ABSTRACT

Title of dissertation: INFORMATION SYNTHESIS ACROSS SCALES IN ATMOSPHERIC STATE ESTIMATION: THEORY AND NUMERICAL EXPERIMENTS

Matthew Kretschmer, Doctor of Philosophy, 2015

Dissertation directed by: Professor Edward Ott
Department of Physics

This thesis studies the benefits of simultaneously considering system information from different sources when performing ensemble data assimilation. In particular, in Chapter 2 we consider ensemble data assimilation using both a global dynamical model and climatological forecast error information, and, in Chapters 3 and 4, using both a global dynamical model and at least one higher-resolution limited-area dynamical model. Focus is given to applying data assimilation for atmospheric state estimation. Introductory material on ensemble forecasting is given in Chapter 1.

In Chapter 2, I first investigate how the forecast background-error climatology can be used to help improve state estimates, and subsequent forecasts initialized from those state estimates. "Climatological perturbations" derived from an estimate of the background-error covariance matrix are added to the dynamic ensemble that has been forecasted from the previous analysis time, enlarging the space of possible analysis increments. Numerical experiments on a one-dimensional toy model test

this method and illustrate that climatologically augmenting the dynamical forecast ensemble during the analysis has a positive impact on state estimation and forecast accuracy.

Chapter 3 studies data assimilation that considers state information from various spatial scales. In practice, it is common for regional-scale weather forecasts to be created using limited-area atmospheric models which have relatively high spatial resolution. Limited-area model forecasts require lateral boundary conditions, which often come from a lower resolution forecast model (with different model physics) defined over a larger, often global, domain. Here I describe how data assimilation may be performed on a composite forecast state containing information from all available forecast models, and show results from numerical experiments that detail the benefits of this approach.

Chapter 4 of this thesis explores forecast model bias, which is the result of uncertain, unknown or incorrect model physics. I adapt a strategy for correcting forecast model bias to use when performing data assimilation using the composite state method described in Chapter 3. In numerical experiments, I test this bias correction strategy for differently biased global and limited-area models, and observe that analysis and forecast accuracy is dramatically improved when compared to forecasts made without bias correction.

INFORMATION SYNTHESIS ACROSS SCALES IN ATMOSPHERIC STATE ESTIMATION: THEORY AND NUMERICAL EXPERIMENTS

by

Matthew Kretschmer

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2015

Advisory Committee:
Professor Edward Ott, Chair
Professor Brian Hunt
Professor Michelle Girvan
Professor Daryl Kleist
Professor James Yorke

Foreword

Chapters 2 to 4 of this thesis are based upon the following publications.

1. Kretschmer, M., Hunt, B. R., Ott, E., Bishop, C.H., Rainwater, S., Szunyogh, I. 2015. A Composite State Method for Ensemble Data Assimilation with Multiple Limited-Area Models. **In Press.** *Tellus A*.

2. Kretschmer, M., Hunt, B. R., Ott, E. 2015. Data Assimilation Using a Climatologically Augmented Local Ensemble Transform Kalman Filter. **Under Review.** *Tellus A*.

3. Kretschmer, M., Hunt, B. R., Ott, E. 2015. Estimating Forecast Model Bias in Coupled Global and Limited-Area Models. **Submitted.** *Tellus A*.

Dedication

For Katie.

Acknowledgments

I first would like to thank my advisors, Professor Edward Ott and Professor Brian Hunt, for their guidance and generosity. I am very grateful to them for introducing me to data assimilation, and giving me the opportunity to pursue challenging and interesting research projects. I would also like to thank Professors Michelle Girvan, Daryl Kleist, and James Yorke for being on my committee and providing valuable feedback on this thesis.

I am thankful to Young-noh Yoon, for providing invaluable help when I was getting started on my first project, which would eventually turn into Chapter 3. Further thanks are owed to my colleagues Sabrina Rainwater, Craig Bishop, and Istvan Szunyogh, for helping me formulate my findings so as to be most useful to operational weather forecasters. Without the generous support of the Office of Naval Research, this work would truly have not been possible.

Throughout my time as a graduate student, I have enjoyed many scientific and non-scientific discussions with friends and colleagues, including Zhixin Lu, Shane Squires, Wai Lim Ku, and Mark Herrera. Along these lines, I also owe thanks to Thomas Rensink, Joshua Parker, Paul Syers, Matthew Harrington, Jeffrey Magill and Brian Calvert, for making graduate school more enjoyable.

I am incredibly grateful for my family, for their unwavering support throughout my life and education. Finally, I am grateful to my girlfriend, Katie, whose love and encouragement I have found invaluable.

Table of Contents

List of Figures

List of Abbreviations

LAM	Limited-Area Model
LETKF	Local Ensemble Transform Kalman Filter
EnKF	Ensemble Kalman Filer
CSM	Composite State Method

Chapter 1: A (Brief) Introduction to Data Assimilation

Since the first attempts at using computers to create weather forecasts by numerically integrating atmospheric equations of motion, numerical weather prediction has made remarkable improvements in accuracy and skill that have lead to better and better forecasts at longer and longer lead times. The availability of more powerful computers has played a large role in these advancements, as it has allowed simulations of the atmosphere with increasingly complex models and ever increasing spatial and temporal resolutions. Crucially, increased computational power also allows the use of more sophisticated data assimilation algorithms, which combine information from an ever growing variety and quantity of observations with prior forecast information. This results in improvements to the initial conditions from which subsequent forecasts are initialized.

This thesis presents new data assimilation methods that consider extra sources of information when combining observations and forecast state estimates. The two main contributions are a novel method that considers both short-term and long-term estimates of the underlying forecast uncertainty when performing ensemble data assimilation, and a data assimilation framework that uses model forecasts for multiple spatial scales. Before these ideas are presented, we first introduce key data

assimilation concepts and terminology, and review the advantages and pitfalls of popular state estimation techniques, including the traditional Kalman filter and its variants.

1.1 Data Assimilation and Bayes' Rule

Forecasting the weather requires three ingredients, a computational model that can be used to forecast an estimate of the state of the atmosphere at different geographic locations from suitable initial conditions, a measurement system [satellites, radiosondes (weather balloons), radars, etc.] and the ability to combine information from recent measurements with prior forecasts to yield an accurate estimate of the current atmospheric state. This third step is known as *data assimilation*, and is the subject of this thesis. Data assimilation is at its heart a direct application of Bayes' rule, as it seeks to determine the posterior probability distribution of possible state vectors $\mathbf{x}$, given a set of observations $\mathbf{y}$. The state vector $\mathbf{x}$ is used to represent a given atmospheric state; its components are the values of atmospheric state variables at the model grid points. Specifically, for data assimilation the posterior probability distribution $p(\mathbf{x}|\mathbf{y})$ given by Bayes' rule takes the form

$$p(\mathbf{x}|\mathbf{y}) = \frac{p(\mathbf{y}|\mathbf{x})p(\mathbf{x})}{p(\mathbf{y})}. \tag{1.1}$$

For the discussion here we assume that the probability distribution functions in eq. (1.1), $p(\mathbf{x})$, $p(\mathbf{x}|\mathbf{y})$, $p(\mathbf{y}|\mathbf{x})$ and $p(\mathbf{y})$, are Gaussian. The posterior distribution $p(\mathbf{x}|\mathbf{y})$ expresses the likelihood of a true state $\mathbf{x}$ as a product of the likelihood of the

observations given that state, $p(\mathbf{y}|\mathbf{x})$, and the prior likelihood $p(\mathbf{x})$ of the state $\mathbf{x}$, as estimated by a forecast model. Here the probability $p(\mathbf{y}|\mathbf{x})$ can be interpreted as characterizing the accuracy of the measurements. In eq. (1.1), the likelihood of the observations, $p(\mathbf{y})$, is regarded as a normalizing constant, and is often disregarded in data assimilation problems.

The distributions $p(\mathbf{x})$ and $p(\mathbf{y}|\mathbf{x})$ in eq. (1.1) can be thought of as characterizing the distributions of possible model forecast and observation errors, respectively. Assuming the model forecast and instrument errors follow Gaussian distributions, $p(\mathbf{x})$ and $p(\mathbf{y}|\mathbf{x})$ can be expressed as $p(\mathbf{x}) \sim e^{-\frac{1}{2}J_B(\mathbf{x})}$ and $p(\mathbf{y}|\mathbf{x}) \sim e^{-\frac{1}{2}J_O(\mathbf{x})}$, where $J_B(\mathbf{x})$ and $J_O(\mathbf{x})$ are the quadratic forms $J_B(\mathbf{x}) = (\mathbf{x} - \mathbf{x}^b)^T(\mathbf{P}^b)^{-1}(\mathbf{x} - \mathbf{x}^b)$ and $J_O(\mathbf{x}) = (\mathbf{y} - h(\mathbf{x}))^T\mathbf{R}^{-1}(\mathbf{y} - h(\mathbf{x}))$. Using this notation, the model forecast and instrument errors are given by $\mathbf{x} - \mathbf{x}^b$ and $\mathbf{y} - h(\mathbf{x})$, respectively. Here, $\mathbf{x}$ and $\mathbf{y}$ are N- and l-dimensional vectors, respectively. Importantly, the observations $\mathbf{y}$ may not be of state variables, and the observation operator h is used to convert from the N-dimensional state space to the l-dimensional observation space. The matrices $\mathbf{P}^b$ and $\mathbf{R}$ are the background-error and observation-error covariance matrices, respectively. The matrix $\mathbf{P}^b$ contains the (pre-analysis) uncertainty in the components of the state vector $\mathbf{x}$, and includes cross-correlations between the different state vector components. The observation operator h and the matrix $\mathbf{R}$ encode correlations between the model state variables and observations, and $\mathbf{R}$ may be thought of as specifying the uncertainties in observational measurements.

The state that maximizes $p(\mathbf{x})$ in eq. (1.1) is referred to as the *background state*, and is denoted using $\mathbf{x}^b$. The background state represents the most likely state

when considering only forecast information. Using the form of the distributions $p(\mathbf{x})$ and $p(\mathbf{y}|\mathbf{x})$ given above under the Gaussian assumption, the posterior probability distribution from eq. (1.1) may be expressed as

$$p(\mathbf{x}|\mathbf{y}) \approx e^{-\frac{1}{2}J(\mathbf{x})}. \tag{1.2}$$

In eq. (1.2), the "cost function" $J(\mathbf{x})$ is given by $J(\mathbf{x}) = J_B(\mathbf{x}) + J_O(\mathbf{x})$,

$$J(\mathbf{x}) = (\mathbf{x} - \mathbf{x}^b)^T (\mathbf{P}^b)^{-1} (\mathbf{x} - \mathbf{x}^b) + (\mathbf{y} - h(\mathbf{x}))^T \mathbf{R}^{-1} (\mathbf{y} - h(\mathbf{x})). \tag{1.3}$$

Given the observations $\mathbf{y}$, estimates of the matrices $\mathbf{P}^b$ and $\mathbf{R}$ and an estimate of $\mathbf{x}^b$, the most likely state is the value of $\mathbf{x}$ that minimizes $J(\mathbf{x})$, and hence maximizes $p(\mathbf{x}|\mathbf{y})$. This state is referred to as the *analysis state*, and is denoted with $\mathbf{x}^a$. The analysis state represents the best guess of the true system state. Any procedure which finds an updated analysis state by applying a correction, derived from observational information, to a background state is conventionally referred to as an *analysis procedure*.

1.2 Kalman Filters

One of the most well known and widely used techniques for state-estimation and data assimilation is the Kalman filter (Kalman, 1960), which finds the most likely state and its error statistics when there is an exact model of the observed system, the observed system has linear dynamics $\mathbf{M}$, the observation operator h is linear (expressed $h(\mathbf{x}) = \mathbf{H}\mathbf{x}$), and the background- and observation-error covariances $\mathbf{P}^b$ and $\mathbf{R}$ are known. Under these assumptions, $J(\mathbf{x})$ in eq. (1.3) can be

re-expressed as $J(\mathbf{x}) = (\mathbf{x} - \mathbf{x}^a)^T(\mathbf{P}^a)^{-1}(\mathbf{x} - \mathbf{x}^a)$, where the analysis state vector $\mathbf{x}^a$ is

$$\mathbf{x}^a = \mathbf{x}^b + \mathbf{K}(\mathbf{y} - \mathbf{H}\mathbf{x}^b), \tag{1.4}$$

and the uncertainty of the analysis state estimate is given by the analysis-error covariance matrix $\mathbf{P}^a$,

$$\mathbf{P}^a = (\mathbf{I} - \mathbf{K}\mathbf{H})\mathbf{P}^b. \tag{1.5}$$

In eqs. (1.4) and (1.5), the matrix $\mathbf{K}$ is referred to as the *Kalman gain* matrix,

$$\mathbf{K} = \mathbf{P}^b\mathbf{H}^T(\mathbf{H}\mathbf{P}^b\mathbf{H}^T + \mathbf{R})^{-1}. \tag{1.6}$$

The correction to the background state estimate, given in eq. (1.4) by $\mathbf{K}(\mathbf{y} - \mathbf{H}\mathbf{x}^b)$, is known as the *analysis increment*.

The Kalman filter uses eqs. (1.4)-(1.6) to find, at time t, the analysis state $\mathbf{x}_t^a$ and its uncertainty $\mathbf{P}_t^a$, given observations $\mathbf{y}_t$, a background state estimate $\mathbf{x}_t^b$, and an estimate of the uncertainty in the background state estimate $\mathbf{P}_t^b$. For the discussion here, we assume that the linear observation operator h and the observation-error covariance matrix $\mathbf{R}$ are time independent. Background state estimates at a time t' in the future are generated by forecasting the analysis state estimate and its uncertainty valid at time t, $\mathbf{x}_t^a$ and $\mathbf{P}_t^a$, using the linear dynamics $\mathbf{M}$ according to

$$\begin{aligned} \mathbf{x}_{t'}^b &= \mathbf{M}\mathbf{x}_t^a, \\ \mathbf{P}_{t'}^b &= \mathbf{M}\mathbf{P}_t^a\mathbf{M}^T, \end{aligned} \tag{1.7}$$

where $\mathbf{M}$ in eq. (1.7) is a matrix that represents the linear dynamics of the forecast model; the background state vector at time t' is found by simply multiplying the

analysis state vector at time t by $\mathbf{M}$. In eq. (1.7), the superscript T is used to denote the transpose operation. The forecasted state estimate $\mathbf{x}^b_{t'}$ and its uncertainty $\mathbf{P}^b_{t'}$ specify a new prior distribution $p(\mathbf{x})$ in eq. (1.1). Together eqs. (1.4)-(1.7) represent the classical Kalman filter. Repeating the analysis and forecast procedures at future times forms a closed cycle of operations called the data assimilation *analysis cycle*.

A modified version of the Kalman filter known as the extended Kalman filter (Jazwinski, 2007) can be utilized to make approximate state estimates when the dynamics or observation operator are nonlinear. However, these estimates are very poor if the linearized observation operator and linearized dynamics provide poor approximations to their respective nonlinear quantities over the analysis window. In operational weather prediction, where the system is very high-dimensional, there is a huge computational cost associated with the extended Kalman filter; so much so that it would be completely impractical in such a setting. Even more importantly, it is thought that linearization of the dynamics is inappropriate, as it prevents the error covariance matrices from capturing dynamical modes and instabilities that exhibit nonlinear growth. A heuristically motivated way of addressing this issue is the ensemble Kalman filter (discussed in the next section).

Before the inception of computationally affordable methods for approximately forecasting the uncertainties of state estimates using the ensemble Kalman filter, the predominant techniques for performing atmospheric data assimilation were the *variational* methods, which continue to be widely used. Variational data assimilation uses a time-independent $\mathbf{P}^b$ in Eq. 1.3. The choice of the static $\mathbf{P}^b$ is optimized to minimize the time-averaged (and model specific) forecast error, and is usually

estimated from a history of forecast errors. Despite having high rank, the time independent estimates of the background-error covariance matrices used in variational methods do not explicitly contain important time and flow dependent structures and correlations between model state variables. Advanced variational methods, such as 4D-Var, can implicitly represent some time dependent correlations over the analysis time window; however the background-error covariance at the beginning of each time window is time-independent. The lack of explicit time dependent correlations represents a major flaw of variational methods, as the use of time dependent background-error covariances has been shown to realize substantial gains in forecast accuracy (Bishop et al., 2001; Ott et al., 2004; Hunt et al., 2007). An additional weakness of variational methods is the need for a linearized version of the forecast model and its adjoint, which are costly to develop and maintain.

1.3 The Ensemble Kalman Filter

When performing state estimation, the uncertainty of a state estimate $\mathbf{x}^b$, given by $\mathbf{P}^b$, evolves dynamically in time over each analysis cycle. However, as discussed above, computational limitations make explicitly forecasting these matrices impossible when the state dimension N is very large, which is very common in geophysical applications, where N is often on the order of 10^8. Fortunately, ensemble Kalman filters utilizing localization, a technique that will be discussed more below, are one potential solution to this problem. Ensemble Kalman filters use a collection ('ensemble') of Monte Carlo state samples to approximate the background and

analysis error probability distributions. The mean of the ensemble of Monte Carlo state samples is interpreted as the most likely state, and its uncertainty is given by the ensemble sample covariance matrix. To forecast the ensemble of state samples, the full nonlinear model dynamics is used to forecast each of the ensemble members forward in time.

The ensemble *perturbations* are important quantities in ensemble data assimilation; each ensemble perturbation is given by the difference between a given ensemble member state and the ensemble mean state. Ensemble Kalman filters use available observations to update the collection of ensemble members so that the analysis and background ensemble means obey eq. (1.4) and the analysis and background ensemble sample covariances (approximately) obey eq. (1.5). Despite being sub-optimal estimators for large, highly non-linear systems like the atmosphere, as a result of not meeting the linearity conditions of the Kalman filter discussed above, ensemble forecasting techniques have enjoyed great success in weather forecasting. Even though ensemble Kalman filters do not, in principle, treat the time dependent evolution of the error covariances exactly, numerical experiments on models of atmospheric dynamics indicate that they still do well at producing accurate forecasts at reasonable computational cost.

Computational affordability severely limits the maximum possible ensemble size of ensemble Kalman filters; a state of the art operational implementation at Environment Canada (the Canadian National Weather Service) has $k = 192$ ensemble members (Houtekamer et al., 2014). Small ensembles, relative to the number of model degrees of freedom, are not able to adequately sample the full distribution of

possible forecast states, leading to covariance estimates suffering from severe *sampling error*. The technique of "localization" (described below) can greatly mitigate sampling error and has proven to be the key idea allowing application of ensemble Kalman filtering to numerical weather forecasting. However, localization is not a perfect solution to the rank deficiency problem, as it requires empirical tuning and can result in analyses with undesirable qualities, such as increased dynamical imbalances (Greybush et al., 2011). In all studies presented in this thesis, data assimilation is performed utilizing ensemble Kalman filtering with localization.

Localization in ensemble Kalman filters was motivated in part by the paper of Patil et al. (2001), which found that in sufficiently small geographic regions and over sufficiently short time scales, the real atmosphere exhibits low-dimensional dynamics. Specifically, Patil et al. (2001) found that forecasts produced by an atmospheric general circulation model used by the National Weather Service exhibited variability in a space of much smaller dimension than the full state space of the global atmospheric model. This finding had implications for ensemble Kalman filters, as it meant that only a few ensemble members would be needed to form a full-rank estimate of the *local* background-error covariance matrix, and calculating the Kalman gain matrix in eq. (1.6) would thus only require inverting correspondingly low-dimensional matrices. Performing data assimilation at each grid point using the local, low-dimensional background-error covariance matrix would then only allow information from observations within a local region to correct the forecast at that grid point. Furthermore, another crucial point is that this allows the analyses for each grid point to be performed independently and in parallel.

The background-error covariance matrix is crucial in determining the analysis increment, as it determines the space of possible corrections that can be applied to forecasts to account for observations. More specifically, the column space of the background-error covariance matrix is the space of all possible analysis increments available to the Kalman filter algorithm. Rank deficiency of the background-error covariance matrix can severely limit the possible analysis increments which may be applied to a background state estimate, restricting the amount of information which can be gained from observations. In addition to localization, the rank of the background-error covariance estimate in ensemble Kalman filters can be increased by enlarging the ensemble size, but this can represent a substantial computational burden, as this increases the required number of dynamical forecasts which must be made.

There are several different variants of the ensemble Kalman filter; this thesis focuses discussion on the Local Ensemble Transform Kalman Filter (LETKF) (Ott et al., 2004; Hunt et al., 2007). The LETKF finds a linear transformation to apply to the background ensemble members such that the resulting analysis ensemble mean and ensemble covariance approximately obey the Kalman Filter equations. The LETKF updates the ensemble at each model grid point, performing localization by only assimilating observations within an empirically tuned geographic radius.

Hybrid data assimilation methods have been developed to estimate background-error covariances that have both a greater degree of flow-dependent features than the static background covariances used by variational methods and a substantially higher rank than those possible with ensemble Kalman filters, while minimizing ad-

ditional computational expense. Hybrid data assimilation typically achieves this by linearly combining ensemble derived flow-dependent estimates of the background-error covariance matrix with the static covariances utilized by variational methods. Often, hybrid data assimilation involves running a pair of ensemble and variational data assimilation systems (Hamill and Snyder, 2000; Lorenc, 2003; Wang et al., 2013). The first step performed in a hybrid analysis procedure uses a variational analysis to minimize the cost function of eq. (1.3), using for the background-error covariance a linear combination of a flow-dependent, ensemble-derived estimate of the background-error covariance matrix and a higher-rank, static estimate of the background-error covariance matrix. Simultaneously, the ensemble perturbations are updated to correct for the most recent observational information. Upon completion of the variational minimization procedure, fully-coupled hybrid methods form updated ensemble members by adding the analysis state that minimizes the variational cost function to each updated ensemble perturbation. The resulting ensemble may then be forecast to the next analysis time, where the cycle can be repeated. In contrast, one-way coupled hybrid data assimilation, which does not re-center the ensemble about the variational analysis state, instead updates the ensemble mean and perturbations independently of the variational update (Wang et al., 2013).

