ABSTRACT

Title of dissertation: ENSEMBLE KALMAN FILTER EXPERIMENTS WITH A PRIMITIVE-EQUATION GLOBAL MODEL

Takemasa Miyoshi, Doctor of Philosophy, 2005

Dissertation directed by: Professor Eugenia Kalnay
Department of Meteorology

The ultimate goal is to develop a path towards an operational ensemble Kalman filtering (EnKF) system. Several approaches to EnKF for atmospheric systems have been proposed but not systematically compared. The sensitivity of EnKF to the imperfections of forecast models is unclear. This research explores two questions: 1. What are the relative advantages and disadvantages of two promising EnKF methods? 2. How large are the effects of model errors on data assimilation, and can they be reduced by model bias correction?

Chapter 2 contains a theoretical review, followed by the FORTRAN development and testing of two EnKF methods: a serial ensemble square root filter (serial EnSRF, Whitaker and Hamill 2002) and a local EnKF (LEKF, Ott et al. 2002; 2004). We reproduced the results obtained by Whitaker and Hamill (2002) and Ott et al. (2004) on the Lorenz (1996) model. If we localize the LEKF error covariance, LEKF outperforms serial EnSRF. We also introduce a method to objectively estimate the optimal covariance inflation.

In Chapter 3 we apply the two EnKF methods and the three-dimensional variational method (3DVAR) to the SPEEDY primitive-equation global model (Molteni 2003), a fast but relatively realistic model. Perfect model experiments show that EnKF greatly

outperforms 3DVAR. The 2-day forecast "errors of the day" are very similar to the analysis errors, but they are not similar among different methods except in low ensemble dimensional regions. Overall, serial EnSRF outperforms LEKF, but their difference is substantially reduced if we localize the LEKF error covariance or increase the ensemble size. Since LEKF is much more efficient than serial EnSRF when using parallel computers and many observations, LEKF would be the only feasible choice in operations.

In Chapter 4 we remove the perfect model assumption using the NCEP/NCAR reanalysis as the "nature" run. The advantage of EnKF to 3DVAR is reduced. When we apply the model bias estimation proposed by Dee and da Silva (1998), we find that the full dimensional model bias estimation fails. However, if instead we assume that the bias is low dimensional, we obtain a substantial improvement in the EnKF analysis.

ENSEMBLE KALMAN FILTER EXPERIMENTS WITH A PRIMITIVE-EQUATION GLOBAL MODEL

by

Takemasa Miyoshi

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
2005

Advisory Commmittee:

Professor Eugenia Kalnay, Chair/Advisor
Professor James Carton
Professor Owen Thompson
Professor Brian Hunt
Dr. Joaquim Ballabrera

ACKNOWLEDGMENTS

First of all, I express my deep gratitude to Prof. Eugenia Kalnay for her advising and supports during the entire processes of my graduate study. She always supported me both academically and mentally, which was indispensable for me to complete this hard task.

Secondly, I thank the Japanese Government, especially the National Personnel Authority, the Ministry of Land, Infrastructure and Transport, and the Japan Meteorological Agency, for supporting my graduate study through the Japanese Government long-term fellowship program.

Thirdly, I truly appreciate Drs. Franco Molteni and Fred Kucharski for kindly having provided the source codes and documents for the SPEEDY model. Their help was essential in the present research.

Many thanks go to Dr. Joaquim Ballabrera and Christopher Danforth for their useful comments and suggestions on the drafts of this thesis. I thank committee members: Prof. Eugenia Kalnay, Prof. James Carton, Prof. Owen Thompson, Prof. Brian Hunt, and Dr. Joaquim Ballabrera for their support.

I am grateful to Prof. Edward Ott and Dr. Istvan Szunyogh, for insightful discussions and kind support. John Harlim has developed LETKF based on the SPEEDY model, corroborating my results of LEKF. I thank students in the chaos group's weather project, especially Dr. Shu-Chih Yang, John Harlim, Christopher Danforth, Junjie Liu, and Hong Li, for many useful discussions. I thank Tomohide Higuchi for helpful discus-

sions on some mathematical details of the recursive filter theory. Kohei Aranami kindly volunteered to let me use his powerful personal computer remotely. Some of the analysis/forecast computations have been done on his computer. I am grateful to all my friends who have supported my life in the U.S. in many ways.

Lastly, I thank my wife Seiko and my son Tetsushi for their support and patience to encourage me even in living separately during my stay away from Japan.

Without tremendous help and support, I could not have achieved the hard task to complete this thesis in the two-year graduate study. My sincere thanks for all the kindness and supports are truly beyond description.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Introduction

1.1 Background

In 2004, ten typhoons hit Japan, bringing tremendous damages to the local economy and people's lives. This was the worst typhoon season in 50 years. Better preparation for such extreme weather would minimize damage and save lives. In order for people to prepare well in advance of extreme events, it is important to predict typical weather events that are directly connected to people's lives as early and precisely as possible.

Modern weather forecasting is greatly dependent on numerical weather prediction (NWP). By integrating momentum equations of the atmosphere with physical processes, we can predict future state of the atmosphere. Since the problem is mathematically an initial value problem, forecasts require an estimate of the initial state of the atmosphere. In the early stages of NWP, well-trained meteorologists created a subjective analysis by hand to provide numerical values of the initial condition. Meteorologists would plot "synoptic" observations all over the world on a single map and draw a weather chart with constant pressure lines, constant temperature lines, etc. However, even among well-trained meteorologists, the analyzed charts are different. For high quality NWP, it is essential to generate an "objective analysis" with less uncertainty given the same amount of information.

Since the initial conditions are crucial to NWP, the objective analysis (data assimilation) method has nearly as long a history as NWP itself, about 50 years. Many data assimilation methods, such as successive corrections, optimal interpolation (OI), and the

three-dimensional variational method (3DVAR) have been used operationally as Daley (1991) and Kalnay (2003) describe in their comprehensive textbooks. NWP models have improved so much recently that they provide forecasts as precise as observations. As a result, the effect of errors in the initial condition is more important. The relative impact of introducing a new data assimilation system has increased, and data assimilation is recognized as one of the most important issues in the development of operational NWP systems. Thus, more advanced data assimilation methods such as four-dimensional variational method (4DVAR) and Kalman filtering (KF) have been explored and some have already been implemented operationally.

Modern methods of data assimilation are based on maximum likelihood estimation using all available information, namely, observations and forecasts. Observations provide direct measurements of the current state of the atmosphere, the model forecasts provide information obtained from the past. Each source has limitations. Observations are subject to errors in measurements as well as the limited number of observing stations. There is no observational instrument without errors. Even if we have a perfect instrument, a grid point value for an NWP model variable will be different from the perfect observation because the grid point value represents a mean state for the surrounding area. The difference is known as the error of representativeness. In addition, we do not have observations for all model grid points all over the world for all meteorological elements. Thus, we cannot have perfect information for the present atmospheric state. As for the model forecasts, no model that simulates Mother Nature can be perfect. In addition, the atmospheric system is nonlinear, which introduces dynamically chaotic behavior. Even if we have a perfect model, a chaotic system "forgets" past information in a relatively short time, so only limited information can be used from the past. In other words, if a perfect model were

known and the system were linear, all past observational information could be transmitted toward the present, and we could reach better and better estimates of the truth with time because of the accumulation of information. Thus, the effects of imperfect observations, an imperfect model, and the nonlinearity of the system introduce difficulties in estimating the atmospheric state.

Furthermore, an important problem in atmospheric data assimilation lies in the large number of degrees of freedom of NWP models. Even if we know a best method for data assimilation in theory, the large number of degrees of freedom can prohibit its explicit implementation in practice. Thus, a lot of simplifying assumptions need to be made in atmospheric data assimilation. The simplifications will decrease as computational capability increases.

Kalman filtering (KF, Kalman 1960) is a theory of data assimilation that optimally combines all the available information for linear systems. The large number of degrees of freedom of NWP models prohibit the direct implementation of KF. However, the recent improvement of computational capability enables its implementation with simplifications. In the ensemble formulation of Kalman filtering, a.k.a. ensemble Kalman filtering (EnKF) initially proposed by Evensen (1994), a low rank of the forecast error covariance is assumed; limited ensemble members are assumed to be able to represent forecast errors. After Evensen (1994), several methods of EnKF have been proposed, the methods can be divided into two groups: a perturbed observation (PO) method and a square root filter (SRF) method. The PO method uses an independent data assimilation cycle for each ensemble member. Unfortunately, this method requires perturbing observational data (Burgers et al. 1998), a source of sampling errors. Alternatively, SRF does not introduce such sampling errors, which is why Whitaker and Hamill (2002) suggested that the SRF is expected to

outperform the PO method.

Several SRF methods have been proposed: an ensemble transform Kalman filter (ETKF, Bishop et al. 2001), an ensemble adjustment Kalman filter (EAKF, Anderson 2001), a serial EnSRF (Whitaker and Hamill 2002), all of which are effective when observational data are assimilated serially, and a local EnKF (LEKF, Ott et al 2002; 2004), where observations are assimilated simultaneously. All of these methods were proposed and tested separately; there has been little research thus far comparing different methods of SRF. In addition, as Tippett et al. (2003) mentioned, the formulation of SRFs is not theoretically unique, and it is not clear if there is any optimal choice among different implementations.

Lastly, an important problem with EnKF in practice lies in the imperfection of NWP models. KF is optimal when the forecasting model is linear and perfect, both of which are not the case in the real NWP. The problem of nonlinearity can be treated by taking a short time interval for the data assimilation cycle if the observations are frequent enough so that the tangent linear assumption is valid. However, advanced data assimilation methods with fewer assumptions tend to be more sensitive to imperfection of models, since they are more strictly optimal under the assumptions (e.g. Miyoshi 2004a). Thus, it is very important to consider model errors in EnKF. Dee and da Silva (1998) discussed a way to treat model biases in data assimilation. Assuming that model biases are another set of prognostic variables and that observational increments, i.e. forecast minus observation, are observations of model biases, we can construct an additional data assimilation system for model biases. It is difficult to derive a time evolving model for biases, and Dee and da Silva (1998) used a "persistent" model that is just the identity.

1.2 Objectives

The ultimate goal of the present research is to develop a path towards an operational ensemble forecast-analysis system that is expected to produce substantially improved weather/climate information. EnKF provides a way to achieve an ensemble forecast-analysis system. However, as mentioned in the previous section, at least four ways to implement EnKF on NWP models have been suggested, but the relative advantages and disadvantages among them are not clear. In addition, EnKF has been tested most extensively in the perfect model scenario, there is little research testing the sensitivities of EnKF to imperfection of forecast models. Thus, to achieve the goal, it is relevant to answer two questions:

1. What are the relative advantages of different implementations of EnKF?

2. What is the effect of model errors on EnKF, and can we handle them?

Whitaker and Hamill (2002) suggested SRF is expected to outperform the PO method, thus, only SRF is investigated in the present research. In addition, SRF can be divided into two groups: serial treatment of observations and simultaneous treatment of observations. Thus, two most efficient methods of each group, a serial EnSRF (Whitaker and Hamill 2002) and LEKF (Ott et al. 2002; 2004), are chosen. The present research aims to investigate the following questions:

1. What are the relative advantages and disadvantages between a serial EnSRF and the LEKF?

2. How large are the effects of model errors on data assimilation, and how does a model bias estimation work in a chosen EnKF method?

1.3 General methodology

To investigate the questions described in the previous section, data assimilation cycle experiments on a primitive-equation global model are performed. The model used in the present research is the SPEEDY model (Simplified Parameterizations, primitivE-Equation DYnamics, by Molteni 2003). The SPEEDY model has simplified physical processes, the resolution is T30L7 corresponding to grid size $96 \times 48 \times 7$ and sigma vertical coordinate ($\sigma = p/p_s$, where p and p_s are pressure and surface pressure, respectively). The prognostic variables are horizontal wind velocity components (u, v), temperature (T), specific humidity (q), and surface pressure (p_s). Molteni (2003) has shown that the SPEEDY model has characteristics similar to state-of-the-art models, though it has larger systematic errors.

The data assimilation experiments in this thesis are based on observational systems simulation experiments (OSSEs), where a true "nature run" is assumed to be known, and observations are simulated by adding random noise to the nature run according to observational errors. To investigate the first question, we first assume a perfect model. Under the perfect model assumption, the nature run is created by the model integration after a 1-year spin-up period. Three data assimilation systems (3DVAR, serial EnSRF, and LEKF) are developed and their performances are compared. 3DVAR is representative of classical methods, and it provides a control for the performance of data assimilation. In the perfect model cases, sensitivity to the parameter values are investigated, and the parameters are adjusted. Three types of observational locations are tested; a regularly distributed dense network, a regularly distributed sparse network, and an irregularly distributed realistic rawinsonde network (few over oceans and many over land in NH). From this experiment, one of the EnKFs (serial EnSRF) is chosen for the latter experiments.

To investigate the second question, we use the NCEP/NCAR reanalysis (Kalnay et

al. 1996; Kistler et al. 2001) to generate "observations". The NCEP/NCAR reanalysis provides a best estimate of the real atmosphere using a recent operational forecast model and a 3DVAR data assimilation system. The SPEEDY model is less realistic than the 1995 NCEP model used in the reanalysis. Thus, we can expect NCEP/NCAR reanalysis fields to be an approximation representative of the true evolution of the atmosphere for the SPEEDY model. We perform the same experiments as in the perfect model case using two data assimilation systems, 3DVAR and serial EnSRF. From the experiments using the NCEP/NCAR reanalysis as "truth", the effects of the model errors become clear. We estimate model bias by including bias variables, assuming the observational increment is an observation of the model bias and the forecast model of the model bias is persistence (i.e. identity), following Dee and da Silva (1998).

1.4 Outline

Chapter 2 describes theory of data assimilation and experiments using the simple Lorenz-96 model (Lorenz 1996; Lorenz and Emanuel 1998). Section 2.2 introduces the theory of data assimilation in a self-contained manner. Data assimilation is introduced as an optimal weighted mean between forecasts and observations. Then, we introduce forecast covariance and the difference between KF and OI. Ensemble formulation is introduced, and serial EnSRF and LEKF are described. Before going to a realistic primitive-equation atmospheric model, Section 2.3 describes data assimilation experiments under the perfect model assumption using the simple Lorenz-96 model. This gives an idea of how each method works and how it is different from each other. In addition, the results ensure the program developed for the present research is correctly coded by comparing the performance with those of Whitaker and Hamill (2002) and Ott et al. (2004). In this section, the

three methods (3DVAR/OI, serial EnKF and LEKF) and full KF are implemented and tested on the Lorenz-96 model, the methods are compared to each other. Here, full KF means that all five KF equations, introduced in Section 2.2, are solved explicitly without the low-rank approximation that is essential in EnKF.

Chapter 3 discusses the first question under the perfect model assumption. In this chapter, 3DVAR, serial EnKF and LEKF are introduced separately, and implementations of each method are described. After descriptions of implementation, the results are shown and compared. Sensitivities to different parameters, observational networks, and localization types are discussed. The characteristics of the analysis and forecast errors are also discussed. For latter experiments, one EnKF method (serial EnSRF) is chosen in this chapter. The choice is based on the performance (accuracy and efficiency) that we obtained using a single computer.

In Chapter 4, we remove the perfect model assumption. Using NCEP/NCAR reanalysis fields as the nature run, we introduce model errors. First, the model error statistics are computed, accumulated by the difference between the reanalysis fields and the model forecasts initiated by the reanalysis fields. Then, data assimilation experiments assuming the reanalysis fields as truth are performed without model bias estimation, which clarifies the effects of the model errors on data assimilation. Finally, we apply the model bias estimation scheme proposed by Dee and da Silva (1998) and discuss the results.

Chapter 5 summarizes the results and gives conclusion of the present research. This chapter also describes possible future directions, based on our new perspective achieved through the present research.

Chapter 2

Theory of data assimilation methods and application on the Lorenz-96 model

2.1 Introduction

We describe a theoretical review on data assimilation methods in Section 2.2. Section 2.3 describes data assimilation experiments on a simple system (the Lorenz-96 model, Lorenz 1996; Lorenz and Emanuel 1998), providing an idea on how each method works and how it is different from each other. Here, we develop the core modules used for this entire research. The results shown in this chapter support the correctness of the Fortran programs.

2.2 Theoretical review

This section describes a review on theory of data assimilation methods that are developed in the present research. Some basics in estimation theory are briefly introduced so that the present dissertation is self-contained. For more comprehensive introduction and details of atmospheric data assimilation, see the textbooks by Daley (1991) and Kalnay (2003, Chapter 5). Bouttier and Courtier (1999) provide a more practical introduction on atmospheric data assimilation. As for Kalman filtering, Jazwinski (1970) provides a more mathematically precise introduction and discussion. Gelb et al. (1974) also provide a more comprehensive and practical introduction on Kalman filtering. Bouttier (1996) introduces Kalman filtering in the context of atmospheric data assimilation.

2.2.1 Optimal weighted mean

Data assimilation is a way to keep the model state close to the nature by assimilating observations. Modern methods of data assimilation are a maximum likelihood estimation using all available information, namely, forecasts $\mathbf{x}^f$ and observations $\mathbf{y}^o$. We look for an analysis state $\mathbf{x}^a$ by a linear weighted mean between $\mathbf{x}^f$ and $\mathbf{y}^o$:

$$\mathbf{x}^a = \mathbf{x}^f + \mathbf{K}(\mathbf{y}^o - H(\mathbf{x}^f)) \tag{2.1}$$

where $\mathbf{K}$ denotes the weight (notation follows Ide et al. 1997). Since observations cannot be taken at all grid points for all prognostic variables, the vectors $\mathbf{y}^o$ and $\mathbf{x}$ lie in different spaces. H converts prognostic variables $\mathbf{x}$ to the space of observational data $\mathbf{y}^o$ and is called an "observational operator". H is not necessarily a linear mapping.

Define differences from the true state $\mathbf{x}^t$:

$$\delta\mathbf{x}^a = \mathbf{x}^a - \mathbf{x}^t \tag{2.2}$$

$$\delta\mathbf{x}^f = \mathbf{x}^f - \mathbf{x}^t \tag{2.3}$$

$$\delta\mathbf{y}^o = \mathbf{y}^o - H(\mathbf{x}^t) \tag{2.4}$$

as residual errors of analysis, forecast, and observation, respectively. Subtracting $\mathbf{x}^t$ from the both sides of eq.(2.1), we get

$$\delta\mathbf{x}^a \approx \delta\mathbf{x}^f + \mathbf{K}(\delta\mathbf{y}^o - \mathbf{H}\delta\mathbf{x}^f) \tag{2.5}$$

where $\mathbf{H}$ denotes the Jacobian of H. The tangent linear approximation made here is accurate as long as $\delta\mathbf{x}^f$ is small. The analysis error covariance $\mathbf{P}^a$ is written as

$$\mathbf{P}^a = \left\langle \delta\mathbf{x}^a (\delta\mathbf{x}^a)^\top \right\rangle \tag{2.6}$$

where $\langle\cdot\rangle$ denotes the statistical expected value. Substituting eq.(2.5) into eq.(2.6), we

note

$$\begin{aligned}
\mathbf{P}^a &= \left\langle \left(\delta\mathbf{x}^f + \mathbf{K}(\delta\mathbf{y}^o - \mathbf{H}\delta\mathbf{x}^f)\right)\left(\delta\mathbf{x}^f + \mathbf{K}(\delta\mathbf{y}^o - \mathbf{H}\delta\mathbf{x}^f)\right)^\top \right\rangle \\
&= \left\langle (\mathbf{I}-\mathbf{KH})\delta\mathbf{x}^f(\delta\mathbf{x}^f)^\top(\mathbf{I}-\mathbf{KH})^\top \right\rangle + \left\langle \mathbf{K}\delta\mathbf{y}^o(\delta\mathbf{y}^o)^\top\mathbf{K}^\top \right\rangle \\
&= (\mathbf{I}-\mathbf{KH})\mathbf{P}^f(\mathbf{I}-\mathbf{KH})^\top + \mathbf{KRK}^\top \qquad (2.7)
\end{aligned}$$

where the forecast error covariance $\mathbf{P}^f$ and observational error covariance $\mathbf{R}$ are defined as

$$\mathbf{P}^f = \left\langle \delta\mathbf{x}^f(\delta\mathbf{x}^f)^\top \right\rangle \qquad (2.8)$$

$$\mathbf{R} = \left\langle \delta\mathbf{y}^o(\delta\mathbf{y}^o)^\top \right\rangle \qquad (2.9)$$

and the forecast error and observational error are assumed to be uncorrelated:

$$\left\langle \delta\mathbf{x}^f(\delta\mathbf{y}^o)^\top \right\rangle = 0 \qquad (2.10)$$

We assume the error probability distribution function (PDF) is Gaussian. Then the maximum likelihood state minimizes the total analysis error variance. That is, the optimal weight gives the minimum of the trace of $\mathbf{P}^a$. Thus,

$$\frac{\partial}{\partial\mathbf{K}}\left(trace(\mathbf{P}^a)\right) = 0 \qquad (2.11)$$

Using the formulas

$$\frac{\partial}{\partial\mathbf{X}}\left(trace(\mathbf{XYX}^\top)\right) = \mathbf{X}(\mathbf{Y}+\mathbf{Y}^\top) \qquad (2.12)$$

$$\frac{\partial}{\partial\mathbf{X}}\left(trace(\mathbf{XY})\right) = \mathbf{Y}^\top \qquad (2.13)$$

(cf. eqs.(2.1-72) and (2.1-73) of Gelb et al. 1974), we can transform eq.(2.11) with eq.(2.7) as

$$-2(\mathbf{I}-\mathbf{KH})\mathbf{P}^f\mathbf{H}^\top + 2\mathbf{KR} = 0 \qquad (2.14)$$

Here, we used the fact that covariance matrices $\mathbf{P}^f$ and $\mathbf{R}$ are symmetric. By solving eq.(2.14) for $\mathbf{K}$, we get the optimal weight (a.k.a. Kalman gain)

$$\mathbf{K} = \mathbf{P}^f\mathbf{H}^\top(\mathbf{H}\mathbf{P}^f\mathbf{H}^\top + \mathbf{R})^{-1} \tag{2.15}$$

For simplicity, define $\mathbf{S}$ as

$$\mathbf{S} = \mathbf{H}\mathbf{P}^f\mathbf{H}^\top + \mathbf{R} \tag{2.16}$$

Substituting eq.(2.15) into eq.(2.7), we get the analysis error covariance

$$\begin{aligned}
\mathbf{P}^a &= (\mathbf{I} - \mathbf{P}^f\mathbf{H}^\top\mathbf{S}^{-1}\mathbf{H})\mathbf{P}^f(\mathbf{I} - \mathbf{P}^f\mathbf{H}^\top\mathbf{S}^{-1}\mathbf{H})^\top + \mathbf{P}^f\mathbf{H}^\top\mathbf{S}^{-1}\mathbf{R}(\mathbf{P}^f\mathbf{H}^\top\mathbf{S}^{-1})^\top \\
&= \mathbf{P}^f - 2\mathbf{P}^f\mathbf{H}^\top\mathbf{S}^{-1}\mathbf{H}\mathbf{P}^f + \mathbf{P}^f\mathbf{H}^\top\mathbf{S}^{-1}\mathbf{H}\mathbf{P}^f\mathbf{H}^\top\mathbf{S}^{-1}\mathbf{H}\mathbf{P}^f + \mathbf{P}^f\mathbf{H}^\top\mathbf{S}^{-1}\mathbf{R}\mathbf{S}^{-1}\mathbf{H}\mathbf{P}^f \\
&= \mathbf{P}^f - 2\mathbf{P}^f\mathbf{H}^\top\mathbf{S}^{-1}\mathbf{H}\mathbf{P}^f + \mathbf{P}^f\mathbf{H}^\top\mathbf{S}^{-1}(\mathbf{H}\mathbf{P}^f\mathbf{H}^\top + \mathbf{R})\mathbf{S}^{-1}\mathbf{H}\mathbf{P}^f \\
&= \mathbf{P}^f - 2\mathbf{P}^f\mathbf{H}^\top\mathbf{S}^{-1}\mathbf{H}\mathbf{P}^f + \mathbf{P}^f\mathbf{H}^\top\mathbf{S}^{-1}\mathbf{S}\mathbf{S}^{-1}\mathbf{H}\mathbf{P}^f \\
&= \mathbf{P}^f - \mathbf{P}^f\mathbf{H}^\top(\mathbf{H}\mathbf{P}^f\mathbf{H}^\top + \mathbf{R})^{-1}\mathbf{H}\mathbf{P}^f \\
&= (\mathbf{I} - \mathbf{K}\mathbf{H})\mathbf{P}^f
\end{aligned} \tag{2.17}$$

Thus, the analysis error covariance is the forecast error covariance decreased by the factor $(\mathbf{I} - \mathbf{K}\mathbf{H})$.

2.2.2 Forecast error covariance

A forecast model M gives a mapping from an analysis state at time i ($\mathbf{x}_i^a$) to a forecast state at time $i+1$ ($\mathbf{x}_{i+1}^f$)

$$\mathbf{x}_{i+1}^f = M(\mathbf{x}_i^a) \tag{2.18}$$

Define the mapping that gives the true time evolution to be M^t. Then

$$\mathbf{x}_{i+1}^t = M^t(\mathbf{x}_i^t) \tag{2.19}$$

A common assumption is that the model M contains random error

$$\mathbf{x}_{i+1}^t = M(\mathbf{x}_i^t) - \eta \tag{2.20}$$

where the random process η has no bias but has covariance $\mathbf{Q}$, that is,

$$\langle \eta \rangle = 0 \tag{2.21}$$

$$\left\langle \eta\eta^\top \right\rangle = \mathbf{Q} \tag{2.22}$$

Then, the forecast error covariance matrix is written as

$$\begin{aligned}
\mathbf{P}_i^f &= \left\langle \delta\mathbf{x}_i^f (\delta\mathbf{x}_i^f)^\top \right\rangle \\
&= \left\langle (\mathbf{x}_i^f - \mathbf{x}_i^t)(\mathbf{x}_i^f - \mathbf{x}_i^t)^\top \right\rangle \\
&= \left\langle \left(M(\mathbf{x}_{i-1}^a) - M^t(\mathbf{x}_{i-1}^t)\right) \left(M(\mathbf{x}_{i-1}^a) - M^t(\mathbf{x}_{i-1}^t)\right)^\top \right\rangle \\
&= \left\langle \left(M(\mathbf{x}_{i-1}^a) - M(\mathbf{x}_{i-1}^t) + \eta\right) \left(M(\mathbf{x}_{i-1}^a) - M(\mathbf{x}_{i-1}^t) + \eta\right)^\top \right\rangle \\
&\simeq \left\langle \left(\mathbf{M}\delta\mathbf{x}_{i-1}^a + \eta\right) \left(\mathbf{M}\delta\mathbf{x}_{i-1}^a + \eta\right)^\top \right\rangle \\
&= \left\langle \mathbf{M}\delta\mathbf{x}_{i-1}^a (\mathbf{M}\delta\mathbf{x}_{i-1}^a)^\top + \mathbf{M}\delta\mathbf{x}_{i-1}^a \eta^\top + \eta(\mathbf{M}\delta\mathbf{x}_{i-1}^a)^\top + \eta\eta^\top \right\rangle \\
&= \left\langle \mathbf{M}\delta\mathbf{x}_{i-1}^a (\delta\mathbf{x}_{i-1}^a)^\top \mathbf{M}^\top \right\rangle + \left\langle \eta\eta^\top \right\rangle \\
&= \mathbf{M}\mathbf{P}_{i-1}^a \mathbf{M}^\top + \mathbf{Q}
\end{aligned} \tag{2.23}$$

where $\mathbf{M}$ (the Jacobian of M) is called the tangent linear model (TLM). The transpose of TLM ($\mathbf{M}^\top$) is called the adjoint model (ADJ). The tangent linear approximation is valid as long as $\delta\mathbf{x}^a$ is small. In the derivation of eq.(2.23), the analysis error $\delta\mathbf{x}^a$ and the random error η are assumed to be uncorrelated, that is,

$$\left\langle \delta\mathbf{x}^a \eta^\top \right\rangle = 0 \tag{2.24}$$

The five equations (2.1), (2.15), (2.17), (2.18), and (2.23) constitute the Kalman filtering algorithm. The optimal weight $\mathbf{K}$ is called the Kalman gain matrix in Kalman

filtering. In optimal interpolation (OI), eqs.(2.1) and (2.15) are used with $\mathbf{P}^f$ constant in time. Usually, the time-independent forecast error covariance used in OI is denoted $\mathbf{B}$ and is obtained from a long-term statistical mean.

2.2.3 Ensemble formulation

The covariance matrix $\mathbf{P}$ of size $N \times N$ is defined as

$$\mathbf{P} = \left\langle \delta\mathbf{x}(\delta\mathbf{x})^\top \right\rangle \tag{2.25}$$

$$\simeq \frac{1}{n-1}\sum_{l=1}^{n} \delta\mathbf{x}^{(l)}(\delta\mathbf{x}^{(l)})^\top \tag{2.26}$$

where n and l denote the number of samples and a sample index respectively, and $\delta\mathbf{x}$ is a vector of length N (the number of degrees of freedom of the model). An infinite number of samples ($n \to \infty$) is required to get a perfect estimate of $\mathbf{P}$. Since the covariance matrix $\mathbf{P}$ is real and symmetric, it has a real square root $\tilde{\mathbf{E}}$:

$$\mathbf{P} = \tilde{\mathbf{E}}\tilde{\mathbf{E}}^\top \tag{2.27}$$

Here, both $\mathbf{P}$ and $\tilde{\mathbf{E}}$ are $N \times N$ matrices. The choice of $\tilde{\mathbf{E}}$ is not unique because $\tilde{\mathbf{E}}\mathbf{U}$ also satisfies eq.(2.27), where $\mathbf{U}$ is a unitary matrix, i.e., one that satisfies $\mathbf{U}\mathbf{U}^\top = \mathbf{I}$. Eq.(2.27) indicates that a given covariance can be perfectly represented by N samples

$$\tilde{\mathbf{E}} = \frac{1}{\sqrt{N-1}}\left[\delta\mathbf{x}^{(1)} \ \cdots \ \delta\mathbf{x}^{(N)}\right] \tag{2.28}$$

Usually N is $\mathcal{O}(10^7)$ in NWP models, which prohibits the explicit estimation of the covariance matrix. However, it is common for the covariance matrix $\mathbf{P}$ to be degenerate (see for example, Dee 1995; Fukumori and Malanotte-Rizzoli 1995; Cane et al. 1996). That is, $\mathbf{P}$ has $m << N$ non-zero eigenvalues. By ignoring zero eigenvalues, the eigenvalue decomposition can be written as

$$\mathbf{P} = \mathbf{S}\mathbf{D}\mathbf{S}^\top \tag{2.29}$$

$$= \mathbf{E}\mathbf{E}^\top \tag{2.30}$$

where $\mathbf{D}$ is an $m \times m$ diagonal matrix whose components are m non-zero eigenvalues, and $\mathbf{S}$ is an $N \times m$ matrix whose columns are the corresponding eigenvectors. $\mathbf{E}$ is an $N \times m$ matrix, which can be written as

$$\mathbf{E} = \frac{1}{\sqrt{m-1}} \left[\delta\mathbf{x}^{(1)} \ \cdots \ \delta\mathbf{x}^{(m)} \right] \tag{2.31}$$

In this way, m samples express the covariance matrix $\mathbf{P}$ fairly well. As was the case for $\tilde{\mathbf{E}}$, the choice of $\mathbf{E}$ is not unique because $\mathbf{EU}$ also satisfies eq.(2.30).

Using the square root of the covariance matrices, eq.(2.23) can be written as

$$\begin{aligned} \mathbf{P}_i^f &= \mathbf{E}_i^f (\mathbf{E}_i^f)^\top \\ &= \mathbf{M}\mathbf{P}_{i-1}^a \mathbf{M}^\top \\ &= \mathbf{M}\mathbf{E}_{i-1}^a (\mathbf{E}_{i-1}^a)^\top \mathbf{M}^\top \\ &= \mathbf{M}\mathbf{E}_{i-1}^a (\mathbf{M}\mathbf{E}_{i-1}^a)^\top \end{aligned} \tag{2.32}$$

where the covariance of the model random process $\mathbf{Q}$ is ignored, i.e. we assume $M = M^t$. Thus,

$$\mathbf{E}_i^f = \mathbf{M}\mathbf{E}_{i-1}^a \tag{2.33}$$

The TLM $\mathbf{M}$ gives the time evolution of the perturbation vector $\delta\mathbf{x}$. By the definition of the TLM, the original nonlinear model can approximate eq.(2.33), thus,

$$\begin{aligned} \mathbf{M}\mathbf{E}_{i-1}^a &= \frac{1}{\sqrt{m-1}} \left[\mathbf{M}\delta\mathbf{x}_{i-1}^{a(1)} \ \cdots \ \mathbf{M}\delta\mathbf{x}_{i-1}^{a(m)} \right] \\ &\simeq \frac{1}{\sqrt{m-1}} \left[M(\mathbf{x}_{i-1}^a + \delta\mathbf{x}_{i-1}^{a(1)}) - M(\mathbf{x}_{i-1}^a) \ \cdots \ M(\mathbf{x}_{i-1}^a + \delta\mathbf{x}_{i-1}^{a(m)}) - M(\mathbf{x}_{i-1}^a) \right] \\ &\simeq \frac{1}{\sqrt{m-1}} \left[M(\mathbf{x}_{i-1}^{a(1)}) - \bar{\mathbf{x}}_i^f \ \cdots \ M(\mathbf{x}_{i-1}^{a(m)}) - \bar{\mathbf{x}}_i^f \right] \end{aligned} \tag{2.34}$$

where $\mathbf{x}^{a(l)}$ and $\bar{\mathbf{x}}^f$ denote the lth ensemble member and the ensemble mean, respectively:

$$\begin{aligned}
\mathbf{x}_i^{a(l)} &= \bar{\mathbf{x}}_i^a + \delta\mathbf{x}_i^{a(l)} && (2.35)\\
\bar{\mathbf{x}}_i^a &= \frac{1}{m}\sum_{l=1}^{m}\mathbf{x}_i^{a(l)} && (2.36)\\
\bar{\mathbf{x}}_i^f &= \frac{1}{m}\sum_{l=1}^{m}M(\mathbf{x}_{i-1}^{a(l)}) && (2.37)
\end{aligned}$$

Eq.(2.34) is equivalent to the ensemble forecast with an ensemble of size m.

In the ensemble formulation, the optimal weight for data assimilation eq.(2.15) can be written

$$\begin{aligned}
\mathbf{K} &= \mathbf{P}^f\mathbf{H}^\top(\mathbf{H}\mathbf{P}^f\mathbf{H}^\top+\mathbf{R})^{-1}\\
&= \mathbf{E}^f(\mathbf{H}\mathbf{E}^f)^\top[\mathbf{H}\mathbf{E}^f(\mathbf{H}\mathbf{E}^f)^\top+\mathbf{R}]^{-1} && (2.38)
\end{aligned}$$

Thus, we can avoid storing the entire $\mathbf{P}$ matrix. We simply store $\mathbf{E}^f$ and $\mathbf{H}\mathbf{E}^f$, which are m members of the forecast ensemble and their equivalents in observational space respectively. Since $\mathbf{H}^\top$ is not required, the nonlinear H can approximate the linearized $\mathbf{H}$ in a similar way as in the nonlinear model (eq.(2.34)):

$$\begin{aligned}
\mathbf{H}\mathbf{E}^f &= \frac{1}{\sqrt{m-1}}\left[\mathbf{H}\delta\mathbf{x}^{f(1)}\ \cdots\ \mathbf{H}\delta\mathbf{x}^{f(m)}\right]\\
&\simeq \frac{1}{\sqrt{m-1}}\left[H(\mathbf{x}^f+\delta\mathbf{x}^{f(1)})-H(\mathbf{x}^f)\ \cdots\ H(\mathbf{x}^f+\delta\mathbf{x}^{f(m)})-H(\mathbf{x}^f)\right]\\
&\simeq \frac{1}{\sqrt{m-1}}\left[H(\mathbf{x}^{f(1)})-H(\bar{\mathbf{x}}^f)\ \cdots\ H(\mathbf{x}^{f(m)})-H(\bar{\mathbf{x}}^f)\right] && (2.39)
\end{aligned}$$

Eq.(2.38) requires the inverse of a $p\times p$ matrix where p denotes the number of observations. If $m<p$, we can simplify computation of eq.(2.38) by expressing it as

$$\mathbf{K}=\mathbf{E}^f[\mathbf{I}+(\mathbf{H}\mathbf{E}^f)^\top\mathbf{R}^{-1}\mathbf{H}\mathbf{E}^f]^{-1}(\mathbf{H}\mathbf{E}^f)^\top\mathbf{R}^{-1} \qquad (2.40)$$

where the inverse of an $m\times m$ matrix is required. We assume here that different observations are uncorrelated to each other, therefore the observational error covariance matrix $\mathbf{R}$ is almost diagonal, and its inverse is trivial.

Finally, the Kalman filtering algorithm requires the computation of $\mathbf{P}^a$ to obtain $\mathbf{P}^f$ at the next analysis time. This process is equivalent to producing an appropriate analysis ensemble in the ensemble formulation or "ensemble update". There are two types of ensemble update methods:

1. a perturbed observation (PO) method

2. a square root filter (SRF)

The PO method applies an independent data assimilation cycle to each ensemble member yielding

$$\mathbf{E}^a = (\mathbf{I} - \mathbf{KH})\mathbf{E}^f \tag{2.41}$$

This gives

$$\mathbf{P}^a = (\mathbf{I} - \mathbf{KH})\mathbf{P}^f(\mathbf{I} - \mathbf{KH})^\top \tag{2.42}$$

where the analysis covariance is too small compared to eqs.(2.7) and (2.17). Thus, this process requires perturbing observational data so that we get the term $\mathbf{KRK}^\top$ in eq.(2.7) (cf. Burgers et al. 1998). Whitaker and Hamill (2002) suggested that the PO method has a disadvantage in perturbing observations because the perturbation introduces an additional source of sampling errors.

Alternatively, SRF solves eq.(2.17) directly. Assume the ensemble update is given by a linear combination of the forecast ensemble perturbations:

$$\mathbf{E}^a = \mathbf{E}^f\mathbf{T} \tag{2.43}$$

Then, eq.(2.17) yields

$$\begin{aligned} \mathbf{P}^a &= \mathbf{E}^f\mathbf{T}\mathbf{T}^\top(\mathbf{E}^f)^\top \\ &= (\mathbf{I} - \mathbf{KH})\mathbf{E}^f(\mathbf{E}^f)^\top \end{aligned} \tag{2.44}$$

By solving for $\mathbf{T}$, we get an ensemble update formula. Since the choice of $\mathbf{T}$ is not unique ($\mathbf{TU}$ also satisfies eq.(2.44)), the ensemble update formula is not unique. Among the several methods of SRF, a serial method by Whitaker and Hamill (serial EnSRF, 2002) and a batch method by Ott et al. (LEKF, 2002; 2004) are described in the present paper. Tippett et al. (2003) and Evensen (2003) also review several methods of SRF.

In summary, the ensemble formulation of Kalman filtering with SRF is realized by the following procedures:

1. Ensemble forecast provides $\mathbf{E}^f$ and $\bar{\mathbf{x}}^f$ (eq.(2.34))
2. Generate Kalman gain matrix $\mathbf{K}$ using eq.(2.38) or (2.40)
3. Compute the analysis ensemble mean $\bar{\mathbf{x}}^a$ using eq.(2.1) with $\mathbf{x}^a$ and $\mathbf{x}^f$ replaced by $\bar{\mathbf{x}}^a$ and $\bar{\mathbf{x}}^f$
4. Get the analysis ensemble perturbation $\mathbf{E}^a$ by an ensemble update formula
5. Get the analysis ensemble members $\mathbf{x}^{a(l)}$ using eq.(2.35)
6. The analysis ensemble members constitute initial conditions for the next ensemble forecast; repeat the processes

2.2.4 Serial ensemble square root filtering (serial EnSRF)

The serial EnSRF (Whitaker and Hamill 2002) assumes an ensemble update of the form

$$\mathbf{E}^a = (\mathbf{I} - \tilde{\mathbf{K}}\mathbf{H})\mathbf{E}^f \tag{2.45}$$

In order for eq.(2.45) to give a solution of eq.(2.17),

$$(\mathbf{I} - \tilde{\mathbf{K}}\mathbf{H})\mathbf{P}^f(\mathbf{I} - \tilde{\mathbf{K}}\mathbf{H})^\top = (\mathbf{I} - \mathbf{K}\mathbf{H})\mathbf{P}^f \tag{2.46}$$

has to be satisfied (cf. eq.(9) in Whitaker and Hamill 2002). In this algorithm, data assimilation is performed only once on the ensemble mean using eq.(2.1) with the optimal weight eq.(2.38), eq.(2.45) is used for the ensemble update. The solution of $\tilde{\mathbf{K}}$ is given by

$$\tilde{\mathbf{K}} = \mathbf{P}^f\mathbf{H}^\top \left[(\mathbf{H}\mathbf{P}^f\mathbf{H}^\top + \mathbf{R})^{-1/2}\right]^\top \left[(\mathbf{H}\mathbf{P}^f\mathbf{H}^\top + \mathbf{R})^{1/2} + \mathbf{R}^{1/2}\right]^{-1} \tag{2.47}$$

(Andrews 1968), cf. eq.(10) in Whitaker and Hamill (2002). This formula allows for correlated observations. If observations are uncorrelated ($\mathbf{R}$ is diagonal) observations are assimilated one at a time (serially), making the terms $\mathbf{H}\mathbf{P}^f\mathbf{H}^\top$ and $\mathbf{R}$ scalar. In this case, eq.(2.47) can be simplified, and assuming $\tilde{\mathbf{K}} = \alpha\mathbf{K}$ where α is a scalar value, one can get the formula for α

$$\alpha = \left(1 + \sqrt{\frac{\mathbf{R}}{\mathbf{H}\mathbf{P}^f\mathbf{H}^\top + \mathbf{R}}}\right)^{-1} \tag{2.48}$$

which was first derived by Potter in 1964, cf. (13) in Whitaker and Hamill (2002). Thus, the explicit computation of $\tilde{\mathbf{K}}$ is avoided, and almost no additional computation is required. In the present research, observations are assumed to be independent of each other, which makes only the computation of eq.(2.48) necessary. Thus, $\mathbf{K}$ and $\mathbf{H}$ are vectors with N dimensions, $\mathbf{R}$ is a scalar. Observations are treated serially, hence the name "serial EnSRF".

In the ensemble formulation, we use only m ensemble members to reproduce the $N \times N$ covariance matrix $\mathbf{P}$, where m is $\mathcal{O}(10^2)$. This approximation is not usually justified because the $\mathbf{P}$ reproduced from the m members of ensemble contains large sampling errors, resulting in artificially large covariances between distant points. Thus, it is necessary to localize the influence of observations. Intuitively, covariances between distant points can be assumed to be zero, since it is reasonable to assume there is no correlation between, for example, the temperature at College Park and the wind at Tokyo. In the serial EnSRF,

localization around observations is usually done using a Schur product with a weighting function similar to Gaussian function but compactly supported (Gaspari and Cohn 1999; Hamill et al. 2001). The Gaussian-like weighting function $S \circ f$ multiplies a factor $S(r)$ to an input function f (the correlation), where r denotes a measure of distance d from the center point. In this way, the input function f is localized around the center point. The factor $S(r)$ is given by a fifth-order piecewise rational function:

$$S(r) = \begin{cases} 1 - \frac{1}{4}r^5 + \frac{1}{2}r^4 + \frac{5}{8}r^3 - \frac{5}{3}r^2 & (r \leq 1) \\ \frac{1}{12}r^5 - \frac{1}{2}r^4 + \frac{5}{8}r^3 + \frac{5}{3}r^2 - 5r + 4 - \frac{2}{3}r^{-1} & (1 < r \leq 2) \\ 0 & (2 < r) \end{cases} \tag{2.49}$$

cf. eq.(14) in Hamill et al. (2001). Here, r is defined as

$$r = \frac{d}{\sqrt{10/3}\sigma} \tag{2.50}$$

where σ denotes the characteristic length scale. Fig.2.1 shows how the weighting function looks compared to the Gaussian function, given the same length scale $\sigma = 10$. The function has a shape similar to a Gaussian function, but it becomes exactly 0 for distant points. In serial EnSRF, the Kalman gain $\mathbf{K}$, now an N dimensional vector, is localized around the observation point. In this way, we force the analysis increment to zero at distant points from the observation, assuming the components of $\mathbf{K}$ in distant points are caused by sampling errors. This method of "gain localization" is equivalent to covariance localization (cf. eq.(5) of Houtekamer and Mitchell 2001) as follows

$$\mathbf{K} = (S \circ \mathbf{P}^f)\mathbf{H}^\top \left(\mathbf{H}(S \circ \mathbf{P}^f)\mathbf{H}^\top + \mathbf{R}\right)^{-1} \tag{2.51}$$

$$\approx S \circ \left(\mathbf{P}^f\mathbf{H}^\top(\mathbf{H}\mathbf{P}^f\mathbf{H}^\top + \mathbf{R})^{-1}\right) \tag{2.52}$$

Here, we used the fact that $\mathbf{H}\mathbf{P}^f\mathbf{H}$ is just a scalar in the serial treatment of observations. If the observation is at a grid point, eq.(2.52) becomes exact.

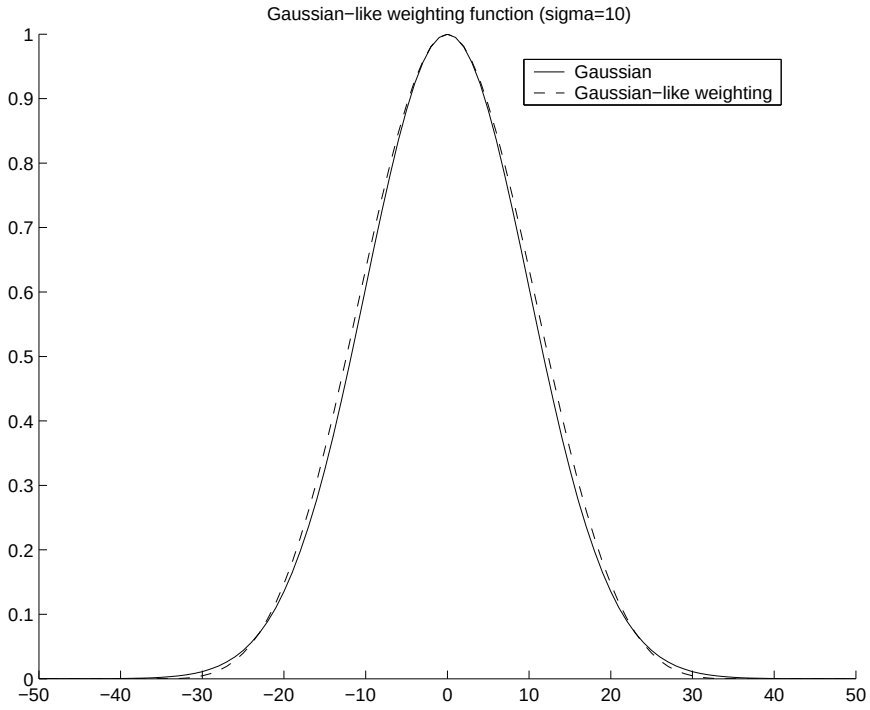

Figure 2.1: The Gaussian-like weighting function (dashed line) and Gaussian function (solid line) when length scale σ is set to 10.

2.2.5 Local ensemble Kalman filtering (LEKF)

LEKF (Ott et al. 2002; 2004; Szunyogh et al. 2005) treats the local patches surrounding every grid point independently, which automatically avoids correlations between distant points. Since the local region is much smaller than the globe, the degrees of freedom of the local patch are much smaller than those of the globe, which is preferable for the low rank assumption of the error covariance. Local patches corresponding to all grid points are analyzed totally independently, and all the analyses are combined at the end to get the global analysis. Hereafter, we consider only a local patch. m ensemble perturbations span a space with r dimensions, where $r \leq m-1$. The eigenvalue decomposition of $\mathbf{P}^f$

$$\mathbf{P}^f = \mathbf{G}\hat{\mathbf{P}}^f\mathbf{G}^\top \tag{2.53}$$

defines a variable transformation $\mathbf{G}$. $\hat{\mathbf{P}}^f$ is a diagonal matrix with just r components different from 0. Ignoring degenerate eigenvectors, we can define $\mathbf{G}$ and $\hat{\mathbf{P}}^f$ as an $N \times r$ matrix and an $r \times r$ diagonal matrix, respectively. $\mathbf{G}^\top$ gives a transformation from the N dimensional physical space to the r dimensional "hat space"

$$\hat{\mathbf{x}} = \mathbf{G}^\top\mathbf{x} \tag{2.54}$$

$$\hat{\mathbf{X}} = \mathbf{G}^\top\mathbf{X}\mathbf{G} \tag{2.55}$$

where $\mathbf{x}$ and $\mathbf{X}$ are an arbitrary N-dimensional vector and an arbitrary $N \times N$ matrix, respectively. The inverse transformation from the "hat space" to the physical space is given by

$$\mathbf{x} = \mathbf{G}\hat{\mathbf{x}} \tag{2.56}$$

In LEKF, physical vectors and matrices are transformed to the r-dimensional "hat space" where data assimilation is considered. After data assimilation processes, the analysis

ensemble in the "hat space" is transformed back into the physical space. Analysis weighted mean equations in the local patch are given by

$$\hat{\mathbf{x}}^a = \hat{\mathbf{x}}^f + \hat{\mathbf{P}}^a \hat{\mathbf{H}}^\top \mathbf{R}^{-1} \mathbf{d} \tag{2.57}$$

$$\hat{\mathbf{P}}^a = \left[(\hat{\mathbf{P}}^f)^{-1} + \hat{\mathbf{H}}^\top \mathbf{R}^{-1} \hat{\mathbf{H}} \right]^{-1} \tag{2.58}$$

$$= \hat{\mathbf{P}}^f \left[\mathbf{I} + \hat{\mathbf{H}}^\top \mathbf{R}^{-1} \hat{\mathbf{H}} \hat{\mathbf{P}}^f \right]^{-1} \tag{2.59}$$

(cf. eqs.(22) and (23) in Ott et al. 2004) which requires an inverse in the low dimensional "hat space". The ensemble update formula is given by

$$\hat{\mathbf{E}}^a = \hat{\mathbf{E}}^f \mathbf{Y} \tag{2.60}$$

cf. eq.(34) in Ott et al. (2004), where $\mathbf{Y}$ is an $m \times m$ matrix that transforms the forecast ensemble into the analysis ensemble in the "hat space". Eq.(2.60) yields

$$\hat{\mathbf{P}}^a = \hat{\mathbf{E}}^f \mathbf{Y} \mathbf{Y}^\top (\hat{\mathbf{E}}^f)^\top \tag{2.61}$$

cf. eq.(37) in Ott et al. (2004). As mentioned before, the choice of $\mathbf{Y}$ is not unique, but an optimal choice of $\mathbf{Y}$ is given by

$$\mathbf{Y} = \left[\mathbf{I} + (\hat{\mathbf{E}}^f)^\top (\hat{\mathbf{P}}^f)^{-1} (\hat{\mathbf{P}}^a - \hat{\mathbf{P}}^f) (\hat{\mathbf{P}}^f)^{-1} \hat{\mathbf{E}}^f \right]^{1/2} \tag{2.62}$$

cf. eq.(41) in Ott et al. (2004). This choice of $\mathbf{Y}$ yields the minimum difference between the forecast ensemble perturbations and the analysis ensemble perturbations. This $\mathbf{Y}$ is advantageous to the smoothness and physical balance of the analysis fields.

2.2.6 Observational error covariance localization

In the LEKF formulation described in the previous section, LEKF localizes around an analyzed point with Heaviside-like step weighting whereas serial EnSRF used Gaussian-like weighting function. Hunt (pers. comm.) suggested a Gaussian-type error covariance

localization in LEKF. Originally, all observations had the same weight in the local patch. However, when the ensemble size is very small, we expect sampling errors cannot be negligible even with the small local patch size, and thus, farther away observations may add large sampling errors. If observations are sparse compared to the local patch size, there is an unfavorable situation that only a single observation far from a local patch center is included in a patch. This may introduce a large sampling error in the analysis at the local patch center.

Thus, Hunt suggested to approximate the localization of $\mathbf{P}^f$ (which is difficult to carry out in the LEKF formulation) by weighting observational error according to the distance from the center of the local patch. Intuitively,

$$\mathbf{K} = (\rho \circ \mathbf{P}^f)\mathbf{H}^\top \left(\mathbf{H}(\rho \circ \mathbf{P}^f)\mathbf{H}^\top + \mathbf{R}\right)^{-1} \tag{2.63}$$

$$\simeq \mathbf{P}^f\mathbf{H}^\top \left(\mathbf{H}\mathbf{P}^f\mathbf{H}^\top + (\rho^{-1} \circ \mathbf{R})\right)^{-1} \tag{2.64}$$

Here, $\rho\circ$ denotes the localization function applied to a state vector or a matrix whose components have correspondence to grid points. $\rho\circ$ multiplies weights on the components of the vector or matrix according to the distance for the localization. Thus, if we want to localize the error covariance around the center of the local patch, we just multiply inverse of the covariance localization weight to the observational error variance. If we apply Gaussian-type weighting, we substitute the observational error variance R with

$$R \leftarrow R \times \exp\left(\frac{d^2}{2\sigma^2}\right) \tag{2.65}$$

where d and σ denote distance from the center of the local patch and localization scale, respectively. We call this localization "observational error covariance localization" originally suggested by Hunt (pers. comm.).

2.2.7 Covariance inflation

Kalman filtering applied in nonlinear systems usually underestimates the forecast error covariance because of the nonlinearity of the system and model errors that we neglected (eq.(2.33)). The forecast error covariance needs to be inflated for stable filtering, so that observations are given enough weight. If not, the process amplifies feedback and eventually results in filter divergence.

A simple way to inflate covariance is to multiply the covariance matrix $\mathbf{P}^f$ or $\mathbf{P}^a$ by a factor slightly larger than 1

$$\mathbf{P}^f \leftarrow \mathbf{P}^f \times (1 + \delta) \tag{2.66}$$

where δ is a small positive number ($\delta << 1$). Using the inflated covariance $\mathbf{P}^f$, observations have larger effects and we prevent filter divergence. This operation is called "multiplicative covariance inflation".

