The impact of observations in convective-scale numerical weather prediction

Tobias Marcel Necker

Munich - June 13, 2019

The impact of observations in convective-scale numerical weather prediction

Tobias Marcel Necker

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

vorgelegt von
Tobias Marcel Necker
aus München

München, 13. Juni 2019

Erstgutachter: Dr. Martin Weissmann
Zweitgutachter: Prof. Dr. Bernhard Mayer
Datum der Abgabe: 13.06.2019
Datum der mündlichen Prüfung: 29.07.2019

Parts of this thesis are included in (see Appendix A):

Necker, T, M, Weissmann, M, Sommer, 2018: The importance of appropriate verification metrics for the assessment of observation impact in a convection-permitting modelling system. *Quarterly Journal of the Royal Meteorological Society*; 144: 1667– 1680. https://doi.org/10.1002/qj.3390

Necker, T, S, Geiss, M, Weissmann, J, Ruiz, T, Miyoshi, G-Y, Lien, 2019a: A convective-scale 1000-member ensemble simulation and potential applications. *Quarterly Journal of the Royal Meteorological Society*; (submitted)

Necker, T, M, Weissmann, Y, Ruckstuhl, J, Anderson, T, Miyoshi, 2019b: Sampling error correction evaluated using a convective-scale 1000-member ensemble. *Monthly Weather Review*; (submitted)

Zusammenfassung

Die Qualität von numerischen Wettervorhersagen wird stark von der Genauigkeit der Anfangsbedingungen bestimmt. Diese werden von Datenassimilationssystemen bereitgestellt, welche Millionen von Beobachtungen mit der neusten Kurzfristvorhersage kombinieren. Erst seit kurzem sind Wetterdienste in der Lage hochauflösende und konvektionserlaubende Vorhersagen durchzuführen. Der Schritt zu einer höheren Modellauflösung ist mit mehreren Herausforderungen in Bezug auf die Beobachtungen und die zugrunde liegenden Datenassimilationsalgorithmen verbunden. Insbesondere das chaotische Verhalten und die eingeschränkte Vorhersagbarkeit von Konvektion erfordern räumlich und zeitlich hoch aufgelöste Beobachtungen. Aktuell gibt es nur begrenzte Kenntnisse darüber, welche Beobachtungen für die hochauflösende numerische Wettervorhersage am wichtigsten sind. Daher ist ein besseres Verständnis des Einflusses verschiedener Beobachtungen auf konvektiver Skala erforderlich, um aktuelle Datenassimilations-, Vorhersage- und Beobachtungssysteme zu verbessern. Darüber hinaus sind Informationen über den Einfluss von Beobachtungen erforderlich, um zukünftige Beobachtungs- und Datenassimilierungsstrategien für die regionale numerische Wettervorhersage zu entwickeln.

Die vorliegende Arbeit untersucht, welche Beobachtungen konvektionserlaubende Ensemblevorhersagen am stärksten beeinflussen. Der Einfluss assimilierter Beobachtungen und der potenzielle Einfluss zukünftiger Beobachtungen wird mit zwei verschiedenen ensemblebasierten Methoden bestimmt. Beide Methoden basieren auf Stichprobenkorrelationen, die mit dem Ensemble geschätzt werden. Ensemblevorhersagesysteme nach aktuellem Stand stellen jedoch nur Ensembles mit 20 bis 250 Vorhersagen zur Verfügung, um die Unsicherheit der Vorhersage und räumliche sowie zeitliche Kovarianzen abzuschätzen. Da die Anzahl der Freiheitsgrade atmosphärischer Modelle jedoch um mehrere Größenordnungen höher ist, werden Stichprobenkorrelationen erheblich durch Stichprobenfehler beeinflusst. Folglich ist die begrenzte Ensemblegröße bei der Bestimmung des Beobachtungseinflusses, sowie in vielen anderen Ensembleanwendungen ein Problem. Aus diesem Grund ist es wichtig, Stichprobenfehler auf konvektiver Skala zu quantifizieren und Methoden zu entwickeln, um sie zu reduzieren. Um den zuvor diskutierten Herausforderungen zu begegnen, zielt diese Dissertation darauf ab, den Einfluss von Beobachtungen auf konvektionserlaubende Vorhersagen abzuschätzen und das Problem von Stichprobenfehlern zu reduzieren.

Im ersten Teil dieser Arbeit wird daher der Einfluss von rund 3 Millionen konventionellen Beobachtungen auf das regionale Vorhersagesystem des Deutschen Wetterdienstes untersucht. Diese Studie stellt die erste Auswertung von ensemblebasierten Schätzungen des Beobachtungseinflusses über einen längeren Zeitraum von sechs Wochen in einem konvektionserlaubenden Modellsystem dar. Nahezu alle früheren Studien verwendeten den

Unterschied zwischen der Vorhersage und einer Analyse desselben Modellsystems zur Verifikation. Diese Art der Verifikation spiegelt jedoch relevante Aspekte der konvektionserlaubenden Vorhersage unzureichend wider. Daher wird der Einfluss von Beobachtungen mit verschiedenen beobachtungsbasierten Verifikationsnormen ausgewertet.

Der zweite Teil dieser Arbeit stellt eine Methode zur Abschätzung des relativen potenziellen Einflusses verschiedener beobachtbarer Größen auf konvektionserlaubende Vorhersagen vor. Diese Methode basiert auf einer Ensemble-Sensitivitätsanalyse und verwendet akkumulierte quadratische räumliche und zeitliche Korrelationen als Näherung für den potenziellen Beobachtungseinfluss. Um zuverlässige Korrelationen zu erhalten, ist jedoch ein sehr großes Ensemble erforderlich. Daher wurde in Zusammenarbeit mit dem RIKEN Institut für Computerwissenschaften ein einzigartiges Ensemble mit 1000 konvektionserlaubenden Vorhersagen berechnet. Diese Simulationen ermöglichen es, die Empfindlichkeit der Methodik auf die Ensemblegröße hin zu untersuchen. Die vorliegende Studie hebt die Skalenabhängigkeit des potenziellen Einflusses hervor und bildet die Grundlage für die Entwicklung besserer Beobachtungs- und Datenassimilationsstrategien.

Im dritten Teil dieser Arbeit wird das Ensemble aus 1000 Vorhersagen als Referenz benutzt, um Stichprobenfehler auf konvektiver Skala zu quantifizieren und eine statistische Stichprobenfehlerkorrektur auszuwerten. Die Stichprobenfehlerkorrektur ist ein einfacher Ansatz auf der Basis einer Nachschlagetabelle und zielt darauf ab, zufällige Korrelationen zu reduzieren. Eine detaillierte Auswertung für räumliche und zeitliche Korrelationen zeigt, dass die Stichprobenfehlerkorrektur Fehler in Korrelationen, die zur Abschätzung des Einflusses von Beobachtungen erforderlich sind, signifikant reduziert. Außerdem unterstreicht die Studie das große Potential der Stichprobenfehlerkorrektur für eine Anwendung in der Datenassimilation, wo sie entfernungsbasierte Lokalisierungstechniken ersetzen und dadurch die Einbindung von Beobachtungen verbessern könnte.

Abstract

The accuracy of the initial conditions strongly determines the skill of numerical weather prediction (NWP). Data assimilation systems combine millions of observations with the latest short-range forecast to provide optimal initial conditions. Only recently, NWP centers are capable of performing high-resolution, convection-permitting forecasts on a regional scale. However, moving to a higher model resolution involves several challenges concerning observations and the underlying data assimilation algorithm. The chaotic nature and limited predictability of convection calls for spatially and temporally high resolved observations. However, limited knowledge exists on which observations are most important for high-resolution NWP. Hence, a better understanding of the impact of different observations on these scales is required to improve current data assimilation, forecasting, and observing systems. Furthermore, knowledge of the potential impact of observations is needed to develop advanced observation and data assimilation strategies for future convective-scale NWP.

This thesis, therefore, investigates the impact of observations in convective-scale ensemble forecasting. The impact of assimilated observation and the potential impact of future observations is evaluated by applying two complementary ensemble-based methods. Both methods rely on sample correlations that are estimated with an ensemble. However, state of the art ensemble prediction systems usually provide ensembles with only 20-250 members for estimating the uncertainty of the forecast and its spatial and temporal covariance. Given that the degrees of freedom of atmospheric models are several magnitudes higher, sample correlations are significantly affected by sampling errors. Therefore, sampling errors pose an issue for the impact assessment and in many other ensemble applications. Thus, it is essential to quantify sampling errors on convective-scales and to find methods to mitigate sampling errors. To address the previously discussed challenges, this dissertation aims to estimate the impact of observations and to reduce the issue of sampling error in convective-scale modeling and ensemble diagnostics.

The first part of this thesis evaluates the impact of about 3 million conventional observations in the regional ensemble forecasting system of Deutscher Wetterdienst. This study presents the first evaluation of ensemble-based estimates of observation impact over an extended period of six weeks in a convection-permitting modeling system. Nearly all previous observation impact studies used the difference between the forecast and subsequent analysis of the same modeling system for verification. However, this kind of verification does not adequately reflect relevant forecast aspects of convective-scale forecasting. Hence, the observation impact is examined for different observation-based verification norms.

The second part introduces an approach for estimating the relative potential impact of different observable quantities in convective-scale modeling. The approach is based on

ensemble sensitivity analysis and uses accumulated squared spatiotemporal correlations as a proxy for the potential impact. To obtain reliable spatiotemporal correlations, a very large ensemble is required. Therefore, an unprecedented convective-scale 1000-member ensemble was computed in collaboration with the RIKEN Institute for computational science. This simulation allows to examine the sensitivity of the approach to the ensemble size. The present study further highlights the scale dependence of the potential impact and provides the basis for developing better observation and data assimilation strategies.

The third part uses the 1000-member ensemble simulation as truth to quantify sampling errors on convective-scales and to evaluate a statistical sampling error correction. The sampling error correction is a simple look-up table based approach and aims to reduce spurious correlations. A detailed evaluation for spatiotemporal correlations shows that the sampling error correction significantly reduces sampling errors in sample correlations that are required for estimating the impact of observations. Additionally, the study demonstrates the great potential of the sampling error correction method for data assimilation where it could replace distance-based localization techniques and thereby increase the impact of observations.

Contents

Chapter 1

Introduction

1.1 Numerical weather prediction

The quality of weather forecasts influences both society and economy on a daily basis. Forecasts of extreme weather events help to protect human life and property. Weather warnings prevent economic losses and are crucial for road traffic and aviation. Forecasts for wind and solar radiation are necessary to predict the contribution of renewable energy sources to the power supply. The list of applications is long, and there are many reasons why various services require accurate weather prediction. Nowadays, operational numerical weather prediction (NWP) centers perform multiple forecasts per day to meet the growing demand. These forecasts range from a lead time of a few hours up to several weeks and are performed for different areas and with different spatial resolutions.
Over the past thirty years, the skill of NWP systems enhanced tremendously. This progress results from both scientific and technological advances in various fields (Bauer et al., 2015): First, new computational resources allowed to move to a higher spatial resolution. Consequently, most regional models nowadays represent deep convection explicitly using a horizontal grid-spacing of a few kilometers. Second, new NWP models with improved physics are available. That includes a better representation of subgrid-scale physical processes by more sophisticated parameterizations. Third, ensemble prediction systems deliver probabilistic forecasts that estimate the uncertainty of a forecast. Finally, advances in data assimilation (DA) made one of the most important contributions to the improvement of NWP. That includes a steadily growing observing system, which provides several millions of observations per day. Furthermore, advanced methods enable to assimilate additional complex observations and to exploit the provided information better.

Data assimilation Already Bjerknes (1904) postulated that weather prediction is an initial value problem. Nowadays, we know that the atmosphere is a non-linear, chaotic system (Lorenz, 1963) and that the accuracy of the initial state strongly determines the practical predictability of weather and the forecast skill. Data assimilation, therefore, aims to provide an optimal estimate of the initial state that is required to initialize a new forecast (Lorenc, 1986). State of the art data assimilation algorithms combine the latest short-range forecast with thousands or millions of observations to generate the best estimate of the initial state. This process is repeated for consecutive assimilation windows and called cycling (Figure 1.1). Operationally, near real-time cycling frequently provides

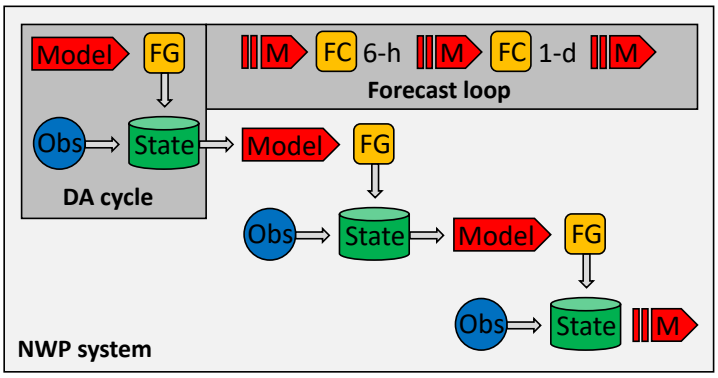


Figure 1.1: *Flow-chart of an NWP system. A data assimilation cycle typically combines a short-range forecast (background or first guess; FG) with observations (Obs) to provide the best initial state for a new forecast. A forecast loop is performed several times per day to provide long-range forecasts (FC) for various applications.*

new initial conditions to start forecasts several times per day. Weather services apply several different data assimilation schemes. The choice of the data assimilation scheme differs depending on the available computational resource, observational information, or the applied modeling system. Furthermore, the choice is a matter of the forecast resolution as the demands on global or regional scales are quite different.

State of the art data assimilation systems are based on variational, ensemble, or hybrid DA approaches. All these approaches are sequential DA methods that combine observations y and a background state x_b to generate a new analysis state x_a accounting for the error of both sources of information. Numerically, the state estimation for x can be solved minimizing a quadratic cost function J

$$J(x) = \frac{1}{2}(x - x_b)^T \mathbf{B^{-1}}(x - x_b) + \frac{1}{2}(y - H(x))^T \mathbf{R^{-1}}(y - H(x)), \tag{1.1}$$

where $\mathbf{B}$ is the background error covariance matrix, $\mathbf{R}$ the observation error covariance matrix and H the non-linear observation operator that maps from the model into observation space. Usually, both observation and background errors are assumed to be Gaussian distributed and unbiased. The background error covariance matrix is often named $\mathbf{P_b}$ (instead of $\mathbf{B}$) if it is estimated using an ensemble. In practice, adequate modeling of the background error covariance matrix is decisive on the performance of a DA algorithm. Insufficient modeling of error covariances leads to imbalances or a sub-optimal weighting of the information from background and observations. Subsequently, these effects likely cause a sub-optimal analysis and degrade the forecast.

Data assimilation methods mainly differ in the way they treat and obtain the background error covariance matrix. For variational schemes (e.g., 3DVAR), a static background error covariance matrix $\mathbf{B}$ is estimated climatologically with additional constraints such as geostrophic balance. Several prediction centers maintain a 4DVAR DA system, which is an extension of 3DVAR that accounts for the temporal evolution of the state within the assimilation window. 4DVAR additionally allows the $\mathbf{B}$ to develop throughout the assimilation window. However, applying variational methods on convective-scales is challenging. Usu-

ally, a climatologically estimated static $\mathbf{B}$ matrix is unsuitable for relatively fast-changing weather situations on these scales (Hohenegger and Schar, 2007). For convective-scale DA many operational centers, therefore, apply an Ensemble Kalman filter (EnKF) (Evensen, 1994). An EnKF uses an ensemble to estimate the error covariance matrix $\mathbf{P_b}$ and exhibits a flow-dependent $\mathbf{P_b}$. This flow-dependence is a significant advantage over a climatologically estimated background error (Houtekamer and Mitchell, 1998).
For global models, many weather services maintain hybrid data assimilation systems that combine the benefits of both variational and ensemble approaches. For example, Deutscher Wetterdienst (DWD) runs a hybrid 3D ensemble variational (3DEnVAR) DA system to obtain the initial conditions for the global model (ICON). For the regional forecast ensemble, DWD implemented a Local Ensemble Transform Kalman Filter (LETKF; Hunt et al. (2007)) system, which is a special type of EnKF. The global observing system delivers observations that are assimilated by all DA systems.

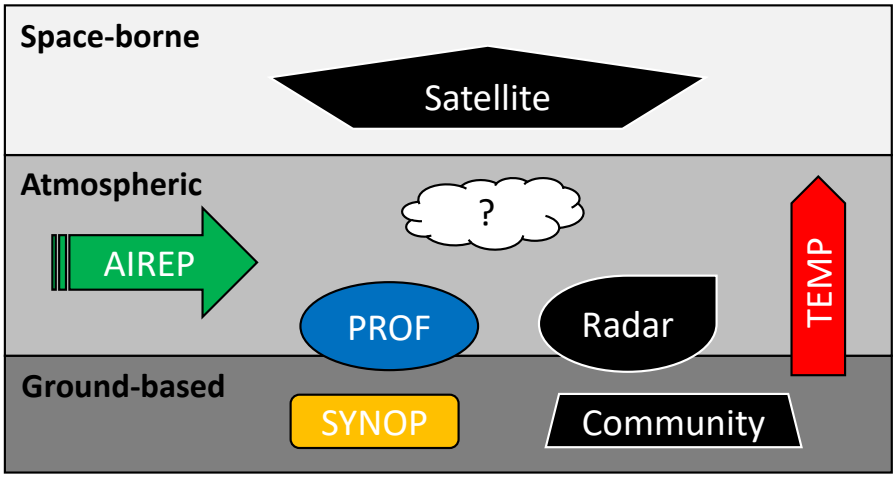


Figure 1.2: *Sketch of the regional observing network of Deutscher Wetterdienst. Conventional and operationally assimilated observations provided by aircraft (AIREP, green), wind profiler (PROF, blue), radiosonde (TEMP, red) and surface stations (SYNOP). Observations that DWD is aiming to assimilate in the future (black): satellites, weather radars, automobiles and smartphones (community data).*

Observing system The global observing system provides several millions of observations per day measured by various instruments (Gustafsson et al., 2018). Remote sensing observations from satellites by far provide the majority of observations. A significant amount of observations are radiances in the infrared, visible and microwave spectral range observed by radiometer or imaging spectrometer. Additionally, global positioning system radio occultation (GPSRO) or wind lidar (Aeolus) measurements provide observations of humidity and wind. However, the operational regional DA system of DWD so far only makes use of radar and conventional observations (Figure 1.2). Radar observations are indirectly assimilated using a latent heat nudging. That means a temperature increment is obtained from the latent heat release that is approximately proportional to the near surface precipitation rate. Conventional observations include direct observations of prog-

nostic model variables such as temperature, wind, humidity, or pressure. Conventional observations can be grouped into four different observational sources: Observations that are measured by aircrafts (AIREP), wind profiler (PROF), radiosondes (TEMP) as well as ships, drifting buoys and surface stations (SYNOP). At the moment, DWD is working on the assimilation of additional observation types for its regional ensemble DA system: solar (Scheck et al., 2018) and thermal (Harnisch et al., 2016) satellite observations, radar reflectivity (Bick et al., 2016) and radar radial velocity.

This thesis presents results for two different regional modeling systems that assimilate conventional observations applying a LETKF: The pre-operational regional forecasting system of DWD (COSMO-KENDA) and the Japanese experimental regional modeling system (SCALE-LETKF). Experiments that are carried out with both systems mainly aim to assess the actual and potential impact of observations in the context of high-resolution NWP. The next section, therefore, discusses several challenges that are crucial for performing convective-scale DA.

1.2 Challenges for convective-scale data assimilation

The chaotic nature and limited predictability of convection pose fundamental challenges in terms of convective-scale DA (Gustafsson et al., 2018). In particular, the higher resolution and low predictability calls for the assimilation of spatially and temporally highly resolved observations. Consequently, major efforts are made to assimilate high-resolution observations as radar reflectivity or cloud-affected satellite observations (Miyoshi et al., 2016b, Harnisch et al., 2016, Scheck et al., 2018, Sawada et al., 2019). However, successfully assimilating such observations requires both accurate parameterizations and observation operators as well as accurate estimates of highly flow-dependent error covariances (Houtekamer and Zhang, 2016). Overall, major challenges can roughly be grouped based on the three components of the data assimilation system: the numerical model, the underlying data assimilation algorithm, and the observational information.

Model-related challenges On convective scales, model-related challenges include random and systematic model errors (Whitaker and Hamill, 2012, Bannister, 2017). Random errors arise from non-linear and stochastic processes such as convection or precipitation and can only be represented using an ensemble. Systematic model errors need to be reduced to enable successful assimilation of, for example, cloud-related observations.

As a consequence, great efforts are made to improve parameterizations that are required to describe clouds and subgrid-scale processes in the model. Overall, given that current models are far from perfect implies model error. Sources for model error, for example, are insufficient parameterizations or the need for discretization. In practice, different relaxation and inflation methods are applied to account for model error (Whitaker and Hamill, 2012, Zeng et al., 2018). Inflating the ensemble perturbations helps to represent the uncertainty in the forecast better and partly accounts for model deficiencies. Nevertheless, existing methods only provide a mitigation to this issue. Consequently, the correct treatment of model error is still a very active research field.

Algorithm-related challenges In convective-scale data assimilation, there are many algorithm-related challenges, such as non-Gaussianity, non-linearity, noise, or mass conservation. Furthermore, under-sampling using a too small ensemble requires methodological solutions (Houtekamer and Zhang (2016)). A too small ensemble, for example, causes spurious correlations estimating error covariances. Error covariances overall determine the weights for combining background and observation as well as decide how information is spread spatially and between model variables. Thus, accurate estimates of error covariances are crucial for optimal initial conditions. To reduce the effect of spurious nonphysical correlations, usually, distance-based localization techniques are applied. However, finding an optimal localization length-scale is an intrinsically difficult task as an optimal scale may differ for different variables, vertical levels, and regions. Therefore, a constant distance-based localization in many cases is inappropriate (Anderson, 2012). Hence, better knowledge of sampling errors and localization could improve many ensemble-based methods and increase the impact of observations. How to deal with sampling error, therefore, will be one topic of this thesis and further discussed in Section 1.4 of the introduction.

Observation-related challenges Observation-related challenges include the optimal design of the observing network, the development of observation operators, and the modeling of observation errors (Gustafsson et al., 2018). Furthermore, quality control and pre-processing of observational information are crucial. Observation errors usually are assumed to be uncorrelated, and in practice, a diagonal observation error covariance matrix is often applied. The observation error for data assimilation is composed of three different error sources: representativity, instrument, and operator errors. A representativity error can appear due to unresolved processes such as missing subgrid-scale variability. The instrument error accounts for deficiencies of the measuring instrument or algorithm. Errors in observation operators, for example, can arise from assumptions that are made to develop sufficiently fast forward operators.
Performing real-time cycling, accurate and fast observation operators are especially necessary to assimilate complex remote sensing observations. Consequently, many studies deal with the development of fast forward operators for radar (Zeng et al., 2016) or satellite observations (Scheck et al., 2018). Overall, improving the observing network and developing strategies on how to design the future observing network are significant challenges for future data assimilation (Gustafsson et al., 2018). Therefore, it is essential to gain knowledge of the actual and potential impact of observations within the forecasting system to answer these questions. Estimating the impact of various observation types will be the primary purpose of this thesis. The next section provides a detailed introduction to this topic.

1.3 Monitoring the impact of observations

The amount and variety of observations that are available for NWP are steadily increasing. Hence, it is essential to understand and monitor the role of various observation types. For that reason, different methods have been developed to estimate the contribution of individual observations or observation types to the reduction of forecast error. This contribution is usually referred to as observation impact.

Observation impact Reliable estimates of observation impact are crucial for many reasons: First, a systematically detrimental impact indicates issues with individual observations or their assimilation. Such knowledge helps to improve the usage of observations and thereby improves the forecast. Second, observation impact can be used to enhance the cost-benefit ratio of the observing system. For instance, if a specific observation type turns out to be very cheap and beneficial, more of these observations could be deployed in the future. Furthermore, observation impact tools are used to estimate the impact of new observing systems within the existing NWP system. Hence, knowledge of observation impact provides essential information for optimizing the observing, data assimilation, and forecasting systems.

Over the last decades several different approaches have been developed to estimate observation impact in variational, ensemble-based or hybrid DA systems. Independent from the DA scheme, observing system experiments (OSEs) such as data denial experiments can be used to assess observation impact (e.g. Bouttier and Kelly (2001); Kelly et al. (2007); Cardinali (2009); Gelaro and Zhu (2009); Benjamin et al. (2010); Weissmann et al. (2011); Harnisch et al. (2011); Lupu et al. (2012); Bauer et al. (2014); Horányi et al. (2015a); Horányi et al. (2015b)). However, OSEs are computationally very demanding as they require to run the full DA system and NWP model for every configuration of interest. In practice, OSEs are therefore only feasible for very few subsets of observations and limited time periods. Another approach is to calculate the influence of observations in the analysis (e.g. Cardinali et al. (2004); Desroziers et al. (2005); Fourrié et al. (2006); Liu et al. (2009); Lupu et al. (2011); Cardinali and Healy (2014); Brousseau et al. (2014)). These are valuable diagnostics for the DA system, but the observation analysis influence can deviate from the forecast impact.

To overcome these limitations, adjoint-based methods for the assessment of observation impact on short-term forecast error emerged with the development of 4DVAR DA schemes. In the beginning of this century, Langland and Baker (2004) first introduced an adjoint-based method to calculate the forecast sensitivity to observation impact (FSOI). Since then, FSOI methods were applied to assess observation impact in several different forecasting systems (e.g. Langland (2005); Cardinali (2009); Gelaro et al. (2010); Weissmann et al. (2012); Lorenc and Marriott (2013); Jung et al. (2013); Holdaway et al. (2014); Privé et al. (2014); Lupu et al. (2015); Janisková and Cardinali (2016); Horányi et al. (2017); Mallick et al. (2017)).

More recently, Liu and Kalnay (2008), Li et al. (2010), and Kalnay et al. (2012) demonstrated the feasibility to approximate the forecast sensitivity to observation impact in an idealized ensemble modeling system (EFSOI). The method is based on the available output of an ensemble data assimilation and forecasting system and thereby avoids the requirement of an adjoint model. Based on that progress, the EFSOI method was then implemented and tested in global and mesoscale modeling systems (e.g., Kunii et al. (2012); Ota et al. (2013)). Recently, Buehner et al. (2018) combined the adjoint and ensemble-based approaches to estimate observation impact in a hybrid DA system.

Whereas FSOI methods are now commonly used in global modeling systems, the assessment of observation impact in high-resolution regional modeling systems received much less attention. Only recently, a few NWP centers (e.g., DWD, UK MetOffice, Meteo France) started developing approaches for FSOI in convective-scale models. Sommer and

Weissmann (2014) first applied the EFSOI method in a convective-scale ensemble system, performed a quantitative evaluation of the method, and showed good agreement of the EFSOI approximation with data denial experiments. However, their experimental period of only 1.5 days was too short for drawing representative conclusions on the impact of various observation types.

Another aspect that received little attention is the choice of the verification metric. Until recently, all FSOI studies used energy-norm verification metrics that are calculated based on differences of a model forecast and subsequent analysis fields. This type of verification may be suitable for synoptic-scale applications as energy-norms combine different forecast aspects. However, analysis fields are not an ideal choice for the verification of short-term forecasts as they may be affected by model biases, and their errors are correlated with those of the forecast. While it seems common knowledge that such a self-verification is potentially dangerous, over 20 studies on FSOI were published without much investigation of the choice of the verification metric and the role of model biases. For the investigation of observation impact in convective-scale modeling systems, the issue of biases gets even more severe as model biases tend to be larger in areas of convective precipitation. Forecasting convective events, however, is usually one of the primary purposes of convective-scale modeling systems. Additionally, total energy seems an inappropriate verification metric as it does not reflect primary forecast quantities as precipitation and wind gusts.

As a first step to overcome these deficiencies, Sommer and Weissmann (2016) reformulated the EFSOI method and introduced an observation-based verification metric. In their study, they used a verification norm based on the departures of all observations assimilated in the subsequent analysis cycle weighted by their errors and investigated the pre-operational regional ensemble DA and modeling system of DWD. Similarly, Cardinali (2018) introduced an observation-based norm for the adjoint-based approximation of FSOI in the global European Centre for Medium-Range Weather Forecasts (ECMWF) modeling system. While the use of observations for the verification of short-term forecasts overall seems advantageous, the limitations of this approach are that the observations are unevenly distributed in space and some observation types may be affected by observational biases. Furthermore, it would be desirable that the verification metric reflects quantities that forecast users are most interested in (e.g., precipitation, wind gusts, surface wind, and temperature or total hours of sunshine). For that reason, this study analyzes EFSOI results using different observation-based verification metrics, investigates the role of potential observational biases and includes independent radar-derived precipitation observations for verification.

Potential impact Another major challenge is the development of observation and data assimilation strategies for high-resolution NWP considering the vast amount of potentially available information in developed countries (Gustafsson et al., 2018): First, NWP centers do not have the human resources to incorporate all these often complex sources of information at the same time. Second, new data selection strategies are required considering the vast amount of unused observations provided by radars, satellites, ground-based profilers, or community observations (e.g., smartphones, webcams, and renewable power production). Last, technological advances have led to novel and much cheaper remote-sensing instruments that could be deployed in the future. Therefore, better knowledge is needed on what observations are most important for convective-scale NWP and where to

set priorities or resources.
This thesis introduces an approach that can be used to develop observing and data assimilation strategies based on ensemble sensitivity analysis (ESA; Ancell and Hakim (2007)). The approach was first introduced by Geiss (2017) and uses spatiotemporal correlations as a proxy for the potential impact of observable quantities. The main focus of this thesis is to assess the potential impact of different quantities on precipitation, which is a primary forecast quantity of convective-scale forecasting systems. The proposed approach strongly depends on the reliability of spatiotemporal correlations derived from an ensemble. For that reason, this thesis also introduces a novel convective-scale 1000-member ensemble simulation. The ensemble is required to achieve reliable impact results based on realistic spatiotemporal correlations.
Ensemble sensitivity analysis itself is an efficient method to explore probabilistic data-sets by investigating linear relations between a forecast metric and initial quantities. ESA has been applied for various synoptic-scale case studies (e.g. Hakim and Torn (2008); Torn and Hakim (2008); Torn and Hakim (2009); Torn (2010); Hanley et al. (2013); Barrett et al. (2015)). Recently, several studies showed that ESA also can provide reasonable results for the analysis of convective-scale simulations (Bednarczyk and Ancell, 2015, Wile et al., 2015, Hill et al., 2016, Berman et al., 2017, Limpert and Houston, 2018). Nevertheless, nearly all previous studies on ESA applied relatively small ensembles for their analysis, which implies sampling error. So far, earlier studies could not quantify potential sampling errors due to spurious correlations as no larger ensemble was available for comparison. These studies attempted to account for under-sampling by applying a confidence test that excludes insignificant correlations (Torn and Hakim, 2008). However, this approach may also exclude small physical correlations, which can lead to systematic effects and is therefore not well-suited for a quantitative analysis of sensitivities. The 1000-member ensemble simulation for the first time enables to quantify the contribution of sampling error for ESA depending on the ensemble size.

1.4 Ensemble prediction and sampling error

Ensemble-based estimates of impact and potential impact crucially rely on accurate spatiotemporal correlations that are estimated with an ensemble. However, operational ensemble sizes only range from about 20 up to 250 members and large ensembles of 250 members are only affordable for very short lead times (Houtekamer et al., 2014, Bannister, 2017, Leutbecher, 2018, Caron and Buehner, 2018, Gustafsson et al., 2018). Given that the number of ensemble members is therefore much smaller than the number of degrees of freedom of the model ($\approx 10^7$) causes several problems: First, the ensemble underestimates variances and does not sample all possible states. Second and more severe, sampling errors significantly affect the estimates of sample correlations leading to spurious correlations. All state of the art ensemble approaches, therefore, have to deal with sampling errors. Hence, it is essential to investigate sampling errors and to find appropriate methods to correct spurious correlations.
Historically, the exploration of the chaotic behavior of weather in the 1960s is the starting point of present-day ensemble prediction (Lorenz, 1963). The European Centre for Medium-Range Weather Forecasts (ECMWF) and the National Centers for Environmen-

tal Prediction (NCEP) produced their first operational ensemble forecasts in the early 1990s (Kalnay, 2003). Nowadays, most operational weather services maintain ensemble systems to gain essential probabilistic information using various ensemble configurations. The applied ensemble size to some extent depends on the primary purpose of the ensemble, for example, estimating forecast uncertainty (variances) or estimating error covariances for data assimilation, but are restricted by the available computing power. That explains why the number of ensemble members is limited, and a trade-off between the required ensemble size and accessible computational resource.

Sampling error correction Mitigation to the issue of under-sampling could provide a statistical sampling error correction (SEC) approach as introduced by Anderson (2012, 2016). The SEC systematically corrects for the over-estimation of correlations due to spurious correlations. It is a look-up table-based approach calculated using a Monte-Carlo technique and therefore, easy to apply. One central part of this thesis explicitly applies the SEC to spatiotemporal correlations to evaluate its potential for ESA, EFSOI, or other ensemble applications. Additionally, the SEC is applied to spatial correlations to evaluate its potential for ensemble or hybrid DA.

Originally, the SEC was designed for covariance localization in ensemble Kalman filter DA. EnKF algorithms or hybrid ensemble approaches rely on accurate estimates of error covariances. Localization techniques usually are applied to reduce the effect of spurious correlations (Houtekamer and Mitchell (1998); van Leeuwen (1999); Houtekamer and Mitchell (2001)). Localization is a physically motivated approach, which cuts off or damps spatial correlations after a certain distance using a tapering function. An example for such a tapering function is the Gaspari-Cohn function (Gaspari and Cohn, 1999) (Figure 1.3a). The effect of the tapering function on spatial correlations is exemplarily shown in Figure 1.3b, which displays the vertical correlation of near-surface temperature with upper-air temperature. Comparing the spatial correlation of a small (40 members) and large ensemble (1000 members) reveals the presence of spurious correlations caused by finite sample size. Applying a perfectly fitted localization function, in this case, can significantly improve the correlation. However, the choice of the localization length scale is an intrinsically difficult task given that physical correlations in the atmosphere can extend horizontally over thousands of kilometers and vertically throughout the troposphere and even into the stratosphere (Caron and Buehner, 2018). Furthermore, different observation types and different model variables presumably require different localization-scales.

Especially vertical localization is a challenging task for data assimilation as several observation types (e.g., passive satellite observations) can be significantly correlated with the full vertical profile of the atmosphere (Lei et al., 2018). Furthermore, satellite observations often cannot be assigned to a single level, which makes a distance-based localization unsuitable. Figure 1.3c shows an example of a vertical correlation of a cloud at 500 hPa with temperature and specific humidity in the tropospheric column. The layer of strong positive correlation indicates that the cloud extends over a broad vertical region and can not be assigned to a single level. Furthermore, the cloud is correlated with the full vertical column due to adiabatic and radiative processes. For example, evaporative cooling at the surface caused by precipitation can lead to strong negative long-range correlations. How to deal with such situations is a challenging task. Hence, to increase the impact of observations,

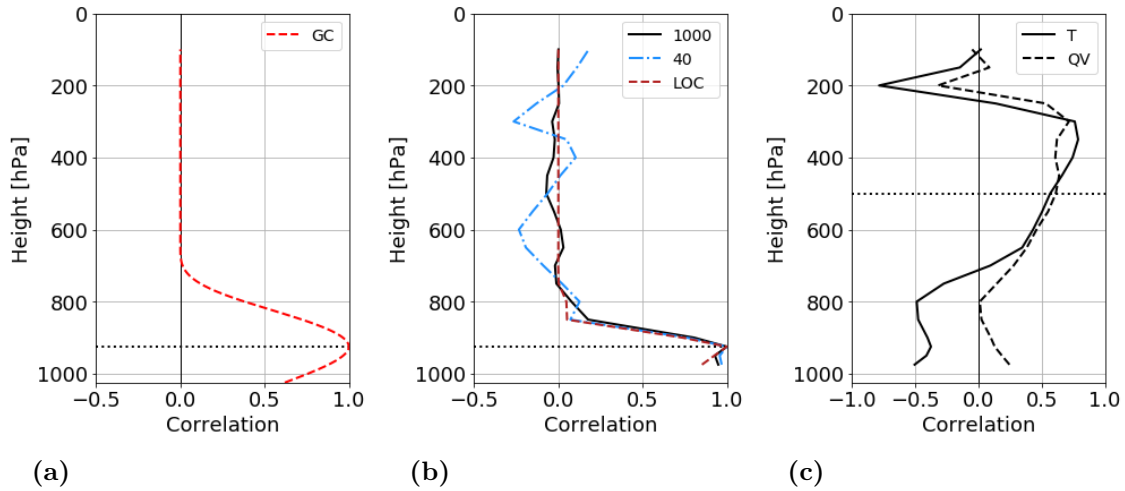


Figure 1.3: *(a) Gaspari-Cohn localization function. (b) Example of a vertical temperature correlation (1000 member, black, solid), spurious correlations (40 member, blue, dash-dotted) and the effect of localization (40 member with localization, red, dashed). (c) Example of long-range vertical correlations of hydrometeors (clouds) at 500 hPa to temperature (T, solid) and specific humidity (QV, dashed) in the entire column. Response level in each sub-figure is indicated by the horizontal dotted line.*

it is crucial to understand error covariances better, to quantify sampling error depending on the ensemble size and to develop improved techniques for sampling error correction and localization.

1000-member ensemble Accurate probabilistic forecasts and therefore large ensembles are explicitly required in convective-scale forecasting, which aims at predicting local weather phenomena and the occurrence of extreme weather events that are often related to atmospheric convection (Gustafsson et al., 2018). Only recently, the latest generation of supercomputers allows performing high-resolution big ensemble forecasts with a frequent update cycling (Miyoshi et al., 2015, 2016a). These advances enable to conduct the convective-scale 1000-member ensemble simulation, which is mandatory to answer the target research questions of this thesis.

The 1000-member ensemble follows upon previous studies that conducted large ensemble simulations using mainly lower-resolution or idealized models. For example, first experiments using a 10240-member global ensemble showed that large ensembles can be applied to learn about sampling error, non-Gaussianity (Miyoshi et al., 2014), or to improve covariance localization (Kondo and Miyoshi, 2016). Furthermore, a study by Jacques and Zawadzki (2015) once computed 1000 convective-scale forecasts to investigate background errors for radar data assimilation. All these studies highlight the potential of large ensemble simulations to investigate error correlations and sampling error.

The present 1000-member ensemble simulation mainly builds upon a setup introduced by Geiss (2017). The main difference in the setups originates from the use of improved ensemble boundary conditions (BC) that lead to more realistic spread properties. The applied 1000-member ensemble simulation is the first of its kind and, therefore, requires a basic evaluation. Consequently, this thesis compares the large ensemble to radar precipitation

observations and the pre-operational COSMO-KENDA 40-member ensemble.
Overall, the large high-resolution ensemble for the first time provides reliable estimates of correlations that serve as truth to quantify sampling errors that would be made with smaller subsets of the full ensemble. The 1000-member ensemble, furthermore, is required to estimate the potential impact of observations for convective-scale DA. Given that the large ensemble is hardly affected by sampling errors enables more detailed evaluation of sampling error and correction methods as it could be achieved beforehand.

1.5 Research questions and outline

This dissertation consist of three parts (Chapter 4, 5, and 6) that intend to estimate the impact of observations in convective-scale NWP. For that purpose, two complementary ensemble-based approaches are applied to assess the actual and potential impact of different observations. The third part evaluates a sampling error correction, as both approaches are ensemble-based and therefore have to deal with sampling errors. Chapter 3 introduces the unique convective-scale 1000-member ensemble simulation that is required for two studies. Figure 1.4 illustrates a schematic overview of the overlap of the different parts of this thesis.

Scientific questions The key scientific questions addressed in this thesis are:

1. How large is the impact of observations in the convective-scale forecasting system of Deutscher Wetterdienst and how strongly does the impact depend on the choice of the verification metric?
(*Observation impact*)

2. How can we estimate the potential impact of observable quantities for convective-scale data assimilation?
(*Potential impact*)

3. Can a statistical sampling error correction approach improve spatial and spatiotemporal correlations that are required for several ensemble applications?
(*Sampling error correction*)

The first part of this thesis evaluates the impact of about 3 million conventional observations in the regional operational forecasting system of Deutscher Wetterdienst. This is the first study assessing the observation impact in a convective-scale regional DA system over an extended period of six weeks. The observation impact is calculated by applying an EFSOI method and using different observation-based verification norms. EFSOI measures the observation impact, which is determined by the configuration of the data assimilation system. Hence, it is a powerful tool for monitoring the KENDA system. Particular emphasis is given to the question of how the verification norm or existing biases affect the estimated observation impact.
The second part introduces an approach that can analyze the potential impact of observable quantities. The proposed approach is very efficient, applicable to a large amount of data and only requires an ensemble forecast. Therefore, it could be applied in any ensemble forecasting system and help to develop improved data assimilation and observing strategies

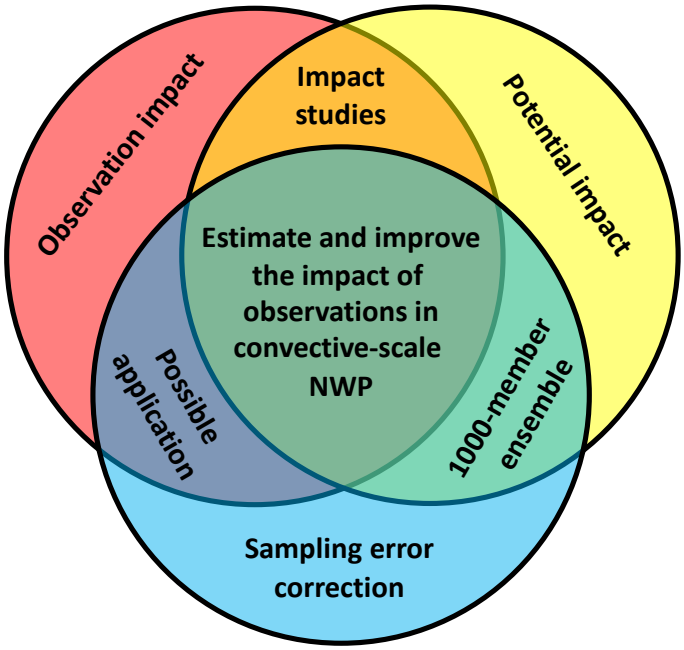


Figure 1.4: *Illustration of the overlap and connection between different parts of this thesis.*

in the future. The potential impact is analyzed using the spatiotemporal correlations as a proxy for the potential impact. The present study mainly focuses on the scale dependence of the approach and the sensitivity of the approach to the ensemble size. Overall, the approach estimates the relative potential impact assuming a perfect data assimilation system and, therefore, provides a different estimate of the impact and complements the EFSOI.
The third part of this thesis evaluates a sampling error correction approach that can help to reduce sampling error performing an ensemble-based impact assessment. The sampling error correction could be applied in various ensemble applications that incorporate sample correlations. Initially, the SEC was introduced as an alternative to constant distance-based localization techniques. This thesis evaluates the SEC for both spatiotemporal and spatial correlations. Reliable spatiotemporal correlations are required for the estimation of observation impacts as well as for ensemble sensitivity analysis. Spatial correlations are needed in data assimilation to spread observational information in the analysis state.
In addition to the three main parts, this thesis introduces and evaluates a novel convective-scale 1000-member ensemble simulation that is applied to quantify sampling error in convective-scale NWP (Chapter 3). The 1000-member ensemble simulation delivers reliable spatial and spatiotemporal correlations. Those correlations are explicitly needed to estimate the potential impact of observations in Chapter 5 and to evaluate the sampling error correction in Chapter 6.