1.4 New Findings

Chapter 2 of this thesis presents a new hybrid-like data assimilation method that modifies the ensemble Kalman filter to incorporate information from a static

background-error covariance matrix while staying in a pure ensemble assimilation framework. The resulting climatologically augmented Local Ensemble Transform Kalman Filter creates additional ensemble members from the static background-error covariance matrix, so that the implicit background-error covariance matrix estimated by the ensemble is approximately a linear combination of static and flow-dependent covariance matrices, in analogy with the traditional hybrid methods discussed above. The addition of these new ensemble members enhances the rank of the background-error covariance estimated by the ensemble, and allows the LETKF to search for analysis increments in state space directions potentially missed by the dynamically forecast ensemble members. We find that this technique provides a potentially attractive way of increasing the accuracy of analyses and forecasts, particularly in settings with constrained computational resources.

Chapters 3 and 4 focus on applying data assimilation to situations where different forecast models provide background information at different spatial scales. When forecasting the weather, estimating the state of the atmosphere at high spatial resolution over an entire global domain can be computationally infeasible. Limited-area models allow for higher spatial resolution (and hence, potentially more accurate) forecasts to be made over limited geographic regions of interest, without the requisite observational and computational resources needed by a fully global high resolution forecast model (which can be beyond reach for the desired resolution). However, limited-area models require boundary conditions, which often come from coarser-resolution models defined on larger domains. For such situations, data assimilation has commonly been carried out separately on limited-area models and the

larger scale, coarser models supplying their lateral boundary conditions. Chapter 3 presents a method that couples limited-area and global models during the analysis procedure, assimilating observations into a "composite" state vector created from all available model forecasts. The updated analysis composite state vector is then used to initialize subsequent forecasts made with all of the models. The composite state method combines state information from different spatial scales in order to best represent an estimate of the true state of the atmosphere. Numerical tests are presented that demonstrate the improvements that can potentially be realized by our composite state method.

The fourth and final chapter of this thesis presents a method for estimating and accounting for forecast model bias in the setting of Chapter 3. Forecast model bias can occur because of many factors, including incorrectly specified (or unknown) model physics parameterizations and the forecast model's discrete grid representation of the atmosphere. These forecast model biases are often observed through the presence of non-zero time-mean forecast errors. One strategy for correcting forecast model errors is through the estimation of their composite effect on forecasts. This view has the benefit of not requiring internal modification of the forecast model. With a suitable model of forecast bias, the cumulative corrections to forecasts needed to account for model error may be estimated through data assimilation. Chapter 4 adapts a method to account for forecast model bias, previously tested on toy models and an atmospheric general circulation model (Baek et al., 2006, 2009), for data assimilation with the composite state method presented in Chapter 3. The work presented in Chapter 4 provides a method that can be used

to account for the pervasive and unavoidable effects of forecast model bias, and will be crucially needed if the composite state method described in Chapter 3 is to be implemented with atmospheric general circulation models.

Chapter 2: The Climatologically Augmented Local Ensemble Transform Kalman Filter (caLETKF)

2.1 Introduction

Data assimilation aims to optimally combine a best-guess forecast of a system state with observations of the true system state. This best-guess is known as the background state, and is typically made with a computer model. Modern data assimilation includes variational and ensemble methods, which update the background state vector using either static (for variational methods) or flow-dependent (for ensemble methods) background-error covariance models. Due to the size of atmospheric models, static covariances between atmospheric variables are typically approximated as spatially homogeneous and isotropic (Wang et al., 2007). Static covariance matrices have high rank, and can effectively encode important relationships and balance constraints. In contrast, flow-dependent covariance matrices, such as those approximated using ensemble methods, can distinguish physically realized space- and time-dependent relationships that may not be accounted for by static covariance matrices. However, such ensemble-derived covariance estimates are typically restricted to much lower rank than the covariance estimates used in variational

methods. The empirical technique of localization (Houtekamer and Mitchell, 1998, 2001; Hamill et al., 2001; Ott et al., 2004) is the most often used method for compensating for the rank deficiency of ensemble-derived covariance estimates (see Section 2 for further discussion).

The aim of this chapter is to propose a method for improving the performance of ensemble methods by effectively increasing the rank of their estimates of background-error covariance relationships. The way in which this is accomplished is through the inclusion of information from a static, climatologically-derived background-error covariance estimate, such as that which is often used in variational methods. One of the simplest and most common ways to combine flow-dependent and static covariance estimates is through a linear combination of their respective covariance matrices, which is essential to the workings of hybrid data assimilation methods (Hamill and Snyder, 2000; Lorenc, 2003). In contrast, here we consider remaining entirely within the ensemble data assimilation framework, and achieve higher rank by using climatological information to construct additional ensemble members. This approach achieves a higher-rank estimate of the background-error covariance matrix without simultaneously increasing the number of forecasted ensemble members.

Referring to the ensemble members corresponding to computed forecasts as the "dynamic" ensemble, at analysis time we augment the ensemble by adding supplementary static "climatological" ensemble members. These climatological ensemble members are formed by adding constant-in-time perturbations to the background dynamic ensemble mean at each analysis time. In our method, these climatologi-

cal perturbations are chosen to be approximately parallel to the directions of the leading eigenvectors of a static, climatological background-error covariance matrix, thus potentially enabling the analysis to allow for additional error directions that may not be well-represented by the dynamic ensemble. After these new ensemble members have been created and the ensemble has been enlarged, assimilation is performed on the collection of both dynamic and climatological ensemble members. The analysis ensemble members which correspond to the updated background (dynamic) ensemble members are then forecasted to the next analysis time, and the cycle is repeated. The accuracy of ensemble data assimilation methods can strongly depend on the number of ensemble members used, and we believe that our method will obtain benefits from increased ensemble size, without correspondingly increasing the number of forecasts carried out.

A related but fundamentally different technique to enhance ensemble perturbations is additive covariance inflation (Hamill and Whitaker, 2005; Wang et al., 2009; Whitaker and Hamill, 2012). This approach adds perturbations randomly sampled from a climatological error distribution to each ensemble member during each analysis cycle. A key difference between this approach and the caLETKF is that the caLETKF increases the size of the ensemble, and hence the rank of the background covariance.

The rest of this chapter is organized as follows. Section 2 provides background on ensemble data assimilation methods, while Section 3 formulates our proposed method. Section 4 describes numerical experiments testing the effectiveness of our approach, and Section 5 contains a review and discussion of the results. Section 6

contains our conclusions and additional discussion.

2.2 Background : The Ensemble Kalman Filter

Ensemble data assimilation methods are newer than variational methods, and show great potential for many geophysical applications. Specifically, we focus here on the ensemble Kalman filter method (Evensen, 1994; Burgers et al., 1998). These filters estimate the time-evolving background-error covariance matrix, $\mathbf{P^b}$, as the sample covariance of an ensemble of k_d model forecasts. Each of these forecasts can be done independently of the others, allowing for a naturally parallel computational method for forecasting background-error covariances. However, due to computational limitations, the ensemble size k_d is typically much smaller than the size N of the model state vector, $k_d << N$. Thus, despite gaining flow-dependence, the estimate of $\mathbf{P^b}$ is severely rank deficient. Empirical techniques such as localization are one way to remedy this rank deficiency. Localization works by suppressing correlations between model variables beyond some spatial distance determined by an empirically tuned localization radius. Though localization can substantially increase the effective rank of background covariances, practical limitations on the ensemble size may still result in information being missed by the ensemble. Decreasing the localization radius in such a situation is not necessarily a solution to this problem, as too much localization can lead to deleterious effects, such as increased imbalances in the analyses (Greybush et al., 2011). On the other hand, larger localization radii may improve state estimates at the cost of requiring a larger ensemble.

There are several variants of the ensemble Kalman filter (e.g. Houtekamer and Mitchell (1998); Anderson (2001); Bishop et al. (2001); Wang et al. (2004); Whitaker and Hamill (2002); Ott et al. (2004)). A class of ensemble filters, known as square-root filters, work by finding transformations that, when applied to the background ensemble members, produce a new collection of ensemble members, whose mean and sample covariance obey the Kalman filter equations. Though we apply our method to one of these Kalman filter variants known as the Local Ensemble Transform Kalman Filter (LETKF) (Hunt et al., 2007), it can also be applied to other ensemble Kalman filter formulations.

The LETKF finds an ensemble of analysis state vectors whose mean and sample covariance match those given by the Kalman filter equations, in the case of a linear observation operator H. Ensemble Kalman filters represent the ensemble of model forecasts as a N-dimensional ensemble mean state vector and a $N \times k_d$ dimensional matrix of ensemble perturbations. The ith column of the matrix of ensemble perturbations contains the difference between ensemble member i and the ensemble mean. The LETKF finds the analysis ensemble mean, $\bar{\mathbf{x}}^{\mathbf{a}}$, and ensemble perturbations, $\mathbf{X}^{\mathbf{a}}$, by transforming the background ensemble mean, $\bar{\mathbf{x}}^{\mathbf{b}}$, and ensemble perturbations, $\mathbf{X}^{\mathbf{b}}$, via

$$\bar{\mathbf{x}}^{\mathbf{a}} = \bar{\mathbf{x}}^{\mathbf{b}} + \mathbf{X}^{\mathbf{b}}\mathbf{w}, \tag{2.1a}$$

$$\mathbf{X}^{\mathbf{a}} = \mathbf{X}^{\mathbf{b}}\mathbf{W}. \tag{2.1b}$$

In the LETKF, the transformations $\mathbf{w}$ and $\mathbf{W}$ are found locally at each model grid point and depend on the observations (and their error covariance matrix $\mathbf{R}$) contained within a local analysis region centered at each model grid point. The transformations in eq. (2.1a) and (2.1b) are applied at each grid point to the ensemble of state vectors at that grid point. The global analysis ensemble is then formed by combining the results of each local analysis.

2.3 A Climatologically Augmented Ensemble Kalman Filter

As mentioned above, we present here a method which incorporates climatologically derived error covariance information into a pure ensemble data assimilation framework. In ensemble Kalman filters, a collection of k_d model forecasts is used to estimate the most likely system state and its uncertainty. Each of these forecasts is integrated using the full non-linear dynamics from one analysis time to the next, and is represented by an N-dimensional state vector, which we denote as $\widetilde{\mathbf{x}}^i_{\mathbf{d}}$. The ensemble mean, given by

$$\overline{\mathbf{x}}_{\mathbf{d}} = \frac{1}{k_d}\sum_{i=0}^{k_d}\widetilde{\mathbf{x}}^i_{\mathbf{d}},$$

is interpreted as the best estimate of the system state. Its uncertainty is estimated by the dynamic ensemble sample covariance matrix,

$$\mathbf{P}^{\mathbf{b}}_{\mathbf{d}} = \frac{1}{k_d - 1}\sum_{i=0}^{k_d}(\widetilde{\mathbf{x}}^i_{\mathbf{d}} - \overline{\mathbf{x}}_{\mathbf{d}})(\widetilde{\mathbf{x}}^i_{\mathbf{d}} - \overline{\mathbf{x}}_{\mathbf{d}})^T = \frac{1}{k_d - 1}\mathbf{X}_{\mathbf{d}}\mathbf{X}_{\mathbf{d}}{}^T,$$

where $\mathbf{X_d}$ is the $N \times k_d$ dimensional matrix of dynamic ensemble perturbations, $\mathbf{x_d^i} = \widetilde{\mathbf{x}}_\mathbf{d}^i - \overline{\mathbf{x}}_\mathbf{d}$, and the superscript T denotes the transpose operation. We incorporate climatological covariance information into this framework by increasing the size of the ensemble at analysis time, by adding to the dynamic ensemble a collection of k_c "climatological" ensemble members, $\widetilde{\mathbf{x}}_\mathbf{c}^j$, for $k_d + 1 \leq j \leq k_d + k_c$. In practice, the size k_c of this climatological ensemble is determined by a number of factors including convenience, empirical tests of forecast effectiveness and the user's computational resources. Each of these N-dimensional climatological ensemble members are created by adding a climatologically-derived perturbation $\mathbf{x_c^j}$ to the dynamic ensemble mean:

$$\widetilde{\mathbf{x}}_\mathbf{c}^j = \overline{\mathbf{x}}_\mathbf{d} + \mathbf{x_c^j}.$$

Thus, the ensemble on which the analysis is performed may be represented as a mean $\bar{\mathbf{x}}_\mathbf{d}$ and a $N \times (k_d + k_c)$ dimensional perturbation matrix $\mathbf{X} = [\mathbf{X_d}\ \mathbf{X_c}]$, where the columns of $\mathbf{X_d}$ are the dynamic ensemble perturbations, and the columns of $\mathbf{X_c}$ are the climatological ensemble perturbations.

In order that the mean of the combined ensemble be the same as the mean of the dynamical ensemble, we require that the mean of the columns of $\mathbf{X_c}$ be 0 (note that the mean of the columns of $\mathbf{X_d}$ is 0 by definition). We view the covariance $\mathbf{P_c^b} = \frac{1}{k_c}\mathbf{X_c}\mathbf{X_c}^T$ of the climatological ensemble as an approximation to a multiple of the true climatological background-error covariance matrix $\mathbf{B}$. In practice, we think that the magnitude of $\mathbf{X_c}$ should be tuned, rather than prescribed a priori. In

our formalism, we use the population covariance formula (with a factor of $1/k_c$) to simplify the following equations. The sample covariance of the combined ensemble, $\mathbf{P^b} = \frac{1}{k_d+k_c-1}\mathbf{X}\mathbf{X}^T$, may then be written as

$$\mathbf{P^b} = \frac{1}{k_d+k_c-1}[(k_d-1)\mathbf{P_d^b} + k_c\mathbf{P_c^b}] = \frac{k_d-1}{k_d+k_c-1}\mathbf{P_d^b} + \frac{k_c}{k_d+k_c-1}\mathbf{P_c^b}. \quad (2.2)$$

Thus, as with many hybrid methods (Hamill and Snyder, 2000; Lorenc, 2003), the implicit background covariance $\mathbf{P^b}$ is a linear combination of a dynamical (flow-dependent) covariance $\mathbf{P_d^b}$ and a climatological covariance $\mathbf{P_c^b}$, with coefficients whose sum is 1.

In most applications, the true climatological background-error covariance matrix $\mathbf{B}$ is not known a priori, and is estimated using various physical arguments and statistical techniques (we denote this estimate by $\mathbf{B}_{est}$). These techniques seek to determine appropriate correlation structures, but the overall covariance magnitude is often empirically tuned by a constant multiplicative factor. The climatological perturbations $\mathbf{x}_j$ in our method are derived from the orthogonal eigenvectors of the background-error covariance estimate $\mathbf{B}_{est}$. Specifically, they are derived from k_c columns of the matrix $\mathbf{A} = \mathbf{V}\mathbf{D^{1/2}}$, where $\mathbf{V}$ is a $N \times N$ dimensional matrix whose columns are the orthonormal eigenvectors of $\mathbf{B}_{est}$, and $\mathbf{D}$ is a $N \times N$ dimensional diagonal matrix of the eigenvalues of $\mathbf{B}_{est}$, so that $\mathbf{B}_{est} = \mathbf{A}\mathbf{A}^T$. The collection of k_c scaled orthonormal eigenvectors, where $k_c \ll N$, correspond to the k_c largest eigenvalues of $\mathbf{B}_{est}$, and we interpret these climatological vectors as representing the k_c directions of greatest climatological error variability. Once the collection of these

k_c columns of $\mathbf{A}$ has been chosen, the mean of this collection is subtracted from each eigenvector, since the columns of $\mathbf{A}$ do not sum to zero. The resulting vectors are scaled by a tuning factor α to form the k_c climatological ensemble "perturbations," which we store in the columns of the $N \times k_c$ dimensional matrix $\mathbf{X_c}$.

In our experiments, we use as an analysis algorithm the Local Ensemble Transform Kalman Filter (LETKF) (Ott et al., 2004; Hunt et al., 2007). We perform the analysis on the collection of $k_d + k_c$ dynamic and climatological ensemble members, and upon completion of the analysis procedure, the ensemble mean is updated according to eq. (2.1a). The analysis also produces a collection of $k_d + k_c$ analysis ensemble perturbations. The 'dynamic' analysis ensemble perturbations are created from the first k_d analysis ensemble perturbations, which for the LETKF, are the ones that are closest to the dynamic background ensemble perturbations (Ott et al. (2004), in particular, see appendix A). The mean of these k_d perturbations is subtracted from each perturbation to yield the dynamic analysis ensemble perturbations. The dynamic analysis ensemble members are then calculated by adding the analysis ensemble mean to each of these dynamic analysis perturbations. This k_d member analysis ensemble is then forecasted forward in time to the next analysis time. Especially for other Ensemble Kalman Filters, it may be fruitful to consider other methods of selecting k_d analysis perturbations from a $(k_d + k_c)$-member analysis ensemble. Our choice of the first k_d perturbations is appropriate for filters like the LETKF and perturbed observations EnKF (Burgers et al., 1998; Houtekamer and Mitchell, 1998), for which there is a natural correspondence between pairs of background and analysis ensemble members.

We think of the new climatological members as representing potential error directions that might not be captured by the dynamically forecasted ensemble members. In this sense, the data assimilation algorithm can search in a higher dimensional space for corrections to the dynamic ensemble. Though the addition of ensemble members increases the cost of the analysis computation, we show in Section 5 that a major benefit of this climatologically augmented Local Ensemble Transform Kalman Filter (caLETKF) is that greater analysis accuracy can be achieved with fewer forecasts. We note that the static, climatological perturbations only need to be calculated once, so that the increased computational cost of our method comes purely from carrying out the analysis in a higher dimensional space.

2.4 Setup of our Numerical Experiments

We apply our climatological ensemble augmentation method in a series of experiments that utilize a one-dimensional, chaotic model which we call Lorenz Model II (Lorenz, 2005). Lorenz Model II represents the flow of an "atmospheric-like" quantity Z around a circle of constant "latitude." Specifically, Z is defined on a lattice of N grid points with periodic boundary conditions. The value of Z at a given grid point n is denoted by Z^n, and its temporal behavior is given by

$$\frac{dZ^n}{dt} = [Z, Z]^{K,n} - Z^n + F. \tag{2.3}$$

The terms on the right hand side of eq. (2.3) are meant to roughly model quantities analogous to spatially averaged non-linear advection, linear dissipation, and constant

forcing, respectively. For our experiments, we set the parameter F to have a value of 15, and take $N = 240$. The first term on the right-hand side of eq. (2.3) is given by

$$[Z, Z]^{K,n} = \frac{1}{K^2} \sum_{j=-J}^{J}{}' \sum_{l=-J}^{J}{}' (Z^{n-K+j-l} Z^{n+K+j} - Z^{n-2K-l} Z^{n-K-j}), \qquad (2.4)$$

where K is an integer valued parameter, and $J = K/2$ if K is even and $J = (K+1)/2$ if K is odd. If K is even, the primed summation notation in eq. (2.4) denotes an ordinary sum with the first and last terms in the sum each multiplied by 1/2. For our experiments, $K = 8$. The result of the spatial averaging present in eq. (2.4) is that the Model II states are characterized by spatially smooth waves.

We construct an estimate, $\mathbf{B}_{est}$, of the background-error covariance matrix through the NMC Method of Parrish and Derber (1992). This method approximates the background-error covariance matrix using a large set (for our experiments $\sim$ 50000) of differences between 6- and 24- hour forecasts that verify at the same time. These forecasts were started from initial conditions generated using a 40-member LETKF. To account for magnitude differences between the estimate of the background-error covariance matrix found using the NMC method and the true climatological background-error covariance matrix, the estimate of the background-error covariance matrix estimated using the NMC method is typically multiplied by a constant scalar factor that is tuned for the particular application (e.g., 3D-Var). To accomplish this tuning within the caLETKF, the eigenvectors of $\mathbf{B}_{est}$, adjusted as described in Section 3, are multiplied by a scalar factor α to form the climatological

perturbations $\mathbf{X_c}$. For the cases explored here, we found that the analysis error was insensitive to the precise value of α, and that the value $\alpha = \sqrt{k_c}$ produced a near-minimum error across a range of values of k_c and k_d. Thus we used $\alpha = \sqrt{k_c}$ for all of the results reported here. Note that this choice of α maintains the contribution of each eigenvector of $\mathbf{B}_{est}$ to the covariance $\mathbf{P_c^b} = \frac{1}{k_c}\mathbf{X_c}\mathbf{X_c}^T$ as k_c increases, but that the relative weight of $\mathbf{P_d^b}$ and $\mathbf{P_c^b}$ in eq. (2.2) still varies with k_c. More specifically, scaling by this choice of α makes the trace of the climatological covariance $\mathbf{P_c^b}$ equal to a sum of the largest k_c eigenvalues of the estimated true error covariance matrix $\mathbf{B}$. Consequently, in the limit of very large climatological ensemble size the error variance represented by the climatological perturbations reproduces the true error variance.

