UNIVERSITY OF READING

School of Mathematical and Physical Sciences

Estimating Forecast Error Covariance Matrices with Ensembles

by

Emine Akkus

August 2014

A dissertation is submitted to the Department of Mathematics in partial ful lment of the requirements for the degree of Master of Science

Abstract

The problem of variational data assimilation for a nonlinear coupled atmosphere and ocean model is formulated as an optimization problem to find the best initial condition. The input data includes errors of observations and background. Therefore, the optimal solution involves error. The modelling of the background error covariance matrix is important in any data assimilation methods in the sense that it determines the spread of the errors. Therefore, in this study we try to estimate the forecast (background) error covariance matrix calculated by the idea of ensemble Kalman Filter (EnKF). To do this, we use a method which is an ensemble of four dimensional variational (4DVar) methods. We generate ensemble members by perturbing the background and observations with different random numbers. Then, we set up different ensembles and investigate how many ensemble members can make the forecast error covariance matrix convergence. We look at the convergence of each component of the matrix with the ensemble size getting bigger, and want to find a sufficient ensemble size which makes all components convergence. Then, we look at the effects of the error correlations of model variables and,

Acknowledgements

I would like to thank my supervisor Dr Amos S. Lawless so much for being helpful and supportive during this dissertation. I would also like to thank all my lecturers in this MSc course especially to Professor Michael Baines and Dr Peter K. Sweby for their guidance and patience.

A special thanks to the God, my family, my ance and my friends. Their supports provide me high motivation during this course.

Declaration

I con rm that this is my own work and the use of all material from other sources has been properly and fully acknowledged.

Emine Akkus

Contents

List of Figures

List of Tables

Chapter 1

Introduction

In general, data assimilation (DA) combines all available information to estimate the state of a system. It uses observations and prior information, which is called the background, and combines them with a computer model to nd the analysis which is an approximation to the observed reality. Lawless (2013) discusses the progress of data assimilation in more detail. Basically, it started being used in the 1940s and with the development of data assimilation techniques its usage has reached a wide spectrum of areas such as environmental modelling, climate monitoring, tra c modelling and numerical weather prediction (NWP). We are interested in its use in NWP. DA aims to provide the most appropriate initial condition for a forecast. In the assimilation window, we have an initial valuve Th27(ums27(e)co)-3045bei-344(anle)-2(e)coe an-3045bg1(c)-28
bei-340(dle014n-1n)d a

1.1 Motivation

For a long time, DA has been applied to the atmosphere and the ocean separately. On the other hand, recent studies show that coupling them gives numerically the most certain outcomes of climate change investigations (Sausen and Voss, 1995). To simulate long-term climatic variations, the coupled atmosphere and ocean models need to be used. In practice, there are some complications. For instance, it needs a lot of computational work because of the big di erence between the scales of atmosphere and ocean. The ocean is much slower than the atmosphere because of the thermal inertia of the ocean (Dubois et al., 1999).

Variational methods solve the assimilation by minimizing a function with the ability of using future observations. These methods involve a background error covariance matrix but they do not provide any information about this matrix. Therefore, this matrix remains constant for di erent assimilations. Ensemble methods, which are sequential so all things in sequence, provide

assimilation, because those statistics become the background error statistics for the next assimilation window.

In this project, the behaviour of the Molteni et al. (1993) coupled atmosphere and ocean model will be investigated with the ensembles of non-incremental 4DVar methods. The ensemble members are generated by using the idea of Isaksen et al. (2010). We perturb the background and observations with di erent random numbers to get a new ensemble member. Then, we try to estimate the forecast error covariance matrix P_f generated by using the idea of EnKF. We use di erent ensembles to be able to estimate P_f. In this project, the research questions are:

(1) How many ensemble members are needed to capture the forecast covariances correctly?

(2) What are the e ects of error correlations on the estimation of forecast covariances?

(3) How these are a ected by observation errors and frequencies?

To investigate these questions, rst we try to estimate P_f with di erent ensemble sizes then try to understand how many ensemble members make the matrix P_f convergence. The convergence of each component of P_f is investigated separately with the ensemble size getting bigger. Then, we look at the e ects of error correlations and the e ects of the accuracies and the numbers of observations on the convergence of P_f.

1.2 Outline

This project is divided into six main chapters. In Chapter 2, we introduce the 4DVar method in mathematical detail with the use of tangent linear and adjoint model. The 4DVar is the foundation of the experiments. Then, as a sequential method the Kalman Filter algorithm is presented to show the formulation of analysis error covariance matrix P_a and forecast error covariance matrix P_f. The idea of ensemble Kalman Filter (EnKF) is adopted to calculate the ensemble based flow dependent background error covariance matrix P_f. Then, we describe the process of Isaksen et al. (2010) for perturbing a system to generate different ensemble members in the hybrid methods section.

Chapter 3 gives the system of Molteni et al.(1993) coupled model, which we use as the toy model in our experiments, and the second order Runge-Kutta method, which we use to discretize this coupled model.