Ott et al. (2004) proposed a more sophisticated way to inflate covariance in the context of LEKF, which is called "enhanced variance inflation". The diagonal components of the analysis error covariance matrix in the "hat space" $\hat{\mathbf{P}}^a$ are inflated as follows:

$$\hat{\mathbf{P}}^a \leftarrow \hat{\mathbf{P}}^a + \frac{\delta \times trace(\hat{\mathbf{P}}^a)}{r}\mathbf{I} \tag{2.67}$$

where δ and r denote a small positive number and the rank of $\mathbf{P}^a$, respectively (cf. eq.(42) in Ott et al. 2004). In this way, the inflation magnitude depends on the total analysis error variance. In addition, the inflation is the same for perturbation directions with any uncertainty; an eigendirection corresponding to a small eigenvalue has the same inflation as an eigendirection corresponding to a large eigenvalue.

2.2.8 Variational formulation

Assuming a Gaussian shape of the error statistics, we can get the probability density function (PDF) with respect to the forecast state as

$$Prob^f(\mathbf{x}) \propto \exp\left(-\frac{1}{2}(\mathbf{x}-\mathbf{x}^f)^\top \mathbf{B}^{-1}(\mathbf{x}-\mathbf{x}^f)\right) \tag{2.68}$$

In the same way, the PDF with respect to the observation is given by

$$Prob^o(\mathbf{x}) \propto \exp\left(-\frac{1}{2}(H\mathbf{x}-\mathbf{y}^o)^\top \mathbf{R}^{-1}(H\mathbf{x}-\mathbf{y}^o)\right) \tag{2.69}$$

The joint probability is written as

$$\begin{aligned} Prob^{b\&o}(\mathbf{x}) &= Prob^b(\mathbf{x}) \cdot Prob^o(\mathbf{x}) \\ &\propto \exp\left(-\frac{1}{2}(\mathbf{x}-\mathbf{x}^b)^\top \mathbf{B}^{-1}(\mathbf{x}-\mathbf{x}^b) - \frac{1}{2}(H\mathbf{x}-\mathbf{y}^o)^\top \mathbf{R}^{-1}(H\mathbf{x}-\mathbf{y}^o)\right) \end{aligned} \tag{2.70}$$

Since data assimilation is a maximum likelihood estimate, the analysis state $\mathbf{x}$ is obtained by maximizing eq.(2.70). This is equivalent to minimizing the cost function

$$J(\mathbf{x}) = \frac{1}{2}(\mathbf{x}-\mathbf{x}^b)^\top \mathbf{B}^{-1}(\mathbf{x}-\mathbf{x}^b) + \frac{1}{2}(H\mathbf{x}-\mathbf{y}^o)^\top \mathbf{R}^{-1}(H\mathbf{x}-\mathbf{y}^o) \tag{2.71}$$

At a minimizer $\mathbf{x}^*$, the gradient of the cost function is equal to 0 , that is,

$$\nabla J(\mathbf{x}^*) = 0 \tag{2.72}$$

3DVAR algorithms look for the solution that satisfies eq.(2.72) using a quasi-Newton algorithm. The solution found by the quasi-Newton minimizer is not necessarily the global minimum of the cost function. However, since we start from a state very close to the global minimum, we assume the local minimum found coincides with the global minimum. This assumption is reasonable considering we have a 6-hour model forecast

that is close to observations, and the most likely state lies in somewhere between the forecast and observations.

In the incremental form, the cost function (eq.(2.71)) can be rewritten as

$$J(\mathbf{x}^b + \delta\mathbf{x}) \approx \frac{1}{2}\delta\mathbf{x}^\top \mathbf{B}^{-1}\delta\mathbf{x} + \frac{1}{2}(\mathbf{H}\delta\mathbf{x} - \mathbf{d})^\top \mathbf{R}^{-1}(\mathbf{H}\delta\mathbf{x} - \mathbf{d}) \tag{2.73}$$

where we defined analysis increment $\delta\mathbf{x}$ and observational increment $\mathbf{d}$ as

$$\delta\mathbf{x} = \mathbf{x} - \mathbf{x}^b \tag{2.74}$$

$$\mathbf{d} = \mathbf{y}^o - H\mathbf{x}^b \tag{2.75}$$

Eq.(2.73) is not an exact equation because of the tangent linear approximation of the generally nonlinear operator H, which is valid as long as $\delta\mathbf{x}$ is small. If the observational operator is linear, eq.(2.73) becomes exact. The gradient of eq.(2.73) is

$$\nabla J(\delta\mathbf{x}) = \mathbf{B}^{-1}\delta\mathbf{x} + \mathbf{H}^\top \mathbf{R}^{-1}(\mathbf{H}\delta\mathbf{x} - \mathbf{d}) = 0 \tag{2.76}$$

Solving eq.(2.76), we get

$$\delta\mathbf{x} = (\mathbf{B}^{-1} + \mathbf{H}^\top \mathbf{R}^{-1}\mathbf{H})^{-1}\mathbf{H}^\top \mathbf{R}^{-1}\mathbf{d} \tag{2.77}$$

$$= \mathbf{B}\mathbf{H}^\top (\mathbf{R} + \mathbf{H}\mathbf{B}\mathbf{H}^\top)^{-1}\mathbf{d} \tag{2.78}$$

which is equivalent to the weighted mean (eq.(2.1)) with the optimal weight (eq.(2.15)). Thus, the variational formulation is equivalent to the weighted mean formulation with linearized $\mathbf{H}$. Since 3DVAR uses the time-independent background error covariance $\mathbf{B}$ obtained from a long-term statistical mean, it is equivalent to OI as long as the observational operator H is linear.

2.3 Data assimilation experiments on a simple system

In this section, we apply four data assimilation methods (3DVAR, serial EnSRF, LEKF, full KF) on the Lorenz-96 model.

2.3.1 The Lorenz-96 model

The Lorenz-96 model (Lorenz 1996; Lorenz and Emanuel 1998) is defined as

$$\frac{dx_i}{dt} = x_{i-1}(x_{i+1} - x_{i-2}) - x_i + F \tag{2.79}$$

Here, $i = 1, \cdots, N$, where the boundary is cyclic, i.e. $x_{-1} = x_{N-1}$, $x_0 = x_N$, and $x_1 = x_{N+1}$. This model behaves chaotically in the case of external forcing $F = 8.0$, in which case a time increment of 0.2 non-dimensional units corresponds to about one day in terms of error growth rate (Lorenz 1996). The first term of the right hand side simulates advection, and this model can be regarded as the time evolution of an arbitrary one-dimensional quantity on a constant latitude circle, that is, the subscript i corresponds to longitude. As in Lorenz (1996), we choose $N = 40$ and $F = 8$. The time step is chosen as $\Delta t = 0.01$, and the time differencing is solved using a 4th-order Runge-Kutta scheme.

2.3.2 Implementation of data assimilation methods

With a linear observational operator, 3DVAR and OI are equivalent, that is, they treat the same problem with different approach as described in Section 2.2.8. Since the Lorenz-96 model has just 40 degrees of freedom, the full (40×40) background covariance matrix can be computed explicitly. Thus, in the simple model, 3DVAR and OI have the same form, and the optimal weight (eq.(2.15)) is computed explicitly. To obtain the background error covariance $\mathbf{B}$, we apply the NMC method (Parrish and Derber 1992), accumulating

differences between 24-hour forecasts and 18-hour forecasts valid at the same time as samples of background errors.

In this simple system, the five equations for Kalman filtering can be solved explicitly. Thus, the "full Kalman filtering" (full-KF) without low-rank assumption has been implemented.

As for the serial EnSRF and LEKF, model-independent core modules have been developed. The details are described in Appendix A. Thus, only interfaces need to be developed according to the model.

The implementation of serial EnSRF is straightforward. The system dimension in core modules is set to 40. The core subroutines are called in the loop of observations because observations are assimilated serially. After assimilating each observation, the background ensemble is updated serially. After assimilating all observations, we retain the final analysis ensemble. In serial EnSRF, we apply the multiplicative covariance inflation described by eq.(2.66).

As for the implementation of LEKF, the local patch is taken as in Fig.2.2. We compute the data assimilation equations for the larger region denoted by l. Since local patches overlap each other, we average the overlapping values inside the smaller region denoted by l_2 to obtain the global analysis. In LEKF, we apply both the multiplicative covariance inflation and the enhanced variance inflation (eq.(2.67)).

2.3.3 Description of experiments

The "nature" or true run is created after long-term integration to avoid the initial spin-up period. All of the data assimilation methods (3DVAR, full-KF, serial EnSRF, and LEKF) are tested in the cases of 20 and 40 observations. We take the origin of the phase space

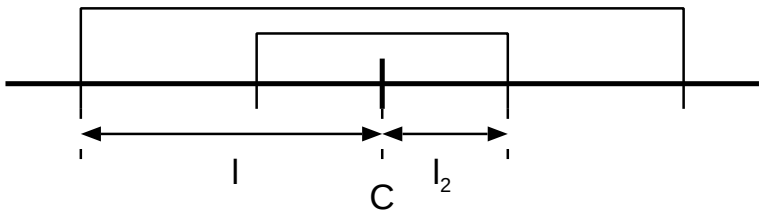

Figure 2.2: Definition of the local regions around a center point (C) for LEKF in the Lorenz-96 model. The larger region denoted by l is the local patch of LEKF, the smaller region denoted by l_2 is the subset of the local patch that is averaged to obtain the global analysis. Since l denotes the length of one side, the size of the local patch becomes $2l+1$, similarly for l_2.

as the initial condition for the analysis experiments. The first 60 days in the analysis experiments are considered as the spin-up period. To obtain the 1-year (365-day) mean of the analysis root mean square error (RMSE), 425-day forecast-analysis cycle has been performed. Observations are taken and analyzed every 6 hours. The observational error standard deviation is 1.0 in the present experiments. The initial ensemble for serial EnSRF and LEKF is chosen as vectors consisting of random numbers with Gaussian distribution. Without data assimilation, the natural variability of the model shows an RMSE of 6.70, which defines the error saturation limit.

2.3.4 Results in the case of 40 observations

Table 2.1 shows summary of the 1-year mean analysis RMSE in the case of 40 observations, that is, all grid points have observations. All methods produce analysis state much closer to the truth than observations. Observational error standard deviation is 1.0, about 15% of the natural variability. All the Kalman filtering methods show the same performance in their optimal cases (RMSE=0.2), whereas 3DVAR has errors with magnitudes about

3DVAR	Full-KF	Serial EnSRF	LEKF
0.40	0.20	0.20	0.20
	($\delta = 0.05$)	($\delta = 0.02$, $\sigma = 5$)	($\delta = 0.015$, $l = 6$, $l_2 = 2$)

Table 2.1: The 1-year mean analysis RMSE of each method with optimal parameter settings on the Lorenz-96 model when the number of observations is 40. Observational error is given as 1.0, and the model natural variability without data assimilation shows RMSE of 6.70. Ensemble size for the EnSRF and LEKF is fixed at 10.

twice as large (RMSE=0.40).

When serial EnSRF uses 40 ensemble members, the filter is stable even without the Schur product to localize the Kalman gain. When ensemble size is reduced, the Schur product is required. There are two parameters for serial EnSRF: the localization length scale σ and the covariance inflation factor δ. The dependence on the parameters is shown graphically in Fig.2.3, which is similar to Fig.3(b) of Whitaker and Hamill (2002). The localization length scale σ is different from the covariance filter length scale L in Whitaker and Hamill (2002). Whitaker and Hamill (2002) defined the length scale L so the covariance becomes 0 when $d > L$ where d denotes distance. Our definition is given by eqs.(2.49) and (2.50), where the distance d that yields zero covariance is given by

$$d > 2 \times \sqrt{\frac{10}{3}} \times \sigma \tag{2.80}$$

Thus, the relationship between the length scale L in Whitaker and Hamill (2002) and the length scale σ used in this experiments is given as

$$\begin{aligned} L &= 2 \times \sqrt{\frac{10}{3}} \times \sigma \\ &\simeq 3.65\sigma \end{aligned} \tag{2.81}$$

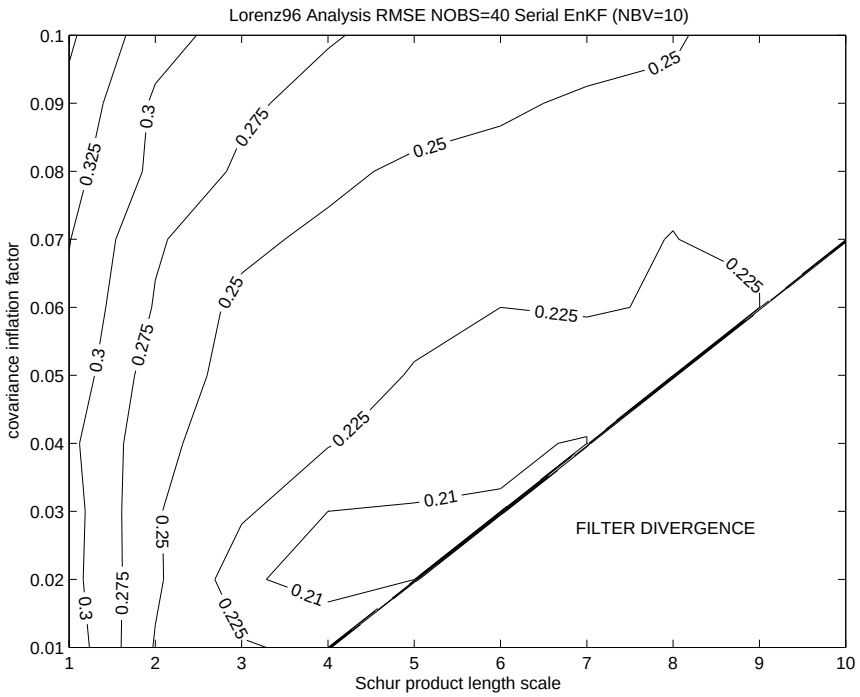

Figure 2.3: Parameter dependence of the analysis RMSE of serial EnSRF with 10 ensemble members on the Lorenz-96 model when the number of observations is 40, cf. Fig.3(b) of Whitaker and Hamill (2002). The horizontal and vertical axes show the localization length scale σ and the covariance inflation paramter δ, respectively. The minimum error 0.20 is observed when $\sigma = 5$ and $\delta = 0.02$. "FILTER DIVERGENCE" denotes the region with RMSE of more than 1.0.

In considering the different measure of the length scale, Fig.2.3 shows almost identical performance as Fig.3(b) in Whitaker and Hamill (2002), which supports the correctness of the Fortran code developed here for the present research. According to Whitaker and Hamill (2002), the minimum error was 0.20 which is identical to our results.

Fig.2.4 shows a similar figure as Fig.2.3 in the case of LEKF. LEKF has three parameters: the local patch size $2l + 1$, the size of the local patch averaged to obtain the global analysis $2l_2 + 1$, and the enhanced variance inflation factor δ. Since no clear dependence

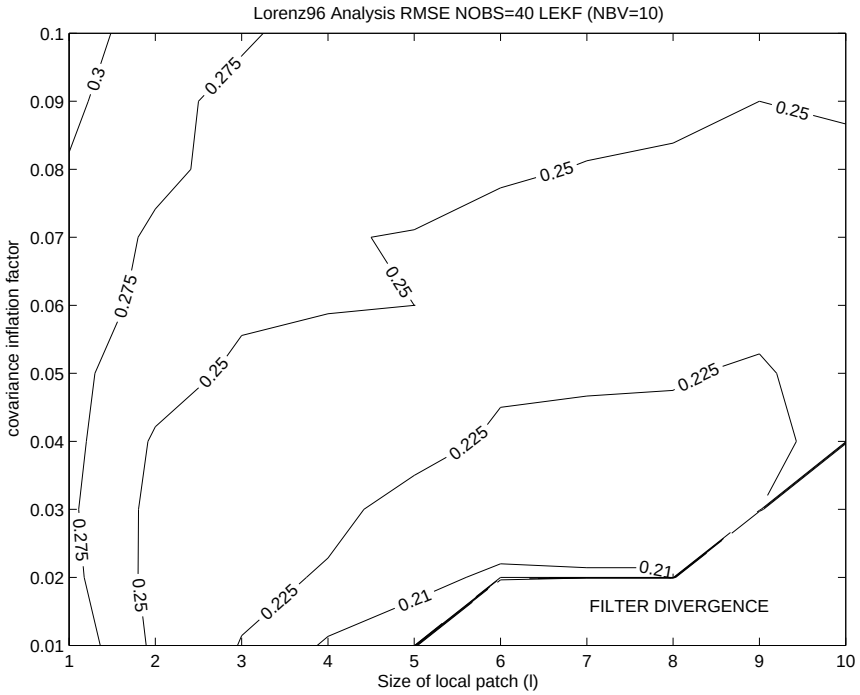

Figure 2.4: The similar as Fig.2.3 but using LEKF with 10 ensemble members using the enhanced variance inflation. The horizontal and vertical axes show the local patch parameter l and the enhanced variance inflation paramter δ, respectively. l_2 is fixed to 1. The minimum error 0.20 is observed when $l = 5$ and $\delta = 0.01$.

has been observed with respect to l_2, l_2 is fixed to 1. Note that the direct comparison between serial EnSRF and LEKF may be misinterpreted since the way of localization is different between serial EnSRF and LEKF, so that the metric in the horizontal axis may be different. The minimum error 0.20 is observed when $l = 5$ and $\delta = 0.01$.

In both EnKF, the following characteristics are observed:

- The smaller the localization length scale, the more stable the filter (i.e., the farther away from the region of filter divergence).

- The larger the covariance inflation factor, the more stable the filter.

3DVAR	Full-KF	Serial EnSRF	LEKF
1.15	0.33	0.33	0.33
($\mathbf{B} = 2.0 \times \mathbf{B}$)	($\delta = 0.1$)	($\delta = 0.04$, $\sigma = 6$)	($\delta = 0.025$, $l = 6$, $l_2 = 2$)

Table 2.2: The same as Table 2.1 but in the case of 20 observations and 10 ensemble members.

- A larger length scale requires a larger covariance inflation factor for stable filtering.
- Filter divergence occurs with larger length scales and smaller inflation factors.
- The minimum error is observed in a medium length scale ($\sim 5, 6$) with a smaller inflation factor (~ 0.02) near the boundary of the filter divergence area.

2.3.5 Results in the case of 20 observations

The results in the case of 20 observations are summarized in Table 2.2 when the ensemble size for EnKF is fixed to 10. The 20 observations are located regularly every other point. Since the error variance is larger than in the case of 40 observations, the background error covariance used in 3DVAR was multiplied by 2.0, which is a good choice among several values we tried, though this may not be strictly optimal. The RMSE of 3DVAR is larger than the observational error, whereas those of Kalman filters are much smaller. The difference between 3DVAR and Kalman filters becomes larger here, thus, Kalman filters show a larger advantage with fewer observations. All the Kalman filtering methods show the same performances (RMSE=0.33) with their optimal settings.

Fig.2.5 shows the same figure as Fig.2.3 but in the case of 20 observations. Compared to the case of 40 observations, the filter becomes considerably more unstable and the region "FILTER DIVERGENCE" becomes larger, although the basic characteristic is similar.

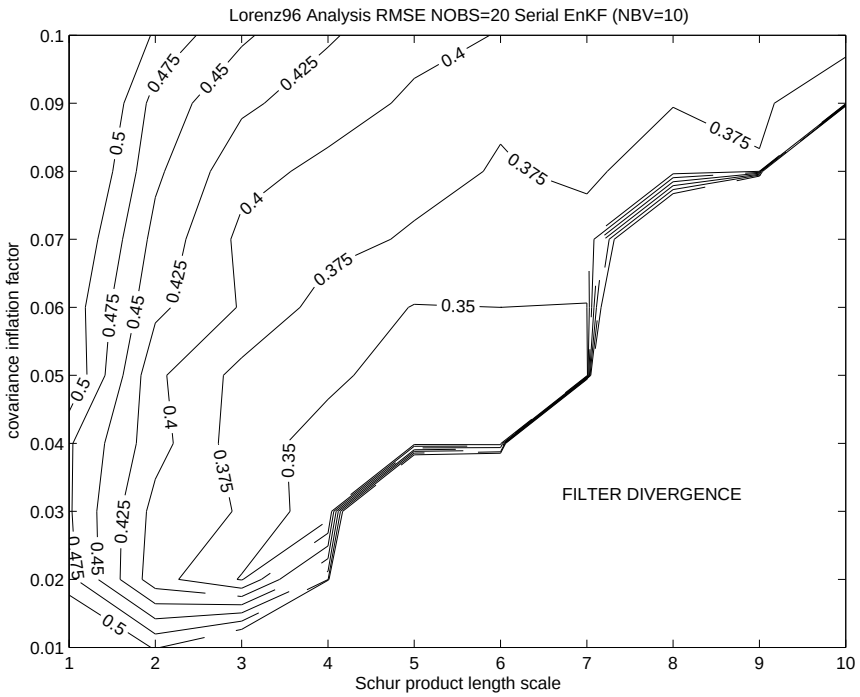

Figure 2.5: The same as Fig.2.3 but in the case of 20 observations and serial EnSRF with 10 ensemble members. The minimum error 0.33 is observed when $\sigma = 6$ and $\delta = 0.04$.

The minimum error 0.33 occurs when $\sigma = 6$ and $\delta = 0.04$.

Fig.2.6 shows the same figure as Fig.2.5 but ensemble size is increased to 20. When ensemble size is increased, the filter becomes more stable. In addition, the error becomes slightly smaller almost everywhere. The minimum error 0.30 occurs when $\sigma = 6$ and $\delta = 0.02$.

Both serial EnSRF and LEKF work with 8 ensemble members. Fig.2.7 shows the same figure as Fig.2.5 but with 8 ensemble members. Fig.2.7 shows a similar structure as Fig.2.5 but with larger "filter divergence" area and with larger errors almost everywhere. Smaller ensemble size requires smaller covariance length scales. The minimum error 0.34 is observed when $\sigma = 5$ and $\delta = 0.04$, which is slightly worse than in the case of 10 ensemble members.

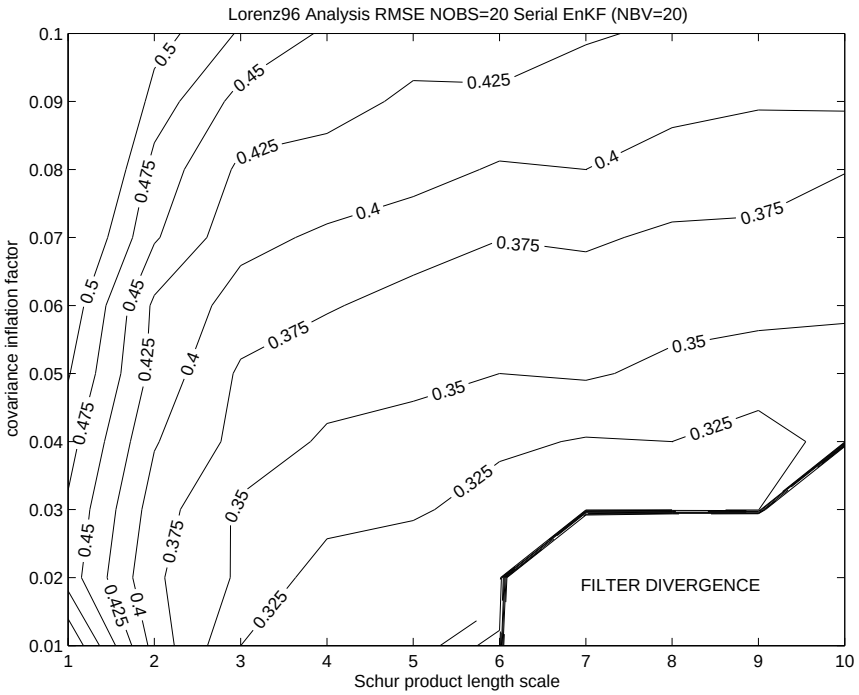

Figure 2.6: The same as Fig.2.5 but serial EnSRF with 20 ensemble members. The minimum error 0.30 is observed when $\sigma = 6$ and $\delta = 0.02$.

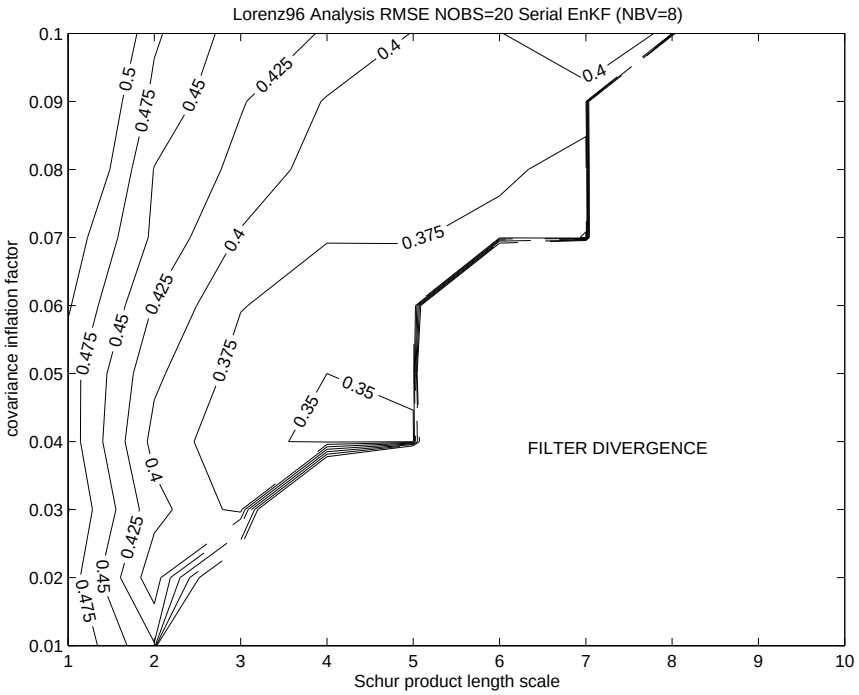

Figure 2.7: The same as Fig.2.5 but serial EnSRF with 8 ensemble members. The minimum error 0.34 is observed when $\sigma = 5$ and $\delta = 0.04$.

l_2 / inflation	0.03	0.04	0.05	0.06	0.07
1	.41	.37	.35	.36	.38
2	.34	.34	.35	.35	.37
3	.82	.34	.34	.38	.35
4	.34	.35	.40	.35	.36
5	.33	.35	.33	.35	.35
6	.81	.35	.35	.38	.36

Table 2.3: Analysis RMSE using LEKF with 8 ensemble members on Lorenz-96 model when the number of observations is 20. Enhanced variance inflation is applied, and the local patch parameter is fixed as $l = 6$. The other two parameters, l_2 and δ, are changed.

Fig.2.8 shows the same figure in the case of LEKF. Here, l_2 is fixed to 0. It seems LEKF has a smaller "filter divergence" area, thus, LEKF seems more stable than serial EnSRF. However, the errors are slightly larger in LEKF. The minimum error 0.35 occurs when $l = 5$ and $\delta = 0.03$, which is slightly worse than serial EnSRF. As shown in Table 2.3, when we change l_2 and δ with l fixed as $l = 6$, we get the minimum error 0.33 in LEKF.

Although a smaller error appeared by changing l_2, no clear dependence with respect to l_2 can be seen in Table 2.3. The fluctuation of the values suggests that the lengths of the runs may not be sufficient to get reliable RMSE values, although the same computations with different random numbers for observational errors performed several times yielded the same overall results. Table 2.3 suggests that the LEKF errors are rather insensitive to l_2, since the only two large RMSE values (about 0.8) occur with the inflation parameter $\delta = 0.03$ close to the area in which the filter becomes divergent (Fig.2.8).

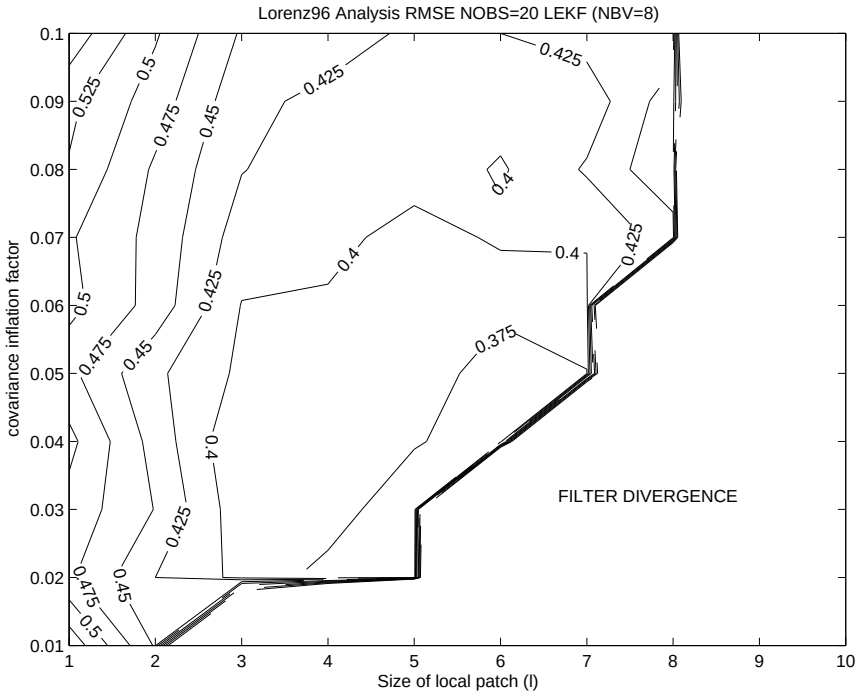

Figure 2.8: The same as Fig.2.4 but LEKF with 8 ensemble members and enhanced variance inflation. l_2 is fixed to 0. The minimum error 0.35 is observed when $l = 5$ and $\delta = 0.03$.

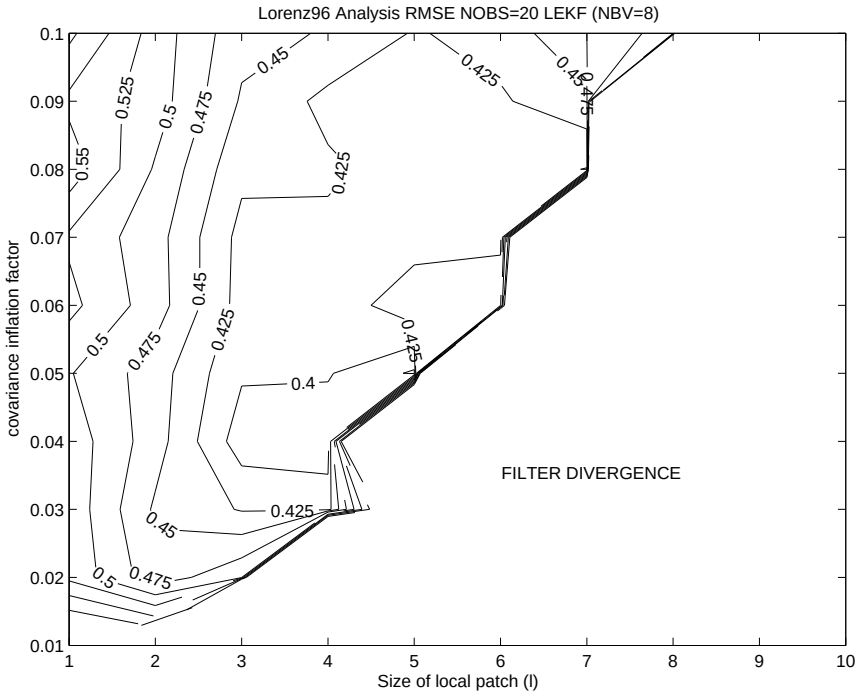

Figure 2.9: The same as Fig.2.8 but LEKF with 8 ensemble members and multiplicative variance inflation. The way to inflate covariance is the same as that of serial EnSRF. The minimum error 0.37 is observed when $\sigma = 5$ and $\delta = 0.05$.

Fig.2.9 shows the same figure as Fig.2.8 but without enhanced variance inflation, replaced by the same multiplicative covariance inflation used for the serial EnSRF. Enhanced variance inflation contributes to stabilize the filter, the minimum error becomes smaller with the smaller inflation factor. In addition, with enhanced variance inflation, the errors are smaller than with multiplicative inflation almost everywhere. Thus, enhanced variance inflation contributes positively to the filter performance.

We also applied observational error covariance localization within LEKF. Fig.2.10 shows the same figure as Fig.2.9 but with observational error covariance localization. Here, we fixed local patch parameters to $l = 15$ and $l_2 = 0$, and varied localization scale σ of the observational error covariance localization. With observational error covariance

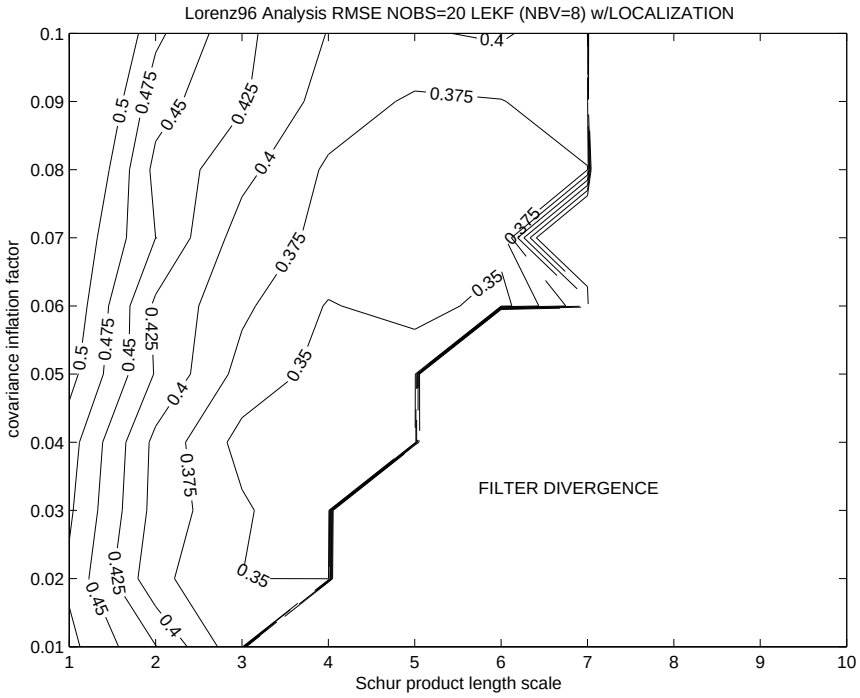

Figure 2.10: The same as Fig.2.9 but LEKF with 8 ensemble members, multiplicative variance inflation, and observational error covariance localization. The way to inflate covariance is the same as that of serial EnSRF. The local patch parameters are fixed to $l = 15$ and $l_2 = 0$. The horizontal axis shows the length scale σ of observational error covariance localization. The minimum error 0.33 is observed when $\sigma = 4$ and $\delta = 0.03$.

localization, we obtain the lowest minimum error of 0.33 among all methods in the case of 8 ensemble members. Comparing Figs.2.10 and 2.7, we see that the use of localization allows LEKF to use larger patches, with more observations, multiplicative inflation (instead of enhanced inflation), and an overall performance at least as good as that of the serial EnSRF with similar parameters. This observation has important implications for the practical implementation of LEKF.

2.3.6 Online estimation of covariance inflation parameter

Kalnay (pers. comm.) suggested the covariance inflation parameter can be estimated by the following statistics:

$$\left\langle \mathbf{d}\mathbf{d}^{\top} \right\rangle = (1+\delta)\mathbf{H}\mathbf{P}^{f}\mathbf{H}^{\top} + \mathbf{R} \tag{2.82}$$

a similar equation is found in eq.(3) of Houtekamer et al. (2005). Here, $\mathbf{d}$ denotes the observational increment defined as

$$\mathbf{d} = \mathbf{y}^{o} - H\mathbf{x}^{f} \tag{2.83}$$

Eq.(2.82) shows that the observational increment is the combination of the observational error and the forecast error. To get the best estimate of the inflation parameter δ, one can run the data assimilation cycle once and compute the statistics eq.(2.82) to update δ. As an analogy of Kalman filtering in a model output statistics (MOS) (see for example, appendix C of Kalnay 2003), we can apply Kalman filtering to make an online estimation of δ. At each analysis time step, we estimate δ from eq.(2.82) to be δ^{o}. We assume error variance σ^{o} in this estimate, and we use the estimate at the previous time as a background δ^{b} with error variance σ^{b}. Then, the analysis equations are given as:

$$\delta^{a} = \frac{\delta^{b}\sigma^{o} + \delta^{o}\sigma^{b}}{\sigma^{b} + \sigma^{o}} \tag{2.84}$$

$$\sigma^{a} = \left(1 - \frac{\sigma^{b}}{\sigma^{b} + \sigma^{o}}\right)\sigma^{b} \tag{2.85}$$

This analyzed δ^{a} can be used for the variance inflation parameter at the current step. δ^{a} and σ^{a} are used as background at the next time step.

We applied online estimation of covariance inflation parameter. Table 2.4 shows analysis RMSE in the case of 20 observations using LEKF with 8 ensemble members. Multiplicative covariance inflation and enhanced variance inflation are applied in both

local patch (l) or localization scale (σ)	1	2	3	4	5	6	7
Multiplicative	.50	.46	.38	.38	.35	.36	.37
Enhanced	.47	.40	.36	.37	.35	.35	.38
Multiplicative w/obs error localization	.47	.38	.37	.34	.33	.34	.36
Enhanced w/obs error localization	.49	.41	.36	.34	.34	.34	.38

Table 2.4: Analysis RMSE using LEKF with 8 ensemble members on Lorenz-96 model when the number of observations is 20. Online estimation of variance inflation parameter is applied. l_2 is fixed to $l_2 = 0$. In the cases with observational error covariance localization, l is fixed to $l = 15$.

cases with and without observational error covariance localization. The online estimation provides quite stable performance, although the offline optimally chosen constant inflation parameter provides similar results. The smallest overall analysis error 0.33 is observed with multiplicative inflation, a local patch parameter of $l = 15$, and an observational error covariance localization size of 5.

This result is important because it provides a new approach to estimate covariance inflation adaptively: In a global model the use of the same inflation for both tropical and midlatitude regions is probably suboptimal since forecast errors (including model errors) are larger in the tropics (e.g., Szunyogh et al, 2005).

2.3.7 Timing results

Computational time on a Linux PC with a 2.7GHz Intel Celeron (Northwood) processor is shown in Table 2.5. The computation is a whole 425-day forecast-analysis cycle, containing 1700 data assimilations. The only difference between 3DVAR and full-KF is covariance

3DVAR	Full-KF	Serial EnSRF	LEKF
2.9	3.1	0.7	8.7

Table 2.5: Computational time in seconds on a Linux PC with an Intel Celeron 2.7GHz processor for a 425-day forecast-analysis cycle on the Lorenz-96 model.

forecasts. Serial EnSRF is faster because it does not require matrix inversion. LEKF is slower than full KF for the Lorenz-96 system.

Note that this timing result is very unrepresentative of realistic atmospheric systems with high dimensionality, large number of observations and massively parallel computers. LEKF has an advantage for parallel computation since each local patch is completely independent, the LEKF computation can be parallelized very efficiently. In addition, LEKF can treat large number of observations with less additional cost. Furthermore, the new local ensemble transform KF (LETKF, Hunt 2005) avoids the eigenvalue decomposition and is several times faster than LEKF (Harlim, Szunyogh, pers. comm.). On the other hand, serial EnSRF cannot be parallelized efficiently. Moreover, serial EnSRF may be expensive in practice since it contains a loop over the number of observations. Implementation of 3DVAR is completely different for large systems, so the timing cannot be generalized.

2.3.8 Summary

The results on the Lorenz-96 model can be summarized as follows:

- The core modules of both serial EnSRF and LEKF are correctly coded since we have been able to reproduce the results of Whitaker and Hamill (2002) and of Ott et al (2004).

- All the Kalman filtering methods (full-KF, serial EnSRF, and LEKF) outperform

3DVAR, and the advantage becomes larger with fewer observations.

- All the Kalman filtering methods (full-KF, serial EnSRF, and LEKF) have comparable performance with their optimal settings.

- Observational error covariance localization contributes to better LEKF performance with little additional computational cost.

- Enhanced variance inflation in LEKF contributes to better filter performance, but multiplicative inflation becomes better when using localization.

- With the same multiplicative variance inflation, serial EnSRF outperforms LEKF, but when localization and larger patches are used, LEKF is equal or better than serial EnSRF with similar parameters.

- Smaller localization length scales and larger covariance inflation factors stabilize EnKF, although both give suboptimal filtering.

- A new method introduced to estimate adaptively the optimal inflation factor works well and gives results similar to the best obtained with offline tuning of a constant inflation. In global models it could be used to estimate adaptive inflation factors depending on location and time.

Chapter 3

Data assimilation on the SPEEDY primitive-equation model

3.1 Introduction

3.1.1 Overview

In Chapter 2, we described theoretical review and applied 3DVAR, serial EnSRF, LEKF, and full KF on the Lorenz-96 model. All KF methods (serial EnSRF, LEKF, and full KF) show similar results on the Lorenz-96 model, outperforming 3DVAR. In this chapter, we describe the implementation and numerical experiments of three data assimilation methods (3DVAR, serial EnSRF, and LEKF) on the SPEEDY primitive-equation model. We explore the following questions in this chapter:

1. How does each method work on the SPEEDY model?
2. What is the relative advantages and disadvantages between serial EnSRF and LEKF?
3. How sensitive is the data assimilation to experimental settings?
4. What are the characteristics of the analysis and forecast error fields?

A goal of this chapter is to investigate and compare the performance of the three methods. In addition, we perform sensitivity experiments with different observational networks, moisture observations, vertical error correlations, error covariance localization, and random perturbation addition to ensemble members. These sensitivity experiments investigate robustness of our results to the experimental settings. Moreover, they suggest possible ways to improve the filter performance. We also see the characteristics of the

analysis and forecast error fields, suggesting how EnKF works.

3.1.2 The SPEEDY model

The SPEEDY model (Molteni 2003) is a recently developed atmospheric general circulation model (AGCM) with a spectral primitive-equation dynamic core and a set of simplified physical parameterization schemes (SPEEDY stands for Simplified Parameterizations, primitivE-Equation DYnamics). The goal of this model is to achieve computational efficiency while maintaining characteristics similar to the state-of-the-art AGCMs with complex physics. The resolution of the model is T30L7 (horizontal spectral truncation of 30 wave numbers and 7 vertical levels), the computational cost is one order of magnitude less than that of state-of-the-art AGCMs at similar horizontal resolution. According to Molteni (2003), the SPEEDY model simulates the general structure of global atmospheric circulation fairly well, and some aspects of the systematic errors are similar to many AGCMs, though the error amplitude is larger than state-of-the-art models.

The SPEEDY model includes basic components of physical parameterizations used in more complex GCMs, such as convection (a simplified mass-flux scheme), large-scale condensation, clouds, short-wave radiation (two spectral bands), long wave radiation (four spectral bands), surface fluxes of momentum and energy (bulk aerodynamic formula), and vertical diffusion. Details of the simplified physical parameterization schemes of the SPEEDY model can be found in Molteni (2003), especially in its Appendix which is available on the website: "http://www.ictp.trieste.it/~moltenif/speedy-doc.html". The boundary conditions of the SPEEDY model includes topographic height and land-sea mask, which are constant, and sea surface temperature (SST), sea ice fraction, surface temperature in the top soil layer, moisture in the top soil layer and the root-zone layer,

and snow depth, all of which are specified by monthly means, and bare-surface albedo and fraction of land-surface covered by vegetation, which are specified by annual-mean fields. The lower boundary conditions such as SST are obtained by ECMWF's reanalysis in the period 1981-90. The incoming solar radiation flux and the boundary conditions (SST etc.), except bare-surface albedo and vegetation fraction, are updated daily.

Since the SPEEDY model is designed for long-term climate variability studies that require ensembles of long-term integrations, the original code contains only time-mean outputs. Thus, the input/output processes of a grid point value at intermittent time steps had to be developed in order to enable short-term integrations of the forecast-analysis cycle. The prognostic variables are zonal and meridional wind velocity components (u, v), temperature (T), specific humidity (q), and surface pressure (p_s). The grid point value output is in the physical space with the grid size of $96 \times 48 \times 7$. Inputs are only taken in sigma levels, but outputs are in both sigma levels and pressure levels. The specific heights of the vertical levels are shown in Table 3.1.

As for the timing results of running the model, a 3-month integration took about 6 minutes on a Linux PC with a 2.7GHz Intel Celeron (Northwood) processor, whereas it took around 8 minutes for 24-hour integration of the Japanese operational global weather prediction model with a reduced resolution (T42) on the same PC. For 6-hour cycle experiments, a 6-hour forecast of the SPEEDY model requires about 2 seconds on the same PC. Thus, a 2-month cycle experiment requires about 8-minute forecast computations for each ensemble member.

Level index	Sigma heights (σ)	Pressure heights (hPa)
1	0.950	925
2	0.835	850
3	0.685	700
4	0.510	500
5	0.340	300
6	0.200	200
7	0.080	100

Table 3.1: Vertical levels of the SPEEDY model outputs. Sigma levels are also used as model levels.

3.2 3DVAR implementation

3.2.1 Theory for practical implementation

The variational formulation has been introduced in Section 2.2.8. Practically, it is impossible to implement the variational formulation explicitly because of the large number of degrees of freedom of the system. Usually, NWP models have prognostic variables with dimensions of $\mathcal{O}(10^7)$, thus, the matrix $\mathbf{B}$ has $\mathcal{O}(10^{14})$ elements, which requires at least $\sim$ 10TB of memory just for storing the matrix. However, most components of the matrix $\mathbf{B}$ are very close to 0, and we can simplify $\mathbf{B}$ very much under reasonable assumptions. We can understand this fact by considering the forecast error of temperature at College Park, for example, has nothing to do with the forecast error of wind in Tokyo.

Parrish and Derber (1992) assumed zero spatial error correlation in the spectral space. With the spectral transformation, they reduced most components of $\mathbf{B}$. Alter-

natively, Barker et al. (2004) did not use spectral transformation in their 3DVAR built for the nonhydrostatic fifth-generation Pennsylvania State University/National Center for Atmospheric Research Mesoscale Model (MM5). We follow a similar approach as Barker et al.

We define a variable transformation $\mathbf{U}$ as follows:

$$\delta\mathbf{x} = \mathbf{U}\delta\mathbf{v} \tag{3.1}$$

Here, $\delta\mathbf{x}$ is defined by eq.(2.75), $\delta\mathbf{v}$ is a vector with the same dimension as $\delta\mathbf{x}$. Substituting eq.(3.1) into eq.(2.73), we get

$$J(\delta\mathbf{v}) = \frac{1}{2}\delta\mathbf{v}^\top\mathbf{U}^\top\mathbf{B}^{-1}\mathbf{U}\delta\mathbf{v} + \frac{1}{2}(\mathbf{HU}\delta\mathbf{v} - \mathbf{d})^\top\mathbf{R}^{-1}(\mathbf{HU}\delta\mathbf{v} - \mathbf{d}) \tag{3.2}$$

In order for the covariance matrix to be the identity,

$$\mathbf{B} = \mathbf{U}\mathbf{U}^\top \tag{3.3}$$

should be satisfied. Then, eq.(3.2) can be rewritten as

$$J(\delta\mathbf{v}) = \frac{1}{2}\delta\mathbf{v}^\top\delta\mathbf{v} + \frac{1}{2}(\mathbf{HU}\delta\mathbf{v} - \mathbf{d})^\top\mathbf{R}^{-1}(\mathbf{HU}\delta\mathbf{v} - \mathbf{d}) \tag{3.4}$$

Differencing eq.(3.4) with respect to $\delta\mathbf{v}$, we get the gradient of the cost function

$$\nabla J(\delta\mathbf{v}) = \delta\mathbf{v} + \mathbf{U}^\top\mathbf{H}^\top\mathbf{R}^{-1}(\mathbf{HU}\delta\mathbf{v} - \mathbf{d}) \tag{3.5}$$

At this point, we can solve the 3DVAR problem for $\delta\mathbf{v}$. Since we start the minimizing process from the background state, initially $\delta\mathbf{v} = 0$. Given the cost function (eq.(3.4)) and its gradient (eq.(3.5)), a quasi-Newton minimizer finds the solution for $\delta\mathbf{v}$, and eventually, eq.(3.1) converts $\delta\mathbf{v}$ to $\delta\mathbf{x}$, the analysis state in the real space. Note that only $\mathbf{U}$ and $\mathbf{U}^\top$ are used in this process; $\mathbf{U}^{-1}$ is not required.

The construction of the variable transformation **U** is the key in the 3DVAR algorithm. Usually it is assumed that **U** can be separated into a spatial correlation component and an inter-variable correlation component, that is,

$$\mathbf{U} = \mathbf{VCA} \tag{3.6}$$

where **A**, **C** and **V** stand for the error standard deviation, the spatial error correlation and the inter-variable error correlation, respectively. Moreover, it is assumed **C** can be separated into horizontal and vertical correlations. Usually, the horizontal correlation is assumed to be a Gaussian shape. Only correlation length scale parameter needs to be stored. Furthermore, the length scale is assumed to be constant in all points at the same vertical level. Quasi-isotropy is also assumed. Full vertical correlation could be considered, but it also can be assumed to be Gaussian.

An important part of 3DVAR is how to express the inter-variable correlation **V**, which transforms prognostic variables into control variables that are assumed to be independent of each other. Some prognostic variables are strongly dependent on each other mostly because of the geostrophic balance. Usually in global models, only the geostrophic balance is considered since it is the strongest and the most important balance in the system.

3.2.2 Implementation on the SPEEDY model

In the previous section, we described a general idea to implement 3DVAR on a practical atmospheric model following the approach by Barker et al. (2004). Here, we describe the design of the variable transformation **U** in the present implementation.

The first part is the error standard deviation **A** in eq.(3.6). Since the present 3DVAR analyzes not in the spectral space but in the physical grid space, we consider the full spatial

dependence of the error standard deviation. Almost no additional computation is required by considering the full spatial dependence.

For the spatial error correlation $\mathbf{C}$ in eq.(3.6), the spatial correlation is separated into vertical and horizontal correlations. The vertical resolution of the SPEEDY model is so coarse that vertical background error correlation is not considered (experiments including 3 levels yielded worse results than one layer). Horizontal error correlation is assumed to be Gaussian, and the 4th order recursive filter (RF) technique (Purser et al. 2003) has been implemented to estimate the Gaussian correlation shape. The 4th order RF is a quasi-Gaussian filter composed of a pair of forwarding and backwarding processes

$$B_i = \beta A_i + \alpha_1 B_{i-1} + \alpha_2 B_{i-2} + \alpha_3 B_{i-3} + \alpha_4 B_{i-4} \tag{3.7}$$

$$C_i = \beta B_i + \alpha_1 C_{i+1} + \alpha_2 C_{i+2} + \alpha_3 C_{i+3} + \alpha_4 C_{i+4} \tag{3.8}$$

where the subscript i denotes the space index of grid points in one direction. In eq.(3.7), that is the forwarding process in the ascending order of i, A is input and the filtered value B is obtained. In eq.(3.8), that is the backwarding process in the descending order of i, B is input and the filtered value C is obtained. Eq.(3.7) and eq.(3.8) are mutually adjoint, so eq.(3.7) may be used in $\mathbf{U}$ whereas eq.(3.8) may be used in $\mathbf{U}^\top$. The coefficients β and α can be obtained by solving a 4th degree equation according to the correlation length scale. The details of the recursive filter technique are described in Appendix B.

For the inter-variable error correlation $\mathbf{V}$ in eq.(3.6), we consider the geostrophic balance and introduce control variables as follows:

$$u_u = u - r_1 u_g(p_s, T) \tag{3.9}$$

$$v_u = v - r_2 v_g(p_s, T) \tag{3.10}$$

where r_1 and r_2 are regression coefficients determined from statistics, and u_g and v_g are

zonal and meridional components of the geostrophic wind computed from p_s and T using the geostrophic balance equation on the sigma coordinate system (cf. Kalnay 2003, p.65, eq.(2.6.27)):

$$f\mathbf{k} \times \mathbf{v}_g = -RT\nabla \ln p_s - \nabla\phi \tag{3.11}$$

where f, $\mathbf{k}$, R, and ϕ denote Coriolis parameter, vertical unit vector, gas constant, and geopotential height, respectively. Other variables (T, q, and p_s) are assumed to be independent of the unbalanced wind (u_u and v_u). Thus, only the geostrophic balance is considered for inter-variable correlation.

3.2.3 Background error statistics - NMC method

The design of variable transformation $\mathbf{U}$ described in the previous section requires background error statistics. We need to specify the error standard deviation, the length scale of horizontal error correlations, and the regression coefficients from the statistics. Background error statistics were computed using the NMC method (Parrish and Derber 1992), accumulating the difference between 24-hour forecast and 18-hour forecast as a sample of the background error. In the present experiment, 87 samples from January 10 to January 31 were accumulated for the statistics. The analysis fields for the first time are obtained using 3DVAR with a reasonable guess of background error statistics, such as 500km for the horizontal background error correlation length scale and the background error standard deviation with the same magnitude as observational error standard deviation. The NMC method provides better statistics. Once we get a more reasonable guess of background error statistics using the NMC method, this process may be repeated.