A key step in our proposed method is the generation of the leading eigenvectors of the static, climatological background-error covariance matrix. For the experiments described here, small system size allowed all of the eigenvectors of the climatological background-error covariance matrix to be easily calculated. However, for large, operational meteorological applications, explicit diagonalization of the complete $\mathbf{B}$ matrix for systems of order $N \sim 10^{8-10}$ is problematic, since $\mathbf{B}$ is too large to be stored in a computer. On the other hand, the approximate effect of multiplying vectors by such matrices is available and is, for example, widely used for preconditioning in variational data assimilation. Furthermore, we only require the leading eigenvectors, not all of them, and methods such as the Lanczos algorithm are capable of calculating the leading $m \ll N$ eigenvectors numerically while requiring only m evaluations of the action of $\mathbf{B}$ on a vector (Golub and Van Loan, 1996).

Exploiting special structure of the background-error covariance matrix can make these numerical procedures even more efficient and tractable. We again emphasize that these operations must only be completed once and can be done offline, as they are the same at each analysis cycle.

We employ a 'perfect model' set-up in our numerical experiments, so that the truth against which our state estimate is compared is generated from a free model integration that uses the same parameter values and dynamics as are used to forecast the ensemble. Observations are generated by adding Gaussian white noise, with mean 0 and standard deviation 1, to the truth state vector at the observation times and locations. We chose a spatially homogeneous, static observation network. Specifically, we assimilate 12 observations at each analysis time, with one observation every 20 model grid points. We perform assimilations every 0.05 model time units, analogous to approximately every 6 hours for the atmosphere (Lorenz, 2005). The initial ensembles used in our numerical experiments are found by randomly sampling widely time-separated states from a long run of the forecast model states.

As a benchmark for comparison, we use results produced when analyses are performed using the standard LETKF algorithm (i.e. without a climatological supplement to the ensemble). Benchmark runs have the same ensemble size, assimilate the same set of observations, and compare against the same truth run as experiments with the climatologically augmented algorithm. In both sets of experiments, the localization radius used by the analysis algorithm is constant at 20 model grid points, with no tapering of the observation influence (Hunt et al., 2007). Additionally, we utilize 3% multiplicative covariance inflation (Anderson and Anderson, 1999) in all

experiments involving the standard LETKF, and 2.75% multiplicative covariance inflation for experiments using the caLETKF. These inflation factors were tuned to minimize analysis root mean square error over the interval of $0 - 7\%$ inflation.

To compare the performance of the caLETKF and the LETKF, we measure the root-mean square error (RMSE) between the truth and forecast ensemble mean for several forecast lead times f. We denote the difference at location n between the ensemble mean of the f-hour forecast ensemble verifying at time r and the truth at the same time as $\epsilon(r, n, f)$. The RMSE of the f-hour forecast ensemble mean verifying at analysis time r is then expressed as

$$RMSE(r, f) = \left\{ \sum_{n=1}^{N} (\epsilon(r, n, f))^2 / N \right\}^{1/2}. \tag{2.5}$$

The temporally averaged root mean square error of the forecast ensemble mean is then found by averaging eq. (2.5) over all c times during which statistics are calculated, to yield the average RMS error for a given forecast lead time f,

$$\langle RMSE(f) \rangle = \sum_{r=1}^{c} \left\{ \sum_{n=1}^{N} (\epsilon(r, n, f))^2 / N \right\}^{1/2} / c, \tag{2.6}$$

where $r = 1$ is chosen to correspond to an analysis time occurring after a sufficiently long spin-up time. We compare the LETKF and caLETKF through the analysis accuracy ($f = 0$ in eq. (2.5) and eq. (2.6)) and for ensemble forecasts of various lead-times ($0 < f \leq 72$ in eq. (2.6)). Though the present work focuses on the usage of the RMSE as a metric of analysis and forecast performance, we recognize that other diagnostics (e.g. the spread/skill relationship) will be useful and necessary for

further explorations of the properties of the caLETKF.

2.5 Experimental Results

The effectiveness of our method was first examined through time series of the analysis root-mean square error. The RMS error of the analysis ensemble mean was recorded at each analysis time, the results of which are shown in Fig. 2.1, for both the caLETKF and standard LETKF. Here the dynamic ensemble size of both the standard LETKF and the caLETKF are both equal to 10, while the caLETKF supplements the dynamic ensemble with 10 additional static, climatological ensemble members at each analysis time. We note that the LETKF case with $k_d = 10$ dynamic ensemble members does not converge. Once our method (red curve) has converged, it produces analysis errors that are significantly smaller than than those produced using the standard LETKF with $k_d = 10$ dynamic ensemble members (blue curve). The standard LETKF, albeit with a substantially larger dynamic ensemble size ($k_d = 20$, black curve), can achieve analysis accuracies similar to those of our method shown in Fig. 2.1. Comparing the results for the standard LETKF with $k_d = 10$ dynamic ensemble members with results for the caLETKF for $k_d = 10$ dynamic and $k_c = 10$ climatological ensemble members, one can see that the inclusion of the additional climatological ensemble members helps to stabilize the filter (i.e. fluctuations in the errors are substantially reduced). The behavior shown in Fig. 2.1 suggests that the climatological members account for realistic errors in directions not captured by the dynamic ensemble members, and are providing value

to the overall assimilation without the expense of being forecasted themselves.

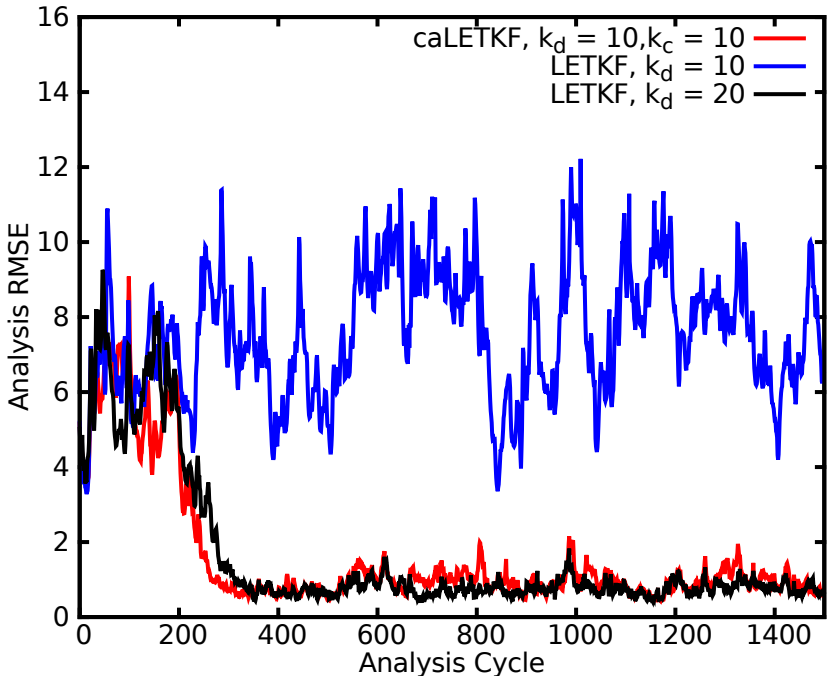

Figure 2.1: An example in which a standard LETKF analysis with insufficient ensemble size (10 dynamic ensemble members, blue curve) is stabilized by augmenting the ensemble with 10 additional climatological ensemble members (red curve). The RMS error of the analysis ensemble mean (eq. (2.5)) is plotted at each analysis cycle over a 1500 analysis cycle period. Both experiments assimilate the same observations, on the same observation network. For comparison, results from an experiment where the standard LETKF has $k_d = 20$ dynamic ensemble members (black curve) are also included.

To better quantify how much of an advantage is gained by our climatological augmentation of the ensemble at analysis time, we also compare how RMS analysis error changes with increasing dynamic ensemble size, for both the augmented and standard analysis methods. At each dynamic ensemble size, the RMS analysis error of the ensemble mean is averaged over 50000 analysis cycles, after an initial 1000 spin-up cycles. For each dynamic ensemble size listed on the x-axis of Fig. 2.2, the caLETKF was computed with $k_c = 10$ climatological ensemble members. As

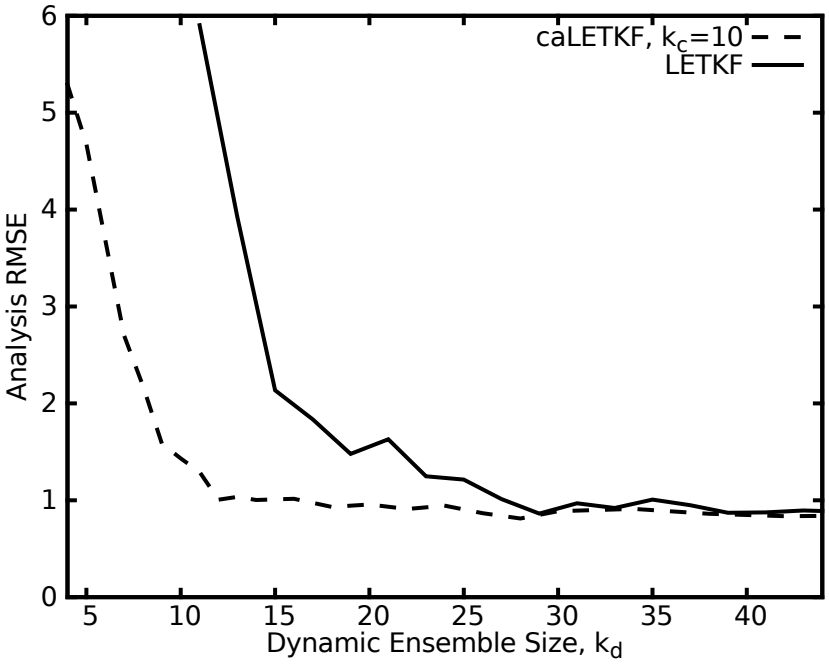

Figure 2.2: A comparison of the RMS error of analysis ensemble means, eq. (2.6), between both the standard LETKF (solid curve) and the caLETKF (dotted curve). After an initial spin-up of 1000 analysis cycles, RMS error is averaged over 50000 analysis cycles, and plotted as a function of dynamic ensemble size. For the experiments shown here, the climatologically augmented method uses $k_c = 10$ climatological ensemble members. Below $k_d = 4$ dynamic ensemble members, we find that the caLETKF is susceptible to filter divergence, while the standard LETKF is susceptible to filter divergence below $k_d = 10$ dynamic ensemble members.

Fig. 2.2 shows, the advantage of the caLETKF is shown in its convergence, at smaller dynamic ensemble size, to the error level reached by the traditional LETKF at larger dynamic ensemble size. Specifically, our method needs approximately 13 dynamic ensemble members to achieve the same error that the LETKF achieves with approximately 28 dynamic ensemble members. Fig. 2.2 further suggests that, if limited by computational resources for calculating ensemble member forecasts, it can be beneficial to consider using climatological ensemble members during the analysis, rather than striving to just increase dynamic ensemble size.

In the experiments detailed above, comparisons between the caLETKF and LETKF analyses are made when the caLETKF uses $k_c = 10$ climatological ensemble members. Fig. 2.3 explores how many climatological ensemble members the caLETKF needs by changing the number of climatological ensemble members, keeping the size of the dynamic ensemble constant at $k_d = 15$. As in Fig. 2.2, the plots are of time-averaged RMS error of the ensemble mean versus climatological ensemble size, and each data point shown here is averaged over 50000 analysis cycles. Fig. 2.3 shows that as few as about 8 climatological ensemble members can be used without a significant loss of analysis accuracy, implying that the caLETKF curve of Fig. 2.2 would look very similar for any $k_c \geq 8$. This suggests that much of the information missed by the dynamic ensemble is found in a relatively small number of state space directions.

The experiments just described measure analysis accuracy of the caLETKF while keeping forecasting costs (i.e., k_d) constant. We next present analysis accuracy results from an experiment which keeps analysis cost constant, by varying the size of the climatological ensemble while keeping constant the sum of the number of dynamic and climatological ensemble members, $k_d + k_c$. Specifically, the total ensemble size on which the analysis is performed is kept constant at $k_d + k_c = 30$ ensemble members. The value $k_d + k_c = 30$ is chosen because, as shown in Fig. 2.2, if a dynamic ensemble and no contributing static members ($k_c = 0$) are used, the LETKF analysis accuracy (solid curve in Fig. 2.2) quickly degrades when the number of dynamic ensemble members falls below 30, $k_d < 30$. Fig. 2.4 shows the time-averaged RMSE of the analysis ensemble mean, averaged over 50000 analysis

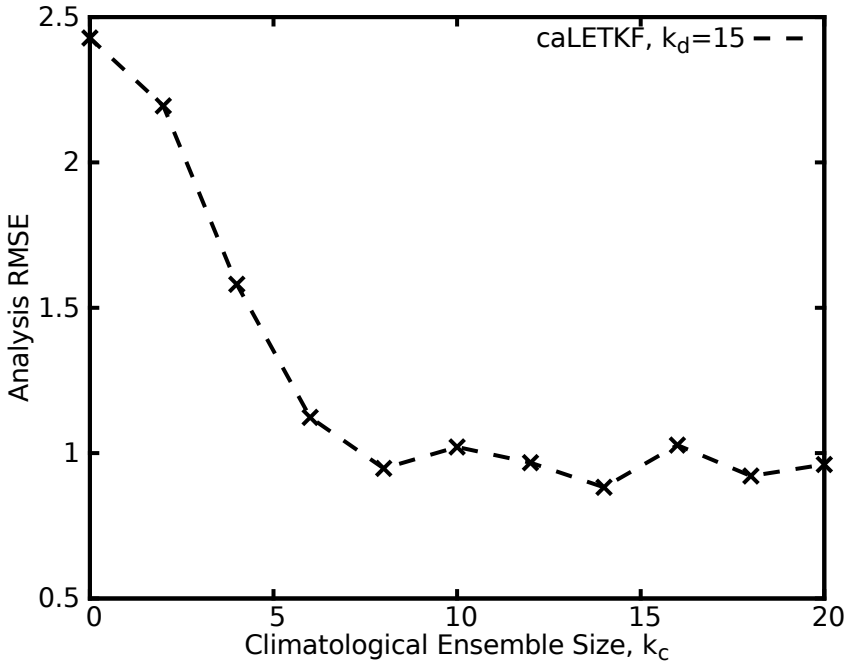

Figure 2.3: Here we compare RMS error of the analysis ensemble mean for the caLETKF, eq. (2.6), as a function of climatological ensemble size k_c. Fifteen dynamic ensemble members ($k_d = 15$) are used, and each trial is averaged over 50000 analysis cycles, discarding the first 1000 cycles as spinup.

cycles as a function of climatological ensemble size. We see that the inclusion of climatological ensemble members allows the caLETKF to achieve the same accuracy with 12 dynamic ensemble members (and $k_c = 18$ climatological members) as that achieved with the standard LETKF with 30 dynamic ensemble members (and $k_c = 0$ climatological ensemble members). For reference, the horizontal line represents the analysis accuracy of the standard LETKF when $k_d = 30$ dynamic ensemble members are used, which is the smallest dynamic ensemble size at which the LETKF converges. Here, the caLETKF is advantageous over the LETKF, as it achieves comparable analysis errors with many fewer (as low as $k_d = 12$) dynamic ensemble members.

To measure ensemble forecast accuracy, we average the accuracy of the forecast ensemble mean over a series of 50000 ensemble forecasts. Forecast accuracy information is stored for a maximum lead-time of 3 model days. The forecast accuracy, as measured by RMSE (eq. (2.6)) is plotted in Fig. 2.5 as a function of forecast lead-time. For this experiment, the caLETKF uses $k_d = 20$ dynamic ensemble members and $k_c = 10$ climatological ensemble members. Forecast results from the caLETKF are compared against results from the LETKF with $k_d = 20$ and $k_d = 30$ dynamic ensemble members. Forecast results from the LETKF with $k_d = 30$ dynamic ensemble members outperform those of the LETKF with $k_d = 20$ dynamic ensemble members, and are on average comparable to the caLETKF with $k_d = 20$ and $k_c = 10$. Fig. 2.5 shows that gains made from more accurate caLETKF analyses persist, and lead to more accurate forecasts. In fact, the caLETKF corresponds to gains in forecast lead-time of 24 to 17 hours: the 1-day forecasts initialized with the 20 dynamic

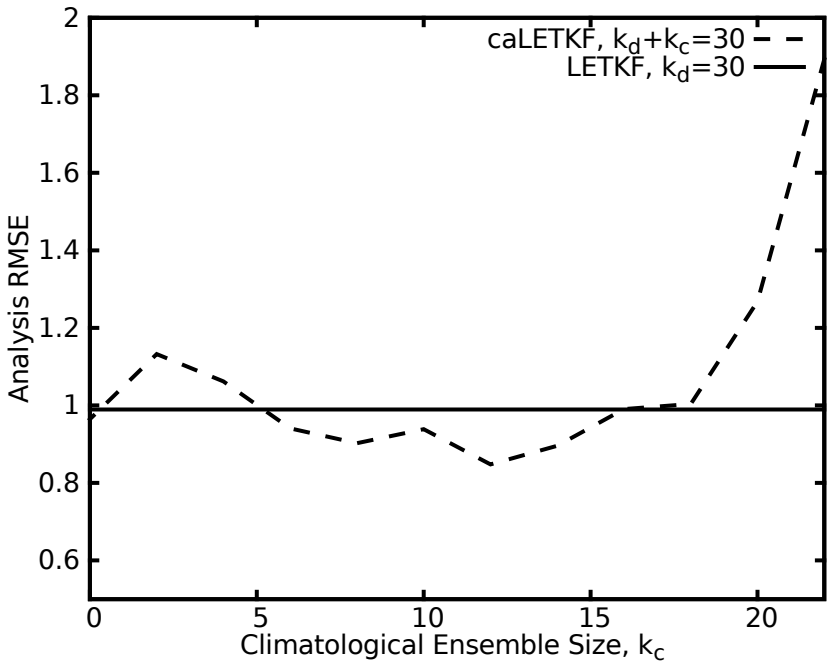

Figure 2.4: Analysis accuracy of the caLETKF at constant analysis cost. Here, the sum of the dynamic and climatological ensemble sizes is kept constant at k_d+$k_c = 30$, and the climatological ensemble size is plotted versus analysis RMS error, eq. (2.6), averaged over 50000 analysis cycles, after discarding 1000 initial cycles. For small dynamic ensemble sizes, $k_d < 8$ and $k_c > 22$, the caLETKF was susceptible to filter divergence.

member caLETKF analysis ensemble are as accurate as the 20 dynamic member LETKF analysis ensemble. Though these gains diminish slightly with time, at constant dynamic ensemble size ($k_d = 20$), 48-hour caLETKF-initialized forecasts are as accurate as 30-hour LETKF-initialized forecasts, and 72-hour caLETKF-initialized forecasts are as accurate as 55-hour LETKF-initialized forecasts.

2.6 Summary and Conclusions

The techniques of data assimilation can broadly be categorized as either variational or ensemble methods. While variational methods have been around for

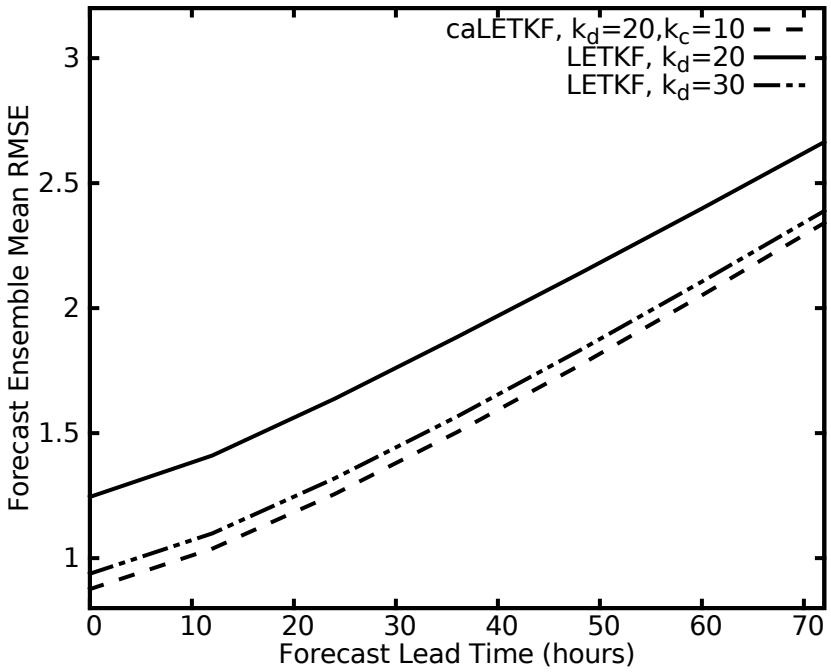

Figure 2.5: Ensemble forecast accuracy as a function of lead time f, for forecasts initialized from caLETKF and LETKF analysis ensembles. Ensembles were forecasted forward in time, and the mean of the forecast ensemble was compared against truth. The resulting errors were averaged over a sample of 50000 forecasts at each lead time, using eq. (2.6). Forecast results initialized from caLETKF analysis ensembles with $k_d = 20$ dynamic and $k_c = 10$ climatological ensemble members are shown as a dashed black curve. For comparison, results of forecasts initialized from LETKF analysis ensembles with $k_d = 20$ and $k_d = 30$ dynamic ensemble members are shown as solid and dot-dashed curves, respectively. The small difference (0.06 in forecast RMSE) between the LETKF $k_d = 30$ result and the caLETKF result is within the level of statistical fluctuations seen in our experimental system (for example, see the variation of the solid and dashed curves in Fig. 2.2 for $k_d \geq 30$).

longer than ensemble methods, and are in widespread use in both operational and research settings, ensemble methods have advantages of their own, such as flow-dependent covariance estimates and natively computationally parallel forecasting. Recently, there has been a push to combine these methods to take advantage of their mutual advantages. The result has been the 'hybrid' methods, which combine the flow-dependent covariance estimates of ensemble methods with the climatological covariance estimates used by variational methods. Inspired by the success of these methods, we present here an ensemble method that aims to take advantage of climatological covariance information while staying in a purely ensemble framework.

