_2 data provided by the Ozone Monitoring Instrument (OMI) and numerical simulations of the atmosphere performed with the regional CTM LOTOS-EUROS. Comparison of the simulated NO_2 columns and those derived from satellite measurements have been performed to validate the quality of representation of major oxidation and transport processes which may play an important role in variability of other photo-oxidants and their precursors. The major problem is the large discrepancy between the modeled and OMI observed NO_2 values. Therefore the main aims of this chapter is to bring a better understanding of modeling system in the presence of persistent bias. To this end a bias aware data assimilation scheme has been applied.

This chapter addresses these issues as in the following overview. In Section 4.2 the OMI satellite data is presented, followed in the next section by a description of the LOTOS-EUROS model used in this study. Section 4.4 summarizes and discusses the results of the comparison between the model simulation and retrieved data. The methodology for including the bias parameter in a sequential data assimilation scheme is described in Section 4.5 and the results of applying this procedure to our settings are examined in Section 4.6. the last section gives the concluding remarks.

4.2 Satellite observations

The Dutch-Finish Ozone Monitoring Instrument (OMI) on board NASA's EOS Aura satellite that was launched on 15 July 2004 is a nadir viewing imaging spectrograph. The OMI's wide field of view which corresponds to a 2600 km wide spatial swath on the Earth's surface allows for observing the air pollution sources with daily global coverage. There are Aura overpasses over Europe once or twice per day; the average overpasses time is 12:45 hours (GMT). OMI measures NO_2 column with pixels of 13 km (along track) by 24 km (cross track) at nadir and 13 by 28 km^2 for the outermost swath angles.

In this study the OMI NO_2 tropospheric columns data is provided by the Royal Netherlands Meteorological Institute (KNMI) in the framework of the SMOGPROG project. Data sets are available on [64]. The preprocessed data contains the observed slant column NO_2, the total vertical column NO_2 and the estimated tropospheric portion of the total column. The algorithm for the retrieval of total column and tropospheric NO_2 from OMI data is described by [9] and [13].

Validation efforts of OMI product have been reported in a number of studies by comparing the retrieved nitrogen dioxide column with ground-based remote sensing measurements ([19]), with aircraft measurements ([10], [14]) and with in situ NO_2 measurements ([75]). A complex study has been performed by [11] in the Dandelions experiment. Retrievals of tropospheric NO_2 column densities from OMI and SCIAMACHY have been compared with ground-based and balloon sonde datasets. This validation study has been performed during two campaigns at the Cabauw Experimental Site for Atmospheric Research. The datasets on NO_2 and ozone are available at Aura Validation Data Center. The results showed good agreement between average OMI and ground nitrogen dioxide concentrations at Cabauw site and in the vicinity. Comparison with RIVM in situ data, and with CHIMERE model have been performed too.

4.3 The version v1.5 of the LOTOS-EUROS model

A regional model of tropospheric chemistry is an important source of information for the vertical distribution of NO_2 as the model simulations are performed to match the time and the location of the satellite information. In this section we give an overview of the LOTOS-EUROS modeling system, version v1.5 used in the present study. While the model domain is designed to cover Europe, in this chapter we use a smaller domain configuration with a horizontal resolution of $0.5° \times 0.25°$, which is approximately 25×25 km over the Netherlands. In the vertical there are three dynamic layers and a surface layer with a fixed depth of 25 m. The lowest dynamic layer is the mixing layer, followed by two reservoir layers. The height of the mixing layer is part of the meteorological input data. The heights of the reservoir layers are determined by the difference between the mixing layer height and the top of the model at 3.5 km above sea level. The transport consists of advection in 3 dimensions, horizontal and vertical diffusion. In this study the advection is driven by meteorological data produced at the the European Center for Medium-range Weather Forecasts (ECMWF). Boundary concentrations for our configuration setup are obtained with the climatological global TM5 that runs for 2006 (see [84]) on resolutions of both $6°\times4°$ and $3° \times 2°$ over Europe up to 60N latitude. Within this study anthropogenic emissions of NO_x have been taken from the new TNO emission database (PAREST project), which is available at a resolution of 0.25° by 0.125 ° for the year 2005 (see [111]). The temporal resolution of the model run is 20 minutes. For a more detailed description of the model including chemical mechanism, emissions and latest developments we refer to [95].

4.4 Model validation

The changes in pollution distribution revealed by OMI NO_2 retrievals are compared to the LOTOS-EUROS simulations to understand their general consistency and differences. We have chosen July 2006 as test period, since it is characterized by a number of episodes with high ozone concentrations and contains a reasonable amount of cloud free days and areas. High temperatures has been measured in this period; on most days it was warmer than 25° and on several days warmer than 30° with weak southerly or easterly winds. Three episodes of enhanced ozone level can be distinguished roughly: from July, 1 to 6, from 17 to 20 and from 25 to 28.

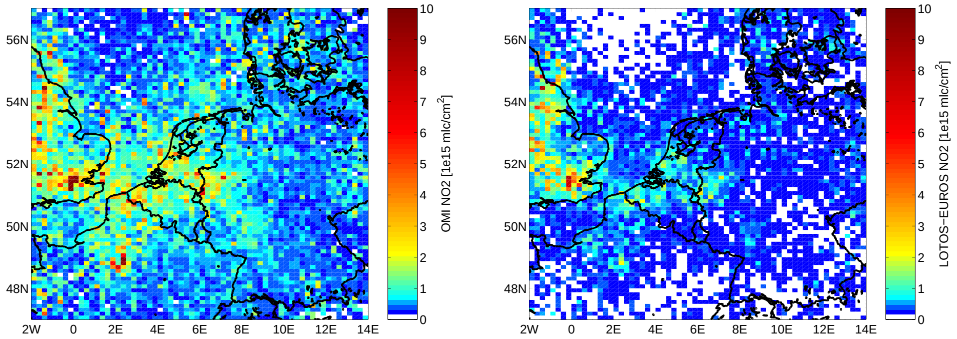

Figure 4.1: Comparison between monthly averaged NO_2 tropospheric columns (10^{15} molecules/cm^2) from OMI (left) and LOTOS-EUROS simulation (right).

The satellite data resolution is 0.25° by 0.25 °. Since the model resolution is larger in latitude, the available OMI data from the same orbit are collected and averaged in each model grid to eliminate the effect of resolution difference between model and retrievals. The model simulations are linearly interpolated in time and space to produce vertical nitrogen dioxide profiles to the center of the OMI pixel. The averaging kernel approach is applied to the model profile in order to be compared with satellite retrieved total column. The model prediction of the retrieved column reduces the errors associated with a priori profile shape assumption needed in the retrieval ([39]). In order to be consistent with satellite retrievals, the simulated NO_2 tropospheric columns were analyzed at two daily passing times around 11h45 and 13h15 GMT respectively, for every grid cell and collocated with OMI measurements. The daily NO_2 columns are used to derived the distributions along the whole month. Cloudy pixels are not used for calculating the tropospheric NO_2 column.

While the observed mean calculated over the whole month and domain is 0.82, the simulated average given by the model shows a lower value of 0.36. This results in a negative bias of -0.46 (see Table 4.1, section 4.6). The spatio-temporal correlation is about 0.76 suggesting a high consistency between OMI data and model simulated columns.

Monthly averages of modeled and OMI NO_2 tropospheric columns are depicted in Figure 4.1, left side. The NO_2 spatial distributions shows a good agreement between OMI and LOTOS-EUROS with qualitatively similar patterns associated to higher and lower NO_2 values. Strong consistency has been found over England. The model is able to capture the main spots around the polluted regions (Ruhr industrial area, Benelux) which indicates that the

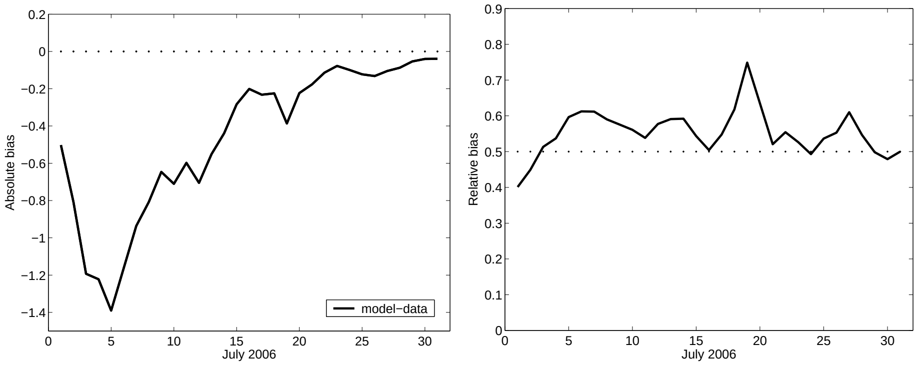

Figure 4.2: Modeled minus observed NO_2 tropospheric columns (10^{15} molecules/cm^2) (left) and relative bias given in percents (right) as a function of time.

anthropogenic sources are correctly located. Note that the white pixels depicted on the top and bottom of this figure represent very low values of the concentration field.

The investigation of the background residuals evolving in time (Figure 4.2, left panel) shows that the model exhibits a persistent negative bias over the whole month. As temporal function the absolute bias is calculated by averaging the differences between model and data for each cells. The first period of July up to six days is characterized by the highest discrepancy and corresponds to the one of the cloud-free period of the month (see Figure 4.3), but the second sunny period (between days 17 and 20) is characterized by much lower bias values corresponding to a lower concentration field measured by OMI. This change in air pollution distribution can be caused by the meteorological input as lower temperature and different wind field. The relative bias (Figure 4.2, right panel) indicates also that the bias is persistent for the second half of July when the measured concentrations are low.

The lower modeled NO_2 values relative to OMI measurements can be explained by several causes that contribute together to the difference between the model and retrieved datasets. As the main source of nitrogen oxides is from the surface emissions, they play an important role in determining the modeled concentrations. The good correlation between the model and data and the correct source locations indicate that the man-made emission inventory used in this study is unlike to be the major cause of the discrepancy. The commonly used 30% of uncertainty in anthropogenic emissions cannot explain the large model underestimation (around 54%). Boundary conditions could play an important role in explaining the underestimation of the modeled

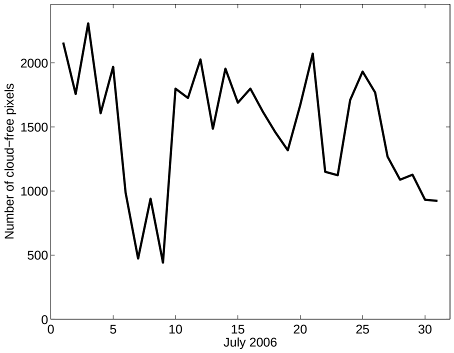

Figure 4.3: Number of cloud-free pixels as a function of time.

concentrations, but is unlike to be considered as the main source of the bias.

Cloud effect is also not an important source of persistent bias since the daily differences between the modeled and observed values does not depend on cloud coverage (as results by comparing Figure 4.2 with Figure 4.3).

Since the long term transport is not important for nitrogen dioxide concentrations at surface, this process has been not considered as a major cause of incorrect modeled concentrations. Also the main physical process, namely vertical mixing, that may largely influence the NO_2 distribution cannot be detected due to the specific nature of data measured as column. Therefore the limitations in the chemistry scheme due to the incorrect chemical lifetime of NO_2 is clearly considered as the major issue. The low modeled tropospheric NO_2 might be explained if the nitrogen oxide lifetime is assumed to be reduced due to relatively high level of OH or N_2O_5 hydrolysis rate.

4.5 State and bias estimation

Although data assimilation rely on the assumption of an unbiased model, incorrect physical parameterizations, boundary conditions and uncertain model inputs can cause the presence of a persistent bias in the forecast model. By assuming that the observations are unbiased, different methods for on-line forecast bias estimation have been addressed. The procedure of augmenting the state vector by adding uncertain parameters that represent the model error or bias term [66] has been often used in data assimilation framework. Another technique designed for the bias estimation in the updating process is the two-stage estimation approach derived in the context of estimation theory [46] and from the Bayesian perspective (e.g. [77]). The bias estimation is treated separately from the computation of a bias-blind estimate of the state.

The methodology for an unbiased assimilation adopted in this chapter is the colored noise model obtained by augmenting the vector state with a bias term that is modeled as a colored noise process. This procedure has been compared with the separate-bias estimation by [31] in a hydrological application.

We start with the description of the deterministic LOTOS-EUROS model as in the following equation:

$$c^f(k+1) = M_{det}(c^a(k)) \tag{4.1}$$

Here, the state space operator of the LOTOS-EUROS model is denoted by M_{det}. This operator computes the concentration vector c, which contains all considered components for each grid cells, at time k hour given the concentration at time k−1. A stochastic representation of the model error is defined by extending the state vector c with a correlated noise processes for NO_2 and VOC emissions, deposition velocities and boundary conditions. The new augmented state vector is written in terms of the new operator M_{stocha} and the forcing term Gw as in the following:

$$x(k+1) = M_{stocha}(x(k)) + Gw(k) \qquad w \sim N(0,1). \tag{4.2}$$

The measurements have white Gaussian errors v with covariance denoted by R:

$$y(k) = H(k)c(k) + v(k), \quad v \sim N(0,R). \tag{4.3}$$

The observation operator H consists of the horizontal interpolation to the observation location, vertical interpolation to the OMI retrieval grid and weighted average using the averaging kernel of the model field to the OMI a priori profile.

The forecast bias represents the expectation of forecast error. The forecast and observational errors being defined, a state-space description of the forecast bias in the presence of an existing bias-free observational system is needed. The persistent model has been considered for propagating the forecast bias in the original bias correction algorithm proposed by [28]. Here we proposed a model for the time evolution of the bias. Then instead of cycling with Equation 4.2 that is an integration of a biased forecast model, we consider the modified version of the model according to:

$$x(k+1) = M_{stocha}(x(k)) - F\beta(k) + Gw(k),$$

where β represents the bias variable written in a scalar form as following:

$$\beta(k) = a\beta(k-1) + \sigma\sqrt{1-a^2}u(k-1), \qquad u \sim N(0,1),$$

The influence of the bias on the evolution of the state is introduced into the model propagation by using a feedback measure F. The matrix F is set as a function of the background state by taking non-negative values for the nitrogen dioxide concentrations and zero for the rest components of the state. The non negative values are calculated by considering a uniform bias on horizontal scale and a proportional dependence between the bias parameter and the vertical distributions of the NO_2 concentrations.