Fig.3.1 shows the zonal mean of the regression coefficients r_1 with respect to u-wind (eq.(3.9)). Almost no geostrophic balance is observed in the tropics, whereas strong

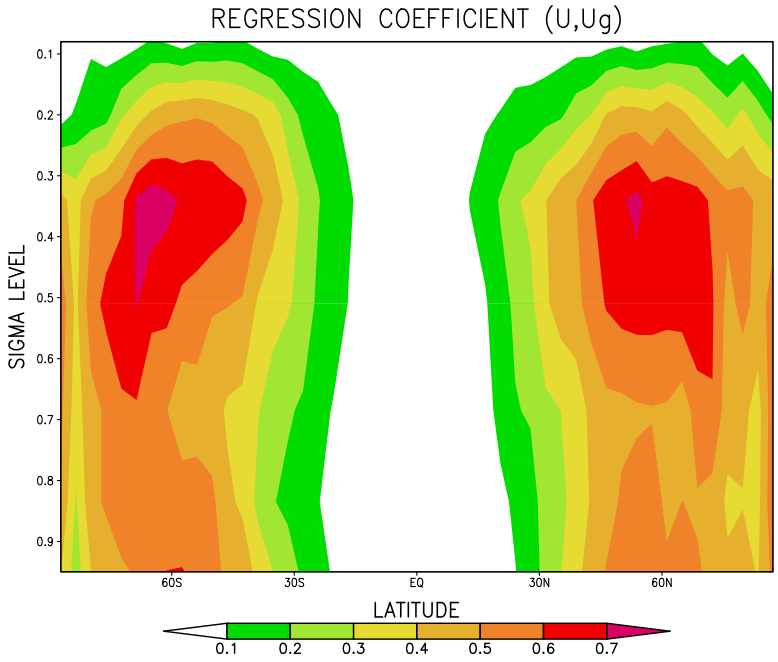

Figure 3.1: Zonal mean of the regression coefficients r_1 of u. The vertical and horizontal axes show height and latitude respectively. The coefficients reflect statistical strength of the geostrophic relationship. The larger values, i.e. stronger geostrophic balance, are observed in mid-latitudes.

geostrophy is observed in the mid-latitudes especially at the jet levels. The regression coefficients r_2 with respect to v-wind show a similar structure. Fig.3.2 shows the error standard deviation of p_s (top panel) and u at the 4th level (bottom panel), indicating strong spatial dependence, especially in zonal structures in error standard deviation. Fig.3.3 shows the vertical structure and latitudinal dependence of horizontal error correlations. There is almost no vertical dependence (top panel), all levels except the top show almost no correlation beyond 2 grid points. There is large latitudinal dependence in grid spacing (bottom panel). This reflects the skewness of the global map, grid spacing is physically more dense in higher latitudes.

The figures shown as background error statistics so far are only small parts of the

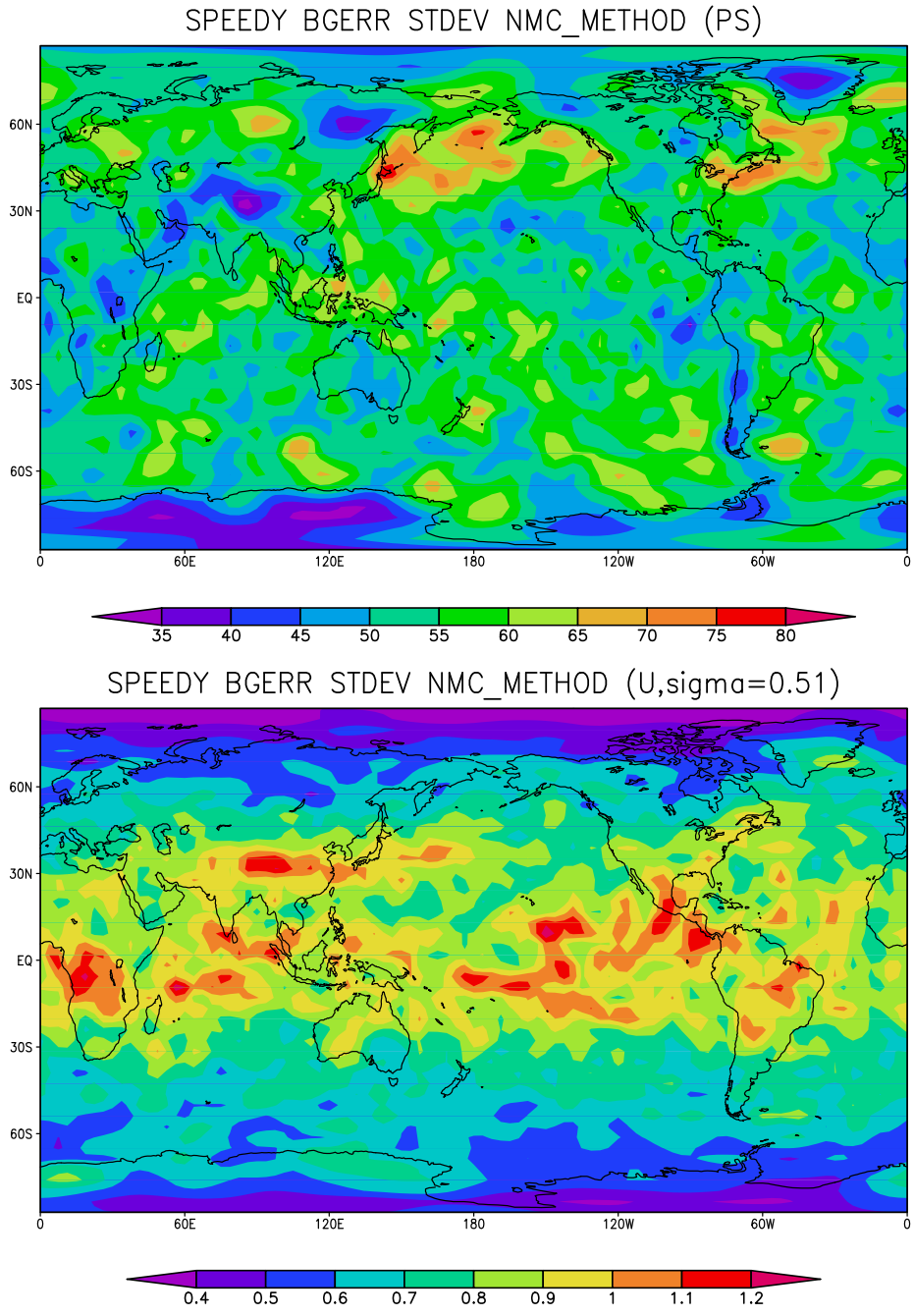

Figure 3.2: Background error standard deviation of p_s (top panel) and u at the 4th level (bottom panel) that are estimated using the NMC method. The units for p_s and u are Pa and m/s, respectively. Strong spatial dependence especially in zonal structures is observed, but their noisiness indicates sampling errors.

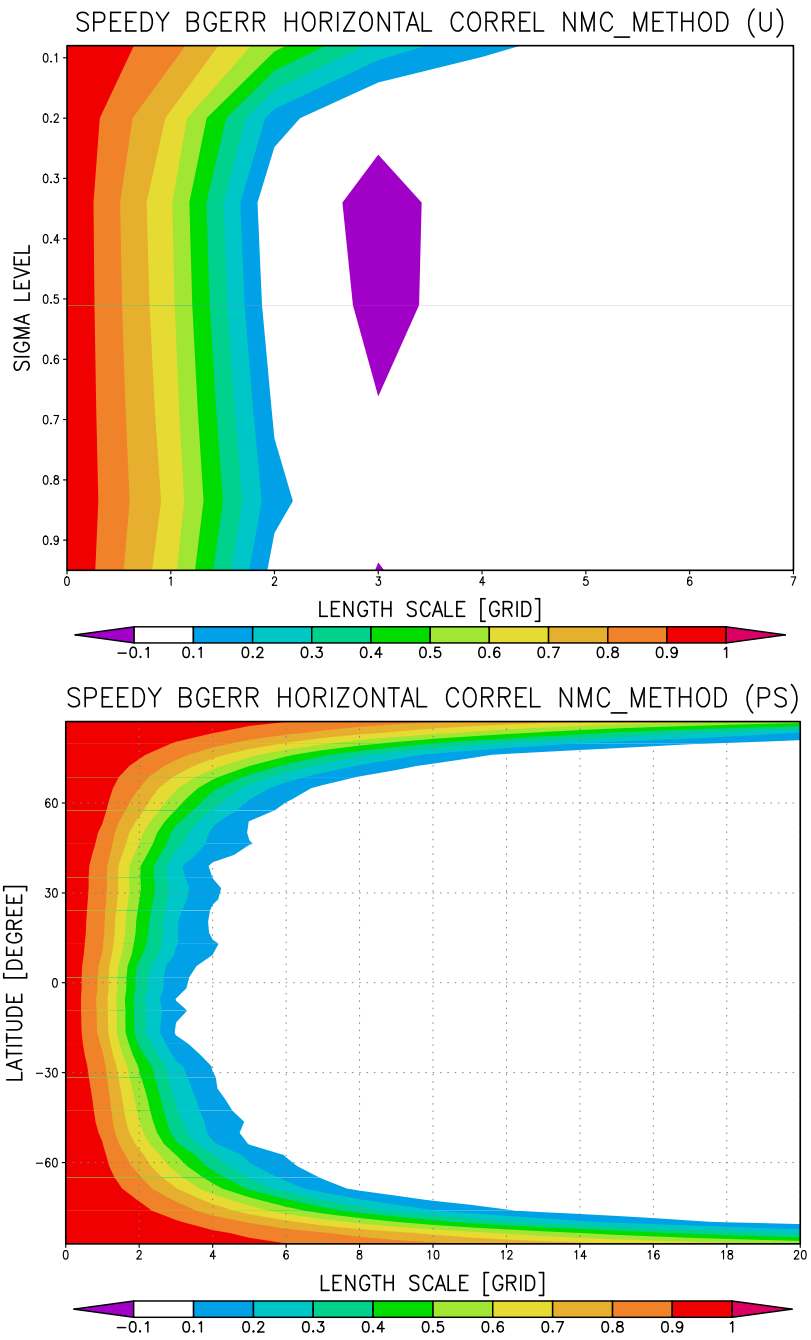

Figure 3.3: Background error horizontal correlations. The top panel shows vertical structure of u in y-direction, where the vertical axis shows sigma levels. The bottom panel shows the latitudinal structure of p_s in x-direction, where the vertical axis shows latitudes. The horizontal axis shows horizontal length scale in grid spacing. The shades show correlation.

background error information that are considered in the present 3DVAR. The present 3DVAR considers all components of standard deviation, the standard deviation has the same size as a grid point value. As horizontal spatial correlation, zonal and meridional components are stored separately. The zonal components are allowed to vary in latitude and height, whereas meridional components are allowed to vary only in height. The regression coefficients measure the strength of the geostrophic balance and are allowed to vary in height and latitudes. The regression coefficients for u (r_1) and v (r_2) are separately stored, although they should be equal in theory.

In order to examine our assumption of vertical independence, we compute vertical correlations averaged horizontally. Fig.3.4 shows vertical correlations using the NMC method. Some vertical correlations are seen at lower levels, but almost no correlation is observed at upper levels. Except for moisture, only the first and second levels have large correlations (about 0.4), other levels show less than 0.2 correlations. Thus, the vertical independence is not a bad assumption in 3DVAR. Moisture, shown by long-short dashed lines with black square, has largest vertical correlation at levels 2, 3, and 4, which may be related to vertical moisture transport by large-scale convective parameterization.

3.2.4 Response tests

We perform a response test using a single observation of u with the observational increment of 1.0m/s. Fig.3.5 shows responses when the observational station is located at mid-latitude (top panel) and high-latitude (bottom panel) at the 4th level. It shows how RF works, and it seems RF expands the observation signal to the Gaussian shape as expected. In the higher latitude, the skewness of the map is considered correctly. Fig.3.6 shows analysis increments of wind vector and T (top panel) and p_s (bottom panel) when

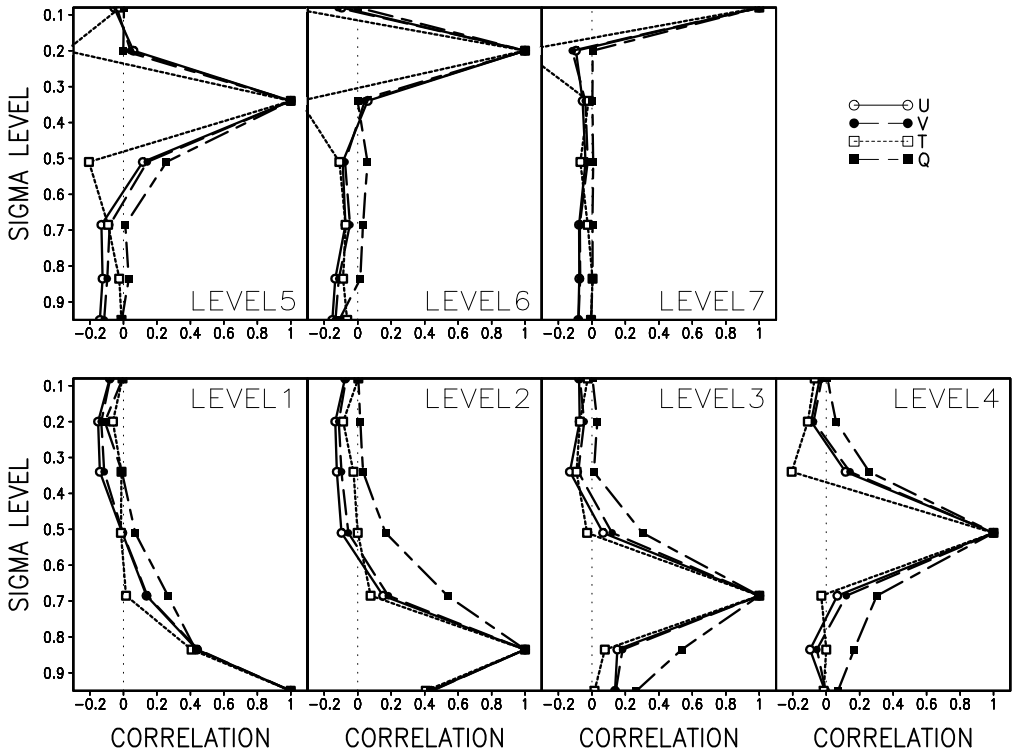

Figure 3.4: Vertical background error correlations of u-wind (solid lines with white circle), v-wind (dashed lines with black circle), temperature (short dashed lines with white square), and moisture (long-short dashed lines with black square). The correlations are averaged horizontally.

the observation is at mid-latitude at the 4th level. Although only u is observed, T and p_s are analyzed through the inter-variable correlation.

Since the single observation response tests show reasonable results, the 3DVAR is tested using the dense observational network (Fig.3.9). Fig.3.7 shows analysis increment (top panel) and the difference between truth and first guess (bottom panel). Analysis increment has a very similar structure as the difference between truth and first guess, which means the 3DVAR scheme succeeds as expected.

3.3 EnKF implementation

3.3.1 Serial EnSRF

We now apply the same core modules used for serial EnSRF in the Lorenz-96 model (Section 2.3) to the SPEEDY model. Appendix A describes the core modules. The core subroutines require one-dimensional arrays for inputs and outputs, all the variables are combined together to form a one-dimensional array. All inter-variable correlations are considered.

As described in Section 2.2.4, localization is very important to avoid sampling errors caused by the limited ensemble size. As for horizontal localization, distance is computed using the grid spacing unit. Thus, skewness of the map in the physical space is not considered. The boundary is cyclic in longitudinal direction, whereas in latitudinal direction, the boundary is treated as a solid wall. The Schur product with the Gaussian-like weighting function (eq.(2.49)) is applied in this grid spacing distance.

As in 3DVAR, errors in different vertical levels are considered to be independent. Surface pressure (p_s) is assumed to be correlated only with variables at the bottom level ($\sigma = 0.950$). Thus, p_s changes because of observations of p_s itself and all other variables

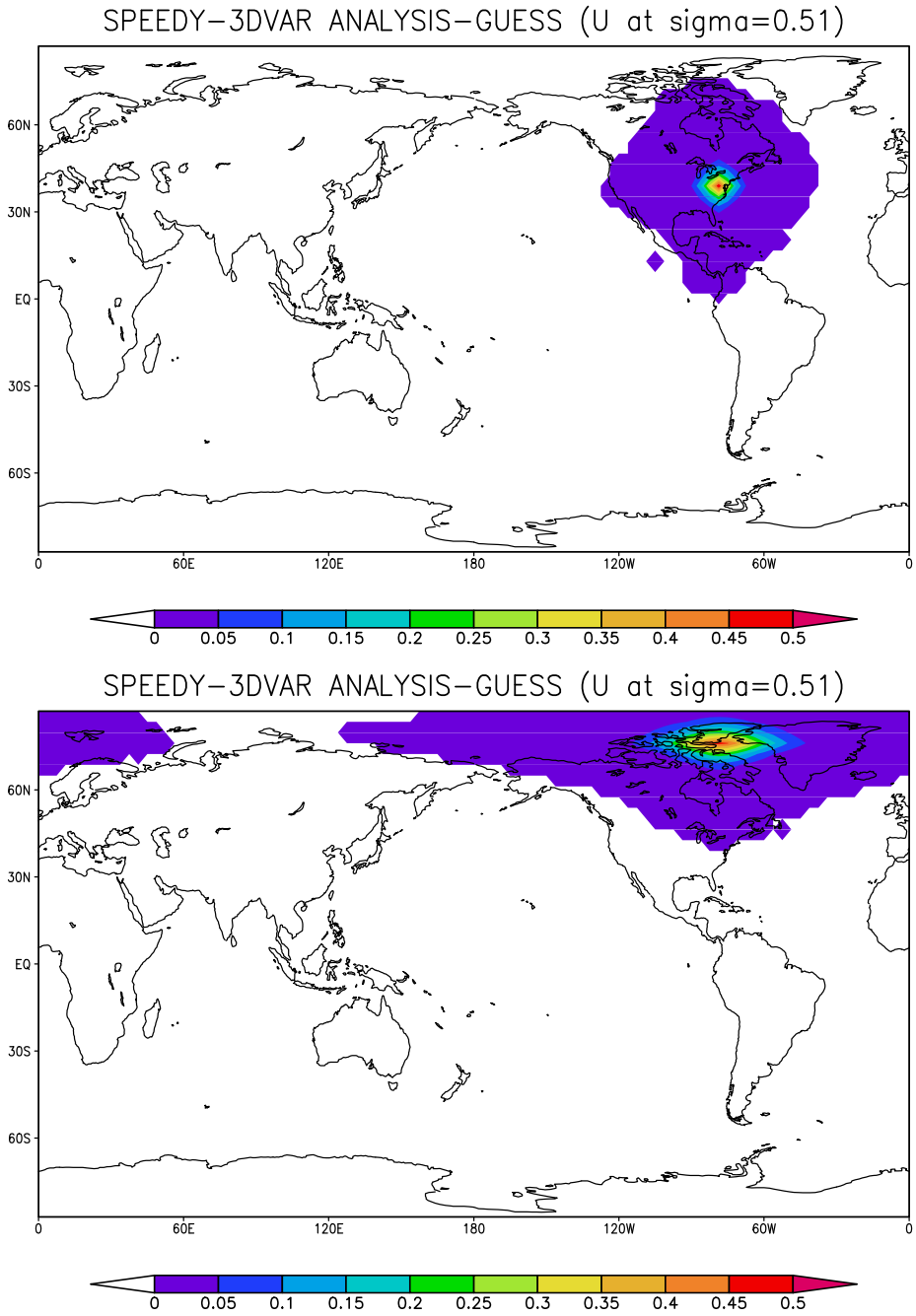

Figure 3.5: Spatial response of the 3DVAR in the case of a single observation with the observational increment of 1.0m/s at mid-latitude (top panel in m/s) and high-latitude (bottom panel in m/s) at the 4th level ($\sigma = 0.51$).

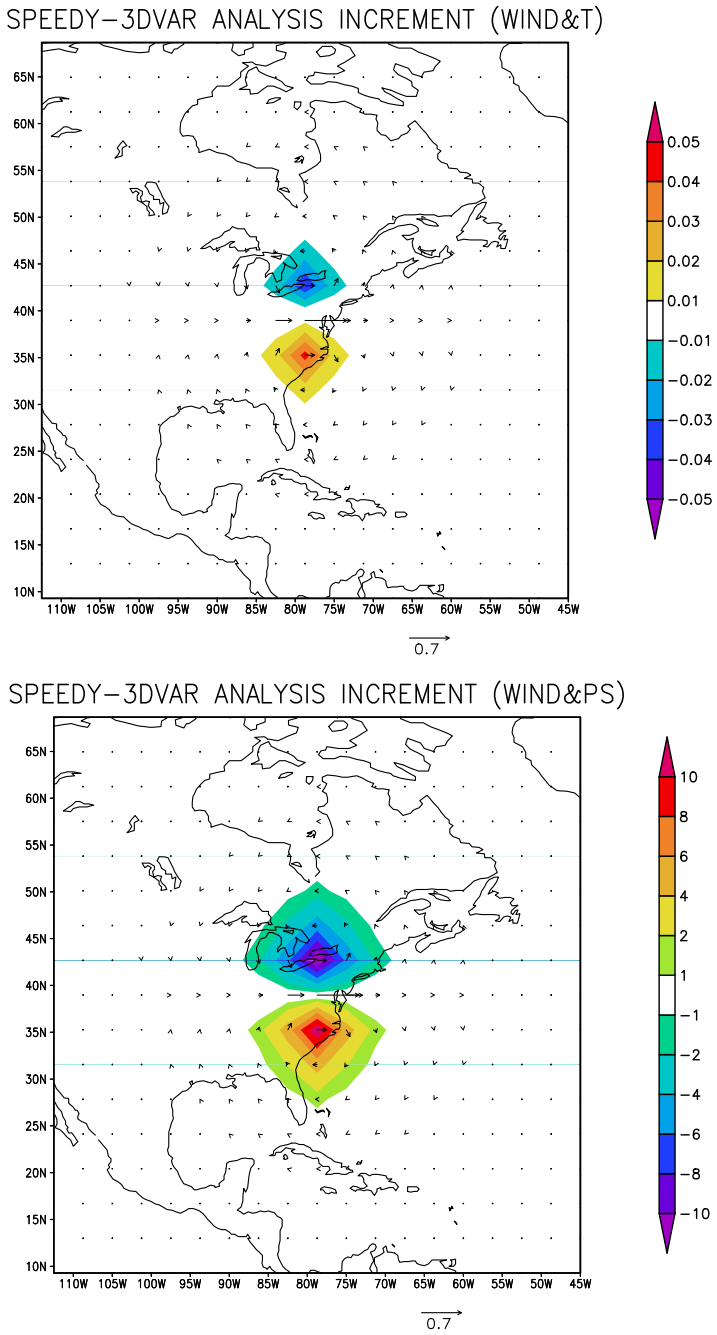

Figure 3.6: Inter-variable response of the 3DVAR in the case of a single observation with the observational increment of 1.0m/s at mid-latitude at the 4th level. Analysis increment of T (top panel in K) and p_s (bottom panel in Pa) are shown.

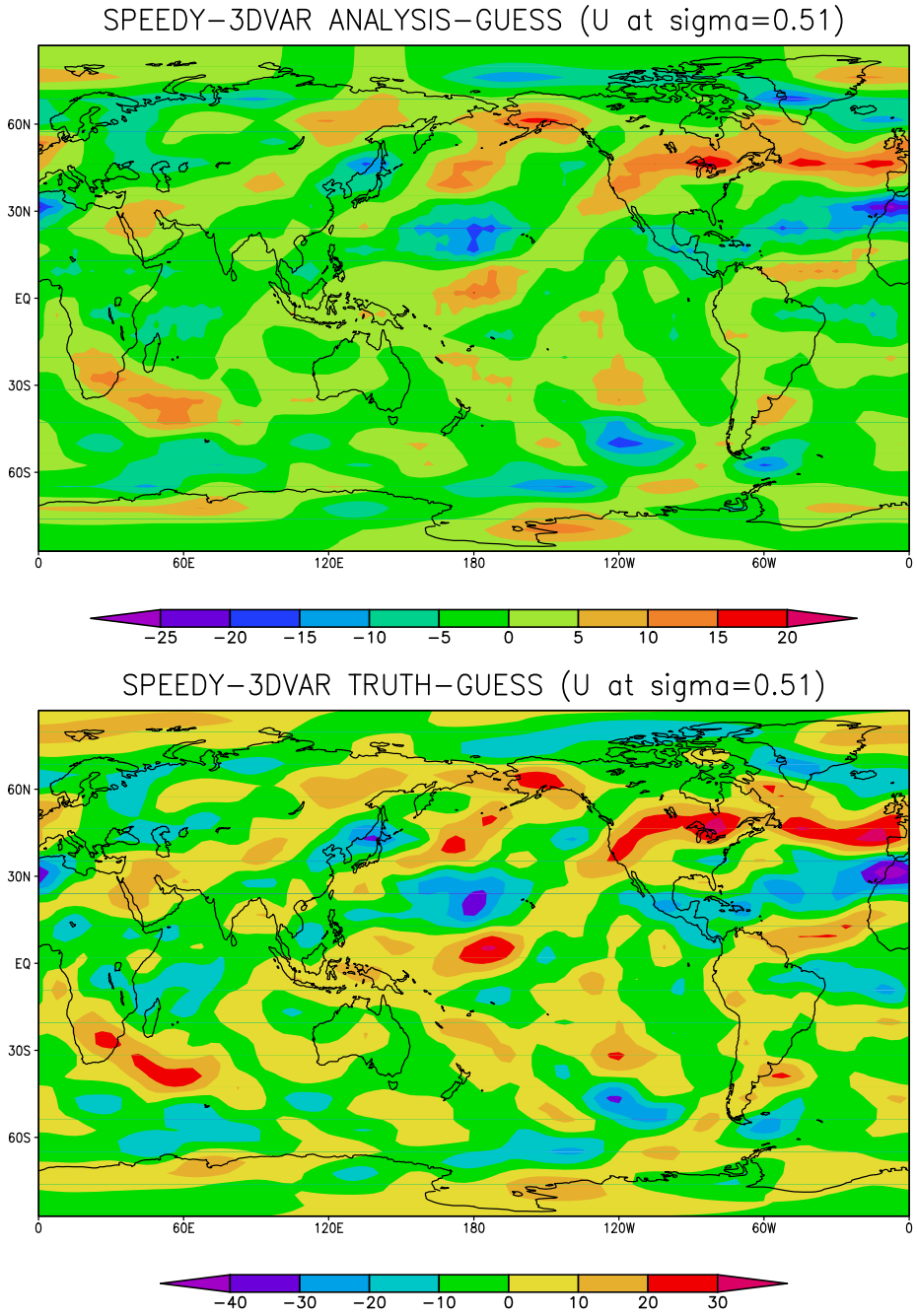

Figure 3.7: Analysis increment of 3DVAR (top panel) and the difference between truth and first guess (bottom panel). The fields are u-wind fields (in m/s) at the 4th level.

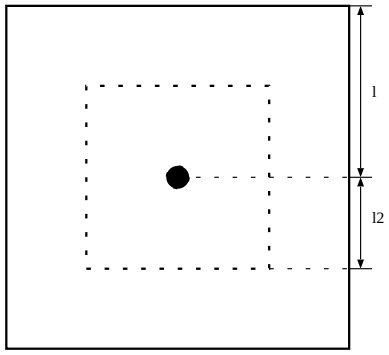

Figure 3.8: Definition of the local regions around a center point for LEKF in the SPEEDY model. The larger region denoted by l is the local patch of LEKF, the smaller region denoted by l_2 is the subset of the local patch that is averaged to obtain the global analysis. Since l denotes a half of a side, the area of the local patch becomes $(2l+1) \times (2l+1)$, similarly for l_2.

(u, v, T, q) at the bottom level, and vice versa.

3.3.2 LEKF

In implementing the LEKF for the SPEEDY model, we use the same core modules from the Lorenz-96 model (Section 2.3). Appendix A describes the core modules. The inputs and outputs of the core subroutines are one-dimensional arrays. Thus, the model variables are projected to the local space, and the local variables are combined together to be stored in a one-dimensional array.

Fig.3.8 shows the local patch, which is a two-dimensional expansion of Fig.2.2 in the case of the Lorenz-96 model. The region is a square in the grid space, the physical shape is not a square because of the skewness of the map in higher latitudes. Since the local patches are overlapped, only the values inside the smaller region denoted by l_2 are averaged to obtain the global analysis.

Cyclic boundary conditions are applied in the longitudinal direction. In the latitudinal direction, the point next to a point (i, j) where i denotes longitude and j denotes the highest latitude is taken as $(i + 180, j)$. In this way, the same size of the local patch is taken everywhere. This treatment is the same as Szunyogh et al. (2005).

As for the vertical localization, similar to the serial EnSRF, errors at different levels are assumed to be independent except surface pressure (p_s) unless otherwise noted. p_s is combined to other variables at all levels and thus, p_s observations affect other variables at all levels. However, only the variables at the bottom level affect p_s. In this way, the combination of p_s and other variables is one-way in upper levels, whereas it is two-way only at the bottom level. This treatment is also the same as Szunyogh et al. (2005).

In general, the vertical independence is not a good assumption, the system is developed so that it can consider vertical correlations. For comparison with 3DVAR and serial EnSRF, vertical correlation is not considered, but the impact of vertical correlation is shown later.

3.4 Experimental setup

So far, we described the implementation of three methods (3DVAR, serial EnSRF, and LEKF). In this section, the condition of numerical experiments is described.

3.4.1 Observational network

We use three types of observational networks in the present research:

1. A regularly distributed dense network (Fig.3.9)
2. A regularly distributed sparse network (Fig.3.10)
3. An irregularly distributed realistic network (Fig.3.11)

Variable	Observational error standard deviation
U-wind (u)	1.0 [m/s]
V-wind (v)	1.0 [m/s]
Temperature (T)	1.0 [K]
Specific humidity (q)	0.0001 [kg/kg]
Surface pressure (p_s)	100.0 [Pa]

Table 3.2: Observational error standard deviation

Fig.3.9 shows the dense network that consists of 1056 total sounding stations (about 23 percent of all grid points) which are chosen regularly as 1 station out of every 4 grid points excluding higher latitudes than 80 degrees north and south. Fig.3.10 shows the sparse network that consists of 264 total sounding stations (about 6 percent of all grid points) which are chosen regularly as 1 station out of every 16 grid points excluding higher latitudes. Fig.3.11 shows the realistic observational network consists of 415 total sounding stations (about 9 percent of all grid points) that simulate the real operational radiosonde stations. The network is realistic in that more stations exist over land, especially in Western Europe, Eastern Asia, and North America, with fewer stations over the ocean.

It is assumed that all sounding stations are located at model grid points so that no interpolation is required. Each sounding station has complete vertical soundings, that is, all variables are observed at all sigma levels. The observational error standard deviations are shown in Table 3.2.

The dense observational network is dense enough for 3DVAR to work almost perfectly as shown later. Thus, in the dense network, there would be no significant difference between EnKFs and 3DVAR. When the number of observations is reduced, advanced

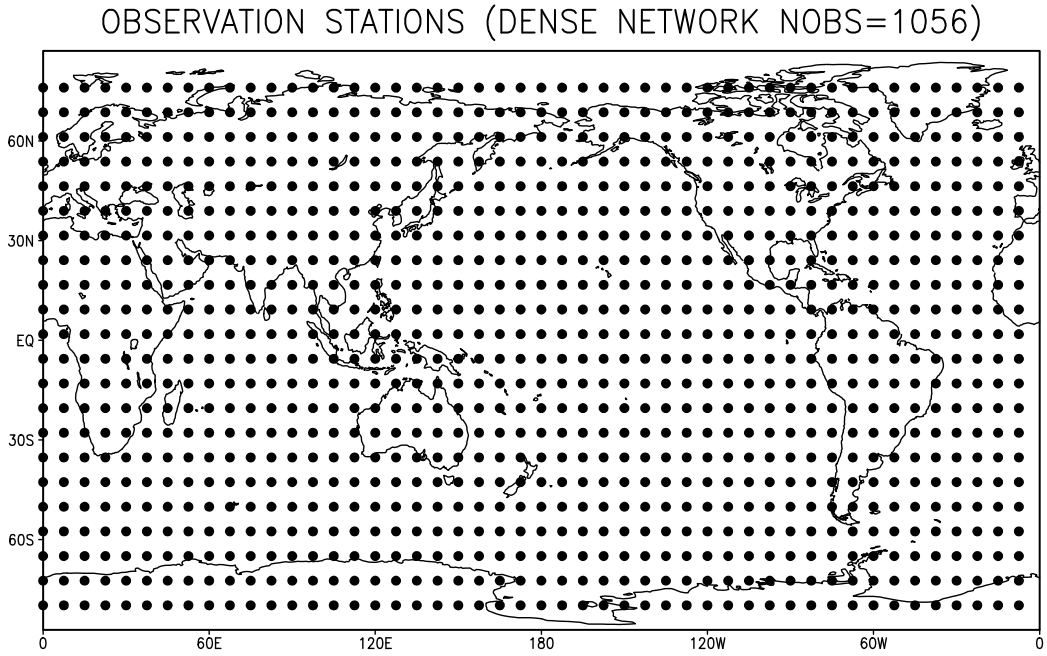

Figure 3.9: A regularly distributed dense observational network. 1056 stations (23 percents of all grid points) are located at one grid point out of every four grid points.

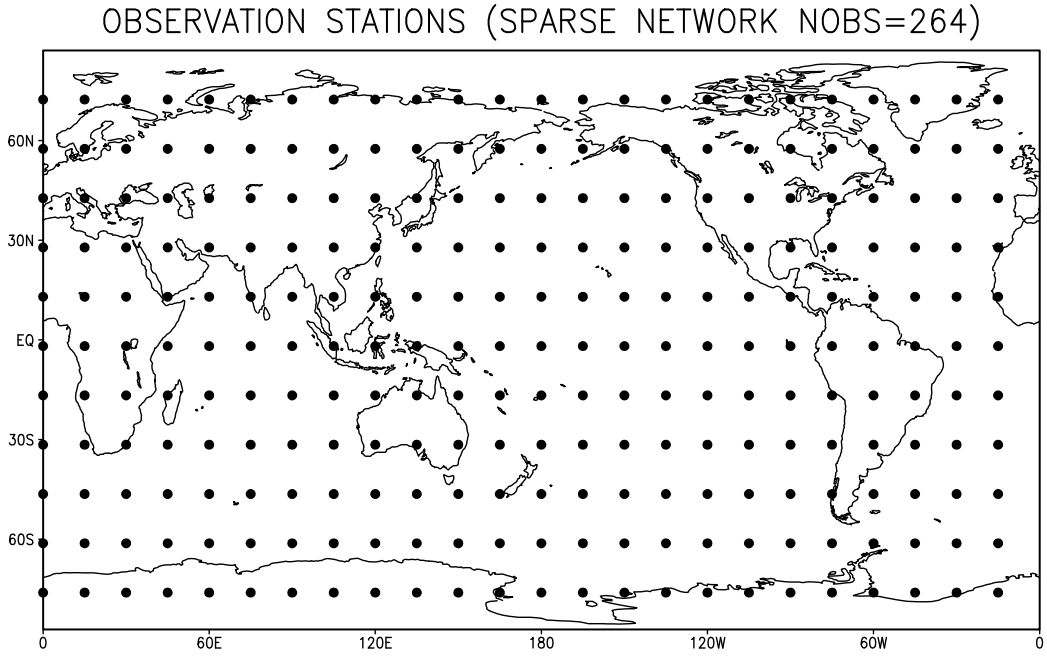

Figure 3.10: A regularly distributed sparse observational network. 264 stations (5.7 percents of all grid points) are located at one grid point out of every 16 grid points.

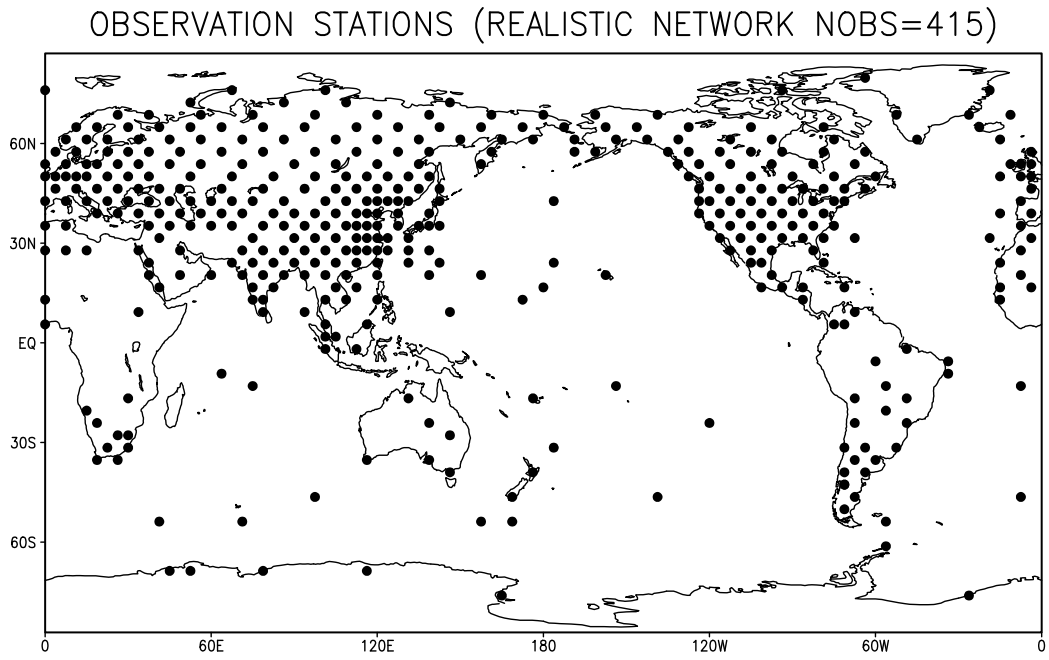

Figure 3.11: An irregularly distributed realistic observational network. 415 stations (9.0 percents of all grid points) are located mostly over continents in the northern hemisphere.

methods are expected to show their advantage. In the sparse network, 3DVAR is expected to show large errors, whereas it is expected that EnKFs show much smaller errors. Therefore, we compare the methods using the sparse network.

3.4.2 Data assimilation experiments

As described in Section 1.3, the data assimilation experiments are based on observational systems simulation experiments (OSSEs) where the nature run is assumed to be known, and observational data are simulated by adding random noise to the nature run according to the observational errors. The nature run is created by the model integration after a 1-year spin-up period. By default, the SPEEDY model starts from the atmosphere at rest ($u = 0$, $v = 0$) on the initial date January 1, 1981. We choose January 1, 1982 as the initial date for data assimilation experiments after 1-year integration. We use the same

nature run for all data assimilation experiments in this chapter.

We generate another independent nature run starting from the atmosphere at rest at a different time. The difference between the two nature runs temporally averaged in February, 1982 is 106.37 meters measured by the RMS difference of 500hPa height field, the error magnitude without data assimilation. We choose fields at arbitrary times from the second nature run to obtain the initial conditions of data assimilation. For 3DVAR, a field is chosen. For EnKF, we choose fields at arbitrary times for the initial ensemble. The same initial ensemble is used for serial EnSRF and LEKF. If the ensemble size is infinitely large, the ensemble mean becomes the climatology. Thus, this is equivalent to beginning the data assimilation cycle with climatology in 3DVAR. This choice is practical in the sense that it is always available just by running the model, even if we have no idea about the nature run.

Fig.3.12 shows a schematic of data assimilation experiments on the SPEEDY model. The forecast field ("FORECAST") by the SPEEDY model gives the first guess field for the data assimilation system ("DA"), and the true field ("TRUTH") with random noise provides observations. "DA" may be 3DVAR, serial EnSRF, or LEKF. The output of "DA" is the analysis field ("ANALYSIS") that provides the initial condition for the next model integration. "ANALYSIS" contains high frequency components, whereas "INITIAL", the output by the SPEEDY model at forecast time 0, does not because the spectral transformation of the model smoothes out the high frequency components. Since the SPEEDY model does not see the high frequency components, the smoothed "INITIAL" field is considered as the analysis field for verification. Thus, the analysis error is defined as the difference between "INITIAL" and "TRUTH".

We use the multiplicative covariance inflation (eq.(2.66)) both for serial EnSRF and

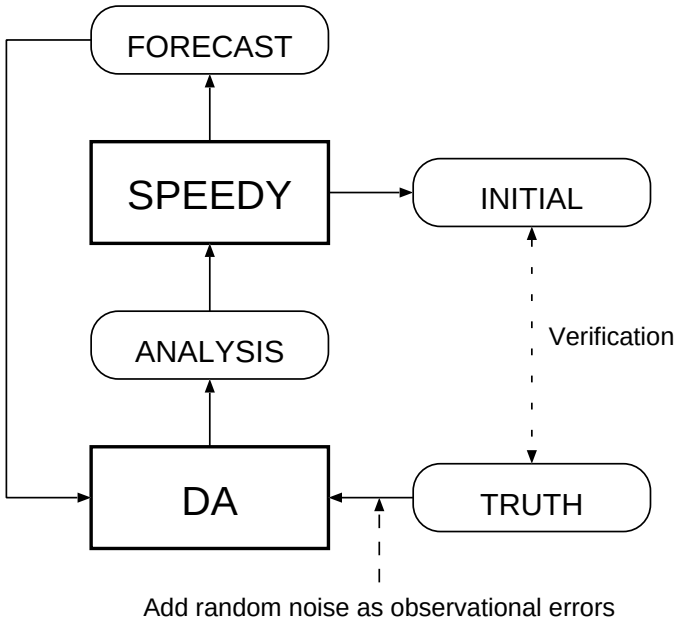

Figure 3.12: Schematic of data assimilation experiments on the SPEEDY model. "DA" denotes the data assimilation system. The "initial" field is the output by the SPEEDY model which is different from the "analysis" field because the high frequency components are smoothed out by the spectral transformation of the model. The difference between "initial" and "truth" is defined as the analysis error.

LEKF. The enhanced variance inflation (eq.(2.67)) has been tested, but the filter quickly diverges in the SPEEDY model. The inflation factor is fixed to $\delta = 0.04$, that is, 4% covariance inflation, the same as in Szunyogh et al. (2005). This choice of inflation factor is reasonable to make the EnKFs stable with reasonable choices of localization scale and ensemble size.

3.5 Results

Following the development of the system, we perform analysis-forecast cycle experiments. In this section we show the results of the three data assimilation systems (3DVAR, serial EnSRF, LEKF) and compare them.

3.5.1 3DVAR

Fig.3.13 shows the analysis RMSE of 500hPa height field. The solid line shows the result in the case of a dense observational network. The dense observational network is dense enough for 3DVAR to perform almost perfectly; the analysis RMSE of 500hPa height is less than 5 meters, much smaller than the operational NWP system for the real atmosphere. RMSE averaged over one month after the initial one-month spin-up period is 3.58 meters. When the observations are decreased and the sparse network is applied, the performance of 3DVAR worsens as expected. The dashed line shows the performance of 3DVAR using the same background error statistics optimal for the dense network. Since the background error standard deviation of the sparse network is expected to be larger than that of the dense network, the error standard deviation is multiplied by 2.0 giving better performance (dotted line). The background error statistics have been regenerated using the sparse network. The new background error information for the sparse network

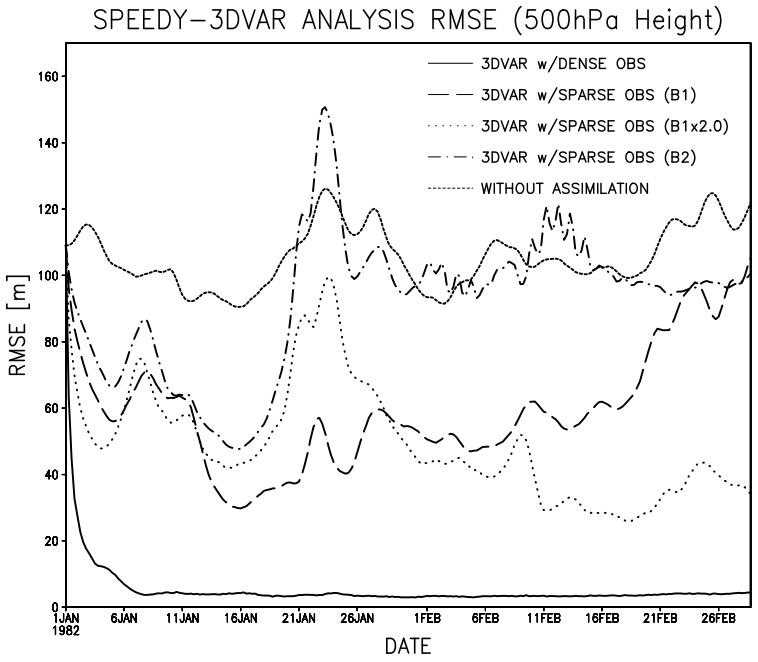

Figure 3.13: Analysis RMSE of 500hPa height field in the case of 3DVAR with the dense observational network (solid line), sparse observational network using the same background error statistics based on the dense network (dashed line), sparse observational network using background error standard deviation multiplied by 2.0 (dotted line), and sparse observational network using background error statistics based on the sparse network (dot-dashed line).

gives worse 3DVAR performance than the original background error information based on the dense network (dot-dashed line). The three lines of 3DVAR with the background statistics based on the dense network always outperform the run with no data assimilation, which supports the usefulness of 3DVAR even in the sparse observational network.

The present 3DVAR system can consider full spatial dependencies of error standard deviation, but we can easily simplify the dependencies by averaging the error standard deviation in space. Fig.3.14 shows analysis RMSE of 500hPa height in the case of the sparse observational network with different spatial dependencies of the error standard deviation.

The solid line shows the results when we consider the full spatial dependencies of the error standard deviation, the same as the dotted line in the previous Fig.3.13. The dashed line shows the results in the case of horizontal mean of standard deviation, in which case the standard deviation is a function of only the vertical levels and variables. The dotted line shows the results in the case of zonal mean of standard deviation, in which case, the standard deviation is a function of latitudes, vertical levels, and variables. All the three performances are similar, although the horizontal mean gives better performance generally. Table 3.3 shows the average performance of the three cases. The RMSE is averaged for one month (112 samples) after the initial one-month spin-up period. The simpler structure of background error standard deviation gives the better 3DVAR performance, suggesting that the full spatial structure of the error covariance obtained with 87 samples is dominated by sampling errors.

Since we observed the best performance (31.14 meters) when we multiplied the background error standard deviation by 2.0 and took the horizontal mean of the original error variance obtained from NMC-method statistics, hereafter, we use this configuration as an optimal setting of 3DVAR.

In order to see the error distributions in space, analysis RMSE of 500hPa height field is plotted in space. Analysis errors are averaged only in time to obtain the spatial distribution. Fig.3.15 shows the horizontal structure of analysis RMSE of 500hPa height field. Shades and contours show analysis RMSE and mean analysis fields, respectively. Dots indicate observational locations. Large errors exist in polar regions where no observations are available. We see a lattice-like pattern generated by the regularity of observational locations. The top panel of Fig.3.16 shows the analysis error bias. We can also see large analysis error bias in the polar region, and we can still find the lattice-like pattern. We

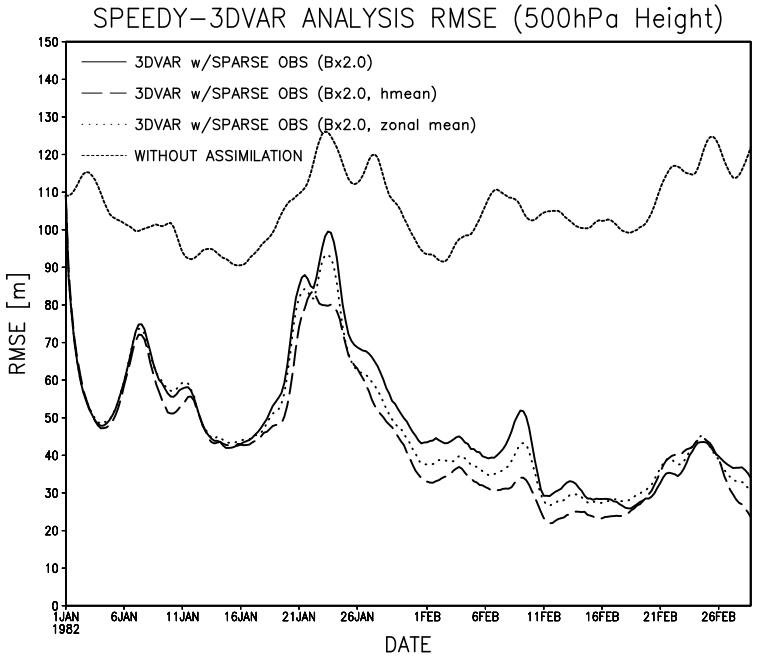

Figure 3.14: Analysis RMSE of 500hPa height field in the case of the sparse observational network with full spatial dependence of background error standard deviation (solid line), with horizontally uniform background error standard deviation (dashed line), and with zonally uniform background error standard deviation (dotted line).

	3DVAR(full)	3DVAR(zonal mean)	3DVAR(hmean)	NO ASSIM
RMSE[m]	36.65	34.62	31.14	106.37

Table 3.3: Analysis RMSE of 500hPa height field in meters using 3DVAR on the SPEEDY model when the sparse observational network is applied. The RMSE is temporally averaged for a month (112 samples) after the initial one-month spin-up period. "3DVAR(full)" denotes the case that full spatial dependence of background error standard deviation is considered. "3DVAR(zonal mean)" and "3DVAR(hmean)" denote the cases that zonally and horizontally uniform background error standard deviation is considered, respectively. "NO ASSIM" denotes the case that no data assimilation is applied.

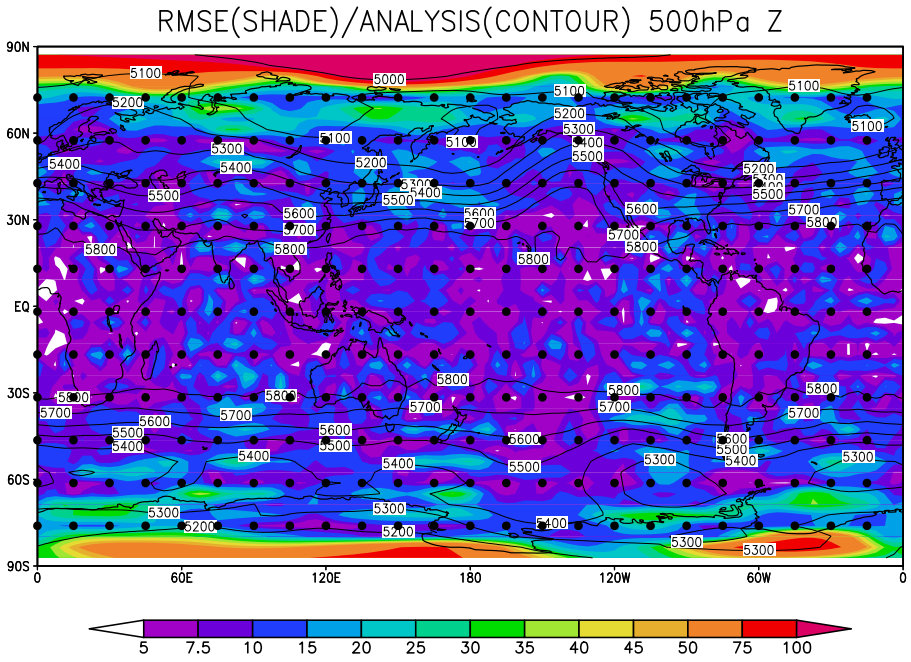

Figure 3.15: Horizontal structure of analysis RMSE of height field in the case of 3DVAR with horizontally uniform background error standard deviation and the sparse observational network. Shades and contours show analysis RMSE (in meters) and mean analysis fields (in meters), respectively. Dots indicate observational locations.

cannot neglect the error bias in this one-month period even if the model is perfect. The bias field shows a clear signal in a certain wavelengths, mostly the wavelength of distance between observations. In the longitudinal direction in higher latitudes, longer wavelengths show a stronger signal. It seems the regular observational network introduces a regular bias structure. If we subtract error bias from the analysis error field and recalculate the analysis RMSE, i.e. standard deviation, the lattice-like pattern of the analysis RMSE is reduced as shown in the bottom panel of Fig.3.16. In addition, the large errors in polar regions are reduced, thus, the error bias explains a large amount of the polar errors.

So far, we have examined the analysis RMSE of 500hPa height field. Although this field is often used as a proxy of the performance of a forecast model and a data

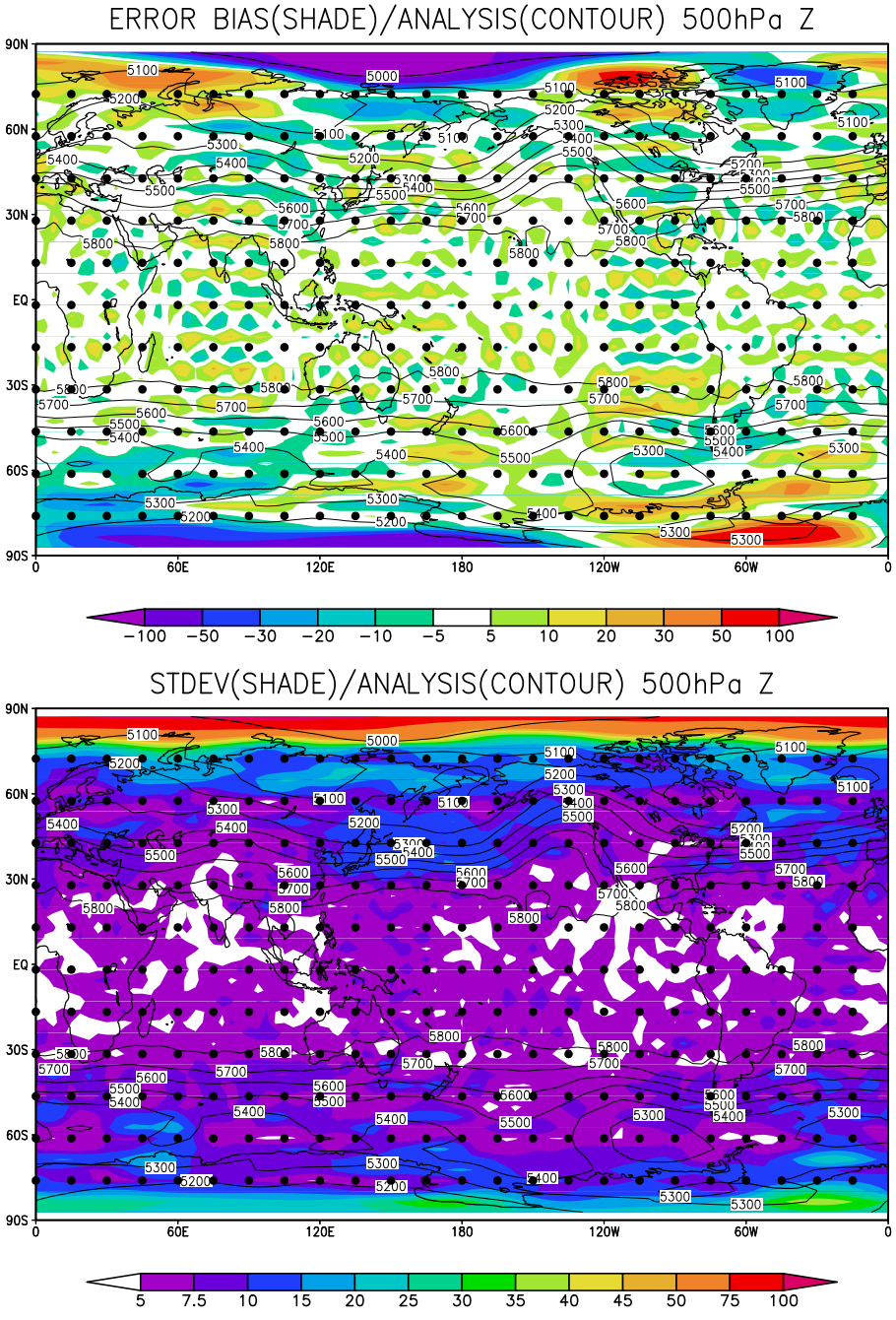

Figure 3.16: Horizontal structure of analysis error bias (top panel) and standard deviation (bottom panel) of height field (in meters) in the same case as Fig.3.15. Shades show analysis error bias (top panel) and standard deviation (bottom panel), contours show mean analysis fields. Dots indicate observational locations.

assimilation system, it is important to see other variables at other levels. Fig.3.17 shows the zonal structures of analysis RMSE of the u-wind (top panel) and temperature (bottom panel). Shades and contours show analysis RMSE and mean analysis fields, respectively. If we ignore the large errors in the polar regions, we see large analysis error in the tropical regions around the 200hPa level in both figures. In addition, both figures show striped pattern, where the minima correpond to observational locations. For the temperature field (bottom panel), we see the tendency for larger errors to appear in upper levels in the tropics. The error is also large in lower levels in middle and high latitudes. This feature is similar to what Szunyogh et al. (2005) obtained.