The method we present here computes the analysis in a higher-dimensional space than standard ensemble Kalman filters, by considering an augmented ensemble during the analysis. At the start of each analysis, the background dynamic ensemble mean is computed, along with the background dynamic ensemble perturbations. Additional ensemble members are created by adding to the background dynamic ensemble mean perturbations derived from a climatological background-error covariance matrix. We generate these climatological perturbations from the eigenvectors that correspond to the largest eigenvalues of the climatologically estimated background-error covariance matrix. These chosen eigenvectors correspond to the directions in model space that climatologically account for the most forecast error variability. Preliminary results (not shown) indicate that generating climatological perturbations from the eigenvectors of a static covariance estimate $\mathbf{B}_{est}$ is advantageous over the simpler approach of using perturbations that are samples taken at each analysis cycle from the archive of forecast errors used to generate $\mathbf{B}_{est}$

and scaled by a factor independent of k_c. For the experimental results reported here, we estimate the background-error covariance matrix through the NMC method of Parrish and Derber (1992). We anticipate that other methods of generating the climatological background-error covariance matrix might be called for in other settings.

We conducted a series of numerical experiments using a one-dimensional chaotic model of Lorenz (Lorenz, 2005). In these experiments, we compared the caLETKF against the standard LETKF analysis algorithm. Initial experiments compared time-series analysis errors generated by both methods at a constant dynamic ensemble size (Fig. 2.1). These experiments showed that the caLETKF with $k_c = 10$ added climatological ensemble members can have a clear advantage over the standard LETKF.

Our next series of experiments investigated how well the new method performs as a function of dynamic ensemble size when the climatological ensemble size is fixed at $k_c = 10$ (Fig. 2.2). This experiment showed that the caLETKF converged and achieved a consistent level of analysis accuracy at approximately 13 dynamic ensemble members, while the LETKF needed approximately 28 dynamic ensemble members to perform as accurately. In this experiment, we observed that both the caLETKF and LETKF produce similar results at large enough dynamic ensemble sizes. In more complex, highly flow-dependent systems, we imagine that a large number of climatological ensemble members could reduce the relative weight given to the dynamical portion of the covariance enough to potentially degrade performance. In these scenarios, one might introduce an additional tunable scaling factor that,

in conjunction with α, would independently control the weights of dynamic and climatological perturbations in the estimation of background covariances.

A third experiment explored the effect on analysis accuracy of varying the climatological ensemble size (Fig. 2.3). This experiment found that fewer than the 10 climatological ensemble members used in the previous experiments could be used during the analysis with negligible loss of analysis accuracy. In addition to comparing our method to the standard LETKF at constant forecasting cost, as just described, the accuracy of the caLETKF was also tested against the standard LETKF at constant analysis cost. Here, the total size of the ensemble is kept constant (i.e., the sum of dynamic and climatological ensemble sizes), while the proportion of the ensemble members that are climatological ensemble members is varied (Fig. 2.4). Results from these experiments indicate that analysis accuracy can be maintained by replacing a significant number of dynamic ensemble members with climatological ensemble members. In applications where the same result applies, this could substantially reduce forecast costs.

Our last series of experiments investigated the accuracy of 1-, 2-, and 3- day ensemble forecasts (Fig. 2.5). This experiment found that the caLETKF analysis gains discussed earlier were retained during the forecasts, as ensemble forecasts initialized from caLETKF analyses were more accurate than forecasts initialized from LETKF analyses. We find the results of these numerical experiments with the caLETKF to be encouraging, and to suggest that it be tested on larger and more realistic atmospheric models. Numerical experiments with larger, more complex models will necessitate other, more advanced diagnostics (e.g. the spread-error relationship) to

measure and quantify the benefits of the caLETKF. The caLETKF could potentially be very useful in these settings, as fewer dynamic ensemble members can be used without loss of analysis accuracy.

Chapter 3: The Composite State Method

3.1 Introduction

Geophysical fluid dynamical forecasts (e.g., for atmospheric or oceanic states) are typically created by integrating a numerical model forward in time from suitable initial conditions. Limited-Area Models (LAMs), which only cover a restricted geographic area, allow high-resolution forecasts for small, sub-global regions of interest to be made when limited resources prevent running a high-resolution global forecast model or assimilating high-resolution global data. While higher spatial resolution does not guarantee that a model will behave more realistically, it is generally helpful in this endeavor. (For simplicity, our numerical experiments are designed to ensure that increased model resolution corresponds to increased model realism.)

LAMs are found in a variety of settings, ranging from research to operational situations. Typically, operational weather centers run LAMs that are defined over the geographic region(s) of interest to that center. Some centers, such as those of the U.S. National Weather Service and the U.S. Navy, run their LAM independently on multiple domains that may or may not overlap, to produce high-resolution regional forecasts. The U.S. Navy, in particular, routinely runs their limited-area COAMPS® model on more than 70 different limited-area domains (Pielke Sr, 2013). Some of

these domains are overlapping, and their union covers more than 20% of the globe. But, in the data assimilation phase, the analysis state of one LAM is normally not directly affected by the state of any of the others, and the analysis state of the global model providing LAM boundary conditions is not affected by the LAM analysis states. The present chapter investigates the possibility of improving forecast performance by allowing interactions between global and LAM states at analysis times in situations where multiple LAMs are employed.

In particular, data assimilation is considered in an ensemble forecasting context. Ensemble data assimilation uses a collection of model forecasts to estimate the probability distribution of the state of the system. We investigate an approach in which each individual LAM is used to integrate its own ensemble of model states. Integrating an ensemble of LAM model states requires an ensemble of boundary conditions that transmit information about the large scale flow features to each of the LAM ensemble members. An ensemble of lateral boundary conditions may be generated in a number of ways (Torn et al., 2006), which includes concurrently running a global forecast ensemble (Merkova et al., 2011; Holt et al., 2013).

Traditionally, data assimilation has been performed on the global and limited-area models separately. However, recently it has been shown that using the global analysis (Guidard and Fischer, 2008) or a short-range forecast based on the global analysis (Dahlgren and Gustafsson, 2012) as an additional constraint on the limited-area state estimate has a positive effect on the limited-area analysis, while allowing communication between the global and limited-area processes can improve both the global and the limited-area analyses (Yoon et al., 2012). This finding is similar

to the one that two-way grid nesting during the forecasting phase can help reduce discrepancies between global and limited-area model states (Harris and Durran, 2010). In situations with multiple LAMs, two-way coupling of the models during the analysis procedure can potentially improve the global analysis. More accurate global analyses would, in turn lead to improvements in the lateral boundary conditions applied to LAMs. Allowing for additional communication between LAMs during the analysis may help alleviate the lateral boundary condition errors that have typically plagued LAMs (Warner et al., 1997).

We present a framework for combining all global and limited-area model state information available at the analysis time, and for performing data assimilation on this composite state. This new approach aims to obtain ensemble mean analyses for all limited-area and global model ensembles that minimize the difference between the updated ensemble means and the truth (at the resolution each model simulates). We do not explicitly aim to find analyses which lie on their respective model attractors. However, treating the model as an approximation to reality, obtaining analyses close to reality will provide initial conditions close to model attractors. Recently, Klocke and Rodwell (2014) suggested that even the most advanced current data assimilation systems provide analyses that are perturbed off the model attractor in the direction of the true (filtered) state. The authors of that paper also showed that the mean short-term initial drift of the model forecast from the analysis state towards the model attractor provides useful information for the diagnosis and correction of forecast model errors.

The new approach presented here is based on the idea that there is only

one true state, and the objective of data assimilation should be to estimate this state as accurately as possible. At locations where a limited-area model is present, it is possible to model the true atmospheric state more accurately and at higher resolution than at locations where only the global model is defined. As a result, the new composite state method introduced in this chapter applies the LETKF to a single, variable resolution state vector. The local resolution of this state vector at a geographic location is determined by the highest resolution state information available at that location, whether from a LAM or the global model. Forecasts for this variable resolution state vector are obtained through linear combination of model forecasts that are defined on domains which overlap at points on the variable resolution composite state grid. Application of the LETKF results in an ensemble of variable resolution analysis state estimates, which are interpolated down to lower resolution models that may cover the same grid area. The result is that when using the composite state method, model forecasts are initialized from analysis states that are informed by all forecast models.

Yoon et al. (2012) used a *joint-state* approach, in which the components of the state vector on which the analysis is performed are the components of the regional and global state vectors. A drawback of this approach is that an *ad-hoc* term had to be added to the state estimation equations to moderate a tendency of the global and regional state estimates to diverge. A theoretically appealing aspect of the composite state approach is that it would deliver the minimum error variance state estimate, provided the correct forecast and observation error covariance matrices were specified and the effects of localization in the LETKF were negligible. In con-

trast, the presence of the ad-hoc term in Yoon et al. (2012)'s formulation precludes the possibility of it delivering the minimum error variance state estimate.

We test our method in a series of simulated observation experiments that utilizes the simple, one-dimensional, chaotic models introduced by Lorenz (2005), and we find that analyses and forecasts produced by a coarse resolution 'global' model and a collection of high-resolution 'LAMs' that cover the entire simulation domain can attain essentially the same accuracy as analyses and forecasts produced by a single high-resolution global model. If only part of the global domain is covered by overlapping LAMs, the analyses and forecasts are essentially as accurate as if the overlapping LAMs were replaced with a single LAM covering the same region, and except near the boundary of this region, the results can be nearly as accurate as using the high-resolution model globally. These results serve as motivation for further investigation, to see if real systems utilizing our method might realize similar forecast improvements. Importantly, unlike the Lorenz models, real-world atmospheric models include interaction between motions at a wide range of scales. Further studies will be necessary to explore the behavior of these scale interactions in model simulations that use the composite state method.

The rest of this chapter is organized as follows. Section 2 introduces the composite state technique. Section 3 describes our numerical experiments with the Lorenz (2005) models. We use a sparse observation network, so that the advantage of the high-resolution LAMs over the global model is primarily in greater forecast accuracy, rather than an ability to resolve the observations. In Section 4, we consider the case where the LAMs cover the global domain. In this case, the composite state

analysis is done on a global high-resolution grid, and any disadvantage compared to using a single global high-resolution model must be due to the decomposition of the global model states into LAM states for the forecast phase of the analysis cycle. We explore the effect of varying LAM region size and global domain size within this context. We find that there is essentially no degradation of the composite state estimate relative to using a single high-resolution model, unless the LAM region size is too small. In Section 5 we consider overlapping LAMs that cover only part of the global domain, so that the analysis is done at coarse resolution on the part of the globe that is not covered by LAMs. As in Section 4, the interface between neighboring LAMs is not a significant source of error; degradation compared with using a global high-resolution model occurs near the boundary with the coarse-resolution region. Section 6 presents our conclusions and further discussion.

Finally, we note that our presentation is in the context of one dimensional models defined on subsets of the same common grid. In practice, one would be interested in three dimensional atmospheric models and a collection of Limited Area Models, each with its own grid. These issues are not directly addressed here, but would have to be dealt with if our method were to be operationally applied (see Sec. 6).

3.2 Data Assimilation and The Composite State Method

Data assimilation is a cyclic alternation between a short-term forecast and a procedure, called the 'analysis,' that seeks to estimate a system's state (and pos-

sibly its error statistics). At the beginning of each cycle, the analysis procedure is performed, combining available observational information with a forward forecast from the end of the previous cycle (called the 'background' estimate) to yield an updated estimate of the system state (the 'analysis state'). The estimates of the uncertainties in the observations and background state estimate are crucial to this step, as they determine the relative weighting of the observations and background in forming the analysis state estimate. Once the analysis procedure is completed, the analysis state, and possibly its uncertainty, can be forecast to the next analysis time, where the cycle is repeated.

Data assimilation based on Ensemble Kalman Filters (Evensen, 1994; Burgers et al., 1998), which will be the basis of our work, has attracted much recent interest because of its ability to form consistent time-dependent uncertainty estimates in the analysis. In an Ensemble Kalman Filter, a collection of system state estimates (the ensemble) is evolved in time and updated with observational information at the start of each analysis. The best guess of the system state is given by the ensemble mean, and the background error covariance matrix $\mathbf{P^b}$, a measure of the state estimate's uncertainty, is approximated by the sample covariance of the background ensemble. Here, we conduct our analysis on the composite state ensemble using the Local Ensemble Transform Kalman Filter (LETKF)(Hunt et al., 2007). The LETKF algorithm seeks an analysis ensemble with mean and sample covariance given by the Kalman Filter update equations. It is equivalent to a localized version of the ETKF of Bishop et al. (2001) with spherical simplex centering (Wang et al., 2004), and may be viewed as a computationally advantageous version of the LEKF method

of Ott et al. (2004). A local ETKF has been successfully implemented to generate ensemble perturbations for an operational global-regional model pair in the UK Met Office's MOGREPS system (Bowler and Mylne, 2009). The LETKF has also been tested, with positive results, using the regional models of the Italian and German weather services (Bonavita et al., 2010; Reich et al., 2011; Lange and Craig, 2014).

In Ensemble Kalman Filters, the background and analysis ensembles are typically expressed as an ensemble mean, and a collection of ensemble perturbations from this mean. The ensemble mean is denoted $\bar{\mathbf{x}}$, an N vector, with N being the dimension of the model state. The ensemble perturbations form an $N \times k$ matrix $\mathbf{X}$, where k is the number of ensemble members. The mth column of this matrix ($m = 1, 2, \ldots, k$) represents the perturbation $(\widetilde{\mathbf{x}}_m - \bar{\mathbf{x}})$ of the mth ensemble member, $\widetilde{\mathbf{x}}_m$, from the ensemble mean. In the LETKF, the analysis ensemble mean and ensemble perturbations are expressed in terms of the background ensemble mean and ensemble perturbations through a weight vector, $\mathbf{w}$, and a weight matrix, $\mathbf{W}$,

$$\begin{aligned} \bar{\mathbf{x}}^{\mathbf{a}} &= \bar{\mathbf{x}}^{\mathbf{b}} + \mathbf{X}^{\mathbf{b}}\mathbf{w}, \\ \mathbf{X}^{\mathbf{a}} &= \mathbf{X}^{\mathbf{b}}\mathbf{W}. \end{aligned} \tag{3.1}$$

Here $\mathbf{X}^{\mathbf{b}}$ and $\mathbf{X}^{\mathbf{a}}$ are the matrices of background and analysis ensemble perturbations, respectively. The background and analysis ensemble means are given by $\bar{\mathbf{x}}^{\mathbf{b}}$ and $\bar{\mathbf{x}}^{\mathbf{a}}$, respectively. The LETKF conducts its analysis in a local ensemble space. At each grid point n, the LETKF uses the observations within an empirically determined local analysis region to determine $\mathbf{w}$ and $\mathbf{W}$ for grid point n, and hence the analysis state value at that grid point.

Although our presentation and numerical examples assume use of the LETKF framework, our method does not rely on the details of the LETKF, and we expect it can be straightforwardly adapted to other versions of Ensemble Kalman Filters (see, e.g., Houtekamer and Mitchell, 1998; Anderson and Anderson, 1999; Anderson, 2001; Bishop et al., 2001; Whitaker and Hamill, 2002).

3.2.1 The Composite State Method

In the following discussion, we consider cases with a global model, whose state is denoted $\widetilde{\mathbf{x}}_0$, and c LAMs, with states labeled $\widetilde{\mathbf{x}}_i$, where $i = 1, 2, \ldots, c$. Our method does data assimilation on all model ensemble states simultaneously by performing the analysis on a 'composite' state ensemble. The composite state is defined on a model grid with potentially non-uniform spatial resolution that, in a given region, matches that of the highest resolution model defined there. More formally, denoting the region where the ith LAM is defined as L_i, the entire global model domain as L_0, and a location in the continuous global domain as A, the composite lattice has the same resolution at A as the global model when A is not in $\bigcup_{i=1}^{c} L_i$. Here the notation $\bigcup_{i=1}^{c} L_i$ denotes the subset of the global domain covered by all LAMs. When A is in $\bigcup_{i=1}^{c} L_i$, the composite state lattice has the same resolution at A as the highest resolution LAM defined there.

We view the state estimates contained in each forecast model ensemble to be approximations to the same true state, albeit with different accuracies. We consider situations in which high spatial-resolution short-term forecasts will be more

accurate than low spatial-resolution short-term forecasts, and thus we value those forecasts produced at higher spatial resolution more when constructing the composite state ensemble. This assumption, and the algorithmic simplifications it permits for performing data assimilation, are what lead us to the specific formulation of the composite state method presented here. In formulating our composite state method, we envision a situation in which there are no persistent forecast model biases in the LAM and global model forecasts, so that the dominant forecast errors inside the LAM domain are caused by spatial truncation errors. (The subject of correcting forecast model biases will be a subject of future work.) The mth ensemble member of the composite state ensemble, $\hat{\mathbf{x}}_m$, is constructed from the mth ensemble members of the global and each LAM ensemble. For brevity, in the following discussion we suppress the ensemble member subscript, m. The composite state $\hat{\mathbf{x}}(n)$ at location n is a linear function of the state vectors of the global and limited-area models whose domains contain n, as given by eq. (3.2),

$$\hat{\mathbf{x}}(n) = \sum_{i=0}^{c} p_i(n) O_n[\widetilde{\mathbf{x}}_i]. \tag{3.2}$$

The functions $p_i(n)$ define the weighting given to the ith LAM model state at location n. The operator O_n interpolates a state vector $\widetilde{\mathbf{x}}_i$ to location n when n does not correspond to a grid point of $\widetilde{\mathbf{x}}_i$, and acts as the identity operator in situations when n corresponds to a grid point of $\widetilde{\mathbf{x}}_i$. If all LAMs whose domains contain a location n have the same spatial resolution, but incommensurate grids, O_n will still need to interpolate, but the choice of which LAM grid to interpolate to is left to

the user. For all $i = 0, 1, \ldots, c$, the functions $p_i(n)$ satisfy $0 \leq p_i(n) \leq 1$ and $\sum_{i=0}^{c} p_i(n) = 1$ at every location n, and $p_i(n) = 0$ if n is outside the domain L_i of the ith LAM. In general, $p_i(n)$ should vary slowly with n, to ensure continuity of $\hat{\mathbf{x}}(n)$. However, if the LAM is evolved with boundary conditions from the global model, so that $\widetilde{\mathbf{x}}_i = \widetilde{\mathbf{x}}_0$ on the boundary of L_i, then $p_i(n)$ and $p_0(n)$ can change discontinuously as n crosses the boundary. Within these restrictions, there is a lot of freedom in picking the values of the $p_i(n)$ functions when LAMs overlap. In Section 3 we specify reasonable $p_i(n)$ choices for the situations we consider in our experiments.

The first step in performing data assimilation using the composite state method is to create a background ensemble of composite state vectors from the background ensembles of global and limited-area model states, using the definition of the composite state, eq. (3.2). Once the background composite state ensemble has been created, the analysis procedure is carried out on it, yielding the analysis composite state ensemble. The global and limited-area model state ensembles that are forecasted to the next analysis time are each constructed from the analysis composite state ensemble, interpolating from the composite state grid to the appropriate global or limited-area lattice when necessary. In the LETKF formulation, the interpolation is best done on the weight vector $\mathbf{w}$ and matrix $\mathbf{W}$, applying the interpolated weights to the model states on their native grid (Yang et al., 2009).

3.3 Numerical Experiments

For ease of presentation, in the following sections we specialize to the context of one spatial dimension, evenly spaced grid points forming a high-resolution grid, and a low-resolution grid of evenly spaced grid points that are a subset of this high-resolution grid. The global model is taken to be defined on the low-resolution grid, and each LAM is defined on a subset interval of the high-resolution grid. For this situation, the operator O_n in eq. (3.2) always acts as the identity, and can thus be omitted. In addition, we consider the case where at most two LAM domains can overlap at any given location. At locations where only the ith LAM is defined, $p_i(n) = 1$, and outside of the domain of the ith LAM, $p_i(n) = 0$. In our experiments, we chose $p_0(n) = 0$ at all locations n where LAMs are defined, and $p_0(n) = 1$ at locations where only the global model is defined. If LAM i overlaps with another LAM, we chose $p_i(n)$ to decrease linearly across the overlap region, to zero at the edge of LAM i. As an illustrative case for the example of two LAMs with domains as shown in Fig. 3.1(a), Figs. 3.1(b-d) show corresponding choices of $p_0(n)$, $p_1(n)$ and $p_2(n)$. We chose this form of weighting function for its simplicity, and because of our observation that boundary-condition errors decreased with increasing distance from the LAM's lateral boundary (see Fig. 3.2). For other models, it may be advantageous to have the global weights $p_0(n)$ taper more continuously from 1 to 0 near the boundary of the LAM domains, or remain positive throughout the LAM domains. Additionally, our experiments use LAMs governed by the same model physics, at the same spatial resolution. More complex scenarios may benefit by

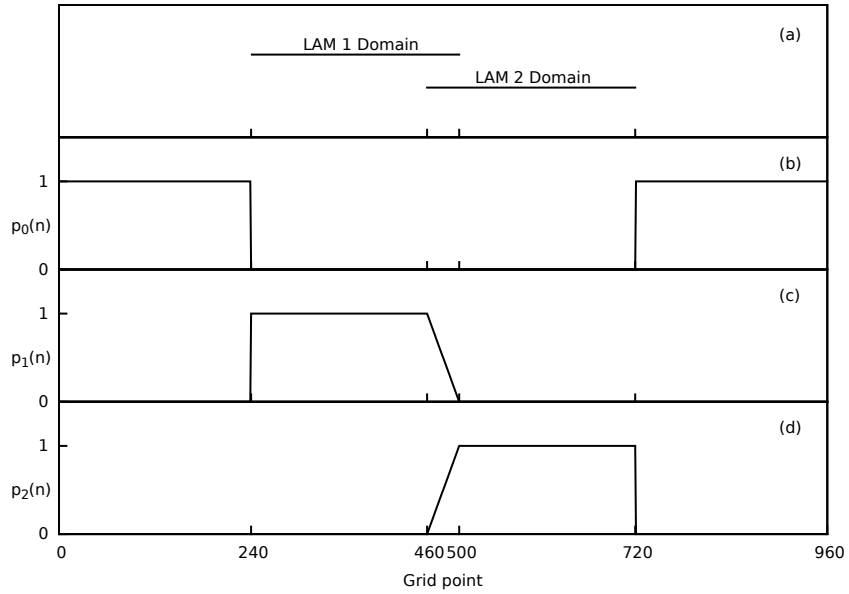

Figure 3.1: An example of the p_i functions for a scenario with two limited-area models, denoted LAM 1 and LAM 2, used in one of our experiments. Part (a) shows the domains on which the LAMs are defined, which cover grid point intervals of $[240, 500]$ and $[460, 720]$. Parts (b-d) show the functional form of the p_i functions for the global model, $p_0(n)$ (shown in panel (b)), and each of the limited-area models, $p_1(n)$ and $p_2(n)$(shown in panels (c) and (d), respectively). All plots show grid point location n on the horizontal axis.

choosing the $p_i(n)$ functions to more heavily weight LAMs with better historical error properties.

