ABSTRACT

Title of Document: APPLICATIONS OF THE LETKF TO ADAPTIVE OBSERVATIONS, ANALYSIS SENSITIVITY, OBSERVATION IMPACT AND THE ASSIMILATION OF MOISTURE

Junjie Liu, Doctor of Philosophy, 2007

Directed By: Professor Eugenia Kalnay
Department of Atmospheric and Oceanic Science

In this thesis we explore four new applications of the Local Ensemble Transform Kalman Filter (LETKF), namely adaptive observations, analysis sensitivity, observation impact, and multivariate humidity assimilation. In each of these applications we have obtained promising results.

In the adaptive observation studies, we found that ensemble spread strategy, where adaptive observations are selected among the points with largest ensemble spread (with the constraint that observations cannot be contiguous in order to avoid clusters of adaptive observations) is very effective and close to optimal sampling. The application on simulated Doppler Wind Lidar (DWL) adaptive observation studies shows that 3D-Var is as effective as LETKF with 10% adaptive observations sampled with the ensemble spread strategy. With 2% adaptive observations, 3D-Var is not as effective as the LETKF.

In the analysis sensitivity study, we proposed to calculate this quantity within the LETKF with low additional computational time. Unlike in 4D-Var (Cardinali et al., 2004), in the LETKF, the computation is exact and satisfies the theoretical value limits (between 0 and 1). The results from simulated experiments show that the trace of analysis sensitivity qualitatively reflects the observation impact obtained from independently computed data addition or data denial OSSE experiments.

In the observation impact study, we derived a formula to estimate the impact of observations on short-range forecasts as in Langland and Baker (2004), but without using an adjoint model. Both methods estimate more than 90% accuracy the actual observation impact on the short-range forecast error improvement. Like the adjoint method, the method we proposed detects observations that have either large random error or unaccounted bias. This method can be easily calculated within the LETKF, and provides a powerful tool to estimate the quality of observations.

Finally, for the first time, we assimilate humidity observations multivariately in both perfect model experiments and real data assimilation. We found that multivariate assimilation is better than univariate assimilation. The assimilation of pseudo-RH (Dee and da Silva, 2003) is better than the choice of specific humidity and relative humidity. The multivariate assimilation of AIRS specific humidity retrievals on NCEP GFS system shows positive impact on the winds analysis.

APPLICATIONS OF THE LETKF TO ADAPTIVE OBSERVATIONS, ANALYSIS SENSITIVITY, OBSERVATION IMPACT AND THE ASSIMILATION OF MOISTURE

By

Junjie Liu

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
2007

Advisory Committee:

Professor Eugenia Kalnay, Chair
Professor Ernesto Hugo Berbery
Professor James Carton
Professor Brian Hunt
Professor Istvan Szunyogh

Acknowledgements

First of all, my deepest gratitude goes to my advisor, Prof. Eugenia Kalnay, for providing me valuable guidance, and always being supportive and encouraging. I especially want to thank her for showing me the perseverance and passion in doing research, and for leading me to discover the joy of doing research. I would also like to thank my committee members: Prof. Ernesto Hugo Berbery, Prof. James Carton, Prof. Brian Hunt, and Prof. Istvan Szunyogh for many suggestions. Also, I would like to express my gratitude to the professors and students in the Chaos group, with whom I had very productive scientific discussions. I want to especially thank Prof. Ed Ott, who patiently read my paper draft and gave valuable suggestions, and Dr. Elana Fertig and Dr. Hong Li, with whom I had enjoyable collaborations in the last three years.

I want to thank Dr. Ricardo Todling for the help in the initial stage of my research. I also want to thank Dr. Takemasa Miyoshi for providing the source code of the SPEEDY model with 3D-Var and LEKF data assimilation schemes, Dr. Chris Barnet and Eric Maddy for providing AIRS humidity retrievals, Dr. Shu-Chih Yang for many useful discussions and encouragements, and Debra Baker for reading one of my papers. I am indebted to the students in our department for providing a stimulating environment to learn and to grow. I am especially grateful to Bin Guan, Chanh Kieu, Can Li, and Haifeng Qian for their helpful discussions in finishing the classes and the research afterwards.

My sincerest gratitude goes to a special person, Malise Dick, the husband of my advisor, who passed away in June 2007. He always gave me the warmest encouragement during my difficult times. He showed me the world beyond research, and with Eugenia invited me as an extended part of their family.

I am also deeply grateful to Prof. Yihui Ding and Prof. Jinhai He for their encouragement and advising when I began my academic career in China. I would like to give my special thanks to my friends Dr. Huiju Zhang and Dr. Si Chen for the emotional support, entertainment and caring they provided during the last four years.

Lastly, and most importantly, I would like to express my deepest gratitude to my parents, my brother and sisters, and especially my husband, Yu Pan. Without their love, supports and encouragements, I would never have gone this far.

Table of Contents

This Table of Contents is automatically generated by MS Word, linked to the Heading formats used within the Chapter text.

List of Tables

List of Figures

Chapter 1 Introduction

Data assimilation is a process combining observation information and model forecast (background) based on their uncertainty estimation (e.g., Kalnay, 2003). Ensemble Kalman Filter (EnKF, Evensen, 1994; Anderson, 2001; Bishop et al., 2001; Houtekamer and Mitchell; 2001; Whitaker and Hamill, 2002; Ott et al., 2004; Hunt et al., 2007) is a type of data assimilation in which the time changing background error covariance is estimated from an ensemble of forecasts. The Local Ensemble Transform Kalman Filter (LETKF, Hunt et al., 2007) is an efficient type of EnKF, which calculates the ensemble analyses in a local patch centered at each grid point. The analysis at each grid point is independent from each other, so the scheme is highly parallel. The analysis mean state in the LETKF is

$$\overline{\mathbf{x}}^a = \overline{\mathbf{x}}^b + \mathbf{X}^b \widetilde{\mathbf{K}}(\mathbf{y}^o - h(\overline{\mathbf{x}}^b)) \tag{1.1}$$

The vectors $\overline{\mathbf{x}}^a$ and $\overline{\mathbf{x}}^b$ are the mean analysis and background field. $\mathbf{y}^o$ is the observation vector, and $h(\cdot)$ is nonlinear observation operator interpolating the mean background to the observation space. $\mathbf{X}^b$ is the matrix whose columns are the ensemble perturbations, which are the difference between ensemble forecasts and ensemble mean state. $\widetilde{\mathbf{K}} = \left[(\mathbf{H}\mathbf{X}^b)^T \mathbf{R}^{-1}(\mathbf{H}\mathbf{X}^b) + (K-1)\mathbf{I}\right]^{-1}(\mathbf{H}\mathbf{X}^b)^T \mathbf{R}^{-1}$ is the Kalman gain in the ensemble perturbation space, with K equal to the number of the ensemble members. $\mathbf{R}$ is the observation error covariance. $\mathbf{H}\mathbf{X}^b$ is the matrix whose columns are the ensemble perturbations in the observation space. The analysis ensemble perturbations in the LETKF are a linear combination of the background ensemble perturbations:

$$\mathbf{X}^a = \mathbf{X}^b \left[(K-1)\widetilde{\mathbf{P}}^a \right]^{\frac{1}{2}} \tag{1.2}$$

where $\widetilde{\mathbf{P}}^a = \left[(\mathbf{H}\mathbf{X}^b)^T \mathbf{R}^{-1} \mathbf{H}\mathbf{X}^b + (K-1)\mathbf{I} \right]^{-1}$ is the analysis error covariance in the ensemble perturbation space. The background error covariance and the analysis error covariance are estimated as:

$$\mathbf{P}^b = \frac{1}{K-1} \mathbf{X}^b \mathbf{X}^{bT} \tag{1.3}$$

$$\mathbf{P}^a = \frac{1}{K-1} \mathbf{X}^b \widetilde{\mathbf{P}}^a \mathbf{X}^{bT} \tag{1.4}$$

Throughout this thesis, we will study adaptive observations, analysis sensitivity, observation impact on the short-range forecast, and assimilation of humidity observations with the LETKF scheme.

1.1 Adaptive observations

Conventional atmospheric observations, such as rawinsondes, are fixed with time, and are concentrated over land. The locations that do not have conventional observations at all, such as most of the ocean areas, were never observed before the advent of satellite data. In the satellite period (from 1979 on), satellites provide global observational coverage, but each location can be at most observed twice a day. Due to cloud contamination and some other reasons, some locations may not have any observations for more than a day. This insufficient observational coverage problem is more severe over ocean than over land. However, the predictability over land is determined by the analysis accuracy of the upstream regions, i.e., over ocean. Therefore, in 1996, Snyder (1996) proposed to allocate limited rawinsonde observation resources adaptively, an approach called "targeted" or "adaptive"

observations. The idea of adaptive observation is to select the location of observations where they can be mostly useful in improving the forecast results.

Later on, some field experiments were carried out to test the effectiveness of adaptive observations, such as the Fronts and Atlantic Storm-Track Experiment (FASTEX), the North Pacific Experiment (NORPEX), Winter Storm Reconaissance Program and Atlantic TOST/TReC (Snyder, 1996; Joly et al., 1997; Emanuel and Langland, 1998; Bergot, 1999; Langland et al., 1999a; Langland et al., 1999b; Pu and Kalnay, 1999; Szunyogh et al., 1999; Majumdar et al., 2002; Toth et al., 2002; Langland, 2005). In most of these field experiments, they aimed to improve the short range forecast over land (verification region represented by $\sum$, grey area in Figure 1.1) by observing a limited area over the targeted area (white area in Figure 1.1 represented by T).

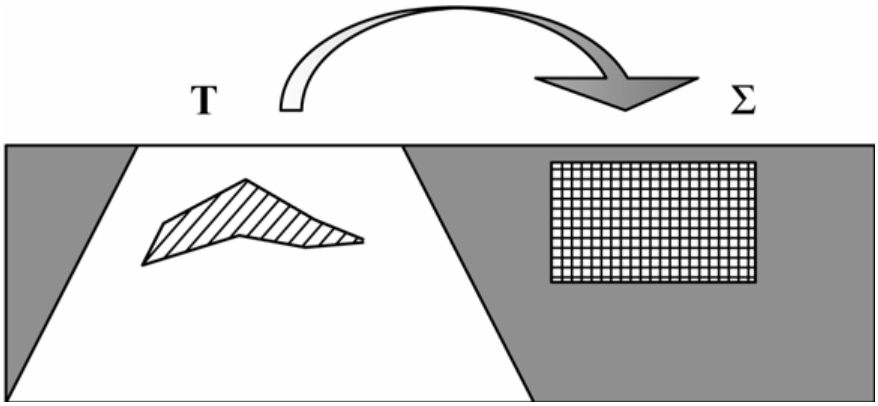


Figure 1.1 Schematic illustration of the concept of the 'adaptive /target observations': the grey areas identify land, while the white region identifies the ocean. T is the target area and $\sum$ is the verification region. (From Buizza et al., 2007)

The concept of adaptive observations has been mostly used in designing dropsonde aircraft routes to improve short range forecasts over some verification region in field campaigns. However, it is also a useful tool to save energy for any satellite instrument designed to "dwell" in regions of high uncertainty rather than providing uniform coverage along the orbit as conventionally done. Doppler Wind Lidar (DWL) is such an instrument which gives 'line of sight' wind estimate by measuring the reflection of a lidar shot on either molecules or aerosols. Detecting such a signal requires a large amount of energy. Therefore, the U.S. DWL will be operated in an adaptive mode, in which the goal is "to obtain 90% improvement from 10% coverage". Shown in Figure 1.2 is an example of targeted DWL observation distribution from Observing System Simulation Experiments (OSSEs). The white symbols are the full coverage, and the red symbols are the adaptive observations. The adaptive observation locations are either the area that the verification region most sensitive to or the areas that have largest uncertainty.

Our study will focus on selecting adaptive observations based on reducing the analysis uncertainty. The central issue in this problem is how to get the dynamical uncertainty estimation. LETKF, like any other EnKF, provides both the background uncertainty as well as analysis uncertainty estimations along the analysis (Equation (1.3) and (1.4)). Therefore, it is straightforward to do adaptive observation within the LETKF data assimilation framework. We will explore the ensemble-based adaptive observation strategies in both a simple model (Lorenz-40 variable model, Lorenz and

Emanuel, 1998) and in a global primitive equation model to sample the simulated DWL observations.

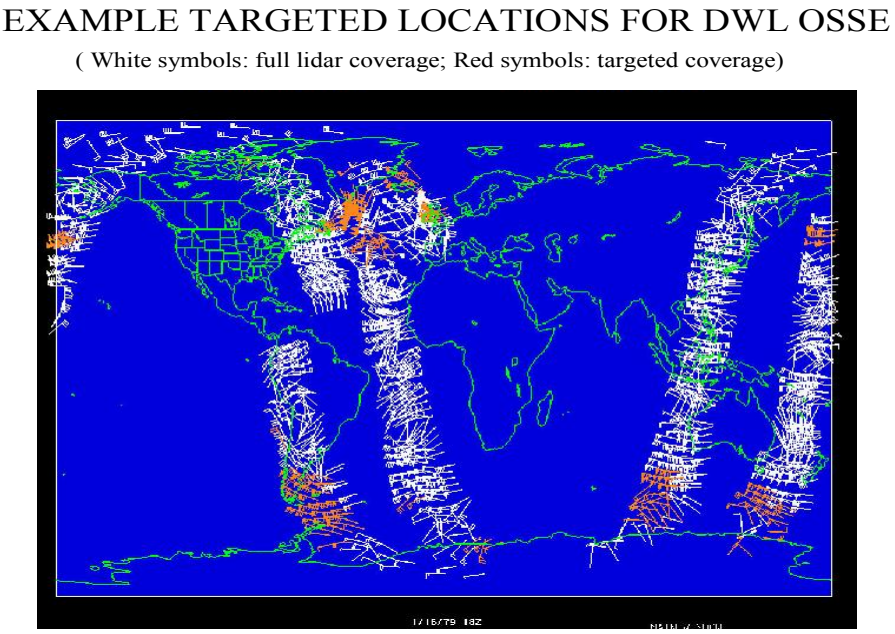


Figure 1.2 Example of targeted locations for DWL OSSE. From a presentation by Mike Hardesty (2006). The white symbols: full lidar coverage; Red symbols: targeted coverage.

In Chapter 2, we will compare several ensemble-based adaptive observation strategies using Lorenz-40 variable model (Lorenz and Emanuel, 1998) following the same experimental setup as previous studies (Lorenz and Emanuel, 1998; Hansen and Smith, 2001; Trevisan and Uboldi, 2004). We will show the performance of each strategy and compare with the best results published so far with this simple model. In Chapter 3, we perform OSSEs with the global primitive equation model known as SPEEDY (Molteni, 2003). We compare different strategies by sampling the simulated DWL observations uniformly, randomly, based on the background uncertainty estimated from the LETKF, and also the climatological uncertainty estimation. We

compare the analysis improvement due to the DWL observations from these different adaptive strategies in both the LETKF data assimilation system and the 3D-Var assimilation system. We further study the effectiveness of 3D-Var and LETKF with both dense sampling (10% of DWL total coverage in six-hour) and sparse sampling scenarios (2% of DWL global coverage in six-hour). This paper has been published in GRL (Liu and Kalnay, 2007).

1.2 Analysis sensitivity and observation impact

Modern operational data assimilation systems have evolved into very complicated systems combining high resolution dynamical model and the observations from both routine network and satellites. With the assimilation of kilo-channel satellite, such as Advanced InfraRed Satellite (AIRS), assimilation systems become more complicate, though only about 300 channels have been assimilated (e.g., Joiner et al., 2004). In such a complex system, it is necessary to monitor the role of each factor, such as how much the information comes from the background, and how much comes from each type of observations. Cardinali et al. (2004) proposed a method to calculate analysis sensitivity in a 4D-Var system, which measures how sensitive the analysis value is to the observations. It is complementary (adding up to 1) to the sensitivity to the background at the observation location. The sum of the analysis sensitivity of each type observation gives the information content of that type observation. The comparison of the information content can show the relative importance of each type observation in the data assimilation system, such as the result obtained by Cardinali et al. (2004) in a 4D-Var system (Figure 1.3). However, in the 4D-Var system, the calculation of analysis error covariance, which is part of the

analysis sensitivity calculation, needs some approximations, which creates some values of the analysis sensitivity outside the 0 to 1 range.

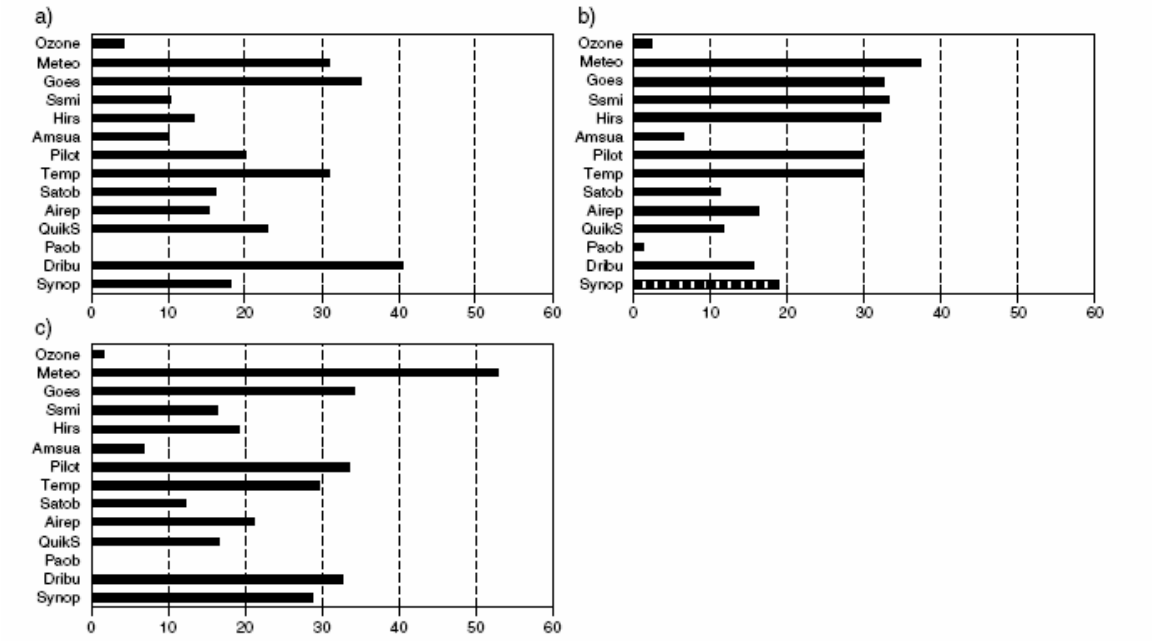


Figure 1.3 Average analysis sensitivity (%) for each of the main observation types (See Table 1 in Cardinali et al. 2004 for the full name of each observation type). (a) for Northern Hemisphere extratropics, (b) for the tropics, (c) for the Southern Hemisphere extratropics. From Cardinali et al. (2004)

Analysis sensitivity allows monitoring the sensitivity of the assimilation system to each component within the data assimilation system (Figure 1.3) and the trace of analysis sensitivity has also been used in selecting the channels from kilo-channel satellite (Fourrie and Thepaut, 2003). However, the diagnostics based on analysis sensitivity can not evaluate the actual quality of observations. Though statistically the assimilation of observations improves the analysis and so it improves the short-range forecast, in some cases, some observations may actually deteriorate the analysis. In addition, analysis sensitivity can only show the relative importance of different observations. It can not show the actual observation impact on the forecast.

The method proposed by Langland and Baker (2004) is pioneering in being able to detect poor observations, and showing the actual impact of each type of observations, even each channel of satellite, on the forecast. As shown in Figure 1.4 is the actual impact of some sensors of AQUA satellite on the improvement of forecast accuracy due to assimilation of the observations at 00hr. The positive values indicate that the observations from those channels actually increase the forecast error. It shows that the assimilation of the radiance from some channels makes the forecast worse, which identifies problems with either observing systems or assimilation systems, and provide the guidance for further improvement. By grouping the observations based on instrument types, it can further compare the actual observation impact of different instrument types on the forecast, as shown in Figure 1.5.

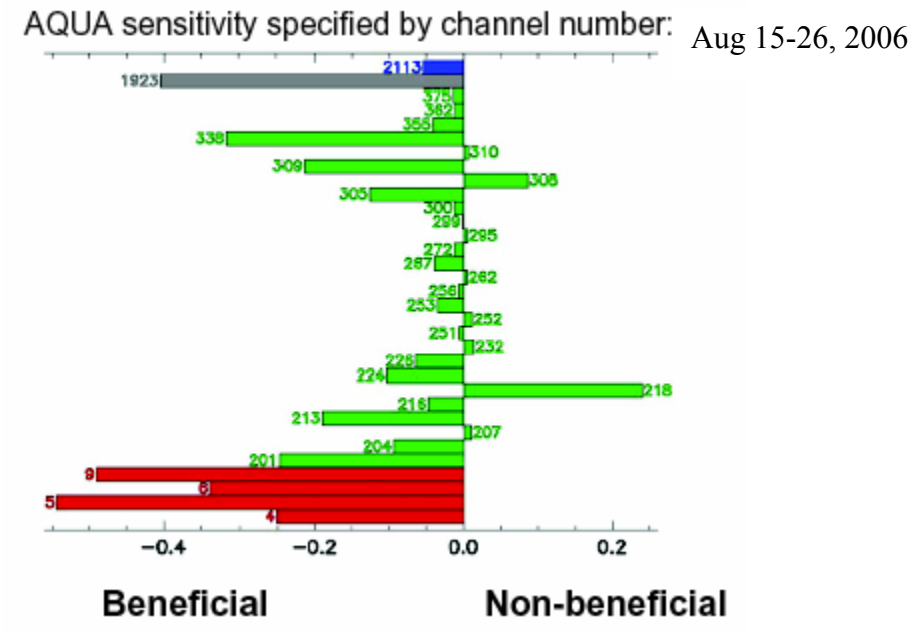


Figure 1.4 Assessments of AQUA sensors. Red: AMSU/A; Green: AIRS longwave14-13μm; Grey: shortwave 4.474μm; Blue: AIRS shortwave 4.180μm (From the presentation by Bishop in University of Maryland, 2007)

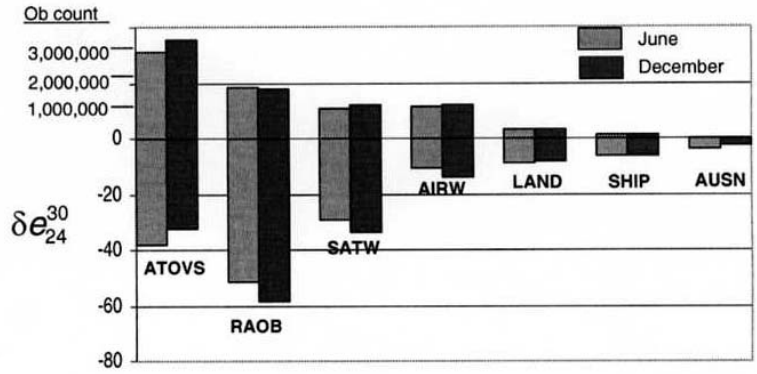


Figure 1.5 Summed global observation impact for June and December 2002, partitioned by instrument type. Includes all observations assimilated at 00UTC. The key is as follows: ATOVS, temperature retrievals; RAOB, rawinsondes; SATW, cloud and feature-track winds; AIRW,, commercial aircraft observations; LAND, land surface observations; SHIP, ship surface observations; AUSN, synthetic sea level pressure data (Southern Hemisphere only). From Langland and Baker (2004)

LETKF provides a framework to calculate analysis sensitivity and obtain the observation impact without using the adjoint model. Since analysis uncertainty is calculated along with the data assimilation in the LETKF (Equation (1.4)), the calculation of analysis sensitivity needs no approximation. In the LETKF, the analysis ensemble perturbations are linear combination of the background ensemble perturbations (Equation (1.4)). The analysis ensemble can also be written as a linear combination of the background ensemble (Chapter 5). When the forecast length is short enough that the perturbations with respect to the ensemble mean grow linearly, we can estimate the ensemble forecasts at the verification time t using the same weights as at the initial time. With this approximation, we derive a new procedure to calculate the observation impact on any short-range forecast using ensemble but without using adjoint (Chapter 5).

In Chapter 4, we give a detailed calculation procedure of the analysis sensitivity without any approximation in the LETKF data assimilation system. We

verify our calculation procedure in the Lorenz-40 variable model, and further explore the usefulness of analysis sensitivity in a global primitive equation model (SPEEDY) by comparing the information content and the results from "data denial" and "add-on" experiments. In Chapter 5, we derive an ensemble method which can calculate the same observation sensitivity as the adjoint method proposed by Langland and Baker (2004), but without using the adjoint model. We compare the results from the ensemble sensitivity method we proposed with the adjoint method by Langland and Baker (2004) in the Lorenz-40 variable model.

1.3 Humidity data assimilation

Due to the exponential variability of atmospheric moisture in latitude and height, the poor quality of humidity observations and the model errors related with moisture parameterizations, the assimilation of humidity observations is a difficult problem. With the improvement of observation quality and parameterization process, currently, most operational centers (NCEP, ECMWF) assimilate humidity observations within their assimilation systems. The assimilation approaches in these operational centers are variational approaches using a constant background error covariance (e.g., Kalnay, 2003). However, unlike the other dynamical variables, the humidity field changes with time and locations abruptly, which makes the constant error variance assumption less valid. Due to the small scale features of the humidity field, it is difficult to obtain the statistical covariance between humidity field and the other dynamical variables. Therefore, operational centers assimilate humidity observations uni-variately.

The humidity field can be represented in several different ways (e.g., dew point depression, specific humidity or relative humidity). This leads to several choices of assimilation variables, such as specific humidity, the logarithm of specific humidity, and the relative humidity. The different choices of variable type results in the different observation error distribution. In most of these choices, the observation error distribution is far from Gaussian. Since the Gaussian observation error distribution is assumed in data assimilation schemes, the choice of assimilated variable type is a central issue in humidity data assimilation. Dee and da Silva (2003) proposed to use pseudo-relative humidity (pseudo-RH) as the observed variable, which is to normalize the observed specific humidity by the saturated specific humidity from the background field. Holm (2002), based on a then unpublished idea of Dee and da Silva (2003), proposed a method to re-formulate the humidity variable. The chosen humidity control variable is a normalized relative humidity normalizing the relative humidity by a polynomial approximation of the background error. In both studies, the proposed variables, either pseudo-RH or normalized relative humidity has a more Gaussian observation error distribution than other choices of humidity variables.

Unlike variational assimilation methods, in the LETKF (or any other EnKF), the background error covariance (Equation (1.3)) is updated each analysis cycle based on the background ensemble forecasts. In addition, the background error covariance automatically couples the error statistics of all the dynamical variables together. Therefore, with an EnKF as a data assimilation scheme to assimilate humidity

observations, it can more accurately capture the time changing error characteristics and can easily couple the humidity field with other dynamical variables.

In Chapter 6, we perform OSSEs using the LETKF to assimilate humidity observations both uni-variately and multivariately in a global primitive equation model. We will compare pseudo-RH with the other choices of humidity observation types when the specific humidity observations have non-Gaussian observation error. In addition, we assimilate AIRS humidity retrievals within the NCEP GFS T64L28 system with specific humidity and pseudo-RH as assimilated humidity variable type in a coupled (multivariate) mode. As far as we know, this is the first time that moisture observations have been assimilated multivariately.

Chapter 2 : Adaptive observation strategies based on the Local Ensemble Transform Kalman Filter using the Lorenz-40 variable model

2.1 Introduction

Strategies to select the location of observations where they can be mostly useful in improving the forecast results are called as "targeted" or "adaptive" observation strategies (Snyder, 1996). The effectiveness of some adaptive strategies has been tested in some field experiments, such as Fronts and Atlantic Storm-Track Experiment (FASTEX), the North Pacific Experiment (NORPEX), Winter Storm Reconaissance Program and Atlantic TOST/TReC (Snyder, 1996; Joly et al., 1997; Emanuel and Langland, 1998; Bergot, 1999; Langland et al., 1999a; Langland et al., 1999b; Pu and Kalnay, 1999; Szunyogh et al., 1999; Majumdar et al., 2002; Toth et al., 2002; Langland, 2005). There are two basic types of adaptive observation strategies. One class is the adjoint based techniques, such as singular vector method (Palmer et al., 1998; Morss and Emanuel, 2001; Langland, 2005). The other is ensemble-based techniques such as the ensemble spread method (Lorenz and Emanuel, 1998; Hamill and Snyder, 2002), the Ensemble Transform Kalman Filtering (ET KF) (Bishop et al., 2001; Majumdar et al., 2002; Hamill and Snyder, 2002), and the quasi-inverse technique (Pu and Kalnay, 1999). The main difference between these two types of methods is the requirement of the adjoint model. The singular vector method uses the adjoint model to propagate the forecast uncertainty in the verification time back to the targeting time. The location with the largest error growth rate is chosen as the adaptive observation location. Ensemble based adaptive

observation methods do not use adjoint model, but use ensemble forecast information to identify the locations with largest uncertainty at the targeting time.

With the development of ensemble data assimilation methods in recent years (Evensen, 1994; Anderson, 2001; Bishop et al., 2001; Whitaker and Hamill, 2002; Ott et al., 2004; Hunt et al., 2007), ensemble based adaptive observation strategies have been proposed (Hamill and Snyder, 2002; Majumdar et al., 2002). In this chapter, we will focus on the ensemble based adaptive observation strategies derived from the LETKF data assimilation scheme. We will discuss the formulation, characteristics and the relationship of the background ensemble spread method, local analysis ensemble spread method and a combined method we proposed (Section 3.3). To test the accuracy of these methods, we use Lorenz-40 variable model, and follow the same experimental design as the previous studies that have used the same model to test adaptive observation strategies (Lorenz and Emanuel, 1998; Berliner et al., 1999; Hansen and Smith, 2000; Trevisan and Uboldi, 2004). We will further compare our results with the best result published so far (Hansen and Smith, 2000) with the same model and same experimental design, but different adaptive observation strategy.

This chapter is organized as follows: Section 2.2 describes the experimental design; Section 2.3 gives the formulation of several adaptive strategies; Section 2.4 illustrates the relationship between background ensemble spread method and local analysis ensemble spread method discussed in Section 2.3; Section 2.5 presents the results from these different adaptive observation strategies; Section 2.6 is a summary.

2.2 Experimental Design

The Lorenz 40-variable model is governed by the following equation:

$$\frac{d}{dt}x_j = (x_{j+1} - x_{j-2})x_{j-1} - x_j + F \qquad (2.1)$$

The variables (x_j, j=1…J) represent a "meteorological" variable on a "latitude circle" with periodic boundary conditions. As in previous studies, J is chosen to be equal to 40. The time step is 0.05, which corresponds to a 6-hour integration interval. F is the external forcing, which is equal to 8 for the nature run, and equal to 7.6 when do the forecast, thus introducing some model error.

Observations are obtained from the nature run (long-term "true evolution") plus Gaussian distribution errors with standard deviation equal to 0.2. Following previous studies (Lorenz and Emanuel, 1998), we observe the variables every six-hour at every "land" grid point (from 21 to 40), and a single adaptive observation from one of the points over "ocean" (grid points 1-20). The analysis is the combination of the six-hour forecast and both routine observations over land and the adaptive observation over ocean. The optimality of this additional observation is evaluated by the analysis error at the observation time and the 10-day forecast error.

We use a 20-member ensemble to estimate the background error covariance, which is used in the data assimilation to represent the background error. In order to compensate for the sampling error due to the insufficient ensemble members, we use a multiplicative inflation method (Anderson and Anderson, 1999) on the background

error covariance, i.e., the background error covariance is multiplied by a number larger than 1. The estimation method is based on the online estimation method proposed by Miyoshi (2005) (see appendix A). It is valid when the observation error statistics reflect the true observation uncertainty (Li, 2007), which is the case in our experimental setup. Unlike Miyoshi (2005), we estimate the inflation factor patch by patch instead of estimating a global inflation factor, since the observation coverage is non-uniformly distributed in our experimental design, and the inflation factor depends strongly on the observation coverage (Whitaker et al., 2007). The inflation factor is larger over the area where there are more observations, such as land and adjacent areas，and smaller inside of the ocean area where the observation is only from adaptive observation, as shown from time-average inflation factor from background ensemble-spread strategy (Figure 2.1). Since we add model error in our forecast model, the inflation factor also partially accounts for model error.

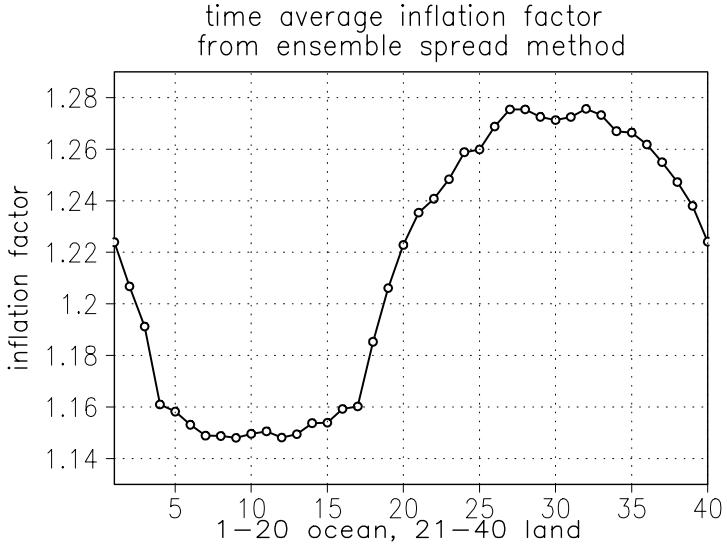


Figure 2.1 Time averaged inflation factor dependence on locations obtained from the background ensemble spread adaptive observation strategy

2.3 Formulation of adaptive observation strategies

The purpose of exploring adaptive observations is to maximize the analysis or the forecast uncertainty reduction with the same amount of observation resources. Since in ensemble data assimilation, the background uncertainty (Equation (1.3)) and the analysis uncertainty (Equation (1.4)) are calculated along with the data assimilation without using the actual observation value, the ensemble data assimilation provides the statistics to guide the adaptive observation network design. In the following discussion, we will focus on how to minimize the analysis error rather than the short-range forecast error with adaptive observation strategies.

The trace of the analysis error covariance has been shown to be an appropriate statistical standard to evaluate the accuracy of the analysis (Berliner, et al., 1999). The optimal adaptive observation is to make the trace of the analysis error covariance, referred to as the analysis ensemble spread, as small as possible. In the EnKF, since the analysis error covariance is proportional to the background error covariance, minimizing the background uncertainty in the background ensemble spread method indirectly minimizes the analysis uncertainty (Section 2.3.1). With a single adaptive observation, minimizing the six-hour forecast uncertainty in the background ensemble spread method also minimizes analysis uncertainty (Section 2.4). In EnKF, since the analysis error covariance is part of the data assimilation, we can directly minimize the trace of analysis error covariance. Unlike other EnKF data assimilation schemes, LETKF calculates the analysis error covariance $\mathbf{P}^a$ locally. Therefore, we call the adaptive method based on the diagonal value of local $\mathbf{P}^a$ "local analysis ensemble

spread" method (Section 2.3.2). Although LETKF allows parallel computing of the analysis ensemble spread, it would still require large computational time if we try to select a large number of adaptive observations. Thus, we combine the economical background ensemble spread method and local analysis ensemble spread method in a combined background-analysis ensemble spread method (Section 2.3.3), taking advantage of both methods. Finally, we discuss one "ideal" adaptive observation strategy (Section 2.3.4), in which we use the truth to find the optimal adaptive observation locations, and use it as an unattainable benchmark.

2.3.1 Background ensemble spread method

In EnKF, the six-hour ensemble forecasts give the estimation of the background error covariance. Ensemble spread is the trace of the background error covariance, defined by

$$S_j = (K-1)^{-1} \sum_{i=1}^{K} (\mathbf{x}_{i,j}^b - \overline{\mathbf{x}}_j^b)(\mathbf{x}_{i,j}^b - \overline{\mathbf{x}}_j^b)^T \tag{2.1}$$

$\mathbf{x}_{i,j}^b$ is the i^{th} background ensemble member at the grid point j, K is the number of ensemble members, $\overline{\mathbf{x}}_j^b$ is the ensemble mean state at the grid point j.

In the background ensemble spread adaptive observation strategy, the adaptive observation location is the location with largest background ensemble spread of all the potential adaptive observation locations over ocean. By putting the observation at the location with largest background ensemble spread, the analysis gives the largest weight to the adaptive observation compared to the other potential adaptive observations. In addition, it improves the analysis accuracy most by assimilating the

observation at the largest background uncertainty location. With a single adaptive observation, the location that minimizes the background ensemble spread also minimizes the analysis spread. But if there are several observations, this is not valid (more details in Section 2.4).

2.3.2 Local analysis ensemble spread method

It is similar to the adaptive observation strategies proposed by Bishop et al. (2001) and Hamill and Snyder (2002) in explicitly minimizing the trace of the analysis error covariance, i.e., the summation of the analysis ensemble spread over all grid points. It differs from these methods in the calculation details and the parallel computation characteristics as discussed below.

In the LETKF (Hunt et al., 2007), the analysis error covariance can be expanded as:

$$\mathbf{P}^a = \mathbf{X}^b[(k-1)\mathbf{I} - (\mathbf{H}\mathbf{X}^b)^T \mathbf{R}^{-1}(\mathbf{H}\mathbf{X}^b)]^{-1}\mathbf{X}^{bT} \tag{2.2}$$

which depends on the background ensemble perturbations $\mathbf{X}^b$ (difference between ensemble forecasts and ensemble mean state), the observation location reflected in the observation operator $\mathbf{H}$, and the observation error covariance $\mathbf{R}$. $\mathbf{H}\mathbf{X}^b$ is the ensemble perturbation matrix at the observation space with the i^{th} column equal to $h(\mathbf{x}^{bi}) - h(\overline{\mathbf{x}}^b)$, where h is a nonlinear observation operator. The dimension of the inverse in the calculation of the analysis error covariance (Equation (2.2)) is the number of ensemble members, which are usually less than 100. Note that the

calculation of the analysis error covariance does not require the actual observation value, so it can be calculated before the observation values are known.