The results using 3DVAR are summarized as follows:

- The 3DVAR data assimilation cycle experiment using a dense observational network shows almost perfect performance: 3.58m RMSE of 500hPa height field.
- The 3DVAR data assimilation cycle experiments using the sparse observational network show much worse performance: 31.14m RMSE of 500hPa height field.
- Horizontally uniform background error standard deviation results in better 3DVAR performance, although the 3DVAR system can consider full spatial dependence, suggesting that a sample of 87 is not enough for the full spatial dependence.
- Background error statistics estimated with the dense observational network have a good structure for data assimilation with the sparse observational network.
- Horizontal distribution of analysis RMSE shows a lattice-like pattern generated by the regularity of observational locations.
- Zonal mean structure of analysis RMSE shows a striped pattern corresponding to the observational locations.

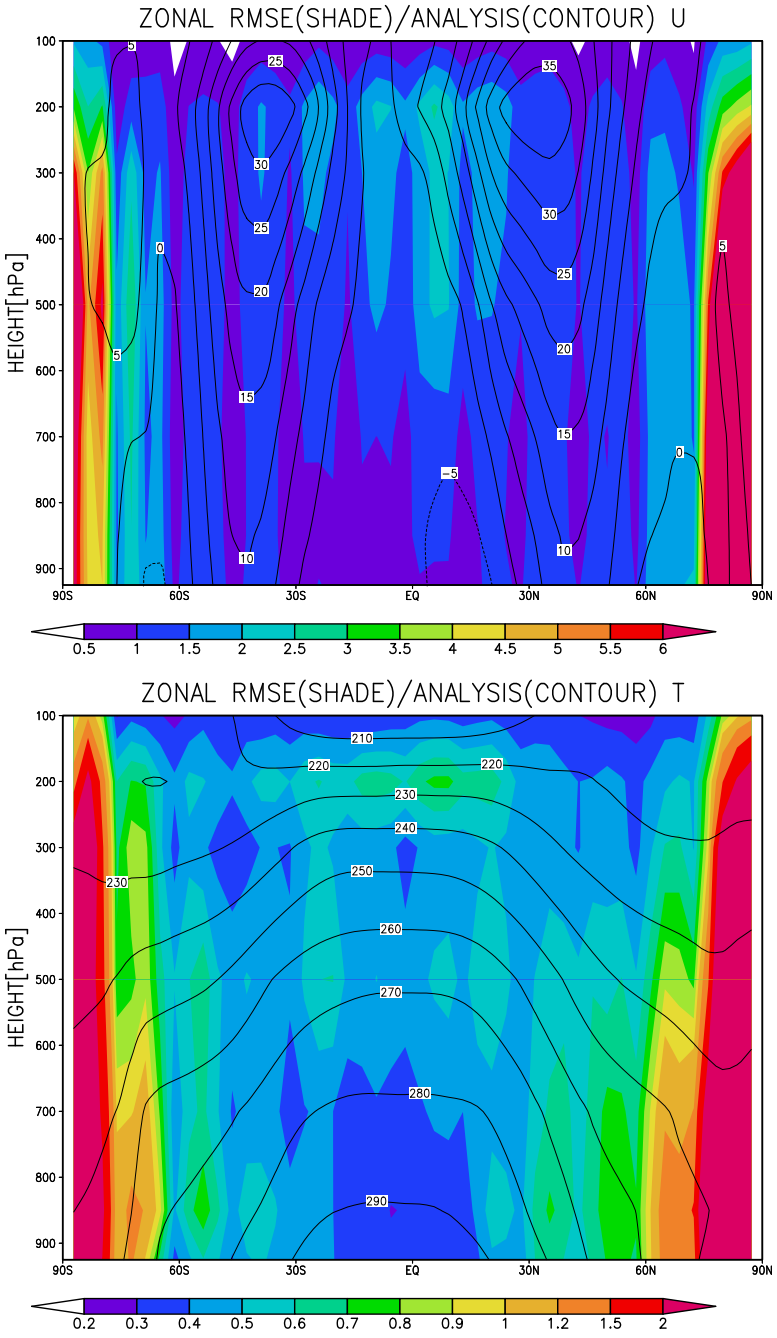

Figure 3.17: Zonal structure of analysis RMSE of u-wind field (top panel, in m/s) and temperature field (bottom panel, in K) in the case of 3DVAR with horizontally uniform background error standard deviation and the sparse observational network. Shades and contours show analysis RMSE and mean analysis fields, respectively.

NBV / Schur	1	2	3	4	5	6
10	11.74	24.60	40.17	D	D	D
20	10.80	10.69	10.99	D	22.37	D
30	15.07	5.47	4.99	8.00	10.48	17.02

Table 3.4: Analysis RMSE of 500hPa height field in meters using serial EnSRF on the SPEEDY model when the sparse observational network is applied. The RMSE is temporally averaged for a month (112 samples) after the initial one-month spin-up period. For comparison, the analysis RMSE of 3DVAR in the same period is 31.14. The ensemble size (NBV) is chosen as 10, 20, or 30, and the horizontal length scale of the Schur product is chosen as integers from 1 to 6. "D" denotes filter divergence. 4% multiplicative covariance inflation is applied.

3.5.2 Serial EnSRF

Table 3.4 shows the analysis RMSE of 500hPa height field. The ensemble size is chosen to be 10, 20, and 30, and the horizontal correlation length scale for the Schur product is chosen from the integers between 1 and 6. Even with 10 ensemble members, serial EnSRF outperforms 3DVAR whose best result shown in the previous section was 31.14m. The best performance of serial EnSRF is seen in the case of 30 ensemble members with the horizontal scale of 3.0, which shows almost perfect performance (as good as 3DVAR with the dense observational network).

A smaller ensemble size requires a smaller horizontal length scale of the Schur product. The data assimilation cycle is more stable for a larger ensemble size. Both results are similar to those of the Lorenz-96 model.

Fig.3.18 shows temporal sequences of the analysis RMSE of 500hPa height field in

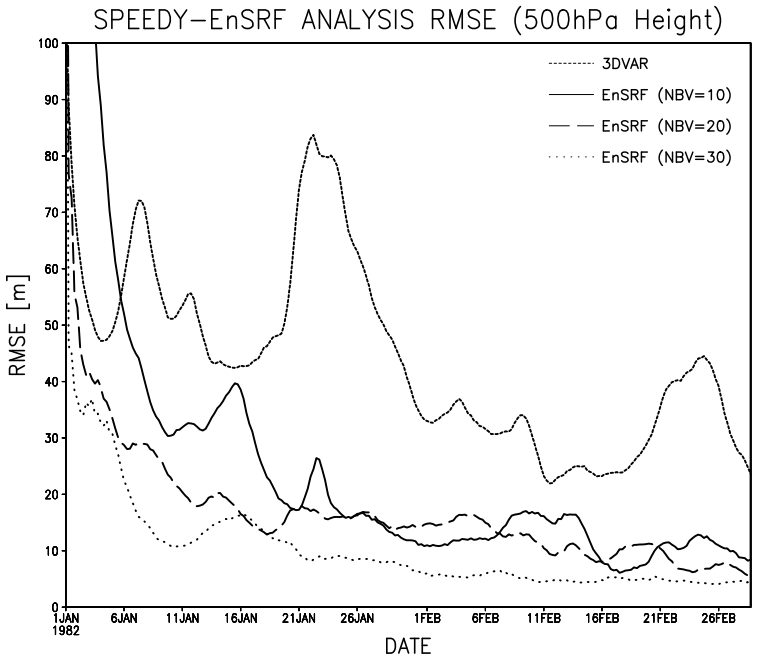

Figure 3.18: Analysis RMSE of 500hPa height field in the case of the sparse observational network using serial EnSRF with 10 (solid line), 20 (dashed line), and 30 ensemble members (dotted line). The localization length scale of the Schur product is chosen as optimal. For comparison, the RMSE of 3DVAR is shown as the short dashed line.

the cases of serial EnSRF with 10 (solid line), 20 (dashed line), and 30 ensemble members (dotted line). The localization length scale of the Schur product is chosen as optimal, that gives the best result in each case in the Table 3.4 (marked by rectangles). For comparison, the RMSE of 3DVAR is shown as the short dashed line. Serial EnSRF clearly outperfoms 3DVAR. The case of 30 ensemble members clearly outperforms others almost everywhere.

In order to see the error distributions in space, analysis RMSE of 500hPa height field is plotted in space, corresponding to Figs.3.15 and 3.16 to be compared with 3DVAR. Fig.3.19 shows the horizontal structure of analysis RMSE of 500hPa height field, similar to Fig.3.15. Large errors exist in polar regions where no observation is available, but the errors are smaller than those in 3DVAR. We can still see a lattice-like pattern cor-

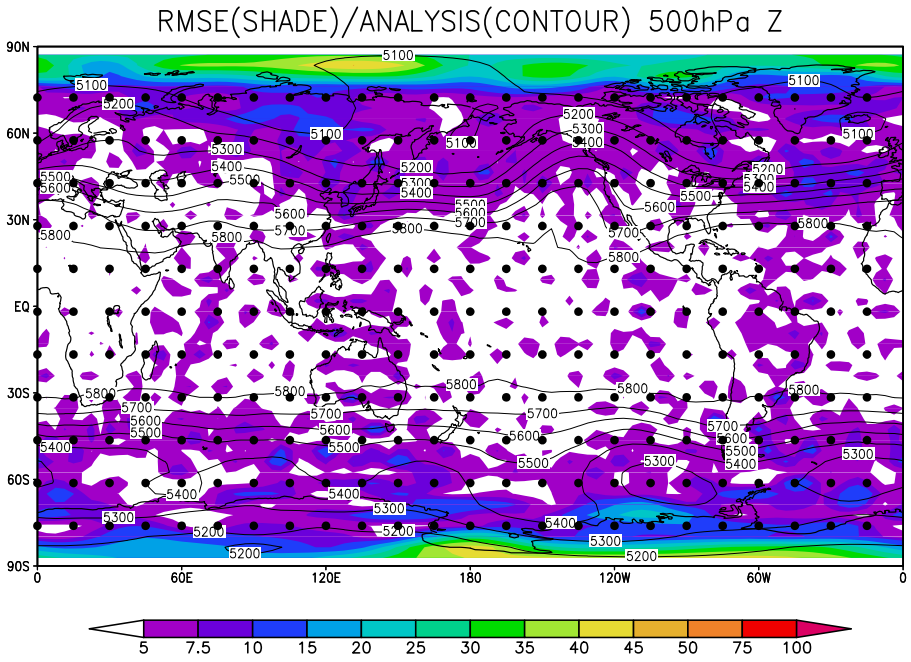

Figure 3.19: Horizontal structure of analysis RMSE of height field, similar to Fig.3.15, in the case of serial EnSRF with 20 ensemble members and localization scale of 2.0. Shades show analysis RMSE (in meters), contours show mean analysis fields (in meters). Dots indicate observational locations.

responding to the observational locations. The top panel of Fig.3.20 shows the analysis error bias. Basic characteristics remain the same as the case of 3DVAR: large errors with low wavenumbers exist in polar regions. We find a lattice-like pattern with the same wavelength as the distance between observational points. The bottom panel of Fig.3.20 shows analysis error standard deviation, it is obtained in the same manner as RMSE but subtracting error bias. The standard deviation field shows almost no lattice-like pattern. Thus, we can attribute the lattice-like pattern to the error bias. In addition, the large errors in polar regions are reduced, thus, the error bias explains the majority of the polar errors.

Fig.3.21 shows the analysis RMSE of four variables at all seven pressure levels. As in

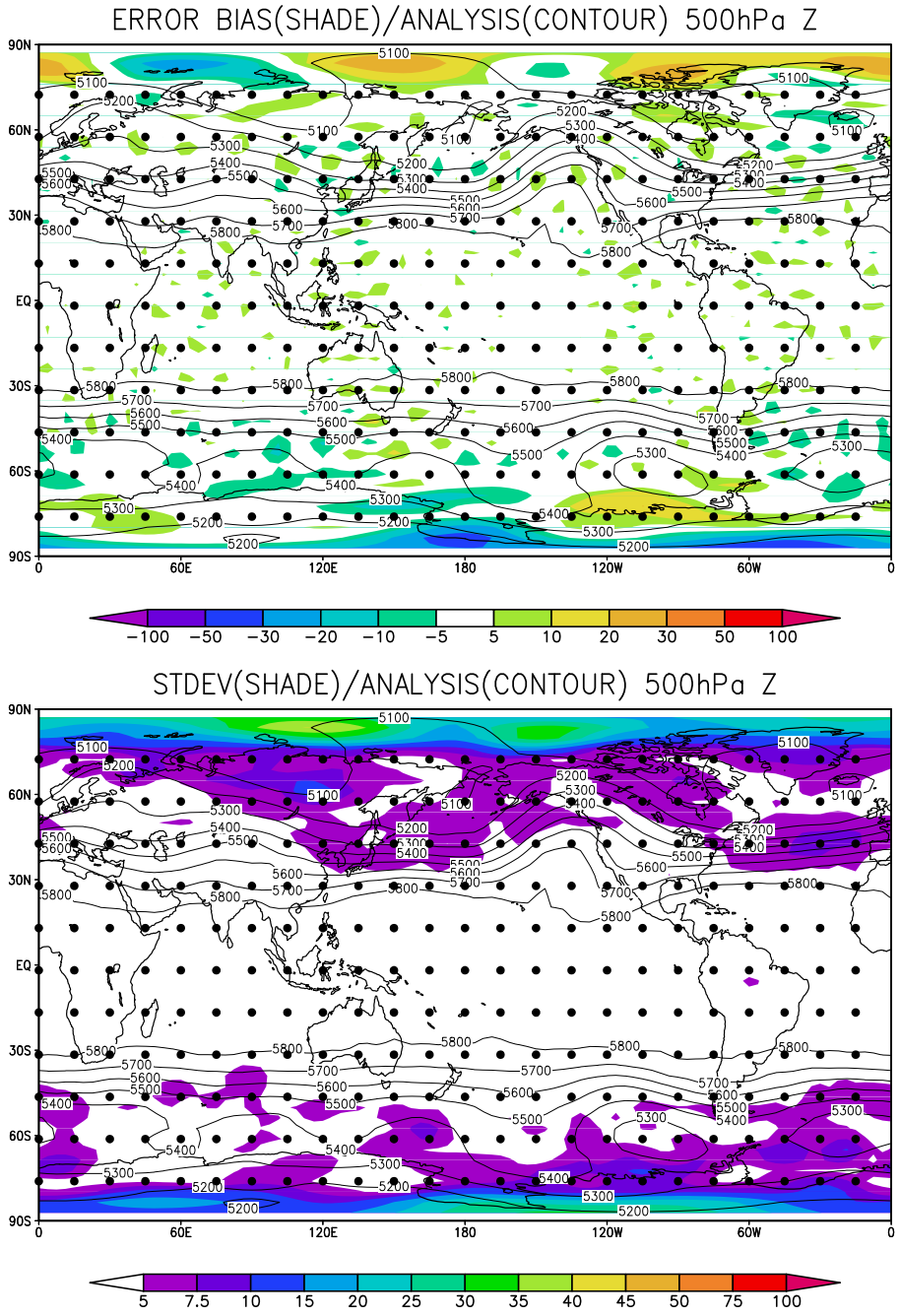

Figure 3.20: Horizontal structure of analysis error bias (top panel) and standard deviation (bottom panel) of height field (in meters), similar to Fig.3.16, in the case of serial EnSRF with 20 ensemble members and localization scale of 2.0. Shades show analysis error bias (top panel) and standard deviation (bottom panel), contours show mean analysis fields. Dots indicate observational locations.

Table 3.4, RMSE is averaged for one month after the initial one-month spin-up period (112 samples), so that we get a single number for one variable at one level. Fig.3.21(a) shows the analysis RMSE of u-wind fields. A smaller RMSE is observed at levels higher than 300hPa, this trend is clearer in 3DVAR. As expected, the larger ensemble size gives better performance everywhere, although the cases of 10 and 20 ensemble members are similar. The case of 30 ensemble members clearly outperforms the others. It is noted that only the case of 30 ensemble members gives smaller RMSE than observational error standard deviation (1.0m/s, shown by the thin solid line) at all levels. Fig.3.21(b) shows the analysis RMSE of height fields, the serial EnSRF clearly outperforms 3DVAR at all levels. Serial EnSRF performs almost constantly at all levels, whereas 3DVAR performs better at upper levels. Fig.3.21(c) shows the analysis RMSE of temperature fields. Generally, higher altitude has smaller RMSE, especially clear in 3DVAR. All cases of serial EnSRF provide smaller RMSE than observational error standard deviation (1.0K, shown by the thin solid line) at all levels. Fig.3.21(d) shows the analysis RMSE of specific humidity fields. The order of the performances is basically the same as the previous figures (a)-(c). Higher altitudes are drier, so the RMSE is smaller in higher altitudes. Still, data assimilation provides useful information, and we can see the difference among different data assimilation schemes and settings in higher altitudes. Lower altitudes contain more humidity, and mostly the RMSE exceeds the observational error standard deviation (10^{-4}kg/kg, shown by the thin solid line).

Fig.3.22 shows the zonal structures of analysis RMSE of the u-wind (top panel) and temperature (bottom panel), corresponding to Fig.3.17 of 3DVAR. Shades and contours show analysis RMSE and mean analysis fields, respectively. It is noted that the errors in the polar regions are much smaller than those in 3DVAR. Basic features seen in 3DVAR

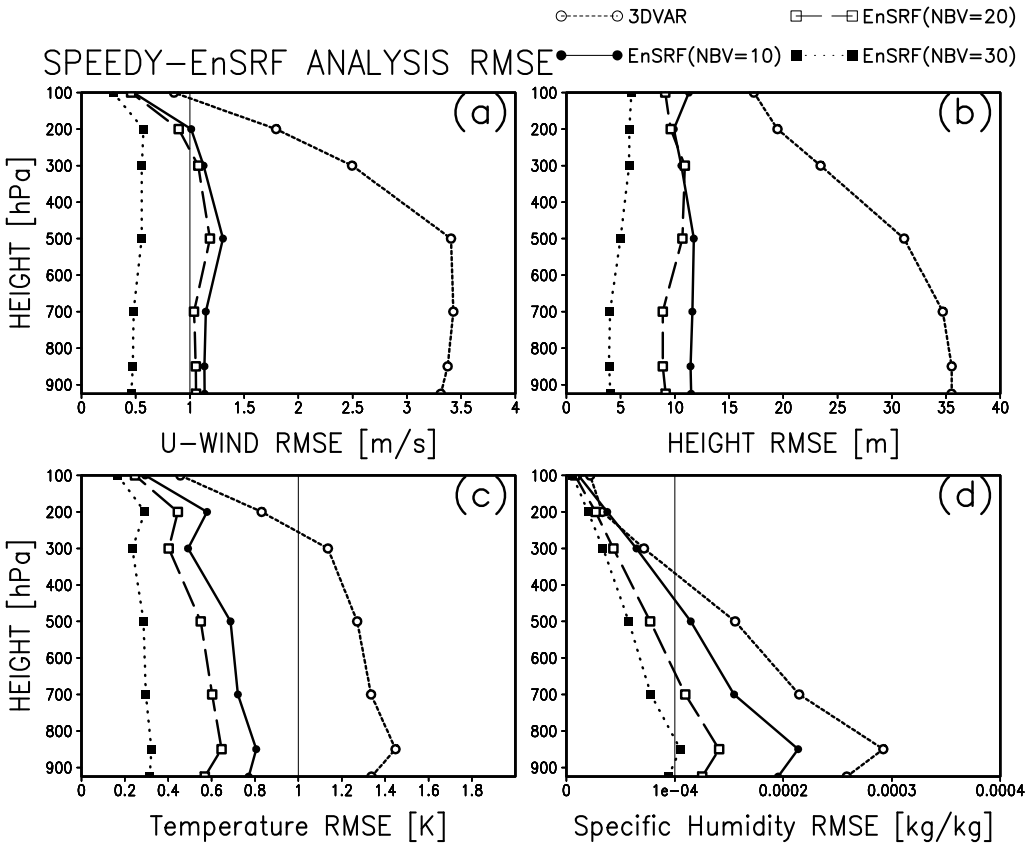

Figure 3.21: Analysis RMSE at the all pressure levels temporally averaged for one month (112 samples) after the initial one-month spin-up period in the case of the sparse observational network using serial EnSRF with 10 (solid line), 20 (dashed line), and 30 ensemble members (dotted line). The four panels (a), (b), (c) and (d) correspond to u-wind, height, temperature (T) and specific humidity (q) fields, respectively. The observational error standard deviations are shown as thin solid lines wherever applicable. The localization length scale of the Schur product is chosen as optimal. For comparison, the RMSE of 3DVAR is shown as the short dashed line.

remain similar, but with smaller amplitudes. The striped pattern seen in 3DVAR also still exists, but with much smaller amplitudes.

The results using serial EnSRF are summarized as follows:

- In the case of the sparse observational network, serial EnSRF clearly outperforms 3DVAR and shows almost perfect performance (4.99m RMSE of 500hPa height field with an optimal setting versus 31.14 for 3DVAR).
- The larger ensemble size, the better the performance.
- A smaller ensemble size requires smaller localization length scale, as observed in the Lorenz-96 model.
- Even with only 10 ensemble members, serial EnSRF outperforms 3DVAR.
- Horizontal and zonal distributions of analysis RMSE show lattice-like and striped patterns similar to 3DVAR but with smaller amplitudes, we can attribute the lattice-like pattern to the analysis error bias.
- Not only 500hPa height fields, but also other fields at other levels show much better performance than 3DVAR.

3.5.3 LEKF

Table 3.5 shows the analysis RMSE of 500hPa height field, similar to Table 3.4 in the case of serial EnSRF. The ensemble size is chosen to be 10, 20, and 30, and the local patch parameter l is chosen from 1 to 6. LEKF outperforms 3DVAR (31.14m at the best) for 20 and 30 ensemble members. The best performance is seen in the case of 30 ensemble members with the local patch parameter $l = 4$.

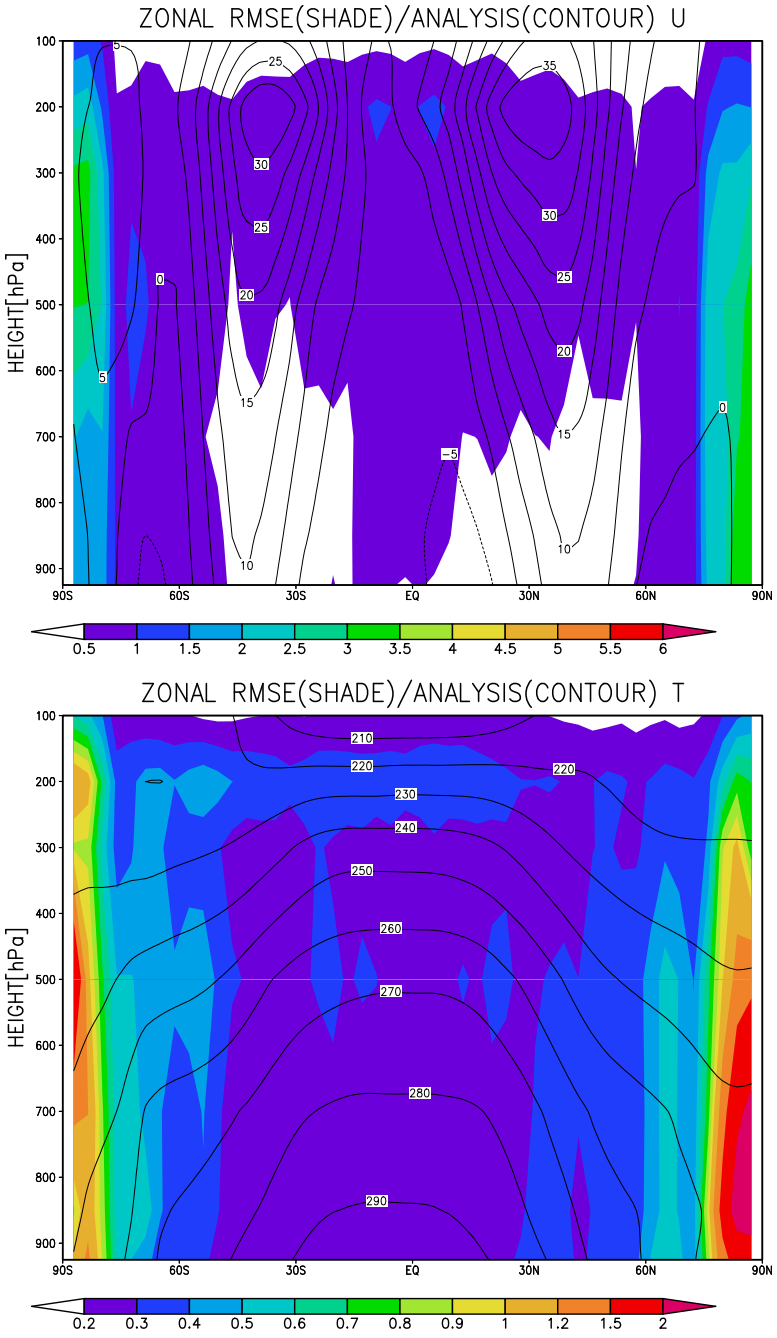

Figure 3.22: Zonal structure of analysis RMSE of u-wind field (top panel, in m/s) and temperature field (bottom panel, in K) in the case of serial EnSRF with 20 ensemble members and localization scale of 2.0. Shades and contours show analysis RMSE and mean analysis fields, respectively.

NBV / l	1	2	3	4	5	6
10	60.38	75.88	D	D	D	D
20	35.45	19.81	16.85	30.84	D	D
30	17.68	17.40	8.67	8.44	13.23	14.50

Table 3.5: Analysis RMSE of 500hPa height field in meters using LEKF on the SPEEDY model when the sparse observational network is applied. The RMSE is temporally averaged for a month (112 samples) after the initial one-month spin-up period. For comparison, the analysis RMSE of 3DVAR in the same period is 31.14. The ensemble size (NBV) is chosen as 10, 20, or 30, and the local patch parameter l is chosen as integers from 1 to 6. "D" denotes filter divergence. 4% multiplicative covariance inflation is applied.

A smaller ensemble size requires a smaller local patch size. The data assimilation cycle is more stable for larger ensemble size. Both results are the same as in the case of the Lorenz-96 model.

Fig.3.23 shows an LEKF version of Fig.3.18, which shows temporal sequences of the analysis RMSE of 500hPa height field in the cases of LEKF with 10 (solid line), 20 (dashed line), and 30 ensemble members (dotted line). The local patch parameter l is chosen as optimal, it gives the best result in each case in the Table 3.5 (surrounded by rectangles). For comparison, the RMSE of 3DVAR is shown as the short dashed line. LEKF with 20 and 30 ensemble members clearly outperfoms 3DVAR. The case of 30 ensemble members outperforms others almost everywhere.

In order to see the error distributions in space, analysis RMSE of 500hPa height field is plotted in space, similar to Figs.3.15, 3.16, 3.19 and 3.20. Fig.3.24 shows the horizontal structure of analysis RMSE of 500hPa height field, similar to Figs.3.15 and

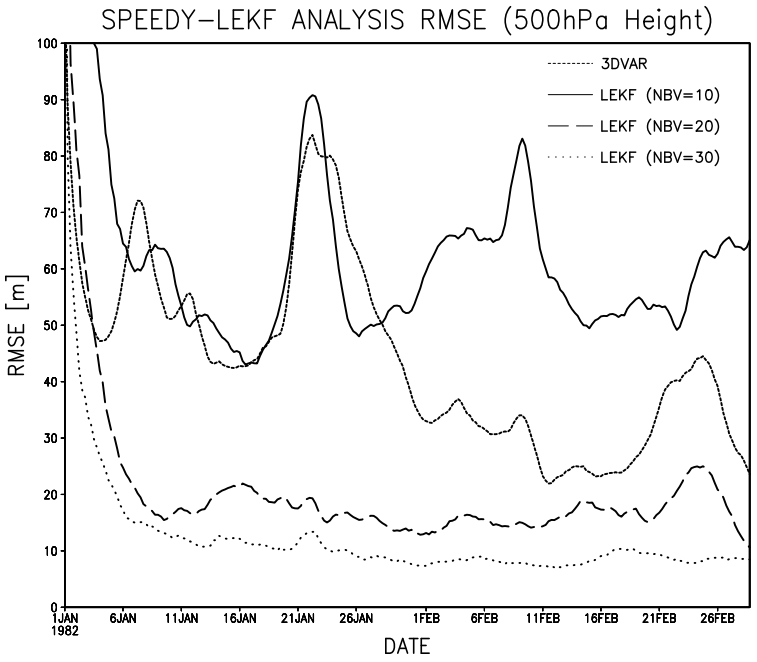

Figure 3.23: Analysis RMSE of 500hPa height field in the case of the sparse observational network using LEKF with 10 (solid line), 20 (dashed line), and 30 ensemble members (dotted line). The local patch parameter l is chosen as optimal. For comparison, the RMSE of 3DVAR is shown by the short dashed line.

3.19. Large errors exist in polar regions where no observations are available, but the errors are slightly larger than those in serial EnSRF. We can see a lattice-like pattern corresponding to the observational locations. The top panel of Fig.3.25 shows the analysis error bias. Basic characteristics remain the same as the case of 3DVAR: large errors with low wavenumbers exists in polar regions and we see a lattice-like pattern with the same wavelength as distance between observational points. The bottom panel of Fig.3.25 shows analysis error standard deviation, which is obtained in the same way as obtaining RMSE but subtracting error bias. The standard deviation field shows almost no lattice-like pattern, as in the case of serial EnSRF. Thus, we can attribute the lattice-like pattern to the error bias. In addition, the large errors in polar regions are reduced, thus, the error bias explains large amount of the polar errors.

As was shown in the case of serial EnSRF (Fig.3.21), the RMSEs of other variables at all levels in the case of LEKF are also shown here. LEKF with 10 ensemble members is worse than 3DVAR everywhere, which is omitted from the following figures. Thus, 3DVAR, LEKF with 20 and 30 ensemble members are shown by short dashed lines, solid lines, and dashed lines, respectively.

Fig.3.26 shows the analysis RMSE of four variables at the all seven pressure levels. Fig.3.26(a) shows the analysis RMSE of u-wind fields. Smaller RMSE is observed at the levels higher than 300hPa. This trend is clearer for 3DVAR. As expected, the larger ensemble size performs better everywhere. It is noted that only the case of 30 ensemble members gives smaller RMSE than observational error standard deviation (1.0m/s, shown by the thin solid line) at the lowest three levels. LEKF with 30 ensemble members clearly outperforms the others everywhere. Fig.3.26(b) shows the analysis RMSE of height fields, where we see that LEKF outperforms 3DVAR, especially at lower levels. At the top level,

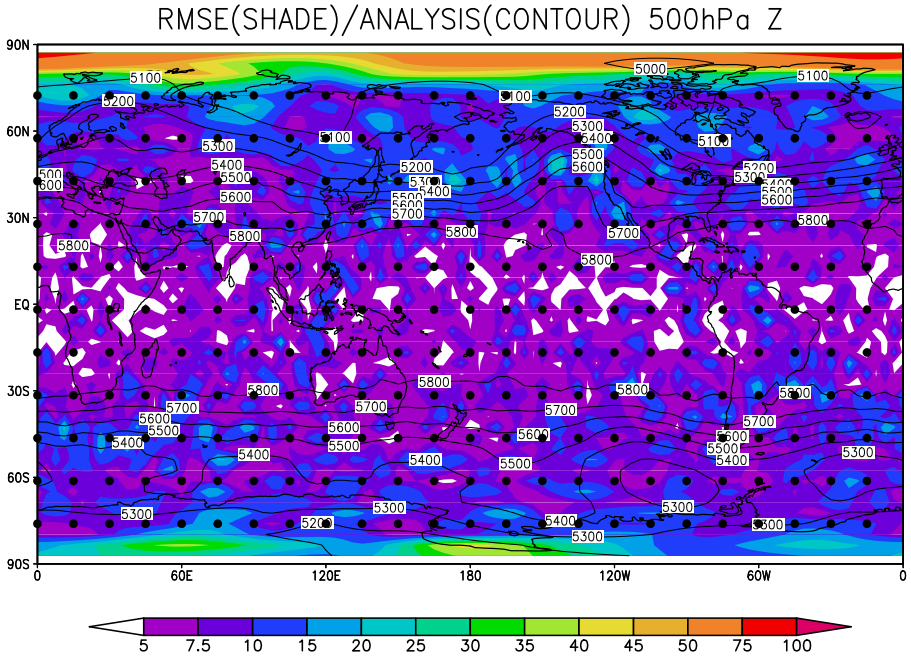

Figure 3.24: Horizontal structure of analysis RMSE of height field, similar to Figs.3.15 and 3.19, in the case of LEKF with 20 ensemble members and local patch parameter $l = 3$. Shades show analysis RMSE (in meters), contours show mean analysis fields (in meters). Dots indicate observational locations.

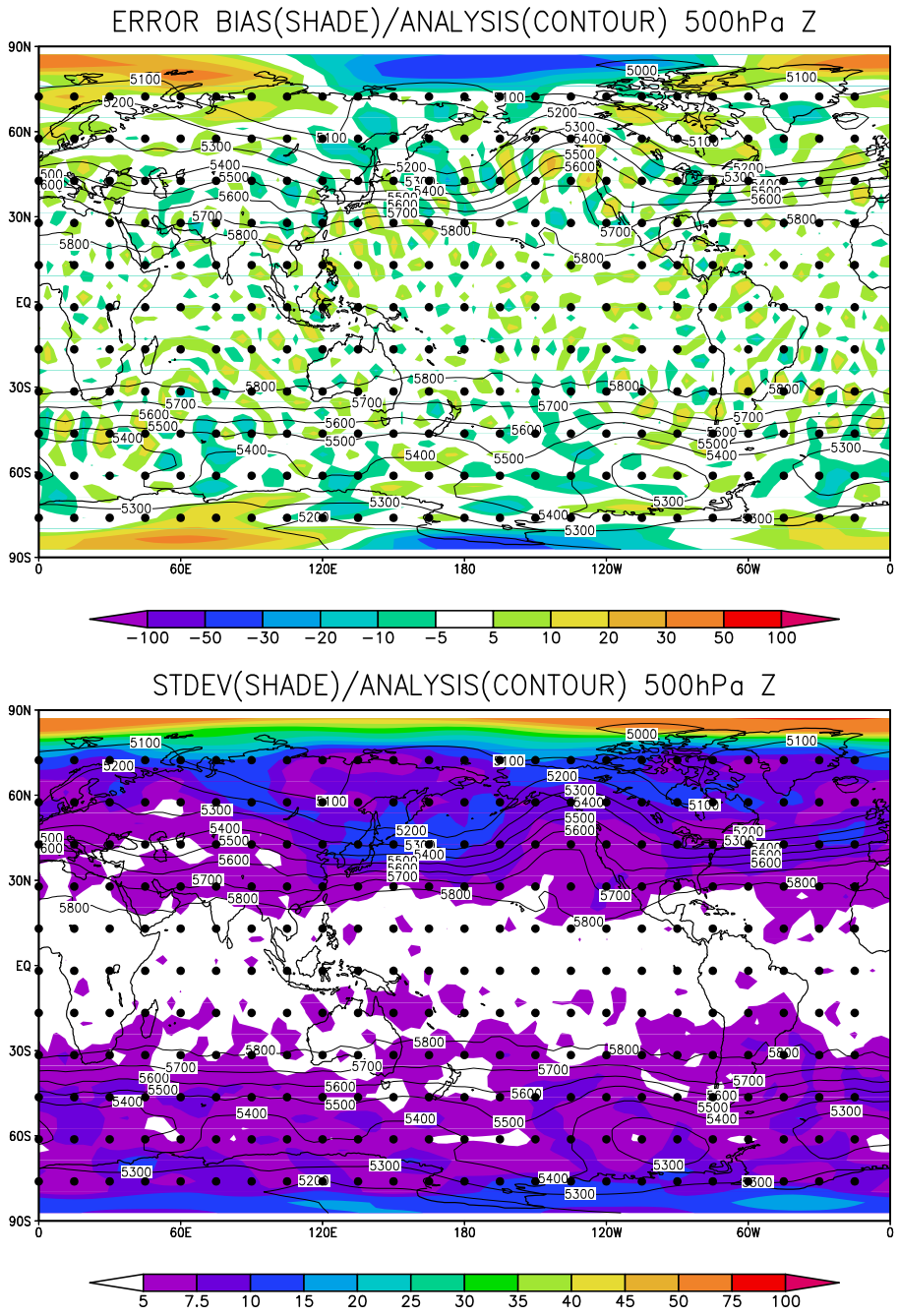

Figure 3.25: Horizontal structure of analysis error bias (top panel) and standard deviation (bottom panel) of height field (in meters), similar to Figs.3.16 and 3.20, in the case of LEKF with 20 ensemble members and local patch parameter $l = 3$. Shades show analysis error bias (top panel) and standard deviation (bottom panel), contours show mean analysis fields. Dots indicate observational locations.

there is almost no difference between 3DVAR and LEKF with 20 ensemble members. LEKF with 30 ensemble members clearly outperforms the others everywhere. Fig.3.26(c) shows the analysis RMSE of temperature fields. Generally, higher altitude gives smaller RMSE, that is clear in 3DVAR. Almost all cases of LEKF provide smaller RMSE than observational error standard deviation (1.0K, shown by the thin solid line). LEKF with 30 ensemble members clearly outperforms the others everywhere, and shows almost constant RMSE in the lower 6 levels. Fig.3.26(d) shows analysis RMSE of specific humidity fields, where we can see that higher altitude is drier, so the RMSE is smaller in higher altitudes. All show similar performances, and no clear difference can be seen, whereas there was clear difference in the case of serial EnSRF (cf. Fig.3.21(d)). Lower altitudes contains more humidity, and the RMSEs exceed the observational error standard deviation (10^{-4}kg/kg, shown in the thin solid line).

Fig.3.27 shows the zonal structures of analysis RMSE of the u-wind (top panel) and temperature (bottom panel), similar to Figs.3.17 and 3.22 of 3DVAR and serial EnSRF. Shades and contours show analysis RMSE and mean analysis fields, respectively. We can see the same basic features seen in 3DVAR and serial EnSRF: large errors in u-wind field exist in upper tropics, large errors in the temperature field exist in upper tropics and lower mid-latitudes. The amplitude of the errors is smaller than 3DVAR, but larger than serial EnSRF. The striped pattern seen in 3DVAR also still exists with much smaller amplitudes.

Table 3.6 shows the algorithm sensitivity to the parameter l_2. The table is created in the same way as Table 3.5, that is, the temporal mean of the analysis RMSE of 500hPa height field is computed using various parameter settings. Here, the local patch parameter is fixed to $l = 2$, and the ensemble size is fixed to 20 or 30. As in the Lorenz-96 model, we see no clear dependence on the parameter l_2, though the best performance in a given en-

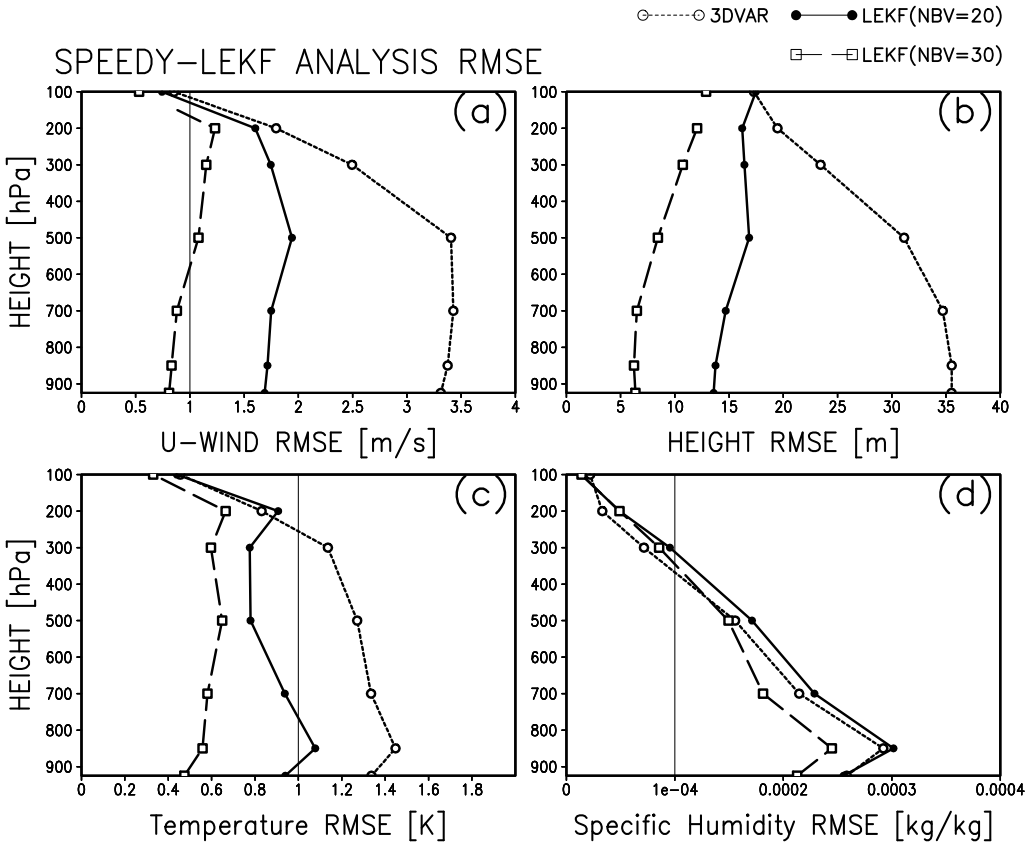

Figure 3.26: Analysis RMSE at all pressure levels temporally averaged for one month (112 samples) after the initial one-month spin-up period in the case of the sparse observational network using LEKF with 20 (solid line) and 30 ensemble members (dashed line). The four panels (a), (b), (c) and (d) correspond to u-wind, height, temperature (T) and specific humidity (q) fields, respectively. The observational error standard deviations are shown as thin solid lines wherever applicable. The localization length scale of the Schur product is chosen as optimal. For comparison, the RMSE of 3DVAR is shown as the short dashed line.

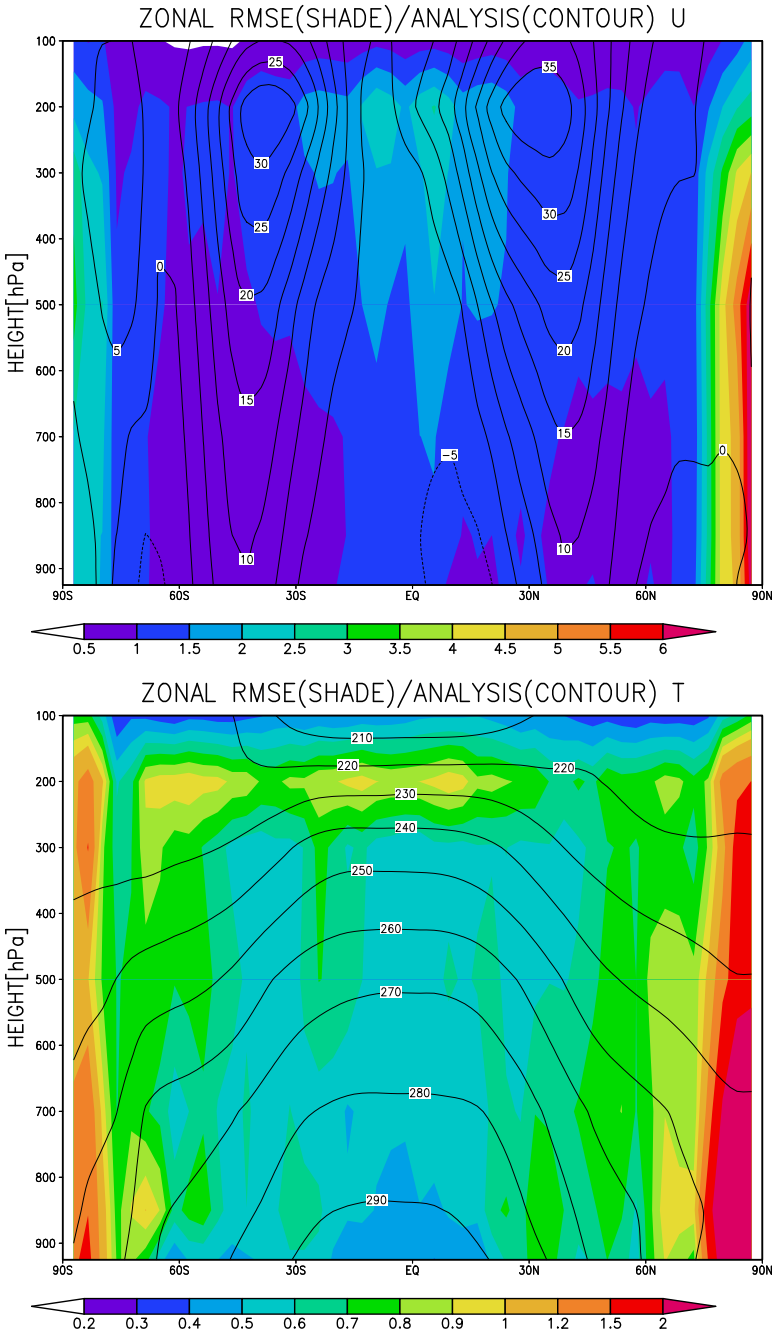

Figure 3.27: Zonal structure of analysis RMSE of u-wind field (top panel, in m/s) and temperature field (bottom panel, in K) in the case of LEKF with 20 ensemble members and local patch parameter $l = 3$. Shades and contours show analysis RMSE and mean analysis fields, respectively.

NBV / l_2	0	1	2
20	19.81	19.22	17.24
30	17.40	23.81	10.70

Table 3.6: Analysis RMSE of 500hPa height field in meters using LEKF on the SPEEDY model when the sparse observational network is applied and the parameter l_2 is changed. The ensemble size (NBV) is chosen to be 20 or 30, the local patch parameter is fixed to $l = 2$. The RMSE is temporally averaged for a month (112 samples) after the initial one-month spin-up period. For comparison, the analysis RMSE of 3DVAR in the same period is 31.14.

semble size is observed at the largest l_2 value (shown by rectangles). The large fluctuations may be caused by the limited run length, although overall results look reasonable.

The results using LEKF are summarized as follows:

- In the case of the sparse observational network, LEKF clearly outperforms 3DVAR and shows very good performance (8.44m RMSE of 500hPa height field with an optimal setting).
- The larger ensemble size, the better the performance.
- A smaller ensemble size requires smaller localization length scale, as in the Lorenz-96 model.
- 10 ensemble members are not enough for the LEKF to outperform 3DVAR. However, with the observational error covariance localization, LEKF with 10 ensemble members outperform 3DVAR.
- We observe no clear dependence on the parameter l_2.

- Horizontal and zonal distributions of analysis RMSE show lattice-like and striped patterns with smaller amplitudes than 3DVAR but with larger amplitudes than serial EnSRF, we can attribute the lattice-like pattern to the analysis error bias.

- Not only 500hPa height fields, but also most other fields at other levels show better performance than 3DVAR in the cases of 20 and 30 ensemble members.

3.5.4 Timing results

In 3DVAR, the computational time depends on the number of iterations of the variational process. Each iteration requires less than a half second on a Linux PC with a 2.7GHz Intel Celeron (Northwood) processor. A typical number of iterations until convergence is about 35, which requires about 5 seconds of computational time.

Table 3.7 shows timing of serial EnSRF for each analysis. Here, the sparse observational network is used. The analysis is performed on two Linux PCs: one with a 2.7GHz Intel Celeron (Northwood) and the other with a 2.53GHz Intel Celeron D (Prescott). The latter is a later model with a twice larger L2-cache memory. Although the clock speed of the former processor (2.7GHz) is larger, the latter one (2.53GHz) performs almost twice as fast.

Computational time of serial EnSRF strongly depends on the number of observations and the ensemble size. However, it does not depend on the horizontal localization scale. The system can be further accelerated by optimizing the code by considering the localization scale.

The computational time of LEKF strongly depends on the ensemble size and the local patch parameters(both l and l_2). However, it does not strongly depend on the number of observations. The system has been accelerated by optimizing the code as much

NBV	10	20	30
Celeron 2.7GHz (Northwood)	2:20	4:30	
Celeron D 2.53GHz (Prescott)	1:15	2:20	3:20

Table 3.7: Timing (min:sec) of serial EnSRF with 10, 20, and 30 ensemble members on the SPEEDY model when the sparse observational network is applied.

NBV(l)	10(1)	20(3)	30(4)
Celeron 2.7GHz (Northwood)	2:50	8:30	
Celeron D 2.53GHz (Prescott)	1:35	6:00	16:30

Table 3.8: Timing (min:sec) of LEKF with 10, 20, and 30 ensemble members on the SPEEDY model when the sparse observational network is applied. The local patch parameter l is chosen as an optimal setting shown in the parenthesis, and l_2 is fixed to 0.

as possible. However, we do not apply the new technique to avoid eigenvalue decomposition (local ensemble transform Kalman filter, LETKF) proposed by Hunt (2005). Note that LEKF can be parallelized very efficiently, thus, LEKF can be accelerated almost perfectly on parallel computers.

Table 3.8 shows timing of LEKF for each analysis. Here, the sparse observational network is used, l_2 is fixed to 0. The local patch parameter l is chosen as an optimal setting that gives the best performance in a given ensemble size. The analysis is performed on two Linux PCs: one with a 2.7GHz Celeron (Northwood) and the other with a 2.53GHz Celeron D (Prescott), as in the case of serial EnSRF (Table 3.7). Similar to the case of serial EnSRF, the latter processor performs faster.

Ensemble size (NBV)	10	20	30
Optimal localization scale σ of serial EnSRF	1	2	3
Optimal local patch parameter l of LEKF	1	3	4

Table 3.9: Optimal localization parameters (length scale σ of serial EnSRF and patch size l of LEKF) with given ensemble sizes. Here, the covariance inflation parameter is fixed to 4%, the parameter l_2 in LEKF is fixed to 0.

3.5.5 Comparison

For comparison of each method, the results of each method are combined and shown in Table 3.10 and Fig.3.28. Each result has already been shown separately in previous sections, the optimal parameters are chosen as shown in Table 3.9. Serial EnSRF is the best among the three methods. This result is the same as that in the case of the Lorenz-96 model when the same multiplicative covariance inflation is applied in both serial EnSRF and LEKF. Note that in the Lorenz-96 model, enhanced variance inflation significantly improved LEKF's performance, LEKF is comparable to serial EnSRF. The RMSE values contain errors due to the limited run length, but we obtained a similar value with doubled run length in the case of serial EnSRF with 10 ensemble members. Fig.3.39 shows that the rms errors of LEKF and EnSRF with 30 ensemble members are essentially identical except poleward of 80 degrees, where the LEKF has larger errors which could be improved with a different formulation for the polar patches in the LEKF.

As for computational time, Table 3.11 shows timing on a Linux PC with a 2.53GHz Intel Celeron D (Prescott) processor. The timing was adopted from Tables 3.7 and 3.8. On a single computer with serial treatment, it is clear that serial EnSRF is faster, especially in the larger ensemble size.

NBV	Serial EnSRF	LEKF	3DVAR	NO ASSIM
10	11.74	60.38	31.14	106.37
20	10.69	16.85	31.14	106.37
30	4.99	8.44	31.14	106.37

Table 3.10: Comparison of analysis RMSE of 500hPa height field in meters on the SPEEDY model when the sparse observational network is applied. For each ensemble size (NBV), best results are chosen for serial EnSRF and LEKF from Tables 3.4 and 3.5. For 3DVAR, the best performance 31.14m is chosen. "NO ASSIM" denotes the case without data assimilation.

