Testing data assimilation methods in idealized models of moist atmospheric convection

Dissertation
an der Fakultät für Physik
der Ludwig-Maximilians-Universität
München

vorgelegt von

Michael Würsch
geboren in Aarau, Schweiz

München, September 2013

Erstgutachter: Prof. Dr. G. C. Craig
Zweitgutachter: Prof. Dr. R. Smith
Tag der mündlichen Prüfung: 2.12.2013

Parts of this thesis are published in:
Craig and Würsch (2013)
Würsch and Craig (2013)

Contents

Zusammenfassung

Das Ziel dieser Arbeit war es, eine rechenkostengünstige aber physikalisch plausible Entwicklungsumgebung für Datenassimilation auf der konvektiven Skala bereitzustellen. Dabei sind die Modelle Teil einer Hierarchie von Testmodellen mit zunehmender Komplexität. Sie stellen die extreme Nichtlinearität und Nicht-Gaußheit dar, welche mit den sich schnell entwickelnden und intermittenten konvektiven Gewittern verbunden sind.

Zuerst wird ein einfaches stochastisches Modell basierend auf einem Poisson "birth-death"-Prozess eingeführt. In diesem Rahmen werden der Ensemble Transform Kalman Filter (ETKF) und der Sequential Importance Resampling (SIR) Filter evaluiert. Zusätzlich werden zwei Strategien zur möglichen Verbesserung der Methoden ausgewertet: Lokalisierung und Beobachtungsmittelung. In ihren Grundformen funktionieren beide Filter schlecht. Der SIR kollabiert schnell und konvergiert danach nur langsam zu den Beobachtungen hin, wobei zufällige Störungen durch das Resampling die Analyse gelegentlich verbessern. Der ETKF assimiliert die richtigen Positionen von Wolken schnell, hat aber trotzdem einen großen Fehler aufgrund des Erzeugens von falschen Wolken. Diese entstehen durch nichtlineare Verstärkung kleiner Datenassimilationsinkremente. Lokalisierung, das heißt Assimilieren von lokalen Beobachtungen, erzeugt eine drastische Verbesserung der Leistung des SIR Filters, während der Fehler beim ETKF nicht kleiner wird. Beobachtungsmittelung, das heißt örtliches Glätten der Beobachtungen vor der Assimilation, macht die Verteilungen gaußförmiger. Dies ist erfolgreich beim SIR Filter und verbessert die Konvergenz auch beim ETKF.

Das zweite Modell ist ein vereinfachtes Modell für Konvektion, basierend auf den Flachwassergleichungen. Zusätzliche Prozesse wie Schwerewellen, bedingte konvektive Instabilität und Niederschlagsbildung sind enthalten. Durch Anpassung von Parameterwerten können die wichtigsten räumlichen und zeitlichen Skalen von Konvektion nachgebildet werden. Das Modell ist in der Lage, den typischen Lebenszyklus einer konvektiven Zelle zu simulieren. Falls mit einer Störungsquelle kleiner Amplitude Konvektion ausgelöst wird, stellt sich im Modell ein statistischer Gleichgewichtszustand ein, der für die Wolkengrößen und -abstände demjenigen entspricht, der mit wolkenauflösenden Modellen unter Annahme des Strahlungs-Konvektions-Gleichgewichts entsteht. Falls Konvektion durch die Überströmung eines orographischen Hindernisses ausgelöst wird, gibt es zwei Zustände. Abhängig von der Windstärke ist die Konvektion entweder begrenzt auf das Gebirge oder bewegt sich stromabwärts.

Das Modell enthält prognostische Variablen für Wind und Regen, die zur Berech-

nung synthetischer Beobachtungen in Datenassimilationsexperimenten benutzt werden können. Solche Experimente wurden mit einem LETKF mit 50 Ensemblemitgliedern und zwei verschiedenen Szenarien für die Beobachtungen durchgeführt. Ein Szenario, in welchem die Beobachtungen hoch aufgelöst sind und eines mit gröberen "Super-Beobachtungen". Die Assimilation wird während den ersten 36 Minuten durchgeführt gefolgt von weiteren 36 Minuten freier Vorhersage. In dem feinen Experiment sind die meisten Regengebiete in der Analyse innerhalb weniger Assimilationszyklen erfasst. Das grobe Experiment gelangt auch zu einem guten Ergebnis, die Regengebiete sind aber etwas verschmierter und es gibt mehr Streuung im Ensemble. Der starke Einfluss der Beobachtungen produziert außerdem unphysikalische Zustände in der feinen Analyse. Dieses Problem kann entweder durch Verwendung der groben Analyse oder eines größeren Lokalisierungsradius gelöst werden.

Während der Phase der freien Vorhersage verliert die feine Analyse ihren Vorteil gegenüber der groben innerhalb von 10 bis 20 Minuten. Damit wird die Notwendigkeit einer hoch aufgelösten Analyse in Frage gestellt. Insgesamt zeigen die Experimente viele Gemeinsamkeiten zu Datenassimilationsläufen welche in einem wolkenauflösenden numerischen Wettervorhersagemodell erzeugt wurden.

In dieser Arbeit werden Testmodelle für Datenassimilation auf konvektiver Skala entwickelt, getestet und verifiziert. Sie bilden die wichtigsten Eigenschaften von konvektiver Bewegung und damit verbundenen Phänomenen nach. Außerdem wird aufgezeigt, was typische Probleme von konventionellen Datenassimilationsmethoden in einer solchen Umgebung sind, und dass eine hoch aufgelöste Analyse nicht unbedingt zu den besten Ergebnissen führt.

Abstract

Two different models are proposed in this thesis. They are intended to provide a computationally inexpensive, but physically plausible, environment for developing methods for convective-scale data assimilation and are part of a model hierarchy with increasing complexity. The models mimic the extreme nonlinearity and non-Gaussianity associated with rapidly developing and intermittent convective storms. They should provide a computationally inexpensive, but physically plausible, environment for developing methods for convective-scale data assimilation and be part of a model hierarchy with increasing complexity.

First, a simple stochastic model which is based on a Poisson birth-death process is introduced. In this framework, the Ensemble Transform Kalman Filter (ETKF) and Sequential Importance Resampling (SIR) Filter are evaluated. The impact of two strategies to improve their performance and efficiency is assessed: localization and observation averaging. In their basic implementations, both filters perform poorly. The SIR filter rapidly collapses, then very gradually converges to the observations as random perturbations introduced by resampling occasionally improve the analysis. The ETKF rapidly assimilates the correct locations of convective storms, but has large errors due to creation of spurious clouds by nonlinear amplification of small data assimilation increments. Localization, i.e. assimilating only local observations to produce the analysis at a given grid point, dramatically improves the performance of the SIR filter, but does not reduce errors in the ETKF. Observation averaging, i.e. spatially smoothing the observations before assimilation and thus making the distribution more Gaussian, is also effective for the SIR filter, and improves convergence of the ETKF.

The second model is a simplified model for cumulus convection based on the shallow water equations. Additional processes, including gravity waves, conditional instability and precipitation formation, are represented, and parameter values are chosen to reproduce the most important space and time scales of cumulus clouds. It is shown that the model is able to reproduce the classic life cycle of an isolated convective storm. When provided with a low amplitude noise source to trigger convection, the model produces a statistically steady state with cloud size and cloud spacing distributions similar to those found in radiative-convective equilibrium simulations using a cloud resolving model. In addition convection triggered by flow over an orographic obstacle is investigated. Depending on the wind speed two regimes are found with convection trapped over the mountain, or propagating downstream.

The model features prognostic variables for wind and rain that can be used to compute synthetic observations for data assimilation experiments. Such experiments

have been performed using an LETKF with 50 ensemble members and two different observation settings. One, where the observations have a high resolution and one with coarser superobservations. The assimilation experiments are done for 36 minutes with another 36 minutes of free forecast. In the fine experiment most of the rain regions are captured in the analysis within a few assimilation cycles. The coarse experiment ends up with a good result as well, but the rain regions are blurred out and there is more spread in the ensemble. The strong influence of the observations in the analysis produces unphysical states in the fine analysis. This issue can be solved by using the coarse analysis or a larger localization radius.

During the free forecast phase, the fine analysis loses the advantage over the coarse run within 10 to 20 minutes, questioning the need for a high resolution analysis. Overall, the experiments show many similarities to data assimilation runs performed with a cloud-resolving numerical weather prediction model. Therefore the modified shallow water model is verified as a useful test model for convective-scale data assimilation.

In this thesis test models for convective-scale data assimilation are developed, tested and verified. They show the most important characteristics of convective motion and related phenomena. It is also illustrated what some of the problems of conventional data assimilation methods are in such an environment and that choosing a high-resolution in the analysis might not necessarily lead to the best result.

Chapter 1

Introduction

Since the first use of computers to describe the evolution of the atmosphere (Charney et al., 1950), the resolution of Numerical Weather Prediction (NWP) models has continuously been enhanced. During the last years a big improvement came from the explicit modeling of convection (Done et al., 2004; Weusthoff et al., 2010) which was possible when the resolutions reached the kilometre-scale. To be able to produce forecasts at this scale, a detailed analysis of the state of the atmosphere including convection is indispensable. Current research shows that nowcasts (short-range forecasts using extrapolated radar and satellite observations) of precipitation are more skillful than Quantitative Precipitation Forecasts (QPF) of models for up to 6 hours after the analysis (Lin et al., 2005; Kober et al., 2012). Forecasting in this temporal and spatial range is important for the release of warnings to minimize the impact of hazardous weather on humans as well as on the economy.

To produce a good weather forecast one needs accurate initial conditions for the model. This was already shown 50 years ago by Lorenz (1963) in his famous experiments on predictability which proved that the evolution of the atmospheric state is very sensitive to the initial conditions. An analysis of the current state of the atmosphere showing small-scale features requires observations of a high spatial and temporal resolution which conventional data types do not provide. Radar is an important additional source of information which contributes a lot of supplementary knowledge which is only used partly or not at all in operational data assimilation until now.

If the observed quantity of a system is not a state variable of the model, which is the case for quantities observed by radar, it must be inferred instead. "Data assimilation" is the general term referring to methods which provide an initial state estimate based on observations and it will subsequently be called convective-scale data assimilation

if used at the kilometre-scale. Latent Heat Nudging (LHN) was the first data assimilation method at these small scales to be used operationally (Jones and Macpherson, 1997; Macpherson, 2001; Leuenberger and Rossa, 2007; Stephan et al., 2008) and to include information derived from radar observations. The impact duration of radar rainfall assimilation using latent heat nudging and its constraints are presented in Craig et al. (2012). It has been shown that the limitation in the predictability is mainly due to the fast growth of errors at the smallest scales which amplify very fast due to nonlinear processes that are present for example in the explicit microphysical parameterizations (Zhang et al., 2003; Hohenegger and Schär, 2007a,b).

A number of problems arise when assimilating radar observations. First, convective cells and their associated observed precipitation fields can be distributed very irregularly. They occupy only a small fraction of the model domain and can have large spatial displacement errors as well as false alarms or missed events in the analyzed background state. This can lead to highly non-Gaussian background error statistics and also produces the well-known "double penalty" problem (Anthes, 1983). When a cloud has a small error in the position, there will be an error at the correct position because the cloud is missing, and an additional error where the cloud is, although there should not be one. A second problem is that since rain particles are only detectable for meteorological radars approximately 15-30 minutes after the initialization of an updraft (Houze, 1993), there is a time lag in the detection which leads to a nonlinearity in the observed field between different observation times. This is the so-called late detection problem and it causes issues because at the time the radar sees the rain the dynamical circulation of the thunderstorm is already fully developed and there is a lag in the collection of information. Another factor enhancing the nonlinearity of the observed field is the rapid development of convective clouds and the nonlinear relationship between observables like radar and dynamical variables like vertical velocity and water content. Therefore error growth between two observation times can be significant and cumulus clouds seem to pop up out of the nowhere. This lack of temporal correlation results in an essentially stochastic evolution of the field, as previously unobserved features suddenly appear as precipitating clouds.

An example of the intermittency of a convective radar field can be seen in Fig. 1.1 which shows one-hourly accumulated rainfall estimated from radar reflectivities. One can see a number of isolated thunderstorms which do not seem to be connected. They vary a lot in size, shape and intensity. Such a field contains strong gradients over distances comparable to a few grid points. One reason why the intermittency

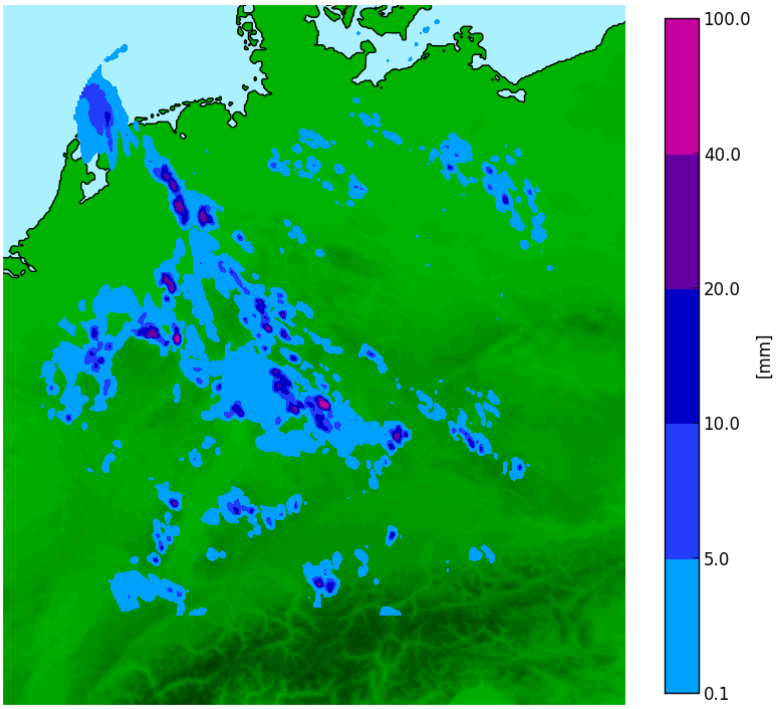

Figure 1.1: One-hourly accumulated rainfall estimated from radar reflectivities of the german radar composite at 15 UTC on June 5th, 2011.

causes problems is that radar has a high spatial resolution comparable to the model resolution, leading to the intermittency happening in the range of only a few grid points. The model resolution of 2.8 km, which will be important in this work, is in the so-called "gray zone", meaning that convection is only partly resolved. Shallow convection for example is still parameterized in contrast to deep convection which is permitted. Processes like turbulence in the convection are still not resolved well enough (Bryan et al., 2003; Bryan and Morrison, 2011) and the parameterization schemes rely on large-scale hypotheses which are also not true anymore due to the too high resolution. This is a known issue since more than 15 years (Kuo et al., 1997). As most of the common data assimilation methods are based on Gaussianity and linearity, the points mentioned above are an important handicap for the performance of these systems. Classic models with lower resolutions do not suffer from these problems because they are focused on a good representation of the large-scale synoptic flow. Therefore considerable effort is now being expended on developing new algorithms and methods to deal with the challenges mentioned in the last Sections. Dance (2004) and Bocquet et al. (2010) discuss recent advances and issues in high resolution limited area data assimilation.

Variational data assimilation methods like Four Dimensional Variational Assimilation (4D-Var) (Talagrand and Courtier, 1987; Bouttier and Rabier, 1998; Bouttier and Kelly, 2001) have proven very succesful in producing good analysis states in global weather models. Variational methods are minimizing a cost function which penalizes deviations of the model to the observations weighted by covariances. The model state at this minimum then is called analysis. The distribution of the background error is assumed unimodal and represented through an error covariance matrix. Although theoretically this matrix has a size which is not manageable by computers, its low rank approximations have become very accurate and are being used operationally by many meteorological centers. Especially on the synoptic scale it is possible to reduce the size of the matrix by balance assumptions and observations of correlations. One reason for this is the smoothness of the fields on these larger scales.

The Kalman Filter (Kalman, 1960; Kalman and Bucy, 1961) is another established data assimilation method where in an advancement called Ensemble Kalman Filter (EnKF) (Evensen, 1994; Houtekamer and Mitchell, 1998, 2001) the error covariance matrix is approximated through a representation with a small ensemble of model runs. In both methods the temporal evolution of the error distribution is represented by approximations. In 4D-Var it is done by a tangent linear model and in the EnKF by the evolution of an ensemble. As shown in the previous paragraph, the dynamics of convective clouds at small scales strongly violate the assumptions these simplifications rely on. The spatial intermittency of convection destroys the spatial correlations and therefore the reduction in the effective number of degrees of freedom of the system (Patil et al., 2001; Ott et al., 2002).

A third type of data assimilation methods are the so-called particle filters (Metropolis and Ulam, 1949; van Leeuwen, 2009) which are sequential Monte Carlo methods. They sample the state by using an ensemble of model forecasts which are weighted by observations and are fully nonlinear both in model evolution and analysis step. Particle filters belong to a family of methods which is believed to have much potential for future use but still needs further development. One existing problem is the so-called “curse of dimensionality” (Bellman, 1957, 1961) which states that the number of possible states of a dynamical system increases exponentially as the dimensionality of the system grows. It has been shown that a small number of model runs, in this context called particles, is very unlikely to be close to a large set of observations if it is shooting at random through state space (Bengtsson et al., 2008; Bickel et al., 2008; Snyder et al., 2008). This is even more true if the observations are in a high-dimensional system. Therefore a general particle filtering method like Sequential

Importance Resampling (SIR) (van Leeuwen, 2009) requires a number of particles which is not feasible for operational use. The gain in generality comes at the cost of computational efficiency.

All the advantages and disadvantages of these data assimilation methods play an important role when they are applied on the convective-scale. They can be summed up as the key aspects of convective-scale data assimilation in two points. First, it is the nature of convection with its nonlinearity and non-gaussianity causing intermittent fields both in time and space. Second, it is the theoretical formulation of data assimilation methods which reach their limit when applied to a convective situation. More details on the different data assimilation methods will be given in Chapter 2.

These aspects indicate that the one data assimilation method for use at the convective-scale has not been identified yet. Therefore when new methods are being developed, a lot of testing is needed. As running full atmospheric models is computationally very expensive, it is a preferred approach to identify basic characteristics of certain methods in an idealized and cheaper model setup first. Popular choices are for example the models of Lorenz (1995, 2005) which contain the coupling between fast and slow variables in a low-dimensional dynamical system and capture key aspects of nonlinearity. A little bit more complex are models based on the quasi-geostrophic equations (Ehrendorfer and Errico, 2008) which have proven useful in their realistic representation of the synoptic-scale structures found in the atmosphere. However, as both types of models are designed for the representation of larger scales, they do not capture the spatial intermittency and other characteristics of the convective-scale. Therefore a suitable test model needs to consider both the nature of the non-Gaussianity and nonlinearity of a convecting atmosphere. Succesful examples so far which represent this setting are simulations of idealized storms in a "perfect model" configuration with EnKF assimilation of radar data (Snyder and Zhang, 2003; Zhang et al., 2004; Tong and Xue, 2005; Potvin and Wicker, 2012). In similar experiments, Lange (2013) used the Kilometer-scale Ensemble Data Assimilation (KENDA) system developed by the German Weather Service (Deutscher Wetterdienst (DWD)) (Reich et al., 2011). In this setup the Local Ensemble Transform Kalman Filter (LETKF) of Hunt et al. (2007) was used to assimilate observations of idealized storms that developed in a convective environment defined by Bischof (2011). There are also simplified experiments based on real events which use EnKF assimilation of high-resolution observations to get better analysis and forecasts of super cell storms (Dowell et al., 2004; Aksoy et al., 2009, 2010; Dowell et al., 2011; Dawson et al., 2012).

All these experiments are still computationally expensive and, although useful, only a limited number of data assimilation methods, observation configurations and initialization settings can be investigated. Therefore there is the need of computationally very cheap models which might be simple enough to trace the origins of typical patterns of errors. Ideally, one would like to have a hierarchy of simple models in growing order of complexity. These models should be designed to represent the dynamics and physics of the atmosphere on the convective-scale, and in particular the key processes of cumulus convection. These models should produce results of data assimilation experiments which still hold when compared to the results of complex atmospheric models.

This thesis introduces two kinds of simple models which try to capture the mentioned characteristics of convection. The first is a stochastic model which is based on a spatial Poisson birth-death process and the second a modified shallow water model. The stochastic model only captures the spatial intermittency and nonlinearity of a developing convective cloud field. It can be regarded as a minimal version of the stochastic cumulus parameterization scheme of Plant and Craig (2008), which is based on a statistical mechanics theory of convective fluctuations (Craig and Cohen, 2006). Their parameterization produces a spectrum of entraining and detraining plumes with a stochastic variability. The mass flux of the clouds is drawn from a Probability Distribution Function (PDF) which defines the chance of finding a plume with a given mass flux in the model grid box. In the stochastic model presented in this work the distribution is produced by the cumulus clouds appearing and disappearing at random locations with given triggering and dissolving frequencies. In the resulting field, the clouds are located at random positions with a random lifetime, but on average the number of clouds and the lifetime of a cloud are determined by the birth and death probabilities.

In this very simple model some physical processes that are important in nature are missing. The most important ones being the processes responsible for triggering convective clouds and the coupling to the larger-scale flow. Therefore the second model presented in this work goes one step further and includes more physics but is still of intermediate complexity and computationally cheap. It is a modified shallow water model which contains spatial correlations between dynamical variables. This behaviour is not present in the stochastic cloud model but is relevant for testing the ability of an assimilation algorithm to accurately represent background error covariances. This second model is based on the equations of fluid motion and will be extended to represent conditional instability. If there happens sufficient upward movement, there

is a representation of a cloud formed. This cloud can also disappear and the whole process emits gravity waves, as there is the underlying water surface. An additional variable called rain water acts as a negative buoyancy and forces the clouds to dissolve. One goal of the modified shallow water model is to not only show a qualitatively resemblance to convective motions, but also to accurately represent the key space and time scales of storm evolution in the real atmosphere. The numerical implementation for such a model can be developed with similar methods as used in NWP models and the presence of time and space as variables leads to better possibilities for comparison to real convective life cycles.

Both models presented in this thesis can also be used to predict their state variables. Observations which are taken from a model run and used in data assimilation experiments will be called synthetic observations henceforth. In the stochastic cloud model this would be the number of clouds. The synthetic observations of the modified shallow water model are for example horizontal velocity, water level or rain water concentration which can be processed into an analogous to Doppler winds and radar reflectivity. Therefore another part of this work focuses on results of data assimilation experiments obtained with the two models. Two data assimilation methods will be applied in their basic and localized forms. The Ensemble Transform Kalman Filter (ETKF), LETKF and the SIR are used for assimilation in the stochastic cloud model and the LETKF in the modified shallow water model. These methods are chosen because they rely on different approximations in trying to solve the same problem. Based on theory, one can expect the particle filter to depend a lot on the effective dimensionality of the system. For the EnKF methods a prediction seems more difficult, as they are applied well outside of their regime of validity and may not even work at all. The localized versions of the filters are used to see whether exploration of local subspaces leads to an improvement by producing more dimensions in total. Averaging is applied to see whether it is beneficial to discard information but improve the Gaussianity. It is important to use more than one test model to see whether the data assimilation algorithms might function more effectively in a more realistic environment and how much the results depend on the model.

Summarized there are three main scientific questions which are treated in this thesis:

- Is it possible to create parts of a model hierarchy where the models are computationally cheap but still represent key features of convective-scale movement?
- How do data assimilation methods like the SIR and ETKF behave in such test

models and to which parameters is their performance sensitive? What are the characteristics of typical problems that these methods encounter? Is a high-resolution analysis needed or is a coarse resolution scheme with averaged observations sufficient?

- Can the results of data assimilation experiments in simple models be used to predict the behaviour of the methods in high-resolution atmospheric models?

It is also worth mentioning a couple of points which are not covered in this work. First, the aim is not to determine which data assimilation method is the best, but rather shed some light on the characteristics of the different methods and identify typical error patterns. Then there are no experiments with sophisticated variants of particle filters like the nudging particle filter (van Leeuwen, 2010). The methods used in this work are established and also serve as a test for the two models. Finally, variational methods are not used, as they rely on a tangent linear model which makes them especially sensitive to the problems of radar assimilation. A thorough comparison between 4D-Var and EnKF techniques for radar data assimilation is provided by Caya et al. (2005).

The thesis is organized as follows: Chapter 2 introduces data assimilation in general together with the specific methods used in this thesis. Chapter 3 presents a description of the stochastic cloud model and the evaluation of the data assimilation experiments obtained with this model. In Chapter 4 the modified shallow water model is introduced and its general behaviour as well as some statistical analysis is presented. The results of data assimilation experiments achieved with the modified shallow water model are analyzed and compared to a high resolution model in Chapter 5. Chapter 6 concludes the thesis with a summary of the main results and an outlook on possibilities for further research.

Chapter 2

Methods

This Chapter presents an overview on data assimilation in general in Section 2.1 as well as the methods used in this work. The overview on Kalman filter methods in Section 2.2 is adapted mainly from the publication of Hunt et al. (2007) while the short overview on particle filter methods in Section 2.3 is based on van Leeuwen (2009) where one can find a much more detailed description of more sophisticated variants. Step-by-step explanations of the implementations used in this work are given for both methods. As model error is not treated in this thesis the reader is refered to the works of Palmer (2001), Mitchell et al. (2002) and Houtekamer et al. (2009) for more insight into the topic of model error. Section 2.4 concludes this Chapter with a short summary.

2.1 Data assimilation

The goal of meteorological data assimilation is to estimate the state of the atmosphere and its uncertainty using as many as possible of the available observations. Usually such an estimate is repeated on an ongoing basis as new observations are being used on a regular basis. Therefore data assimilation is alternating between a forecast step and a state estimation step. This approach is also called sequential data assimilation. The estimated state is often called "analysis" and displayed on the grid of a NWP model. Therefore not only information of observations can be used but also background information coming from previous analyses. In the most simple case the observations would only be interpolated onto the model grid. But this procedure neglects the errors that are present both in the observations and in the model. More sophisticated methods apply a statistical procedure to combine observations and background information in

Table 2.1: Vectors, operators and matrices used in Data Assimilation

Variable	Name
$\mathbf{d}$	innovation
$\mathbf{x}$	m-dimensional State vector
$\mathbf{y}^o$	l-dimensional observation vector
$\boldsymbol{\epsilon}$	observation error
H	observation operator
J	cost or penalty function
M	model operator
$\mathbf{H}$	linearized observation operator
$\mathbf{I}$	Identity matrix
$\mathbf{K}$	Kalman gain matrix
$\mathbf{M}$	linearized model operator
$\mathbf{P}$	error covariance matrix
$\mathbf{Q}$	model error covariance
$\mathbf{R}$	observational error covariance

Table 2.2: Super- and Subscripts

Symbol	Definition
$(\cdot)^b$	Background, also called prior or first guess
$(\cdot)^a$	Analysis
$(\cdot)^{a(i)}$	Analysis ensemble member i with $i = 1, 2, ..., k$
$(\cdot)_n$	Situation at time t_n

an optimal way.

Before the different methods are presented, the most important terms and their notation are explained in tables 2.1 and 2.2. They are based on Ide et al. (1997) where a unified notation for data assimilation is proposed. Vectors and matrices are always written in bold font. Vectors with lowercase letters and matrices with capital ones.

Usually the model system where data assimilation is applied, is governed by an

ordinary differential equation:

$$\frac{d\mathbf{x}}{dt} = F(t, \mathbf{x}) \tag{2.1}$$

Observations of this system at various times lead to a trajectory $\mathbf{x}^a(t)$ of states that "best" fit them. In a mathematical formulation of this problem, "best" needs to be defined first. To get to such a definition, the relationship between observations and the true model state needs to be specified as:

$$\mathbf{y}_j^o = H_j(\mathbf{x}(t_j)) + \boldsymbol{\epsilon}_j,$$

The observation error $\boldsymbol{\epsilon}$ is assumed to be a random variable with mean $\mathbf{0}$ and covariance matrix $\mathbf{R}$. H_j is called observation operator and therefore the observation equals the model state transformed into the observation space plus the observation error. In the simple case considered here the model error is neglected. Of course this is not true in reality as the model in 2.1 is an imperfect image for the atmosphere. But as the model experiments presented in this work are so-called "perfect model" experiments, and for simplicity, model error will be neglected from now on. Especially since the size and nature of the model error can deviate a lot from Gaussian statistics.

With these assumptions one can formulate the analysis as the "most likely" model state. This is the so-called variational approach where the maximum likelihood is estimated for a trajectory of 2.1 that is the best fit to the observations over certain observation times. This likelihood is proportional to the product of all Gaussian distributions at the observation times. Therefore the maximum likelihood is achieved for a maximum of this product (being the place with the highest probability) which is equivalent to the minimum of the exponents. The term which needs to be minimized is called "cost function" and takes the following form:

$$J(\mathbf{x}(t)) = \sum_{j=1}^{n} [\mathbf{y}_j^o - H_j(\mathbf{x}(t_j))]^T \mathbf{R}_j^{-1} [\mathbf{y}_j^o - H_j(\mathbf{x}(t_j))]. \tag{2.2}$$

The difference between the observations and the model state in observation space $\mathbf{d}_j = [\mathbf{y}_j^o - H_j(\mathbf{x}(t_j))]$ is also called innovation. It is then weighted with the errors of the individual observations. The most likely trajectory is therefore the one which best fits the observations in a least square sense.

As the model states at the different times need to be calculated, one needs to introduce a map that propagates a state from time t to time t'. This map is defined

as $M_{t,t'}$ and can be used to rewrite $\mathbf{x}(t_j) = M_{t,t_j}(\mathbf{x})$ which only leaves the initial model state in the equation.[1] In this formulation it becomes clearer to see how the information of all past observations are used.

A couple of problems arise when one is trying to calculate the minimum to find a state estimate. First, atmospheric models are nonlinear and it is possible that there might be several minima in the cost function. Second, evaluating J and finding a minimum at all is computationally very expensive as the state space has dimension of the number of observations which can go up to 30 million per day. A simplification of the process is possible if both the model and the observation operators are assumed to be linear. Additionally instead of calculating the minimization for the whole trajectory at each time, it can be done iteratively. The Kalman filter (Kalman, 1960; Kalman and Bucy, 1961) is the method which does so and is described now.

2.2 Kalman Filter

Instead of using the full operators as in the previous Section, they are assumed to be linear in the Kalman filter and therefore can be written as matrices: $M_{t,t'}(\mathbf{x}) = \mathbf{M}_{t,t'}\mathbf{x}$ and $H_j(\mathbf{x}) = \mathbf{H}_j\mathbf{x}$. An analysis at a time t_n is now being calculated by using the analysis of time t_{n-1} and the observations at time t_n. These times will be used for the upcoming formulation. But one still needs to take into account the information of the uncertainty of all prior observations which has been propagated to time n. Therefore the most likely state and its uncertainty is still required. These contain the required information of all previous observations and the cost function now includes only a term with the analysis and its uncertainty as well as the latest observations and their uncertainty. The sum of all previous observation times (as in equation 2.2) will therefore vanish in the final formulation of the cost function.

With these basics it is now possible to write down precise statements for the calculation of a new analysis state. As a starting point it is assumed that the state estimate $\bar{\mathbf{x}}^a_{n-1}$ and the associated covariance matrix $\mathbf{P}^a_{n-1}$ are known from a previous analysis. These values are the outcome of the maximum likelihood method mentioned before and therefore represent the mean and covariance of a Gaussian probability distribution.

As a next step the background state can be calculated by propagating the analysis

[1] In this framework $\mathbf{x}$ is just the system state at general time t.

and the the error covariance:

$$\bar{\mathbf{x}}_n^b = \mathbf{M}_{t_{n-1},t_n}\bar{\mathbf{x}}_{n-1}^a, \tag{2.3}$$

$$\mathbf{P}_n^b = \mathbf{M}_{t_{n-1},t_n}\mathbf{P}_{n-1}^a\mathbf{M}_{t_{n-1},t_n}^T. \tag{2.4}$$

As the model is considered linear, this transformation of a Gaussian mean and its covariance leads to another Gaussian distribution. Otherwise, the uncertainty is not represented correctly as the nonlinear terms are missing. Therefore one might add an additional term onto the right side of 2.4 to represent this model error. More details on methods for a better representation of errors can be found in Hamill and Whitaker (2005). Wu et al. (2012) presents new results for an adaptive estimation procedure for such errors. But it will not be mentioned in this work.