The coefficients a for each state variable represent the time correlation parameter. Setting a to zero, we obtain a white noise sequence with zero mean and variance σ. The case $a = 1$ is equivalent to the persistent model used in separate-bias estimation approach, but without any random component.

The Ensemble Kalman Filter (EnKF) ([40], [61] is applied to the concatenated stochastic model with the bias parameter:

$$z = \begin{bmatrix} x \\ \beta \end{bmatrix}. \tag{4.4}$$

An ensemble of N states realizations, $\xi_1, \xi_2, \ldots \xi_N$ assumed to be a sample out of a distribution of the true state is propagated through the augmented model. The model state is represented by the predicted ensemble mean z^f:

$$z^f = \frac{1}{N} \sum_{j=1}^{N} \xi_j^f. \tag{4.5}$$

The forecast error covariance matrix P^f is assumed to be carried by the ensemble of perturbations L^f:

$$L^f = \left[\xi_1^f - z^f, \ldots, \xi_N^f - z^f \right], \tag{4.6}$$

$$P^f = \frac{1}{N-1} L^f {L^f}^\top. \tag{4.7}$$

When the measurements become available, the ensemble replicates are updated:

$$\xi_j^a = \xi_j^f + K * \left[y - H\xi_j^f + v_j \right], \tag{4.8}$$

where v_j represent the realizations of the white noise processes v and K is the Kalman gain:

$$K \quad = \quad P^f H^T \left[HP^f H^T + R \right]^{-1}. \tag{4.9}$$

	Simulation	Bias aware assimilation
calculated mean	0.36	0.57
absolute bias	-0.46	-0.25
rms error	0.97	0.79
spatial-temporal correlation	0.76	0.83

Table 4.1: Statistical comparison of modeled and assimilated NO_2 concentrations in the bias assimilation experiment. All statistics have been averaged over all grid cells.

4.6 Assimilation results and discussions

The EnKF algorithm with 15 ensemble members has been used for the assimilation of OMI tropospheric columns over one summer month July 2006. It is assumed that the uncertainty in the measurements is defined using a fixed standard deviation of 10%. The stochastic version of the model has been built by considering four uncertain parameters, namely NOx and VOC emissions, ozone top boundary condition and deposition velocities. The parameters have been modeled by using correlated noise factors. The standard deviation of the colored noise process for all noise factors is set to 30% and the time correlation parameter is set to 1 day. Since the nitrogen dioxide lifetime parameter is produced by different processes implicitly, the bias cannot be reduced by simply tuning this model parameter. The major question what we should do with systematic model error forces us to include a bias correction scheme into the assimilation procedure. This scheme that consists of augmenting the model state with the uncertain bias parameter distributes the bias according to the vertical profile of nitrogen dioxide concentrations. The modeled uncertainties by using a colored noise process are set to the same factors as the other three parameters of the model. Table 4.1 summarizes the statistical results averaged over July and our domain for the model output and assimilation experiment. The observed mean is 0.82 while the calculated average from the model output shows a lower value of 0.36. The root mean square (rms) error is about 0.97. By applying the bias aware assimilation of NO_2 data the ensemble Kalman filter is able to compensate this bias with 45%. The rms error is reduced with about 19%.

The monthly averages of NO_2 tropospheric column resulting after assimilation are illustrated in Figure 4.4, left panel. One notes an increased concentration level over the regions with very low simulated NO_2 values, e.g. over the southern part of our domain in France and Germany and over the coastal regions. The impact of the local shipping emissions is probably larger than has been estimated and our results indicate that these emissions could be too

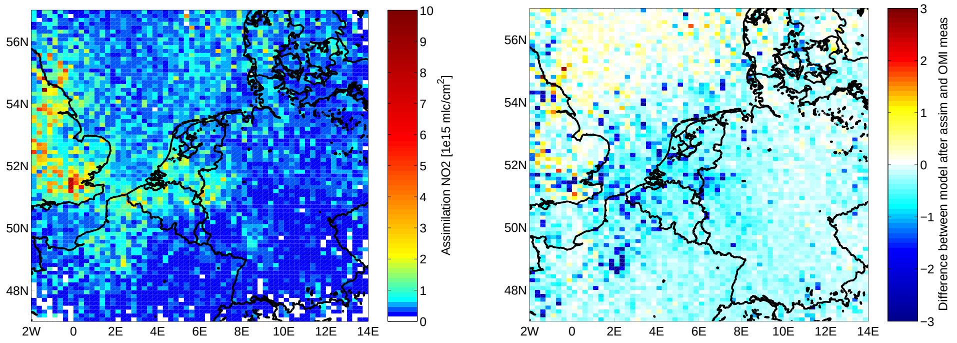
56N
54N
52N
50N
48N
2W
0
2E
4E
6E
8E
10E
12E
14E
0
1
2
3
4
5
6
7
8
9
10
Assimilation NO2 [1e15 mlc/cm2]
−3
−2
−1
0
1
2
3
Difference between model after assim and OMI meas

Figure 4.4: Monthly average of simulated NO_2 columns (10^{15} molecules/cm^2) (left) and the difference between the model and data after assimilation (right).

low. On the right panel it is shown the difference between the assimilation output and measurements averaged over the whole month. Over large parts of our domain a negative bias still remains, even after assimilation is performed.

Figure 4.5 shows the observed, simulated and assimilated NO_2 tropospheric columns evolving in time. The assimilation has a positive impact on reduction of the bias over the period. The benefit of using the assimilation scheme is significant, especially for the fist half of the month.

The assimilation performance on improved NO_2 concentrations should be confronted with the results obtained for the ozone component as the two species are strongly related. One month period used in this chapter is not representative for an averaged summer, but from operational perspective it may be especially relevant to forecast high ozone values correctly. Figure 4.6 illustrates the changes on the monthly averaged ozone concentrations modeled by LOTOS-EUROS as they result from assimilation of OMI tropospheric NO_2. An overall increased O_3 concentrations of about 8% has been found. One notes the shipping track that is visible on the top of the this figure. One may see here that the influence of boundary conditions on the results is not very important.

From other studies it is known that the current model overestimates the ozone mean and underestimates the higher ozone episodes. Previous simulations have shown a rather flat model behavior with problems in description of the variability of ozone. Time-series of O_3 concentrations given by the model before and after the assimilation are compared with ozone ground measurements at Cabaw rural site in the Netherlands in Figure 4.7. The monitoring data is collected from the national air quality network as operated by RIVM. On the left panel the daily mean concentrations of ozone before and after

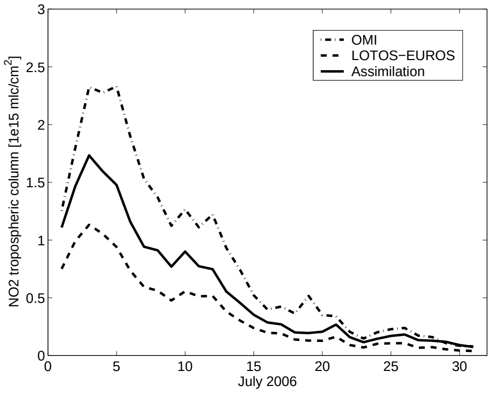

Figure 4.5: Observed, modeled and assimilated NO_2 tropospheric columns (10^{15} molecules/cm^2).

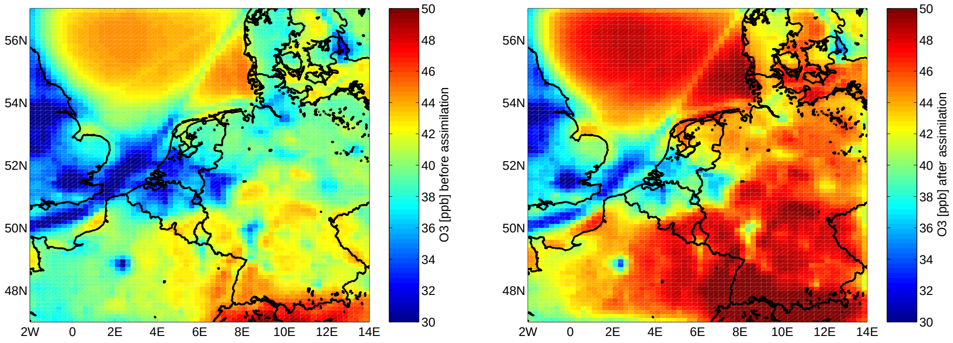

Figure 4.6: Comparison between monthly averaged O_3 (10^{15} molecules/cm^2) before (left) and after (right) assimilation of OMI NO_2 tropospheric columns, respectively.

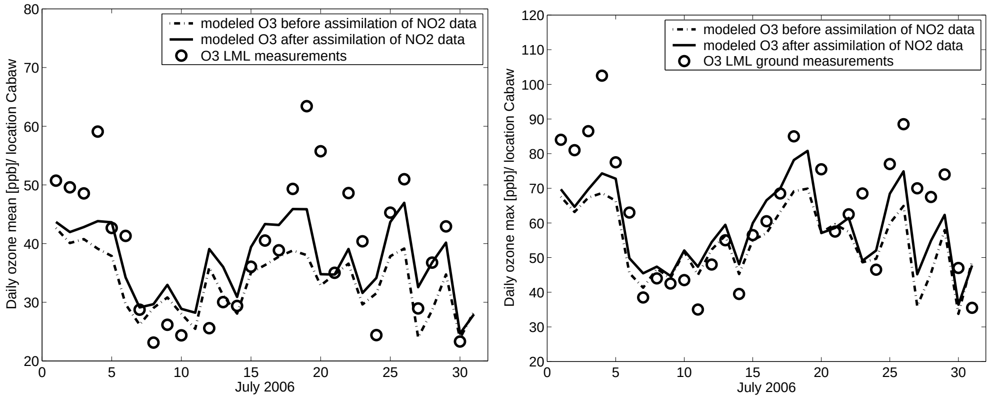

Figure 4.7: Daily averaged ozone (ppb) (left) and maximum ozone (ppb) (right) at Cabaw station.

assimilating nitrogen dioxide derived from OMI is depicted. Here in situ measurements are represented by the circles. In the first six days, the negative bias in the modeled O_3 concentrations is slightly reduced by the assimilation. The next period up to July 14 is characterized by the ability of the system to capture the low O_3 concentrations. In the last period of the month (15-31 July) the underestimation of the observed concentrations is larger. We have noticed an increase of O_3 by using the assimilation of OMI NO_2 data, especially in the period with high ozone episodes.

The model overestimates low ozone daily maxima, whereas often underestimates the peaks significantly. On average, the daily ozone maximum is increased by the assimilation of OMI NO_2 with 10%. This increase can be noticed also at Cabaw station (right panel of Figure 4.7). The model cannot capture the peaks for the first episode with enhanced level of ozone pollution, even after assimilation. For the remaining two periods with high ozone values, the assimilation can better reach the extremes, especially for the second peak (between 17 and 20 July).

4.7 Conclusions

A non-zero mean of background residuals or innovations indicates the presence of a systematic bias in the observations and/or the model giving the information of the combined effect of the measurements and model. One cannot attribute the discrepancy to the observed or modeled system without any additional information. In order to separate the two sources of bias and detect the main bias source, spatial and temporal information has been analyzed.

A number of studies including ours that made a comparison between satellite and modeled NO_2 columns found that the forecast bias is a commune characteristic of different models. The spatial pattern of nitrogen dioxide given by LOTOS-EUROS has been shown to be in a good agreement with the OMI dataset, but having a large negative bias for the region selected in this study. The main sources of this bias have been found to be the limited chemistry scheme due to the incorrect NO_2 lifetime parameter.

Since a biased forecast causes always a biased update, the systematic bias should be included in a successful data assimilation scheme. In our approach the bias problem is treated by considering the bias variables as part of the stochastic model state. The ensemble Kalman filter applied to a stochastic versions of LOTOS-EUROS is directed to assimilation of NO_2 tropospheric column by using OMI data for July 2006. It has been shown that by talking into account a more realistic description of the stochastic version of the model, the EnKF algorithm is able to reduce the bias by almost 50%.

However, the remaining bias could be a result of the other incorrect chemical processes in the model or missing sink of nitrogen oxides. An unified study of comparing several models with OMI NO_2 tropospheric column could be a meaningful methodology for improving the LOTOS-EUROS system. It would be useful also to extend the model validation to a longer time period for a complete seasonal analysis. Such an intercomparison is currently undertaken in order to obtain insight in the differences in modelled and observed nitrogen dioxide and resulting changes in ozone and OH concentrations and their possible impact on climate.

A multi-component assimilation scheme in which nitrogen dioxide data provided by OMI and ground-based ozone measurements are simultaneously incorporated into the model could contribute to an increased performance of the system and to accurate predictions of air quality.

CHAPTER 5

Square root filters

Abstract: In order to avoid an underestimation of the error covariance and possibly filter divergence of the Ensemble Kalman Filter (EnKF) random noise is added to the observations during assimilation. Although it was shown that this procedure results in an unbiased estimate of the error covariance for large ensemble sizes, this procedure also introduces additional sampling errors. To eliminate these sampling errors a number of Ensemble Square Root Filters (ESRFs) have been proposed. In a number of recent papers it was shown that some versions of the ESRF type introduce a bias. Moreover, it was shown that there exists a symmetric version, that is unique and unbiased. Finally all other unbiased ESRFs can be generated from the symmetric solution by multiplication with a unitary matrix that has the vector with all ones as an eigenvector.

For a large number of models the system noise or model uncertainty plays an important role in data assimilation. In many of the papers applying ESRFs, the system noise is either treated in a simplified form by using covariance inflation or introduced by random perturbations as in the Ensemble Kalman Filter.

In this chapter we propose a method for handling the model error in a manner similar to the ESRF analysis. This method is based on earlier work by Heemink and Verlaan and their Reduced Rank Square Root (RRSQRT) algorithm. Here, we will emphasize the similarities with the ESRF analysis and extend the work on the symmetric version for ESRF analysis to handling the system noise. It will be shown that the symmetric version of the ESRF and RRSQRT filter introduces the smallest increments of the state. For a non-linear model small increments have the advantage of introducing fewer problems in the model, e.g. an unbalanced state, negative layer thickness or negative concentrations. Finally, the experiments in this chapter indicate that this algorithm can be more accurate than the EnKF, ESRF or original

RRSQRT filter.