In Chapter 4, we describe the methodology we follow step by step. It starts with generating truth state and continues with generating background and observations, which we assimilate with 4DVar to find the best fit initial state. Then, we perturb the system by using the idea of Isaksen et al. (2010) to get different ensemble members. Thus, we have an ensemble of analyses at the initial time. Then, we forecast each of them to get a forecast error covariance matrix P_f.

Chapter 5 demonstrates our experiments and their results. In this chapter, we try to investigate our research questions described in Section (1.1).

In Chapter 6, we give a brief summary with conclusions of this project and propose some ideas for future work.

Chapter 2

Background

There are two types of data assimilation methods which are sequential and variational. Sequential data assimilation only considers observations made in the past until the time of analysis, while the variational one can use observations from the future. It can be seen from Figure (2.1)(b) that variational data assimilation methods produce a continuous analysis trajectory (the black line) by using the background trajectory (the red line with the initial value x_0^b) and the observations y_i, for $i = 1$

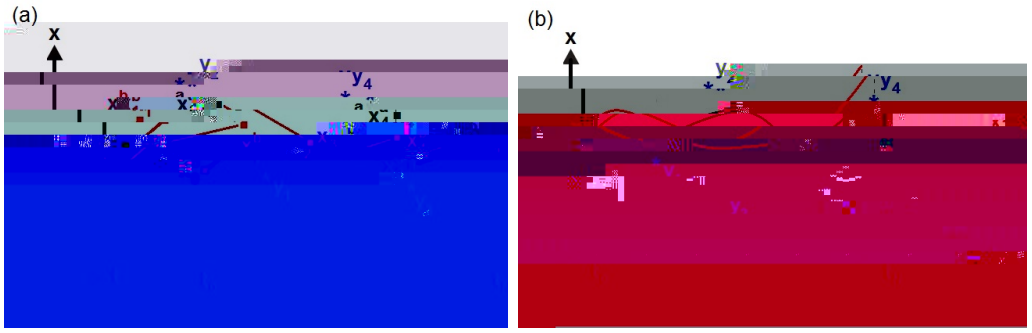


Figure 2.1: The processes of sequential (a) and variational (b) data assimilation methods over the assimilation window with the range $[t_0; t_N]$.

Nowadays, the use of ensemble methods has a signi cant part in NWP. Operational centres have started to seek di erent methods that consist of an

by trying to put their advantages together and call the method En4DVAR. It uses ow-dependent background error covariance matrix P from EnKF and tries to nd variational solution likewise 4DVar. This method does not use tangent linear and adjoint models to produce an analysis. In their experiments, the method produces an analysis result which is similar to the analysis produced with tangent linear and adjoint models, which need high computational work. Fairbairn et al. (2014) call the same method the 4DEnVar. From now on we use the abbreviation 4DEnVar for this method not to cause any confusion. By taking inspiration from the idea of 4DEnVar, in this study we use the ensemble of non-incremental 4DVar methods and produce the forecast error covariance matrix.

2.1 Assimilation Methods

2.1.1 The Four-Dimensional Variational Data Assimilation (4DVar)

4DVar method produced by Le Dimet and Talagrand (1986) is an extension of 3DVar with the consideration of observations in time. After the 2000s, the method has the most use in operational centres (Fairbairn et al., 2013). This method helps us get the best trajectory of the system by running the

where x_0 is the state vector of the system at time t_0 and is a vector of random unbiased Gaussian errors with known covariance matrix B and we have observations y_k at time t_k that satisfy

$$y_k = H_k(x_k) + \quad _k \tag{2.2}$$

where x_k

the right hand side (RHS) of Equation (2.4) has an error term). Equation (2.4) can also be written as

$$x_k = M_{k-1} M_{k-2} \cdots M_0(x_0) \quad (2.5)$$

$$= M_{0:k-1}(x_0) \quad (2.6)$$

Therefore, 4DVar is a nonlinear constrained optimization problem so in general it is hard to solve (Bouttier et al., 2002). To deal with this problem, the tangent linear hypothesis is developed. Under the hypothesis, we assume that the observation operator H can be made approximately linear as follows

$$H_0(x_0^b) \; H_0(x_0) \quad H_0(x_0^b)(x_0^b - x_0) \quad (2.7)$$

where $H_0(x_0^b)$ is the di erential of H_0 at x_0^b.

In addition, under this hypothesis it is assumed that the model operator M can be linearised as

$$y_k \; H_k M_{0:k-1}(x_0) \quad y_k \; H$$

analysis x_0^a satis es rJ $(x_0^a) = 0$ at the initial time. The aim of 4DVar is to nd x_0^a by minimizing the cost function, J .

2.1.2 The Kalman Filter (KF)

The KF method was produced by Kalman (1960) as a data assimilation (DA) method to estimate the current state. It is a sequential method and its process is explained in Figure (2.1)(a). Basically, the lter works just for linear systems. Jazwinski (1970) discusses the lter in detail. We just present here a brief explanation of how the KF works. In the KF, the system have a forecast step x_k^f with a forecast error covariance matrix P_k^f and an analysis step x_k^a with an analysis error covariance matrix P_k^a at time t_k. The state forecasts are explained as follows

$$x_k^f = M_{k\ 1} x_{k\ 1}^a \tag{2.10}$$

where

where K_k are the Kalman gain computations given by

$$K_k = P_k^f H_k^T (H_k P_k^f H_k^T + R_k)^{-1} \tag{2.14}$$

where H_k are the linear observation operators and R_k are the observation covariance matrices. As we mention before, this method is for linear systems. The extended Kalman Filter (EKF) applies KF to nonlinear systems by using tangent linear and adjoint models.