The cost function 2.2 can be rewritten with equations 2.4 and 2.3, together with the background state estimate and the observations at time t_n. The linearized operators are also introduced and as the hypothetical state $\mathbf{x}$ was established at time t_{n-1} it needs to be transformed into time t_n. Therefore it can be written as $\mathbf{M}_{t_n,t_{n-1}}\mathbf{x} = \mathbf{M}_{t_{n-1},t_n}^{-1}\mathbf{x}$ which leads to the simplification of $\mathbf{M}_{t_{n-1},t_j}\mathbf{M}_{t_{n-1},t_n}^{-1}\mathbf{x} = \mathbf{M}_{t_n,t_j}\mathbf{x}$ in the sum over j. Algebraically this reads

$$\sum_{j=1}^{n-1}[\mathbf{y}_j^o - \mathbf{H}_j\mathbf{M}_{t_n,t_j}\mathbf{x}]^T\mathbf{R}_j^{-1}[\mathbf{y}_j^o - \mathbf{H}_j\mathbf{M}_{t_n,t_j}\mathbf{x}] = [\mathbf{x} - \bar{\mathbf{x}}_n^b]^T(\mathbf{P}_n^b)^{-1}[\mathbf{x} - \bar{\mathbf{x}}_n^b] + c, \tag{2.5}$$

with some constant c and the cost function finally is:

$$J_{t_n}(\mathbf{x}) = [\mathbf{x} - \bar{\mathbf{x}}_n^b]^T(\mathbf{P}_n^b)^{-1}[\mathbf{x} - \bar{\mathbf{x}}_n^b] + [\mathbf{y}_n^o - \mathbf{H}_n\mathbf{x}]^T\mathbf{R}_n^{-1}[\mathbf{y}_n^o - \mathbf{H}_n\mathbf{x}] + c. \tag{2.6}$$

This formulation of J using the background state and the new observations has to equal the cost function using the new analysis and its error covariance which is

$$J_{t_n}(\mathbf{x}) = [\mathbf{x} - \bar{\mathbf{x}}_n^a]^T(\mathbf{P}_n^a)^{-1}[\mathbf{x} - \bar{\mathbf{x}}_n^a] + c', \tag{2.7}$$

where c' is another constant. The two cost functions have to be equal except for a constant because they both are minimal for $\mathbf{x} = \bar{\mathbf{x}}_n^a$. To get the solution of this equation, first all terms have to be expanded. Then all terms which are of second degree in $\mathbf{x}$ build one equation. This leads to

$$\mathbf{P}_n^a = [(\mathbf{P}_n^b)^{-1} + \mathbf{H}_n^T\mathbf{R}_n^{-1}\mathbf{H}_n]^{-1}. \tag{2.8}$$

The same can be done for the terms of first degree:

$$\bar{\mathbf{x}}_n^a = \mathbf{P}_n^a[(\mathbf{P}_n^b)^{-1}\bar{\mathbf{x}}_n^b + \mathbf{H}_n^T\mathbf{R}_n^{-1}\mathbf{y}_n^o]. \tag{2.9}$$

These terms would be a lot simpler if the observations would be located at the position of the model grid points ($\mathbf{H} = \mathbf{I}$). In that case the analysis state is a weighted average of the background and the new observations, both weighted with their inverse covariance:

$$(\mathbf{P}_n^a)^{-1}\bar{\mathbf{x}}_n^a = (\mathbf{P}_n^b)^{-1}\bar{\mathbf{x}}_n^b + \mathbf{R}_n^{-1}\mathbf{y}_n^o.$$

Or put in other words, if the background has a large error, the analysis will lie closer to the observations.

For actual calculations there are many different ways to evaluate equations 2.8 and 2.9. Their initial formulations are not useable in practice because of the large number of matrix inversions appearing. Therefore by using 2.8 and the Woodburry-Matrix-Identity it is possible to eliminate $(\mathbf{P}_n^b)^{-1}$ from 2.9 to get:

$$\bar{\mathbf{x}}_n^a = \bar{\mathbf{x}}_n^b + \mathbf{P}_n^a\mathbf{H}_n^T\mathbf{R}_n^{-1}(\mathbf{y}_n^o - \mathbf{H}_n\bar{\mathbf{x}}_n^b). \tag{2.10}$$

Here the matrix $\mathbf{P}_n^a\mathbf{H}_n^T\mathbf{R}_n^{-1}$ is called the Kalman gain matrix $\mathbf{K}_n$. It is multiplied by the difference between the observations and the predicted observation values from the background. This vector is then added to the background state to build the analysis. Therefore if the Kalman gain is close to zero (as a $m \times l$ matrix), the observation errors are large and the information of the new observations is mainly discarded. The other extreme is if the Kalman gain ends up as $\mathbf{H}_n^T$ and the analysis will just be the observations transformed onto the model grid without using any background information. To get a better form of $\mathbf{P}_n^a$ without the need to calculate the inverse of the background covariance, the right side of 2.8 is multiplied with the Identity matrix $\mathbf{I} = (\mathbf{P}_n^b)^{-1}\mathbf{P}_n^b$ to yield

$$\mathbf{P}_n^a = (\mathbf{I} + \mathbf{P}_n^b\mathbf{H}_n^T\mathbf{R}_n^{-1}\mathbf{H}_n)^{-1}\mathbf{P}_n^b. \tag{2.11}$$

2.2.1 Extended Kalman Filter

The Extended Kalman Filter (EKF) (Jazwinski, 1970) is a variant of the Kalman Filter where the full nonlinear model is used to propagate the state estimate to the next time step and also nonlinear observation operators are allowed. This is beneficial

if one is interested in a sequence of analyses, since the errors introduced by a linear approximation would amplify after a sequence of assimilations and lead to bad results. To do nonlinear propagation with the analysis covariance, a tangent linear model is used which is a linearization of M_{t_{n-1},t_n} around $\bar{\mathbf{x}}^a_{n-1}$. The same is done for the observation operator H_n around $\bar{\mathbf{x}}^b_n$. By doing so, one assumes that the uncertainties in the state estimate can be approximated by linearized dynamics. The EKF is the standard assimilation method used in the Global Positioning System (GPS) navigation (Zhao et al., 2003).

Although the final formulations of $\bar{\mathbf{x}}^a_n$ and $\mathbf{P}^a_n$ in equations 2.10 and 2.11 contain only a single matrix inversion, in high-dimensional global weather models the number of variables m can go up to the order of 100 millions. Therefore already normal computations involving the $m \times m$ covariances are very expensive. A multiplication of two $10^8 \times 10^8$ matrices requires 10^{24} floating-point operations. The fastest supercomputers at the present time (2013) can sustain around 10^{16} floating-point operations per second and would need more than 3 years for one multiplication. This is the most obvious problem in the application of the EKF to weather models. Another one is that it can be difficult to determine the tangent linear model. Also its use can lead to linear instability if the initial estimate of the state is too far away from the truth. In this case the minimization fails and the filter diverges, meaning the observations are not able to correct the forecast and make it worse instead.

2.2.2 Ensemble Kalman Filter

Kalman Filter methods which can be applied to weather models therefore have to rely on low-rank approximations of the covariance matrices. One of these methods is the reduced-rank Kalman filter (Cohn and Todling, 1996; Fisher, 1998). It has been shown that this method does not work well in practice and therefore will not be investigated here (Fisher and Andersson, 2001; Beck and Ehrendorfer, 2003). Instead, the focus is on ensemble methods where the main step of the EnKF (Evensen, 1994, 2003) is to replace the analysis and its error covariance with the mean and spread of an ensemble of states. This ensemble mean is defined as $\bar{\mathbf{x}}^a_{n-1}$ where $n-1$ indicates the time. Then these ensembles can be propagated using the nonlinear model. The background covariance $\mathbf{P}^b_n$ at the next time can be computed using the resulting ensemble. Therefore like in the EKF the uncertainty is propagated from one analysis to the next.

The biggest difference to the former mentioned methods is that the uncertainty is now quantified in the space spanned by the ensemble. As the number of ensemble

members k is much smaller than the number of model variables (equal the dimension of the state vector) the computations involving the covariances are much cheaper. Of course there is also a computational restriction on the number of ensemble members to be used and there can be a rank difference between the error covariance and its estimate (Hamill et al., 2001). Other problems are the sensitivity to model error because the dynamics are used to propagate the error statistics (Houtekamer et al., 2009) and the need of an algorithm to generate a new ensemble after every analysis. The sampled variance of the ensemble is only the same as the background error in the theoretical limit that each ensemble member is a random sample of the true state. This is not true in reality as the members are not independent from each other (Anderson, 2001; Zupanski, 2009).

In the methods used by Hunt et al. (2007), the ensemble mean represents the best state estimate. As from now on only analysis time t_n is treated, the subscript n is skipped for readability while the superscript i is introduced for the k different ensemble members. Therefore the mean and covariance of the background ensemble are defined as:

$$\bar{\mathbf{x}}^b = k^{-1} \sum_{i=1}^{k} \mathbf{x}^{b(i)}, \tag{2.12}$$

$$\begin{aligned} \mathbf{P}^b &= (k-1)^{-1} \sum_{i=1}^{k} (\mathbf{x}^{b(i)} - \bar{\mathbf{x}}^b)(\mathbf{x}^{b(i)} - \bar{\mathbf{x}}^b)^T \\ &\equiv (k-1)^{-1} \mathbf{X}^b (\mathbf{X}^b)^T, \end{aligned} \tag{2.13}$$

with the $m \times k$ matrix $\mathbf{X}^b$. The rank of $\mathbf{P}^b$ is at most $k-1$ as the sum of all columns is $\mathbf{0}$.[2] Therefore the rank of the background covariance is limited by the ensemble size. Equivalent definitions are given for the analyzed state and covariance:

$$\bar{\mathbf{x}}^a = k^{-1} \sum_{i=1}^{k} \mathbf{x}^{a(i)}, \tag{2.14}$$

$$\begin{aligned} \mathbf{P}^a &= (k-1)^{-1} \sum_{i=1}^{k} (\mathbf{x}^{a(i)} - \bar{\mathbf{x}}^a)(\mathbf{x}^{a(i)} - \bar{\mathbf{x}}^a)^T \\ &\equiv (k-1)^{-1} \mathbf{X}^a (\mathbf{X}^a)^T. \end{aligned} \tag{2.15}$$

Now when the analyzed mean and covariance are specified, there are a lot of possi-

[2] This is the case, as the ensemble mean is also used. Then one ensemble member can be calculated from the other members together with the ensemble mean.

bilities how the ensemble members are built around the mean. One can get the analysis members by applying the Kalman Filter update in equation 2.10 separately to all background members instead of only to the mean (Houtekamer and Mitchell, 1998). Another possibility are different forms of an Ensemble Square Root Filter (EnSRF) (Tippett et al., 2003). They use deterministic algorithms to generate the analysis ensemble and owe their name to the definition of $\mathbf{X}^a$ as a matrix square root of $\mathbf{P}^a$.

The methods presented in this thesis follow this deterministic approach. A huge advantage of the EnKF is that no tangent linear and adjoint operators are needed anymore which saves a lot of computation time.

2.2.3 Local Ensemble Kalman Filter

Because the background covariance matrix from 2.13 can only span a k-dimensional subspace, the adjustments of the global analysis can only happen in this subspace. But usually there are not enough degrees of freedom available to fit the huge number of observations which is in the order of millions. Therefore a new analysis cannot correct for all of the growing errors of the high-dimensional model space. But if the analysis were to be performed in small local subspaces the ensemble would only need to represent the errors in this smaller region. Also different linear combinations of the ensemble members can be used in the different regions and overall the global analysis explores a space with much higher dimension. The method which applies this type of localization is called Local Ensemble Kalman Filter (LEKF) and was developed by Ott et al. (2004).

It is not only the gain in dimensionality that makes localization useful. Due to the limited sample size, the covariances will introduce spurious correlations between distant locations in the global filter. These may for example be the cause that the analysis in Europe is influenced by an observation over Greenland, which in general is unrealistic. This problem can be corrected by localization if a region is not larger than a characteristic distance over which a correlation seems realistic. Then any correlation which goes beyond this distance will not show up in the analysis. Another advantage is that the model might explore a smaller number of dimensions in small local regions compared to the full model domain (Patil et al., 2001). Finally localization is computationally efficient as it can be done in parallel for every region. This is an advantage for the highly parallel structure of supercomputers (Ott et al., 2002).

There are two possibilities on how to exactly do the localization. In the explicit way, only the observations from a region surrounding the analysis location are considered

(Keppenne, 2000; Ott et al., 2004). In the implicit way the entries of $\mathbf{P^b}$ are multiplied by a distance-dependent function that decays to zero at a certain distance. By doing so, the observations do not affect the analysis beyond that distance. This method is also called covariance localization (Hamill et al., 2001; Houtekamer and Mitchell, 2001; Whitaker and Hamill, 2002). In this work the explicit approach is used and an analysis is computed at every grid point of the model separately, using observations only from a neighborhood of the grid point.

There are many options selecting the observations which are used and making their influence decay with distance. Often Gaspari-Cohn functions are used for the horizontal localization (Gaspari and Cohn, 1999). Vertical localization is still an unsolved problem, as it is difficult to assign observations like satellite radiances to a specific model level (Anderson, 2007b; Miyoshi and Sato, 2007; Campbell et al., 2010). The reason for this is that satellite observations are non-local and should be applied to multiple model layers. Vertical localization is not considered in this thesis, as the models that will be used do not contain a vertical coordinate. Another problem of localization is that due to the small regions larger scales may be analyzed poorly and difficulties in producing balanced analyses arise (Kepert, 2009; Greybush et al., 2010).

2.2.4 Local Ensemble Transform Kalman Filter

Finally the LETKF is one of the data assimilation methods used in this thesis. The difference to the LEKF is that it contains an efficient way to transform the background ensemble into an appropriate analysis ensemble. This transformation was introduced by Bishop and Toth (1999). It is efficient in the sense that most of the operations take place in the k-dimensional ensemble space and only a few in the higher-dimensional model or observation space.

As mentioned in Section 2.2.2 there are different possibilities on how to choose an ensemble. The first problem is that the $m \times m$ background covariance matrix $\mathbf{P}^b = (k-1)^{-1}\mathbf{X}^b(\mathbf{X}^b)^T$ has at most rank $k-1$ and is therefore not invertible[3]. But as the last equation shows, it is a symmetric matrix and therefore one-to-one (injective) on its column space S which is also the column space of $\mathbf{X}^b$. In equation 2.13 one can see that the ith column of this matrix is $\mathbf{x}^{b(i)} - \bar{\mathbf{x}}^b$ and there are k columns. $(\mathbf{P}^b)^{-1}$ is well-defined in the subspace S and the minimization of the cost function in 2.6 can be carried out in this subspace. The advantage of this transformation is computational efficiency.

[3]For a $k \times k$ matrix $\mathbf{A}$ being invertible is equivalent to having rank k.

To perform the analysis in this space, one needs an appropriate coordinate system and independent basis vectors. To do so, $\mathbf{X}^b$ is regarded as a linear transformation from a k-dimensional space $\tilde{S}$ onto S. If $\mathbf{w}$ (for weight) is a vector in $\tilde{S}$, then $\mathbf{X}^b\mathbf{w}$ is in S and the corresponding model state is $\mathbf{x} = \bar{\mathbf{x}}^b + \mathbf{X}^b\mathbf{w}$. One can now construct a cost function $\tilde{J}$ in the following form:

$$\tilde{J}(\mathbf{w}) = (k-1)\mathbf{w}^T\mathbf{w} + [\mathbf{y}^o - H(\bar{\mathbf{x}}^b + \mathbf{X}^b\mathbf{w})]^T\mathbf{R}^{-1}[\mathbf{y}^o - H(\bar{\mathbf{x}}^b + \mathbf{X}^b\mathbf{w})]. \tag{2.16}$$

If it is minimized by the mean of the weights $\bar{\mathbf{w}}^a$, $\bar{\mathbf{x}}^a = \bar{\mathbf{x}}^b + \mathbf{X}^b\bar{\mathbf{w}}$ will also minimize the original cost function J. But this is only the case if $\mathbf{w}$ is a random vector with mean $\mathbf{0}$ and covariance $(k-1)^{-1}\mathbf{I}$. In that case $x = \bar{\mathbf{x}}^b + \mathbf{X}^b\bar{\mathbf{w}}$ is also Gaussian with mean $\bar{\mathbf{x}}^b$ and covariance $\mathbf{P}^b = (k-1)^{-1}\mathbf{X}^b(\mathbf{X}^b)^T$ which was required.

There are different possibilities to allow for a nonlinear observation operator H. A simple one is to evaluate it in the ensemble space as it is not needed in the other spaces. Therefore H is applied to each ensemble member $x^{b(i)}$ and then interpolated. To do so, first an ensemble of background observation vectors is defined by

$$\mathbf{y}^{b(i)} = H(\mathbf{x}^{b(i)}), \tag{2.17}$$

with mean $\bar{\mathbf{y}}^b$. The $l \times k$ matrix $\mathbf{Y}^b$ is defined through its columns $\mathbf{y}^{b(i)} - \bar{\mathbf{y}}^b$. The observations are then approximated with:

$$H(\bar{\mathbf{x}}^b + \mathbf{X}^b\mathbf{w}) \approx \bar{\mathbf{y}}^b + \mathbf{Y}^b\mathbf{w}. \tag{2.18}$$

When all the formulations of the last three paragraphs are put together into 2.16, one yields the quadratic cost function

$$\tilde{J}^\star(\mathbf{w}) = (k-1)\mathbf{w}^T\mathbf{w} + [\mathbf{y}^o - \bar{\mathbf{y}}^b - \mathbf{Y}^b\mathbf{w}]^T\mathbf{R}^{-1}[\mathbf{y}^o - \bar{\mathbf{y}}^b - \mathbf{Y}^b\mathbf{w}]. \tag{2.19}$$

$\mathbf{Y}^b$ acts as the observation operator in the subspace $\tilde{S}$. The background mean is $\bar{\mathbf{w}}^b = 0$ and the background covariance is $\tilde{\mathbf{P}}^b = (k-1)^{-1}\mathbf{I}$. This now leads to the analogues of the analysis equations 2.9 and 2.8 which are

$$\bar{\mathbf{w}}^a = \tilde{\mathbf{P}}^a(\mathbf{Y}^b)^T\mathbf{R}^{-1}(\mathbf{y}^o - \bar{\mathbf{y}}^b), \tag{2.20}$$

$$\tilde{\mathbf{P}}^a = [(k-1)\mathbf{I} + (\mathbf{Y}^b)^T\mathbf{R}^{-1}\mathbf{Y}^b]^{-1}. \tag{2.21}$$

These equations can finally be transformed back into the model space where the anal-

ysis mean and covariance are

$$\bar{\mathbf{x}}^a = \bar{\mathbf{x}}^b + \mathbf{X}^b \bar{\mathbf{w}}^a, \tag{2.22}$$

$$\mathbf{P}^a = \mathbf{X}^b \tilde{\mathbf{P}}^a (\mathbf{X}^b)^T. \tag{2.23}$$

This only fixes the analysis mean for the LETKF while a method to choose the analysis ensemble is still missing. As mentioned before, the only restriction is that the ensemble mean and covariance have to be equal to $\bar{\mathbf{x}}^a$ and $\mathbf{P}^a$. Therefore one has to choose a matrix $\mathbf{X}^a$ such that the sum of its columns is zero. Then the analysis ensemble is given by adding the mean to each of the columns of this matrix.

In the LETKF described by Hunt et al. (2007) and used in this work, the analysis ensemble is characterized by the matrix $\mathbf{X}^a = \mathbf{X}^b \mathbf{W}^a$, with

$$\mathbf{W}^a = [(k-1)\tilde{\mathbf{P}}^a]^{1/2}. \tag{2.24}$$

The 1/2 power in this definition stands for the symmetric square root. This root can be calculated as $\mathbf{P}$ is a symmetric matrix. Then one can solve for $\tilde{\mathbf{P}}^a = (k-1)^{-1}\mathbf{W}^a(\mathbf{W}^a)^T$ and by putting this into 2.23 it can be shown that the condition of 2.15 still holds. The choice of a symmetric square root has two advantages. First, it ensures that the sum of the columns of $\mathbf{X}^a$ is zero, as required. Second, it ensures that $\mathbf{W}^a$ depends continuously on $\tilde{\mathbf{P}}^a$. This is an important property to make sure that analyses from neighbouring grid points cannot differ too from each other and produce unphysical situations.

The analysis ensemble is finally built in $\tilde{S}$ by adding $\bar{\mathbf{w}}^a$ to each of the columns of $\mathbf{W}^a$. The vectors $\mathbf{w}^{a(i)}$ are the resulting columns of this calculation. These "weight" vectors define which linear combinations of the background ensemble perturbations are added to the background mean. The resulting sum is the final analysis ensemble in model space:

$$\mathbf{x}^{a(i)} = \bar{\mathbf{x}}^b + \mathbf{X}^b \mathbf{w}^{a(i)} \tag{2.25}$$

From this ensemble the analysis mean can easily be computed.

Covariance inflation

If an Ensemble Kalman Filter strictly follows the filter equations 2.10 and 2.11 it is still not sure that it is able to follow the "true" trajectory. One reason is that the

model error is not taken into account. But experiments have shown that also in a perfect model environment, the filter underestimates the uncertainty in the analysis (Whitaker and Hamill, 2002). This behavior is also known as overconfidence because the analysis does not show the true uncertainty. In such a case the ensemble does not have enough spread and if this effect amplifies over a couple of analysis cycles, it is possible that in the end the observations do not get any weight at all and the filter diverges.

A simple approach to cope with this problem is to inflate either the background or analysis covariance during each data assimilation cycle. There are two different ways of doing this. One is "multiplicative inflation" (Anderson and Anderson, 1999) where the background covariance matrix is multiplied by a constant factor larger than one. In the other one, "additive inflation" (Ott et al., 2004), a small multiple of the identity is added to the covariance. Both methods are applied during each cycle. Additive inflation needs substantial additional computation because the adjustment has to be calculated in the k-dimensional subspace $\tilde{S}$. Even more advanced are methods where the inflation is determined adaptively (Anderson, 2007a, 2009; Li et al., 2009; Miyoshi, 2011). In these methods instead of using trial and error, the amount of inflation is calculated at each time step by using innovation statistics and varies spatially and temporally. As these statistics are not perfectly known, Li et al. (2009) proposed to estimate the inflation factor and observational errors simultaneously within the EnKF.

In this work a simple multiplicative inflation is applied as the focus is not on getting the best inflation. The easiest approach would be to multiply $\mathbf{W}^a$ by $\sqrt{\rho}$ which is equivalent to multiplying the analysis covariance by ρ. If one wants to inflate the background ensemble instead, multiplication of $\mathbf{X}^a$ by $\sqrt{\rho}$ is also possible. In that case the background ensemble has to be adjusted accordingly as it needs to have the correct spread. If H is linear this is identical to replacing $(k-1)\mathbf{I}$ by $(k-1)\mathbf{I}/\rho$ in 2.21 since it is the inverse of the background covariance in the subspace $\tilde{S}$. This procedure is used in this work.

Multiplicative inflation can also be thought of as weakening the influence of past observations on the current analysis. As inflation is applied at every time step, the effect accumulates and the influence of past observations decays exponentially with time. The inflation factor together with the analysis frequency determines the time scale over which observations still have an impact on the analysis. Thus, the analysis is also localized in time. This effect is desirable in the presence of model error because past observations would propagate to a wrong current state. Inflation is even more important when there are a lot of observations used because more observations usually

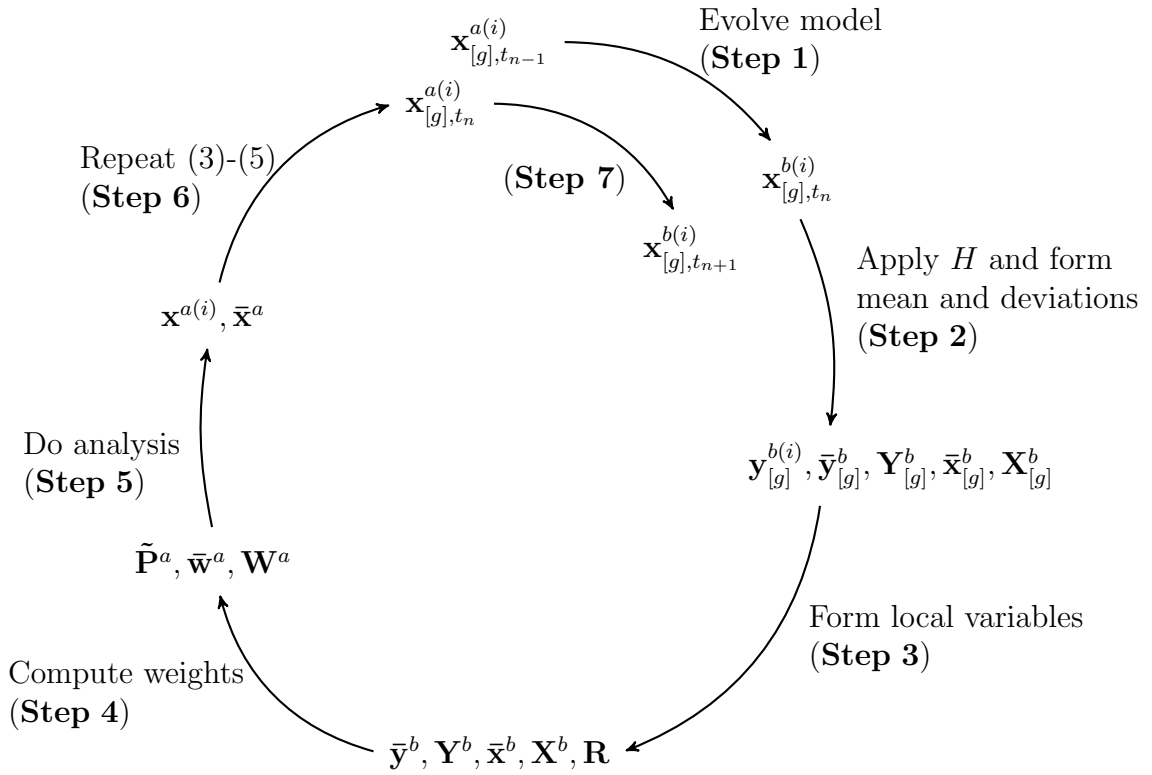

Figure 2.1: Illustration of the step-by-step implementation of the LETKF. Adapted from Ott et al. (2004)

means more trust in the observations. Another possibility to deal with this problem is to chose the localization radius depending on the number of observations available in the region. The effective number of observation is kept constant, while the localization radius is varying. This is achieved by keeping the sum of the observation weights constant and therefore changing the individual weight of the observations.

Implementation

In this Subsection a step-by-step description of the implementation of the filter as it is used in this work is presented. A sketch of these steps is also shown in Figure 2.1 to help envisioning the procedure. This illustration was inspired by Figure 1 of Ott et al. (2004). Because there is an analysis computed for every grid point separately, the subscript $[g]$ is introduced to show where the global model state or observations are meant. If this subscript is omitted, only the specific localized variables are used. The input for an assimilation cycle is the analysis ensemble of the model state at time $n-1$: $\{\mathbf{x}^{a(i)}_{[g],t_{n-1}}\colon\ i = 1, 2, ..., k\}$, a set of l observations $\mathbf{y}^o_{[g],t_{n-1}}$ and their corresponding

error covariance matrix $\mathbf{R}_{[g]}$. In the ETKF that is also used in this work, all operations are done in the same way described here with the exception that the steps 3 and 6 are left out.

1. First, the analysis ensemble is advanced to the next time to create a background ensemble. The time index is skipped from step 2 to step 6 as they all are applied at the same time.

2. The global observation operator is applied to the background ensemble to form the background observation ensemble. These observations are also called synthetic observations and their mean is calculated. Then this vector is subtracted from the observation ensemble members to form the columns of the matrix $\mathbf{Y}^b_{[g]}$. These are the ensemble deviations from the mean. The same is done with the ensemble of background model states.

3. During this step all the localized variables are produced. Therefore the necessary data for the analysis at a certain grid point is chosen. This is done for the model state and the synthetic observations. The observation error covariance matrix is also localized by weighting the observation errors with the Gaspari-Cohn function. Therefore observations closer to the analysis grid point get a higher weight than the ones further away. This smoothens the spatial influence of the observations.

4. $\bar{\mathbf{w}}^a$ and $\mathbf{W}^a$ are computed according to equations 2.20 and 2.24. To do so a matrix $\mathbf{C} = (\mathbf{Y}^b)^T\mathbf{R}^{-1}$ is calculated first. In case of a diagonal matrix $\mathbf{R}$, this calculation is simple. This is true for the observations which will be used in this work. Then $\tilde{\mathbf{P}}^a = [(k-1)\mathbf{I}/\rho + \mathbf{C}\mathbf{Y}^b]^{-1}$ as in 2.21 can be computed. As mentioned in the previous Section, ρ is a multiplicative covariance inflation factor. The multiplicative inverse of the square matrix in the brackets can be computed by using predefined algorithms or a singular value decomposition. This might be useful as the eigenvalues and eigenvectors of $[(k-1)\mathbf{I}/\rho + \mathbf{C}\mathbf{Y}^b]$ can also be used in the calculation of $\mathbf{W}^a = [(k-1)\tilde{\mathbf{P}}^a]^{1/2}$, because the square root can be constructed using them.

5. Finally the analysis ensemble can be computed using equation 2.25.

6. Steps 3 to 5 have to be repeated for all grid points and all local analyses can be put together to form the global analysis ensemble $\mathbf{x}^{a(i)}_{[g],t_n}$ and the assimilation cycle is finished.

7. Step 7 is identical to step 1 and a new cycle starts.

The specific description of quantities like the Gaspari-Cohn localization radius or observation errors is given separately for the different experiments in the following Chapters that present data assimilation results.

2.3 Particle Filter

All Kalman Filter methods are Gaussian methods in the sense that they assume that the PDF of the error is described by a mean and a covariance. Often this leads to good results although the assumptions are not satisfied in reality. But if the nonlinearity and non-Gaussianity are too large, these methods often fail. As described in the introduction of this work this is often the case for convective precipitation. Therefore it seems obvious to formulate non-Gaussian methods which do not make any of the false assumptions. Particle filters are a class of such methods as well as nudging, but the focus here lies on particle filters. Doucet et al. (2001) gives a good overview on the theory with some examples from the statistical science.

In particle filters the term particle stands for an ensemble member as used in the EnKF. The full nonlinear model and observation operators are used and no assumptions are made on the shape of the error distribution. The PDF is simply approximated by a discrete distribution

$$p(\mathbf{x}) = \frac{1}{k} \sum_{i=1}^{k} \delta(\mathbf{x} - \mathbf{x}_i), \tag{2.26}$$

where $\mathbf{x}_i$ is the i-th particle that is randomly drawn from an ensemble of k particles and δ is the Dirac delta function.

2.3.1 Basic Importance Sampling

The basic implementation of importance sampling is done by just implementing Bayes theorem with

$$p(\mathbf{x}|\mathbf{y}) = \sum_{i=1}^{k} w_i \delta(\mathbf{x} - \mathbf{x}_i), \tag{2.27}$$

where the weights w_i are given by

$$w_i = \frac{p(\mathbf{y}|\mathbf{x}_i)}{\sum_{j=1}^{k} p(\mathbf{y}|\mathbf{x}_j)}. \tag{2.28}$$

Here $p(\mathbf{y}|\mathbf{x}_i)$ stands for the PDF of the observations given the model state $\mathbf{x}_i$, which is a conditional probability. This is still a distribution as different values for a specific observation are possible given a fixed model state. The meaning of $p(\mathbf{x}|\mathbf{y})$ seems clearer, as this is the distribution of possible model states given the observations. If the observation errors were Gaussian, $p(\mathbf{y}|\mathbf{x}_i)$ could be written as

$$p(\mathbf{y}|\mathbf{x}_i) = A \cdot \exp\left\{-\frac{1}{2}[\mathbf{y} - H(\mathbf{x}_i)]^T R^{-1}[\mathbf{y} - H(\mathbf{x}_i)]\right\}, \tag{2.29}$$

with a term similar to the known cost function from Kalman filtering in the exponent. Other forms are also possible depending on the error distribution. Each particle has a weight according to the likelihood that its state has in the distribution of the observations.