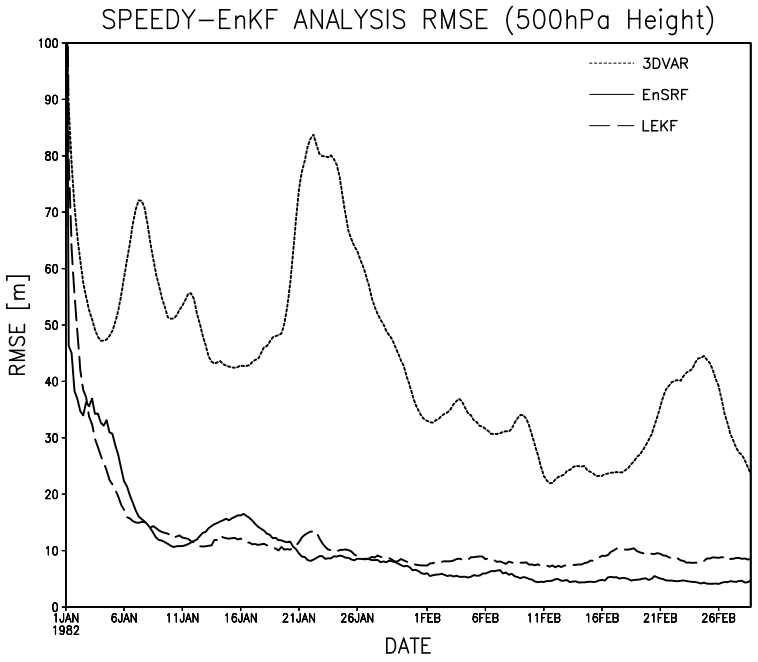

Figure 3.28: Analysis RMSE of 500hPa height field in the case of the sparse observational network using 3DVAR (short dashed line), serial EnSRF (solid line) and LEKF (dashed line). The ensemble size for serial EnSRF and LEKF is 30. Parameters are optimal, that is, best results are shown for each method.

NBV	10	20	30
Serial EnSRF	1:15	2:20	3:20
LEKF	1:35	6:00	16:30

Table 3.11: Timing (min:sec) of serial EnSRF and LEKF with 10, 20, and 30 ensemble members on the SPEEDY model when the sparse observational network is applied.

However, LEKF can be parallelized very efficiently, thus, in parallel computers with tens of computational nodes, LEKF could be much faster than serial EnSRF. Furthermore, timing of serial EnSRF strongly depends on the number of observations unlike LEKF. More satellite data will be available in the near future. With larger number of observations, LEKF has an important advantage in terms of computational efficiency.

Simulating the realistic situation, we estimate timing when 100 times more observations (26400 observations) are assimilated using a parallel computer with 100 computational nodes. We assume each node has the same performance as a 2.53GHz Intel Celeron D (Prescott) processor. Let the number of computational nodes and parallelization efficiency denote N and α, respectively. The computation is accelerated as

$$T' = T \times \frac{1}{N\alpha} \tag{3.12}$$

where T and T' denote timings using 1 node and N nodes, respectively. LEKF is efficiently parallelized, we assume 80% efficiency. Serial EnSRF could be parallelized by treating distant observations separately by virtue of Schur product localization, we assume 20% efficiency. LEKF can treat large number of observations more efficiently, we assume 50 times more computation caused by 100 times more observations. By contrast, Serial EnSRF simply increases the number of loops, we assume 100 times more computation caused by 100 times more observations. In summary, we multiply the timing results of

NBV	10	20	30
Serial EnSRF	6:15	11:40	16:40
LEKF	1:00	3:45	10:19

Table 3.12: Estimated timing (min:sec) of serial EnSRF and LEKF with 10, 20, and 30 ensemble members when 100 times more observations are assimilated using a parallel computer with 100 computational nodes. Parallelization efficiency is assumed as 80% and 20% for LEKF and serial EnSRF, respectively. We assume the increased observations require 50 and 100 times more computation for LEKF and serial EnSRF, respectively.

Table 3.11 as follows:

$$T'_{\mathrm{LEKF}} = T_{\mathrm{LEKF}} \times \frac{5}{8} \tag{3.13}$$

$$T'_{\mathrm{EnSRF}} = T_{\mathrm{EnSRF}} \times 5 \tag{3.14}$$

Table 3.12 shows the estimated timimg. By the efficient parallelization and large number of observations, LEKF is expected to be faster everywhere. If we apply the new LETKF, several times faster than LEKF, we can further accelerate the process.

3.6 Sensitivity experiments

In this section, we investigate the robustness of our results to the experimental settings. We apply a different observational network and investigate the impact of moisture observations. Then, we investigate the impact of vertical error correlations and a different type of error covariance localization. Finally, we apply random perturbation addition to ensemble members. The sensitivity results also suggest possible ways to improve the filter performance.

3.6.1 Response with a different observational network

So far, we applied mostly the sparse observational network, and we saw lattice-like patterns in the analysis error fields. Since the observations were available uniformly in space, the error fields looked uniform. However, in reality we have more rawinsonde observations over land and less over the ocean. Thus, we expect non-uniform distribution of error fields. In addition, it is not clear how sensitive each data assimilation method is to the choice of observational network. We expect EnKF to be more robust with respect to the choice of the observational network than 3DVAR. To see the response with a different observational network, we apply a realistic observational network (Fig.3.11) to 3DVAR and serial EnSRF as a representative of EnKF methods.

Fig.3.29 shows analysis RMSE of 500hPa height field in the cases of 3DVAR (short dashed line) and serial EnSRF (solid line) with 30 ensemble members and localization scale of 3.0. Clearly serial EnSRF outperforms 3DVAR. The temporal averages are 45.6 and 11.6 meters for 3DVAR and serial EnSRF, respectively.

Fig.3.30 shows the spatial distribution of analysis RMSE of 500hPa height fields of 3DVAR (top panel) and serial EnSRF (bottom panel). Both figures have a similar structure with different amplitudes. Larger errors appear in data-poor regions. The southern hemisphere shows larger errors than the northern hemisphere. The largest error, except the polar area, is observed over south of the Pacific ocean. The Northern Atlantic ocean contains the largest error part in the northern hemisphere, except the polar area. Clearly 3DVAR shows larger errors everywhere, the effect of the irregular observational network is larger in 3DVAR. As for the lattice-like pattern, we can see some high frequency components especially in the tropics in 3DVAR, but we cannot see clear correspondence to the observational locations. Thus, the error fields look more realistic; data-rich regions

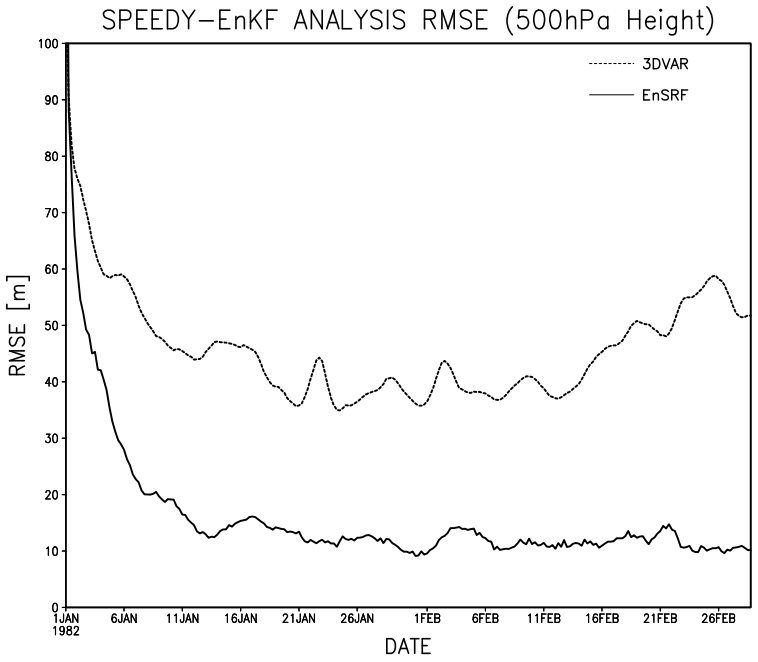

Figure 3.29: Analysis RMSE of 500hPa height field in the case of the realistic observational network using 3DVAR (short dashed line) and serial EnSRF (solid line). The ensemble size for serial EnSRF is 30, the localization scale is 3.0.

show smaller errors, data-poor regions show larger errors.

Figs.3.31 and 3.32 show zonal structure of analysis RMSE of u-wind and temperature fields in the cases of 3DVAR and serial EnSRF, respectively. In all figures, it is clear that the southern hemisphere has much larger errors than the northern hemisphere. By comparing 3DVAR and serial EnSRF, both figures show very similar structure with much larger amplitudes in 3DVAR. However, there is a structural difference between 30 and 60 degrees north in the lower to middle atmosphere. Here 3DVAR shows larger errors, serial EnSRF does not, even if we consider the difference in the error amplitudes. This is clearer in the temperature field (bottom panel). Although many observations are available in the northern hemisphere, there are data-poor regions over the Atlantic and Pacific ocean. Since 3DVAR is more sensitive to the lack of data, 3DVAR tends to be affected more over the data-poor regions. This could explain why 3DVAR shows larger errors in the northern hemisphere. RMSE of temperature fields shows large errors over the Eastern Atlantic especially at lower levels, no other signal such as orographic shape is observed.

3.6.2 Usefulness of moisture observations

Moisture fields are strongly related to highly nonlinear and complicated physical processes, moisture fields tend to show a noisy structure compared to other variables. Since the moisture fields represent a complicated and noisy physical processes, it is not clear if the moisture observations actually provide useful information in data assimilation, especially in the case of sparse observational networks. In fact, Szunyogh et al. (2005) assumed moisture was not observed in their LEKF experiments using the NCEP global model. In spite of these facts, we included moisture observations in the present research, considering the SPEEDY model has much simpler physical processes than state-of-the-art models. It

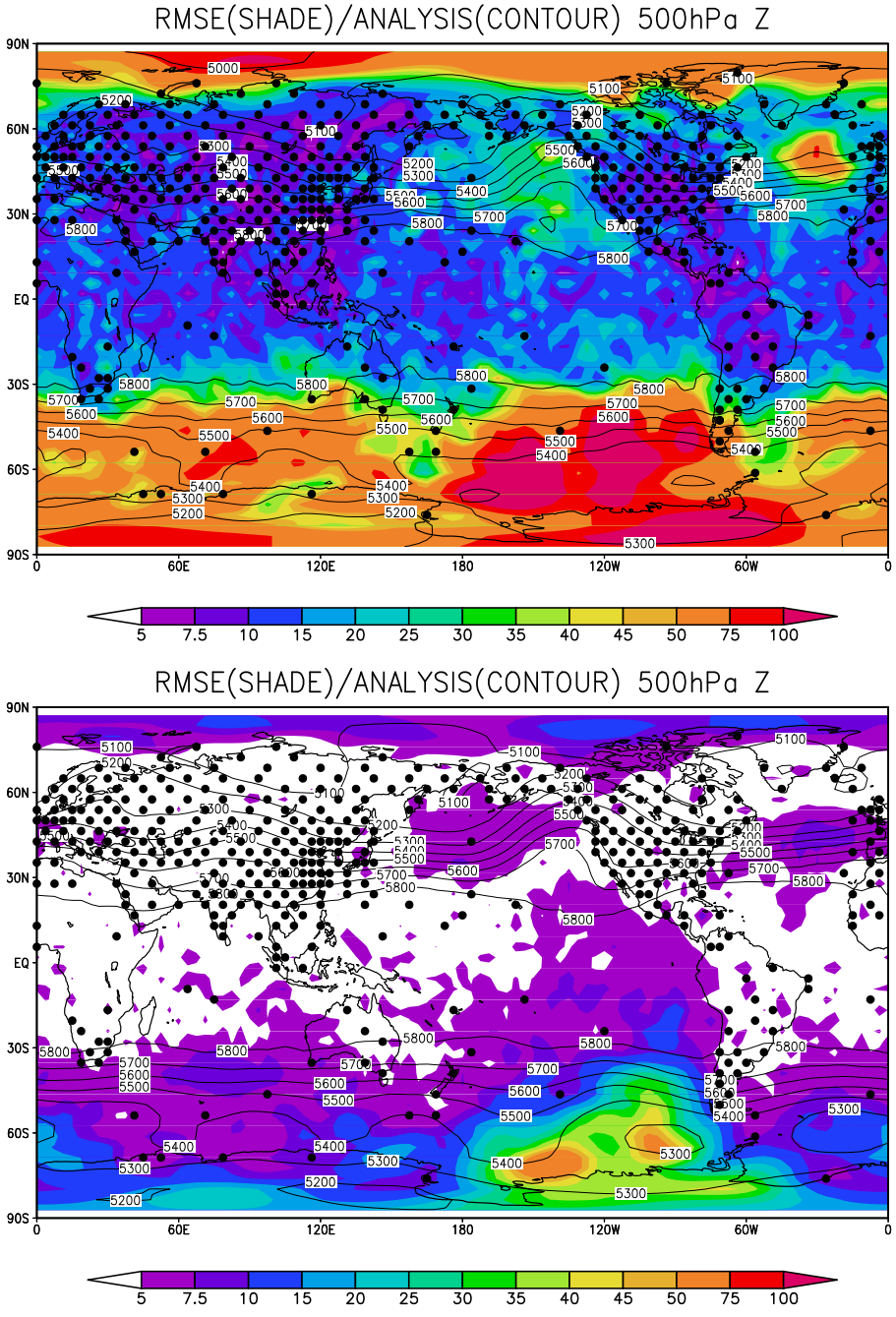

Figure 3.30: Spatial distribution of analysis RMSE of 500hPa height fields (in meters) in the case of the realistic observational network using 3DVAR (top panel) and serial EnSRF (bottom panel). The ensemble size for serial EnSRF is 30, and the localization scale is 3.0. Dots indicate the observational locations.

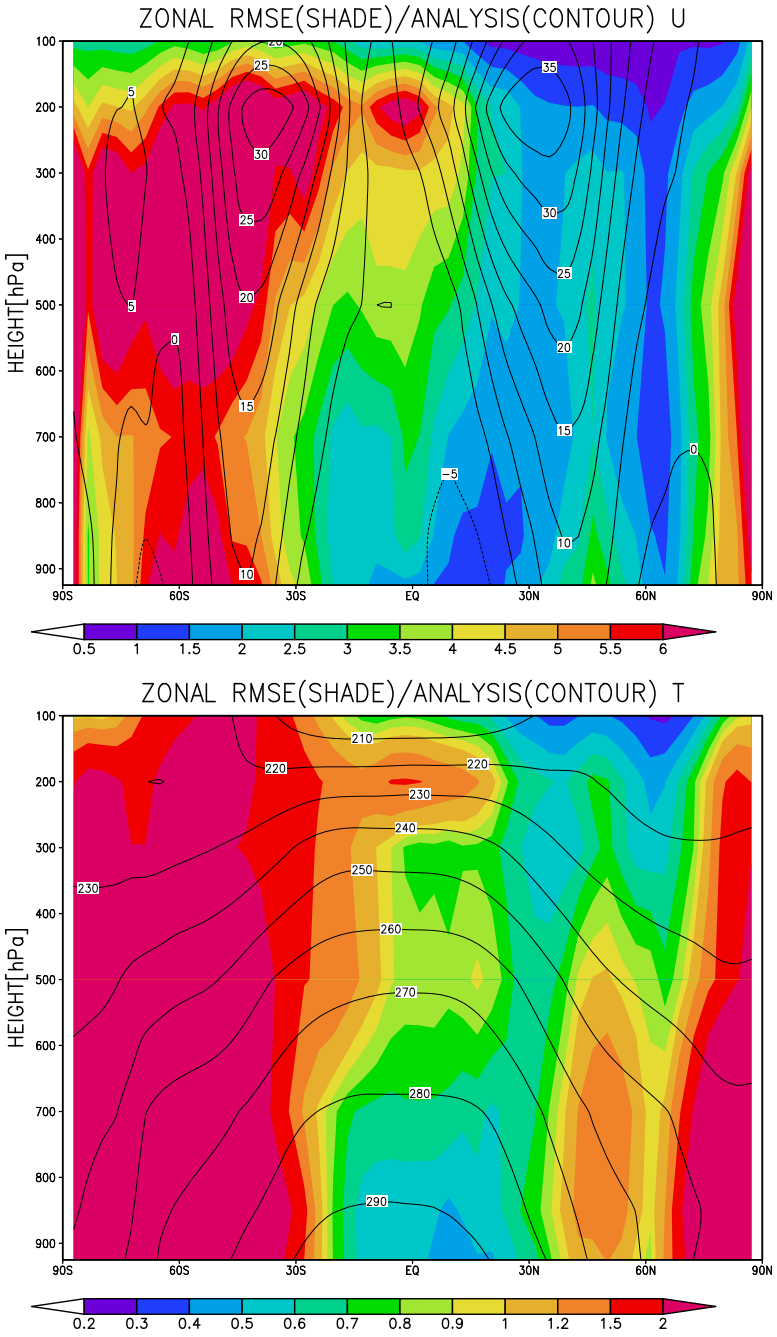

Figure 3.31: Zonal structure of analysis RMSE of u-wind field (top panel) and temperature field (bottom panel) in the case of 3DVAR with the realistic observational network. Shades and contours show analysis RMSE and mean analysis fields, respectively.

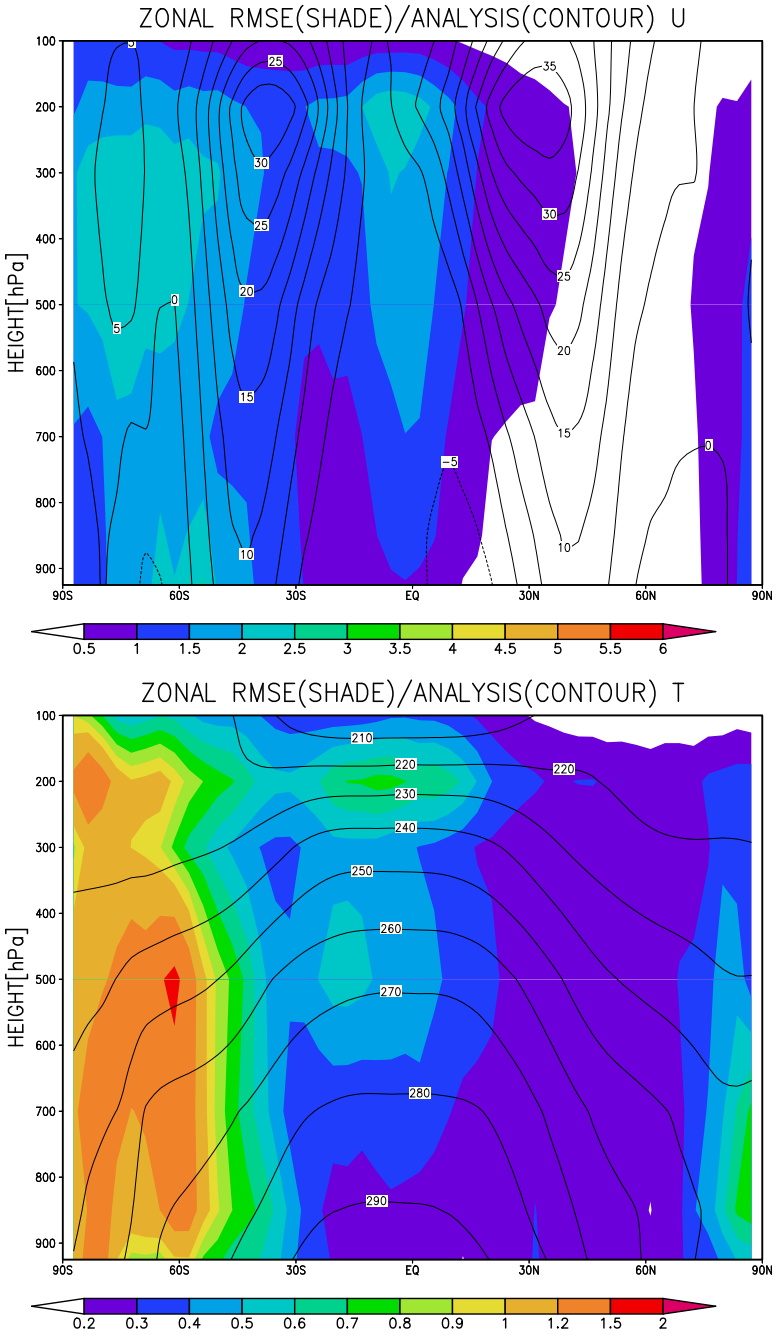

Figure 3.32: The same figure as Fig.3.31 but in the case of serial EnSRF with 30 ensemble members and localization scale of 3.0.

is important to investigate the impacts of moisture observations, so that we confirm our results are not strongly affected by the noisy moisture observations.

In order to investigate the impact of moisture observations, we perform data assimilation cycle experiments excluding moisture observations and compare with the previous results including moisture observations. Fig.3.33 shows analysis RMSE using serial EnSRF with 10 ensemble members with and without moisture observations. For the moisture variable (specific humidity, q), we obtain large improvement by moisture observations as expected. Importantly, we see positive impact by the moisture observations for other variables. We also see improvements by moisture observations using LEKF, supporting moisture observations have positive impact. Thus, it is reasonable to include moisture observations in our experiments, suggesting our results are not adversely affected by the moisture observations.

Note that the SPEEDY model has only very simple physics, whereas the physical processes in more sophisticated models are much more complicated. As a result, it is not straightforward to determine in the operational context if moisture observations cause such big improvement.

3.6.3 Vertical error correlations

So far, vertical error correlations were not considered in any data assimilation method (3DVAR, serial EnSRF, and LEKF). However, it is known that atmosphere has an important vertical structure, especially with respect to synoptic instability such as baroclinic instability. Thus, we expect that the inclusion of the vertical error correlation might contribute a certain amount in data assimilation.

To see the impact of vertical correlation, vertical correlation is introduced in LEKF

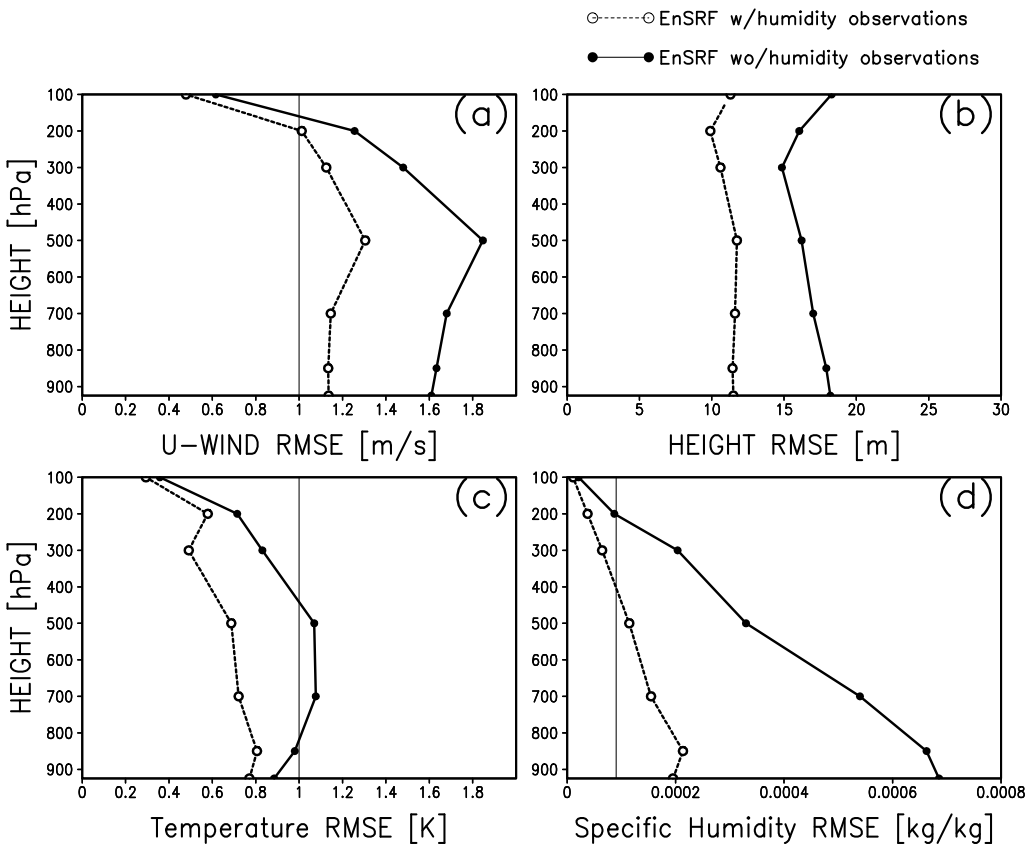

Figure 3.33: Analysis RMSE of the u-wind field (a), hight field (b), temperature field (c) and specific humidity field (d) in the cases of serial EnSRF with 10 ensemble members with and without humidity observations shown by broken and solid lines, respectively. The thin solid line shows observational error standard deviation wherever applicable.

by taking local patches combined with their closest levels. For example, the bottom level is combined with the second level, thus, the local patch becomes twice larger at the bottom level, which is the same at the top level. The second level is combined with the bottom and the third levels, thus, the local patch becomes three times larger at the second level.

Fig.3.34 shows the impact of the vertical correlation in the cases of 10 and 20 ensemble members. The local patch parameters are fixed to $l = 1$ and $l_2 = 0$, respectively. With 10 ensemble members, the vertical correlation has a clear positive impact: 46.82m RMSE of 500hPa height in average with the vertical correlation (black solid line), while 60.38m without the vertical correlation (black short-dashed line). Thus, the vertical correlation indeed shows a positive impact. With 20 ensemble members, however, it has slight negative impact in average: 38.35m RMSE of 500hPa height with the vertical correlation (blue dotted line), while 35.45m without the vertical correlation (blue long-dashed line). However, the difference is not statistically significant.

At the other extreme, we can construct the LEKF with full vertical correlations, i.e. without localization in the vertical. In this case, we combine all the vertical levels and take a vertical column as analyzed points in a local patch. So far, we needed a three-dimensional loop for the local patch center to cover the entire three-dimensional grids. However, now with the full vertical correlation, we need only a two-dimensional horizontal loop.

Fig.3.35 shows the impact of the full vertical correlation in the cases of 10 and 20 ensemble members. Here, the local patch parameters are chosen to be $l = 1$ and $l_2 = 0$, and the line legends are the same as the previous Fig.3.34. With 10 members, the full vertical correlation shows a clear negative impact. A possible reason is as follows: The more combined vertical levels, the larger the dimension of a local patch becomes, which is

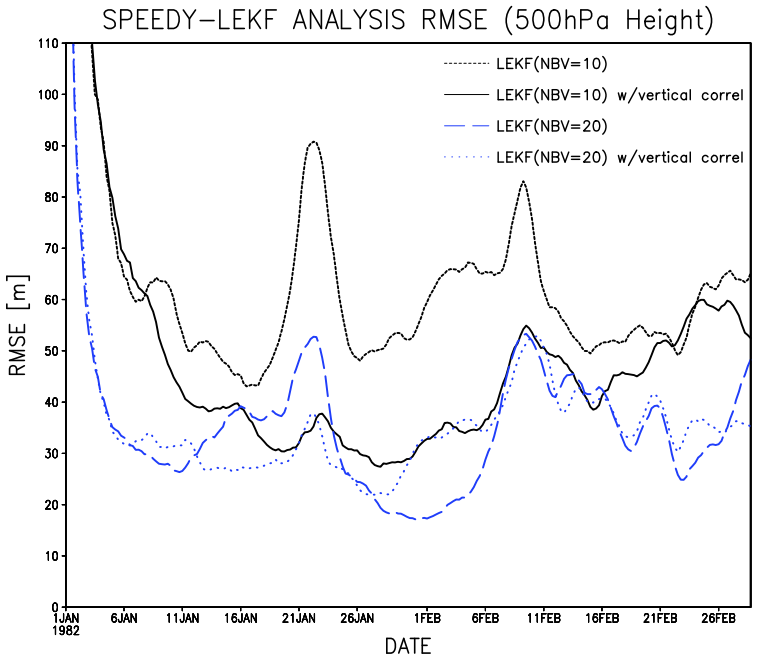

Figure 3.34: Impact of the vertical correlation in the analysis RMSE of 500hPa height field using LEKF with 10 (black lines) and 20 (blue lines) ensemble members. The black solid line and black short-dashed line show the cases with and without the vertical correlation, respectively. The blue dotted line and blue long-dashed line show the cases with and without the vertical correlation, respectively.

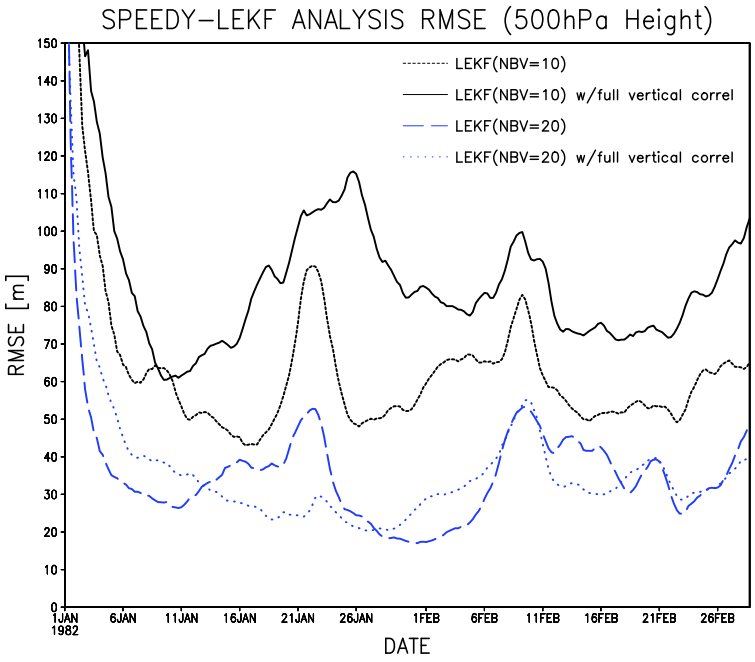

Figure 3.35: The same figure as Fig.3.34 but for the full vertical correlation.

why sampling errors in far away levels due to the limited ensemble size get larger. With 20 members, the impact of the full vertical correlation is not clear. It may be possible that the vertical correlation provides useful information, but sampling errors are also introduced, even with 20 ensemble members. Thus, the positive and negative effects are cancelled out and the net effect is almost zero.

3.6.4 Error covariance localization

Since the serial EnSRF with 10 ensemble members has outperformed 3DVAR, the ensemble size of 10 is large enough for EnKF to work in the given configuration. However, LEKF with 10 ensemble members has performed worse than 3DVAR. The difference between EnSRF and LEKF is in the way of localization. Serial EnSRF localizes error covariance around observations with Gaussian-like weighting, LEKF localizes around an analyzed point with Heaviside-like step weighting. Thus, we applied observational error covariance

Localization scale (σ)	∞	1.0	$1/\sqrt{2}$	0.5	0.25
500Z RMSE [m] ($l = 1$)	60.4	48.2	34.5	25.9	42.0
500Z RMSE [m] ($l = 2$)	75.9	29.5		21.6	

Table 3.13: Analysis RMSE of 500hPa height field using LEKF with 10 ensemble members with various localization scale of observational error covariance localization. The local patch parameter is chosen to be $l = 1, 2$.

localization in LEKF.

Table 3.13 shows analysis RMSE of 500hPa height field using LEKF with 10 ensemble members and local patch parameter $l = 1$ with various localization scales of observational error covariance localization. We can see a large improvement of LEKF with the observational error covariance localization in the case of 10 ensemble members. If we choose localization scale of 0.5, the analysis RMSE (25.9) is smaller than that of 3DVAR (31.1). When we increase local patch parameter l to 2, we observe better performance than in the case of local patch parameter $l = 1$. Thus, by increasing the local patch size, we include distant observations that still provide useful information.

With 30 ensemble members, LEKF showed 8.44m RMSE of 500hPa height when local patch parameter $l = 4$ and 8.67m when $l = 3$. Both of them are much smaller than 3DVAR. The localization is applied for 30 ensemble members to see the impact of the localization when the filter works very well. We chose $l = 3$ and $\sigma = 1.5$, then the RMSE became 7.94m, slightly better than the case without localization.

An important characteristic is observed in the temporal changes of analysis RMSE. Fig.3.36 shows the temporal sequences of analysis RMSE of 500hPa height fields with different localization settings. After the initial one-month transient period, the three lines

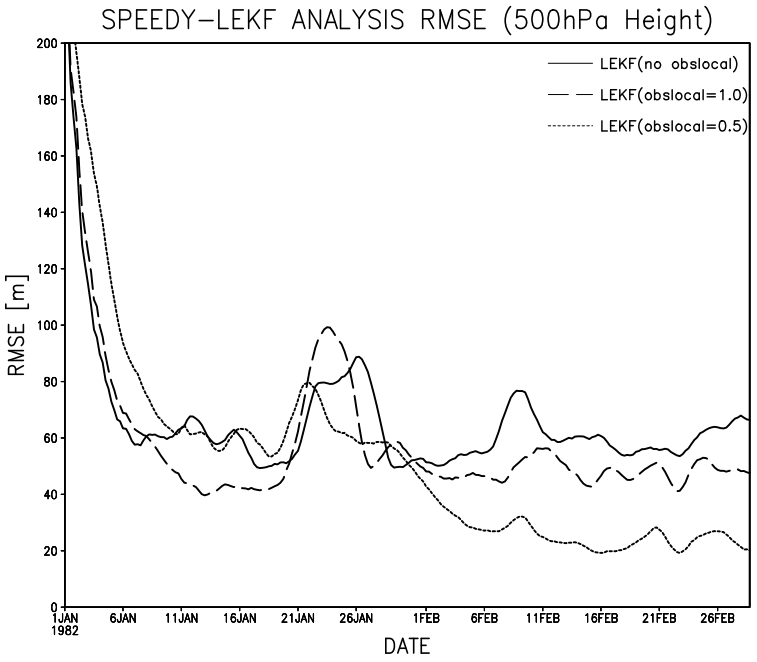

Figure 3.36: Analysis RMSE of 500hPa height field using LEKF with 10 ensemble members. Here, observational error covariance localization is applied. The solid, long-dashed and short-dashed lines correspond to no localization, with localization scales $\sigma = 1.0$, and $\sigma = 0.5$ respectively.

clearly separate. The order of the performance is already described in Table 3.13. In the very initial transient period (first 10 days or so), the three lines also show clear separation, but with the opposite order. With large initial errors, distant observations provide useful information, but the localization reduces their effect. Thus, the more localized, the smaller the correction of the initial large error becomes. After the transient period, the sampling errors in the distant points dominate, therefore, the more localized, the smaller the analysis errors.

So far, we modified the localization in LEKF. We can modify the localization of serial EnSRF using Heaviside-like step weighting function. Houtekamer and Mitchell (1998) applied this kind of localization around the observations by taking "radius of influence".

However, LEKF takes local patch with square shape. Thus, we take a square around an observation and force zero analysis increment outside of the square so that we simulate localization weighting function similar to LEKF. The impact of the different localization weighting function turned out to be very small. When we took 1×1 squares around observations, analysis RMSE of 500hPa height was 14.0m, which is just slightly worse than 11.7m that was obtained using the more natural Schur product with the length scale of 1. Here, we used 10 ensemble members.

In EnKF formulations, we applied globally the same localization length scale in grid spacing. However, as we could see in the background error statistics using the NMC method, the correlation length scale is not constant (Fig.3.3). We consider the latitudinal dependence and apply the length scale given by the NMC method in serial EnSRF. Then, analysis RMSE of 500hPa height was 22.1m, almost twice as large as 11.7m obtained by the constant localization scale of 1. The large error comes mainly from high latitudes, suggesting large length scales introduce sampling errors in the estimation of the error covariance with 10 ensemble members.

3.6.5 Random perturbation addition as "stochastic seeding"

Corazza et al. (2002) kept the bred vectors (BVs) "young" by applying stochastic perturbations to the BVs used in a 3DVAR data assimilation where the background error covariance is augmented with BVs in order to account for "errors of the day". They found that the augmentation of the background error covariance with BVs reduced analysis errors by 20%, including stochastic seeding on the BVs at the beginning of the breeding integration increased the improvement to about 40%. Miyoshi and Kalnay (2005) investigated the effect of the "stochastic seeding" in a breeding cycle on the Lorenz-96 model,

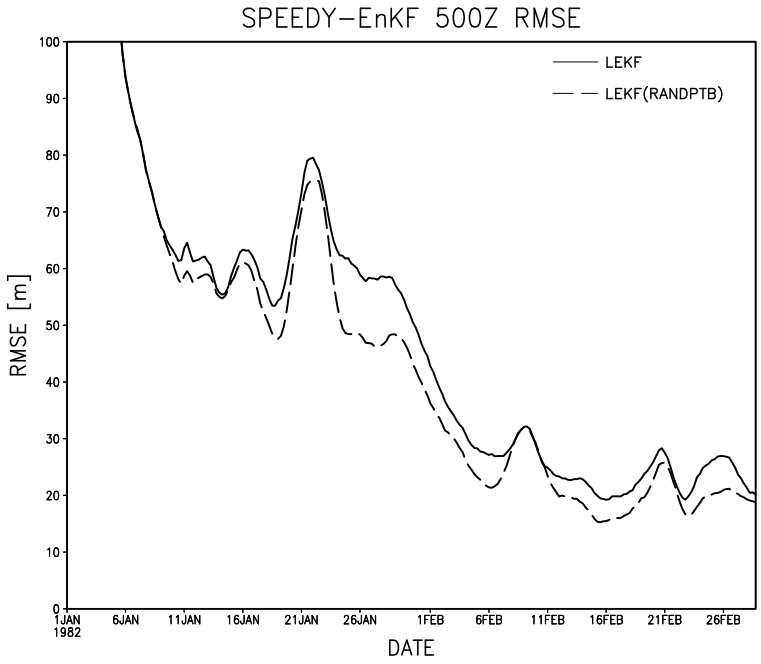

Figure 3.37: Analysis RMSE of 500hPa height field using LEKF with 10 ensemble members. Here, observational error covariance localization is applied. Dashed and solid lines show the cases with and without random perturbation addition, respectively.

and they found stochastic seeding lets BVs grow secondary instabilities that were ignored in leading BV but captured in higher order BVs with orthogonalization processes. In EnKF, the ensemble members do not converge to a single direction, but still, it may be possible that ensemble members lose signals of instabilities.

Expecting that the random perturbation addition or "stochastic seeding" has similar effects to let EnKF capture possible instabilities, we apply it on LEKF with 10 ensemble members. Random perturbation consists of random numbers with Gaussian distribution and 1% of the observational error standard deviation. Fig.3.37 shows analysis RMSE of 500hPa height field using LEKF with and without random perturbation addition. With random perturbation addition, the RMSE is smaller almost everywhere or at least is not worse than the case without random perturbation addition.

3.6.6 Summary

In this section, we confirmed that our results in the previous section were reasonably robust with the experimental settings. Using an irregularly distributed realistic observational network, both 3DVAR and serial EnSRF show larger errors than the regularly distributed sparse network. 3DVAR is more strongly affected by the irregular distribution of observations, the advantage of EnKF is larger. Moisture observations provide useful information for data assimilation on the SPEEDY model, supporting that we included moisture observations in our experiments. In LEKF with 10 ensemble members, accounting for local vertical correlation improve the results, but full vertical correlation worsens the filter. With 20 ensemble members, the effect of considering vertical correlations is not significant. These results suggest that the vertical error independence is a reasonable choice in the SPEEDY model. LEKF shows significant improvements by the observational error covariance localization, especially with fewer ensemble members. However, serial EnSRF with 10 ensemble members is not sensitive to the choice of localization weighting function, suggesting that the shape of localization weighting function does not have strong influences. When adding random noise as "stochastic seeding", LEKF showed better results, suggesting possible further improvements in EnKF.

3.7 Characteristics of the analysis and forecast errors

3.7.1 Structures of the analysis increment

Carrying out perfect model experiments, where we know the "truth", allows us to analyze the structure of analysis errors (which are not known in real data assimilation) and to compare them with forecast errors. It is also important to study the relationship be-

tween EnKF estimated errors and observed errors as previously done by Corazza with a quasi-geostrophic model (Kalnay et al. 2005). We compare forecast errors and analysis increments for 3DVAR and EnKF.

Fig.3.38 shows the background error fields (forecast minus truth, shaded) and the analysis increment fields (forecast minus analysis, contour) at an arbitrary time in the case of 3DVAR (top panel) and serial EnSRF (as representative of EnKF) with 30 ensemble members (bottom panel). Here, the localization scale of serial EnSRF is chosen as 3.0, which is an optimal. The results we obtain with a primitive-equation model are very similar to those obtained with the quasi-geostrophic model. The analysis increment field of EnKF captures flow-dependent structures of the background error field, e.g. positive pressure increment at the lower left corner of the bottom panel. In contrast, 3DVAR cannot capture the flow-dependent structures at all. 3DVAR analysis increments look like circles around observations, and only in some wide error regions with large amplitudes, increments by neighbor observations look connected, e.g. large positive and negative increments in higher latitudes. The observations are so sparse that basically 3DVAR cannot capture the structure between observations.

Importantly, we see a striped pattern in the background error field of 3DVAR, whereas there is no such pattern in the case of EnKF. This is also seen in the zonal RMSE structure shown in Fig.3.39, where 3DVAR shows lower RMSE every 4 point corresponding to observational location. Although correlations between the analysis increment and background error are small (around 0.3) both in 3DVAR and EnKF, EnKF shows a clear advantage because of better background fields.

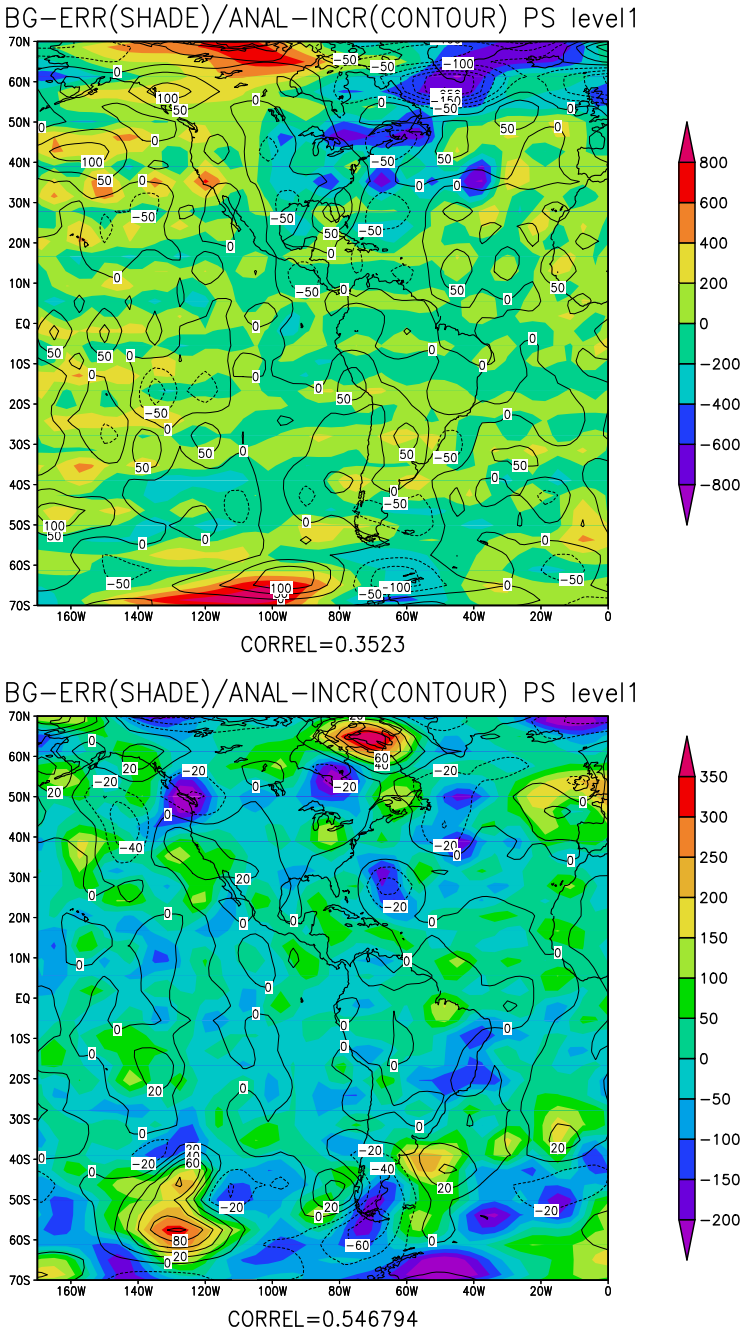

Figure 3.38: Background error field (shaded) and analysis increment field (contour) of surface pressure (p_s) at an arbitrary time in the case of 3DVAR (upper panel) and serial EnSRF with 30 ensemble members and the localization scale of 3.0 (bottom panel).

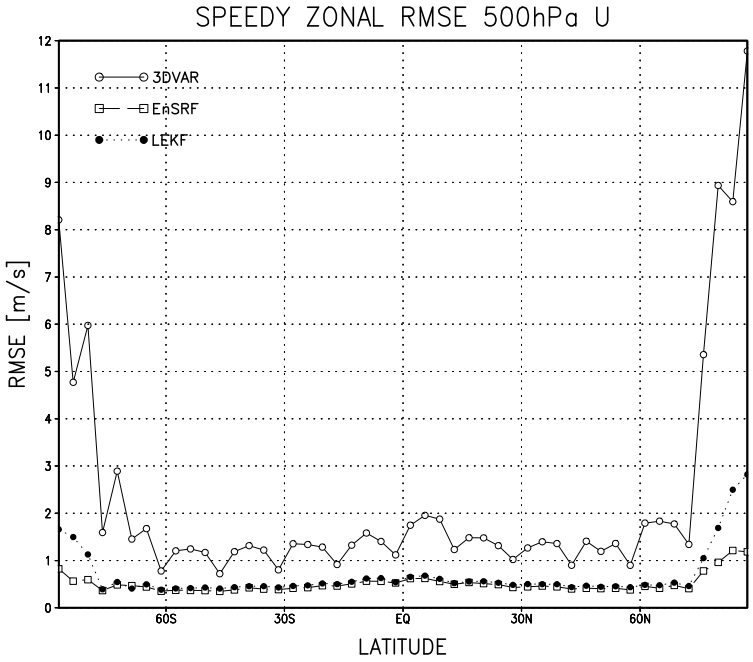

Figure 3.39: Analysis zonal RMSE of 500hPa u-wind fields in the cases of 3DVAR, serial EnSRF with 30 ensemble members and the localization scale of 3, and LEKF with 30 ensemble members, the local patch parameter $l = 3$ and covariance localization scale $\sigma = 1.5$.

3.7.2 Analysis error and ensemble perturbation growth

5-day forecasts were computed using serial EnSRF and 3DVAR data assimilation cycles. Here, 10 ensemble members are used in serial EnSRF with the localization length scale of 1. Fig.3.40 shows perturbation growths of analysis error and ensemble perturbations. Here, the energy norm (Houtekamer et al. 2005; Ehrendorfer and Errico 1995) is used to measure the perturbation amplitude:

$$E = \frac{1}{2}\left(u^2 + v^2 + \frac{c_p}{T_r}T^2 + R_d T_r \left(\frac{p_s}{p_0}\right)^2\right) \tag{3.15}$$

where E is energy, and constant-pressure heat capacity $c_p = 1005.7\mathrm{Jkg}^{-1}\mathrm{K}^{-1}$, dry-air gas constatnt $R_d = 287\mathrm{Jkg}^{-1}\mathrm{K}^{-1}$, reference temperature $T_r = 270\mathrm{K}$, and reference pressure $p_0 = 1000\mathrm{hPa}$ are constants. E is averaged over three-dimensional grids.

Fig.3.40 shows RMS values of forecast/analysis errors and ensemble perturbations. "RMSE" shows forecast/analysis RMSE, indicating the growth of the initial analysis errors. "SPREAD" shows forecast/analysis ensemble perturbation spreads, indicating the growth of ensemble perturbations. The perturbation amplitudes are averaged in time using whole February (112 samples) so that the figure shows one-month averages.

It is noted that 3DVAR analysis RMSE is almost equal to 5-day forecast RMSE of serial EnSRF. Thus, serial EnSRF provides much better analysis and forecast than 3DVAR in the perfect model case.

Both 3DVAR and serial EnSRF analysis errors grow as expected, and the ensemble perturbations of serial EnSRF also grow. The question of whether ensemble perturbations generated by EnKF data assimilation grow or not is not trivial. Houtekamer et al. (2003; 2005) indicated the ensemble perturbations in their EnKF scheme decay in the first few days. However, in their EnKF scheme, they added a model error covariance term to each

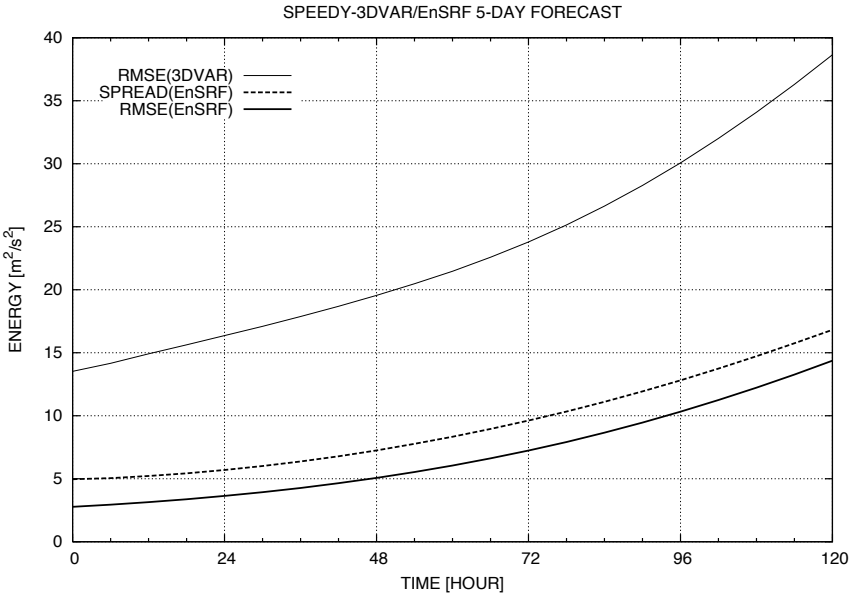

Figure 3.40: The growth of the initial analysis errors and ensemble perturbations in the cases of 3DVAR and EnKF. Thin solid line shows the growth of the initial 3DVAR analysis errors (RMSE(3DVAR)). Thick solid line and thick broken line show the growth of the initial EnKF analysis errors (RMSE(EnSRF)) and ensemble perturbations (SPREAD(EnSRF)), respectively.

ensemble. The perturbations due to the model error covariance term are obtained by the statistical average and are unrelated to the daily flow. Thus, we expect these perturbations are not dynamically fast growing perturbations, and because of their relatively large amplitude, they dominate the initial error decay in Houtekamer's results. Importantly in our results, the curves of ensemble perturbation and analysis error growth of EnKF are almost parallel. Thus, ensemble perturbation and analysis error growths show very similar growth rates (doubling in about three days), which implies both perturbation fields are dominated by similar dynamically growing errors.

The ensemble spread is larger than the analysis error, which may be because of the choice of covariance inflation parameter, (4% multiplicative inflation). This situation is opposite to what Houtekamer et al. (2005) indicated in their perfect model experiments (Fig.4 of Houtekamer et al. 2005). They observed that the ensemble spread is smaller than the analysis error. In the present configuration, it seems 4% inflation is too large, which is why the ensemble spread is larger. To confirm this anticipation, we applied online estimation of the inflation parameter to satisfy eq.(2.82) (cf. Section 2.3.6). We observed ensemble spread 2.07 and analysis error 2.425 in the energy norm (originally the spread was 4.85 and analysis error was 2.25). Thus, if we choose smaller covariance inflation, the analysis ensemble spread becomes smaller, the relative size between the analysis ensemble spread and analysis error can be adjusted through the covariance inflation parameter. However, although the ensemble spread was decreased to the analysis errors by changing the inflation parameter, the analysis error got slightly larger.

3.7.3 Error fields and ensemble spread

Since EnKF works as expected, ensemble members are expected to capture the error structures. Thus, we plot the error fields and ensemble spread in the case of serial EnSRF with 30 ensemble members and the Schur product length scale of 3.0. The top panel of Fig.3.41 shows the analysis error field at an arbitrary time, and the bottom panel shows the forecast error field at the same time. The analysis and forecast error fields are defined as analysis/forecast minus the true field. The shaded field in Fig.3.41 shows the analysis/forecast error field, the contour line shows the analysis/forecast ensemble spread. The ensemble spread represents the analysis/forecast error standard deviation captured by EnKF. Some areas (indicated by red rectangles) show similar structures between the error field and the ensemble spread. However, the correspondence is not very good, the pattern correlations between error fields and ensemble spread are 0.25 in the analysis and 0.18 in the 6-hour forecast. By contrast, the pattern correlation between analysis and forecast ensemble spreads are very high (0.96). Similarly, the pattern correlation between analysis and forecast error fields are also high (0.74).