The special characteristic of the method we discuss here is the calculation efficiency resulting from parallel implementation, as in the LETKF data assimilation scheme itself. Like the localization scheme used in the LETKF assimilation scheme, the analysis error covariance can be calculated independently for each grid point based on the information within a local patch centered at that grid point. The average of the analysis ensemble spread of this analysis error covariance is regarded as the analysis ensemble spread of the center grid point. The final global analysis ensemble spread is the sum of the analysis ensemble spread at each grid point. The adaptive observation is the one that makes the global analysis ensemble spread smallest. Due to the independence of the analysis error covariance calculation in each local patch, the calculation is highly parallel, and could save a lot of computation time when dealing with large systems, such as realistic Observing System Simulation Experiments (OSSEs).

When more than one adaptive observation is to be chosen, the adaptive observation has to be selected serially, so that the impact from previous observations has already been taken into account before selecting the next adaptive observation. The process is as follows: the analysis ensemble perturbation (Equation (1.4)) based on the routine observations is calculated first, and regarded as the background ensemble perturbation in the first adaptive observation selection. Each potential

adaptive observation has a different observation operator, so each potential adaptive observation will get different analysis ensemble spread (Equation (2.2)). The one that makes the global analysis ensemble spread smallest is the first adaptive observation. After the first adaptive observation point is selected, the analysis ensemble perturbations are updated based on the new adaptive observation, and used as the background ensemble perturbations in the next adaptive observation selection. Since these processes are all highly parallel, different potential adaptive observations can be tested independently at the same time. This process repeats until all the adaptive observations are selected. In implementing on Lorenz 40 variable model, since we only need to select one adaptive observation, it is not necessary to use serial selection. We directly calculate the global analysis ensemble spread based on 20 possible adaptive observation locations. The adaptive observation is the observation that makes the magnitude of the analysis ensemble spread smallest.

2.3.3 Combined background-analysis ensemble spread method

Compared to the background ensemble spread method, the local analysis ensemble spread method has the advantage of considering the observation error, background covariance between grid points, and the impact from the observations that have already been chosen (discussed in more detail in Section 2.4), but it requires much more computational time even with parallel computations. The background ensemble spread method, on the other hand, considers only the background ensemble variance, and it is available at no cost within an ensemble Kalman filter. Therefore, we propose a method combining both methods by first choosing a small portion of the potential adaptive observation locations based on the background ensemble spread,

and then applying the local analysis ensemble spread method only on the grid points with the largest background ensemble spread. In this way, we combine the advantage of background ensemble spread method and local analysis ensemble spread method. We call this method as combined background-analysis ensemble spread method, abbreviated it as combined method. We expect that the combined method will show significant computational advantage when dealing with the whole atmosphere and at the same time, retain the optimality of local analysis ensemble spread method. In the implementation on Lorenz 40-variable model, five grid points with largest ensemble spread are first picked out from 20 grid points over ocean. Then, we only compare global analysis ensemble spread based on these five potential observation locations. The grid point that makes the expected global analysis uncertainty smallest is the adaptive observation point. It saves more than half of the computation time compared to local analysis ensemble spread method. In a global model, the advantage would be proportionally much larger.

2.3.4 Ideal method

In this method, we calculate the ensemble uncertainty using the true state, i.e., the ensemble spread is the difference between background ensemble and the true state, instead of the mean forecast state. The adaptive observation is at the location with largest true ensemble spread. In reality, it is impossible to know the true state of the atmosphere, so we call this method as 'ideal method'. The performance of this method sets an optimal unattainable benchmark for the other methods.

2.4 The relationship between background ensemble spread method and local analysis ensemble spread method

In the background ensemble spread method, we assume that the analysis error variance increases with the background error variance. By putting the adaptive observation at the location with largest background error variance, we indirectly minimize the analysis error variance. In local analysis ensemble spread method, we directly minimize the analysis error variance. Both methods try to minimize the analysis error variance, and both are related with the background error variance, so they must have some relationship. Here, we will use two simple examples to illustrate the relationship between background ensemble spread method and local analysis ensemble spread method.

Suppose we have three grid points, x_1, x_2 and x_3, whose error standard deviations are σ_1, σ_2 and σ_3, and the background error covariance $\mathbf{P}^b = \begin{pmatrix} \sigma_1^2 & \sigma_1\sigma_2 & \sigma_1\sigma_3 \\ \sigma_2\sigma_1 & \sigma_2^2 & \sigma_2\sigma_3 \\ \sigma_3\sigma_1 & \sigma_3\sigma_2 & \sigma_3^2 \end{pmatrix}$. We will select one adaptive observation from them based on the trace of the analysis error covariance. Suppose the adaptive observation is at the first grid point x_1 with error variance of r^2, then the observation operator is $\mathbf{H} = \begin{pmatrix} 1 & 0 & 0 \end{pmatrix}$. The Kalman gain matrix $\mathbf{K} = \dfrac{\mathbf{P}^b\mathbf{H}^T}{\mathbf{H}\mathbf{P}^b\mathbf{H}^T + \mathbf{R}}$, can be written

as $\begin{pmatrix} \frac{\sigma_1^2}{\sigma_1^2 + r^2} \\ \frac{\sigma_1\sigma_2}{\sigma_1^2 + r^2} \\ \frac{\sigma_1\sigma_3}{\sigma_1^2 + r^2} \end{pmatrix}$. The analysis error covariance $\mathbf{P}^a = (\mathbf{I} - \mathbf{KH})\mathbf{P}^b$, can be written as $\frac{1}{\sigma_1^2 + r^2}\begin{pmatrix} \sigma_1^2 r^2 & \sigma_1\sigma_2 r^2 & \sigma_1\sigma_3 r^2 \\ \sigma_2\sigma_1 r^2 & \sigma_2^2 r^2 & \sigma_2\sigma_3 r^2 \\ \sigma_3\sigma_1 r^2 & \sigma_3\sigma_2 r^2 & \sigma_3^2 r^2 \end{pmatrix}$. The trace of the analysis error covariance is

$$tr(\mathbf{P}^a) = \frac{(\sigma_1^2 + \sigma_2^2 + \sigma_3^2)r^2}{\sigma_1^2 + r^2} \tag{2.3}$$

$\sigma_1^2 + \sigma_2^2 + \sigma_3^2$ is the summation of the ensemble spread at all grid points, which is independent of the adaptive observation location. The denominator $(\sigma_1^2 + r^2)$ is the ensemble variance and observation error variance at the observation location, which depends on the adaptive observation location. Assuming that all the observations have the same error variance r^2, minimizing the analysis error variance is equivalent to maximizing the denominator, which means that the analysis error variance will be minimized when the observation is at the location with the largest forecast ensemble spread. Therefore, for a single adaptive observation, the background ensemble spread method is equivalent to the local analysis ensemble spread method, if the observation is of the same type as the model variable, collocated with a grid point and the observation error standard deviations are same for all the potential adaptive observation locations.

In the following example, we consider the case when only one adaptive observation is to be selected, but the adaptive observation is going to be placed in the middle of two grid points. There are a total of three grid points, and two potential adaptive observation locations. We can define the observation operator as $\mathbf{H} = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} & 0 \end{pmatrix}$. Following the same derivation as equation (2.3), the trace of the analysis error covariance is

$$tr(\mathbf{P}^a) = \frac{(\sigma_1^2 + \sigma_2^2 + \sigma_3^2)r^2}{0.25 \times (\sigma_1^2 + \sigma_2^2 + \sigma_3^2 + 2\sigma_1\sigma_2) + r^2} \tag{2.4}$$

To minimize the trace of analysis error covariance in local analysis ensemble spread method, it is again equivalent to maximize the denominator. However, in this case, it is not only dependent on the ensemble spread, but also on the background covariance $\sigma_1\sigma_2$. Even if the potential adaptive observation is assumed to be at a grid point, it can be related with more than one dynamical variable. In that case, the local analysis ensemble spread method is not equivalent with the background ensemble spread method anymore, since minimizing of the analysis ensemble spread requires not only the variance, but the covariance terms.

With more than one adaptive observation locations chosen, the local analysis ensemble spread method may give different result from background ensemble spread method since the background ensemble perturbations used in the calculation of $\mathbf{P}^a$ will be updated each time after a new adaptive observation is selected. Furthermore, the background ensemble spread method is very likely to pick adjacent grid points as

adaptive observations since grid points having large ensemble spread tend to be clustered together. On the other hand, the local analysis ensemble spread method will less likely to pick two adjacent grid points as adaptive observations since the updated uncertainty at the grid points around the adaptive observation will be mostly reduced. We will discuss more about how to deal with this problem in the background ensemble spread method in Chapter 3.

In summary, the background ensemble spread and the local analysis ensemble spread method are related to each other. Under some special conditions (a single adaptive observation of the same type as the dynamical variable and constant observation error variance), these two methods are equivalent. But in most cases, the local analysis ensemble spread method is more advanced, and the choice of adaptive observation location is more optimal than that from background ensemble spread method.

2.5 Results

2.5.1 Analysis RMS error comparison among different adaptive observation strategies

Figure 2.2 shows that local analysis ensemble spread method, background ensemble spread method, and combined method show similar performance over both ocean and land. Such result could be explained from the discussion in Section 2.4, because the observation error is assumed to be independent of location and there is a single adaptive observation. The small analysis RMS error differences among these methods may be due to the sampling error of the observations and to tiny differences

in the estimated inflation factors. Since the background ensemble spread method gives the same result as the more complicated local analysis ensemble spread method in our experimental setup, we only discuss the result from the background ensemble spread method here. With a single adaptive observation from background ensemble spread method, the analysis RMS error is greatly reduced compared to no observation, one random observation and one constant observation over ocean (Lorenz and Emanuel, 1998). The RMS error from ensemble spread method is only slightly larger than the 'ideal' method.

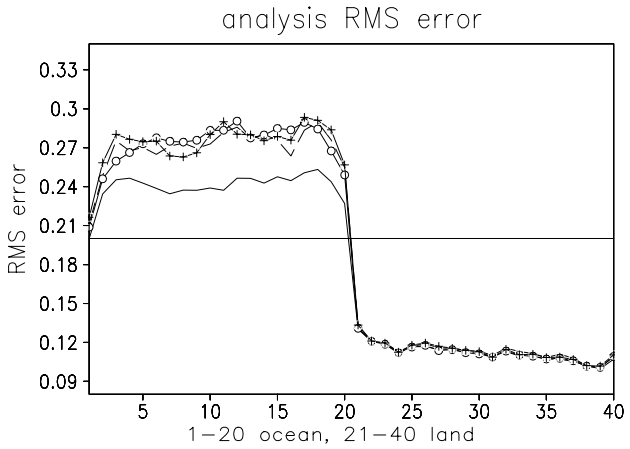


Figure 2.2 Five-year-average analysis RMS error for different adaptive observation strategies (the straight line is the observation error standard deviation; the solid line without marks: 'ideal' method, the dashed line: local analysis ensemble spread method, the solid line with open circles: background ensemble spread method, the solid line with cross: combined method.)

The analysis sensitivity (discussed in Chapter 4) with respect to that single adaptive observation is about 0.85 (Figure 2.3), which means that 85 % of the information of the analysis comes from the observation at the adaptive observation location. The analysis sensitivity with respect to the routine observation is only about 0.2, much smaller than that of the adaptive observation. The main reason is due to the difference of observation density between ocean and land. The sparser observation

distribution makes the adaptive observation more important. Whereas over land, the background is accurate and provides about 80% of the information. The result underlines the importance to have adaptive observations in vast unobserved areas.

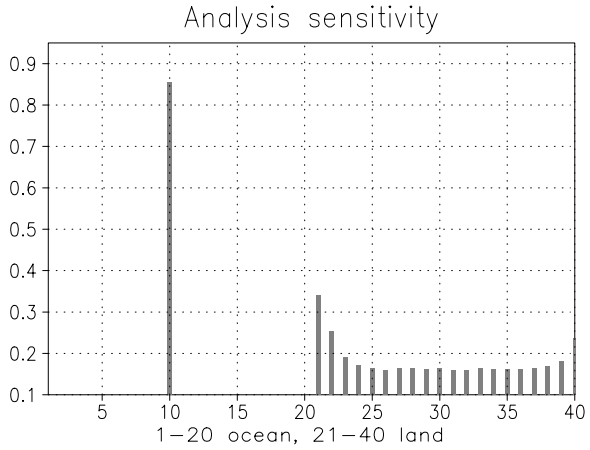


Figure 2.3 Analysis sensitivity with respect to both the routine observations over land and a single adaptive observation over ocean (we use 10th grid point to represent the adaptive observation locations).

2.5.2 10-day forecast RMS error

Figure 2.4 shows that it takes about one day for the forecast RMS error from background ensemble spread method to reach a level of 0.5 over ocean. This result is much better than the best result (Hansen and Smith, 2000) published with a similar experimental setup with this model. In Hansen and Smith (2002) (Figure 2.5), using the singular vector method and 1024-member ensemble Kalman filter, the forecast RMS error gets to 0.5 after only about 0.2 day, whereas it takes over one day to reach this level in the LETKF ensemble spread method.

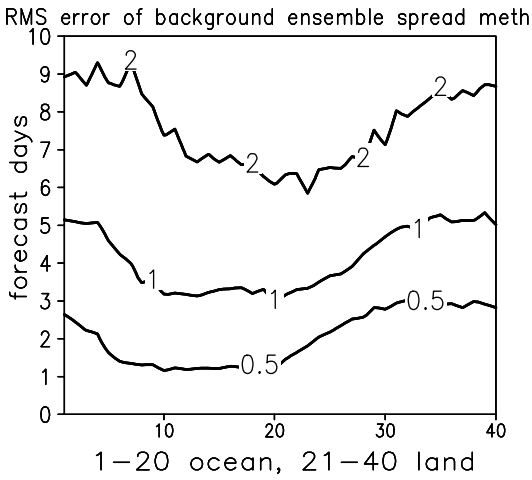


Figure 2.4 Five-year-average forecast errors from ensemble spread method.

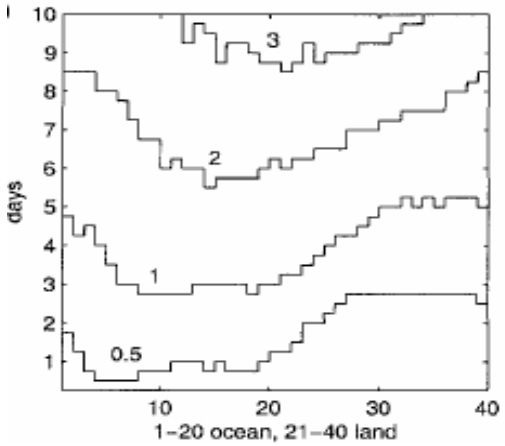


Figure 2.5 10-day forecast RMS error from Hansen and Smith (2000), singular vector adaptive observation strategy is used in this result.

2.6 Summary

In this chapter, we illustrated several ensemble-based adaptive observation strategies using the LETKF data assimilation scheme, namely, the background ensemble spread method, local analysis ensemble spread method, and combined background-analysis ensemble spread method. We also introduced one 'ideal' method which is used as the optimal benchmark for the other adaptive observation strategies. In the background ensemble spread method, the adaptive observation is at the

location with the largest background ensemble spread. It indirectly minimizes the analysis error variance. Local analysis ensemble spread method directly minimizes the analysis error variance, which can be computed in parallel. The combined method combines the advantages of both the background ensemble spread method and local analysis ensemble spread method, trying to utilize the free computation characteristics of background ensemble spread method and the consideration of the covariance, observation error and the determined observation locations in local analysis ensemble spread method. Using two simple examples, we have shown that the background ensemble spread method gives the same result as local analysis ensemble spread method when only one adaptive observation is to be selected from the grid point, and all the potential adaptive observations are the same type as the model variable and have the same accuracy. Otherwise, the result from these two methods is different.

Following the same experimental setup as Lorenz and Emanuel (1998), we show that the background ensemble spread method, local analysis ensemble spread method and combined method give the same result, only slightly worse than the 'ideal' method, and better than the best result published so far in the literature. The analysis sensitivity with respect to that single adaptive observation over ocean is much larger than that of the routine observations, which underlines the importance of having adaptive observation over vast unobserved region.

Chapter 3 Simplified Doppler Wind Lidar (DWL) adaptive observations in a primitive equations model (shorter version published in GRL, 2007)

3.1 Introduction

Within the next few years, the first Doppler Wind Lidar (DWL) will be deployed in space by the European Space Agency (ESA, see, http://www.congrex.nl/06c05/). In addition, in its recent Decadal Survey Report, the National Research Council recommended a US global winds mission in the coming decade. Because the operation of DWL is strongly constrained by energy resources (Rishojgaard and Atlas, 2004), a frequently stated qualitative goal is to get about 90% of the total effectiveness from just 10% coverage with adaptive observations. Here, 10% coverage means making measurements in only 10% of the total footprints that the DWL can possibly scan in a certain interval such as 6 hours. Unlike the applications of adaptive dropsonde observing in field experiments (FASTEX, NORPEX, Joly et al., 1997; Bergot, 1999; Langland et al., 1999a; Langland et al., 1999b; Pu and Kalnay, 1999; Szunyogh et al., 1999; Majumdar et al., 2002; Toth et al., 2002; Langland 2005), which attempt to optimize the 2-3 days forecast within a specified verification region (e.g, Europe, or North America), the goal in our study is to optimize the six-hour global analysis by optimally distributing the limited DWL observation resources. As pointed out by Lorenz and Emanuel (1998) and in Section 2.4, if a single adaptive observation is made at the locations with largest background uncertainty, the global analysis error will be most reduced as compared to other

locations. The question we address in this chapter is how to represent the background dynamical uncertainty and choose adaptive observation locations accordingly.

The Ensemble Kalman Filter (EnKF) (Evensen, 1994; Anderson, 2001; Houtekamer and Mitchell, 2001; Bishop et al., 2001; Whitaker and Hamill, 2002; Ott et al., 2004; Hunt et al., 2007), a relatively new data assimilation approach, provides an estimate of the background dynamical uncertainty. We call the diagonal value of an EnKF-computed background error covariance matrix for a given variable the ensemble spread for that variable. Locations with large ensemble spread are those in which dynamical instabilities of the evolving flow will result in large background (forecast) error and therefore where observations can be most useful, as discussed in the last chapter. The different observation location selection strategies that we compare are (a) one based on the LETKF ensemble spread, (b) a uniform observation distribution, (c) one based on the climatological background uncertainty, (d) random locations, and (e) an "ideal" strategy based on assumed knowledge of the true forecast error. We compare the impacts of adaptive observations selected with these different methods by assimilating them with two different data assimilation schemes, 3D-Var and Local Ensemble Transform Kalman Filter (LETKF). We test both 10% and 2% adaptive observations coverage, allowing for relatively dense and sparse adaptive observation scenarios. Comparison of these two scenarios will show the sensitivity of data assimilation schemes to the amount of adaptive observations.

This chapter is organized as follows: Section 3.2 describes the model, observations and data assimilation schemes we will use. Section 3.3 gives the detail of the adaptive observation strategies and the distributions of the simulated DWL observations. In Section 3.4, we will show the results from both 10% and 2% adaptive observation strategies assimilated by 3D-Var and the LETKF data assimilation schemes. Section 3.5 is the summary and conclusion.

3.2 Model, observation and data assimilation schemes

In this study, we use the Simplified Parameterizations, primitivE Equation DYnamics (SPEEDY) model, developed by Molteni (2003) and adapted for data assimilation by Miyoshi (2005). It has a simplified but complete set of physical processes, seven vertical levels, 96 longitudinal grid points, and 48 latitudinal grid points. We follow a "perfect model" Observing System Simulation Experiments (OSSEs) setup, in which the simulated "truth" (long model integration) is generated with the same atmospheric model as the one used in data assimilation. In such an "ideal twin" experimental setup, we avoid the complications of model error, and the only source of forecast errors comes from the initial conditions. Observations are obtained from the "truth" with added Gaussian random perturbations. The observation error standard deviations assumed for wind components (u, v), temperature (T), specific humidity (q) and surface pressure (p_s) are 1.0m/s, 1.0K, 0.1g/kg, and 1.0hPa, respectively.

To test the sensitivity of the impacts of adaptive observations to data assimilation methods, we use both 3D-Var (Parrish and Derber, 1998, Miyoshi, 2005)

and LETKF (Ott et al., 2004; Hunt et al., 2007). 3D-Var uses a constant background error covariance, which is calculated as in Parrish and Derber (1998). LETKF, a newly developed scheme belonging to EnKF family, employs the time evolving error covariance estimated from the forecast ensemble. It automatically gives the estimation of the forecast uncertainty. The application of LETKF on the SPEEDY model follows Hunt et al. (2007).

3.3 Adaptive strategies and the distribution of the simulated DWL observations

We mimic satellite tracks and DWL observations assuming that the satellite scans half hemisphere "orbits" in each six-hour analysis cycle. The basic observations (u, v, T, q, p_s) assimilated in all our experiments are simulated rawinsonde, shown as closed circles in Figure 3.1 (six-hour "orbits" are shown separated by vertical dashed lines). Figure 3.1 also shows an example of the 10% adaptive observation distribution (crosses) from the ensemble spread strategy (defined below) at 1200 UTC. At 0000 UTC, the satellite scans the same half hemisphere orbit as at 1200 UTC, and the other half hemisphere orbit is scanned at 0600 UTC and 1800 UTC. Thus, we assume that each grid point can be observed twice a day (this is too optimistic because we neglect the impact of clouds). Since the characteristics of the forecast uncertainties are different in different regions (e.g., Kalnay, 2003), the adaptive DWL observations are distributed into seven sub-regions, the equatorial region, the northern and southern tropics, and northern and southern mid- and high-latitudes (separated by horizontal dashed lines in Figure 3.1). Each sub-region is allotted a number of adaptive observations proportional to its area. The latitude ranges and the number of the adaptive observations in each sub-region are listed in Table 3.1. At the selected

adaptive DWL locations, both zonal wind and meridional wind are observed at all vertical levels. This is also over-optimistic because the lidar wind component that is actually observed is its projection on the line-of-sight direction (Stoffelen et al., 2005).

Table 3.1Adaptive observation distribution in seven latitude bands

Latitude range	53.81N-87.26N	31.55N-50.10N	9.28N-27.83N	5.57S-5.57N	24.12S-9.27S	50.10S-27.83S	87.26S-53.81S
Adaptive obs. #	22	33	35	52	35	33	22

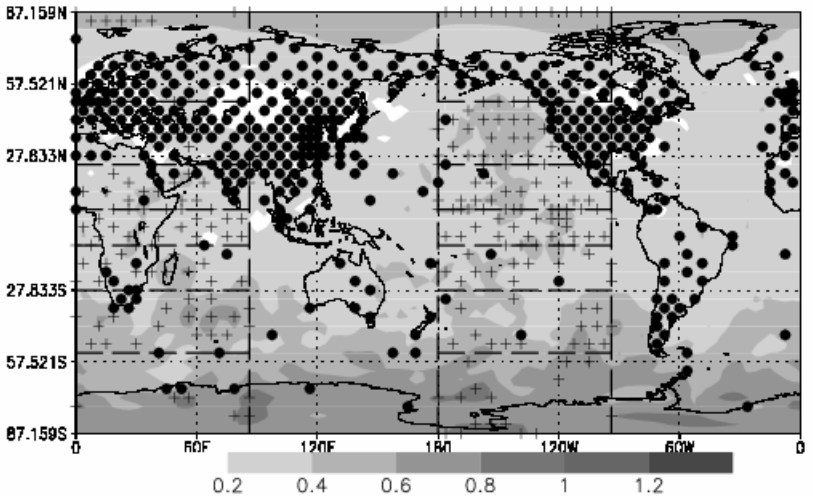


Figure 3.1 Example of the distribution of adaptive observations (crosses) from the ensemble spread sampling strategy at 1200 UTC February 03. The closed circles represent rawinsonde observation locations. Shades represent the average ensemble spread of zonal and meridional wind at 500hPa at that time. Horizontal dashed lines divide the whole globe into seven latitude bands. Vertical dashed lines separate the globe into four sub-regions representing two "orbits".

In all of the five adaptive observation strategies we tested, we impose a horizontal separation constraint to minimize possible observation redundancy, namely that the adaptive observations have to be at least two grid points apart in both

longitude and latitude directions. Hamill and Snyder (2002) account for observation redundancy by selecting the observations serially in minimizing the analysis error variance. However, directly minimizing the analysis error variance is much more expensive than computing ensemble spread and applying the separation constraint, especially when selecting adaptive observations from a very large pool of observation locations (Chapter 2). Moreover, by selecting adaptive observations at the locations with large ensemble spread in ensemble spread strategy, we approximately minimize the analysis error variance, as we discussed in Section 2.4. The separation constraint is done by first ordering the average six-hour forecast ensemble spread of wind at 500hPa from largest to smallest in each region. Within each region, the location with largest ensemble spread is selected as the first adaptive observation location. Then, we delete the locations adjacent to the first adaptive observation location in both zonal and meridional direction from the potential adaptive observation queue. The second adaptive observation location is where the ensemble spread in the remaining queue is largest. This process is repeated until all the adaptive observation locations are selected. If all the observations are either selected or deleted before the allotted number of adaptive observations are picked out, the remaining adaptive observations are the locations with largest ensemble spread that were deleted from the queue. A similar separation constraint is applied in all of the other strategies. In the climatological spread method, the climatological background ensemble spread is obtained from LETKF analyses of rawinsondes observations, and the adaptive observations are at the locations with largest climatological ensemble spread. In the ideal strategy, the adaptive observations are located where the background error (i.e.,

the absolute difference between six-hour forecasts of 500hPa wind and the true 500hPa wind field) is largest. Since this strategy requires knowing the "truth", it cannot be implemented in practice. The adaptive observation locations from ensemble spread, random location and the ideal strategy change with time, whereas the locations are fixed for uniform distribution and climatological ensemble spread strategies. In order to test whether the forecast ensemble spread truly represents forecast uncertainty, we use the same adaptive observation locations for both 3D-Var and LETKF in the ensemble spread and climatological ensemble spread strategies, even though they are both derived from LETKF assimilations.

3.4 Results

We examine the effectiveness of these five adaptive observation strategies by computing the analysis Root Mean Square (RMS) errors and comparing them to extremes of both 0% DWL coverage (i.e., rawinsondes only), and full (100%) DWL coverage. The percentage improvement for each strategy is defined as $PI = \frac{RMS - RMS^{0\%}}{RMS^{100\%} - RMS^{0\%}} \times 100\%$, where RMS is the time mean global average RMS error of the adaptive strategy, $RMS^{100\%}$ and $RMS^{0\%}$ are the time mean global average RMS error of full DWL coverage and no DWL coverage, respectively.

3.4.1 10% adaptive observation RMS error comparison different adaptive observation strategies

Figure 3.2 shows the time evolution of the 500hPa global averaged zonal wind analysis RMS errors for 3D-Var (left) and LETKF (right) with 0% coverage (dashed line) and 100% coverage (solid line), as well as the five adaptive strategies using 10%

coverage. The time averaged RMS error for the second month is presented in Table 3.2. Not surprisingly, the ideal strategy (dot dashed line) has the smallest errors, and is close to the 100% coverage. The LETKF-based ensemble spread strategy (solid line with open squares) is the best of the adaptive strategies that are feasible in practice, and is very close to the ideal strategy even for the 3D-Var analysis. The random location (solid line with crosses) is better than the uniform distribution strategy (solid line with closed circles). The worst results are obtained from the climatological ensemble spread distribution (solid line with triangles) because there are no adaptive observations over vast areas (not shown). The adaptive strategies with time-changing locations (ensemble spread, random location, ideal strategy) are all better than the constant observation distributions (uniform distribution, climatological ensemble spread), a conclusion consistent with previous results (Lorenz and Emanuel, 1998; Hamill and Snyder, 2002).

Table 3.2 500hPa time average (over February) of zonal wind global mean RMS errors and percentage improvement (PI) of 10% adaptive observations for both 3D-Var and LETKF.

Data assimilation	Experiment	Rawinsonde (0%)	Climatology (10%)	Uniform (10%)	Random (10%)	Spread (10%)	Ideal (10%)	100%
3D-Var	RMS error (m/s)	4.04	2.36	0.92	0.74	0.43	0.36	0.30
	PI	N/A	45%	83%	88%	97%	98%	N/A
LETKF	RMS error (m/s)	1.18	0.38	0.36	0.33	0.32	0.29	0.23
	PI	N/A	84%	84%	89%	91%	94%	N/A

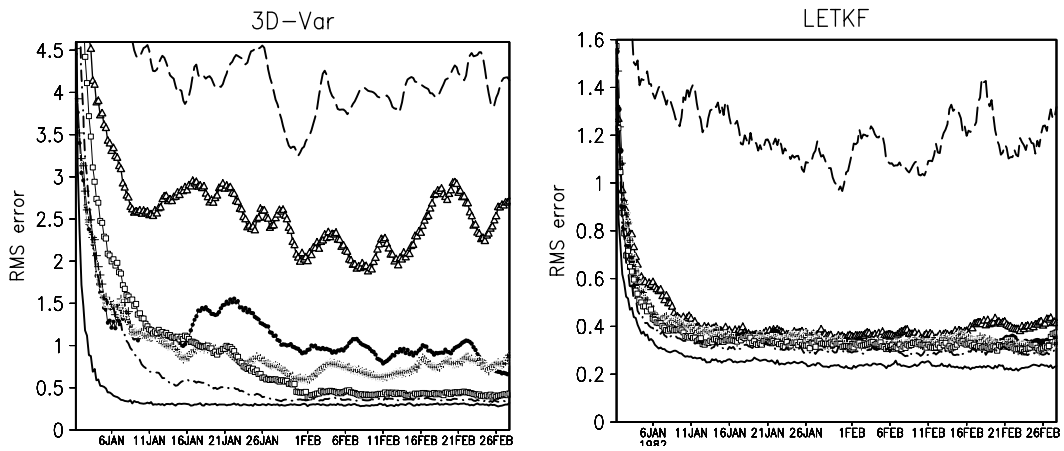


Figure 3.2 2-month evolution of 500hPa globally averaged zonal wind analysis RMS errors for 3D-Var (left panel) and LETKF (right panel) from 10% adaptive observations assimilation. From top to bottom their order is dashed line: rawinsonde observation (0% DWL) assimilation; solid line with triangles: climatological spread; solid line with closed circles: uniform distribution; solid line with crosses: random locations; solid line with open squares: ensemble spread adaptive strategy; dot dashed line: ideal sampling; solid line without marks: 100% adaptive observation coverage over half hemisphere.

Ensemble spread method and 'ideal' adaptive observation strategy are both based on the 500hPa statistics. To check the optimality of the adaptive observation over the other vertical levels, we further check the wind RMS error time evolution at 200hPa (Figure 3.3). Compared to 500hPa zonal wind RMS error time evolution, the 200hPa zonal wind shows a similar RMS error difference between different adaptive strategies for both 3D-Var and LETKF. The adaptive observations from ensemble spread method are as effective as in 500hPa. Not only in 200hPa, but in all other vertical levels, the ensemble spread adaptive observation is the most effective among all the operational possible sampling strategies (Figure 3.4). As shown more clearly in Figure 3.5, the RMS error percentage improvement from 10% ensemble spread based adaptive observation is more than 90% for 3D-Var, and more than 80% for LETKF.

The percentage improvement for 3D-Var is higher than that of LETKF because of the special characteristics of the data assimilation scheme itself. 3D-Var uses constant error covariance, so the analysis at a grid point is not accurate when there is no observation because of the poor estimation of the error correlation. On the other hand, the LETKF utilizes the time changing error covariance and better updates the analysis even where there is no observation. When the new observations are introduced in the 10% adaptive observation case, the LETKF analysis from the rawinsonde observation assimilation is already relatively accurate, so the impact of the new observations is not as significant as in 3D-Var.

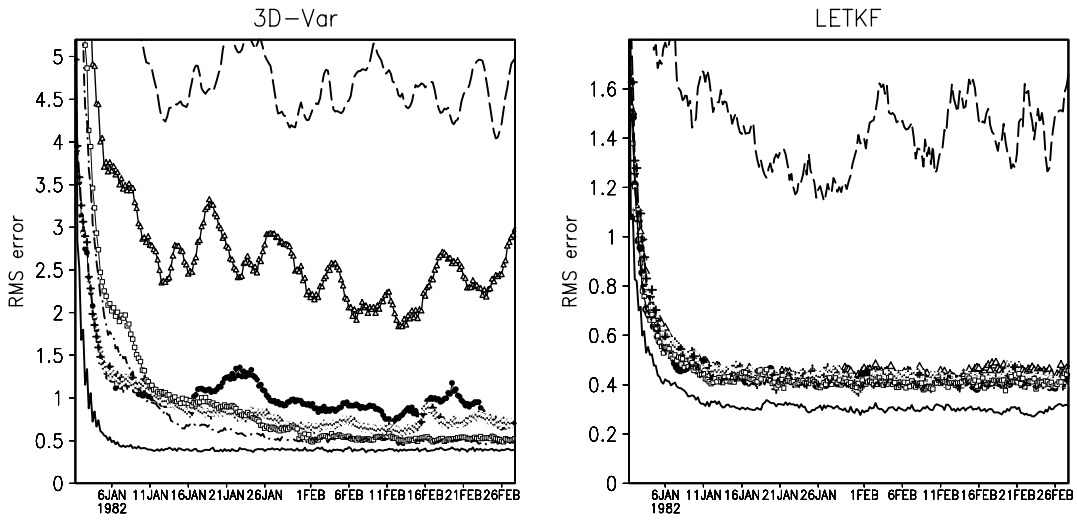


Figure 3.3 Same as Figure 3.2 except this is for 200hPa zonal wind RMS error (m/s) time evolution.

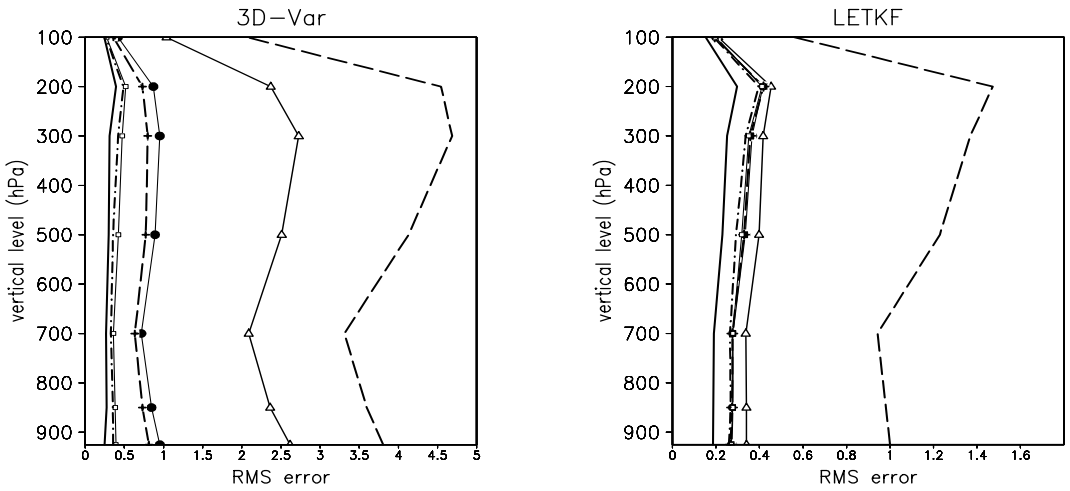


Figure 3.4 Time average (over the last half month analysis cycle) of zonal wind RMS error (m/s) over all the vertical levels for both 3D-Var (left panel) and LETKF (right panel) (Line notation is same with Figure 3.2)

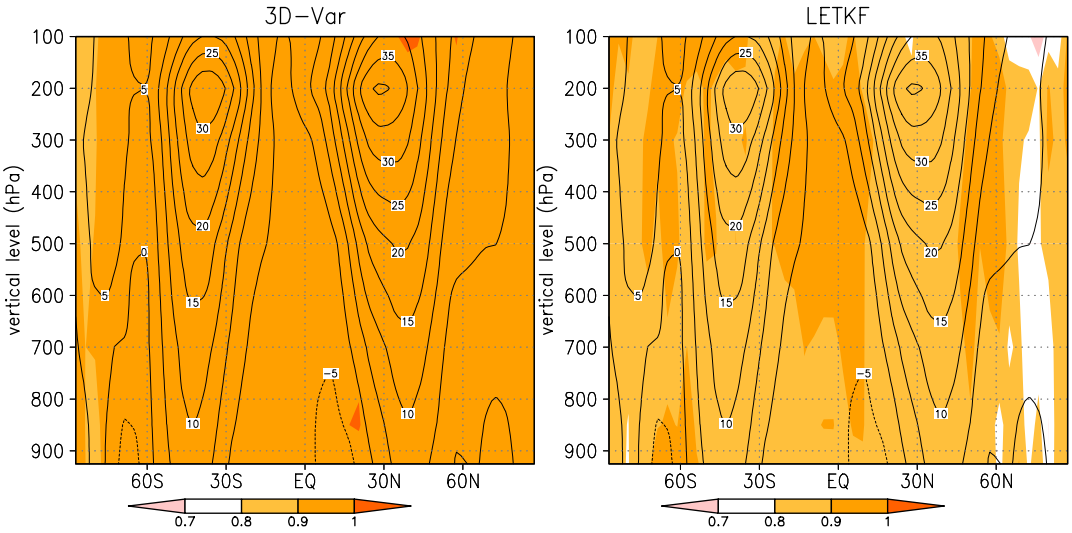


Figure 3.5 RMS error percentage improvement from 10% adaptive observations based on ensemble spread strategy (3D-Var: left panel; LETKF: right panel)

Through the covariance between winds and the other variables in background error covariance, the wind observations improve the analysis of the other variables as well, such as geopotential height (Figure 3.6). The different adaptive observation strategies have the same ranking as for the wind analysis.

The advantage of ensemble spread adaptive observation strategy persists with time (Figure 3.7). The ranking among these different sampling strategies also remains the same as in Figure 3.2. Since our experiments are based on a perfect model experimental setup, the improvement in the initial condition will persist with time. When there is model error involved, this may change and require further study.