The procedure for an assimilation in importance sampling is as follows (see also Fig. 2.2 for a schematic):

1. k particles $\mathbf{x}_i$ from the initial model probability density $p(\mathbf{x}_{t_0})$ are sampled.
2. All particles are integrated with the model up to the observation time.
3. The weights of the particles are calculated according to equation 2.28.
4. Steps 2 and 3 are repeated sequentially until all measurements are used.

After the assimilation, the analysis state can be calculated using all particles and their final weights:

$$\mathbf{x}^a = \bar{\mathbf{x}} = \int \mathbf{x}\, p(\mathbf{x})\, d\mathbf{x} \approx \sum_{i=1}^{k} w_i \mathbf{x}_i. \tag{2.30}$$

One problem of this method is that filter degeneracy can happen very fast. This means that after some time the weight of one particle is close to one while all the other particles have almost no weight anymore ($t = 20$ in Fig. 2.2). If that happens there is no distribution sampled anymore but just a delta function present. Another problem is that the particles are not modified and if they are not able to follow the

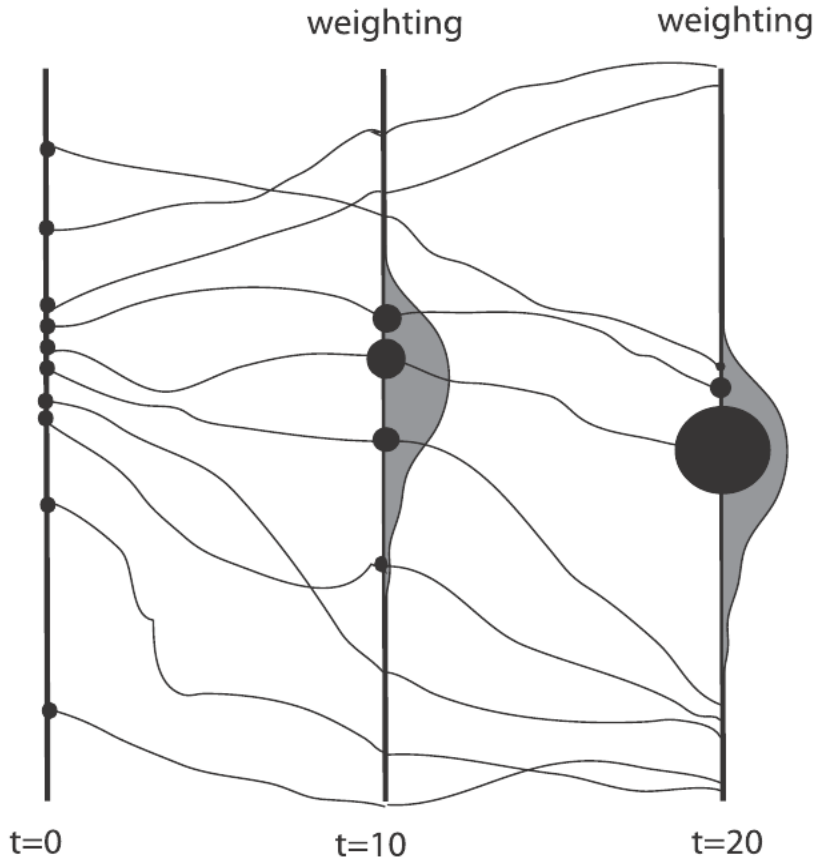

Figure 2.2: Fig. 1 from van Leeuwen (2009), showing a schematic of two assimilation cycles with a standard particle filter with importance sampling. The vertical axis denotes the model variable, while the size of a circle corresponds to the weight of a particle. Time is in the horizontal axis, with observations every 10 units. All particles start with identical weight. The total likelihood at time 10 is displayed in the gray distribution. At time 20 one member has almost all weight.

observations due to model error, they would only end up closer to the observations by chance.

2.3.2 Sequential Importance Resampling

The particle filter method that will be used for assimilation experiments with the stochastic model is the SIR as introduced by van Leeuwen (2009) for geophysical applications. Previous work on the topic was done by Liu and Chen (1998), Gordon et al. (1993) and Doucet et al. (2001). The goal of this method is to reduce the variance in the weights. This is done by resampling the posterior ensemble such that all weights become approximately equal. This is achieved by abandoning particles with very low

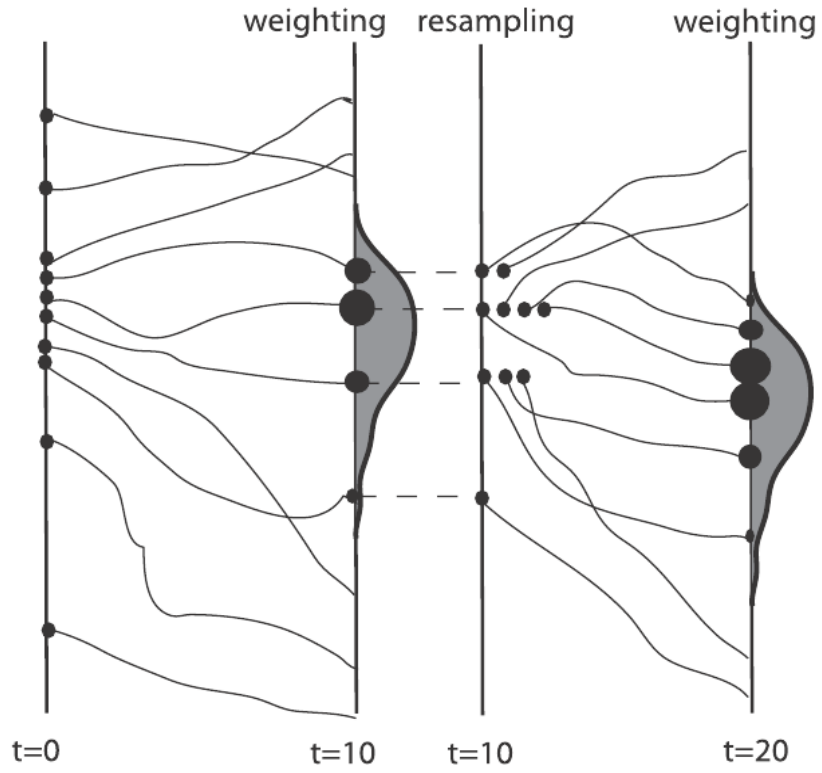

Figure 2.3: Fig. 2 from van Leeuwen (2009), showing a schematic of two assimilation cycles with resampling. All particles start with identical weight. At time 10 they are first weighted according to the likelihood and then resampled so that at time 20 the filter has not collapsed.

weights and replacing them with perturbed copies of particles with high weight.

The order of operations is similar to the one of importance sampling. There is only one additional resampling step after the calculation of the weights (also sketched in Fig. 2.3):

1. k particles x_i from the initial model probability density $p(\mathbf{x}_{t_0})$ are sampled.
2. All particles are integrated with the model up to the observation time.
3. The weights of the particles are calculated according to equation 2.28.
4. The particles are resampled such that the weights are equal to $1/k$.
5. Steps 2 to 4 are repeated sequentially until all measurements are used.

There are many different possibilities how step 4 can be performed but only the so-called probabilistic resampling will be used in this work.

Propabilistic resampling means sampling directly randomly from the density given by the weights (Gordon et al., 1993). This is easy since the density is a discrete,

one-dimensional number already calculated is already calculated. However when a small number of particles is used, this procedure introduces sampling noise due to the random character of the sampling. Kim et al. (2003) have shown that SIR with probabilistic resampling is superior to basic importance sampling. But Zhou et al. (2006) estimated a 684-dimensional state vector in a land-surface model and needed 800 particles to converge versus only 80 with an EnKF. They also mention the need for even more particles when increasing the numbers of measurements used or increasing the state space. The problem of this curse of dimensionality for the particle filter is described in detail in Bengtsson et al. (2008); Snyder et al. (2008). Therefore the SIR indeed performs better than basic importance sampling, but still relates badly to the dimensionality of the system.

Particle filters are still under development and being improved but recent variants like the nudging particle filter of van Leeuwen (2010) are not considered in this work.

Implementation

The exact implementation of the SIR filter used for data assimilation experiments with the stochastic model follows the five steps described above, with probabilistic resampling. In the analysis step at time t_n, each ensemble member i is assigned a weight, $w_i(t_n)$, according to

$$w_i(t_n) = w_i(t_{n-1}) \cdot \exp[-rms_i/\sigma], \tag{2.31}$$

where rms_i is the root mean square difference between the ensemble member i and the truth run, and $\sigma = 0.05$ is the square root of the observation error variance. The value of 0.05 used later is the result of a simple tuning. Since the weight of a given ensemble member is obtained by updating its previous value, it continues to be influenced by previous observations. Rearranging the above equation shows that the impact of the observations at a given time decays exponentially at a rate proportional to the ratio rms/σ.

Therefore σ can be interpreted as providing a memory timescale for the weights and if a particle has a large error at one time it will not be punished instantly but rather if the error persists for a sequence of analyses. After the calculation of the updated weights using 2.31 the weights are normalized so that the sum over all ensemble members is equal to one as defined in 2.28.

Finally the new analysis ensemble has to be formed by resampling the old particles. New ensemble members are chosen by randomly drawing from the old ensemble,

where each member has a probability of being chosen according to its weight. If one member has a sufficiently high weight, it is possible that all other members will be replaced by copies of this high weighted particle. Of course it is not the goal to have k identical members, therefore to maintain diversity in the ensemble all members are then perturbed with an additive noise of the form $a \cdot \epsilon \cdot rms_i$. In this definition a is an amplitude factor set to 0.1 for the global particle filter and 0.25 for the local version, and ϵ is a random number drawn from a uniform distribution between -0.5 and 0.5.

The localized version of the SIR filter is obtained by dividing the domain into equally sized subregions (usually single grid points), and performing the analysis and resampling steps described above on each subregion individually. Therefore localization works in principle in the same way as for the LETKF.

2.4 Summary

This Chapter presented the two main data assimilation methods that are used in this work, the Kalman filter and the particle filter. The Kalman filter is a well established method which has been used for a long time in many different applications. In contrast, particle filtering methods have just startet to find their way into the geophysical sciences and are currently undergoing a rapid development. The advantage of the LETKF is the easy numerical implementation and possibility to parallelize the computation. As it is based on linearity and Gaussianity it is not well-known how it will behave in a convective environment. This is the advantage of particle filters which are not based on any assumption but are known to have problems in higher-dimensional spaces.

Details on the specific settings and application of the two methods in the test models are given in the respective subsequent Chapters, starting with the stochastic cloud model in the next Chapter.

Chapter 3

Stochastic cloud model

The following Chapter presents the stochastic cloud model and data assimilation experiments that were achieved using this model. It is the first and simplest one in the model hierarchy for data assimilation on the convective-scale. The stochastic cloud model is based on a simplified version of the stochastic convection parameterization scheme of Plant and Craig (2008). It is designed to capture the extreme nonlinearity and non-Gaussianity associated with convection while still trying to be as simple as possible. Data assimilation experiments are performed in this model with the aim of identifying basic characteristics of data assimilation methods in such an environment. Additionally localization and averaging are applied to detect possibilities to improve the methods. The model definition is shown in Section 3.1 and the results from data assimilation experiments are given in Section 3.2 followed by a short discussion in Section 3.3.

3.1 Model description

A simple stochastic process is used to produce a changing number of clouds. These clouds are the only variable the model contains and they can appear at every one of a set of n grid points. At each of these points an integer number of clouds is defined to be present. The model only allows for integer states and the convective dynamics is specified as a birth-death process which is defined by the two parameters λ and μ. λ is the probability of a cloud being initiated at a grid point and μ the probability for each existing cloud to disappear. These probabilities do not change during the simulation, and are determined by specifying a mean cloud half life (hl) and average density of clouds per grid point (ρ) using the relations $\lambda = \rho(1 - 0.5^{1/hl})$ and $\mu = 1 - 0.5^{1/hl}$.

So either hl and ρ, or λ and μ can be defined. Therefore the only thing going on in this model is clouds appearing and disappearing at everyone of a given number of grid points.

In this initial study it is assumed that the grid points are arranged on a one-dimensional line, but since cloud positions are uncorrelated the grid points could equally have been arranged in a two-dimensional array without changing anything of the results. The only influence the total number of grid points have is to change the dimensionality of the system. But this affects only the data assimilation experiments and not the behaviour of the model itself.

A step by step description of the model integration is as follows:

a) Initialize each grid point i by assigning a number of clouds m_i by randomly drawing from a Poisson distribution with mean ρ:

$$p(m_i) = \rho^{m_i} e^{-\rho} / m_i!$$

b) The time step calculation at each grid point is executed by the calculation of

   i) death: remove each of the m_i clouds with probability μ

   ii) birth: add a new cloud with probability λ

Step b) is now being repeated for the desired amount of times. To do data assimilation experiments one realisation of this model can be integrated as a "truth" simulation, and an ensemble of k simulations is used for ensemble data assimilation. In these ensemble members, the data assimilation method applies increments at each time step after the birth and death in the model is calculated. The specific computation of the increments for each of the two assimilation algorithms will be described below in Section 3.2.1.

This is a perfect model scenario in the statistical sense that all ensemble members are governed by the same rules as the truth run, but since the random numbers are the only thing different in each member and the truth, model unpredictability still plays a key role. The random numbers are important for the birth-death process while all the other parameters are identical in all ensemble members.

For the experiments in this thesis, the number of grid points is fixed at $n = 100$, and the birth and death probabilities are chosen to give a mean cloud density of $\rho = 0.1$ clouds per grid point, thus providing a realistic degree of intermittency. A density of 0.1 means that at approximately 90% of the grid points there is no cloud present and at the rest of the grid points there is one cloud or very rarely two or more clouds.

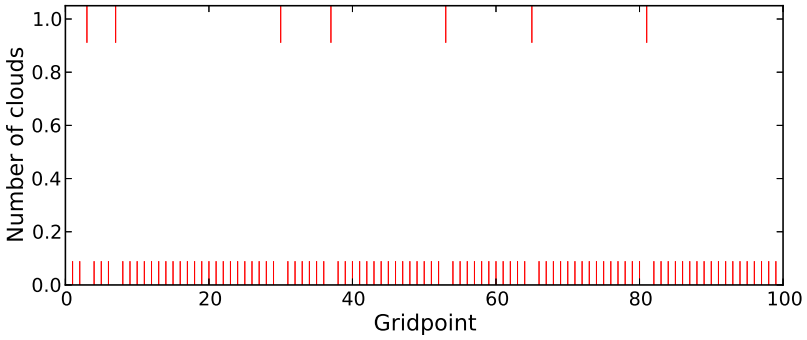

Figure 3.1: Model state at a random situation. The positions of clouds are marked with red vertical lines.

A random model state is displayed in Fig. 3.1 where the red vertical lines indicate locations of clouds. Although these lines stretch over some distance they stand for an integer number. This style is chosen for better readability in Figures 3.3 to 3.5. In the model state of Fig. 3.1 there are only 7 clouds in the whole domain and no grid point contains more than one cloud. Depending on the choice of the cloud half life this situation will change faster or slower.

3.2 Data assimilation

The stochastic cloud model presented in the previous Section is now being used for data assimilation experiments. The methods used are the SIR and the ETKF together with their localized versions and variants where the observations are averaged. First the general model and data assimilation settings are explained in Section 3.2.1. Section 3.2.2 explains the general behaviour of the methods followed by the dependence on the ensemble size in Section 3.2.3. To see whether changes to the methods improve the results, localization is applied in Section 3.2.4 and averaging in Section 3.2.5.

3.2.1 Settings

As the speed at which the model state changes purely depends on the half life setting, each model time step can correspond to an observation time and assimilation time as well. One observation contains the complete state of the truth run at that instant, i.e. the number of clouds present at each grid point, with no added error. Note that this includes points where the number of clouds is zero, so that observations of “no cloud” are also assimilated. This is a procedure which is also applied in radar assimilation.

Although the observations are not explicitly perturbed, their error characteristics must be specified.

Consistent with the dynamical development of the stochastic model, it is assumed that the observation errors are independent of location and uncorrelated, so that the observation error covariance matrix is proportional to the identity matrix. The constant of proportionality, corresponding to the (constant) error variance, does not affect the ETKF results which depend only on the relative weighting of ensemble members. The situation is slightly more complicated for the SIR filter, where the error variance does affect the length of time during which a given set of observations have influence on the current assimilation. This is discussed in more detail in Section 2.3 of the previous Chapter.

The observation strategy is motivated by the characteristics of network radar observations that provide a spatially complete view of precipitation, but relatively little information about the dynamical variables associated with its evolution. Indeed, for the stochastic dynamics used here, the observations contain no information about which clouds will die out in future or where new clouds will later be initiated. If radar observations were available every 5 minutes, the mean cloud lifetime of $hl = 30$ steps would correspond to 2.5 hours, making it possible, in principle, for the data assimilation to lock on to an observed cloud before it randomly disappears.

For comparison, experiments were also done with a mean cloud lifetime of $hl = 3000$ observation times, which will be referred to as a stationary cloud field since the mean lifetime is much longer than the duration of the experiments. Therefore this large half life is used to show what would be possible for a certain method while the shorter half life corresponds to a more realistic situation. For the runs with observation averaging, the observations are computed as the total number of clouds in non-overlapping subregions with the size of 10 grid points. The modeled observations are computed from the ensemble members in the same way.

To construct a new analysis ensemble, an updated ensemble mean is produced, and the updated ensemble members are constructed by adding linear combinations of the deviations. The cloud number acquired in this way will contain non-integer values which must be converted to integers to produce a valid model state. This is done by treating the non-integer part of the analyzed cloud number as a probability for a cloud being present, and randomly choosing to assign a cloud or not according to this probability. This procedure adds another nonlinearity to the system which in a similar way also happens in a full weather model with moisture values around saturation. Another problem which can also occur is that the assimilation produces

negative cloud values. These values are just put back to zero. As this does not happen to a large degree the mean cloud number is not affected in a significant way. No covariance inflation factor is used, except for the simulations with observation averaging, where it was found convenient to use a deflation of 0.7 to avoid an excess of spread due to the probabilistic conversion to integer cloud numbers.

In the local version of the filter, the size of the local region used is one grid point, as in Hunt et al. (2007), but only one observation is used for each region. This is contrary to the recommendation of Hunt et al. (2007), who suggest using observations from a region centred on the grid point being updated in order to ensure that the increments at neighbouring grid points vary smoothly. However, concerns about smoothness do not arise for the stochastic birth-death process used here, since it produces fields that are uncorrelated between grid points. In this case using more than one grid point would degrade the result.

As mentioned in Chapter 1 the goal of this work is not to judge whether one assimilation method is better than the other. The focus lies on an investigation of generic behaviour of the two methods and therefore no attempt has been made to tune parameters or otherwise optimize the two schemes. Such a tuning would of course have improved the results, but it would not change the qualitative conclusions that were obtained.

3.2.2 Convergence for stationary and time-varying cloud fields

The ability of the two assimilation schemes to converge to the observed state is considered first. In this initial test the basic forms of the methods are applied with a representative ensemble size of 50 members. The Root Mean Square (RMS) error of the individual ensemble members is computed at every time step and the average RMS error of all ensemble members is plotted (not the RMS error of the ensemble mean, except where noted). The spread is defined as the mean squared difference of the ensemble members from the ensemble mean. To reduce the noise level in the figures and increase the statistical significance, the errors are averaged over 100 repetitions of the experiment with different realisations of the stochastic process. In the case of the SIR hl 30 experiment 400 realisations were done because 100 was still leading to a noisy result. The average error is then scaled so that a value of one corresponds to the RMS average difference between two randomly chosen realisations of the model state. As the ensembles are initialized with random states of the correct cloud density, they

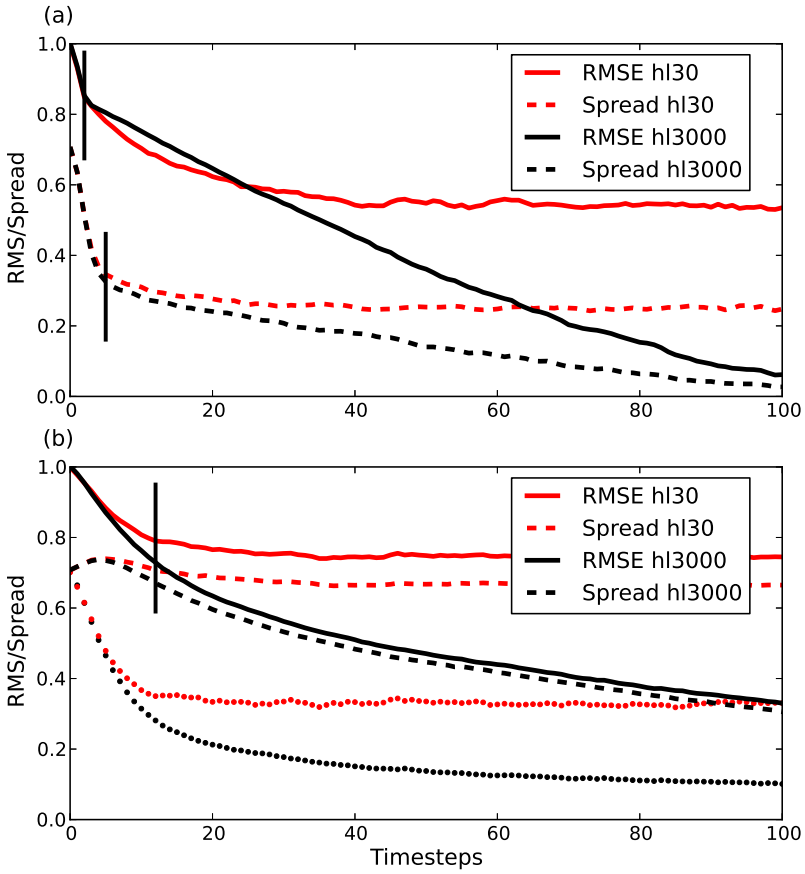

Figure 3.2: Root mean squared error (solid lines) and ensemble spread (dashed lines) as function of time, for an ensemble size of 50, for (a) SIR and (b) ETKF with half-lives of 3000 (black) and 30 (red) time steps. Vertical lines indicate regime boundaries (see text). Dotted lines show RMS of the ensemble mean.

start with exactly this error. This can be seen in Fig. 3.2.

The continuous black line in Fig. 3.2a shows the evolution of the mean error of the SIR filter for the stationary cloud field. The SIR filter converges, although rather slowly, and the decrease in error continues beyond 100 time units, eventually saturating at a value close to zero. The ensemble spread (black dashed line) shows an initial rapid decrease to a fixed fraction of the error, which is maintained through the rest of the experiment.

One can identify three phases in this process, separated by the vertical lines in Fig. 3.2a. In the first stage, resampling removes members with no correct clouds, replacing them with perturbed copies of members with a correct cloud (with the parameters used here it is rare to obtain a member with more than one correct cloud in the initial ensemble). By the end of the first phase, the ensemble consists of descen-

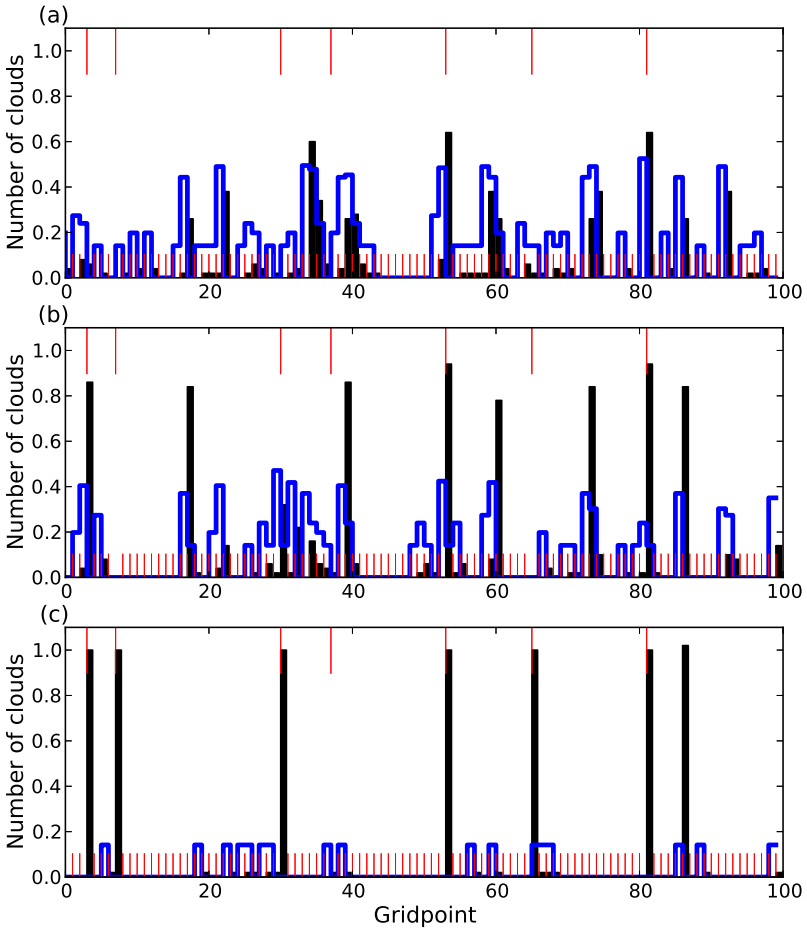

Figure 3.3: Ensemble mean cloud number (black bars) and spread (blue) for a sample run of the SIR filter with stationary cloud field and 50 ensemble members at time step (a) 3, (b) 7, and (c) 45. Thin vertical lines indicate locations of clouds in the truth.

dents of only a few members of the initial ensemble with only one correctly assimilated cloud.

An example realisation at the end of the first phase is shown in Fig. 3.3a after 3 assimilation cycles, where a set of three clouds is seen to be present in about 60% of the members, showing their common parentage. Of these three clouds two are assimilated correctly while the third one is at a position where no cloud is observed. There are also a number of grid points where there is a correct no-cloud state assimilated. The remaining grid points have little spread with only a few members containing (mainly wrong) clouds. The average error and spread decrease rapidly during this phase as the members with no correct cloud are eliminated.

The second stage in Fig. 3.2a is characterized by a slower rate of decrease in

error, but a continuing rapid decrease in spread. Since the correct clouds are part of the subset common to most members, differences in error between the members are determined by the number of incorrect clouds. Resampling removes the members with the largest error, which are those with the largest number of clouds where no cloud should be. In this short phase between time step 3 and 7, the number of grid points without a cloud in any member increases, corresponding to a rapid drop in spread. By the end of this phase, the filter has essentially collapsed, with spread only being maintained by the stochastic perturbations introduced at each resampling stage.

The example state displayed in Fig. 3.3b shows that nearly all members have a common subset of eight clouds (three in correct locations), and at many locations there are no clouds in any ensemble member. The mean error decreases slowly during this phase, since occasionally a perturbation during the resampling will change the subset of clouds common to all members, producing a member that has an additional correct cloud, or one that does not have an incorrect cloud that all other members have. Descendents of this better member take over the ensemble within a few resampling steps, but since such events are rare, the average error decreases only slowly.

During the third phase, the mean error and spread both decrease slowly (as seen in Fig. 3.2a). The rate of creation of new clouds by resampling perturbations approximately balances the tendency to reduce the number of clouds by selectively removing the members with the most clouds, and further changes in error and spread are associated only with occasional changes to the common subset of clouds. Each time a new correct cloud is produced by the resampling perturbations it rapidly spreads to the rest of the ensemble, but since the number of possible locations is large, many time steps are required until all the correct cloud locations are found. In the example analysis state in Fig. 3.3c after 45 cycles, most members have the correct solution at most locations, with one incorrect cloud and occasional additional clouds at random locations.

The rapid collapse of the SIR filter to include a common subset of clouds in all members is as expected for small ensemble sizes. Subsequently, the filter behaves as a Markov Chain Monte Carlo simulation, randomly exploring the state space, but retaining correct information from previous time steps. The effectiveness of this behaviour depends crucially on the strategy for perturbing duplicate members introduced during resampling. In this simple model, the resampling perturbations have been chosen consistently with the stochastic model dynamics, and the filter continues to converge, albeit slowly.

While the SIR filter eventually produces a good analysis for a stationary cloud field,

Figure 3.2a shows that the time to converge is too long compared to the physically motivated mean cloud lifetime of up to 30 time steps. This suggests that, for this ensemble size, the SIR filter will not be able to track changes in a time-varying cloud field. As seen from the continuous red line in Fig. 3.2a, the error initially decays at a rate similar to that for stationary clouds, but reaches a minimum value of about 55% of the error of a random field. The initial behaviour of the filter is similar, with the initial random noise disappearing within a few time steps and a common subset of clouds appearing in the majority of ensemble members. However, the evolution of the common subset is too slow to track changes in the evolving observed state and the error remains large.

Fig. 3.2b shows the corresponding results for the ETKF. For the stationary cloud field, the mean error eventually approaches a small value (well beyond the 100 time steps shown in the figure), with spread very similar to the error. Also shown with dotted lines is the error in the ensemble mean, which in theory should represent the "best estimate" of the observed state. One can see that this is in fact the case, although one has to note that the initial error of the ensemble mean is already smaller. For the ETKF there are only two stages identifiable in the figure, distinguished mainly by the behaviour of the ensemble spread. While the error decreases continuously, the spread remains roughly constant for approximately 15 time steps, then decreases together with the error.

Snapshots of model states during these two phases are plotted as for the SIR. As shown for an example realisation at time step 15 in Fig. 3.4a, by the end of the first phase clouds appear at the correct locations in the majority of the ensemble members. However, the variability at other locations, associated with incorrect clouds in random members, remains largely unchanged. Since the incorrect clouds are at different locations in the different ensemble members (in contrast to the common subset in the SIR filter), the error of the ensemble mean (also shown in Fig. 3.2b) is smaller than the mean of the errors of the individual ensemble members. This reflects the well-known reduction of RMS error for smoother fields, and does not imply that the ensemble mean is a satisfactory estimate of the observed state since the presence of low rain rates everywhere in the domain is not consistent with the dynamics of the physical model. There is a big difference between an analysis where there is drizzle throughout the domain or one where there exist a couple of isolated thunderstorms.

The gradual decrease of spread and error in the second phase is associated with the disappearance of incorrect clouds from more and more members. This process is much slower than the introduction of clouds at the correct locations. As a result even

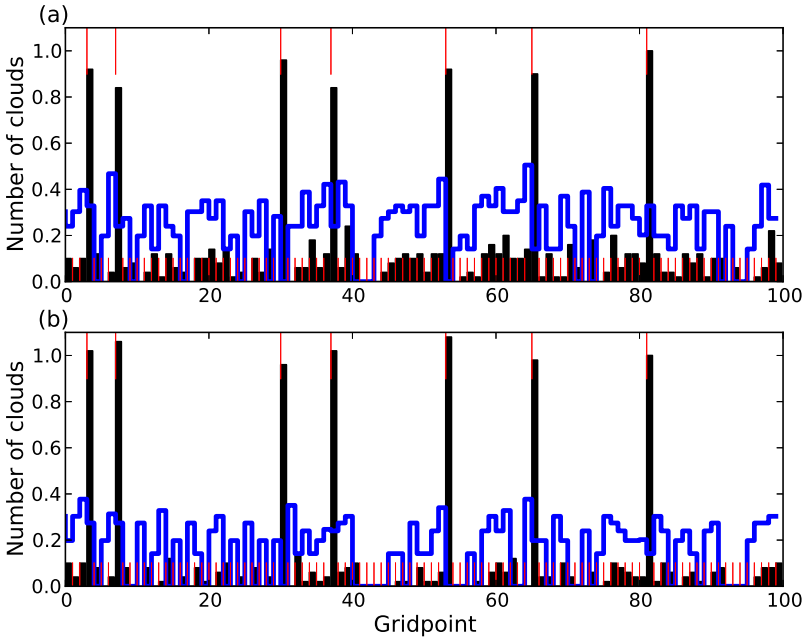

Figure 3.4: As Fig. 3.3, but for the ETKF at time step (a) 15, and (b) 45.

at time step 45 (Fig. 3.4b) there is a lot of spread at most of the grid points although the clouds are now correct in almost all ensemble members. The ETKF analysis step takes each ensemble member and adds positive and negative perturbations from other members. Since clouds occupy a small fraction of grid points in each member, there is more chance of introducing a new incorrect cloud than to remove an existing one. The slow convergence is thus a direct result of the non-Gaussianity of the errors once the locations of the observed clouds have been captured by the ensemble.