5.1 Introduction

Square Root Filters (SRF) are a powerful tool that allow the Kalman Filter approach to be applied to large scale data assimilation applications. We classify the SRF into three subclasses as the cycle of alternating forecast and analysis steps are performed: stochastic SRF, where both steps are influenced by sampling for each ensemble member using a pseudo-random generator, semi-deterministic SRF, where only the forecast step is based on sampling, and deterministic SRF, where both steps of the assimilation cycle are deterministic. The first two classes contain the ensemble-based filters. An example of a deterministic SRF is given by the Reduced Rank Square Root (RRSQRT) filter introduced by [110]. Also an Ensemble Square Root filter where the model error is prescribed by covariance inflation (as e.g. used by [91]), can be considered semi-deterministic.

The standard Ensemble Kalman Filter (EnKF) introduced by [40] is a Monte Carlo approach to the Kalman Filter. It is based on the representation of the probability density of the state estimate by a finite number of randomly generated system states and is able to handle the nonlinearities of the models. In the first version of the EnKF, all ensemble members were updated with the same observations and as consequence, the analyzed covariance matrix was systematically underestimated. Therefore, now an independent set of perturbed observations obtained by adding random noise to the actual measurements is used in the analysis step to prevent the collapse of the ensemble [17].

An alternative way to solve the update step of the ensemble-based filter is represented by the deterministic analysis that is not sensitive to the observational sampling errors associated with the use of perturbed observations. This method referred to as ensemble square root filter (ESRF) is placed in a unified framework by [104] and [79]. Also, it allows for the classification of existing semi-deterministic filters. Several authors [79, 91] restricted the class of all possible solutions of the square root decomposition of the analysis error covariance to those that do not introduce a systematic bias in the updated ensemble mean. Here we will refer to these algorithms as unbiased ensemble square root filter (UESRF).

The goal of this chapter is twofold. The first is to investigate the performance of several square root filters applied to two simplified cases. The background is that earlier papers, e.g. [91], suggested that the symmetric ESRF

This chapter is a slightly revised version of [6].

and the version with mean-preserving random rotations have similar properties, although [88], [91] note that the symmetric version has the property that the analysis increments are smallest with respect to P^f and P^a norms. In their experiments the two unbiased algorithms exhibit a similar performance. However, an experiment with a simple 2D pollution model shows that this is not always the case. The symmetric ESRF results in better estimates of the state than the ESRF with mean-preserving random rotations. This is shown to be related to a property of the symmetric ESRF that introduces the smallest analysis increments for an arbitrary compatible norm. These results, together with earlier results of [91], [79] suggest that the symmetric ESRF is likely to be most accurate ESRF for most applications.

Secondly, this study considers the introduction of errors due to sampling of system noise. The model error may be of high importance for many models, such as atmospheric chemistry models (with uncertain pollution sources), hydrological models or models used in reservoir engineering that are nonlinear, but stable. Contrary to meteorological or oceanography systems that are highly unstable and for which the application of a covariance inflation factor represents the solution of preventing the filter divergence, for the other models covariance inflation will often be too simple approach.

It is shown that for a simple 2D pollution model, these sampling errors for the system noise significantly affect the assimilation. New algorithms based on the RRSQRT filter and similarities with the ESRFs are proposed. The symmetric version of the RRSQRT is shown to work quite well for the two examples in this study.

These chapter consists of three topics. Section 5.2 introduces the background of ensemble-based methods followed in section 5.3 by the description of four reduced rank square root filters in a framework similar with that for the ESRF filters. Section 5.5 presents the setup of our experiments with a 2D pollution model and Lorenz 40-variables model. The assimilation results applying the involved algorithms are presented in section 5.6. Finally, the last section contains a summary and conclusions.

5.2 Ensemble-based filters

For application of the filter algorithms to a dynamical model, a stochastic representation should be written in a state-space form according to:

$$x(k+1) \quad = \quad M(x(k)) + w(k). \tag{5.1}$$

The state-space operator M describes the time evolution from the time k to k+1 of the state vector x. The random forcing term w is drawn from the normal distribution $N(0, Q)$ with Q the covariance matrix. The state of the observational network is defined by the observation operator H that maps state variables x to observations y. We further assume that the measurements have white Gaussian errors v with covariance denoted by R:

$$y(k) = H(x(k)) + v(k), \quad v \sim N(0, R). \tag{5.2}$$

5.2.1 The Ensemble Kalman Filter

The Ensemble Kalman Filter as a stochastic method is based on the representation of the probability density of the state estimate in an ensemble of N states, $\xi_1, \xi_2, \ldots \xi_N$. Each ensemble member is assumed to be a single sample out of a distribution of the true state. Whenever necessary, statistical moments are approximated with sample statistics. In the first step of the algorithm an ensemble of N states $\xi^a(0)$ is generated to represent the uncertainty in $x(0)$. In the second step, the *forecast*, the stochastic model propagates the distribution of the true state from the time k to $k+1$:

$$\xi_j^f(k+1) = M(\xi_j^a(k)) + w_j(k), \tag{5.3}$$

$$x^f = \frac{1}{N} \sum_{j=1}^{N} \xi_j^f. \tag{5.4}$$

The ensemble covariance matrix of forecast errors P^f is assumed to be carried at time k by the ensemble of perturbations denoted by L.

$$L^f = \frac{1}{\sqrt{N-1}} \left[\xi_1^f - x^f, \xi_2^f - x^f, \ldots, \xi_N^f - x^f\right], \tag{5.5}$$

$$P^f = L^f \left(L^f\right)^T. \tag{5.6}$$

When the measurements become available, the mean and the covariance are replaced with equivalent ones in the *analysis step* using the ensemble Kalman gain:

$$K = P^f H^T \left(H P^f H^T + R\right)^{-1}, \tag{5.7}$$

The stochastic analysis step defines the standard EnKF with perturbed observations. The analysis ensemble involves the update of each ensemble members the and their ensemble covariance matrix.

$$\xi_j^a = \xi_j^f + K\left[y - H\xi_j^f + v_j\right], \tag{5.8}$$

$$P^a = (I - KH) P^f, \tag{5.9}$$

where the ensemble of state vectors is generated by the realizations w_j and v_j of the white noise processes w and v, respectively.

The EnKF analysis can be written in a square root form [42]:

$$L^a = L^f T^{EnKF}, \tag{5.10}$$

where $P^a = L^a \left(L^a\right)^T$, L^a represents the analysis ensemble perturbations and T depends on the random numbers generated for v_j.

5.2.2 Ensemble Square Root Filters

To reduce the sampling errors introduced by adding random numbers v_j to the observations, several square root type filters (ESRF) were proposed. Following [79] we call an ensemble filter to be semi-deterministic if its analysis step is deterministic. The ESRF approach allows to classify existing semi-deterministic filters. Using the following notations:

$$\begin{aligned} Y &= HL^f, & (5.11)\\ S &= YY^T + R. & (5.12) \end{aligned}$$

the updated covariance matrix becomes:

$$\begin{aligned} P^a &= L^a(L^a)^T \\ L^a(L^a)^T &= L^f \left(I - Y^T S^{-1} Y\right) L^{fT}. & (5.13) \end{aligned}$$

The solution of (5.13) is obtained by:

$$L^a = L^f T, \tag{5.14}$$

where T is a $N \times N$ matrix which satisfies:

$$TT^T = I - Y^T S^{-1} Y\cdot \tag{5.15}$$

It can easily be shown that there is a unique symmetric positive definite solution to (5.15) defined as the square root of the symmetric positive definite matrix from the brackets.

$$T^s = \left[I - Y^T S^{-1} Y\right]^{\frac{1}{2}}, \tag{5.16}$$

By using the eigenvalue decomposition, the matrix T^s has the following form:

$$T^s = C\Lambda^{\frac{1}{2}}C^T. \tag{5.17}$$

Following [91] and [88] we will refer to T^s as the symmetric solution. The symmetric algorithm defined above introduces the smallest analysis increments for an arbitrary compatible norm. This property, proved in the section 5.3, is related to the good performance of the symmetric ESRF obtained by [91] and [88] and in our experiments.

With a formal definition, an ESRF is an ensemble filter in which the analysis ensemble is updated by using an ensemble transform matrix (ETM) T which satisfies (5.15). Consequently, every semi-deterministic filter belongs to the class of the ESRF. In addition, the set of all solutions T which characterize the ESRF class is described in terms of the orthogonal matrix group $O(\mathrm{N})$. Then, a general form of T that satisfies (5.15) is:

$$T \;=\; T^s U, \tag{5.18}$$

where T^s is the symmetric solution and U is an arbitrary orthonormal $N \times N$ matrix. The ESRF method encompasses filters with a ETM which matches the exact analyzed covariance, but the update perturbations could change the ensemble mean. An example is provided by the Ensemble Transform Kalman Filter (ETKF) introduced by [7] whose ETM denoted by T^o is obtained by the multiplication of (5.17) with the orthogonal matrix C:

$$T^o = T^s C = C\Lambda^{\frac{1}{2}} \tag{5.19}$$

We will refer to the solution (5.19) as the one-sided formulation of the ESRF. A random rotation U^r was added to the solution (5.19) to prevent the ETKF tendency of producing high variance outliers [42].

$$T \;=\; T^o U^r \tag{5.20}$$

It has been found by [78] and [79] that the statistics of the update ensemble were still inconsistent with the actual error. A valid analysis ensemble should satisfy the zero-centered condition:

$$L^a \mathbf{1} \;=\; 0, \tag{5.21}$$

where $\mathbf{1}$ is the vector with all elements being 1. Due to the fact that the forecasted ensemble perturbations do not perturb the ensemble mean, a sufficient condition for an analyzed ensemble to preserve the ensemble mean is that the ensemble transform matrix T verifies (up to a scalar constant λ) the following mean-preserving condition:

$$T\mathbf{1} \;=\; \lambda \mathbf{1}, \tag{5.22}$$

We will refer to an ESRF with the ETM satisfying (5.22) as an unbiased ensemble square root filter (UESRF). [91] and [112] have shown that the symmetric transformation does not introduce a bias. Therefore, for providing a mean-preserving solution, it is sufficient to find a rotation matrix U^p such that the vector $\mathbf{1}$ is an eigenvector of U^p. To obtain an arbitrary orthogonal transformation with the desired propriety, we need to construct an orthonormal basis B whose first orthonormal vector is $e_1 = \frac{1}{\sqrt{N}}\mathbf{1}$ by using the Gram-Schmidt procedure. Consequently, the required rotation matrix has the following form:

$$U^p = B \begin{bmatrix} 1 & 0 \\ 0 & U_1 \end{bmatrix} B^T, \tag{5.23}$$

where U_1 is a random $(N-1) \times (N-1)$ orthonormal matrix obtained from the singular value decomposition of a generated pseudo-random matrix. In conclusion, the ESRF with the matrix transformation T^s is unbiased and, if the vector $\mathbf{1}$ is an eigenvector of a rotation matrix U^p, the new transformation matrix written as:

$$T = T^s U^p. \tag{5.24}$$

is a mean-preserving solution. The ETM from (5.24) defines an UESRF algorithm.

5.3 Reduced-rank square root filter

Much of the research for square root filtering has been devoted to the analysis step. For strongly unstable dynamics this can be motivated, but for many applications also the system noise plays an important role. In the EnKF the system noise is added with the introduction of random numbers (5.15), but as in the analysis, this scheme introduces sampling errors.

To avoid the sampling errors, several approaches have been proposed, e.g. the reduced-rank square root (RRSQRT) filter by [110]. The algorithm belongs to the deterministic SRFs (both forecast and analysis are deterministic) methods. It is based on a factorization of the covariance matrix P of the state estimate according to $P = LL'$, where L is a matrix with the N leading eigenvectors l_i (scaled by the square root of the eigenvalues), $i = 1, ..., N$, of P as columns. The algorithm starts with an initial step where the initial covariance matrix is approximated by the leading eigenvectors truncation. The forecast step is represented by an extended matrix $\tilde{L}^f$ with the square root matrix $Q^{1/2}$ of the model error covariance which causes an increased size with each cycle of the algorithm.

$$\tilde{L}^f = [l_1^f, \cdots, l_N^f, Q^{1/2}] \tag{5.25}$$

Each new column introduces a new direction for the uncertainty of the state vector. Therefore the number of columns is reduced to N after every forecast step by computing the eigenvalue decomposition of the $(N+m)$ squared matrix $\tilde{L}^T\tilde{L}^f$ and discarding the m components with the smallest variance.

$$\left(\tilde{L}^f\right)^T \tilde{L}^f = CDC^T \tag{5.26}$$

Let us assume the following decomposition of the matrix C in four matrix blocks:

$$C = \begin{bmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{bmatrix}, \tag{5.27}$$

In order to keep an unified framework with the semi-deterministic version of SRFs, we denote by T^o the matrix built from the two matrix blocks C_{11} and C_{21} as $(N+m) \times N$ matrix. The reduction of the number of columns to N completes the truncation step of the RRSQRT filter. The way of this truncation is performed may define four versions of the RRSQRT filter corresponding to the four ESRF algorithms described in the previous section. If the model error is small, but without losing the generality, we can assume that the largest values in the columns of T^o are situated in the $N \times N$ block C_{11}. Otherwise, the largest elements can be placed in the matrix C_{11} by changing the rows.

The standard RRSQRT filter can be viewed as the one-sided deterministic formulation and is obtained by the following computation:

$$L^f(k) = \tilde{L}^f T^o \tag{5.28}$$

By considering the singular value decomposition of $C_{11} = U\Gamma V^T$, the new algorithm called symmetric RRSQRT filter is obtained by the multiplication:

$$L^f(k) = \tilde{L}^f T^o V U^T \tag{5.29}$$

Note that the transformation is only approximately symmetric because of the truncation. Similar to the previous ESRF algorithms, the orthogonal transformations U^p or U^r have been added to the symmetric RRSQRT to produce "rotated" RRSQRT filters. The difference between the two algorithms that use random rotations is not relevant since the central forecast has been used to perform the RRSQRT filters.