3.3.1 The Lorenz Models

To test our data assimilation framework we perform a series of numerical experiments. These tests require three models: a global high-resolution model which generates the simulated 'nature' or 'truth,' a high-resolution LAM model, and a low-resolution global model that has only large spatial-scale behavior.

With these considerations in mind, we utilize in our experiments two models

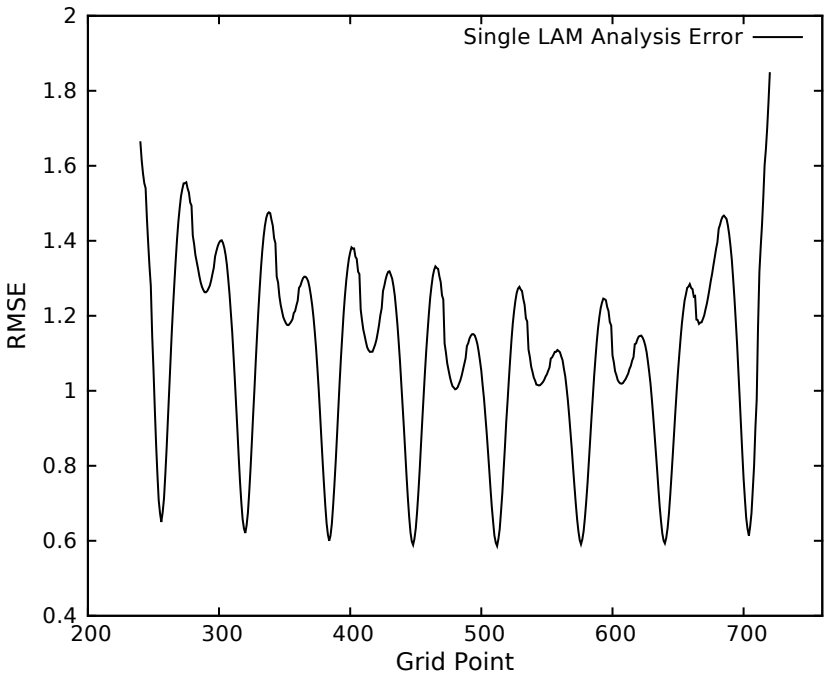

Figure 3.2: RMS analysis errors of the ensemble mean when the LETKF is performed using only the state information of a single LAM. The LAM is defined over grid points [240,720]. The RMS errors shown are averaged over 2×10^4 analysis cycles. Boundary condition errors can be seen in the increase in RMS error at grid points near the LAM boundary.

that are described in Lorenz (2005), known as Lorenz Models II and III. Lorenz Model II describes the spatiotemporal behavior of a quantity, Z, defined on a one dimensional lattice with periodic boundary conditions, analogous to a ring of constant latitude. A superscript n is used to index the value of Z at each grid point on this lattice. Model II exhibits spatially extended and smooth waves, and is given by

$$\frac{dZ^n}{dt} = [Z, Z]^{K,n} - Z^n + F. \tag{3.3}$$

The terms on the right hand side of eq. (3.3) are analogous to nonlinear advection, dissipation, and forcing, respectively. The bracket term in eq. (3.3) is a function of

inputs X and Y given by

$$[X,Y]^{K,n} = \sum_{j=-J}^{J}{}' \sum_{l=-J}^{J}{}' (-X^{n-2K-l}Y^{n-K-j} + X^{n-K+j-l}Y^{n+K+j})/K^2. \tag{3.4}$$

The forcing parameter, F, and smoothing parameter, K, in eq. (3.3) and eq. (3.4) are chosen by the experimenter, with $J = K/2$ when K is even, and $J = (K-1)/2$ when K is odd. The primed summation, $\sum'$, represents a modified summation, with the first and last terms divided by 2 when K is even. If K is odd, $\sum'$ denotes a normal summation.

As noted in Lorenz (2005), the waves of Model II have long spatial scales, making it fitting to represent a coarse 'global' model. Smaller spatial-scale behavior can be added to the dynamics of the quantity Z by modifying the equations of Model II, to arrive at Model III,

$$\frac{dZ^n}{dt} = [X,X]^{K,n} + b^2[Y,Y]^{1,n} + c[Y,X]^{1,n} - X^n - bY^n + F. \tag{3.5}$$

In Model III, the quantity Z varies spatially with long and short spatial scale components, denoted by X and Y, respectively. The first two terms on the right hand side (RHS) of eq. (3.5) represent spatially smoothed nonlinear advection of the long spatial scale component X and nonlinear advection of the short spatial scale component Y, respectively. The long and short scale components are coupled through the third term on the RHS of eq. (3.5). The last terms on the RHS of eq. (3.5) represent linear damping and constant forcing. The parameters b and c control the amplitude of short scale waves and the coupling between scales, respectively. We

use Model III to represent both the LAM and 'nature' dynamics. The long spatial scale component of a Model III state is found at grid point n via spatial averaging of Z,

$$X^n = \sum_{l=-I}^{I}{}' (\alpha - \beta|l|) Z^{n+l},$$

and the short scale component is found by $Y^n = Z^n - X^n$. The summation limit I is chosen by the experimenter. The α and β quantities are functions of I, whose exact form can be found in Lorenz (2005). These are chosen so that $X^n = Z^n$ whenever Z^n varies quadratically within $\pm I$ grid points of n. Like Model II, Model III is also defined on a 1-dimensional lattice with periodic boundary conditions, although typically at a higher spatial resolution. As noted by Lorenz, one model time unit of these models is analogous to approximately 5 days in the atmosphere.

3.3.2 Experimental Parameter and Domain Details

In our experiments, we use Model III with parameter values of $K = 32$, $b = 10$, $c = 0.6$, $F = 15$, and $I = 12$ to govern the LAM and 'nature' model dynamics. The global model is Model II, with parameters $F = 15$ and $K = 8$, and 1/4 the spatial resolution of the 'nature' model. In both Models II and III, K is a smoothing parameter which controls the spatial resolution of the long wavelength waves. Additional experiments used $c = 2.5$ in the LAM and nature model dynamics, enhancing the difference between the global and limited-area/nature model dynamics. Despite more dramatic differences between the global and limited-area model attractors, the

composite state method performed similarly to the experiments reported on here. The 'nature' model lies on a lattice of $n = 960$ grid points, which are indexed from $n = 0, ..., 959$. The LAM domains are not constant from one experiment to the next, but are always defined on continuous subsets of the 'nature' grid, such as $n = 0, 1, ..., 540$ or $n = 480, 481, ..., 60$, for example. In most of our experiments, we consider a global model defined on a lattice of 240 grid points, each corresponding to every 4th 'nature' model grid point, $n = 0, 4, 8, 12, ...956$. In experiments where the global model resolution is lowered by a factor of two, K is adjusted to $K = 4$. To avoid ambiguity, we index all model grid points by their corresponding 'nature' grid point index.

The global and LAM ensembles have 40 ensemble members in our experiments. The initial Model II global ensemble is sampled from a free run of the global model, after 600 model 'days' of spin up time. At the beginning of each experiment, each LAM ensemble member is produced by interpolating a corresponding initial global Model II ensemble member onto the finer LAM grid, at locations where the LAM is defined. The initial conditions of the 'nature' model are found by initializing its grid points with random numbers uniformly distributed in the interval $[0, 1]$, and allowing the model to spin up for 600 model days.

During our numerical experiments, we perform data assimilation every $dt = 0.05$ model time units (the equivalent of about every 6 h in real time, according to Lorenz (2005)). At each analysis time, observations of the 'nature' model at 15 equally spaced observation locations, located at grid points $0, 64, ..., 896$, are created by adding Gaussian noise with mean 0 and standard deviation 1 to the

'nature' model values. A key parameter of the LETKF is the size of the local state-vector patch on which the analysis is performed. For our experiments, we use a local patch size that is 81 grid points wide, so that at least 1 observation is assimilated at every grid point. The influence of observations inside of a local analysis region was not tapered for the experiments reported on here; we expect that results could improve when that technique is implemented in the LETKF. Also, multiplicative covariance inflation is used in our experiments to prevent filter divergence (Anderson and Anderson, 1999). At each analysis cycle, the composite ensemble background covariance matrix is inflated by the constant factor $\rho = 1.048$, which was found through empirical tuning to minimize RMS error.

3.3.3 Numerical Integration

The global and regional ensembles are forecasted simultaneously, using a fourth order Runge-Kutta scheme, breaking the '6 hour' forecast between analysis times into 36 time steps. The boundary conditions required by each LAM ensemble member are provided by its corresponding global ensemble member. Specifically, interpolation of the state values of the appropriate global ensemble member is used to provide state values needed in eq. (3.4) that are outside of the LAM domain when forecasting the LAM states using eq. (3.5).

During the integration, we utilize Davies Relaxation (Davies, 1983). We define 'sponge regions' at both boundaries of each LAM domain, having a length of 10 LAM grid points. After the ensembles have been integrated forward in time by 1 time

step, the state of an ensemble member of LAM i at a grid point in a sponge region is updated to a linear combination of the corresponding forecasted LAM i and global model ensemble members at that grid point, according to

$$\widetilde{\mathbf{x}}_i(n) \to (1 - \gamma(n))\widetilde{\mathbf{x}}_i(n) + \gamma(n)\widetilde{\mathbf{x}}_0(n). \tag{3.6}$$

Here, $\widetilde{\mathbf{x}}_i(n)$ and $\widetilde{\mathbf{x}}_0(n)$ are the state values at grid point n of a member of limited-area ensemble i and a global model ensemble member, respectively, and $\gamma(n)$ is a spatially dependent weighting function. In our experiments, $\gamma(n)$ decreases linearly over the sponge region, from a value of 1 at the outer sponge region boundary to 0 at the inner sponge region boundary. At grid points in the sponge region at which the global model is undefined, the global state value is linearly interpolated onto the finer LAM mesh, and this value is used for $\widetilde{\mathbf{x}}_0(n)$ in eq. (3.6).

3.3.4 Verification Details

The results presented below use the temporally averaged Root-Mean Square Error (RMSE) between the ensemble mean, $\bar{\mathbf{x}}$, and the truth $\mathbf{x}^t$

$$\text{RMSE}(n) = \sqrt{\frac{1}{T}\sum_{q=1}^{T}(\bar{\mathbf{x}}_q(n) - \mathbf{x}_q^t(n))^2},$$

as a measure of the effectiveness of our method. Here the subscript q indexes analysis cycle. The RMSE of the ensemble mean at grid point n is the average, over T analysis cycles, of the squared difference between the ensemble mean and the truth at n, square-rooted. Errors are calculated at each analysis time, as well

as for single deterministic forecasts. These forecasts use analysis ensemble means as initial conditions.

We test the composite state method in two situations, one where the LAMs collectively cover the global domain, $L_0 = \bigcup_{i=1}^{c} L_i$, and one where they do not. In each case, we generate forecasts, initialized from the analysis ensemble mean, for both the global and limited-area models.

We compare results using the composite state method to high- and low-resolution forecasts made from a pair of models that are defined over the entire experimental domain. These benchmark high- and low-resolution ensemble forecasts are created using the same high- and low-resolution models used to create the composite state method forecasts, Lorenz Models III and II, respectively. The benchmark ensemble forecasts assimilate the same set of observations as the composite state forecasts, using the same algorithm (the LETKF) with the same number of ensemble members. The benchmark ensembles are integrated for 6 model hours between analysis cycles. The composite state analysis ensemble mean is verified against the benchmark analysis ensemble mean after each analysis. For forecasts longer than 6 hours, both benchmark and composite state forecast estimates are found by integrating the benchmark and composite state analysis ensemble means, using the appropriate forecast models.

For the experiments in which the LAMs do not collectively cover the entire experimental domain, we compare composite state analyses and forecasts to those from the 'joint-state method' of Yoon et al. (2012). The joint-state method performs data assimilation simultaneously on both global and limited-area models. It

accomplishes this by using an observation function that predicts the observations by using information from both the global model state and the LAM state, as well as by including a constraint term in the cost function that penalizes large differences between global and LAM model states. We compare the high-resolution limited-area forecasts of the joint-state method to similar high-resolution forecasts created using the composite state method.

Our method differs from the joint-state method in a number of ways, one of which is that the joint-state method does not constrain the global and LAM analyses to be identical. More specifically, the joint-state method performs data assimilation by minimizing a local cost function, $J^n(\widetilde{\mathbf{x}})$, at each grid point. This local cost function depends upon the local background ensemble, local observations, and their respective error covariances, and is given by

$$J^n(\widetilde{\mathbf{x}}) = (\widetilde{\mathbf{x}} - \bar{\mathbf{x}}_b)^T (\mathbf{P}^b)^{-1} (\widetilde{\mathbf{x}} - \bar{\mathbf{x}}_b) + (\mathbf{y} - h(\widetilde{\mathbf{x}}))^T \mathbf{R}^{-1} (\mathbf{y} - h(\widetilde{\mathbf{x}})) + \kappa (\widetilde{\mathbf{x}}_g - \widetilde{\mathbf{x}}_r)^T (\widetilde{\mathbf{x}}_g - \widetilde{\mathbf{x}}_r). \tag{3.7}$$

Here $\bar{\mathbf{x}}_b$ is the local background ensemble mean, $\mathbf{P}^b$ is the local ensemble sample covariance, $\mathbf{y}$ is a vector of local observations, $\mathbf{R}$ is the local observation error covariance matrix, and $\widetilde{\mathbf{x}}_g$ and $\widetilde{\mathbf{x}}_r$ are vectors which contain the global and regional (LAM) model state values, respectively, at grid points at which both the global and regional ensembles are defined. The κ in eq. (3.7) is a parameter. Importantly, the forward operator h used in the joint-state method depends on location. Inside of the local analysis region, it maps a linear combination of global and regional

model states to observation space, and depends upon a parameter λ. Both of these parameters, λ and κ, must be tuned for optimal performance in any application of the joint-state method. In applications with d limited-area models, the joint-state method would necessitate the empirical tuning of $2d$ of these parameters, in addition to parameters associated with other empirical techniques, such as covariance inflation. Additionally, the computational cost of finding the minimum of eq. (3.7) grows quickly with the number of LAMs present in a local analysis region, as each of the d LAMs would contribute a term equivalent to the third term in eq. (3.7). Both of these qualities make implementing the joint-state method in any context with multiple LAMs exceedingly complicated. The composite state method proposed here presents a simpler approach to performing data assimilation in the multiple LAM context that is a natural extension of the joint-state method which avoids the necessity of empirically tuning a large number of adjustable parameters, and allows a considerably simpler cost function to be minimized. The composite state method corresponds to the case where $\lambda = 1$ and $\kappa = \infty$ in eq. (3.7). For additional details on the joint-state method, see Yoon et al. (2012).

As a diagnostic of the composite state method, we conducted ensemble forecasts, verified at several lead times, measuring the relationship between ensemble spread and ensemble RMS error while using the composite state method, and found that the global model ensemble spread adjusts appropriately to match the decreased RMS ensemble forecast error of the composite state method. Globally averaging over space and time, the ensemble spread for 6 hour forecasts was found, for the cases considered, to be approximately equal to the ensemble RMS error. Specifically,

these quantities differed by approximately 2 percent.

3.4 Results for Global LAM Coverage

As a first test of the composite state method, we apply it to a situation where there are two LAMs, whose domains together cover the entire 'global' domain of our experiments. Both LAMs are driven at their boundaries by the global model dynamics. When there is no communication between global and LAM ensembles during the analysis it is typical that boundary conditions supplied this way lead to an increase in errors near LAM boundaries, as a result of mismatches in state information at these locations. To see how such errors can be eliminated by the composite state method, Fig. 3.2 provides an illustrative example of lateral boundary-induced LAM errors, for the case of a single LAM defined over the interval $[240, 720]$ of our $n = 0$ to 960 grid, when assimilation is performed separately on the limited-area and global model states with no feedback between limited-area and global model state information during the analysis. While both ensembles of model states assimilate the same observations, the ensemble of LAM states is used as the background ensemble only for the LAM assimilation, and the ensemble of global states is used as the background ensemble only for the global assimilation. Fig. 3.2 shows how error near both LAM boundaries can be dramatically larger than the error closer to the LAM domain interior.

In contrast to the separate analysis method shown in Fig. 3.2, we find that applying the composite state method in multiple LAM situations allows LAM analysis,

1-day and 5-day forecast errors to rival those of a globally high-resolution perfect model, as demonstrated in Figs. 3.3-3.5, for a situation with two overlapping LAMs that cover the entire $n = 0$ to 960 domain. For the case shown in Figs. 3.3-3.5, the LAMs are defined over the intervals $[0, 520]$ and $[480, 40]$, overlapping at 41 grid points near each of their boundaries. Similar LAM analysis and forecast accuracies were achieved with other overlap values, as shown in Fig. 3.6. In the experiments whose results are shown in Fig. 3.6, analysis accuracy of the composite state analysis ensemble mean is calculated when two equally sized LAMs, whose domains collectively cover the entire experimental domain, are used to construct the composite state ensemble. The analysis RMS error is calculated as a function of LAM domain overlap, with larger overlap corresponding to larger LAM domain size. As the size of the LAM domain overlap grows, we see from Fig. 3.6 that there is virtually no change in analysis or forecast errors, indicating that there is little benefit to large LAM domain overlap.

In Figs. 3.3 and 3.4, we can see that the RMS analysis and 1-day forecast errors obtained using the composite state method (red and green curves, respectively) are virtually the same as those obtained using a global high-resolution perfect model (black curve). Figure 3.5 shows 5-day forecast statistics, produced by limited-area and global models initialized to the composite state analysis ensemble mean. For comparison, curves of the RMS error of 5-day forecasts produced by global low- and high-resolution ensembles that do not use the composite-state method are also shown. Comparing the 5-day forecast errors produced by the low-resolution global model initialized from composite state (blue curve) and control (orange curve) anal-

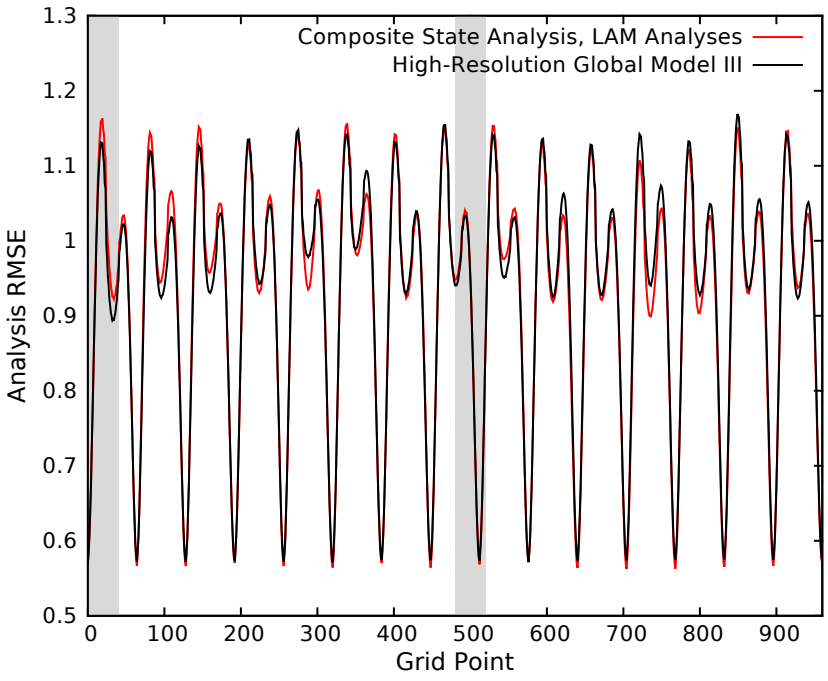

Figure 3.3: RMS analysis errors of the composite state ensemble mean (red curve). For comparison, the analysis error of the ensemble mean for a global high-resolution perfect model LETKF analysis (black curve) is also shown. LAMs are defined over grid points [0,520] and [480,40], and statistics are taken over 10^5 analysis cycles, discarding the first 10^3 cycles. The shaded areas indicate the domain where both LAMs are defined.

yses, the composite state method substantially improves the global model 5-day forecasts, indicating that much of the difference between the high and low-resolution global model forecasts (black and orange curves, respectively) is due to initial condition errors.