Outline The outline of this thesis is as follows: **Chapter 2** gives a general introduction to ensemble Kalman filter data assimilation and different observation impact methods. Furthermore, ensemble sensitivity analysis and the approach for estimating the potential impact of observable quantities are introduced. As the last step, the evaluated sampling error correction is described. **Chapter 3** provides details on the operational convective-scale

ensemble DA system (KENDA/ COSMO-DE) of Deutscher Wetterdienst and the 1000-member ensemble simulation, which is based on the SCALE-LETKF system. Initially, 1000-member ensemble forecasts are compared to observations and forecasts performed with the COSMO-KENDA system. Afterwards, the value of the large ensemble simulation to quantify sampling error for variances and covariances is discussed. **Chapter 4** is the first main part of this thesis and analyzes the observation impact for a 6-week high-impact weather period in the summer of 2016. The observation impact is evaluated using different observation-based verification norms. **Chapter 5** presents an approach for estimating the potential impact of different observable quantities. In particular, the sensitivity of the approach to the ensemble size and the scale dependence of the potential impact are discussed. **Chapter 6** examines a statistical sampling error correction, which could mitigate sampling error during the impact assessment. This study evaluates the SEC separately for spatial and spatiotemporal correlations. A summary with conclusions follows in **Chapter 7**.

Chapter 2

Theory and methods

Chapter 2 presents an overview of methods that are applied to answer the target research questions. Additionally, this chapter provides the reader with essential theoretical knowledge to allow a better understanding of performed experiments. Section 2.1 introduces the basic principles of data assimilation and motivates the use of LETKF in convective-scale NWP. Section 2.2 presents different approaches that are available for estimating observation impact. In addition, the derivation of the ensemble-based observation impact approach is presented. This approach is applied to evaluate a regional forecasting system using observations for verification. In this context, the advantages and disadvantages of different verification metrics are discussed. Section 2.3 introduces and refines the approach that is used to estimate the potential impact of observable quantities. The approach is based on ESA and uses the sensitivity of the forecast to initial conditions as a proxy for the potential impact. Finally, two different methods are presented that are evaluated for reducing sampling error (Section 2.4): A statistical sampling error correction and a confidence test.

2.1 Sequential data assimilation

As already introduced in Section 1.1, numerical weather prediction is an initial value problem. Data assimilation systems estimate the initial state by minimizing a quadratic cost-function (Eq. 1.1). The initial state combines atmospheric observations $\mathbf{y}$ with a background state $\mathbf{x}_b$ weighting both sources of information with their errors. Sequential data assimilation methods obtain the initial state (analysis) $\mathbf{x}_a$ by adding a correction (increment) to the background state (first guess). The increment is determined by the difference between background and observation in observation space (innovation) multiplied with the optimal weight matrix $\mathbf{K}$:

$$\mathbf{x_a} = \mathbf{x_b} + \mathbf{K}(\mathbf{y} - \mathbf{H}\mathbf{x_b}). \tag{2.1}$$

The optimal weight is given by the Kalman gain matrix $\mathbf{K}$

$$\mathbf{K} = \mathbf{P_b}\mathbf{H}^T \left(\mathbf{H}\mathbf{P_b}\mathbf{H}^T + \mathbf{R}\right)^{-1}, \tag{2.2}$$

where $\mathbf{P_b}$ is the background error covariance matrix, $\mathbf{R}$ is the observation error covariance matrix and $\mathbf{H}$ the linearized forward operator that transforms from model to observation

space. The subscript a stands for the analysis and the subscript b for the background state, respectively. The observation error covariance matrix $\mathbf{R}$ contains variances and covariances of measurements and representativity errors. In practice, usually a diagonal $\mathbf{R}$ matrix is applied assuming that observation errors are uncorrelated. The Kalman gain $\mathbf{K}$ depends on the ratio of the background and observation error covariance matrices. The optimal weight minimizes the cost-function (Eq. 1.1) or equivalently the mean analysis error over all grid points. According to Eq. 2.2, a small observation error compared to the error of the model background results in a large increment and the analysis will be close to the observation.
A data assimilation cycle usually is composed of two steps: An *analysis step* that generates a new analysis and a *forecast step* in which the forecast is propagated forward in time. The propagation of the state $\mathbf{x_a}$ from time $t-1$ to t is done using a full non-linear numerical model $\mathcal{M}$

$$\mathbf{x_b^t} = \mathcal{M}^{t-1,t}(\mathbf{x_a^{t-1}}). \tag{2.3}$$

The analysis error covariance matrix $\mathbf{P_a}$ can be computed using the Kalman gain, the background error covariance matrix, and the observation operator

$$\mathbf{P_a} = (I - \mathbf{KH})\mathbf{P_b}. \tag{2.4}$$

Table 2.1 provides a summary of variables definitions and corresponding dimensions for the entire Section 2.1.

Name	Variable	Dimension
Model state vector	$\mathbf{x_{a/b}}$	$n \times 1$
Observation state vector	$\mathbf{y}$	$m \times 1$
Background error covariance matrix	$\mathbf{P_b}$	$n \times n$
Observation error covariance matrix	$\mathbf{R}$	$m \times m$
Linear observation operator	$\mathbf{H}$	$m \times n$
Non-linear observation operator	H	$m \times n$
Kalman gain matrix	$\mathbf{K}$	$n \times m$
Ensemble perturbations in model space	$\mathbf{X_{a/b}}$	$n \times N$
Ensemble perturbations in ensemble space	$\mathbf{Y_{a/b}}$	$n \times N$
Analysis error covariance matrix	$\mathbf{P_a}$	$n \times n$
Analysis error covariance matrix in ensemble space	$\tilde{\mathbf{P}}_\mathbf{a}$	$N \times N$
Weight vector	$\mathbf{w_a}$	$N \times 1$
Weight matrix	$\mathbf{W_a}$	$N \times N$

Table 2.1: *Variables and their dimensions: Number of ensemble members N, number of observations m and number of state variables n.*

Frequently applied sequential data assimilation schemes in NWP are variational, and Ensemble Kalman Filter (EnKF) type approaches. Variational approaches (e.g., 3DVAR) perform the minimization of the cost function iteratively using, for example, a conjugate gradient method. Historically, the application of variational methods enabled the direct assimilation of satellite radiances, which led to a massive improvement in forecast skill. In general, 3DVAR assumes all observational information to be observed at analysis time.

On convective scales, this assumption can be unsuitable considering rapidly changing weather situations. A 4DVAR solves this problem by assimilating observational information at the correct time. Furthermore, 4DVAR analysis fields are more balanced due to the integration of the model over the assimilation window. Overall, using variational approaches is associated with several challenges. For example, applying 4DVAR, a tangent linear adjoint model is required, which is particularly challenging for convection-permitting models.
One main difference between variational and ensemble DA schemes exists in the way they obtain the background error covariance matrix $\mathbf{P_b}$. For example, maintaining a 3DVAR approach can be sub-optimal as it applies a constant background error covariance matrix (Kalnay, 2003). For 3DVAR, the background error is usually estimated using a forecast climatology and physical constraints such as the geostrophic or hydrostatic balance. However, balance relations are especially limited for convective-scale applications making the estimation of the background error demanding. A significant advantage of ensemble algorithms is that the estimated the background error covariance matrix is based on an ensemble of forecasts. This feature makes the background error flow dependent, which is crucial for convective-scale NWP. For that reason, many NWP centers apply hybrid and EnKF approaches for convective-scale DA.

Ensemble Kalman Filter An ensemble Kalman filter uses an ensemble to calculate the uncertainty of the background and analysis error covariance (Evensen, 1994). The ensemble of N forecasts is used to estimate the background error covariance matrix $\mathbf{P_b}$. This means, the background error covariance matrix is calculated as sample covariance using the ensemble deviations of each member from the ensemble mean state

$$\mathbf{P_b} = \frac{1}{N-1}\sum_{n=1}^{N}(\mathbf{x_b^n} - \overline{\mathbf{x_b}})(\mathbf{x_b^n} - \overline{\mathbf{x_b}}))^T \tag{2.5}$$

$$= \frac{1}{N-1}\mathbf{X_b}(\mathbf{X_b})^T \tag{2.6}$$

where

$$\overline{\mathbf{x}_b} = \frac{1}{N}\sum_{n=1}^{N}\mathbf{x_b^n} \tag{2.7}$$

is the ensemble mean background state. According to Eq. 2.6, the uncertainty of a forecast, is quantified by the ensemble with $(N-1)$ degrees of freedom. The performance of an EnKF, therefore, depends on the available ensemble size. To improve the filter performance usually inflation and localization techniques are applied (Whitaker and Hamill, 2012). Inflation methods increase ensemble perturbations and account for model error. Localization reduces sampling errors by damping spurious correlations.

ETKF In general, the problem can be reformulated by solving the problem in ensemble space. This modification yields the Ensemble Transform Kalman Filter (ETKF) (Bishop et al., 2001). *Transform* means that the minimization of the cost-function is performed in an N-dimensional subspace S for a vector $\mathbf{w}$ where $\mathbf{X_b}$ is the linear transform onto the sub-

space S (Rhodin et al., 2013). In other words, the analysis is obtained by re-weighting the information from each ensemble member based on departures from observations. However, this also means that the solution is limited by the degrees of freedom in the ensemble. For an ETKF, the optimal weights are calculated in ensemble space S and not in observation space, which reduces the dimension of the problem significantly. Furthermore, it makes the ETKF algorithm cheap and allows the filter to perform stably in complex atmospheric applications. The next paragraph provides a summary on the ETKF algorithm applying a similar notation as used by Rhodin et al. (2013).
In the ETKF, the analysis mean $\overline{\mathbf{x_a}}$ is a linear combination of the weighted background ensemble member

$$\overline{\mathbf{x_a}} = \overline{\mathbf{x_b}} + \mathbf{X_b}\overline{\mathbf{w_a}}. \tag{2.8}$$

where $\mathbf{X_b w_a}$ represents the linear combination of ensemble perturbations. Note that the sequential formulation in Eq. 2.8 is equivalent to the formulation in Eq. 2.1. Here, the optimal weight $\overline{\mathbf{w_a}}$ is given by

$$\overline{\mathbf{w_a}} = \tilde{\mathbf{P}}_\mathbf{a}(\mathbf{Y_b})^T\mathbf{R}^{-1}(\mathbf{y} - \mathbf{H}\overline{\mathbf{x_b}}) \tag{2.9}$$

with

$$\tilde{\mathbf{P}}_\mathbf{a} = \left[(N-1)I + (\mathbf{Y_b})^T\mathbf{R}^{-1}\mathbf{Y_b}\right]^{-1} \tag{2.10}$$
$$= (N-1)^{-1}\mathbf{W_a}(\mathbf{W_a})^T. \tag{2.11}$$

The observation operator $\mathbf{H}$ again is assumed to be linear mapping from model to observation space. The analysis error covariance matrix in model space $\mathbf{P_a}$ is given by the re-transformation using the background perturbation matrix $\mathbf{X_b}$ and the analysis error covariance matrix in ensemble space $\tilde{\mathbf{P}}_\mathbf{a}$:

$$\mathbf{P_a} = \mathbf{X_b}\tilde{\mathbf{P}}_\mathbf{a}\mathbf{X_b}^T. \tag{2.12}$$

The analysis ensemble perturbations $\mathbf{X_a}$ can be calculated using a symmetry square root method and the weight matrix $\mathbf{W_a}$

$$\mathbf{X_a} = \mathbf{X_b}\mathbf{W_a} \tag{2.13}$$

with

$$\mathbf{W_a} = \left[(N-1)\tilde{\mathbf{P}}_\mathbf{a}\right]^{\frac{1}{2}}. \tag{2.14}$$

A specific implementation of the ETKF is the Local Ensemble Transform Kalman Filter (LETKF; Hunt et al. (2007)) that often is used for atmospheric data assimilation due to its computational efficiency.

LETKF For regional data assimilation, Deutscher Wetterdienst maintains the Kilometer Scale Ensemble Data Assimilation (KENDA; Schraff et al. (2016)) system that is based on a LETKF. A detailed description of the KENDA system is available in the DA system

documentation of DWD (Rhodin et al., 2013). The term *Local* describes one main difference compared to an ETKF. Local means that the analysis is performed locally and for each grid point separately. This localization increases the number of degrees of freedom of which the analysis is composed and therefore significantly improves the initial conditions. In KENDA, the analysis is obtained for a coarser grid and interpolated afterward to a finer grid, which further reduces the computational cost. Overall, the LETKF is easy to parallelize as a local analysis is computed separately for each grid point to obtain a global analysis state.

In atmospheric applications, the number of observations m as well as the number of state variables in the model n is much larger than the degrees of freedom provided by the ensemble ($N << m << n$). This under-sampling implies several challenges. Principally, a limited ensemble size leads to sampling error. Sampling error, to some extent, can be addressed applying localization and inflation techniques. Inflation increases the variance of the ensemble to account for model error. Localization reduces spurious error correlations between distant points (Houtekamer and Zhang, 2016). KENDA applies localization by calculating the weights $\mathbf{w_a}$ separately for each grid point of the analysis. In practice, the number of observations that can affect a grid point is also limited using an observation space localization. This **R**-localization reduces the number of assimilated observations affecting the local state estimate to observations in the vicinity of the grid point. Usually, localization is done using a distance-based damping function such as the Gaspari-Cohn function (Gaspari and Cohn, 1999).

2.2 Estimating observation impact

General concepts

Observing system experiment There are several ways to assess the impact of observations in a data assimilation system. For every NWP system, the impact of observations can be analyzed by performing observing system experiments (OSEs) (Bouttier and Kelly, 2001, Kelly et al., 2007). An OSE can be both a data addition or data denial experiment. Performing a data denial experiment the impact of an observational subset $\mathbf{d}'$ can be estimated by removing observations $\mathbf{d}'$ from the full set of observations $\mathbf{d}$ and repeating the analysis and forecast step. The observation impact can then be quantified using a scalar forecast metric J

$$J(\mathbf{d}') = \left|\mathbf{e}_f^{\mathbf{d}}\right|^2 - \left|\mathbf{e}_f^{\mathbf{d}-\mathbf{d}'}\right|^2 \tag{2.15}$$

Here, $\mathbf{d} - \mathbf{d}'$ defines the full set of observations $\mathbf{d}$ leaving out the subset $\mathbf{d}'$. $\mathbf{e}$ is the forecast error of forecast f.

In NWP, only OSEs can be used to measure the impact on long-range forecasts as data denial experiments apply the full non-linear model in each configuration. Overall, performing OSEs usually is an expensive task as the forecast system needs to run in every configuration of interest. This fact led to the development of different approaches that can be used to approximate the impact of observations.

In Chapter 4, an ensemble-based approach is applied to estimate the observation impact.

This approach is a reformulation of the forecast sensitivity to observation impact (FSOI) (Langland and Baker, 2004). The original FSOI approach combines different components (Figure 2.1). Each of these components, to some degree, can be used on their own to examine the impact of observations. All these components have in common that they analyze sensitivities in the data assimilation and forecasting system. A sensitivity gradient, for example, can be computed using the adjoint of the forecast model or data assimilation scheme. Similarly, an ensemble can be used to compute sensitivity gradients.

The observation impact usually is measured as the reduction in forecast error. Based on the definition of observation impact in Eq. 2.15 a detrimental impact of observations $\mathbf{d}$ corresponds to a positive observation impact $J(\mathbf{d}')$. The same definition is applied for the ensemble-based observation impact experiments performed in this thesis as well as for the majority of impact studies (e.g., Langland and Baker (2004), Kalnay et al. (2012), Sommer and Weissmann (2014)). That means the goal is a negative observation impact as it indicates a reduction in forecast error – a forecast improvement. Subsequently, the components that make up the FSOI are presented.

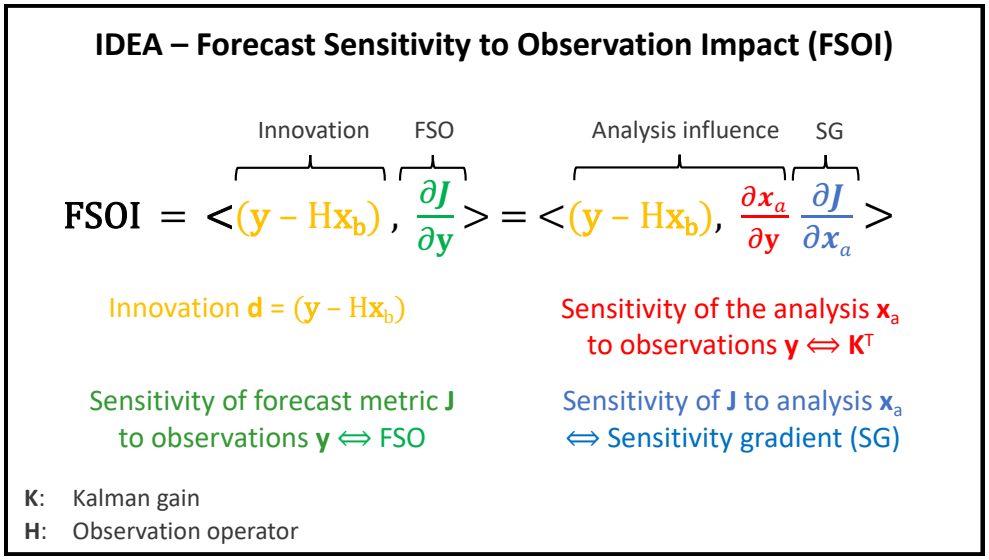


Figure 2.1: *Components of the FSOI approach for estimating the impact of observations.*

Forecast sensitivity to observations At the beginning of this century, Baker and Daley (2000) introduced the forecast sensitivity to observations (FSO) to estimate the impact of observations in an adjoint-based DA system. According to Baker and Daley (2000), the sensitivity of a scalar function J to observations $\mathbf{y}$ can be expressed by

$$\frac{\partial J}{\partial \mathbf{y}} = \frac{\partial J}{\partial \mathbf{x}_a}\frac{\partial \mathbf{x}_a}{\partial \mathbf{y}} \tag{2.16}$$

using the derivative chain rule. The forecast metric J can be any scalar function of interest. As can be seen from Eq. 2.16, the FSO is composed of two sensitivity gradients that each on its own can be used as impact diagnostic.

Sensitivity gradient The sensitivity gradient is the sensitivity of a forecast response function J with respect to the initial conditions $\mathbf{x}_a$. The sensitivity gradient can be expressed using the adjoint of the tangent linear model $(M)^T$ (Rabier et al., 1996)

$$\frac{\partial J}{\partial \mathbf{x}_a} = (M)^T \frac{\partial J}{\partial \mathbf{x}_f}. \tag{2.17}$$

Another possibility to approximate the sensitivity gradient is to apply an ensemble (see Section 2.3). In general, the aim is to find regions with a large sensitivity as a change in the initial conditions in these regions will likely have a large impact on the forecast error. Sensitivity gradients are often used for observation targeting as they indicate sensitive regions in the initial conditions (Ancell and Hakim, 2007, Majumdar et al., 2011).

The second component of the FSO is the observational influence in the analysis. The sensitivity of the analysis with respect to observations $\partial \mathbf{x}_a / \partial \mathbf{y}$ is given by the adjoint of the data assimilation scheme, which corresponds to the transpose of the Kalman gain matrix

$$\frac{\partial \mathbf{x}_a}{\partial \mathbf{y}} = \mathbf{K}^T \tag{2.18}$$

with

$$\mathbf{K}^T = \left(\mathbf{H}\mathbf{P}_\mathbf{b}\mathbf{H}^T + \mathbf{R}\right)^{-1} \mathbf{H}\mathbf{P}_\mathbf{b}. \tag{2.19}$$

Compared to the adjoint of the forecast model, the adjoint of the DA scheme is easier to obtain as it only requires a rearrangement of matrices that are already available from the DA cycling.

Analysis influence An often used measure to quantify the impact of observations in the analysis is the analysis influence (AI) (Cardinali et al., 2004, Liu et al., 2009). The analysis influence combines the observational influence in the analysis $\partial \mathbf{x}_a / \partial \mathbf{y}$ with the first guess departure

$$AI = \frac{\partial \mathbf{x}_a}{\partial \mathbf{y}} \mathbf{d} = \frac{\partial \mathbf{x}_a}{\partial \mathbf{y}} \left(\mathbf{y} - \mathbf{H}\mathbf{x}_\mathbf{b}\right). \tag{2.20}$$

The analysis influence (see also Figure 2.1) is a valuable diagnostic of the DA system as it quantifies the influence of observations in the analysis step. However, the observation impact on the forecast may be different as a large influence on the analysis does not necessarily coincide with a reduction in forecast error.

Forecast sensitivity to observation impact Using the components presented above, we can obtain the formulation for the FSOI (Figure 2.1). The FSOI combines the analysis influence with the sensitivity gradient to quantify the reduction in forecast error provided by each observation. Following Langland and Baker (2004) the reduction in forecast error

$J(\mathbf{d}')$ is given by the inner product of the innovation and the FSO

$$J(\mathbf{d}') \approx \left\langle (\mathbf{y} - \mathbf{H}\mathbf{x}_b), \mathbf{K}^T (\frac{\partial e_f}{\partial \mathbf{x}_a} + \frac{\partial e_g}{\partial \mathbf{x}_b}) \right\rangle . \tag{2.21}$$

Note: The formulation of Langland and Baker (2004) (Eq. 2.21) translates to the formulation in Figure 2.1 using the subsequent relation (Langland and Baker, 2004, Cardinali, 2009)

$$\frac{\partial J}{\partial \mathbf{x}_a} = \frac{\partial e_f}{\partial \mathbf{x}_a} + \frac{\partial e_g}{\partial \mathbf{x}_b}, \tag{2.22}$$

where the sensitivity gradient is expressed by the sum of two sensitivity gradients from two different forecast trajectories.
According to Eq. 2.21, the implementation of adjoint-based FSOI approach by Langland and Baker (2004) requires the calculation of two trajectories f and g (Figure 2.2). This means we need the forecast error e_f of the forecast trajectory f and the error e_g of the background trajectory g that excludes all observations. The majority of FSOI studies computes the forecast error with respect to a verifying analysis $\mathbf{x}_v$ based on an energy norm (Ehrendorfer et al., 1999, Langland and Baker, 2004, Cardinali, 2009, Gelaro et al., 2010). The subscript v indicates the verification time. Usually, a quadratic measure of the forecast error is chosen such that

$$\mathbf{e}_f = \langle (\mathbf{x}_f - \mathbf{x}_\mathrm{v}), \mathbf{C}(\mathbf{x}_f - \mathbf{x}_\mathrm{v}) \rangle , \tag{2.23}$$

where the matrix $\mathbf{C}$ is a symmetric weight matrix in model space that can be used to restrict the energy norm to regions of interest (Park and Xu, 2009). Most frequently, a dry-kinetic energy norm is applied, which uses the zonal wind component $\mathbf{u}$ and meridional wind component $\mathbf{v}$ to calculate a kinetic energy difference (e.g., $\frac{1}{2}(\mathbf{u}_f - \mathbf{u}_v)^2$).
Overall, observation impact diagnostics are sensitive to the choice of the verification norm as it is a critical element of the diagnostic. Depending on the verification norm, FSOI evaluates different forecast aspects. Recently, Janisková and Cardinali (2016) highlighted the sensitivity of the adjoint-based FOSI to the choice of the verification norm comparing dry and moist energy norms. So far, only a few studies avoided energy norms by using observations for verification. Observations seem to be the better choice for convective-scale applications, where an energy norm might not properly reflect relevant forecast aspects (e.g., precipitation). Sommer and Weissmann (2016) first applied observations for verification using the ensemble-based FSOI. Recently, Cardinali (2018) introduced an observation-based verification norm in the context of the adjoint-based FSOI.

Ensemble-based FSOI Based on Langland and Baker (2004), Kalnay et al. (2012) introduced an approach to estimate the observation impact in an ensemble data assimilation system. Computing the ensemble forecast sensitivity to observation impact (EFSOI), sensitivity gradients are estimated with the ensemble instead of the adjoint. The EFSOI approach generally is similar to the one from Langland and Baker (2004) and requires the computation of two forecast trajectories. Both approaches provide the observation impact per observation and are significantly cheaper than performing OSEs. Nevertheless, as men-

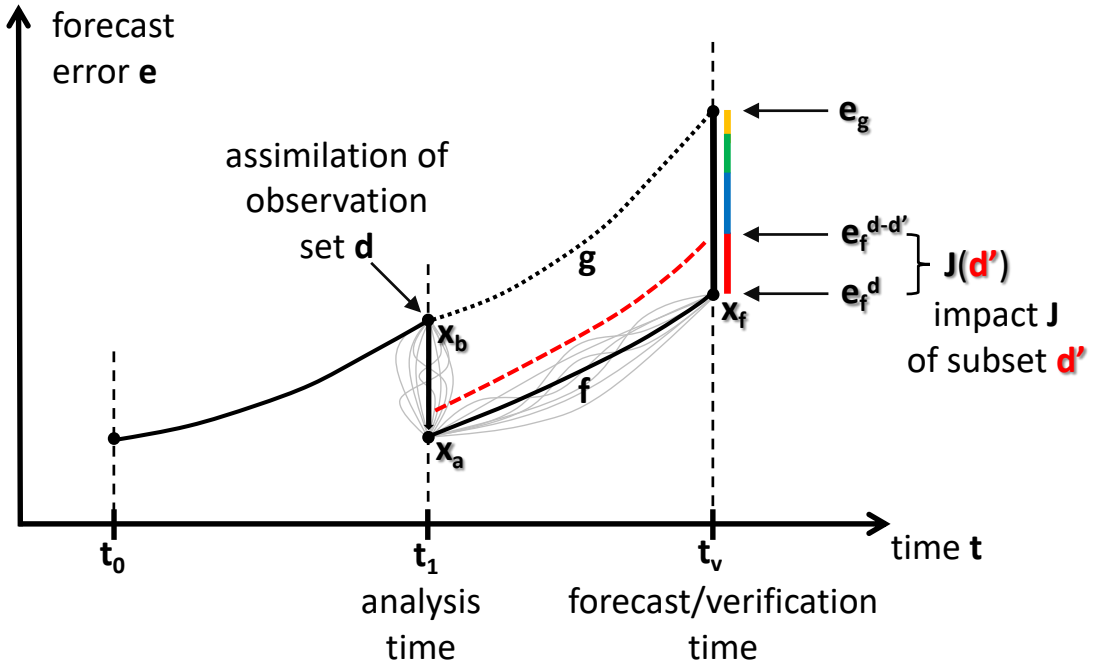


Figure 2.2: *Illustration of the forecast sensitivity to observation impact (FSOI) approach. FSOI aims to estimate the contribution (colored bars) of observational subsets* $\mathbf{d}'$ *to the reduction in forecast error* $\mathbf{e}$*. Usually, FSOI requires the computation of two forecast trajectories* f *and* g*. Gray trajectories indicate the ensemble of forecasts that is required for the ensemble-based FSOI computation. The approximation of EFSOI by Sommer and Weissmann (2016) omits trajectory* g*.*

tioned earlier, both approaches are limited by linearity constraints. Therefore, the forecast lead time usually does not exceed 24 h on synoptic scales or 6 h on convective-scales.

Sommer and Weissmann (2014) first applied the EFSOI approach of Kalnay et al. (2012) in a regional modeling system. Their study compared the EFSOI results to data denial experiments and showed that the approach provides reasonable estimates of observation impact for short-range forecasts using KENDA. Nevertheless, they highlighted that a model-state-based verification norm is especially sub-optimal for verifying short-range forecasts. The verifying analysis can be strongly correlated to the initial analysis, which calls for an independent verification norm. Sommer and Weissmann (2016) reformulated the EFSOI approach of Kalnay et al. (2012) to be able to use observations for verification. Additionally, the reformulation is cheaper as it does not require to compute the trajectory g (Figure 2.2). Sommer and Weissmann (2016) initially applied conventional observations for verification in a short experimental period. This thesis extends the approach using independent remote sensing observations and compares the impact of different verification norms in a 6-week summer period. Subsequently, the EFSOI approach of Sommer and Weissmann (2016), as well as different observation-based verification norms, are discussed in detail.

Implementation in KENDA

Derivation of the EFSOI equation According to the definition of Kalnay et al. (2012), the impact J of observations $\mathbf{d}'$ is given by the squared forecast error difference (Eq. 2.15). For small subset of $\mathbf{d}'$, J can be approximated by the linearization around 0 using a Taylor expansion (Sommer and Weissmann, 2016)

$$J(\mathbf{d}') = J(\mathbf{0}) + \left.\frac{d}{d\mathbf{d}'}\right|_{\mathbf{d}'=\mathbf{0}} J(\mathbf{d}')\mathbf{d}' + \mathcal{O}\left(|\mathbf{d}'|^2\right). \tag{2.24}$$

The first term in Eq. 2.24 vanishes and the last term is neglected. Using Eq. 2.15 gives us

$$J(\mathbf{d}') \approx \left.\frac{d}{d\mathbf{d}'}\right|_{\mathbf{d}'=\mathbf{0}} (\left|\mathbf{e}_f^{\mathbf{d}}\right|^2 - \left|\mathbf{e}_f^{\mathbf{d}-\mathbf{d}'}\right|^2)\mathbf{d}'. \tag{2.25}$$

The derivation of $\left|\mathbf{e}_f^{\mathbf{d}}\right|^2$ with respect to $\mathbf{d}'$ is zero. Performing the first part of the derivation of $\left|\mathbf{e}_f^{\mathbf{d}-\mathbf{d}'}\right|^2$ yields

$$J(\mathbf{d}') \approx -2\mathbf{e}_f^{\mathbf{d}} \left(\left.\frac{d}{d\mathbf{d}'}\right|_{\mathbf{d}'=0} \mathbf{e}_f^{\mathbf{d}-\mathbf{d}'} \right) \mathbf{d}'. \tag{2.26}$$

Here, the forecast error is defined relative to the verifying observations $\mathbf{y}_{\text{veri}}$

$$\mathbf{e}_f^{\mathbf{d}} = \overline{\mathbf{H}(\mathbf{x}_f^{\mathbf{d}})} - \mathbf{y}_{\text{veri}}, \tag{2.27}$$

where $\mathbf{H}$ stands for the corresponding observation operator and the overbar for the ensemble mean.

Before we obtain the final equation for the observation impact we need to solve the remaining derivative in Eq. 2.26. This can be done substituting $\mathbf{e}_f^{\mathbf{d}-\mathbf{d}'}$ using the definition from Eq. 2.27 and the following relationships from Kalnay et al. (2012):

$$\mathcal{M}\mathbf{X}_a^{\mathbf{d}} \approx \mathbf{X}_f^{\mathbf{d}} \tag{2.28}$$

$$\mathbf{H}\mathbf{X}_a^{\mathbf{d}} = \mathbf{Y}_a^{\mathbf{d}} \tag{2.29}$$

$$\left.\frac{d}{d\mathbf{d}'}\right|_{\mathbf{d}'=0} \overline{\mathbf{X}_f^{\mathbf{d}-\mathbf{d}'}} = -\mathcal{M}\mathbf{K}. \tag{2.30}$$

Consequently,

$$\begin{aligned}
\left.\frac{d}{d\mathbf{d}'}\right|_{\mathbf{d}'=0} \mathbf{e}_f^{\mathbf{d}-\mathbf{d}'} &= \left.\frac{d}{d\mathbf{d}'}\right|_{\mathbf{d}'=0} \left(\overline{\mathbf{H}(\mathbf{X}_f^{\mathbf{d}-\mathbf{d}'})} - \mathbf{y}_{\text{veri}}\right) \\
&= \left.\frac{d}{d\mathbf{d}'}\right|_{\mathbf{d}'=0} \overline{\mathbf{H}(\mathbf{X}_f^{\mathbf{d}-\mathbf{d}'})} \\
&= -\mathbf{H}\mathcal{M}\mathbf{K}
\end{aligned}$$

Using the relationships from above including the Kalman gain $\mathbf{K}$ (Eq. 2.2) yields

$$\left.\frac{d}{d\mathbf{d}'}\right|_{\mathbf{d}'=0} \mathbf{e}_f^{\mathbf{d}-\mathbf{d}'} = -\frac{1}{N-1}\mathbf{H}\mathcal{M}\mathbf{X}_a^{\mathbf{d}}(\mathbf{X}_a^{\mathbf{d}})^T\mathbf{H}^T\mathbf{R}^{-1} \tag{2.31}$$

$$= -\frac{1}{N-1}\mathbf{Y}_f^{\mathbf{d}}(\mathbf{Y}_a^{\mathbf{d}})^T\mathbf{R}^{-1}. \tag{2.32}$$

Finally, we can combine Eq. 2.26 with Eq. 2.32 to obtain a formula for $J(\mathbf{d}')$. Reorganized, this gives us the ensemble-based approximation for the FSOI in an LETKF:

$$J(\mathbf{d}') \approx \frac{2}{N-1}\mathbf{e}_f^{\mathbf{d}} \cdot \mathsf{Y}_f^{\mathbf{d}}(\mathsf{Y}_a^{\mathbf{d}})^{\intercal}\mathsf{R}^{-1}\mathbf{d}',. \tag{2.33}$$

EFSOI in KENDA Following Sommer and Weissmann (2016), the observation impact J of any subset of observations $\mathbf{d}'$ is given by Eq. 2.33, where the subscript a stands for the analysis state and f for the forecast to the next analysis time. The superscript $\mathbf{d}$ stands for the set of observations that have been used to compute the analysis or to initialize the forecasts. Furthermore, the following notation is used:

N : Number of ensemble members,
$\mathsf{Y}_f^{\mathbf{d}}$: Forecast ensemble perturbations in observation space,
$\mathsf{Y}_a^{\mathbf{d}}$: Analysis ensemble perturbations in observation space,
R : Localized observation error covariance matrix,
$\mathbf{d}'$: Innovation vector of a small subset of observations.

Localization is applied in the assimilation step as well as for the calculation of the forecast error (Eq. 2.27). Furthermore, the scalar product in Eq. (2.33) is defined through a metric that includes a normalization with the observation error σ_o as well as the number of verifying observations N_v to give equal weight to situations with varying observation density. Due to the normalization with the observation error all presented observation impacts have no unit:

$$\left|\mathbf{e}_f^{\mathbf{d}}\right|^2 := \frac{1}{N_v}\sum_{l=1}^{N_v}\left(\frac{\mathbf{e}_f^{\mathbf{d}}}{\sigma_o}\right)_l^2 \tag{2.34}$$

Depending on the verification norm, the forecast error is calculated with different sets of observations $\mathbf{y}_{\text{veri}}$. This study uses direct observations of model variables (so-called conventional observations) as well as a novel approach based on independent remote sensing observations. As the forecast error differs depending on the applied norm, only the relative observation impact can be compared directly for different norms.

Verification norms

Following Eq. (2.33), all quantities that are required to compute the observation impact are already available in KENDA, which makes the computation of observation impact cheap and efficient. The only additional quantities required are forecast departures (observations minus their model equivalents at the respective time) when observation impact is verified using observations that are not assimilated.

This study mainly compares two different verification metrics: First, conventional observations are used for verification (metric abbreviated with *CONV* in the following) as introduced by Sommer and Weissmann (2016). If not stated otherwise, all assimilated conventional observation types are also used for verification. Only observations within the second and third hour after the analysis are used for verification to avoid spin-up effects. However, sensitivity studies showed that the results hardly change including the first hour after the analysis. As in previous studies, this thesis uses the same horizontal and vertical localization for the assimilation and the calculation of observation impact. Sommer and Weissmann (2014) showed that using a static localization leads to reasonable results for lead times up to three hours.

Second, independent radar-derived precipitation observations are used for verification (abbreviated with *PREC*). Forecast errors are calculated for 3-h accumulated precipitation rates. Verification is only done if the precipitation amount exceeded a threshold of 0.1 mm/h to exclude drizzle in both forecast and observations. No vertical localization is applied in the verification as processes that contribute to the formation of precipitation take place in the entire tropospheric column. A coarse-graining is applied for both precipitation observations and forecast fields (see also Section 4.2). This is advantageous for several reasons: The coarse-graining reduces the effect of double penalty errors in the verification. Additionally, the effective resolution is lower than the model resolution (2.8 km). For this reason, the *PREC* metric includes a coarse-graining of 10×10 grid cells (28×28 km), which roughly corresponds to the lowest forecast resolution of DWD weather warnings.

The European radar precipitation product used for the *PREC* metric covers approximately three-quarters of the COSMO-DE domain, excluding areas over the sea in the northwestern and northeastern corner of the domain as well as the Alps over Austria. Besides the European radar product, a second radar precipitation product is used for verification, which has been adjusted by rain gauge observations. The second product covers a smaller area slightly larger than Germany, but not the outermost parts of the model domain. This should minimize possible boundary effects introduced by the nesting. Furthermore, the German radar product is expected to be more accurate due to the homogeneity of the radar systems within Germany. However, the German radar product covers a smaller area and therefore provides fewer observations than the European radar product, which requires a longer experimental period to obtain results with similar reliability.

In general, using independent remote sensing observations for verification provides several advantages: First, they are independent of the analysis, which mitigates potential issues of systematic and temporally correlated model or observation errors (issue of self-verification). Second, precipitation is a primary forecast quantity of convective-scale models, whereas total energy does not sufficiently reflect quantities that forecast users are interested in. Thirdly, radar observations have good and nearly homogeneous coverage of the model domain. However, using observations for verification can have disadvantages. Particularly,

observational biases are an issue for every kind of verification. Furthermore, observations usually are not homogeneously distributed and only available for specific time windows (e.g., radiosondes).

2.3 Estimating the potential impact of observations

Ensemble sensitivity analysis

Ancell and Hakim (2007) introduced ensemble sensitivity analysis (ESA) as the sensitivity S of a forecast response function $\mathbf{J}$ to the initial conditions $\mathbf{x}$

$$\frac{\partial \mathbf{J}}{\partial \mathbf{x}} \approx \frac{cov(\mathbf{J}, \mathbf{x})}{var(\mathbf{x})} := S, \tag{2.35}$$

where $\mathbf{J}$ and $\mathbf{x}$ are ensemble estimates of shape $1 \times m$ with ensemble size m. Here, *cov* denotes the covariance between two quantities and *var* denotes the variance of one quantity. Note that this approach is similar to estimating the sensitivity gradient using the model adjoint (see Section 2.2). A normalization of the sensitivity S with the ratio of the ensemble spread of the forecast response function $\mathbf{J}$ to the spread of the state variable of interest $\mathbf{x}$ provides the dimensionless sample correlation $\hat{r}$ that can be compared for different variables

$$\hat{r} = S\frac{\sqrt{var(\mathbf{x})}}{\sqrt{var(\mathbf{J})}} = \frac{cov(\mathbf{J}, \mathbf{x})}{\sqrt{var(\mathbf{J})var(\mathbf{x})}}. \tag{2.36}$$

In this thesis, mainly hourly precipitation averaged over a box as well as grid point temperature are used as forecast metrics $\mathbf{J}$ to compute sensitivities or correlations. The state variable $\mathbf{x}$ can be any quantity of interest. Overall, ESA assumes a linear relation between response function and state variable and therefore is not able to account for non-linear effects.

Potential impact

One goal of this dissertation is to estimate the relative potential impact of observable quantities using spatiotemporal correlations computed with a large ensemble. The approach takes the squared correlations accumulated over the evaluation domain and all response functions of interest as a proxy for the potential impact of the respective observable quantity on a precipitation forecast. This gives the accumulated squared correlation (ASC):

$$ASC = \sum_{i=1}^{n}\sum_{I=1}^{N} \left(\hat{r}_{i,I}\right)^2. \tag{2.37}$$

where

$$\hat{r}_{i,I} = \frac{cov(\mathbf{J}_i, \mathbf{x}_I)}{\sqrt{var(\mathbf{J}_i)var(\mathbf{x}_I)}}. \tag{2.38}$$

with index $i = 1...n$ (n - number of forecast response functions) and index $I = 1...N$ (N - number of grid points). Here, the precipitation forecast spatially averaged over kernels of 40×40 grid points is used as forecast response function $\mathbf{J}$. The kernels do not overlap and cover the entire ESA domain. The focus is to estimate the potential impact of observable quantities on precipitation, which is a primary forecast quantity of convective-scale ensemble forecasts. For simplicity, this study assumes that each grid point can be observed equally well. Therefore, sensitivities at all grid points in the domain are taken into account calculating the ASC. Finally, it is presumed that correlations obtained with a 1000-member ensemble are hardly affected by spurious correlations and can be used as a proxy for the potential impact. Later it will be shown that this assumption is reasonable. Two different approaches are tested for small ensembles to mitigate the effect of sampling errors.