Assimilation step for each of the RRSQRT versions is performed in the same way as the deterministic analysis step described in the section 5.2 for the ESRFs involved. Similarly, the new matrix is updated by:

$$L^a = L^f T^s, \tag{5.30}$$

where T^s is the symmetric matrix which satisfies (5.15).

5.4 Smallest analysis increments

In order to demonstrate that the symmetric ESRF and RRSQRT algorithm introduce the smallest analysis increments, let us consider a symmetric positive defined matrix T, an orthogonal matrix U and the identity matrix in the N dimensional space of the squared matrices denoted by I.
Then the following propriety shows that the symmetric ETM minimizes the norm inequality and ensures that the analysis perturbations L^a are closest to the forecast perturbations L^f:

$$\| TU - I \| \geq \| T - I \|, \tag{5.31}$$

where $\| \, . \, \|$ denotes a matrix norm (e.g. Frobenius norm).

It is sufficient to prove that the statement holds for any positive definite diagonal matrix D and orthogonal matrix V. To see this, let B be another orthogonal matrix. Then the statement becomes:

$$\| B(DV - I)B^T \| \geq \| B(D - I)B^T \| \tag{5.32}$$

For the orthogonal matrix $V = BUB^T$, from (5.32) we have the following:

$$\begin{aligned} \| B(DB^TUB - I)B^T \| &\geq \| B(D - I)B^T \| && (5.33) \\ \| BDB^TU - I \| &\geq \| BDB^T - I \| && (5.34) \\ & && (5.35) \end{aligned}$$

Without losing the generality, by considering T as in the following decomposition:

$$T = BDB^T, \tag{5.36}$$

the inequality is proved. It remains to show that the propriety is valid for any positive definite diagonal matrix D given by the following:

$$D = diag\,(\lambda_1, \lambda_2 \ldots \lambda_N) \tag{5.37}$$

where λ_i, $i \in 1,\bar{N}$ represent the positive eigenvalues of D. The left hand term of the norm inequality becomes:

$$\begin{aligned} \| DU - I \|^2 &= \sum_{i=1}^{N} (\lambda_i u_{ii} - 1)^2 - \sum_{i=1}^{N} {\lambda_i}^2 u_{ii}^2 \\ &+ {\lambda_1}^2 \sum_{i=1}^{N} u_{1i}^2 + \ldots + {\lambda_N}^2 \sum_{i=1}^{N} u_{Ni}^2 \\ \| DU - I \|^2 &= \sum_{i=1}^{N} {\lambda_i}^2 - 2 \sum_{i=1}^{N} \lambda_i u_{ii} + N. && (5.38) \end{aligned}$$

The right hand term is calculated according to:

$$\begin{aligned} \| D - I \|^2 &= \sum_{i=1}^{N} (\lambda_i - 1)^2 \\ \| D - I \|^2 &= \sum_{i=1}^{N} {\lambda_i}^2 - 2\sum_{i=1}^{N} \lambda_i + N. \end{aligned} \tag{5.39}$$

By subtracting 5.38 from 5.39 we have the following:

$$\begin{aligned} \| DU - I \|^2 - \| D - I \|^2 &= 2\left(\sum_{i=1}^{N} \lambda_i - \sum_{i=1}^{N} \lambda_i u_{ii}\right) \\ \| DU - I \|^2 - \| D - I \|^2 &= 2\sum_{i=1}^{N} \lambda_i \,(1 - u_{ii}). \end{aligned} \tag{5.40}$$

Due to the fact that each λ_i is positive and each element of U is less than one, the conclusion is proved. Then, for any symmetric positive definite matrix T and orthogonal matrix U:

$$\| TU - I \| \geq \| T - I \| \,. \tag{5.41}$$

5.5 Experimental setup

The numerical experiments in this section are intended to show the performance of the stochastic filter (EnKF), the four semi-deterministic filters obtained by using different solutions of the equation 5.15 and four analogous RRSQRT filters. Our data assimilation experiments use two stochastic models, namely the weekly nonlinear 2D pollution model and the strongly nonlinear Lorenz 40-variables model.

5.5.1 2D pollution model

The first model ([110]) is based on the 2D advection diffusion equation for the transport of a pollutant:

$$\frac{\partial c}{\partial t} + u\frac{\partial c}{\partial x} + v\frac{\partial c}{\partial y} = \nu\frac{\partial^2 c}{\partial x^2} + \nu\frac{\partial^2 c}{\partial y^2} + e, \tag{5.42}$$

with the $[0, 30] \times [0, 31]$ domain and zero initial conditions. Here, c is the concentration of the pollutant, $[u, v]$ is the velocity field, ν is the dispersion coefficient set to 0.2, and e is the source term represented by the emissions.

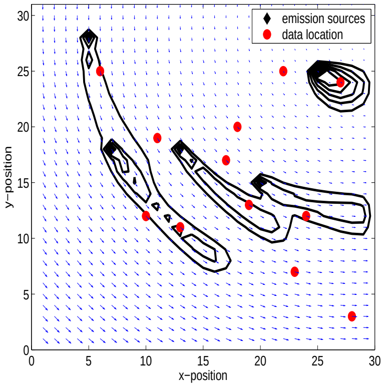

Figure 5.1: 2D pollution model; reference simulation of concentrations and wind velocity. Emission sources and data locations are represented by diamonds and circles, respectively.

For all experiments the velocity field is assumed to be known and constant in time. A reference simulation was performed by inserting randomly generated emissions at five grid cells (see Figure 5.1).

Observations are measured in twelve locations of the domain and they are simulated using the true concentrations to which a zero mean Gaussian observation noise was added with standard deviation 0.1. The emissions e are treated as uncertain with only their mean value and statistics known. They are modeled according to:

$$e = \max(\bar{e} + \lambda_e, 0), \tag{5.43}$$

where e contains the truncated value to zero, $\bar{e}$ is the deterministic value of the emission in a grid cell and λ_e distributes a correlated noise over the emission array. The correlated noise process has the following equation in scalar form:

$$\begin{aligned} \lambda_e(k+1) = \alpha\lambda_e(k) + \sqrt{1-\alpha^2}w(k), \\ w(k) \sim N(0, 10). \end{aligned} \tag{5.44}$$

The coefficient $\alpha \in [0,1]$ represents the time correlation parameter $\alpha = \exp(-\delta t/\tau)$ for a given time correlation length. We used $\alpha = 0.9$ and time step $\delta t = 1.0$. The stochastic model state is formed by augmenting the state vector with the noise process λ_e. Note that the truncation introduces a nonlinearity in this otherwise linear model.

It is important to note that the use of system noise is crucial for this application as the uncertain emissions are the only source of uncertainty. Covariance inflation is also not applicable as the structure of the error covariance is determined by the interaction of system noise and model dynamics, whereas covariance inflation ignores the influence of the system noise structure.

5.5.2 Lorenz 40-variables model

The Lorenz 40-variables model ([82]) is a strongly nonlinear system designed to produce a very simplified simulation of a scalar meteorological parameter around a latitude circle. This model is governed by the following equations and circular boundary conditions:

$$\frac{dX_j}{dt} = (X_{j+1} - X_{j-2})\, X_{j-1} - X_j + F, \qquad j = 1, ..., 40. \tag{5.45}$$

$$X_{-1} = X_{39}, \quad X_0 = X_{40}, \quad X_{41} = X_1. \tag{5.46}$$

The equations obtained with the forcing term $F = 8$ are integrated by using the forth-order Runge-Kutta solver with a time step of 0.05. The reference solution was generated by 20000 steps model integration. The observations were generated from the reference model variables after the first 20000 steps by intervals of 4 steps by adding the normal zero mean noise with the standard deviation of 0.5.

By contrast to the use of the deterministic Lorenz 40-variables model with the covariance inflation scheme in several testing data assimilation applications, e.g. [1], [113], a system noise of zero mean and standard deviation of 0.1 has been added per time unit. Since the inflation factor should be tuned for each experiment, the use of system noise avoids the need for covariance inflation which simplifies the experiments. Also, we feel that the errors introduced by the algorithms are masked with the rather arbitrary increase of the error covariance.

The assimilation run covers a time window of 40000 time steps by using an initial condition obtained from 20000 time steps integration of the model and initial standard deviation of the error set to 6.0 for spinning up the Kalman filters.

5.6 Results and discussions

The experiments were designed to examine the performances of the ensemble-based and deterministic algorithms in terms of the root mean square (RMS) error. For evaluating the assimilation performance we compute the averaged RMS error over 10 independent model simulations of the filters involved. This allows for easy computation of the significance of the differences found.

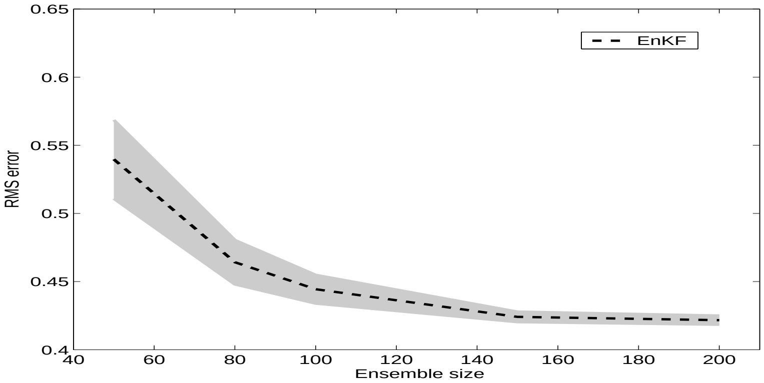

Figure 5.2: 2D pollution model; average RMS error over 500 time steps and 10 realizations versus ensemble size for the EnKF with observational error standard deviation of 0.1. The standard deviation between the independent simulations is represented by the grey band.

The assimilation results depend on the setup of the experiments: stochastic model used, observation error specification and the accuracy of the algorithms. For each of two models the nine filters described above were applied and compared. Experiments were performed with various sizes of the ensemble and different numbers of modes, where modes represent the leading eigenvalues needed for matrix truncation.

5.6.1 2D pollution model

The convergence of the EnKF with the number of ensemble members is studied. Figure 5.2 shows the RMS error of the forecast errors average for the 10 simulations. The grey bands denote the standard deviation due to the sampling errors for the EnKF as computed from the difference between the repetitions. It can be noticed that by using an EnKF with 100 ensembles a good performance is obtained, although there is further improvement achieved with ensemble from 100 to 200 members. Note that no inflation factor was needed which significantly simplifies these experiments.

The next experiment compares the EnKF with the various semi-deterministic ESRFs. The RMS error for the four variants of the ESRF is illustrated in Figure 5.3, top panel. The performance of the ESRF with random and mean-preserving rotations and the ESRF with one-sided ETM are very similar and show higher RMS values compared to the ESRF with the symmetric ETM. Also in [91] and [79] has been found that the symmetric ESRF performed better than one-sided and ESRF with random rotations. Contrary

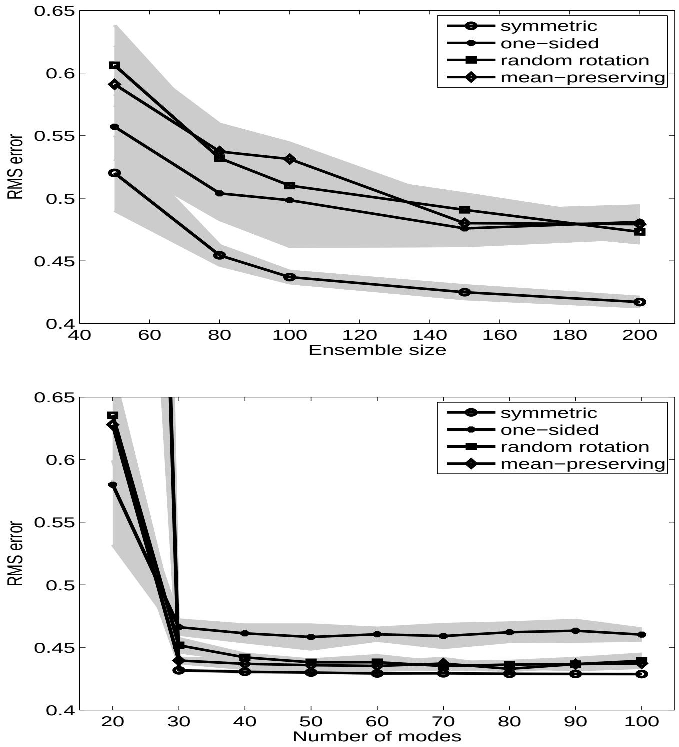

Figure 5.3: 2D pollution model; average root mean square error over 500 time steps and 10 realizations as function of ensembles and modes, respectively for the 4 ESRFs (top) and 4 RRSQRT (bottom) filters using an observational error standard deviation of 0.1. The standard deviation between the independent simulations is represented by the grey bands.

to their results, in these experiments the UESRF does not perform similar to the symmetric ESRF, but significantly worse. [91] and [79] explained the good performance of the unbiased ESRF with mean-preserving random rotations by the fact that this algorithm do not introduce an additional bias into the estimate. To investigate this further, another experiment has been performed where the amount of truncation due to negative concentrations has been stored. Negative concentrations may result because of the linear transformation used in the analysis. Within the model negative concentrations are set to 0 at the beginning of each forecast. In addition we computed the average analysis increments for each of the ESRFs. Table 5.1 shows a summary of the statistics obtained from an experiment with various filters. All simulations

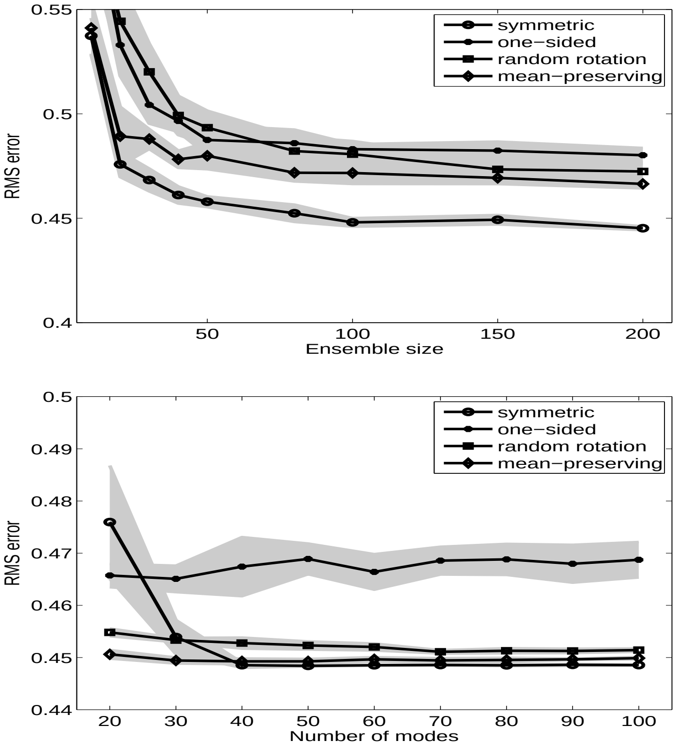

Figure 5.4: 2D pollution model; average root mean square error over 500 time steps and 10 realizations as function of ensembles and modes, respectively for the ESRFs (top) and RRSQRT (bottom) filters using an observational standard deviation error of 1.0. The standard deviation between the independent simulations is represented by the grey bands.