The 5-day LAM forecast accuracies are able to approach those of high-resolution global model forecasts (black curve in Fig. 3.5) in the interior of the LAM domains. The effects of imperfect boundary information coming from the global model can be seen in Fig. 3.5, as LAM forecast errors rise near LAM boundaries. The size of the effected region is dictated by the flow of imperfect state information into the

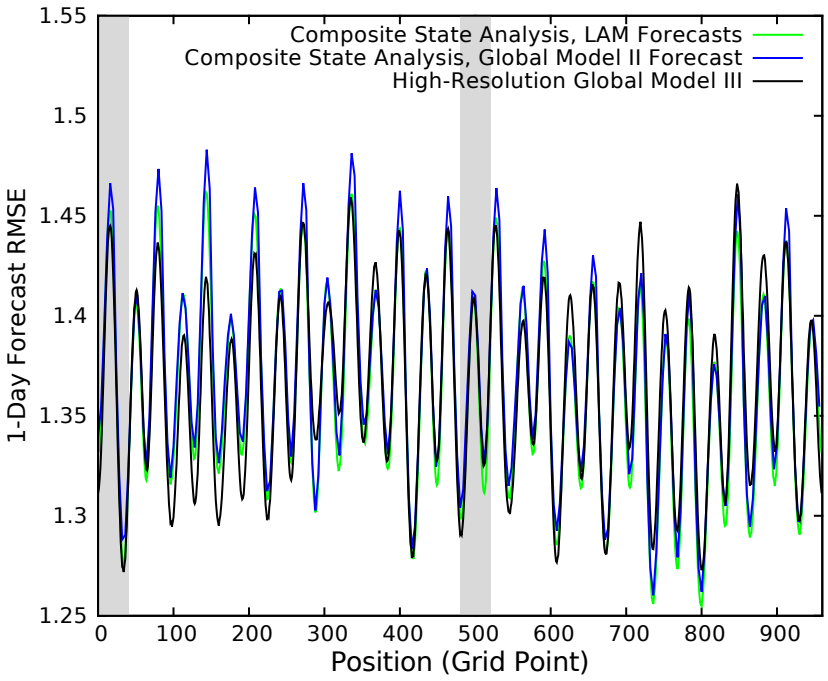

Figure 3.4: RMS 1-day forecast errors, initialized using the composite state analysis ensemble mean. The green curve shows errors of forecasts produced by the LAMs, while the blue curve shows errors for forecasts produced by the low-resolution, imperfect global model. For comparison, forecast errors produced by a global high-resolution perfect model initialized from an LETKF analysis are shown as a black curve. These results are from experiments under the conditions described in Fig. 3.3. The shaded areas indicate regions of LAM domain overlap.

LAM from the lower resolution global model. As shown by Yoon et al. (2010), this information moves predominantly 'eastward' (direction of increasing n), at a rate of 1.4 grid points per 'hour' in the Lorenz models. Thus for a 5-day (120 hour) forecast, on the order of 160 grid points will be adversely affected by boundary condition errors, which is approximately what can be seen in Fig. 3.5, in the grid point intervals $[0, 160]$ and $[480, 640]$. Also, compatible with the predominantly 'eastward' propagation of information in the Lorenz models (Yoon et al., 2010), we see that the adversely affected region of the LAM domain is larger to the 'east' of a LAM boundary than to the 'west' of a boundary. Choosing LAM domains with greater

overlap can help mitigate the effect of LAM boundary errors on forecasts, as with sufficient overlap grid points affected by boundary conditions in one LAM might correspond to more accurate, interior grid points of another.

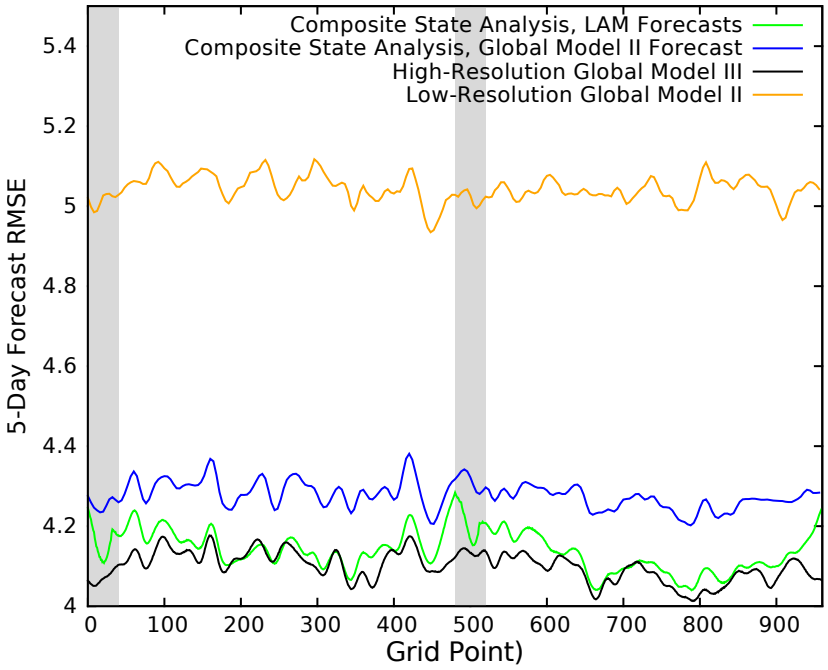

Figure 3.5: RMS 5-day forecast errors, initialized using the composite state analysis ensemble mean. The green curve shows errors of forecasts produced by the LAMs, while the blue curve shows errors for forecasts produced by the low-resolution, imperfect global model. For comparison, forecast errors produced by a global high-resolution perfect model and a global low-resolution imperfect model, initialized from an LETKF analysis are shown as black curve and orange curves, respectively. These results are from experiments under the conditions described in Fig. 3.3. The shaded areas indicate regions of LAM domain overlap.

To see how the global model resolution influences the accuracy of these results, we apply the composite state method in an experiment with two LAMs defined over the intervals $[0, 540]$ and $[480, 60]$, but do so using a global model with $K = 4$ that is defined on every 8 of the nature model grid points, or half of the resolution of the previously used global models. To quantify how much the RMS forecast and analysis

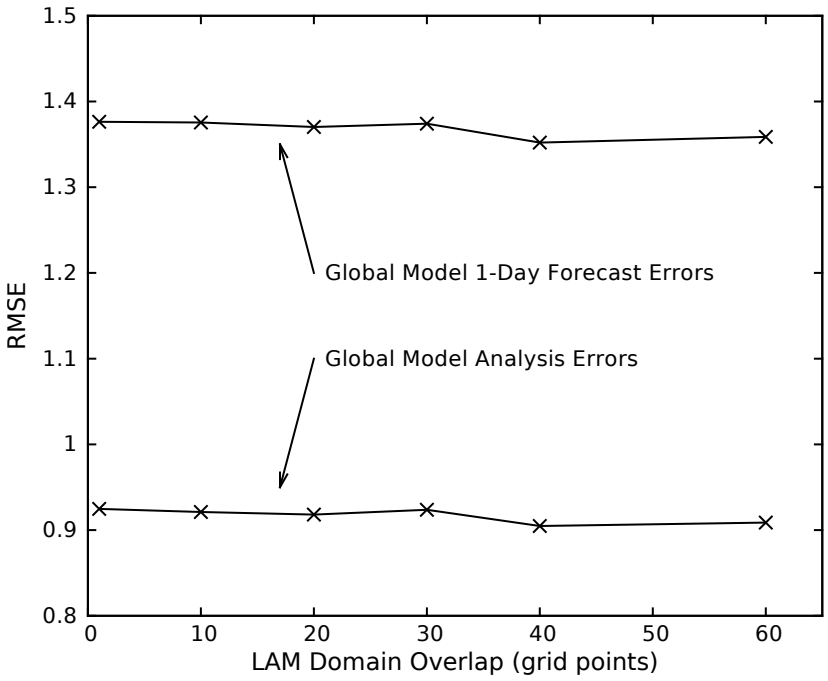

Figure 3.6: RMS analysis and 1-day forecast errors of the global model ensemble mean, averaged over all grid points and time (10^5 analysis cycles), discarding 10^3 initial spin-up cycles, and performing the analysis using the composite state method. The results shown above are for two LAMs whose domains tile the globe. The x-axis shows the number of grid points that the LAM domains have in common. For these models, there appears to be no benefit to large LAM domain overlap.

error change when the global model resolution is lowered, we compare spatiotemporally averaged RMS errors of composite state analyses and 1-day forecasts made using global models at both of these spatial resolutions. The lowering of resolution causes the RMSE of the composite state analysis mean to increase by approximately 2.9%, and the 1-day forecast RMSE of the LAMs to increase by approximately 1.7%, while decreasing global forecasting costs by 50%, as a result of decreased function evaluations. Thus, for our original setup, we see that the resolution of the global model may be lowered without much loss of accuracy.

We now test the composite state method on multiple LAMs defined over a

larger experimental domain that consists of twice as many grid points (1920) as in the experiments just described, but maintains spatial resolution and model II and III parameter values (e.g. $K = 32$ for LAM and nature dynamics). Thus, there are now four LAMs in this 'Large World' experiment, each defined over 541 grid points. The observation density is also held constant. The 1-day forecast results from this experiment are shown in Fig. 3.7. The composite state method is again seen to achieve performance that is virtually the same as that of a global high-resolution perfect model. In this scenario, the errors are approximately constant across LAM domains and there are no large deviations in RMSE at the LAM boundaries (shaded regions) similar to those seen at the LAM boundaries in Fig. 3.2, indicating that lateral boundary errors have been minimized. The results of this experiment lead us to believe that the composite state method is scalable with LAM number, and similar forecast accuracies may be achieved with a much larger number of LAMs, in a proportionately much larger global domain.

Another condition we investigate is the size of LAM domains. In order to test this we consider an experiment on our original $n = 0$ to 960 grid that uses an increasing number of LAMs. Specifically, we calculate errors of the composite state ensemble mean when two, four, eight, and sixteen identically sized LAMs are independently forecasted during each analysis cycle; results of these experiments are shown in Fig. 3.8. These LAMs have sizes of 521, 261, 131, and 65 grid points, with overlaps of 41, 21, 11, and 5 grid points, respectively. The concurrently running global model ensemble provides boundary conditions to each of these individual LAMs. Despite the smaller domain sizes of the experiments with four or eight

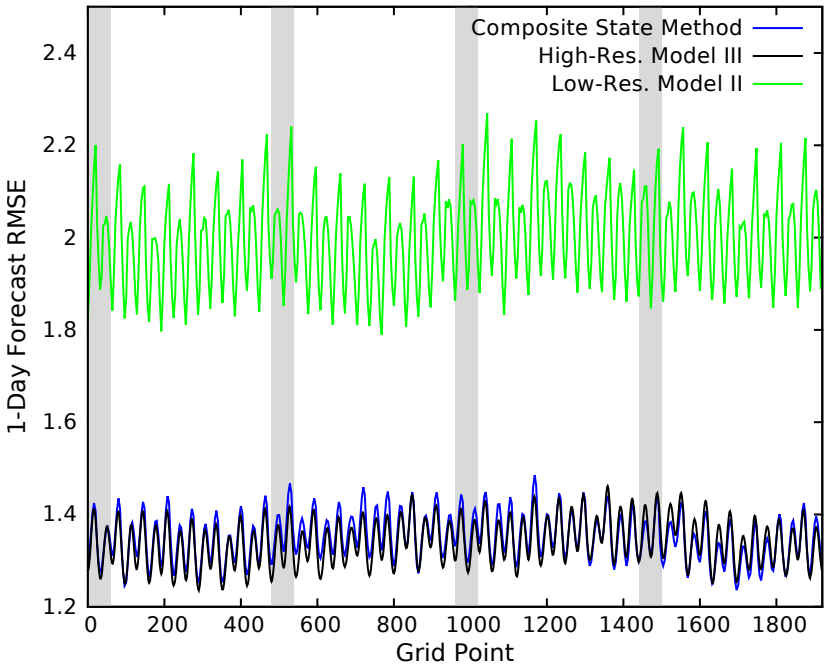

Figure 3.7: RMS 1 day forecast errors, averaged over 10^5 analysis cycles, in the 'Large World' scenario. The experimental domain runs from $n = 0$ to 1920, and LAMs are defined over grid point intervals $[0, 540]$, $[480, 1020]$, $[960, 1500]$ and $[1440, 60]$. The blue curve shows forecasts, initialized using the composite state analysis ensemble mean, made with the low-resolution global model. Observations are located at every 64 grid points, and the shaded areas indicate grid point intervals where more than one LAM domain is defined.

LAMs, Fig. 3.8 shows that there is not much loss of accuracy with these domain sizes. In Fig. 3.8 we can see that errors begin to increase when the LAM size is small enough such that on the forecast phase of the analysis cycle, boundary errors from the driving global dynamics are able to affect a larger proportion of the LAM domain. This increase in error is not because of smaller LAM overlap, as Fig. 3.6 shows that forecast accuracy is not strongly dependent on LAM domain overlap. As the speed of information flow in the Lorenz models is approximately 1.4 grid points per hour (Yoon et al., 2010), in one 6-hour analysis cycle information travels about 10 grid points. For a 24-hour forecast time, information from the boundaries would

affect LAM grid points up to approximately 30 grid points inside the LAM domain, and it is unsurprising that LAM domains large enough such that these boundary regions represent a small fraction of overall size would exhibit similar forecast results.

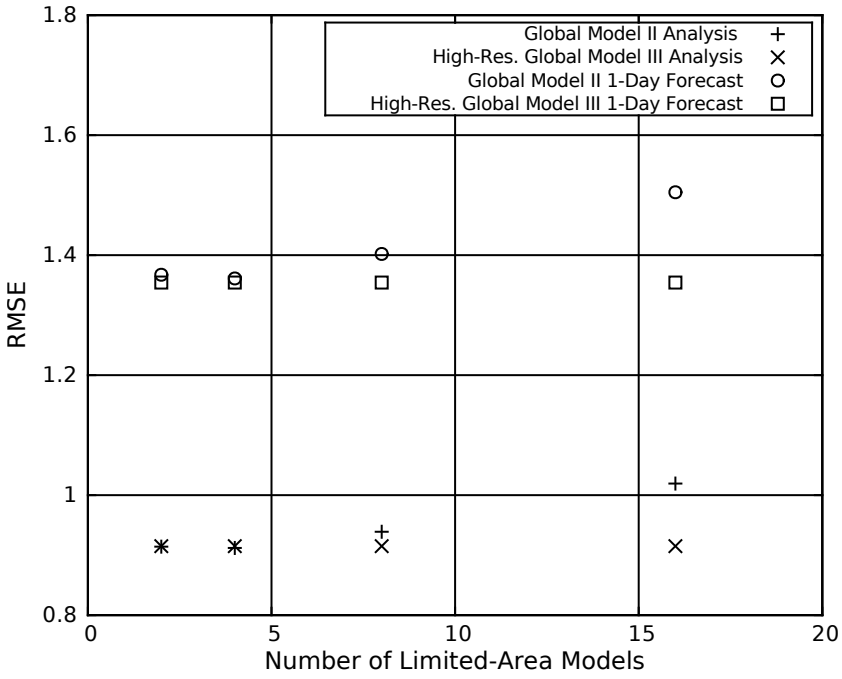

Figure 3.8: RMS analysis errors of the ensemble mean and 1-day forecast errors calculated using the composite state method, when the entire simulation domain is divided amongst different numbers of LAM domains. In a given experiment, each of the LAM domains are the same size, so that the LAM domains are 521 grid points long in the two LAM case, 261 grid points long in the four LAM case, 131 grid points long in the eight LAM case, and 65 grid points long in the sixteen LAM case. Errors begin to increase as the area influenced by boundary condition errors becomes a larger part of the total LAM domain. Statistics are averaged first over 10^5 analysis cycles, discarding the first 10^3 cycles, then over all grid points.

3.5 Results for Incomplete LAM Coverage of Global Domain

As we have seen, analyses and forecasts produced using the composite state method can rival those produced by a high-resolution ensemble of perfect model

states when there is a collection of LAMs which cover the entire experimental 'globe.' We now show that forecasting systems comprised of a single limited-area and global model pair can realize dramatic benefits to analysis and forecast accuracy if state information from an additional LAM is included in the analysis procedure. To do this, we first find analysis errors for a single limited-area and global model pair, when the analysis is performed using both the composite state method and the joint-state method of Yoon et al. (2012). These errors are then compared to those calculated using the composite state method when there are two limited-area models and a global model. In this second situation, the limited-area models are defined over the domains $[240, 720]$ and $[720, 240]$, collectively covering the globe, but with overlap only at the two boundary points.

Figure 3.9 shows the results of these experiments as curves of the RMS error of the analysis composite state ensemble mean, for the LAM defined over the interval $[240, 720]$ of a $n = 0$ to 960 grid point domain. This result demonstrates that for the single LAM case, the composite state analysis error (blue curve) is approximately the same as that calculated using the joint-state method (brown curve). However, when we add another LAM to the analysis the accuracy of the composite state method analysis (green curve in Fig. 3.9) becomes competitive with the high-resolution global perfect model ensemble (black curve), as shown above in Section 4 (see Fig. 3). The previous strong increase of analysis error near the left LAM boundary is nearly eliminated when this extra LAM state information is considered in the analysis procedure. (The asymmetry between left and right boundaries is a result of the eastward (direction of increasing n) 'group velocity' of waves in the

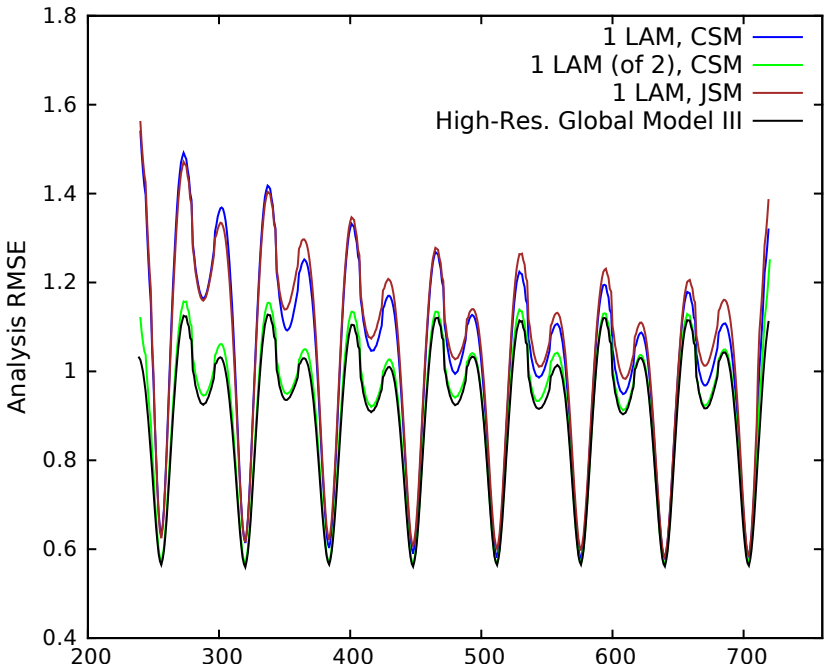

Figure 3.9: Ensemble mean analysis accuracy gains with the addition of a second LAM. Statistics are gathered over 10^5 analysis cycles, after 10^3 cycles of spin-up time. In both the one and two LAM situations the LAM domain of interest runs from grid points [240,720]. The second LAM is added on the domain [720,240]. Curves denoted 'CSM' are calculated using the composite state method, and those denoted 'JSM' are found when using the joint-state method of Yoon et al. (2012). The results from the perfect model ensemble are included over the LAM domain of interest as a benchmark for comparison.

Lorenz models, which is analogous to the westerly atmospheric flow in the Northern Hemisphere mid-latitudes.) Comparing the blue curve in Fig. 3.9 to the curve in Fig. 3.2 shows that the composite state method helps alleviate increases in boundary error near the right LAM boundary as well.

Even more dramatic effects are seen in Fig. 3.10, which shows the benefits to 1-day forecast accuracy when state information from an additional LAM ensemble is considered in the analysis. The composite state method considers this extra information in the analysis and can produce an analysis ensemble whose mean allows

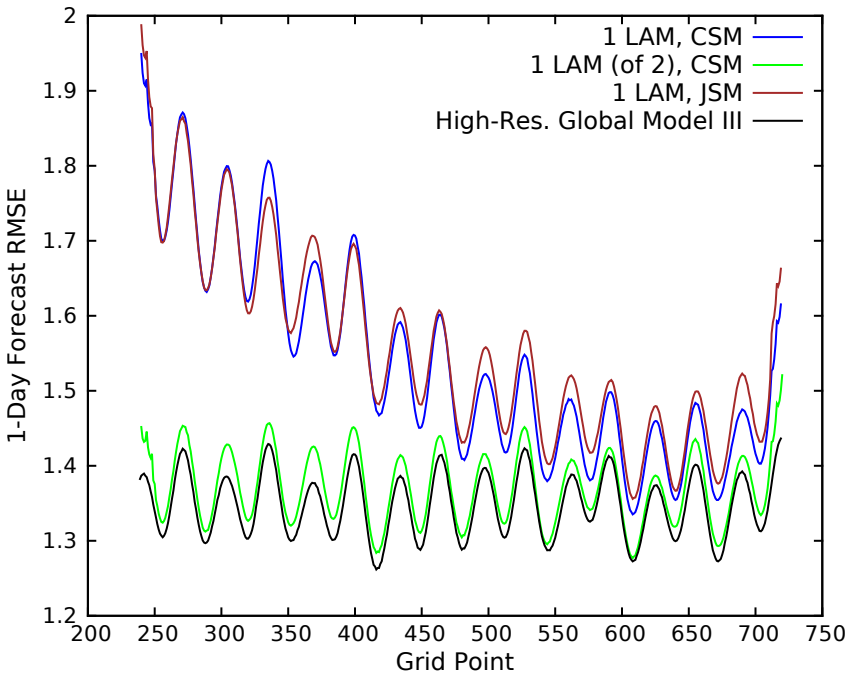

Figure 3.10: Ensemble mean 1-day forecast accuracy gains with the addition of a second LAM, for the experiment described in Fig. 3.9. Single deterministic forecasts are initialized with the LAM analysis ensemble mean.

more accurate LAM and global model forecasts than would otherwise be possible. We note that the addition of a LAM allows the composite state method to greatly decrease the forecast error near the left-most ('western') LAM boundary. An additional LAM also allows forecasts made by the lower resolution global model to improve, as shown in Fig. 3.11. Here we compare the 1-day forecast accuracy of forecasts produced by the global model, initialized with a composite state analysis mean calculated using 1 or 2 LAMs (blue and gold curves, respectively). The addition of an LAM helps to more accurately initialize global model forecasts across the entire simulation domain, allowing forecast accuracies to approach those made by a higher resolution perfect model over the whole 'globe,' rather than only at certain locations inside the LAM domain, which may be far from its boundaries.