For time-varying cloud fields (red lines in Fig. 3.2b), a significant diversity is retained in the ensemble, but the error saturates at a high level since the filter is not able to remove the noise within the half-life of the clouds. In general it is clear that a method applied to the half life 30 case can only achieve a result which is similar to the one produced in the half life 3000 case within 20 to 30 time steps.

Fig. 3.5 shows how both methods end up in the constant case after 100 time steps. The particle filter has a perfect analysis with zero spread in the ensemble while the ETKF has all the clouds assimilated but still some spread at grid points where no cloud should be.

Although both the SIR filter and ETKF have large errors with a time-varying cloud field, the nature of the errors is quite different. The error in the SIR filter comes primarily from wrongly positioned clouds that are present in almost all ensemble

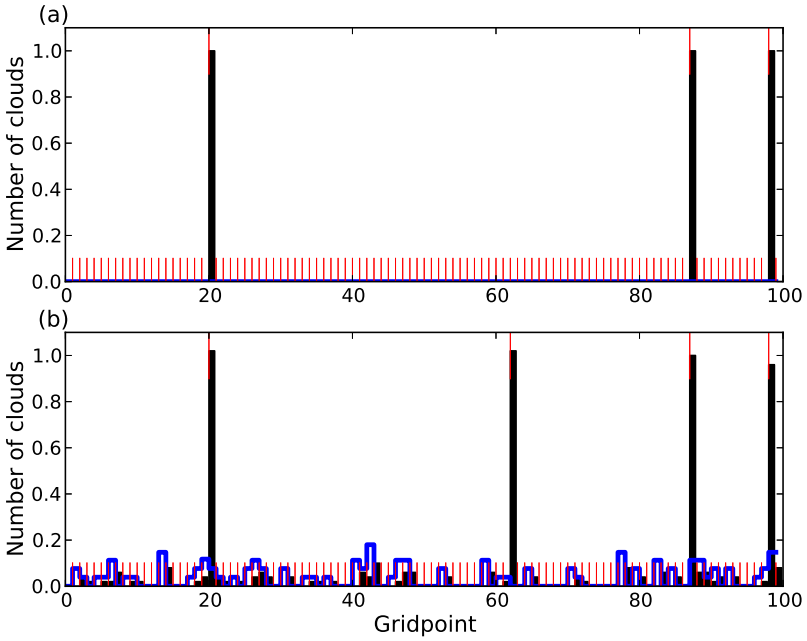

Figure 3.5: As Fig. 3.3, but at time step 100 for (a) SIR and (b) ETKF

members, while the error in the ETKF is associated with a high level of background noise. This also shows that in such a situation the RMS is not the perfect error measure as identical values correspond to totally different situations. This suggests that the optimal method for producing probabilistic predictions from the two ensembles will be different.

3.2.3 Ensemble size

With an ensemble size of 50, neither data assimilation method is able to converge to the time-varying cloud field with any degree of accuracy. To illustrate how the results change with ensemble size, a *final error* was computed for each experiment. This was estimated as the error after 500 time steps for experiments with a stationary cloud field and 100 time steps for the 30 time step lifetime cases, since changes in error after this time were found to be negligible. In the following plots this final error is drawn against the ensemble size.

In Fig. 3.6a it can be seen that any ensemble size greater than about 10 is sufficient for the SIR filter to converge for stationary clouds, while for the time-varying cloud field the error decreases rapidly with increasing ensemble size up to about 20 members, after which the decrease in error is slow. Except for very small ensemble sizes, the spread is about half the magnitude of the error.

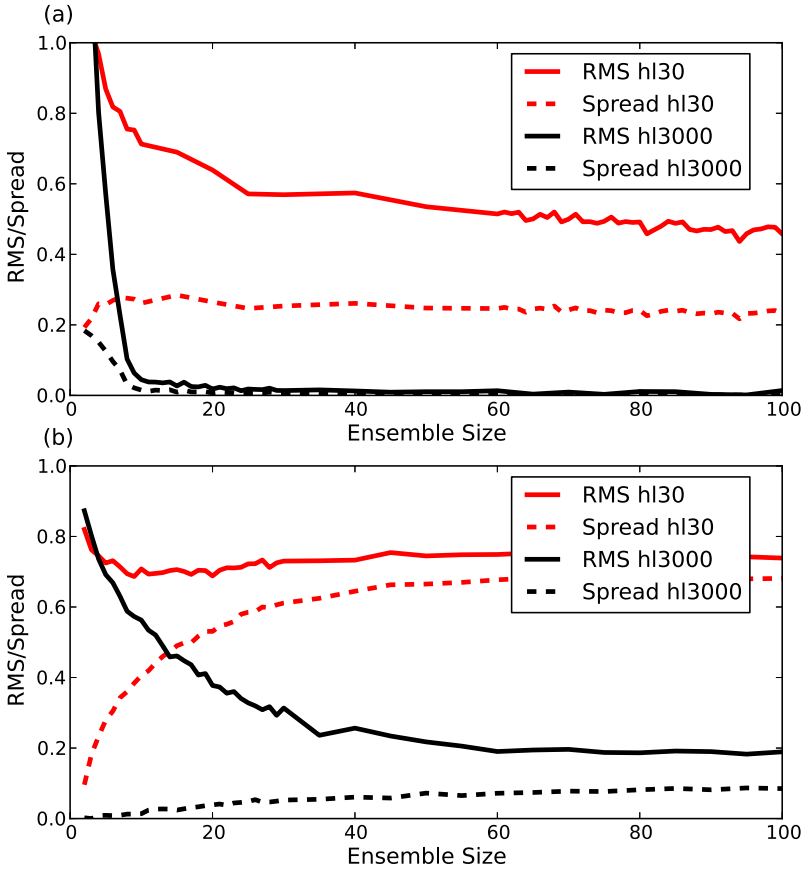

Figure 3.6: Final error (solid, as defined in text) and ensemble spread (dashed) as function of ensemble size for (a) SIR and (b) ETKF with half-lives of 3000 (black) and 30 (red) time steps.

The ETKF collapses for small ensembles with a stationary cloud field (Fig 3.6b). Collapse in this context means vanishing of the spread. Therefore for these small ensembles the ETKF is not able to represent the error of the analysis to any degree. This situation where the ensemble spread is much smaller than the actual error is also called overconfidence, meaning the method thinks it is better than it actually is. The results improve significantly with increasing ensemble size, up to about 40 members, with final error reaching about 20% of the random value and ensemble spread about half the final error.

For the time-varying cloud field there is almost no improvement with ensemble size, and even for an ensemble size of 100, the final error is only about 5% better than for an ensemble size of 15. Therefore the results for the half-life 30 case do not depend much on the ensemble size.

One can identify an increase in the error of the half life 30 case for the ETKF with

ensemble size larger than 20. This is the result of the lack of tuning. For the stochastic cloud model the integer states can produce too much spread under certain conditions. In this case deflation would improve the results. But as the effect would not be huge and it would change the generality of the results, it is not applied. The behaviour of ensemble spread relative to error shown in Fig. 3.6 is also found in the subsequent experiments, therefore spread will not be plotted on the remaining figures.

3.2.4 Localization

As assimilating data in local regions independently can drastically reduce the number of degrees of freedom in the system, this was also applied here. There is a lot of potential in improving the performance of the filter already for small ensemble sizes. As can be seen in Fig. 3.7a, localization has a major effect on the performance of the SIR filter (difference between the two red lines), leading to convergence with even smaller ensemble sizes than achieved by the non-local filter for the stationary cloud field.

A major reduction in final error is found for the time-varying cloud field, with errors less than 20% for ensemble sizes larger than 20. This result is not surprising: On average, the observations have 10 clouds scattered over 100 possible locations, for a total of 100^{10} possible states. This is vastly larger than the ensemble sizes used here and it does not matter much whether 50 or 100 ensembles are used as the difference to 100^{10} is huge. Localization decomposes the domain into 100 subdomains, each with only 2 likely states (cloud or no-cloud), for a total of about $100 \cdot 2$ possibilities. An ensemble of a few tens of members can sample this space within the cloud half-life of 30 time steps.

On the other hand, the effects for the LETKF are modest, with only small improvements in final error, particularly for smaller ensemble sizes. As shown in the previous Section, the primary source of error for the ETKF is the continual creation of clouds at incorrect locations by the assimilation increments. This is not affected in a systematic way by localization. As the stochastic model does not contain any correlations between grid points, localization only changes the dimensionality, but not the nonlinearity or non-Gaussianity.

3.2.5 Averaging

In contrast to localization, observation averaging should have the effect of making the distribution of the errors more Gaussian, potentially leading to better results for

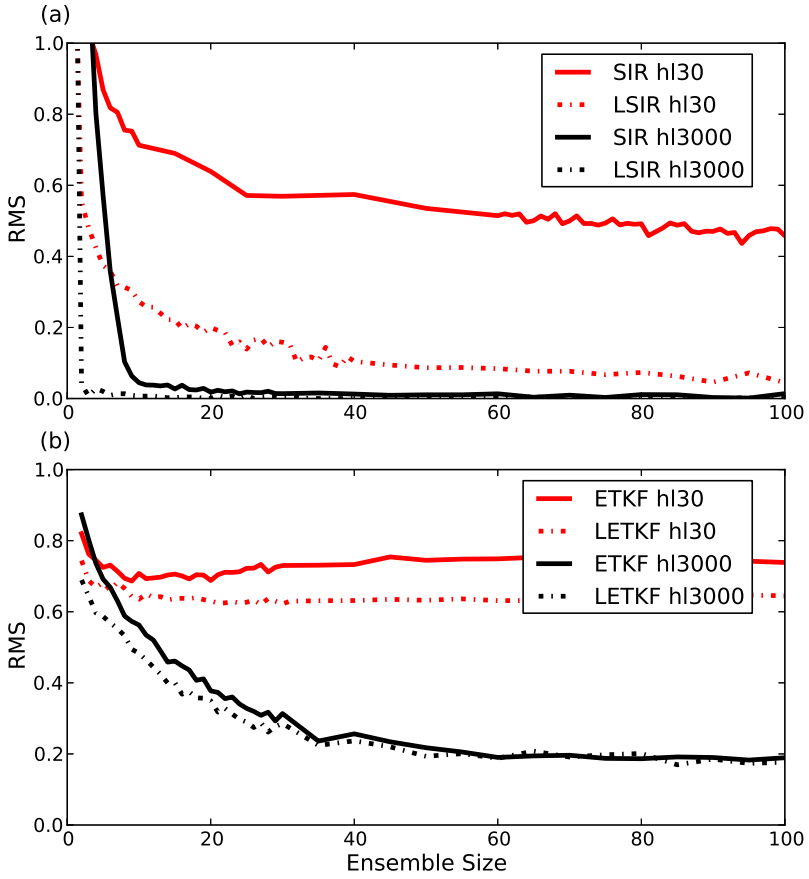

Figure 3.7: Final error as function of ensemble size for (a) SIR and (b) ETKF, with half-life 30 (red) and localization (dash-dotted) and half-life 3000 (black).

the ETKF. The adding of clouds in regions of 10 grid points produces smoother fields and reduces the intermittency. Figure 3.8 shows the final errors for experiments with observation averaging. For this figure RMS errors have been computed from averaged observations and model states, and normalized by the difference between two random states in this measure. An error of zero would thus imply that the number of clouds within each 10 grid point region was correct, but not necessarily their locations. The errors thus reflect the ability of the methods to solve the simpler problem of producing the correct density of convective clouds over subregions, and are not directly comparable to the errors shown on the previous figures.

With averaged observations the SIR filter converges even with small ensemble sizes (Fig. 3.8a) for the dynamic situation. As with localization, the problem is decomposed into a smaller number of problems as there are now only 10 of these subregions present. Although these subregions have now more than one possible state, this problem is eas-

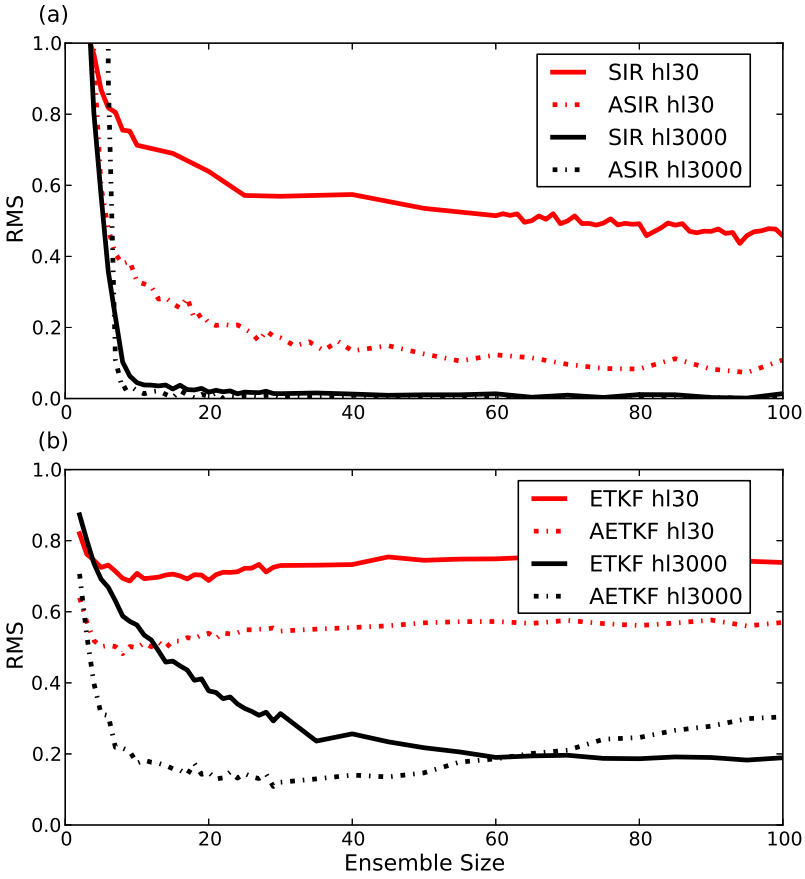

Figure 3.8: As Fig. 3.7 but with dash-dotted lines indicating experiments with observation averaging.

ier to solve than the full problem. Unlike localization, averaging changes the statistical character of the smaller problems, but the reduction in dimensionality is similar.

For a stationary cloud field, the ETKF with averaged observations (AETKF) (Fig. 3.8b) shows a large reduction in the ensemble size required to obtain small errors in the region of small ensembles (10-20 members). The increase in error for larger ensemble sizes is surprising, but additionally performed experiments (not shown) suggest that this could be corrected by increasing the covariance deflation factor with increasing ensemble size to produce errors at least as small as those for smaller ensembles. This is a similar effect to the one mentioned at the end of Section 3.2.3.

Even for a time-varying cloud field the performance of the ETKF is significantly improved by averaging, although for the relatively small averaging regions used here the errors remain large. Averaging is not able to solve the problem of the stochasticity of the fast evolving cloud field.

3.3 Discussion

The aim of this Chapter was to test convective-scale data assimilation algorithms using a simple model based on a stochastic birth-death process. The stochastic dynamics model the extreme nonlinearity of the convecting atmosphere, where a storm can appear from one observation time to the next. The spatial Poisson distribution in the simple model represents the intermittency and lack of correlation characteristic of remote sensing observations of convective clouds, where the space and time scales of changes in the convection are comparable to the resolution of the observations.

Since the model does not include any dynamical balances that might or might not be present at the convective-scale, it constitutes an extreme test for data assimilation algorithms. Both, the SIR filter and ETKF fail to produce good results for realistic parameter values of cloud density and lifetime. The performance could undoubtedly be improved somewhat by optimizing parameters such as the covariance inflation or the resampling probability in the particle filter, but the results would be unlikely to change qualitatively.

Although the size of the ensembles tested here is comparable to the number of grid points in the model ($n = 100$), they are small in comparison to the number of possible states of the system (roughly $2^n \approx 10^{30}$). So the results are not likely to be improved by using a larger ensemble size unless it gets to the order of system states.

The SIR filter rapidly collapses to a state in which variance is only maintained by the perturbations associated with resampling after members are eliminated. Interestingly, the filter can eventually converge to the observed state, since any correct cloud locations found by the random perturbations are retained in the ensemble. Since the rate of convergence is controlled by the resampling perturbations, rather than the importance weighting, strategies like an improved proposal distribution (van Leeuwen, 2009) may have the greatest potential to improve the filter performance. In any case the true statistical properties of the collapsed ensemble need to be taken into account in generating probabilistic forecast products, since the weights of the ensemble members produced by the filter provide little information.

Localization and observation averaging both produced dramatic improvements in the performance of the SIR filter since they both drastically reduce the dimensionality of the space that must be explored by the resampling perturbations. In more realistic models, however, these methods might cause problems by violating dynamical balances that couple different spatial regions.

The ETKF collapses for small ensemble sizes, but otherwise captures the correct

cloud locations quickly. However, the ensemble is plagued by large numbers of incorrect clouds that contaminate the ensemble mean. This occurs because negative clouds are not possible, so that the nonlinear dynamics rectify the analysis increments, producing a non-Gaussian distribution of variability in the ensemble.

Averaging of observations has the potential to correct this problem, since the Poisson distribution will converge to Gaussian as the averaging region becomes large enough to contain many clouds. Indeed, significant improvement with averaging was found, although for the relatively small averaging region used here, the errors remain large. Localization, on the other hand, seems to have no benefit for the ETKF in this environment.

Caution is necessary in extrapolating the results of this simple model to operational data assimilation in convective-scale numerical weather prediction. The model proposed in this Chapter represents extreme non-Gaussianity and nonlinearity. It is not a representative model of the range of physical processes that occur in high-resolution data assimilation, but rather a model of the particular additional processes that distinguish the convective-scale problem from the better-understood synoptic scale problem. The performance of a particular method in a particular weather regime will be influenced, and often dominated, by processes other than those described in the simple stochastic model.

However, the types of errors that show up in the idealized model are likely to appear, at least in some situations, in the full problem. The simple model should provide some insight into which modifications or improvements are likely to help ameliorate these errors, and are thus worthy of investigation in the full system.

The idealized model proposed here is a starting point for a hierarchy of models, where a stochastic process representing convection is introduced into simple dynamical models (e.g. shallow water, or quasi-geostrophic), and the convection is coupled to the large-scale dynamics by a simple closure assumption, where the rate parameter depends on the large-scale fields. Conversely, the simplicity of the current model lends itself to a more formal mathematical analysis. There is an extensive literature on spatial birth-death processes (e.g. Cox and Isham, 2000), including extensions to nearest-neighbour interactions (Møller and Sørenson, 1994). As these models as well as the stochastic model have no dynamical basis for initiation and interaction of clouds, the next step in the hierarchy of models is to create a model based on the shallow water equations. This modified shallow water model will be presented in the next Chapter.

Chapter 4

Modified shallow water model

In this Chapter the more sophisticated model in the hierarchy is presented. It is a modified model based on the shallow water equations and therefore intended to capture the dynamical interaction between clouds. The characteristics of the model clouds are designed to be a closer match to real convection than the clouds of the stochastic model. The model should therefore be a much more realistic simplification of the atmosphere. The modified shallow water model is introduced in Section 4.1. Different properties like the cloud size distribution and life time are analyzed in Section 4.2 and its interaction with orography is shown in Section 4.3. Finally a brief discussion of this Chapter is given in Section 4.4. The modified shallow water model is then used for data assimilation experiments with an LETKF. These results are shown in the next Chapter 5.

4.1 Model description

The stochastic toy model used for data assimilation experiments in the previous Chapter misses some features of convection. Therefore the model presented here is designed to represent more of the convective processes, including the classic life cycle of a convective storm. The mass and temperature perturbations that cumulus clouds introduce into their environment are communicated in space by radiation of gravity waves. Neighbouring clouds may be initiated or enhanced by uplift, or suppressed by downward motion associated with these waves. On shorter space scales or longer time scales the effect of convection on humidity is also important, but this will be ignored here.

The primary importance of gravity waves in convection is where the focus lies and

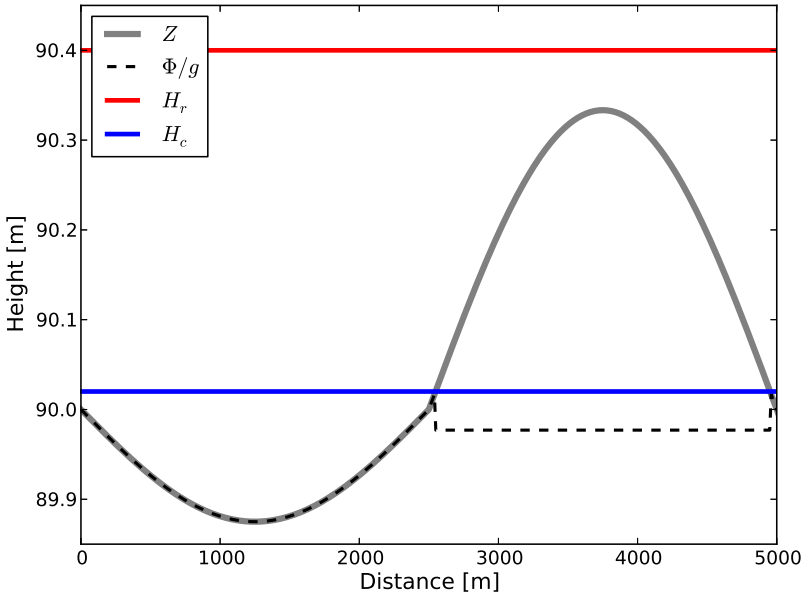

Figure 4.1: Sketch showing the modification of the geopotential ϕ in a cloud which is still below the rain threshold.

therefore suggests a model based on the shallow water equations:

$$\frac{\partial u}{\partial t} + u\frac{\partial u}{\partial x} + \frac{\partial \phi}{\partial x} = K\frac{\partial^2 u}{\partial x^2}, \tag{4.1}$$

$$\frac{\partial h}{\partial t} + \frac{\partial (uh)}{\partial x} = K\frac{\partial^2 h}{\partial x^2}, \tag{4.2}$$

where u is the fluid velocity, h the depth of the fluid, H the height of the topography which will be used later, and the geopotential $\phi = g(H + h)$ is the absolute fluid layer height $(H + h) = Z$ multiplied by the gravitational acceleration g. A diffusion term with constant K has been included in both equations.

To provide a representation of cumulus convection, the shallow water equations must be extended to include conditional instability, i.e. positive buoyancy due to latent heat release in ascending, saturated air. This will be accomplished by modifying the geopotential ϕ. The standard definition of ϕ is based on the height of the fluid surface, and its gradient provides a momentum forcing away from regions of increased surface height. This will be altered when the height exceeds a threshold H_c, representing the level of free convection. At these positions, the geopotential is replaced by a relatively low constant value ϕ_c, as illustrated in Fig. 4.1. The gradient of geopotential will thus force fluid into the region of decreased geopotential, increasing the fluid depth there. As can be seen in the sketch, the forcing only happens at the border of the

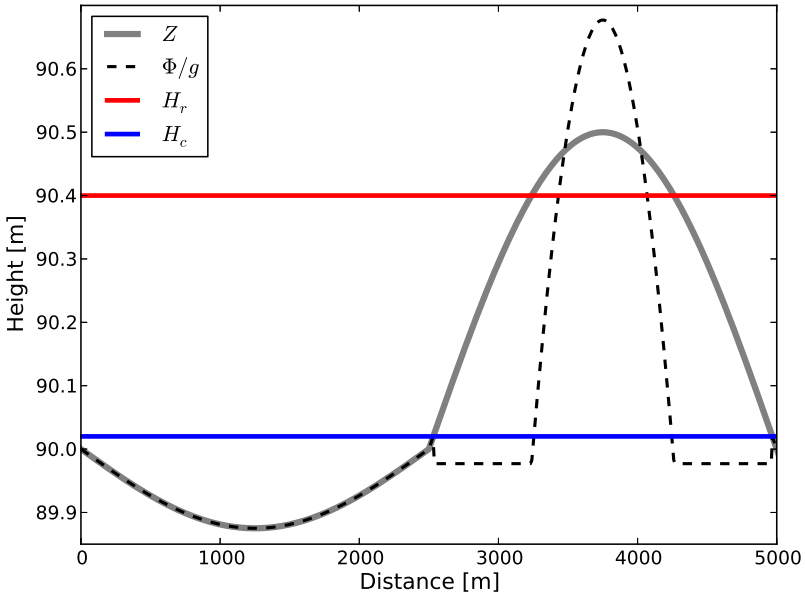

Figure 4.2: Sketch showing the modification of the geopotential ϕ in a cloud which is above the rain threshold. Only the influence of the rain is plotted here, not the rain itself.

threshold. Therefore quite large gradients can exist which more sophisticated designs of the potential could disarm. But for simplicity the model presented here runs with the large gradients.

Once such a "cloud" forms, the width of the region with altered geopotential will collapse until limited by diffusion, as occurs for cumulus clouds in simulations with kilometre-scale weather prediction models. This is not entirely realistic, because in nature the lifetime of a cumulus cloud in an unsheared environment is limited by the formation of heavy precipitation particles that eventually overcome the postive buoyancy and turn the updraft into a downdraft.

This effect will be mimicked by adding a rain water mass potential r to the geopotential, which can be seen in Fig. 4.2 where the potential increases at the grid points where the water level is high enough and instantly produces rain. As soon as there is enough rain present its influence on the potential is large enough to force the water level down again, as can be seen in the center of the cloud. Therefore a separate conservation equation for r is then added to the model equations, with the following source and sink terms. Rainwater is produced when the fluid level exceeds a threshold value H_r, and is rising (u has positive convergence). This ensures that rain is only built when the cloud is still growing. The rain production threshold is set higher than the threshold for buoyancy to ensure that rain production is delayed relative to the

onset of the cloud circulation. Removal of rain by precipitation is modeled by a simple linear relaxation towards zero.

Finally the modified shallow water equations are:

$$\frac{\partial u}{\partial t} + u\frac{\partial u}{\partial x} + \frac{\partial(\phi + \gamma r)}{\partial x} = K\frac{\partial^2 u}{\partial x^2} + F, \tag{4.3}$$

where

$$\phi = \begin{cases} \phi_c + gH, & Z > H_c \\ g(H + h), & \text{otherwise.} \end{cases} \tag{4.4}$$

A forcing term F has been added to the momentum equation and will be discussed further in Section 4.1.2. The continuity equation reads

$$\frac{\partial h}{\partial t} + \frac{\partial(uh)}{\partial x} = K\frac{\partial^2 h}{\partial x^2}, \tag{4.5}$$

and the equation for rain water is given by

$$\frac{\partial r}{\partial t} + u\frac{\partial r}{\partial x} = K_r\frac{\partial^2 r}{\partial x^2} - \alpha r - \begin{cases} \beta\frac{\partial u}{\partial x}, & Z > H_r \text{ and } \frac{\partial u}{\partial x} < 0 \\ 0, & \text{otherwise,} \end{cases} \tag{4.6}$$

where α and β are constants discussed below. The constant γ is used to convert rain into a potential and therefore is just 1 m^2 s^{-2}. This is needed as rain is kept nondimensional for simplicity. Note that the equations are written for a single horizontal dimension x, but could be trivially extended to two horizontal dimensions. Coriolis force could also be taken into account but will only have an effect on larger scales which are not considered here.

4.1.1 Parameter selection

The values for the physical parameters are chosen to produce realistic space and time scales for the model clouds, as explained in the following list. The parameters are chosen from a mixture of estimating from relating the model to the real atmosphere and tuning to get closer to the desired scales.

- $H_0 = 90$ m is the initial water level and chosen to give a gravity wave speed of roughly 30 m/s which is a value typical for the gravest internal mode in the troposphere (Gill, 1982).

- $H_c = 90.02$ m and $\phi_c = 899.77$ m^2 s^{-2} give a reasonable cloud fraction of roughly

five percent, and a time for the cloud to develop to full height of around half an hour.

- $H_r = 90.4$ m and $\beta = 3$ imply a lag between cloud and rain formation of about 15 min, and a cloud lifetime of about 1-2 hr.

- $\alpha = 2.5 \cdot 10^{-4}$ s^{-1} corresponds to a half-life for the influence of rain of roughly 1 hr. This gives persistence of rain and associated negative buoyancy even after the collapse of the height perturbation.

Getting a desired behaviour of the model is often possible by tuning different parameters. For example the time between initiation of a cloud and first rain can be influenced by either changing H_r, H_c or ϕ_c. The values of K and K_r are chosen mainly for numerical smoothness and depend on the resolution of the discretized equations but also have an important effect on the size of the clouds which will be discussed later. These aspects are discussed in Section 4.1.3 below.

4.1.2 Initiation of convection

In nature convective clouds are often initiated by low-level disturbances that rise boundary layer air to its level of free convection. The modified shallow water model presented here also requires a trigger mechanism to initiate convection. This is done by considering two processes explicitely here: perturbations originating below cloud base associated with the convective boundary layer, and flow over orographic obstacles.

In addition, secondary initiation by existing storms can also trigger clouds if they emit gravity waves with a large amplitude or influence the water level due to mass conservation. It will be shown later that gravity waves radiating from convective clouds can provide sufficient lifting to initiate new convection even in this simple model. Another secondary initiation mechanism, lifting of air over cold pools generated by evaporation of precipitation, will not be considered.

Convective clouds are often triggered by buoyant plumes in the dry convective boundary layer beneath cloud base which provide vertical displacements that can initiate convective updrafts. This process can be modeled by adding random convergent wind perturbations into the model field. These perturbations are then able to elevate the fluid field above the cloud threshold. In particular, perturbations F_n of the form

$$F_n = \bar{u} \frac{\partial}{\partial x} \left(e^{-(x-x_n)^2/l^2} \right), \tag{4.7}$$

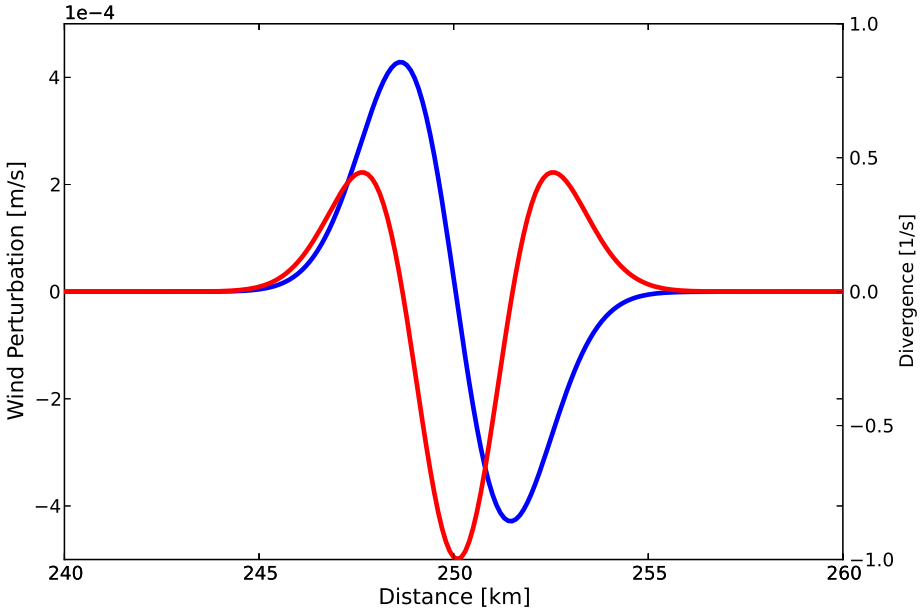

Figure 4.3: Wind perturbation (in blue) and associated divergence field (in red) for the perturbation described in the text. The center of the perturbation is located at km 250.

are added to u, with an amplitude $\bar{u}$ and length scale l, centered at location x_n. In the simulations that include this method of convective initiation, perturbations are added each time step at a random number of locations with random positions. For the results presented in this work, $l = 2000$ m, $\bar{u} = 0.005$ ms^{-1}, and the number of locations is drawn from a Poission distribution with mean 4.

The shape of such a perturbation is illustrated in Fig. 4.3. The divergence field has a minimum (=convergence) at the center of the perturbation and is decaying to zero within 2 km. It is accompanied with two weaker divergence zones to ensure that the mean wind is not changed. This length scale is related to the size of boundary layer eddies, which in nature scale with the depth of the subcloud layer. This scale is in the range of 1 km. The perturbations described here are added at random locations x_n at a rate of 1.6×10^{-6} m^{-1}s^{-1}, which corresponds to approximately one perturbation in a 1 km region every 10 minutes. This value is in the range of a typical eddy turnover time for the convective boundary layer.