3.7.4 Forecast errors, analysis errors, and bred vectors

In order to see the shape of the error fields, we plot the 2-day forecast error and analysis error fields valid at the same time. Fig.3.42 shows 2-day forecast error field (shaded) and analysis error field (contour) of u-wind at the 4th level ($\sigma = 0.51$) at an arbitrary time in the case of serial EnSRF with 10 ensemble members and the localization scale of 1. It is clear that the two fields have extremely similar shapes, confirmed by their high pattern correlation of 0.76. 2-day forecast error field is regarded as a dynamically growing error, the analysis field contains growing errors. Even with EnKF that considers dynamical error

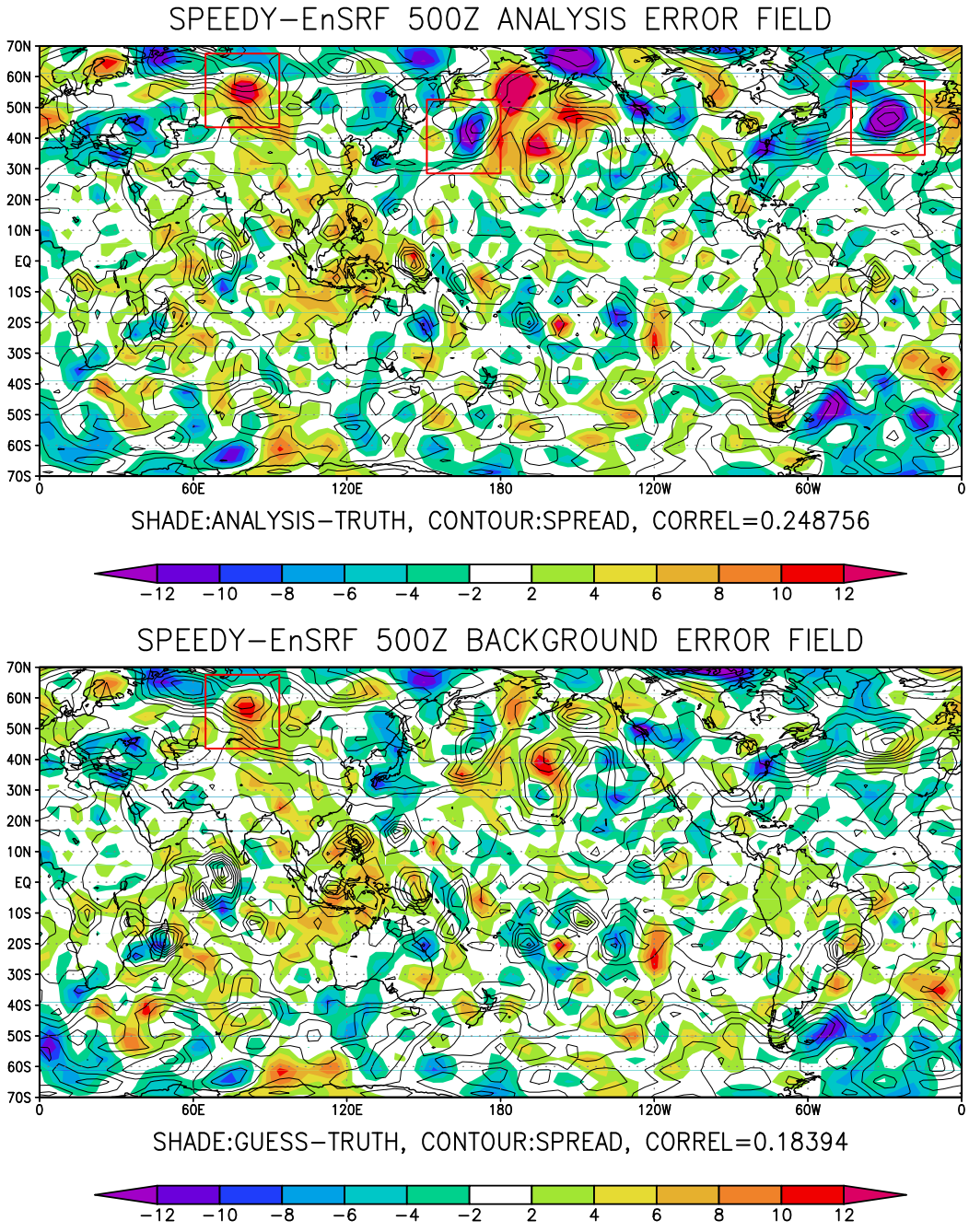

Figure 3.41: The error fields (ananlysis/forecast minus truth; shaded) and the ensemble spread of 500hPa height field (contour) at an arbitrary time. The top and bottom panels show analysis and background (6-hour forecast), respectively, both at the same time. Some parts have similar structures indicated by rectangles.

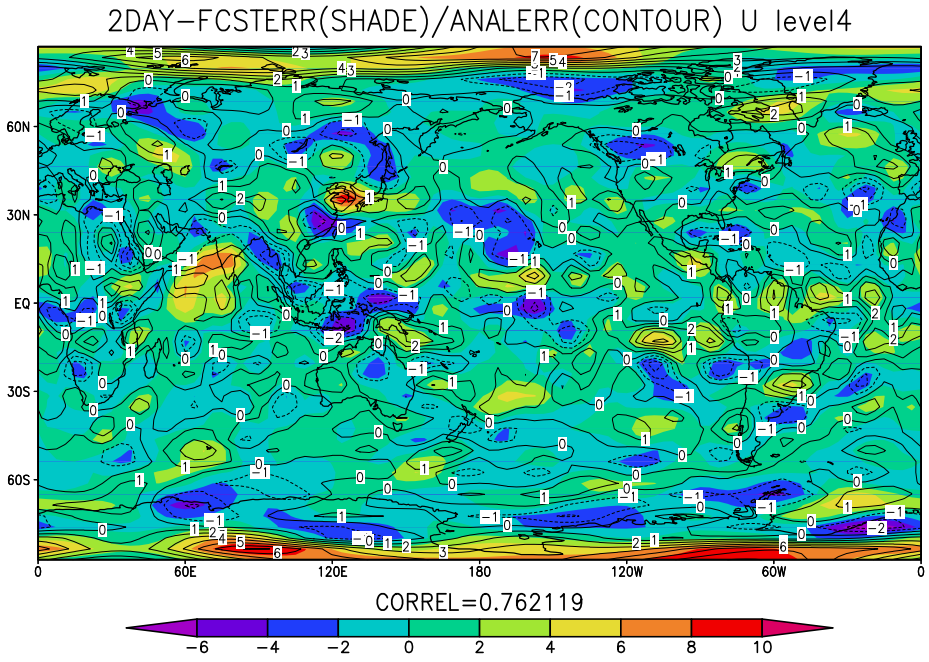

Figure 3.42: 2-day forecast error field (shaded) and analysis error field (contour) of u-wind at the 4th vertical level ($\sigma = 0.51$) at an arbitrary time in the case of serial EnSRF with 10 ensemble members and the localization scale of 1. The correlation between the two fields are 0.76.

structure, the analysis fields still contain growing errors. Fig.3.43 shows the same figure in the case of 30 ensemble members, which still shows high correlation (0.79). Fig.3.44 shows the same figure in the case of 3DVAR, which also shows high correlation (0.74).

The three figures show the same time with the same truth, so it is expected the dynamically growing error has similar shape. However, all the three figures show different shapes, which implies there are many possible dynamical "errors of the day" rather than a dominance by bred vectors.

One can observe high correlation of the two fields not only at this moment for the specific variable, but also at every time for all variables. Fig.3.45 shows temporal mean of the correlation coefficients for four model variables (u, v, T, q), all of which show high

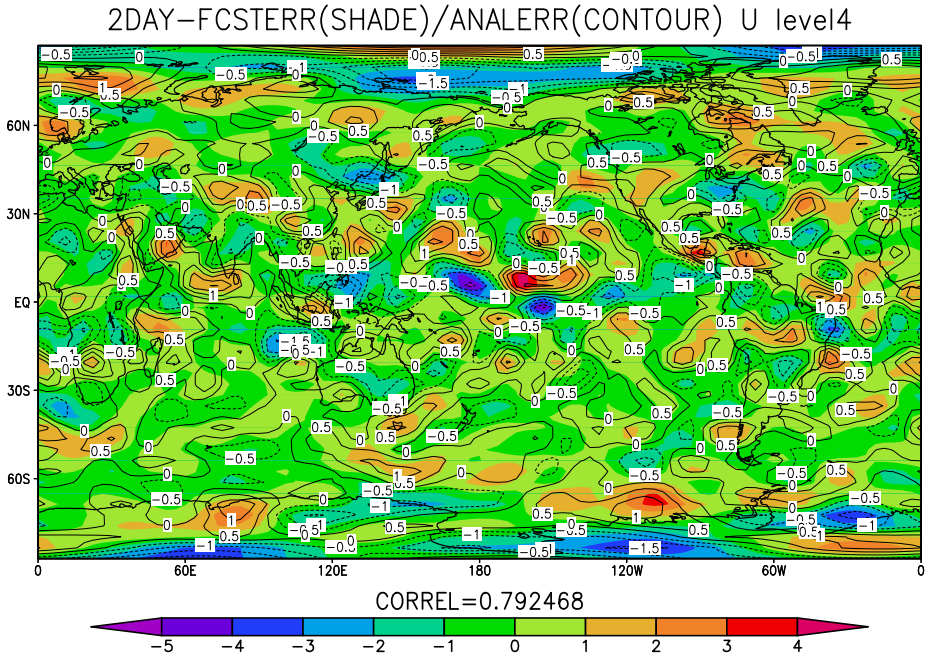

Figure 3.43: The same as Fig.3.42, but in the case of serial EnSRF with 30 ensemble members and the localization scale of 3. The correlation between the two fields are 0.79.

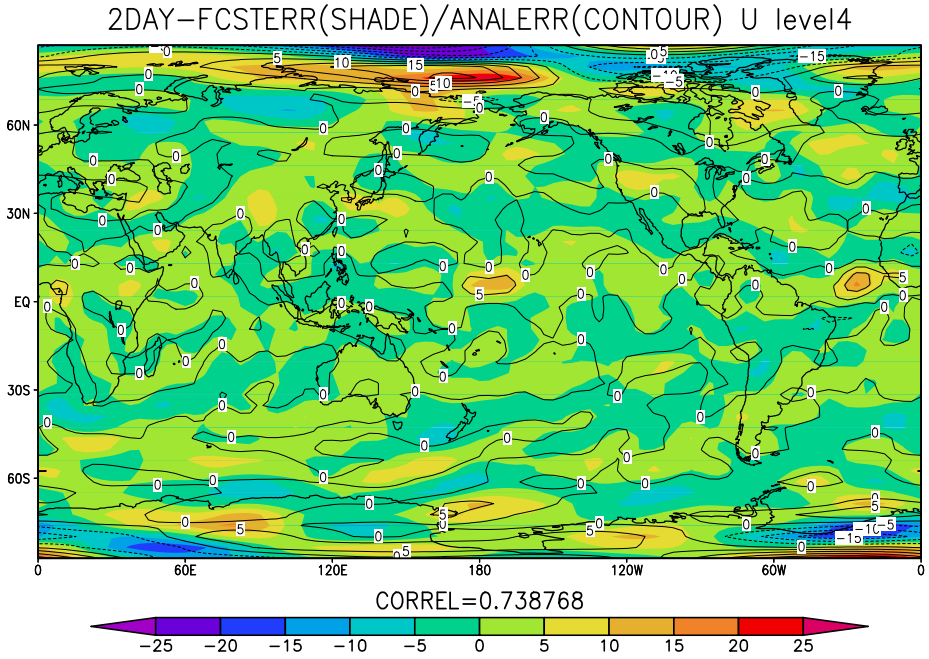

Figure 3.44: The same as Fig.3.42, but in the case of 3DVAR. The correlation between the two fields are 0.74.

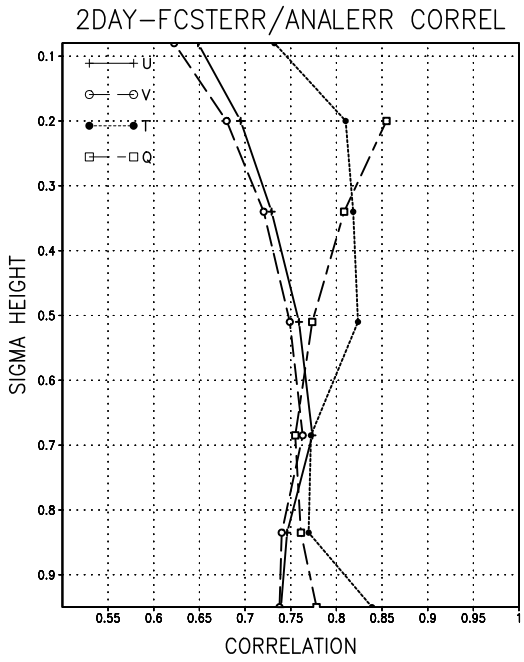

Figure 3.45: Temporal mean of correlation coefficients between 2-day forecast error and analysis error fields for four model variables (u, v, T, q) in the case of serial EnSRF with 10 ensemble members and the local patch parameter $l = 1$.

correlations around 0.75.

In order to further examine the shape of the error fields, 5-day and 2-day forecast errors, analysis error, E-dimension, and bred vector (BV) fields are shown in Fig.3.46. E-dimension, initially proposed by Patil et al. (2001), is a measure of substantial dimensionality of the space spanned by ensemble members, the definition is shown in the next section. Patil et al. suggested the usefulness of low E-dimensionality. BVs are generated by a breeding cycle starting from a field consisting of random numbers. The time interval of the breeding cycle is chosen as 6 hours, the rescaling is 3 m^2/s^2 with the energy norm (eq.(3.15)), though BVs are robust to the choice of norm. Fig.3.46 (g) shows E-dimension computed in a 5×5 local patch for a single variable without considering inter-variable correlations, (h) shows BV. There are some areas showing low E-dimension and large sig-

nals in BV such as the north west coast of the United States, the north western corner of Pacific ocean, western Siberia, and the southern end of south America. We find similar signals in 5-day forecasts (a) and (b) in these areas. In the 2-day forecasts (c) and (d), the signal in western Siberia is very clear, but others are not so clear. 2-day forecast may not be long enough to form growing modes. In the analysis error fields (e) and (f), we see a similar signal in the western Siberia with smaller amplitudes. The growing errors shown in BV and forecast fields almost disappear in analysis error fields.

3.7.5 E-dimension and explained variance

Szunyogh et al. (2005) have shown a strong anti-correlation between E-dimension and explained variance. E-dimension ϕ is a measure of substantial dimensionality of the space spanned by ensemble members, defined as

$$\phi = \frac{(\sum_{i=1}^{m} \lambda_i)^2}{\sum_{i=1}^{m} \lambda_i^2} \tag{3.16}$$

where m and λ denote the ensemble size and eigenvalues of an $m \times m$ matrix $\mathbf{E}^\top\mathbf{E}$, cf. eq.(1) of Patil et al. (2001). Here, $\mathbf{E}$ is an $N \times m$ matrix consisting of ensemble members. Explained variance is a measure showing how well the true error is spanned by the ensemble members. Let $\mathbf{G}$ denote a matrix consisting of eigenvectors of the matrix $\mathbf{E}^\top\mathbf{E}$, explained variance ξ is defined as

$$\xi = \frac{||\mathbf{G}\mathbf{G}^\top(\mathbf{x}^f - \mathbf{x}^t)||^2}{||(\mathbf{x}^f - \mathbf{x}^t)||^2} \tag{3.17}$$

The top and bottom panels of fig.3.47 show zonal averages of the time-mean E-dimension and explained variance, similar to Figs.5 and 6 of Szunyogh et al. (2005). Here, E-dimension is computed using a 3×3 local patch for all variables, inter-variable correlations are considered. Similar structure as Szunyogh et al. (2005) is observed: low

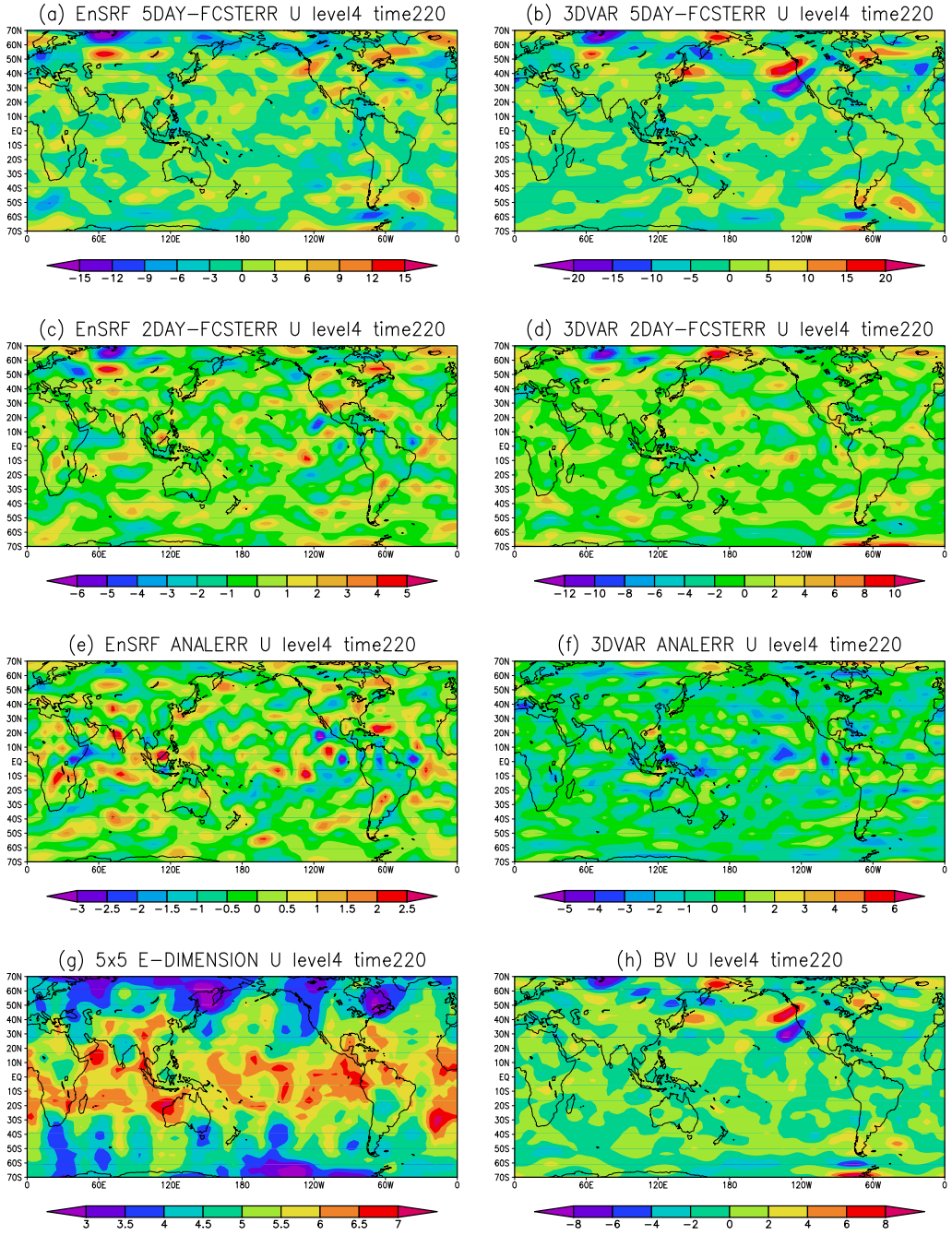

Figure 3.46: 5-day forecast error ((a) and (b)) and 2-day forecast error ((c) and (d)), analysis error ((e) and (f)), E-dimension (g), and bred vector (h) of u-wind at the 4th level all valid at the same time chosen arbitrarily, in the cases of serial EnSRF with 10 members and 3DVAR.

E-dimension and high explained variance appear in higher latitudes, high E-dimension and low explained variance appear in tropics, especially at middle levels and high level. Fig.3.48 shows the scatter plot between the zonal mean of the time-mean E-dimension and explained variance, indicating high anti-correlation, similar to Fig.7 of Szunyogh et al. (2005).

Without the zonal mean, Fig.3.49 shows the time-mean E-dimension (top panel) and explained variance (bottom panel) at the 4th level ($\sigma = 0.51$). Fig.3.50 shows the scatter plot, indicating a high anti-correlation of -0.66. The zonal mean is not essential to obtain the high anti-correlation between E-dimension and explained variance.

However, if we fix the time, we do not observe the strong anti-correlation. Fig.3.51 shows the same figure as Fig.3.49 but at a fixed time. Fig.3.52 shows the scatter plot between E-dimension and explained variance shown in Fig.3.51, indicating low anti-correlation of -0.26. There are a few points with low E-dimension and small explained variance, which explains the weak anti-correlation. If we focus on several regions showing especially low E-dimension, e.g. east of Australia, north-eastern Indian ocean, east coast of middle South America, and northern Pacific, we find correspondence of low E-dimension and large explained variance. In these several parts, EnKF could capture the true error shape and contribute to reduce the growing errors; we saw growing errors indicated by BV appear in such areas in the previous section.

3.7.6 Summary

- The analysis increment field of EnKF captures flow-dependent structures of the background error field, 3DVAR cannot capture the flow-dependent structures at all.
- 3DVAR showed a striped pattern in error fields, they almost disappeared in EnKF.

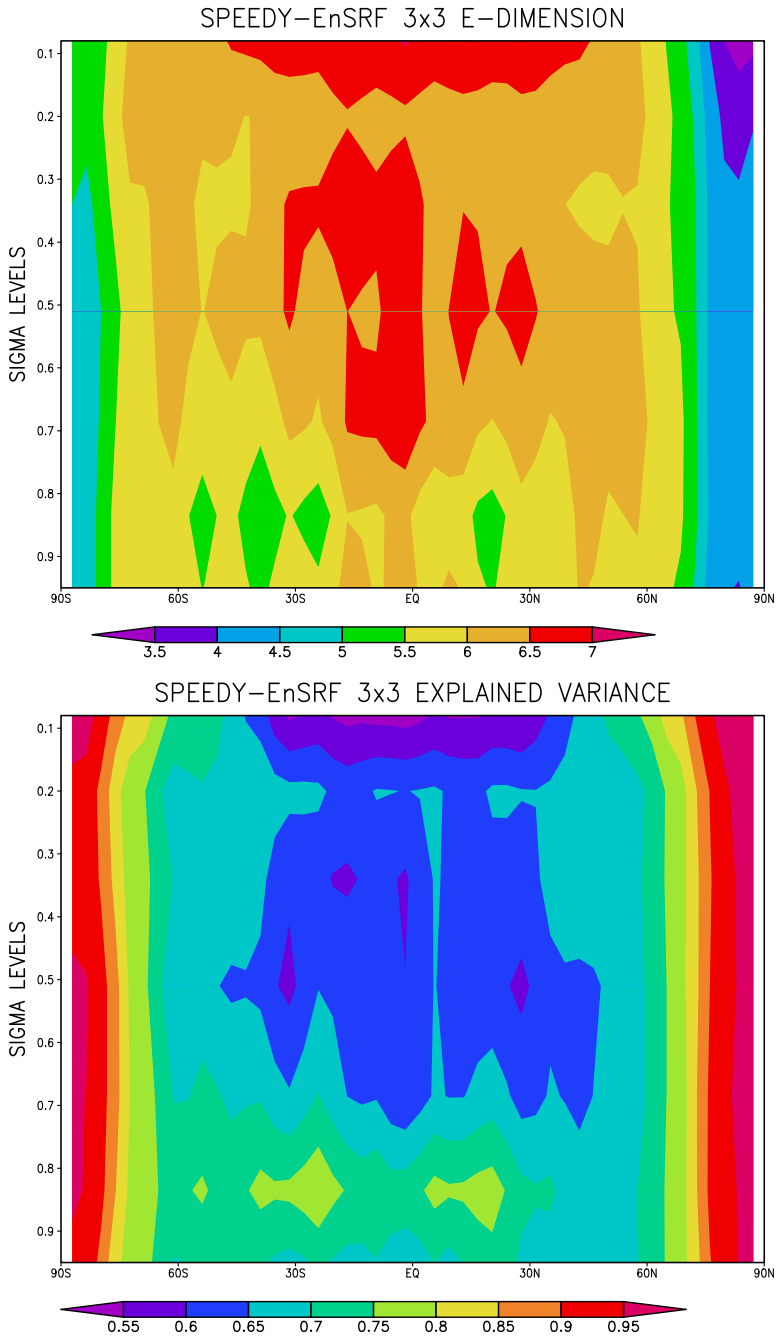

Figure 3.47: Zonal mean of the E-dimension (top panel) and explained variance (bottom panel) in the case of serial EnSRF with 10 ensemble members temporally averaged for one month (112 samples).

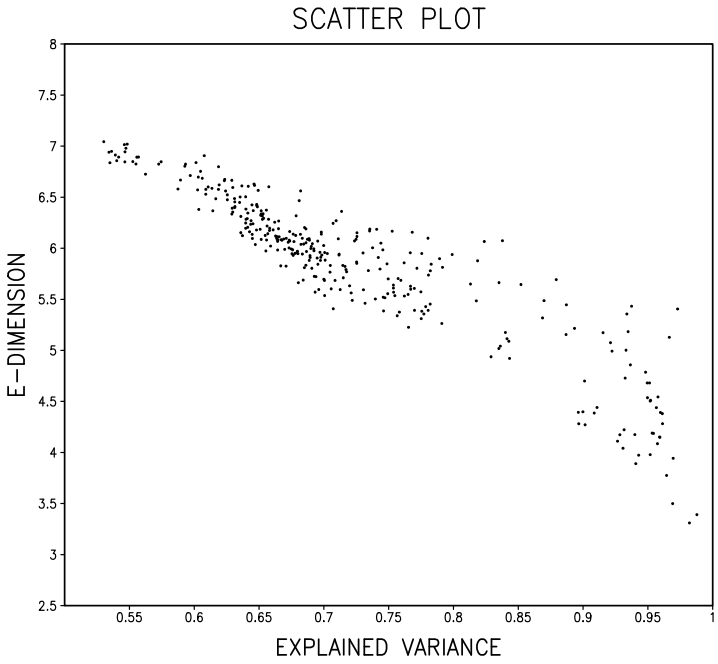

Figure 3.48: Scatter plot of the zonal mean of the time-mean E-dimension and explained variance, indicating high anti-correlation.

- Ensemble perturbations generated by serial EnSRF grow with a rate similar to the growth rate of the analysis error.
- Error and ensemble spread fields show similar signals in a few areas, but with low pattern correlations of about 0.2.
- 2-day forecast error and analysis error fields have very similar shapes, with correlations about 0.75. However, the error fields are different with different data assimilation schemes, suggesting the possibility of multiple "errors of the day".
- 5-day forecast error, 2-day forecast error and bred vector fields show similar signals in low E-dimensional areas.
- Time-mean fields of explained variance and E-dimension are anti-correlated with pattern correlation of about -0.66. Without the time-mean, only a few low E-dimension

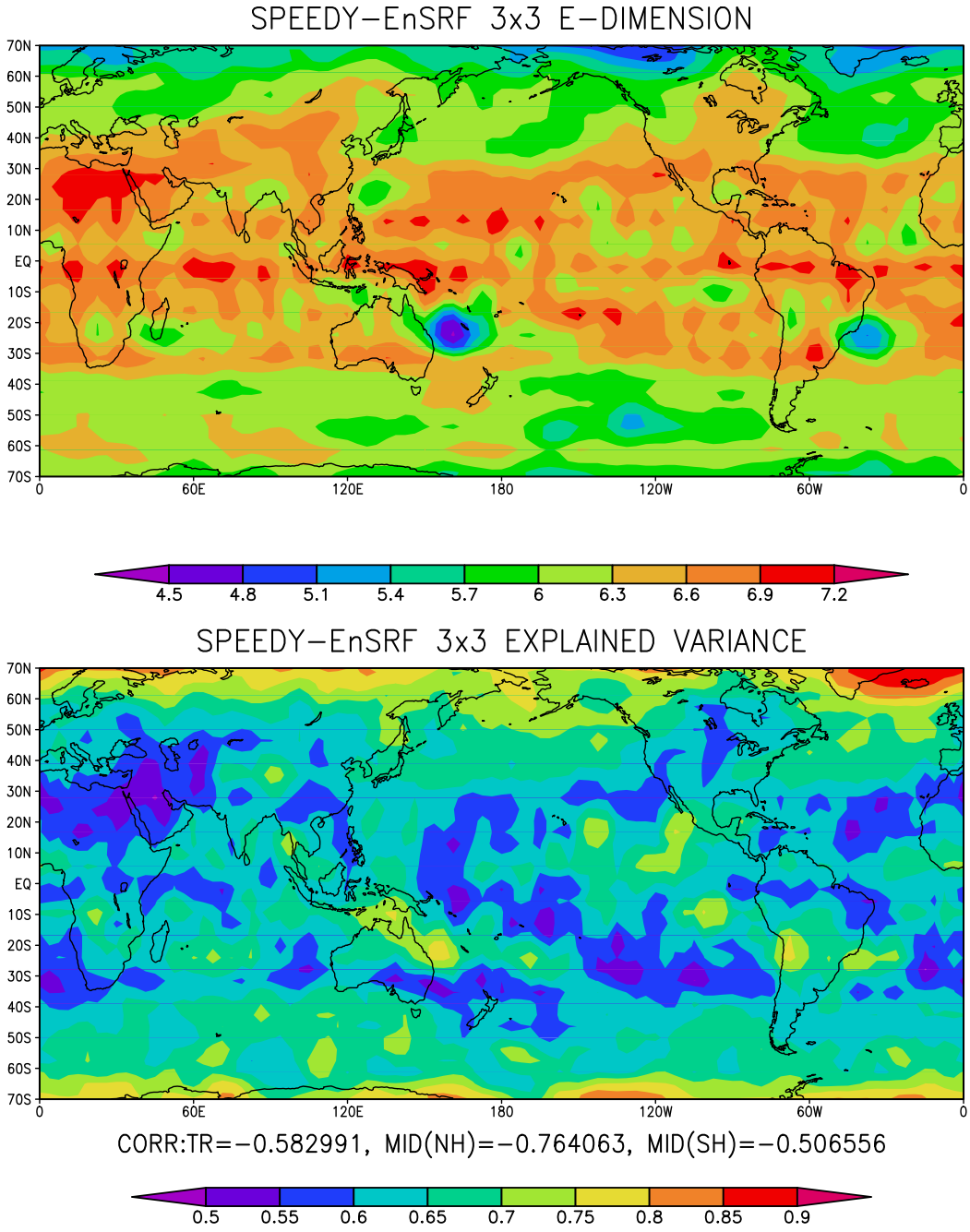

Figure 3.49: The E-dimension (top panel) and explained variance (bottom panel) at the 4th level in the case of serial EnSRF with 10 ensemble members temporally averaged for one month (112 samples).

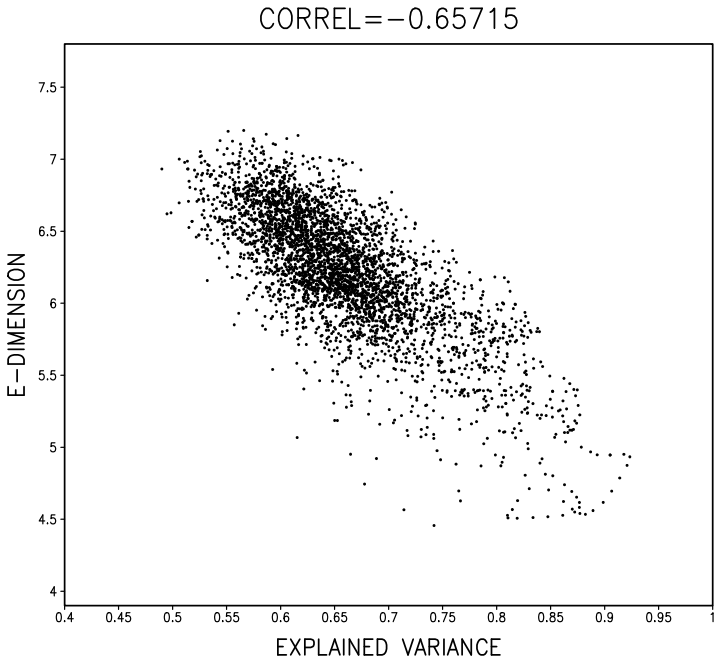

Figure 3.50: Scatter plot of the time mean E-dimension and explained variance at the 4th level, indicating high anti-correlation -0.66.

areas show good anti-correlation, with global pattern correlation about -0.26.

3.8 Discussion

In this chapter, three methods (3DVAR, serial EnSRF, LEKF) are implemented and applied on the SPEEDY model under the perfect model assumption. Our results show that the EnKF clearly outperforms 3DVAR. One surprising result is that for different methods and number of ensemble members the 2-day forecast "errors of the day" are very similar to the analysis errors. However, they are not necessarily similar among different methods, suggesting there are multiple growing "errors of the day". In low E-dimensional regions, however, the errors tend to have similar structures. Bred vectors also show similar structures in those regions. The initial analysis errors and ensemble perturbations show very similar growth rates, suggesting the fields contain growing "errors of the day".

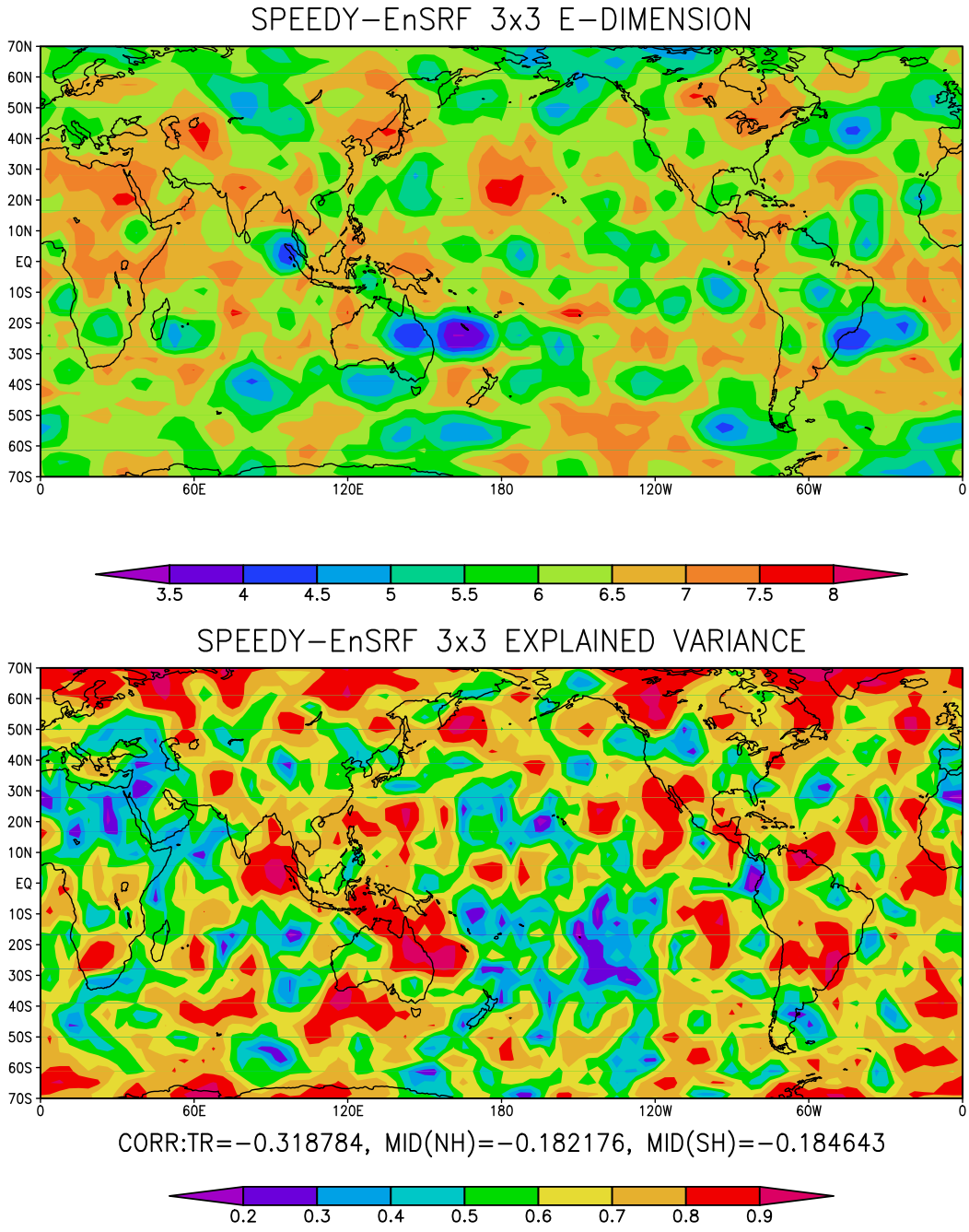

Figure 3.51: The E-dimension (top panel) and explained variance (bottom panel) at the 4th level in the case of serial EnSRF with 10 ensemble members at an arbitrary time.

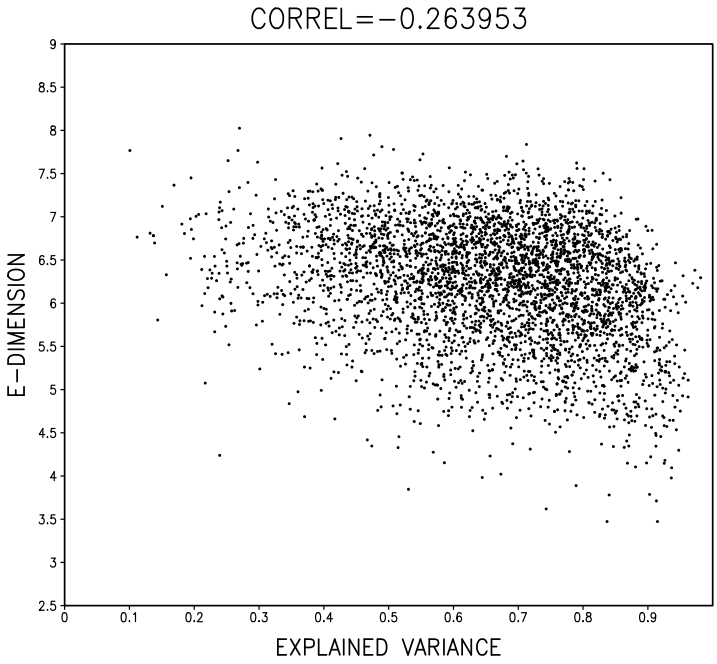

Figure 3.52: Scatter plot of E-dimension and explained variance at the 4th level at an arbitrary time, indicating anti-correlation -0.26.

Overall, our results suggest that serial EnSRF outperforms LEKF, but this advantage is substantially reduced when we localize the observational error covariance or increase the number of ensemble members. A possible reason of the disadvantage of LEKF is that there are many possible ways to combine overlapped local patches, our implementation of LEKF may not be optimal. We average the overlapped local analyses in the smaller local patch denoted by l_2 using Heaviside-like step weighting. We see no clear dependencies on l_2. A natural way is to use Gaussian-like weighting according to the distance from the center point. Further investigation is required to optimize the way to combine overlapped local patches in the LEKF. It is possible that the formulation of the local patches over the poles may not be optimal in the LEKF.

In a realistic case with a parallel computer and a large number of observations, LEKF should be much faster than the serial EnSRF, especially with the new LETKF

developed by Hunt (pers. comm.). As a result, in an operational setup, LEKF/LETKF may be the only computationally feasible choices. However, in the present experiments, we used a single computer and a small number of observations. For this case, LEKF has shown no advantage against serial EnSRF; serial EnSRF performs better with less computational time. Thus, we choose serial EnSRF for the last part of the experiments.

Chapter 4

Data assimilation in the presence of model errors

4.1 Introduction

4.1.1 Overview

In Chapter 3, three data assimilation methods (3DVAR, serial EnSRF, and LEKF) are applied on the SPEEDY model. EnKF clearly outperforms 3DVAR under the perfect model assumption. In this chapter, we remove the perfect model assumption using the NCEP/NCAR reanalysis (NNR) fields (Kalnay 1996) as the nature run. We explore the following questions in this chapter:

1. What do model error bias and EOFs (empirical orthogonal functions, e.g. Wilks 1995) look like?

2. What are the effects of model errors on data assimilation?

3. How does model bias estimation (Dee and da Silva 1998) work with EnKF?

We introduce NNR and compute model error statistics by taking differences between NNR fields and the SPEEDY model 6-hour forecasts initialized by NNR fields. Then, the effects of model errors on data assimilation are investigated. We perform data assimilation cycle experiments using the same data assimilation systems as in Chapter 3, but replacing the nature run by NNR fields. We describe theory on model bias estimation following Dee and da Silva (1998) and apply the estimation method with EnKF. We also apply low-order bias estimation, assuming the bias field is spanned by a limited number of base fields. As base fields, we use the model error bias field and EOFs.

4.1.2 OSSE methodology

So far, in this research, we followed OSSEs (Observing Systems Simulation Experiments) as described in sections 1.3 and 3.4. In OSSE methodology, we generate observations using the nature run generated by the model forecasts. So far, we have used the same model for both generating observational data and performing data assimilation. This perfect model experiment is over optimistic partly because such observational data do not contain the error of representativeness. Alternatively, Atlas (1997) suggested to generate observations using a more sophisticated and higher resolution model, different from the model used in data assimilation. This considers the error of representativeness as well as the imperfection of the model. Similarly in this chapter, we use NNR fields to generate observational data, the NNR fields are generated by a relatively recent sophisticated model (1995 NCEP model) and data assimilation system. The difference from the Atlas's approach is that we consider the real nature. Real observations are assimilated in NNR fields, NNR fields simulate the evolution by the unknown "true nature" model.

4.1.3 NCEP/NCAR reanalysis (NNR)

The NCEP/NCAR reanalysis (NNR) fields (Kalnay 1996; Kistler et al. 2001) are a fairly good estimate of the true atmospheric state for the past 50 years. They are generated by a relatively recent (1995) sophisticated model and data assimilation system. We assume the NNR fields as a proxy for the truth. The validity of this assumption is beyond the scope of this research. We randomly perturb the NNR fields to simulate observations with the same standard deviation (Table 3.2), we verify our analysis fields against the NNR fields.

NNR data are freely available on the Internet. The NNR data on pressure levels

have a different grid structure ($144 \times 73 \times 17$) than the SPEEDY model ($96 \times 48 \times 7$). The NNR fields need to be interpolated to obtain the fields on the SPEEDY model grids. The horizontal interpolation is a first order linear interpolation. No vertical interpolation is required to obtain pressure level data because the reanalysis data have all seven pressure levels: 925, 850, 700, 500, 300, 200, and 100hPa. A first order linear interpolation is applied to obtain sigma level data. Using horizontally interpolated p_s of the reanalysis data, $p_s \times \sigma = p$ is computed to get the target pressure level for the interpolation. The reanalysis field is interpolated for the target level. $\log p$ is used as a vertical coordinate for the linear interpolation.

Since the original NNR fields have a higher resolution, they contain wavenumber components higher than those which the SPEEDY model can resolve. Wavenumber components higher than 30 cannot be detected in the SPEEDY model since it has spectral triangular truncation of 30 wavenumbers. The reanalysis fields are smoothed out by the spectral transformation of the SPEEDY model.

For the present experiments, we downloaded the NNR fields from the NCAR's website and generated the interpolated and smoothed fields during the period from January 1, 1982 to February 28, 1982. This period coincides with the satellite era that started in 1979. For simplicity, hereafter we call the interpolated and smoothed NNR fields "reanalysis fields".

4.2 Model error statistics

In this section, model error statistics (bias and EOFs) are computed and shown.

4.2.1 Model error biases

In order to estimate the model error against the reanalysis fields, the reanalysis fields are given as initial conditions to the SPEEDY model. Since the reanalysis fields are evolved by an unknown model expected to simulate the true atmospheric dynamics, the differences between the SPEEDY model 6-hour forecasts and reanalysis fields valid at the same time represent the model errors. We assume the difference is a sample of model errors.

Fig.4.1 shows the differences of the three prognostic variables (u, v and T) between the SPEEDY model 6-hour forecasts and reanalysis fields, temporally averaged for two months using 235 samples in the period from January 1, 1982 to February 28, 1982. Temporally averaged differences are interpreted as the model bias. Large areas are not shaded by colors, suggesting that the SPEEDY model simulates the real atmosphere fairly well. The largest model bias of the wind fields can be seen in polar regions. Some differences are caused by the orographic effect, most clearly seen in the u-wind field at 200hPa level, where there is a clear negative bias downstream of the Himalayas. As for temperature fields, a strong negative bias is observed at lower levels over Antarctica, though almost no clear bias can be seen at 500hPa level. Strong biases are observed mostly over land at lower levels.

Fig.4.2 shows the same figure as Fig.4.1 but for the variables q (specific humidity), z (height) and p_s (surface pressure). q biases are large in the tropics. Since z is a diagnostic variable obtained from p_s and T, it is clear that z biases are strongly affected by the p_s bias field. It seems that the orographic difference in the reanalysis field and the SPEEDY model affects the p_s model bias. However, it still shows some biases over the ocean, e.g. positive bias over the western tropical pacific basin, which disappears in the z bias field in the upper troposphere.

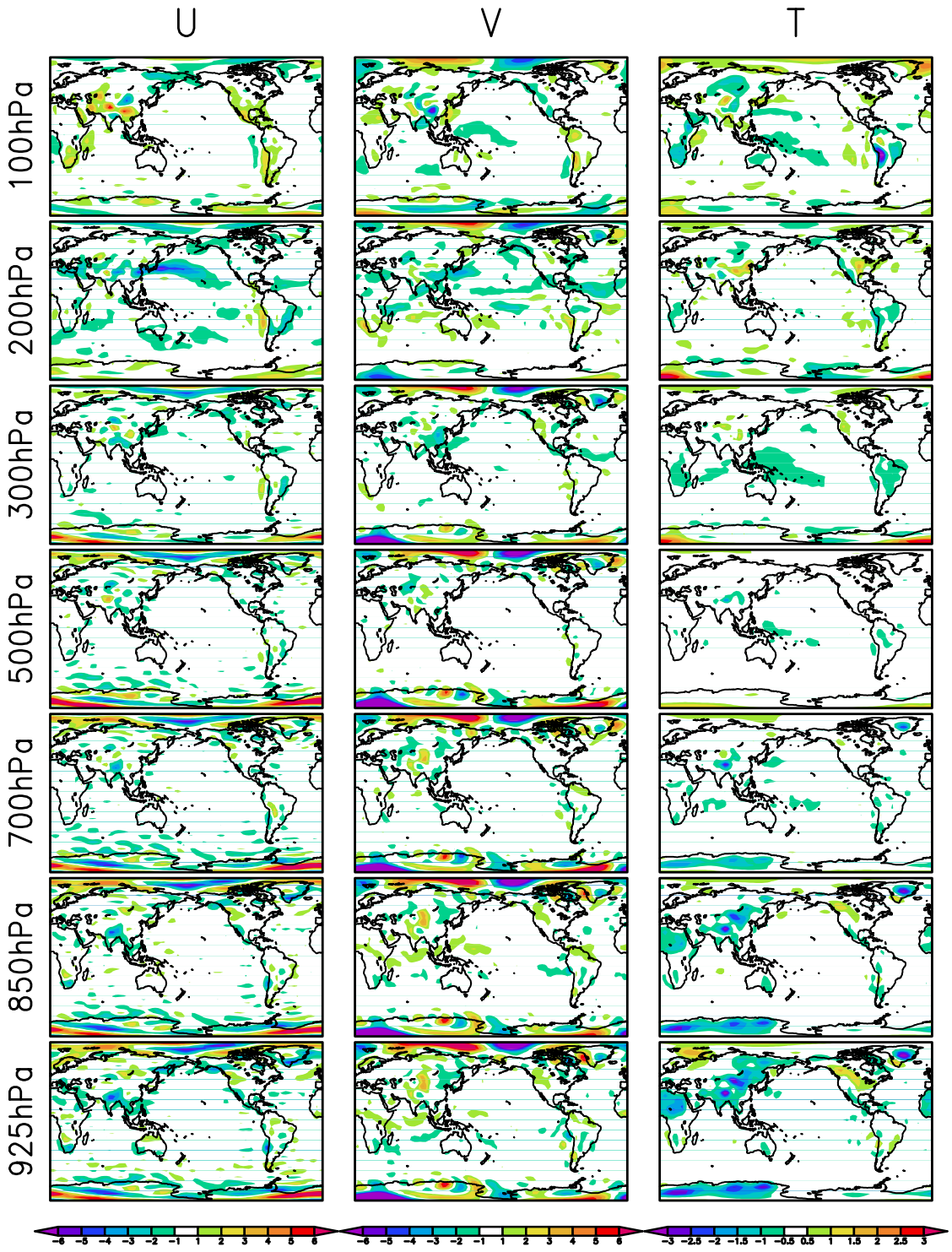

Figure 4.1: SPEEDY model biases of u, v [m/s] and T [K] relative to the NCEP/NCAR reanalysis fields.

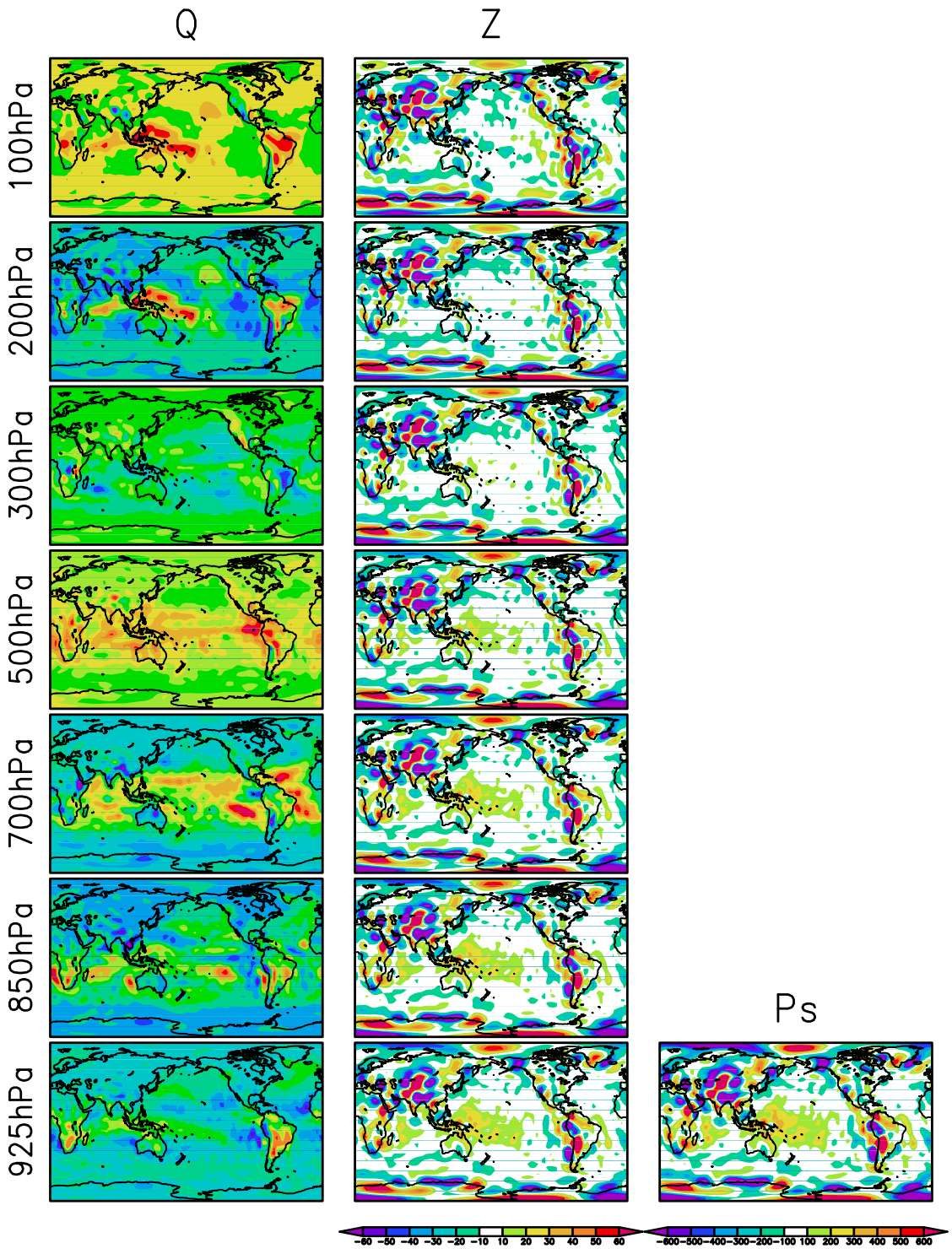

Figure 4.2: SPEEDY model biases of q, z [m] and p_s [Pa] relative to the NCEP/NCAR reanalysis fields. Each vertical level has different scales for q fields, it is omitted. Red color shows larger values.

4.2.2 EOFs

We computed empirical orthogonal functions (EOFs, see for example, Wilks 1995) of the two-month model error samples. EOFs are defined as eigenvectors of the covariance matrix $\mathbf{C}$ of the model error anomaly $\mathbf{x}'(t,s)$ where t and s denote time and space indices, respectively:

$$\begin{aligned} \mathbf{C} &= \mathbf{x}'(t,s)^\top \mathbf{x}'(t,s) \\ &= \mathbf{V}\mathbf{D}\mathbf{V}^\top \end{aligned} \tag{4.1}$$

The model error anomaly $\mathbf{x}'(t,s)$ is defined by subtracting the time-mean model bias field $\bar{\mathbf{x}}(s)$ from time series of model error fields $\mathbf{x}(t,s)$:

$$\mathbf{x}'(t,s) = \mathbf{x}(t,s) - \bar{\mathbf{x}}(s) \tag{4.2}$$

$\mathbf{V}$ is an $N \times N$ matrix composed of eigenvectors, $\mathbf{D}$ is an $N \times N$ diagonal matrix composed of eigenvalues, where N denotes the number of the space index. Usually, N is a number of $\mathcal{O}(10^4)$ or larger, the number of time index m is smaller. As a result, the covariance matrix is degenerate. Ignoring zero eigenvalues, we have $\mathbf{V}$ as an $N \times m$ matrix and $\mathbf{D}$ as an $m \times m$ diagonal matrix. We write the singular value decomposition of $\mathbf{x}'(t,s)$

$$\mathbf{x}'(t,s) = \mathbf{U}\mathbf{D}^{1/2}\mathbf{V}^\top \tag{4.3}$$

where $\mathbf{U}$ is an $m \times m$ matrix composed of left singular vectors, $\mathbf{D}^{1/2}$ is an $m \times m$ diagonal matrix composed of singular values, and $\mathbf{V}$ is an $N \times m$ matrix composed of right singular vectors. Since singular vectors are orthonormal, $\mathbf{U}^\top\mathbf{U} = \mathbf{I}$ is satisfied. Using eq.(4.3), we get EOFs as follows:

$$\begin{aligned}
\mathbf{x}'(t,s)^\top \mathbf{x}'(t,s) &= \left(\mathbf{U}\mathbf{D}^{1/2}\mathbf{V}^\top\right)^\top \left(\mathbf{U}\mathbf{D}^{1/2}\mathbf{V}^\top\right) \\
&= \mathbf{V}\mathbf{D}^{1/2}\mathbf{U}^\top\mathbf{U}\mathbf{D}^{1/2}\mathbf{V}^\top \\
&= \mathbf{V}\mathbf{D}\mathbf{V}^\top \qquad (4.4)
\end{aligned}$$

$\mathbf{V}$ is equal to that in eq.(4.1), containing EOFs. There is an algorithm to find the largest singular values and corresponding singular vectors, we obtain first few EOFs with a little computation.