2.4 Reducing sampling errors

Sampling error correction (SEC)

Following Anderson (2012) the sampling error corrected correlation r_{sec} can be obtained by

$$\hat{r}_{sec} = \gamma_{m,p}\hat{r}, \tag{2.39}$$

where γ is provided by a look-up table for a given ensemble size m and a prior distribution p given the sample correlation $\hat{r}$. $\gamma_{m,p}$ is calculated with an offline Monte Carlo technique assuming that all the sampling error comes from the correlation coefficient and that the prior distribution from which all correlation coefficients are drawn is U[-1, 1]. The offline calculation generates a look-up table for the targeted ensemble size, which can be used to correct sampling errors. The only input for the SEC are the ensemble size and the calculated sample correlation $\hat{r}$ and therefore no additional prior information is needed (Anderson, 2012). The SEC statistically corrects for the overestimation of correlations due to spurious correlations. Figure 2.3 shows the sampling error corrected correlation $\hat{r}_{sec}$ as a function of the sample correlation $\hat{r}$ for different ensemble sizes. This study uses the SEC table provided by the Data Assimilation Research Testbed (DART; Anderson et al. (2009)).
The present study assumes that sampling errors in the 1000-member ensemble are negligible and the large ensemble, therefore, can be seen as 'truth' to assess the performance of SEC. Originally, Anderson (2012) designed the SEC as localization to reduce sampling error in EnKF data assimilation. This thesis evaluates the SEC for both spatial correlations, which are required in data assimilation, and for spatiotemporal correlations calculated using ensemble sensitivity analysis.

Application to ensemble sensitivity analysis The sampling error corrected sensitivity S_{sec} can then be obtained using a look-up table by substituting $\hat{r}$ with the sampling error corrected correlation $\hat{r}_{sec}$ in Eq. 2.36

$$S_{sec} = \hat{r}_{sec}\frac{\sqrt{var(\mathbf{J})}}{\sqrt{var(\mathbf{x})}}. \tag{2.40}$$

Application to ensemble and hybrid data assimilation NWP data assimilation schemes combine observations with a short-term model forecast to achieve an optimal estimate of the atmospheric state. How the spatially sparse observational information is distributed in space is determined by sample correlations that are obtained from the ensemble. However, the ensemble size is usually too small to sample all possible states. Consequently, spurious correlations caused by under-sampling significantly degrade the initial conditions. In this context, the sampling error correction of Anderson (2012, 2016) provides an alternative to constant covariance localization length scales that are usually applied. In this study, a 1000-member ensemble is used to evaluate the sampling error correction applied to spatial correlations for different variables and ensemble sizes.
Note that the SEC is not applied during data assimilation but to spatial correlations, which are contained in the background error covariance matrix $\mathbf{P_b}$. In general, applying the SEC for a LETKF is more complicated than using an EnKF where $\mathbf{P_b}$ is computed explicitly (see Eq. 2.6). In KENDA, correlations are contained in the weights (see Eq. 2.9) that are used to compute the linear combination of the ensemble perturbations. However, this characteristic is only essential for application during data assimilation and consequently, does not affect the analysis performed in this thesis.

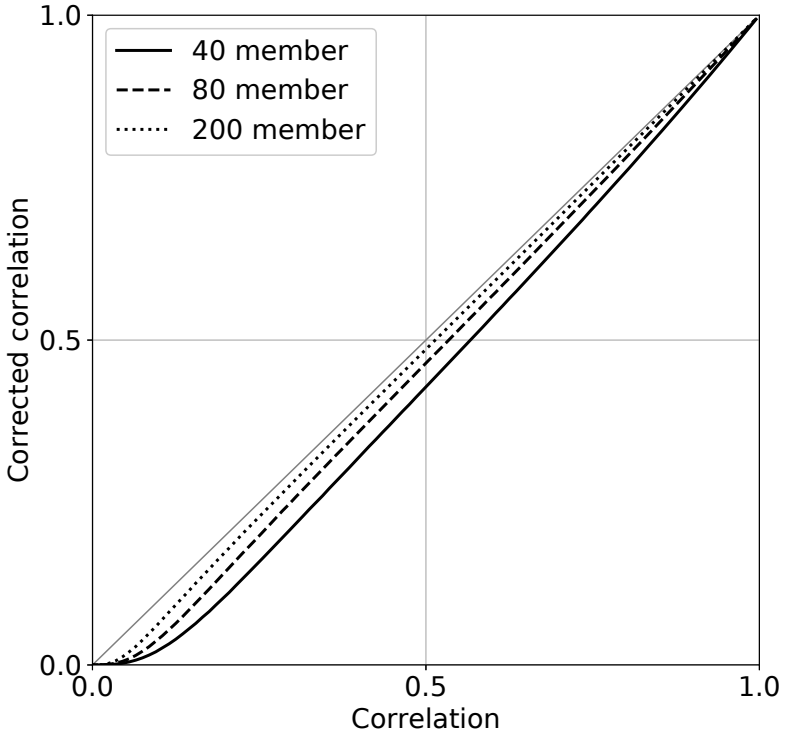


Figure 2.3: *Absolute sampling error corrected correlation $|\hat{r}_{sec}|$ as a function of absolute sample correlation $|\hat{r}|$ using different ensemble sizes.*

Confidence test (T95)

For spatiotemporal correlations, the SEC is compared to a confidence test, which is a common approach to reduce spurious correlations for ESA. The confidence test is applied to detect and exclude insignificant correlations (Torn and Hakim, 2008). This requires to

evaluate if a state variable **x** is able to cause a statistically significant change in the forecast response function **J**

$$\left|\frac{\partial \mathbf{J}}{\partial \mathbf{x}}\right| > \delta_s, \tag{2.41}$$

where δ_s is the confidence interval on the linear regression coefficient. For a given sample size, we compute the value of $|S_{conf}|$ that allows us to reject the null hypothesis with a defined confidence (i.e. that there is no correlation between the response function and the state variable X). $|S_{conf}|$ values below that threshold are not considered in the computation of the ASC. In this thesis, a 95% confidence interval (T95) will be used.

Chapter 3

Ensemble simulations

This chapter introduces and compares the two main ensemble simulations. Section 3.1 describes the 40-member ensemble simulations performed with the pre-operational forecasting system of Deutscher Wetterdienst. The COSMO-KENDA ensemble simulation covers two experimental periods: two weeks in 2014 and six weeks in 2016. These periods are mainly applied to assess the observation impact within the DWD forecasting system in Chapter 4. Section 3.2 presents the experimental setup of the convective-scale 1000-member ensemble simulation. This large ensemble simulation is required to compute realistic spatial and spatiotemporal correlations for the analysis performed in Chapter 5 and 6. Overall, the 1000-member ensemble simulation covered five days in 2016. The short period has been chosen because of exceptionally strong summertime convective precipitation. Both ensemble simulations have an overlapping time-window and roughly cover the same experimental domain. Section 3.3 gives a synoptic overview of the five-day period. This period is used to compare ensemble mean and spread of the two simulations investigating the most relevant quantities (Section 3.4). The comparison of the simulations is needed to evaluate the performance of the unique 1000-member ensemble simulation. At the end of this chapter, two basic experiments are presented that examine the added value by using a large ensemble simulation (Section 3.5). The first experiment quantifies the sampling error of variances for different ensemble subsets. The second experiment evaluates spatial correlations as they are crucial for data assimilation and discusses the need for localization. In all studies, sampling errors are quantified using the 1000-member ensemble as truth. Several sections of this chapter have been submitted for publication by Necker et al. (2019a) (see Appendix A).

3.1 COSMO-DE 40-member ensemble

Observation impact experiments are carried out with the Km-scale ENsemble Data Assimilation (KENDA; Schraff et al. (2016)) system of Deutscher Wetterdienst. KENDA is a 40-member Localized Ensemble Transform Kalman Filter (LETKF) coupled with the non-hydrostatic limited-area forecast model COSMO-DE (Baldauf et al., 2011). COSMO-DE has a horizontal grid spacing of 2.8 km, 50 vertical levels and covers approximately 1200×1300 km of Central Europe with Germany in the center of its domain (Figure 3.1). Boundary conditions are provided by the global model ICON (Zängl et al., 2015), which has a horizontal resolution of about 16 km.

The experimental setup is largely the same as in the observation impact study of Sommer

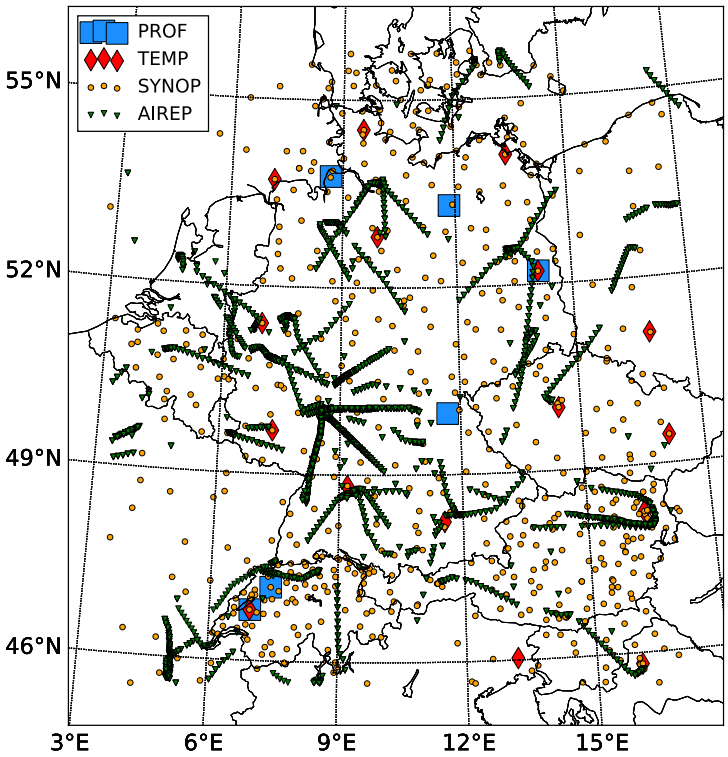


Figure 3.1: *KENDA/COSMO-DE domain and spatial distribution of the assimilated observation types on 29 May 2016 1200 UTC.*

and Weissmann (2016). Compared to the operational KENDA setup of DWD, Latent Heat Nudging (LHN) is switched off as it is not possible to estimate the impact of LHN with EFSOI. Furthermore, the localization is fixed to 100 km in the horizontal and $0.3ln(p)$ in the vertical (Gaspari-Cohn function; Gaspari and Cohn (1999)). Each experiment is given a spin-up time of one day. An analysis has been computed every 3 h, which served as the initialization for the next 3-h forecast. The present study covers two experimental periods. The first one ranges from 17 May 2014 0000 UTC to 30 May 2014 2100 UTC and the second from 26 May 2016 0000 UTC to 01 July 2016 0000 UTC. Especially the longer period in 2016 is assumed to be representative for a typical summer season as it showed several cases with synoptic and local forcings of precipitation (Piper et al., 2016). The shorter period in 2014 is mainly used for sensitivity experiments.
So far, KENDA only assimilates conventional observations that consist of four groups (Figure 3.1):

(i) Temperature, wind and humidity observations from aircraft (AIREP)

(ii) Wind observations from wind profilers (PROF)

(iii) 10-m wind and surface pressure observations from synoptic surface stations (SYNOP)

(iv) Temperature, wind and humidity observations from radiosondes (TEMP)

Following the operational KENDA setup, 10-m wind observations are only assimilated in areas with an elevation lower than 100 meters. Higher elevations likely exhibit orographic

features that can cause large representativity errors. Except for radiosondes, all conventional observations are available within almost every assimilation window.
Currently, DWD and the HErZ DA branch are working on the assimilation of additional observation types in the KENDA system, as for example solar satellite observations (Kostka et al. (2014); Scheck et al. (2016); Scheck et al. (2018)), thermal satellite observations (Harnisch et al., 2016), global navigation satellite systems (GNSS) total delay observations, radar reflectivity (Bick et al., 2016) and radar radial velocity. The inclusion of all these observations emphasizes the need for reliable estimates of observation impact in KENDA in the future.

3.2 SCALE-RM 1000-member ensemble

The large ensemble simulation comprises a set of ten 1000-member forecasts during summer 2016 with forecast lead times of 14h. Forecasts are generated coupling two domains through an offline nesting approach (see flow-chard in Figure 3.2a). The outer domain is used for the 15-km grid spacing cycled ensemble data assimilation and driven by global forecast system ensemble (GFSE) boundary conditions (BC). Initial conditions for the inner domain are obtained by downscaling from 15-km to 3-km grid spacing. The convective-scale forecasts are driven by additional forecasts performed in the outer domain.
In detail, initial conditions for the 15 km cycled experiment on 28 May 2016 00 UTC are taken from a previous 1000-member DA experiment over the same domain that has been spun-up for one week. Perturbed boundary conditions are provided every 6 hours and are generated, combining the GFSE 20-member analysis ensemble with additional random perturbations. Random perturbations at time t are obtained as the difference between two random atmospheric states that correspond to the same season and time of the day. This difference is re-scaled by a multiplicative factor, which in this simulation is equal to 0.1. Boundary conditions for each ensemble member are generated using different random perturbations (e.g., the number of generated random perturbations is equal to the ensemble size). At time t+6 hours, random perturbations are obtained from the fields corresponding to 6 hours later with respect to the random dates used in the computation of the perturbations at time t. This transition guarantees a smooth evolution of the random perturbations. Atmospheric states for the computation of random perturbations are obtained from the Climate Forecast System Reanalysis (CFSR) data-set (Saha et al., 2010) in the period between 2006 and 2009. Sensitivity studies with the modeling system have shown that the use of the GFSE perturbations at the boundaries significantly improves the ensemble spread within both domains compared to using random perturbations only.
The 1000-member ensemble simulation applies the SCALE-LETKF DA system (Lien et al. (2017)). The SCALE-LETKF system combines the open source Scalable Computing for Advanced Library and Environment - Regional Model (SCALE-RM; version 5.1.2) (Nishizawa et al., 2015, Sato et al., 2015, Nishizawa and Kitamura, 2018) and a Localized Ensemble Transform Kalman Filter (LETKF) (Hunt et al., 2007). The LETKF assimilates conventional observations using a 3-hourly assimilation window on the 15 km grid. The localization is done with an R-localization approach (Greybush et al., 2011) using a Gaussian function with a fixed localization scale of 120 km in the horizontal and $0.3ln(p)$ in the vertical and a cut-off radius equal to $2\sqrt{10/3}$ times the localization scale. The ensemble

spread is inflated using relaxation to prior spread (RTPS) with a relaxation coefficient of 0.8 (Whitaker and Hamill, 2012). Figure 3.2b shows the cycling domain that is centered over Germany. The outer domain extends over an area of 100×100 grid points (15 km mesh size) and exhibits 31 vertical levels. The model physics configuration is similar as in (Lien et al., 2017, Honda et al., 2018). All the experiments use the Tomita (2008) single-moment bulk microphysics scheme, the Mellor-Yamada-Nakanishi-Niino 2.5 closure boundary layer scheme (Nakanishi and Niino, 2004), the Model Simulation Radiation Transfer code for the representation of radiative fluxes (Sekiguchi and Nakajima, 2008") and the Beljaars-type surface model (Beljaars and Holtslag, 1991) for the computation of soil variables and surface fluxes.

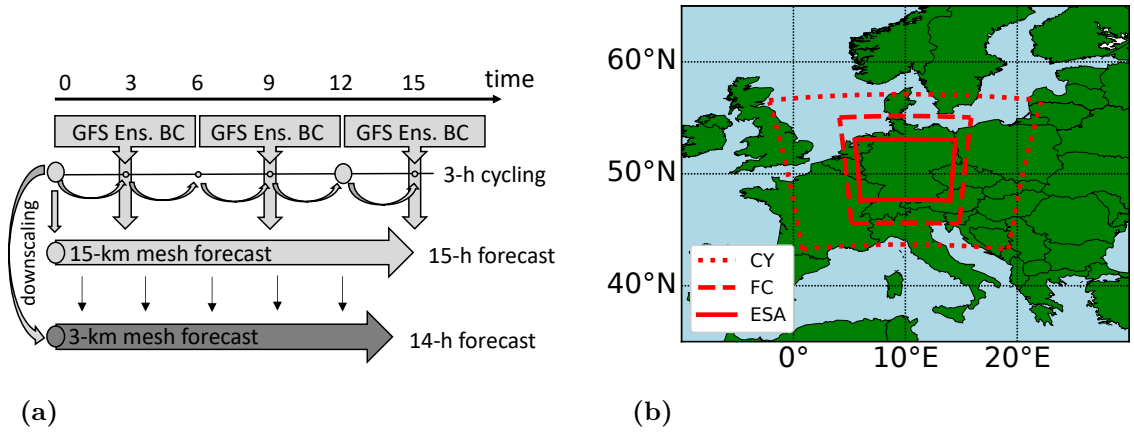


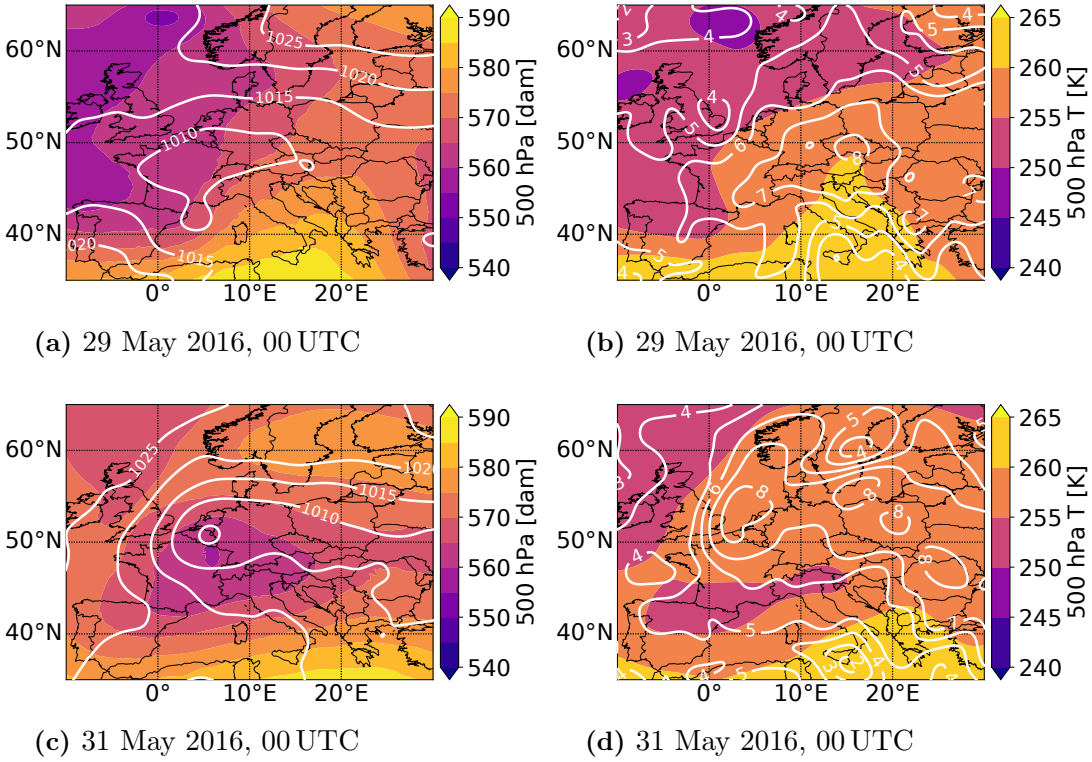
Figure 3.2: *(a) Flow-chart of the 1000-member ensemble simulation setup. (b) Experimental domains used for 15 km cycling and forecasts (CY - dotted), for 3 km forecasts (FC - dashed), and for ensemble sensitivity analysis (ESA - solid).*

The convective-scale 14-h forecasts are computed in the inner forecast domain (Figure 3.2b). The convective-scale domain measures 350×250 grid points with a 3 km mesh size and 30 vertical levels. The initial conditions for each forecast are down-scaled from the 15 km analysis to 3 km mesh size (cold-start approach). Lower-resolution 14-h (15 km mesh size) forecasts are performed in the cycling domain and provide frequent (hourly) boundary conditions for the convective-scale forecasts. This thesis analyzes a total of ten 1000-member ensemble forecasts. These forecasts are initialized every 12 hours from 00 UTC 29 May to 12 UTC 02 June 2016. All simulations have been performed on the K-computer in Kobe, Japan (Miyoshi et al., 2016a,b).

3.3 Synoptic overview

Figure 3.3 shows the general weather situation during the overlapping five-day experimental period. The short period was mainly determined by an atmospheric blocking over the Atlantic ocean leading to a fairly stationary weather situation over central Europe (Piper et al., 2016). An upper-level trough accompanied by a shallow surface low was located over the experimental domain (Figure 3.3a). The low-pressure system stayed almost stationary over France and Germany and reached its minimum pressure on 30 May (Figure 3.3c).

This weather situation led to a highly unstable environment with weak pressure gradients and synoptic-scale flow that changed from southerly (29./30. May) to easterly (31. May and 1./2. June).



(a) 29 May 2016, 00 UTC **(b)** 29 May 2016, 00 UTC

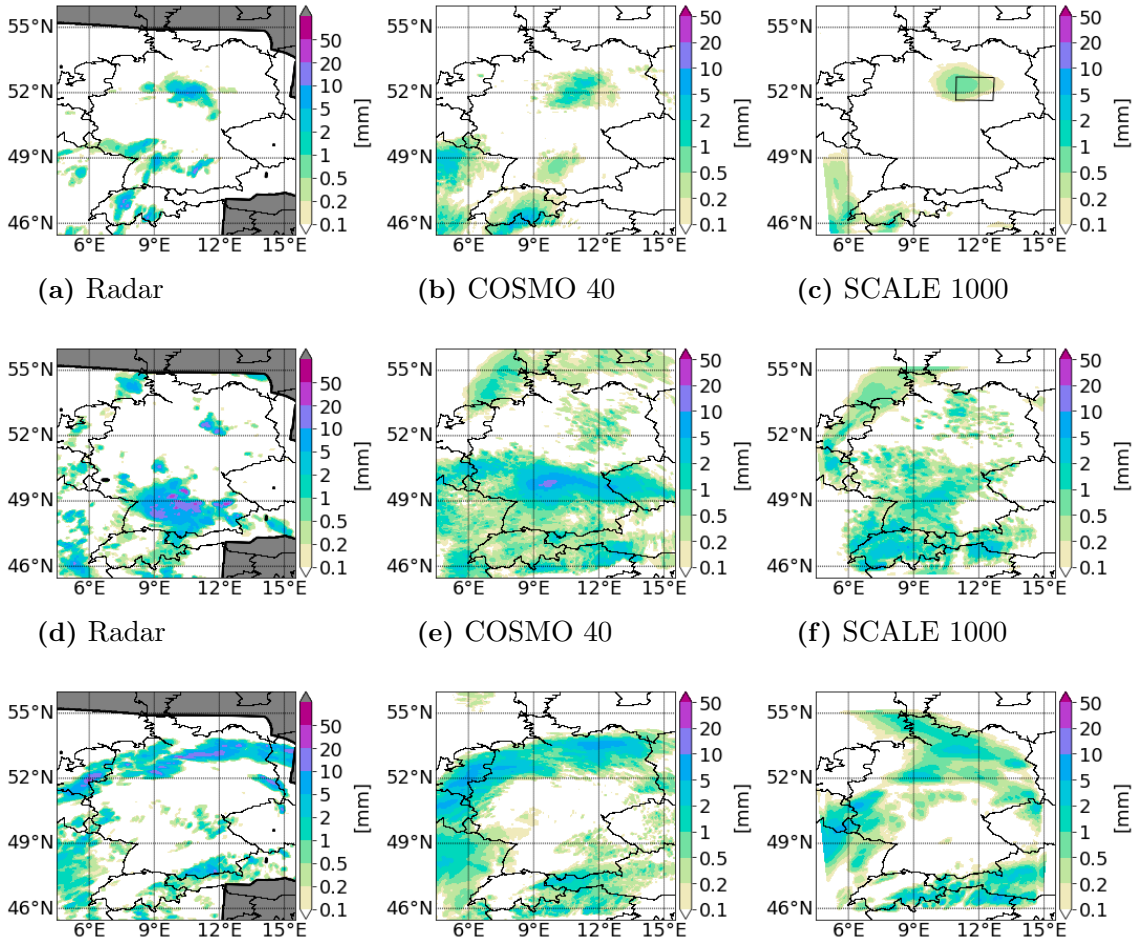
(c) 31 May 2016, 00 UTC **(d)** 31 May 2016, 00 UTC

Figure 3.3: *(a, c) ECMWF IFS analysis of geopotential height at 500 hPa (shaded, dam) and contour lines of sea level pressure (white contour, hPa). (b, d) ECMWF IFS analysis of temperature at 500 hPa (shaded, K) and contour lines of specific humidity (white contour, g/kg).*

At the beginning of the experimental period (Figure 3.3b and 3.3d), the low-level advection of moist and warm air masses from southern Europe increased the thermal instability over Germany. Both intense surface heating, as well as convective instability, forced the development of deep convection and thunderstorms on all five days. In addition, low wind speeds at 500 hPa led to several slow-moving cells causing locally extreme precipitation. The highest number of severe precipitation events occurred on 29 May 2016 (Figure 3.4d) producing flash floods, landslides, hail, and tornadoes over southern Germany. In some regions, the rainfall exceeded an amount of 100 mm per day. Observed thunderstorms showed a distinct diurnal cycle peaking in the late afternoon. Overall, the five-day period and adjacent days were characterized by intense convective precipitation. This provides a unique period that was also investigated as a test case in several other studies using the COSMO-KENDA system (Rasp et al., 2018, Keil et al., 2019, Baur et al., 2019, Bachmann et al., 2019).

3.4 Comparison of the ensemble simulations

The 1000-member ensemble forecasts are compared against COSMO-DE forecasts and independent radar-derived precipitation observations. The radar-based precipitation product (RADOLAN; EY-product) covers most parts of central Europe and delivers time-frequent observations over Germany (Figure 3.4a). The main focus of the comparison is to assess if the 1000-member ensemble captures the precipitation and provides realistic spread as this is crucial for estimating the potential impact using ensemble correlations in Chapter 5.



(a) Radar **(b)** COSMO 40 **(c)** SCALE 1000

(d) Radar **(e)** COSMO 40 **(f)** SCALE 1000

(g) Radar **(h)** COSMO 40 **(i)** SCALE 1000

Figure 3.4: *Hourly accumulated precipitation as estimated by the radar network (a,d,g) as well as COSMO-DE 40-member (b,e,h) and SCALE-RM 1000-member (c,f,i) ensemble mean precipitation for 29 May 2016 04 UTC (top row), 18 UTC (middle row) and 30 May 2016 16 UTC (bottom row). The forecast lengths for the model simulations (COSMO-DE and SCALE-RM) are 4-h (b,c,h,i) and 6-h (e,f), respectively.*

Precipitation

Examples of regional distribution of precipitation Figure 3.4a displays a map of the 1-h accumulated precipitation observations for 29 May 2016, 04 UTC. The radar composite shows a precipitation event over northern Germany and several scattered smaller

cells over France, Switzerland, and southern Germany. The 4-h COSMO-DE forecast captures the precipitation event over northern Germany while the precipitation over France and Switzerland is slightly overestimated and the precipitation in south-west Germany is underestimated (Figure 3.4b). The SCALE-RM 1000-member ensemble forecast also predicts precipitation over northern Germany (Figure 3.4c) but exhibits smaller precipitation amounts. Furthermore, SCALE-RM does not reveal enough precipitation over southern Germany even though some individual members showed precipitation in this area.
Figure 3.4d shows the precipitation observations for 29 May 2016 18 UTC, which was the strongest precipitation event occurring in the entire experimental period. On this day, both mesoscale and synoptic-scale lifting led to the development of severe thunderstorms that produced hail and rain-rates locally exceeding 20 mm per hour. The main precipitation event took place over southern Germany, although, additional cells have been observed all over central Europe. The 6-h COSMO-DE ensemble mean precipitation forecast (Figure 3.4e) covers the region of maximum precipitation but also predicts precipitation in many other parts of the domain. The region of severe precipitation is smaller, weaker, and slightly shifted to the north compared to the radar observation. SCALE-RM (Figure 3.4f) underestimates the intensity of this unique event even more and shows stronger precipitation over Switzerland and Austria. However, some members were at least able to produce precipitation rates close to the ensemble mean precipitation of COSMO-DE in the area of the maximum observed precipitation (not shown).
Figure 3.4g shows the precipitation observations at 30 May 2016, 16 UTC. At that time, an elongated precipitation region is visible over northern Germany. COSMO-DE can predict the approximate structure and intensity of the precipitation event, but there is some uncertainty on the exact position among the ensemble members (Figure 3.4h). In the SCALE-RM simulation (Figure 3.4i), the precipitation band moved too slowly and is located approximately 100 km south of the observed position. Nevertheless, SCALE-RM is able to capture the precipitation band as well as the precipitation over France and Austria.
In summary, COSMO-DE provides slightly more accurate precipitation forecasts than SCALE-RM, which is likely due to the high-resolution data assimilation incorporated in COSMO-KENDA and better tuning for the region of interest. Nevertheless, the SCALE-RM forecasts overall provide realistic precipitation amounts and patterns, which is an essential prerequisite for studying spatial and spatiotemporal correlations based on this data set.

Temporal evolution Figure 3.5a shows the temporal evolution of the domain mean precipitation during all ten forecasts. Both ensemble mean and spread are investigated for the innermost domain (see Figure 3.2b). The radar-derived domain mean precipitation is again used as a reference for both ensemble simulations. All five days featured strong precipitation events and showed a diurnal cycle in the precipitation amount peaking in the afternoon. As discussed previously, most severe thunderstorms occurred in the afternoon of 29 May 2016 indicated by the highest domain average precipitation. COSMO-DE well captures the temporal evolution of the precipitation peaking in the afternoon at a similar time and with a similar intensity as in the radar observation. Nevertheless, the COSMO-DE ensemble is not able to predict the intensity of the severe rainfall on 29 May 2016.
SCALE-RM reproduces the diurnal cycle of precipitation similarly to COSMO-DE but

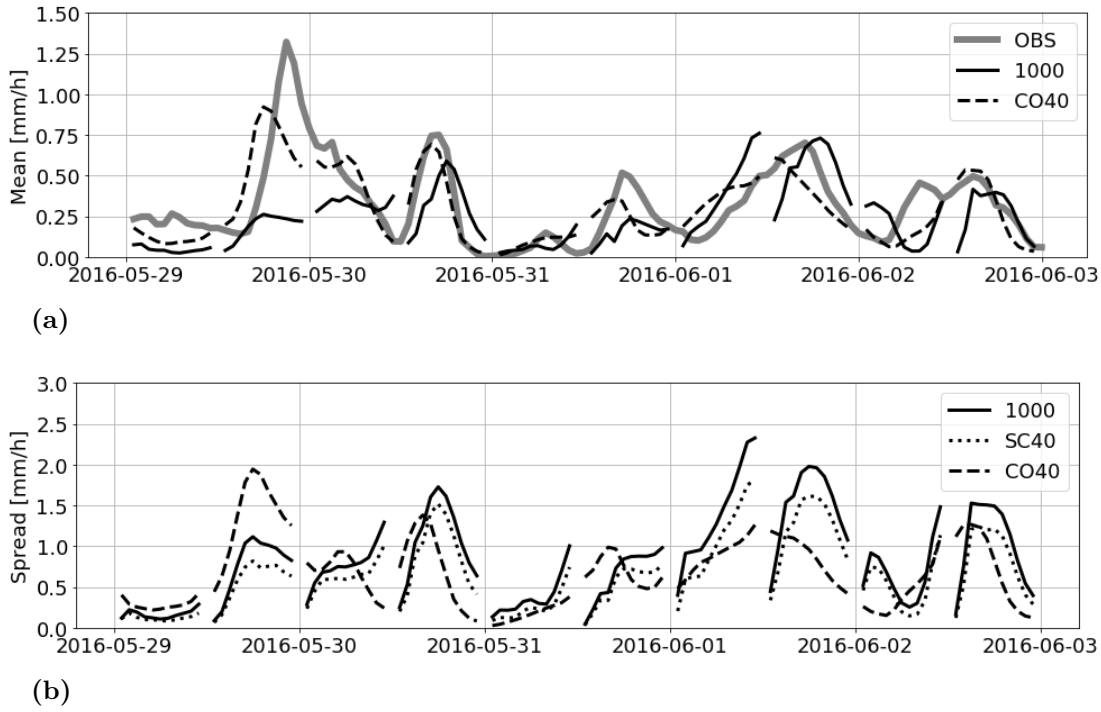


Figure 3.5: *(a) Domain mean hourly accumulated radar derived precipitation observation (solid grey) as well as SCALE-RM 1000-member (solid black) and COSMO-DE 40-member (dashed black) 12-h ensemble precipitation forecasts, 29 May 2016 00 UTC till 03 June 2016 00 UTC. (b) Domain mean spread of the hourly accumulated precipitation forecasts for different ensemble samples and the same period (SCALE 40-member ensemble, dotted black).*

less precisely. Both timing and amplitude of the peaks are slightly different from the radar observation, especially at the beginning of the experimental period. As discussed previously, one reason is that SCALE-RM was not able to fully predict the correct intensity of the severe thunderstorms over southern Germany. Additionally, most members exhibited its strongest precipitation over the Alps, but this region is not included in the verification domain (Figure 3.2b and 3.4f). Nevertheless, some members revealed a three times stronger precipitation as the ensemble mean.

Overall, the first forecast hour after each analysis should be treated with caution (see Figure 3.5). The re-initialization of a new forecast in some cases leads to an underestimation of precipitation at the beginning of the forecast. This spin-up effect is especially visible for the SCALE-LETKF system at the beginning of June. The cause of this effect seems to originate from the downscaling, which is required to obtain the high-resolution initial conditions. SCALE-RM requires a few model iterations to develop small-scale structures, as well as to diagnose sufficient precipitation from the prognostic variables, while in the COSMO-KENDA system, the precipitation amount is almost preserved after the analysis. This is reasonable recalling that for the COSMO-KENDA system the analysis is obtained on the same resolution as the forecast (warm-start initialization; 2.8 km), while the SCALE-RM simulation is based on a cold-start approach that includes a down-scaling of the initial condition from a coarser grid.

Figure 3.5b shows the temporal evolution of the domain mean spread of hourly precipita-

tion. The diurnal cycle is visible in the spread peaking in the late afternoon and showing a smaller amplitude during the night. Except for the first day of the experimental period, the SCALE-RM 1000-member ensemble exhibits the largest spread of all ensemble simulations. A random 40-member subset of the 1000-member ensemble is additionally included in the comparison to assess if the ensemble spread changes with the ensemble size as well as to compare COSMO-DE and SCALE-RM using an equal ensemble size. The SCALE-RM 40-member ensemble reveals a smaller spread in all forecasts resulting from an under-sampling of the true variance and is often closer to the spread of COSMO-DE. As discussed previously, initializing SCALE-RM from the down-scaled analysis reduces the spread at the beginning of most forecasts.
Overall, the SCALE-RM 1000-member ensemble delivers fairly realistic precipitation forecasts regarding ensemble mean and spread. The amount and timing of precipitation events do not necessarily need to coincide with an operational forecasting system or observations as ensemble sensitivity analysis, or the analysis of sampling errors does not incorporate observations and therefore only requires realistic scenarios. The first forecast hour has been ignored for the ensemble evaluation to exclude potential spin-up effects originating from downscaling. For this reason, the 1-h forecast is used as an initial state x to compute ensemble sensitivities to precipitation.

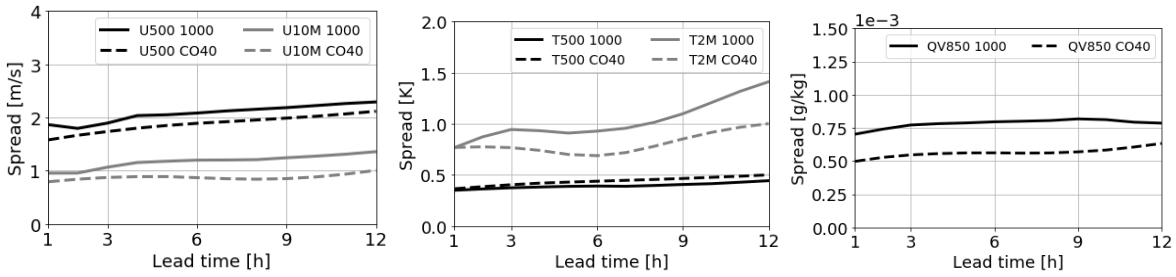
Growth of spread for prognostic model variables

Figure 3.6 displays the evolution of the domain mean ensemble spread with forecast lead time for different prognostic model variables. For simplicity and as the focus is on the growth of the ensemble spread, the available ten forecasts have been averaged. The 1000-member ensemble spread of 10-m and 500 hPa zonal wind (Figure 3.6a) is slightly larger than for COSMO-DE. For both simulations, the ensemble spread increases equally fast throughout the forecast, while the upper-air spread is larger than that close to the surface. The ensemble spread of 500 hPa temperature (Figure 3.6b) hardly increases with lead time and coincides roughly for both simulations.
In contrast to the zonal wind, the ensemble spread close to the surface is larger than in the middle troposphere. Initially SCALE-RM and COSMO-DE exhibit a similar surface temperature spread, which increases stronger in the 1000-member ensemble simulation. The ensemble spread for 850 hPa specific humidity (Figure 3.6c) is higher for the SCALE-RM 1000-member ensemble, while the COSMO-DE ensemble spread increases slightly faster. For all prognostic model variables, the evolution of the ensemble spread is reasonable and well simulated by the SCALE-RM ensemble.

3.5 Added value by large ensemble size

Variance error as a function of ensemble size

Figure 3.7 displays the temporally and spatially averaged relative variance error for three different representative variables as a function of the ensemble size. For the investigation of variances in this subsection, multiple small ensembles are obtained by random subsampling from the full 1000-member ensemble without using additional constraints. The



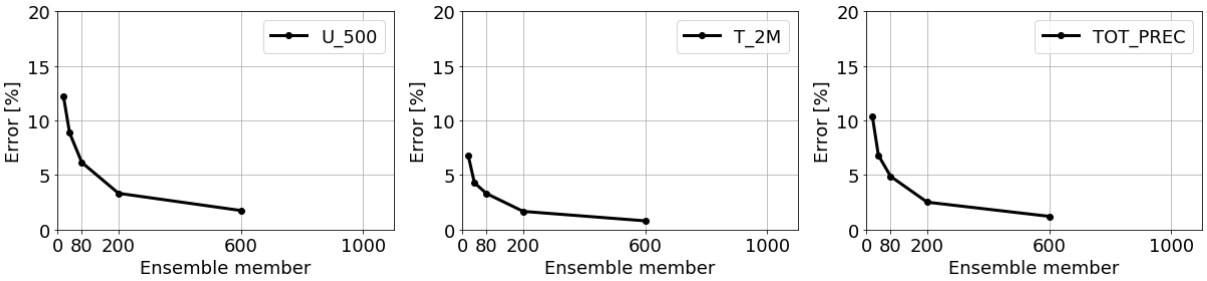
(a) Zonal wind **(b)** Temperature **(c)** Specific humidity

Figure 3.6: *Time-averaged domain mean spread as function of lead time for different ensembles and prognostic variables, 29 May to 03 June 2016. (a) 500 hPa and 10-m zonal wind (b) 500 hPa and 2-m temperature and (c) 850 hPa specific humidity.*

1000-member ensemble variance serves as a reference to calculate relative errors at each grid point. Results for smaller ensemble sizes are additionally averaged over multiple subsets to increase the robustness of the results. Figure 3.7a shows the relative variance error of 500 hPa zonal wind. The 20-member ensemble variance on average exhibits a sampling error of approximately 12%. Increasing the ensemble size by a factor four from 20 to 80 members almost halves the sampling error to about 6%. Using 200 members again halves the relative error to about 3%. A similar error reduction is found increasing the ensemble size to 600 members. However, the relative reduction using a 600-member ensemble is small, considering the computational effort and the overall small sampling error. Figure 3.7b shows the relative variance error for 2-m temperature. The sampling error of surface temperature and a 20-member ensemble is smaller compared to zonal wind. Nevertheless, the relative reduction of the sampling error with ensemble size is similar. Again, increasing the ensemble size to 80 members (or from 80 to 200 members) almost halves the sampling error. A similar dependence of the sampling error on the variance is found for precipitation (Figure 3.7c) and various other variables (not shown). These results show that increasing the ensemble size from 20 to 80 members (or from 80 to 200 members, respectively) approximately halves the sampling error in the variance taking the 1000-member ensemble simulation as a benchmark. Considering larger ensemble sizes (e.g., 600 members), sampling errors are relatively small, and the error curve flattens.

Spatial correlations and sampling errors

Below, horizontal and vertical spatial correlations are investigated. Such correlations are crucial for hybrid and ensemble DA systems, which rely on accurate correlation estimates for spreading the information from observations spatially and among different variables. Results are discussed for three different ensemble sizes: A small (40 members), medium-sized (200 members) and large (1000 members) ensemble. The subsets are drawn from the 1000-member ensemble with the constraints that each member of the 20-member GFSE BC is represented equally often and that the 40-member subset is included in the 200-member subset. Spatial correlations between different grid points and variables are calculated using the 1-h forecast state, which is similar to using the first guess during hourly cycling. Ensemble subsets applied for the studies in Chapter 6 and 5 are generated identically.