Examples of the initiation of convection by orography will be presented in Section 4.3.

4.1.3 Numerical implementation

The numerical implementation of the shallow water equations follows Gohm and Mayr (2004), based on previous work by Schär and Smith (1993a,b). The equations are discretized by standard second-order centered differences on a staggered grid. A Robert-Asselin-Williams (RAW) Filter (Williams, 2009, 2011) is used for time-smoothing of all model variables. This filter is an improvement to the standard Robert-Asselin Filter.

For the simulations shown in this paper, the domain size L is set to 500 km, the horizontal resolution $dx = 500$ m, and the time step $dt = 5$ s. The diffusion constant is $K = 25000$ m^{2}s^{-1} for h and u and a smaller $K_r = 200$ m^{2}s^{-1} for the rain variable. As with models that simulate convection by solving the full equations of motion with kilometre-scale resolution, the resolution and diffusion control the size of the model clouds. As a result, the values of the physical parameters identified in Section 4.1.1 might have to be reconsidered if a different numerical implementation or resolution were used. The advantage of this is the possibility to tune the model if one wants different cloud sizes to do resolution experiments.

4.2 Randomly triggered convection

As a first example, results will be shown for convection triggered at random in space and time by boundary layer disturbances. In these experiments no mean wind is imposed and no orography is present. The model is initiated with a constant surface at 90 m. The triggering disturbances are as described in Section 4.1.2, and lead to a statistically steady ensemble of clouds throughout the domain. The background noise starts with the first time step and stays active for the whole simulation. The results should be comparable to radiative-convective equilibrium simulations of convection (Tompkins and Craig, 1998; Cohen and Craig, 2006) produced with a Cloud Resolving Model (CRM). First, the life cycle of a single cloud within the simulation is considered in the next Section and after that in Section 4.2.2 the cloud size distribution and a measure of clustering are presented.

4.2.1 Life cycle of a convective cloud

Figures 4.4 and 4.5 show an example of the temporal evolution of the fluid level, rain content and divergence fields in a typical convective cloud. The plots focus on a small subdomain of only 14 grid points (7 km) and each differently colored line corresponds to a different time with the interval being 6 minutes. In Fig. 4.4 one can see the

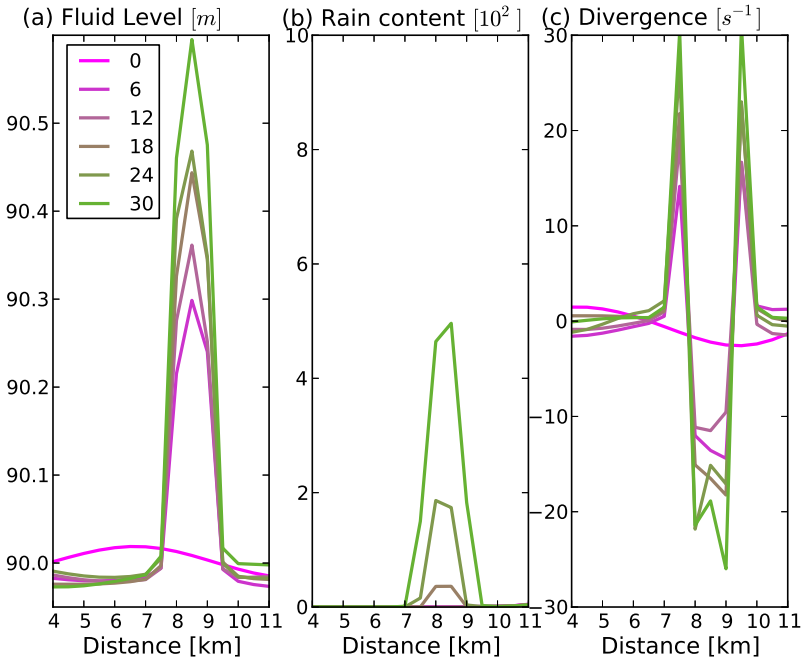

Figure 4.4: Updraft phase of a cloud. The different lines correspond to different times in minutes.

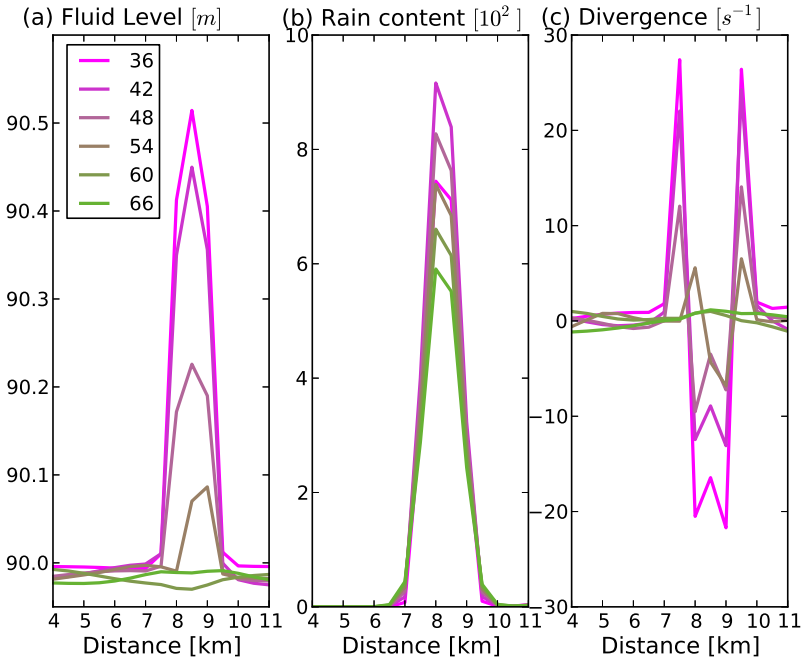

Figure 4.5: Downdraft phase of a cloud. The different lines correspond to different times in minutes.

building and strengthening of an updraft. At time 0 there is just a small positive perturbation in the fluid level. This initial peak may have come from the background

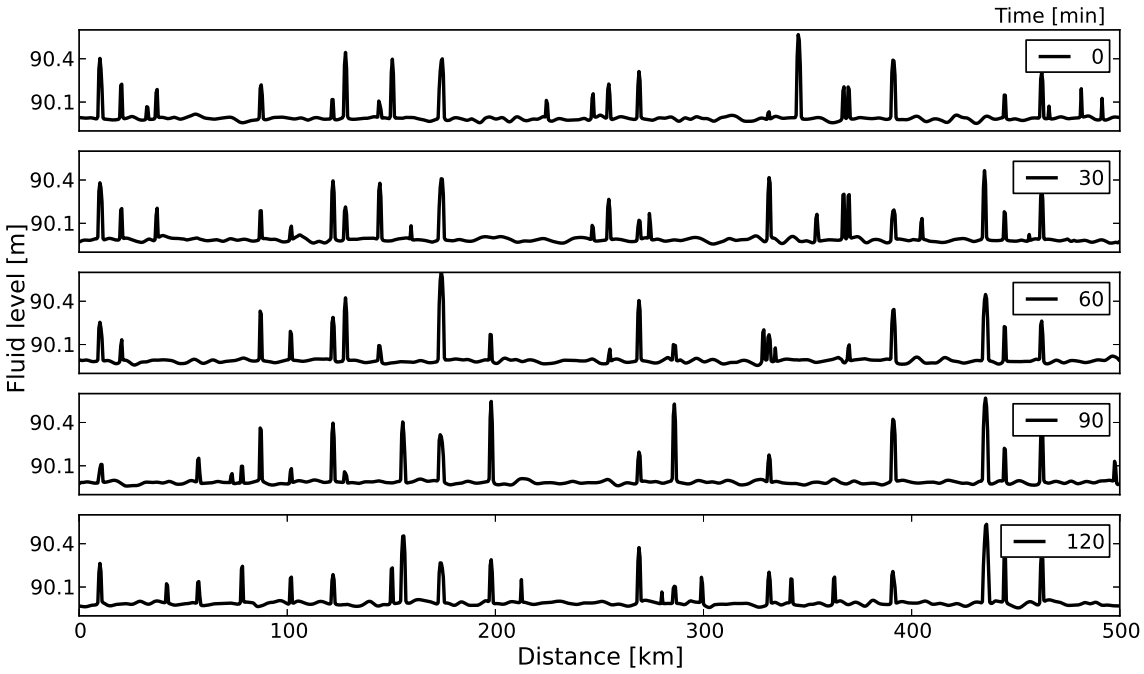

Figure 4.6: Typical evolution of the cloud field in the whole model domain.

perturbations or from the influence of neighboring updrafts. This wave peak moves a little to the right and reaches above H_c until the next plotted time. Therefore the modifed geopotential causes fluid to flow into the updraft of the cloud region and the fluid level continues to rise.

The divergence plot shows a steady convergence into the cloud, with sharp positive divergence peaks on both sides of the updraft. It also shows that the convergence was already located a little bit to the right of the fluid level maximum at time 0. Rain starts to accumulate after about 15 minutes when the fluid level reaches above 90.4 m and is still converging. In this example cloud, the downward force associated with the rain becomes strong enough to prevent further increase in the fluid level after about 30 minutes, although the accumulated rain maximum is not reached until minute 42. After another 30 minutes the height perturbation has collapsed leaving gravity wave perturbations that propagate away. The weakening of the convergence zone during the collapse of the cloud is clearly visible in Fig. 4.5c. As there is still a significant amount of rain present at 60 minutes, it makes this location unfavourable for convection for some time afterwards. Remember that the parameter α is the only one influencing the time it takes for the rain to be removed.

A view of the ensemble of clouds over the full 500 km domain is shown in Fig. 4.6 which is produced from the same model run as the cloud pictures above. The fluid level

is displayed at 30 minute intervals over a period of 2 hours to focus on the distribution of clouds. There is a wide range of differently sized clouds with different life cycles visible. Some last only for a short time while others persist throughout the 2 hour period.

In addition to the simple life cycle illustrated in Figs. 4.4 and 4.5, another typical cloud behaviour occurs when the rain starts to push down the fluid level. This happens when the rain does not manage to completely kill the updraft. When the water level is getting pushed down and between the two thresholds, the upward forcing is still active. During this phase rain is also being removed. If too much rain is removed, the upward forcing will be stronger again and the cloud gets pushed upward again. This is visible for example in the first cloud from the left or the cloud around kilometre 130 in Fig 4.6 which still exists after 2 hours. In such cases clouds can undergo multiple phases of growth and decay until they are finally destroyed either by the rain, the background perturbations or the influence of gravity waves emitted from other clouds.

The number and location of clouds in the domain will be determined by a combination of effects: triggering of clouds by the imposed wind perturbations or upward displacements associated with gravity waves, and suppression by downward gravity wave displacements or negative buoyancy due to rain that remains from an earlier cloud. The effect of these processes on the spatial distribution of clouds will be looked at in the next Section.

4.2.2 Statistics of convection

Statistics of the size and relative location of clouds have been calculated using output every 30 minutes from a long model integration. This almost 10 year long integration proved useful to get reliable statistics. Figure 4.7 shows a logarithmic histogram of the cloud size where a cloud is defined as a region where $Z > 90.04$ m. This threshold is chosen to avoid small perturbations around the cloud threshold of $H_c = 90.02$ m being identified as a cloud. The distribution approximately matches the exponential form (dashed black line) expected from theory (Craig and Cohen, 2006) and obtained in cloud resolving models in a radiative-convective equilibrium (e.g. Figure 2 of Cohen and Craig (2006)).

Some measures calculated from these statistics are the average cloud size around 3.4 grid points, with the most common value being about 2 grid points and values of more than 10 grid points being very rare. The average number of clouds present in the domain at a given time is 14.9, which leads to an average convective area fraction

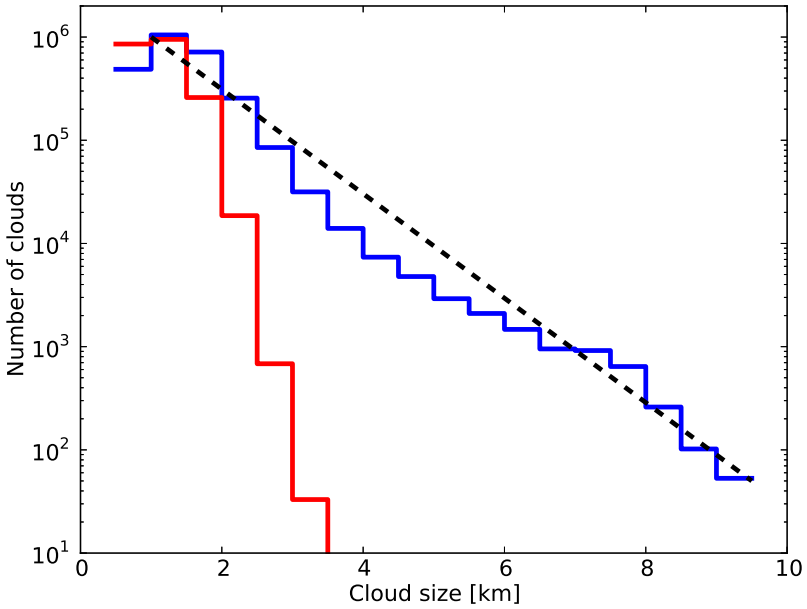

Figure 4.7: Logarithmic cloud size distribution for clouds with a height above 90.04 (in blue) and clouds which do not precipitate (in red).

of about 5% of the domain. The steeper part of the distribution between 1 and 3.5 km is based only on small clouds which do not precipitate and are plotted in red. This distribution is very sharp and goes down from almost 10^6 to 10 in this range. The rest of the larger clouds is usually associated with rain and decays slower. The red line shows more small clouds than the blue line as the algorithm sometimes mistakenly counts one big cloud with rain in its center as two smaller clouds with no rain.

In Fig. 4.8 a histogram of the distance between pairs of clouds is displayed. This distance is calculated between all clouds present in one snapshot where the clouds are defined as stated above. The distance is defined as going from one center of a cloud to another one and the calculation is getting looped over all clouds. The dashed line indicates the uniform distribution that would be expected if the positions of the clouds were completely uncorrelated (a spatial Poisson process). Cloud separations of less than 3 km are uncommon since this is in the range of the average width of individual clouds and their presence permits other clouds at this distance. However there is a pronounced peak in the distribution at a distance of approximately 3.5 km, which appears to be associated with strong gravity wave perturbations initiating new clouds near existing ones. The effect captured here is from gravity waves emitted from the presence of the clouds, not the presence of the downdrafts. To analyze the effect of the influence of the rain one would have to look at the distance to clouds some time after the center cloud has disappeared. Separation distances of 6 to 18 km are less

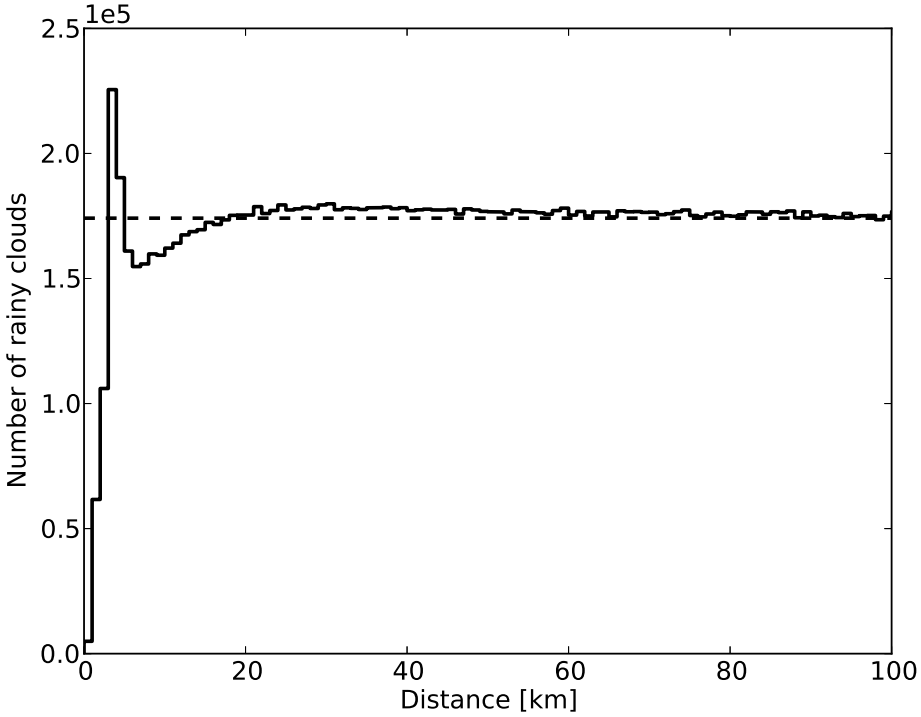

Figure 4.8: Distribution of distance between two different clouds.

common, reflecting a suppression of convection due to reduced average fluid levels in the vicinity of a cloud. This is a similar minimum than the one at 0 km.

At long distances the separation frequency approaches the value expected for uncorrelated cloud locations and is almost constant for distances larger than 100 km. This pattern of enhanced frequency of clouds at very short separations, with a region of reduced frequency at somewhat larger distance, is also present in Figures 6a and 8a of Cohen and Craig (2006).

Finally the behaviour of the model is assessed by looking at the power spectrum of the wind field in Fig. 4.9. The spectrum is also calculated from a 10 year long model integration. Two maxima can be found at roughly 3 and 20 km with two minima at 5 and 60 km. The maxima are likely related to the preferred separation distances between clouds and the minimum at 5 km to the minimum in Fig. 4.8. These results are similar to the findings of Lane and Zhang (2011) who looked at the spectrum in an idealized cloud-system-resolving model. The gradient in the spectrum (i.e. spectral slope) is close to the one of an exponential with the power (5/3) as indicated by the blue line. This slope is also found in the atmosphere (Nastrom and Gage, 1985) as well as in model simulations (Bierdel et al., 2012).

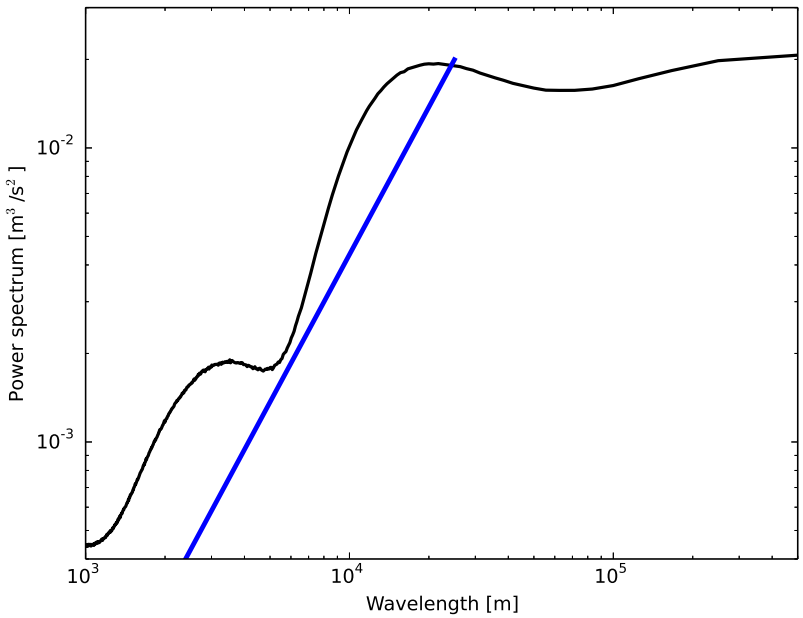

Figure 4.9: Mean power spectrum of the wind field (black) and reference line of an exponential with the power (5/3).

4.3 Orographically triggered convection

In this Section, results are presented for convection triggered by flow over a bell shaped mountain with a half-width of 10 km and a height of 0.2 m. This time no random perturbations are put into the model, but a positive mean wind is introduced. For sufficiently large wind speeds the resulting flow can displace the fluid level above H_c over the mountain and initiate a cloud. The orography and the mean wind pose problems to the numerical stability and it was found necessary to reduce the time step to 1 s and increase the diffusion to 30000 $\mathrm{m^2 s^{-1}}$.

Results for two values of initial mean wind speed are shown in the form of space-time (Hovmöller) plots indicating where the fluid depth exceeds 90.05 m. Fig. 4.10 shows a 30 $\mathrm{ms^{-1}}$ wind simulation and 40 $\mathrm{ms^{-1}}$ was used for Fig. 4.11. In case of a mean wind of 30 $\mathrm{ms^{-1}}$, clouds are present over the mountain almost at all times, growing a small distance on the upwind side and decaying some kilometres downwind. Sometimes clouds or small clusters of clouds are initiated a larger distance upwind from the mountain by gravity waves (Fig. 4.10) but they never move further than 40 km away from the center of the mountain.

A different behaviour occurs for wind speeds larger than the intrinsic gravity wave speed of the fluid. For a mean wind of 40 $\mathrm{ms^{-1}}$ (Fig. 4.11) cloud initiation happens in a similar way than for the slower mean wind. But although most of the clouds decay

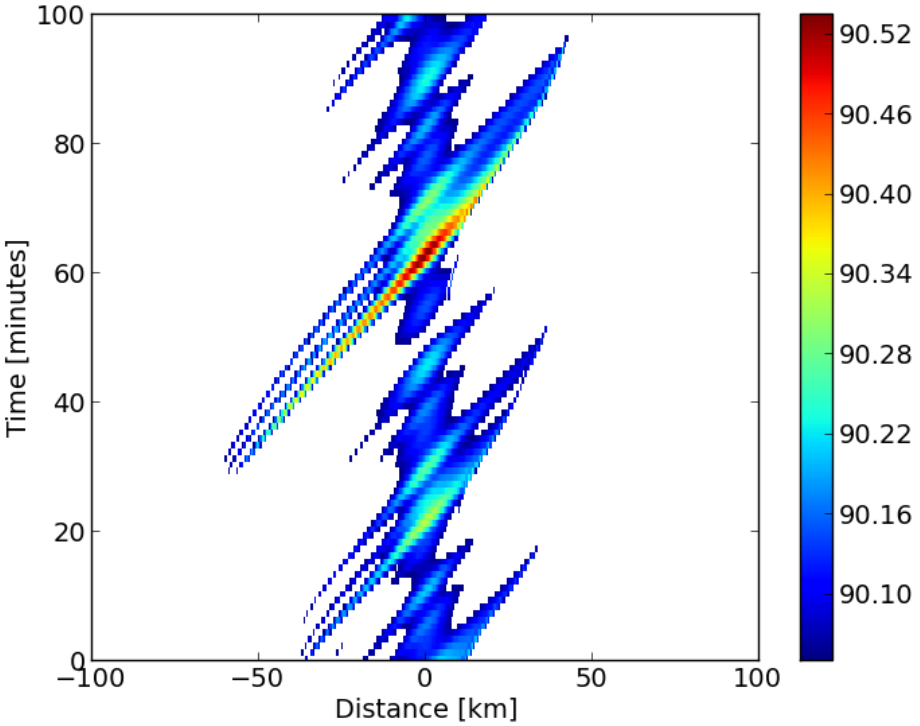

Figure 4.10: Hovmöller diagram for a mean wind of 30 ms^{-1}. Marked are only grid points where the water level is above 90.05 m.

on the downwind side, some are advected downstream away from the mountain. These interact and merge, eventually leading to a sequence of propagating convective storms with a separation distance of roughly 75 km, moving at a speed of about 35 ms^{-1}. Due to the setting of the model, none of the possible influences are strong enough to bring down these clouds and they enter a steady state. The propagation speed of the clouds is inverse proportional to the slope of the slope of the lines in the two figures, which is a little bit steeper for the slower wind speed.

These two regimes compare favorably with simulations of flow over a ridge by Chu and Lin (2000). The behaviour for a wind speed of 30 ms^{-1} is comparable to the quasi-stationary convective system in their regime (II), with cells developing on the upstream side of the ridge and decaying on the downstream side. The results for a windspeed of 40 ms^{-1} correspond to regime (III) with quasi-stationary and downstream propagating systems.

It should be noted however that the simple model here does not reproduce the full range of phenomena described by Chu and Lin (2000) or in the more recent study of Miglietta and Rotunno (2009), which involves complex interactions with evaporation-driven cold pools that are not represented here.

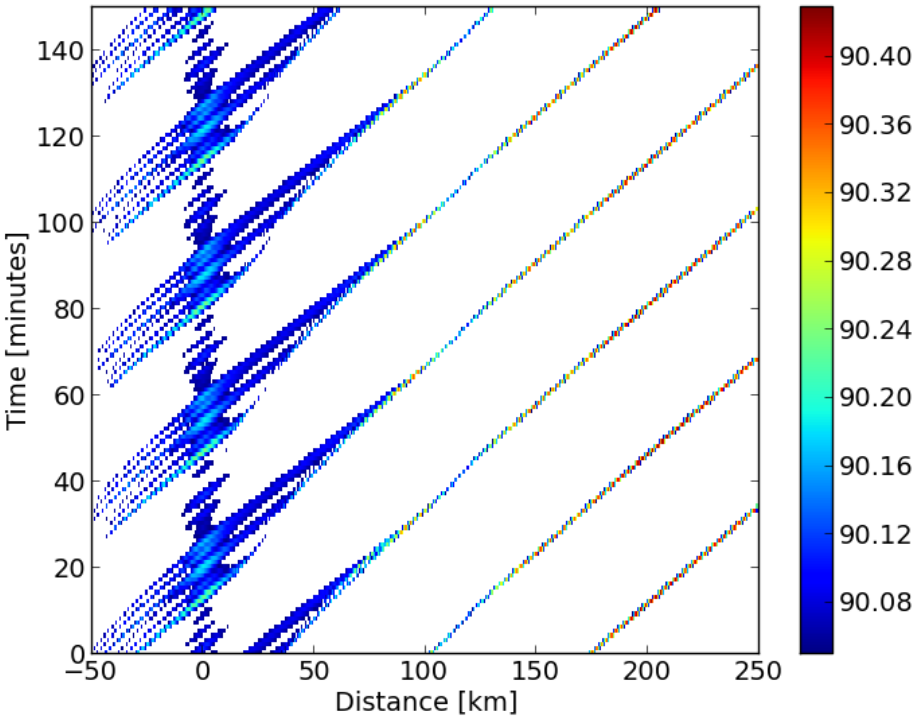

Figure 4.11: Same as Fig. 4.10, but with a mean wind of 40 ms^{-1}.

4.4 Discussion

A simplified dynamical model which is designed to represent key features of cumulus convection has been presented in this Chapter. The model is based on the shallow water equations and allows propagation of gravity waves. Conditional instability is represented by modifying the geopotential to produce convergence when the fluid level exceeds a height threshold. A rain variable is introduced that provides a downward force that can lead to the destruction of a cloud.

Two mechanisms for initiation of convection have been considered. The first is a stochastic noise term representing subcloud layer disturbances, and the second is flow over an orographic obstacle. With a continuous stochastic forcing, the model produces an ensemble of clouds that is comparable to radiative-convective equilibrium simulations with complex atmospheric models. In particular, the clouds undergo a realistic life cycle with a rapid intensification, formation of precipitation, followed by collapse.

The cloud size distribution roughly corresponds to the exponential form predicted by Craig and Cohen (2006), and the cloud spacing shows the effects of local triggering and more distant suppression of convection by gravity waves. Even the power spectrum

of the wind field reproduces the observed (5/3) power law of the atmosphere up to a wavelength of 20 km.

With orographic triggering, two regimes of convection are found, depending on the background wind speed. With wind speeds up to the intrinsic gravity wave speed, convection is trapped over the mountain, while for larger wind speeds, clouds develop over the orography but periodically move downstream.

While the simple model reproduces many behaviours of real convection, some aspects are excluded. First the model is assumed to be one dimensional. A two dimensional version could easily be constructed, although the choices of parameter values would need to be re-examined. Secondly the interaction of convection with synoptic or larger scales in the atmosphere is not represented. The use of constant values for H_c and ϕ_c implies that the conditional instability of the atmosphere is uniform in space and time. In principle these variables could be coupled to the dynamical equations, but many details would need to be addressed.

The absence of a representation of convective feedbacks on the subcloud layer, most importantly the creation of and the interaction with cold pools, implies that the simple model will not produce important modes of mesoscale convective organization, such as squall lines. Adding such a representation is not trivial and would probably require the introduction of one or more additional prognostic variables.

The aim of this Chapter was to provide the next contribution to a hierarchy of convection models for data assimilation research, that fills the gap between the very simple stochastic model presented in Chapter 3, and the full solution of the fluid equations in kilometre-scale weather prediction models. In contrast to the stochastic cloud model which ignored interactions between convective clouds, the model presented in this Chapter includes gravity waves, allowing initiation and inhibition effects, leading to a realistic spatial distribution of clouds.

The availability of wind and rain variables that correspond to observable atmospheric quantities should make it possible to define background and observation error covariance matrices that contain properties which are similar to those arising in full NWP systems.

The next Chapter presents results from data assimilation experiments achieved with this modified shallow water model.

Chapter 5

Data assimilation in the modified shallow water model

This Chapter presents the results from data assimilation experiments with an LETKF in the modified shallow water model. One focus of this Chapter is to show whether the modified shallow water model is useful as a test model for convective-scale data assimilation. Therefore the data assimilation experiments mimic the behaviour of a real atmospheric model and are compared to similar experiments that were done with idealized Consortium for Small-scale Modeling (COSMO) simulations by Lange and Craig (2013). To be able to relate the results of the two different models to each other, the modified shallow water model is changed a little bit compared to the setting presented in the last Chapter. These changes are described in Section 5.1 where also the observation types and the setting of Lange and Craig (2013) are specified. The results of example assimilation runs and averaged scores from 100 random runs are shown in Section 5.2. The comparison of the results of the two models is presented in Section 5.3. Data assimilation in the modified shallow water model has also been compared to the stochastic model in a Bachelor Thesis (Kigle, 2013). Section 5.4 concludes this Chapter with a short summary and discussion.

5.1 Methods

5.1.1 Model

The model used is the modified shallow water model presented in the last Chapter, with one dimension. To make its behaviour more comparable to the idealized COSMO experiments, some parameters have been changed. This ensures that the clouds have

Table 5.1: Model settings for the data assimilation experiments.

Parameter	**Model only**	**Assimilation**
α	$2.5 \cdot 10^{-4}\ s^{-1}$	$5 \cdot 10^{-4}\ s^{-1}$
β	3	10
H_0	90.0 m	90.0 m
H_c	90.2 m	90.0 m
H_r	90.4 m	90.2 m
ϕ_c	899.77 $m^2 s^{-2}$	899.8 $m^2 s^{-2}$
K	25000 $m^2 s^{-1}$	20000 $m^2 s^{-1}$
K_r	200 $m^2 s^{-1}$	200 $m^2 s^{-1}$

a longer life time and are a bit bigger, closer to the large and long-lived thunderstorms in the comparison. The model is therefore used with the parameter settings shown in Table 5.1, leading to an average size of the rain regions of roughly 4 grid points (2 km) and a life time of almost an hour. The updraft threshold H_c can be at 90.0 since when clouds are built, the water level in cloudless regions will go down. Therefore there will still be a positive perturbation in the water level needed to overcome the threshold. The clouds in the height field are a little bit larger and also live for almost an hour. The distribution is quite broad and also a lot of clouds live for a longer time and the ones which produce rain are larger than the short-lived small perturbations, as was shown in Fig. 4.7 in the last Chapter.

The model runs with 1000 grid points, a resolution of 500 m, a time step of 5 seconds and periodic boundaries. Additionally, the background perturbations which are added onto the wind field are only added at one grid point per time step and completely turned off during the free forecast phase. This is done to better see the impact of the interactions between the clouds during this period.