	RMS $(x^a - x^f)$		**Mean truncation**		**RMS error**	
	mean	std	mean	std	mean	std
EnKF	0.23	0.01	-0.0022	0.00028	0.51	0.05
ESRF rand rot	0.62	0.01	-0.0028	0.00008	0.55	0.03
ESRF symmetric	0.18	0.01	-0.0018	0.00014	0.46	0.02
ESRF one-sided	0.30	0.02	-0.0022	0.00023	0.58	0.06
ESRF mean-pres rot	0.62	0.02	-0.0029	0.00023	0.55	0.05

Table 5.1: Magnitude of analysis increments and the amount of truncated negative concentrations and its impact.

has been performed from $t = 0$ until $t = 50$ and 100 ensemble members has been used for all experiments. Each run has been repeated 10 times to check the uncertainty of the estimates. In the table 'mean' denotes the average over these runs and 'std' the standard deviation. The RMS of the analysis increments $(x^a - x^f)$ is significantly smaller for the symmetric ESRF. The random rotations add considerably to the magnitude of the analysis increments. The average change due to truncation of negative values appears to increase with the analysis increments. Finally the RMS error seems to deteriorate with larger amounts of truncation. These results are consistent with the idea that the analysis increments should be as small as possible, while still correcting sufficiently towards the true state.

The next experiment compares the semi-deterministic ESRFs to the RRSQRT algorithms. The illustration of the performance of the RRSQRT filters is given in Figure 5.3, bottom panel. The most striking feature is that all the RRSQRT filters need fewer members than ESRFs for convergence. The saturation of the RMS error versus number of modes is reached after 30 modes. Apparently, the sampling errors introduced by the system noise are significant in this experiment. Moreover, the RRSQRT filters are effective in reducing these sampling errors. Similar to the previous experiments, the symmetric formulation of the deterministic algorithm results in better state estimates than the traditional one-sided algorithm and the two algorithms with random rotations, although the difference in RMS error with the latter two is quite small. Note that the RRSQRT filters use a central forecast for the state estimate and not the ensemble mean. This implies that preserving the mean for the rotations does not add to the accuracy in this case.

The convergence of the ESRF algorithms with the ensemble size is influenced by the parameters of the problem. In the next experiment the standard deviation of the errors for the observations is increased from 0.1 to 1.0. Figure 5.4, top panel shows that all ESRF algorithms converge faster for this larger observation error. The most likely explanation for this is that the Kalman filter becomes less sensitive to errors in L. As one can see from equation 5.16, the ETM will be close to identity when the observational errors are large, but it may contain amplifications if these are small relative to the background errors L. The relative differences between the four ESRFs investigated are similar to the previous experiments.

The bottom panel of Figure 5.4 illustrates the results for the 4 RRSQRT filters with the increased observation errors. Also in this experiment the RMS error values have been reduced for small mode sizes. An unexpected result is the very good performance for the two algorithms with random rotations in the ETM.

5.6.2 Lorenz 40-variables model

Figure 5.6 illustrates the performance of the ESRF and RRSQRT filters applied to the Lorenz 40-variables model. Our experiments confirm the out performance of the one-sided ESRF version with a substantially large RMS error. Also the narrow standard deviation band suggests that the one-sided variant is not able to provide more accurate results. The differences between the results presented by [91] and here, are that the RMS values obtained with the one-sided ETM are larger and the convergence of the algorithm is slower in our case. We think the differences are caused by the variable covariance inflation used by [91]. By selecting an optimal inflation factor for each experiment separately, the covariance inflation can compensate for errors introduced by the algorithm. Our results are consistent with this difference in the experimental setup.

The filters with random and mean-preserving rotations exhibit an unstable behavior for an ensemble of 100 members. Further investigation of this effect shows that the algorithms are sensitive in respect to random rotations for several ensemble sizes around 100. In more detail, we see that the larger RMS values are caused by filter divergence. The experiments do not recover from divergence, most likely due to the absence of the covariance inflation. Our results are again consistent with the findings of [91] (their figure 3) where for a fixed inflation factor a number of algorithms also did not improve monotonously with increasing ensemble size. This aspect was hidden somewhat by compensation with a larger inflation factor for these experiments.

More accurate and stable results are obtained with the symmetric ESRF. The convergence is reached using 50 ensemble members in contrast to that shown by the classical EnKF (see Figure 5.5). The EnKF needs 100 ensembles to provide a comparable performance with the symmetric ESRF.

In the bottom panel of Figure 5.6 the RMS errors for the 4 RRSQRT variants are shown. The convergence is a little faster than for the ESRFs. However, it should be noted that this model has only 40 state variables, so

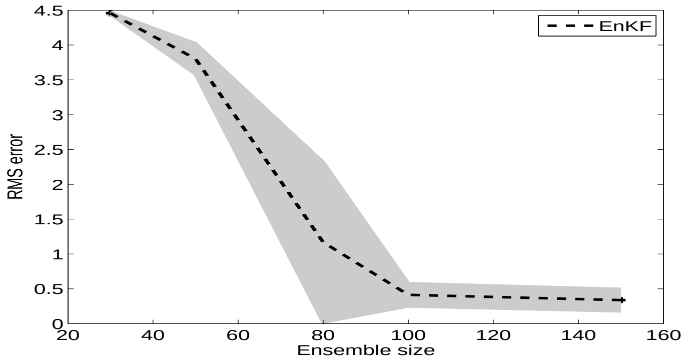

Figure 5.5: Lorenz model; average RMS error over 500 time steps and 10 realizations versus ensemble size for the EnKF. The standard deviation between the independent simulations is represented by the grey band.

that there is no real truncation error for the RRSQRT filters for N larger than 40. This behavior is of course, not representative for real applications. The small differences between the algorithms in this case are not statistically significant.

5.7 Conclusions

In this chapter we have introduced a new deterministic algorithm that represents the symmetric variant of the RRSQRT filter. Our goals were to investigate the performance of stochastic, semi-deterministic and deterministic filters with two models and to gain insight into the behavior of the symmetric algorithms by enhancing the understanding of the propriety of the smallest increments of the state.

The results obtained in our study confirm that ESRFs converge more quickly with the size of the ensemble than the classical EnKF. This fact is due to the sampling errors introduced in the analysis of the EnKF needed there to produce unbiased analysis errors.

In our experiments, the symmetric version of the ESRFs which has the smallest analysis increments over all semi-deterministic filters provides the more accurate solution compared to the other three versions (one-sided variant, random rotations and mean-preserving random rotations), as it has been already demonstrated by several authors. Contrary to the experiments shown by [91], the results obtained with the mean-preserving random rotation algorithm are not close to those of symmetric algorithm in our first experiment.

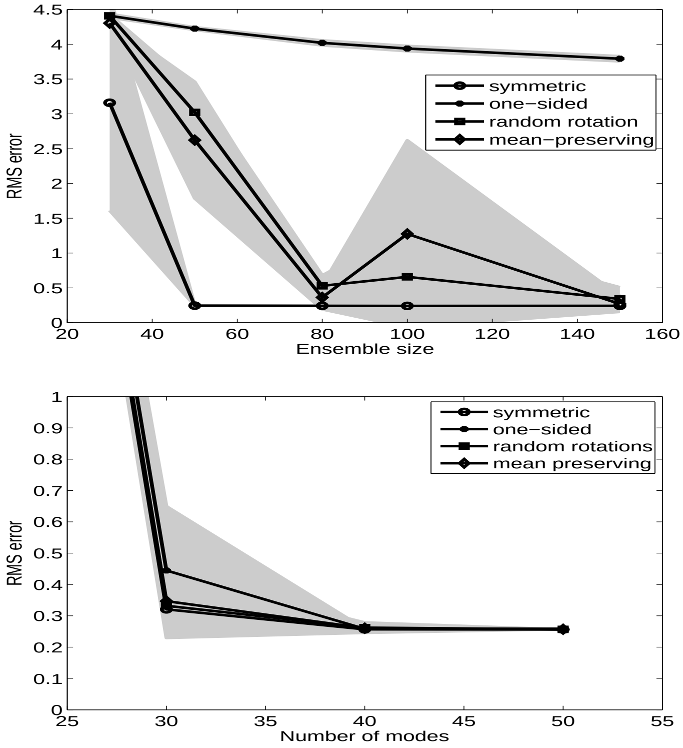

Figure 5.6: Lorenz model; average root mean square error over 2000 time steps and 10 realizations as function of ensembles and modes, respectively for the ESRFs (top) and RRSQRT (bottom) filters using an observational standard deviation error of 0.1. The standard deviation between the independent simulations is represented by the grey bands.

This can be explained from negative concentrations introduced by the additional increments from random rotation. This would imply that the influence of the random rotations could depend on the experiment. In several applications adding random rotations may provide an improved result upon a one-sided solution. Furthermore, mean-preserving random rotations may improve the estimation upon a random rotation variant, but this is not necessarily the case for all applications. What seems consistent is that the symmetric ESRF is always among the best performing algorithms of the 4 ESRFs. In addition, the computational requirements are not significantly larger for the symmetric ESRF, nor is it difficult to implement. Therefore, we advise to at least try the symmetric ESRF when applying ESRF type algorithms.

In analogy to the symmetric ESRF, a symmetric RRSQRT filter can be defined. The results obtained with this algorithm are much accurate than the original (one-sided) version. From the implementation point of view, the new algorithm is not difficult and does not require a significant amount of computation over the RRSQRT filter.

For applications with model errors that are not negligible the symmetric RRSQRT filter provides a good alternative to the ESRF. The convergence is often faster and results are reproducible because no sampling errors are introduced for the model error.

CHAPTER 6

Summary, Conclusions and Outlook

6.1 Overview

The atmosphere is a complex system which includes physical, chemical and biological processes. Many of these processes affecting the atmosphere are subject to various interactions and can be highly nonlinear. This complexity makes it necessary to apply computer models in order to understand the natural behavior of the atmosphere. A mathematical model is a representation of all relevant atmospheric processes. In addition to the chemical and physical processes it requires detailed information on the emissions, deposition and transport of trace constituents.

One of the most important message the atmospheric scientific community send us is that man is changing the atmosphere. Changing the chemical composition of the atmosphere will have a large impact on the human environment. Examples include the effect of aerosols on health, climate and visibility.

Data assimilation represents a crucial tool for estimating and predicting the chemical processes in the atmosphere. It refers to several techniques that aim to combine the information from various sources to provide unified and consistent description of an atmospheric chemical system. A large number of data assimilation schemes are available: optimal interpolation, kriging, variational methods and ensemble-based techniques.

The aim of this thesis was to investigate the use of several sequential data assimilation techniques in both ideal and real settings in the context of atmospheric chemistry. The increased complexity of models together with the different types of information about pollutants have the potential of contributing to a better understanding of chemical processes. Therefore, an optimal selection of the data assimilation techniques adapted to the new challenges in the atmospheric chemistry is required. Moreover, this work intends to give methodologies on how to use a specific data assimilation method.

6.2 Summary and conclusions

The research questions addressed in this thesis were:

- How can we use the retrospective data assimilation with an atmospheric chemistry model?
- How accurate can we estimate sulphur dioxide and sulphate by using a single component setup compared to a combined assimilation procedure?
- How can the OMI satellite product contribute to a better understanding of LOTOS-EUROS capabilities in predicting the tropospheric ozone?
- Can we select an optimal algorithm from the data assimilation perspective?

The first question has been related to the concept of filtering in contrast with smoothing. Several algorithms which combine the ensemble techniques and the Kalman smoother method have been tested in a twin experiment before using the algorithms for real-life atmospheric chemistry data assimilation problems. The goal is to improve the estimates provided by the Kalman filter by making use of future observations in the analysis time. The quality of the smoothed estimates is determined by how accurate the smoother computes the covariance matrix. Accurate computations are complicated because of the existent non-linearities in the model and the use of approximate covariance matrices in the case of large scale problems. Insight has been gained in the efficiency by showing that the computational effort required by using the FIFO lag algorithm is independent of the lag length. Therefore the FIFO algorithm may be more efficient than the EnKS.

To answer the second question, in Chapter 3 the sulphur cycle has been evaluated over Europe for the year 2003. Comparison of the deterministic model results with in situ data derived from EMEP database showed that the annual average simulated sulphate was systematically underestimated. For sulphur dioxide the model generally overestimated the annual mean concentrations. Other uncertainties than emissions should be consider to explain the difference between the model and measurements. A stochastic model based on a combination of uncertain emissions and the reaction rate was shown to be beneficial for the assimilation of sulphate aerosol and its precursor gas. Therefore, the joint assimilation of multiple species is advised in the case of large model uncertainties which can be attributed to several causes. We have demonstrated that with a more accurate description of the model error using two noisy parameters (emissions and reaction rate) instead of one (emissions) the multi-component assimilation performs better. The experiments showed that the filter technique was able to correct the model error, if the uncertain model parameters were specified correctly. The reconstruction of sulphur dioxide emissions has been treated by using the ensemble Kalman smoother methodology.

The third question is the topic of Chapter 4. The NO_2 tropospheric data provided by the Ozone Monitoring Instrument has been used with the LOTOS-EUROS model. The spatial pattern of nitrogen dioxide given by model simulations has been shown to be in a good agreement with the OMI dataset, but has a large negative bias. The main reason for this bias has been found to be the limited chemistry scheme due to the incorrect NO_2 lifetime parameter. It has been shown that including a bias aware data assimilation scheme, the discrepancy between the simulated and observed NO_2 tropospheric columns is reduced significantly. The changes in nitrogen dioxide determine an overall increase of ozone values, with positive impact on the O_3 maxima. From operational perspective this fact may be especially relevant to forecast high ozone values that correspond to enhanced pollution episodes.