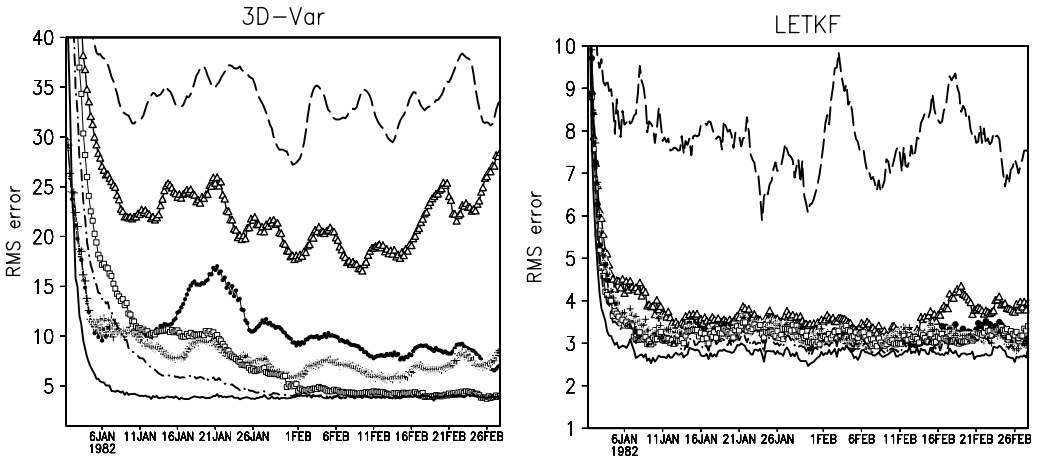
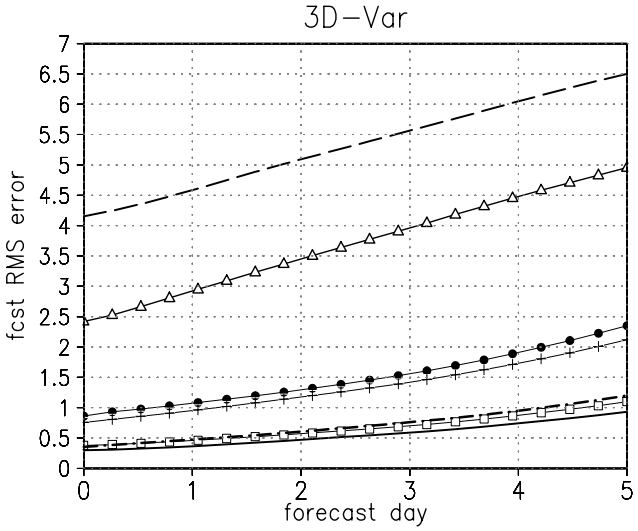


Figure 3.6 Same with Fig 4.2, except this is for 500hPa geopotential height (m).



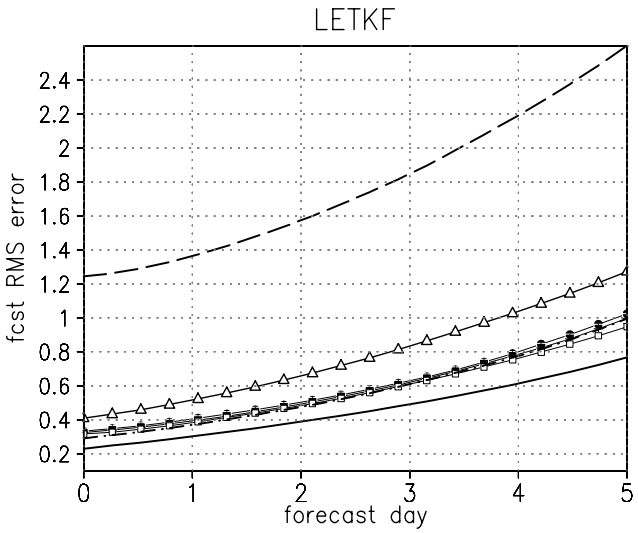


Figure 3.7 5-day forecast from different adaptive observation strategies for 3D-Var (top panel) and LETKF (bottom panel). (The line notation is same with Figure 3.2)

3.4.2 The comparison among adaptive observation locations from ensemble spread method, the background error and the analysis increment

A striking result is that the RMS error of LETKF (last section) shows a much smaller difference among the adaptive strategies than that of 3D-Var, although their relative ranking is the same. This is because 3D-Var, with a constant background error covariance, is much more sensitive to the choice of observations. With less optimal adaptive strategies, such as uniform distribution, the large background errors are not effectively reduced due to lack of observations around some locations with large background error (right panel in Figure 3.8). On the other hand, with the ensemble spread strategy, the adaptive observations are near the locations with large background errors (left panel in Figure 3.8). Therefore, the assimilation of these adaptive observations is equivalent to providing the information of the time-changing large background errors to 3D-Var. As a result, the analysis increments in 3D-Var have a shape more similar (but with opposite sign) to the background error (Figure

3.8, left panel) than in any other feasible method. By contrast, LETKF, whose background error covariance already includes information on the "errors of the day", is more efficient in extracting information from the observations even if their locations are not optimal, so that all the strategies give similarly small analysis errors. As shown in Figure 3.9, the analysis increment lines up, but have opposite sign with the background error even in the uniform observation distribution (right panel), though analysis increment and background error have a better agreement in ensemble spread sampling strategy (left panel).

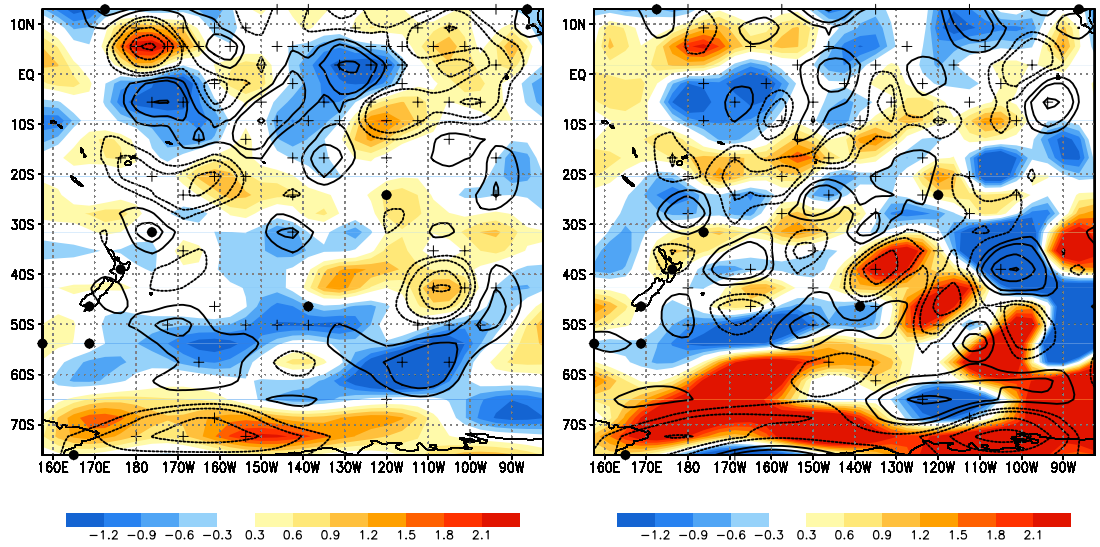


Figure 3.8 3D-Var zonal wind analysis increments (contour interval 0.3m/s), background error (shaded) and adaptive observation distribution (crosses) from the ensemble spread sampling strategy (left panel) and from uniform distribution (right panel) at 1200 UTC February 03. The closed circles are rawinsonde observation locations.

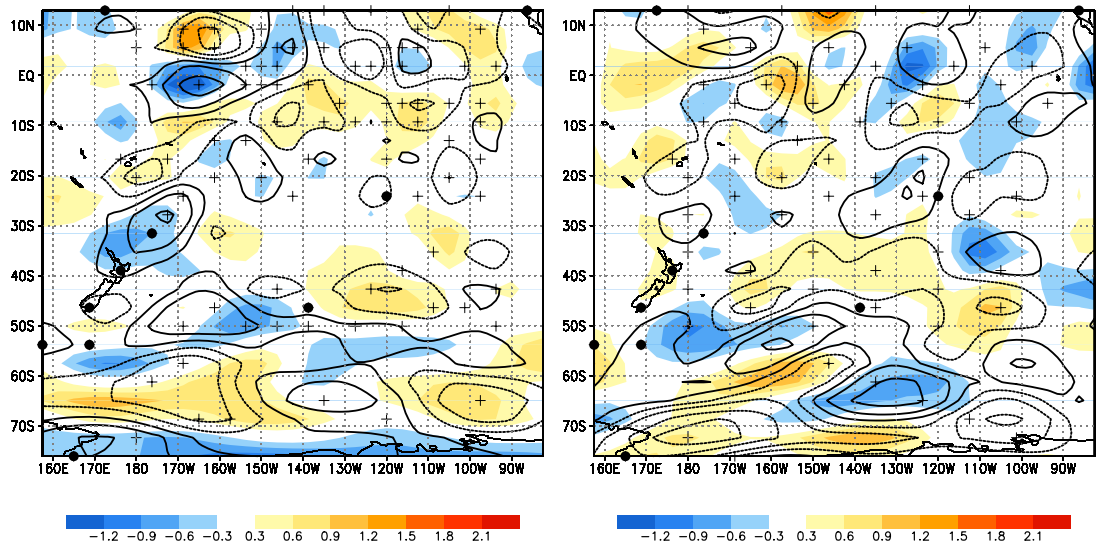


Figure 3.9 Same as Figure 3.8, except this is form LETKF data assimilation scheme.

3.4.3 2% adaptive observation RMS error comparison

It is clear from Figure 3.2 (left panel) and Table 3.2 that 3D-Var attains more than 90% of the improvements between 0% and 100% coverage from just 10% adaptive observations determined with the ensemble spread strategy. The percentage improvement of ensemble spread strategy in LETKF is somewhat smaller than for 3D-Var, and, as discussed above, all adaptive strategies are similarly successful (Table 3.2). This seems to contradict to the conclusions based on the previous adaptive observation field experiments that adaptive observations would be more effective with more advanced data assimilation schemes, such as 4D-Var or EnKF (Langland, 2005). However, we used relatively dense adaptive observation coverage in our experiments with 10% observed every 6 hours over half the globe. To make our results more compatible with previous field experiments, we now use the same adaptive observation strategies but substantially reduce the number of observation

locations to only 2% of the full coverage. With this small number of adaptive observations, the analysis errors of the adaptive strategies in 3D-Var are much larger, and even the most effective strategies, random location and ensemble spread, are only able to reduce the errors by less than 30% (left panel in Figure 3.10 and Table 3.3). By contrast, the LETKF still obtains 77% improvements from just 2% adaptive observations (right panel in Figure 3.10 and Table 3.3). The difference in performance among the five adaptive observation strategies is much more evident for LETKF, but with the same ranking as before. This result shows that with fewer adaptive observations, the data assimilation scheme plays a more important role in determining the effectiveness of adaptive observations. More advanced data assimilation schemes, such as the LETKF, use more efficiently small amounts of observation information, which is consistent with previous field experiments (Langland, 2005). The small number of observations is not enough to provide enough global information on the "errors of the day" needed for the improvement of 3D-Var, while in the LETKF, it is possible to estimate the evolving error structures even with few observations.

Table 3.3 500hPa time average (over February) of zonal wind global mean RMS errors and percentage improvement (PI) of 2% adaptive observations for both 3D-Var and LETKF.

Data assimilation	Experiment	Rawinsonde (0%)	Climatology (2%)	Uniform (2%)	Random (2%)	Spread (2%)	Ideal (2%)	100%
3D-Var	RMS error (m/s)	4.04	3.26	3.53	3.00	3.11	1.68	0.30
	PI	N/A	21%	14%	28%	25%	63%	N/A
LETKF	RMS error (m/s)	1.18	0.67	0.59	0.51	0.45	0.41	0.23
	PI	N/A	54%	62%	71%	77%	81%	N/A

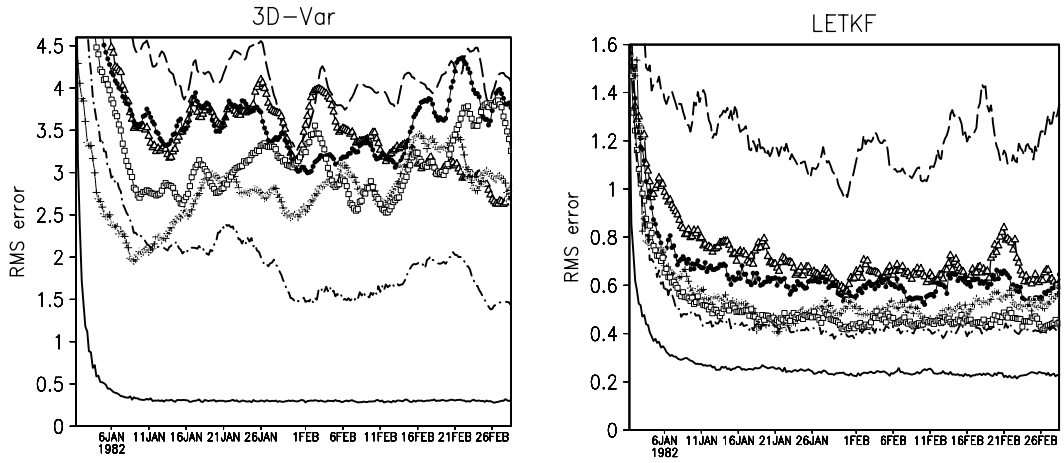


Figure 3.10 Same with Figure 3.2, except this is from 2% adaptive observation distribution.

3.5 Conclusion and discussion

In this chapter, we showed the potential of a simple ensemble spread strategy for adaptive observations in the context of minimizing the energy required by DWL laser firings. The same adaptive strategy could be used for any satellite instrument designed to "dwell" in regions of high uncertainty rather than providing uniform coverage along the orbit as conventionally done.

We compared ensemble spread with several other adaptive observation strategies (uniform distribution, random distribution, climatological ensemble spread) and found that the six-hour LETKF forecast ensemble spread gives a useful estimate of background uncertainty and dynamical instabilities. With 10% adaptive DWL observations, the ensemble spread sampling strategy gives the best result in both 3D-Var and LETKF, attaining more than 90% effectiveness of the full observation

coverage. 3D-Var is more sensitive to adaptive strategies than the LETKF. Since the latter already includes information on the "errors of the day", different adaptive strategies have closer performances.

We found that the sensitivity of adaptive observation effectiveness to data assimilation schemes is related to the amount of adaptive observations to be determined. With a relatively dense number of adaptive wind observations, such as 10% of the maximum coverage, 3D-Var can be as effective as LETKF, a more advanced data assimilation schemes. With only 2% coverage, 3D-Var is not as effective as LETKF even when using the LETKF ensemble spread locations.

Although our results are indicative of the potential for adaptive observations in remote sensing, we made several simplifying assumptions, using a perfect model scenario, a low resolution global model, an extreme simplification of satellite orbits and DWL observations, assuming uncorrelated Gaussian observation errors, and neglecting the effect of clouds. As a result, the actual percentage improvements from assimilating DWL adaptive observations may be overoptimistic. Experiments with state-of-the-art OSSE systems should be carried out to verify whether our results are valid in a more realistic setup. We believe that the main results, which states that the EnKF-based uncertainty estimation gives valuable guidance to allocate limited observation resources along the satellite track, and that the effectiveness of data assimilation schemes is sensitive to the amount of adaptive observations, would be valid even in a realistic experimental setup.

Chapter 4 : Analysis sensitivity calculation within an ensemble Kalman filter

4.1 Introduction

Modern atmospheric data assimilation systems (e.g., 3D-Var operational system in NCEP and 4D-Var operational system in ECMWF) usually include a high-dimension dynamical model with about 10^8 degrees of freedom, and assimilate the observations from both space and ground-based observation sources. In addition, operational centers frequently improve the model and introduce new observations into the data assimilation system. In such a complicated and continuously changing system, it is necessary to use some measures to monitor the influence of each factor on the performance of the system: how much information content does a new observation system have? How spatially different is the impact of the same type observations on the analysis? And what is the relative influence of the background and observation on the analysis?

Since 3D-Var, 4D-Var and Ensemble Kalman Filter (EnKF), the most commonly used data assimilation methods in both operational NWP centers and in research community, are special cases of least square problems (e.g., Kalnay, 2003), the diagnostic methods used for monitoring statistical multiple regression analyses can also be used to measure these data assimilation systems. The influence matrix is such a diagnostic whose element indicates the data influence on the regression fit of the analysis. Cardinali et al. (2004) proposed an approximate method to calculate the analysis sensitivity, which is the diagonal value of the influence matrix, within 4D-

Var data assimilation framework. They showed that the relative importance of different type observations based on the summation of analysis sensitivity was in good qualitative agreement with the observation impact from other studies.

In this chapter, based on Cardinali et al. (2004), we derive a method to calculate analysis sensitivity and the related diagnostics within the LETKF (Ott et al., 2004; Hunt et al., 2007), and study the properties and possible applications of these diagnostics. This chapter is organized as follows: the derivation is in Section 4.2. In Section 4.3, with a geometrical interpretation method adapted from Desroziers et al. (2005), we will show that the analysis sensitivity is proportional to the analysis accuracy and decreases with observation errors. In section 4.4, we verify the calculation method in Lorenz-40 model variable (Lorenz and Emanuel, 1998), and in Section 4.5, we use a primitive equation model to examine the effectiveness of the trace of analysis sensitivity in assessing the observation impact in the data assimilation.

4.2 Calculation of the influence matrix and analysis sensitivity within the LETKF

The LETKF, as explained in Chapter 1, combines background (n-dimension vector) and observations (p-dimension vector) based on the time changing weighting matrix $\mathbf{K}$. It can be expressed as:

$$\mathbf{x}_a = \mathbf{K}\mathbf{y} + (\mathbf{I}_n - \mathbf{K}\mathbf{H})\mathbf{x}_b \tag{4.1}$$

The vector $\mathbf{x}_a$ is the analysis. The gain matrix $\mathbf{K}(n \times p)$ considers the respective accuracies of background vector $\mathbf{x}_b$ and observation vector $\mathbf{y}$ by $\mathbf{P}^b$ and $\mathbf{R}$.

Following the derivation in Cardinali et al. (2004), we project the analysis into observation space, equation (4.1) becomes

$$\hat{\mathbf{y}} = \mathbf{H}\mathbf{x}_a = \mathbf{H}\mathbf{K}\mathbf{y} + (\mathbf{I}_p - \mathbf{H}\mathbf{K})\mathbf{H}\mathbf{x}_b \tag{4.2}$$

The analysis in observation space ($\hat{\mathbf{y}}$) is a linear combination of the observation vector ($\mathbf{y}$) and the background vector at observation space ($\mathbf{H}\mathbf{x}_b$). Then, the analysis sensitivity with respect to observations is:

$$\mathbf{S} = \frac{\partial \hat{\mathbf{y}}}{\partial \mathbf{y}} = \mathbf{K}^T\mathbf{H}^T = \mathbf{R}^{-1}\mathbf{H}\mathbf{P}^a\mathbf{H}^T, \tag{4.3}$$

and the sensitivity with respect to the background is given by

$$\frac{\partial \hat{\mathbf{y}}}{\partial \mathbf{H}\mathbf{x}_b} = \mathbf{I}_p - \mathbf{K}^T\mathbf{H}^T = \mathbf{I}_p - \mathbf{S} \tag{4.4}$$

where $\mathbf{P}^a$ is the analysis error covariance. The matrix $\mathbf{S}$ is called as the influence matrix (Cardinali et al., 2004), because the elements of the matrix reflect how much influence of the observations on the analysis. Similarly $\frac{\partial \hat{\mathbf{y}}}{\partial \mathbf{H}\mathbf{x}_b}$ reflects how much influence the background has on the analysis. The diagonal element of the matrix $\mathbf{S}$ is the analysis sensitivity, also called as self-sensitivity, which measures the sensitivity of the analysis at the observation location with respect to the corresponding observation. The sensitivity of the analysis with respect to the observation and with respect to the background is complementary (i.e., they add up to one) if the observation and the background are of the same type and at the same location. The Kalman gain is the ratio between background error covariance and the sum of observation error covariance and the background error covariance at the observation location, and the influence matrix is the adjoint of Kalman gain matrix in

observation space, so that the self-sensitivity has no units and its value is between zero and one.

In the variational data assimilation schemes, the Kalman gain and analysis error covariance are not explicitly calculated. However, in the LETKF, $\mathbf{P}^a\mathbf{H}^T\mathbf{R}^{-1}$ is explicitly calculated as:

$$\mathbf{P}^a\mathbf{H}^T\mathbf{R}^{-1} = \mathbf{X}^b[(\mathbf{H}\mathbf{X}^b)^T\mathbf{R}^{-1}(\mathbf{H}\mathbf{X}^b) + (K-1)\mathbf{I}]^{-1}(\mathbf{H}\mathbf{X}^b)^T\mathbf{R}^{-1} \tag{4.5}$$

where $\mathbf{X}^b$ is the background ensemble perturbation matrix with the i^{th} ensemble perturbation $\mathbf{X}^{bi} = \mathbf{x}^{bi} - \overline{\mathbf{x}}^b$, $\mathbf{x}^{bi}$ is the i^{th} ensemble forecast and $\overline{\mathbf{x}}^b$ is the mean background state. Since the influence matrix is a symmetric matrix, it can be written as

$$\mathbf{S}^T = \mathbf{H}\mathbf{X}^b[(\mathbf{H}\mathbf{X}^b)^T\mathbf{R}^{-1}(\mathbf{H}\mathbf{X}^b) + (K-1)\mathbf{I}]^{-1}(\mathbf{H}\mathbf{X}^b)^T\mathbf{R}^{-1} \tag{4.6}$$

Comparing equation (4.5) and equation (4.6), it is clear that the influence matrix $\mathbf{S}$ can be calculated in the LETKF by replacing the first element $\mathbf{X}^b$ in equation (4.5) with $\mathbf{H}\mathbf{X}^b$. It needs little additional computational time, and in addition, requires no approximations, which guarantees the self-sensitivity calculated in the LETKF satisfies the value limit (between zero and one). In 4D-Var (Cardinali et al., 2004), by contrast, the analysis error covariance is calculated from a truncated eigenvector expansion with vectors obtained through the Lanczos algorithm (Cardinali et al., 2004), which introduces some spurious values larger than one.

Equation (4.6) calculates the analysis sensitivity with respect to the observations, which can be calculated along with the LETKF. However, in the

LETKF, since each grid point is updated independently based on the observation and background information only within a local patch centered at that grid point (Ott et al., 2004; Hunt et al., 2007), each observation is used more than once during data assimilation. The self-sensitivity with respect to the same observation will be different in the different local patches. As a result, we propose to average the self-sensitivity with respect to the same observation in different local patches, and obtain the final self-sensitivity for that observation In Section 4.4, with Lorenz-40 variable model, we test the validity of this computation procedure in the LETKF by comparing it with the self-sensitivity calculated from a global ETKF (where each observation is only used once). Since in other versions of EnKF (Evensen, 1994; Anderson, 2001; Bishop et al., 2001; Houtekamer and Mitchell, 2001; Whitaker and Hamill, 2002), the Kalman gain is also explicitly calculated, it should be possible to calculate the influence matrix and the self-sensitivity in these schemes in a similar way.

Based on self-sensitivity, there are two other diagnostics which can show the characteristics of the analysis system. One is information content, which is the trace of self-sensitivity Tr(S), added for each subset of observations. It can be interpreted as a measure of the amount of information extracted from a particular set of observations. The other is relative information content. We define it as $\frac{tr(S_i)}{tr(S)}$, which is the ratio between the i^{th} type of observation information content and the information content of all the observations.

Since the larger the analysis sensitivity to the observation, the more important that observation is, the deletion of that observation will result in the larger change in the analysis value compared to the deletion of the other observations. Based on the assumption that the change of the analysis due to the assimilation of observations makes the analysis more accurate, which is true when the observation error statistics reflect the actual observation error, the deletion of an observation with larger analysis sensitivity will result in a worse analysis. Therefore, the analysis sensitivity can qualitatively reflect the change of analysis accuracy when part of the observations is denied from the assimilation without actually carrying out the data denial experiments. Based on the same assumption, we can evaluate the improvement of the analysis accuracy due to the addition of some observations without actually carrying out the "add-on" experiments.

We will test above arguments in Section 4.5 by comparing the trace of self-sensitivity $Tr(S_i)$ of type i^{th} observation with the actual analysis error change due to the deletion of that type observation, comparing the trace of self-sensitivity of the future possible observations with the actual observation impact from the assimilation of these observations in the system. When computing self-sensitivity, we assume the observation error statistics are accurate, i.e., they reflect the actual observation error standard deviation. However, in realistic assimilation cases, there may be some observations with larger observation error than assumed in the observation error statistics. We will discuss a method to detect such problem as in Langland and Baker

(2004) and show the actual quantitative observation impact on the forecast in next Chapter.

4.3 Geometric interpretation of the self-sensitivity

Equations (4.3) and (4.6) show that the analysis sensitivity is related with the background uncertainty, analysis uncertainty and observation error. In this section, we adapt the geometrical interpretation method of Desroziers et al. (2005) further to examine the relationship among the analysis sensitivity, the analysis accuracy and the observation accuracy in the space of eigenvectors $\mathbf{V}$ of the matrix $\mathbf{HK}$. Following the same notation as Desroziers et al. (2005), we rewrite the equation (4.2) by subtracting $\mathbf{H}(\mathbf{x}^t)$ on both sides of the equation,

$$\hat{\mathbf{y}} - \mathbf{H}(\mathbf{x}^t) = \mathbf{HK}(\mathbf{y} - \mathbf{H}(\mathbf{x}^t)) + (\mathbf{I}_p - \mathbf{HK})(\mathbf{H}(\mathbf{x}^b) - \mathbf{H}(\mathbf{x}^t)) \tag{4.7}$$

where $\mathbf{H}(\mathbf{x}^t)$ is the true state at the observation space. We define $\delta\hat{\mathbf{y}} = \hat{\mathbf{y}} - \mathbf{H}(\mathbf{x}^t)$, $\delta\mathbf{y} = \mathbf{y} - \mathbf{H}(\mathbf{x}^t)$ and $\mathbf{H}(\delta\mathbf{x}^b) = \mathbf{H}(\mathbf{x}^b) - \mathbf{H}(\mathbf{x}^t)$ respectively. Equation (4.7) can be written as,

$$\delta\hat{\mathbf{y}} = \mathbf{HK}\delta\mathbf{y} + (\mathbf{I}_p - \mathbf{HK})\mathbf{H}(\delta\mathbf{x}^b) \tag{4.8}$$

After eigenvalue decomposition, $\mathbf{HK} = \mathbf{V}\Lambda\ \mathbf{V}^T$, where Λ is the diagonal matrix of the eigenvalues of $\mathbf{HK}$,

$$\delta\hat{\mathbf{y}} = \mathbf{V}\Lambda\mathbf{V}^T\delta\mathbf{y} + \mathbf{V}(\mathbf{I}_p - \Lambda)\mathbf{V}^T\mathbf{H}(\delta\mathbf{x}^b) \tag{4.9}$$

Projecting $\delta\hat{\mathbf{y}}$ onto the eigenvector space, which is given by $\mathbf{V}^T\delta\hat{\mathbf{y}}$, above equation is written as,

$$\delta\vec{\hat{\mathbf{y}}} = \Lambda\delta\vec{\mathbf{y}} + (\mathbf{I}_p - \Lambda)\delta\vec{\mathbf{x}}^b \tag{4.10}$$

where $\delta\vec{\hat{\mathbf{y}}}$, $\delta\vec{\mathbf{y}}$ and $\delta\vec{\mathbf{x}}^b$ are the projections of $\delta\hat{\mathbf{y}}$, $\delta\mathbf{y}$ and $\mathbf{H}(\delta\vec{\mathbf{x}}^b)$ onto the eigenvector ($\mathbf{V}$) space. $\delta\vec{\hat{\mathbf{y}}}$, $\delta\vec{\mathbf{y}}$ and $\delta\vec{\mathbf{x}}^b$ are the analysis error, observation error and the background error at the eigenvector ($\mathbf{V}$) space respectively. When these vectors are projected onto a particular eigenvector $\mathbf{V}_i$ with corresponding eigenvalue equal to λ_i, the above equation is written as,

$$\delta\vec{\hat{\mathbf{y}}}_i = \lambda_i \delta\vec{\mathbf{y}} + (1-\lambda_i)\delta\vec{\mathbf{x}}_i^b \tag{4.11}$$

Therefore, in the space of eigenvector $\mathbf{V}_i$, the analysis sensitivity with respect to the observation is λ_i , and with respect to the background is (1- λ_i). They are complementary, which means that the more sensitivity of the analysis to the observations, the less sensitivity to the background.

Schematically, all the elements in equation (4.11) are shown in Figure 4.1. Following Desroziers et al. (2005), we define $\vec{d}_{bi}^{\,o}$ as the observation increment (the difference between observation and the background in the observation space), which is the line connecting the observation and the background. The angle between $\vec{d}_{bi}^{\,o}$ and $\delta\vec{\mathbf{y}}$ is α . The observation error ($\delta\vec{\mathbf{y}}$) and the background error ($\delta\vec{\mathbf{x}}^b$) in the eigenvector $\mathbf{V}_i$ space are perpendicular, which means that they are not correlated. The analysis error ($\delta\vec{\hat{\mathbf{y}}}$) is also perpendicular to the line connecting the observation and the background (Desroziers et al., 2005). Therefore, the projection of $\delta\vec{\hat{\mathbf{y}}}$ onto $\delta\vec{\hat{\mathbf{y}}}$ is,

$$\delta\vec{\hat{\mathbf{y}}} \cdot \delta\vec{\mathbf{y}}_i = \lambda_i \delta\vec{\mathbf{y}}_i \cdot \delta\vec{\mathbf{y}}_i \tag{4.12}$$

$$\left\|\delta\vec{\hat{\mathbf{y}}}_i\right\| \left\|\delta\vec{\mathbf{y}}_i\right\| \cos(90° - \alpha) = \lambda_i \left\|\delta\vec{\mathbf{y}}_i\right\|^2 \tag{4.13}$$

and $\lambda_i = \left\|\delta\vec{\hat{\mathbf{y}}}_i\right\| \left\|\delta\vec{\mathbf{y}}_i\right\| \cos(90° - \alpha) / \left\|\delta\vec{\mathbf{y}}_i\right\|^2$. Since $\sin(\alpha) = \left\|\delta\vec{\hat{\mathbf{y}}}_i\right\| / \left\|\delta\vec{\mathbf{y}}_i\right\|$, $\lambda_i = \sin^2(\alpha)$, and $(1-\lambda_i) = \cos^2(\alpha)$. The smaller the angle α is, the smaller is the analysis sensitivity with respect to the observations, and the larger analysis sensitivity to the background.

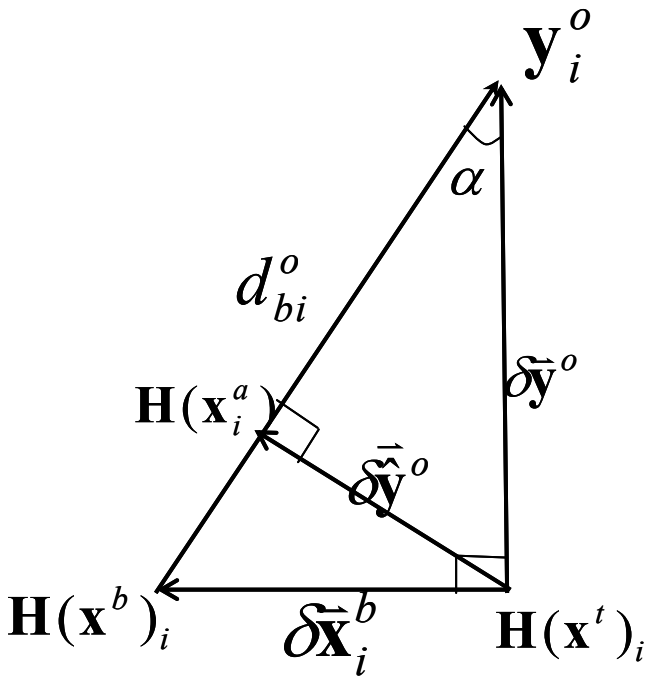


Figure 4.1 Geometrical representation of the elements in equation (4.11) (each element is explained in the text). The analysis sensitivity with respect to the observations is $\sin^2\alpha$ (after Desroziers et al., 2005).

From this geometrical representation, we can conclude that the analysis sensitivity per observation, $\sin^2(\alpha)$, is related to the observation error, analysis error and the background error. With constant observation error, the analysis sensitivity per

observation is proportional to the analysis error. With the analysis error unchanged, the analysis sensitivity per observation decreases with the size of the observation error. These properties are related with the adjustment of Kalman gain in the data assimilation system. When the observation error is larger, the analysis system gives less weight to the observation by changing the Kalman gain matrix. When the analysis is very accurate, the analysis system gives more weight to the background, so the self-sensitivity with respect to that observation is smaller. Therefore, the analysis sensitivity reflects the characteristics of the analysis system, reflecting the importance of the observation and the background. **However, these conclusions are only valid when the statistics used in the data assimilation approximately reflect the true background and observation error.**

4.4 Validation of the self-sensitivity calculation method with Lorenz 40-variable model

4.4.1 Lorenz-40 variable model and experimental setup

As in Chapter 2, we use the same parameter setup of Lorenz-40 variable model (Lorenz and Emanuel, 1998) (Equation (2.1)) to test the calculation procedure of self-sensitivity within the LETKF data assimilation scheme.

Since the self-sensitivity based on equation (4.6) is valid by itself, and the peculiar characteristic of our proposed procedure is the averaging scheme used in the LETKF, we test this procedure by comparing the self-sensitivity calculated in the global ETKF without averaging with that of LETKF. We carry out this comparison in the case of several uniform observation coverage scenarios, namely 10, 20, 30, and 40

observations. The experiments with different observation coverage allow us to examine the relationship among the analysis sensitivity per observation, the observation coverage, and the analysis accuracy. The local patch size is 39 in LETKF, i.e., 19 grid points on each side of the central grid point. With such large local patch size, it is equivalent to performing an ETKF on each grid point separately. Therefore, the accuracy of LETKF and ETKF should be similar, and so should be the self-sensitivity. With such large local patch in LETKF, and global ETKF, we use 40 ensemble members in both LETKF and ETKF to avoid filter divergence. The assimilation interval is every six-hour, which is equivalent to 0.05 output time interval in the model. We run each experiment for 7560 analysis cycles, and the time average is over the last 6560 analysis cycles.

4.4.2 Results

Figure 4.2 shows that the averaged self-sensitivity calculated from LETKF (closed circles) is almost identical with the self-sensitivity calculated from ETKF (plus signs), which indicates that the averaging scheme we used to calculate the self-sensitivity in LETKF is valid. The self-sensitivity increases with the increasing of the analysis RMS error, which is consistent with the geometrical interpretation in the Section 4.3. Since the analysis error is anti-correlated with the observation coverage, so is the self-sensitivity (Figure 4.2). The analysis sensitivity per observation becomes larger when the observation coverage becomes sparser. The analysis sensitivity per observation is about 0.28 when all the grid points are observed, which indicates that 28% of the information of the analysis comes from the observation at each location. Since the analysis sensitivity with respect to the background is complementary to the

analysis sensitivity with respect to the observation (Section 4.3), 72% information of the analysis comes from the background in this observation coverage scenario. When only 10 grid points have observations, about 53% information of the analysis comes from the observation at the observation locations, which indicates that deletion of one observation in dense observation coverage will do less harm to the analysis system than deletion of one observation in a sparse coverage case, which is consistent with field experiments (e.g., Kelly et al., 2007).

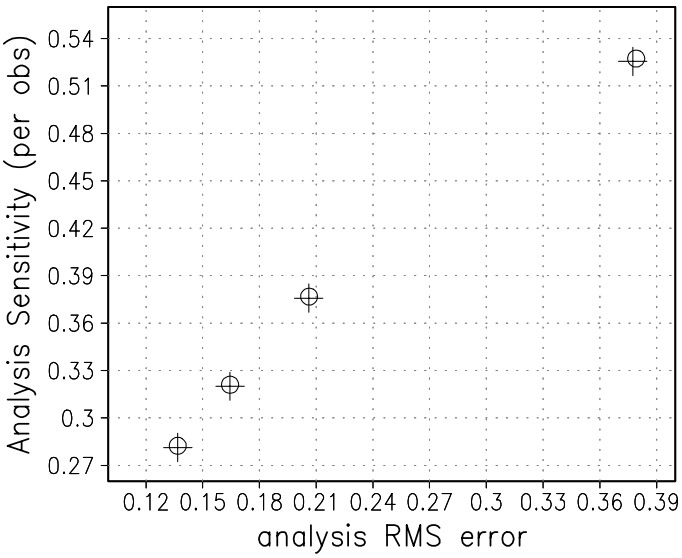


Figure 4.2 The scatter plot of the time averaged analysis sensitivity per observation (y-axis) and the analysis RMS error (x-axis) for the LETKF (open circles) and the ETKF (plus signs) with different observation coverage (from bottom to the top, the points correspond to 40 observations, 30 observations, 20 observation, and 10 observations).

4.5 Results with an idealized simplified primitive equation model

The results in the last section verify the validity of our proposed method to calculate self-sensitivity within the LETKF framework, and show that the self-

sensitivity is proportional to the analysis RMS error when the statistics used in the data assimilation is accurate. In section 4.2, we argued that the trace of the self-sensitivity of a subset of observations can qualitatively indicate the improvement of the analysis accuracy due to the assimilation of these observations. In this section, we will explore the validity of these arguments by comparing the trace of the self-sensitivity from a control experiment and the actual observation impact in data denial experiments, and comparing the trace of self-sensitivity of potential possible observations with the actual observation impact from "add-on" experiments, in which new observations are added.

4.5.1 Experimental setup

We use the Simplified Parameterizations primitivE Equation DYnamics (SPEEDY, Molteni, 2003) model that has been used in Chapter 3. As in Chapter 3, we follow a "perfect model" Observing System Simulation Experiments (OSSEs) setup, in which the simulated truth is generated with the same atmospheric model as the one used in data assimilation. Observations are the truth with added Gaussian random perturbations. The observation error standard deviations assumed for winds and specific humidity is about 30% natural variability of each dynamical variable, shown in Figure 4.3. The specific humidity is only observed in the lowest five vertical levels, which corresponds to the level below 300hPa. Since temperature variability does not change much with vertical levels, we assume the observation error standard deviation is 0.8K in all vertical levels. The error standard deviation for surface pressure is 1.0hPa.