5.1.2 Setup of Lange and Craig (2013)

To confirm that the model is a good test model for convective-scale data assimilation, it will be compared to the results of Lange and Craig (2013). Therefore the data assimilation setting used in the modified shallow water model is chosen to be close to the setting Lange and Craig (2013) used in their work. They ran experiments with the LETKF in an idealized setting of the non-hydrostatic COSMO model. This system is called KENDA and currently being developed at the DWD. The model has a resolution of 2 km and is initialized with a horizontally homogeneous sounding and

random white noise in the temperature and vertical velocity fields. The sounding has a large Convective Available Potential Energy (CAPE) value of 2200 J kg^{-1}. The model spin-up takes from 6 to 14 UTC until large convective systems have appeared. These systems have a life time of a couple of hours. Data assimilation starts at 14 UTC and runs for three hours with 5 (fine run) and 20 (coarse run) minute cycles, then a three hour free forecast period starts. The full COSMO model physics is used with an active radiation scheme. The observations used for data assimilation are simulated Doppler radar, radial wind and no-rain observations which were thresholded. They also used a fine and coarse analysis scheme which are set up similar to the one presented in this work. The fine experiment has a cut-off radius of 7.3 grid points (or 14.6 km) and is called R4 while the coarse experiment has a cut-off radius of 29 grid points and is called R16. The final results have been obtained from averages of five random initializations. Further details concerning the model and data assimilation setting can be found in the original publication.

5.1.3 Data Assimilation

The LETKF of Hunt et al. (2007) is used for data assimilation. The observations are generated from a truth run which uses the same settings as the ensemble members. Except for the randomness of the background noise, this is a perfect model experiment. Used are rain observations above a given threshold (thr_{rain}) to which a Gaussian random error is added to represent the error of the instruments and representativity (σ_{rain}). Below this threshold the rain value is set to 0 and the same random noise ($\sigma_{no-rain}$) is added. These are the so-called no-rain observations which are used to suppress spurious convection. The last observed field is the radial wind, i.e. wind is observed at all grid points where there is rain present. An observation error σ_u is assigned to the wind field. The observation errors are chosen large, to still represent a realistic amount of uncertainty in the observations. The wind error is in the range of 10-20% of the natural wind variance, while for rain error this value is roughly 25%.

Two different observation resolutions are used to create a fine and a coarse experiment. In the former, the cut-off radius in the localization (r_{Loc}) is set to 10 grid points (5 km) and the observations are taken at every grid point. The localization is done with a Gaspari-Cohn correlation function (Gaspari and Cohn, 1999) in the observation error covariance matrix $\mathbf{R}$. This fine experiment is called R10 as observations inside a radius of 10 grid points have non-zero weight. The second one is done where the observations are averaged over a region of 5 grid points. This average is then assigned as an

Table 5.2: Settings of the LETKF.

	R10	**R50**
r_{Loc}	5 km	25 km
$\Delta x_{obs/ana}$	0.5 km	2.5 km
Δt_{ass}	1 min	1 min
$\mathbf{R}_u$	$(0.01\ \mathrm{ms}^{-1})^2$	$(0.01\ \mathrm{ms}^{-1})^2$
$\mathbf{R}_{rain}$	$(1 \cdot 10^{-5})^2$	$(1 \cdot 10^{-4})^2$
$\mathbf{R}_{no-rain}$	$(1 \cdot 10^{-5})^2$	$(1 \cdot 10^{-4})^2$
thr_{rain}	0.005	0.0025
σ_{rain}	0.005	0.005
$\sigma_{no-rain}$	0.005	0.005
σ_u	$0.01\ \mathrm{ms}^{-1}$	$0.01\ \mathrm{ms}^{-1}$
ρ	1.05	1.03

artificial observation to the centre of these 5 points and used for the calculation of the weight matrix in the LETKF algorithm. The weight matrices of all local regions are then interpolated onto the real grid and used to calculate the analysis. This method was first used by Yang et al. (2009). The localization cut-off radius still contains the same amount of (virtual) analysis grid points and therefore this experiment is called R50. Such observations are also called superobservations and have been used e.g. by Alpert and Kumar (2007) and Zhang et al. (2009).

Instead of putting the true observation errors (σ_u, σ_{rain} and $\sigma_{no-rain}$) into the observation error correlation matrix $\mathbf{R}$, smaller values are used to force the analysis state more rigorously towards the observed values. $\mathbf{R}$ is a diagonal matrix, with the values corresponding to the different observation types called $\mathbf{R}_u$, $\mathbf{R}_{rain}$ and $\mathbf{R}_{no-rain}$. This strong weighting of the observations is needed to ensure that the method captures the state of the system. Sensitivity experiments have shown that less forcing would lead to results where the state of the system was not well caught in the ensemble.

The settings of the different data assimilation parameters for the R10 and R50 experiments can be found in table 5.2. The rain threshold is chosen lower in the R50 experiment because the averaged observations have smaller values. The multiplicative covariance inflation factor ρ is also slightly smaller for the R50 experiment, because it contains a smaller effective number of observations and the analysis therefore has a smaller chance of being overconfident.

The two different assimilation settings are chosen to see whether a high-resolution analysis is really needed for a good weather forecast. It is possible that forcing the

model towards one specific state is not beneficial for producing reliable forecasts, although the analyzed state might be better. This topic has been looked at in Lange and Craig (2013) and therefore the setting in the work presented here is chosen in accordance to the setting in the idealized COSMO experiments.

The only difference between the ensemble members and the nature run are the random numbers used for the positioning of the background noise. Both ensemble and truth have a spin-up of 1000 time steps before the assimilation starts to reach a steady state in the sense of ongoing generation and decay of clouds. This also ensures that the ensemble is totally decorrelated and has no prior knowledge of the position of clouds. The assimilation is done with an ensemble size of $k = 50$ members and runs for 36 minutes with observation intervals of one minute. Then the assimilation and the background noise are turned off and a free forecast is run for another 36 minutes. During this phase the absence of noise increases the life-time of clouds and also the gravity waves emitted from the clouds cause more interaction because they are not superimposed by the background noise.

5.2 Results

In this Section the data assimilation results are presented. First, the assimilated states are analyzed by using example snapshots. Then, averaged results from 100 different random assimilation experiments are used to have a look at some error measures. Both, the data assimilation and free forecast phase are analyzed like that. Finally, the results are compared to the idealized COSMO-KENDA experiments of Lange and Craig (2013).

5.2.1 General Performance

Fig. 5.1 shows the assimilated rain states after 36 cycles. The top picture shows the truth run in red and the averaged observations, obtained by averaging over 5 grid points in blue. The dashed line shows the rain/no-rain threshold for the R10 case. One can see that the observations for the R50 case still contain all the rain regions, but with a smaller peak value. The clouds appear with a larger size and a broader base, but this base is not present in the final observation as it is removed into the no-rain observation if below the threshold. The ensemble mean in the middle plot shows that most of the regions where there is rain present are captured in the ensemble mean of both runs, with only a few regions missing or only little rain present mainly in R50. A

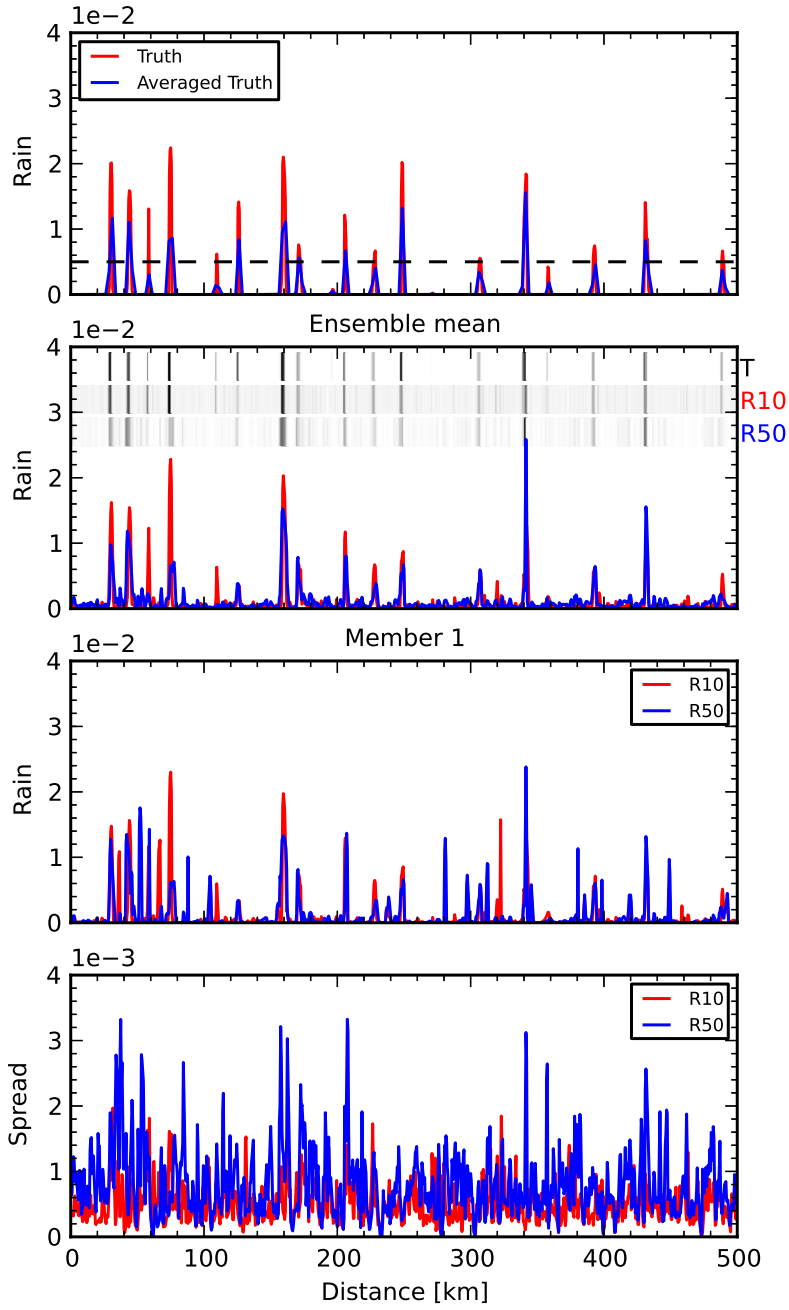

Figure 5.1: Panels from top to bottom, 1: Rain field in the truth run (red) and the observation field averaged over 5 grid points (blue). The threshold for distinction between rain and no-rain in the R10 experiment is marked with a dashed black line. 2: Ensemble mean of the R10 (red) and R50 (blue) experiments. The horizontal bands show the rain amount in a gray scale going from 0 (white) to the maximum rain (black) for truth (T) and both assimilation runs. 3: Ensemble Member 1 for R10 (red) and R50 (blue) 4: Ensemble spread for both assimilation runs. All fields are displayed in the state after 36 assimilation cycles.

couple of regions contain some spurious rain, but its amount is not significant due to the no-rain observations. The middle box in Fig. 5.1 shows a gray scale plot of the rain amount to better visualize the correspondence of the rain regions. The concurrence of dark gray or black vertical stripes is large and in the ensemble mean there is small amplitude noise present as can be seen in the light gray shading and the bottom plot. The rain maxima are in general closer to the true values in the R10 experiment and a little bit broader in R50. This is reasonable as with the superobservations there is no forcing to a precise position. It is also visible that the blurring effect in the ensemble mean is not very large which also corresponds to the averaged observations in the top panel of Fig. 5.1. In the R50 run there is a little more spread in the ensemble and large spread is mainly connected to regions with rain. In ensemble member 1 this can be partially seen, as in some regions the member looks exactly as the ensemble mean, while in other regions it looks completely different. In R10 the ensemble member 1 looks almost identical to the ensemble mean. Overall the R10 experiment captures the rain better, but the regions close to km 340 or 440 show that R10 is not necessarily always superior.

The assimilated height field of the same situation (Fig. 5.2) looks a lot different. Both experiments contain a significant smaller number of clouds in all members and the spread is much larger. The spread in the height field is in the range of 25% of the variation in the field compared to roughly 10% spread in the rain field. This is also clearly visible when looking at member 1, which differs very much from the ensemble mean. A reason for this is that the height field is not observed directly. Therefore any information comes from the correlation to the other fields. As rain only starts to get produced when the height field reaches 90.2 m, all clouds which have not reached that height are not observed, producing a delay in the assimilation. There is also the possibility that a cloud has already disappeared in the truth and is still present in the ensemble. Another reason for the bad match in the ensemble mean is the positional spread in the single members which averages out the clouds in the ensemble mean. Although the field in the single member looks more like the truth, there are fewer clouds and smaller amplitudes present. Therefore the averaging effect cannot be the only error. The R50 experiment seems to have captured more clouds compared to R10 but does not reach as high peaks. The reason for this is that in the R10 the height field is not conserved which is apparent in the two middle plots of Fig. 5.2 where the base of the red line is a little bit lower than the one of the blue line. This loss of mass will be considered further in the next Section.

As last variable, the wind field in Fig. 5.3 is examined. The ensemble mean is

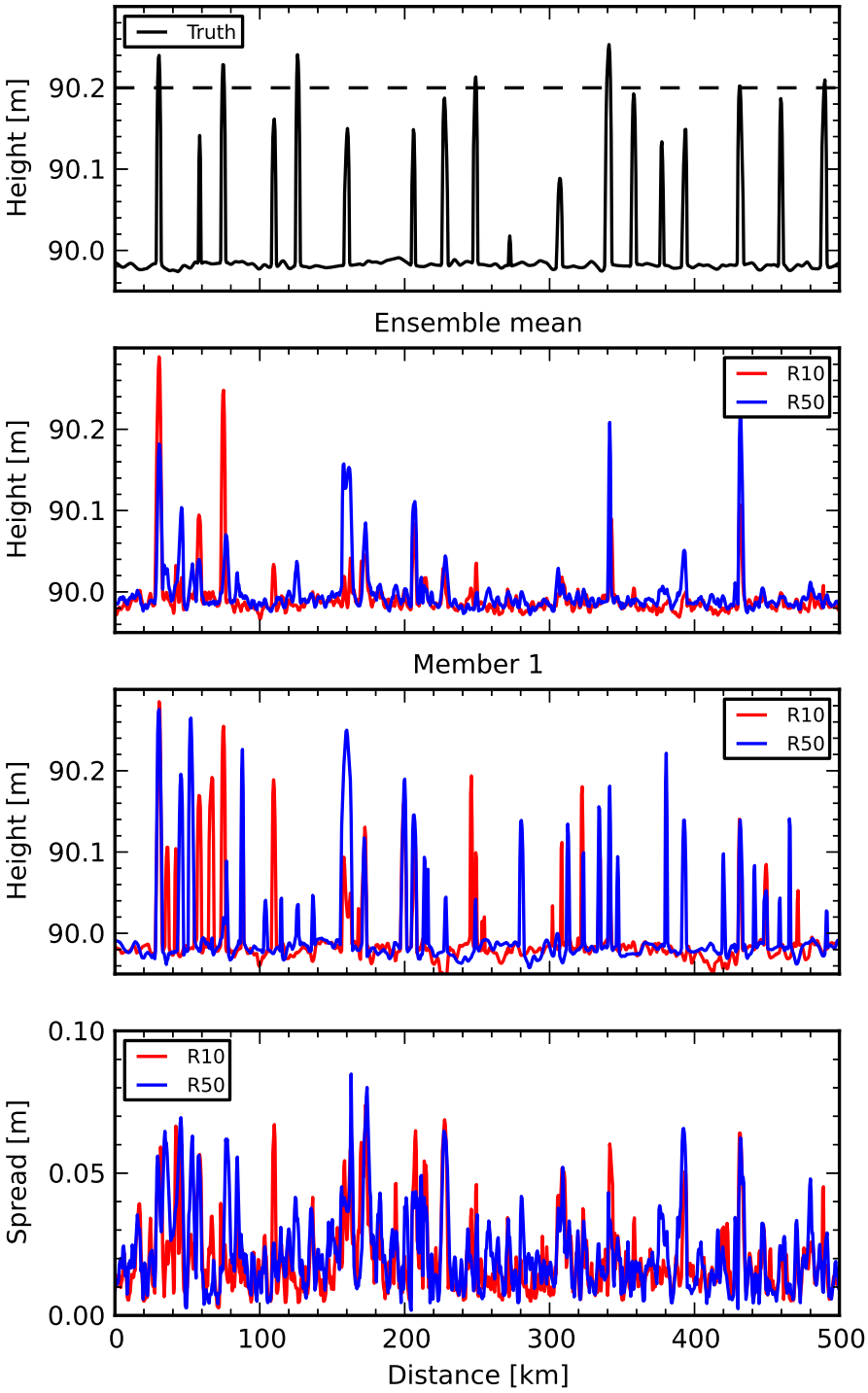

Figure 5.2: From top to bottom, 1: Height field in the truth run. Dashed line marks the rain threshold. 2: Ensemble mean of the R10 (red) and R50 (blue) experiments. 3: Ensemble Member 1 for R10 (red) and R50 (blue) 4: Ensemble spread for R10 (red) and R50 (blue). All fields are displayed in the state after 36 assimilation cycles.

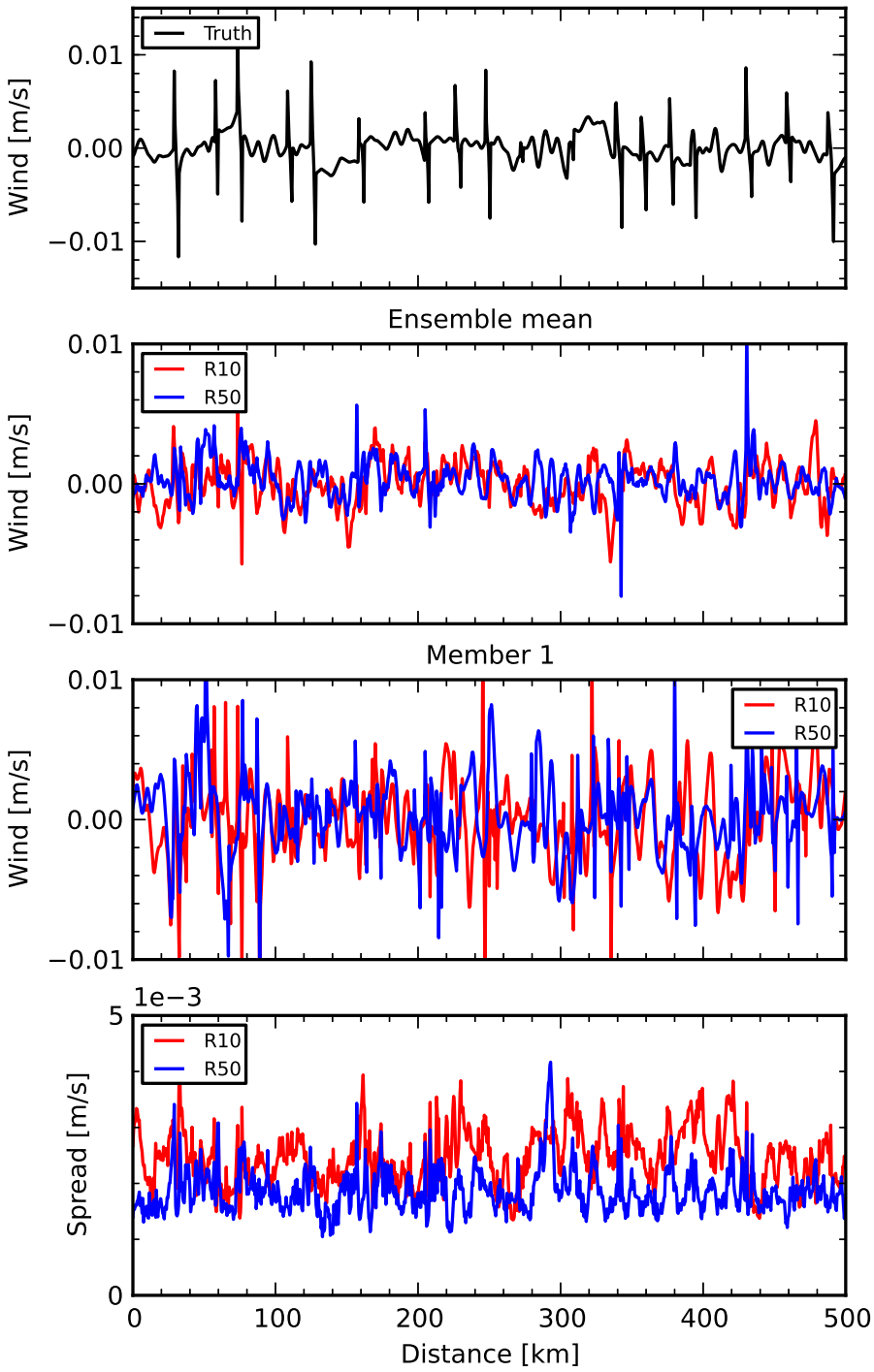

Figure 5.3: From top to bottom, 1: Wind field in the truth run. 2: Ensemble mean of the R10 (red) and R50 (blue) experiments. 3: Ensemble Member 1 for R10 (red) and R50 (blue) 4: Ensemble spread for R10 (red) and R50 (blue). All fields are displayed in the state after 36 assimilation cycles.

clearly far from a physical model state if compared to the nature run. It does not show the specific convergent features in the center of clouds that are present in the nature run. This result is consistent with results of Lange and Craig (2013). Similarly to the height field, R50 seems to be more physical as it shows more of the characteristic peaks but they do not appear very often. As the single ensemble members still show this feature (e.g. ensemble member 1), but every one at a slightly different position, the ensemble mean will average out the information and look unphysical. The R10 run has much more spread which also explains at least partially the bad ensemble mean. These different fields show nicely that one certain state in the rain field may be generated in the analysis by different states in the other variables. Therefore it is difficult for the data assimilation to capture the updrafts correctly. One part of this problem is the timing of convection, as a specific rain amount might belong to a cloud which has just appeared, or one which is already dying. Additionally, until the data assimilation has produced the correct rain amount, the cloud might already have disappeared in the truth.

5.2.2 Analysis of averaged experiments

The error scores that have been calculated for this Section are averages over 100 random realizations of noise in the nature and ensemble runs. This ensures the statistical robustness of the results. All grid points are used for the calculations of the error.

Root mean squared error

Fig. 5.4 shows the Root Mean Squared (RMS) errors and ensemble spread for all three model variables. A big impact of the assimilation during the first few cycles can be seen in all the fields, although it is a negative impact in the wind field. This positive influence is mainly due to the clouds which are captured in the ensemble within a few cycles. The removal of wrong clouds only works to a certain degree as the background noise continuously introduces new perturbations which can also lead to the generation of new clouds. Therefore there is no more improvement in the ensemble mean after roughly 10 cycles in the rain field. As this is the field where most of the information comes from, it is also the main field to do the evaluation. The rain RMS at dry grid points is almost the same in R10 and R50 (not shown) and significantly smaller but with a similar temporal behaviour as the error of the free forecast at these points. Therefore the main improvement of R10 compared to R50 is a better representation of the rain at rainy grid points.

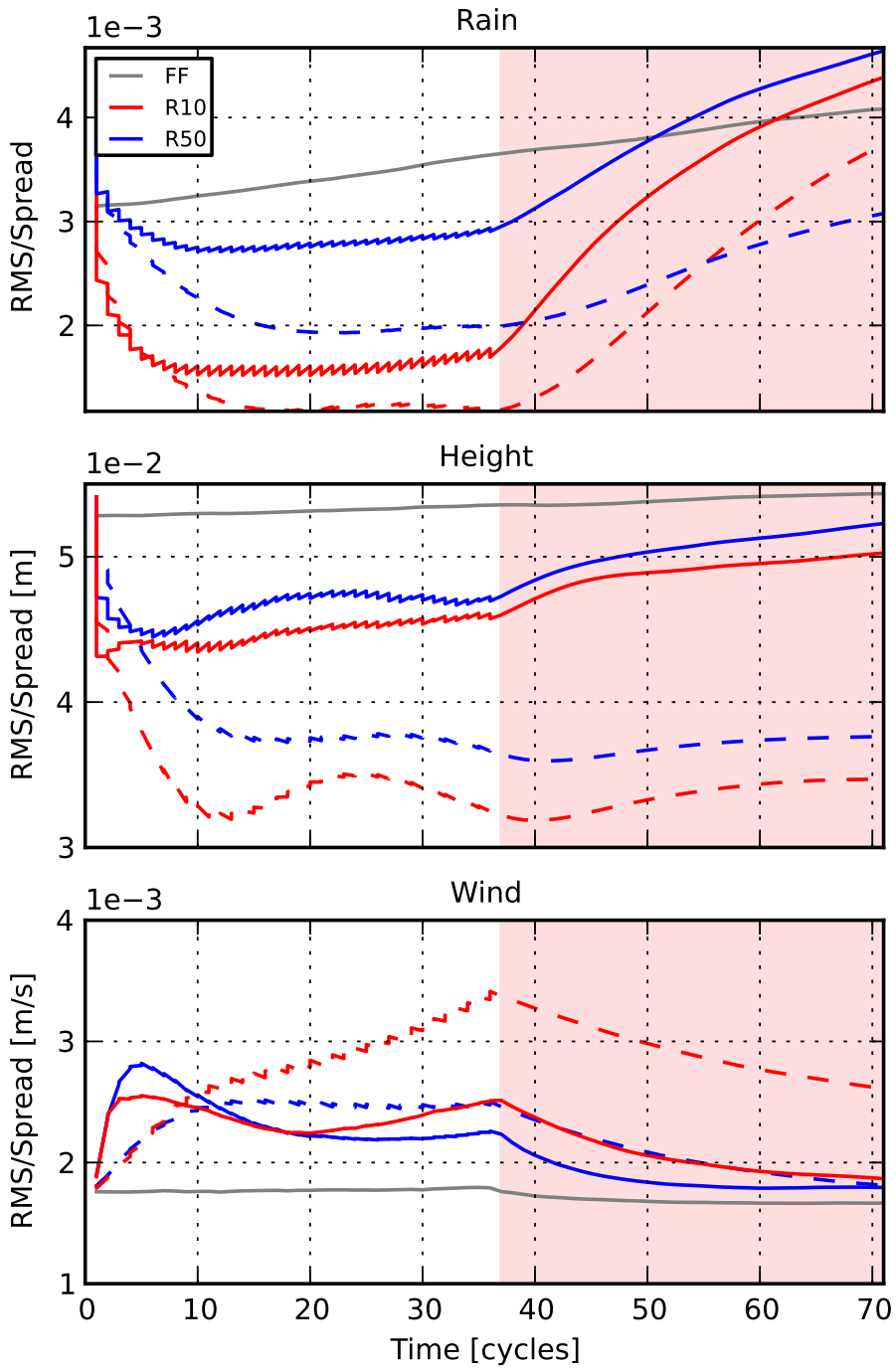

Figure 5.4: Top: RMS (continuous) and spread (dashed) lines for the R10 (red) and R50 (blue) of the rain field. The gray line shows the RMS of the free forecast. The red transparent area marks the beginning of the free forecast in the ensembles. Middle: The same plot for the height field. Bottom: The same plot for the wind field. As stated in Table 5.2 1 cycle is equivalent to 1 minute.

The sawtooth pattern of the curves comes from the fact that the rain error increases a little bit during every forecast step and is corrected by the assimilation. This pattern is not present in the RMS of the dry grid points (not shown). The ratio of spread to RMS is better in the R10 experiment, but tuning with multiplicative inflation has shown that a better ratio for R50 does not necessarily lead to better results in the other variables.

In the height field the main improvement also comes from a correct representation of cloudy regions. Although a lot of clouds are missed, the ones that are captured still produce an improvement compared to the free reference ensemble. The difference between the fine and the coarse resolution is almost zero and the R50 even has a better spread to skill ratio. In the wind field both experiments end up worse than the free reference ensemble. One reason for this is the strong forcing onto a correct rain field. As could be seen in Fig. 5.3, this does not produce correct wind fields in the ensemble mean. But as the zero mean wind is still present in the ensemble mean, and the error is on a small level, this is not a significant problem affecting the other fields in having a good analysis.

During the free forecast phase the general behaviour one can expect is the RMS lines approaching the gray line of the free forecast. This happens quite slowly in the height and wind field, but rather fast in the rain field where the RMS even gets larger than the free forecast. This is an indication that although the rain was assimilated quite well, the assimilation was moving the model into a state where not the natural behaviour is present anymore. One can also see that the RMS in the R10 experiment increases much faster than in the R50 case, especially its spread overtakes the R10 after 20 minutes.

Object based scores

RMS based measures are only partially useful for fields like precipitation as they encounter for example the double penalty problem. Therefore the object-based SAL score of Wernli et al. (2008) is used to further analyze the rain field. The location part of the SAL is shown in Fig. 5.5 where 0 means that all the objects are located at the correct position. Because in the R50 experiment the rain amounts in the ensemble mean are a little bit smaller due to the averaged observations, the plot shows results from two different object identification thresholds of 0.0025 (continuous) for the mean and 0.005 (dashed) for the single members. These two thresholds are used to investigate the difference between the ensemble mean and the single members.

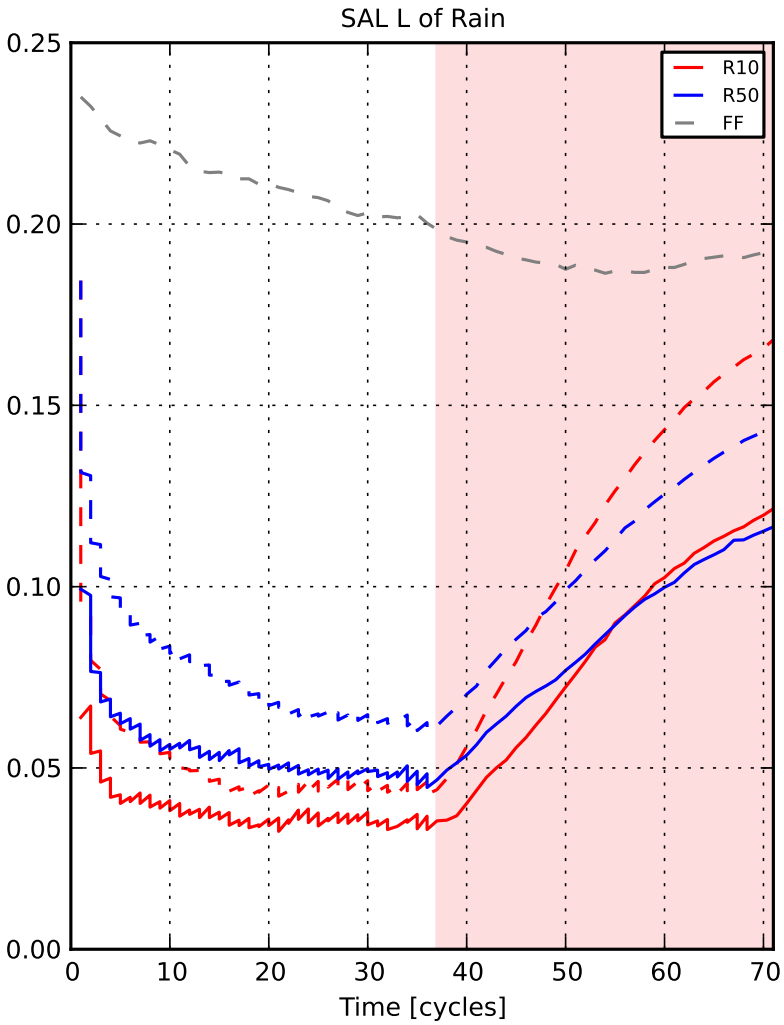

Figure 5.5: SAL location score of the ensemble mean with a threshold of 0.0025 for the rain object identification (continuous) and mean score of the single members with a threshold of 0.005 (dashed) for the R10 (red) and R50 (blue) experiments. The gray line shows the mean score of the single members for the free forecast with a threshold of 0.0025.