Chapter 5 focuses on the last question. Our goals were to investigate the performance of the ensemble square root filters (ESRFs) with two stochastic models. We demonstrated that the symmetric version of the ESRFs has the smallest analysis increments over all square root filters. From our experiments and those reported by other authors, we may conclude that the symmetric ESRF is likely to provide the most accurate results for a large number of applications when compared to other ESRFs. In analogy to the symmetric ESRF, a symmetric RRSQRT filter has been introduced. The results obtained with this algorithm are much more accurate than the original (one-sided) version. It is advisable for the existing and new implementations of the RRSQRT algorithm to modify the code to the symmetric form as it is clearly more accurate, more reproducible and less susceptible to numerical errors. Moreover, the changes needed are not difficult to implement and do not require a significant amount of computation.

6.3 Outlook

The work described in this thesis leads to several important issues and points out that challenging problems are envisaged in the future.

For many models, such as hydrological and land models with uncertainties in the atmospheric forcing, or atmospheric chemistry models with uncertain pollution sources, the model error represents a key factor of successfully applying DA techniques. Therefore, more effort should be dedicated to the detailed characterization of uncertainties in modeling, and for this purpose, data assimilation may be considered as a beneficial tool.

More work is required to completely understand the use of ensemble data assimilation to reduce uncertainties in emission inventories. One challenge arises from the long integration time needed to develop meaningful correlations between the emissions rates and concentration fields. Another challenge is posed by large spurious correlations which lead the filter to correct the emission factors in order to compensate for other sources of error.

Another problem that has to be taken into account is to jointly assimilate observations of several different species in order to get improvements over all chemical system. It has been shown that one should move from single component applications of data assimilation to multi-component applications. The increased complexity associated with this move requires a very careful specification of the system configuration, which can be one of the main challenges for the future.

The number of measurement sites and the density of the measurement network play an important role. Due to the inhomogeneous density of data corroborated with the localization effect imposed by the ensemble-based filter, the influence of assimilation on model state and parameters is limited to certain areas. Hence, by including a more extensive set of monitoring data the spatial impact of the assimilation can be improved. Future studies dedicated to provide accurate concentration maps over a given domain should encompass a larger number of observation sites. A more extensive dataset could be obtained using data from national networks in combination with rigorous quality assurance to ensure the use of a dataset representative for the model resolution.

Furthermore, satellite data becomes an important source of information. The use of nitrogen dioxide satellite measurements derived from OMI is relatively new. Apart from investigating model weaknesses, the satellite dataset can significantly contribute to forecast activities.

The symmetric ESRF is likely to provide the most accurate results for a large number of applications when compared to other ESRFs. In addition, the computational requirements are not significantly larger. Therefore, we advise to use the symmetric ESRF when applying ESRF type algorithms with LOTOS-EUROS model.

We have proposed a method for handling the model uncertainty based on a deterministic Kalman filter. This method has been tested in a twin experiment, but not yet in a large scale application which can be a challenging problem.

Appendix

The LOTOS-EUROS model

LOTOS-EUROS describes the distributions of oxidant and aerosols over Europe. The model is based on a discretization of the advection diffusion equation:

$$
\begin{aligned}
\frac{\partial c_s}{\partial t} &= -\nabla \cdot (u_h c_s) + \nabla \cdot (\mu_h \nabla c_s) \\
+\frac{\partial}{\partial v}\left(\mu_v \frac{\partial c_s}{\partial v}\right) &+ E_s + C(c_s) - D(c_s) + V(c_s),
\end{aligned}
$$

where c_s is the concentration field of the trace gas, u_h is the horizontal velocity field in two dimensions, μ_h and μ_v represent the horizontal and vertical diffusion coefficients, and the source terms E, C, D, and V account for emissions, chemistry, deposition and mean vertical exchange, respectively.

Domain

The model domain used in this study is bound at 35° and 70° North and 10° West and 40° East covering Europe from the western part of Russia to the Atlantic Ocean and from the Mediterranean Sea to Scandinavia (see Figure 3.1). The projection is normal longitude-latitude and the grid resolution is 0.5° longitude × 0.25° latitude, approximately 25 × 25 km. In the vertical there are a surface layer of 25 m and three dynamic layers with a top at 3.5 km above sea level. The lowest dynamic layer represents the variable mixing layer with the height obtained from the meteorological input. The upper two layers are reservoir layers with equal thickness and a minimum of 100 m.

Transport

The transport consists of advection in 3 dimensions, horizontal and vertical diffusion. The calculation of the gas and cloud phase chemistry in the model requires meteorological input fields. The advection is driven by meteorological data produced at the Free University of Berlin employing a diagnostic meteorological analysis system ([71]). The vertical wind speed is calculated by the model as a result of the divergence of the horizontal fields.

Chemistry

The gas phase chemistry in LOTOS-EUROS is described by the TNO CBM-IV (Carbon Bond Mechanism) scheme ([95]), which is a modified version of the original CBM-IV ([114]). The mechanism was tested against the results of an inter-comparison presented in [73] and found to be in good agreement with the results obtained with the other mechanisms. The complete chemistry scheme includes 28 species and 66 reactions, including 12 photolytic reactions. In this thesis the focus is on sulphur and nitrogen oxides chemistry which are described in details in the 3 and 4, respectively.

Emissions

In order to understand the role of different species in atmosphere, there is important to quantify their sources and sinks. The largest sources are over the continent and produce pollutants at the surface layer mainly from fossil fuel combustion and biomass burning. Emissions databases provide total emissions in terms of tones per year for each grid cells based on inventories by local environmental agency. Area sources are injected into the lowest layer, whereas the emissions from point sources are injected according to stack height ([16]). The temporal variation of the emissions is represented by time factors. The annual emission totals are translated to hourly emissions in LOTOS-EUROS using prescribed temporal factors ([16]).

Removal processes

Important processes for pollutants are the removal processes as dry deposition on the surface and wet deposition. Dry deposition for each species is modeled by using a combination of three parameters describing the atmospheric, viscous and surface resistance. Wet deposition is calculated using simple coefficients for below cloud scavenging.

Bibliography

[1] J. L. Anderson. An ensemble adjustment Kalman filter for data assimilation. *Mon. Wea. Rev.*, 129:2884–2903, 2001.

[2] C. Andersson, J. Langner, and R. Bergstrom. Inter-annual variation and trends in air pollution over Europe due to climate variability during 1958-2001 simulated with a regional CTM coupled to the ERA40 reanalysis. *Tellus B*, 59, 2007.

[3] A. L. Barbu, R. G. Hanea, M. Schaap, and A. W. Heemink. Estimation of sulphur emissions using ensemble smoothers. *Proceedings of the 28th NATO/CCMS International Technical Meeting on Air Pollution Modeling and its Applications*, 2006.

[4] A. L. Barbu, A. J. Segers, M. Schaap, A. W. Heemink, and P. J. H Builtjes. Bias aware assimilation of OMI no_2 tropospheric columns with LOTOS-EUROS model. *to be submitted*.

[5] A. L. Barbu, A. J. Segers, M. Schaap, A. W. Heemink, and P. J. H Builtjes. A multi-component data assimilation experiment directed to sulphur dioxide and sulphate. *Atmos. Environ.*, 43:1622–1631, 2009.

[6] A. L. Barbu, M. Verlaan, and A. W. Heemink. Square root filters in the presence of model error. *submitted to Mon. Wea. Rev.*

[7] C. Bishop, B. J. Etherthon, and S. J. Majumdar. Adaptive sampling with the ensemble transform Kalman filter. *Mon. Wea. Rev.*, 129:420–436, 2001.

[8] N. Blond, K. F. Boersma, H. J. Eskes R. J. van der A, M. van Roozendael, I. De Smedt, G. Bergametti, and R. Vautard. Intercomparison of SCIAMACHY nitrogen dioxide observations, in situ measurements and air quality modeling results over Western Europe. *J. Geophys. Res.*, 112:D10311, 2007.

[9] K. F. Boersma, H. J. Eskes, J. P. Veefkind, E.J. Brinksma, R. J. van der A, M. Sneep, G.H.J. van den Oord, P. F.Levelt, P. Stammes, J. F. Gleason, and E. J. Bucsela. Near-Real Time Retrieval of Tropospheric no_2 from OMI. *Atmos. Chem. and Phys.*, 7:2103–2118, 2007.

[10] K. F. Boersma, D. J. Jacob, H. J. Eskes, R. W. Pinder, J. Wang, and R. J. van der A. Intercomparison of SCIAMACHY and OMI tropospheric NO2 columns–observing diurnal evolution chemistry and emissions from space. *J. Geophys. Res.*, 113:D16S26, 2008.

[11] K. F. Brinksma and et all. Near real time retrievals of tropospheric NO2 from OMI. *J. Geophys. Res*, 112:D10311, 2006.

[12] B. Brunekreef. Air pollution and life expectancy: Is there a relation? *Occup. Environ. Med.*, 54:781–784, 1997.

[13] E. J. Bucsela, E. A. Celarier, M. O. Wenig, J. F. Gleason, J. P. Veefkind, K. F. Boersma, and E. J. Brinksma. Algorithm for NO2 vertical column retrieval from the Ozone Monitoring Instrument. *IEEE Trans. Geosci. Remote Sens.*, 44:1245–1258, 2006.

[14] E. J. Bucsela and et all. Comparison of NO2 in situ aircraft measurements with data from the Ozone Monitoring Instrument. *J. Geophys. Res*, 113:D16S31, 2008.

[15] P. J. H. Builtjes. The LOTOS–Long Term Ozone Simulation–project, summary, report. *TNO–MW– R report*, 240, 1992.

[16] P.J.H. Builtjes and M. van Loon. Project on the modeling and verification of ozone reduction strategies: contribution of TNO–MEP. *Journal of Sound and Vibration*, 263:679–699, 2003.

[17] G. Burgers, P. J. Leewen, and G. Evensen. Analysis scheme in the ensemble Kalman filter. *Mon. Wea. Rev.*, 126:1719–1724, 1998.

[18] G. R. Carmichael, A. Sandu, T. Chai, D. N. Daescu, E. M. Constantinescu, and Y. Tang. Predicting air quality: Improvements through advanced methods to integrate models and measurements. *J. Comp. Physics*, 227:3540–3571, 2008.

[19] E. A. Celarier and et all. Validation of Ozone Monitoring Instrument nitrogen dioxide columns. *J. Geophys. Res.*, 112:D15S15, 2008.

[20] R. J. Charlson, S. E. Schwarz, J. M. Hales, R. D. Cess, J. A. Coackley, J. E. Hansen, and D. J. Hofmann. Climate forcing by anthropogenic aerosols. *Science*, 255:423–430, 1992.

[21] M. P. Chipperfield, B. V. Khattatov, and L. Lary. Sequential assimilation of stratospheric chemical observations in a three–dimensional model. *J. Geophys. Res.*, 2007.

[22] S. E. Cohn. An introduction to estimation theory. *NASA Goddart Space Flight Center Data assimilation Office note*, 97, 1997.

[23] S. E. Cohn and D. F. Parish. The behaviour of forecast error covariances for a Kalman filter in two dimensions. *Mon. Wea. Rev.*, 119:1757–1785, 1991.

[24] E. M. Constantinescu, A. Sandu, T. Chai, and G. R. Carmichael. Ensemble–based chemical data assimilation I: General approach. *Quart.J. Roy. Meteor. Soc.*, 133, 2007.

[25] E. M. Constantinescu, A. Sandu, T. Chai, and G. R. Carmichael. Ensemble–based chemical data assimilation II: Covariance localization. *Quart.J. Roy. Meteor. Soc.*, 133:1245–1256, 2007.

[26] P. Courtier. Dual formulation of four–dimensional variational assimilation. *Quart. J. Roy. Meteor. Soc.*, 123:2449–2461, 1997.

[27] F. A. A. M. de Leeuw and H. J. van Rheineck Leyssius. Modeling study of sox and nox during the januari 1985 smog episode. *Water, Air and Soil Pollution*, 51:357–371, 1990.

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

[29] D. P. Dee and R. Todling. Data assimilation in the presence of forecast bias: The GEOS moisture analysis. *Mon. Weather Rev.*, 128:3268–3282, 2000.

[30] B. Denby, M. Schaap, A. J. Segers, P. J. H. Builtjes, and J. Horalek. Comparison of two data assimilation methods for assessing PM10 exceedances on the European scale. *Atmos. Environ*, 42:7122–7134, 2007.

[31] J. P. Drecourt, H. Madsen, and D. Rosbjerg. Bias aware Kalman filters: Comparison and improvements. *Adv. Water Resour.*, 29:707–718, 2006.

[32] H. Elbern, H. Schmidt, and A. Ebel. Variational data assimilation for tropospheric chemistry modeling. *J. Geophys. Res.*, 102:15967–15985, 1997.

[33] H. Elbern, H. Schmidt, O. Talagrand, and A. Ebel. 4d–variational data assimilation with an adjoint air quality model for emission analysis. *Atmos. Chem. Phys.*, 7:3749–3769, 2000.

[34] H. Elbern, A. Strunk, H. Schmidt, and O. Talagrand. Emission rate and chemical state estimation by 4–dimensional variational inversion. *Environ. Modeling and Software*, 15:539–548, 2007.

[35] EMEP. $http://www.nilu.no/projects/ccc/$.

[36] R. J. Engelen and A.P. McNally. Estimating atmospheric CO2 from advanced infrared satellite radiance within an operational four–dimensional variational data assimilation system: results and validation. *J. Geophys. Res.*, 110, 2005.

[37] J. W. Erisman. Evaluation of a surface resistance parameterization of sulphur dioxide. *Atmos. Environ.*, 28:2583–2594, 1994.

[38] J. W. Erisman and M. Schaap. The need for ammonia abatement with respect to secondary PM reductions in Europe. *Environ. pollution*, 129:159–163, 2004.

[39] H. J. Eskes and K. F. Boersma. Averaging kernels for DOAS total columns satellite retrievals. *Atmos. Chem. and Phys.*, 3:1285–1291, 2003.