(a) 500 hPa zonal wind **(b)** 2-m temperature **(c)** Precipitation

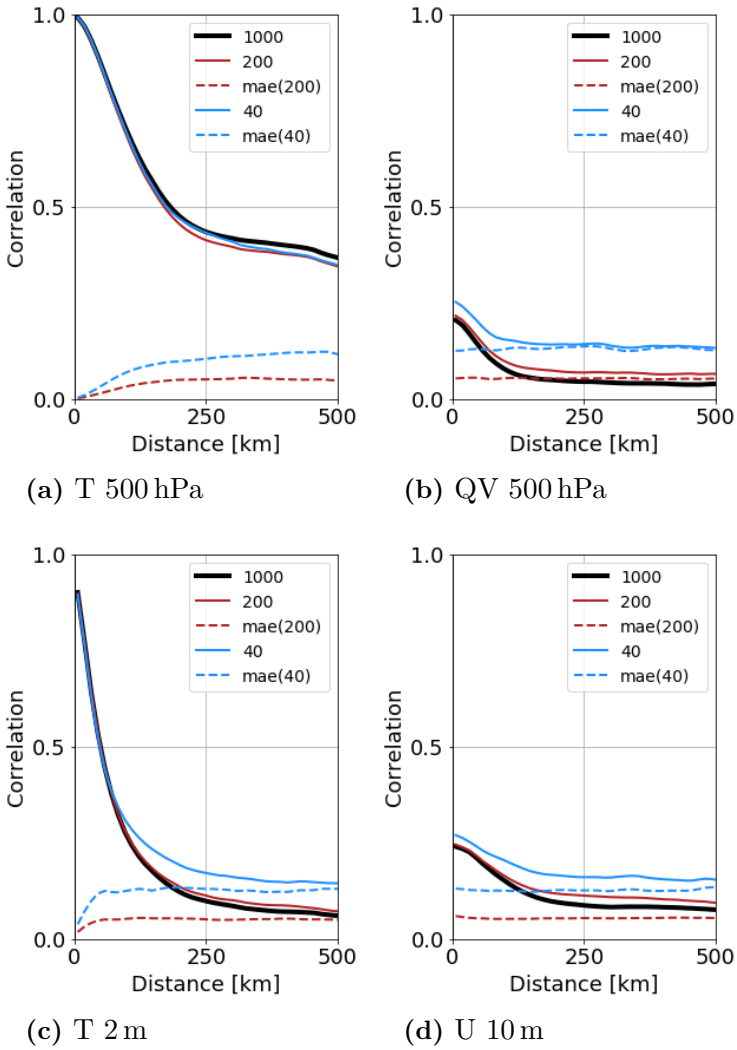
Figure 3.7: *Time-averaged relative error of variance as function of ensemble size for different variables using all ten 12-h forecasts. Relative error averaged over 20 (for 20 and 40 members), 10 (for 80 and 200 members) and 5 (for 600 members) randomly sub-sampled ensembles.*

Horizontal correlation Figure 3.8 displays the mean absolute correlation (MAC) and error (MAE) as a function of spatial distance. The error here is the error in the correlation assuming the 1000-member correlation as the truth. For each forecast, we evaluate correlations from nine different grid points to all other grid points in the domain. These nine grid points are evenly distributed in the domain and lie at least 150 km apart from each other. The error at each grid point is calculated with respect to the 1000-member ensemble correlation, and results are averaged over all forecasts.

Figure 3.8a shows the mean absolute correlation of the full 1000-member ensemble (MAC_{1000}) of 500 hPa temperature to 500 hPa temperature itself as a function of horizontal distance. On short distances, tropospheric temperatures are highly correlated. The MAC_{1000} decreases to 0.5 at a distance of 200 km. From 200 to 500 km, the spatial de-correlation continues but with a weaker gradient. Both, the MAC_{40} and MAC_{200} coincide with the MAC_{1000}, slightly underestimating the correlation at large distances. The 200-member subset exhibits approximately half the mean absolute error (MAE) of the 40-member subset, but the error of both samples is much smaller than the MAC. The MAE_{200} increases slower with distance than MAE_{40} and seems to be saturated after about 150 km.

Cross-correlations of temperature to specific humidity (Figure 3.8b) are significantly weaker. The MAC_{1000} exhibits a maximum value of about 0.22, decreases up to a distance of approximately 100 km and remains constant farther away. The 200-member ensemble roughly estimates the shape of the MAC_{1000}, while 40 members significantly overestimate the true correlation due to spurious correlations. The MAC_{40} is approximately three times larger compared to the MAC_{1000} after a distance of 150 km. For the 40-member ensemble, correlations of distances longer than 50 km are not trustworthy as their error exceeds the absolute value. This cross-over point roughly indicates a suitable choice as localization scale in data assimilation. Using 200 members almost doubles the distance of this cross-over point compared to 40 members.

Figure 3.8c shows the spatial correlation of 2-m temperature to 2-m temperature itself. As for upper air temperatures, the MAC_{1000} is large on short distances but decreases faster and is weaker at longer distances. The 200-member ensemble almost coincides with the MAC_{1000} and the error does not exceed the MAC_{1000} before reaching 500 km distance. The MAC_{40} agrees with the MAC_{1000} up to a distance of about 100 km, but around 200 km the



(a) T 500 hPa **(b)** QV 500 hPa

(c) T 2 m **(d)** U 10 m

Figure 3.8: *Mean absolute correlation (MAC; solid) and error (MAE; dashed) as function of spatial distance [km] for 1000, 200 and 40 members. Correlations of 500 hPa temperature to (a) 500 hPa temperature and (b) 500 hPa specific humidity as well as correlations of 2-m temperature to (c) 2-m temperature and (d) 10-m zonal wind.*

error starts to get larger than the absolute value.

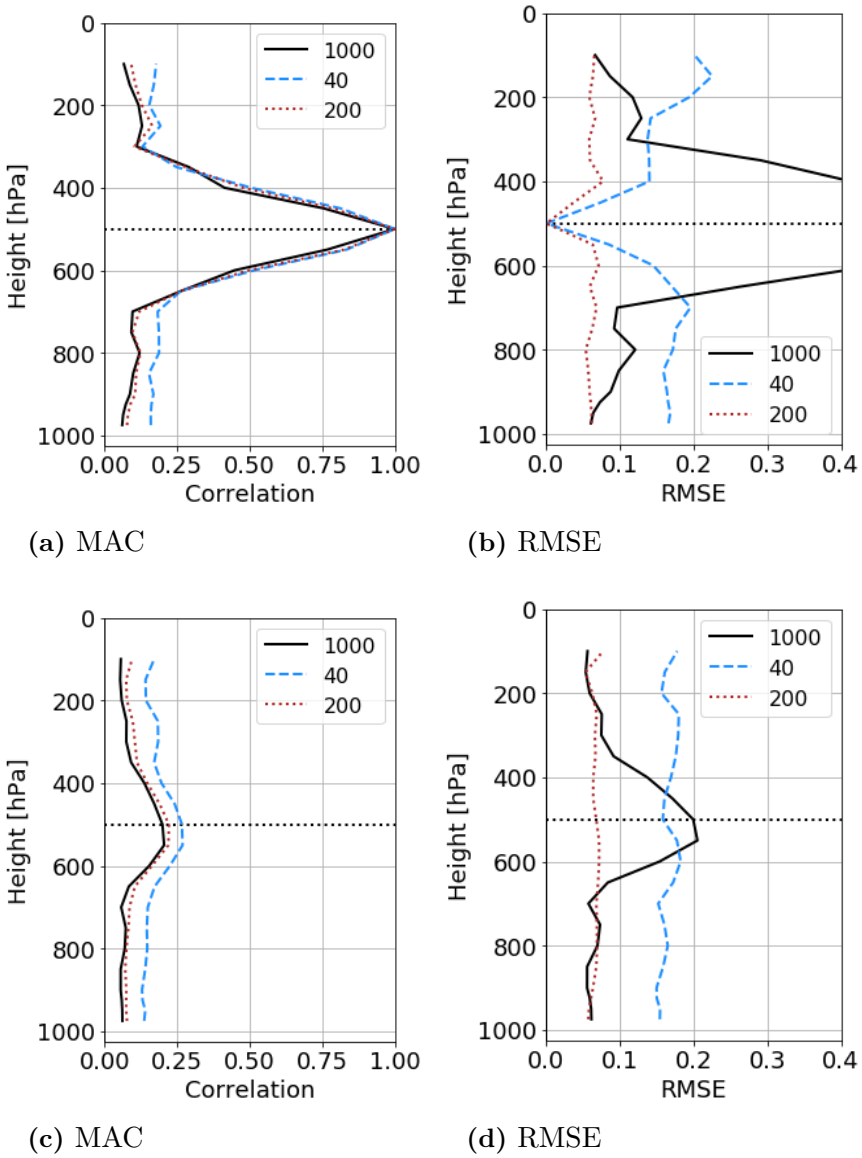
The MAE of cross-correlations of 2-m temperature to 10-m zonal winds (Figure 3.8d) is fairly constant at all distances, while the corresponding MAC is much smaller than for the correlation of 2-m temperature to 2-m temperatures. As a consequence, the MAE_{40} exceeds the MAC_{1000} at a distance of slightly over 100 km, while the MAE_{200} remains below up to a distance of 500 km.

In summary, these examples show that the 1000-member ensemble can be used to quantify sampling errors and to investigate suitable choices for localization length scales in convective-scale NWP. The different results for different variables highlight that it would be desirable to select very different scales for different model variables and combinations of variables. Such a variable-dependent localization, however, is not straight forward to implement in some variations of ensemble DA such as the LETKF for example, where localization is applied in observation space. Finally, the results show a significant advantage

of correlations from the 200-member subset compared to 40 members.

Vertical correlation Vertical correlations are evaluated using one 1000-member ensemble forecast at 30 May 2016 and vertical profiles at all grid points in the domain. Figure 3.9a shows the mean absolute correlation of temperature at 500 hPa to temperature at other levels. The MAC_{1000} exhibits a correlation of 1 at the response level and rapidly decreases to a value of 0.25 reaching a vertical distance of 150 hPa. Levels close to the ground and tropopause are hardly correlated with 500 hPa. The 200-member ensemble again roughly coincides with 1000 members with only a slight overestimating of the true correlation. The 40-member ensemble gives similar results close to the response level but overestimates the correlation above and below by a factor two.
Next, the average root mean square error (RMSE) of the correlation is examined as a function of height, assuming the 1000-member correlation as the truth (Figure 3.9b). The 200-member ensemble exhibits a small sampling error of about 0.05, except for the response level and the two neighboring levels above and below. The RMSE_{40} exhibits a significantly higher error with values that are up to three times higher at distances of more than 100 hPa. Comparing the amplitudes of sampling errors for both subsets with the MAC_{1000}, the RMSE_{200} does hardly exceed the true correlation. In contrast, the RMSE_{40} increases faster with distance and exceeds the MAC_{1000} 200 hPa above and below the response level. Consequently, using a 40-member ensemble would require a narrow vertical localization to reduce the impact of spurious correlations. For temperature, the width (in hPa) of the required vertical localization does hardly change with the height of the chosen response level (not shown).
Figure 3.9c displays vertical cross-correlations of 500 hPa temperature to specific humidity (Figure 3.9d) at other levels. The MAC_{1000} is generally weak and exhibits a maximum of 0.2 around 500 hPa. The 40-member ensemble overestimates the MAC by 0.1 independently of the height. The 200-member ensemble only slightly overestimates the MAC. As for temperature, the vertical extent of the area of increased correlation is approximately 300 hPa. Nevertheless, the relative error is larger for cross-correlations as the 1000-member correlation is much weaker (Figure 3.9d). Using 200 members, correlations for distances larger than about 150 hPa are not trustworthy. For the 40-member ensemble, the error significantly exceeds the 1000-member correlation at nearly all levels except for a narrow band around the response level. Thus a strong localization would be required to reduce sampling errors.
Overall, 200 members appear sufficient to estimate vertical correlations of temperature, while 40 members require a narrow vertical localization of less than 200 hPa vertical extent. In general, estimating vertical cross-correlations is more demanding than estimating horizontal correlations. Especially, spurious correlations in combination with weak correlations are an issue as significant relative errors emphasize the need for localization for both investigated ensemble sizes. Principally, localizing after a certain distance is potentially dangerous in case of long-range correlations. As discussed in Chapter 1, clouds and hydrometeors can exhibit long-range vertical correlations. Similarly, satellite observations often provide integrated information of the vertical profile. A possible solution for this issue is the statistical sampling error correction (SEC) approach that aims to correct for spurious correlations without damping correlations after a certain distance (Anderson, 2012, 2016).

Further investigation of this approach is provided in Chapter 6.



(a) MAC **(b)** RMSE

(c) MAC **(d)** RMSE

Figure 3.9: *Mean absolute correlation (MAC) and mean RMSE as function of vertical distance [hPa] for differently sized ensembles. Correlations are calculated from 500 hPa response level (horizontal dotted line) to all other levels. Correlations of 500 hPa temperature to (a) temperature and to (c) specific humidity. Corresponding errors for temperature (b) and specific humidity (d), respectively. Note: The black solid line in (b,d) displays the* MAC_{1000} *as shown in (a,c).*

Chapter 4

Observation impact

This chapter presents the first part of the results and deals with observation impact. As elaborated in Section 1.3 and 2.2, estimates of observation impact are crucial to improve the data assimilation, forecasting, and observing system. However, Deutscher Wetterdienst so far does not monitor the impact of observations in the convective-scale NWP system. To overcome this deficit, Sommer and Weissmann (2014) and Sommer and Weissmann (2016) developed a method to estimate the ensemble forecast sensitivity to observation impact (EFSOI) in the COSMO-KENDA system. The EFSOI approach applies an observation-based verification norm and efficiently reuses available information from the data assimilation cycling. Nevertheless, previously conducted studies by Sommer and Weissmann (2016) have been too short of drawing representative conclusions on the actual impact. Furthermore, Sommer and Weissmann (2016) applied a verification norm that was based on similar observations as are assimilated during the cycling. In general, performing a self-verification is potentially dangerous as it could affect the impact measure. Therefore, the present study extends the approach of Sommer and Weissmann (2016) by using independent remote sensing observations for verification. Additionally, a 6-week summer period is analyzed to obtain representative estimates of observation impact for the convective-scale forecasting system of Deutscher Wetterdienst.

Chapter 4 is outlined as follows: The first section examines the impact of conventional observations during the extended summer period. The observation impact is evaluated for different verification norms. Mainly the norm applied by Sommer and Weissmann (2016) is compared against a new norm, which is based on radar precipitation observations. Furthermore, several sensitivity studies with the FSOI method (Section 4.2) are presented. A major focus is given to the issue of biases in assimilation and verification. In this context, the connection between biases and the fraction of beneficial observations is discussed, which turns out to be strongly connected (Section 4.3). Section 4.4 examines the contribution of high-impact observations to the reduction in forecast error. Afterward, a simple approach that can be used to verify the representativity of the impact results is presented (Section 4.5). At the end of this chapter, Section 4.6 compares the observation impact of two Swiss wind profilers to results from a European observing network impact study. The last section provides the reader with a summary and conclusions. Overall, most results from the following pages have been published by Necker et al. (2018) (see Appendix A).

4.1 Observation impact for a 6-week period

Figure 4.1 shows the number of assimilated observations during the 6-week high-impact weather period in summer 2016. Treating the u- and v-component of wind as two single observations, wind profilers provide almost a million observations followed by aircraft wind ($\approx 840,000$) and surface pressure observations ($\approx 490,000$). The number of aircraft humidity observations is comparably small as only a few commercial aircrafts are equipped with humidity sensors. Overall, there are ten times more aircraft than radiosonde observations as the latter are usually only launched at 0000 and 1200 UTC. The smallest number of observations is therefore available for radiosonde temperature ($\approx 45,000$) and humidity ($\approx 30,000$).

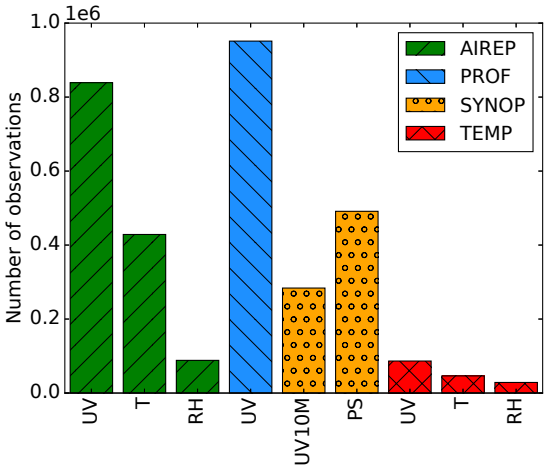


Figure 4.1: *Number of assimilated observations for different observation types and variables during the 6-week summer period 2016: Aircraft (AIREP), wind profilers (PROF), surface stations (SYNOP) and radiosonde (TEMP). Wind (UV), temperature (T), humidity (RH), 10-m wind (UV10M) and surface pressure (PS).*

Verification with conventional observations (CONV)

Figure 4.2a shows the mean observation impact per observation verified with the *CONV* metric for the investigated 6-week period. The majority of all observation types exhibits a beneficial impact, except radiosonde temperature. Surface pressure shows by far the largest impact per observation followed by aircraft temperature (note that the surface pressure impact exceeds the scale by five times). The impact of wind observations is similar for all instruments. In contrast, aircraft and radiosonde temperature show a very different impact, which seems to be related to biases in the verification metric (see Section 4.1 for further investigation and discussion of this result).
A two-sided t-test is performed to obtain the 95 % confidence interval, which is indicated by grey whiskers. For the 6-week period and the verification with the *CONV* metric, the confidence is fairly high for most observation types, and the sign of the estimated impact appears reliable for all observation types. Only for the smaller sample of radiosonde temperature and humidity observations, the confidence interval is of similar magnitude as the estimated impact.

Figure 4.2b shows the corresponding total impact of different observation types and variables. Surface pressure contributes most to the overall reduction in forecast error followed by aircraft temperature. Wind profilers exhibit the third largest impact. Radiosonde, aircraft humidity, and surface wind observations only represent a small proportion of the total impact due to their comparably small number.

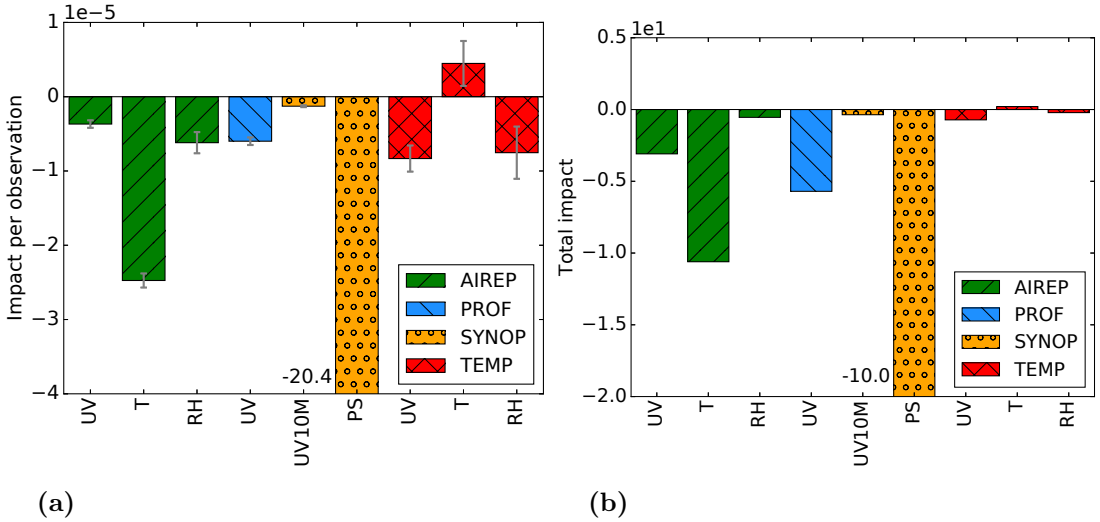


Figure 4.2: *(a) Temporally averaged observation impact per observed variable and type verified with the CONV metric. Grey whiskers display the 95 % confidence interval obtained with a two-sided t-test. (b) Corresponding total impact per observed variable and type. The bar of surface pressure exceeds the regime scale by five times in both (a) and (b).*

Biases in the verification

Given that radiosondes are reliable sensors for temperature, their detrimental impact found in Figure 4.2a is surprising. This behavior can be explained by looking at systematic differences between observations and their model equivalents. Figure 4.3a shows mean vertical profiles of first-guess departures (observations minus their model equivalents; $\mathbf{D} = \mathbf{O} - \mathbf{FG}$) of temperature observations during the 6-week period. Usually, radiosondes $\mathbf{O}_{TEMP}$ are close to the truth as they are very reliable temperature observations. In contrast, aircraft temperature observations $\mathbf{O}_{AIREP}$ are known to exhibit some systematic errors, which are difficult to correct as they tend to depend on the individual aircraft and can even differ for different segments of the flight (ascent, cruise level, descent). Due to these difficulties, the KENDA system currently has no bias correction scheme for aircraft observations. Assuming that radiosonde temperature $\mathbf{O}_{TEMP}$ is unbiased leads to the conclusion that both the model first-guess $\mathbf{FG}$ and aircraft temperature $\mathbf{O}_{AIREP}$ exhibit a warm bias (Figure 4.3c). In Figure 4.3a, mean radiosonde departures $\mathbf{D}_{TEMP}$ are negative at all heights (indicating a model warm bias), and they are of a larger magnitude than mean aircraft departures $\mathbf{D}_{AIREP}$ at all heights (indicating an aircraft warm bias). Below 700 hPa, both aircraft and radiosonde first-guess departures have a negative sign and above 700 hPa, the aircraft first-guess departures reveal an opposite/negative sign.

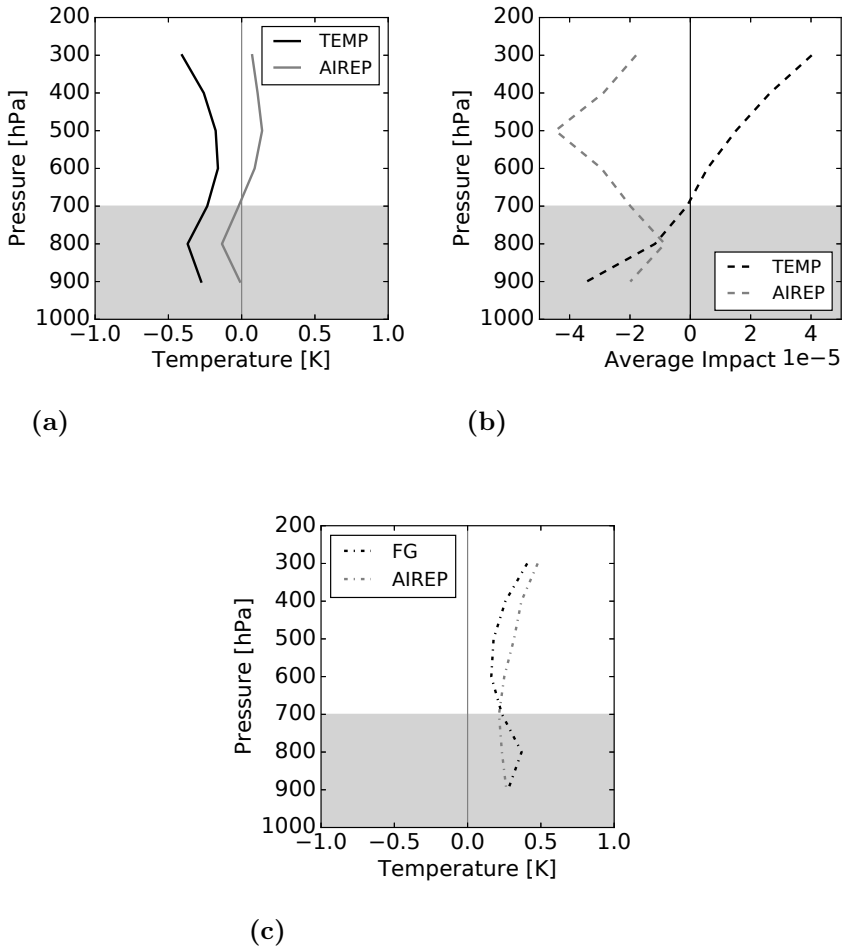


Figure 4.3: *(a) Vertical profiles of mean first-guess departures (observation minus model equivalents;* $\mathbf{D} = \mathbf{O} - \mathbf{FG}$*) of aircraft (*$\mathbf{D}_{AIREP}$*; solid gray) and radiosonde (*$\mathbf{D}_{TEMP}$*; solid black) temperature observations. (b) Vertical profiles of the corresponding mean temperature observation impacts of aircraft (dashed gray) and radiosonde observations (dashed black). (c) Vertical profiles of the mean temperature first-guess bias (dash-dotted black) as well as aircraft bias (dash-dotted gray) assuming radiosonde observations are unbiased.*

Figure 4.3b displays the corresponding vertical profiles of the mean estimated impacts calculated for bin widths of 100 hPa. Considering the large number of aircraft observations (see Figure 4.1), radiosondes are mainly verified with presumably biased aircraft observations. Above 700 hPa, where mean aircraft departures $\mathbf{D}_{AIREP}$ show the opposite sign of mean radiosonde departures $\mathbf{D}_{TEMP}$, the estimated radiosonde impact is detrimental (positive impact values). Below 700 hPa, where both aircraft and radiosonde temperature are on average colder than the model, the estimated radiosonde impact is beneficial (negative impact values). This result indicates that the estimated radiosonde impact is systematically affected by the verification with unreliable aircraft temperature observations $\mathbf{O}_{AIREP}$ and that even a moderate bias of $0.2-0.5$ K can significantly influence the estimated impact. Section 4.2 presents bias sensitivity experiments for temperature to explore the effect of biases further.

Assessing the bias of other observation types, it turns out that the surface pressure observations also exhibit biases that are large enough to affect the EFSOI results. Figure 4.4b shows the diurnal cycle of the surface pressure impact verified with two different verification metrics. Verified with the *CONV* metric, surface pressure exhibits its maximum impact in the early afternoon. The diurnal cycle vanishes if surface pressure is excluded from the *CONV* metric, and the estimated impact is then close to zero. Figure 4.4a displays the diurnal cycle of mean surface pressure first guess and analysis departures. Surface pressure first-guess departures exhibit a clear diurnal cycle with a maximum in the afternoon. The assimilation of surface pressure observations effectively corrects the model bias and leads to mean analysis departures close to zero. Hence, the magnitude of this surface pressure bias correction in the analysis and its diurnal cycle coincides with the estimated impact.

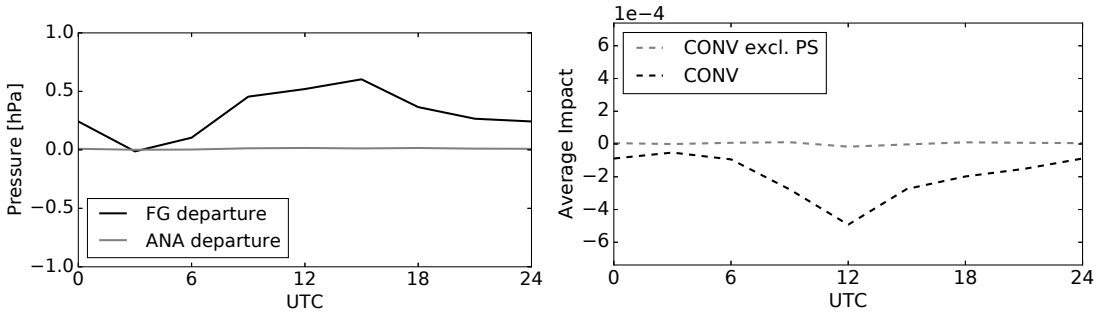


Figure 4.4: *6-week temporally averaged diurnal cycles of surface pressure (PS) observation minus first guess departures, observation minus analysis departures (a) and surface pressure impact verified with the CONV metric in-/excluding surface pressure (b).*

These findings emphasize the importance of monitoring potential biases in the verification metric as there is further indication that even moderate biases (approximately 0.5 hPa at its maximum in the afternoon) strongly influence the estimated impact. Presumably, this effect is particularly pronounced when the same observation type is used for assimilation and verification - or if the impact calculation is based on an analysis that contains information from the short-term forecast. As a consequence, it seems desirable to use independent observations for verification in the impact calculation as presented in Section 4.1.
Figure 4.5 shows the mean observation impact excluding both biased aircraft temperature and surface pressure observations from verification. The confidence is slightly reduced compared to the *CONV* metric, as fewer observations are used for verification. Surface pressure and aircraft temperature, in this case, show fairly large detrimental impacts in contrast to the verification including these observations (Figure 4.2a). Radiosonde temperature still shows a moderate detrimental impact, but the confidence of the estimate is low, and one should keep in mind that there are only very few temperature observations left in the verification metric. Results for the other observations types also change slightly, but the sign of their impact is unaffected. Sensitivity experiments for excluding other types from the verification metric (not shown) generally revealed a decreased impact of the observation type that is excluded from verification, but changes are usually smaller than the changes for aircraft temperature and surface pressure in Figure 4.5.

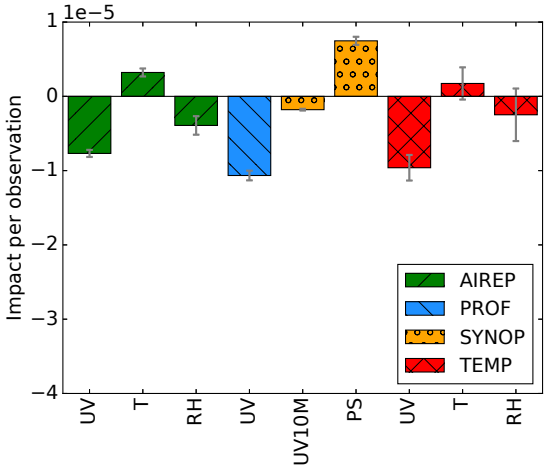


Figure 4.5: *Temporally averaged observation impact per observed variable and type verified with the CONV metric excluding biased surface pressure and aircraft temperature observations from verification. Grey whiskers display the 95 % confidence interval.*

Verification with remote sensing observations (PREC)

Comparing the results for different verification metrics is important to achieve a complete and reliable picture of observation impact – and generally for any evaluation of forecasts. Furthermore, it is advantageous to use metrics that are independent from the assimilation and metrics that reflect primary forecast parameters. Figure 4.6a shows the temporally averaged observation impact verified with independent radar-derived precipitation observations. A coarse-graining for 10×10 grid cells is applied to account for the resolvable scale of precipitation. Compared to the *CONV* metric, all observation types and variables exhibit a beneficial impact. Aircraft temperature now exhibits a smaller impact per observation than radiosonde temperature. This change is likely related to the aircraft temperature bias discussed in Section 4.1. Surface pressure still shows the largest beneficial impact, which indicates that the correction of the model surface pressure bias through the assimilation of surface observations is also very beneficial for precipitation forecasts. The neutral impact of radiosonde wind observations is caused by very few observations with large detrimental impact compensating the majority of beneficial observations (see Section 4.4). The results for all radiosonde as well as aircraft humidity observations should be treated with caution as they exhibit low confidence.

Generally, remote sensing observations provide a better spatial and temporal coverage than conventional observations. However, compared to the *CONV* metric, the confidence is reduced as verification is only done when precipitation occurs. Achieving similar confidence as for the *CONV* metric would, therefore, requires an extended period with more precipitation events.

Figure 4.6b shows the total impact verified with the *PREC* metric. Surface pressure again exhibits the largest beneficial impact, followed by aircraft wind observations. The impact of profiler wind observations, in this case, is smaller compared to the *CONV* metric that favors temporally continuous observations as these always have observations for verification at the same location (Sommer and Weissmann, 2016). The impact of 10-m wind is comparably small as for other verification metrics. This is likely related to large representativity errors

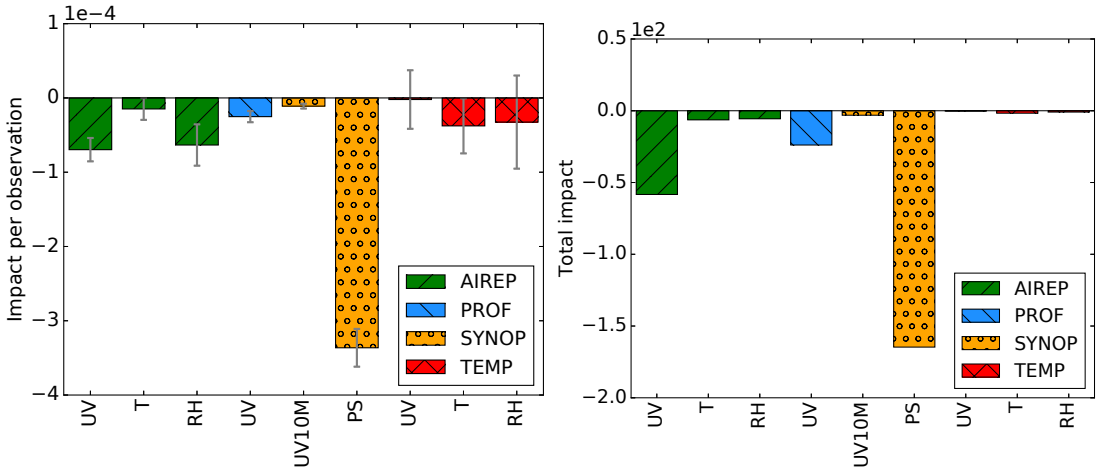


Figure 4.6: *(a) Temporally averaged observation impact per observed variable verified with the PREC metric. Grey whiskers display the 95 % confidence interval. (b) Corresponding total impact per observed variable.*

assigned during data assimilation. Due to their low number, the total impact of aircraft humidity and radiosonde temperature/humidity is comparably small while their impact per observation is similar to that of aircraft wind observations.

Figure 4.7 displays the temporally averaged observation impact verified with the German radar product that has been adjusted with rain-gauges. The overall picture is similar as for the *PREC* metric based on the European radar product, which confirms the results obtained previously. The confidence is slightly lower due to the smaller verification domain and the reduced number of verifying observations. The surface pressure impact increases and is still the observation type with the largest impact per observation. The strongest change appears for radiosonde humidity, but the difference lies within the confidence interval. In general, a longer period is required for reliable estimates of the impact of radiosonde observations independent of the product used for verification.

4.2 Sensitivity studies

Recent studies performed sensitivity experiments considering different aspects of the EFSOI method. Among other parameters, different localization length scales and the sensitivity to the number of ensemble member were investigated (Kalnay et al. (2012); Ota et al. (2013); Sommer and Weissmann (2014)). So far, only one study investigated the sensitivity of the EFSOI method to observation biases. Liu and Kalnay (2008) assimilated one single biased observation in a Lorenz 40-variable model approach and showed that a bias could change the estimated impact of a biased observation and the unbiased observations in its vicinity. As already shown in this study, biases seem to affect the estimated impact, particularly. For that reason, different bias sensitivity experiments are presented in the following.

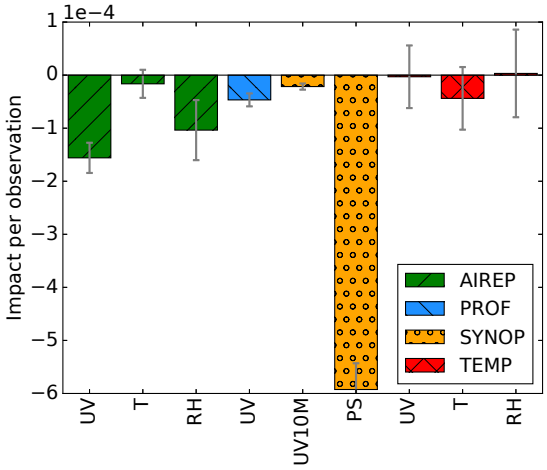


Figure 4.7: *Temporally averaged observation impact per observed variable and type verified with the German gauge adjusted precipitation product. Grey whiskers display the 95 % confidence interval.*

Biases in assimilation and verification

The first experiment investigates the sensitivity of the impact estimate to different artificially added aircraft temperature biases using the *CONV* metric for verification (Figure 4.8). The reference observation impact without added bias is calculated for two randomly chosen days in May 2014. More than 10.000 aircraft temperature observations are assimilated on these days. Apart from aircraft humidity, all other observation types provide at least 2.500 and up to 40.000 observations during the experiment. Two artificial temperature biases of 0.1 K and 0.5 K are added to every aircraft temperature observation to evaluate the sensitivity. The *CONV* metric excluding surface pressure is chosen for verification, and the same artificial bias is added to aircraft temperature observations for both assimilation and verification.
The introduced artificial bias leads to a larger aircraft temperature impact, which increases with bias. Already a small bias of 0.1 K has a significant influence on the impact. Furthermore, the verification with biased aircraft temperature observations affects the impact of related observation types. The second strongest influence is seen for radiosonde temperature, and also surface pressure impact is affected. The impact of radiosonde temperature is here detrimental for a bias of 0.5 K, as the verification is performed with biased temperature observations. The impact of surface pressure is more beneficial if aircraft temperature is warmer than the model. Besides the sensitivity experiments for aircraft temperature, experiments for surface pressure have been performed (not shown here), which lead to similar conclusions. In those experiments, already a bias of 5 Pa affected the observation impact for surface pressure.

Artificial bias added to verifying observations

The second bias sensitivity experiment is performed for the verification with radar-derived precipitation observations. For this experiment, the artificial bias is only added in the verification metric as radar or precipitation observations are not assimilated. A coarse-graining of 10×10 grid cells is applied as previously. The bias is added such that the

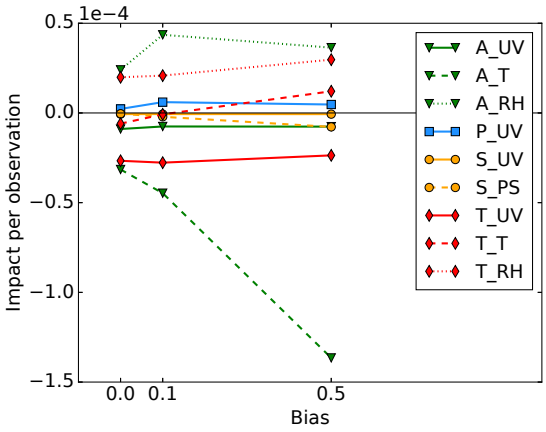


Figure 4.8: *Temporally averaged observation impact per observed variable verified with the CONV metric excluding surface pressure for different aircraft temperature biases (K). Wind (solid), humidity (dotted), pressure, and temperature (dashed). AIREP/A_ (triangle), PROF/P_ (square), SYNOP/S_ (circle) and TEMP/T_ (diamond).*

precipitation bias increases with precipitation amount (multiplication of observed precipitation data with factor 0.9, 0.95, 1.05, and 1.1). In general, COSMO-DE systematically underestimates the amount of precipitation, compared to the European radar product by 13 % (10 % for the German radar product). The entire 6-week period is evaluated, as the observation impact calculation is inexpensive, and the bias is only added in the verification. Only observation types that exhibit sufficient confidence are considered. For this reason, radiosonde wind/temperature/humidity, as well as aircraft humidity observations, are not displayed.

Figure 4.9 shows the change in the estimated observation impact for ±5 % and ±10 % precipitation bias. For a negative sign, the amount of observed precipitation is reduced, and for a positive sign, the precipitation bias is increased. A bias of ±5 % has some effect on the estimated impact, but the order of the impact for different observation types is the same, and the changes are within the estimated confidence interval (not shown). Even for a large bias of ±10 %, the changes in the impact are not that large in relative terms compared to the magnitude of the introduced artificial bias. Only for -10 % the order changes as aircraft temperature exhibits a smaller impact than 10-m wind. Overall, a precipitation bias does influence the results but appears to be less harmful than, for instance, an aircraft temperature bias that acts in both assimilation and verification. This finding underlines the benefits of independent observation-based verification metrics for a reliable observation impact assessment.

Coarse graining

The last sensitivity experiment investigates the effect of different coarse-graining kernel sizes. As already mentioned above, all results presented previously are obtained for a coarse-graining kernel size of 10 × 10 grid cells. Figure 4.10 shows the observation impact for 5 × 5 (effective COSMO model resolution) as well as 10 × 10 (resolvable scale for precipitation), 20 × 20 and 30 × 30 grid cells. Again, the small samples of radiosonde and aircraft humidity observations are not displayed due to their low confidence. The

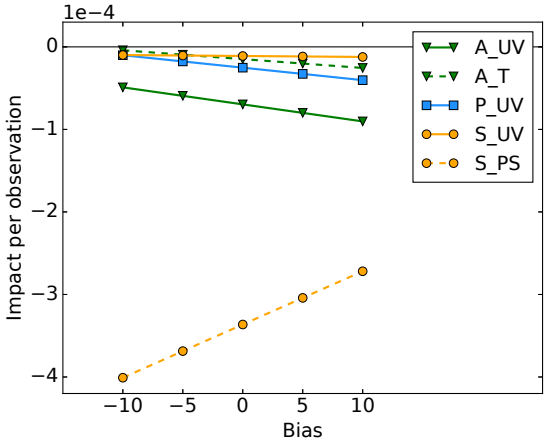


Figure 4.9: *Temporally averaged observation impact per observed variable verified with the PREC metric for different precipitation biases (%).*

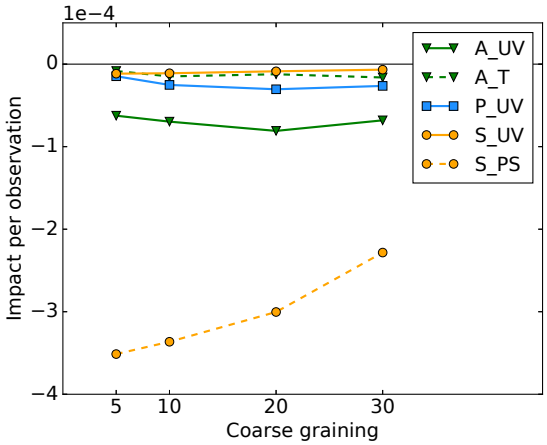


Figure 4.10: *Temporally averaged observation impact per observed variable verified with the PREC metric for different coarse-graining kernel sizes.*

strongest changes are seen for kernel sizes smaller than 10×10 grid cells. The changes for larger kernel sizes are smaller, and the order remains the same from the beneficial to detrimental observations. In general, the sensitivity to the kernel size used for verification seems to be smaller compared to the sensitivity to a bias of observations that are used for both assimilation and verification (see Section 4.2). Furthermore, a coarse-graining is a reasonable choice to reduce double penalty errors and to account for the resolvable scale of precipitation.