Here, we take space index s by combining variables (u, v, T, and q) but separating levels. In this way, inter-variable correlations are considered but no correlations between levels are considered. Fig.4.3 shows the first EOF (top panel) and the second EOF (bottom panel) of temperature fields at the bottom level (shades) and surface pressure fields (contour). There is a clear signal corresponding to diurnal variation with wave number 1 over longitudes. We also see the phase difference between the first and second EOFs. The SPEEDY model considers only daily updated forcings and no diurnal variation. Temperature at the bottom level is strongly affected by insolation, showing strong diurnal signals over land. Other variables at other levels do not show this kind of clear structures. The explained variances are 20%, 15%, 12%, 10% in the first four EOFs. Fig.4.4 shows time series of expansion coefficients of the first four EOFs for a week. We see a clear signal of the diurnal cycle for the first and second EOFs. The second EOF shows diurnal fluctuation with a slightly smaller amplitude and a different phase than the first EOF. Fig.4.5 shows time series of expansion coefficients of the first EOF (top panel) and the third and fourth EOFs (bottom panel) in a longer time range. Diurnal fluctuation is dominant in the first EOF, but we also see longer time fluctuation with a small amplitude. We see

a very similar time series for the second EOF. In the third and fourth EOFs, there are longer time fluctuations spanning positive and negative values with periods of about 1-2 weeks. Since the training period was only 50 days, it is clearly insufficient to estimate in a robust way variable errors with a time scale not much smaller than the training period. Although the explained variances do not show a clear jump after the second EOF, the fields themselves and expansion coefficients indicate that with this short training period, only the first two EOFs (representing the error due to the absence of diurnal cycle) can be estimated in a statistically significant fashion.

4.3 Effects of model errors on data assimilation

While the previous section showed bias and EOFs of the model errors, the effects of systematic model errors on data assimilation are not clear. To see the effects, we perform the same data assimilation experiments as the perfect model experiments using the reanalysis fields as truth. In other words, the same data assimilation systems are applied, the only difference is the nature run is replaced by the reanalysis fields. In this chapter, only serial EnSRF is used as representative of EnKF as described in Chapter 3, we will refer to it as "EnKF" hereafter. The parameter values are chosen as optimal, shown in Table 3.9, Chapter 3.

Fig.4.6 shows the analysis RMSE of 500hPa height field. Even EnKF with 30 ensemble members, the best among the three, shows about 60 meters of analysis RMSE. 3DVAR shows the largest temporal fluctuation and largest mean errors. However, the performance of EnKF is more than 10 times worse than in the perfect model case, and the ratio of 3DVAR to EnKF is clearly smaller than in the perfect model case. This means that EnKF is more strongly affected by the model errors.

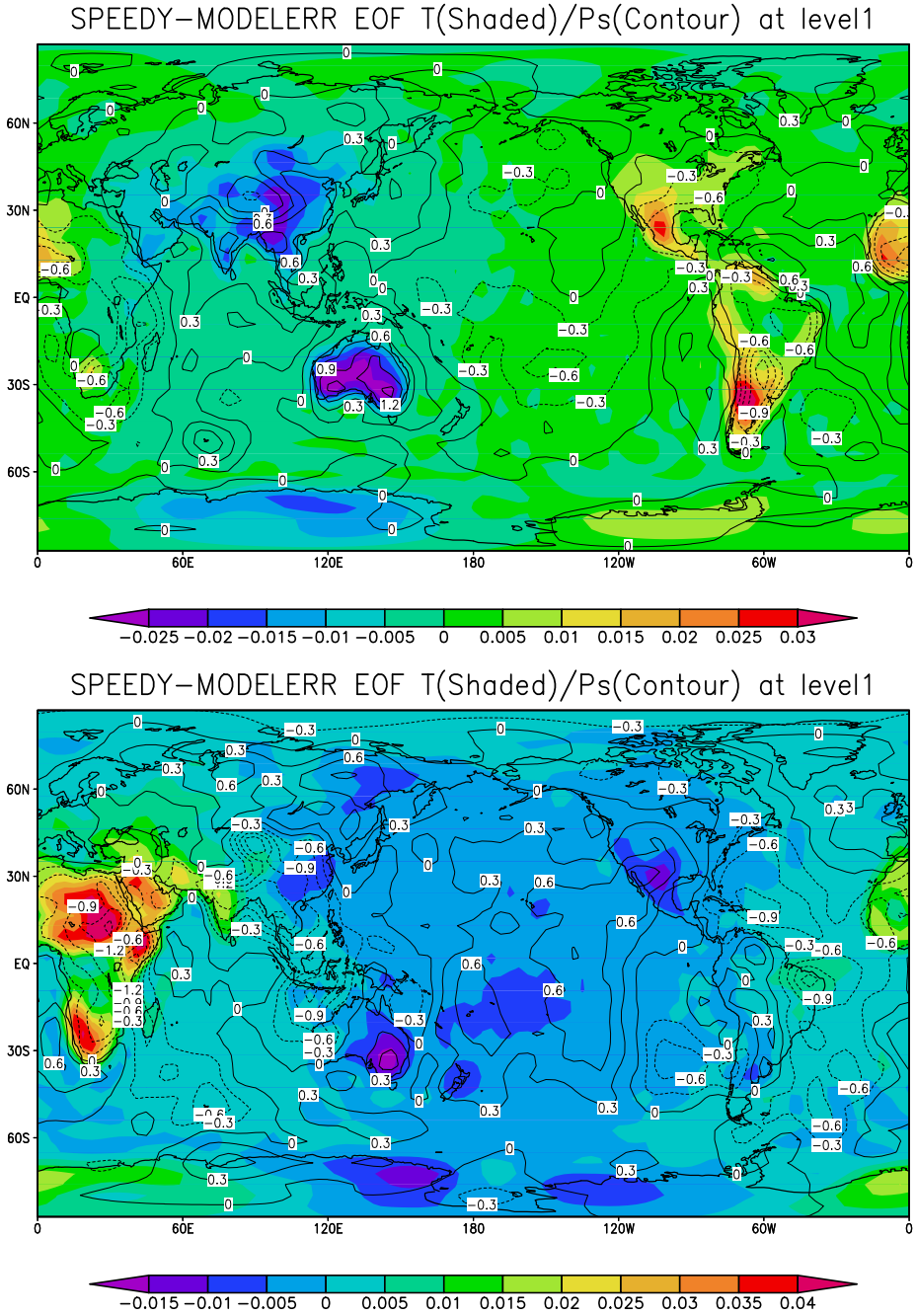

Figure 4.3: EOFs of two-month model error samples from January 1, 1982 to February 28. The top and bottom panels show temperature fields at the bottom level (shades) and surface pressure fields (contour) of the first and second EOFs, respectively.

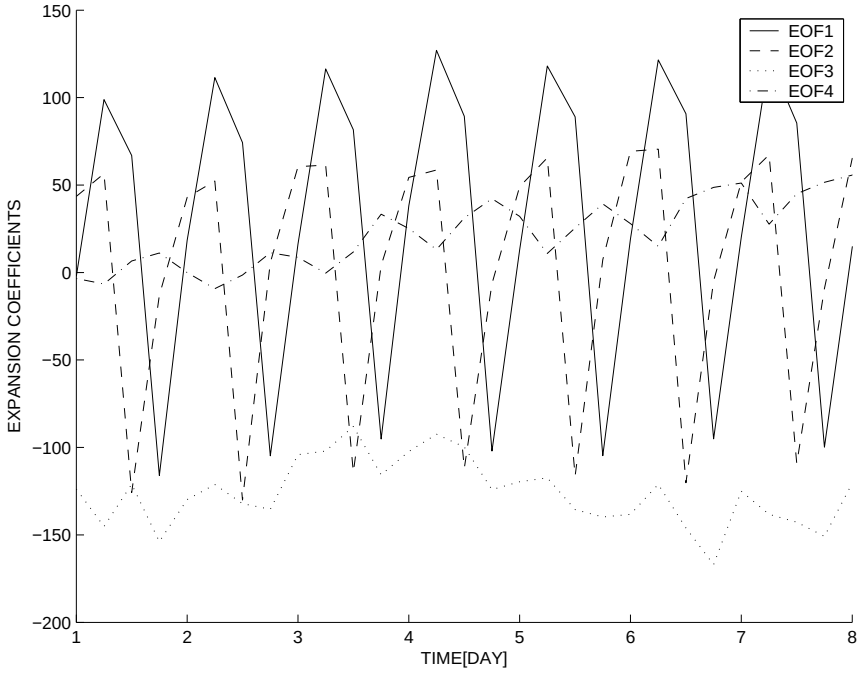

Figure 4.4: Time series of expansion coefficients of the first to fourth EOFs for a week. Solid, dashed, dotted, dash-dotted lines show first, second, third, and fourth EOFs, respectively.

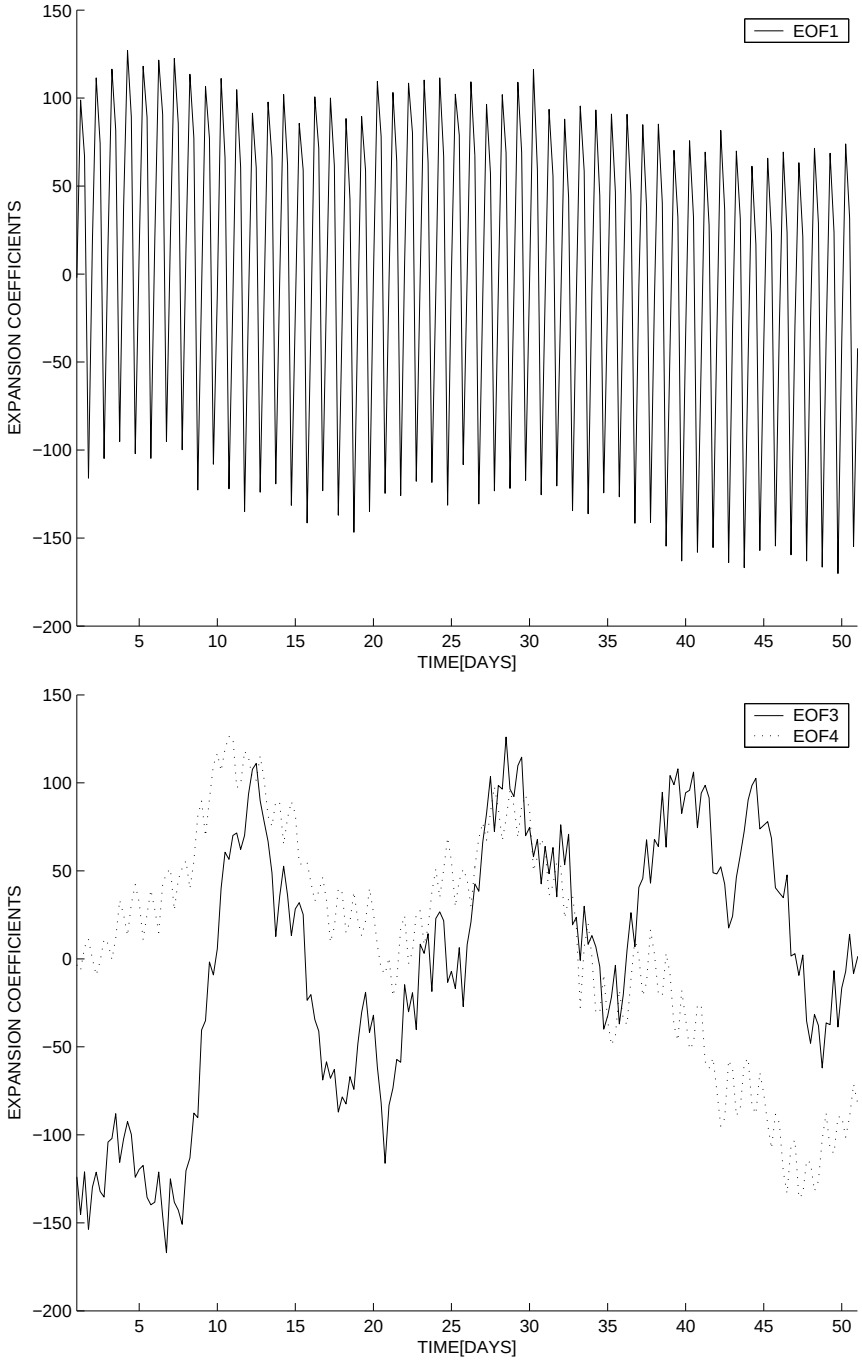

Figure 4.5: Time series of expansion coefficients of the first EOF (top panel) and the third (solid line) and fourth (dotted line) EOFs (bottom panel) for 50 days.

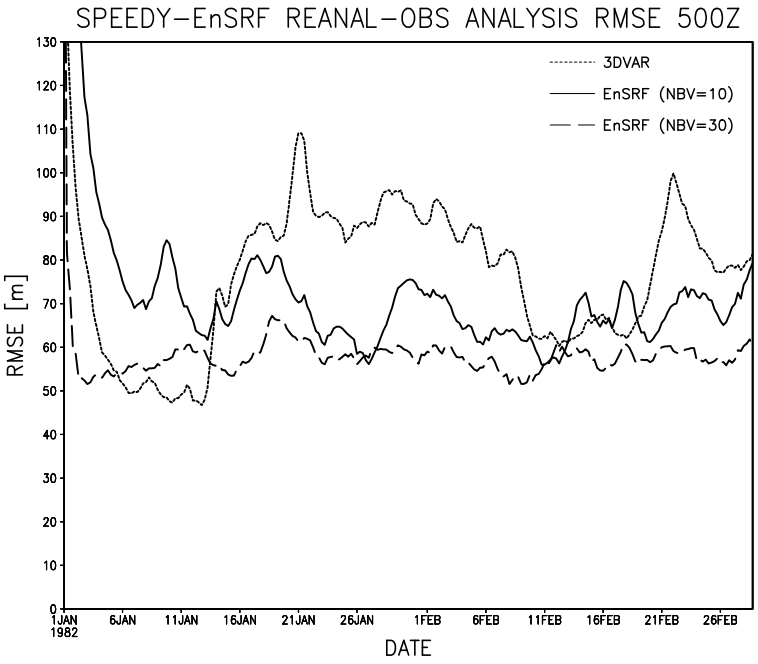

Figure 4.6: Analysis RMSE of 500hPa height field of 3DVAR (short dashed line) and EnKF with 10 ensemble members (solid line) and 30 ensemble members (long dashed line) in the case of the regular sparse observational network (Fig.3.10) in the presence of model errors.

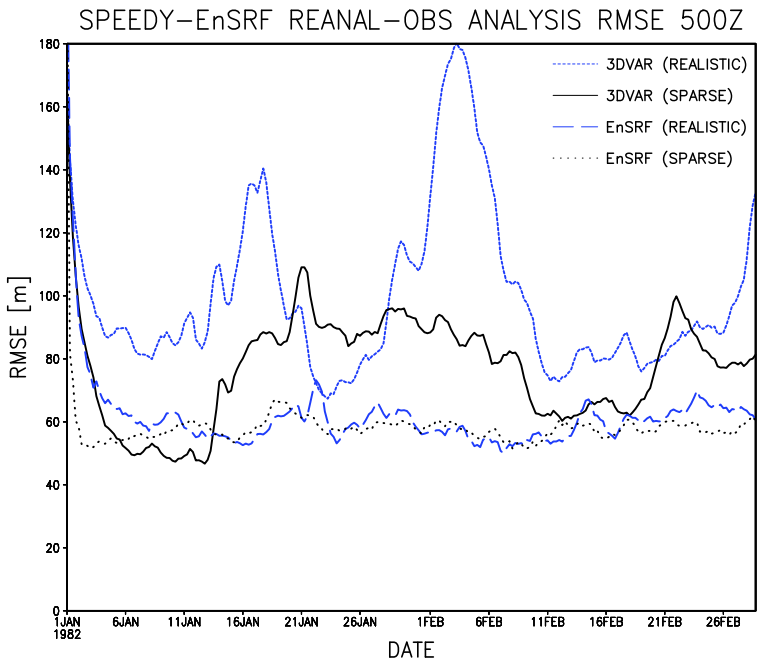

Figure 4.7: Analysis RMSE of 500hPa height field of 3DVAR and EnKF with 30 ensemble members in the case of the sparse and realistic observational networks in the presence of model errors. Blue lines show the cases of realistic observational network, black lines show the cases of sparse observational network. As for blue lines, short dashed line and long dashed line show 3DVAR and EnKF, respectively. As for black lines, solid line and dotted line show 3DVAR and EnKF, respectively.

When the realistic observational network (Fig.3.11) is applied, the advantage of EnKF with respect to 3DVAR becomes larger. Fig.4.7 shows the analysis RMSE of 500hPa height field in the cases of both the sparse and realistic observational networks. Clearly, the difference between 3DVAR and EnKF is larger in the case of the realistic observational network. 3DVAR is strongly affected by the different observational network, the temporal fluctuation is larger with the realistic network. By contrast, EnKF is not sensitive to the choice of the observational network.

Fig.4.8 shows time-mean RMSEs of other variables at other levels. Importantly,

3DVAR and EnKF with 10 ensemble members show similar results. Significant difference is only seen in height and temperature fields at a few levels; EnKF shows advantage in height fields at middle and upper levels, 3DVAR shows advantage in temperature fields at 500, 200, and 100hPa levels. The fact that EnKF is less sensitive to the choice of observational network than 3DVAR is generally valid for other variables at other levels except temperature and specific humidity, and height fields at upper levels. There is almost no difference among the five data assimilation methods in specific humidity fields.

4.4 Model bias estimation

The previous section showed that model errors affect data assimilation, especially in EnKF. The advantage of EnKF to 3DVAR became smaller than the perfect model case. In this section, model bias estimation (Dee and da Silva 1998) is introduced and applied to improve the filter performance in the presence of model errors.

4.4.1 Theory

We follow the same approach as Dee and da Silva (1998) for bias estimation within KF. In Section 2.2, we assumed a perfect model. Model imperfections introduce additional suboptimality to the data assimilation. Thus, we need to modify eq.(2.18) to include a model bias:

$$\tilde{\mathbf{x}}^f_{i+1} = M(\tilde{\mathbf{x}}^a_i) - \mathbf{b}^a_i \tag{4.5}$$

Here, $\tilde{\mathbf{x}}$ denotes an unbiased state. Since the model M has a bias $\mathbf{b}$, it is necessary to subtract the bias to get the unbiased forecast. The bias also evolves in time:

$$\mathbf{b}^f_i = M_b(\mathbf{b}^a_{i-1}) \tag{4.6}$$

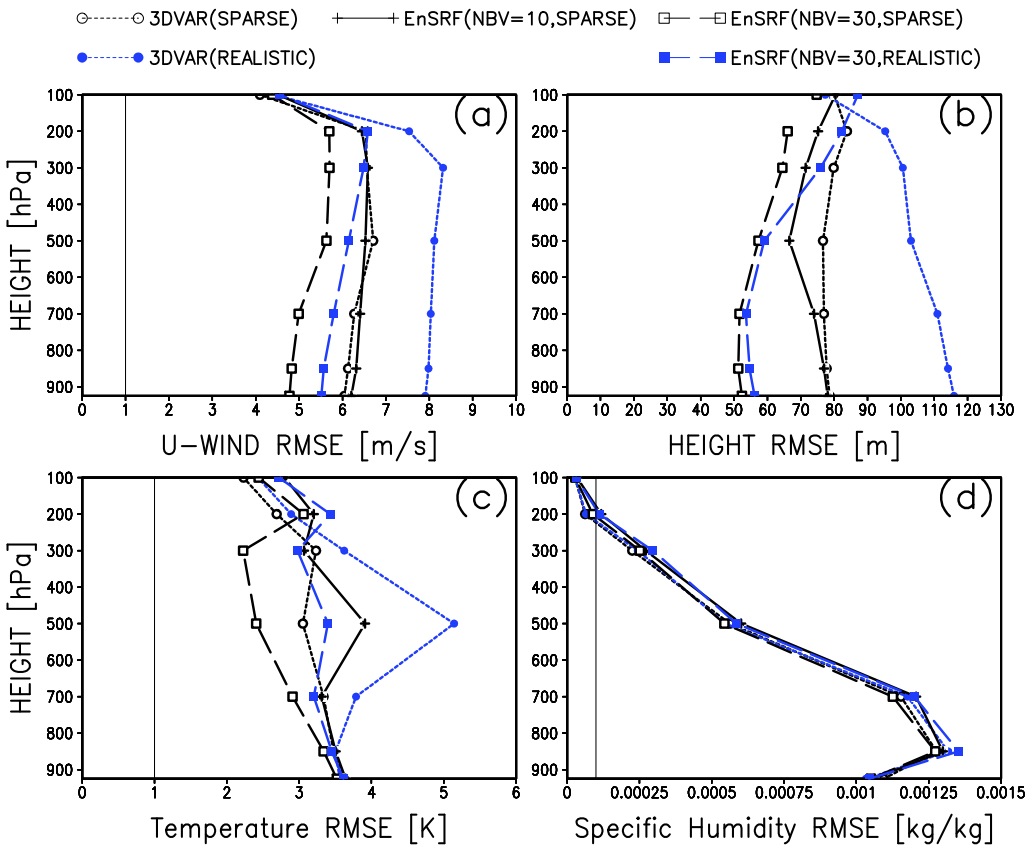

Figure 4.8: Analysis RMSE at the all pressure levels temporally averaged for one month after the initial one-month spin-up period (112 samples) in the cases of 3DVAR (short-dashed lines), EnKF with 10 ensemble members (solid line), EnKF with 30 ensemble members (long-dashed lines). Black and blue lines show the cases with the regular sparse observational network and the realistic observational network, respectively. The four panels (a), (b), (c) and (d) correspond to u-wind, height, temperature (T) and specific humidity (q) fields, respectively. The observational error standard deviations are shown as thin solid lines wherever applicable.

where M_b denotes the time evolving model of the bias. Now, we have two separate variables: the model state variable $\mathbf{x}$ and the bias variable $\mathbf{b}$.

For the unbiased variable $\tilde{\mathbf{x}}$, we can apply the same filtering equations as in Section 2.2, thus,

$$\tilde{\mathbf{x}}_i^a = \tilde{\mathbf{x}}_i^f + \mathbf{K}(\mathbf{y}^o - H\tilde{\mathbf{x}}_i^f) \tag{4.7}$$

where the optimal weight $\mathbf{K}$ is given as

$$\mathbf{K} = \mathbf{P}^f\mathbf{H}^\top(\mathbf{H}\mathbf{P}^f\mathbf{H}^\top + \mathbf{R})^{-1} \tag{4.8}$$

cf. eq.(2.15). Here, the error covariance matrix $\mathbf{P}^f$ is defined for the unbiased variable as

$$\mathbf{P}^f = \left\langle (\tilde{\mathbf{x}}^f - \mathbf{x}^t)(\tilde{\mathbf{x}}^f - \mathbf{x}^t)^\top \right\rangle \tag{4.9}$$

The covariance analysis is given by

$$\mathbf{P}^a = (\mathbf{I} - \mathbf{K}\mathbf{H})\mathbf{P}^f \tag{4.10}$$

cf. eq.(2.17). With the ensemble formulation, the covariance forecast is derived from the ensemble forecast, and the ensemble update is given by the SRF (square root filtering) to satisfy eq.(4.10).

Similarly for the bias, we define a weighted mean of the bias forecast and the bias observation:

$$\mathbf{b}_i^a = \mathbf{b}_i^f - \mathbf{K}_b(\mathbf{y}^o - H\mathbf{x}_i^f + H\mathbf{b}_i^f) \tag{4.11}$$

Similar to eqs.(2.2) and (2.3), we define the differences from the true bias $\mathbf{b}^t$:

$$\delta\mathbf{b}^a = \mathbf{b}^a - \mathbf{b}^t \tag{4.12}$$

$$\delta\mathbf{b}^f = \mathbf{b}^f - \mathbf{b}^t \tag{4.13}$$

Subtracting $\mathbf{b}^t$ from both sides of eq.(4.11), we get

$$\begin{aligned}
\delta\mathbf{b}^a &= \delta\mathbf{b}^f - \mathbf{K}_b\left(\delta\mathbf{y}^o - \mathbf{H}\delta\mathbf{x}^f + \mathbf{H}(\delta\mathbf{b}^f + \mathbf{b}^t)\right) \\
&= \delta\mathbf{b}^f - \mathbf{K}_b\left(\delta\mathbf{y}^o - \mathbf{H}(\delta\mathbf{x}^f - \mathbf{b}^t) + \mathbf{H}\delta\mathbf{b}^f\right) \\
&= \delta\mathbf{b}^f - \mathbf{K}_b\left(\delta\mathbf{y}^o - \mathbf{H}\delta\tilde{\mathbf{x}}^f + \mathbf{H}\delta\mathbf{b}^f\right) \qquad (4.14)
\end{aligned}$$

where $\delta\mathbf{x}^f$ is defined as the same as eq.(2.3) in Section 2.2, and we assumed unbiased forecast error:

$$\begin{aligned}
\delta\tilde{\mathbf{x}}^f &= \mathbf{x}^f - \mathbf{x}^t - \mathbf{b}^t \qquad (4.15) \\
&= \tilde{\mathbf{x}}^f - \mathbf{x}^t \qquad (4.16) \\
&= \delta\mathbf{x}^f - \mathbf{b}^t \qquad (4.17)
\end{aligned}$$

Bias analysis error covariance $\mathbf{P}_b^a$ is written

$$\begin{aligned}
\mathbf{P}_b^a &= \left\langle \delta\mathbf{b}^a(\delta\mathbf{b}^a)^\top \right\rangle \\
&= \left\langle \left(\delta\mathbf{b}^f + \mathbf{K}_b(\delta\mathbf{y}^o - \mathbf{H}\delta\tilde{\mathbf{x}}^f + \mathbf{H}\delta\mathbf{b}^f)\right)\left(\delta\mathbf{b}^f + \mathbf{K}_b(\delta\mathbf{y}^o - \mathbf{H}\delta\tilde{\mathbf{x}}^f + \mathbf{H}\delta\mathbf{b}^f)\right)^\top \right\rangle \\
&= \left\langle \left((\mathbf{I} - \mathbf{K}_b\mathbf{H})\delta\mathbf{b}^f - \mathbf{K}_b\delta\mathbf{y}^o + \mathbf{K}_b\mathbf{H}\delta\tilde{\mathbf{x}}^f\right)\left((\mathbf{I} - \mathbf{K}_b\mathbf{H})\delta\mathbf{b}^f - \mathbf{K}_b\delta\mathbf{y}^o + \mathbf{K}_b\mathbf{H}\delta\tilde{\mathbf{x}}^f\right)^\top \right\rangle \\
&= (\mathbf{I} - \mathbf{K}_b\mathbf{H})\mathbf{P}_b^f(\mathbf{I} - \mathbf{K}_b\mathbf{H})^\top + \mathbf{K}_b\mathbf{R}\mathbf{K}_b^\top + \mathbf{K}_b\mathbf{H}\mathbf{P}^f\mathbf{H}^\top\mathbf{K}_b^\top \qquad (4.18)
\end{aligned}$$

where bias forecast error covariance is defined as

$$\mathbf{P}_b^f = \left\langle \delta\mathbf{b}^f(\delta\mathbf{b}^f)^\top \right\rangle \qquad (4.19)$$

and all the cross covariances are assumed to be zero:

$$\begin{aligned}
\left\langle \delta\tilde{\mathbf{x}}^f(\delta\mathbf{y}^o)^\top \right\rangle &= 0 \qquad (4.20) \\
\left\langle \delta\mathbf{b}^f(\delta\mathbf{y}^o)^\top \right\rangle &= 0 \qquad (4.21) \\
\left\langle \delta\tilde{\mathbf{x}}^f(\delta\mathbf{b}^f)^\top \right\rangle &= 0 \qquad (4.22)
\end{aligned}$$

In order for the weighted mean eq.(4.11) to be optimal, we minimize the total bias analysis error variance, i.e. the trace of $\mathbf{P}_b^a$. Thus,

$$\frac{\partial}{\partial \mathbf{K}_b}\left(trace(\mathbf{P}_b^a)\right) = 0 \tag{4.23}$$

is satisfied (cf. eq.(2.11)). Using the formula (2.12) and (2.13), we get the optimal weight

$$\mathbf{K}_b = \mathbf{P}_b^f \mathbf{H}^\top (\mathbf{H}\mathbf{P}_b^f \mathbf{H}^\top + \mathbf{H}\mathbf{P}^f \mathbf{H}^\top + \mathbf{R})^{-1} \tag{4.24}$$

cf. eqs.(40) and (53) in Dee and da Silva (1998). Substituting the optimal weight (4.24) into eq.(4.18), we can get

$$\mathbf{P}_b^a = (\mathbf{I} - \mathbf{K}_b\mathbf{H})\mathbf{P}_b^f \tag{4.25}$$

cf. eq.(54) in Dee and da Silva (1998).

The dimension N of the bias variable $\mathbf{b}$ is the same as that of the model variable $\mathbf{x}$, and the bias error covariance $\mathbf{P}_b$ is an $N \times N$ matrix. To avoid the explicit computation of the large matrix, we need a low dimensional assumption to apply the ensemble formulation. We expand all the covariances with a limited number of ensemble members in a similar way as in Section 2.2. Thus, we use a bias ensemble to represent $\mathbf{P}_b$.

Since we do not have good information about the bias forecast model M_b, persistence is assumed as the bias forecast model:

$$M_b = \mathbf{I} \tag{4.26}$$

which assumes the model bias should be intrinsically constant in time. Then, eq.(4.6) is just

$$\mathbf{b}_i^f = \mathbf{b}_{i-1}^a \tag{4.27}$$

In this case, the bias has no dynamics, and the low dimensional assumption may not be reasonable.

Alternatively, we can choose a limited number, say k, of base fields to span the bias $\mathbf{b}$ (Dee and da Silva 1998; Danforth, Kalnay, pers. comm.):

$$\mathbf{b} = \mathbf{T}\beta \tag{4.28}$$

Here, $\mathbf{T}$ is an $N \times k$ matrix where each column denotes base fields, and β denotes a k-dimensional vector. In this way, β is the variable to be estimated, $\mathbf{T}$ is fixed. For the best estimate of β, we substitute eq.(4.28) into the previous equations derived for $\mathbf{b}$. Here, the error covariance of $\mathbf{b}$ is substituted as

$$\begin{aligned}
\mathbf{P}_b &= \left\langle \delta\mathbf{b}(\delta\mathbf{b})^\top \right\rangle \\
&= \left\langle \mathbf{T}\delta\beta(\mathbf{T}\delta\beta)^\top \right\rangle \\
&= \mathbf{T}\left\langle \delta\beta(\delta\beta)^\top \right\rangle \mathbf{T}^\top \\
&= \mathbf{T}\mathbf{P}_\beta\mathbf{T}^\top
\end{aligned} \tag{4.29}$$

Thus, substituting eqs.(4.28) and (4.29), we construct the Kalman filtering equations for β as follows:

$$\beta_i^f = \beta_{i-1}^a \tag{4.30}$$

$$\mathbf{P}_\beta^f = (1+\delta)\mathbf{P}_\beta^a \tag{4.31}$$

$$\beta_i^a = \beta_i^f - \mathbf{K}_\beta(\mathbf{y}^o - H\mathbf{x}_i^f + H\mathbf{T}\beta_i^f) \tag{4.32}$$

$$\mathbf{K}_\beta = \mathbf{P}_\beta^f\mathbf{T}^\top\mathbf{H}^\top(\mathbf{H}\mathbf{T}\mathbf{P}_\beta^f\mathbf{T}^\top\mathbf{H}^\top + \mathbf{H}\mathbf{P}^f\mathbf{H}^\top + \mathbf{R})^{-1} \tag{4.33}$$

$$\mathbf{P}_\beta^a = (\mathbf{I} - \mathbf{K}_\beta\mathbf{H}\mathbf{T})\mathbf{P}_\beta^f \tag{4.34}$$

where δ denotes a covariance inflation parameter for $\mathbf{P}_\beta$. If the dimension of β is small enough, we can compute the above equation explicitly.

In summary, we introduced two bias estimation schemes:

1. EnKF for the full dimensional bias variable (Dee and da Silva 1998)

2. KF for the low dimensional bias variable (suggested by Kalnay and by the results of Danforth et al. 2005)

4.4.2 Numerical experiments

First, we applied the full dimensional bias variable estimate using the ensemble formulation of Kalman filtering. We use the serial EnSRF system with minimal modifications. The difference is in the gain computation. Since we need to consider not only $\mathbf{P}_\beta$ but also $\mathbf{P}^f$ in the gain computation, we cannot apply exactly the same system for the bias variable. However, because we have the forecast ensemble as well as the bias ensemble in memory, the modification is not complicated.

The initial bias ensemble is generated using Gaussian random numbers. A random perturbation field is constructed by putting a Gaussian random number at each grid point and smoothing high frequency components using a Lanczos filter. The amplitudes are normalized using observational error size, so that each random variable has an appropriate order of magnitude. We used 10 ensemble members, where 5 random fields are generated and their plus-minus pairs constitute 10 members. Thus, initially the bias is assumed to be zero, i.e. ensemble mean is zero.

When we applied Dee and da Silva's full dimensional bias estimation, the filter was unstable and diverged rapidly. In order for the filter to be more stable in the initial transient period, the bias correction eq.(4.5) is modified as

$$\tilde{\mathbf{x}}^f_{i+1} = M(\tilde{\mathbf{x}}^a_i) - \alpha \mathbf{b}^a_i \tag{4.35}$$

where α is a scalar parameter that weights the bias correction. Even with $\alpha = 0.01$, one percent bias correction, we still observed unstable filtering. Fig.4.9 shows the analysis

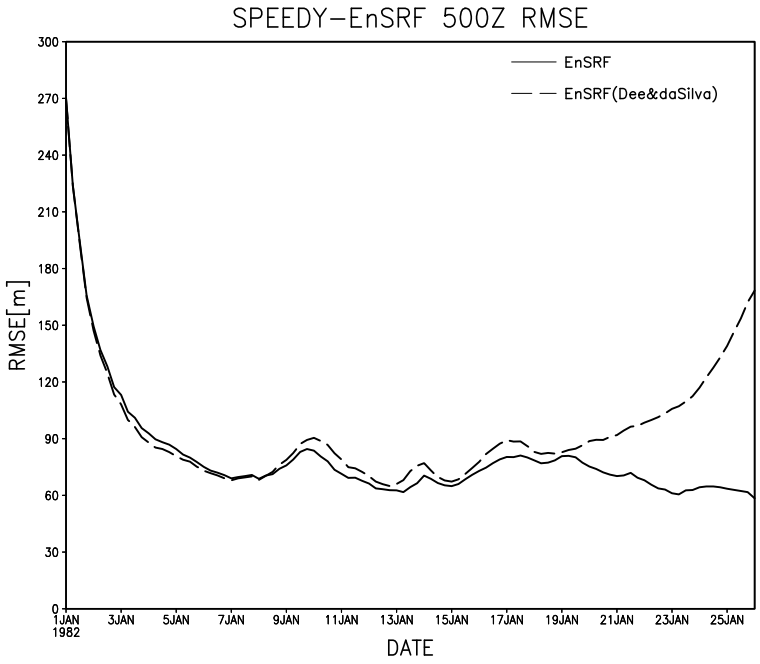

Figure 4.9: Analysis RMSE of 500hPa height field of EnKF with 10 ensemble members in the case of the sparse observational networks in the presence of model errors. The dashed and solid lines show the cases with and without Dee and da Silva's full dimensional bias estimation, respectively.

RMSE of 500hPa height field with/without bias estimation. Here, we used $\alpha = 0.01$. The time period finishes before February, because the filter diverges just after the final time of the figure even with $\alpha = 0.01$. In the first week, the bias estimate makes analyses better, but eventually it gets worse.

Possible reasons are as follows:

- Too small ensemble size
- Too few observational stations

As mentioned before, the bias forecast is assumed to have no dynamics. Thus, the bias field cannot have dynamically growing modes. Therefore, the low-dimensional assumption for

ensemble formulation may not be a good assumption. Even with localization, 10 ensemble members may be too few to capture all the possible directions. As for the number of observations, larger number of observations helps the filter converge faster. With fewer observations and without dynamics, the initial spin-up period of the filtering becomes longer.

Since the full dimensional estimation made the filter unstable, we applied low-dimensional bias variable estimation. We used the mean bias fields shown in Figs.4.1 and 4.2 to span the model bias. Although variables are combined, each level is separated. For comparison, we show the case with constant bias subtraction, i.e. the bias variable β is fixed to 1. In addition, we also show the cases of 3DVAR with and without constant bias subtraction. Fig.4.10 shows analysis RMSE of 500hPa height field. Thin red line shows 6-hour forecast errors initiated from the reanalysis fields. These forecasts provide background errors purely caused by model errors. Clearly, EnKF shows smaller analysis RMSE with bias estimation. There is no significant difference between low order bias estimation (green short-dashed line) and constant bias subtraction (blue dotted line). 3DVAR (blue lines) shows slightly better performance with constant bias subtraction, but the improvement is smaller than with EnKF. Fig.4.11 shows the values of all the seven components of the bias variable. The seven components correspond to vertical levels. After the initial spin-up, all values are close to one with small temporal variations which explains why the constant bias subtraction shows similar performance.

Fig.4.12 shows the spatial distribution of analysis RMSE of 500hPa height fields with and without bias estimation. Clearly, the amplitude is reduced more in the case with bias estimation (bottom panel) than the case without bias estimation (top panel). The large difference can be seen for example over the Andes of South America, Himalayas,

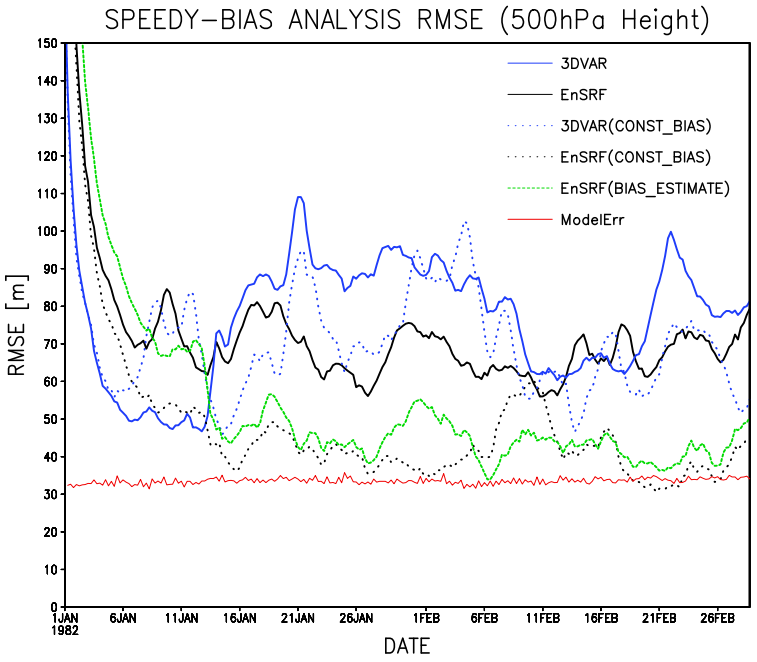

Figure 4.10: Analysis RMSE of 500hPa height field for EnKF with low order bias estimation (green short-dashed line), EnKF with constant bias subtraction (black dotted line), 3DVAR with constant bias subtraction (blue dotted line). Solid lines show the cases without bias estimation in the cases of EnKF (black line) and 3DVAR (blue line), the same as solid and long-dashed lines of Fig.4.6. Thin red line shows 6-hour forecast errors initiated from the reanalysis fields.

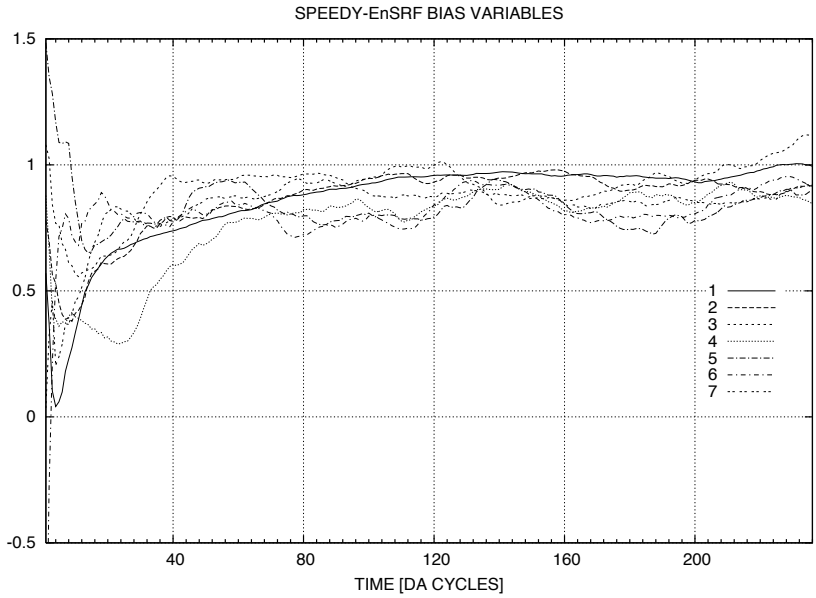

Figure 4.11: Values of all the seven components of the bias variable when the mean bias field is used as a basis to span the bias where each vertical level is assumed to be independent. The seven lines correspond to vertical levels.

and southern Persian Gulf, where the top panel shows areas with large errors while the bottom panel does not. In such areas, large model error biases are evident in Fig.4.2.

To see what caused the large error spots, the analysis error bias fields are shown in Fig.4.13. The case without bias estimation (top panel) shows much larger amplitudes than the case with bias estimation (bottom panel), large error spots exist in similar locations. If we subtract the error bias from the error fields and take the RMS, i.e. analysis error standard deviation, the difference between the cases with and without bias estimation becomes small. Fig.4.14 shows that the difference between the top and bottom panels are smaller than the previous figures. Here, the top and bottom panels show the cases without and with bias estimation, respectively. Importantly, the large error spots disappear in the top panel. Thus, we deduce that the large error spots are caused by the persistent model bias, which is greatly reduced by the bias estimation.

Fig.4.15 shows analysis RMSE of other variables at all vertical levels. Within EnKF, the bias estimation generates better analysis states for all variables at all levels. The difference is 30-40 meters in height fields, 2-3 m/s in wind fields, 1-3 K in temperature fields, all of which are fairly large improvements. There is almost no difference between the cases with low order bias estimation (green short-dashed line) and with constant bias subtraction (black dotted line), only specific humidity RMSE shows significant difference. 3DVAR and EnKF show similar performance without bias subtraction (solid lines). However, with bias subtraction, EnKF shows an advantage especially in the wind and height RMSEs. In the temperature RMSE, EnKF shows an advantage of less than 0.5 K. In the specific humidity, the situation is opposite, 3DVAR shows smaller RMSE than EnKF. With low order bias estimation, EnKF shows almost identical performance as 3DVAR with constant bias subtraction. Concentrating on dynamics, we obtain much better per-

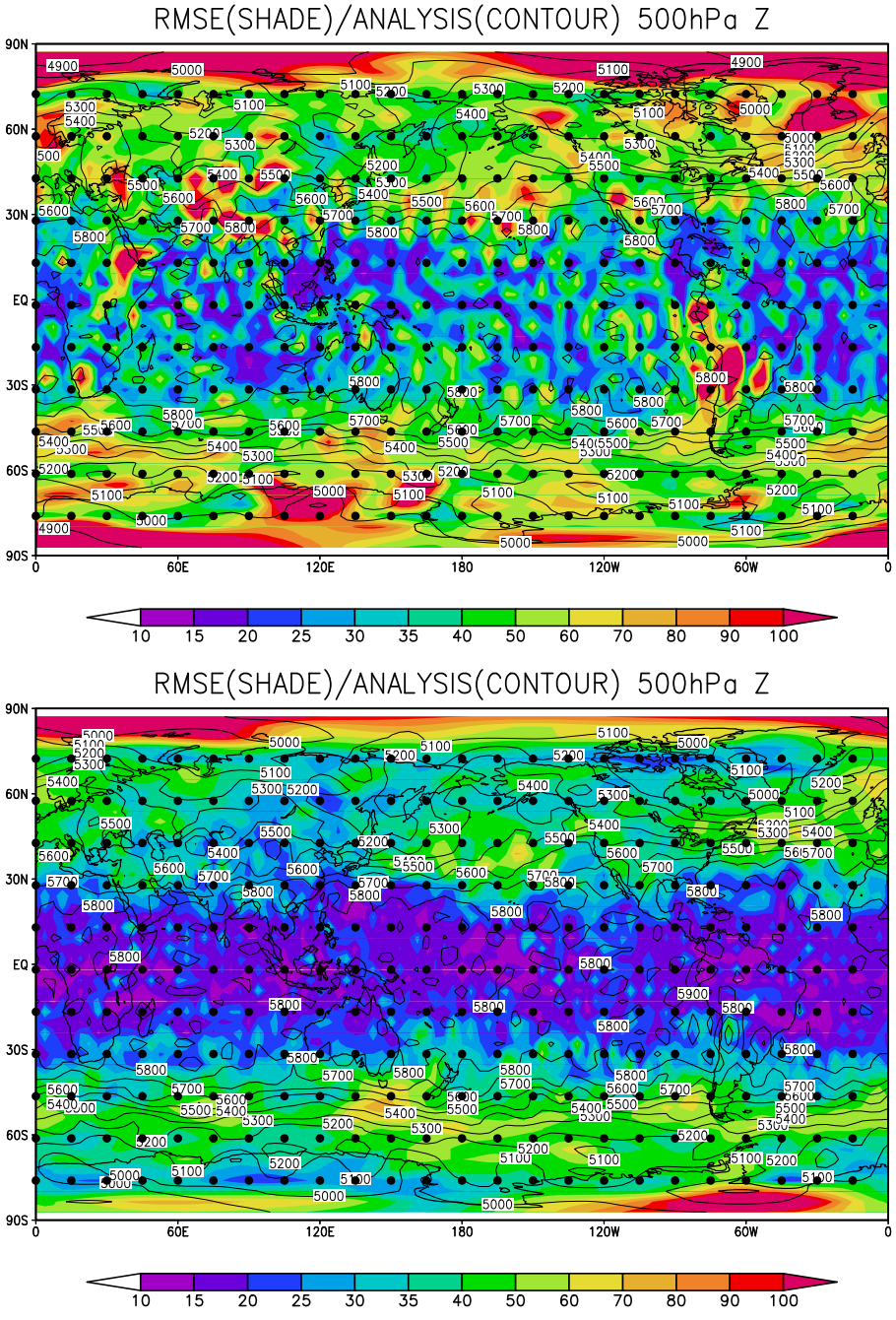

Figure 4.12: Spatial distribution of analysis RMSE of 500hPa height fields (shades in meters) in the case of imperfect model experiments using EnKF with 10 ensemble members. Contours show mean analysis fields (in meters). Top and bottom panels show the cases without and with bias estimation, respectively. Dots indicate the observational locations.

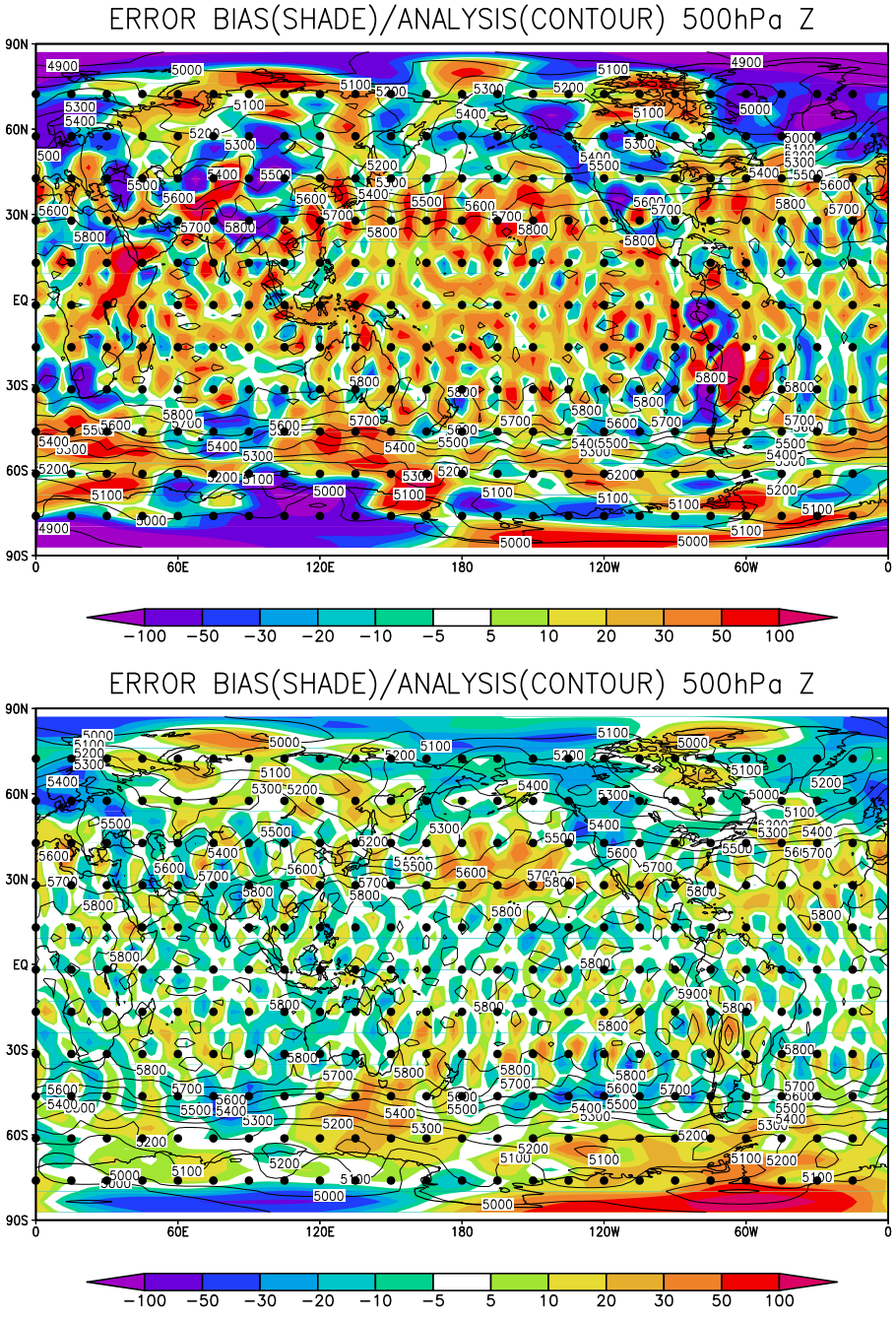

Figure 4.13: Analysis error bias fields of 500hPa height (shades in meters) in the case of imperfect model experiments using EnKF with 10 ensemble members. Contours show mean analysis fields (in meters). Top and bottom panels show the cases without and with bias estimation, respectively. Dots indicate the observational locations.

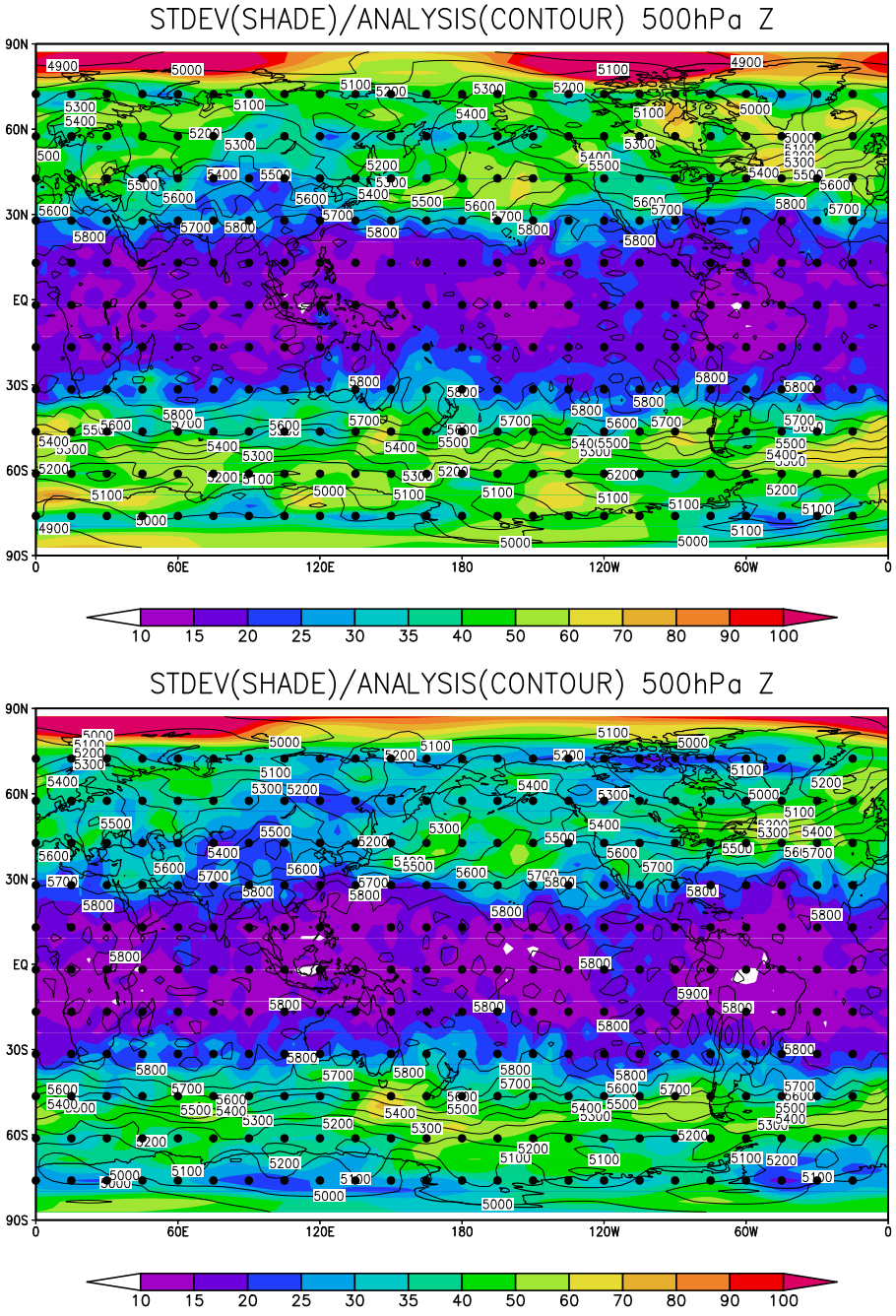

Figure 4.14: Analysis error standard deviation fields of 500hPa height (in meters) in the case of imperfect model experiments using EnKF with 10 ensemble members. Contours show mean analysis fields (in meters). Top and bottom panels show the cases without and with bias estimation, respectively. Dots indicate the observational locations.

formance using EnKF with bias estimation and constant bias subtraction in wind and height fields. However, for variables strongly related to physical processes, the advantage of EnKF to 3DVAR is less significant. There is significant improvement of humidity fields with low order bias estimation compared to constant bias subtraction, similar to 3DVAR with constant bias subtraction.