[40] G. Evensen. Sequential data assimilation with a nonlinear quasi–geostrophic model using Monte–Carlo methods to forecast error statistics. *J. Geophys. Res.*, 99:10143–10162, 1994.

[41] G. Evensen. The ensemble Kalman filter: theoretical formulation and practical implementation. *Ocean Dynamics*, 53:343–367, 2003.

[42] G. Evensen. Sampling strategies and square root analysis schemes for the EnKF. *Ocean Dynamics*, 54:539–560, 2004.

[43] G. Evensen and P. J. van Leeuwen. An ensemble Kalman smoother for nonlinear dynamics. *Mon. Wea. Rev.*, 128:1852–1867, 2000.

[44] H. Fagerli, D. Simpson, and W. Aas. Chapter 1: Model performance for sulphur and nitrogen compounds for the period 1980 to 2000. *EMEP report 1/2003, Part II. Transboundary acidification, eutrophication and ground level ozone in Europe*, 2003.

[45] B. F. Farrell and P. J. Ioannou. Data assimilation in the presence of forecast bias. *J. Atmos. Sci*, 58:2771–2789, 2001.

[46] B. Friedland. Treatment of bias in recursive filtering. *IEEE Trans Automatic Control*, pages 359–367, 1969.

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

[48] M. Ghil. Meteorological data assimilation for oceanographers: Part I Description and theoretical framework. *Dyn. Atmos. Oceans*, 13:171–218, 1989.

[49] M. Ghil and P. Malanotte-Rizzoli. Data assimilation in meteorology and oceanography. *Adv. Geophys.*, 33:141–266, 1991.

[50] T. M. Hamill, J. S. Whitaker, and C. Snyder. Distance-dependent filtering of background error covariance estimates in an ensemble Kalman filter. *Mon. Wea. Rev.*, 129:2776–2790, 2001.

[51] P. Hammingh, H. Tho, F. de Leeuw, F. Sauter, A. van Pul, and J. Matthijsen.

[52] R. G. Hanea, G. J. M. Velders, and A. W. Heemink. Data assimilation of ground–level ozone in Europe with a Kalman filter and chemistry transport model. *J. Geophys. Res.*, 109:D10302, 2004.

[53] R. G. Hanea, G. J. M. Velders, A. J. Segers, M. Verlaan, and A. W. Heemink. A hybrid Kalman filter algorithm for large scale atmospheric–chemistry data assimilation. *Mon. Wea. Rev.*, 135:140–151, 2007.

[54] H. Hass, H. J. Jacobs, and M. Memmesheimer. Analysis of a regional model (EURAD) near surface gas concentration predictions using observations from networkslarge scale atmospheric–chemistry data assimilation. *Meteor. Atmos. Phys*, 57:173–200, 1995.

[55] A. W. Heemink, K. Bolding, and M. Verlaan. Sorm surge forecasting using Kalman filtering. *J. Meteo Soc. Japan*, 75:305–318, 1995.

[56] A. W. Heemink and A. J. Segers. Modeling and prediction of environmental data in space and time using Kalman filtering. *Stochastic Environ. Res. and Risk Asses.*, 16, 2002.

[57] A. W. Heemink, M. Verlaan, and A. J. Segers. Variance reduced ensemble Kalman filter. *Mon. Wea. Rev.*, 129:1718–1728, 2001.

[58] A. G. Hjellbrekke. Data report 2003: Acidifying and eutrophying compounds. *EMEP/CCC-Report*, 3, 2005.

[59] A. Hodzik, S. Madronich, B. Bohn, S. Massie, L. Menut, and C. Wiedinmyer. Wildfire particulate matter in Europe during summer 2003: meso-scale modeling of smoke emissions, transport and radiative effects. *Atmos. Chem. Phys. Discuss.*, 7:4705–4760, 2007.

[60] G. Hoek, B. Brunekreef, S. Goldbohm, P. Fisher, and P. A. van den Brandt. Association between mortality and indicators of traffic-related air pollution in the Netherlands: a cohort study. *Lancet*, 360:1203–1209, 2002.

[61] P. L. Houtekamer and H. L. Mitchell. Data assimilation using an ensemble Kalman filter technique. *Mon. Wea. Rev.*, 126:796–811, 1998.

[62] P. L. Houtekamer and H. L. Mitchell. A sequential ensemble Kalman filter for atmospheric data assimilation. *Mon. Wea. Rev.*, 129:123–137, 2001.

[63] S. Houweling. Global modeling of atmospheric methane sources and sinks. *PhD thesis*, 2000.

[64] http://www.temis.nl/luchtkwaliteit/.

[65] C. M. J Jacobs and W. A. J. van Pul. Long–range atmospheric transport of persistant Organic Pollutants, I: Description of surface–atmosphere exchange modules and implementation in EUROS. *Report 722401013, National Institute of Public Health and Environmental Protection (RIVM), Bilthoven*, 1996.

[66] L. H. Jazwinski. Stochastic Procceses and Filtering Theory. *Academic Press, New York*, 1970.

[67] K. Kaeresen and D. Hirst. Statistical estimation of emission and deposition of sulphur using the EMEP 50km model. *EMEP/MSC-W Note*, 2, 1999.

[68] R. E. Kalman. A new approach to linear filtering and prediction problem. *Transaction of the ASME–Journal of Basic Engineering*, 82:35–45, 1960.

[69] E. Kalnay. Atmospheric modeling, data assimilation and predictability. *Cambridge University Press*, 2005.

[70] P. Kasibatla, W. L. Chameides, and J. S. John. A three-dimensional global model investigation of seasonal variations in atmospheric burden of anthropogenic sulphate aerosols. *J. Geophys. Res.*, 102:3737– 3759, 1997.

[71] A. Kerschbaumer and E. Reimer. Preparation of meteorological input data for the RCG–model. *UBA-Report*, 299, 2003.

[72] I. B. Konovalov, M. Beekmann, A. Richter, and J. P. Burrows. Inverse modelling of the spatial distribution of NOx emissions on a continental scale using satellite data. *Atmos. Chem. Phys.*, 6:1747–1770, 2006.

[73] M. Kuhn, P. J. H. Builtjes, D. Poppe, and et all. Inter–comparison of the gas–phase chemistry in several chemistry and transport models. *Atmos. Environ.*, 32:693–709, 1998.

[74] J. F. Lamarque, B. V. Khattatov, V. Yudin, and et all. Application of bias estimator for the improved assimilation of measurements of pollution in the troposphere (MOPITT) carbon monoxide retrievals. *J. Geophys. Res.*, 109:D16304, 2004.

[75] L. N. Lamsal, R. V. Martin, A. van Donkelaar, M. Steinbacher, E. A. Celarier, E. J. Bucsela, E. J. Dunlea, and J. Pinto. Ground level NO2 concentrations infered from the satellite-borne Ozone Monitoring Instrument. *J. Geophys. Res.*, 113:D16308, 2008.

[76] A. Lauer, M. Dameris, A. Richter, and J. P. Burrows. Tropospheric NO2 columns:a comparison between model and retrieved data from GOME measurements. *Atmos. Chem. and Phys.*, 2:67–78, 2002.

[77] D. J. Lea, J. P. Drecourt, K. Haines, and M. J. Martin. Ocean altimeter assimilation with observational- and model-bias correction. *Quart. J. Royal Meteo. Soc.*, 134:1761–1774, 2008.

[78] O. Leeuwenburgh, G. Evensen, and L. Bertino. The impact of ensemble filter definition on the assimilation of temperature profiles in the tropical pacific. *Q. J. Meteorol. Soc.*, 131, 2005.

[79] D. M. Livings, S. L. Dance, and N. K. Nichols. Unbiased ensemble square root filters. *Physica D: Nonlinear Phenomena*, 237, 2008.

[80] A. C. Lorenc. Analysis methods for numerical weather prediction. *Quart. J. Roy. Meteor. Soc.*, 112:1177–1194, 1986.

[81] A. C. Lorenc. Modelling of error covariances by 4D–Var data assimilation. *Quart. J. Roy. Meteor. Soc.*, 129:3167–318, 2003.

[82] E. N. Lorenz and K. A. Emanuel. Optimal sites for suplementary weather observations: Simulation with a small model. *J. Atmos. Sci.*, 55:399–414, 1998.

[83] M. Marchand, S. Bekki, A. Hauchecorne, and J.-L. Bertaux. Validation of the self–consistency of GOMOS NO3, NO2 and O3 data using chemical data assimilation. *Geophys. Res. Lett.*, 31:L10107, 2004.

[84] E. Marmer, F. Dentener, J. v. Aardenne, F. Cavalli, and E. Vignatti. What can we learn about ship emission inventories from measurements of air pollution over Mediteranean Sea? *Atmos. Chem. and Phys. Discuss.*, 9:7155–7211, 2009.

[85] R. N. Miller, E. Carter, and S. Blue. Data assimilation into non–linear stochastic models. *Tellus*, 1999.

[86] R. N. Miller, M. Ghil, and F. Gautiez. Advanced data assimilation in strongly nonlinear dynamical systems. *J. Atmos. Sci.*, 51:1037–1056, 1999.

[87] D. Orrell, L. Smith, J. Barkmeijer, and T. Palmer. Model error in weather forcasting. *Nonlin. Processes Geophys.*, 2001.

[88] E. Ott, B. R. Hunt, I. Szunyogh, A. V. Zimin, E. J. Kostelich, M. Corazza, E. Kalnay, D. J Patil, and J.A. Yorke. A local ensemble Kalman filter for atmospheric data assimilation. *Tellus*, 56, 2004.

[89] D. T. Pham, J. Verron, and M. C. Roubaud. A singular evolutive extended Kalman filter for data assimilation in oceanography. *J. Mar. Syst.*, 16:323–340, 1998.

[90] S. Ravela and D. McLaughlin. Fast Ensemble Smoothing. *Ocean Dynamics*, 57, 2007.

[91] P. Sakov and P. R. Oke. Implication of the form of the ensemble transformation in the ensemble square root filters. *Mon. Wea. Rev.*, 54:539–560, 2007.

[92] A. Sandu, E. M. Constantinescu, L. Wenyuan, G. R. Carmichael, C. Tianfeng, J. Seinfeld, and D. N. Daescu. Ensemble–based data assimilation for atmospheric chemical transport models. *Lecture Notes on Comp. Sci.*, 3515:648–655, 2005.

[93] N. H. Savage, K. S. Law, J. A. Pyle, A. Richter, H. Nuess, and J. P. Burrows. Using GOME NO2 satellite data to examine regional differences in

TOMCAT model performances. *Atmos. Chem. and Phys.*, 4:1895–1912, 2004.

[94] M. Schaap and et all. Evaluation of long term aerosol simulations from seven regional air quality models and their ensemble in the EURODELTA study. *Atmosph. Environ.*, 43:4822–4832, 2009.

[95] M. Schaap, F. Sauter, R. M. A. Timmermans, M. Roemer, G. Velders, J. Beck, and P. J. H. Builtjes. The LOTOS-EUROS model: description, validation and the latest developments. *Int. J. Environ. and Pollution*, 32:270–290, 2008.

[96] M. Schaap, H. L. C.Denier van Der Gon, F. J. Dentener, L. J. H. Visschedijk, M. van Loon, H. M. Ten Brink, J. P. Putaud, B. C. Guillaume, B. Liousse, and P. J. H. Builtjes. Anthropogenic Black Carbon and Fine Aerosol Distribution over Europe. *J. Geophys. Res.*, 109:D18201, 2004.

[97] M. Schaap, M. van Loon, H. M. ten Brink, F. D. Dentener, and P. J. H. Builtjes. Secondary inorganic aerosol simulations for Europe with special attention to nitrate. *Atmos. Phys. Chem.*, 4:857–874, 2004.

[98] H. C. Schmidt, C. Derognat, R. Vautard, and M. Beekmann. A comparison of simulated and observed ozone mixing ratios for the summer 1998 in western Europe. *Atmos. Environ.*, 35:6277–6297, 2001.

[99] A. J. Segers. Data assimilation in atmospheric chemistry models. *PhD thesis*, 2001.

[100] D. Simpson, H. Fagerli, J. Jonson, S. Tsyro, and P. Wind. Transboundary acidification, eutrophication and ground level ozone in Europe, Part I. Unified EMEP model description. *EMEP Status Report*, 2003.

[101] R. Stern, P. J. H. Builtjes, M. Schaap, R. M. A. Timmermans, P. Vautard, A. Hodzic, M. Memmesheimer, H. Feldmann, E. Renner, R. Wolke, and A. Kerschbaumer. A model inter–comparison study focussing on episodes with elevated PM10 concentrations. *Atmos. Environ.*, 10:1016, 2008.

[102] O. Talagrand and P. Courtier. Variational assimilation of meteorological observations with the adjoint vorticity equation. I: Theory. *Quart. J. Roy. Meteor. Soc.*, 113:1311–1328, 1989.

[103] L. Tarrason and T. Iven. Modelling intercontinental transport of atmospheric sulphur in the northern hemisphere. *Tellus B*, 50, 1998.

[104] M. K. Tippett, J. L. Anderson, C. H. Bishop, T. M. Hamill, and J. S. Whitaker. Ensemble square root filters. *Mon. Wea. Rev.*, 131:1485–1490, 2003.

[105] P. J. van Leeuwen and G. Evensen. Data assimilation and inverse methods in terms of a probabilistic formulation. *Mon. Wea. Rev.*, 124:2898–2913, 1996.

[106] M. van Loon. Numerical smog prediction, I: The physical and chemical model. *CWI research report NM–R9411*, 1994.

[107] M. van Loon. Numerical smog prediction, II: grid refinement and its application to the Dutch smog prediction model. *CWI research report NM–R9523*, 1995.

[108] M. van Loon. Data assimilation of ozone and aerosols. *Proceedings 5th GLOREAM workshop*, 2001.

[109] M. van Loon, P. J. H. Builtjes, and A. J. Segers. Data assimilation applied to LOTOS: First experiences. *Environ. Modeling and Software*, 15:603–609, 2000.

[110] M. Verlaan and A. W. Heemink. Tidal flow forecasting using reduced–rank square root filter. *Stochastic Hydro. Hydraul.*, 11:349–368, 1997.