4.3 Fraction of beneficial observations

Several previous studies on FSOI and EFSOI discussed the fraction of beneficial observations (e.g. Gelaro et al. (2010); Lorenc and Marriott (2013); Sommer and Weissmann (2014); Jung et al. (2013); Hotta et al. (2017); Lien et al. (2018); Cardinali (2018)). Most studies reported that the ratio of beneficial observations is in the range of about 50 % to 55 %, but some studies showed a notably larger fraction of beneficial observations. Jung et al. (2013), for example, reported 66 % to 72 % beneficial observations for 6-h forecasts

impacts verified with the model analysis and attributed the larger fraction to a stronger correlation of analysis and verification state at this short lead time. Hotta et al. (2017) discovered that the fraction becomes larger as the lead time gets shorter, which supports the argument that correlation of the analysis and verification state leads to a large fraction of beneficial observations. Furthermore, Lien et al. (2018) found fractions of up to 70 % under certain conditions for assimilating precipitation, which may be related to systematic model deficiencies. Overall, most studies showing a significantly larger fraction indicate somehow that this may be related to the correlation of the analysis and verification state or systematic model deficiencies that would spuriously enhance such a correlation.
In the present experiments, a fraction of beneficial observations around 52 % is found for the majority of observation types. The fraction of beneficial observations varies only by a few percents except for surface pressure and temperature observations. Figure 4.11a displays the fraction of beneficial observations per observed variable verified with the *CONV* metric. For surface pressure, 67 % of the observations show a beneficial impact. This fraction seems to be related to the previously discussed pressure bias that is systematically corrected during the assimilation. As the surface pressure impact is verified with later surface pressure observations (self-verification), the majority of observations appear beneficial. The same is evident for aircraft temperature observations, which exhibit the second largest fraction of beneficial observations. Only radiosonde temperature observations have a larger fraction of detrimental observations as their verification is mainly performed with biased aircraft temperature observations.

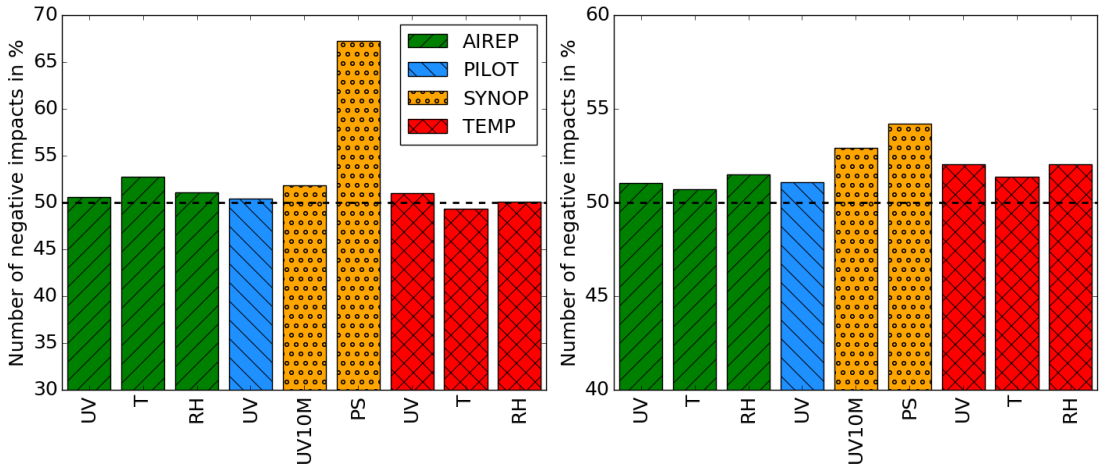


Figure 4.11: *Fraction of beneficial observations per observed variable verified with (a) the CONV metric and (b) the PREC metric. Note: The vertical axis for subplot (a) and (b) is different.*

The bias sensitivity experiments performed in Section 4.2 emphasize this conclusion. Adding a bias of 0.5 K to the aircraft temperature observations increased the fraction of beneficial aircraft temperature observations from 57 % to 67 % (not shown). These results indicate that the fraction of beneficial observations may be used to identify biases through an exceptionally small or large number of beneficial observations.
For the *PREC* metric (Figure 4.11b), the fraction of beneficial observations for all obser-

vation types ranges from 51 % to 54 %. The largest fraction is seen for surface pressure observations, whereas aircraft temperature observations exhibit the lowest fraction. The result that the fraction lies close to 50 % for all types supports the assumption that correlation between the analysis and verification state is less of an issue when using independent observations.

4.4 Influence of high-impact observations

The distribution of impact values is remarkably wide compared to its calculated mean impact. Sommer and Weissmann (2016) discovered that the distribution roughly resembles an asymmetrically stretched exponential. Taking this into account, it is obvious that there is only a small number of extreme impact values compared to a large number of moderate impact values. Based on the distribution of observation impact values, it is possible to investigate the relative contribution of a few high-impact observations versus the majority of observations with low impacts.

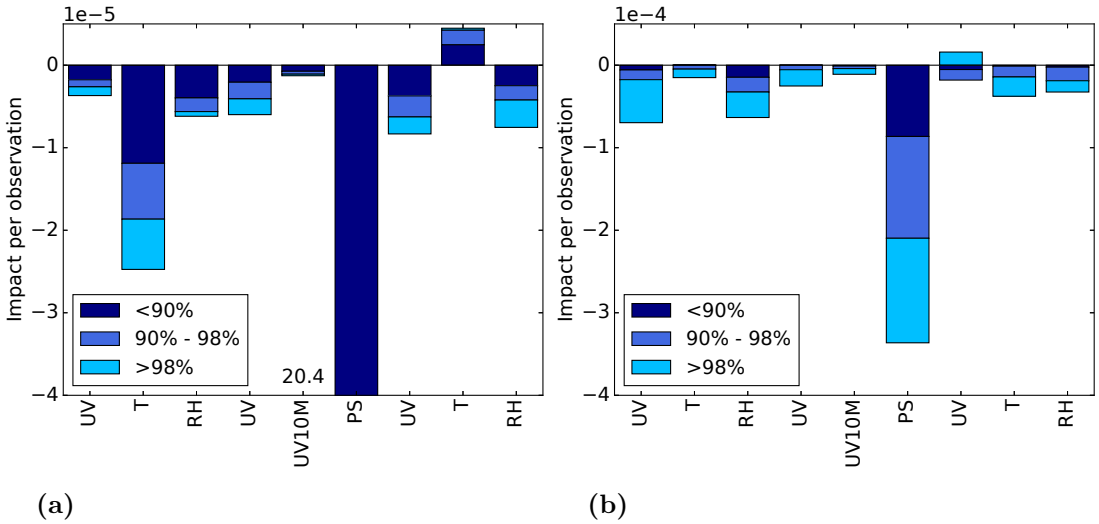


Figure 4.12: *Temporally averaged observation impact per observed variable verified with the CONV metric (a) and PREC metric (b). The differently shaded bars represent the average impact for defined percentile ranges. For instance, the bar with the brightest shading shows the average impact of the 2 % most extreme impact values. The order of observation types is the same as in Figure 4.11a. Note: the bar of surface pressure in (a) exceeds the regime scale by five times. (a) corresponds to Figure 4.2a and (b) to Figure 4.6a.*

Figure 4.12a (corresponding to Figure 4.2a) shows the averaged observation impact per observed variable verified with the *CONV* metric. The bar of each observation type is now additionally divided into the average impact of three different percentile ranges: moderate impacts (smallest 90 %), large impacts (90 % to 98 %) and extreme impacts (most extreme 2 %). Each percentile range contains both positive and negative impacts, but the classification is done regardless of the sign. Comparing the reduction of forecast error by moderate, large and extreme impact values, it turns out that approximately the 10 % observations with the highest impacts contribute as much as the other 90 % with lower impacts. The

ratio of the contribution of large and extreme impact values is more or less the same for most observation types using the *CONV* metric. For the *PREC* metric (Figure 4.12b corresponding to Figure 4.6a) the contribution of few observations with extreme impact values even increases. The extreme impact values now contribute a significant fraction, whereas for the *CONV* metric the 10 % most extreme impact values contributed about half of the total impact. This difference could be related to the non-Gaussian distribution of precipitation with a large fraction of outliers.
Considering the stacked impact of radiosonde wind observations (Figure 4.12b) helps understanding the almost neutral impact obtained using the *PREC* metric (Figure 4.6a). The 98 % radiosonde wind observations with smallest impacts are on average beneficial, but as the 2 % most extreme impact values exhibit a detrimental impact of the same amplitude, the average impact of all radiosonde wind observations is neutral. A similar but reverse effect can be observed for aircraft temperature observations. Here, moderate impact values on average show a neutral impact, whereas large and extreme impact values are on average beneficial. This difference could be related to the previously mentioned aircraft temperature bias. However, for the radiosonde wind observations, the small sample size seems to be relevant for the different behavior of moderate and extreme impact values.
In general, all this relates to the idea of target observations in particularly sensitive areas, which have been subject of several recent studies and field campaigns (Majumdar et al., 2011). Observational efforts could be significantly reduced if these high impact observations could be identified a priori through sensitive area calculations and observation targeting. However, further research is required to understand observational needs for convective-scale NWP better and develop approaches for observation targeting on these scales (Gustafsson et al., 2018).

4.5 Representativity of the results for summer period

The applied two-sided t-test showed that the results for most observation types are reliable for a 6-week period. Another option to evaluate the representativity of the results is to look at the temporally ordered partial mean observation impact (Figure 4.13). The idea is to check if the temporally averaged observation impact is saturated for a certain subsample of observations and use this information as a representativity measure. This approach also provides information if the results vary within a longer period and if there is variability due to different weather regimes. A larger difference between the partial sum and the period average impact indicates a larger variability. Fast convergence of the ordered partial sum towards the temporal average corresponds to lower temporal variability of the estimated impact.
For most observations types the mean impact is reached before the end of period (not all shown). This indicates that the results are representative for a high-impact summer weather period with various different weather regimes. For verification with the *PREC* metric (Figure 4.13b & 4.13d) the partial mean impact converges slower than for the verification with conventional observations (Figure 4.13a & 4.13c), which is in agreement with the smaller confidence obtained using the t-test. Furthermore, it is seen that for radiosonde observations (see for example Figure 4.13c & 4.13d) a larger sample size should be considered to obtain robust results. Finally, it needs to be mentioned that the estimated

impact to a certain extent depends on the considered experimental period and for instance could change for a different period, weather regime or season.

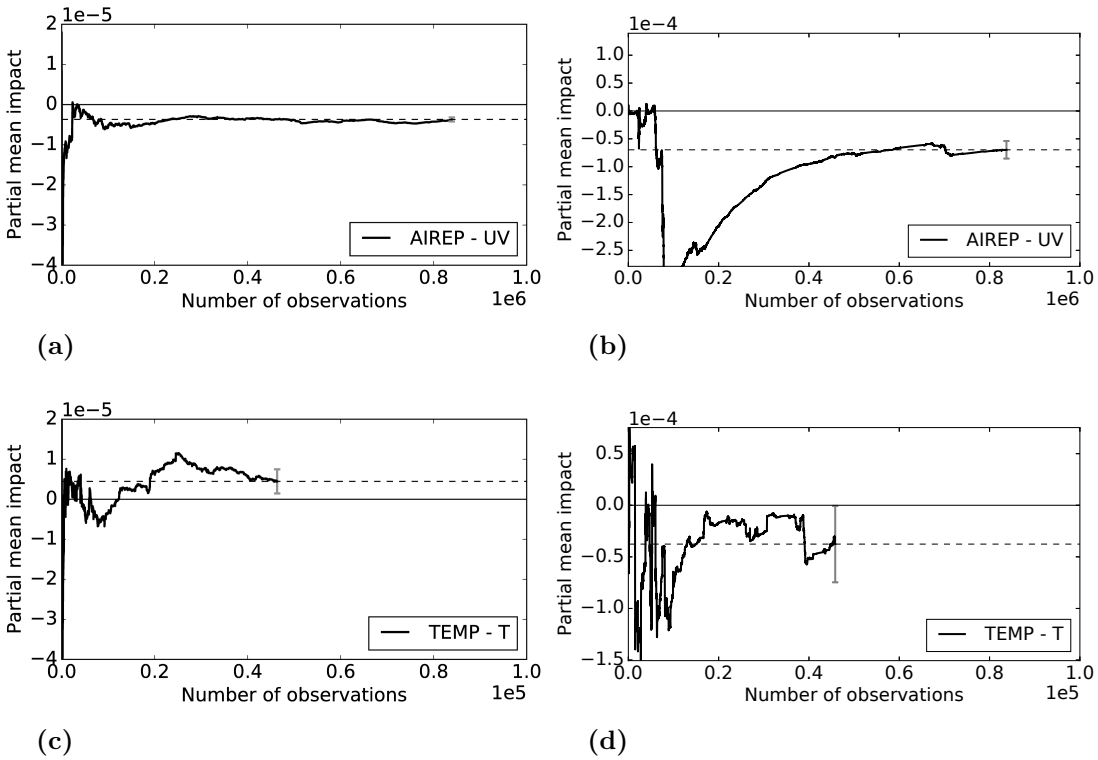


Figure 4.13: *Partial mean observation impact (solid) and temporally averaged observation impact (dashed) for verification with the CONV (a, c) and the PREC (b, d) metric. Aircraft wind (a, b) and radiosonde temperature (c, d). Grey whiskers display the 95 % confidence interval obtained with a two-sided t-test.*

4.6 Wind profiler inter-comparison

Recently, several impact studies initiated by the European Meteorological Network (EUMETNET) assessed the impact of various observation types provided by the European observing network. One of these studies particularly focused on the E-PROFILE network that consists of aircraft, wind profiler, and lidar/ceilometer observations. This impact study was conducted in collaboration by the Swiss weather service (MeteoSwiss), Deutscher Wetterdienst, and the ECMWF. Both aircraft and wind profiler impacts were assessed using the routinely monitored observation impact within the ECMWF forecasting system. In the ECMWF system, the impact is calculated applying an adjoint-based FSOI approach for 24-h forecasts verified with a dry kinetic energy norm (Cardinali, 2009).
Overall, Ruefenacht et al. (2018) found that the wind profiler network significantly contributes to the reduction in forecast error, especially in regions with few aircraft wind measurements. However, two wind profilers in Switzerland exhibited a mainly neutral (Schaffhausen) or detrimental (Payerne) observation impact. Therefore, the Swiss wind profilers are evaluated for the summer period in 2016 to determine if this behavior is also seen in the convective-scale COSMO-KENDA system and for short-range forecasts.

Figure 4.14 shows a comparison of all wind profilers that are located in the COSMO-DE domain (see also map in Figure 3.1). Each wind profiler provides a significant amount of wind observations and at least 7000 assimilated observations per station (Figure 4.14a). This means each wind profiler contributes a similar amount of or even more wind observations as provided by all radiosonde wind measurements together. Figure 4.14b examines the total impact per wind profiler verified with the *CONV* metric. Performing a self-verification, the wind profiler in Payerne exhibits the largest beneficial impact of all stations. However, if verified with independent radar precipitation observations (Figure 4.14c), only the Payerne wind profiler shows a significant detrimental impact. This discrepancy between both verification metrics is similar to the impact results found for aircraft temperature. It seems that systematic effects for the wind profiler in Payerne cause a detrimental impact. This finding confirms the results from the European wind profiler inter-comparison. The second Swiss wind profiler in Schaffhausen exhibits a beneficial impact independent of the verification metric.

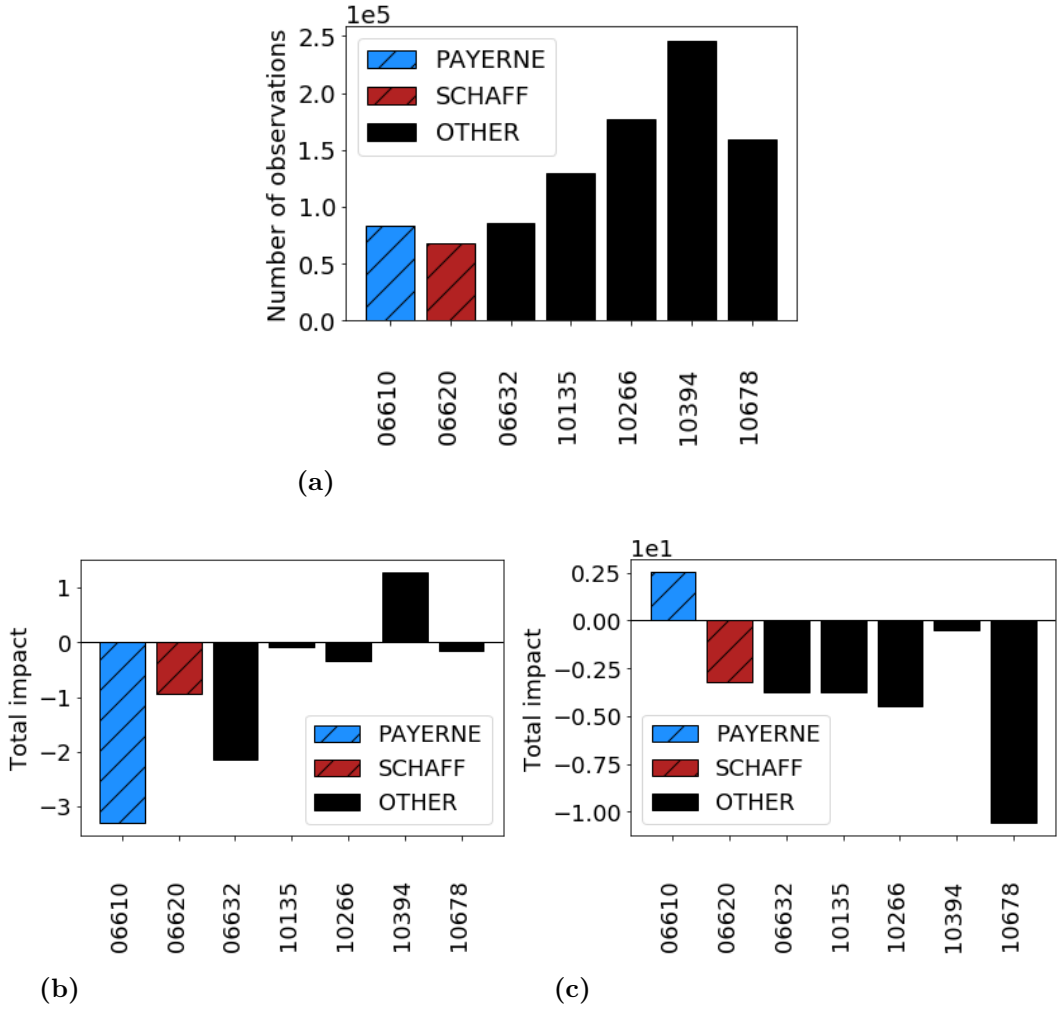


Figure 4.14: *Number of assimilated wind profiler observations per station during the 6-week summer period 2016. Corresponding total observation impact per wind profiler verified with (b) CONV and (c) PREC metric. Wind profiler over Switzerland (colorized) and Germany (black). The station number is plotted on the x-axis.*

4.7 Summary and discussion

This study presents the first evaluation of observation impact in a convection-permitting modeling system over an extended period of six weeks. The short-range observation impact of about 3.3 million conventional observations is analyzed in a 6-week summer period using the regional ensemble data assimilation system of DWD. The observation impact was evaluated with the EFSOI method that was initially published by Kalnay et al. (2012) and reformulated by Sommer and Weissmann (2016). Moreover, the observation impact is computed using several different observation-based verification metrics, including independent radar observations.

The results highlighted the sensitivity of the calculated impact on the choice of the verification metric, particularly in the presence of model or observational biases. Sensitivity experiments showed that already small biases can significantly affect the estimated impact if the verification metric is based on observation types that have been assimilated (self-verification). Nearly all previous studies on EFSOI and FSOI used analysis-based verification metrics, which are potentially also affected by biases that lead to a correlation of the analysis and verification state. Especially for the impact assessment of short-term impact during strongly convective events, the use of analysis-based verification norms, therefore, seems potentially dangerous. Furthermore, the use of a total energy norm as verification metric seems inappropriate for convective-scale modeling systems as such a norm does not reflect primary forecast quantities (as, e.g., precipitation). Hence, the observation impact should be assessed comparing results for different verification metrics and including independent observations for verification. Additionally, the assessment of observation impact needs to be accompanied by careful monitoring of potential biases and exclusion of biased observations (or analysis fields) from the verification metric.

In this study, radar-derived and rain-gauge adjusted precipitation fields were used as an additional independent verification metric that represents a specific, but particularly important aspect of the forecast. Sensitivity studies showed that a bias of such independent observations used for verification has a smaller effect on observation impact results than if assimilated observation types are used. This finding is of interest for the implementation of quality control schemes that are mostly based on short-range FSOI and aim to detect and exclude detrimental observations (Hotta et al. (2017); Lien et al. (2018)).

Using independent radar precipitation fields for verification, surface pressure showed the largest beneficial impact followed by aircraft wind observations. Wind profiler observations had the third largest total impact as they are the most frequent observation type. However, they exhibited a smaller impact per observation compared to the verification with conventional observations that favors observations with a higher measurement frequency (Sommer and Weissmann, 2016). The average impact per observation of radiosonde temperature and humidity, as well as aircraft humidity, was comparable to aircraft winds, but results should be treated with caution given the comparably small number of evaluated observations for these observation types. Aircraft temperature exhibited a much smaller impact per observation compared to the verification with conventional observations, which is likely associated with uncorrected observational biases. The development of a bias correction scheme for aircraft temperature observations is currently ongoing at DWD.

Furthermore, it was demonstrated that EFSOI could be used to identify observation and model biases. One indicator for biases is a significant difference in the impact of different

verification metrics. Additionally, an exceptionally high or low fraction of beneficial observations appears to be an indicator of biases in the assimilation or verification as discussed for surface pressure and aircraft temperature.
For all experiments, a two-sided t-test was performed to calculate the 95 % confidence interval of the calculated observation impact to quantify the reliability of the results. Additionally, a new approach to estimate the representativity of the calculated impact was presented, which uses the partial mean of the impact as a reliability measure. Both approaches suggest that a larger sample size of radiosonde wind/temperature/humidity, as well as aircraft humidity observations, is needed to make a reliable statement about the actual observation impact, while for the other quantities the results are already reliable.
Recently, the presented EFSOI results could confirm an observation impact inter-comparison study performed for the E-PROFILE network based on the ECMWF FSOI system. Ruefenacht et al. (2018) found a systematically detrimental impact for the Swiss wind profiler located in Payerne. The present study for summer 2016 similarly diagnosed a detrimental impact for Payerne for the COSMO-KENDA system. Notably, the difference in the impact seen for the *CONV* and *PREC* metric suggests that systematic differences between model and observations likely caused the detrimental impact. Currently, MeteoSwiss is investigating the cause of this deficiency.

Chapter 5

Potential impact of observable quantities

Nowadays, a vast amount of observational information is available in developed countries. Additionally, technological developments enable to install new or additional observing systems. However, more observations are available than can be assimilated due to limited human and computational resources. Given those developments, advanced data selection strategies are crucial to optimize data assimilation processes in the future. In general, knowledge of the potential impact of observations can help to focus on the assimilation of observations that will likely have a significant impact on the forecast skill. Hence, such knowledge is expected to increase the cost-benefit ratio of the NWP system.

The goal of this chapter is to provide guidance toward improved data assimilation strategies. This chapter, therefore, presents an efficient approach that uses spatiotemporal correlations as a proxy for the relative potential impact of observable quantities. The approach has initially been proposed by Geiss (2017) and is based on ensemble sensitivity analysis (ESA). Compared to Geiss (2017), the present study applies an improved 1000-member ensemble simulation. Moreover, the relative potential impact is computed using a refined methodology. This refinement, among other things, includes a scale analysis. As has been outlined, this thesis focuses on challenges in convective-scale NWP. Section 5.3, therefore, will explain that it is crucial to distinguish the relative potential impact depending on the spatial scale of correlations. The present study aims to provide the reader with the central concept of the approach. Some parts of this chapter have been submitted for publication by Necker et al. (2019a) (see Appendix A).

Chapter 5 starts with an investigation of a case study examining ensemble sensitivities for different quantities. The same case study is discussed in Section 3.4 and 6.1 but in a different context. Furthermore, the first section qualitatively addresses the issue of undersampling comparing correlations obtained for different quantities and ensemble subsets of the 1000-member ensemble. Section 5.2 exemplarily presents the relative potential impact for seven different quantities. Additionally, this section includes an analysis examining the sensitivity of the relative potential impact on the applied ensemble size. In this context, the SEC and T95 are applied to reduce sampling error. Both approaches will be evaluated in detail in Section 6.1 and can reduce sampling error applied to ESA. Finally, Section 5.3 highlights the scale-dependence of the potential impact.

5.1 Ensemble sensitivity analysis

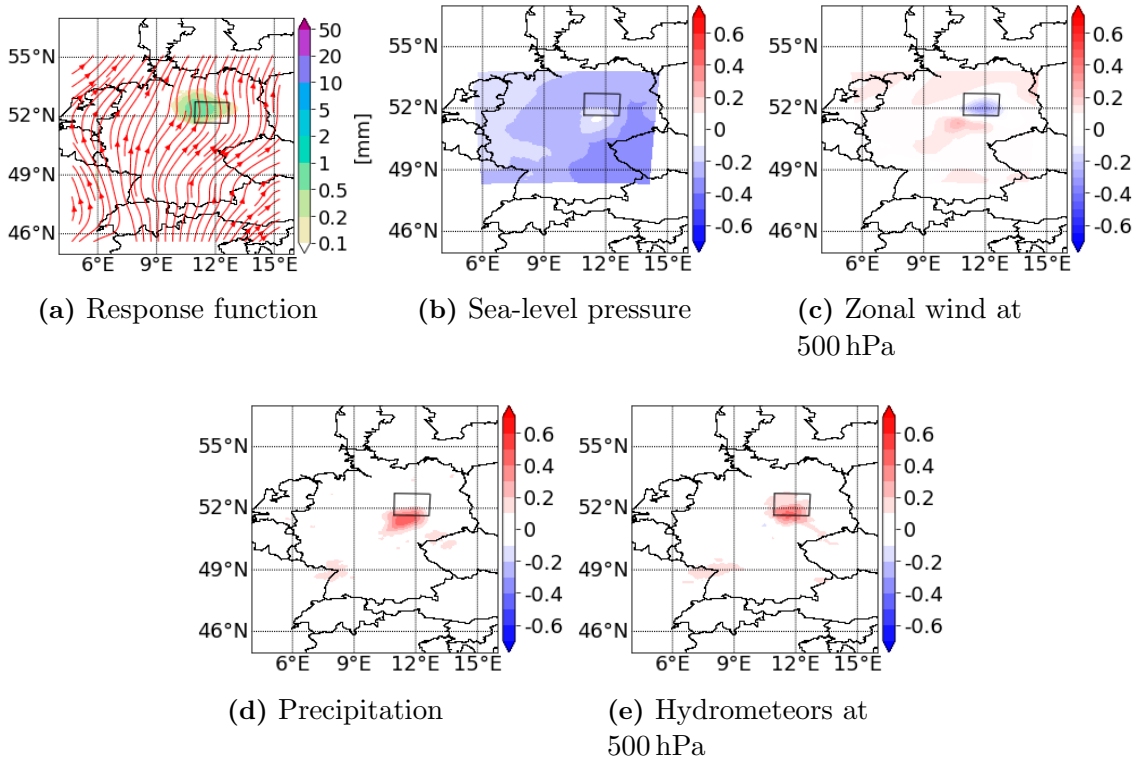
Example of correlation fields Ensemble sensitivity analysis is used to compute spatiotemporal correlations for the 1000-member ensemble as well as two random subsets. Ensemble subsets are generated identically as for spatial correlations (Section 3.5). All results in this chapter focus on short-range forecasts with a lead time of 3-h. The response function is fixed at 4-h lead time, and the 1-h forecast is used as the initial state. Figure 5.1a shows the 1000-member ensemble mean precipitation forecast at 29 May 2016 4 UTC including streamlines of 500 hPa wind. The black box marks the position of the response function that is used to calculate the spatiotemporal correlations for the investigation of the sensitivity of this precipitation system.

The precipitation forecast and initial sea-level pressure field are negatively correlated (Figure 5.1b). This correlation reveals that lower pressure coincides with stronger precipitation in the ensemble. A small-scale structure with correlation values near zero is embedded slightly south of the response function within the relatively smooth large-scale correlation field. This small-scale structure roughly matches the position of the precipitating system at the beginning of the forecast and likely corresponds to surface cooling due to evaporating precipitation. The correlation field of initial 500 hPa zonal wind (Figure 5.1c) exhibits a dipole structure. In this case, the dipole seems to indicate stronger cyclonic shear in the south of the box.

Figure 5.1d shows the spatiotemporal correlation of precipitation inside the response function to earlier precipitation. Precipitation is positively correlated with itself as initially stronger precipitation correlates with increased precipitation three hours later. A similar correlation signal can be observed for hydrometeors (Figure 5.1e). Here, hydrometeors are composed of specific cloud water, rain, ice, snow, and graupel content. For hydrometeors, the region of maximum correlation is slightly shifted northwards compared to the precipitation. This shift could originate from accumulation of the precipitation over one forecast hour or the fact that hydrometeors appear before precipitation is observed at the ground. Furthermore, both precipitation and hydrometeors exhibit a weak positive correlation signal over south-east Germany, which is caused by precipitation in this region in some of the 1000 members.

Similar to sea-level pressure, the upper-air temperature (Figure 5.2a) reveals a rather smooth and large-scale correlation pattern with negative values, but positive correlation values in the vicinity of the precipitating system that are likely related to the release of latent heat in the precipitating system. The correlation of the specific humidity at 850 hPa (Figure 5.2d) is weaker and only extends over a smaller area compared to temperature. The elongated tail roughly marks the track of the precipitating system during the night indicated by the streamlines in Figure 5.1a. It seems that the humidity signal reflects precipitation that took place already before analysis time. The maximum correlation is located in the same region as for temperature, showing a positive correlation of specific humidity and precipitation intensity.

Overall, the correlations obtained from the 1000-member ensemble depict physical processes that contribute to the evolution of the precipitating system. Consequently, these correlations are suitable as a proxy for the relative potential impact. In this case, if the initial conditions are warmer, more humid and exhibit a higher amount of hydrometers at 500 hPa the resulting precipitation is more intense. The same applies to initially lower



(a) Response function **(b)** Sea-level pressure **(c)** Zonal wind at 500 hPa

(d) Precipitation **(e)** Hydrometeors at 500 hPa

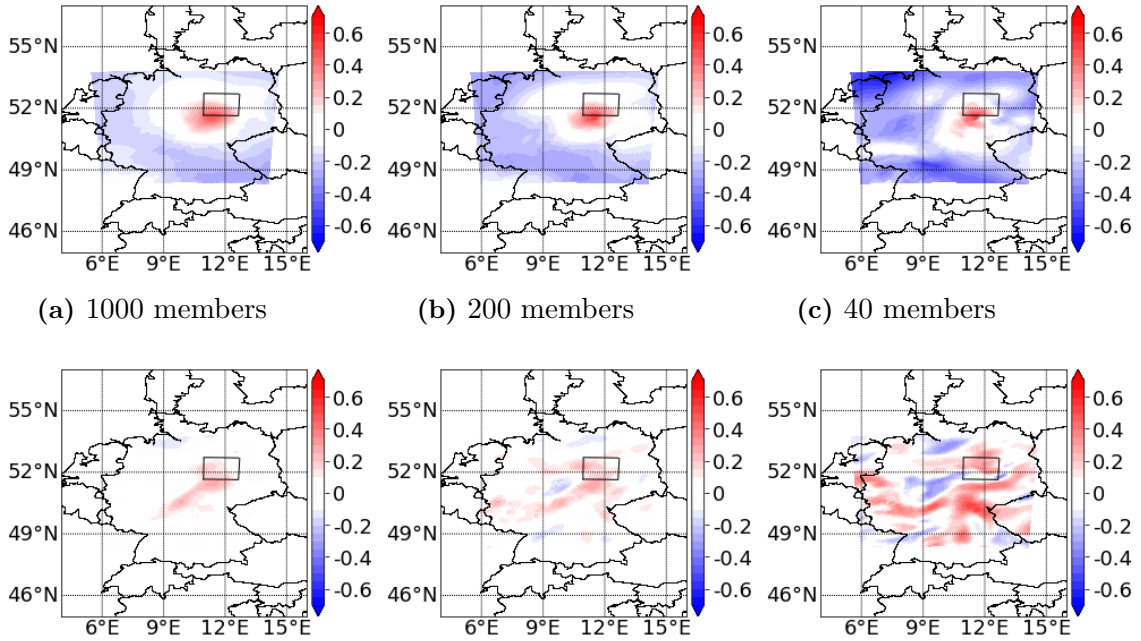
Figure 5.1: *Sensitivity of the 3 h precipitation forecast inside the box to different initial model fields using the 1000-member ensemble, 29 May 2016 01 UTC.*

pressure and stronger precipitation.

Properties of sampling errors The subsequent paragraph presents examples of sampling errors for spatiotemporal correlations of precipitation to two representative variables (500 hPa temperature and 850 hPa specific humidity) using two ensemble subsets (40 and 200 members). The 500 hPa temperature correlation pattern is exemplary for other variables with large-scale correlation patterns (e.g., pressure), whereas 850 hPa specific humidity is representative for variables that exhibit small-scale structures in the correlation field (as, e.g., surface quantities or hydrometeors and precipitation). The 1000-member ensemble again serves as a reference.

Reducing the ensemble size from 1000 to 200 members-only leads to moderate changes for the correlation to 500 hPa temperature (Figure 5.2b). The region of positive correlation still looks fairly similar regarding position and magnitude, but negative correlations farther away are systematically larger in magnitude due to spurious oscillations. Differences moving to a 40-member ensemble are significantly larger (Figure 5.2c). The local positive correlation pattern lost its shape and negative correlations farther away intensified even further due to sampling errors. Nevertheless, the 40-member ensemble still provides qualitative information as it captures the overall structure and sign of the correlation field for 500 hPa temperature.

Figure 5.2e shows the sensitivity of precipitation to specific humidity at 850 hPa using a 200-member ensemble. Similar to temperature, weak spurious correlations appear in large parts of the domain, but the region of maximum correlation as well as its elongated tail



(a) 1000 members **(b)** 200 members **(c)** 40 members

(d) 1000 members **(e)** 200 members **(f)** 40 members

Figure 5.2: *Same as Figure 5.1, but for the sensitivity of the 3 h precipitation forecast to temperature at 500 hPa (top row) and specific humidity at 850 hPa (bottom row) for different ensemble sizes, 29 May 2016 01 UTC.*

are well-captured. Lowering the ensemble size to 40 members (Figure 5.2f) significantly increases sampling errors, and spurious correlations now dominate the correlation field. These results suggest that the 40-member ensemble can provide qualitative information for large-scale patterns, but struggles to estimate correlations for more variable fields as, 850 hPa humidity, hydrometeors, or surface temperature. The 200-member ensemble provides reasonable correlation patterns for all variables, but the fields are still affected by spurious correlations. Caution is especially necessary when using correlations in a quantitative sense as in the following.

5.2 Estimating the relative potential impact

This section introduces an approach for investigating the relative potential impact of observable quantities for improving precipitation forecasts. Again, the focus is on spatiotemporal correlations of precipitation obtained for 3-h lead time forecasts. As discussed in the Section 2.3, the accumulated squared correlation (ASC) is used as a proxy for the relative potential impact. Accordingly, the squared correlations of the precipitation forecast with an initial condition variable are accumulated over the whole domain and all available forecasts to estimate the relative importance of that variable for data assimilation.

Senstivity to ensemble size Figure 5.3 shows the time-averaged ASC as a function of ensemble size between the 3-h precipitation forecast and zonal wind at 500 hPa (Figure 5.3a), 2-m temperature (Figure 5.3b) and precipitation (Figure 5.3c), respectively, at

the initial time. For all variables, the ASC evaluating small ensembles is significantly overestimated due to spurious correlations. For instance, the ASC_{1000} is overestimated by more than 200 % using a 40-member ensemble. Generally, the ASC strongly decreases with increased ensemble size, but hardly changes from 600 to 1000 members. Given this saturation for large samples, spatiotemporal correlations obtained with a 1000-member ensemble are presumably reliable estimates.

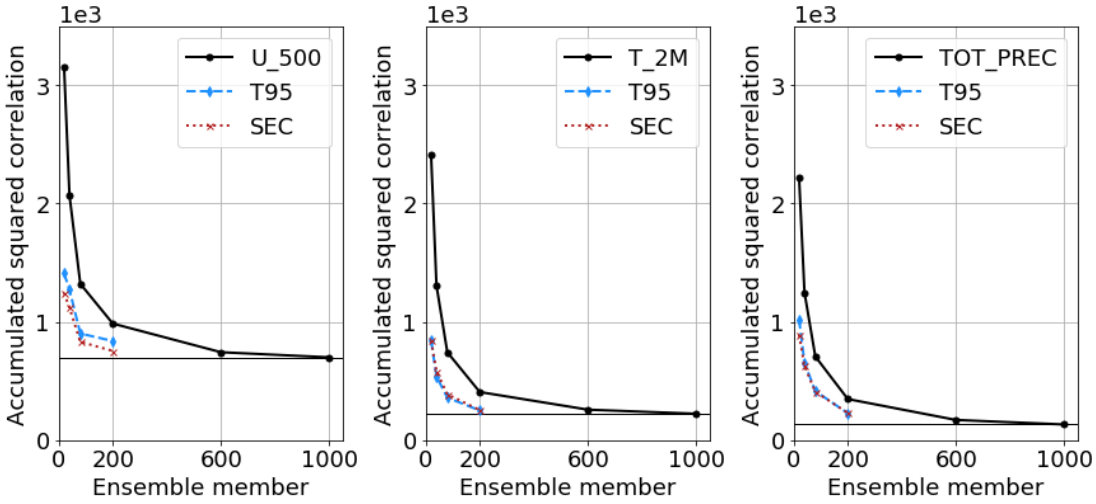
Chapter 6 shows that the SEC significantly reduces sampling errors for spatiotemporal correlations as they are calculated within ESA and for spatial correlations, as they are required in data assimilation. As the potential impact approach is based on ESA, it is also possible to mitigate sampling errors computing the ASC. Sampling errors are addressed using two different approaches. The first is the confidence test (T95) that excludes insignificant correlations (Torn and Hakim, 2008). The second is the statistical sampling error correction (SEC; Anderson (2012)). In general, compared to the T95, the SEC does not fully exclude small correlations, which is favorable for quantitative analysis. Moreover, the SEC systematically reduces correlation values to account for the overestimation of correlations due to spurious correlations, which corrects the magnitude bias due to spurious correlations. A detailed evaluation of both approaches for spatiotemporal correlations follows in Section 6.1.

Examining 500 hPa zonal wind (Figure 5.3a), both approaches significantly improve the ASC estimate for small samples. The SEC performs slightly better compared to the T95 and results for the sampling error corrected 200-member ensemble are already close to the ASC_{1000}. Improvements are similar for other variables (Figure 5.3b and Figure 5.3c). Overall, 200 members including SEC seem to be a reasonable choice for estimating the ASC if no 1000-member ensemble is available. However, it should be noted that there is a small remaining error in the estimate and that the error differs for different variables as sampling errors tend to be higher for smaller-scale fields. This effect can lead to a systematic over- or underestimation of the relative potential of the respective variable. For smaller ensembles (e.g., 40 members), this effect is even larger, and it seems questionable if smaller ensembles are applicable for such a quantitative evaluation of correlations.

Relative potential impact Figure 5.4a shows the time-averaged ASC_{1000} for seven different variables using the 1000-member ensemble. Before the discussion, it should be noted that the primary purpose here is the evaluation of the appropriate ensemble size for such an application. Additionally, the need for a scale-decomposition is highlighted in the last section of this chapter. The relative potential impact of various observable quantities needs to be analyzed in more detail by consecutive studies.

Sea-level pressure (PS) exhibits the largest ASC_{1000}, followed by wind at 500 hPa, and 10-m height. Precipitation has a smaller sensitivity to initial perturbations of temperature and humidity. The smallest ASC is found for precipitation. Applying a confidence test to the 1000-member correlations hardly changes the ASC (Figure 5.4a). This small change confirms the reliability of the results obtained for the 1000-member ensemble.

Using 200 members, the ASC is overestimated for all variables (Figure 5.4b). The largest differences are visible for wind and precipitation. As found before, both the T95 and the SEC significantly improve the ASC (Figure 5.4b and 5.4c). Again, the SEC performs slightly better than the T95. The results for 200 members including the SEC are close to



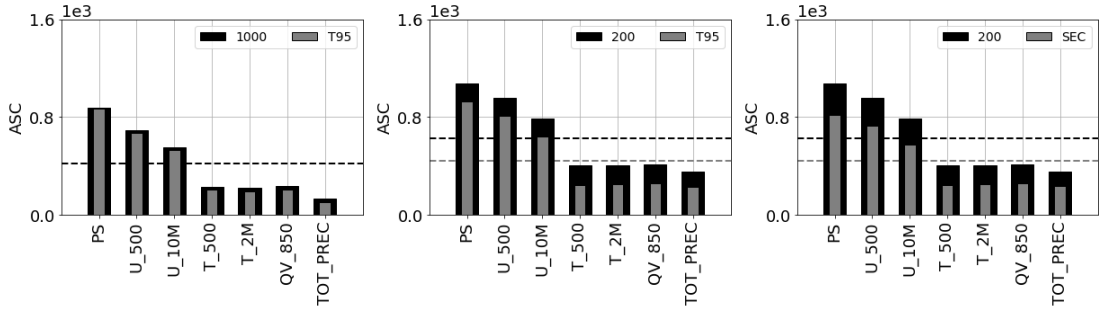
(a) 500 hPa zonal wind **(b)** 2-m temperature **(c)** Precipitation

Figure 5.3: *Time-averaged accumulated squared correlation as function of ensemble size for different variables, 29 May to 03 June 2016. Spatiotemporal correlations of precipitation to (a) 500 hPa zonal wind, (b) 2-m temperature and (c) precipitation.*

the ASC_{1000} but there are still some small differences, for example, an overestimation of the ASC for precipitation.
Using a 40-member ensemble (Figure 5.4d), the ASC is significantly overestimated and the ranking changes compared to examining the ASC_{1000}, even when the SEC is included. For example, the ASC for precipitation now has an equally large or higher impact as specific humidity or surface temperature. Overall, under-sampling causes an overestimation of the ASC, even if the T95 or the SEC are applied (Figure 5.4d and 5.4e). The 40-member ensemble is therefore not able to estimate the ASC or relative ASC reliable. To some extent, the small ensemble can provide qualitative guidance.