Fig.4.16 shows the zonal structure of analysis RMSE of u-wind and temperature fields in the case without bias estimation. Fig.4.17 shows the same figure but in the case with bias estimation. Again, the amplitude is larger in the case without bias estimation. Both cases have similar structures in the zonal mean errors. u-wind error fields show a clear striped pattern according to the observational locations. We see a large reduction of temperature errors in the lower altitudes in the northern hemisphere in the case with bias estimation.

We include the first few EOFs as bases to span the bias variables. Table 4.1 shows analysis RMSE of 500 hPa height field with various numbers of EOFs. The mean bias field and first two EOFs give the best performance, in agreement with the argument in Section 4.2.2 the higher order EOFs do not contain statistically significant information due to the shortness of the training period. Since the diurnal cycle components and the bias are well defined, even with a short training period, their online correction leads to significant improvements. The addition of EOFs beyond the diurnal cycle (not shown) makes the results slightly worse.

4.5 Summary and discussion

The bias estimation method proposed by Dee and da Silva (1998) did not succeed for full dimensionality. Thus, we applied a low-dimensional expansion of bias fields, as suggested

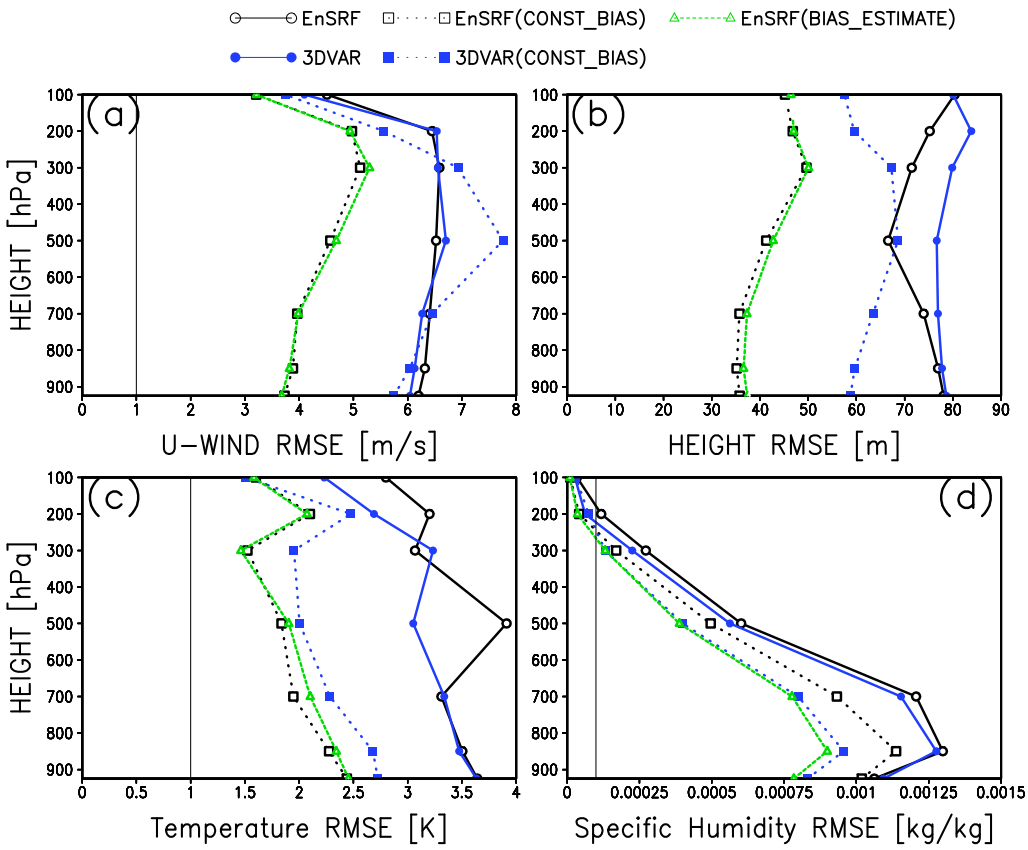

Figure 4.15: Analysis RMSE at the all pressure levels temporally averaged for one month after the initial one-month spin-up period (112 samples) in the cases of EnKF with low order bias estimation (green short-dashed line), EnKF with constant bias subtraction (black dotted line), 3DVAR with constant bias subtraction (blue dotted line). Solid lines show the cases without bias estimation in the cases of EnKF (black line) and 3DVAR (blue line), the same as the black solid line and the black short-dashed line of Fig.4.8. The four panels (a), (b), (c) and (d) correspond to u-wind, height, temperature (T) and specific humidity (q) fields, respectively. The observational error standard deviations are shown as thin solid lines wherever applicable.

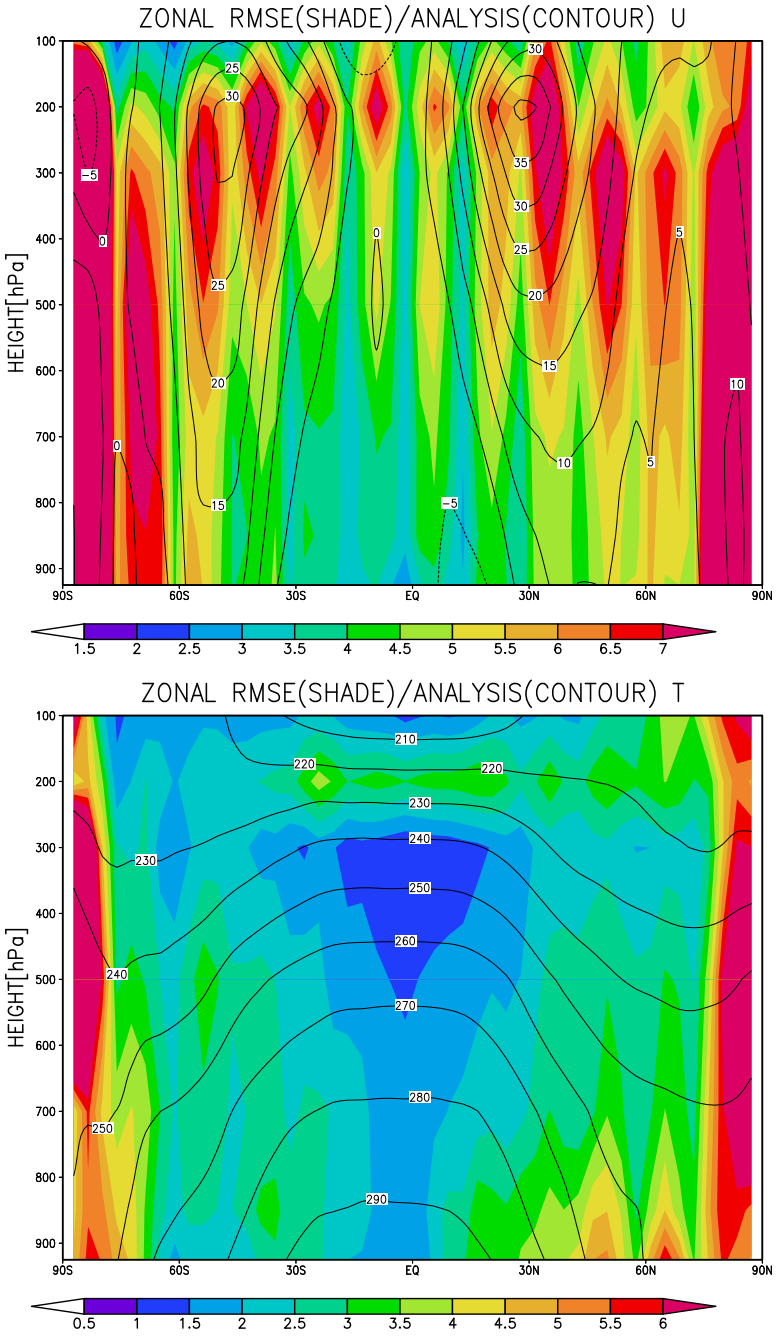

Figure 4.16: Zonal structure of analysis RMSE of u-wind field (top panel in m/s) and temperature field (bottom panel in K) in the case of imperfect model experiments using EnKF with 10 ensemble members where no bias estimation is applied. Shades and contours show analysis RMSE and mean analysis fields, respectively.

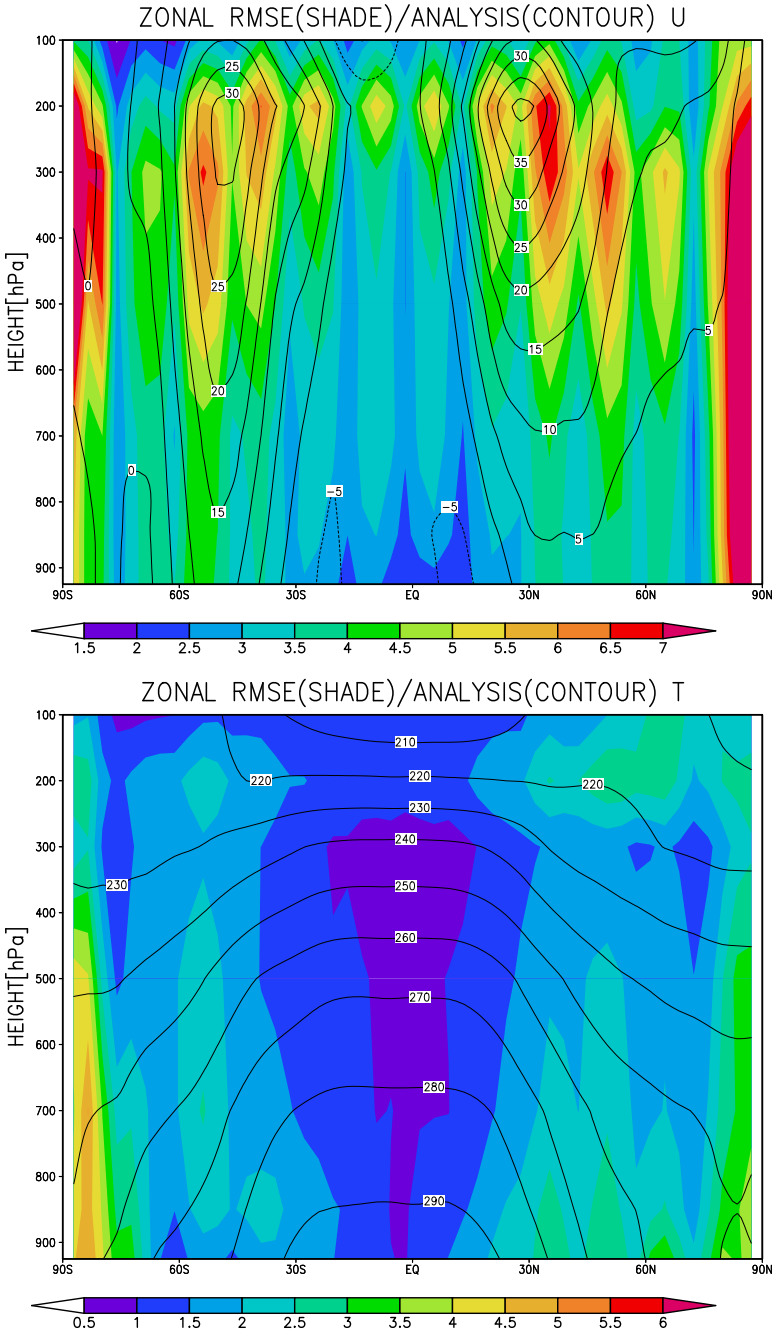

Figure 4.17: The same figure as Fig.4.16 but with bias estimation.

Bias correction	Uncorrected	Mean bias	EOF1	EOF2
500Z RMSE [m]	66.6	42.8	40.6	39.2

Table 4.1: Analysis RMSE of 500hPa height field using EnKF with 10 ensemble members and low order bias estimation using various numbers of bias basic fields. "Uncorrected" denotes the case without bias correction. "Mean bias", "EOF1", and "EOF2" denotes the cases with low-order bias correction using the mean bias field, using the mean bias field and first EOF, and using the mean bias field and first and second EOFs, respectively.

by Kalnay and by the results of Danforth et al. (2005). The theory of low-dimensional expansion was also described in Dee and da Silva (1998).

We computed statistics using model errors sampled by the difference between the reanalysis fields and 6-hour forecasts started from the reanalysis fields. We computed the mean bias and EOFs from the model error samples, and used them as bases of model bias. Then, we found the low order bias estimation has a positive impact even with a single statistical bias field. Constant bias subtraction has also shown similar performance, although it resulted in larger RMSE in specific humidity fields. Although 3DVAR and EnKF have similar performance without bias estimation or constant bias subtraction, constant bias subtraction showed larger positive impact within EnKF than within 3DVAR, except in specific humidity. When we applied low order bias estimation within EnKF, EnKF outperformed 3DVAR everywhere and for all fields. It is important to note that we used the dependent sample for the statistics, suggesting we may have overestimated the performance. It is expected that if we sample with other terms or years to obtain the basic bias fields, our results would be worse than what we show in the present research. However, other experiments have shown similar basic bias fields even using sampling of

other years (Danforth et al. 2005).

Another problem is that the SPEEDY model has much larger errors than sophisticated operational models. We used the NNR fields as a best estimate of the true atmosphere and took the difference from the SPEEDY model forecast to obtain samples of model errors. In practice, we use a state-of-the-art model both for estimating the true nature (i.e. reanalysis) and for forecasts. If we use the same model for both reanalyses and forecasts, the differences between the forecasts and analyses by the same model, i.e. the analysis increments, will be samples of model errors. The model bias field obtained from the analysis increments was very different from the one obtained from the NNR fields. Using the analysis increments as model error samples, we obtained worse results (not shown). Thus, obtaining good samples of model errors may be a big challenge in practice, requiring longer training periods, and/or a multi-model approach; we may obtain a best estimate of the true nature by an optimal combination with separate analysis fields such as NNR and ECMWF (European Centre on Medium-range Weather Forecasting) reanalysis fields.

Chapter 5

Summary and future directions

5.1 Summary

Our goal was to develop a path towards the operational application of ensemble Kalman filtering (EnKF), which is expected to produce substantially improved weather/climate information. Several approaches to EnKF for atmospheric systems have been suggested, but were not previously directly compared. In addition, there has been little research published on the sensitivity of EnKF to the imperfections of forecast models. The present research aimed to explore two basic questions:

1. What are the relative advantages and disadvantages of the two most promising EnKF methods?

2. How large are the effects of model errors on data assimilation, and can they be reduced by model bias correction?

In Chapter 2 we described a theoretical review of EnKF, followed by the FORTRAN development and testing on the Lorenz-96 model (Lorenz 1996; Lorenz and Emanuel 1998) of two EnKF methods: 1) a serial ensemble square root filter (serial EnSRF, Whitaker and Hamill 2002) that assimilates observations sequentially and localizes the background error covariance multiplying it with a Gaussian-like function; and 2) a local EnKF (LEKF, Ott et al. 2002; 2004) that assimilates all the observations within a local volume surrounding a grid point simultaneously. We reproduced the results obtained by Whitaker and Hamill (2002) and Ott et al (2004). With observational error covariance localization (multiplying

the observational error covariance by the inverse of the Gaussian function similar to the one used in serial EnSRF suggested by Hunt, pers. comm.), the optimal size of the LEKF patches becomes larger, allowing the use of more observations. The overall LEKF performance is then equal or better than that of the serial EnSRF with similar parameters. We also introduced a new method to objectively estimate the optimal covariance inflation that could be performed during the data assimilation, allowing for inflation dependent on location and time.

In Chapter 3 the two EnKF methods as well as the three-dimensional variational method (3DVAR) were applied to the SPEEDY primitive-equation global model (Molteni 2003). The SPEEDY model is a fast but relatively realistic model allowing a comparison of methods that address the first question. Our results show that in a perfect model scenario the EnKF greatly outperforms 3DVAR. Surprisingly, for each of the methods the 2-day forecast "errors of the day" are very similar to the analysis errors, but they are not similar among different methods except in regions identified as having low ensemble dimension (E-dim), where they are also similar to bred vectors. Our results suggest that for the SPEEDY model serial EnSRF outperforms LEKF, but that their difference is substantially reduced if we either localize the LEKF error covariance or increase the ensemble size. Using stochastic perturbations further improved the performance of the LEKF. Since LEKF is much more efficient than serial EnSRF when using parallel computers and many observations, LEKF (or the more efficient algorithm LETKF, Hunt et al, 2005) would be the only feasible choices in an operational setup.

In Chapter 4 we remove the perfect model assumption and investigate the second question, using the NCEP/NCAR reanalysis (Kalnay et al. 1996) as the "nature" run. The advantage of EnKF with respect to 3DVAR is greatly reduced. When we apply

the model bias estimation proposed by Dee and da Silva (1998), we find that the full dimensional model bias estimation fails. However, if instead we assume that the bias is low dimensional, we obtain a substantial improvement in the EnKF analysis.

5.2 Future directions

Since the present research aims to develop a path towards an operational application of EnKF, we would like to conclude this thesis by describing possible future directions.

Possible further investigations relevant in operations are as follows:

- The relative advantages and disadvantages between EnKF and the four-dimensional variational method (4DVAR)
- Application of four-dimensional EnKF (4D-EnKF, Hunt et al. 2004) to treat asynoptic observations
- The effect of using a nonlinear observational operator H
- The effect of non-diagonal components of observational error covariance $\mathbf{R}$
- Application to real observations

4DVAR is the mainstream in current operational data assimilation; following the ECMWF, several major operational weather centers including the Meteo France, the Japan Meteorological Agency (JMA), the UK MetOffice (UKMO), and the Canadian Meteorological Centre (CMC) applied 4DVAR operationally. However, there is no research thus far comparing EnKF with 4DVAR. There are many asynoptic observations in practice, 4DVAR can treat them intrinsically. Alternatively, Hunt et al. (2004) proposed 4D-EnKF, assuming ensemble forecasts can propagate observational increments in time. When the

observational operator H is nonlinear, EnKF has an advantage over variational methods in that it does not require a linearized operator or its adjoint. In addition, LEKF can consider non-diagonal components of the observational error covariance $\mathbf{R}$ with less additional computation than other methods. Since more satellite observations with nonlinear observational operator and correlated errors were available recently, it is important to investigate these effects.

As for model errors, research suggests that cross-correlation between forecast error and bias error may play an important role (H. Li, pers. comm.). Since our investigation on the model bias estimation has limitations, more investigation is necessary. If a model bias correction scheme using both forecast and bias ensembles shows advantages, EnKF would be much advantageous over 4DVAR. In addition, there is another issue related to model errors that we did not discuss so far: model error covariance. Ensemble members are expected to represent error structures, but model error covariance could affect them even in the absence of model biases. This may be important with real observations. Furthermore, it is important to improve not only the analysis but also the forecast by considering model errors. Danforth et al. (2005) have shown improved forecasts just by keeping a constant bias as a nudging term in the forecast model using a quasi-geostrophic global model (Marshall and Molteni 1993) and the SPEEDY model. If we estimate model biases adaptively using a bias estimation scheme with EnKF, we may apply the same approach as Danforth to improve forecast skill. The better forecasts may feedback in error representation of an ensemble, improving analysis as well. However, we need to consider the effect of the improved forecast in the bias estimation, since the forecast model is different every time.

There are further applications of ensembles representing error structures:

- Application in targeted or adaptive observations
- Advantage in ensemble forecasting
- Application in spatially weighted inflation parameter

If we select observing locations adaptively (targeted or adaptive observations, e.g. Lorenz and Emanuel 1998), we can use the information of errors represented by an ensemble to select the best locations to observe. It is important to observe in large error areas, an ensemble can suggest the appropriate locations.

The ensemble generated by EnKF should be an optimal choice for ensemble forecasting. Ensemble members generated by EnKF are not positive-negative pairs, Wang et al. (2004) have shown a centered spherical simplex ensemble provides better ensemble forecasts than an ensemble of positive-negative pairs with a limited ensemble size. The ensemble generated by EnKF represents analysis errors, Wang and Bishop (2003) showed bred vectors weighted by the estimated analysis errors using observational locations (masked breeding scheme, Wang and Bishop 2003) provided better ensemble forecasting.

We could also use the information of E-dimension to weight the covariance inflation parameter in space, as suggested by Kalnay (pers. comm.). Low E-dimension areas show dynamically important areas with good agreement of error fields generated separately. We can make the covariance inflation smaller in those areas, since we can trust the ensemble more in these low E-dimensional areas.

A big challenge in operations is in applying EnKF in regional/mesoscale models. Operationally, regional/mesoscale models provide important weather products, more directly connected to peoples lives. It is important that we apply better methods in regional/mesoscale prediction systems. There are not many results on ensemble forecasting

or EnKF on mesoscale models, and few ensemble prediction systems are operational. Since the regional models require lateral boundary conditions, their treatment introduces a new problem. A global ensemble generated by EnKF may provide a good set of lateral boundary conditions for regional ensemble members. In addition, the scales of weather phenomena expressed by regional/mesoscale models are different from global models. Asynoptic observations capturing such mesoscale phenomena are more important and should be treated appropriately. Meso-EnKF could be an important possible direction in operations.

Appendix A

The model-independent core modules of serial EnSRF and LEKF

A.1 Introduction

Since the methods of serial EnSRF and LEKF are general, they can be applied to any dynamical system with observations, including the simple Lorenz-96 model and the realistic SPEEDY model. Thus, the core parts are developed in a model-independent way so that they can be tested using a simple model and applied to a realistic model. This appendix describes details in the model-independent modules that are written in Fortran90.

A.2 Serial EnSRF

Three subroutines are developed for serial EnSRF:

1. `enkf_anal0(nx,m,hdxf,r,dxf,gain)`

2. `enkf_serial(nx,m,hdxf,r,dxf,gain,dxa)`

3. `enkf_schur(scale,dist,factor)`

The first subroutine `enkf_anal0` computes the Kalman gain using eq.(2.38), i.e.,

$$\mathbf{K} = \mathbf{E}^f(\mathbf{H}\mathbf{E}^f)^T[\mathbf{H}\mathbf{E}^f(\mathbf{H}\mathbf{E}^f)^T + \mathbf{R}]^{-1} \tag{A.1}$$

As inputs, $\mathbf{E}^f$, $\mathbf{H}\mathbf{E}^f$ and $\mathbf{R}$ are given, and $\mathbf{K}$ is returned as an output. Table A.1 gives all the arguments of the subroutine `enkf_anal0`.

Argument	IN/OUT	Entity
`nx`	IN	dimension of the system
`m`	IN	ensemble size
`hdfx(m)`	IN	$\mathbf{HE}^f$
`r`	IN	observational error variance
`dxf(nx,m)`	IN	forecast ensemble perturbations $\mathbf{E}^f$
`gain(nx)`	OUT	Kalman gain $\mathbf{K}$

Table A.1: Arguments of the subroutine enkf_anal0 that computes Kalman gain.

The second subroutine `enkf_serial` computes analysis ensemble perturbations using eqs.(2.45) and (2.48), that is,

$$\mathbf{E}^a = (\mathbf{I} - \alpha\mathbf{KH})\mathbf{E}^f \tag{A.2}$$

$$= \mathbf{E}^f - \alpha\mathbf{KHE}^f \tag{A.3}$$

$\mathbf{E}^f$, $\mathbf{HE}^f$, $\mathbf{R}$ and $\mathbf{K}$ are given as inputs, and $\mathbf{E}^a$ is returned as an output. Table A.2 gives all the arguments of the subroutine `enkf_serial`.

The third subroutine `enkf_schur` computes the factor $S(r)$ of the Schur product using eq.(2.49). Table A.3 gives all the arguments of the subroutine `enkf_schur`.

The Schur product is applied just after computing $\mathbf{K}$ by `enkf_anal0`, and the input of $\mathbf{K}$ for the subroutine `enkf_serial` is localized. Analysis equation eq.(2.1), that is,

$$\bar{\mathbf{x}}^a = \bar{\mathbf{x}}^f + \mathbf{Kd} \tag{A.4}$$

is solved w.r.t. the ensemble mean state, where the localized Kalman gain $\mathbf{K}$ is applied to avoid sampling errors in distant points. Since there is no matrix inversion or square root computation in the above equations, no subroutine for eigenvalue decomposition or other

Argument	IN/OUT	Entity
`nx`	IN	dimension of the system
`m`	IN	ensemble size
`hdfx(m)`	IN	$\mathbf{HE}^f$
`r`	IN	observational error variance
`dxf(nx,m)`	IN	ensemble perturbations $\mathbf{E}^f$
`gain(nx)`	OUT	Kalman gain $\mathbf{K}$
`dxa(nx,m)`	OUT	analysis ensemble perturbations $\mathbf{E}^a$

Table A.2: Arguments of the subroutine enkf_serial that computes analysis ensemble perturbations.

Argument	IN/OUT	Entity
`scale`	IN	length scale l_c
`dist`	IN	distance d
`factor`	OUT	Schur product factor $S(r)$

Table A.3: Arguments of the subroutine enkf_schur that computes the factor of the Schur product for localization.

Argument	IN/OUT	Entity
`h(nobsl,ndiml)`	IN	observational operator in a local patch
`rdiag(nobsl)`	IN	observational error variance in a local patch
`dep(nobsl)`	IN	observational increment $\mathbf{d}$ in a local patch
`xb(ndiml,m)`	IN	forecast ensemble $\mathbf{E}^f + \bar{\mathbf{x}}^f$ in a local patch
`xa(ndiml,m)`	OUT	analysis ensemble perturbation $\mathbf{E}^a$ in a local patch
`xabar(ndiml)`	OUT	analysis ensemble mean $\bar{\mathbf{x}}^a$

Table A.4: Arguments of the subroutine lekf_core. ndiml, nobsl, and m denote the dimension of the local patch, the number of observations in the local patch, and ensemble size, respectively.

matrix-related procedures is required for this EnKF module.

A.3 LEKF

A core subroutine `lekf_core(h,rdiag,dep,xb,xa,xabar)` is developed for LEKF. Since the analysis equations are solved in local patches, the dimension of the local patch (`ndiml`) and the number of observations in the local patch (`nobsl`), which are defined as global variables of the module, need to be determined in advance. Table A.4 gives arguments of the subroutine `lekf_core`, where `m` denotes ensemble size.

Inside the subroutine, subroutines for real-symmetric matrix inverse, square root, and eivenvalue decomposition are called. Each needs to be prepared separately.

Appendix B

Recursive filter technique (Purser et al. 2003a)

B.1 Introduction

As introduced in Section 3.2.2, recursive filter (RF) is applied to realize the Gaussian shape of the horizontal correlation in 3DVAR. RF is a quasi-Gaussian filter which has an important advantage in computational cost.

3DVAR requires forward and adjoint transformations. Thus, given Gaussian transformation $\mathbf{G}$, the Cholesky decomposition

$$\mathbf{G} = \mathbf{L}\mathbf{L}^T \tag{B.1}$$

provides a solution that can be used in 3DVAR. Here, $\mathbf{L}$ is a lower triangular matrix which gives a forward transformation, and $\mathbf{L}^T$ gives its adjoint. Although this method is currently used in some operational 3DVAR systems, this matrix application is computationally expensive.

Alternatively, RF is much cheaper. RF consists of a set of forward and backward processes mutually adjoint:

$$B_i = \beta A_i + \sum_{j=1}^{n} \alpha_j B_{i-j} \tag{B.2}$$

$$C_i = \beta B_i + \sum_{j=1}^{n} \alpha_j C_{i+j} \tag{B.3}$$

which are nth-order RF (nth-RF), cf. eqs.(3.7) and (3.8) in the case of 4th-order. Here, A is an input signal, and C is the filtered output signal. The subscript i denotes the space index of grid points in one direction. Eq.(B.2) gives forward transformation, and eq.(B.3)

gives its adjoint. The key to implement RF is to find coefficients α and β. This appendix introduces RF theory based on Purser et al. (2003a) and describes details on how to find the RF coefficients especially in the case of 4th-RF.

B.2 nth-order recursive filter

Define a differential operator $\mathcal{D}_{(n)}$ as

$$\mathcal{D}_{(n)} = 1 - \frac{\sigma^2 \delta x^2}{2}\frac{d^2}{dx^2} + \frac{1}{2!}\left(\frac{\sigma^2 \delta x^2}{2}\frac{d^2}{dx^2}\right)^2 + \cdots + \frac{1}{n!}\left(-\frac{\sigma^2 \delta x^2}{2}\frac{d^2}{dx^2}\right)^n \tag{B.4}$$

where σ and δx denote a length scale measured in numbers of grids and a grid spacing, respectively, i.e. $\sigma\delta x$ denotes a length scale in a physical unit. $\mathcal{D}_{(\infty)}$ gives

$$\mathcal{D}_{(\infty)} = \exp\left(-\frac{\sigma^2 \delta x^2}{2}\frac{d^2}{dx^2}\right) \tag{B.5}$$

Let $A(x)$ and $C(x)$ be input and output signals of $\mathcal{D}_{(\infty)}^{-1}$, respectively. Consider their spectral representations $\tilde{A}(k)$ and $\tilde{C}(k)$:

$$A(x) = \frac{1}{2\pi}\int \tilde{A}(k) e^{ikx} dk \tag{B.6}$$

$$C(x) = \frac{1}{2\pi}\int \tilde{C}(k) e^{ikx} dk \tag{B.7}$$

Then,

$$\mathcal{D}_{(\infty)} C(x) = \frac{1}{2\pi}\int \tilde{C}(k) \exp\left(-\frac{\sigma^2 \delta x^2}{2}\frac{d^2}{dx^2}\right) e^{ikx} dk \tag{B.8}$$

$$= \frac{1}{2\pi}\int \tilde{C}(k) \exp\left(\frac{\sigma^2 \delta x^2}{2} k^2\right) e^{ikx} dk \tag{B.9}$$

Thus, the spectral representation of the pseudo-differential operator $\mathcal{D}_{(\infty)}$ becomes

$$\tilde{\mathcal{D}}_{(\infty)} = \exp\left(\frac{\sigma^2 \delta x^2}{2} k^2\right) \tag{B.10}$$

For an input signal with the form of delta function

$$A(x) = \delta(x) = \frac{1}{2\pi}\int e^{ikx} dk \tag{B.11}$$

the transformation by $\tilde{\mathcal{D}}^{-1}_{(\infty)}$ is written as

$$\tilde{\mathcal{D}}^{-1}_{(\infty)}\delta(x) = \frac{1}{2\pi}\int \exp\left(-\frac{\sigma^2\delta x^2}{2}k^2\right)e^{ikx}dk \tag{B.12}$$

$$= \frac{1}{2\pi}\exp\left(-\frac{x^2}{2\sigma^2\delta x^2}\right) \tag{B.13}$$

which is nothing but a Gaussian function with the length scale $\sigma\delta x$. Thus, the transformation $\tilde{\mathcal{D}}^{-1}_{(\infty)}$ gives a Gaussian filter, and $\tilde{\mathcal{D}}^{-1}_{(n)}$ is the nth-order approximation, that is,

$$C = \mathcal{D}^{-1}_{(n)}A \tag{B.14}$$

gives an nth-order quasi-Gaussian filter that transforms the input signal A to get the smoothed output C with the Gaussian shape with the length scale $\sigma\delta x$.

Define a finite difference operator K as

$$K\psi_i = -\psi_{i-1} + 2 - \psi_{i+1} \tag{B.15}$$

where ψ denotes an arbitrary variable with one-dimensional space index i (cf. eq.(3.1) in Purser et al. 2003a). Second order derivative with finite differencing is written as

$$\frac{d^2}{dx^2}\psi_i = \frac{\psi_{i-1} - 2\psi_i + \psi_{i+1}}{\delta x^2} \tag{B.16}$$

$$= -\frac{K}{\delta x^2}\psi_i \tag{B.17}$$

where uniform grid spacing is assumed. Using eq.(B.17), we can rewrite eq.(B.4) as

$$\mathcal{D}_{(n)} = 1 + \frac{\sigma^2 K}{2} + \frac{1}{2!}\left(\frac{\sigma^2 K}{2}\right)^2 + \cdots + \frac{1}{n!}\left(\frac{\sigma^2 K}{2}\right)^n \tag{B.18}$$

In fact, eq.(B.18) is not as accurate as what can be obtained using spectral representation, that is given by eq.(3.12) in Purser et al. (2003a). As shown in Purser et al. (2003a), small corrections to the coefficients of the terms $K^j (j = 1, \cdots, n)$ give better results. In either

way, $\mathcal{D}_{(n)}$ is written in the form of an nth-degree polynomial w.r.t. K. The polynomial can be written in a different form as

$$\mathcal{D}_{(n)} = \prod_{j=1}^{n} \left(1 - \frac{K}{\kappa_j}\right) \tag{B.19}$$

cf. eq.(A.1) in Purser et al. (2003a).

Define a shift operator Z as

$$Z\psi_i = \psi_{i+1} \tag{B.20}$$

cf. eq.(A.2) in Purser et al. (2003a). Then, eq.(B.15) can be written as

$$K = -Z^{-1} + 2 - Z \tag{B.21}$$

cf. eq.(A.3) in Purser et al. (2003a). Each factor of eq.(B.19) can be written as

$$1 - \frac{K}{\kappa_j} = \left(\frac{1 - \zeta_j Z^{-1}}{1 - \zeta_j}\right)\left(\frac{1 - \zeta_j Z}{1 - \zeta_j}\right) \tag{B.22}$$

cf. eq.(A.5) in Purser et al. (2003a), where ζ_j satisfies

$$\zeta_j^2 - (2 - \kappa_j)\zeta_j + 1 = 0 \tag{B.23}$$

Thus, the quasi-Gaussian transformation (eq.(B.14)) is rewritten as

$$C_i = \prod_{j=1}^{n} \left(\frac{1 - \zeta_j}{1 - \zeta_j Z^{-1}}\right)\left(\frac{1 - \zeta_j}{1 - \zeta_j Z}\right) A_i \tag{B.24}$$

We can separate the transformation process (eq.(B.24)) as follows:

$$B_i = \prod_{j=1}^{n} \left(\frac{1 - \zeta_j}{1 - \zeta_j Z^{-1}}\right) A_i \tag{B.25}$$

$$C_i = \prod_{j=1}^{n} \left(\frac{1 - \zeta_j}{1 - \zeta_j Z}\right) B_i \tag{B.26}$$

which are equivalent to

$$B_i = \beta A_i + \sum_{j=1}^{n} \alpha_j Z^{-j} B_i \tag{B.27}$$

$$C_i = \beta B_i + \sum_{j=1}^{n} \alpha_j Z^{j} C_i \tag{B.28}$$

where

$$\alpha_1 = \zeta_1 + \zeta_2 + \cdots + \zeta_n \tag{B.29}$$

$$\alpha_2 = \zeta_1\zeta_2 + \zeta_1\zeta_3 + \cdots + \zeta_{n-1}\zeta_n \tag{B.30}$$

$$\alpha_3 = \zeta_1\zeta_2\zeta_3 + \zeta_1\zeta_2\zeta_4 + \cdots + \zeta_{n-2}\zeta_{n-1}\zeta_n \tag{B.31}$$

$$\vdots$$

$$\alpha_n = \zeta_1\zeta_2\zeta_3\cdots\zeta_n \tag{B.32}$$

$$\begin{aligned}\beta &= (1-\zeta_1)(1-\zeta_2)\cdots(1-\zeta_n) \\ &= 1 - \alpha_1 - \alpha_2 - \cdots - \alpha_n\end{aligned} \tag{B.33}$$

Eqs. (B.27) and (B.28) are nothing but forward and backward processes of the nth-RF (eqs.(refeq:nrff2) and (refeq:nrfb2)). Thus, solving the quadratic equation (B.23) for ζ_j, we get coefficients for the nth-RF using eqs.(B.29)-(B.33). To solve the quadratic equation (B.23), κ_j are required. Since the polynomials (B.18) and (B.19) should be identical, we get κ_j by comparing the coefficients. The polynomial (B.18) gives the numerical values of the coefficients according to the length scale σ. The other polynomial (B.19) gives the coefficients

$$\begin{aligned}\prod_{j=1}^{n}\left(1-\frac{K}{\kappa_j}\right) = 1 \quad &+ \quad \left(-\frac{1}{\kappa_1} - \frac{1}{\kappa_2} - \cdots - \frac{1}{\kappa_n}\right)K \\ &+ \quad \left(\frac{1}{\kappa_1\kappa_2} + \cdots + \frac{1}{\kappa_{n-1}\kappa_n}\right)K^2 \\ &\vdots \\ &+ \quad (-1)^n\left(\frac{1}{\kappa_1\kappa_2\cdots\kappa_n}\right)K^n\end{aligned} \tag{B.34}$$

In the nth-RF, n roots of an nth-degree equation give κ_j. The details to get κ_j are described in the next section in the case of the 4th-RF.

B.3 4th-order recursive filter

Since 4th-RF has been implemented in the 3DVAR system on the SPEEDY model, this section describes specific solutions on the 4th-RF. Using the corrected polynomial given by eq.(3.12) in Purser et al. (2003a), eq.(B.18) in the 4th-order case is written as

$$\begin{aligned}\mathcal{D}_{(4)} = 1 \quad &+ \quad \frac{\sigma^2}{2}K + \left(\frac{\sigma^4}{8} + \frac{\sigma^2}{24}\right)K^2 + \left(\frac{\sigma^6}{48} + \frac{\sigma^4}{48} + \frac{\sigma^2}{180}\right)K^3 \\ &+ \quad \left(\frac{\sigma^8}{384} + \frac{\sigma^6}{192} + \frac{\sigma^4}{1920} + \frac{\sigma^2}{1120}\right)K^4 \end{aligned} \tag{B.35}$$

Alternatively, eq.(B.34) in the 4th-order case gives

$$\begin{aligned}\mathcal{D}_{(4)} = \prod_{j=1}^{4}\left(1 - \frac{K}{\kappa_j}\right) = 1 \quad &+ \quad \left(-\frac{1}{\kappa_1} - \frac{1}{\kappa_2} - \frac{1}{\kappa_3} - \frac{1}{\kappa_4}\right)K \\ &+ \quad \left(\frac{1}{\kappa_1\kappa_2} + \frac{1}{\kappa_1\kappa_3} + \frac{1}{\kappa_1\kappa_4} + \frac{1}{\kappa_2\kappa_3} + \frac{1}{\kappa_2\kappa_4} + \frac{1}{\kappa_3\kappa_4}\right)K^2 \\ &+ \quad \left(-\frac{1}{\kappa_1\kappa_2\kappa_3} - \frac{1}{\kappa_1\kappa_2\kappa_4} - \frac{1}{\kappa_1\kappa_3\kappa_4} - \frac{1}{\kappa_2\kappa_3\kappa_4}\right)K^3 \\ &+ \quad \left(\frac{1}{\kappa_1\kappa_2\kappa_3\kappa_4}\right)K^4 \end{aligned} \tag{B.36}$$

From the eqs.(B.35) and (B.36),

$$a = \frac{\sigma^2}{2} = -\frac{1}{\kappa_1} - \frac{1}{\kappa_2} - \frac{1}{\kappa_3} - \frac{1}{\kappa_4} \tag{B.37}$$

$$b = \frac{\sigma^4}{8} + \frac{\sigma^2}{24} = \frac{1}{\kappa_1\kappa_2} + \frac{1}{\kappa_1\kappa_3} + \frac{1}{\kappa_1\kappa_4} + \frac{1}{\kappa_2\kappa_3} + \frac{1}{\kappa_2\kappa_4} + \frac{1}{\kappa_3\kappa_4} \tag{B.38}$$

$$c = \frac{\sigma^6}{48} + \frac{\sigma^4}{48} + \frac{\sigma^2}{180} = -\frac{1}{\kappa_1\kappa_2\kappa_3} - \frac{1}{\kappa_1\kappa_2\kappa_4} - \frac{1}{\kappa_1\kappa_3\kappa_4} - \frac{1}{\kappa_2\kappa_3\kappa_4} \tag{B.39}$$

$$d = \frac{\sigma^8}{384} + \frac{\sigma^6}{192} + \frac{\sigma^4}{1920} + \frac{\sigma^2}{1120} = \frac{1}{\kappa_1\kappa_2\kappa_3\kappa_4} \tag{B.40}$$

Since eqs.(B.37)-(B.40) yield

$$\left(x - \frac{1}{\kappa_1}\right)\left(x - \frac{1}{\kappa_2}\right)\left(x - \frac{1}{\kappa_3}\right)\left(x - \frac{1}{\kappa_4}\right) = x^4 + ax^3 + bx^2 + cx + d = 0 \tag{B.41}$$

the four complex roots of the 4th-degree equation (B.41) give $1/\kappa_j$.

Provided κ_j, ζ_j are computed by solving the quadratic equation (B.23):

$$\zeta_j = 1 - \frac{\kappa_j}{2} \pm \sqrt{\left(1 - \frac{\kappa_j}{2}\right)^2 - 1} \tag{B.42}$$

As with ζ_j, the ones with the smaller absolute value are chosen. From eqs.(B.29)-(B.33), coefficients for the 4th-RF is given as

$$\alpha_1 = \zeta_1 + \zeta_2 + \zeta_3 + \zeta_4 \tag{B.43}$$

$$\alpha_2 = \zeta_1\zeta_2 + \zeta_1\zeta_3 + \zeta_1\zeta_4 + \zeta_2\zeta_3 + \zeta_2\zeta_4 + \zeta_3\zeta_4 \tag{B.44}$$

$$\alpha_3 = \zeta_1\zeta_2\zeta_3 + \zeta_1\zeta_2\zeta_4 + \zeta_1\zeta_3\zeta_4 + \zeta_2\zeta_3\zeta_4 \tag{B.45}$$

$$\alpha_4 = \zeta_1\zeta_2\zeta_3\zeta_4 \tag{B.46}$$

$$\beta = 1 - \alpha_1 - \alpha_2 - \alpha_3 - \alpha_4 \tag{B.47}$$

all of which are real numbers.

BIBLIOGRAPHY

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

[2] Anderson, J. L., 2001: An Ensemble Adjustment Kalman Filter for Data Assimilation. *Mon. Wea. Rev.*, **129**, 2884-2903.

[3] Andrews, A., 1968: A square root formulation of the Kalman covariance equations. *AIAA J.*, **6**, 1165-1168.

[4] Atlas, R., 1997: Atmospheric observations and experiments to assess their usefulness in data assimilation. *J. Meteor. Soc. Japan*, **75**(1B), 111-130.

[5] Barker, D. M., W. Huang, Y.-R. Guo, A. J. Bourgeois, and Q. N. Xiao, 2004: A Three-Dimensional Variational Data Assimilation System for MM5: Implementation and Initial Results. *Mon. Wea. Rev.*, **132**, 897-914.

[6] Bouttier, F., 1996: The Kalman Filter. *Proc. Seminar on Predictability*, Vol.1., ECMWF, Reading, Berkshire, UK, 221-245.

[7] Bouttier, F. and P. Courtier, 1999: *Data assimilation concepts and methods*. Meteorological Training Course Lecture Series, ECMWF, 75pp.

[8] Burgers, G., P. J. van Leeuwen, and G. Evensen, 1998: Analysis Scheme in the Ensemble Kalman Filter. *Mon. Wea. Rev.*, **126**, 1719-1724.

[9] Bishop, C. H., B. J. Etherton, and S. J. Majumdar, 2001: Adaptive Sampling with Ensemble Transform Kalman Filter. Part I: Theoretical Aspects. *Mon. Wea. Rev.*, **129**, 420-436.

[10] Cane, M. A., A. Kaplan, R. N. Miller, B. Tang, E. C. Hackert, and A. J. Busalacchi, 1996: Mapping tropical Pacific sea level: Data assimilation via a reduced state space Kalman filter. *J. Geophys. Res.*, **101**(C10), 22,599-22,617.

[11] Corazza, M., E. Kalnay, D. J. Patil, E. Ott, J. Yorke, I. Szunyogh, M. Cai, 2002: Use of the breeding technique in the estimation of the background error covariance matrix for a quasigeostrophic model. AMS Symposium on Observations, Data Assimilation and Probabilistic Prediction, Orland, Florida, 154-157.

[12] Corazza, M., E. Kalnay, D. J. Patil, I. Szunyogh, and S.-C. Yang, 2003: Keeping the bred vectors young: Impact on data assimilation. The presentation file available at http://www.atmos.umd.edu/ ekalnay/.

[13] Daley, R., 1991: *Atmospheric Data Analysis*. Cambridge University Press, 457pp.

[14] Danforth, C., E. Kalnay, and T. Miyoshi, 2005: Estimating and Correcting Global Weather Model Error. In preparation.

[15] Dee, D. P., 1995: On-line Estimation of Error Covariance Parameters for Atmospheric Data Assimilation. *Mon. Wea. Rev.*, **123**, 1128-1145.

[16] Dee, D. P. and A. M. da Silva, 1998: Data assimilation in the presence of forecast bias. *Quart. J. Roy. Meteor. Soc.*, **126**, 269-295.

[17] Ehrendorfer, M. and R. M. Errico, 1995: Mesoscale Predictability and the Spectrum of Optimal Perturbations. *J. Atmos. Sci.*, **52**, 3475-3500.

[18] Etherton, B. J. and C. H. Bishop, 2004: Resilience of Hybrid Ensemble/3DVAR Analysis Schemes to Model Error and Ensemble Covariance Error. *Mon. Wea. Rev.*, **132**, 1065-1080.

[19] 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.

[20] Fukumori I. and P. Malanotte-Rizzoli, 1995: An approximate Kalman filter for ocean data assimilation: An example with an idealized Guls Stream model. *J. Geophys. Res.*, **100** (C4), 6777-6793.

[21] Gaspari, G. and S. E. Cohn, 1999: Construction of correlation functions in two and three dimensions. *Quart. J. Roy. Meteor. Soc.*, **125**, 723-757.

[22] Gelb, A., J. F. Kasper, R. A. Nash, C. F. Price, and A. A. Sutherland, 1974: *Applied Optimal Estimation*. The M.I.T. Press, 374pp.

[23] Hamill, T. M. and C. Snyder, 2000: A Hybrid Ensemble Kalman Filter-3D Variational Analysis Scheme. *Mon. Wea. Rev.*, **128**, 2905-2919.

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

[25] Hamill, T. M., 2003: Ensemble-based Atmospheric Data Assimilation: A Tutorial. University of Colorado and NOAA-CIRES Climate Diagnostics Center, Boulder, Colorado, 46pp.

[26] Houtekamer, P. L. and H. L. Mitchell, 1998: Data Assimilation Using an Ensemble Kalman Filter Technique. *Mon. Wea. Rev.*, **126**, 796-811.

[27] Houtekamer, P. L. and H. L. Mitchell, 1999: Reply. *Mon. Wea. Rev.*, **127**, 1378-1379.

[28] Houtekamer, P. L. and H. L. Mitchell, 2001: A Sequential Ensemble Kalman Filter for Atmospheric Data Assimilation. *Mon. Wea. Rev.*, **129**, 123-137.

[29] Houtekamer, P. L. and H. L. Mitchell, 2003: Practical Ensemble Data Assimilation. Seminar on Recent developments in data assimilation for atmosphere and ocean, 8 to 12 September 2003, ECMWF, 193-210.

[30] Houtekamer, P. L., H. L. Mitchell, G. Pellerin, M. Buehner, M. Charron, L. Spacek, and B. Hansen, 2005: Atmospheric Data Assimilation with an Ensemble Kalman Filter: Results with Real Observations. *Mon. Wea. Rev.*, **133**, 604-620.

[31] Hunt, B. R., E. Kalnay, E. J. Kostelich, E. Ott, D. J. Patil, T. Sauer., I. Szunyogh, J. A. Yorke and A. V. Zimin, 2004: Four-dimensional ensemble Kalman filtering. *Tellus*, **56A**, 273-277.

[32] Ide, K., P. Courtier, M. Ghil, and A. C. Lorenc, 1997: Unified notation for data assimilation: Operational, sequential, and variational. *J. Meteor. Soc. Japan*, **75** (1B), 181-189.

[33] Jazwinski, A. H., 1970: *Stochastic Processes and Filtering Theory*. Academic Press, 376pp.

[34] Kalman, R. E., 1960: A New Approach to Linear Filtering and Prediction Problems. *J. Basic Eng., Trans. ASME*, 35-45.

[35] Kalnay, E. and Z. Toth, 1994: Removing growing errors in the analysis cycle. Preprints, Tenth Conference on Numerical Weather Prediction, Amer. Meteor. Soc., 212-216.

[36] Kalnay, E., M. Kanamitsu, R. Kistler, W. Collins, D. Deaven, L. Gandin, M. Iredell, S. Saha, G. White, J. Woollen, Y. Zhu, A. Leetmaa, B. Reynolds, M. Chelliah, W. Ebisuzaki, W. Higgins, J. Janowiak, K. C. Mo, C. Ropelewski, J. Wang, R. Jenne and D. Joseph: The NCEP/NCAR 40-Year Reanalysis Project. *Bull. Amer. Meteor. Soc.*, **77**(3), 437-471.

[37] Kalnay, E., 2003: *Atmospheric modeling, data assimilation and predictability.* Cambridge University Press, 341pp.

[38] Kistler, R., E. Kalnay, W. Collins, S. Saha, G. White, J. Woollen, M. Chelliah, W. Ebisuzaki, M. Kanamitsu, 2001: The NCEP-NCAR 50-Year Reanalysis: Monthly Means CD-ROM and Documentation. *Bull. Amer. Meteor. Soc.*, **82**(2), 247-267.

[39] Lorenz, E., 1996: Predictability - A problem partly solved. *Proc. Seminar on Predictability*, Reading, United Kingdom, ECMWF, 1-18.

[40] Lorenz, E. and K. Emanuel, 1998: Optimal sites for supplementary weather observations. *J. Atmos. Sci.*, **55**, 399-414.

[41] Marshall, J. and F. Molteni, 1993: Toward a Dynamical Understanding of Planetary-Scale Flow Regimes. *J. Atmos. Sci.*, **50**, 1792-1818.

[42] Miyoshi, T., 2004: Classical methods tour of advanced data assimilation using Lorenz 96 model. A final report submitted to METO658E, Department of Meteorology, University of Maryland, College Park.

[43] Miyoshi, T., 2004: A review of the deterministic ensemble Kalman filtering methods and related techniques. Masters Scholarly Paper. Department of Meteorology, University of Maryland, College Park, 20pp.

[44] Miyoshi, T. and E. Kalnay, 2005: Ensemble forecasting using orthogonalized and stochastically seeded bred vectors in the Lorenz 96 model with orography. In preparation.

[45] Molteni, F., 2003: Atmospheric simulations using a GCM with simplified physical parameterizations. I: model climatology and variability in multi-decadal experiments. *Clim. Dyn.*, **20**, 175-191.

[46] Nicolis, C., 2003: Dynamics of Model Error: Some Generic Features. *J. Atmos. Sci.*, **60**, 2208-2218.

[47] Ott, E., B. R. Hunt, I. Szunyogh, A. V. Zimin, E. J. Kostelich, M. Corazza, E. Kalnay, D. J. Patil, and J. A. Yorke, 2002: Exploiting local low dimensionality of the atmospheric dynamics for efficient Kalman filtering. ArXiv:arc-ive/paper 020358, http://arxiv.org/abs/physics/020358.

[48] Ott, E., B. R. Hunt, I. Szunyogh, A. V. Zimin, E. J. Kostelich, M. Corazza, E. Kalnay, D. J. Patil, and J. A. Yorke, 2004: A local ensemble Kalman filter for atmospheric data assimilation. *Tellus*, **56A**, 415-428.

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

[50] Parrish, D. F. and J. C. Derber, 1992: The National Meteorological Center's Spectral Statistical-Interpolation Analysis System. *Mon. Wea. Rev.*, **120**, 1747-1763.

[51] Purser, R. J., W. S. Wu, D. F. Parrish, and N. M. Roberts, 2003a: Numerical Aspects of the Application of Recursive Filters to Variational Statistical Analysis. Part I:

Spatially Homogeneous and Isotropic Gaussian Covariances. *Mon. Wea. Rev.*, **131**, 1524-1535.

[52] Purser, R. J., W. S. Wu, D. F. Parrish, and N. M. Roberts, 2003b: Numerical Aspects of the Application of Recursive Filters to Variational Statistical Analysis. Part II: Spatially Inhomogeneous and Anisotropic General Covariances. *Mon. Wea. Rev.*, **131**, 1536-1548.

[53] Szunyogh, I., E. J. Kostelich, G. Gyarmati, D. J. Patil, B. R. Hunt, E. Kalnay, E. Ott, and J. A. Yorke, 2004: Assessing a local ensemble Kalman filter: Perfect model experiments with the NCEP global model. Tellus, in print.

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

[55] van Leeuwen, P. J., 1999: Comment on "Data Assimilation Using an Ensemble Kalman Filter Technique". *Mon. Wea. Rev.*, **127**, 1374-1377.

[56] Whitaker, J. S. and T. M. Hamill, 2002: Ensemble Data Assimilation without Perturbed Observations. *Mon. Wea. Rev.*, **130**, 1913-1924.

[57] Wang, X. and C. H. Bishop, 2003: A Comparison of Breeding and Ensemble Transform Kalman Filter Ensemble Forecast Schemes. *J. Atmos. Sci.*, **60**, 1140-1158.

[58] Wang, X., C. H. Bishop, S. J. Julier, 2004: Which Is Better, an Ensemble of Positive-Negative Pairs or a Centered Spherical Simplex Ensemble?. *Mon. Wea. Rev.*, **132**, 1590-1605.

[59] Wilks, D. S., 1995: *Statistical Methods in the Atmospheric Sciences*, Academic Press, 467pp.

[60] Zupanski, D., 1997: A General Weak Constraint Applicable to Operational 4DVAR Data Assimilation Systems. *Mon. Wea. Rev.*, **125**, 2274-2292.