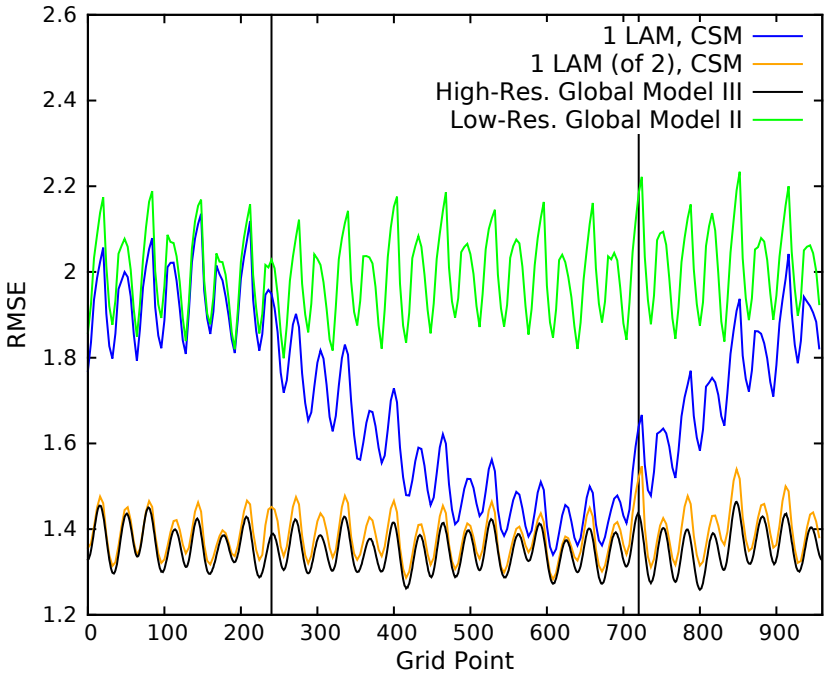

Figure 3.11: Ensemble mean forecast accuracy of the global model, for the conditions described for Fig. 3.9. The addition of a second LAM dramatically lowers global model forecast accuracy over the entire global domain. Results from perfect and imperfect global model ensemble forecasts are included as a benchmark for comparison, and vertical black lines demarcate the LAM boundaries.

Figure 3.12 shows results for our final experiment, which compares state estimates produced for two scenarios. The first has two small overlapping LAMs, covering grid point intervals $[240, 500]$ and $[460, 720]$ (red curve), and the second has a single larger LAM defined over the union of these domains, the interval $[240, 720]$ (blue curve). Data assimilation is performed in both of these situations using the composite state method. Overall, the accuracy of the state estimates produced in both scenarios, when averaged over time, is almost the same. This is a somewhat surprising result, as the two LAMs are driven by the imperfect global dynamics. At regions of LAM overlap and near LAM boundaries, the composite state method eliminates much of the boundary condition errors that would otherwise be present at these locations (for example near the LAM boundaries in Fig. 3.2). Overall, we conclude from our experiments that by considering all relevant LAM state information during the analysis, the composite state method helps to decrease, and in some cases virtually eliminate, boundary condition errors, and that this improved analysis state translates to better forecast performance.

3.6 Summary and Conclusions

While the advantages of using limited-area models for short-term (about 48h and shorter) weather forecasting have been known for some time, data assimilation has been traditionally performed solely on either the limited-area model or the global model that provides lateral boundary conditions. However, recent results indicate that this may not be the most optimal course of action (Guidard and Fischer, 2008;

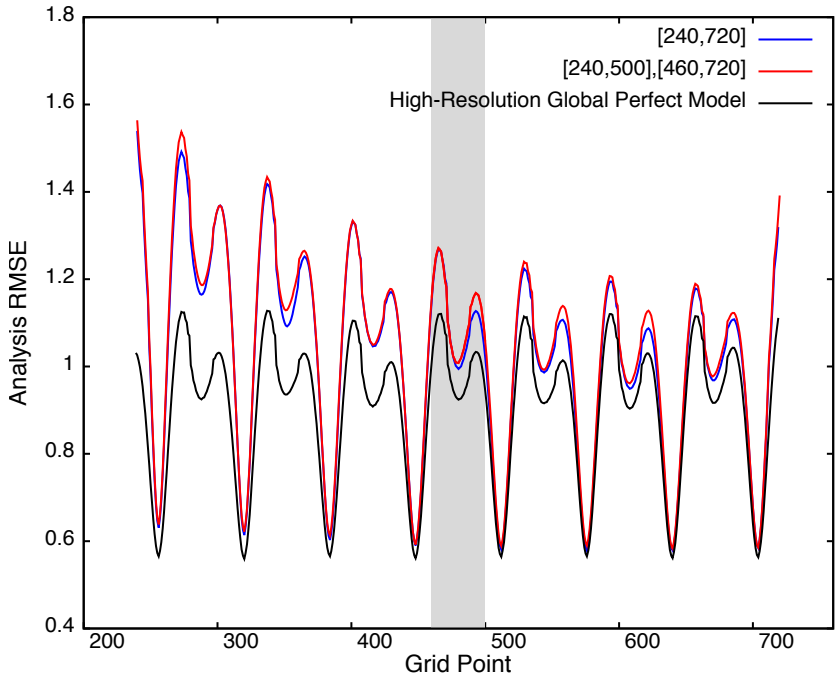

Figure 3.12: RMS analysis errors of the ensemble mean calculated using the composite state method for two model scenarios. The first scenario has a single LAM defined over the grid point interval [240,720] (blue curve), and the second has two LAMs defined over the intervals [240,500] and [460,720] (red curve). The analysis error of the ensemble mean of a global high-resolution perfect model LETKF (black curve) is included as a benchmark for comparison. Statistics are gathered over 10^5 analysis cycles, discarding the first 10^3 cycles. The shaded area indicates the domain where both LAMs are defined.

Dahlgren and Gustafsson, 2012; Yoon et al., 2012). Rather, forecast accuracy can increase by allowing the global model to influence limited-area models, and vice versa, in the analysis. Motivated by these findings, and the fact that some large weather centers run limited-area models for multiple regions, we present an ensemble data assimilation scheme that allows a global and several limited-area models to influence one another during the analysis procedure.

Using numerical experiments conducted under idealized test conditions, we show that our method has the potential to improve forecast accuracy. When ap-

plied to multiple LAMs, the first step of our analysis algorithm creates an ensemble of high-resolution 'composite' states from the ensembles of global and all limited-area model states. We then perform an ensemble analysis procedure (in the case considered in this chapter, the LETKF), to arrive at a composite state analysis ensemble. Next, we use the composite state analysis ensemble to construct global and limited-area model ensembles to be employed as initial conditions for forecasts that provide the background ensemble for the next analysis cycle. We note that, through the analysis procedure and the boundary conditions supplied by the global model ensemble, the composite state method effectively allows observational information from outside of a limited-area model domain to influence the update of the estimate of the state at grid points inside of its domain. Additionally, we note that this general composite state method does not depend on the particulars of the LETKF, allowing for flexibility in the choice of the analysis algorithm.

For experiments with LAMs that cover the entire global domain, the composite state method is shown to be capable of producing forecasts with accuracies that are almost as good as those produced in the ideal case of assimilations using a global perfect model. Additionally, in situations where the collection of LAMs cover a subsection of the global domain, we show that there is a clear benefit to allowing the LAM states to influence one another during the data assimilation process. The experimental results shown here suggest that, in real weather forecasting situations, high-resolution state estimates provided by many (potentially overlapping) limited-area models can be used to greatly improve global atmospheric state analysis estimates. Even including information about the state from limited-area models

with non-overlapping domains in the composite state analysis has a positive effect on global model state estimates outside of a limited-area model domain (see blue and orange versus green curves in Fig. 3.11). The composite state method analysis technique could be a useful tool for organizations like the United States Navy, which currently runs the COAMPS® limited-area model for many, often overlapping, domains, as it presents a straightforward method for utilizing a collection of disparate state estimates in the analysis. The composite state method would allow for short-term high-resolution forecasts produced by these regional models to improve the global analyses, in turn allowing for further improvements to the limited-area model forecasts through improved boundary conditions.

Our presentation of the composite state method has utilized simple one-dimensional models, each defined on subsets of the same grid. Atmospheric models are defined over three spatial dimensions, and different LAMs will not, in general, share common grid points, even if their domains happen to cover common geographic areas. Thus, in further tests of the composite state method on more realistic systems, it will be necessary to choose appropriate $p_i(n)$ functions, and to specify the interpolations that the operator O_n in eq. (3.2) is to perform. These choices will be dependent on the number and relative layout of LAM domains, and can be empirically adjusted to yield the most desirable results. Our choice of $p_i(n)$ used no information from the low-resolution forecast wherever a high-resolution forecast was available. It remains a possible subject for further study whether it is advantageous to retain some of this information in other scenarios. Overall, we are encouraged by our results from the one-dimensional models to speculate that the composite state method might offer

a potential means of obtaining forecast improvements in real weather forecasting situations.

Chapter 4: Composite State Data Assimilation with Forecast Model Bias

Limited-area models are commonly used to generate regional high-resolution atmospheric forecasts. Correcting systematic forecast model errors, known as *forecast model bias*, can be complicated in such systems by the limited-area model's need for lateral boundary conditions. Lateral boundary conditions are typically provided by a concurrently running, nested limited-area or global model, each of which frequently exhibits model errors of its own. In operational weather forecasting, data assimilation is used to correct global and limited-area forecasts toward recent observations, but this is typically done separately for each model. Performing data assimilation simultaneously on both global and limited-area model states has recently been shown to be beneficial to both global and limited-area model states (Yoon et al., 2012). However, simultaneous assimilation introduces additional coupling between model states, further allowing errors in the forecast models to affect each other.

The composite state framework discussed in Chapter 3 forms a state estimate from a combination of limited-area and global model forecasts. The purpose of this chapter is to illustrate the possibility of applying bias correction when using

the composite state method, and further, to illustrate that doing this may have substantial benefits to analysis and forecast accuracy. Correcting composite state forecast model bias accounts for the overall effect of model biases present in coupled global and limited-area forecast models. For this purpose, we will use a bias correction scheme initially presented in Baek et al. (2006).

Data assimilation techniques have been successfully applied to estimate the effect of, and to correct for, forecast model biases using variational and ensemble techniques (Dee and Da Silva, 1998; Carton et al., 2000; Dee, 2005; Keppenne et al., 2005; Li et al., 2009). Often, this is accomplished through state-vector augmentation techniques (Jazwinski, 1970; Cohn, 1997; Anderson and Anderson, 1999). The technique presented here aims to estimate the cumulative effect of model error on short-term model forecasts. Note that this bias correction method only modifies the analysis procedure, and not the forecast model equations. While many studies explore and account for the effects of forecast model bias on global models, many fewer have investigated how to account and correct for forecast model bias when limited-area models are involved (e.g., see the review by Meng and Zhang (2011)).

4.1 Background

4.1.1 Forecast Model Biases

We test our bias correction scheme in numerical experiments with biased forecast models, similar to the forecast model biases of Baek et al. (2006). For this

discussion, and in our experiments, the "truth" model dynamics are denoted as

$$\frac{d\mathbf{x}^t}{dt} = \mathbf{M}(\mathbf{x}^t), \tag{4.1}$$

where $\mathbf{M}$ represents the truth model dynamics, and the vector $\mathbf{x}^t$ represents the truth model state vector. We implement forecast model error by using forecast model dynamics that is a modified version of the "truth" dynamics, so that the forecast model obeys

$$\frac{d\mathbf{x}}{dt} = \mathbf{M}(\mathbf{x}) + \boldsymbol{\beta}. \tag{4.2}$$

The vector $\mathbf{x}$ here represents the model state estimate, and the vector $\boldsymbol{\beta}$ is the N-dimensional "model error."

4.1.2 Forecast Model Bias Correction

This chapter adapts the composite state method to correct for forecast model errors. Baek et al. (2006) presented a series of "bias models" to account for forecast bias resulting from imperfect forecast model dynamics in ensemble forecasting systems. Data assimilation is performed on a chosen bias model to adaptively estimate both the model state and forecast bias. This is accomplished by augmenting the state vector $\mathbf{x}$ with a vector of "bias corrections."

Here, we consider one of the bias models proposed by Baek et al. (2006). In particular, this bias model assumes that, after spin-up, forecasts are approximately initialized at time t_{m-1} to the truth state at t_{m-1}, $\mathbf{x}^t_{m-1}$. Forecast errors then arise because of deviations between the truth and forecast model dynamics. Using data

assimilation, the vector $\mathbf{b}$ is estimated so that the truth at time t_m is estimated by

$$\mathbf{x}_m^t \approx \mathbf{F}(\mathbf{x}_{m-1}^t) + \mathbf{b}_m^b, \tag{4.3}$$

where $\mathbf{F}$ forecasts from time t_{m-1} to t_m, using the forecast model in eq. (4.2). Data assimilation is performed on the corrected forecasted state, $\mathbf{x}_m^b = \mathbf{F}(\mathbf{x}_{m-1}^a) + \mathbf{b}_m^b$ and the current bias correction estimate $\mathbf{b}_m^b$. This yields updated analysis estimates $\mathbf{x}_m^a$ and $\mathbf{b}_m^a$, which may then be forecast to the next analysis time, t_{m+1}. The vector of bias corrections is forecasted to time t_{m+1} via $\mathbf{b}_{m+1}^b = \mathbf{G}^b(\mathbf{b}_m^a)$. The exact form of the bias correction time evolution operator $\mathbf{G}^b$ depends upon the properties of the detected forecast errors, and may be empirically tuned as necessary.

4.1.3 Composite State Forecasting

We account for forecast model bias in a coupled system of global and limited-area models. Data assimilation is performed on a background composite state vector, which is assumed to provide an optimal state estimate. This composite state is formed by combining information from the global and limited-area model forecasts, and is essentially created from the highest spatial resolution forecast information available at a given geographical location. Upon completion of the analysis procedure, the analysis composite state estimate $\mathbf{x}_m^a$ is forecast from analysis time t_m to time t_{m+1} using

$$\mathbf{x}_{m+1}^b = \mathbf{F}(\mathbf{x}_m^a). \tag{4.4}$$

The operator $\mathbf{F}$ initializes the global and limited-area models from $\mathbf{x}^a_m$, runs those models (each of which takes the form of eq. (4.2)) from time t_m to t_{m+1}, and from these forecasts forms the composite state $\mathbf{x}^b_{m+1}$. Chapter 3 gives additional details regarding the construction of the composite state vector and the form of $\mathbf{F}$ used below.

4.2 Composite State Bias Correction

We consider coupled data assimilation performed between limited-area and global model states. When forecast models and data assimilation are coupled, errors in either of the forecast models can affect both limited-area and global model forecasts. Instead of accounting for model error separately in each of these forecast models, we consider errors in the composite state forecast which represent cumulative effects of limited-area and global forecast model biases.

We approximate the effect of errors in the composite state forecast model, $\mathbf{F}$ in eq. (4.4), with an additive correction, using the composite state vector when applying eq. (4.3). Here $\mathbf{x}^t_m$ represents the truth at time m, and both $\mathbf{b}^b_m$ and $\mathbf{x}^t_m$ have the same spatial resolution as the composite state. As we shall see, correcting the forecast model biases of the composite state system allows the composite state, the best estimate of the true system state, to be dramatically more accurate in the presence of forecast model error than it would otherwise be.

4.3 Experimental Details

4.3.1 The Lorenz Models

Our numerical experiments use the one-dimensional chaotic models of Lorenz (Lorenz, 2005). The dynamics of Lorenz's Model II and III are given, at grid point n, by

$$\frac{dZ^n}{dt} = [Z, Z]^{K,n} - Z^n + F, \text{ (Model II)} \tag{4.5a}$$

$$\frac{dZ^n}{dt} = [X, X]^{K,n} + b^2[Y, Y]^{1,n} + c[Y, X]^{1,n} - X^n - bY^n + F. \text{ (Model III)} \tag{4.5b}$$

Here, the fields X and Y in eq. (4.5b) represent the long and short-scale components of the state variable Z, with $Z = X + Y$. See Lorenz (2005) for formulas defining X and Y as functions of Z, and the definition of the square bracket notation, which represents advective coupling over a length scale of K grid points.

For our numerical experiments, Model II describes the global model dynamics, and Model III describes both limited-area and truth model dynamics. The global and limited-area models are defined on different subsets of the $N = 960$ grid point lattice on which the truth model state is defined. The grid points of the truth lattice are indexed from 0, and all model state values are referenced by their index on the truth grid. The global model state is defined on 240 grid points, corresponding to every fourth point of the $N = 960$ truth grid, and the limited-area model is defined on a subset of the truth model grid, over the grid point interval $[240, 720]$.

The "true" state is generated using Model III with $b = 10$, $c = 0.6$, $F = 15$, and $K = 32$. The bias $\boldsymbol{\beta}$ that we introduce below into the global and limited-area models is added to F. The global model uses $K = 8$ to match the length scale of $K = 32$ used in Model III. Below, we use $\mathbf{x}_m$ to represent the values of Z at time m at the composite state grid points, consisting of every grid point in $[240, 720]$ and every 4th grid point elsewhere.

4.3.2 Data Assimilation

In our numerical experiments, data assimilation is performed on a 32-member composite state ensemble using the Local Ensemble Transform Kalman Filter (Ott et al., 2004; Hunt et al., 2007), with a local analysis patch size of 81 grid points. To prevent filter divergence, constant multiplicative covariance inflation is utilized (Anderson and Anderson, 1999). We achieve optimal results when composite state and bias correction parameter ensembles are inflated by different amounts. For the results with spatially constant bias reported below, we find that composite state bias correction works best when inflating the covariances of the composite state and bias correction ensembles by 4% and 7% each cycle, respectively. When the bias was spatially varying, we found inflating the state vectors and bias correction parameters by 6% and 3%, respectively, to be best.

Observations are created by adding Gaussian white noise with standard deviation 1 and mean 0 to the true state at observation locations, and are directly compared to the value of the composite state at the observation location. The nu-

merical experiments utilize a homogenous observation network, with observations generated and assimilated every 8 grid points.

4.3.3 Bias Correction

We find that evolving the bias corrections in time using weak, numerical spatial diffusion allows for larger time-steps and smaller ensemble sizes, in addition to speeding convergence, as previously reported (Baek et al., 2006). In our experiments, $\mathbf{b}$ is evolved in time via

$$\mathbf{b}^b_{m+1}(n) = (1 - 2D_b(n))\mathbf{b}^a_m(n) + D_b(n)\mathbf{b}^a_m(n-1) + D_b(n)\mathbf{b}^a_m(n+1). \qquad (4.6)$$

The parenthetical notation (n) in eq. (4.6) denotes the value of the given field at grid point n. We empirically tune the diffusion coefficient $D_b(n)$ to minimize RMS analysis error for each bias $\boldsymbol{\beta}$. The strength of diffusion changes with location n, to account for the composite state's variable resolution, as well as the possible spatial dependence $\boldsymbol{\beta}(n)$ of the bias. In all experiments, the components of the forecast model bias $\boldsymbol{\beta}(n)$ are constant in time. When global and limited-area forecast model biases are biased differently, their model biases are denoted with the subscripts g and r, respectively.

4.3.4 Error Metric

We use the root-mean square error (RMSE) of the analysis ensemble mean as a metric of analysis accuracy, and the root-mean square error of the forecast ensemble mean to measure forecast accuracy. Ensemble forecasts are initialized

from the analysis ensemble. We define $\epsilon(m,n,f) = \bar{\mathbf{x}}_m^f(n) - \mathbf{x}_m^t(n)$ as the difference at location n and time m between the f-hour lead time forecast ensemble mean and the truth, respectively. The root-mean square error averaged over time is

$$RMSE(n,f) = \left\{ \sum_{m=1}^{c} (\epsilon(m,n,f))^2 / c \right\}^{1/2}. \tag{4.7}$$

The RMSE of the analysis ensemble corresponds to $f = 0$ in eq. (4.7). Unless noted, results shown use $c = 20000$ in eq. (4.7). In eq. (4.7), $f = 0$ corresponds to the analysis RMSE.

4.4 Results and Discussion

Our first numerical experiment biases the global and limited-area models differently, with $\boldsymbol{\beta}_g(n) = -2$ and $\boldsymbol{\beta}_r(n) = -1$, respectively, in eq. (4.2). The global model is biased additionally by the different form of eq. (4.5a) relative to eq. (4.5b).