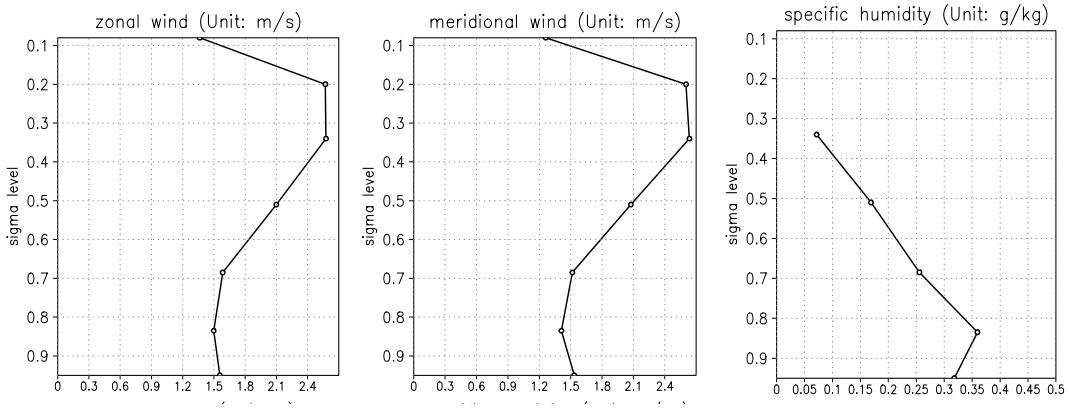


Figure 4.3 The observation error standard deviation for zonal wind (Unit: m/s, left panel), meridional wind (Unit: m/s, middle panel) and specific humidity (Unit: g/kg, right panel).

We carry out both data denial experiments and "add-on" experiments. In the data denial experimental setup, the control experiment is called as *all-obs* experiment, in which the observations are full coverage (Figure 4.4). In each observation location, all the dynamical variables are observed. In the sensitivity experiments, part of the dynamical variables are denied from the locations with red plus signs in Figure 4.4 , and only observed in the rawinsonde locations (closed circles in Figure 4.4). For instance, in the *no-u* sensitivity experiment, zonal wind observations are not observed in the locations with red plus signs, and only observed in the rawinsonde locations. We carry out two other sensitivity experiments, *no-T*, and *no-q*, in which temperature and specific humidity are not observed in the locations with red plus signs. We will compare the trace of self-sensitivity over the locations with red plus signs calculated in the *all-obs* run with the analysis error difference between the data denial experiment (e.g., *no-u*) and *all-obs.* For example, we compare the trace of zonal wind self-sensitivity over the observation locations with red plus signs calculated from *all-obs* experiment with the analysis error difference between *no-u* and *all-obs*

experiment. Ideally, the larger the trace of the self-sensitivity, the larger is the error difference between *no-u* and *all-obs* experiment.

In the "add-on" experimental setup, the control experiment is called *raob-only*, in which only the observations at the rawinsonde locations (closed circles in Figure 4.4) are assimilated. In the sensitivity experiment, we add one type of dynamical variable observed in the locations with red plus signs to the control observation network. For example, in the *add-u* experiment, the zonal wind observations are assimilated in both rawinsonde locations and the locations with red plus signs, and the other variables are only available at rawinsonde observations. The trace of these future possible observations calculated along the control run will be compared to the analysis error difference between the sensitivity experiment and the control experiment. For example, the trace of the zonal wind observation over the locations with red plus signs calculated along with the *raob-only* experiment will be compared with the analysis error difference between *raob-only* and *add-u* experiment. In the "add-on" experiments, since each potential set of observations has a different observation operator, the analysis error covariance has to be recalculated before calculating the self-sensitivity based on equation (4.6). However, the self-sensitivity can also be calculated after finishing *raob-only* experiment. In this way, the self-sensitivity of different possible additional types of observations can be calculated in parallel based on the background ensemble forecasts from *raob-only* experiment, which can save the computational time. The self-sensitivity calculated in "add-on" type experiments provides an estimate of the usefulness of potential future observations, while the self-

sensitivity calculated in data-denial type experiments can evaluate the analysis sensitivity to the existing observations.

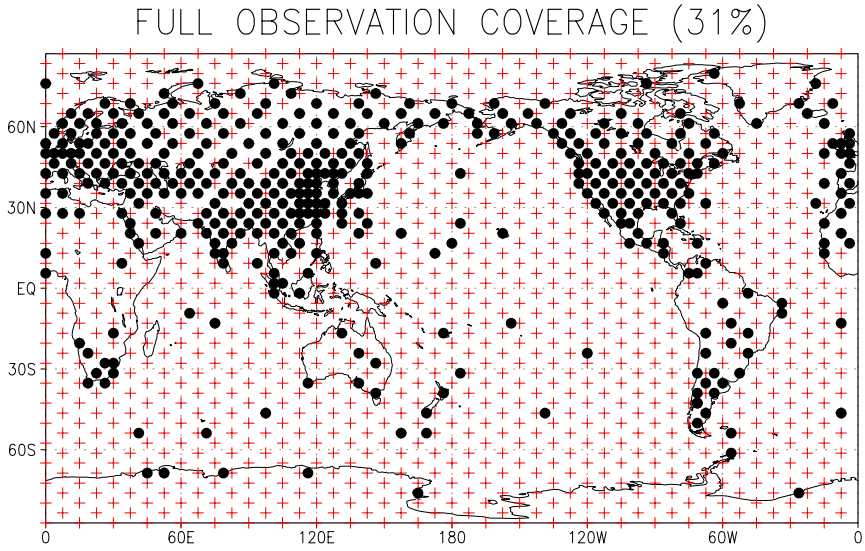


Figure 4.4 Full observation distribution (closed dots: rawinsonde observation network; red plus signs: dense observation network), each observation location is at the grid point.

4.5.2 Comparison between information content (abbreviated as InC) and the actual observation impact from the data denial experiments

In this sub-section, we will compare the information content (trace of self-sensitivity) calculated along with *all-obs* experiment and the actual observation impact given through the traditional data denial experiments, and examine whether the information content can qualitatively show the observation impact without carrying out the data denial experiments.

The left panel of Figure 4.5 shows the zonal mean zonal wind analysis RMS error difference (contours) between *no-u* and *all-obs* experiment and the information

content (shaded) of zonal wind over the locations with red plus signs calculated along *all-obs* experiment. The information content is the trace of zonal wind self-sensitivity at the locations with red plus signs in each latitude circle, which reflects the information extracted from the dense zonal wind observations at that latitude circle. The right panel of Figure 4.5 is the temperature analysis RMS error difference (contour) between *no-T* and *all-obs* and information content (shaded) of temperature at the locations with red plus signs calculated from *all-obs* experiment. Quantitatively, the analysis RMS error difference (contour) between *all-obs* and *no-u* experiment have the largest value over the tropics, and have smallest value over the mid-latitude Northern Hemisphere (NH). Qualitatively, the information content distribution agrees with the RMS error difference, also showing the largest values over the tropics and smallest values in the mid-latitude of the NH. Interestingly, the zonal wind observations have relatively small impact over the mid-latitude in the Southern Hemisphere (SH), even though the rawinsonde coverage is sparse over that region. The reason lies in the fact that the mass field, such as temperature and surface pressure, updates the zonal wind analysis in the mid-latitude of the SH through geostrophic balance in *no-u* experiment. The information content basically reflects this feature, showing relatively small values over that region.

For the temperature sensitivity experiment (right panel in Figure 4.5), the largest RMS error difference between *no-T* and *all-obs* experiment are over the high latitudes, and the spatial distribution of the information content agrees well with the RMS error difference in this region. In the upper level of tropics, however, the

information content has a large value center, and the RMS error difference between *no-T* and *all-obs* experiment is relatively smaller in that region, which is due to the strong multivariate update of the temperature field by the other observations in *no-T* experiment. The multivariate influence is shown more clearly in Figure 4.6.

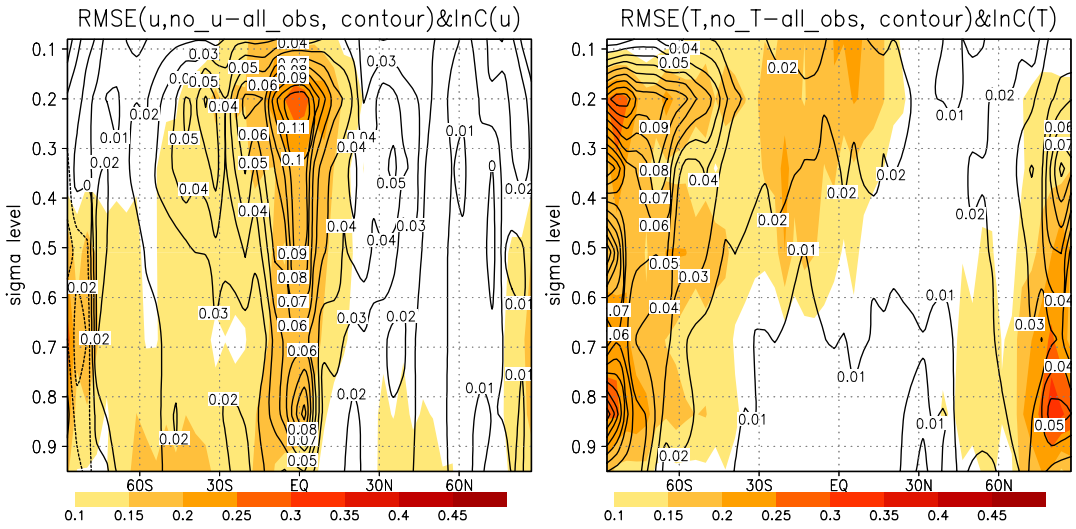


Figure 4.5 RMS error difference (contour) between sensitivity experiment and control experiment, and information content (shaded) (Left panel: between *no-u* and *all-obs*, zonal wind RMS error (Unit : m/s), zonal wind information content; right panel: between *no-T* and *all-obs*, temperature RMS error difference (Unit: K), temperature information content)

Figure 4.6 shows that the specific humidity has the largest information content (shaded area) in the tropics, so does the specific humidity RMS error difference between *no-q* and *all-obs* experiment. The information content is largest over high levels, which is due to relatively small assigned observation error in that region (Figure 4.3). Though specific humidity has smaller absolute value over high levels than that of lower levels, it still has relatively large RMS error difference over high levels where information content is largest. It is important to note that the information content of specific humidity reflects not only the impact of the deletion of specific

humidity observations on the humidity analysis, but also the impact on the other dynamical variables, such as zonal wind, which originates from multivariate characteristics in *all-obs* experiment. The specific humidity observation linearly affects winds through the covariance in the data assimilation process, and this effect is maximized in the tropical upper troposphere (right panel in Figure 4.6) (see also Chapter 6).

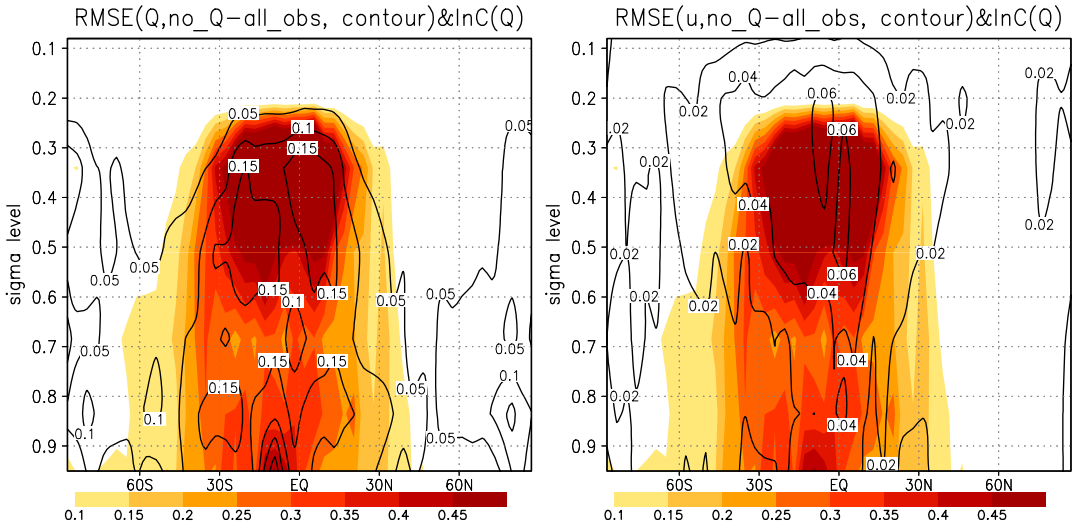


Figure 4.6 RMS error difference (contour) between *no-q* and *all-obs* experiment, and specific humidity information content (shaded) (Left panel: specific humidity RMS error difference (Unit: 10^{-1}g/kg); right panel: winds RMS error difference (Unit: m/s))

The qualitative consistency between the information content calculated from control experiment and the actual observation impact from data denial experiments verifies that we can qualitatively examine the observation impact on the analysis without carrying out data denial experiments when the error statistics used in the data assimilation system is accurate.

4.5.3 The results from "add-on" experiments

In the data-denial experiments, we try to examine the impact of the assimilated observations on the analysis. In the "add-on" scenario, we want to evaluate the impact of future possible observations on the analysis. Traditionally, this is done with OSSE's by actually adding the simulated observations into the data assimilation, and examining the error difference between the control experiment and the add-on sensitivity experiment. With self-sensitivity, we can qualitatively estimate the observation impact without actually knowing the observation value. In this sub-section, we will verify this argument by comparing the information content with the actual observation impact from the "add-on" experiments.

The left panel in Figure 4.7 shows that the information content calculated along with the *raob-only* experiment based on future dense zonal wind observations reflects the actual observation impact from the "add-on" experiments. Note, as stated before, the self-sensitivity can also be calculated after finishing *raob-only* experiment. This assumes that the background error will not change in the "add-on" experiments, so the information content is only an approximation of actual observation impact from adding these observations which reduces the background error covariance in the presence of additional observations. The larger value center of information content is collocated with the larger error difference center. The same is true for the temperature (right panel in Figure 4.7). Not surprisingly, the addition of dense observations into the rawinsonde observation network improves the analysis mostly in the tropics and the SH where there are not much rawinsonde observations. This also verifies that the

information content qualitatively gives the observation impact of the future observations, and can be used in observation network designs. When the possible observations are specific humidity, the information content has some problem reflecting the actual impact in the higher level tropics (Figure 4.8), which may be due to the nonlinearity of the humidity field. However, in the lower levels, the information content agrees well with the analysis error difference between *raob-only* experiment and *add-q* experiment.

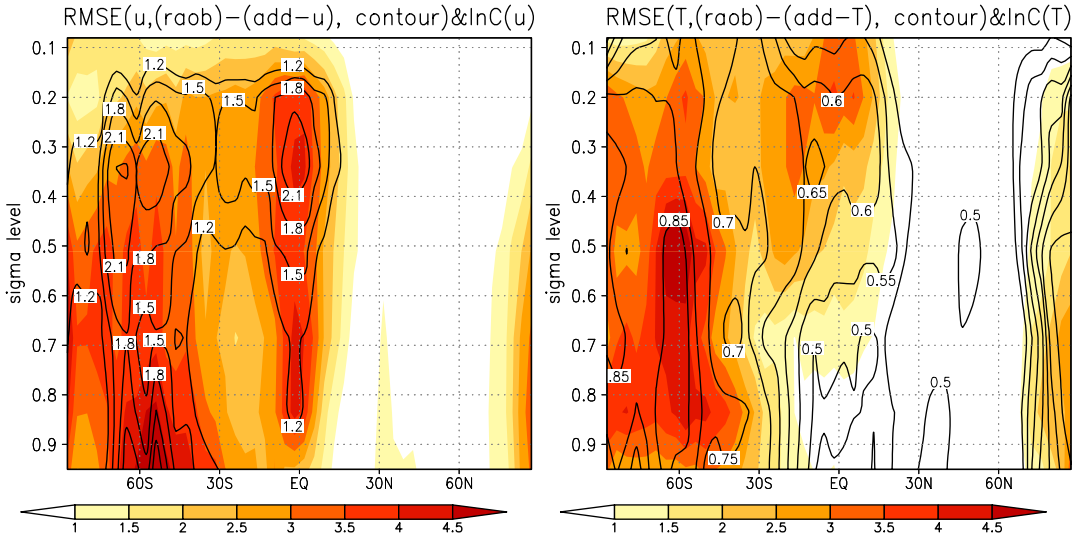


Figure 4.7 RMS error difference (contour) between control experiment and sensitivity experiment, and information content (shaded) (Left panel: between *raob-only* and *raob-u* zonal wind RMS error (Unit : m/s), zonal wind self-sensitivity; right panel: between *raob-only* and *raob-T*, temperature RMS error difference (Unit: K), temperature information content)

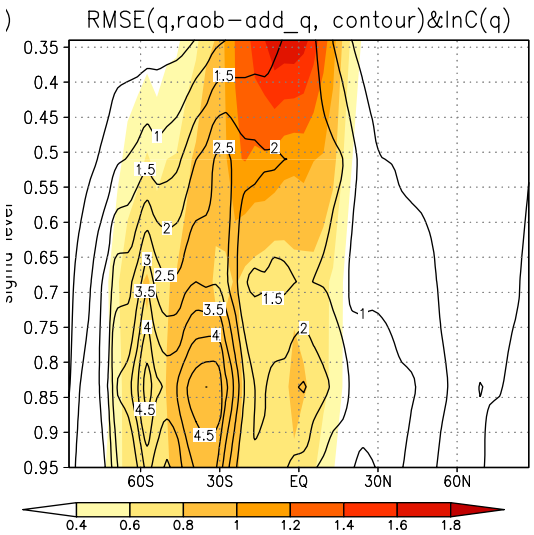


Figure 4.8 RMS error difference (contour, unit: g/kg) between control experiment and sensitivity experiment, and information content (shaded) (between *raob-only* and *raob-q* specific humidity RMS error (Unit : kg/kg), specific humidity information content)

4.5.4 Relative information content of different type observations in different regions

Comparison between the information content and RMS error difference in the data denial and "add-on" experiments clearly shows that information content gives qualitative measure of the impact of the same type observations on the analysis accuracy without doing data denial or "add-on" experiments. Then, can we estimate as well the relationship between the relative observation impact and information content comparison of different types of observations? For this, we compare the relative information content of different type observations in *all-obs* experiment in three latitude bands, which are the mid-latitudes in both Southern and Northern Hemisphere, and the tropics. Different type of observations here are not from different instruments, but different dynamical variables. The relative information content only reflects the information content below the fifth model level since specific humidity observations are only up to that level. Because surface pressure

observations are used to update the dynamical variables in each vertical level in our LETKF implementation, the information content of surface pressure is also the self-sensitivity summed over the lowest five vertical levels. Figure 4.9 shows that specific humidity has the highest relative information content among all the dynamical variables in both the tropics and the SH mid-latitude. All dynamical variables have comparable information content in the NH mid-latitudes, but if we observe both horizontal wind components, the total wind information is larger than that of mass variables. Whether the relative information content among these different dynamical variables can be interpreted as the relative importance of these variables during data assimilation requires further investigation. However, we can at least use the information content comparison to compare the effectiveness of the instruments that measure the same type of observations.

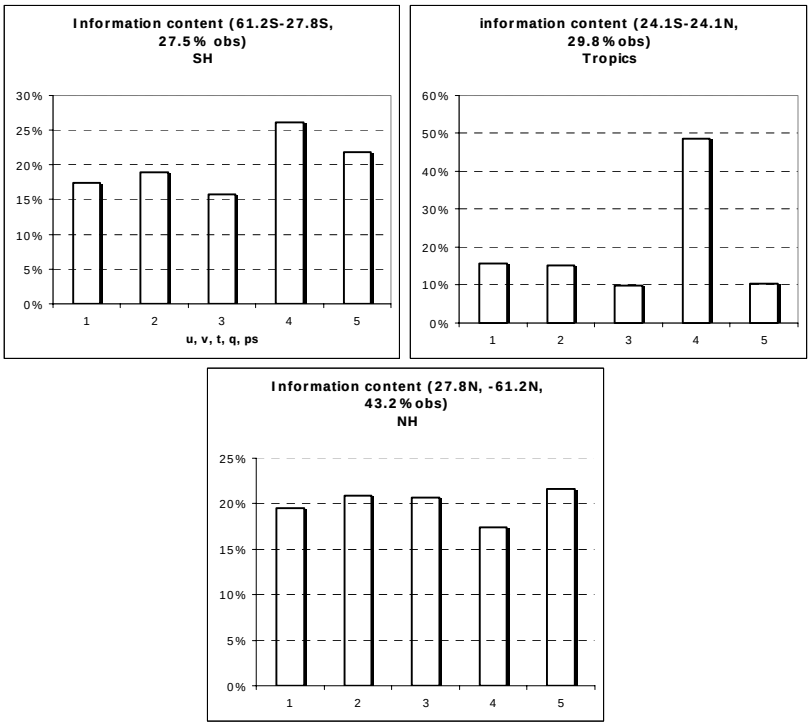


Figure 4.9 Information content of five dynamical variables (1: zonal wind; 2: meridional wind; 3: temperature; 4: specific humidity; 5: surface pressure) over three regions (upper left panel: mid-latitude of the SH; upper right panel: the Tropics; bottom panel: mid-latitude of the NH)

4.6 Conclusions and discussion

The influence matrix reflects the regression fit of the analysis to the observation data, and self-sensitivity gives a measure of the sensitivity of analysis to observations. These measures show the analysis sensitivity to the observations, and can further show the relative impact of the same type observations on the performance of the analysis system when the statistics used in the data assimilation reflects the true uncertainty of each factor.

Following Cardinali et al. (2004), we propose a method to calculate the influence matrix and the self-sensitivity within the LETKF data assimilation scheme. Since the Kalman gain is part of the LETKF scheme, and the influence matrix is the transformation of the Kalman gain to the observation space, it does not require much additional computation time. In the LETKF, each observation is used more than once in different local patches, therefore, we propose to calculate the self-sensitivity in each local patch independently based on the independent influence matrix in each local patch, and the final value is the average of the self-sensitivity over the times that particular observation is being repeatedly assimilated in different local patches. Unlike the self-sensitivity calculation in 4D-Var (Cardinali et al., 2004), the influence matrix and self-sensitivity calculated along with the LETKF is computed exactly so the self-sensitivity satisfies the theoretical value limits (between 0 and 1).

By comparing the self-sensitivity of a global ETKF and the LETKF on Lorenz-40 variable model, we verified the averaging scheme of the self-sensitivity calculation in the LETKF. In addition, the analysis sensitivity per observation increases when the observation coverage is reduced. In agreement with a geometrical interpretation, we showed experimentally that the self-sensitivity is proportional to the analysis error, and is anti-correlated with the observation error.

With a primitive equation model, we carried out two comparisons. One is to compare the information content from the *all-obs* control experiment and the quantitative observation impact calculated from the data denial experiments. The results show that the information content qualitatively reflects the spatial observation impact. The other is to compare the information content calculated in the *raob-only* control experiment based on the possible future observation locations with the actual observation impact from the "add-on" experiments. The results show that the information content can also qualitatively reflect the observation impact in the "add-on" experiments. It implies that the spatial information content can be utilized in the observation design experiment, and can also be used to compare the information content of the instruments that measure the same type of observations. The agreement between self-sensitivity estimates and the actual impacts due to denial or adding on of observations is quite reasonable, especially considering that the self-sensitivity does not take into account the feedback changes in the background error when the observations are added on denied.

Chapter 5 Observation impact study without using adjoint in an ensemble Kalman filter

5.1 Introduction

In recent years, operational NWP centers are assimilating more satellite observations, such as kilo-channel Advanced InfraRed Satellite (AIRS), in addition to in situ observations. Statistically, the assimilation of new observations improves the accuracy of short-range forecasts (e.g. Joiner et al., 2004). However, the value added to the system by different observations depends on the instrument type, observation type, and observation locations, as well as the presence of other observations. The knowledge of the impact that different observations have on the analyses and forecasts is important to better use the observations which have large impact on the forecasts, and avoid using observations which have no impact or even negative impact on the forecasts.

Traditionally, the observation impact has been estimated by carrying out experiments in which part of observations used in the control experiment were not included in the data-denial experiments (e.g., Zapotocny et al., 2000). However, this requires much computational time since a new analysis/forecast experiment has to be carried out for any subset of observations that needs to be evaluated. Langland and Baker (2004, LB hereafter) proposed an adjoint-based procedure to assess observation impact on short-range forecasts without carrying out data-denial experiments. This adjoint-based procedure can evaluate the impact of any or all observations used in the data assimilation and forecast system on a selected measure

of short-range forecast error. In addition, it can be used as a diagnostic tool to monitor the quality of observations, showing which observations make the analysis or the forecast worse, and can also give an estimate of the relative importance of observations from different sources. However, this procedure requires using the adjoint of the forecast model, which is complicated to develop for a comprehensive numerical weather forecast model, and not always available. In this chapter, we propose an ensemble-based sensitivity method to assess the observation impact as in LB but without using the adjoint model, and compare the observation impact calculated from ensemble sensitivity method with the results from the adjoint method, and further compare the impacts from both methods with the actual observation impact. This chapter is organized as follows: Section 5.2 is the derivation of the ensemble sensitivity method and an alternative formula derivation is in Appendix. The experimental design is discussed in Section 5.3 and the results are in Section 5.4. Section 5.5 contains the conclusions.

5.2 Derivation of the ensemble sensitivity method to calculate the observation impact without the adjoint of the NWP model

5.2.1 The sensitivity of forecast error to the observations

We follow the study by LB and calculate the sensitivity of a forecast error at time t to the observations assimilated at time t=00hr (Figure 5.1). LB defined a cost function at time t as the difference between the energy norm of the forecast error from initial condition at t=00hr (at a time when observations $\mathbf{y}_0^o$ were assimilated) and from initial conditions at t=-6hr that did not benefit from the use of the observations $\mathbf{y}_0^o$.

Without loss of generality, and since we will test our calculation procedure in Lorenz-40 variable model, instead of an energy forecast error norm difference, we define the square of the error difference between the forecasts started at 00hr and -6hr and verified at time t as:

$$J = \frac{1}{2}(\boldsymbol{\varepsilon}_{t|0}^T \boldsymbol{\varepsilon}_{t|0} - \boldsymbol{\varepsilon}_{t|-6}^T \boldsymbol{\varepsilon}_{t|-6}) = \frac{1}{2}(\boldsymbol{\varepsilon}_{t|0}^T + \boldsymbol{\varepsilon}_{t|-6}^T)(\boldsymbol{\varepsilon}_{t|0} - \boldsymbol{\varepsilon}_{t|-6}) \quad \textbf{(5.1)}$$

where $\boldsymbol{\varepsilon}_{t|0} = \overline{\mathbf{x}}_{t|0}^f - \overline{\mathbf{x}}_t^a$, and $\boldsymbol{\varepsilon}_{t|-6} = \overline{\mathbf{x}}_{t|-6}^f - \overline{\mathbf{x}}_t^a$. $\overline{\mathbf{x}}_t^a$ is the verification analysis at time t[1]. We follow Bishop's (2007) notation, with the first sub-index indicating the verification time, and the second sub-index, separated by a vertical bar, indicating the time of the initial conditions of a forecast or forecast error, so that $\overline{\mathbf{x}}_{t|0}^f$ and $\overline{\mathbf{x}}_{t|-6}^f$ are the ensemble mean forecast valid at time t, initialized at 00hr and -6hr respectively. Here, K is the number of ensemble members, $\overline{\mathbf{x}}_{t|0}^f = \frac{1}{K}\sum_{i=1}^{K} M_{t|0}(\mathbf{x}_0^{ai})$, $\overline{\mathbf{x}}_{t|-6}^f = \frac{1}{K}\sum_{i=1}^{K} M_{t|-6}(\mathbf{x}_{-6}^{ai})$, and $M_{t|0}$ and $M_{t|-6}$ represent the nonlinear model initialized with the analysis at t=00hr and t=-6hr respectively. Substituting the definitions of $\boldsymbol{\varepsilon}_{t|0}$ and $\boldsymbol{\varepsilon}_{t|-6}$ into the above equation, the cost function can be written as:

$$J = \frac{1}{2}(2\boldsymbol{\varepsilon}_{t|-6}^T + \overline{\mathbf{x}}_{t|0}^f - \overline{\mathbf{x}}_{t|-6}^f)(\overline{\mathbf{x}}_{t|0}^f - \overline{\mathbf{x}}_{t|-6}^f) \quad (5.2)$$

In the following derivation, we aim to express the forecast difference ($\overline{\mathbf{x}}_{t|0}^f - \overline{\mathbf{x}}_{t|-6}^f$) valid at time t, as a function of the observation increments $\mathbf{v}_0 = \mathbf{y}_0^o - \overline{\mathbf{y}}_{0|-6}^b$ at 00hr (Figure 5.1), so that the sensitivity of the cost function to the observations $\frac{\partial J}{\partial \mathbf{v}_0}$ will be

[1] The verification time should be short enough that perturbations grow linearly. Following LB, we define t=24hr.

a function of the observations at time 00hr. $\overline{\mathbf{y}}^{b}_{0|-6} = h(\overline{\mathbf{x}}^{b}_{0|-6})$ is the prediction of the observations at t=00hr, with $h(\cdot)$ the nonlinear observation operator.

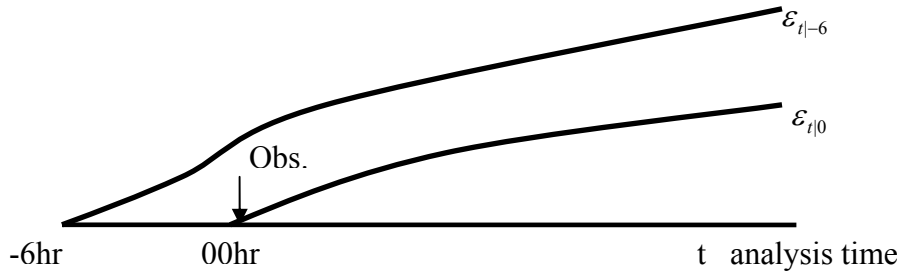


Figure 5.1 Schematic plot of the time relationship of the observation impact on the forecast error at time t. (After Langland and Baker, 2004, Fig 1.)

Following Hunt et al. (2007), the i^{th} analysis ensemble member $\mathbf{x}^{ai}_{0}$ can be written as:

$$\mathbf{x}^{ai}_{0} = \overline{\mathbf{x}}^{b}_{0|-6} + \mathbf{X}^{b}_{0|-6}\widetilde{\mathbf{K}}_{0}\mathbf{v}_{0} + \delta\mathbf{x}^{ai}_{0} \tag{5.3}$$

where $\mathbf{X}^{b}_{0|-6} = \left[\delta\mathbf{x}^{b1}_{0|-6} \;\middle|\; \cdots \;\middle|\; \delta\mathbf{x}^{bK}_{0|-6}\right]$ is a matrix whose K columns are background ensemble perturbations with the i^{th} column given by $\delta\mathbf{x}^{bi}_{0|-6} = \mathbf{x}^{bi}_{0|-6} - \overline{\mathbf{x}}^{b}_{0|-6}$,and $\widetilde{\mathbf{K}}_{0} = \widetilde{\mathbf{P}}^{a}_{0}\mathbf{Y}^{bT}_{0}\mathbf{R}^{-1}_{0}$ is the Kalman gain matrix in the ensemble subspace spanned by the forecasts. $\tilde{\mathbf{P}}^{a}_{0} = \left[(K-1)\mathbf{I} + \mathbf{Y}^{bT}_{0}\mathbf{R}^{-1}_{0}\mathbf{Y}^{b}_{0}\right]^{-1}$ is the analysis error covariance matrix in the ensemble subspace spanned by the forecasts. The i^{th} column of the analysis perturbation matrix $\mathbf{X}^{a}_{0}$ is the analysis perturbation vector $\delta\mathbf{x}^{ai}_{0} = \mathbf{x}^{ai}_{0} - \overline{\mathbf{x}}^{a}_{0}$, and $\mathbf{X}^{a}_{0} = [(K-1)\widetilde{\mathbf{P}}^{a}_{0}]^{\frac{1}{2}}$. $\mathbf{Y}^{b}_{0}$ is a matrix whose i^{th} column $h(\mathbf{x}^{bi}_{0|-6}) - h(\overline{\mathbf{x}}^{b}_{0|-6})$ is the vector of ensemble perturbations in observation space. $\mathbf{R}_{0}$ is the observation error covariance. An over-bar represents an average over the K ensemble members, a tilde

indicates that a vector or matrix is represented in the subspace of ensemble forecasts, and δ represents the difference between an ensemble member value and the ensemble mean.

Based on equation (5.3), the forecast $\overline{\mathbf{x}}^f_{t|0} = \frac{1}{K}\sum_{i=1}^{K} M_{t|0}(\mathbf{x}^{ai}_0)$ initialized at t=00hr is written as:

$$\overline{\mathbf{x}}^f_{t|0} = \frac{1}{K}\sum_{i=1}^{K} M_{t|0}(\overline{\mathbf{x}}^b_{0|-6} + \mathbf{X}^b_{0|-6}\widetilde{\mathbf{K}}_0\mathbf{v}_0 + \delta\mathbf{x}^{ai}_0) \tag{5.4}$$

Note that although in the following derivation we make a linearization, the actual computation does not require the tangent linear or adjoint model. We linearize equation (5.4) around the background mean state $\overline{\mathbf{x}}^b_{0|-6}$, and define the tangent linear model as $\mathbf{M}_{t|0}$, so that:

$$\overline{\mathbf{x}}^f_{t|0} \cong M_{t|0}(\overline{\mathbf{x}}^b_{0|-6}) + \frac{1}{K}\sum_{i=1}^{K}\mathbf{M}_{t|0}(\mathbf{X}^b_{0|-6}\widetilde{\mathbf{K}}_0\mathbf{v}_0 + \delta\mathbf{x}^{ai}_0) \tag{5.5}$$

From Hunt et al. (2007), $\delta\mathbf{x}^{ai}_0 = \mathbf{X}^b_{0|-6}\delta\mathbf{w}^{ai}_0$, wher $\delta\mathbf{w}^{ai}_0$ is the i^{th} column of a symmetric K by K matrix of weight perturbations $\mathbf{W}^a_0 = [(K-1)\widetilde{\mathbf{P}}^a_0]^{1/2}$ with element δw^{aji}_0. Substituting it into equation (5.5),

$$\overline{\mathbf{x}}^f_{t|0} \cong M_{t|0}(\overline{\mathbf{x}}^b_{0|-6}) + \frac{1}{K}\sum_{i=1}^{K}\mathbf{M}_{t|0}\left[\mathbf{X}^b_{0|-6}(\widetilde{\mathbf{K}}_0\mathbf{v}_0 + \delta\mathbf{w}^{ai}_0)\right] \tag{5.6}$$

Since $M_{t|0}(\overline{\mathbf{x}}^b_{0|-6}) = M_{t|0}(\frac{1}{K}\sum_{i=1}^{K} M_{0|-6}(\mathbf{x}^{ai}_{-6}))$, $M_{t|0}(\overline{\mathbf{x}}^b_{0|-6}) = \frac{1}{K}\sum_{i=1}^{K} M_{t|-6}(\mathbf{x}^{ai}_{-6})$, so that $\overline{\mathbf{x}}^f_{t|-6} = M_{t|0}(\overline{\mathbf{x}}^b_{0|-6})$. In addition, $\mathbf{X}^f_{t|-6} = \mathbf{M}_{t|0}\mathbf{X}^b_{0|-6}$, where

$\mathbf{X}^f_{t|-6} = \left[\delta\mathbf{x}^{f1}_{t|-6} \;\; | \;\; \cdots \;\; | \;\; \delta\mathbf{x}^{fK}_{t|-6}\right]$ is a matrix whose K columns are forecast ensemble perturbations with the i^{th} column $\delta\mathbf{x}^{fi}_{t|-6} = \mathbf{x}^{fi}_{t|-6} - \overline{\mathbf{x}}^f_{t|-6}$. Then, equation (5.6) becomes:

$$\begin{aligned}\overline{\mathbf{x}}^f_{t|0} &= \overline{\mathbf{x}}^b_{t|-6} + \frac{1}{K}\sum_{i=1}^{K}\mathbf{X}^f_{t|-6}(\widetilde{\mathbf{K}}_0\mathbf{v}_0 + \delta\mathbf{w}_0^{ai}) \\ &= \overline{\mathbf{x}}^b_{t|-6} + \mathbf{X}^f_{t|-6}(\widetilde{\mathbf{K}}_0\mathbf{v}_0) + \frac{1}{K}\sum_{j=1}^{K}\mathbf{X}^{fj}_{t|-6}\sum_{i=1}^{K}\delta w_0^{aji}\end{aligned} \tag{5.7}$$

The last term in equation (5.7) vanishes since the perturbation weights summed over either the K columns or the K rows are equal to one: $\sum_{j=1}^{K}\delta w_0^{aji} = \sum_{i=1}^{K}\delta w_0^{aji} = 1$ (Appendix B.1), and the average of the forecast ensemble perturbations is equal to zero, so that $\overline{\mathbf{x}}^f_{t|0} = \overline{\mathbf{x}}^b_{t|-6} + \mathbf{X}^f_{t|-6}\widetilde{\mathbf{K}}_0^a\mathbf{v}_0$. Therefore, we can write

$$\boldsymbol{\varepsilon}_{t|0} - \boldsymbol{\varepsilon}_{t|-6} = \overline{\mathbf{x}}^f_{t|0} - \overline{\mathbf{x}}^f_{t|-6} = \mathbf{X}^f_{t|-6}\widetilde{\mathbf{K}}_0\mathbf{v}_0 \tag{5.8}$$

Similarly,

$$\begin{aligned}\boldsymbol{\varepsilon}_{t|0} + \boldsymbol{\varepsilon}_{t|-6} &= \overline{\mathbf{x}}^f_{t|0} - \overline{\mathbf{x}}^a_t + \overline{\mathbf{x}}^f_{t|-6} - \overline{\mathbf{x}}^a_t \\ &= \overline{\mathbf{x}}^f_{t|-6} - \overline{\mathbf{x}}^a_t + \overline{\mathbf{x}}^f_{t|-6} - \overline{\mathbf{x}}^a_t + \overline{\mathbf{x}}^f_{t|0} - \overline{\mathbf{x}}^f_{t|-6} \\ &= 2\boldsymbol{\varepsilon}_{t|-6} + \mathbf{X}^f_{t|-6}\widetilde{\mathbf{K}}_0\mathbf{v}_0\end{aligned} \tag{5.9}$$

The cost function of equation (5.1) is then written as:

$$\begin{aligned}J &= \frac{1}{2}(\boldsymbol{\varepsilon}^T_{t|0} + \boldsymbol{\varepsilon}^T_{t|-6})(\boldsymbol{\varepsilon}_{t|0} - \boldsymbol{\varepsilon}_{t|-6}) \\ &= \frac{1}{2}[2\boldsymbol{\varepsilon}_{t|-6} + \mathbf{X}^f_{t|-6}\widetilde{\mathbf{K}}_0\mathbf{v}_0]^T\mathbf{X}^f_{t|-6}\widetilde{\mathbf{K}}_0\mathbf{v}_0\end{aligned} \tag{5.10}$$

If the model is nonlinear, J is a linear approximation of the cost function of equation (5.1). Since the error $\boldsymbol{\varepsilon}_{t|-6}$ is not correlated with the observations assimilated at t=00hr, the sensitivity of the forecast error to the observations can be written as:

$$\frac{\partial J}{\partial \mathbf{v}_0} = \left[\widetilde{\mathbf{K}}_0^T \mathbf{X}_{t|-6}^{fT}\right]\boldsymbol{\varepsilon}_{t|-6} + \mathbf{X}_{t|-6}^{f}\widetilde{\mathbf{K}}_0 \mathbf{v}_0\Big] \tag{5.11}$$

Note that the sensitivity of the cost function J to the observations (Equation (5.11)) can be calculated "on the fly" based on the matrix of weights calculated in the data assimilation at 00hr, the observation increment at 00hr, and the ensemble forecasts valid at time t initialized at -6hr, and it does not require the adjoint model. This ensemble sensitivity method is different from Ancell and Hakim (2007), who also proposed a method to calculate the forecast sensitivity to the observations without using adjoint model. In their approach, the sensitivity is a function of the inverse of the analysis error covariance, and they calculate it one observation at a time. In the Appendix (B.3), we give another derivation of the sensitivity of the cost function J to the observations without linearization, which gives similar results as those calculated from equation (5.11).