Similar to the RMS plot there is a big impact during the first few assimilation steps when the rain regions are captured. In this score, the difference to the free forecast seems much larger than for the RMS. The ensemble mean SAL is better than the mean SAL of the single members, which is due to the fact that the single members are distributed around the correct location. This puts the ensemble mean in almost the right position, while the individual members vary somewhat around it. The difference between the ensemble mean and the single members is a little bit larger in the R50 experiment compared to R10. This shows that the R50 members fluctuate stronger

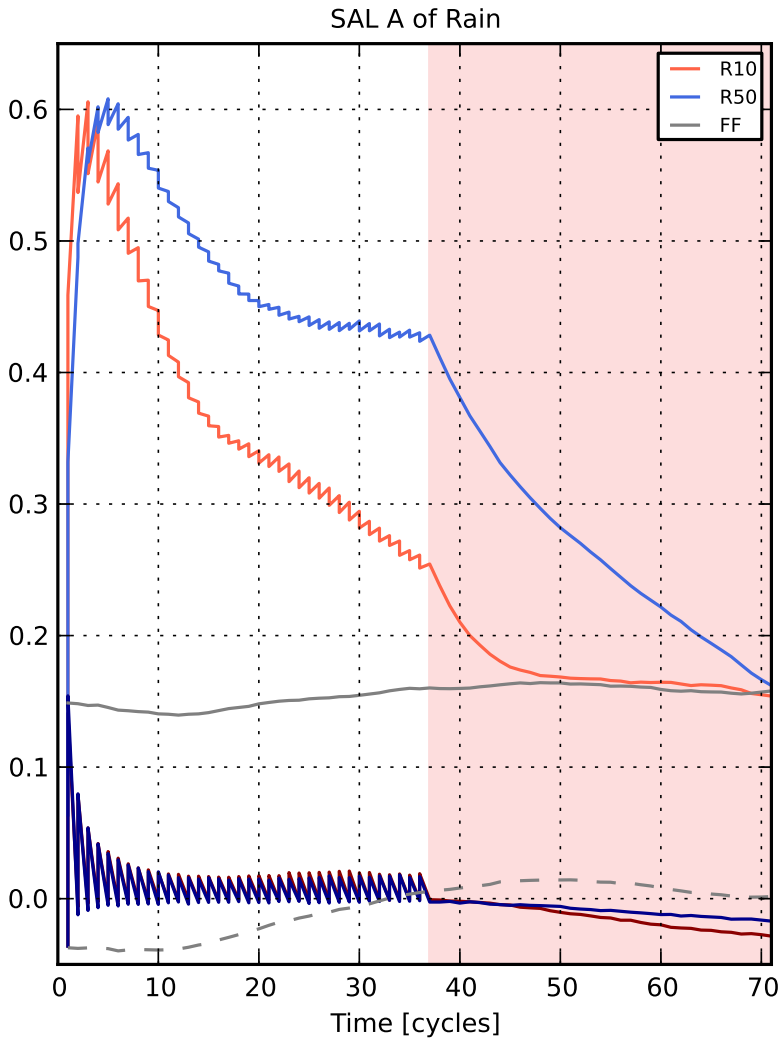

Figure 5.6: SAL amplitude score of the ensemble mean in the R10 (red) and R50 (blue) experiment. The gray line shows the mean in the free forecast run (continuous) and the mean of the single members (dashed). The dark red and dark blue are the difference lines of the score of the ensemble mean minus the mean score of the single members.

around the correct position, leading to this stronger increase in the error.

The R10 experiment is only marginally better after the 36 assimilation cycles and loses the advantage of the better positioning within only a few free forecast minutes. The error grows much faster in R10 and is worse than the R50 after roughly 10 to 20 minutes. This larger error growth is present in both scores throughout the whole forecast time.

The SAL amplitude score in Fig. 5.6 is a measure for the amount of rain present in the whole domain. With a perfect score being 0 and a positive value meaning there

is too much rain in the ensemble. The first assimilation steps introduce a lot of rain leading to a strong overestimation. The plot additionally shows the difference of the score of the ensemble mean to the mean score of the single members. One can see that the assimilation always makes this difference larger, while the evolution of the model brings it towards zero again. This is the case because the assimilation puts a lot of rain into the ensemble members. If this rain is not supported by the height field, it will get removed during the model run. But as during the model run, background noise is introduced, the ensemble mean still has a growing amount of rain. This can be seen in the sawtooth pattern where the rain amount increases during the short forecast phases of the assimilation because the background noise puts in some new rain regions. This rain gets then removed at every assimilation step which works better in the R10 case. But when interpreting the pattern in the amplitude it is important to keep in mind that the amplitude score is normalized by one half of the sum of the total rain in the ensemble plus the truth. If this sum changes during the free forecast phase, so will the score.

During this free forecast phase, both experiments start to relax towards the free forecast value. In contrast to the assimilation period, there is no more background noise and therefore no significant introduction of rain. The experiments start to lose the clouds which were assimilated correctly. A correct value in the amplitude only means, that the overall amount of rain is correct. This does not necessarily mean that these regions are positioned correctly, as can be seen in the location score in Fig. 5.5. The error of the free forecast ensemble is not zero as it is the score of the ensemble mean of 50 simultaneous simulations. The mean score of the single members in the free forecast ensemble is located around zero.

The third SAL component is the structure of the rain regions. In this score zero means that the structure of the regions are identical, while a positive score means that they are too large. This is the case for both experiments as presented in Fig. 5.7, but the R10 still has a good result, with the single members being close to the truth. The difference from the single members to the ensemble mean is much larger for R50, clearly showing that the ensemble members spread more around the center of the rain, producing larger rain regions. This is an additional proof for this behaviour which has already been indicated in the location score. One can also see from the sawtooth shape that the assimilation tries to produce sharper objects, while the model blurs them out again. This goes on during the free forecast when the rain regions smear out for 10 to 15 minutes until they either die out or are not captured as a single cell anymore. The single members in the R10 experiment get worse in the first minutes of the free

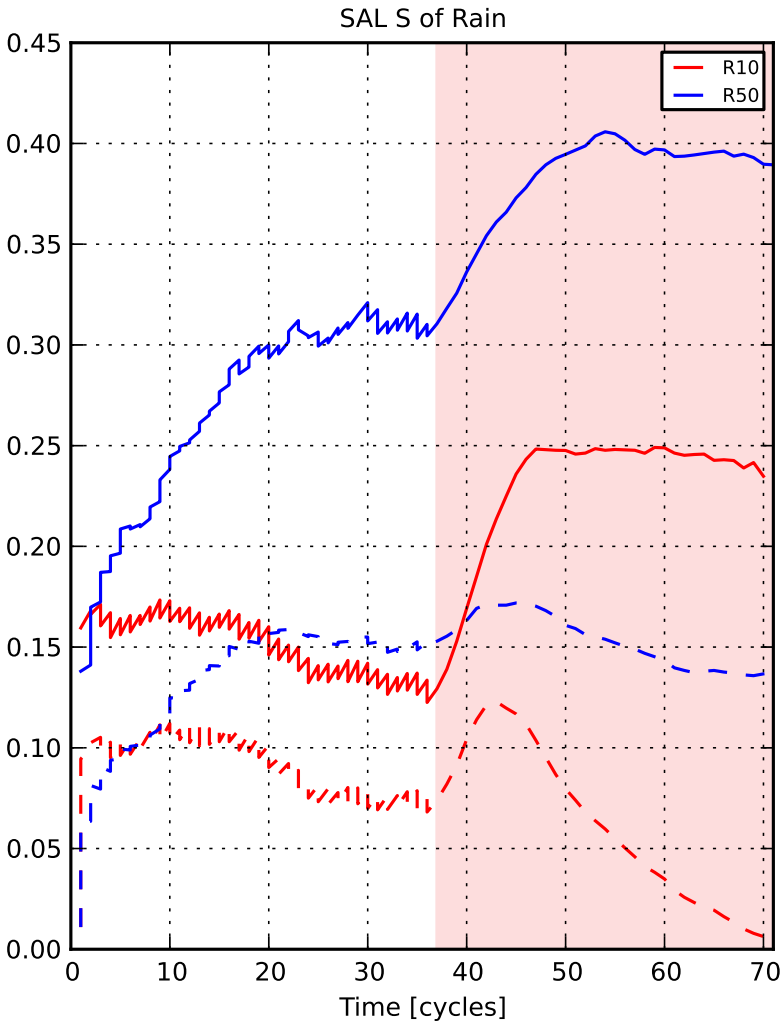

Figure 5.7: Structure score of the R10 (red) and R50 (blue) experiment. The continuous line is the score of the ensemble mean, the dashed line the mean score of the single members.

forecast, showing that they were not in a balanced state at the end of the assimilation. But after a few minutes they start approaching 0 and reach it even faster than the R50 experiment showing how fast they reach the random shapes of the nature run.

Further analysis

To explain some of the problems of the R10 experiment and the non-physical states, one can have a look at the mean water level in Fig. 5.8. During the first assimilation cycle, a lot of mass is introduced when the correct clouds are captured by the assimilation, but still all the wrong ones are present. During the rest of the assimilation the wrong

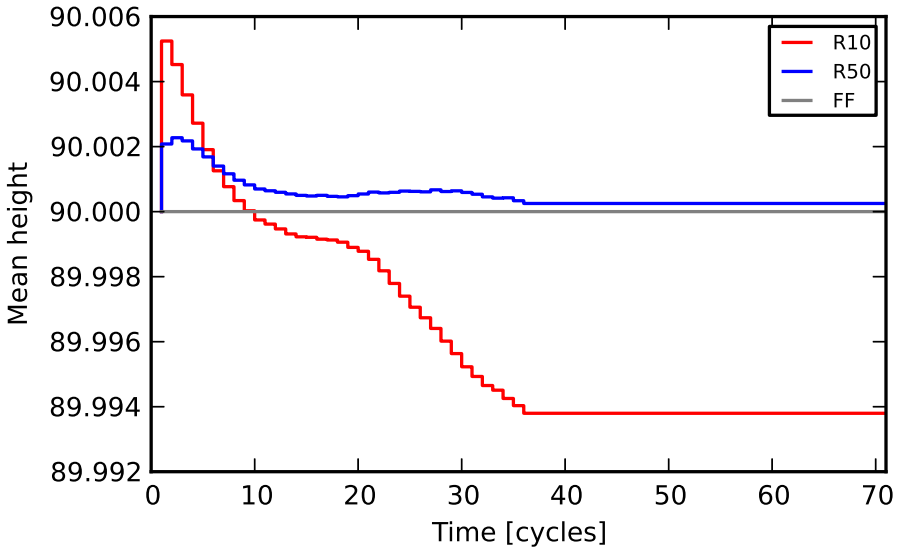

Figure 5.8: Mean height of the water level in the R10 (red) and R50 (blue) experiment as well as the free forecast.

ones get removed, which is a never ending process as the background noise always introduces new perturbations which can grow into full clouds. The second part is keeping the correct clouds and catching a now and then appearing new cloud from the nature run.

The combination of these two aspects is constantly removing mass from the model during the later cycles which is a larger problem in the R10 experiment. With the loss of mass, it will be harder for the ensemble members to exceed over the cloud and rain thresholds. This leads to a worse analysis the longer the assimilation operates. The averaging of the observations together with the larger analysis regions ensure that there are smaller and more physical increments used in R50 and the mass is closer to being conserved. To ensure that the problem is not numerical, the free forecast ensemble is also shown in the plot and clearly conserving the height field.

A couple of experiments have been carried out to test the sensitivity of the height conservation. A changed R10 experiment where height was observed at every 10th grid point showed that this reduces the loss of mass to a significant amount. Using a larger localization radius also improves the mass conservation, which of course does not necessarily lead to a better result in the other scores. The same result could be observed when using a model correction that introduced a small part of the missing mass at every grid point at every time step. A standard ETKF run (not shown) absolutely conserved the mass, showing that it is the localization causing the problem.

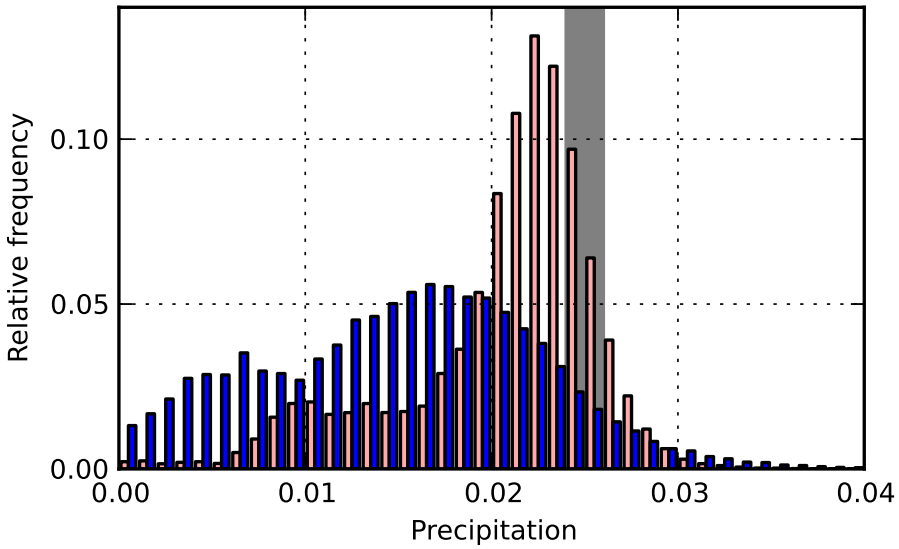

Figure 5.9: Relative frequency of rain amount in the ensemble members of R10 (red) and R50 (blue) at grid points where the truth run has a value in the range of 0.0025±0.001 (marked by gray column). The analysis is done after assimilation cycle 36.

If the analysis is obtained globally, the linear combination of other ensemble members can only draw from members with the correct mass and will not lead to conservation issues.

In Fig. 5.9 one can see the distribution of rain amount in the ensemble members at grid points where the truth run has rain of 0.0025±0.001. This is a large rain amount which is happening rarely, but the distribution looks similar for smaller values. The distribution in R10 is much sharper and closer to the real rain amount while R50 has a significantly smaller peak at 0.017. Both experiments have a second peak below 0.01 which is much more pronounced in the coarse experiment. This peak might be related to decaying rain at this grid point or rain that is present but not in an environment which is able to reproduce these large amounts of rain. Another possibility is that it comes from the background noise. A check on the data assimilation increments that are produced at these grid points with a large amount of rain present showed that they still are quite small.

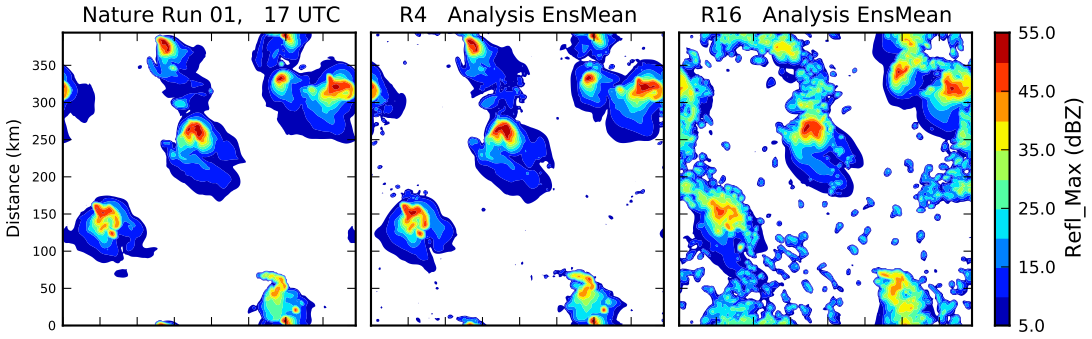

Figure 5.10: Top panel of Fig. 4 in Lange and Craig (2013) showing maximum reflectivity of the nature run and analysis means of R4 and R16 at 17 UTC.

5.3 Comparison to an idealized convection permitting model

The setting of the data assimilation experiments in the modified shallow water model was chosen similar to the setting in the COSMO model of Lange and Craig (2013). This allows to do a comparison of the fine and coarse assimilation runs in the two models. The main differences in model characteristics are that in the shallow water model the clouds are a little bit shorter lived and smaller than the large convective systems of the COSMO model. To compensate for this effect the modified shallow water model has a higher resolution and assimilates more frequently. Therefore the comparison in the next paragraphs is focusing on the number of cycles, not on the real time as in both models 36 cycles (9 for R16) are run and this time corresponds roughly to the life time of one cloud. As rain in the shallow water model is most comparable to reflectivity in the COSMO model the focus lies on these two variables. For better readability, some of the figures of Lange and Craig (2013) are reproduced in this work.

First, the rain ensemble mean in Fig. 5.1 of this work is compared to the analyzed maximum reflectivity after all assimilation cycles in Fig. 5.10. In both fine experiments, the observed field is assimilated with a very good agreement. Both coarse experiments also show most of the observed features, but are not as sharp in the positioning and have more spread than the fine experiment. This is visible in the areas with moderate reflectivity values around 15 to 30 dBZ. Figs. 7 and 8 in Lange and Craig (2013) show example members of the same analyzed state. In the fine case all the members look alike while they show more variability in the coarse case. But the single members in both experiments look very similar as well, proving that the drizzle in the ensemble mean is present due to the averaging of the different locations of thunderstorms. These

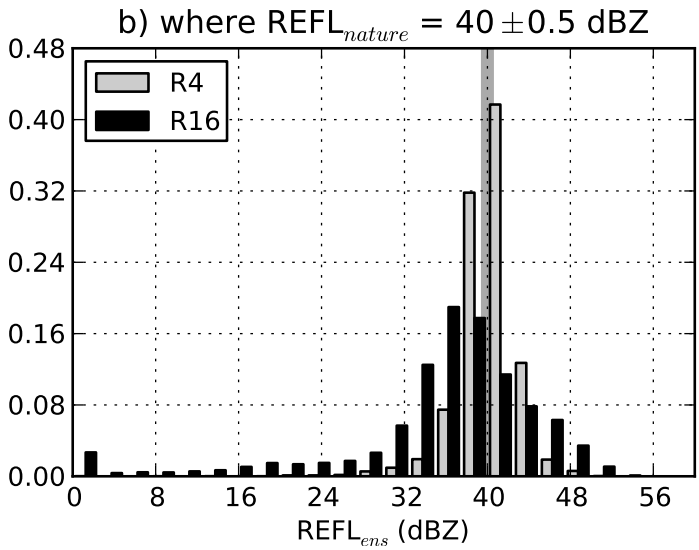

Figure 5.11: Fig. 9b of Lange and Craig (2013) showing the relative frequencies of model values within the ensemble members of R4 (black bars) and R16 (gray bars) at locations of the domain where the reflectivity of the nature run has a value of 40 ± 0.5 dBZ. The analysis is done at 17 UTC.

differences between the fine and coarse experiment are also found in the results achieved in the modified shallow water model.

The equivalent to Fig. 5.9 from Lange and Craig (2013) is shown in Fig. 5.11 where instead of rain the relative frequency of reflectivity is analyzed and the result is only coming from one experiment. The similarity of the two plots is apparent. The fine resolution experiment closely matches the nature run, while the peak of the coarse resolution is at too small values with only half of the relative frequency. Additionally the coarse resolution has a significant amount of rain at values close to zero. Both fine experiments look more like a Gaussian while the coarse experiments seem multimodal.

The general behaviour of the fine resolution having a good analysis after 36 cycles but losing this advantage over the coarse resolution within only a short time into the free forecast can be seen for example in Figures 5, 13 and 14 of Lange and Craig (2013). Fig. 13 is reproduced in this thesis in Fig. 5.12 showing the distance component of the object-based DAS measure from Keil and Craig (2009). This measure for the distance error of objects is similar to the SAL location score used in Fig. 5.5 of this work. The three SAL components of the maximum reflectivity field are shown in Fig. 5.13. The location score contains more spread between the five different runs and one can identify the need of a larger number of experiments to get a more reliable result. The DAS-DIS has less spread and one can clearly see that within one hour into the free

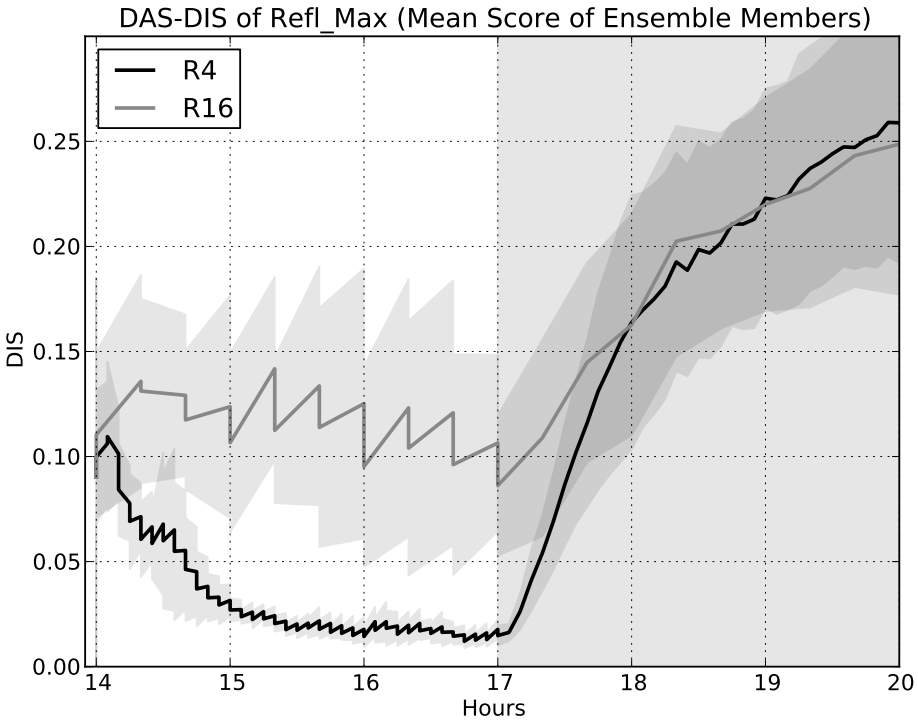

Figure 5.12: Fig. 13 of Lange and Craig (2013) showing the DIS-component of the DAS-score of the maximum column reflectivity > 10 dBZ. Displayed is the mean DIS-score of the ensemble members of R4 (black) and R16 (gray). The error lines are the average over five experimental repetitions. An envelope of the empirical standard deviation is given by the shaded areas.

forecast the fine resolution has lost its advantage over the coarse run. Finally with the random initialization of ensemble members in the correct atmospheric environment, in both models the clouds are captured already after only a few assimilation cycles. This leads to the introduction of rain mass into the models which after these cycles have a good location score, but too much total rain.

The conservation of quantities like mass and energy has been shown to be a problem in the idealized COSMO experiments as well as in the modified shallow water model. Although the consequences are not identical in the two models, the sensitivity to a lack of conservation is present in both.

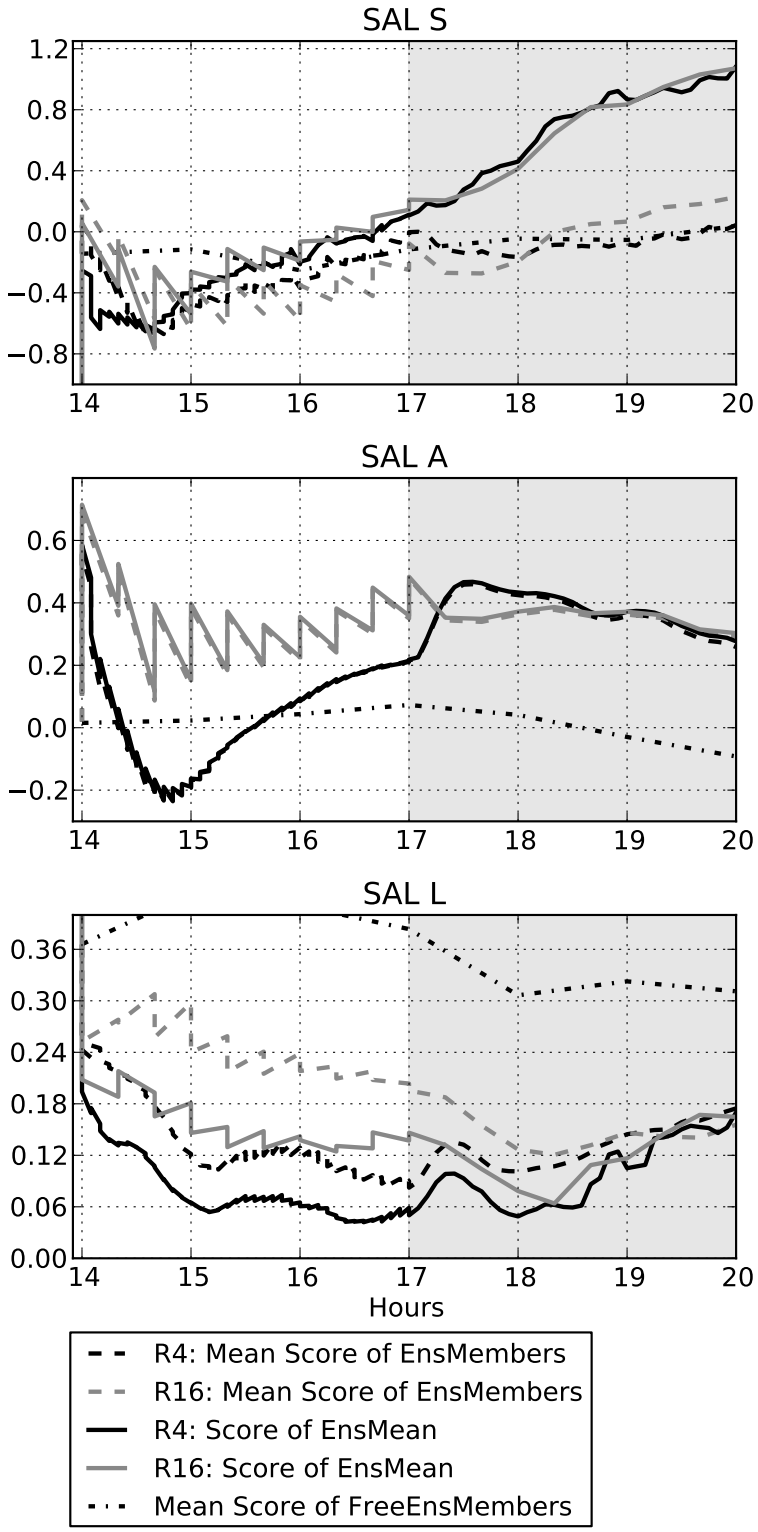

Figure 5.13: Fig. 14 of Lange and Craig (2013) showing the three SAL components of the maximum reflectivity > 10 dBZ averaged over five repetitions.

5.4 Discussion

Data assimilation experiments with an LETKF in a modified one-dimensional shallow water model have been performed. The model was set up to represent the behaviour of appearing and disappearing convective clouds which interact through gravity waves. The convection was forced with random background perturbations which add a random portion to the otherwise perfect model experiment. The observations types rain, no-rain and radial wind have been generated using a nature run and were used in a fine (R10) and a coarse (R50) resolution analysis scheme. 36 assimilation cycles with one minute observation intervals have been calculated followed by 36 minutes of free forecast.

The R10 experiment identified the rain regions very fast and got good results in the rain and height field, with problems in the wind field. Although the spread to skill relationship is not perfect, inflating the spread did not lead to a smaller RMS error. The location score of the SAL also showed that most of the rain regions are captured and placed very close to the correct place. The strong forcing towards the rain observation produced unphysical fields in the analysis, especially the water level was not conserved anymore. This effect is reduced either by using a larger localization radius, a weaker forcing or the R50 setting. A larger localization radius is not desired, as this would restore problems like spurious correlations, which led to the use of localization in the first place. The problem of mass conservation in data assimilation is not explored well, but it is known and some research has started to have a look at it lately. An example would be Cohn (2010) where the *principle of energetic consistency* is formulated and a good theoretical introduction is given.

Experiments in the idealized COSMO-KENDA system also indicated that there might be issues with mass conservation. The height field in the modified shallow water model is comparable to CAPE in the real atmosphere. In a complex atmospheric model surface fluxes can adjust the CAPE when convection is happening. There is no mechanism to have similar processes in the shallow water model, which possibly enhances the problem of mass conservation and needs to be further addressed, especially in more complex models.

The coarse resolution experiment R50 ends up worse than R10 in most of the fields and scores, but still captures most of the rain regions although with less intensity and precision. As the positions of the rain regions are blurred out a little bit, the double penalty problem appears in the RMS of the rain field. In the location score of the SAL the coarse resolution is almost as good as the fine one.

During the free forecast the advantage of the fine resolution analysis compared to the coarse one is gone within only 10 to 20 minutes. In all scores the error of the fine resolution increases faster. This indicates that such a good analysis might have problems in representing the uncertainty associated with the evolution of a convective field. Also a numerical model might not be able to represent the atmosphere with such precision and forcing it to do so may cause it to reach unbalanced states. A coarse analysis as the one in R50 is not able to reproduce a detailed state, but can show the broader probability of convection. A drawback of this is the possibility that such a state might also be unphysical, as due to the large spread in the ensemble there can be "drizzle" over a large area, while the single members all contain thunderstorms. The influence of model error in complex weather models might even enlarge the problems in creating a useful fine analysis.

In summary one can say that the R10 forecast is not optimal, as the strongly forced analysis puts the ensemble members in unphysical states. Additionally the removal of mass produces a different convective environment. Therefore during the forecast phase the model will get into a balanced state first, which not necessarily is close to the analyzed state. Due to the loss of mass, the potential of gravity waves to create new clouds decreases which also leads to a degradation. The R50 on the other hand has a more physical behaviour. Although the analyzed clouds are not exactly at the correct positions, they show the potential of the future evolution and also the uncertainty associated with the forecast.

Comparing the results of the data assimilation experiments to an idealized setting of the COSMO-KENDA system, showed many similarities. In both models, the fine resolved analysis is better, but loses this advantage soon in the free forecast phase. The characteristic way rain regions are analyzed and differ between the coarse and fine analysis is also comparable. Overall the behaviour of the two models is similar and the modified shallow water model can be viewed as a suitable test model for convective-scale data assimilation.

The modified shallow water model also compares well to the stochastic model. In both models, the EnKF methods capture the correct clouds first and then slowly start to remove the wrong ones. More details on the comparison of these two simple models can be found in Kigle (2013).

Chapter 6

Conclusions and outlook

Because NWP models start to resolve convection, this thesis set out to explore the performance of data assimilation methods in a convective environment simulated by test models specifically designed for this purpose. As no idealized models representing the key features of convection have existed so far simple test models are needed. Two possible models have been identified and tested in this work: A simple stochastic birth-death model and a more sophisticated one-dimensional shallow water model modified to show convection-like characteristics. The random part in the evolution of these models is important for representing the non-Gaussianity and nonlinearity that are caused by convective motion as well as radar observations with their high temporal and spatial resolution.

The simple stochastic model focused on the randomness of the convective evolution as well as the non-Gaussianity caused by the observing system. It proved that the suppression of spurious clouds is a difficult task and as has been indicated by recent publications (Aksoy et al., 2009; Lange and Craig, 2013), there is a need to use no-rain observations to get reasonable results. Problems arise since there is no possibility to assimilate negative rain amounts. The stochastic model is an extreme case of the issue of non-Gaussianity. Lien et al. (2013) propose to assimilate gaussian transformated precipitation observations instead of normal rain to get a more Gaussian distribution. This could also help in the stochastic model and would be an interesting future application. Linearity and Gaussianity are mainly a concern for Ensemble Kalman Filters, as particle filters do not rely on any assumptions. In the stochastic model it has been shown that a particle filter can give good results even under extreme nonlinearity and non- Gaussianity. But the "curse of dimensionality" is still an open problem on the application of these methods to models with more degrees of freedom.

In the hierarchy of models, the modified shallow water model introduces more complexity and correlations between grid points. While the stochastic model can be used to do very basic research on new data assimilation methods and gives hints on potential issues, the modified shallow water model is much more realistic. The higher dimensionality, correlation between grid points and influence of gravity waves introduce the physical processes desired for allowing a better comparison to atmospheric models. The model is able to reproduce various atmospheric characteristics as convective life-cycles, orographically driven convection and cloud size-distributions. This makes the model a useful tool to do further research with data assimilation methods aimed for the convective-scale. Its computational efficiency allows to run hundreds of different experiments and choose the interesting ones to be rerun in a full atmospheric model.

The results of data assimilation experiments in the modified shallow water model confirmed the problems and characteristics known from the stochastic model, but additional ones appeared. Unphysical analysis states and the lack of mass conservation have been identified as two core problems and fit into current knowledge of convective-scale data assimilation. Still, the graveness at which these deficiencies appear is alarming and its sensitivities have been illustrated.

It has been shown that an analyzed state from a high-resolution assimilation only has an advantage over an assimilation with superobservations for a short amount of free forecast lead time. As the fine resolution has a higher computational cost, its use in full atmospheric models might be questioned or at least restricted to the ultra short range. The difficulty in using RMS errors in analyzing intermittent fields is a known issue and has been confirmed to be the case especially for an analyzed rain amount with a lot of spread as the one of the coarse resolution experiment. In such a case the effect of the double penalty is too large to make a reliable statement. Object-based measures gave more conclusive results and in future research one might also need to consider the use of probabilistic scores to investigate the forecast fields.