5.3 Scale-dependence of the potential impact

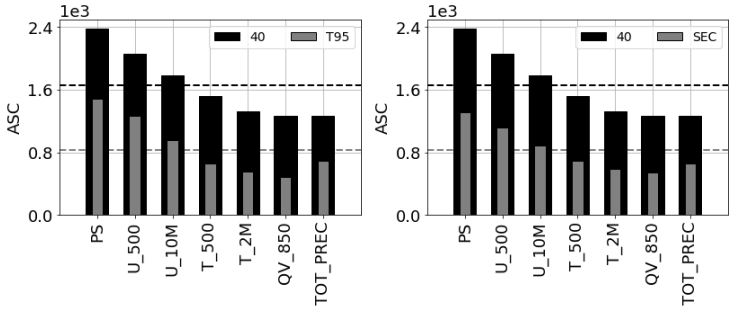
Convective-scale data assimilation requires observations on rapidly changing weather situations. Additionally, the high model resolution calls for dense observational information that resolves clouds or convection. Observations, therefore, need to be able to capture spatial and temporal scales that are targeted by convective-scale DA (Fabry and Sun, 2010, Fabry, 2010). Hence, it is important to distinguish and to account for the spatial scale of observations estimating the relative potential impact. Subsequently, *small* and *large* scales will be divided using a border at 300 km. Accordingly, *small-scale* structures that are of interest range from the effective model resolution of 15 km up to mesoscales of 300 km extent.
As introduced previously, the potential impact measure uses accumulated squared correlations as a proxy for the relative potential impact. For this purpose, spatiotemporal correlations are computed in a regional domain. In general, the impact measure accounts for linear relations combining signals from all spatial scales. Consequently, a sensitivity pattern could be dominated by signals that are also influenced by scales that are not the



(a) 1000 members + T95

(b) 200 members + T95

(c) 200 members + SEC



(d) 40 members + T95

(e) 40 members + SEC

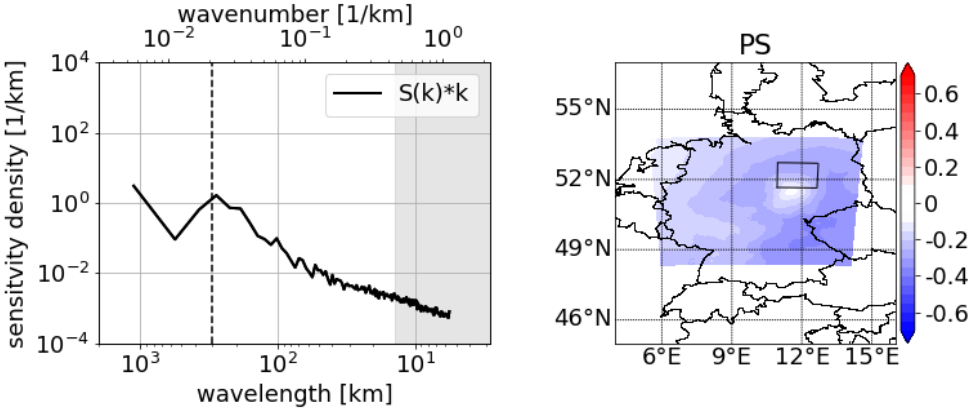
Figure 5.4: *Time-averaged ASC using a precipitation response function for all 3-h forecasts, 29 May to 03 June 2016. 1000-member ensemble (a) as well as sub-sampled 200 (b,c) and 40 (d,e) member ensemble including sampling error correction (SEC) or confidence test (T95). Average ASC of all variables (dashed line).*

target of convective-scale DA. For example, this is the case for sea-level pressure. A pressure gradient within the domain is mainly determined by the boundary conditions, which are provided by the global model. Hence, to obtain a representative impact measure for convective-scale DA, it is necessary to exclude large-scale effects and characterize variables that show the strongest small-scale signals, which are relevant for convective-scale DA.

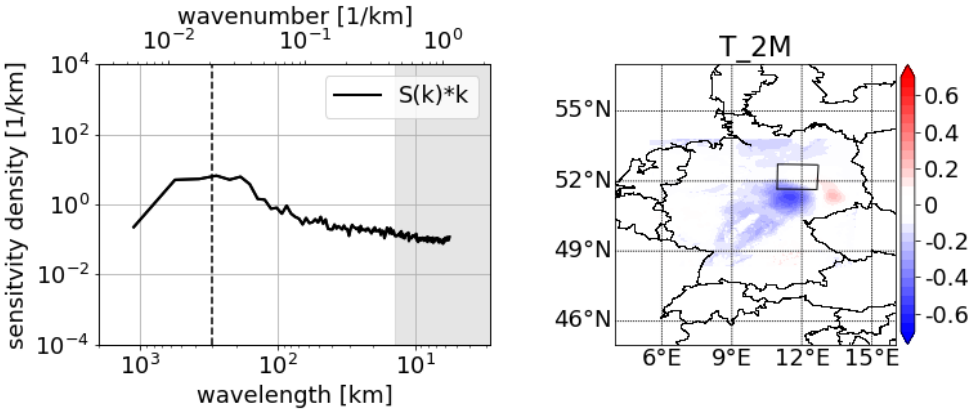
The scale analysis is done using a discrete Fourier transformation and aims to distinguish if a correlation signal is a small or large-scale signal. The scale analysis helps to characterize variables that are dominated by large-scale or boundary effects. Additionally, the scale dependence of the signal intensity can provide information on which observation network density would be desirable. For example, considering a quantity that exhibits a strong small-scale variability, a sparse observation network would not be able to capture relevant small-scale changes.

Figure 5.5a analyzes the correlation of a 3-h precipitation forecast to its initial pressure field. The left graph in Figure 5.5a displays the scale dependence of the correlation signal using a compensated sensitivity density spectrum. Compensated, in this case, means that the sensitivity density is weighted with the wavenumber k (Craig and Selz, 2018). The weighting is done as the results are displayed in logarithmic scales. For sea-level pressure, mainly large-scale effects dominate the sensitivity density spectrum. The wavelength that would fit the pressure signal best is even longer than the domain size. This result agrees

with the large-scale pressure gradient that is visible in the correlation field.
In contrast, the correlation of precipitation to 2-m temperature has a significantly stronger sensitivity density on scales smaller than 200 km (Figure 5.5b). Thus, for scales smaller than 200 km, the precipitation forecast has a higher sensitivity to the 2-m temperature field than to sea-level pressure. Consequently, 2-m temperature exhibits a higher relative potential impact for convective-scale DA than sea-level pressure.



(a) Sea-level pressure



(b) 2-m temperature

Figure 5.5: *Sensitivity of the 3 h precipitation forecast to (a) sea-level pressure and (b) 2-m temperature for 1000 members, 29 May 2016 01 UTC. (left) Compensated sensitivity density spectrum calculated using a Fourier transformation. Grey shading indicates wavelength smaller than the effective model resolution of 14 km. The vertical dashed line marks the border between small and large scales. (right) Corresponding correlations field.*

The goal is now to obtain a similar plot, as shown in Figure 5.4a, but accounting only for scales that are most relevant for convective-scale DA. This goal can be achieved using Parseval's theorem. According to Parseval's theorem, the integral of a squared function is equal to the integral of the squared transform in spectral space. Accordingly, the potential impact defined as the accumulated squared correlation is equivalent to the sum (integral) of squared sensitivities in spectral space. Now, the Fourier transformation can be applied to separate the scales of interest in spectral space, which is similar to applying a bandpass filter (BPS).

Note that both spectra that are displayed in Figure 5.5 do not display the background mean or infinity mode. Especially for sea-level pressure, the mean background correlation dominates the ASC of sea-level pressure in Figure 5.4a. The background mean also is a large-scale effect and, therefore, should be excluded from the analysis. When a bandpass filter is applied, it excludes both the background mean as well as the large-scale correlation signal.
Finally, the most relevant quantities for convective-scale DA can be distinguished. Figure 5.6 displays the ASC after applying a BPS. In this case, only sensitivities that have a scale smaller than 300 km are accumulated, and all large-scale signals are neglected. Sensitivity studies indicate that results hardly change using a threshold of for example 250 km or 200 km. For convective-scale DA, 2-m temperature exhibits the largest bandpass-filtered ASC followed by specific humidity at 850 hPa. Furthermore, assimilating precipitation has stronger relative potential impact compared to other quantities. For short-range precipitation forecasts, wind and temperature at 500 hPa show a relatively small relative potential impact. The same applies to sea-level pressure.

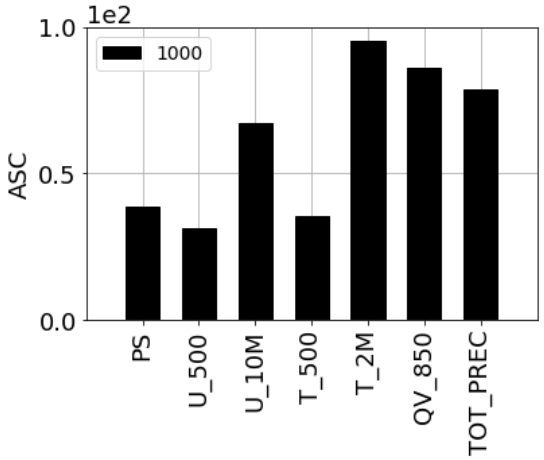


Figure 5.6: *Same as Figure 5.4a, but illustrating the band-pass-filtered ASC accounting only for scales smaller than 300 km.*

In general, results suggest to apply a scale analysis only for a sufficiently large ensemble sample. Spurious correlations mainly act on small scales and thus have a strong influence on estimating the relative potential impact while performing a scale analysis. Studies for a 200-member ensemble including SEC and BPS could provide qualitative guidance but failed to estimate the ASC aiming for a quantitative statement.

5.4 Summary and discussion

This chapter introduces how spatiotemporal correlations can be used to gain knowledge on the potential impact of different observable quantities for data assimilation. In detail, the accumulated squared correlation (ASC) is used as a proxy for the relative potential impact. The ASC is based on spatiotemporal correlations, which are the main ingredient for ESA. The present study computes correlations using a 1000-member ensemble. The ASC is evaluated for precipitation forecasts comparing seven different observable atmospheric

quantities. Furthermore, the sensitivity of the ASC to the ensemble size and its scale dependence is investigated.

Initially, a case study is presented to examine if applied correlations are reliable and meaningful. Therefore, spatiotemporal correlations of precipitation with different initial condition variables are discussed for a nocturnal precipitation event on 29 May 2016. The example shows that ESA using a 1000-member ensemble can return realistic spatiotemporal correlations with respect to precipitation. The 1000-member ensemble can highlight small-scale features that are traceable in space and time. Sensitivity studies on the ensemble size suggest that a 40-member ensemble can provide some qualitative guidance for large-scale patterns. However, more than 200 members are required to detect small-scale structures reliably.

Moreover, sensitivity studies on ensemble size show that a 1000-member ensemble returns reliable estimates of the ASC. A 200-member ensemble can provide sufficient estimates of the ASC if a confidence test or sampling error correction is included. However, some differences to the 1000-member ensemble still occurred for variable fields such as precipitation. Smaller ensembles are not able to estimate the correct amplitude of the ASC but were able to distinguish variables with considerable differences of the ASC. Overall, if a scale analysis is applied to distinguish the relative potential impact for convective-scale DA, only the 1000-member ensemble is able to return reliable estimates.

As discussed in Section 5.3, a scale analysis is required to determine the relative potential impact for convective-scale DA. Spatiotemporal correlations are composed of signals from both small-scale and large-scale features. However, as the focus here is on convective-scale DA, signals from different scales need to be separated. This separation is achieved by applying a bandpass filter. Including a scale analysis, highly variable fields such as 2-m temperature or specific humidity at 850 hPa exhibit the largest relative potential impact for precipitation forecasts. Furthermore, the assimilation of precipitation will likely have a substantial impact on the short-range precipitation forecast. In contrast, the relative potential impact of sea-level pressure, upper tropospheric wind, and temperature is dominated by large-scale correlations that are determined by boundary conditions.

Overall, this study provides the basis for subsequent research on observing and data assimilation strategies for convective-scale NWP. Further investigation is particularly required regarding the scale dependence of the potential impact. The border to separate small and large-scale signals in this study is chosen to be 300 km. However, a different threshold might be required performing the analysis in a system with other properties (e.g., different domain size, or resolution).

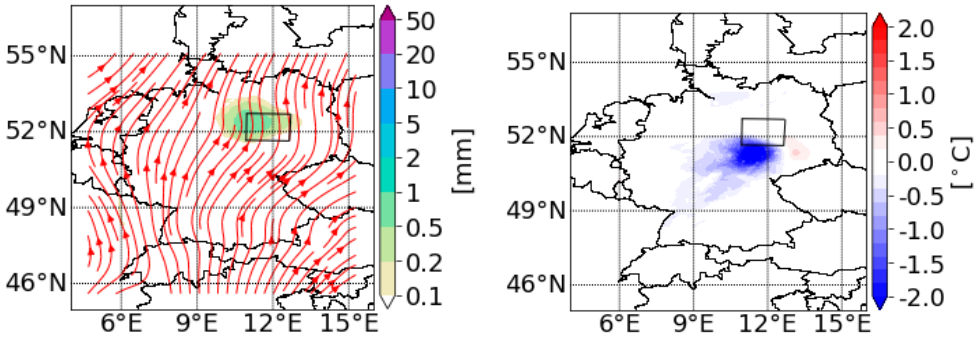
Chapter 6

Sampling error correction

As mentioned in Section 1.4, sampling errors pose an issue estimating a sample correlation with a limited ensemble size. However, many ensemble applications rely on accurate estimates of correlations. Reliable spatiotemporal correlations are required for ensemble sensitivity or impact studies as performed in Chapter 4 and 5. Furthermore, achieving better error correlations can improve the impact of observations in data assimilation significantly. For these reasons, this chapter aims to investigate if the statistical sampling error correction (SEC) introduced by Anderson (2012) can reduce sampling error for various applications. Both spatial and spatiotemporal correlations are computed for different ensemble subsets of the 1000-member ensemble simulation. The sub-sampling is done similarly as introduced in Section 3.5. Overall, it is assumed that the 1000-member ensemble itself is not affected by sampling error and can be seen as truth to verify the performance of the SEC.

Chapter 6 consists of two parts. The first part (Section 6.1) evaluates the SEC for application to ensemble sensitivity analysis (ESA). Spatiotemporal correlations as they are calculated within ESA provide the basis to estimate the potential impact of observable quantities in Chapter 5 or the EFSOI in Chapter 4. Previous ESA studies often applied a confidence test using a 95 % confidence interval (T95) to reduce the effect of spurious correlations (Torn and Hakim, 2008). This study, therefore, compares the SEC with the T95 approach. The SEC is evaluated qualitatively using a case study and quantitatively comparing different variable combinations and ensemble subsets. Here, the analysis focuses on spatiotemporal correlations with respect to precipitation as these correlations are used for subsequent analysis in Chapter 5.

The second part (Section 6.2) evaluates the SEC for application to ensemble and hybrid data assimilation. Anderson (2012) originally designed the SEC to reduce the need for localization in EnKF data assimilation. Anderson (2012) and Anderson (2016) successfully showed that the approach can compete with classical localization approaches that apply constant tampering functions. So far, the SEC approach has not been evaluated for a convective-scale modeling system. Subsequently, the SEC is examined for spatial correlations between different prognostic quantities. The evaluation is performed separately for horizontal and vertical correlations. All results from this chapter are submitted for publication by Necker et al. (2019b) (see Appendix A).



(a) Precipitation **(b)** Anomaly

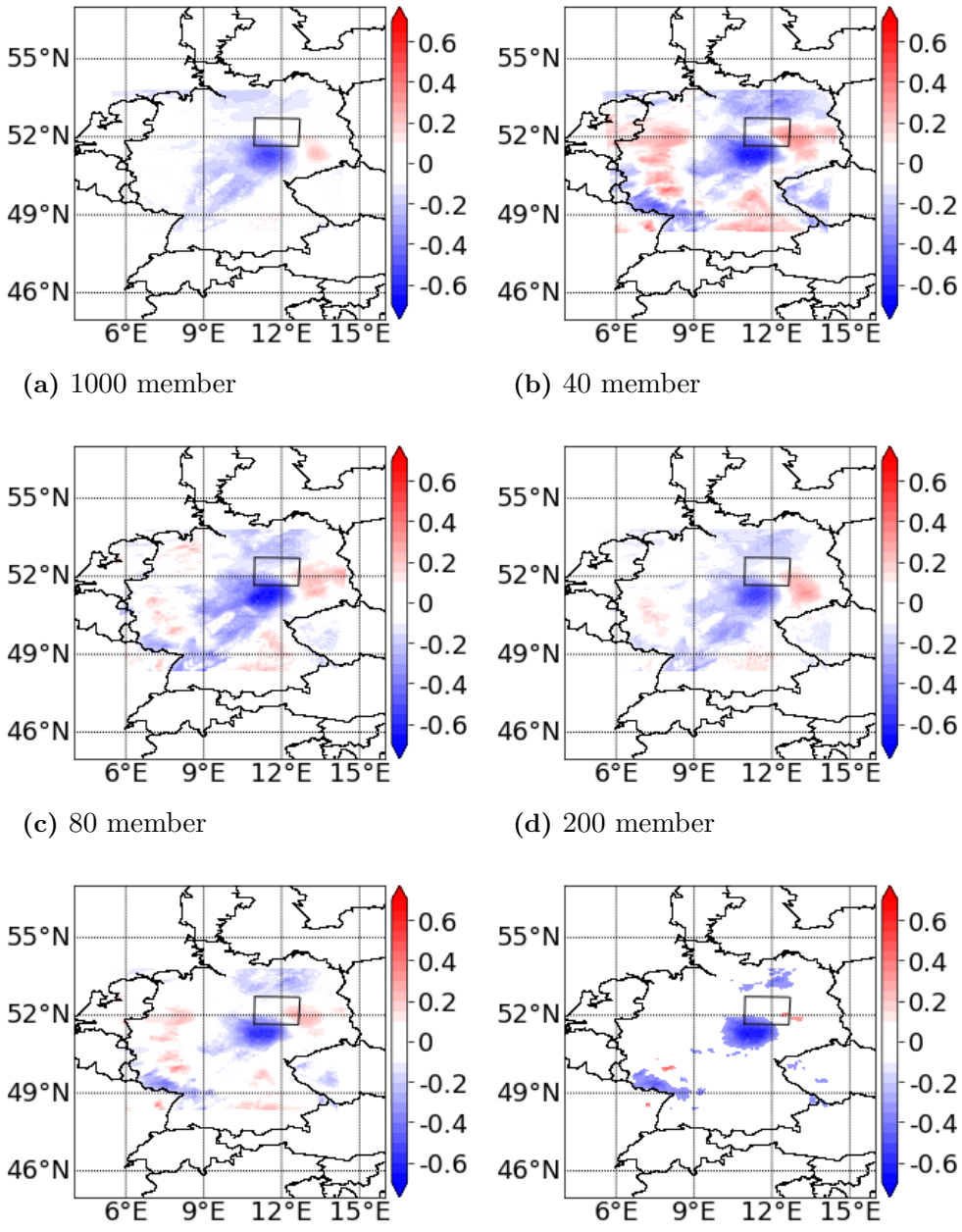
Figure 6.1: *(a) 1000-member ensemble mean precipitation and streamlines of 500 hPa wind (a, 29 May 2016 04 UTC). (b) Initial 2-m temperature anomaly calculated comparing the ensemble mean 2-m temperature of the 100 members with strongest and 100 member with weakest precipitation inside the forecast response function (29 May 2016 01 UTC).*

6.1 Spatiotemporal correlations

Example of spatiotemporal correlations

This section starts with a qualitative analysis of spatiotemporal correlations for the first forecast initialized at 29 May 2016 00 UTC. Figure 6.2 displays sensitivities of the 3-h precipitation forecast (Figure 6.1a) to the initial 2-m temperature field calculated for different ensemble sizes and with different sampling error approaches. The differences compared to the 1000-member ensemble correlation (Figure 6.2a) illustrate the effect of sampling errors. The 1000-member ensemble shows strong negative correlations of precipitation to the initial 2-m temperatures in a region south-west of the response function. These negative correlations are related to evaporative-cooling caused by precipitation resulting in colder surface temperatures in this area. Clustering each the 100 members with the strongest and weakest precipitation inside the response function reveals a temperature anomaly in the initial surface temperature field (Figure 6.1b) that matches the area of negative sensitivities. The southwesterly tail of negative correlation roughly marks the track of the precipitating systems during the night (Figure 6.1a). This feature corresponds to the southwesterly wind indicated by streamlines in Figure 6.1a. The region with positive correlations south-east of the response function is related to a westward shift of precipitation in some of the ensemble members. This effect is stronger for shorter lead times (not shown).
In contrast, the 40-member ensemble correlation field (Figure 6.2b) exhibits various spurious correlations in the south and west of the domain. Furthermore, the small ensemble systematically overestimates the amplitude of sensitivities in several locations. Increasing the ensemble size to 80 or 200 ensemble member (Figure 6.2c & 6.2d) systematically reduces the number of spurious correlations at larger distances from the precipitation event. However, some small positive spurious correlations are still visible for the 200-member ensemble.
Figure 6.2e shows the 40-member ensemble correlation field corrected with the SEC. The SEC can remove several spurious correlations and also corrects the amplitude of the strongest negative correlations. However, it partly removes the tail south-west of the area

of maximum correlation. Applying the confidence test (T95) to the 40-member ensemble correlation field (Figure 6.2f) removes all correlations approximately smaller than ±0.25 and returns an incomplete correlation field. Compared to the SEC, the confidence test eliminates nearly all positive correlations and also removes the entire tail. Nevertheless, some spurious correlations at the French-German border remain as those exhibit comparably large correlation values. Furthermore, T95 does not correct the amplitude of the strongest correlation. Results for other variables are overall similar (not shown).



(a) 1000 member **(b)** 40 member

(c) 80 member **(d)** 200 member

(e) 40 + SEC **(f)** 40 + T95

Figure 6.2: *Correlation of the 3 h precipitation forecast to the initials 2-m temperature field at 29 May 2016 01 UTC for different ensemble configurations: (a) 1000 member, (b) 40 member, (c) 80 member, (d) 200 member, (e) 40 member with SEC, and (f) 40 member with T95.*

Correlation distribution

Figure 6.3 shows four different correlation frequency distributions. The histograms are calculated using correlations from all ten 3 h lead-time forecasts and 2-m temperature as the target state variable. The distribution of the 40-member ensemble nearly re-samples the shape of a normal distribution peaking slightly shifted towards negative values. The 1000-member ensemble distribution peaks at a similar position but showing an approximately three times higher amplitude combined with a smaller width. Applying the SEC to the 40-member ensemble correlations improves the distribution significantly. The width and the amplitude of the peak are now similar to the 1000-member ensemble but slightly shift towards zero. The shift of the peak originates from the assumed uniform prior U(-1,1) and may be reduced by using a more informed prior assumption when calculating the systematic error correction off-line. A different prior would especially make sense for highly positively or negatively correlated fields. Both, a climatological prior (Anderson, 2016) or a prior obtained from 1000-member ensemble could be used to generate a more specified table.
Filtering all unreliable 40-member ensemble correlations using the confidence test (T95) changes the distribution fundamentally. The confidence test removes all sensitivities smaller than approximately ±0.25 and therefore discards the majority of correlations. Comparing both approaches, the SEC significantly improves the distribution, whereas the application of the T95 leads to an unrealistic distribution of correlations. The effect is similar for correlation distributions of other variables (not shown).

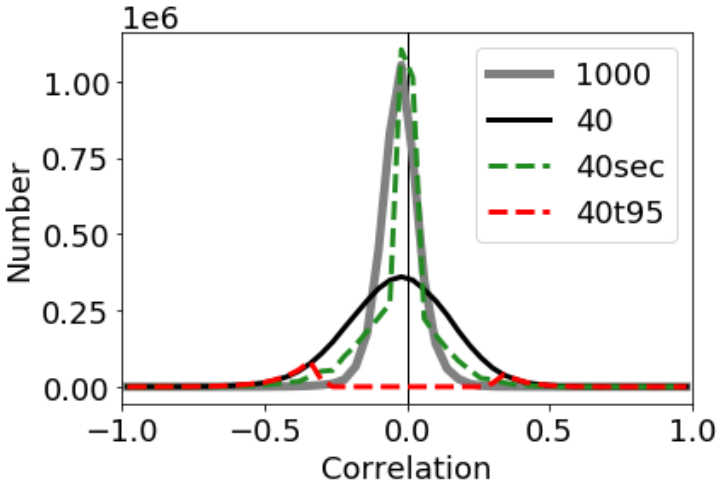
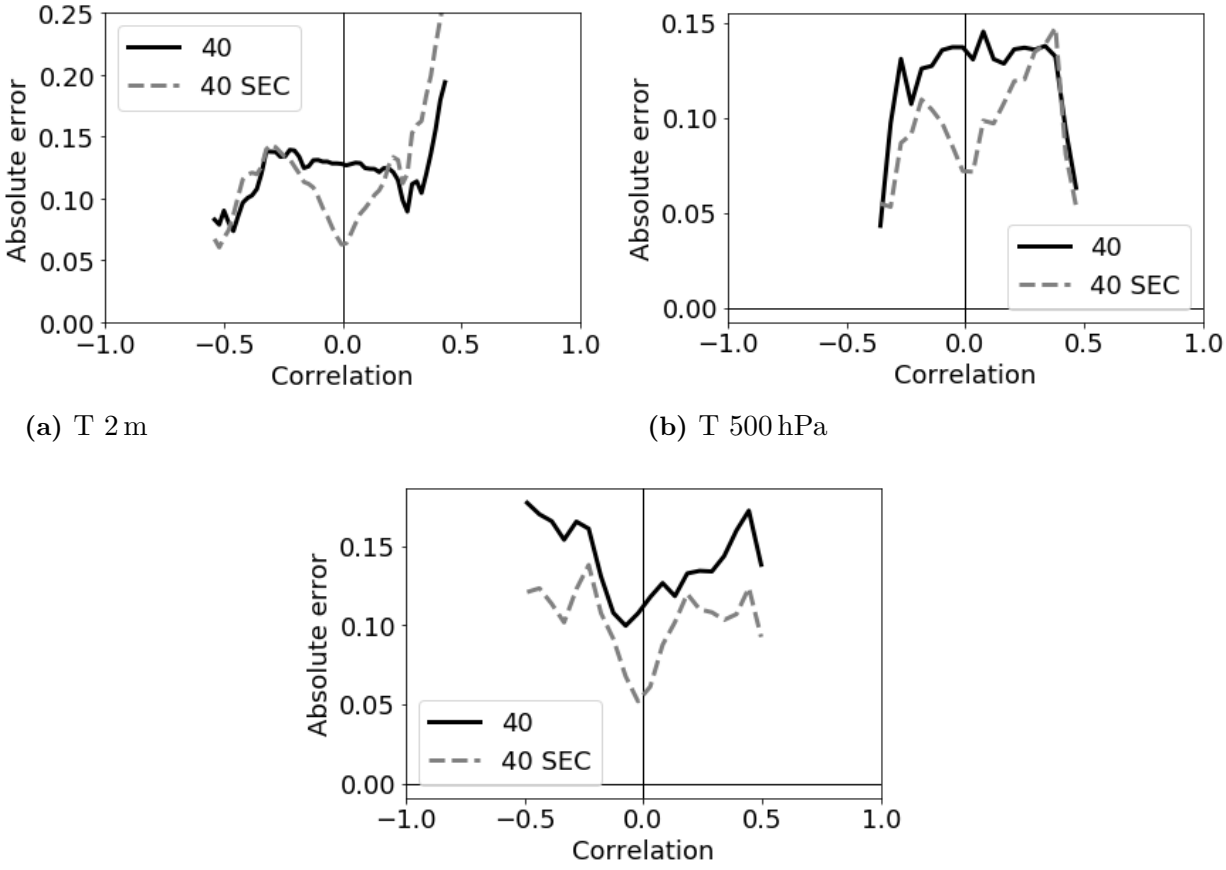


Figure 6.3: *Frequency distributions for correlations of the 3 h precipitation forecast to the initial 2-m temperature field using all ten forecasts. 1000-member ensemble correlations (bold solid grey) and 40-member ensemble (solid black) including SEC (green dashed) or T95 (red dashed).*

Sampling error as a function of correlation value

Figure 6.4 presents the mean absolute correlation error as a function of correlation value. The sampling error of 2-m temperature using a 40-member ensemble (Figure 6.4a) is smallest for large negative correlation values and largest for strong positive correlations. Applying the SEC significantly reduces the error for small correlation values but slightly

degrades the performance of a few high positive correlations. However, results for large correlation values should be treated with caution as there are only a few data points (see frequency distribution in Figure 6.3). The absolute error obtained for correlations of precipitation with 500 hPa temperature (Figure 6.4b) looks similar as for surface temperature. Again, the SEC mainly improves small correlation values, whereas for 500 hPa zonal wind (Figure 6.4c) improvements are visible for the entire range of correlation values. For all variables, the SEC has its strongest effect on small correlation values, which seems reasonable considering the correction function displayed in Figure 2.3. For very small correlation values, the SEC almost halves the sampling error. For larger correlation values, the effect of the SEC is smaller and differs depending on the considered variable. Nevertheless, one should keep in mind the relatively small sample of large correlation values.



(a) T 2 m

(b) T 500 hPa

(c) U 500 hPa

Figure 6.4: *Mean absolute error of the sample correlation (solid black) and sampling error corrected correlation (grey dashed) as a function of the sample correlation using the 40-member ensemble. Correlations of the 3 h precipitation forecast to initial (a) 2-m temperature, (b) 500 hPa temperature and (c) 500 hPa zonal wind using all ten forecasts.*

Sensitivity to ensemble size

Figure 6.5a presents the time-averaged root mean square error (RMSE) of correlations as a function of ensemble size and investigates the same correlations as shown in the previous

two sections (precipitation correlated with 2-m temperature). Here, the RMSE of a 40-member ensemble is given by

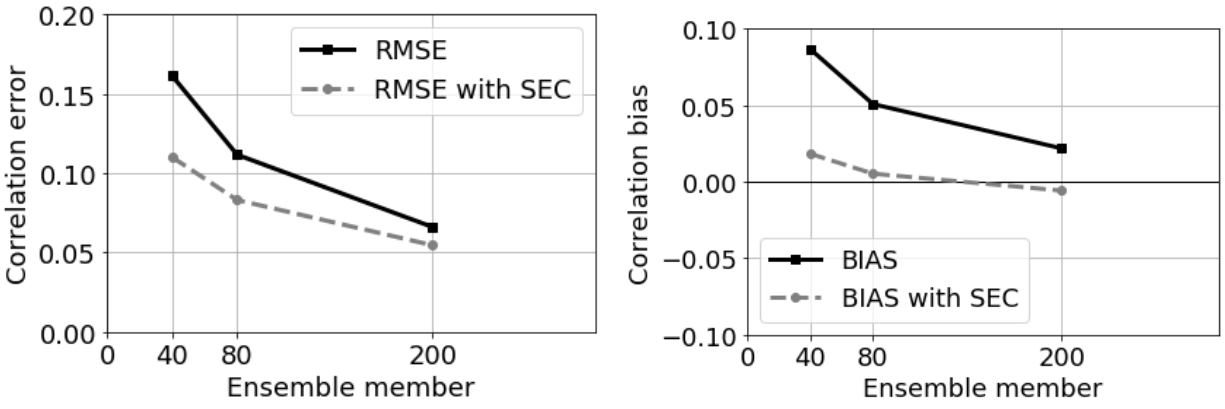
$$RMSE_{40} = \sqrt{\frac{1}{N}\sum_{n=1}^{N}\left(r_{40,n} - r_{1000,n}\right)^2},$$

where N is the number of grid points in the domain. The RMSE is calculated using correlations obtained for the full 1000-member ensemble for verification. The RMSE of the 40-member ensemble is approximately 0.16. Doubling the sample size up to 80 members reduces the RMSE by about 30 % whereas increasing the sample size by the factor five up to 200 members lowers the RMSE by more than 50 %. For small ensemble samples, the SEC strongly improves the performance. Applying the SEC to the 40-member ensemble subset even achieves slightly better results than doubling the ensemble size. The reduction of RMSE due to the SEC decreases with increasing ensemble size. Nevertheless, the 200-member RMSE is still reduced by about 15 % by the SEC.

Figure 6.5b shows the corresponding time-averaged difference of the mean absolute correlation (BIAS) compared to the 1000-member ensemble for all six configurations. Here,

$$BIAS_{40} = \frac{1}{N}\left(\sum_{n=1}^{N}|r_{40,n}| - \sum_{n=1}^{N}|r_{1000,n}|\right).$$

Similar to the RMSE, the BIAS decreases with increasing ensemble size and applying the SEC significantly reduces the BIAS. For nearly all subsets, the BIAS almost vanishes. For larger subsets, the SEC also reduces the bias causing a change in sign. Nevertheless, the improvements due to the SEC are substantial and visible for all variables. Different prior assumptions used for computing the SEC table could presumably improve the results further.



(a) **(b)**

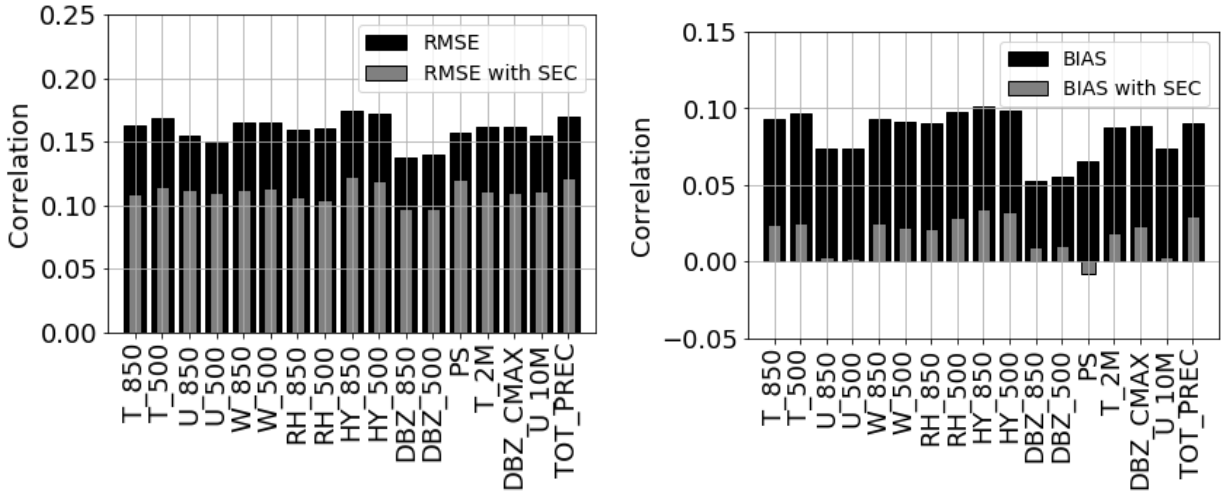
Figure 6.5: *Time-averaged root mean square error (a) and magnitude bias (b) of correlations with/without SEC compared to 1000 members evaluated for different ensemble subsets. Spatiotemporal correlations of precipitation to 2-m temperature.*

Sensitivity to variable

Figure 6.6a presents the RMSE for 40-member correlations of precipitation to various initial quantities. The black and grey bars displayed for 2-m temperature coincide with the markers of the 40-member ensemble shown in Figure 6.5a. The RMSE for all variables ranges from approximately 0.13 to 0.18. As discussed for 2-m temperature, correcting the correlations using the SEC significantly reduces the RMSE independent of the chosen variable. The improvements range from about 20 % to 30 %, and are smallest for sea-level pressure (PS).

Examining the BIAS (Figure 6.6b), sea-level pressure is the only variable that exhibits a change in sign of the bias. This characteristic is likely related to the structure of the correlation field, which is homogeneously distributed over the entire domain as sea-level pressure exhibits a fairly smooth large-scale field consisting of mainly small negative correlations. The SEC systematically reduces the BIAS for all variables and works most efficiently for zonal wind. Examining the impact of the SEC on the 80 and 200-member ensemble correlations (not shown), the systematic reduction of the BIAS relatively increases with increasing ensemble size leading to changes in sign as discussed for 2-m temperature (Figure 6.5b). Nevertheless, the reduction of BIAS is significant for all investigated ensembles sizes and variables.

Further sensitivity studies have been conducted that are not shown in this manuscript. These experiments targeted the sensitivity of the SEC to the precipitation metric kernel size, the choice of the ensemble subset as well as the dependence on forecast lead time. However, these sensitivity studies are not discussed here, as these experiments did not reveal any fundamentally different results.



(a) (b)

Figure 6.6: *Time-averaged root mean square error (a) and magnitude bias (b) of 40-member precipitation correlation to various variables on different height levels with and without SEC. (T: Temperature; U: Zonal wind; W: Vertical wind; RH: Relative humidity; HY: Hydrometeors; DBZ: Radar reflectivity; PS: Sea-level pressure; DBZ_CMAX: Column maximum radar reflectivity; TOT_PREC: Precipitation)*

6.2 Spatial correlations

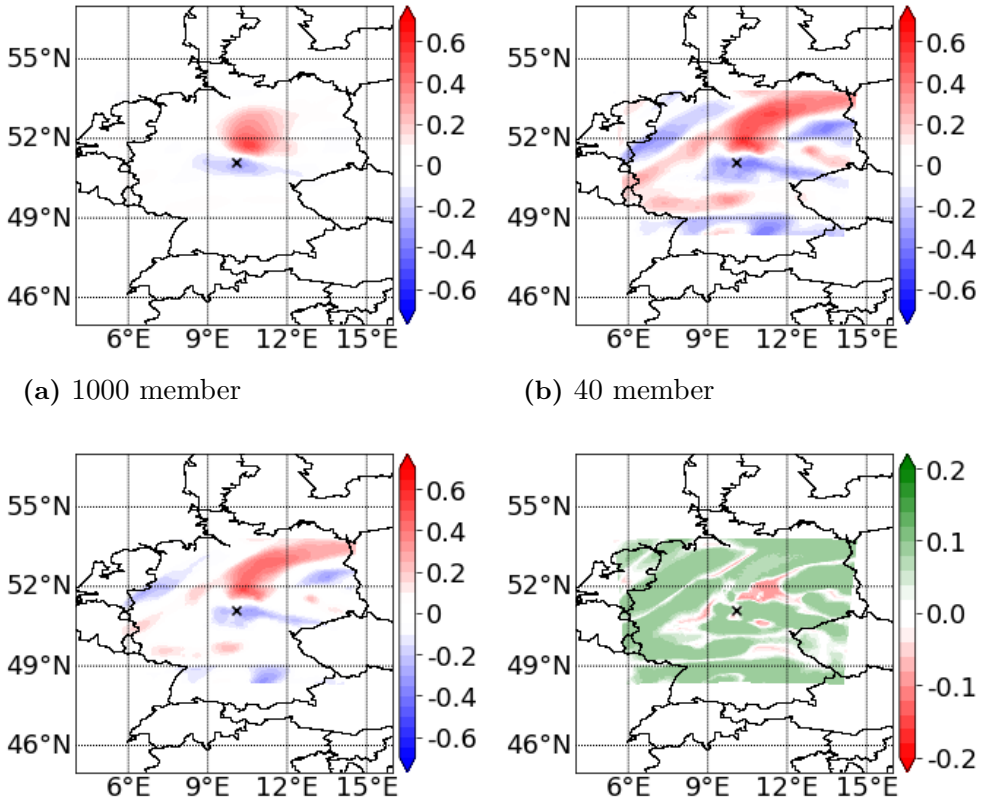
This section investigates the impact of the SEC on spatial correlations that are crucial for ensemble or hybrid DA. Results are shown for the correlation of temperature to various model variables. Spatial correlations are calculated using 1-h forecasts, which is similar to taking the first guess during hourly cycling.

Example of spatial correlations

Figure 6.7a displays horizontal cross-correlations of 500 hPa temperature to 500 hPa specific humidity at every grid point in the domain. The correlation pattern is a dipole showing a negative correlation in the vicinity and a positive correlation to the north of the response function. Except for the dipole, no other considerable correlations are visible. Examining the 40-member correlations (Figure 6.7b), various spurious correlations show up all over the domain similar as discussed for spatiotemporal correlation in Subsection 6.1. To some degree, the dipole is still indicated by the strongest correlations. Applying the SEC (Figure 6.7c) reduces the number of spurious correlations significantly and reveals the dipole more distinctly. Overall, the SEC can reduce the sampling error for the majority of grid points (Figure 6.7d), showing slightly increased errors only in some small areas. The improvements are consistent for spatial correlations to other variables (not shown) and agree with the results obtained for spatiotemporal correlations considering a precipitation-based response function.

Horizontal correlation

Below, horizontal correlations are averaged using the ten available 1-h 1000-member forecasts. Each ensemble forecast is evaluated with nine grid point size metrics that are evenly distributed in the domain with a distance of 50 grid points (150 km) to neighboring metrics and boundaries. In total, 90 correlation fields are examined for each variable pair.
Figure 6.8 displays the mean absolute correlation and error as a function of spatial distance (in km) using differently sized ensembles. Sampling errors are again calculated using the 1000-member ensemble as truth. Figure 6.8a shows the mean absolute correlation of 2-m temperature to 2-m temperatures and Figure 6.8b the corresponding error with and without SEC. The 1000-member ensemble exhibits a correlation of nearly 1 in the close vicinity of the response function, dropping to a value of about 0.4 reaching a distance of 100 km. Up to 100 km, the 40-member ensemble correlation coincides with the 1000-member ensemble correlation. Farther away, the 40-member ensemble systematically overestimates the mean absolute correlation due to spurious correlations. The mean absolute correlation error (Figure 6.8b) strongly increases up to a distance of 100 km, which roughly matches the radius of horizontal localization applied in regional DA systems. For distances larger than 100 km, the sampling error keeps increasing, but slower compared to the vicinity of the response function. Applying the SEC increases the error close to the response function slightly, but significantly reduces the error at larger distances. Similar changes are visible for the mean correlation. Especially for distances larger than 150 km, the sampling error corrected 40-member mean absolute correlation almost coincides with the 1000-member correlation.