5.2.2 Observation impact on the forecast

As discussed in LB, the observation sensitivity can be used to examine the actual observation impact on the forecast. The forecast error difference between $\boldsymbol{\varepsilon}_{t|0}$ and $\boldsymbol{\varepsilon}_{t|-6}$ is solely due to the assimilation of the observations at 00hr. As a result (Appendix B.2), using the observation sensitivity gradients $\frac{\partial J}{\partial \mathbf{v}_0}$, the observation impact on the forecast can be written as:

$$J = \frac{1}{2}(\boldsymbol{\varepsilon}_{t|0}^T \boldsymbol{\varepsilon}_{t|0} - \boldsymbol{\varepsilon}_{t|-6}^T \boldsymbol{\varepsilon}_{t|-6}) = \left\langle \mathbf{v}_0, \frac{\partial J}{\partial \mathbf{v}_0} \right\rangle \tag{5.12}$$

With a nonlinear model, equation (5.12) is an approximation of equation (5.1) due to the use of tangent linear and adjoint model in the derivation of equation (5.12) (Appendix B.2). Though the derivation of equation (5.12) is based on tangent linear and adjoint model, the actual calculation in the ensemble sensitivity method does not require either of them.

The equation (5.12) expresses the forecast error difference as a function of observations. When the assimilated observations improve the forecast at time t**,** the forecast error difference is negative, and so is the value calculated from equation (5.12). When the assimilated observations deteriorate the forecast, the value calculated from equation (5.12) will be positive. Furthermore, the cost function J can be expressed as the sum of J^l, the observational impact caused by the l^{th} subset of the observations $\mathbf{y}_0^o = [(\mathbf{y}_0^{o1})^T \quad \cdots \quad (\mathbf{y}_0^{oL})^T]^T$, if the observation errors of observation subsets are not correlated:

$$J = \sum_{l=1}^{L} J^l = \sum_{l=1}^{L} \mathbf{v}_0^l \cdot \frac{\partial J}{\partial \mathbf{v}_0^l} \tag{5.13}$$

where $v_0^l = \mathbf{y}_0^{ol} - h^l(\overline{\mathbf{x}}_{0|-6}^b)$. Based on equation (5.13), we can calculate the observation impact from any subset of observations without conducting data denial experiments, and can also compare the importance of observations from different sources.

In Chapter 4, we discussed the calculation of the self-sensitivity within the LETKF data assimilation, which reflects how sensitive of the analysis value is to the change of observations. The self-sensitivity can only qualitatively reflect the

observation impact on the analysis assuming that the observation errors correctly reflect the statistics of the assimilated observations, but cannot show the actual quality of the observations. The observation impact discussed in this chapter provides a quantitative estimation of the actual observation impact on the short-range forecasts. It can also be calculated along with the LETKF data assimilation scheme once the short-range ensemble forecasts initialized at -6hr are computed. The calculation procedure is same as the calculation method of self-sensitivity discussed in Chapter 4.

In the following sections, we will examine whether the adjoint method and the ensemble sensitivity method we proposed can actually detect bad observations whose errors do not satisfy the Gaussian assumption $\varepsilon^o \to N(0,R)$, and compare the measured observation impact, the observation impact calculated from the adjoint method (LB), and from the ensemble sensitivity method we derived here. The comparison is carried out in the Lorenz-40 variable model.

5.3 Experimental design

As in Chapter 2, we use Lorenz-40 variable model with the forcing F equal to 8 for the nature run, and 7.6 for the forecasts, allowing for some model error in the system.

Following LB, we estimate the impact of the observations (assimilated at 00hr) on the forecast valid at t=24hr, an interval almost enough that perturbations remain approximately linear, so that in equation (5.1), the cost function is defined as the difference of the forecast errors between a 24-hour forecast (initialized at 00hr) and a

30-hour forecast (initialized at -6hr). The difference between these two forecasts is due only to the assimilation of the observations at 00hr in the initial condition of 24-hour forecast. The observations are observed at every grid point. We present experiments with a "normal" case, a "larger random error" and a "biased observation" cases. In the normal case, the assumed observation error standard deviation 0.2 doses represent the actual error statistics for every observation obtained from the nature run plus a Gaussian random perturbation. In the "larger random error" case, the observation at a single grid point (the 11^{th} grid point) has four times larger random error standard deviation than the other observations. However, in the data assimilation process, we still use the error standard deviation 0.2 to represent the error statistics for every observation, including the 11^{th} grid point. Such an experiment simulates real cases when some observations may have larger random errors than assumed in the data assimilation system. In addition, real observations may also have biases, something especially common when we assimilate satellite observations (e.g., Derber and Wu, 1998). Therefore, in the "biased" case experiment, we include a bias equal to 0.5 in the observation at the 11^{th} grid point, but still assume the observation is non-biased during data assimilation.

We run each experiment for 7500 analysis cycles with the LETKF data assimilation schemes. The time average statistics shown in the next section is the average over the last 7000 analysis cycles. Through out these experiments, we check whether our ensemble sensitivity method is comparable with the adjoint method of

LB in assessing the observation impact on the forecast error, and compare the ability of both methods to detect poor quality observations.

5.4 Results

5.4.1 Normal case

Figure 5.2 shows the observation impact calculated from the adjoint method (red line with crosses), the ensemble method (green line with closed circles) and the actual forecast error difference (black line with open circles) between the analysis cycle 5700 and 5780 for the "normal" case. It shows that the observation impact calculated from ensemble sensitivity method is similar to the result from adjoint sensitivity method, and both methods succeed in capturing the actual forecast improvement due to the assimilation of the observations at 00hr. Both explain more than 90% of the day-to-day variations in forecast improvement.

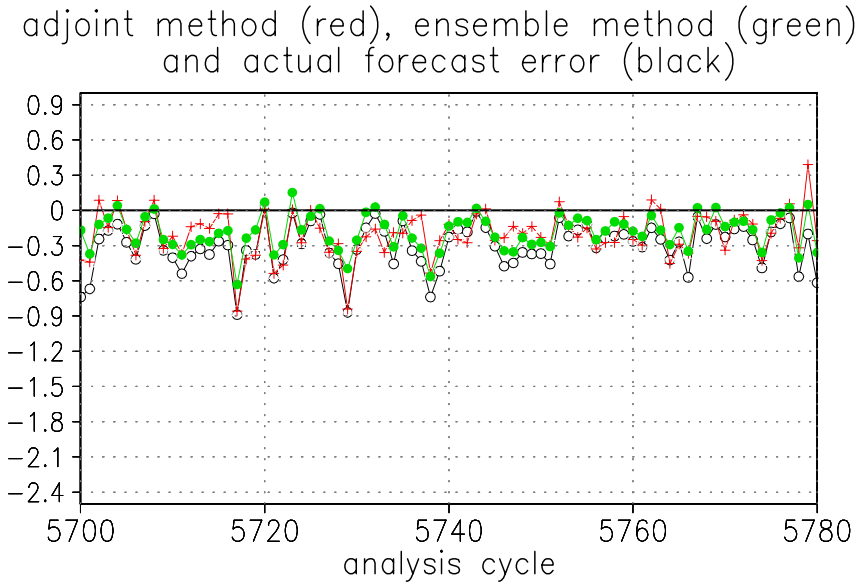


Figure 5.2 Snapshots (between analysis cycles 5700 and 5780) of forecast error difference and the observation impact from the normal case (black line: the actual forecast error difference between 24-hour forecast and the 30-hour forecast; red line:

the observation impact calculated from adjoint method; green line: the observation impact calculated from the ensemble method; black solid line: zero line, i.e., no impact)

5.4.2 Larger random error case

When the observation at the 11th observation location has four times larger random error standard deviation than the other observations, both the ensemble sensitivity method and the adjoint method show that assimilation of this observation increases the forecast error (Figure 5.3). The signal from ensemble sensitivity method at the 11^{th} grid point is larger than that of the adjoint method, but elsewhere, both methods have similar values. It is interesting to note that the observations of the adjacent observation locations improve the forecast most, because they partially correct the impact of the faulty observation at the 11^{th} grid point. Snapshots of the spatially summed impact show that the observation impact calculated from both methods reflects the actual forecast error difference (Figure 5.4) even when one of the observations has erroneous error statistics. Because of the poor quality of the observation at the 11^{th} observation location, the domain averaged observation impact has some large spikes (Figure 5.4).

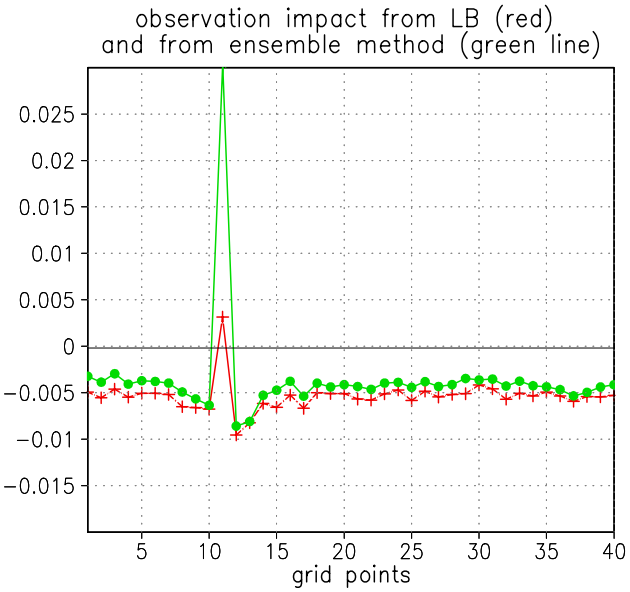


Figure 5.3 Time average (over the last 7000 analysis cycles) of the observation impact from the larger random error case (four times larger random error at the 11th grid point). Green line with closed circles is from ensemble method, and the red line with crosses is from adjoint method, and the black solid line is zero line.

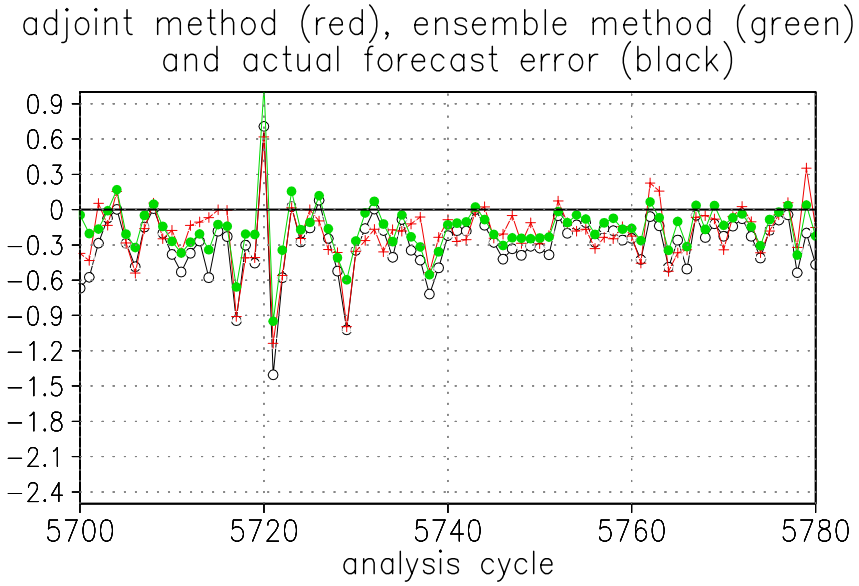


Figure 5.4 Snapshots (between analysis cycle 5700 and 5780) of forecast error difference and the observation impact from the larger random error case (the notation is same as in Figure 5.2)

5.4.3 Biased case

When the 11th observation location has a bias, the ensemble sensitivity method indicates (Figure 5.5), like the adjoint method, that the assimilation of this observation increases the forecast error. Again, the negative observation impact at 11th grid point makes the positive observation impact (reduction of forecast error) of assimilating the adjacent observation locations larger.

These examples show that the ensemble sensitivity method gives observation impact similar with that from adjoint method, and both methods reflect more than 90% of the actual forecast error reduction due to assimilation of the observations at 00hr. Like the adjoint method, the ensemble method can detect observation which has poor quality either with larger random error or bias, and the signal detected by the ensemble sensitivity method is stronger.

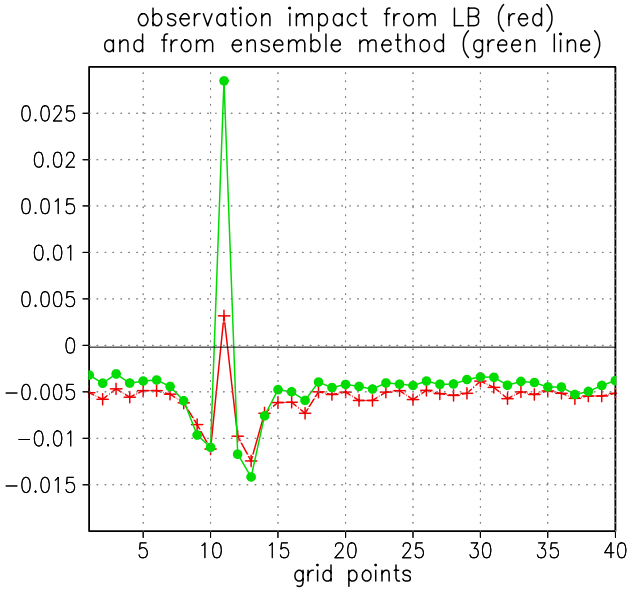


Figure 5.5 The biased case with the bias equal to 0.5 at 11th grid point. The line notation is same with Figure 5.3.

5.5 Summary and conclusions

The observations are the central information introduced into the forecast system during data assimilation. However, the quality and impact of the observations is always different due to the magnitude of observation error, observation locations and the model dynamics. Accurately monitoring the quality and impact of the observations assimilated in the system can help to delete the observations that routinely deteriorate the forecast, and can better use the observations that have larger impact on the forecast than the other observations. In the past, monitoring has been based on observational increments, but we have found that observation sensitivity approach is more effective in detecting poor observations.

In this chapter, following Langland and Baker (2004), we proposed an ensemble sensitivity method to measure the observation impact on the error difference between the forecasts initialized from 00hr and -6hr. Unlike the adjoint method by Langland and Baker (2004), the ensemble sensitivity method we propose does not need the adjoint model. We compared the ensemble sensitivity method we proposed to the adjoint model using Lorenz-40 variable model. The results show that the ensemble sensitivity method gets similar results as the adjoint method, and both explain more than 90% forecast error differences on a day-to-day basis in our experimental setup. Both methods can detect the "bad" observations that are of poor quality, with either larger random errors or with bias, and the ensemble sensitivity method shows stronger signal in such scenarios. Like the adjoint method by LB, this

method can be applied in the observation quality control as well as to compare the importance of different type observations. It could be routinely calculated within the data assimilation, thus providing a powerful tool to understand cases of forecast failure and to tune the observation error statistics.

Chapter 6 Humidity data assimilation with the Local Ensemble Transform Kalman filter

6.1 Introduction

As stated in Chapter 1, humidity data assimilation is a difficult problem due to its highly changing error characteristics, both spatially and temporally, and is especially difficult in variational assimilation methods which assume constant background error covariance. Some variational approaches overcome the constant error assumption by re-formulating the assimilated humidity variables based on the background forecast, such as in Dee and da Silva (2003) and Holm et al. (2002). The re-formulated variable not only introduces the time changing error into the data assimilation system, but also has more Gaussian error distribution than other choices of the humidity variables. Nevertheless, the variational approaches used in operational centers still assimilate humidity variable uni-variately.

EnKF provides a unique assimilation method for multivariate humidity observations, since it estimates the background error covariance in each analysis cycle, and automatically couples all the dynamical variables together. However, as in variational assimilation approaches, it assumes a Gaussian error distribution, so that the choice of humidity variable type is still very important. In this chapter, based on the OSSE experimental setup, we first will compare several choices for the assimilated humidity variable type when the specific humidity has non-Gaussian observation error. The tested humidity variables are logarithm of specific humidity, specific humidity, relative humidity and the pseudo-RH proposed by Dee and da Silva (2003). Since we create the humidity observation by adding Gaussian random

error to the logarithm of specific humidity, the logarithm of specific humidity is the only humidity variable with perfect error statistics.

Compared to the development of data assimilation of the other dynamical variables, one of the obstacles for humidity data assimilation is the poor quality of humidity observations. With the kilo-channel AIRS satellite launched in 2002, more and more high quality humidity observations are available. However, so far, most results only show neutral or negative impacts from assimilation of AIRS humidity information (e.g. Joiner et al., 2004) from radiation in channels with water vapor bands. Since AIRS is a high spectral instrument, humidity retrievals have very high quality (Susskind et al., 2003). At the end of this chapter, we will also show some preliminary results from assimilating AIRS humidity retrievals (Chris Barnet, personal communication) in multivariate mode in the NCEP Global Forecast System (GFS).

This chapter is organized as follows: from Section 6.2 to Section 6.5, we will compare several choices of the assimilated humidity variable types assimilated both uni-variately and multivariately, i.e., coupled with the other dynamical variables, with the LETKF data assimilation scheme in a global primitive equation model. Section 6.6 shows preliminary results from assimilating AIRS humidity retrievals, and finally in Section 6.7, we draw conclusions.

6.2 Model and simulated observations

We use the same primitive equation model as the one we used in Chapter 3, which is the SPEEDY model (Molteni, 2003) (discussed in detail in Section 3.2). In this study, we follow a "perfect model" Observing System Simulation Experiments (OSSEs) setup, in which the simulated truth is generated with the same atmospheric model as the one used in data assimilation. The winds, temperature and surface pressure observations are the truth with added Gaussian random perturbations. The observation error standard deviations assumed for winds are about 30% of the natural variability, shown in the left two panels of Figure 4.3 (Chapter 4). Temperature observation error standard deviation is assumed to be 0.8K at each vertical level. The error standard deviation for surface pressure is 1hPa. We will discuss the observation error characteristics of several choices of humidity variables individually.

Specific humidity is the most commonly used humidity variable. However, in reality, it has non-Gaussian observation error, and an approximately logarithmic vertical distribution. Therefore, we create simulated specific humidity observations by first adding the Gaussian random perturbations to the logarithm of the true specific humidity, and then, transforming the logarithm of specific humidity to specific humidity. The Gaussian random error standard deviation for logarithm specific humidity is shown in Figure 6.1. Since SPEEDY is a spectral model, it can create negative specific humidity values (e.g., Kalnay, 2003). In that case, the true specific humidity is set to a very small positive value before calculating the logarithm specific humidity. The observation error standard deviation for specific humidity observations

is shown in the top panel in Figure 6.2, and has a magnitude similar to that assumed in the operational data assimilations. The error varies significantly with vertical levels. However, the error distribution of specific humidity observations is not Gaussian anymore. As shown in the bottom panel of Figure 6.2, the actual observation error distribution (crosses) is far from its Gaussian fit (open circles) of the observation error for the third model level, and this is also true for all the other levels.

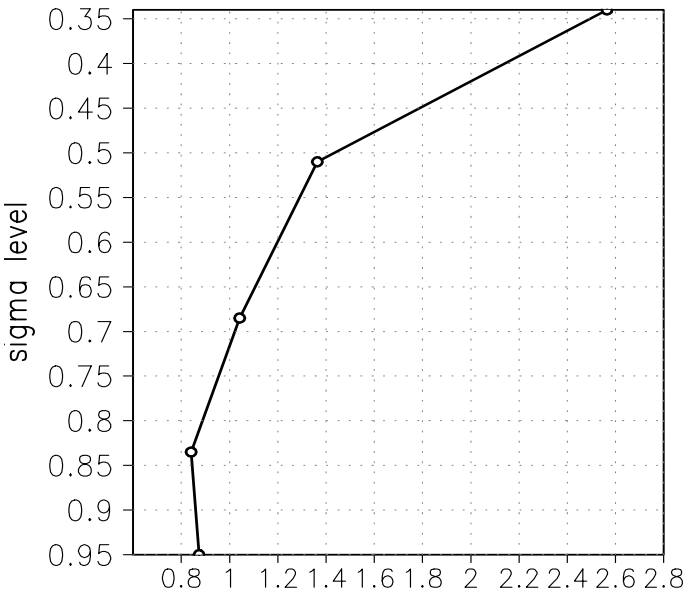


Figure 6.1 The observation error standard deviation for the logarithm specific humidity (unit: 0.1)

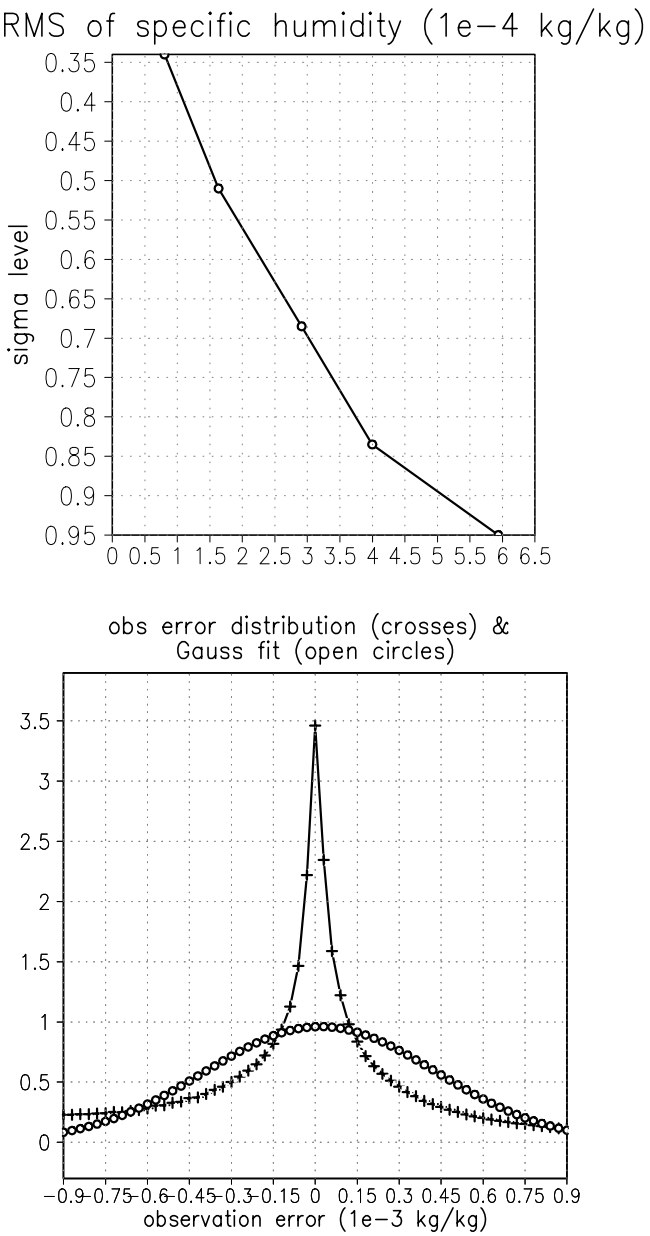


Figure 6.2 Top panel: The observation error standard deviation as function of the vertical levels for specific humidity (Unit: 10^{-4}kg/kg); Bottom panel: The actual observation error distribution (10^{-3}kg/kg, solid line with crosses) and the Gaussian fit of the observation error distribution (10^{-3}kg/kg , open circles) for the third sigma level.

Relative humidity (RH) is another choice of humidity variable. We create relative humidity observations by dividing the observed specific humidity by the saturated humidity calculated from temperature and surface pressure observations.

The observation error is calculated against the true relative humidity, which is shown in the top left panel of Figure 6.3. It does not vary as much as the specific humidity observations (Figure 6.2) with vertical levels. Compared to the specific humidity observations, the RH observation error distribution (crosses) has a much better Gaussian fit (open circles) (left bottom panel in Figure 6.3), indicating that the relative humidity observation error distribution is more Gaussian. However, since RH observations are a function of the observed temperature and surface pressure, they have the disadvantage of large error correlations with these variables. This is also true for the real atmospheric relative humidity and this correlation with temperature and pressure makes assimilation of RH much harder.

Pseudo relative humidity (pseudo-RH) was proposed by Dee and da Silva (2003) with the purpose of maintaining the more Gaussian observation error characteristics of relative humidity observations, and at the same time, avoiding the disadvantage of a high correlation between the relative humidity and temperature observations. The pseudo-RH is defined as the ratio between the observed specific humidity and the background saturated specific humidity. By dividing the specific humidity observations by the saturated specific humidity from background, it has an error distribution similar to the relative humidity observations, as shown in the right panel of Figure 6.3. At the same time, since the saturated humidity comes from the background, the pseudo-RH error is not correlated with the temperature and surface pressure observation errors. As far as we know, our experiments are the first using pseudo-RH within an EnKF formulation.

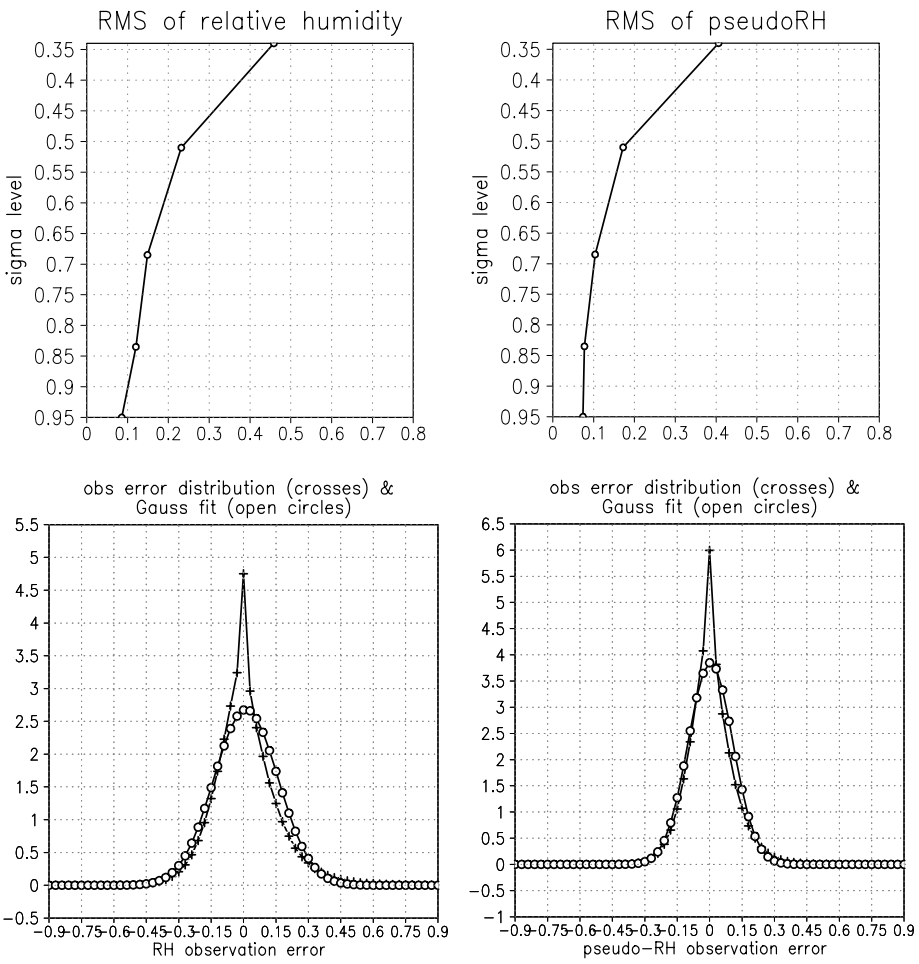


Figure 6.3 The observation error standard deviation for relative humidity (top left panel) and pseudo-RH (top right panel). The actual observation error distribution (crosses) and the Gaussian fit observation error distribution (open circles) for relative humidity (bottom left panel) and pseudo-RH (bottom right panel) at the third sigma level.

6.3 Experimental design

In our formulation, we assume that for the winds, temperature and surface pressure observations, there is one observation every three grid points in both latitude and longitude, so total observation coverage is about 11%, as shown in Figure 6.4 (top panel). We set the humidity observations at the center of two adjacent grid points, as shown in Figure 6.4 (bottom panel). In addition, to make the impact from the assimilation of humidity observations more significant, the humidity observation

coverage is denser than the other dynamical variables, with 25% coverage. As in Chapter 4, the humidity observations are only up to the fifth model level (~300hPa).

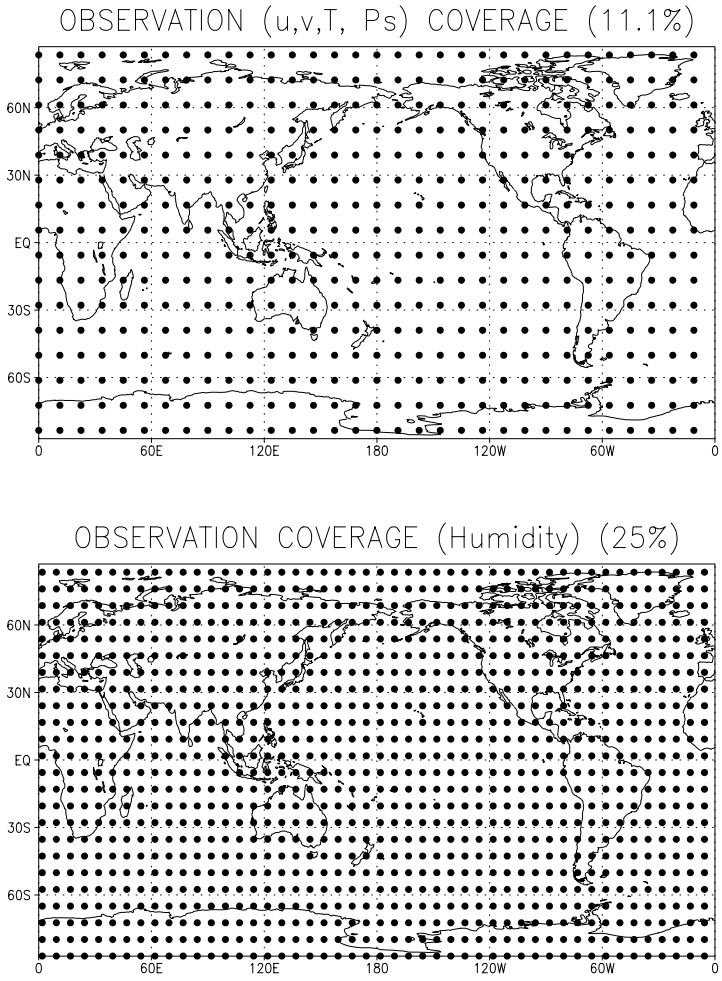


Figure 6.4 Top panel: the observation coverage for winds, temperature and surface pressure; Bottom panel: the observation coverage of humidity observations.

We have two types of experiments. In the first type, the humidity is updated by itself uni-variately, which means that the humidity does not interact with the other dynamical variables during the data assimilation, while the other dynamical variables (i.e., winds, temperature and surface pressure) are coupled together multi-variately. The humidity only interacts with the other dynamical variables during the forecast process. We call this type of experiments as *uni-q* experiment, which is the way that

the operational centers do the humidity assimilation. The other is the fully-coupled experiments, in which the humidity is fully coupled to the other dynamical variables through the coupled background error covariance in the data assimilation. We call it *coupled (multivariate)* experiment. In this type of experiment, the humidity variable is used to update the other dynamical variables, and the humidity analysis is updated by the other dynamical variables as well. In each type of experiment, we carry out four experiments with different choices of humidity variables, which are the logarithm of specific humidity ($\ln(q)$), specific humidity (q), relative humidity (*RH*), and pseudo relative humidity (pseudo-*RH*). We note that the observations are derived from $\ln(q)$ Gaussian errors, so that $\ln(q)$ has a distinct advantage over the other variables. The *control* experiment is the one that does not have the humidity observations assimilated at all. We will compare both *uni-q* and *coupled* experiments with the *control* run results, and compare the performance of different choices of humidity variable types within both *uni-q* experiment and *coupled* experiment.

6.4 Formulation of the assimilation of different choices of humidity variables within LETKF data assimilation scheme

As shown in Chapter 1, in the LETKF, the analysis mean state and the ensemble perturbations are calculated from equation (1.1) and equation (1.2),

$$\overline{\mathbf{x}}^{a} = \overline{\mathbf{x}}^{\mathrm{b}} + \mathbf{X}^{\mathrm{b}}\widetilde{\mathbf{K}}[\mathbf{y}^{\mathrm{o}} - \mathrm{h}(\overline{\mathbf{x}}^{\mathrm{b}})] \tag{6.1}$$

$$\mathbf{X}^{a} = \mathbf{X}^{b}\left((K-1)\left[(\mathbf{H}\mathbf{X}^{b})^{T}\mathbf{R}^{-1}\mathbf{H}\mathbf{X}^{b} + (K-1)\mathbf{I}\right]^{-1}\right)^{\frac{1}{2}} \tag{6.2}$$

$\widetilde{\mathbf{K}} = \left[(\mathbf{HX}^b)^T \mathbf{R}^{-1}(\mathbf{HX}^b) + (K-1)\mathbf{I})\right]^{-1}(\mathbf{HX}^b)^T \mathbf{R}^{-1}$ is the Kalman gain in the ensemble perturbation space, with K equal to the number of the ensemble members. (See details in Chapter 1). The different choices of humidity variables affect the formulation of the observation increment $\mathbf{y}^o - h(\overline{\mathbf{x}}^b)$, the ensemble perturbations in the observation space $\mathbf{HX}^b$, and equation for the analysis mean state. In the following subsections, we give the detailed formulation for each choice of humidity variable.

6.4.1 Assimilation of specific humidity (q)

The assimilation of specific humidity is the easiest one among all these choices of humidity variables since specific humidity is directly available from model output. The assimilation of specific humidity is straightforward:

$$\overline{\mathbf{q}}^a = \overline{\mathbf{q}}^{\mathrm{b}} + \mathbf{X}^{\mathrm{b}}\widetilde{\mathbf{K}}[\mathbf{q}^{\mathrm{o}} - h(\overline{\mathbf{q}}^{\mathrm{b}})] \tag{6.3}$$

$\overline{\mathbf{q}}^a$ and $\overline{\mathbf{q}}^b$ are the analysis and background mean state for the specific humidity field. The observational increment is $\mathbf{q}^o - h(\overline{\mathbf{q}}^b)$, where the observation operator $h(\cdot)$ is just linear interpolation since the model dynamical variable is also specific humidity. During the calculation of the observational increment, we set $h(\overline{\mathbf{q}}^b)$ equal to a very small positive value when it is negative since specific humidity observations are the exponential of the $\ln(q)$ observations, and they are positive definite. The same formulation is used in both *uni-q* experiment and *coupled* experiment. $\mathbf{X}^b$ is a matrix of the specific humidity ensemble perturbations with each column equal to the difference between ensemble forecast and the mean state. $\mathbf{HX}^b = h(\mathbf{q}^b) - h(\overline{\mathbf{q}}^b)$ is

the background specific humidity perturbation in the observation space. The analysis perturbations are directly calculated from equation (6.2)

6.4.2 Assimilation of logarithm specific humidity ($\ln(q)$)

Unlike the assimilation of specific humidity, the observation operator $h(\cdot)$ is not linear in the assimilation of $\ln(q)$, but instead has a logarithmic relationship with the background. The observational increment is equal to $\ln(\mathbf{q}^o) - \ln(\mathbf{H}^l \overline{\mathbf{q}}^b)$, where $\mathbf{H}^l$ is the linear interpolation operator. In the calculation of observational increment, we first horizontally interpolate the background humidity to the observation locations with the linear observation operator $\mathbf{H}^l$, then do the logarithm transformation. The ensemble perturbation at the observation locations $\mathbf{HX}^b = \ln(\mathbf{H}^l \mathbf{q}^b) - \ln(\mathbf{H}^l \overline{\mathbf{q}}^b)$ is calculated in a similar way. When the value of $\overline{\mathbf{q}}^b$ or $\mathbf{q}^b$ is negative, it is set to equal to a very small positive value before the logarithm calculation. The updated analysis variable is specific humidity, so the ensemble perturbations $\mathbf{X}^b$ are still the specific humidity ensemble perturbations. Therefore, the analysis mean state is equal to

$$\overline{\mathbf{q}}^a = \overline{\mathbf{q}}^b + \mathbf{X}^b \widetilde{\mathbf{K}} [\ln(\mathbf{q}^o) - \ln(\mathbf{H}^l \overline{\mathbf{q}}^b] \tag{6.4}$$

The analysis perturbations are directly calculated from equation (6.2).