[111] A. J. H. Visschedijk and H. A. C. Denier van der Gon. Gridded European anthropogenic emission data for NOx, SOx, NMVOC, NH3, CO, PM10, PPM2.5, CH4 for the year 2000. *NO–Rep BO–AR*, 2005:TNO Apeldoorn The Netherlands, 2005.

[112] X. Wang, C. H. Bishop, and S. J. Julier. Which is better an ensemble of positive–negative pairs or a centered spherical simplex ensemble? *Mon. Wea. Rev.*, 132:1590–1605, 2004.

[113] J. S. Whitaker and T. M. Hamill. Ensemble data assimilation without perturbed observations. *Mon. Wea. Rev.*, 130:1913–1924, 2002.

[114] G. Z. Whitten, H. Hogo, and J. P. Killus. The carbon bond mechanism: A condensed kinetic mechanism for photochemical smog. *Environ. Sci. Tech.*, 14:690–700, 1980.

[115] X. F. Zhang, A. W. Heemink, L. Janssen, P. Janssen, and P. Sauter. A computationally efficient Kalman smoother for the evaluation of the CH4 budget in Europe. *Appl. Math. Modelling.*, 23:109–129, 1999.

[116] D. Zupanski and M. Zupanski. Moder error estimation employing an ensemble data assimilation approach. *Mon. Wea. Rev.*, 134:1337–1354, 2006.

Samenvatting

De atmosfeer is een complex systeem dat fysische, chemische en biologische processen omvat. Data assimilatie is een cruciaal instrument voor het modelleren en voorspellen van dergelijke processen in de atmosfeer. Het verwijst naar een aantal technieken die tot doel hebben waarnemingen te combineren uit verschillende bronnen en naar een mathematisch model voor een uniforme en consistente beschrijving van atmosferische chemie systemen. Het doel van dit proefschrift was om het gebruik te onderzoeken van verschillende sequentiële data assimilatie technieken in zowel ideale als reële opstellingen in het kader van atmosferische chemie. De toegenomen complexiteit van de modellen samen met de informatie over verschillende soorten verontreinigers hebben de potentie om bij te dragen aan het begrip van chemische processen die belangrijk zijn voor mogelijke veranderingen in de samenstelling van de atmosfeer. Een optimale selectie van data-assimilatie technieken is vereist, welke aangepast is aan de nieuwe uitdagingen in de atmosferische chemie. Bovendien is het de bedoeling om manieren aan te geven voor het gebruik van een specifieke data assimilatie methode.

De onderzoeksvragen in dit proefschrift zijn:

1. Hoe kunnen we gebruik maken van de retrospectieve data assimilatie met een model over de atmosferische chemie?
2. Hoe nauwkeurig kunnen de zwaveldioxide en sulfaat concentraties worden geschat met behulp van een enkele component opstelling ten opzichte van de gecombineerde assimilatie procedure?
3. Hoe kan het OMI satelliet product bijdragen tot een beter begrip van

LOTOS-EUROS capaciteiten in het voorspellen van de troposferische ozon?

4. Kunnen we kiezen voor een optimale algoritme vanuit het perspectief van data assimilatie?

De eerste vraag is gerelateerd aan het concept van filtering in tegenstelling tot smoothing. Verschillende algoritmen die de ensemble technieken en het Kalman smoother methode combineren zijn getest in een twin experiment voordat de algoritmen gebruikt werden voor real-life, atmosferische chemie data assimilatie vraagstukken.Van belang is de verbetering van een schatting door het Kalman filter gebruikmakend van toekomstige waarnemingen in de analyse tijd. Inzicht is opgedaan in de efficiëntie door aan te tonen dat de benodigde rekenkracht bij het gebruik van de FIFO lag-algoritme onafhankelijk is van de lag-lengte en daarom efficiënter kan zijn dan de EnKS. Om de tweede vraag te beantwoorden is, in hoofdstuk 3, de zwavel cyclus over Europa geëvalueerd voor het jaar 2003. De vergelijking van het deterministisch model met in situ gegevens afkomstig van de EMEP database heeft laten zien dat de jaarlijkse gemiddelde gemodelleerde sulfaat concentratie systematisch is onderschat. Voor zwaveldioxide worden in het algemeen de jaarlijkse gemiddelde concentraties door het model iets overschat. Andere onzekerheden behalve emissies moeten in acht worden genomen om het verschil tussen het model en de metingen te verklaren. Het is aangetoond dat een stochastisch model op basis van een combinatie van onzekere emissies en reactiesnelheid heeft assimilatie van sulfaat aërosol en haar voorloper gas verbeterd. De assimilatie van meerdere chemische stoffen wordt geadviseerd in het geval van grote onzekerheden in het model die aan verschillende oorzaken kunnen worden toegeschreven. We hebben aangetoond dat de multi-component assimilatie beter presteert met een meer nauwkeurige beschrijving van de modelfout met twee parameters voor de ruis (emissies en reactiesnelheid) in plaats van één (emissies). De experimenten toonden aan dat de filtertechniek in staat is om de modelfout te corrigeren, indien de onzekere modelparameters correct opgegeven waren. Ook het probleem van de reconstructie van zwaveldioxide-emissies en de correctie van reactiesnelheid is behandeld met behulp van het ensemble smoother Kalman methodologie door te tonen dat een smoother kan helpen filteren om de nauwkeurigheid te verhogen.

Meer werk is nodig om het gebruik van ensemble data assimilatie in het verminderen van de onzekerheden in emissie-inventarissen volledig te begrijpen. Een uitdaging vloeit voort uit de lange tijd die nodig is om de programma's te runnen ten einde zinvolle verbanden te ontwikkelen tussen de emissiesnelheden en de concentratie velden. Een andere uitdaging is veroorzaakt door grote onechte correlaties die de filter tot een correctie van de emissie factoren leidt om te compenseren voor andere bronnen van fouten.

Het aantal meetpunten en de dichtheid van het meetnet spelen een belangrijke rol. Vanwege de niet-homogene dichtheid van de gegevens in samenhang met de lokalisatie effect veroorzaakt door het ensemble-filter, wordt de invloed van assimilatie op de toestand en de parameters van het model beperkt tot bepaalde gebieden. Dus kunnen de ruimtelijke effecten van de assimilatie worden verbeterd door het opnemen van een meer uitgebreide set van controle gegevens. Toekomstige studies gewijd aan het bieden van nauwkeurige concentratiekaarten over het beschouwde domein dienen een groter aantal observatiepunten te bevatten. Een meer uitgebreide set gegevens kan worden verkregen met behulp van metingen van de nationale netwerken in combinatie met een strenge kwaliteitsbewaking om ervoor te zorgen dat het gebruik van een observatieset representatief is voor de resolutie van het model.

De derde vraag is het onderwerp van hoofdstuk 4. De NO_2 troposferische gegevens van het Ozone Monitoring Instrument zijn gebruikt met de LOTOS-EUROS model. Het is aangetoond dat de ruimtelijke patroon van stikstofdioxide die door modelsimulatie is berekend een goede overeenkomst heeft met de OMI dataset, maar met een grote negatieve bias. De belangrijkste bronnen van deze bias zijn gevonden in de beperkte kennis van de NO_2 chemie, in het bijzonder de levensduur parameter. Het is aangetoond dat de discrepantie tussen de gesimuleerde en de waargenomen troposferische NO_2 kolommen aanzienlijk wordt verminderd door het opnemen van een data assimilatie schema die voor de bias corrigeert. De veranderingen in stikstofdioxide bepalen een algemene toename van ozon waarden, met positief effect op de O_3 maxima. Vanuit operationeel oogpunt kan dit feit met name relevant zijn voor de prognose van hoge ozon waarden die overeenkomen met grotere vervuiling episodes.

Hoofdstuk 5 richt zich op de laatste vraag. Onze doelen waren om de prestaties van stochastische, semi-deterministische en deterministische filters met twee modellen te onderzoeken en om inzicht te krijgen in het gedrag van de symmetrische algoritmen door vergroting van het begrip van de eigenschap van de kleinste toename van de toestand. De symmetrische versie van de ESRFs met de kleinste analyse toename over alle semi-deterministische filters is de meest nauwkeurige oplossing ten opzichte van een eenzijdige, willekeurige rotatie variant of een variant van willekeurige rotatie met behoud van gemiddelde. In verschillende toepassingen kan het toevoegen van willekeurige rotaties een beter resultaat geven ten opzichte van een eenzijdige oplossing. Bovendien kunnen willekeurige rotaties met behoud van gemiddelde een verbetering van de schatting betekenen ten opzichte van de variant met willekeurige rotaties, maar dit is niet noodzakelijkerwijs het geval voor alle toepassingen. Het blijkt dat de symmetrische ESRF zich altijd onder de best presterende algoritmen bevindt van de vier hier geteste ESRFs. Bovendien zijn de rekenkrachteisen niet significant groter voor de symmetrische ESRF, noch is het moeilijker om te implementeren. Daarom adviseren wij het gebruik van de symmetrische ESRF bij de toepassing van ESRF-type algoritmen met het LOTOS-EUROS model.

Naar analogie van de symmetrische ESRF, een symmetrische RRSQRT filter is ingevoerd. De resultaten verkregen met dit algoritme zijn veel nauwkeuriger dan de originele (eenzijdige) versie. Het is raadzaam voor de bestaande en nieuwe implementaties van de RRSQRT algoritme om de code te wijzigen naar de symmetrische vorm doordat deze duidelijk nauwkeuriger, meer reproduceerbaar en minder vatbaar voor numerieke fouten is. Bovendien zijn de noodzakelijke veranderingen niet moeilijk en vereisen geen significante hoeveelheid rekenkracht. Voor toepassingen waarbij de modelfouten niet te verwaarlozen zijn biedt de symmetrische RRSQRT filter een goed alternatief voor de ESRF. Deze methode is getest in een twin-experiment, maar nog niet in een grote schaal toepassing. Dit kan een uitdagend probleem zijn. In veel van de data assimilatie toepassingen werd de onzekerheid van het model in een vereenvoudigde vorm behandeld. Voor vele modellen, zoals hydrologische modellen, land modellen met onzekerheden in de atmosferische forcing, of atmosferische chemie modellen met onzekere bronnen van verontreiniging speelt de modelfout een belangrijke rol in data assimilatie. Daarom moet meer aandacht worden besteed aan gedetailleerde karakterisering van onzekerheden in het modelleren en voor dit doel kan data assimilatie worden beschouwd als een nuttig instrument.

Acknowledgements

This work was supported of by the Netherlands Institute for Space Research (SRON) under the EO-063 project. The research presented in this thesis was carried out in the section of Mathematical Physics at Delft Institute of Applied Mathematics in cooperation with TNO Built Environment and Geosciences, Department of Air Quality and Climate, Utrecht.

I am so much indebted to so many people and in so many ways, that it would be impossible to name them all there. I just hope I manage to mention most of those who have made it possible for me to finish this thesis.

First of all, I would like to thanks both of my advisors Arnold Heemink and Peter Builtjes for the opportunity they have given me to pursue the PhD study. I am deeply grateful to Arnold Heemink for his knowledgeable and kind guidance and his understanding of every problems I had in my professional and personal life. I am also grateful to Peter Builtjes for his support and encouragement. I thank them as well for the careful reading and constructive comments on the manuscript.

Working with Martin Verlaan has been a rewarding experience, and a number of issues presented in this thesis are related to it. Special thanks go to him for the interesting and informal discussions during our weekly meetings that helps me to discover how beautiful topic data assimilation is.

I would like to express my gratitude to Remus Hanea who motivated me to start this research and helped me find my own way.

A list that has far too many names on it to mention separately is that of all the co-workers, group members, and roommates that I have worked, talked, and lunched with over the years. My gratitude goes out to all these colleagues and former colleagues at the DIAM group.

I am especially grateful to Martijn Schaap and Arjo Segers for their skilful guidance during my work at TNO. Martijn contributed to my work with a lot of energy, inspiration and attention. I probably owe most of my scientific writing skills from him. Special thanks go to Arjo for his countless contribution, patience and support. I attribute to his assistance that I was able to use and modify LOTOS-EUROS codes for my experiments. It was very encouraging for me that even during their busy time, Arjo and Martijn remained accessible and answered any issues that I raised.

I would also like to thank all people from the TNO group for our fruitful interactions. I really enjoyed the time spent at TNO.

I would like to thank Evelyn Sharabi and Dorothee Engering for their readiness to help with bureaucratic documents and procedures. I am indebted to Kees Lemmens for his initiative of organising lectures on programming, but especially for getting me out of technical troubles.

Many thanks go to my Romanien friends spread out in the world, but next to me when needed. The community gathered around the Romanian Orthodox Church from the Netherlands has always provided a pleasant atmosphere.

I am especially grateful to all those who offered me precious help and advice in day-to-day life.

I am, of course, particularly indebted to my parents and my family for their constant support and encouragement. *Le multumesc parintilor mei pentru sacrificiu si incredere. Le multumesc copiilor mei, Andrei and Ştefan pentru bucuria si optimismul pe care mi le-au insuflat.*

Delft, *Alina Barbu*
April 2010

Curriculum Vitae

Alina Lavinia Barbu was born in Găeşti, Romania, on February 25, 1972. In 1990 she started her studies at the Department of Mathematics, Faculty of Mathematics, University of Bucharest and received the Bachelor's degree five years later. In 1996 she received the MSc. degree in the field of Differential Geometry from the same faculty. After graduation she joined as a staff member of the Department of Mathematics at the Technical University of Civil Engineering, Bucharest. In 2003 she defended the second Master thesis: " Develop an automatic classification of directional data for the non-destructive control of the spherical reservoirs " and obtained M.Sc degree in Applied Mathematics at the Technical University of Compiégne, France.

In September 2004 Alina Barbu started a PhD reseach program at Delft Insitute of Applied Mathematics, Delft University of Technology. She worked on a project between the TU Delft and TNO Built Environment and Geosciences, Department of Air Quality and Climate, Utrecht. The focus of her study was on the application of data assimilation for large scale atmospheric chemistry models, under the supervision of Prof. Arnold W. Heemink and Prof. Peter J.H. Builtjes. This work results in this thesis to be defended on 22th April 2010. Since 2009, she has been working as a post-doc in the Mesoscale Meteorology Division (GMME) of the National Research Meteorological Center (CNRM) at Météo France. She has been developing a data assimilation system of remote sensing over land surfaces for meteorology.