Comparing the two test models, they show a couple of similarities. In both models the LETKF captures the correct clouds in the first few assimilation cycles and then slowly removes the wrong ones. The experiments with superobservations have shown that there is potential for improving some aspects of the filter behaviour. Averaging in the stochastic model has clearly revealed that the Gaussianity of the error distribution increased and produced a better analysis. This could also be seen in the modified shallow water model during the free forecast phase. The error of the fine analysis increased much faster than the error of the coarse analysis and after only 10-20 minutes

the coarse forecast was better.

To conclude on the main scientific questions one can state that part of a model hierarchy representing key features of convective-scale movement has been created. The models are computationally cheap and useful for data assimilation experiments. But it is not the whole hierarchy, as there is the COSMO-KENDA system which can be used in an idealized or full mode. These are the highest levels of complexity in the hierarchy but they are computationally much more expensive. Additionally the usefulness of the stochastic model turned out to be questionable. The results can be reproduced in the modified shallow water model but it does not have as many possibilities and the same realism. As the modified shallow water model is still computationally very cheap, it seems to be the preferred choice for data assimilation experiments.

Concerning the second scientific objective the data assimilation methods have been examined and typical advantages and disadvantages identified. Some light has been shed on the need of a high-resolution analysis compared to a coarser scheme with averaged observations. Of course the results of this thesis are limited to the methods that have been tested. There are a lot of different methods and modifications possible to be used. As all the experiments performed are perfect model runs, the missing model error is an important shortage compared to full atmospheric models which contain model error a priori.

The last scientific target of this thesis was to predict the behaviour of data assimilation methods in high-resolution atmospheric model. It has been demonstrated that while still being computationally cheap, data assimilation results from the modified shallow water model compare well to similar experiments which have been accomplished in an atmospheric model by Lange and Craig (2013). Therefore the modified shallow water model can be a powerful tool in the development, understanding and tuning of new data assimilation methods. But the model has only been used to reproduce results from the COSMO-KENDA system so far and going the other way could be an additional step in verifying the model as a useful test model. The results presented in this thesis look promising enough to expect that the modified shallow water model can be used to predict the behaviour of data assimilation methods in high-resolution atmospheric models.

There are two different lines, along which future research could develop. One is concerning the models themselves, while the other one relates to data assimilation. One obvious thing would be to run data assimilation experiments in the modified

shallow water model with a mean wind and orography instead of the background noise. As the mountain forces the initiation of convection, such a situation should be assimilated quite well. Additionally the resolution of the model used for generating the nature run could have a lower resolution than the one used for assimilation. By doing so, one could investigate the influence of model error on the performance of data assimilation methods. If the assimilation does not resolve the mountain good enough, convection might be difficult to be reproduced in the analysis. Such an error is especially important for convection in mountainous regions which can be triggered by local small-scale convergence zones generated over complex terrain. It would be important for future research to know how large the effect of the orography and its model representation is on the quality of the analysis.

Another way to use the versatility of the shallow water model would be to create different cloud sizes and life-times by tuning the parameters. This could help find out how the methods behave in different convective environments. Further possibilities for model variations would be to modify the model even more and enhance its complexity. One opportunity would be to expand the model to two horizontal dimensions. This would enable to get the influence of additional scales and one could add a large-scale forcing to drive the initiation of convection. Another option would be to include an additional control variable for the precipitation as an equivalence to CAPE. The cloud threshold H_c could also be varied locally which would correspond to the convective inhibition (CIN). Therefore one could generate regions which are preferred to have convective activity and also different phases similar to the diurnal cycles or frontal forcing. These results could then be related to the convective time-scale which is a measure of the type of convective forcing (Keil and Craig, 2011; Zimmer et al., 2011). This could distinguish the influence of predictability on the performance of data assimilation systems. This knowledge can be important when it comes to warnings issued for thunderstorms. When the predictability varies depending on the type of convection and the forcing, it is important to take this information into account before warnings are released to the public.

Developing superior data assimilation methods and testing them ahead of their operational use is important and can be performed faster using the hierarchy of models presented in this thesis. As mass conservation turned out to be an important issue, a first thing to do concerning data assimilation could be to test different approaches to create methods which are better conserving. One possibility is to use an Iterated Kalman Filter as proposed by Zupanski (2005) which is a type of Maximum Likelihood

Ensemble Filter (MLEF). Another one would be to do a multistep analysis which is currently being developed at the DWD. This would allow to use different localization radii for different observation types which could help create more physically consistent fields and better conservation.

The large increments which are applied in the first few assimilation cycles are part of the same problem. Kalnay and Yang (2010) proposed a method accelerating the spin-up while still creating physically consistent fields. This method is called Running In Place (RIP) and has so far only been tested in a quasi-geostrophic model and a Three-variable Lorenz model (Yang et al., 2012) while a model representing a convective environment is still missing. The method only helps spinning up an ensemble and might not help much in a situation with a lot of convective activity. Therefore this method could also be tested in the modified shallow water model.

The same could be done for testing various other data assimilation methods like the one described in Desroziers et al. (2005). They are using an adaptive estimation of appropriate observation errors for observation error correlation matrix $\mathbf{R}$. Another possibility would be to assimilate precipitation only at locations where at least a couple of ensemble members have precipitation. This is done for global precipitation in a simplified general circulation model by Lien et al. (2013) and still needs to be tested for the convective-scale.

Finally methods which are still being developed like the nudging particle filter (van Leeuwen, 2010) could also be tested and prepared for their use in full atmospheric models.

In summary, this thesis has shown how data assimilation experiments are possible in models designed especially for the representation of the convective-scale. The characteristics of the different methods have been analyzed as well as the usefulness and relevance of the two models.

List of abbreviations

4D-Var	Four Dimensional Variational Assimilation
AETKF	ETKF with averaged observations
CAPE	Convective Available Potential Energy
COSMO	Consortium for Small-scale Modeling
CRM	Cloud Resolving Model
DWD	Deutscher Wetterdienst
EKF	Extended Kalman Filter
EnKF	Ensemble Kalman Filter
EnSRF	Ensemble Square Root Filter
ETKF	Ensemble Transform Kalman Filter
GPS	Global Positioning System
hl	half life
KENDA	Kilometer-scale Ensemble Data Assimilation
LEKF	Local Ensemble Kalman Filter
LETKF	Local Ensemble Transform Kalman Filter
LHN	Latent Heat Nudging
MLEF	Maximum Likelihood Ensemble Filter
NWP	Numerical Weather Prediction
PDF	Probability Distribution Function
QPF	Quantitative Precipitation Forecasts
RAW	Robert-Asselin-Williams
RIP	Running In Place
RMS	Root Mean Square
SAL	Structure, Amplitude and Location
SIR	Sequential Importance Resampling

Bibliography

Aksoy, A., D. C. Dowell, and C. Snyder, 2009: A Multicase Comparative Assessment of the Ensemble Kalman Filter for Assimilation of Radar Observations. Part I: Storm-scale Analyses. *Mon. Weather Rev.*, **137**, 1805–1824.

Aksoy, A., D. C. Dowell, and C. Snyder, 2010: A Multicase Comparative Assessment of the Ensemble Kalman Filter for Assimilation of Radar Observations. Part II: Short-range Ensemble Forecasts. *Mon. Weather Rev.*, **138**, 1273–1292.

Alpert, J. C. and V. K. Kumar, 2007: Radial Wind Super-Obs from the WSR-88D Radars in the NCEP Operational Assimilation System. *Mon. Weather Rev.*, **135**, 1090–1109.

Anderson, J. L., 2001: An Ensemble Adjustment Kalman Filter for Data Assimilation. *Mon. Weather Rev.*, **129**, 2884–2903.

Anderson, J. L., 2007a: An adaptive covariance inflation error correction algorithm for ensemble filters. *Tellus A*, **59**, 210–224.

Anderson, J. L., 2007b: Exploring the need for localization in ensemble data assimilation using a hierarchical ensemble filter. *Physica D: Nonlinear Phenomena*, **230**, 99 – 111.

Anderson, J. L., 2009: Spatially and temporally varying adaptive covariance inflation for ensemble filters. *Tellus A*, **61**, 72–83.

Anderson, J. L. and S. L. Anderson, 1999: A Monte Carlo Implementation of the Nonlinear Filtering Problem to Produce Ensemble Assimilations and Forecasts. *Mon. Weather Rev.*, **127**, 2741–2758.

Anthes, R. A., 1983: Regional Models of the Atmosphere in Middle Latitudes. *Mon. Weather Rev.*, **111**, 1306–1335.

Beck, A. and M. Ehrendorfer, 2003: *Data assimilation and covariance dynamics in a quasigeostrophic model*. Facultas Verlag.

Bellman, R., 1957: *Dynamic programming*. Rand Corporation research study, Princeton University Press.

Bellman, R., 1961: *Adaptive control processes: a guided tour*. A Rand Corporation Research Study Series, Princeton University Press.

Bengtsson, T., P. Bickel, and B. Li, 2008: Curse-of-dimensionality revisited: Collapse of the particle filter in very large scale systems. *IMS Collections*, **2**, 316–334.

Bickel, P., B. Li, and T. Bengtsson, 2008: Sharp failure rates for the bootstrap particle filter in high dimensions. Pushing the Limits of Contemporary Statistics. *Contributions in Honor of Jayanta K. Ghosh*, **3**, 318–329.

Bierdel, L., P. Friederichs, and S. Bentzien, 2012: Spatial kinetic energy spectra in the convection-permitting limited-area NWP model COSMO-DE. *Meteorol. Z.*, **21**, 245–258.

Bischof, M., 2011: Ensemble Simulations of Convective Storms. Master's thesis, ETH Zürich.

Bishop, C. and Z. Toth, 1999: Adaptive Sampling with the Ensemble Transform Kalman Filter. Part I: Theoretical Aspects. *Mon. Weather Rev.*, **56**, 1748–1765.

Bocquet, M., C. A. Pires, and L. Wu, 2010: Beyond Gaussian Statistical Modeling in Geophysical Data Assimilation. *Mon. Weather Rev.*, **138**, 2997–3023.

Bouttier, F. and G. Kelly, 2001: Observing-system experiments in the ECMWF 4D-Var data assimilation system. *Q. J. R. Meteorol. Soc.*, **127**, 1469–1488.

Bouttier, F. and F. Rabier, 1998: The operational implementation of 4D-Var. *ECMWF Newsletter*, **78**, 2–5, available online at http://ecmwf.int/publications/newsletters/pdf/78.pdf.

Bryan, G. H. and H. Morrison, 2011: Sensitivity of a Simulated Squall Line to Horizontal Resolution and Parameterization of Microphysics. *Mon. Weather Rev.*, **140**, 202–225.

Bryan, G. H., J. C. Wyngaard, and J. M. Fritsch, 2003: Resolution Requirements for the Simulation of Deep Moist Convection. *Mon. Weather Rev.*, **131**, 2394–2416.

Campbell, W. F., C. H. Bishop, and D. Hodyss, 2010: Vertical Covariance Localization for Satellite Radiances in Ensemble Kalman Filters. *Mon. Weather Rev.*, **138**, 282–290.

Caya, A., J. Sun, and C. Snyder, 2005: A Comparison between the 4DVAR and the Ensemble Kalman Filter Techniques for Radar Data Assimilation. *Mon. Weather Rev.*, **133**, 3081–3094.

Charney, J. G., R. Fjörtoft, and J. Von Neumann, 1950: Numerical Integration of the Barotropic Vorticity Equation. *Tellus*, **2**, 237–254.

Chu, C.-M. and Y.-L. Lin, 2000: Effects of Orography on the Generation and Propagation of Mesoscale Convective Systems in a Two-Dimensional Conditionally Unstable Flow. *J. Atmos. Sci.*, **57**, 3817–3837.

Cohen, B. G. and G. C. Craig, 2006: Fluctuations in an Equilibrium Convective Ensemble. Part II: Numerical Experiments. *J. Atmos. Sci.*, **63**, 2005–2015.

Cohn, S. E., 2010: The Principle of Energetic Consistency in Data Assimilation. *Data Assimilation - Making Sense of Observations*, W. Lahoz, R. Swinbank, and B. Khattatov, Eds., Springer, 137–216.

Cohn, S. E. and R. Todling, 1996: Approximate Data Assimilation Schemes for Stable and Unstable Dynamics. *J. Met. Soc. Japan*, **74**, 63–75.

Cox, D. R. and V. Isham, 2000: *Point Processes*. Chapman & Hall/CRC, New York.

Craig, G. C. and B. G. Cohen, 2006: Fluctuations in an Equilibrium Convective Ensemble. Part I: Theoretical Formulation. *J. Atmos. Sci.*, **63**, 1996–2004.

Craig, G. C., C. Keil, and D. Leuenberger, 2012: Constraints on the impact of radar rainfall data assimilation on forecasts of cumulus convection. *Q. J. R. Meteorol. Soc.*, **138**, 340–352.

Craig, G. C. and M. Würsch, 2013: The impact of localization and observation averaging for convective-scale data assimilation in a simple stochastic model. *Q. J. R. Meteorol. Soc.*, **139**, 515–523.

Dance, S., 2004: Issues in high resolution limited area data assimilation for quantitative precipitation forecasting. *Physica D: Nonlinear Phenomena*, **196**, 1 – 27.

Dawson, D. T., L. J. Wicker, E. R. Mansell, and R. L. Tanamachi, 2012: Impact of the Environmental Low-Level Wind Profile on Ensemble Forecasts of the 4 May 2007 Greensburg, Kansas, Tornadic Storm and Associated Mesocyclones. *Mon. Weather Rev.*, **140**, 696–716.

Desroziers, G., L. Berre, B. Chapnik, and P. Poli, 2005: Diagnosis of observation, background and analysis-error statistics in observation space. *Q. J. R. Meteorol. Soc.*, **131**, 3385–3396.

Done, J., C. A. Davis, and M. Weisman, 2004: The next generation of NWP: explicit forecasts of convection using the weather research and forecasting (WRF) model. *Atmos. Sci. Lett.*, **5**, 110–117.

Doucet, A., N. de Freitas, and N. Gordon, 2001: An Introduction to Sequential Monte Carlo Methods. *Sequential Monte Carlo Methods in Practice*, Springer New York, 3–14.

Dowell, D., L. Wicker, and C. Snyder, 2011: Ensemble Kalman Filter Assimilation of Radar Observations of the 8 May 2003 Oklahoma City Supercell: Influences of Reflectivity Observations on Storm-Scale Analyses. *Mon. Weather Rev.*, **139**, 272–294.

Dowell, D. C., F. Zhang, L. J. Wicker, C. Snyder, and N. A. Crook, 2004: Wind and Temperature Retrievals in the 17 May 1981 Arcadia, Oklahoma, Supercell: Ensemble Kalman Filter Experiments. *Mon. Weather Rev.*, **132**, 1982–2005.

Ehrendorfer, M. and R. Errico, 2008: An atmospheric model of intermediate complexity for data assimilation studies. *Q. J. R. Meteorol. Soc.*, **134**, 1717–1732.

Evensen, G., 1994: Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics. *J. Geophys. Res.*, **99**, 10 143–10 162.

Evensen, G., 2003: The Ensemble Kalman Filter: theoretical formulation and practical implementation. *Ocean Dynamics*, **53**, 343–367.

Fisher, M., 1998: *Development of a Simplified Kalman Filter*. ECMWF technical memorandum, ECMWF Technical Memo. No. 260.

Fisher, M. and E. Andersson, 2001: *Developments in 4D-Var and Kalman filtering*. ECMWF Technical Memo. No. 347.

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

Gill, A., 1982: *Atmosphere-Ocean Dynamics*. International Geophysics Series, Academic Press.

Gohm, A. and G. J. Mayr, 2004: Hydraulic aspects of föhn winds in an Alpine valley. *Q. J. R. Meteorol. Soc.*, **130**, 449–480.

Gordon, N. J., D. J. Salmond, and A. F. Smith, 1993: Novel approach to nonlinear/non-Gaussian Bayesian state estimation. *IEE Proc.*, **140**, 107–113.

Greybush, S. J., E. Kalnay, T. Miyoshi, K. Ide, and B. R. Hunt, 2010: Balance and Ensemble Kalman Filter Localization Techniques. *Mon. Weather Rev.*, **139**, 511–522.

Hamill, T. M. and J. S. Whitaker, 2005: Accounting for the Error due to Unresolved Scales in Ensemble Data Assimilation: A Comparison of Different Approaches. *Mon. Weather Rev.*, **133**, 3132–3147.

Hamill, T. M., J. S. Whitaker, and C. Snyder, 2001: Distance-Dependent Filtering of Background Error Covariance Estimates in an Ensemble Kalman Filter. *Mon. Weather Rev.*, **129**, 2776–2790.

Hohenegger, C. and C. Schär, 2007a: Atmospheric predictability at synoptic versus cloud-resolving scales. *Bull. Am. Meteor. Soc.*, **88**, 1783–1793.

Hohenegger, C. and C. Schär, 2007b: Predictability and error growth dynamics in cloud-resolving models. *J. Atmos. Sci.*, **64**, 4467–4478.

Houtekamer, P., H. L. Mitchell, and X. Deng, 2009: Model error representation in an operational ensemble Kalman filter. *Mon. Weather Rev.*, **137**, 2126–2143.

Houtekamer, P. L. and H. L. Mitchell, 1998: Data Assimilation Using an Ensemble Kalman Filter Technique. *Mon. Weather Rev.*, **126**, 796–811.

Houtekamer, P. L. and H. L. Mitchell, 2001: A Sequential Ensemble Kalman Filter for Atmospheric Data Assimilation. *Mon. Weather Rev.*, **129**, 123–137.

Houze, R. A., 1993: *Cloud Dynamics*. Academic Press.

Hunt, B. R., E. J. Kostelich, and I. Szunyogh, 2007: Efficient data assimilation for spatiotemporal chaos: A local ensemble transform kalman filter. *Physica D*, **230**, 112–126.

Ide, K., P. Courtier, M. Ghil, and A. Lorenz, 1997: Unified Notation for Data Assimilation: Operational, Sequential and Variational. *J. Met. Soc. Japan,*, **75**, 181–189.

Jazwinski, A., 1970: *Stochastic Processes and Filtering Theory*. Mathematics in Science and Engineering, Elsevier Science.

Jones, C. and B. Macpherson, 1997: A latent heat nudging scheme for the assimilation of precipitation data into an operational mesoscale model. *Met. Apps*, **4**, 269–277.

Kalman, R., 1960: A new approach to linear filtering and prediction problems. *Transactions of the ASME–Journal of Basic Engineering*, **82**, 35–45.

Kalman, R. and R. Bucy, 1961: New results in linear filtering and prediction theory. *Transactions of the ASME–Journal of Basic Engineering*, **83**, 95–107.

Kalnay, E. and S.-C. Yang, 2010: Accelerating the spin-up of Ensemble Kalman Filtering. *Q. J. R. Meteorol. Soc.*, **136**, 1644–1651.

Keil, C. and G. C. Craig, 2009: A displacement and amplitude score employing an optical flow technique. *Wea. Forecasting*, **24**, 1297–1308.

Keil, C. and G. C. Craig, 2011: Regime-dependent forecast uncertainty of convective precipitation. *Meteorol. Z.*, **20**, 145–151.

Kepert, J. D., 2009: Covariance localisation and balance in an Ensemble Kalman Filter. *Q. J. R. Meteorol. Soc.*, **135**, 1157–1176.

Keppenne, C. L., 2000: Data Assimilation into a Primitive-Equation Model with a Parallel Ensemble Kalman Filter. *Mon. Weather Rev.*, **128**, 1971–1981.

Kigle, S., 2013: Datenassimilationsexperimente an einem Modifizierten Flachwassermodell mit dem Local Ensemble Transform Kalman Filter. Bachelorarbeit, LMU München.

Kim, S., G. L. Eyink, J. M. Restrepo, F. J. Alexander, and G. Johnson, 2003: Ensemble Filtering for Nonlinear Dynamics. *Mon. Weather Rev.*, **131 (11)**, 2586–2594.

Kober, K., G. C. Craig, C. Keil, and A. Dörnbrack, 2012: Blending a probabilistic nowcasting method with a high-resolution numerical weather prediction ensemble for convective precipitation forecasts. *Q. J. R. Meteorol. Soc.*, **138**, 755–768.

Kuo, Y.-H., J. Bresch, M.-D. Cheng, J. Kain, D. B. Parsons, W.-K. Tao, and D.-L. Zhang, 1997: Summary of a mini-workshop on cumulus parameterization for mesoscale models. *Bull. Am. Meteor. Soc.*, **78**, 475–492.

Lane, T. P. and F. Zhang, 2011: Coupling between gravity waves and tropical convection at mesoscales. *Journal of the Atmospheric Sciences*, **68**, 2582–2598.

Lange, H., 2013: Data assimilation of severe convection using a Local Ensemble Kalman Filter. Master's thesis, LMU München.

Lange, H. and G. C. Craig, 2013: On the Benefits of a High-Resolution Analysis for Convective Data Assimilation of Radar Observations using a Local Ensemble Kalman Filter, submitted.

Leuenberger, D. and A. Rossa, 2007: Revisiting the latent heat nudging scheme for the rainfall assimilation of a simulated convective storm. *Meteorol. Atmos. Phys.*, **98**, 195–215.

Li, H., E. Kalnay, and T. Miyoshi, 2009: Simultaneous estimation of covariance inflation and observation errors within an ensemble Kalman filter. *Q. J. R. Meteorol. Soc.*, **135**, 523–533.

Lien, G.-Y., E. Kalnay, and T. Miyoshi, 2013: Effective assimilation of global precipitation: simulation experiments. *Tellus A*, **65**.

Lin, C., S. Vasić, A. Kilambi, B. Turner, and I. Zawadzki, 2005: Precipitation forecast skill of numerical weather prediction models and radar nowcasts. *Geophys. Res. Letters*, **32**, L14 801.

Liu, J. S. and R. Chen, 1998: Sequential Monte Carlo Methods for Dynamic Systems. *J. Am. Statist. Assoc.*, **93**, 1032–1044.

Lorenz, E., 1963: Deterministic Nonperiodic Flow. *J. Atmos. Sci.*, **20**, 130–141.

Lorenz, E. N., 1995: Predictability: A problem partly solved. *In Proc. Sem. Predictability*, **1**, 1–18.

Lorenz, E. N., 2005: Designing Chaotic Models. *J. Atmos. Sci.*, **62**, 1574–1587.

Macpherson, B., 2001: Operational experience with assimilation of rainfall data in the Met Office Mesoscale model. *Meteorol. Atmos. Phys.*, **76**, 3–8.

Metropolis, N. and S. Ulam, 1949: The Monte Carlo Method. *J. Am. Stat. Assoc.*, **44**, 335–341.

Miglietta, M. M. and R. Rotunno, 2009: Numerical Simulations of Conditionally Unstable Flows over a Mountain Ridge. *J. Atmos. Sci.*, **66**, 1865–1885.

Mitchell, H. L., P. Houtekamer, and G. Pellerin, 2002: Ensemble Size, Balance, and Model-Error Representation in an Ensemble Kalman Filter. *Mon. Weather Rev.*, **130**, 2791–2808.

Miyoshi, T., 2011: The Gaussian approach to adaptive covariance inflation and its implementation with the local ensemble transform Kalman filter. *Mon. Weather Rev.*, **139**, 1519–1535.

Miyoshi, T. and Y. Sato, 2007: Assimilating Satellite Radiances with a Local Ensemble Transform Kalman Filter (LETKF) Applied to the JMA Global Model (GSM). *SOLA*, **3**, 37–40.

Møller, J. and M. Sørenson, 1994: Statistical analysis of a spatial birth-and-death process model with a view to modelling linear dune fields. *Scand. J. Statist.*, **21**, 1–19.

Nastrom, G. and K. S. Gage, 1985: A climatology of atmospheric wavenumber spectra of wind and temperature observed by commercial aircraft. *J. Atmos. Sci.*, **42**, 950–960.

Ott, E., et al., 2002: Exploiting Local Low Dimensionality of the Atmospheric Dynamics for Efficient Ensemble Kalman Filtering. Tech. rep. URL `http://arxiv.org/abs/physics/0203058v3/`.

Ott, E., et al., 2004: A local ensemble Kalman filter for atmospheric data assimilation. *Tellus A*, **56**, 415–428.

Palmer, T. N., 2001: A nonlinear dynamical perspective on model error: A proposal for non-local stochastic-dynamic parametrization in weather and climate prediction models. *Q. J. R. Meteorol. Soc.*, **127**, 279–304.

Patil, D. J., B. R. Hunt, E. Kalnay, J. A. Yorke, and E. Ott, 2001: Local Low Dimensionality of Atmospheric Dynamics. *Phys. Rev. Lett.*, **86**, 5878–5881.

Plant, R. S. and G. C. Craig, 2008: A Stochastic Parameterization for Deep Convection Based on Equilibrium Statistics. *J. Atmos. Sci.*, **65**, 87–105.

Potvin, C. K. and L. J. Wicker, 2012: Comparison between Dual-Doppler and EnKF Storm-Scale Wind Analyses: Observing System Simulation Experiments with a Supercell Thunderstorm. *Mon. Weather Rev.*, **140**, 3972–3991.

Reich, H., A. Rhodin, and C. Schraff, 2011: LETKF for the nonhydrostatic regional model COSMO-DE. Newsletter 11, Consortium for Small-scale Modelling, 27–31 pp. URL `http://www.cosmo-model.org/content/model/documentation/newsLetters/newsLetter11`.

Schär, C. and R. Smith, 1993a: Shallow Water Flow Past Isolated Topography Part I: Vorticity Production and Wake Formation. *J. Atmos. Sci.*, **50**, 1373–1400.

Schär, C. and R. Smith, 1993b: Shallow Water Flow Past Isolated Topography. Part II: Transition to Vortex Shedding. *J. Atmos. Sci.*, **50**, 1401–1412.

Snyder, C., T. Bengtsson, P. Bickel, and J. Anderson, 2008: Obstacles to High-Dimensional Particle Filtering. *Mon. Weather Rev.*, **136**, 4629–4640.

Snyder, C. and F. Zhang, 2003: Assimilation of Simulated Doppler Radar Observations with an Ensemble Kalman Filter. *Mon. Weather Rev.*, **131**, 1663–1677.

Stephan, K., S. Klink, and C. Schraff, 2008: Assimilation of radar-derived rain rates into the convective-scale model COSMO-DE at DWD. *Q. J. R. Meteorol. Soc.*, **134**, 1315–1326.

Talagrand, O. and P. Courtier, 1987: Variational Assimilation of Meteorological Observations With the Adjoint Vorticity Equation. I: Theory. *Q. J. R. Meteorol. Soc.*, **113**, 1311–1328.

Tippett, M. K., J. L. Anderson, C. H. Bishop, T. M. Hamill, and J. S. Whitaker, 2003: Ensemble Square Root Filters. *Mon. Weather Rev.*, **131**, 1485–1490.

Tompkins, A. M. and G. C. Craig, 1998: Radiative–convective equilibrium in a three-dimensional cloud-ensemble model. *Q. J. R. Meteorol. Soc.*, **124**, 2073–2097.

Tong, M. and M. Xue, 2005: Ensemble Kalman Filter Assimilation of Doppler Radar Data with a Compressible Nonhydrostatic Model: OSS Experiments. *Mon. Weather Rev.*, **133**, 1789–1807.

van Leeuwen, P. J., 2009: Particle Filtering in Geophysical Systems. *Mon. Weather Rev.*, **137**, 4089–4114.

van Leeuwen, P. J., 2010: Nonlinear data assimilation in geosciences: an extremely efficient particle filter. *Q. J. R. Meteorol. Soc.*, **136**, 1991–1999.

Wernli, H., M. Paulat, M. Hagen, and C. Frei, 2008: SAL-A novel quality measure for the verification of quantitative precipitation forecasts. *Mon. Weather Rev.*, **136**, 4470–4487.

Weusthoff, T., F. Ament, M. Arpagaus, and M. W. Rotach, 2010: Assessing the benefits of convection-permitting models by neighborhood verification: Examples from MAP D-PHASE. *Mon. Weather Rev.*, **138**, 3418–3433.

Whitaker, J. S. and T. M. Hamill, 2002: Ensemble Data Assimilation without Perturbed Observations. *Monthly Weather Review*, **130**, 1913–1924.

Williams, P. D., 2009: A Proposed Modification to the Robert–Asselin Time Filter. *Mon. Weather Rev.*, **137**, 2538–2546.

Williams, P. D., 2011: The RAW Filter: An Improvement to the Robert–Asselin Filter in Semi-Implicit Integrations. *Mon. Weather Rev.*, **139**, 1996–2007.

Wu, G., X. Zheng, L. Wang, S. Zhang, X. Liang, and Y. Li, 2012: A new structure for error covariance matrices and their adaptive estimation in EnKF assimilation. *Q. J. R. Meteorol. Soc.*, doi:10.1002/qj.2000.

Würsch, M. and G. C. Craig, 2013: A simple dynamical model of cumulus convection for data assimilation research, in review.

Yang, S.-C., E. Kalnay, and B. Hunt, 2012: Handling nonlinearity in an ensemble Kalman filter: Experiments with the three-variable Lorenz model. *Mon. Weather Rev.*, **140**, 2628–2646.

Yang, S.-C., E. Kalnay, B. Hunt, and N. E Bowler, 2009: Weight interpolation for efficient data assimilation with the Local Ensemble Transform Kalman Filter. *Q. J. R. Meteorol. Soc.*, **135**, 251–262.

Zhang, F., C. Snyder, and R. Rotunno, 2003: Effects of Moist Convection on Mesoscale Predictability. *J. Atmos. Sci.*, **60**, 1173–1185.

Zhang, F., C. Snyder, and J. Sun, 2004: Impacts of Initial Estimate and Observation Availability on Convective-Scale Data Assimilation with an Ensemble Kalman Filter. *Mon. Weather Rev.*, **132**, 1238–1253.

Zhang, F., Y. Weng, J. A. Sippel, Z. Meng, and C. H. Bishop, 2009: Cloud-Resolving Hurricane Initialization and Prediction through Assimilation of Doppler Radar Observations with an Ensemble Kalman Filter. *Mon. Weather Rev.*, **137**, 2105–2125.

Zhao, L., W. Y. Ochieng, M. A. Quddus, and R. B. Noland, 2003: An Extended Kalman Filter Algorithm for Integrating GPS and Low Cost Dead Reckoning System Data for Vehicle Performance and Emissions Monitoring. *The Journal of Navigation*, **56**, 257–275.

Zhou, Y., D. McLaughlin, and D. Entekhabi, 2006: Assessing the Performance of the Ensemble Kalman Filter for Land Surface Data Assimilation. *Mon. Weather Rev.*, **134**, 2128–2142.

Zimmer, M., G. Craig, C. Keil, and H. Wernli, 2011: Classification of precipitation events with a convective response timescale and their forecasting characteristics. *Geophys. Res. Letters*, **38**.

Zupanski, M., 2005: Maximum likelihood ensemble filter: Theoretical aspects. *Mon. Weather Rev.*, **133**, 1710–1726.

Zupanski, M., 2009: Theoretical and Practical Issues of Ensemble Data Assimilation in Weather and Climate. *Data Assimilation for Atmospheric, Oceanic and Hydrologic Applications*, Springer, 67–84.

Acknowledgements

At first, I would like to thank Prof. George Craig for supervising this thesis. He always had valuable suggestions on how to improve the thesis and contributed a lot to the development of the models. During our fruitful discussions I learned a lot about creativity, perseverance and scientific work in general.

I would also like to thank Prof. Roger Smith for being the second reviewer of this thesis.

Additional thanks go to Dr. Christian Keil for providing help and support concerning all possible problems.

Dr. Daniel Leuenberger of MeteoSchweiz always was enthusiastic about this work and provided some rewarding discussions.

Special thanks go to all the people at the Meteorological Institute Munich that assisted in the development of this work in various different ways, especially Lucas Fischer, Heiner Lange and Dr. Tobias Selz. Their supporting activities were many coffee and lunch breaks, proofreading, encouragement, discussions and a good atmosphere.

This research was funded by LMU Munich and carried out in the Hans Ertel Centre for Weather Research. This research network of Universities, Research Institutes and the Deutscher Wetterdienst is funded by the BMVBS (Federal Ministry of Transport, Building and Urban Development). I want to thank all the people within the Data Assimilation Branch of HErZ for many discussions and meetings. I also would like to thank the German Weather Service for giving me the possibility to take part of this research network.

Finally I want to express my deepest gratitude to my family and friends who always supported me in many ways. Without this loyalty finishing this thesis would have been much harder.