The RMSE of the analysis composite state ensemble mean, calculated for constant forecast model bias, is shown in Fig. 4.1. The composite state method is able to achieve significantly improved results when applying bias correction (blue curve) as compared to the non-bias corrected case (red curve). The bias-corrected composite state analysis corrects for both the global model bias, $\boldsymbol{\beta}_g$, and the lower resolution and imperfect global model dynamics. We note that, without bias correction, the limited-area model has substantial error near the left boundary of its domain, and this feature is not seen with bias correction. We interpret this as evidence that the bias-corrected composite state improves lateral boundary conditions to the limited-

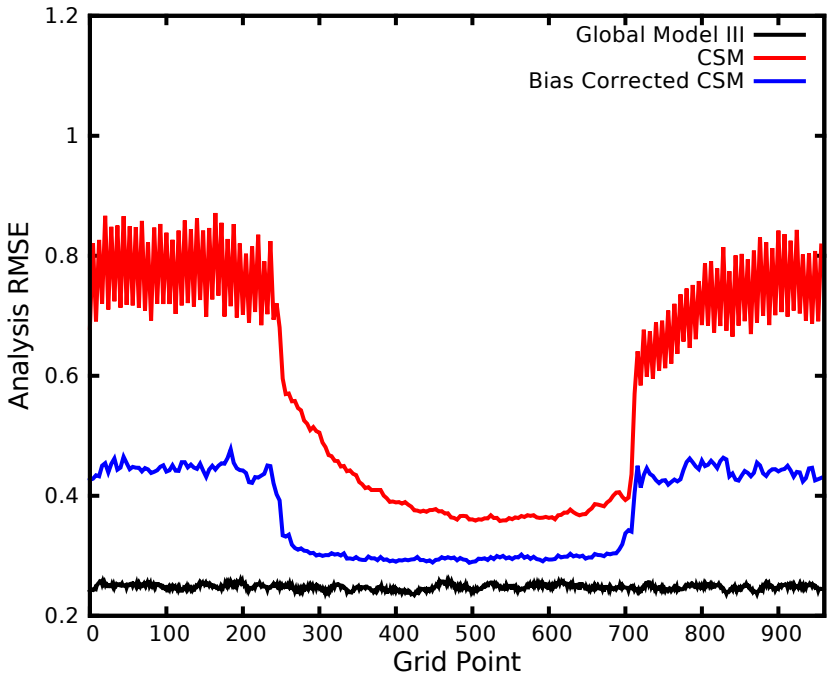

Figure 4.1: RMSE of the composite state analysis ensemble mean (eq. (4.7)). Bias correction (blue curve) significantly increases analysis accuracy compared to the analysis without bias correction (red curve), and approaches perfect (unbiased) global forecast model results (black curve).

area model, as the Lorenz models exhibit right-ward information propagation (Yoon et al., 2010).

For model bias given by eq. (4.2), we expect the bias corrections to converge to

$$\mathbf{b} \approx -\boldsymbol{\beta}\Delta t, \tag{4.8}$$

where Δt denotes the assimilation window (Baek et al., 2006), which for our experiments has the value $\Delta t = 0.05$. For the constant model biases in this experiment, eq. (4.8) predicts $\mathbf{b}$ should converge to $\mathbf{b} \approx 0.1$ where the global model is defined and $\mathbf{b} \approx 0.05$ where the limited-area model is defined. The bias correction parameters estimated in this experiment are shown in Fig. 4.2 (gold curve), and closely

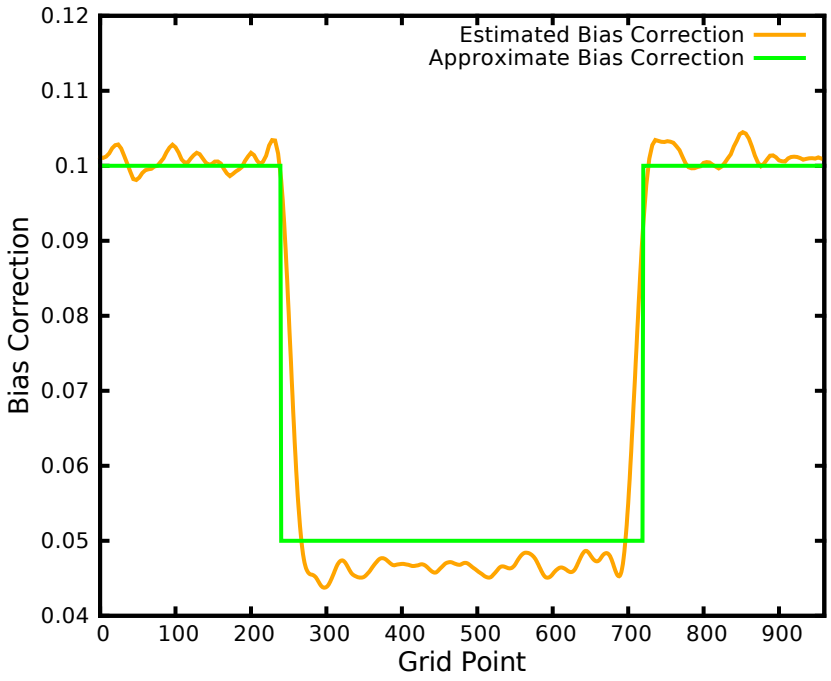

Figure 4.2: Spatial dependence of the time-averaged estimated bias correction **b** (gold curve) and the estimate provided by eq. (4.8) (green curve), for constant forecast model bias as in Fig. 4.2.

approximate the predicted asymptotic values of **b** shown in green.

In practice, it is likely that model errors will have some spatial dependence, and to investigate how well such biases may be corrected in composite state forecasts, we allow the forecast model biases to vary according to $\boldsymbol{\beta}_g(n) = \boldsymbol{\beta}_r(n) = \sin(2\pi\frac{n}{960})$. Figure 4.3 shows the RMS analysis error of the composite state when the forecast model bias varies spatially in this way. As in Fig. 4.1, the bias controlled composite state analysis (blue curve) again outperforms the uncorrected composite state analysis (red curve). Bias correction accounts for the imposed model error in $\boldsymbol{\beta}$ as well as the improper global forecast model dynamics, as evidenced by the relatively flat behavior of the analysis RMSE curve.

The increased accuracy of the bias corrected analysis implies that the effect

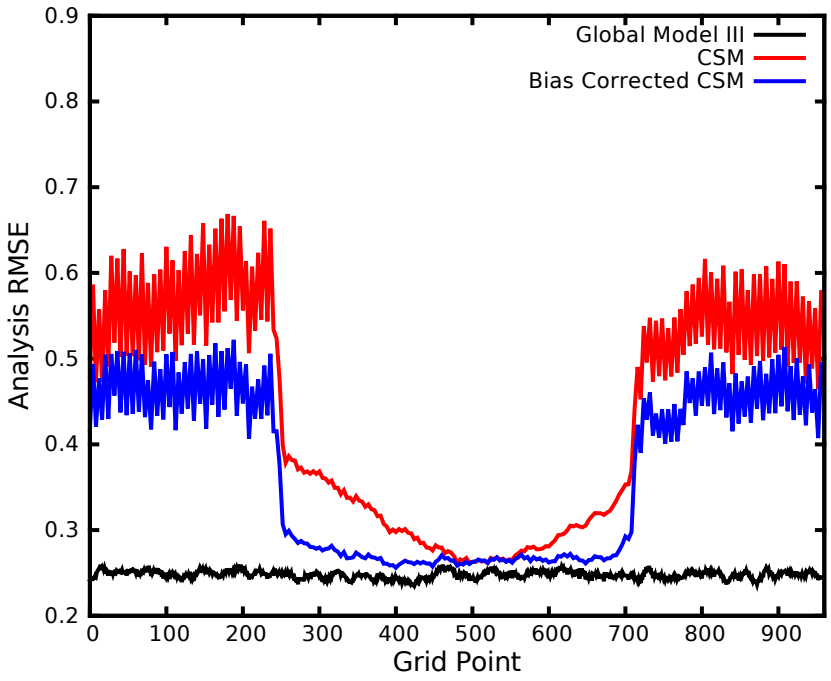

Figure 4.3: Same as Fig. 4.1, but with spatially dependent bias $\boldsymbol{\beta}$. Bias correction leads to decreased analysis RMSE compared to the composite state analysis without bias correction (blue versus red curves, respectively).

of the model bias is being accurately estimated. According to eq. (4.8), the bias corrections are expected to converge to $\mathbf{b} = -0.05\sin(2\pi\frac{n}{L})$, and this estimate is plotted in Fig. 4.4 as a green curve, along with the time-averaged value of $\mathbf{b}$ (gold curve). Their close agreement illustrates how the bias correction scheme is effective when forecast model bias is spatially varying.

Bias correction can potentially improve forecast results as well. We consider forecast lead times that are multiples of the assimilation window, Δt. Every Δt hours after initialization at time t_m, each forecast ensemble member $\mathbf{x}^F$ is adjusted according to $\mathbf{x}^F \rightarrow \mathbf{x}^F + \mathbf{b}^a_m$. Figure 4.5 shows RMSE of 2-day forecasts resulting from integrating an ensemble of global model states, initialized from the composite state analysis ensemble, while applying this methodology. Correcting the forecast

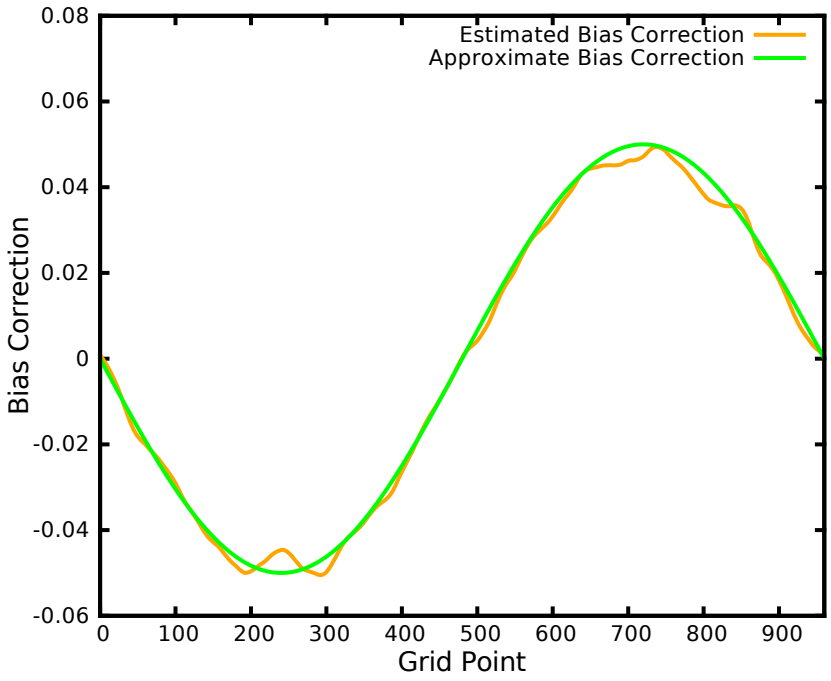

Figure 4.4: Averaged bias correction **b** (gold curve) versus the value predicted by eq. (4.8) (green curve), for spatially dependent bias, as in Fig. 4.3.

model bias, even in this rudimentary fashion, clearly improves forecast accuracy, even out as far as 2-day lead times.

Current atmospheric forecast models contain myriad possible sources of error that can contaminate forecasts. When forecasting with coupled limited-area and global forecast models, these errors can affect the output forecasts in complicated and non-trivial ways, especially when performing coupled data assimilation on limited-area and global model states. Using a simple, illustrative setup we present here evidence that an appropriate method can account for cumulative forecast model errors in a coupled forecasting system, and that our results suggest that the composite state method with bias correction may be useful for producing forecasts with imperfect model dynamics.

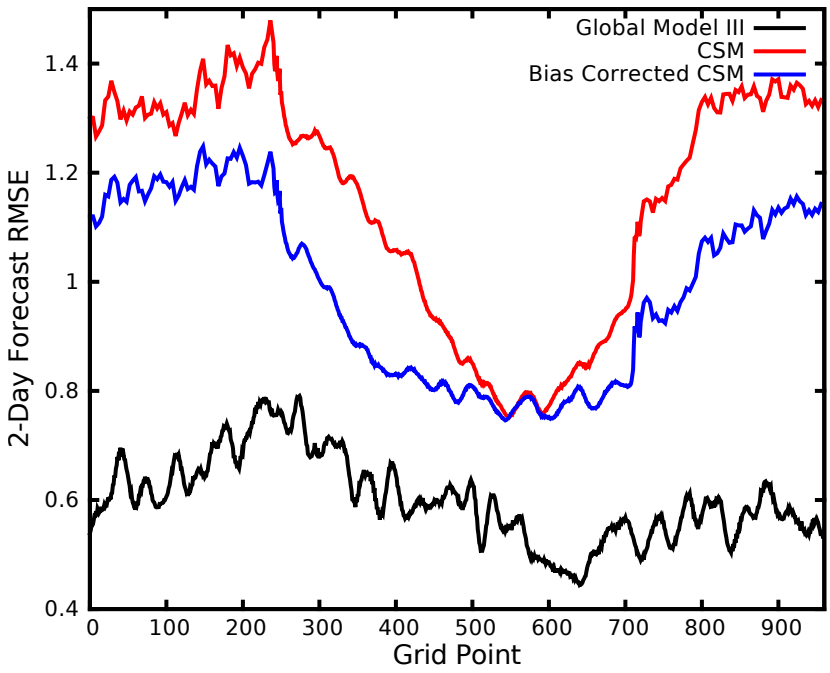

Figure 4.5: RMSE of 2-day forecast ensemble mean, for spatially dependent bias as in Fig. 4.3. Blue and red curves compare forecasts with and without bias correction, respectively. The black curve shows ensemble forecast RMSE when forecasting with a global perfect model.

References

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

Anderson, J. L. and Anderson, S. L. 1999. A Monte Carlo implementation of the nonlinear filtering problem to produce ensemble assimilations and forecasts. *Mon. Wea. Rev.* **127**, 2741—-2758.

Baek, S.-J., Hunt, B. R., Kalnay, E., Ott, E. and Szunyogh, I. 2006. Local ensemble Kalman filtering in the presence of model bias. *Tellus.* **58A**, 293–306.

Baek, S.-J., Szunyogh, I., Hunt, B. R. and Ott, E. 2009. Correcting for surface pressure background bias in ensemble-based analyses. *Mon. Wea. Rev.* **137**, 2349–2364.

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

Bonavita, M., Torrisi, L. and Marcucci, F. 2010. Ensemble data assimilation with the CNMCA regional forecasting system. *Q. J. Roy. Meteorol. Soc.* **136**, 132–145.

Bowler, N. E. and Mylne, K. R. 2009. Ensemble transform Kalman filter perturbations for a regional ensemble prediction system. *Q. J. Roy. Meteorol. Soc.* **135**, 757–766.

Burgers, G., Jan van Leeuwen, P. and Evensen, G. 1998. Analysis scheme in the ensemble Kalman filter. *Mon. Wea. Rev.* **126**, 1719–1724.

Carton, J. A., Chepurin, G., Cao, X. and Giese, B. 2000. A simple ocean data assimilation analysis of the global upper ocean 1950-95. Part I: methodology. *J. Phys. Ocean.* **30**, 294–309.

Cohn, S. E. 1997. An introduction to estimation theory. *Met. Soc. Jap.* **75**, 147–178.

Dahlgren, P. and Gustafsson, N. 2012. Assimilating host model information into a limited area model. *Tellus.* **64A**, 1–17.

Davies, H. C. 1983. Limitations of some common lateral boundary schemes used in regional NWP models. *Mon. Wea. Rev.* **111**, 1002–1012.

Dee, D. P. 2005. Bias and data assimilation. *Q. J. Roy. Meteorol. Soc.* **131**, 3323–3343.

Dee, D. P. and Da Silva, A. M. 1998. Data assimilation in the presence of forecast bias. *Q. J. Roy. Meteorol. Soc.* **124**, 269–295.

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

Golub, G. H. and Van Loan, C. F., *Matrix Computations*, volume 3. JHU Press, 1996.

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

Guidard, V. and Fischer, C. 2008. Introducing the coupling information in a limited-area variational assimilation. *Q. J. Roy. Meteorol. Soc.* **134**, 723–735.

Hamill, T. M. and Snyder, C. 2000. A hybrid ensemble Kalman filter-3D variational analysis scheme. *Mon. Wea. Rev.* **128**, 2905–2919.

Hamill, T. M. and Whitaker, J. S. 2005. Accounting for the error due to unresolved scales in ensemble data assimilation: A comparison of different approaches. *Mon. Wea. Rev.* **133**, 11, 3132–3147.

Hamill, T. M., Whitaker, J. S. and Snyder, C. 2001. Distance-dependent filtering of background error covariance estimates in an ensemble Kalman filter. *Mon. Wea. Rev.* **129**, 11, 2776–2790.

Harris, L. M. and Durran, D. R. 2010. An idealized comparison of one-way and two-way grid nesting. *Mon. Wea. Rev.* **138**, 2174—2187.

Holt, C. R., Szunyogh, I. and Gyarmati, G. 2013. Can a moderate-resolution limited-area data assimilation system add value to the global analysis of tropical cyclones? *Mon. Wea. Rev.* **141**, 1866–1883.

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

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

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

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

Jazwinski, A. H., *Stochastic processes and filtering theory*. Academic Press, Inc., San Diego, 1970.

Jazwinski, A. H., *Stochastic processes and filtering theory*. Courier Corporation, 2007.

Kalman, R. E. 1960. A new approach to linear filtering and prediction problems. *Journal of Fluids Engineering* **82**, 35–45.

Keppenne, C., Rienecker, M., Kurkowski, N., Adamec, D. et al. 2005. Ensemble Kalman filter assimilation of temperature and altimeter data with bias correction and application to seasonal prediction. *Non. Proc. Geophys.* **12**, 491–503.

Klocke, D. and Rodwell, M. 2014. A comparison of two numerical weather prediction methods for diagnosing fast-physics errors in climate models. *Q. J. Roy. Meteorol. Soc.* **140**, 517–524.

Lange, H. and Craig, G. C. 2014. The impact of data assimilation length scales on analysis and prediction of convective storms. *Mon. Wea. Rev.* **142**, 3781–3808.

Li, H., Kalnay, E., Miyoshi, T. and Danforth, C. M. 2009. Accounting for model errors in ensemble data assimilation. *Mon. Wea. Rev.* **137**, 3407–3419.

Lorenc, A. C. 2003. The potential of the ensemble Kalman filter for NWP – A comparison with 4D-Var. *Q. J. Roy. Meteorol. Soc.* **129**, 3183–3203.

Lorenz, E. N. 2005. Designing chaotic models. *J. Atmos. Sci.* **62**, 1574–1587.

Meng, Z. and Zhang, F. 2011. Limited-area ensemble-based data assimilation. *Mon. Wea. Rev.* **139**, 2025–2045.

Merkova, D., Szunyogh, I. and Ott, E. 2011. Strategies for coupling global and limited-area ensemble Kalman filter assimilation. *Nonlin. Proc. Geophys.* **18**, 415–430.

Ott, E., Hunt, B. R., Szunyogh, I., Zimin, A. V., Kostelich, E. J. et al. 2004. A local ensemble Kalman filter for atmospheric data assimilation. *Tellus.* **56A**, 415–428.

Parrish, D. F. and Derber, J. C. 1992. The National Meteorological Center's spectral statistical-interpolation analysis system. *Mon. Wea. Rev.* **120**, 1747–1763.

Patil, D. J., Hunt, B. R., Kalnay, E., Yorke, J. A. and Ott, E. 2001. Local low dimensionality of atmospheric dynamics. *Phys. Rev. Lett.* **86**, 5878–5881.

Pielke Sr, R. A., *Mesoscale meteorological modeling*, volume 98. Academic press, 2013.

Reich, H., Rhodin, A. and Schraff, C. 2011. LETKF for the nonhydrostatic regional model COSMO-DE. *COSMO Newsletter* **11**, 27–31.

Torn, R., Hakim, G. and Snyder, C. 2006. Boundary conditions for limited-area ensemble Kalman filters. *Mon. Wea. Rev.* **134**, 2490–2502.

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

Wang, X., Hamill, T. M., Whitaker, J. S. and Bishop, C. H. 2009. A comparison of the hybrid and EnSRF analysis schemes in the presence of model errors due to unresolved scales. *Mon. Wea. Rev.* **137**, 10, 3219–3232.

Wang, X., Parrish, D., Kleist, D. and Whitaker, J. 2013. GSI 3DVar-based ensemble–variational hybrid data assimilation for NCEP Global Forecast System: Single-resolution experiments. *Mon. Wea. Rev.* **141**, 4098–4117.

Wang, X., Snyder, C. and Hamill, T. M. 2007. On the theoretical equivalence of differently proposed ensemble-3DVAR hybrid analysis schemes. *Mon. Wea. Rev.* **135**, 222–227.

Warner, T. T., Peterson, R. A. and Treadon, R. E. 1997. A tutorial on lateral boundary conditions as a basic and potentially serious limitation to regional numerical weather prediction. *Bull. Amer. Met. Soc.* **78**, 2599–2617.

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

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

Yang, S.-C., Kalnay, E., Hunt, B. and E Bowler, N. 2009. Weight interpolation for efficient data assimilation with the local ensemble transform kalman filter. *Q. J. Roy. Meteorol. Soc.* **135**, 251–262.

Yoon, Y.-N., Hunt, B. R., Ott, E. and Szunyogh, I. 2012. Simultaneous global and limited-area ensemble data assimilation using joint states. *Tellus.* **64A**, 18407.

Yoon, Y.-N., Ott, E. and Szunyogh, I. 2010. On the propagation of information and the use of localization in ensemble Kalman filtering. *J. Atmos. Sci.* **67**, 3823–3834.