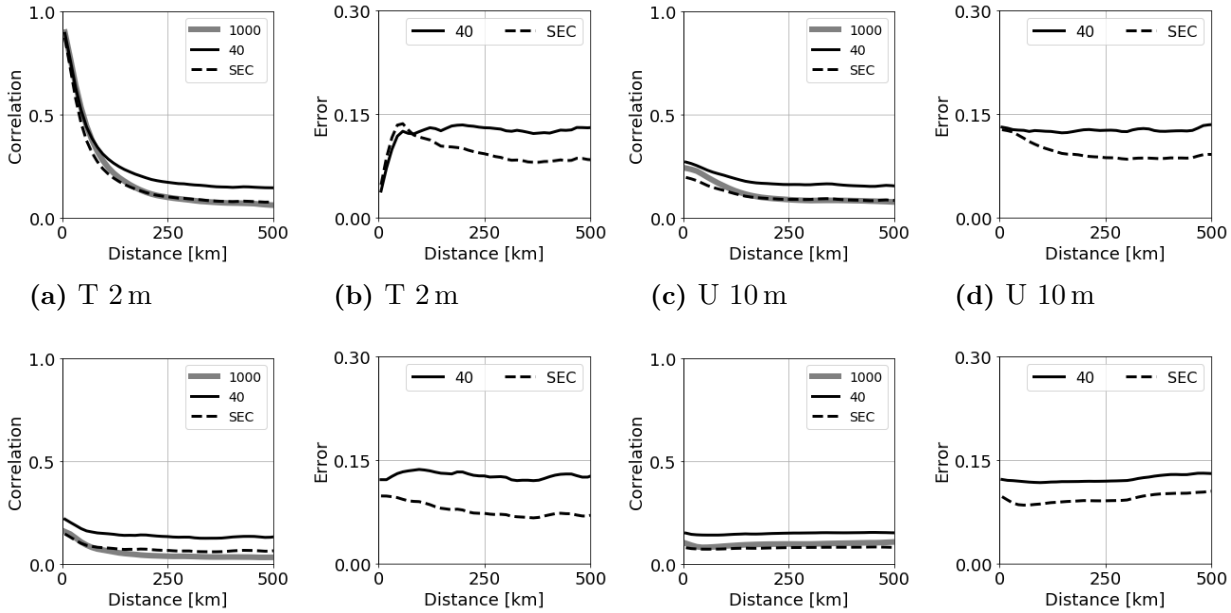


(a) 1000 member **(b)** 40 member

(c) 40 + SEC **(d)** Error reduction

Figure 6.7: *Cross-correlation of 500 hPa temperature (black marker) to 500 hPa specific humidity at 29 May 2016 01 UTC for (a) 1000 member, (b) 40 member, and (c) 40-member including SEC as well as (d) changes in correlation field due to the SEC (green - error reduction). Metric position is indicated by a black cross.*

The mean absolute cross-correlation of 2-m temperature to 10-m zonal wind (Figure 6.8c) and 2-m temperature to near-surface humidity (Figure 6.8e) show similar results. Both variables exhibit the strongest correlation in the near vicinity dropping to a constant value of approximately 0.2 at a distance of 150 km. The mean absolute errors (Figure 6.8d; Figure 6.8f) slightly change with distance showing a similar absolute value for large distances as found in Figure 6.8b. However, the relative error is larger considering the weak mean absolute correlation for these pairs. Including the SEC significantly improves both the mean and error of the spatial cross-correlations. The SEC performs best for distances larger than 100 km reducing the error of the humidity cross-correlation by up to 40 %.
The correlation of 2-m temperature to sea-level pressure (Figure 6.8g) is weaker compared to spatial correlations discussed previously. Mean absolute correlation and error (Figure 6.8h) hardly change with distance. Due to sampling errors, the 40-member mean correlation is twice as large as the 1000-member mean correlation. The SEC significantly improves the 40-member mean correlation, which is now close to the 1000-member mean correlation. The absolute error decreases by approximately 20 %.
Figure 6.9 shows the mean absolute correlation and error as a function of horizontal distance using correlations of 500 hPa temperature to different upper-tropospheric variables. Both the 1000 and 40-member ensemble correlation decline consistently examining spatial

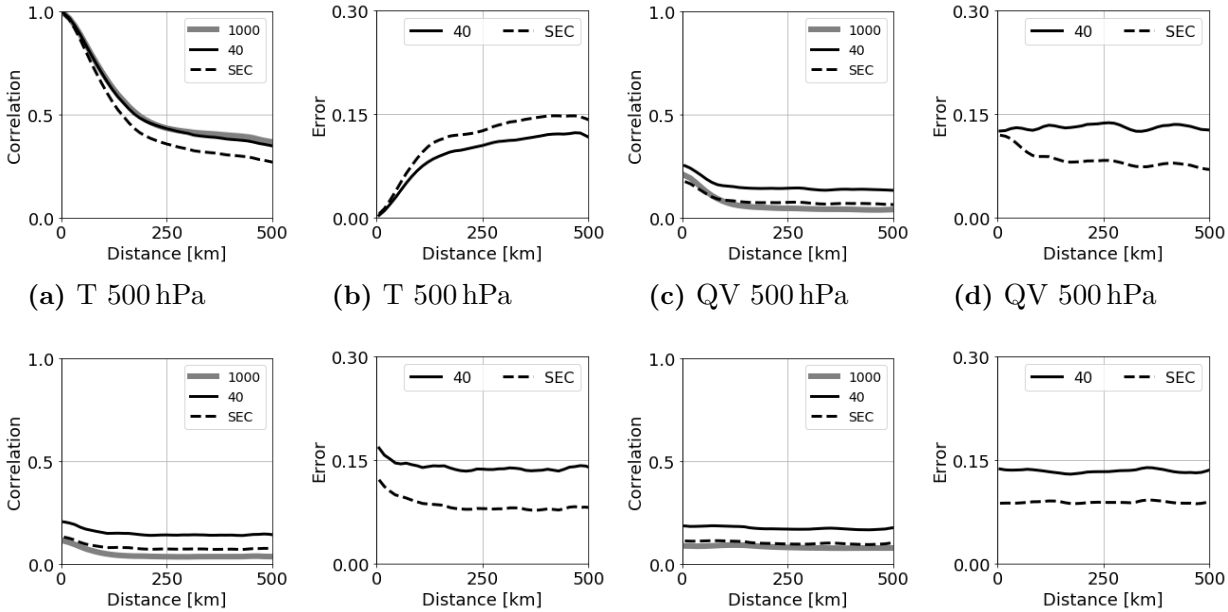


(a) T 2 m **(b)** T 2 m **(c)** U 10 m **(d)** U 10 m

(e) QV 925 hPa **(f)** QV 925 hPa **(g)** PS **(h)** PS

Figure 6.8: *Mean absolute correlation (first and third column) and error (second and fourth column) as a function of spatial distance [km] for different ensembles and with or without SEC. Correlation of 2-m temperature to (a,b) 2-m temperature, (c,d) 10-m zonal wind, (e,f) 925 hPa specific humidity and (g,h) sea-level pressure.*

correlations of 500 hPa temperature (Figure 6.9a). The magnitude of the correlation is larger than for all other discussed quantities. Furthermore, the 40-member mean absolute correlation error is smaller, grows less rapidly, and does not appear saturated at a horizontal distance of 500 km (Figure 6.9b). In contrast to other variables, including the SEC degrades the performance for the entire spatial range. The mean absolute correlation is now underestimated, and the error increases correspondingly. The negative impact of the SEC is likely related to an insufficient prior assumption, which is fundamental for the behavior of the SEC. In this case, a uniformly distributed prior appears to be unsuitable. As suggested by Anderson (2012), a different prior of, for example, U(0.5,1.0) could be used for strongly positively correlated variables and also a climatically estimated prior (Anderson, 2016) could be suitable to improve the performance of the SEC is such cases.
Figure 6.9c and 6.9d analyze horizontal cross-correlations of 500 hPa temperature to 500 hPa specific humidity. Again, the mean absolute correlation decreases with increasing distance. The SEC reduces both mean and error showing an improved performance far from the response function. Cross-correlations of 500 hPa temperature to 500 hPa hydrometeors (Figure 6.9e) are weaker compared to cross-correlations of temperature and humidity. As before, the SEC significantly reduces the error (Figure 6.9f) while it slightly over-adjusts the mean absolute correlation. The results for cross-correlations of 500 hPa temperature to 500 hPa zonal wind (Figure 6.9g and Figure 6.9h) are similar as discussed for cross-correlations of 2-m temperature to sea-level pressure (Figure 6.8) although the mean absolute cross-correlations and errors are slightly larger in this case.
Overall, the SEC reduces the sampling error for the majority of horizontal (cross-)correlations using a uniformly distributed prior as done in this study. Furthermore, the SEC performs



(a) T 500 hPa **(b)** T 500 hPa **(c)** QV 500 hPa **(d)** QV 500 hPa

(e) HY 500 hPa **(f)** HY 500 hPa **(g)** U 500 hPa **(h)** U 500 hPa

Figure 6.9: *Same as Figure 6.8, but spatial correlation of 500 hPa temperature to (a,b) 500 hPa temperature, (c,d) 500 hPa specific humidity, (e,f) 500 hPa hydrometeors and (g,h) 500 hPa zonal wind.*

best for cross-correlations and distances larger than 100 km. Only strongly correlated variables show ambiguous results. This problem could be addressed by a different prior assumption or the exclusion of these variables from the correction.

Vertical correlation

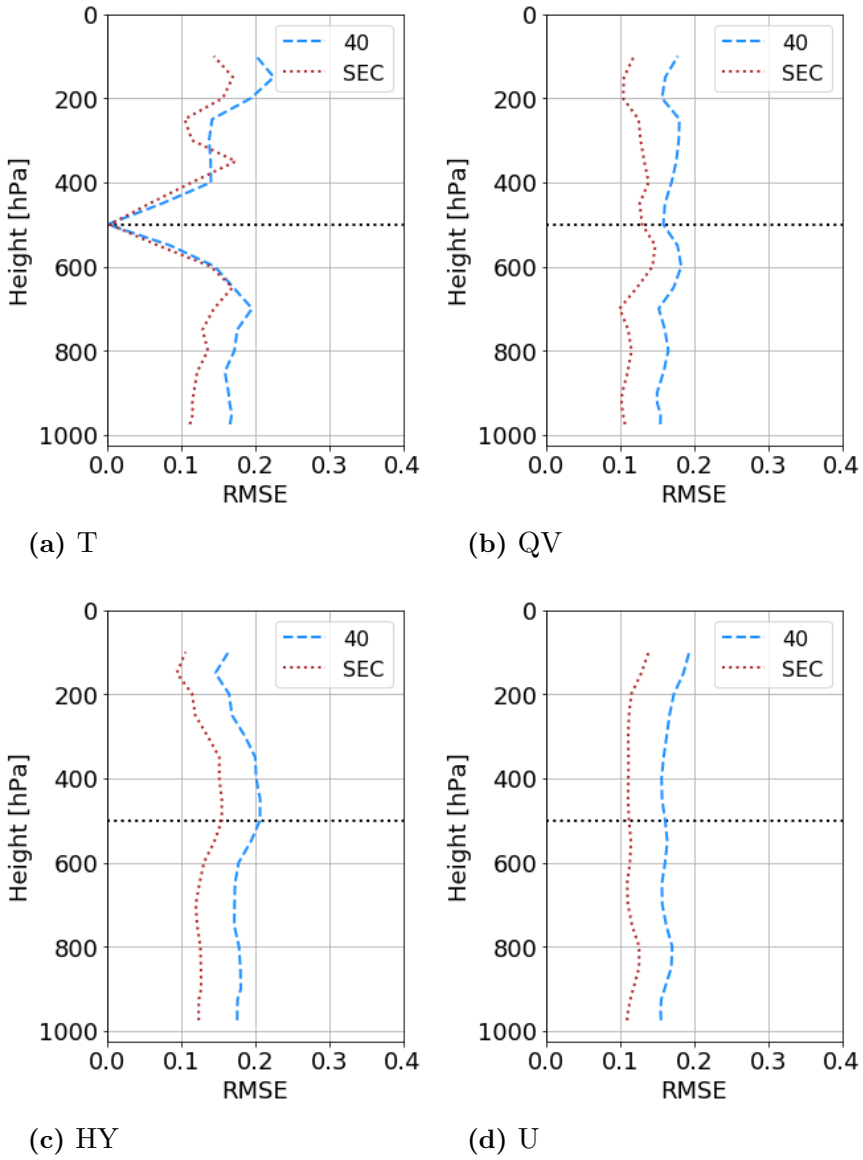
Vertical correlations are evaluated using a single 1000-member ensemble forecast at 30 May 2016 13 UTC and in total 40.000 vertical profiles. For vertical correlations, the focus is on spatial correlations of 500 hPa temperature to 20 different pressure levels and four different variables. Figure 6.10 shows the RMSE of vertical temperature correlations with and without SEC comparing the 40 and 1000-member ensemble. The RMSE of the temperature correlated with itself is zero at 500 hPa (Figure 6.10a) as both 40 and 1000 members exhibit a correlation of 1. The RMSE of the 40-member ensemble correlation increases to a value of 0.15 reaching a vertical distance of 100 hPa and seems to be saturated for distances larger than 150 hPa. The error, including the SEC, increases slower and saturates earlier reducing the relative error far from the response level up to 30 %. Only at 350 hPa, the SEC increases the RMSE as the 40-member ensemble subset on average slightly underestimates the true correlation (not shown).

Figure 6.10b shows the RMSE for vertical cross-correlations of temperature at 500 hPa to specific humidity in the entire tropospheric column. Compared to the previous example, the RMSE for the 40-member ensemble does not exhibit a local minimum at 500 hPa, and hardly changes with height. Adding the SEC reduces the RMSE at all levels, but the reduction is smallest at 500 hPa response level. The RMSE reduction increases up to a vertical distance of 150 hPa, and again hardly changes far from the response level.

Evaluating vertical cross-correlations of temperature to hydrometeors (Figure 6.10c) or zonal wind (Figure 6.10d) the effect of the SEC is independent of the vertical distance and the SEC significantly reduces the RMSE at all levels by about 30 %.
In general, the SEC seams to perform slightly better for vertical than for horizontal correlations. The impact of the SEC is most significant for vertical cross-correlations and far from the response level. The error is roughly symmetric comparing results above and below the response level. On average, the SEC efficiently reduces the overestimation of the true correlation due to spurious correlations. The SEC should perform best if correlations extend over the full vertical profile of the atmosphere (e.g., for passive satellite observations). In such situations, localization techniques are potentially dangerous as they damp or eliminate correlations after a certain distance. The same applies to cloud information, which can affect the surface as well as the entire tropospheric column by modified radiative processes.

Sampling error correction as a function of correlation value

Figure 6.11 displays the 2D correlation frequency distribution comparing the 1000-member ensemble spatial correlations with corresponding spatial correlations obtained for ensemble subsets. Each analysis includes approximately 38 million spatial correlations of temperatures at 500 hPa to various other variables. Each frequency distribution exhibits a maximum at small correlation values. Positive correlations range from 0 up to 1, while the largest negative 1000-member correlation is approximately -0.5. For the 40-member ensemble (Figure 6.11a), the maximum around zero is elongated in the horizontal direction indicating the overestimation of small correlations due to spurious correlations. Applying the SEC reduces this overestimation systematically and changes the pattern of the frequency distribution (Figure 6.11b). The maximum, including SEC, is narrow and extends vertically. The Pearson correlation coefficient between both correlation samples is displayed in the corner of each sub-figure to facilitate the comparison. Plotting the linear regression line (dashed line) reveals the impact of the SEC as it improves both the slope and the intersection as the SEC reduces the magnitude bias. Overall, the SEC improves the performance of the 40-member ensemble by about 5 % using the Pearson correlation as a measure.
Figure 6.11c shows the frequency distribution comparing 200 with 1000 members. Using 200 members significantly reduces sampling errors for the entire range of correlation values. Increasing the ensemble size by a factor of five especially improves the estimation of small correlation values. The 200-member ensemble exhibits a maximum offset of approximately 0.4, which is significantly less than found for 40 members. Adding the SEC (Figure 6.11d) again improves the frequency distribution systematically. The absolute impact is smaller compared to 40 members, but the improvements are particularly visible for small correlations as well as in the slope of the linear regression line.
Considering that the SEC showed ambiguous results for some highly correlated variables, it is important to assess if the SEC systematically fails for large correlations values. Figure 6.12a shows the change in the absolute correlation error caused by the SEC as a function of the 40-member absolute correlation value. The frequency distribution again reveals the most significant improvements for small correlation values. Both negative and positive impacts mainly exhibit the strength of the maximum possible adjustment that is indicated



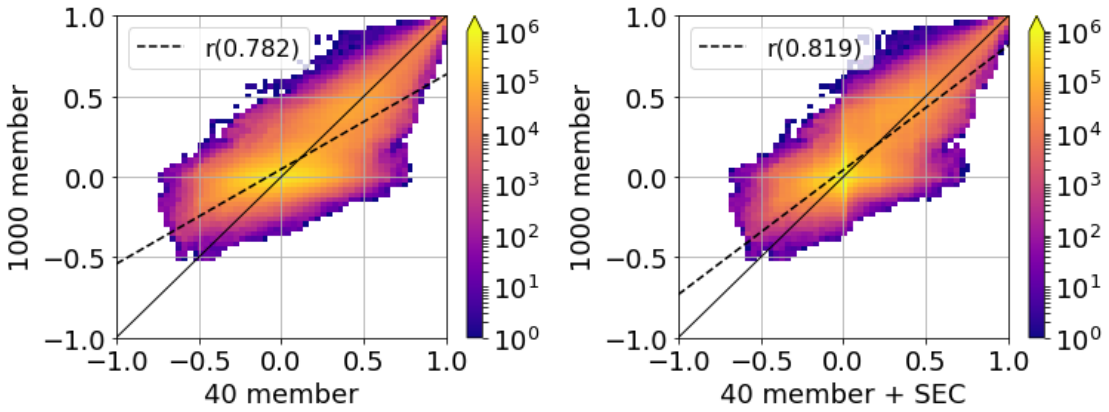
(a) T (b) QV

(c) HY (d) U

Figure 6.10: *Root mean square error of the 40-member correlation compared to the 1000-member correlation with (red, dotted) and without (blue, dashed) SEC. Correlation of 500 hPa temperature to (a) temperature, (b) specific humidity, (c) hydrometeors, and (d) zonal wind at different height levels. RMSE averaged over 40.000 vertical profiles. Note: (a) corresponds to Figure 3.9b and (b) corresponds to Figure 3.9d.*

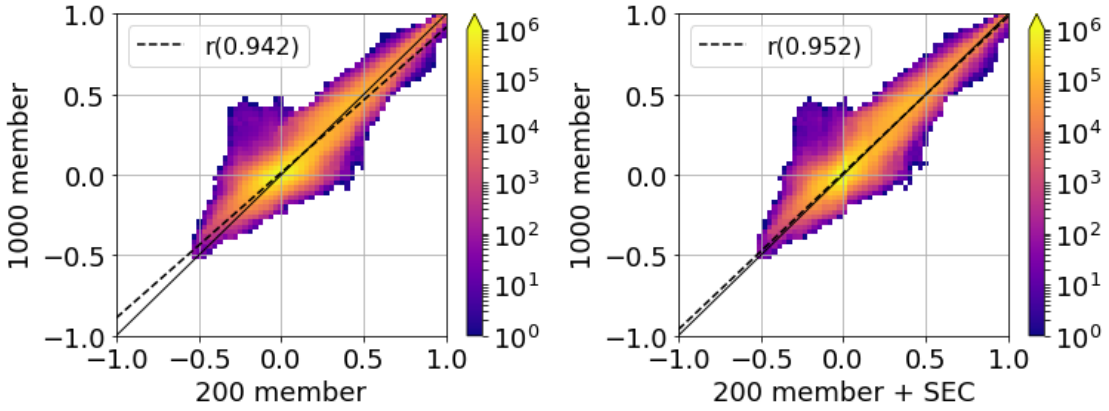
by the dashed line and derived from the correcting function. In general, the beneficial impact of the SEC could be improved, if the correction for small correlation values would be stronger. Examining the average improvement, the SEC systematically improves the results independent of the amplitude of the 40-member correlation value. Overall, the SEC improves about three-quarter of the correlations.

Figure 6.12b shows the same data as before but now distributed as a function of the 1000-member absolute correlation value. Again, the main improvements are observed for small correlation values, and the overall impact is beneficial. However, the impact of the SEC seems to be detrimental for 1000-member correlation values larger than 0.25. Similar



(a) 40 member

(b) 40 member + SEC



(c) 200 member

(d) 200 member + SEC

Figure 6.11: *2d frequency histogram of 1000 member and corresponding sub-sample correlations with and without SEC. The analysis includes about 38 million spatial correlations of temperature at 500 hPa to temperature, specific humidity, hydrometeors, zonal wind, sea-level pressure, and precipitation. Slope of the linear regression fit (dashed line).*

behavior is seen for vertical correlations (not shown). However, as the true correlation is usually unknown, it is difficult to use this behavior to improve such cases. Overall, results suggest that based on the available information from the small ensemble (Figure 14a), the SEC should be applied to all correlations.

6.3 Summary and discussion

The sampling error correction (SEC) described by Anderson (2012) is evaluated applying the first convective-scale 1000-member ensemble simulation over central Europe. This unique data-set consists of ten 1000-member ensemble forecasts with 3 km mesh size and has been computed using the Japanese SCALE-RM model and a LETKF based DA system. The SEC is a simple look-up table based approach, which is calculated using a Monte-Carlo technique. If the look-up table is already computed for a target ensemble size, only the sample correlation and no additional prior information is needed to correct for sampling

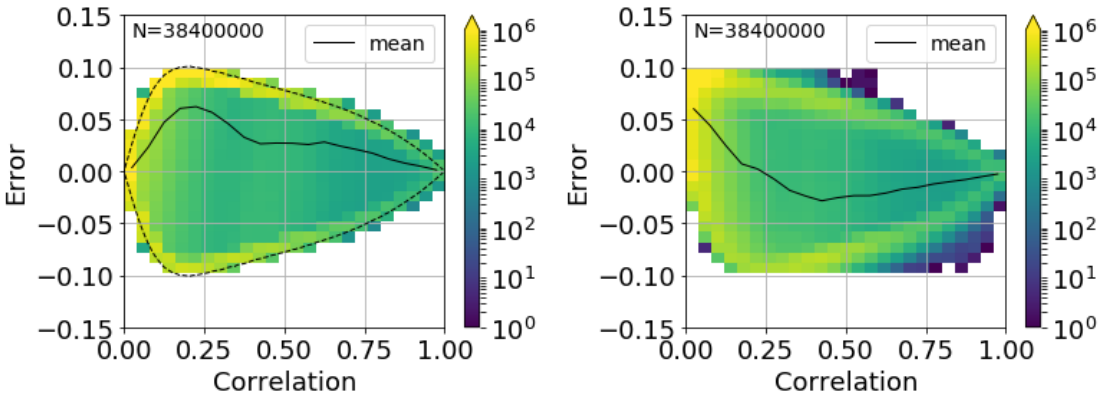


Figure 6.12: *Frequency distribution of error reduction δe applying the SEC to a 40-member ensemble as a function of the absolute value of the (a) 40-member or (b) 1000-member correlation. The solid black line shows the average change and the dashed line sketches the maximum expected adjustment, which is restricted by the correction function. The analysis considers the same correlations as in Figure 6.11. The SEC improves δe for 72.3 % of the correlations ($\delta e = |(r_{40} - r_{1000})| - |(r_{40+SEC} - r_{1000})|$).*

errors.

This study evaluates the SEC for spatiotemporal correlations that are important for ensemble sensitivity analysis (ESA, Ancell and Hakim (2007)) and for spatial correlations that are crucial for ensemble and hybrid data assimilation systems. The 1000-member ensemble correlations are taken as a reference to assess the performance of the sampling error correction as well as of a confidence test (T95, Torn and Hakim (2008)), which is a commonly used approach to exclude spurious correlations in ESA. Furthermore, different subsets of the 1000-member ensemble are used to quantify sampling errors in a convective-scale NWP modeling system.

Examples of correlation fields demonstrate that the 1000-member ensemble provides physically meaningful correlations that are hardly affected by sampling errors while smaller subsets reveal spurious correlations. The 40-member ensemble subset is able to indicate regions of maximum correlation in short-range convective-scale forecasts qualitatively. However, small ensembles overestimate the magnitude of the majority of correlations due to spurious correlations. Increasing the ensemble size up to 80 or 200 members significantly reduces the number of spurious correlations. This finding agrees with the results of Wile et al. (2015) who performed ESA on 4 km resolution using a 96-member ensemble and different subsets.

A confidence test can eliminate some spurious correlations by rejecting small insignificant correlations. However, it also eliminates small true correlations. This behavior is especially visible examining the frequency distribution of correlation values. While this is useful for a qualitative analysis of temporal correlations, the associated removal of weak correlations can lead to systematic errors and is therefore not optimal for quantitative analysis. In contrast to the t-test, the SEC is able to reduce spurious correlations while still allowing for small correlations. The SEC corrects spurious correlations independently of the strength of the correlation and by this significantly improves the frequency distribution. Similar to the

confidence test, the SEC has its most significant impact on small correlations. Overall, the SEC is appropriate for both the qualitative and quantitative interpretation of correlations. The SEC is beneficial for all evaluated ensemble sizes and variable combinations. The mean absolute correlation bias, as well as the RMSE of correlations, are reduced essentially independently of the sample size. For spatiotemporal correlations, the 40-member ensemble including SEC even outperforms the 80-member ensemble as the RMSE is reduced by up to 30 % and the magnitude bias almost vanishes.

Spatial correlations are calculated to investigate sampling errors in ensemble DA. In the vertical, the SEC systematically reduces the RMSE in the entire tropospheric column independently of height. The reduction is largest far from the response level, the impact slightly decreases for distances smaller than 150 hPa, and is smallest close to the response level. Compared to operational localization techniques, which damp or exclude long-range correlations, the SEC allows for correlations far from the response level. This characteristic is crucial for the assimilation of non-local observations (e.g., cloud, satellite radiance, or pressure).

Horizontally, the SEC efficiently improves the estimation of the mean absolute correlation and mitigates the overestimation of the absolute correlation using small ensembles. Furthermore, it reduces the mean absolute error for most variable pairs and performs best on large distances. However, the uniform prior U(-1,1), which is assumed in the calculation of this specific SEC table, appears unsuitable for highly correlated variables. For instance, horizontal correlations of temperatures in the upper troposphere are already sufficiently well estimated by a small ensemble sample and therefore hardly affected by sampling errors. However, the SEC performance could be improved by changing the prior assumption. An improved prior could be estimated climatologically as described by (Anderson, 2016) or obtained from a large ensemble sample, e.g., the 1000-member ensemble used in this study.

Sensitivity studies on the ensemble size show that sampling error corrected correlations using 200 members are already very close to correlations obtained for 1000 members. For horizontal correlations, the SEC increases the correlation between the 40-member and 1000-member ensemble approximately by 5 % and by 1 % using 200 members, respectively. Using 200 members to estimate error covariances in convective-scale DA seems to be a reasonable choice thinking of the achieved accuracy and the computational cost compared to 1000 members.

The SEC improves correlations regardless of their amplitude and has its largest beneficial impact if correlations are small. In general, the SEC should be applied to all correlations regardless of their strength. As the true correlation is unknown in an operational application, it is not beneficial to exclude strong correlations from the SEC.

Overall, the results strongly encourage to use the evaluated sampling correction for ensemble data assimilation systems and ensemble sensitivity analysis. Similarly, it could be applied in the framework of calculating ensemble forecast sensitivity to observation impact (Kalnay et al., 2012, Sommer and Weissmann, 2014, 2016, Buehner et al., 2018). As the method is already implemented in DART, its application is technically simple. Further improvements could be achieved by using more informed prior assumptions, which should and will be the subject of future studies.

Chapter 7

Conclusion

The latest generation of NWP systems reached a convection-permitting resolution. To provide the optimal initial conditions, convective-scale DA systems need to assimilate spatially dense observations that can provide accurate information on small-scale and rapidly changing weather situations. However, limited knowledge exists on which observations are most important for convective-scale data assimilation. The overarching goal of this dissertation is to estimate the impact of observations in high-resolution NWP. For that purpose, several convective-scale ensemble simulations are conducted using two different modeling systems. The first modeling system is the pre-operational convective-scale 40-member ensemble system of Deutscher Wetterdienst (COSMO-KENDA), the second an experimental Japanese convective-scale ensemble system (SCALE-LETKF) with 1000 members.

The SCALE-LETKF simulation was computed in collaboration with the RIKEN Institute for computational science in Japan. The simulation uses a novel setup, which allows generating a unique convective-scale 1000-member ensemble over central Europe. This large ensemble is required to obtain realistic spatial and temporal correlations to be able to answer the target research questions. However, due to its novelty, the convective-scale 1000-member ensemble simulation requires basic evaluation. A comparison of the ensembles to radar observations shows that the SCALE-RM 1000-member ensemble provides realistic precipitation patterns and sufficiently reproduces the diurnal cycle of precipitation. The 1000-member ensemble exhibits a realistic evolution of ensemble spread of precipitation and other variables. Overall, the 1000-member ensemble provides an unprecedented data-set for various applications.

All simulations are applied to examine the observation impact (EFSOI) and the potential impact of observable quantities (ASC) in convective-scale modeling. The two impact estimation approaches provide a complementary perspective on the impact of observations in convective-scale NWP. EFSOI measures the observation impact while accounting for the configuration of the data assimilation system. This characteristic is crucial as the actual impact of an observation strongly depends on the applied data assimilation algorithm, the localization length scale, and the composition of the observing network. EFSOI, therefore, is a useful tool to monitor the forecasting system and to detect deficiencies that exist during the assimilation process.

In contrast, the ASC provides information on the relative potential impact of an observation assuming a perfect data assimilation system and that each quantity can be observed equally well. Hence, it reveals the impact of an observation given an optimal data assimilation

process. The potential impact provides guidance on the amplitude of sensitivities within the forecasting system, which is essential for the identification of observations that can have a significant impact on the analysis and forecast. Furthermore, the ASC can indicate sensitive regions or height levels, where observations have the most significant impact on the forecast. Substantial differences between both impact measures for a specific observation type can be used to identify deficiencies in the data assimilation process.
This dissertation assesses the impact and potential impact using ensemble-based methods that rely on sample correlations. However, ensemble systems can only afford a limited ensemble size and have to deal with sampling error. Hence, sampling errors pose an issue for the ensemble-based impact assessment. The last part of this dissertation, therefore, evaluates a sampling error correction (SEC) that can mitigate sampling errors during the impact assessment and for many other ensemble applications.
In summary, this yields three connected studies: 1.) The assessment of observation impact in the convective-scale ensemble forecasting system of Deutscher Wetterdienst. 2.) The estimation of the potential impact of observations for future convective-scale data assimilation. 3.) The evaluation of a sampling error correction for different ensemble applications. Subsequently, the major findings from each part are summarized. A closing remark will highlight and discuss implications for future applications and research.

Observation impact Knowledge of observation impact is particularly required to detect deficiencies in the data assimilation process. Thereby, observation impact studies can improve the data assimilation, forecasting, and observing system. So far, Deutscher Wetterdienst, as well as other centers, do not monitor the impact of observations in their convective-scale systems operationally. The present and previous studies (Sommer and Weissmann, 2014, 2016), therefore, aim towards an operational impact assessment in the future.
The present study represents the first evaluation of short-range observation impact in a convection-permitting modeling system over an extended period of six weeks. The observation impact of about 3.3 million conventional observations is assessed using an ensemble forecast sensitivity to observation impact (EFSOI) method (Kalnay et al., 2012, Sommer and Weissmann, 2016). The EFSOI approach is an efficient tool to quantify the impact of observations using the available ensemble information from the data assimilation cycling. Previous studies by Sommer and Weissmann (2014, 2016) reformulated and tested the EFSOI approach for the COSMO-KENDA system using conventional observations for verification. However, the experimental period was too short for drawing representative conclusions. For that reason, the present study evaluates the observation impact for a 6-week summer period. Moreover, previous studies indicated a sensitivity of the impact on the chosen verification norm (Janisková and Cardinali, 2016, Sommer and Weissmann, 2016). Therefore, this thesis extends the EFSOI approach by including independent radar precipitation observations for verification.
The evaluation of assimilated conventional observations with independent radar observations shows that the dense surface pressure (SYNOP) network provides the largest reduction in forecast error followed by aircraft wind and wind profiler observations. Radiosonde wind, temperature, and humidity observations exhibit a similar impact per observation as found for aircraft observations of the same type. However, due to the relatively small num-

ber of launched radiosondes, aircraft observations have a significantly larger total impact on the regional short-range forecasts of COSMO-DE.
The verification with independent precipitation observations identified the sensitivity of the observation impact on biases and the choice of the verification norm. For example, aircraft temperature observations exhibited a much smaller impact per observation verified with precipitation compared to the verification with conventional observations. This difference turned out to be associated with uncorrected observational biases. As a consequence, the development of a bias correction scheme for aircraft temperature observations is currently ongoing at Deutscher Wetterdienst.
The present study shows that an exceptionally high or low fraction of beneficial observations is an indicator of biases in the assimilation or verification. This means future studies can perform a bias monitoring by investigating the fraction of beneficial observations. Additionally, several sensitivity experiments highlighted that even small biases significantly affect the observation impact. Consequently, future FSOI studies should perform careful monitoring of potential biases and exclude biased observations or analysis fields from verification.
In general, convective-scale short-range forecasts should be verified with independent observations and verification norms that reflect primary forecast quantities. A self-verification is potentially dangerous as biases or correlations between analysis and verification state can significantly affect the observation impact. Recently, a subsequent study for a global modeling system by Kotsuki et al. (2019) comes to similar conclusions comparing different observation-based and model-based verification norms. All these findings are especially of interest for pro-active quality control schemes (Hotta et al. (2017); Lien et al. (2018)). Such systems perform a short-range FSOI step during the cycling to exclude detrimental observations from the assimilation. By re-computing, the analysis without detrimental observations, a more accurate analysis can be achieved. This analysis is then used to initialize long-range forecasts.

Potential impact Another major challenge for NWP is the development of data assimilation and observing system strategies. Considering the vast amount of unused observations and the limited amount of human and computational resources to incorporate all these observations calls for improved data selection strategies. Particularly, better knowledge of the potential impact of observations could help to address these challenges. Therefore, the second study introduces an approach for estimating the potential impact of observable quantities. The approach first was proposed by Geiss (2017) and initially considered accumulated absolute correlations. The updated approach uses accumulated squared correlation as a proxy for the potential impact of different observable quantities. The refinement supports a scale separation that is required to obtain the relative potential impact for convective-scale data assimilation.
The potential impact is based on sample correlations. Therefore, a basic evaluation of spatiotemporal correlations was performed for a short-range precipitation forecast comparing seven different observable atmospheric quantities. The results show that the 1000-member ensemble can return realistic and meaningful spatiotemporal correlations. These correlations exhibit small-scale features, which are traceable in space and time.
Additionally, sensitivity studies on the ensemble size using all ten forecasts show that

the amplitude of potential impact converges reaching an ensemble size of 1000 members. Hence, a 1000-member ensemble can return reliable estimates of the potential impact. A 200-member ensemble can generate sufficient estimates of the potential impact if sampling errors are reduced by applying a sampling error correction. A 40-member ensemble provides some qualitative guidance for quantities with large-scale correlations. Smaller ensembles, however, are not able to estimate the correct amplitude of the potential impact.

To assess the relative potential impact for convective-scale data assimilation, the scale dependence of correlations is additionally taken into account. In principle, a correlation field contains signals from both small-scale and large-scale structures. As convective-scale data assimilation mainly aims to correct small-scale structures, relevant scales need to be filtered. By applying a bandpass filter, large-scale signals can be removed, which provides the relative potential impact for convective-scale data assimilation. The bandpass-filtered ASC shows that highly variable fields such as surface temperature or specific humidity exhibit a strong relative potential impact for short-range precipitation forecasts. Sea-level pressure, upper tropospheric wind, and temperature are dominated by large-scale correlations and, thus, are a task for global data assimilation.

Sampling error correction Usually, ensemble systems can only afford a limited number of ensemble members, which leads to under-sampling. Sampling errors, therefore, are a severe issue for various ensemble applications that rely on sample correlations: First, the EFSOI approach depends on sample correlations (sensitivity gradients). Second, ensemble sensitivity analysis (ESA) is based on spatiotemporal sample correlations. The same applies to the introduced potential impact approach. Finally, sampling errors significantly affect the data assimilation process considering ensemble DA schemes that rely on accurate background error covariance matrices. Under-sampling causes spurious correlations, which significantly degrade the filter performance.

All these examples emphasize the relevance of sampling error for state of the art NWP. The 1000-member ensemble simulation for the first time provides a unique data-set to quantify sampling error on convective scales. Hence, the large ensemble is applied to evaluate a statistical sampling error correction that could improve various ensemble application, including the two approaches that are applied to assess the impact of observations. Here, the 1000-member ensemble is taken as a reference to verify the reduction of sampling error for subsets of the full ensemble. The examined sampling error correction is a simple look-up table based approach and computed applying a Monte-Carlo technique. The SEC returns the sampling error corrected correlation, which only is a function of the sample correlation, the ensemble size, and a prior distribution of correlation values.

The third study examines the SEC for application to ESA using spatiotemporal correlations as well as for application to data assimilation assessing spatial correlations. For spatiotemporal correlations, the SEC can reduce spurious correlations for all evaluated variable combinations and ensemble sizes. Results showed that the SEC corrects correlations of all strength and by this significantly improves the frequency distribution of observed correlation values. For ESA, the 40-member ensemble including SEC even outperformed the 80-member ensemble. This result highlights that the SEC efficiently corrects the overestimation of the correlations due to spurious correlations. Overall, the SEC should be applied to all correlations regardless of their strength even though it has its most signifi-

cant influence on small correlations.
For ESA, the SEC additionally is compared to a confidence test (Torn and Hakim, 2008). A confidence test was applied by several previous ESA studies to mitigate sampling error. The results indicate that the SEC outperforms the confidence test as it significantly reduces spurious correlations while allowing for small correlations. This actively encourages to apply the SEC in applications that rely on spatiotemporal correlations such as ESA or EFSOI.
In the context of data assimilation, the SEC similarly proved to be beneficial by reducing sampling error for spatial correlations. In all cases, the error reduction is largest for long-range correlations. For vertical correlations, the SEC systematically reduced the error in the entire tropospheric column independently of the height level. Compared to operational localization techniques, which damp or exclude long-range correlations, the SEC allows for correlations far from the response level. Such long-range correlations are crucial for the assimilation of non-local or integrated observations (e.g., cloud, satellite radiance or pressure). Horizontally, the SEC reduced the error for nearly all variable pairs. However, the uniform prior, which is applied in the present study, appears sub-optimal for highly positively correlated variables.
Overall, the SEC significantly reduced sampling error for all evaluate ensemble applications. According to Anderson (2016), the SEC performance could even be improved by changing the uniform prior to a more informed prior. Hence, finding suitable priors (or look-up tables) for various situations and applications is an important challenge. Anderson (2016), for example, used previous forecasts to estimate the prior distribution with a climatology. Another possibility would be to estimate the prior from the introduced 1000-member ensemble. Such a prior could be evaluated using an independent second large ensemble simulation.

Implications for future research The two impact estimation approaches provide a complementary perspective on the impact of different observations in convective-scale modeling. Both approaches are applied to estimate the impact of observations on precipitation forecasts. Precipitation is chosen as it is a primary forecast quantity of convective-scale forecasting. For COSMO-KENDA, surface pressure observations exhibit the largest beneficial impact on the short-range precipitation forecast followed by wind, temperature, and humidity. Similar results are found for the potential impact in a different convective-scale forecasting system but without accounting for the scale dependence of the potential impact and only evaluating a single height level. However, when filtering relevant scales for convective-scale DA, highly variable fields with small-scale structures such as near-surface temperature or specific humidity show the most substantial relative potential impact.
This significant relative potential impact for highly variable fields emphasizes the need for high-resolution and dense observing systems for convective-scale DA. For example, radar observations already showed a substantial impact on the practical predictability of convective precipitation (Bachmann et al., 2018). Furthermore, previous experiments performed at MeteoSwiss showed that the assimilation of surface temperature and humidity could have a significant impact within KENDA (Necker and Leuenberger, 2016). However, Deutscher Wetterdienst does not yet assimilate these observations operationally mainly due to representativity errors. Overall, these results call for operational assimilation of

the spatially dense observations such as surface temperature and humidity in KENDA to improve the initial conditions.

The investigation of sampling errors emphasizes the issue of under-sampling during the impact assessment. These results strongly encourage to use the sampling error correction for the estimation of observation impact, for ensemble sensitivity analysis and covariance localization in ensemble data assimilation. Additionally, the SEC could be applied to other ensemble-based applications that rely on sample correlations. Especially in the framework of EFSOI, the SEC could improve the performance considering that the COSMO-KENDA system consists of a relatively small 40-member ensemble. Currently, sampling errors are addressed by applying localization during assimilation and verification, but the SEC could be applied in addition. Overall, the application of the SEC is technically simple, and it is already implemented in DART. As demonstrated, even assuming a rather simple prior turned out to be beneficial for nearly all evaluated variable combinations. With a small effort, the SEC could be improved by using more informed prior, which, for example, could be obtained from the 1000-member ensemble simulation.