The reason that we use specific humidity as analysis variable instead of $\ln(q)$ is that the choice of $\ln(q)$ as analysis variable will make the specific humidity analysis positive definite, a disadvantage that it will introduce bias into the system since the forecast field of specific humidity has negative values in the SPEEDY

model. In addition, when the analysis variable is $\ln(q)$, the analysis value will be close to zero when either the background or the observation increment is close to zero, which could produce serious problem in the high latitude or upper vertical levels (Dee and da Silva, 2003). Unlike the other choices of humidity observations in our experimental setup, the logarithm of humidity has perfect Gaussian error distribution. Therefore, it is a standard for the other choices of humidity observational variables to attain.

6.4.3 Assimilation of relative humidity (rh)

As in the assimilation of the $\ln(q)$ observations, the updated analysis variable is still the specific humidity. Different from $\ln(q)$ assimilation, the observation operator is linear. The observation increment is the difference between the observed relative humidity and the relative humidity from background. The observation increment is equal to $\mathbf{rh}^o - \mathbf{H}^l(\overline{\mathbf{rh}}^b)$. The ensemble perturbation at the observation locations is $\mathbf{HX}^b = \mathbf{H}^l(\mathbf{rh}^b) - \mathbf{H}^l(\overline{\mathbf{rh}}^b)$. The analysis mean state is equal to

$$\overline{\mathbf{q}}^a = \overline{\mathbf{q}}^b + \mathbf{X}^b\widetilde{\mathbf{K}}[\mathbf{rh}^o - \mathbf{H}^l(\overline{\mathbf{rh}}^b)] \tag{6.5}$$

The analysis perturbations are also directly calculated from equation (6.2)

6.4.4 Assimilation of pseudo-Relative Humidity (pseudo-RH)

As stated earlier, pseudo-RH is the ratio between the observed specific humidity and the saturated specific humidity from background. So far, it has only been applied to variational approaches (Dee and da Silva, 2003; Holm et al., 2002). In the ensemble Kalman filter, since we have an ensemble of possible saturated specific

humidity from background forecast, we normalize the specific humidity observations by the mean saturated specific humidity $\overline{\mathbf{q}}^{sb}$ at the observation locations, then

$$\mathbf{y}^o = \mathbf{E}^{-1}\mathbf{q}^o, \quad \mathbf{E} = diag(h^l(\overline{\mathbf{q}}^{sb})) \tag{6.6}$$

where $h^l(\cdot)$ is the linear interpolation operator. The corresponding background pseudo-RH is equal to

$$h^l(\overline{\mathbf{x}}^b) = h^l(\frac{1}{K}\sum_{i=1}^{K}\mathbf{D}^{-1}\mathbf{q}_i^b), \ \mathbf{D} = diag(\overline{\mathbf{q}}^{sb}) \tag{6.7}$$

The specific humidity ensemble perturbations at the observation locations are normalized by the mean saturated specific humidity, expressed as follows:

$$\mathbf{HX}_i^b = h^l[\mathbf{D}^{-1}(\mathbf{q}_i^b - \overline{\mathbf{q}}^b)] \tag{6.8}$$

The background ensemble perturbations are still the perturbations of specific humidity. In applying the observation operator, we first normalize the specific humidity perturbations by the background mean saturated specific humidity. The reason lies in the fact that spatial variability of relative humidity is less than specific humidity, so the spatial interpolation of pseudo-RH is more accurate than that of specific humidity. Following the derivation of Dee and da Silva (2003), it is easy to get the analysis mean state as:

$$\overline{\mathbf{q}}^a = \overline{\mathbf{q}}^{\mathrm{b}} + \mathbf{DX}^{\mathrm{b}}\widetilde{\mathbf{K}}[\mathbf{y}^o - h^l(\overline{\mathbf{x}}^b)] \tag{6.9}$$

The analysis perturbations are also directly calculated from equation (6.2)

6.5 Results

We first present the results from *uni-q* experiments with different humidity variables and then the results from *coupled* experiments.

6.5.1 Assimilation results from *uni-q* experiments

Figure 6.6 shows 700hPa global average specific humidity Root Mean Square (RMS) error comparison between *control* run and the different choices of humidity variables. Though the specific humidity analysis RMS error from *control run* (black line) is much smaller than the specific humidity observation error (Figure 6.3) in this OSSE experimental setup, the improvement from assimilating humidity observations is still significant with an appropriate choice of humidity variable type. Not surprisingly, the choice of $\ln(q)$ gives the best result (red line in Figure 6.6), since it has perfect Gaussian observation error distribution. However, in reality, $\ln(q)$ does not necessarily have a perfect Gaussian error distribution. In our current experimental setup, it is an ideal (optimal) result that the other choices of humidity variable types are aiming for. Among the other choices of humidity variable types, the best result is from pseudo-RH assimilation (the blue line in Figure 6.6). As shown in Figure 6.3, the error distribution of pseudo-RH is more Gaussian than specific humidity observations. It has similar error distribution as the relative humidity observations, but unlike relative humidity observations, it has no error correlation with the other observation variables. Therefore, the performance of pseudo-RH assimilation is better than both the relative humidity and the specific humidity observations. With the choices of specific humidity (green line) and relative humidity (purple line) variable types, the performance is similar to the *control* run, i.e., there is little improvement on the moisture analysis even though the moisture observations were used.

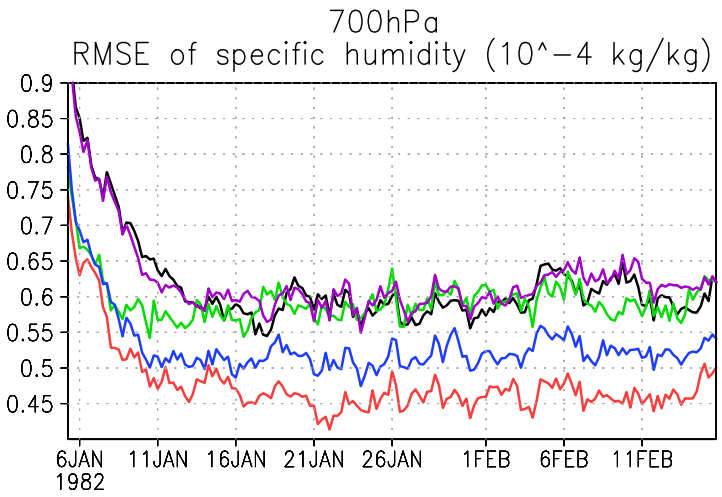


Figure 6.5 700hPa specific humidity RMS error comparison between different choices of the humidity observational type (black line: control run; green line: specific humidity; purple: relative humidity; blue line: pseudo-RH; red line: ln(q))

In *uni-q* experiments, though the humidity observations do not update the other dynamical variables during assimilation, the updated humidity field does have an influence on the other dynamical variables during the forecast through the parameterization processes. The specific humidity directly affects temperature forecast through the condensation and radiation process. It also affects winds through surface flux processes and interaction between planetary boundary layer and lower Troposphere. The analysis results for temperature (top panel in Figure 6.6) and zonal wind (bottom panel in Figure 6.6) have the same ranking as for the specific humidity, though the difference between different choices of humidity variable types is small except for the ln(q) case. It seems that only when the specific humidity analysis is much better than the control run (which happens only for the ideal ln(q) variable choice in the *uni-q* experiment), can it have significant impact on the other variables during the forecast process.

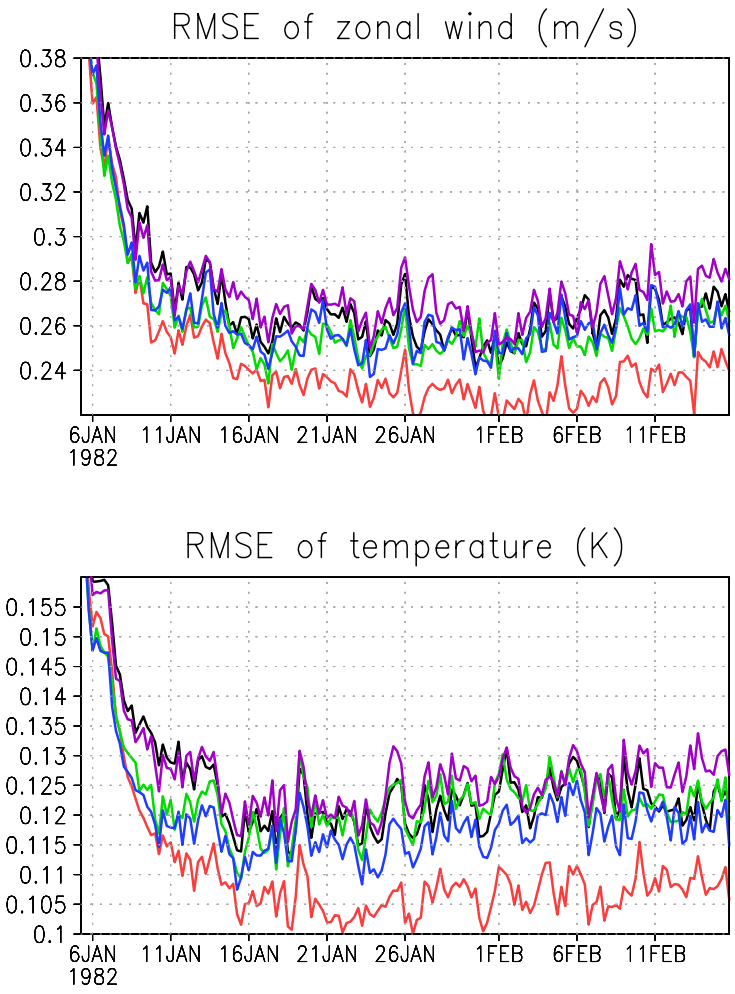


Figure 6.6 700hPa RMS error comparison between different choices of the observed humidity variables. Top panel: zonal wind (Unit: m/s); bottom panel: temperature (Unit: K). The line notation is same with Figure 6.5

The time average (over the last 20 days of analysis cycle) of the uni-variate analysis RMS error as function of vertical levels is shown in top panel of Figure 6.7. The zonal wind and temperature are the bottom two panels. The ranking of different choices of humidity variables is same as the RMS time series shown in Figure 6.5 and Figure 6.6, which is true in all vertical levels.

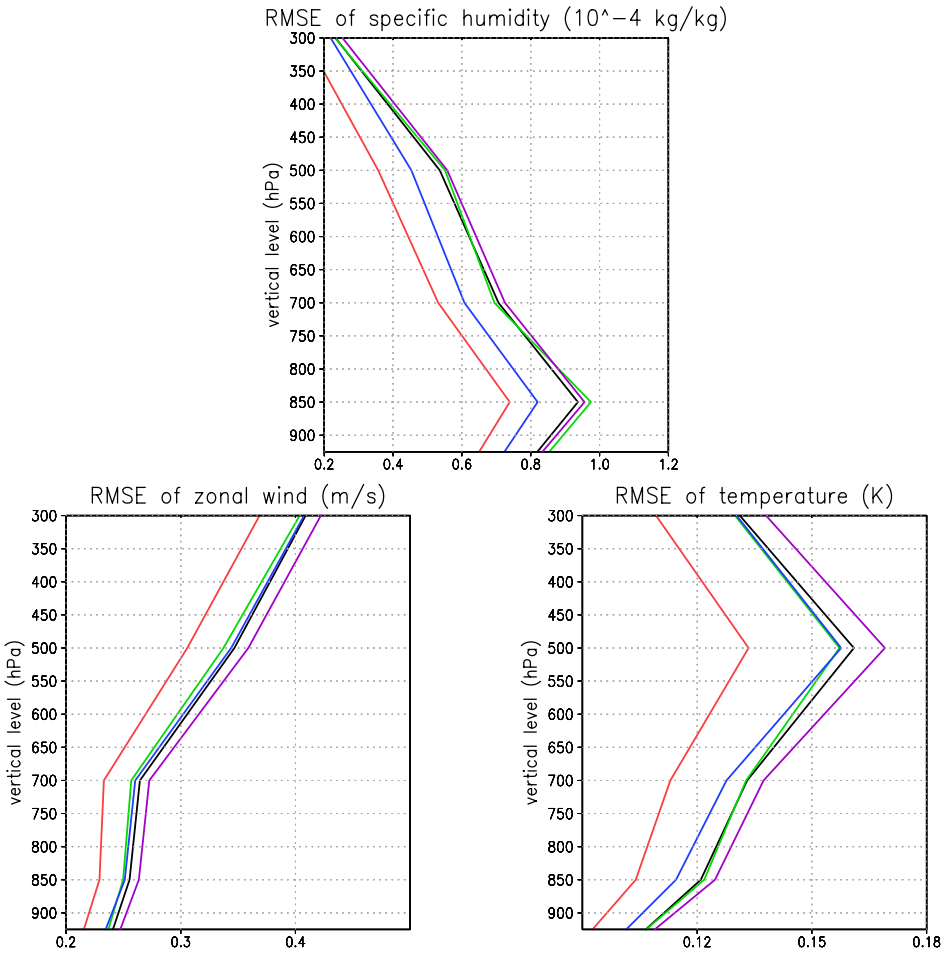


Figure 6.7 Uni-variate assimilation time average RMS error as function of vertical levels for specific humidity (Unit: 10^{-4} kg/kg, top panel), zonal wind (Unit: m/s, left bottom panel) and temperature (Unit: K, right bottom panel).

Spatially, assimilation of both $\ln(q)$ and pseudo-RH shows positive significant impact on the accuracy of specific humidity analysis over the tropics and the mid-latitudes (top two panels in Figure 6.8). The assimilation of relative humidity has neutral impact (bottom left panel in Figure 6.8), and the assimilation of specific humidity only makes the result slightly better in the mid-latitudes, but makes the results worse in the other regions (bottom right panel in Figure 6.8). This spatial pattern is related with both the observation error characteristics of each humidity

variable type and also the value distribution of specific humidity field itself. Since $\ln(q)$ observations have uniform observation error standard deviation over all the latitudes in the same level, and the observation error distribution is Gaussian, it has perfect error statistics. Therefore, the positive impact from assimilating $\ln(q)$ is largest. The specific humidity has larger error in the tropics than the other latitudes, so does the error reduction from assimilating the perfect $\ln(q)$ observations**.** The observation error of pseudo-RH does not change much spatially in the same level, so the single observation error standard deviation for each vertical level is reasonable. Therefore, the error reduction spatial pattern from assimilating pseudo-RH is similar with the assimilation of $\ln(q)$ observations. On the other hand, the actual observation error of specific humidity changes abruptly with latitude, but we still use a single value to represent the observation error statistics in each vertical level. Thus, the assimilation of specific humidity observations only has positive impact on the mid-latitudes. Though relative humidity has more uniform observation error distribution, it has strong error correlation with temperature and pressure that we do not consider during data assimilation, and, as a result, it has a neutral impact.

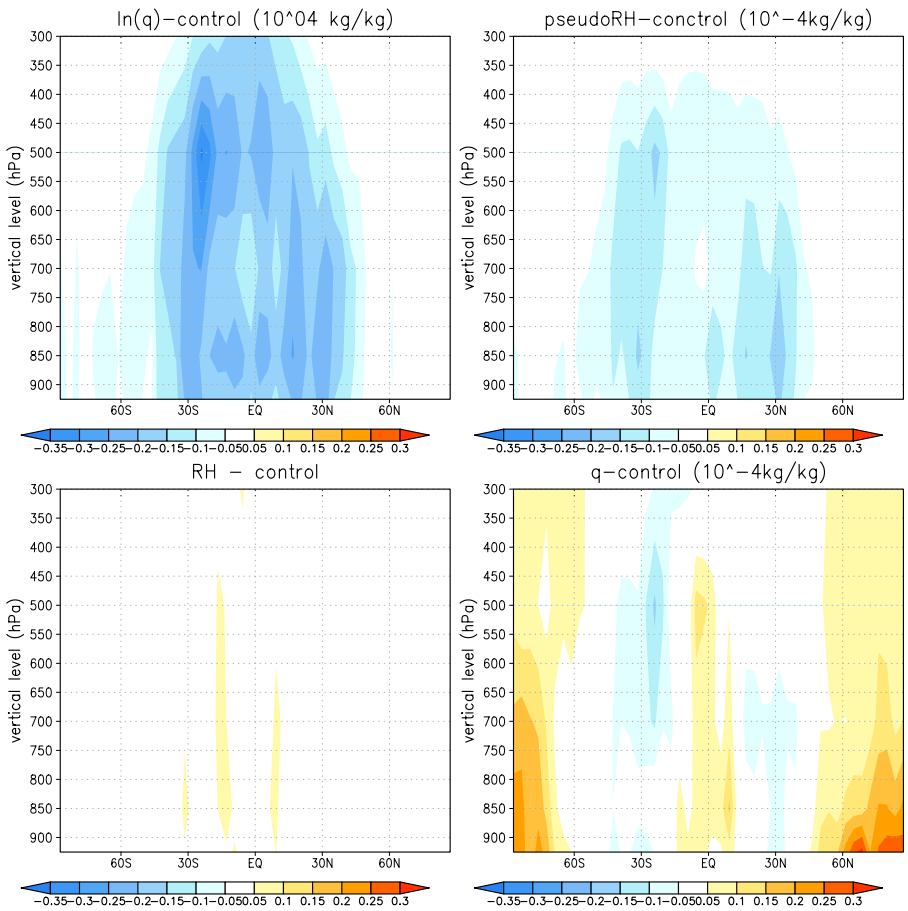


Figure 6.8 Zonal mean specific humidity analysis RMS error difference (Unit: 10^{-4} kg/kg) between different choices of humidity variable type and the control run (top left panel: ln(q); top right panel: pseudo-RH; bottom left panel: RH; bottom right panel: q).

One of the main purposes of assimilating humidity variables is to improve the precipitation forecast. The time averaged six-hour precipitation forecast error difference between *uni-q* experiment and *control* run for both large-scale precipitation and convective precipitation are shown in Figure 6.9 and Figure 6.10 respectively. Large-scale precipitation mainly concentrates over the storm track region, while convective precipitation mainly happens over the tropics. The results show that the assimilation of $\ln(q)$ observations has the largest positive impacts on both large scale and convective precipitation 6-hour forecast results. Though the positive impact from

assimilation of pseudo-RH is not as much as the assimilation of $\ln(q)$, it is still significant. In most areas, the assimilation of specific humidity and the relative humidity makes the 6-hour precipitation forecast worse.

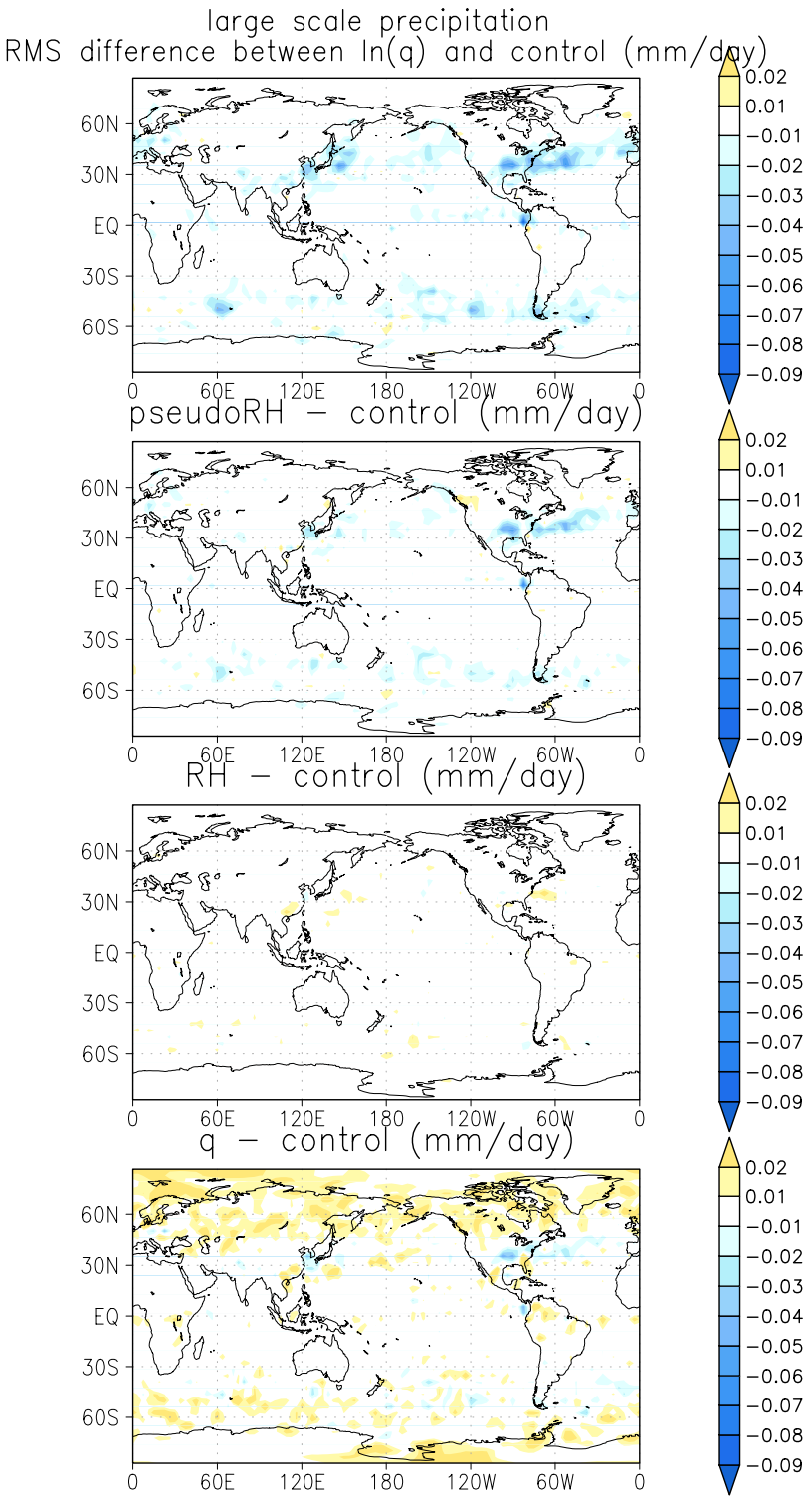


Figure 6.9 Time average (last twenty days) of large scale precipitation RMS error difference (Unit: mm/day) between different choices of the humidity variable types and the *control* run. (The first panel: ln(q)-*control*; second panel: pseudo-RH-*control*; third panel: RH-*control*; fourth panel: q-*control*).

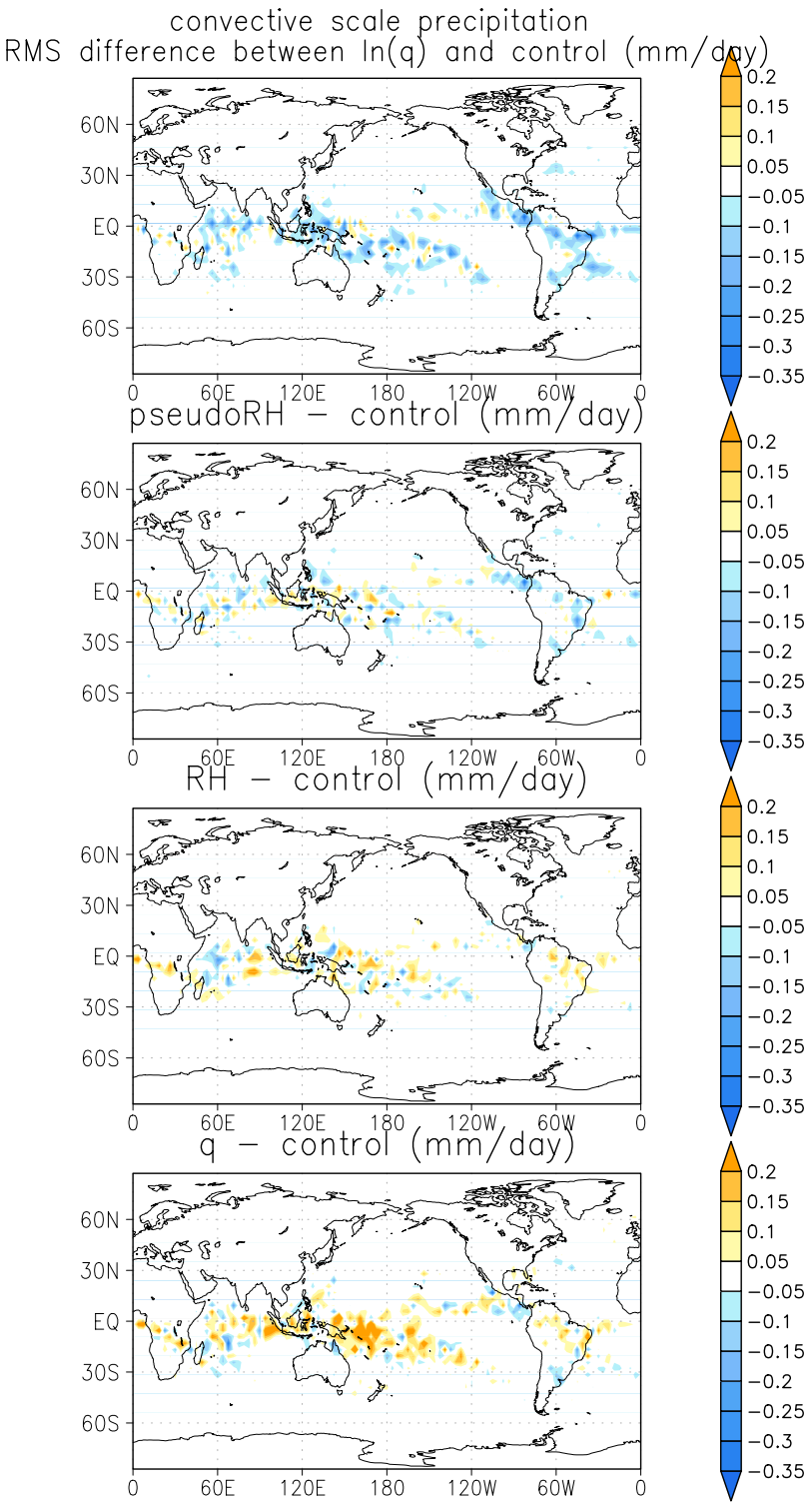


Figure 6.10 Time average (the last twenty days) of convective precipitation RMS error difference (Unit: mm/day) between different choices of humidity variable types and the *control* run. The sequence of the figure is same with Figure 6.9.

6.5.2 Assimilation results from *coupled* (multivariate) experiments

Specific humidity acts like a tracer except in condensation and evaporation processes, so we expect that the coupling between winds and specific humidity would have impact on both the winds analysis and the specific humidity analysis result. In addition, both the temperature and the specific humidity field are mass fields and closely related with each other, so that the coupling between them and their errors should also have impact on each other.

Figure 6.11 shows the specific humidity analysis RMS error comparison between *uni-q* experiment (light blue) and *coupled* experiment (magenta) for each choice of humidity variable type. For reference, we also include the result from control run (black line). It shows that the coupling improves the specific humidity analysis accuracy with every choice of humidity variable type except for the choice of relative humidity. For the relative humidity, the coupling shortens the spin-up time, but after the spin-up time, the performance is similar with *uni-q* experiment. This is due to the strong error correlation between relative humidity observations and the temperature observations that as is customary, we neglect during the data assimilation process. Unlike the other choices of humidity analysis types, the observational operator is nonlinear when the observation is $\ln(q)$. The nonlinear relationship between $\ln(q)$ observations and dynamical variables lengthens the spin-up time, but after the spin-up time, the *coupled* experiment has slightly better performance than the *uni-q* experiment. The coupling between humidity variable and the other dynamical variables significantly improves the specific humidity analysis result when

the assimilated humidity variable type is either specific humidity (bottom right panel) or pseudo-RH (top right panel).

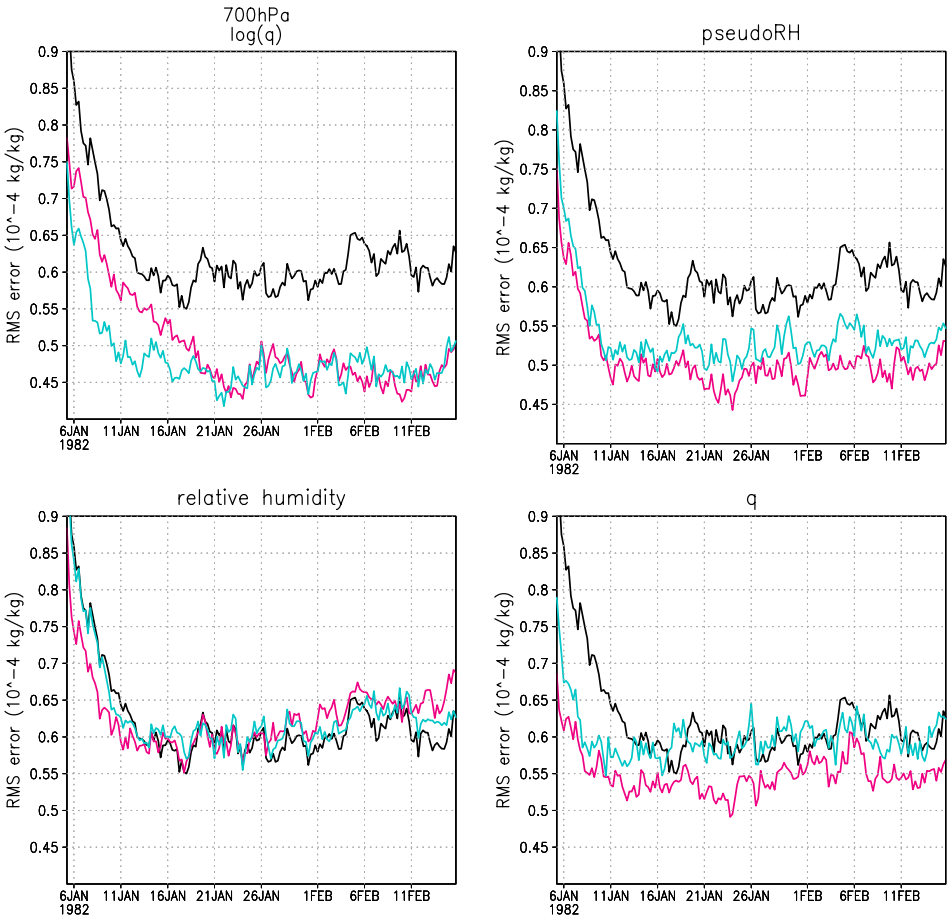


Figure 6.11 700hPa specific humidity RMS error (Unit: 10^{-4}kg/kg) comparison between the *uni-q* experiment (light blue) and the *coupled* experiment (magenta) for different choices of assimilated humidity variable types (top left: ln(q); top right: pseudo-RH; bottom left: RH; bottom right: q). The black line is from *control* run.

The coupling has an impact not only on the specific humidity field, but also on the other dynamical variables, such as zonal wind (Figure 6.12). Figure 6.12 shows that the coupling improves the zonal wind analysis accuracy for each choice of humidity variable, but it has largest improvement when the assimilated humidity variable type is pseudo-RH. It has only a slight impact with the other choices of

humidity analysis variables. The small impact with the choice of $\ln(q)$ is due to the nonlinear relationship between observations and the dynamical variables, while for specific humidity and relative humidity, it is related to the observation error characteristics that we discussed earlier.

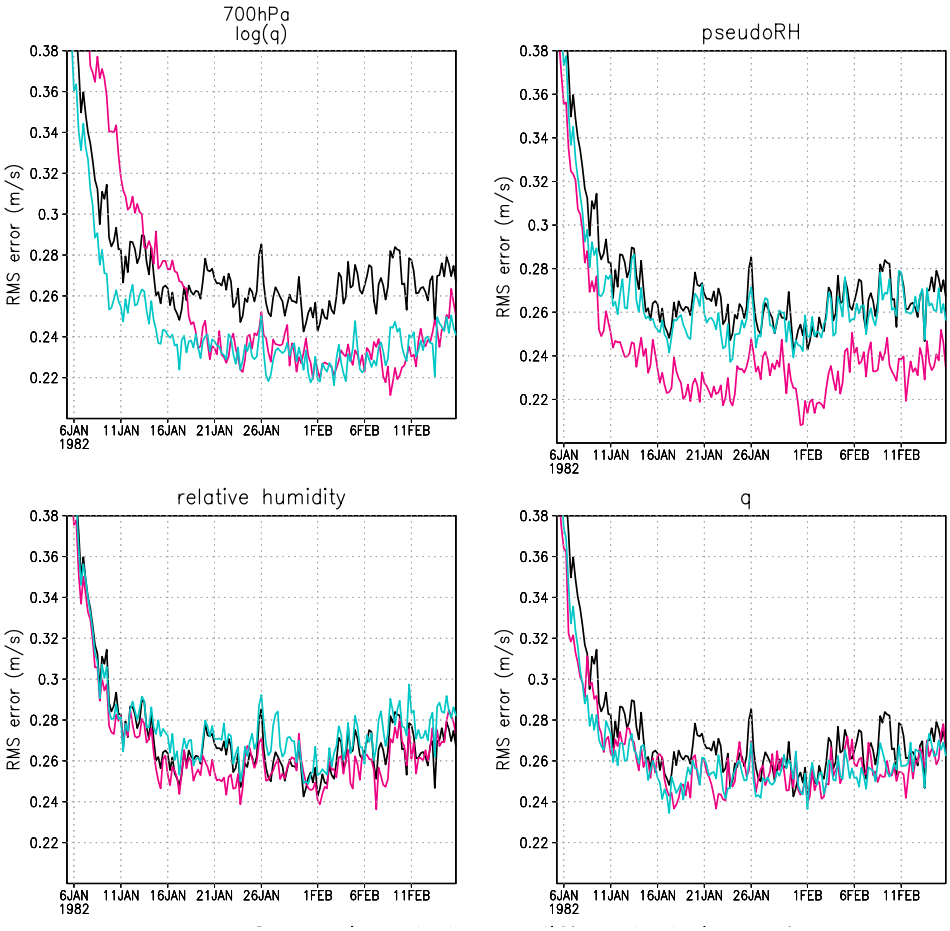
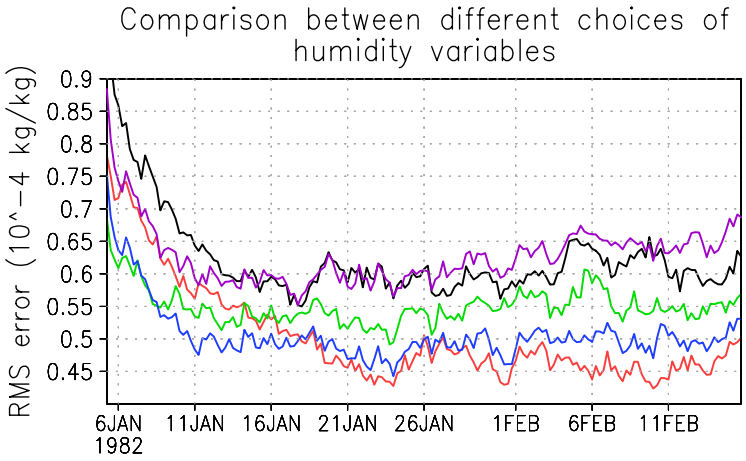


Figure 6.12 700hPa zonal wind RMS error (Unit: m/s) comparison between *uni-q* (light blue) and *coupled* experiment (magenta) for different choices of assimilated humidity variable types. The black line is from control run. The sequence is same with Figure 6.11.

Figure 6.13 shows 700hPa analysis accuracy comparison among different choices of humidity variable types for both specific humidity field (top panel) and the

zonal wind field (bottom panel) in *coupled* experiments. For both variables, the result from $\ln(q)$ (red line) is still the best. However the difference between the assimilation of pseudo-RH (blue line) and the assimilation of $\ln(q)$ becomes smaller compared to the difference in *uni-q* experiments (Figure 6.5). In addition, the choice of pseudo-RH is the best among all the other choices of humidity variables, i.e., specific humidity and relative humidity.

Figure 6.14 shows the RMS error comparison over all the vertical levels for specific humidity (top panel), zonal wind (bottom left panel) and temperature (bottom right panel). It shows that the ranking among different choices of humidity variable types over all the vertical levels is same with that of 700hPa (Figure 6.13) for specific humidity. However, for temperature and zonal wind, the choice of pseudo-RH is slightly better than $\ln(q)$ over the high levels.



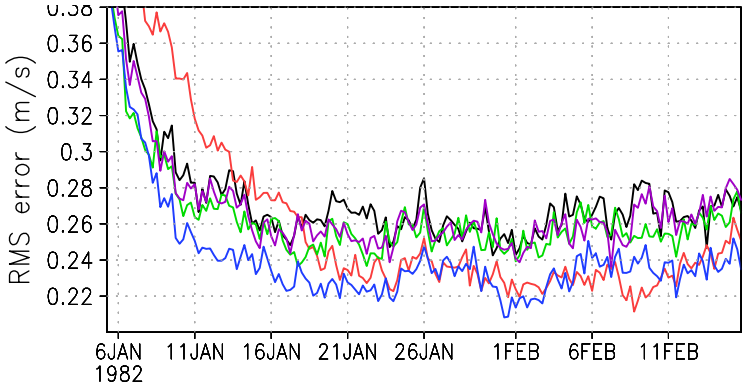


Figure 6.13 700hPa RMS error comparison from *coupled* experiments of different choices of assimilated humidity variable types (purple: RH; green: q; blue: pseudo-RH; red: ln(q); black: control run) for specific humidity (Unit: 10^{-4}kg/kg, top panel) and zonal wind (Unit: m/s, bottom panel)

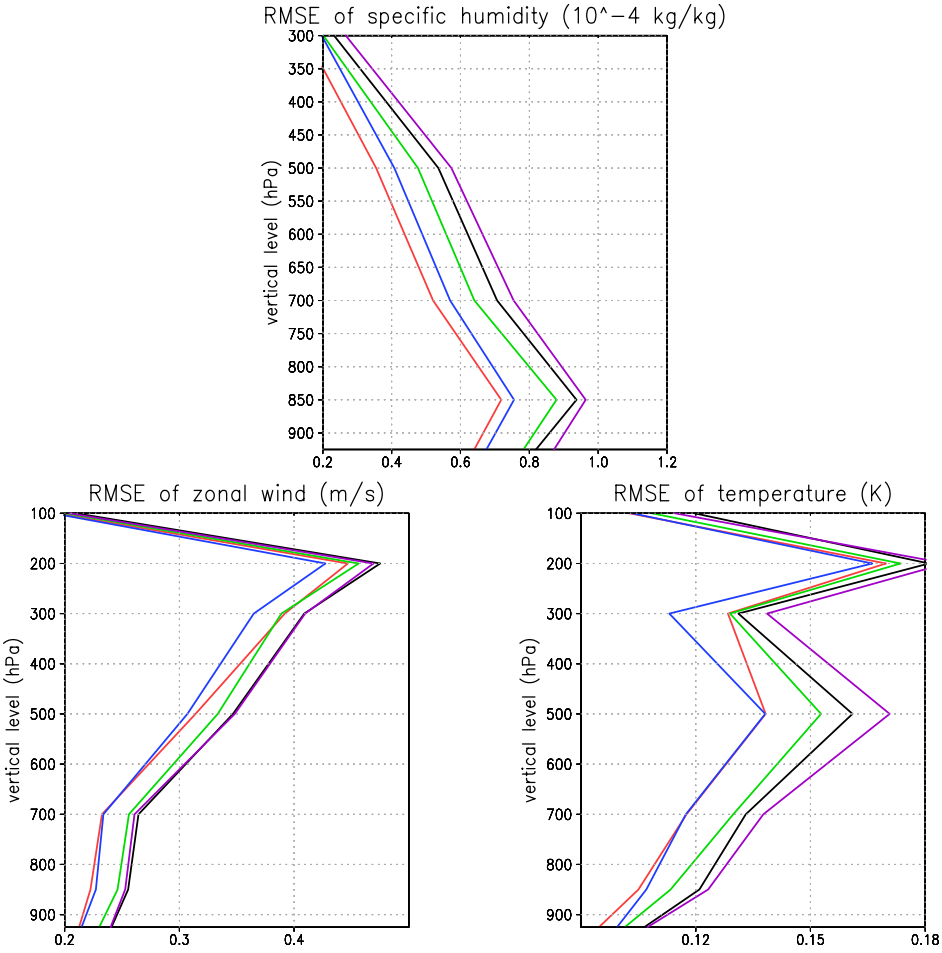


Figure 6.14 Multivariate analysis time average (last twenty days analysis cycle) RMS error as function of vertical levels for specific humidity (Unit: 10^{-4} kg/kg, top panel), zonal wind (Unit: m/s, left bottom panel) and temperature (Unit: K, right bottom panel). The line notation is same with Figure 6.13.

The coupling between humidity and the other dynamical variables during the assimilation process has an impact on the analysis accuracy of specific humidity and also on the other dynamical variables, which further affects the precipitation forecast accuracy. Since the coupling improves the analysis of the specific humidity and the other dynamical variables most with the pseudo-RH observations, the improvement of 6-hour precipitation forecast accuracy is also the largest with the choice of pseudo-RH. After coupling, the accuracy of precipitation forecast with the choice of pseudo-RH observations (second panel in Figure 6.15 and Figure 6.16) is comparable with the precipitation forecast from the choice of $\ln(q)$ (first panel in Figure 6.15 and Figure 6.16). With the choice of relative humidity (third panel in Figure 6.15 and Figure 6.16), the coupling slightly improves the precipitation forecast in the tropics, but makes the forecast worse in high latitudes. As stated earlier, the relative humidity error has strong correlation with temperature error, which is more significant in the higher latitude than in the tropics. With the choice of specific humidity variable type (bottom panel in Figure 6.15 and Figure 6.16), the coupling slightly improves the forecast compared to the forecast from the *uni-q* experiment (bottom panel in Figure 6.9 and Figure 6.10).

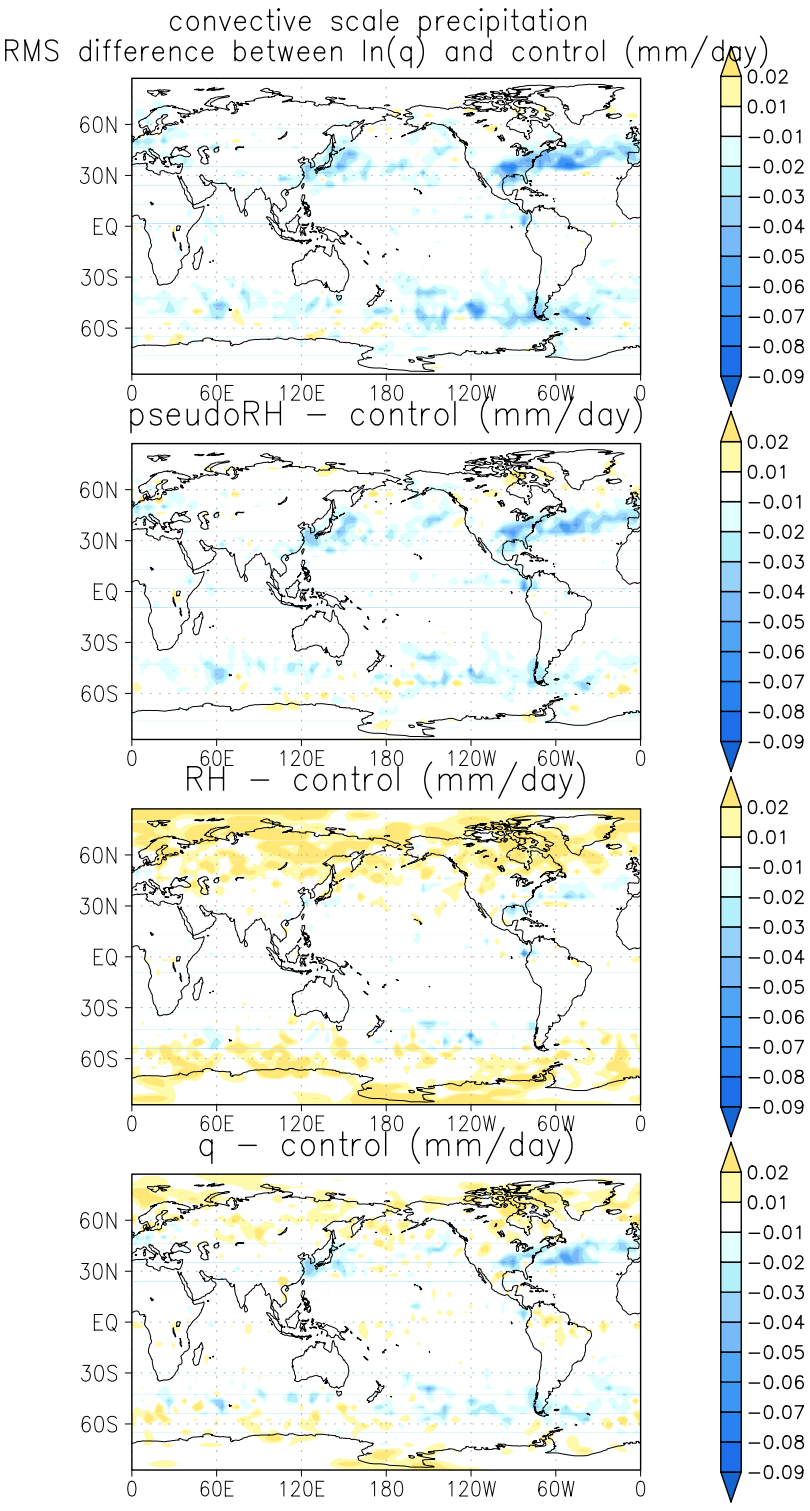


Figure 6.15 Time average of large scale precipitation RMS error difference (Unit: mm/day) between different choices of humidity variable type in the *coupled* experiments and the *control* run. (The first panel: ln(q)-*control*; second panel: pseudo-RH-*control*; third panel: RH-*control*; fourth panel: q-*control*).

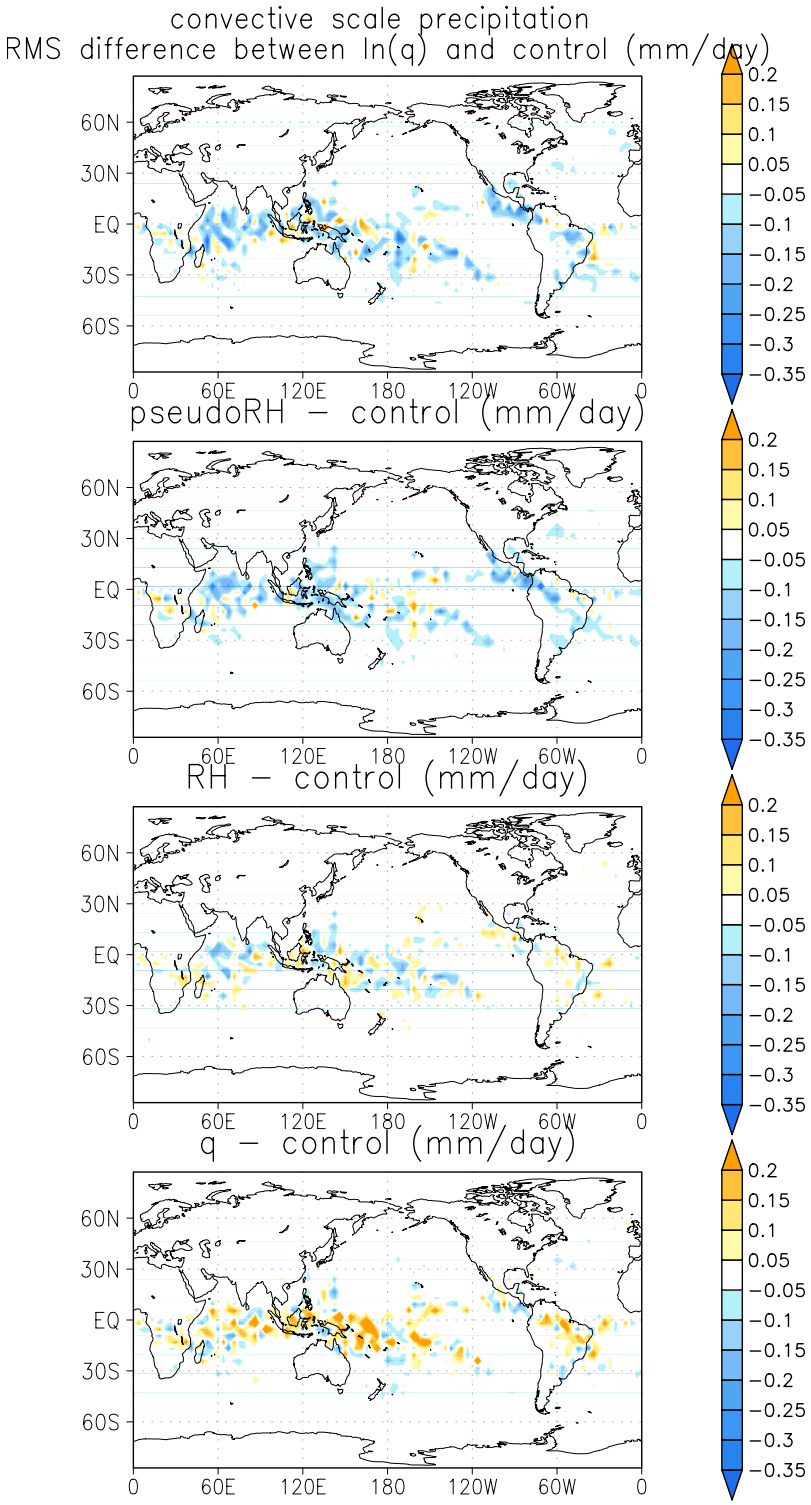


Figure 6.16 Same as Figure 6.15, except this is for the convective precipitation field.

6.6 Assimilation of AIRS humidity retrievals into the GFS LETKF data assimilation system

In this section, we show preliminary results from the assimilation of AIRS humidity retrievals provided by Chris Barnet (personal communication) with the choice of both specific humidity and pseudo-RH variable types. The previous simulation results reveal that the analysis accuracy from the *coupled* experiment (multivariate) is better than *uni-q* experiment. Therefore, we fully coupled the humidity variable types with the other dynamical variables. As far as we know, there have been few if any experiments with multivariate assimilation of humidity before this one.

6.6.1 Experimental design

The dynamical model is T64 resolution NCEP GFS system with 28 vertical levels. The data assimilation scheme is the 4D-LETKF (Hunt et al., 2004) with 6-hour assimilation window centered at the central time (Szunyogh et al., 2007). We have both the *control* run and the *humidity* run with 31 days analysis cycle. The assimilation period is January 2004. In the *control* run, the observation types include all the operational non-radiance non-humidity observations (Szunyogh et al., 2007) and the AIRS temperature retrievals. During data assimilation, the humidity dynamical variable is not updated. Since there is no humidity observation, the other dynamical variables are not updated by the humidity observations either. In the *humidity* run, we add the AIRS specific humidity retrievals as part of the observations between 30°S and 30°N, with the error standard deviation shown in Figure 6.17. We

only assimilate the specific humidity between 30°S and 30°N is because the error standard deviation was specially tuned for this area. In addition, we experimentally found unreasonable large analysis increments in the high latitudes with the choice of specific humidity variable type when we include the specific humidity as part of the dynamical variable in agreement with the simulation experiments. We will explore the reasons more in detail in the near future. Since we neglect the error correlations between different vertical levels when we assimilate specific humidity retrievals with the choice of specific humidity variable type, we double the error standard deviations in the data assimilation process. The same is true when we assimilate the temperature retrievals. The verification is done against the high resolution operational analysis, which is the NCEP GFS T254L64 operational analysis system with the assimilation of all operational observation data set. We will compare the RMS error between the *control* run and the *humidity* run with the choice of both specific humidity and pseudo-RH.

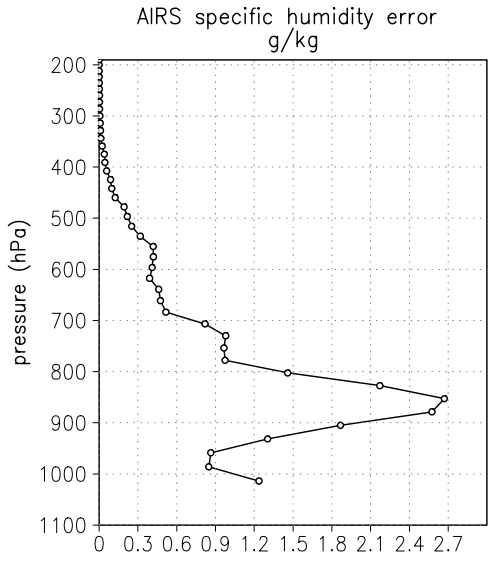


Figure 6.17 AIRS specific humidity retrievals error standard deviation (Unit: g/kg) as function of vertical levels (provided by Eric Maddy and Chris Barnet).

6.6.2 Results

Figure 6.18 shows that the assimilation of specific humidity with the choice of specific humidity variable type mainly affects the relative humidity analysis within 30°S and 30°N where there are observations. It improves the relative humidity result in the upper troposphere of the tropics, but makes the result worse in the lower levels of the tropics. The impact is neutral in the other regions. The worse result in the tropical lower levels may be due to the data quality (Figure 6.17), the non-Gaussian observation error characteristics of the specific humidity observations, or the model errors related with the parameterization process, and needs further investigation. The assimilation of AIRS specific humidity retrievals in the coupled mode has little impact on the temperature analysis, slightly improving the analysis in the higher tropics and the high latitudes of the Northern Hemisphere (Figure 6.19). With the choice of specific humidity variable type, it has a larger positive impact on the zonal wind analysis result (left panel of Figure 6.20), improving the zonal wind analysis accuracy in the most of tropics, and even improving the analysis in the high latitudes of the Northern Hemisphere where we do not assimilate specific humidity retrievals, which may be due to the coupling interaction between specific humidity and the other dynamical variables during the data assimilation propagated by the dynamics. With the choice of pseudo-RH (right panel of Figure 6.20), the wind analysis accuracy is further improved, with a positive impact on winds analysis almost everywhere.

The assimilation of AIRS specific humidity retrievals is preliminary, but it is consistent with the SPEEDY results which show neutral impact with the choice of specific humidity variable type, and significant positive impacts on wind analysis with the choice pseudo-RH.

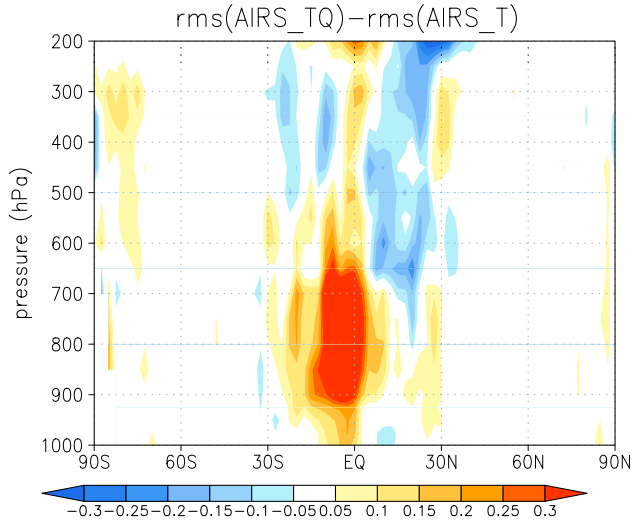


Figure 6.18 Relative humidity RMS error difference (Unit: 10%) between the *humidity* run and the control run.

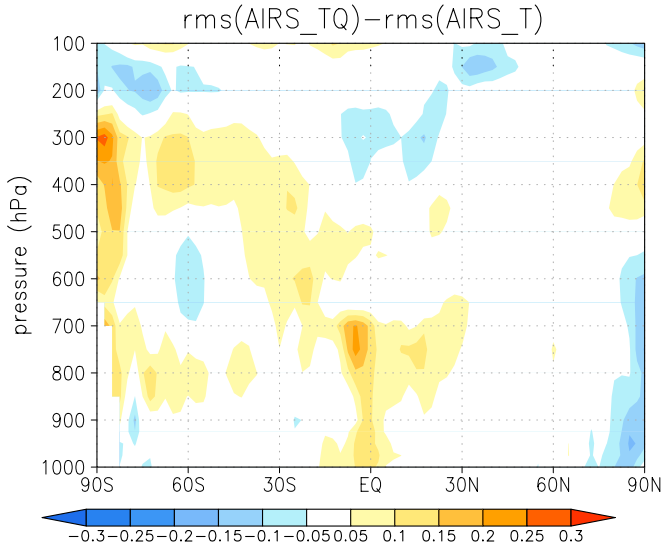


Figure 6.19 Zonal mean time average (averaged over the last twenty days analysis cycle) RMS error difference between *humidity* run and the *control* run for temperature (Unit: K, top panel).

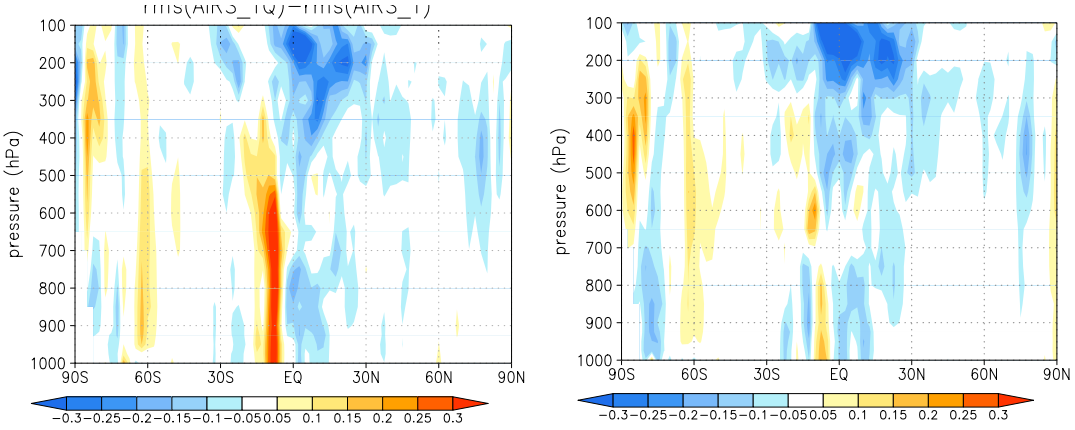


Figure 6.20 Zonal mean time average (averaged over the last twenty days analysis cycle) RMS error difference between *humidity* run and the *control* run for zonal wind (Unit: m/s, assimilated variable, left panel: specific humidity, right panel: pseudo-RH)

6.7 Conclusions and discussion

Due to the highly variable, both spatially and temporally, and non-Gaussian error characteristics of humidity variables, the assimilation of humidity observations is a challenging problem. So far, it has been assimilated uni-variately in operational centers.

The LETKF, as any other EnKF, estimates the time-changing background error covariance and at the same time, automatically couples all the dynamical variables together. Therefore, it is a good choice for the multivariate assimilation of humidity variables. The LETKF, as most other assimilation schemes, assumes Gaussian observation error distribution, while the humidity has the least Gaussian error distribution among all the dynamical variables. Therefore, the choice of humidity observational type is very important. In this Chapter, we compared several choices of humidity variable types when the specific humidity has non-Gaussian

observation error in both *uni-q* experiments and the *coupled* experiments with the SPEEDY model.

By adding Gaussian random error to the logarithm specific humidity, we create simulated specific humidity observations with non-Gaussian observation error distribution, as well as other choices of humidity variables, such as relative humidity and the pseudo-RH proposed by Dee and da Silva (2003). Since the logarithm specific humidity has perfect error statistics, its results are an optimal goal for the other choices of humidity variables to attain. Statistically, pseudo-RH and relative humidity have more Gaussian observation error distribution than specific humidity observations. Compared to the choices of relative humidity or specific humidity, the choice of pseudo-RH has a better analysis result for both specific humidity and the other dynamical variables in both *uni-q* experiment and the *coupled* (multivariate) experiment. It has a performance similar with the choice of logarithm of specific humidity in the *coupled* experiment. The poor result from the assimilation of relative humidity is due to the high correlation between the relative humidity and the temperature observation errors, which we neglect during the data assimilation. For the choice of specific humidity observations, the poor performance is due to the highly spatially variable error characteristics and the significant non-Gaussian observation error characteristics. Overall, this OSSEs experiment shows that pseudo-RH is a better choice for the assimilation of humidity observations, and at the same time, the automatically coupled assimilation between humidity and the other dynamical

variable with the LETKF data assimilation scheme improves the analysis compared to the *uni-q* experiment.

We assimilated real AIRS specific humidity retrievals with NCEP GFS 4D-LETKF assimilation system with the choice of both specific humidity and pseudo-RH variable types in a coupled (multivariate) mode. The preliminary results show that, with the choice of specific humidity variable type, the assimilation of AIRS specific humidity retrievals has a positive impact on the upper tropics of the relative humidity field, neutral impact on the temperature field and positive impact on the zonal wind field in most of the tropics and the Northern Hemisphere. With the choice of pseudo-RH, the analysis accuracy is further improved, and the impact on winds analysis is positive almost everywhere. These results are very promising though we still need to further explore the reason for the poor performance in the lower level tropics.

Chapter 7 Summary and future plans

Our work covered four application areas of the LETKF data assimilation scheme, and in each of these, we obtained encouraging new results.

7.1 Adaptive observations

A straightforward application of the LETKF is to do adaptive observations since the LETKF outputs the background and local analysis uncertainty along with the data assimilation scheme. The background ensemble spread adaptive strategy, which minimizes the trace of the background error covariance, is cost-free but not optimal for more than one adaptive observation. The local analysis ensemble spread method, which can be computed in parallel, minimizes the trace of the analysis ensemble spread. It is optimal for multiple adaptive observations, but can be very expensive even with parallel computation. The combined-background-analysis ensemble spread method selects a few "promising" observations first based on the background ensemble spread, and then the local analysis ensemble spread method is only applied to the observations selected by the background ensemble spread method. It combines the advantages of both methods.

We first compared background ensemble spread method, local analysis ensemble spread method, combined-background-analysis ensemble spread method, and an 'ideal' method based on the truth on the Lorenz-40 variable model. The background ensemble spread method, local analysis ensemble spread method and combined-background-analysis ensemble spread give the same accuracy when only

one adaptive observation is chosen, and are all better than the best result (Hansen and Smith, 2000) published so far with the same experimental setup. Based on two simple examples, we show that the background ensemble spread method is equivalent with the local analysis ensemble spread method only when only one adaptive observation is to be selected and it is the same type with the dynamical variable, and also at a grid point. Otherwise, the results from these two methods would be different, and only the analysis ensemble spread method would be optimal.

DWL is an active sensor strongly constrained by energy resources, and the U.S. instrument is planned to be operated in an adaptive mode. An often stated goal is to 'get 90% improvement from 10% observation coverage'. We adaptively sampled simulated Doppler Wind Lidar (DWL) observations in both 3D-Var and the LETKF assimilation system in a global primitive equation model. We compared the background ensemble spread method with several other sampling strategies, namely, uniform distribution, random sampling, climatological ensemble spread, and an 'ideal' method based on the 'truth'. The LETKF-based ensemble spread method avoiding the choice of neighboring observations gives the best result among the operational possible adaptive methods we tested. With 10% adaptive observations obtained from the LETKF-based ensemble spread, both 3D-Var and LETKF can get more than 90% improvement, showing that the LETKF-selected locations correspond to the areas of instability where errors grow faster. 3D-Var is as effective as the LETKF with 10% coverage, but the LETKF is more effective when only 2% DWL footprints are selected. With 10% adaptive observations, it is sufficient to give

information about the 'error of the day' to 3D-Var, while 2% adaptive observations are not sufficient. On the other hand, since the LETKF scheme already knows 'error of the day', it is not so sensitive to the adaptive observation strategies with 10% adaptive observations. The ensemble spread method is superior to the other methods within the LETKF when only 2% adaptive observations are observed.

7.2 Self-sensitivity

Self-sensitivity is the diagonal value of the influence matrix which is the Kalman gain in observation space and can be computed at little additional cost within LETKF scheme. Self-sensitivity reflects how sensitive of the analysis to observations, so that it is also known as analysis sensitivity. Self-sensitivity is complementary to the analysis sensitivity to the background (the sum is equal to one). Following the formulation of Cardinali et al. (2004), we proposed to calculate self-sensitivity within the LETKF. However, since the LETKF produces a local analysis, and the observations are used in several local patches, the self-sensitivity for a given observation would be different if it is in different local patches. Therefore, we proposed to average the self-sensitivity with respect to the same observation in different local patches together. We verified our averaged scheme by comparing the self-sensitivity calculated from the LETKF and the global ETKF which does not require averaging. The results show that the averaging scheme gives similar results as the self-sensitivity calculated from ETKF. Unlike the self-sensitivity calculated in the 4D-Var system (Cardinali et al., 2004), the self-sensitivity within the LETKF is not approximated, and satisfies the theoretical value limits (between 0 and 1). In agreement with a geometrical analysis, we showed experimentally that the self-

sensitivity is proportional to the analysis error, and that is anti-correlated with the observation error.

The trace of self-sensitivity of any subset observations is the information content of that subset. It can be used to assess the spatial importance of the same type observations. With the SPEEDY model, we compared the information content from *all-obs* control experiment and the quantitative observation impact calculated from data denial experiments, and showed that the information content qualitatively reflects the spatial observation impact calculated from data denial experiments. By comparing the information content calculated in *rawinsonde-only* control experiment based on the possible future observation locations with the actual observation impact from the "add-on" experiments, we showed that information content also qualitatively reflects the observation impact in the "add-on" experiments. This implies that the spatial information content can be utilized in observation design experiments (without carrying out data impact experiments), and can also be used to compare the information content of the instruments that measure the same type of observations.

7.3 Observation impact

Langland and Baker (LB, 2004) pioneered an approach to monitor the observation impact in a derivation based on the adjoint model. The observation impact can help identify the observations that deteriorate the forecast, and better use the observations that have large impact on the forecast.

Following LB (2004), we proposed an ensemble sensitivity method to measure the observation impact on the error difference between the forecasts initialized from 00hr and -6hr. Unlike the adjoint method by LB (2004), the ensemble sensitivity method we propose does not require the adjoint model. We compared the ensemble sensitivity method we proposed to the adjoint model using Lorenz-40 variable model. The results show that the ensemble sensitivity method gives results similar to the adjoint method, and both can explain more than 90% forecast error difference in our experimental setup. Both methods can detect "bad" observations that are of poor quality, with either larger random errors than specified or with bias, and the ensemble sensitivity method shows stronger signal in such scenarios. Like the adjoint method by LB, this method can be applied in the observation quality control as well as comparing the importance of different type observations. It can be used to quantitatively estimate the impact on the forecast of a certain observation type or locations. It could be routinely calculated as part of the analysis cycle, thus providing a powerful tool to understand cases of forecast failure and a tool to tune the observation error statistics.

7.4 Humidity assimilation

Because humidity is highly variable, both spatially and temporally, and with non-Gaussian error characteristics, the assimilation of humidity observations is a challenging problem. So far, it has been assimilated uni-variately in operational NWP centers with variational data assimilation schemes. However, unlike the variational data assimilation schemes, the LETKF, as any other EnKF, estimates the time-changing background error covariance and at the same time, automatically couples all

the dynamical variables together. Therefore, it is a good choice for the assimilation of humidity variables, and automatically coupling humidity variable with the other dynamical variables in the data assimilation.

Since humidity variable is the least Gaussian variable type, the choice of assimilated variable is very important. We compared several choices of humidity variable type when the specific humidity has non-Gaussian observation error in both *uni-q* experiment and the *coupled* experiment with the SPEEDY model. In *uni-q* experiment, the humidity variable is updated by itself, which is the way it is done in operational NWP centers. In *coupled* (multivariate) experiment, the humidity variable is fully coupled with the other dynamical variables. The humidity variable types include specific humidity, logarithm specific humidity, relative humidity and pseudo-RH proposed by Dee and da Silva (2003). As far as we know, this is the first attempt to assimilate pseudo-RH within an EnKF.

By adding the Gaussian random error to the logarithm specific humidity, we created simulated specific humidity observations with non-Gaussian observation error distribution, as well as other choices of humidity variables, such as relative humidity and pseudo-RH. Since the logarithm specific humidity has perfect error statistics, its results are an optimal goal for the other choices of humidity variables to attain. Statistically, pseudo-RH and relative humidity have more Gaussian observation error distribution than specific humidity observations. Compared to the choices of relative humidity or specific humidity, the choice of pseudo-RH has a better analysis result

for both specific humidity and the other dynamical variables in both *uni-q* experiment and the *coupled* (multivariate) experiment. It has a performance similar with the choice of logarithm specific humidity in the *coupled* experiment. The poor result from assimilation of relative humidity is due to the high correlation between the relative humidity and the temperature observation errors that are neglected during data assimilation. For the choice of specific humidity observations, the poor performance is due to the highly spatially variable error characteristics and the significant non-Gaussian observation error characteristics. Overall, this OSSEs experiment shows that pseudo-RH is a better choice for the assimilation of humidity observations, and at the same time, the automatically coupled assimilation between humidity and the other dynamical variable with the LETKF data assimilation scheme improves the analysis compared to the *uni-q* experiment.

We assimilated real AIRS specific humidity retrievals with NCEP GFS 4D-LETKF assimilation system with the choice of both specific humidity and pseudo-RH variable types in a coupled (multivariate) mode. The preliminary results show that, with the choice of specific humidity variable type, the assimilation of AIRS specific humidity retrievals has a positive impact on the upper tropics of the relative humidity field, neutral impact on the temperature field and positive impact on the zonal wind field in most of the tropics and the Northern Hemisphere. With the choice of pseudo-RH, the analysis accuracy is further improved, and the impact on winds analysis is positive almost everywhere. These results are very promising though we still need to further explore the reason for the poor performance in the lower level tropics.

7.5 Future plans

Since many studies in my thesis are the first attempts to do research in that area with the LETKF scheme, they based on the simulated experimental setup in a simple model. We will further explore the applications of these theoretical studies in a more realistic system, especially, the application of observation impact, and the assimilation of humidity variables.

Appendix A Local Online Inflation Estimation Scheme

Most of these equations are from Miyoshi (2005), and the approach is valid if the error of the observation is accurate (Li, 2007). The covariance of the observational innovation **d** (difference between forecast and observation) has the statistical relationship (Houtekamer et al. 2005):

$$\left\langle \mathbf{d}\mathbf{d}^T \right\rangle = \mathbf{H}\mathbf{P}^f\mathbf{H}^T + \mathbf{R} \qquad \text{(A.1)},$$

Where H and R is the linearized observation operator and the observational error covariance respectively. $\langle \bullet \rangle$ represents the statistical mean state. Inflating equation (A.1) on the background error covariance, it becomes:

$$\left\langle \mathbf{d}\mathbf{d}^T \right\rangle = (1+\delta)\mathbf{H}\mathbf{P}^f\mathbf{H}^T + \mathbf{R} \qquad \text{(A.2)}$$

Summing over the trace of the error covariance, the inflation factor δ is estimated as:

$$\delta = 1 - \frac{sum(\left\langle \mathbf{d}\mathbf{d}^T \right\rangle) - sum(\mathbf{R})}{sum(\mathbf{H}\mathbf{P}^{\mathbf{f}}\mathbf{H}^T)} \qquad \text{(A.3)}$$

To avoid unreasonable values, we restrict δ within a reasonable range, which is between -0.1 and 0.48.

To reduce the sampling error, a simple scalar Ensemble Kalman filter is used to estimate the final inflation factor. δ estimated from equation (A.3) is used as observation δ^o in the Kalman filter estimation. When there is no observation in the local patch, δ^o is set to be the same with δ^f . The final inflation factor is:

$$\delta^a = \frac{v^o\delta^f \_ v^f\delta^o}{v^f + v^o} \qquad \text{(A. 4)}$$

$v^{f/o}$ is the forecast/observational error variance. The analysis error variance is obtained by:

$$v^a = (1 - \frac{v^f}{v^f + v^o})v^f \tag{A.5}$$

In the forecast step, the inflation factor and the error variance are both updated as:

$$\delta_{i+1}^f = \delta_i^a \tag{A.6}$$

$$v_{i+1}^f = (1 + \Delta)v_i^a \tag{A.7}$$

where Δ is the forecast factor to evolve the analysis error covariance.