In particular, the present EFSOI study shows that ensemble-based estimates of observation impact are a powerful and efficient tool for monitoring the convective-scale forecasting system of Deutscher Wetterdienst. In early 2019, for example, the present study confirmed results from a European wind profiler monitoring study indicating that the Swiss wind profiler in Payerne seems to exhibit a systematically detrimental impact. The search for the cause is still ongoing. Certainly, this study made the next step towards operational monitoring of observation impact in the regional forecasting system of Deutscher Wetterdienst. Currently, Deutscher Wetterdienst is implementing an operational framework for EFSOI.

The approach for estimating the relative potential impact of observations provides a basis for subsequent research on observing and data assimilation strategies for convective-scale NWP. The approach is efficient, and only requires an ensemble of forecasts, which makes it applicable to every ensemble system. Future studies should evaluate the relative potential impact for a broad set of quantities, including synthetic remote sensing data (e.g., synthetic satellite images). Furthermore, the impact time should be taken into account as some quantities seem to have a longer-lasting impact than others (Geiss, 2017). To gain a complete picture of the impact of different quantities, future studies will investigate other forecast aspects that are relevant for regional forecasting such as hub-height winds or solar radiation.

Finally, the unique convective-scale 1000-member ensemble simulation turned out to be an exceptional data-set for performing basic research. Currently, two subsequent projects reuse the 1000-member ensemble, and additional projects are planned. A bachelor thesis investigates frequency distributions of different quantities concerning non-Gaussian nature and multi-modality. Furthermore, the investigation of vertical localization for satellite data assimilation is ongoing in a concurrent project that builds upon the simulation presented in this dissertation.

Appendix A

Contribution of journal publications to this dissertation

This dissertation includes results from three journal articles. The first article (Necker et al., 2018) has already been published in the Quarterly Journal of the Royal Meteorological Society. The second (Necker et al., 2019a) and third (Necker et al., 2019b) article have been submitted for publication to different peer review journals. Table A.1 provides an overview of the contribution of each article to this dissertation. This table lists all sections that are based on content from the publications.

Necker et al. (2018)	**Necker et al. (2019a)**	**Necker et al. (2019b)**
Section 1.3	Section 3.2	Section 6.1
Section 3.1	Section 3.3	Section 6.2
Section 4.1	Section 3.4	Section 6.3
Section 4.2	Section 3.5	
Section 4.3	Section 5.1	
Section 4.4	Section 5.2	
Section 4.5		
Section 4.7		

Table A.1: *Contribution of journal publications to this dissertation.*

List of Abbreviations

Abbreviation	Description
3DVAR	Three Dimensional Variational Data Assimilation
4DVAR	Four Dimensional Variational Data assimilation
AI	Analysis Infuence
AIREP	Aircraft observations
ASC	Accumulated Squared Correlation
BC	Boundary Conditions
BPS	Band-Pass Filter
CFSR	Climate Forecast System Reanalysis
CONV	Conventional
COSMO	COnsortium for Small-scale MOdeling
COSMO-DE	COSMO limited-area model for Germany
COSMO-KENDA	Operational regional forecasting system of DWD
DA	Data Assimilation
DART	Data Assimilation Research Testbed
DBZ	Radar reflectivity
DWD	Deutscher Wetterdienst
E-PROFILE	EUMETNET Profiling Programme for wind observations
ECMWF	European Centre for Medium-Range Weather Forecasts
EnKF	Ensemble Kalman Filter
EPS	Ensemble Prediction System
ESA	Ensemble Sensitivity Analysis
ETKF	Ensemble Transform Kalman Filter
EUMETNET	European Meteorological Network
FC	Forecast
FG	First Guess
FSO	Forecast Sensitivity to Observation
FSOI	Forecast Sensitivity to Observation Impact
EFSOI	Ensemble Forecast Sensitivity to Observation Impact
GFSE	Global Forecasting System Ensemble
GNSS	Global Navigation Satellite Systems
GPSRO	Global Positioning System Radio Occultation
HErZ	Hans-Ertel Centre for Weather Research
HY	Hydrometeors
ICON	Icosahedral Nonhydrostatic global model
IFS	Integrated Forecasting System

IR	Infrared
KENDA	Kilometer-Scale Ensemble-based Data Assimilation
LETKF	Local Ensemble Transform Kalman Filter
LHN	Latent Heat Nudging
MAC	Mean Absolute Correlation
MAE	Mean Absolute Error
METEOSAT	Meteorological Satellite
MeteoSwiss	Swiss Federal Office of Meteorology and Climatology
MFASIS	Method for FAst Satellite Image Simulation
MODIS	Moderate resolution Imaging Spectroradiometer
MW	Microwave
NCEP	National Centers for Environmental Prediction
NIR	Near Infrared
NWP	Numerical Weather Prediction
OBS	Observation
OSE	Observing System Experiment
OSSE	Observation System Simulation Experiment
PREC	Precipitation
PROF	Wind profiler observations
PS	Pressure
QV	Specific humidity
RADOLAN	Radar-based precipitation product
RH	Relative Humidity
RMSE	Root Mean Square Error
RTPS	Relaxation To Prior Spread
SCALE	Scalable Computing for Advanced Library and Environment
SCALE-LETKF	SCALE data assimilation system
SCALE-RM	SCALE - Regional Model
SEC	Sampling Error Correction
SEVIRI	Spinning Enhanced Visible and Infrared Imager
SG	Sensitivity gradient
SYNOP	Surface observations
T	Temperature
T95	T-test with 95 % confidence interval
TEMP	Radiosonde/Sounding observations
TOT_PREC	Total Precipitation
UTC	Coordinated Universal Time

List of Figures

Bibliography

Ancell, B. and G. J. Hakim, 2007: Comparing Adjoint- and Ensemble-Sensitivity Analysis with Applications to Observation Targeting. *Mon. Wea. Rev.*, **135 (12)**, 4117–4134.

Anderson, J., T. Hoar, K. Raeder, H. Liu, N. Collins, R. Torn, and A. Avellano, 2009: The data assimilation research testbed: A community facility. *Bulletin of the American Meteorological Society*, **90 (9)**, 1283–1296.

Anderson, J. L., 2012: Localization and Sampling Error Correction in Ensemble Kalman Filter Data Assimilation. *Mon. Wea. Rev.*, **140 (7)**, 2359–2371.

Anderson, J. L., 2016: Reducing correlation sampling error in ensemble kalman filter data assimilation. *Mon. Wea. Rev.*, **144 (3)**, 913–925.

Bachmann, K., C. Keil, G. C. Craig, M. Weissmann, and C. A. Welzbacher, 2019: Predictability of deep convection in idealized and operational forecasts: Effects of radar data assimilation, orography and synoptic weather regime. *Q. J. R. Meteorol. Soc. (submitted)*.

Bachmann, K., C. Keil, and M. Weissmann, 2018: Impact of radar data assimilation and orography on predictability of deep convection. *Q. J. R. Meteorol. Soc.*, **145 (718)**, 117–130.

Baker, N. L. and R. Daley, 2000: Observation and background adjoint sensitivity in the adaptive observation-targeting problem. *Q. J. R. Meteorol. Soc.*, **126 (565)**, 1431–1454.

Baldauf, M., A. Seifert, J. Förstner, D. Majewski, M. Raschendorfer, and T. Reinhardt, 2011: Operational convective-scale numerical weather prediction with the COSMO model: description and sensitivities. *Mon. Wea. Rev.*, **139 (12)**, 3887–3905.

Bannister, R. N., 2017: A review of operational methods of variational and ensemble-variational data assimilation. *Q. J. R. Meteorol. Soc.*, **143 (703)**, 607–633.

Barrett, A. I., S. L. Gray, D. J. Kirshbaum, N. M. Roberts, D. M. Schultz, and J. G. Fairman, 2015: Synoptic versus orographic control on stationary convective banding. *Q. J. R. Meteorol. Soc.*, **141 (689)**, 1101–1113.

Bauer, P., G. Radnóti, S. Healy, and C. Cardinali, 2014: GNSS Radio Occultation Constellation Observing System Experiments. *Mon. Wea. Rev.*, **142 (2)**, 555–572.

Bauer, P., A. Thorpe, and G. Brunet, 2015: The quiet revolution of numerical weather prediction. *Nature*, **525**, 47.

Baur, F., C. Keil, and G. C. Craig, 2019: Soil moisture - precipitation coupling over central europe: Interactions between surface anomalies at different scales and its dynamical implication. *Q. J. R. Meteorol. Soc.*, **144 (717)**, 2863–2875.

Bednarczyk, C. N. and B. C. Ancell, 2015: Ensemble Sensitivity Analysis Applied to a Southern Plains Convective Event. *Mon. Wea. Rev.*, **143 (1)**, 230–249.

Beljaars, A. C. M. and A. A. M. Holtslag, 1991: Flux parameterization over land surfaces for atmospheric models. *Journal of Applied Meteorology*, **30 (3)**, 327–341.

Benjamin, S. G., B. D. Jamison, W. R. Moninger, S. R. Sahm, B. E. Schwartz, and T. W. Schlatter, 2010: Relative Short-Range Forecast Impact from Aircraft, Profiler, Radiosonde, VAD, GPS-PW, METAR, and Mesonet Observations via the RUC Hourly Assimilation Cycle. *Mon. Wea. Rev.*, **138 (4)**, 1319–1343.

Berman, J. D., R. D. Torn, G. S. Romine, and M. L. Weisman, 2017: Sensitivity of northern great plains convection forecasts to upstream and downstream forecast errors. *Mon. Wea. Rev.*, **145 (6)**, 2141–2163.

Bick, T., C. Simmer, S. Troemel, K. Wapler, H.-J. Hendricks Franssen, K. Stephan, U. Blahak, C. Schraff, H. Reich, Y. Zeng, and R. Potthast, 2016: Assimilation of 3D radar reflectivities with an ensemble Kalman filter on the convective scale. *Q. J. R. Meteorol. Soc.*, **142 (696)**, 1490–1504.

Bishop, C. H., B. J. Etherton, and S. J. Majumdar, 2001: Adaptive sampling with the ensemble transform Kalman filter Part I: Theoretical aspects. *Mon. Wea. Rev.*, **129 (3)**, 420–436.

Bjerknes, V., 1904: Das Problem der Wettervorhersage, betrachtet vom Standpunkte der Mechanik und der Physik. *Meteorol. Z.*

Bouttier, F. and G. Kelly, 2001: Observing-system experiments in the ECMWF 4D-Var data assimilation system. *Q. J. R. Meteorol. Soc.*, **127 (574)**, 1469–1488.

Brousseau, P., G. Desroziers, F. Bouttier, and B. Chapnik, 2014: A posteriori diagnostics of the impact of observations on the AROME-France convective-scale data assimilation system. *Q. J. R. Meteorol. Soc.*, **140 (680)**, 982–994.

Buehner, M., P. Du, and J. Bedard, 2018: A new approach for estimating the observation impact in ensemble variational data assimilation. *Mon. Wea. Rev.*, **146 (2)**, 447–465.

Cardinali, C., 2009: Monitoring the observation impact on the short-range forecast. *Q. J. R. Meteorol. Soc.*, **135 (638)**, 239–250.

Cardinali, C., 2018: Forecast sensitivity observation impact with an observation-only based objective function. *Q. J. R. Meteorol. Soc.*, **144 (716)**, 2089–2098.

Cardinali, C. and S. Healy, 2014: Impact of GPS radio occultation measurements in the ECMWF system using adjoint-based diagnostics. *Q. J. R. Meteorol. Soc.*, **140 (684)**, 2315–2320.

Cardinali, C., S. Pezzulli, and E. Andersson, 2004: Influence-matrix diagnostic of a data assimilation system. *Q. J. R. Meteorol. Soc.*, **130 (603)**, 2767–2786.

Caron, J.-F. and M. Buehner, 2018: Scale-dependent background error covariance localization: Evaluation in a global deterministic weather forecasting system. *Mon. Wea. Rev.*, **146 (5)**, 1367–1381.

Craig, G. C. and T. Selz, 2018: Mesoscale dynamical regimes in the midlatitudes. *Geophysical Research Letters*, **45 (1)**, 410–417.

Desroziers, G., P. Brousseau, and B. Chapnik, 2005: Use of randomization to diagnose the impact of observations on analyses and forecasts. *Q. J. R. Meteorol. Soc.*, **131 (611)**, 2821–2837.

Ehrendorfer, M., R. M. Errico, and K. D. Raeder, 1999: Singular-vector perturbation growth in a primitive equation model with moist physics. *Journal of the Atmospheric Sciences*, **56 (11)**, 1627–1648.

Evensen, G., 1994: Sequential data assimilation with a nonlinear quasi-geostrophic model using monte carlo methods to forecast error statistics. *Journal of Geophysical Research: Oceans*, **99 (C5)**, 10 143–10 162.

Fabry, F., 2010: For How Long Should What Data Be Assimilated for the Mesoscale Forecasting of Convection and Why? Part II: On the Observation Signal from Different Sensors. *Mon. Wea. Rev.*, **138 (1)**, 256–264.

Fabry, F. and J. Sun, 2010: For How Long Should What Data Be Assimilated for the Mesoscale Forecasting of Convection and Why? Part I: On the Propagation of Initial Condition Errors and Their Implications for Data Assimilation. *Mon. Wea. Rev.*, **138 (1)**, 242–255.

Fourrié, N., D. Marchal, F. Rabier, B. Chapnik, and G. Desroziers, 2006: Impact study of the 2003 North Atlantic THORPEX regional campaign. *Q. J. R. Meteorol. Soc.*, **132 (615)**, 275–295.

Gaspari, G. and S. E. Cohn, 1999: Construction of correlation functions in two and three dimensions. *Q. J. R. Meteorol. Soc.*, **125 (554)**, 723–757.

Geiss, S., 2017: Ensemble sensitivity analysis and the potential impact of observed quantities on precipitation forecast. *Master thesis at LMU Munich*.

Gelaro, R., R. H. Langland, S. Pellerin, and R. Todling, 2010: The THORPEX observation impact intercomparison experiment. *Mon. Wea. Rev.*, **138 (11)**, 4009–4025.

Gelaro, R. and Y. Zhu, 2009: Examination of observation impacts derived from observing system experiments (OSEs) and adjoint models. *Tellus A*, **61 (2)**, 179–193.

Greybush, S. J., E. Kalnay, T. Miyoshi, K. Ide, and B. R. Hunt, 2011: Balance and ensemble kalman filter localization techniques. *Mon. Wea. Rev.*, **139 (2)**, 511–522.

Gustafsson, N., T. Janjić, C. Schraff, D. Leuenberger, M. Weissman, H. Reich, P. Brousseau, T. Montmerle, E. Wattrelot, A. Bučánek, M. Mile, R. Hamdi, M. Lindskog, J. Barkmeijer, M. Dahlbom, B. Macpherson, S. Ballard, G. Inverarity, J. Carley, C. Alexander, D. Dowell, S. Liu, Y. Ikuta, and T. Fujita, 2018: Survey of data assimilation methods for convective-scale numerical weather prediction at operational centres. *Q. J. R. Meteorol. Soc.*, **144 (713)**, 1218–1256.

Hakim, G. J. and R. D. Torn, 2008: Ensemble Synoptic Analysis. *Meteorological Monographs*, **33 (55)**, 147–162.

Hanley, K. E., D. J. Kirshbaum, N. M. Roberts, and G. Leoncini, 2013: Sensitivities of a Squall Line over Central Europe in a Convective-Scale Ensemble. *Mon. Wea. Rev.*, **141 (1)**, 112–133.

Harnisch, F., M. Weissmann, C. Cardinali, and M. Wirth, 2011: Experimental assimilation of dial water vapour observations in the ecmwf global model. *Q. J. R. Meteorol. Soc.*, **137 (659)**, 1532–1546.

Harnisch, F., M. Weissmann, and A. Perianez, 2016: Error model for the assimilation of cloud-affected infrared satellite observations in an ensemble data assimilation system. *Q. J. R. Meteorol. Soc.*, **142 (697)**, 1797–1808.

Hill, A. J., C. C. Weiss, and B. C. Ancell, 2016: Ensemble sensitivity analysis for mesoscale forecasts of dryline convection initiation. *Mon. Wea. Rev.*, **144 (11)**, 4161–4182.

Hohenegger, C. and C. Schar, 2007: Atmospheric predictability at synoptic versus cloud-resolving scales. *Bulletin of the American Meteorological Society*, **88 (11)**, 1783–1794.

Holdaway, D., R. Errico, R. Gelaro, and J. G. Kim, 2014: Inclusion of Linearized Moist Physics in NASA's Goddard Earth Observing System Data Assimilation Tools. *Mon. Wea. Rev.*, **142 (1)**, 414–433.

Honda, T., T. Miyoshi, G.-Y. Lien, S. Nishizawa, R. Yoshida, S. A. Adachi, K. Terasaki, K. Okamoto, H. Tomita, and K. Bessho, 2018: Assimilating all-sky himawari-8 satellite infrared radiances: A case of typhoon soudelor (2015). *Mon. Wea. Rev.*, **146 (1)**, 213–229.

Horányi, A., C. Cardinali, and L. Centurioni, 2017: The global numerical weather prediction impact of mean-sea-level pressure observations from drifting buoys. *Q. J. R. Meteorol. Soc.*, **143 (703)**, 974–985.

Horányi, A., C. Cardinali, M. Rennie, and L. Isaksen, 2015a: The assimilation of horizontal line-of-sight wind information into the ECMWF data assimilation and forecasting system. Part I: The assessment of wind impact. *Q. J. R. Meteorol. Soc.*, **141 (689)**, 1223–1232.

Horányi, A., C. Cardinali, M. Rennie, and L. Isaksen, 2015b: The assimilation of horizontal line-of-sight wind information into the ECMWF data assimilation and forecasting system. Part II: The impact of degraded wind observations. *Q. J. R. Meteorol. Soc.*, **141 (689)**, 1233–1243.

Hotta, D., T.-C. Chen, E. Kalnay, Y. Ota, and T. Miyoshi, 2017: Proactive QC: A Fully Flow-Dependent Quality Control Scheme Based on EFSO. *Mon. Wea. Rev.*, **145 (8)**, 3331–3354.

Houtekamer, P. L., X. Deng, H. L. Mitchell, S.-J. Baek, and N. Gagnon, 2014: Higher resolution in an operational ensemble kalman filter. *Mon. Wea. Rev.*, **142 (3)**, 1143–1162.

Houtekamer, P. L. and H. L. Mitchell, 1998: Data assimilation using an ensemble kalman filter technique. *Mon. Wea. Rev.*, **126 (3)**, 796–811.

Houtekamer, P. L. and H. L. Mitchell, 2001: A sequential ensemble kalman filter for atmospheric data assimilation. *Mon. Wea. Rev.*, **129 (1)**, 123–137.

Houtekamer, P. L. and F. Zhang, 2016: Review of the ensemble kalman filter for atmospheric data assimilation. *Mon. Wea. Rev.*, **144 (12)**, 4489–4532.

Hunt, B. R., E. J. Kostelich, and I. Szunyogh, 2007: Efficient data assimilation for spatiotemporal chaos: A local ensemble transform kalman filter. *Physica D: Nonlinear Phenomena*, **230 (1-2)**, 112–126.

Jacques, D. and I. Zawadzki, 2015: The Impacts of Representing the Correlation of Errors in Radar Data Assimilation. Part II: Model Output as Background Estimates. *Mon. Wea. Rev.*, **143 (7)**, 2637–2656.

Janisková, M. and C. Cardinali, 2016: On the impact of the diabatic component in the forecast sensitivity observation impact diagnostics. *Data Assimilation for Atmospheric, Oceanic and Hydrologic Applications*, **(Vol. III)**, 483–511. Springer, Cham.

Jung, B.-J., H. M. Kim, T. Auligne, X. Zhang, X. Zhang, and X.-Y. Huang, 2013: Adjoint-Derived Observation Impact Using WRF in the Western North Pacific. *Mon. Wea. Rev.*, **141 (11)**, 4080–4097.

Kalnay, E., 2003: Atmospheric modeling, data assimilation, and predictability. *Cambridge University Press.*

Kalnay, E., Y. Ota, T. Miyoshi, and J. Liu, 2012: A simpler formulation of forecast sensitivity to observations: application to ensemble Kalman filters. *Tellus A*, **64 (1)**, 18 462.

Keil, C., F. Baur, K. Bachmann, S. Rasp, L. Schneider, and C. Barthlott, 2019: Relative contribution of soil moisture, boundary layer and microphysical perturbations on convective predictability in different weather regimes. *Q. J. R. Meteorol. Soc. (submitted).*

Kelly, G., J. Thepaut, R. Buizza, and C. Cardinali, 2007: The value of observations. I: Data denial experiments for the Atlantic and the Pacific. *Q. J. R. Meteorol. Soc.*, **133 (628)**, 1803–1815.

Kondo, K. and T. Miyoshi, 2016: Impact of removing covariance localization in an ensemble kalman filter: Experiments with 10 240 members using an intermediate agcm. *Mon. Wea. Rev.*, **144 (12)**, 4849–4865.

Kostka, P., M. Weissmann, R. Buras, B. Mayer, and O. Stiller, 2014: Observation Operator for Visible and Near-Infrared Satellite Reflectances. *J. Atmos. Oceanic Technol.*, **31 (6)**, 1216–1233.

Kotsuki, S., K. Kurosawa, and T. Miyoshi, 2019: On the properties of ensemble forecast sensitivity to observations. *Q. J. R. Meteorol. Soc.*, doi:10.1002/qj.3534.

Kunii, M., T. Miyoshi, and E. Kalnay, 2012: Estimating the Impact of Real Observations in Regional Numerical Weather Prediction Using an Ensemble Kalman Filter. *Mon. Wea. Rev.*, **140 (6)**, 1975–1987.

Langland, R. H., 2005: Observation Impact during the North Atlantic TReC—2003. *Mon. Wea. Rev.*, **133 (8)**, 2297–2309.

Langland, R. H. and N. L. Baker, 2004: Estimation of observation impact using the NRL atmospheric variational data assimilation adjoint system. *Tellus A*, **56 (3)**, 189–201.

Lei, L., J. S. Whitaker, and C. Bishop, 2018: Improving assimilation of radiance observations by implementing model space localization in an ensemble kalman filter. *Journal of Advances in Modeling Earth Systems*, **10 (12)**, 3221–3232.

Leutbecher, M., 2018: Ensemble size: How suboptimal is less than infinity? *Q. J. R. Meteorol. Soc.*, **144**, 1218–1256.

Li, H., J. Liu, and E. Kalnay, 2010: Correction of 'Estimating observation impact without adjoint model in an ensemble Kalman filter'. *Q. J. R. Meteorol. Soc.*, **136 (651)**, 1652–1654.

Lien, G.-Y., D. Hotta, E. Kalnay, T. Miyoshi, and T.-C. Chen, 2018: Accelerating assimilation development for new observing systems using EFSO. *Nonlin. Processes Geophys*, **25 (1)**, 129–143.

Lien, G.-Y., T. Miyoshi, S. Nishizawa, R. Yoshida, H. Yashiro, S. A. Adachi, T. Yamaura, and H. Tomita, 2017: The near-real-time scale-letkf system: A case of the september 2015 kanto-tohoku heavy rainfall. *SOLA*, **13**, 1–6.

Limpert, G. L. and A. L. Houston, 2018: Ensemble sensitivity analysis for targeted observations of supercell thunderstorms. *Mon. Wea. Rev.*, **146 (6)**, 1705–1721.

Liu, J. and E. Kalnay, 2008: Estimating observation impact without adjoint model in an ensemble Kalman filter. *Q. J. R. Meteorol. Soc.*, **134 (634)**, 1327–1335.

Liu, J., E. Kalnay, T. Miyoshi, and C. Cardinali, 2009: Analysis sensitivity calculation in an ensemble Kalman filter. *Q. J. R. Meteorol. Soc.*, **135 (644)**, 1842–1851.

Lorenc, A. C., 1986: Analysis methods for numerical weather prediction. *Q. J. R. Meteorol. Soc.*, **112 (474)**, 1177–1194.

Lorenc, A. C. and R. T. Marriott, 2013: Forecast sensitivity to observations in the Met Office Global numerical weather prediction system. *Q. J. R. Meteorol. Soc.*, **140 (678)**, 209–224.

Lorenz, E. N., 1963: Deterministic nonperiodic flow. *Journal of the Atmospheric Sciences*, **20 (2)**, 130–141.

Lupu, C., C. Cardinali, and A. P. Mcnally, 2015: Adjoint-based forecast sensitivity applied to observation-error variance tuning. *Q. J. R. Meteorol. Soc.*, **141 (693)**, 3157–3165.

Lupu, C., P. Gauthier, and S. Laroche, 2011: Evaluation of the Impact of Observations on Analyses in 3D- and 4D-Var Based on Information Content. *Mon. Wea. Rev.*, **139 (3)**, 726–737.

Lupu, C., P. Gauthier, and S. Laroche, 2012: Assessment of the Impact of Observations on Analyses Derived from Observing System Experiments. *Mon. Wea. Rev.*, **140 (1)**, 245–257.

Majumdar, S. J., S. Aberson, C. Bishop, C. Cardinali, J. Caughey, A. Doerenbecher, P. Gauthier, R. Gelaro, T. Hamill, R. Langland, A. Lorenc, T. Nakazawa, F. Rabier, C. Reynolds, L. Saunders, Y. Song, Z. Toth, C. Velden, M. Weissmann, and C. Wu, 2011: Target observations for improving numerical weather precidiction: an overview. *World Weather Research Programme / THORPEX Publication*, **15**, 37 pp.

Mallick, S., D. Dutta, and K. H. Min, 2017: Quality assessment and forecast sensitivity of global remote sensing observations. *Adv. Atmos. Sci.*, **34 (3)**, 371–382.

Miyoshi, T., K. Kondo, and T. Imamura, 2014: The 10,240-member ensemble kalman filtering with an intermediate agcm. *Geophysical Research Letters*, **41 (14)**, 5264–5271.

Miyoshi, T., K. Kondo, and K. Terasaki, 2015: Big ensemble data assimilation in numerical weather prediction. *Computer*, **48 (11)**, 15–21.

Miyoshi, T., M. Kunii, J. Ruiz, G.-Y. Lien, S. Satoh, T. Ushio, K. Bessho, H. Seko, H. Tomita, and Y. Ishikawa, 2016a: Big data assimilation revolutionizing severe weather prediction. *Bulletin of the American Meteorological Society*, **97 (8)**, 1347–1354.

Miyoshi, T., G. Lien, S. Satoh, T. Ushio, K. Bessho, H. Tomita, S. Nishizawa, R. Yoshida, S. A. Adachi, J. Liao, B. Gerofi, Y. Ishikawa, M. Kunii, J. Ruiz, Y. Maejima, S. Otsuka, M. Otsuka, K. Okamoto, and H. Seko, 2016b: "big data assimilation" toward post-petascale severe weather prediction: An overview and progress. *Proceedings of the IEEE*, **104 (11)**, 2155–2179.

Nakanishi, M. and H. Niino, 2004: An improved mellor–yamada level-3 model with condensation physics: Its design and verification. *Boundary-Layer Meteorology*, **112 (1)**, 1–31.

Necker, T., S. Geiss, M. Weissmann, J. Ruiz, T. Miyoshi, and G.-Y. Lien, 2019a: A convective-scale 1000-member ensemble simulation and potential applications. *Q. J. R. Meteorol. Soc. (submitted)*.

Necker, T. and D. Leuenberger, 2016: Assimilation of surface observations in KENDA. *COSMO Activity – Technical report*.

Necker, T., M. Weissmann, Y. Ruckstuhl, J. Anderson, and T. Miyoshi, 2019b: Sampling error correction evaluated using a convective-scale 1000-member ensemble. *Mon. Wea. Rev. (submitted)*.

Necker, T., M. Weissmann, and M. Sommer, 2018: The importance of appropriate verification metrics for the assessment of observation impact in a convection-permitting modelling system. *Q. J. R. Meteorol. Soc.*, **144 (714)**, 1667–1680.

Nishizawa, S. and Y. Kitamura, 2018: A surface flux scheme based on the monin-obukhov similarity for finite volume models. *Journal of Advances in Modeling Earth Systems*, **10 (12)**, 3159–3175.

Nishizawa, S., H. Yashiro, Y. Sato, Y. Miyamoto, and H. Tomita, 2015: Influence of grid aspect ratio on planetary boundary layer turbulence in large-eddy simulations. *Geoscientific Model Development*, **8 (10)**, 3393–3419.

Ota, Y., J. C. Derber, E. Kalnay, and T. Miyoshi, 2013: Ensemble-based observation impact estimates using the NCEP GFS. *Tellus A*, **65 (1)**, 20 038.

Park, S. K. and L. Xu, 2009: Data assimilation for atmospheric, oceanic and hydrologic applications. *Springer-Verlag*.

Piper, D., M. Kunz, F. Ehmele, S. Mohr, B. Mühr, A. Kron, and J. Daniell, 2016: Exceptional sequence of severe thunderstorms and related flash floods in May and June 2016 in Germany – Part 1: Meteorological background. *Natural Hazards and Earth System Sciences*, **16 (12)**, 2835–2850.

Privé, N. C., R. M. Errico, and K.-S. Tai, 2014: The Impact of Increased Frequency of Rawinsonde Observations on Forecast Skill Investigated with an Observing System Simulation Experiment. *Mon. Wea. Rev.*, **142 (5)**, 1823–1834.

Rabier, F., E. Klinker, P. Courtier, and A. Hollingsworth, 1996: Sensitivity of forecast errors to initial conditions. *Q. J. R. Meteorol. Soc.*, **122 (529)**, 121–150.

Rasp, S., T. Selz, and G. C. Craig, 2018: Variability and clustering of midlatitude summertime convection: Testing the craig and cohen theory in a convection-permitting ensemble with stochastic boundary layer perturbations. *Journal of the Atmospheric Sciences*, **75 (2)**, 691–706.

Rhodin, A., H. Lange, R. Potthast, and T. Janjic-Pfander, 2013: Documentation of the dwd data assimilation system. Deutscher Wetterdienst, 402 pp.

Ruefenacht, R., M. Hervo, A. Haefele, L. Isaksen, and S. Hafner, 2018: Case4: FSO-based impact study of E-PROFILE wind observations in comparison to AMDAR. *Deutscher Wetterdienst - Internal report*.

Saha, S., et al., 2010: The ncep climate forecast system reanalysis. *Bulletin of the American Meteorological Society*, **91 (8)**, 1015–1058.

Sato, Y., S. Nishizawa, H. Yashiro, Y. Miyamoto, Y. Kajikawa, and H. Tomita, 2015: Impacts of cloud microphysics on trade wind cumulus: which cloud microphysics processes contribute to the diversity in a large eddy simulation? *Progress in Earth and Planetary Science*, **2 (1)**, 23.

Sawada, Y., K. Okamoto, M. Kunii, and T. Miyoshi, 2019: Assimilating every-10-minute himawari-8 infrared radiances to improve convective predictability. *Journal of Geophysical Research: Atmospheres*, **124 (5)**, 2546–2561.

Scheck, L., P. Frerebeau, R. Buras-Schnell, and B. Mayer, 2016: A fast radiative transfer method for the simulation of visible satellite imagery. *J. Quant. Spectrosc. Radiat. Transfer*, **175**, 54 – 67.

Scheck, L., M. Weissmann, and B. Mayer, 2018: Efficient Methods to Account for Cloud-Top Inclination and Cloud Overlap in Synthetic Visible Satellite Images. *J. Atmos. Oceanic Technol.*, **35 (3)**, 665–685.

Schraff, C., H. Reich, A. Rhodin, A. Schomburg, K. Stephan, A. Perianez, and R. Potthast, 2016: Kilometre-scale ensemble data assimilation for the COSMO model (KENDA). *Q. J. R. Meteorol. Soc.*, **142 (696)**, 1453–1472.

Sekiguchi, M. and T. Nakajima, 2008": A k-distribution-based radiation code and its computational optimization for an atmospheric general circulation model. *Journal of Quantitative Spectroscopy and Radiative Transfer*, **109 (17)**, 2779 – 2793.

Simmer, C., G. Adrian, S. Jones, V. Wirth, M. Goeber, C. Hohenegger, T. Janjic´, J. Keller, C. Ohlwein, A. Seifert, S. Troemel, T. Ulbrich, K. Wapler, M. Weissmann, J. Keller, , M. Masbou, S. Meilinger, N. Riß, A. Schomburg, A. Vormann, and C. Weingaertner, 2016: HErZ: The German Hans-Ertel Centre for Weather Research. *Bull. Amer. Meteor. Soc.*, **97 (6)**, 1057–1068.

Sommer, M. and M. Weissmann, 2014: Observation impact in a convective-scale localized ensemble transform Kalman filter. *Q. J. R. Meteorol. Soc.*, **140 (685)**, 2672–2679.

Sommer, M. and M. Weissmann, 2016: Ensemble-based approximation of observation impact using an observation-based verification metric. *Tellus A*, **68 (1)**, 27 885.

Tomita, H., 2008: New microphysical schemes with five and six categories by diagnostic generation of cloud ice. *Journal of the Meteorological Society of Japan. Ser. II*, **86A**, 121–142.

Torn, R. D., 2010: Ensemble-Based Sensitivity Analysis Applied to African Easterly Waves. *Weather and Forecasting*, **25 (1)**, 61–78.

Torn, R. D. and G. J. Hakim, 2008: Ensemble-Based Sensitivity Analysis. *Mon. Wea. Rev.*, **136 (2)**, 663–677.

Torn, R. D. and G. J. Hakim, 2009: Initial Condition Sensitivity of Western Pacific Extratropical Transitions Determined Using Ensemble-Based Sensitivity Analysis. *Mon. Wea. Rev.*, **137 (10)**, 3388–3406.

van Leeuwen, P. J., 1999: Comment on data assimilation using an ensemble kalman filter technique. *Mon. Wea. Rev.*, **127 (6)**, 1374–1377.

Weissmann, M., R. H. Langland, C. Cardinali, P. M. Pauley, and S. Rahm, 2012: Influence of airborne Doppler wind lidar profiles near Typhoon Sinlaku on ECMWF and NOGAPS forecasts. *Q. J. R. Meteorol. Soc.*, **138 (662)**, 118–130.

Weissmann, M., F. Harnisch, C.-C. Wu, P.-H. Lin, Y. Ohta, K. Yamashita, Y.-H. Kim, E.-H. Jeon, T. Nakazawa, and S. Aberson, 2011: The influence of assimilating dropsonde data on typhoon track and midlatitude forecasts. *Mon. Wea. Rev.*, **139 (3)**, 908–920.

Weissmann, M., M. Goeber, C. Hohenegger, T. Janjic, J. Keller, C. Ohlwein, A. Seifert, S. Troemel, T. Ulbrich, K. Wapler, C. Bollmeyer, and H. Deneke, 2014: Initial phase of the hans-ertel centre for weather research – a virtual centre at the interface of basic and applied weather and climate research. *Meteor. Z.*, **23 (3)**, 193–208.

Whitaker, J. S. and T. M. Hamill, 2012: Evaluating methods to account for system errors in ensemble data assimilation. *Mon. Wea. Rev.*, **140 (9)**, 3078–3089.

Wile, S. M., J. P. Hacker, and K. H. Chilcoat, 2015: The Potential Utility of High-Resolution Ensemble Sensitivity Analysis for Observation Placement during Weak Flow in Complex Terrain. *Weather and Forecasting*, **30 (6)**, 1521–1536.

Zängl, G., D. Reinert, P. Rípodas, and M. Baldauf, 2015: The ICON (ICOsahedral Non-hydrostatic) modelling framework of DWD and MPI-M: Description of the non-hydrostatic dynamical core. *Q. J. R. Meteorol. Soc.*, **141 (687)**, 563–579.

Zeng, Y., U. Blahak, and D. Jerger, 2016: An efficient modular volume-scanning radar forward operator for NWP models: description and coupling to the COSMO model. *Q. J. R. Meteorol. Soc.*, **142 (701)**, 3234–3256.

Zeng, Y., T. Janjic, A. de Lozar, U. Blahak, H. Reich, C. Keil, and A. Seifert, 2018: Representation of model error in convective-scale data assimilation: Additive noise, relaxation methods, and combinations. *Journal of Advances in Modeling Earth Systems*, **10 (11)**, 2889–2911.

Acknowledgements

Allen voran gilt mein größter Dank meinem Betreuer und Doktorvater, Dr. Martin Weissmann. Die gemeinsamen Jahre waren eine tolle Erfahrung, die ich nicht missen möchte. Martin, Du hast es geschafft die Arbeit in die richtige Richtung zu lenken und mir trotzdem großen Freiraum gelassen, um eigene Wege zu gehen, was ich sehr zu schätzen weiß. Danke für viele gute Ratschläge, unzählige beantwortete Fragen und Deine stete Unterstützung auf meinem Weg – fachlich wie menschlich. Wie scherzhaft von Dir angemerkt, gehe ich vielleicht als der Student in die Geschichte ein, der sich bei der Wahl der Masterarbeit am längsten Zeit ließ. Heute kann ich jedoch zurückblicken und mit Überzeugung behaupten, damals die richtige Entscheidung getroffen zu haben.

Ein besonderer Dank gilt auch meinem ehemaligen Betreuer Dr. Matthias Sommer. Lieber Matthias, auch wenn Du bereits im zweiten Jahr meiner Promotion die Uni verlassen hast, wäre diese Arbeit ohne Deine Hilfe nicht in dieser Form zustande gekommen. Danke für deine Unterstützung zu Beginn meiner Promotion. Ich hoffe sehr unsere Wege kreuzen sich ein weiteres Mal.

Meinem Zweitgutachter, Prof. Dr. Bernhard Mayer, möchte ich sehr für die hilfreichen Ratschläge, sowie für die Begutachtung dieser Doktorarbeit danken. Als begeisterter Fußballer, freue ich mich besonders mit dieser Arbeit einen lupenreinen Zeitgutachter-Hattrick geschafft zu haben.

Zwei weiteren Personen, die einen wichtigen Beitrag zu dieser Arbeit geleistet haben, gilt ebenso mein Dank: Stefan Geiss und Yvonne Ruckstuhl. Lieber Stefan, danke für viele Gespräche und die Grundlagen, die Du mit Deiner Masterarbeit gelegt hast. Du hast mir wertvolle Tipps mit auf den Weg gegeben, wodurch mir die Zusammenarbeit mit Japan sehr leicht gefallen ist. Beste Yvonne, ook bedankt voor vele discussies. Je hebt me veel goede ideeën gebracht die veel hebben bijgedragen aan het derde deel van dit proefschrift. Bedankt ook voor het proeflezen.

Ein großer Dank gilt zudem meinen (Ex-)Büro-Kollegen: Fabi, Kathrin, Kevin und Matze. Ohne Euch hätten die letzten 4 Jahre nicht annähernd so viel Spaß gemacht. Vielen Dank für die gemeinsame Zeit, die mir immer in guter Erinnerung bleiben wird.

Keine Arbeit der Welt würde Spaß machen, ohne die passenden Kollegen. Danke an alle Mitarbeiter und Freunde in HErZ und am Institut: Leonhard, Stephan, Miriam, Yuefei, Robert, Anne, Tobi K. - um nur einige zu nennen. Seit meinem Studienbeginn im Jahr 2008 fallen mir zu viele Personen ein, um hier alle persönlich zu erwähnen. An dieser Stelle sollte sich daher jeder angesprochen fühlen, der nicht namentlich erwähnt wurde. Danke für viele schöne Erlebnisse, inspirierenden Diskussionen, spannende Dienstreisen und eine unvergessliche Zeit.

Liebe Barbara, danke auch Dir, dass Du die letzten Jahre immer ein offenes Ohr hattest,

auch bei Themen, die nicht direkt die Arbeit betrafen. Du warst und bist die gute Seele des Instituts.

Ein weiterer Dank gilt auch mehreren Mitarbeitern vom Deutschen und Schweizer Wetterdienst: U.a. Roland Potthast, Hendrik Reich, Klaus Stefan und Daniel Leuenberger. Danke für viele gute Ratschläge und Diskussionen. Ohne die bereitgestellten Daten und Rechenzeit, sowie ohne die Hilfsstellungen beim Aufsetzen der Simulationen, wäre der erste Teil dieser Doktorarbeit nicht realisierbar gewesen.

I would also like to acknowledge all co-authors from my publications that have not been mentioned elsewhere: Takemasa Miyoshi for providing access to the K Computer. Jeffrey Anderson for guidance with the sampling error correction. Juan Ruiz and Guo-Yan Lien for important discussions and contributions, which helped to improve the 1000-member ensemble simulation.

Arigato ko sai mas to all people from the data assimilation group at the RIKEN Advanced Institute for Computational Science. I really enjoyed my two visits in Japan during the course of this PhD. Thank you for your warm welcome and your assistance with the K computer system. Your help made it easy to handle almost 8×10^6 CPU hours of computing time and to transfer more than 100 terabytes of data. It was a great experience to be able to work on one of the most powerful supercomputers in the world.

Zu allerletzt und am wichtigsten: Meine Freunde, Familie und Freundin. Danke Mario, Halli und Julius für eure Unterstützung und die gemeinsame Zeit in der G7. Danke an meine Eltern und Brüder, dass Ihr jederzeit für mich da seid. Danke Julia, dass du immer an mich glaubst und mich so liebst wie ich bin. Es gibt kein schöneres Gefühl als dich an meiner Seite zu wissen.

HErZ This doctoral thesis was carried out in the Hans-Ertel Centre for Weather Research (HErZ; Weissmann et al. (2014); Simmer et al. (2016)), which is funded by the BMVI (Federal Ministry of Transport and Digital Infrastructure). This German research network includes universities, research institutes, and Deutscher Wetterdienst. HErZ aims to better connect fundamental meteorological research and teaching at German universities supporting the German national weather service. The HErZ branch in Munich is one out of five branches and has a particular focus on data assimilation and ensemble prediction in convection-permitting modelling systems.

RIKEN Finally, it should be noted that the 1000-member ensemble simulation used computational resources of the K computer in Kobe, Japan provided by the RIKEN Center for Computational Science through the HPCI System Research project (Project ID:ra000015, ra001011).