Appendix B

B.1 Perturbation weights averaged over the ensemble

The following derivation is based on Hunt et al. (2007). We define a column vector of K ones: $\mathbf{v} = (1, \ 1, \ \cdots, \ 1)^T$. $\mathbf{v}$ is an eigenvector of $\widetilde{\mathbf{P}}^a$ with eigenvalue $(K-1)^{-1}$: $(\widetilde{\mathbf{P}}^a)^{-1}\mathbf{v} = \left[(K-1)\mathbf{I} + (\mathbf{Y}^b)^T\mathbf{R}^{-1}\mathbf{Y}^b\right]\mathbf{v} = (K-1)\mathbf{v}$ because the sum of the columns of $\mathbf{Y}^b$ is zero. Therefore,

$$\frac{1}{K-1}(\widetilde{\mathbf{P}}^a)^{-1}\mathbf{v} = \mathbf{v} \tag{B.1}$$

In addition, the matrix of analysis weights is given by $\mathbf{W}_0^a = [(K-1)\widetilde{\mathbf{P}}_0^a]^{1/2}$, so that

$$\mathbf{W}_0^a\mathbf{W}_0^{aT} = (K-1)\widetilde{\mathbf{P}}_0^a \tag{B.2}$$

Multiplying both sides by the vector $\mathbf{v}$, we get $\mathbf{W}_0^a\mathbf{W}_0^{aT}\mathbf{v} = \mathbf{v}$, so that $\mathbf{v}$ is an eigenvector of $\mathbf{W}_0^a\mathbf{W}_0^{aT}$ matrix with eigenvalue equal to 1. Based on the properties of

a symmetric matrix, $\mathbf{v}$ is also an eigenvector of $\mathbf{W}_0^a$ matrix with the eigenvalue equal to 1:

$$\mathbf{W}_0^a \mathbf{v} = \mathbf{v} \tag{B.3}$$

Since $\mathbf{v}$ is a column vector of K ones, $\sum_{i=1}^{K} \delta w_0^{a(ji)} = 1$, where $\delta w_0^{a(ji)}$ is an element of the $\mathbf{W}_0^a$. $\mathbf{W}_0^a$ is a symmetric matrix, therefore, we have the following equation:

$$\sum_{j=1}^{K} \delta w_0^{a(ji)} = \sum_{i=1}^{K} \delta w_0^{a(ji)} = 1 \tag{B.4}$$

B.2 Derivation of the observation impact

This derivation is based on Bishop (2007) and Langland and Baker (2004, referred as LB hereafter). As in the derivation by LB, we define a cost function which is the error difference between the short range forecast initialized from the analysis time and a forecast initialized from 6-hour earlier:

$$J = \frac{1}{2}(\boldsymbol{\varepsilon}_{t|0}^T \boldsymbol{\varepsilon}_{t|0} - \boldsymbol{\varepsilon}_{t|-6}^T \boldsymbol{\varepsilon}_{t|-6}) \tag{B.5}$$

where $\boldsymbol{\varepsilon}_{t|0} = \mathbf{x}_{t|0}^f - \mathbf{x}_t^a$, and $\boldsymbol{\varepsilon}_{t|-6} = \mathbf{x}_{t|-6}^f - \mathbf{x}_t^a$. As in LB, we verified the forecast at time t against the analysis at that time $\mathbf{x}_t^a$. Then equation (B.5) can be written as:

$$\begin{aligned} J &= \frac{1}{2}(\boldsymbol{\varepsilon}_{t|0}^T \boldsymbol{\varepsilon}_{t|0} - \boldsymbol{\varepsilon}_{t|-6}^T \boldsymbol{\varepsilon}_{t|-6}) \\ &= \frac{1}{2}(\boldsymbol{\varepsilon}_{t|0}^T + \boldsymbol{\varepsilon}_{t|-6}^T)(\boldsymbol{\varepsilon}_{t|0} - \boldsymbol{\varepsilon}_{t|-6}) \\ &= \frac{1}{2}[(\boldsymbol{\varepsilon}_{t|0}^T + \boldsymbol{\varepsilon}_{t|-6}^T)\mathbf{M}_{t|0}(\boldsymbol{\varepsilon}_0 - \boldsymbol{\varepsilon}_{0|-6})] \end{aligned} \tag{B.6}$$

$\mathbf{M}_{t|0}$ is the tangent linear model starting at time 00hr. We define $\boldsymbol{\varepsilon}_0 = \mathbf{x}_0^a - \mathbf{x}_0^{truth}$ and $\boldsymbol{\varepsilon}_{0|-6} = \mathbf{x}_{0|-6}^b - \mathbf{x}_0^{truth}$, then:

$$\begin{aligned}
&\mathbf{M}_{t|0}(\boldsymbol{\varepsilon}_0 - \boldsymbol{\varepsilon}_{0|-6}) \\
&= \mathbf{M}_{t|0}\left[(\mathbf{x}_0^a - \mathbf{x}_0^{truth}) - (\mathbf{x}_{0|-6}^f - \mathbf{x}_0^{truth})\right] \\
&= \mathbf{M}_{t|0}\left[\mathbf{x}_0^a - \mathbf{x}_{0|-6}^f\right] \\
&= \mathbf{M}_{t|0}\mathbf{K}_0\left[\mathbf{y}_0^o - h(\overline{\mathbf{x}}_{0|-6}^f)\right] \\
&= \mathbf{M}_{t|0}\mathbf{K}_0\mathbf{v}_0
\end{aligned} \tag{B.7}$$

where $\mathbf{K}_0 = \mathbf{X}_{0|-6}^b\widetilde{\mathbf{K}}_0$ is Kalman gain matrix. Based on this equation, the cost function can be written as:

$$\begin{aligned}
J &= \frac{1}{2}(\boldsymbol{\varepsilon}_{t|0}^T + \boldsymbol{\varepsilon}_{t|-6}^T)\mathbf{M}_{t|0}\mathbf{K}_0\mathbf{v}_0 \\
&= \frac{1}{2}(\boldsymbol{\varepsilon}_{t|0}^T - \boldsymbol{\varepsilon}_{t|-6}^T + 2\boldsymbol{\varepsilon}_{t|-6}^T)\mathbf{M}_{t|0}\mathbf{K}_0\mathbf{v}_0 \\
&= \frac{1}{2}(\mathbf{v}_0^T\mathbf{K}_0^T\mathbf{M}_{t|0}^{\mathrm{T}} + 2\boldsymbol{\varepsilon}_{t|-6}^T)\mathbf{M}_{t|0}\mathbf{K}_0\mathbf{v}_0 \\
&= \frac{1}{2}\mathbf{v}_0^T\mathbf{K}_0^T\mathbf{M}_{t|0}^T\mathbf{M}_{t|0}\mathbf{K}_0\mathbf{v}_0 + \boldsymbol{\varepsilon}_{t|-6}^T\mathbf{M}_{t|0}\mathbf{K}_0\mathbf{v}_0
\end{aligned} \tag{B.8}$$

Since we used linear tangent model to get equation (B.6), this is an approximation of equation (B.5) when the model is nonlinear. The sensitivity of the cost function to the observation increments $\mathbf{v}_0$ is

$$\frac{\partial J}{\partial \mathbf{v}_0} = \mathbf{K}_0^T\mathbf{M}_{t|0}^T\mathbf{M}_{t|0}\mathbf{K}_0\mathbf{v}_0 + \mathbf{K}_0^T\mathbf{M}_{t|0}^T\boldsymbol{\varepsilon}_{t|-6} \tag{B.9}$$

We define $e_{t|-6} = \left\langle \boldsymbol{\varepsilon}_{t|-6}^f, \boldsymbol{\varepsilon}_{t|-6}^f \right\rangle$, then $\frac{\partial e_{t|-6}}{\partial \mathbf{x}_{t|-6}^f} = 2\boldsymbol{\varepsilon}_{t|-6}^f$. Multiplying both sides by the adjoint model $\mathbf{M}_{t|0}^T\frac{\partial e_{t|-6}}{\partial \mathbf{x}_{t|-6}^f} = 2\mathbf{M}_{t|0}^T\boldsymbol{\varepsilon}_{t|-6}^f$, then it becomes:

$$\frac{\partial e_{t|-6}}{\partial \mathbf{x}^f_{0|-6}} = 2\mathbf{M}^T_{t|0}\boldsymbol{\varepsilon}^f_{t|-6} \tag{B.10}$$

We define $e_{t|0} = \left\langle \boldsymbol{\varepsilon}^f_{t|0}, \boldsymbol{\varepsilon}^f_{t|0} \right\rangle$ and the error changes $\delta e_{t|0}$ due to a perturbation $\delta \mathbf{x}^f_{t|0}$ as :

$$\delta e_{t|0} = \left\langle (\mathbf{x}^f_{t|0} + \delta\mathbf{x}^f_{t|0} - \mathbf{x}^a_t), (\mathbf{x}^f_{t|0} + \delta\mathbf{x}^f_{t|0} - \mathbf{x}^a_t) \right\rangle - \left\langle (\mathbf{x}^f_{t|0} - \mathbf{x}^a_t), (\mathbf{x}^f_{t|0} - \mathbf{x}^a_t) \right\rangle \tag{B.11}$$

After simplification, $\delta e_{t|0} = \delta\mathbf{x}^{fT}_{t|0}\delta\mathbf{x}^f_{t|0}$, where we used $\left\langle \delta\mathbf{x}^f_{t|0}, \mathbf{x}^f_{t|0} - \mathbf{x}^a_t \right\rangle = 0$. Since $\delta\mathbf{x}^f_{t|0} = \mathbf{M}_{t|0}\delta\mathbf{x}^a_0$, then $\delta e_{t|0} = \delta\mathbf{x}^a_0\mathbf{M}^T_{t|0}\mathbf{M}_{t|0}\delta\mathbf{x}^a_0$, and $\frac{\partial e_{t|-6}}{\partial \mathbf{x}^a_0} = 2\mathbf{M}^T_{t|0}\mathbf{M}_{t|0}\delta\mathbf{x}^a_0$. Because $\delta\mathbf{x}^a_0 = \mathbf{K}_0\mathbf{v}_0$,

$$\frac{\partial e_{t|-6}}{\partial \mathbf{x}^a_0} = 2\mathbf{M}^T_{t|0}\mathbf{M}_{t|0}\mathbf{K}_0\mathbf{v}_0 \tag{B.12}$$

Based on equation (B.10) and (B.12), the sensitivity of the cost function to the observation increments is then written as:

$$\frac{\partial J}{\partial \mathbf{v}^o_0} = \frac{1}{2}\mathbf{K}^T_0\left(\frac{\partial e_{t|-6}}{\partial \mathbf{x}^f_{0|-6}} + \frac{\partial e_{t|0}}{\partial \mathbf{x}^a_{0|0}}\right) \tag{B.13}$$

The cost function J defined as the error difference between $e_{t|0}$ and $e_{t|-6}$ is only due to the assimilation of the observation at time t=00hr. In the following, we will try to express the cost function J as a function of the observations assimilated at t=00hr. We substitute the definitions of $\boldsymbol{\varepsilon}^f_{t|0}$ and $\boldsymbol{\varepsilon}^f_{t|-6}$ into the cost function:

$$\begin{aligned} J &= \frac{1}{2}\left\langle (\mathbf{x}^f_{t|0} - \mathbf{x}^t_a), (\mathbf{x}^f_{t|0} - \mathbf{x}^t_a) \right\rangle - \left\langle (\mathbf{x}^f_{t|-6} - \mathbf{x}^t_a), (\mathbf{x}^f_{t|-6} - \mathbf{x}^t_a) \right\rangle \\ &= \frac{1}{2}\left\langle (\mathbf{x}^f_{t|0} - \mathbf{x}^f_{t|-6} + \mathbf{x}^f_{t|-6} - \mathbf{x}^t_a), (\mathbf{x}^f_{t|0} - \mathbf{x}^f_{t|-6} + \mathbf{x}^f_{t|-6} - \mathbf{x}^t_a) \right\rangle - \\ &\left\langle (\mathbf{x}^f_{t|-6} - \mathbf{x}^t_a), (\mathbf{x}^f_{t|-6} - \mathbf{x}^t_a) \right\rangle \end{aligned} \tag{B.14}$$

Expand the first inner product term in equation (B.14), it is easy to get the following equation:

$$\begin{aligned} J &= \frac{1}{2}\left\langle (\mathbf{x}^f_{t|0} - \mathbf{x}^f_{t|-6}), (\mathbf{x}^f_{t|0} - \mathbf{x}^t_a + \mathbf{x}^f_{t|-6} - \mathbf{x}^t_a) \right\rangle \\ &= \frac{1}{2}\left\langle \mathbf{M}_{t|0}(\mathbf{x}^a_0 - \mathbf{x}^f_{0|-6}), (\mathbf{x}^f_{t|0} - \mathbf{x}^t_a + \mathbf{x}^f_{t|-6} - \mathbf{x}^t_a) \right\rangle \end{aligned} \tag{B.15}$$

Due to the using of tangent linear model, equation (B.15) is also an approximation of equation (B.5) when the model is nonlinear. Based on the characteristics of the inner product, and $\frac{\partial e_{t|-6}}{\partial \mathbf{x}^f_{t|-6}} = \mathbf{x}^f_{t|-6} - \mathbf{x}^t_a$, $\frac{\partial e_{t|0}}{\partial \mathbf{x}^f_{t|0}} = \mathbf{x}^f_{t|0} - \mathbf{x}^a_t$, equation (B.15) becomes:

$$\begin{aligned} J &= \frac{1}{2}\left\langle (\mathbf{x}^a_0 - \mathbf{x}^f_{0|-6}), \mathbf{M}^T_{t|0}\left(\frac{\partial e_{t|0}}{\partial \mathbf{x}^f_{t|0}} + \frac{\partial e_{t|-6}}{\partial \mathbf{x}^f_{t|-6}}\right) \right\rangle \\ &= \frac{1}{2}\left\langle (\mathbf{x}^a_0 - \mathbf{x}^f_{0|-6}), \left(\frac{\partial e_{t|0}}{\partial \mathbf{x}^a_0} + \frac{\partial e_{t|-6}}{\partial \mathbf{x}^f_{0|-6}}\right) \right\rangle \\ &= \left\langle (\mathbf{y}^o_0 - h(\mathbf{x}^f_{0|-6}), \frac{1}{2}\mathbf{K}^T\left(\frac{\partial e_{t|0}}{\partial \mathbf{x}^a_0} + \frac{\partial e_{t|-6}}{\partial \mathbf{x}^f_{0|-6}}\right) \right\rangle \end{aligned} \tag{B.16}$$

Substitute equation (B.13) into (B.16), we obtain,

$$J = \left\langle (\mathbf{y}^o_0 - h(\overline{\mathbf{x}}^f_{0|-6}), \frac{\partial J}{\partial \mathbf{v}_0}) \right\rangle \tag{B.17}$$

The difference between ensemble sensitivity method and the adjoint method (Langland and Baker, 2004) is how the observation sensitivity $\frac{\partial J}{\partial \mathbf{v}^o_0}$ is calculated. In the ensemble sensitivity method, $\frac{\partial J}{\partial \mathbf{v}^o_0}$ is directly calculated based on the ensemble forecast and the weighting matrix (Equation (5.11)). In the adjoint method, it is based on the Equation (B.13).

B.3 Derivation of the sensitivity of the cost function to the observations without using linearization

Unlike the derivation in the text (Section 5.2), this derivation does not use linearization. It is based on the assumption that the forecast length t is short enough that we estimate the i^{th} ensemble forecast at time t initialized at t=00hr with the ensemble forecasts initialized at t=-6hr using the same weights as at the analysis time. Though it does not require linearization, it neglects the correlation between the error due to this assumption and the observations assimilated at t=00hr. The sensitivity formulation based on this derivation gives essentially identical result as equation (5.11), so we report the derivation here.

In this derivation, we first find the dependence of $\left(\boldsymbol{\varepsilon}_0 - \boldsymbol{\varepsilon}_{0|-6}\right)$ (the error difference between analysis and background at the analysis time) on the observational increments $\mathbf{v}_0$, following the LETKF formulation of Hunt et al. (2007):

$$\boldsymbol{\varepsilon}_0 - \boldsymbol{\varepsilon}_{0|-6} = \overline{\mathbf{x}}_0^a - \mathbf{x}_0^{truth} - (\overline{\mathbf{x}}_{0|-6}^b - \mathbf{x}_0^{truth}) = \mathbf{X}_{0|-6}^b \overline{\mathbf{w}}_0^a = \mathbf{X}_{0|-6}^b \widetilde{\mathbf{K}}_0 \mathbf{v}_0 \tag{B.18}$$

where $\mathbf{X}_{0|-6}^b = \left[\delta\mathbf{x}_{0|-6}^{b1} \;\; | \;\; \cdots \;\; | \;\; \delta\mathbf{x}_{0|-6}^{bK}\right]$ is a matrix whose K columns are background ensemble perturbations with the i^{th} column $\delta\mathbf{x}_{0|-6}^{bi} = \mathbf{x}_{0|-6}^{bi} - \overline{\mathbf{x}}_{0|-6}^b$, equation (B.18) indicates that the analysis increments are the linear combination of the background ensemble perturbations with the weighting matrix $\overline{\mathbf{w}}_0^a = \widetilde{\mathbf{K}}_0 \mathbf{v}_0$. $\widetilde{\mathbf{K}}_0 = \widetilde{\mathbf{P}}_0^a \mathbf{Y}_0^{bT} \mathbf{R}_0^{-1}$ and $\tilde{\mathbf{P}}_0^a = \left[(K-1)\mathbf{I} + \mathbf{Y}_0^{bT} \mathbf{R}_0^{-1} \mathbf{Y}_0^b\right]^{-1}$ are Kalman gain and analysis error covariance matrices in the ensemble subspace spanned by the forecasts. $\mathbf{Y}_0^b$ is a matrix whose i^{th}

column is the ensemble perturbations in the observation space equal to $h(\mathbf{x}_{0|-6}^{bi}) - h(\overline{\mathbf{x}}_{0|-6}^{b})$. $\mathbf{R}_0$ is the observation error covariance. We verified the analysis and 6-hour forecast valid at t=00hr against the true state $\mathbf{x}_0^{truth}$. An over-bar represents an average over the K ensemble members, a tilde indicates that a vector or matrix is represented in the subspace of ensemble forecasts, and δ represents the difference between an ensemble member value and the ensemble mean.

We need to compute the impact of analysis change at t=00hr due to assimilation of observations on the average forecast at time t. For this, consider the analysis at time 00hr, the i^{th} analysis ensemble member is given by (Hunt et al., 2007, eq. 25):

$$\mathbf{x}_0^{ai} = \overline{\mathbf{x}}_{0|-6}^{b} + \mathbf{X}_{0|-6}^{b}\mathbf{w}_0^{ai} \tag{B.19}$$

where $\mathbf{w}_0^{ai} = \overline{\mathbf{w}}_0^{a} + \delta\mathbf{w}_0^{ai}$ is a vector with K dimension, whose element is $w_0^{aji} = \overline{w}_0^{aj} + \delta w_0^{aji}$, j is from 1 to K. $\delta\mathbf{w}_0^{ai}$ is the i^{th} column of the K by K matrix $\mathbf{W}_0^{a} = [(K-1)\widetilde{\mathbf{P}}_0^{a}]^{1/2}$ with the elements δw_0^{aji}. We also note that the perturbation weights summed over either the K columns or the K rows are equal to one: $\sum_{j=1}^{K} \delta w_0^{aji} = \sum_{i=1}^{K} \delta w_0^{aji} = 1$ (Appendix B.1).

We need to express $\mathbf{x}_0^{ai}$, the i^{th} analysis ensemble member at t=00hr, as a weighted average of the background ensemble member $\mathbf{x}_{0|-6}^{bj}$, j is from 1 to K. In

order to do this, we expand the terms on the right hand side of equation (B.19) based on the definitions of each term,

$$\begin{aligned}
\mathbf{x}_0^{ai} &= \overline{\mathbf{x}}_{0|-6}^{b} + \mathbf{X}_{0|-6}^{b}\mathbf{w}_0^{ai} \\
&= \sum_{j=1}^{K}\left(\frac{1}{K}\right)\mathbf{x}_{0|-6}^{bj} + \sum_{j=1}^{K}\left[\mathbf{x}_{0|-6}^{bj} - \overline{\mathbf{x}}_{0|-6}^{b}\right]w_0^{aji} \\
&= \sum_{j=1}^{K}\left(\frac{1}{K}\right)\mathbf{x}_{0|-6}^{bj} + \sum_{j=1}^{K}\left[\mathbf{x}_{0|-6}^{bj} - \overline{\mathbf{x}}_{0|-6}^{b}\right]\left(\overline{w}_0^{aj} + \delta w_0^{aji}\right) \\
&= \sum_{j=1}^{K}\mathbf{x}_{0|-6}^{bj}\left(\frac{1}{K} + \overline{w}_0^{aj} + \delta w_0^{aji}\right) - \overline{\mathbf{x}}_{0|-6}^{b}\sum_{j=1}^{K}\left(\overline{w}_0^{aj} + \delta w_0^{aji}\right)
\end{aligned} \tag{B.20}$$

Since $\sum_{j=1}^{K}\delta w_0^{aji} = \sum_{i=1}^{K}\delta w_0^{aji} = 1$, $\overline{\mathbf{x}}_{0|-6}^{b} = \frac{1}{K}\sum_{j=1}^{K}\mathbf{x}_{0|-6}^{bj}$, and we define $\overline{\overline{w}}_0^{a} = \frac{1}{K}\sum_{j=1}^{K}\overline{w}_0^{aj}$, where $\overline{w}_0^{aj}$ is the j^{th} element of the mean weight vector $\overline{\mathbf{w}}_0^{a}$, so that the i^{th} analysis ensemble member at t=00hr is a linear combination of the background ensemble forecast expressed as:

$$\begin{aligned}
\mathbf{x}_0^{ai} &= \sum_{j=1}^{K}\mathbf{x}_{0|-6}^{bj}\left(\frac{1}{K} + \overline{w}_0^{aj} + \delta w_0^{aji}\right) - \left(\overline{\overline{w}}^{a} + \frac{1}{K}\right)\sum_{j=1}^{K}\mathbf{x}_{0|-6}^{bj} \\
&= \sum_{j=1}^{K}\mathbf{x}_{0|-6}^{bj}\left(\overline{w}_0^{aj} - \overline{\overline{w}}^{a} + \delta w_0^{aji}\right)
\end{aligned} \tag{B.21}$$

We assume that the forecast length t is short enough that the perturbations with respect to the ensemble mean grow linearly, so that we estimate the i^{th} ensemble forecast at time t initialized at t=00hr with the ensemble forecasts initialized at t=-6hr using the same weights as at the analysis time:

$$\mathbf{x}_{t|0}^{i} = \sum_{j=1}^{K}\mathbf{x}_{t|-6}^{fj}\left(\overline{w}_0^{aj} - \overline{\overline{w}}^{a} + \delta w_0^{aij}\right) + error \tag{B.22}$$

where 'error' represent the error from this approximation. We take an ensemble average of these forecasts initialized at t=00hr:

$$\begin{aligned}\overline{\mathbf{x}}_{t|0}^{f} &= \frac{1}{K}\sum_{i=1}^{K}\mathbf{x}_{t|0}^{i} \\ &= \frac{1}{K}\sum_{i=1}^{K}\sum_{j=1}^{K}\mathbf{x}_{t|-6}^{fj}(\overline{w}_0^{aj} - \overline{\overline{w}}^{a} + \delta w_0^{aij}) + \overline{\mathbf{e}}_{t|0}\end{aligned} \tag{B.23}$$

We denote the error made with this approximation as $\overline{\mathbf{e}}_{t|0}$. Since $\frac{1}{K}\sum_{i=1}^{K}\sum_{j=1}^{K}\mathbf{x}_{t|-6}^{fj}\delta w_0^{aij} = \frac{1}{K}\sum_{j=1}^{K}\mathbf{x}_{t|-6}^{fj}\sum_{i=1}^{K}\delta w_0^{aij} = \frac{1}{K}\sum_{j=1}^{K}\mathbf{x}_{t|-6}^{fj}$, and we define $\delta\overline{w}_0^{aj} = \overline{w}_0^{aj} - \overline{\overline{w}}$, so that equation (B.23) can be written as:

$$\begin{aligned}\overline{\mathbf{x}}_{t|0}^{f} &= \frac{1}{K}\sum_{j=1}^{K}\mathbf{x}_{t|-6}^{fj}(\overline{w}_0^{aj} - \overline{\overline{w}}^{a} + 1) + \overline{\mathbf{e}}_{t|0} \\ &= \overline{\mathbf{x}}_{t|-6}^{f} + \sum_{j=1}^{K}\mathbf{x}_{t|-6}^{fj}\delta\overline{w}_0^{aj} + \overline{\mathbf{e}}_{t|0}\end{aligned} \tag{B.24}$$

We note that, although very small, the error $\overline{\mathbf{e}}_{t|0} = \overline{\mathbf{x}}_{t|0}^{f} - \overline{\mathbf{x}}_{t|-6}^{f} - \sum_{j=1}^{K}\mathbf{x}_{t|-6}^{fj}\delta\overline{w}_0^{aj}$ (Figure B.1) cannot be neglected in order to obtain accurate observation sensitivity. It is can be calculated once we have the ensemble forecasts initialized from both t=00hr and t=-6hr. Though both $\overline{\mathbf{x}}_{t|0}^{f}$ and $\delta\overline{w}_0^{aj}(\mathrm{j}=1,\ \cdots\ K)$ is function of observations assimilated at t=00hr (approved in next paragraph), we neglect this correlation in the later derivations.

We will show that the vector $\delta\overline{\mathbf{w}}_0^{a}$ with the element $\delta\overline{w}_0^{aj}(\mathrm{j}=1,\ \cdots\ K)$ can be expressed in terms of the increments $\mathbf{v}_0$. Based on Hunt et al. (2007), $\overline{w}_0^{aj} = \sum_{p=1}^{P}\widetilde{K}_0^{jp}v_0^{p}$, where $\widetilde{K}_0^{jp}$ is an element of K by P matrix $\widetilde{\mathbf{K}}_0$, P is the number of observations, so that:

$$
\begin{aligned}
\delta\overline{w}_0^{aj} &= \overline{w}_0^{aj} - \overline{\overline{w}}^{a} \\
&= \sum_{p=1}^{P} \widetilde{K}_0^{jp} v_0^p - \frac{1}{K} \sum_{j=1}^{K} \sum_{p=1}^{P} \widetilde{K}_0^{jp} v_0^p \\
&= \sum_{p=1}^{P} \widetilde{K}_0^{jp} v_0^p - \sum_{p=1}^{P} \left[\frac{1}{K} \sum_{j=1}^{K} \widetilde{K}_0^{jp} \right] v_0^p
\end{aligned}
\tag{B.25}
$$

We define $\overline{\overline{\widetilde{K}}}_0^{p} = \frac{1}{K} \sum_{j=1}^{K} \widetilde{K}_0^{jp}$ and $\delta\widetilde{K}_0^{jp} = \widetilde{K}_0^{jp} - \overline{\overline{\widetilde{K}}}_0^{p}$, then above equation can be written as

$$
\begin{aligned}
\delta\overline{w}_0^{aj} &= \sum_{p=1}^{P} (\widetilde{K}_0^{jp} - \overline{\overline{\widetilde{K}}}_0^{p}) v_0^p \\
&= \sum_{p=1}^{P} \delta\widetilde{K}_0^{jp} v_0^p = \delta\widetilde{\mathbf{K}}_0 \mathbf{v}_0
\end{aligned}
\tag{B.26}
$$

where $\delta\widetilde{\mathbf{K}}_0$ is a K by P matrix whose element is $\delta\widetilde{K}_0^{jp}$.

Based on equation (B.24) and (B.26), $\overline{\mathbf{x}}_{t|0}^{f} = \overline{\mathbf{x}}_{t|-6}^{f} + \mathbf{X}_{t|-6}^{f} \delta\widetilde{\mathbf{K}}_0 \mathbf{v}_0 + \overline{\mathbf{e}}_{t|0}$, where $\mathbf{X}_{t|-6}^{f} = \left[\mathbf{x}_{t|-6}^{f1} \quad | \quad \cdots \quad | \quad \mathbf{x}_{t|-6}^{fK} \right]$ is a matrix whose K columns are background ensemble forecasts. Note that this notation is different from that in the text. So that

$$
\begin{aligned}
\boldsymbol{\varepsilon}_{t|0} - \boldsymbol{\varepsilon}_{t|-6} &= \overline{\mathbf{x}}_{t|0}^{f} - \overline{\mathbf{x}}_{t|-6}^{f} \\
&= \mathbf{X}_{t|-6}^{f} \delta\widetilde{\mathbf{K}}_0 \mathbf{v}_0 + \overline{\mathbf{e}}_{t|0}
\end{aligned}
\tag{B.27}
$$

Similarly,

$$
\begin{aligned}
\boldsymbol{\varepsilon}_{t|0} + \boldsymbol{\varepsilon}_{t|-6} &= \overline{\mathbf{x}}_{t|0}^{f} - \overline{\mathbf{x}}_{t}^{a} + \overline{\mathbf{x}}_{t|-6}^{f} - \overline{\mathbf{x}}_{t}^{a} \\
&= \overline{\mathbf{x}}_{t|-6}^{f} - \overline{\mathbf{x}}_{t}^{a} + \overline{\mathbf{x}}_{t|-6}^{f} - \overline{\mathbf{x}}_{t}^{a} + \overline{\mathbf{x}}_{t|0}^{f} - \overline{\mathbf{x}}_{t|-6}^{f} \\
&= 2\boldsymbol{\varepsilon}_{t|-6} + \mathbf{X}_{t|-6}^{f} \delta\widetilde{\mathbf{K}}_0 \mathbf{v}_0 + \overline{\mathbf{e}}_{t|0}
\end{aligned}
\tag{B.28}
$$

The cost function is then written as:

$$
\begin{aligned}
J &= \frac{1}{2}(\boldsymbol{\varepsilon}_{t|0}^{T} + \boldsymbol{\varepsilon}_{t|-6}^{T})(\boldsymbol{\varepsilon}_{t|0} - \boldsymbol{\varepsilon}_{t|-6}) \\
&= \frac{1}{2}[2\boldsymbol{\varepsilon}_{t|-6} + \mathbf{X}_{t|-6}^{f}\delta\widetilde{\mathbf{K}}_{0}\mathbf{v}_{0} + \overline{\mathbf{e}}_{t|0}]^{T}(\mathbf{X}_{t|-6}^{f}\delta\widetilde{\mathbf{K}}_{0}\mathbf{v}_{0} + \overline{\mathbf{e}}_{t|0})
\end{aligned}
\tag{B.29}
$$

We assume that the error $\boldsymbol{\varepsilon}_{t|-6}$ is not related with the observations assimilated at t=00hr, then the sensitivity of the forecast error to observations is written as:

$$
\frac{\partial J}{\partial \mathbf{v}_{0}} = \left[\delta\widetilde{\mathbf{K}}_{0}^{T}\mathbf{X}_{t|-6}^{fT}\right]\left[\boldsymbol{\varepsilon}_{t|-6} + \mathbf{X}_{t|-6}^{f}\delta\widetilde{\mathbf{K}}_{0}\mathbf{v}_{0} + \overline{\mathbf{e}}_{t|0}\right] \tag{B.30}
$$

The sensitivity to the observations in equation (B.30) can be directly calculated based on the weighting function from data assimilation at 00hr, the observation increment at 00hr, and the ensemble forecast initialized at -6hr. As mentioned previously, though this derivation does not need linearization, it neglects the correlation between $\overline{\mathbf{e}}_{t|0}$ and the observations assimilated at t=00hr. The results obtained with this formulation are undistinguishable from those reported in Chapter 5.

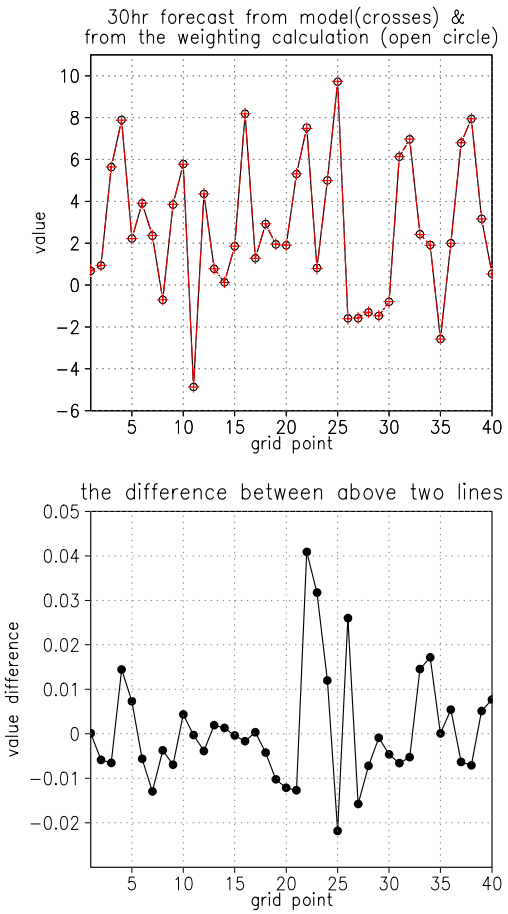


Figure B.1 Top panel: 24-hour forecast initialized at 00hr (red line with crosses) and the 24-hour forecast calculated from the linear combination of the 30-hour forecast initialized at -06hr (black line with open circles) at an arbitrary time; Bottom panel: the difference $\bar{e}_{t|0}$ between the actual forecast and the forecast calculated from the linear combination.

Bibliography

Ancell, B. and G. J. Hakim, 2007: Comparing adjoint and ensemble sensitivity analysis with applications to observation targeting. *Mon. Wea. Rev*. (submitted).

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

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.

Bergot, T., 1999: Adaptive observations during FASTEX: a systematic survey of upstream flights. *Quart. J. Roy. Metero. Soc.* **125***,* 3271-3298

Berliner, L. M., Z.-Q. Lu, and C. Snyder, 1999: Statistical design for adaptive weather observations. *J. Atmos. Sci.,* **56***,* 2536-2552.

Bishop, C. H, 2007: Introduction to data assimilation research at NRL and flow adaptive error covariance localization. http://www.weatherchaos.umd.edu/workshop/Bishop_UMD_workshop.pdf

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

Buizza, R., C. Cardinali, G. Kelly and J-N Thepaut, 2007: The value of targeted observations. Part II: The value of observations taken in singular vectors-based target areas. ECMWF Tech. Memo., **512**, pp33.

Cardinali, C., S. Pezzulli, and E. Andersson, 2004: Influence-matrix diagnostic of a data assimilation system. *Quart. J. Roy. Metero. Soc.* **130***,* 2767-2786

Dee, D. P. and A. M. da Silva, 2003: The choice of variable for atmospheric moisture analysis. *Mon. Wea. Rev.,* **131***,* 155-171

Derber, J. C. and W.-S. Wu, 1998: The use of TOVS cloud-cleared radiances in the NCEP SSI analysis system. *Mon. Wea. Rev.,* **126,** 2287-2299.

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

Emanuel, K. A., and R. Langland, 1998: FASTEX adaptive observations workshop. *Bull. Amer. Meteor. Soc.,* **79**, 1915-1919.

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), 10 143-10 162.

Fourrie, N., and J.-N. Thepaut, 2003: Evaluation of AIRS near-real-time channel selection for application to numerical weather prediction. *Quart. J. Roy. Metero. Soc.,* **129**, 2425-2439.

Hamill, M. T. and C. Snyder, 2002: Using improved background-error covariances from an ensemble Kalman filter for adaptive observations. *Mon. Wea. Rev*., **130**, 1552-1572.

Hansen, J. A., and A. L. Smith, 2000: The role of operational constrains in selecting supplementary observations. *J. Atmos. Sci*., **57**, 2859-2871.

Hardesty, M., 2006: Satellite wind lidar for observing global wind fields. Presentation.

Holm, E., E. Anderson, A. Beljaars, P. Lopez, J-F Mahfouf, A. J. Simmons and J-N Thepaut, 2002: Assimilation and modeling of the hydrological cycle: ECMWF's status and plans. ECMWF Tech. Memo., **383**, pp55.

Houtekamer, P. L. and H. L. Mitchell, 2001: A sequential Ensemble Kalman Filter for atmospheric data assimilation. *Mon. Wea. Rev.*, **129**, 123-137.

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.

Hunt, B. R., E. J., Kostelich, and I., Szunyogh (2007), Efficient Data Assimilation for Spatiotemporal Chaos: a Local Ensemble Transform Kalman Filter. *Physics D*. (in press)

Joiner, J., P. Poli, D. Frank and H. C. Liu, 2004: Detection of cloud-affected AIRS channels using an adjacent-pixel approach. *Quart. J. Roy. Metero. Soc.,* **130**, 1469-1487.

Joly, A., and Coauthors, 1997: The Fronts and Atlantic Storm-Track Experiment (FASTEX): Scientific objectives and experimental design. *Bull. Amer. Meteor. Soc*.**78**, 1917-1940.

Kalnay, E., 2003: Atmospheric modeling, data assimilation and predictability, Cambridge University Press, 341pp.

Kelly, G. J-N Thepaut, R. Buizza and C. Cardinali, 2007: The value of targeted observations part I : the value of observations taken over the oceans. ECMWF's status and plans. ECMWF Tech. Memo., 511, pp27.

Langland, R. H., 2005: Issues in targeted observing. *Quart. J. Roy. Meter. Soc.,* **131**, 3409-3425.

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

Langland, R. H., R. Gelaro, G. D. Rohaly and M. A. Shapiro (1999a), Targeted observations in FASTEX: Adjoint-based targeting procedures and data impact experiments in IOP17 and IOP18. *Quart. J. Roy. Metero. Soc.,* **125**, 3241-3270.

Langland, R. H., and Coauthors (1999b), The North Pacific Experiment (NORPEX-98): Targeted Observations for Improved North American Weather Forecasts. *Bull. Amer. Meteor. Soc., 80*, 1363-1384

Li, H. 2007: Local Ensemble Transform Kalman Filter with realistic observations. Ph. D thesis, University of Maryland

Liu, J. and E. Kalnay, 2007: Simple Doppler Wind Lidar adaptive observation experiments with 3D-Var and an ensemble Kalman filter in a global primitive equations model, *Geophys. Res. Lett.*, **34**, L19808, doi: 10.1029/2007GL030707.

Lorenz, E. N. and K. A. Emanuel (1998), Optimal sites for supplementary observations: Simulation with a small model. *J. Atmos. Sci.,* **55***,* 399-414.

Majumdar, S. J., C. H. Bishop, and B. J. Etherton, 2002: Adaptive sampling with the ensemble transform Kalman filter. Part II: Field program implementation. *Mon. Wea. Rev*., **130**, 1356-1369.

Miyoshi, T. (2005), Ensemble Kalman filter experiments with a primitive-equation global model, Ph. D thesis, University of Maryland.

Molteni, F., 2003: Atmospheric simulations using a GCM with simplified physical parametrizations. I: Model climatology and variability in multi-decadal experiments. *Climate Dyn.*, **20**, 175-191.

Morss, R. E., and K. A. Emanuel, 2001: Idealized adaptive observation strategies for improving numerical weather prediction. *J. Atmos. Sci.,* **58,** 210-232.

Morss, R. E., and K. A. Emanuel, 2002: Influence of added observations on analysis and forecast errors: Results from idealized systems. *Quart. J. Roy. Meter. Soc*., **128,** 285-322.

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

Palmer, T. N., R. Gelaro, J. Barkmeijer and R. Buizza, 1998: Singular vectors, metrics, and adaptive observations. *J. Atmos. Sci*., **643**, 633-653

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

Pu, Z., and E. Kalnay (1999), Targeting observations with the quasi-inverse linear and adjoint NCEP global models: Performance during FASTEX. *Quart. J. Roy. Meter. Soc.,* **125**, 3329-3338

Rishojgaard, L. P., R. Atlas (2004), The impact of Doppler Lidar Wind observations on a Single-Level meteorological analysis. *J. of Applied Meteorology,* **43**, 810-820.

Snyder, C., 1996: Summary of an informal workshop on adaptive observation and FASTEX. *Bull. Amer. Meteor. Soc.*, **77**, 953-965.

Stoffelen, A. and Coauthors (2005), The atmospheric dynamics mission for global wind field experiments. *Bull. Amer. Meteor. Soc.,* **86***,* 73-87

Susskind, J., C. Barnet and J. Blaisdell 2003: Retrieval of atmospheric and surface parameters from AIRS/AMSU/HSB data in the presence of clouds. *IEEE Transactions on Geoscience and Remote sensing.*

Szunyogh, I., Z. Toth, K. A. Emanuel, C. H. Bishop, C. Snyder, R. E. Morss, J. S. Woolen, and T. P. Marchok (1999), Ensemble-based targeting during FASTEX: the impact of dropsonde data from the LEAR jet. *Quart. J. Roy. Metero. Soc.,* **125**, 3189-3217.

Szunyogh, I., E. J. Kostelich, and G. Gyarmati, 2007: Assessing a local ensemble Kalman filter: assimilating observations with the NCEP global model. *Tellus*, accepted.

Toth, Z. and Coauthors (2002), Adaptive observations at NCEP: Past, present and future. Preprints of the Symposium on Observations, Data assimilation and Probabilistic Prediction, *Orlando FL, 13-17 January 2002, 185-190.*

Trevisan, A., and F. Uboldi, 2004: Assimilation of standard and targeted observations within the unstable subspace of the Observation-Analysis-Forecast cycle system. *J. Atmos. Sci*., **104,** 103-113

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

Whitaker, J. S., T. M. Hamill, X. Wei, Y. Song and Z. Toth, 2007: Ensemble data assimilation with the NCEP Global Forecast System. *Mon. Wea. Rev.* (accepted)

Zapotacny, T. H., W. P. Menzel, J. P. Nelson and J. A. Jung, 2002: An impact study of five remotely sensed and five in situ data types in the eta data assimilation system. *Weather and forecasting,* **17,** 263-285