2.1.3 The Ensemble Kalman Filter (EnKF)

The EnKF presented by Evensen (1994) is an ensemble method and approximates the EKF for large systems. We do not give the whole algorithm of the method, only we give here the part of producing the ensemble error covariance matrix. Assume, we have m ensemble members which are denoted x_i for $i = 1; :::; m$, where each x_i is a state vector of the system, then the mean of ensemble members is

$$\bar{x} = \frac{1}{m} \sum_{i=1}^{m} x_i \tag{2.15}$$

The ensemble perturbation matrix X can be derived by

$$X = \frac{1}{\sqrt{m-1}} (x_1 - \bar{x}; x_2 - \bar{x}; :::; x_m - \bar{x}) \tag{2.16}$$

and the ensemble covariance matrix P is given by

$$P = \frac{1}{m-1} \sum_{i=1}^{m} (x_i - \bar{x})(x_i - \bar{x})^T \tag{2.17}$$

which can be also written in terms of the ensemble perturbation matrix in Equation(2.16) as follows

$$P = XX^T \tag{2.18}$$

The ensemble covariance matrix P is ow-dependent and using it as the background error covariance matrix makes the system more realistic. P is a symmetric and square matrix with the size of $n \quad n$ where n is the size of the state vectors x_i. Hence, the form of P is

$$P = \begin{pmatrix} Var(e_1) & Cov(e_1;e_2) & \cdots & Cov(e_1;e_n) \\ Cov(e_2;e_1) & Var(e_2) & \ddots & \vdots \\ \vdots & \ddots & \ddots & Cov(e_{n-1};e_n) \\ Cov(e_n;e_1) & \cdots & Cov(e_n;e_{n-1}) & Var(e_n) \end{pmatrix} \tag{2.19}$$

where e_i for $i = 1;:::;n$ denote each variable of the system which have n variables.

In the EnKF, the smoothness of analysis depends highly on the ensemble size m. It should be su ciently large. If it is not, this leads P to be low rank. To be able to make the system fully observable, the background error

2.1.4 Hybrid Methods

Hybrid methods are combinations of variational and ensemble methods. They try to apply the best features of both methods. These methods solve variational problems with the ow-dependent background error covariance matrix P instead of xed background error covariance matrix B. Ideally, the background error covariance matrix should depend on the current ow. For example, consider the pressure in the real atmosphere, sometimes high pressure takes the place of low pressure quickly. This kind of alterations should e ect the background error covariance matrix. Additionally, variational solutions use observations more e ciently than sequential ones, because sequential ones only use observations in the past, while variational ones can also use observations from the future. Therefore, hybrid methods are created. Most of them are produced by a combination of the EnKF and the incremental 4DVar and called Four-Dimensional Ensemble Variational DA referred as 4DEnVar in literature. In this project, we use an ensemble of non-incremental 4DVar methods producing ensemble members by perturbing the background and observations. We do these perturbations by choosing di erent random numbers for each ensemble members similar to the idea of Isaksen et al. (2010). They have perturbed observations by adding random noise from the probability density function (pdf) of observation error, and added further perturbations to account for model error during the forecasts. There is no direct perturbations applied to the background. The Isaksen et al. (2010) system can be described as follows:
Consider the following linear system:

$$\begin{aligned} x_k^a &= x_k^b + K_k(y_k \quad H_k x_k^b) \\ x_{k+1}^b &= M_k(x_k^a) \end{aligned} \tag{2.20}$$

where k denotes analysis cycle, y_k is the vector of observations, x_k^a is the analysed state, x_k^b, K_k and M_k are matrices, and K_k is a general gain matrix (not specifically the Kalman gain). Then the covariance matrices for this system are

$$\begin{aligned} P_k^a &= (I - K_k H_k) P_k^b (I - K_k H_k)^T + K_k R_k K_k^T \\ P_{k+1}^b &= M_k P_a^k M_k^T + Q_k \end{aligned} \tag{2.21}$$

where R_k is the observation error covariance matrix and Q_k is the model error covariance matrix.

Then we perturb the system (2.20) as follows (denoted by)

$$\begin{aligned} \tilde{x}_k^a &= \tilde{x}_k^b + K_k(y_k + \ _k - H_k \tilde{x}_k^b) \\ \tilde{x}_{k+1}^b &= M_k(\tilde{x}_k^a) + \ _k \end{aligned} \tag{2.22}$$

where $\ _k$ and $\ _k$ are perturbations with covariance matrices R_k and Q_k respectively.

Then subtracting the perturbed (2.22) and unperturbed (2.20) systems as follows:

$$\begin{aligned} {}^a_k &= {}^b_k + K_k(\ _k - H_k\ {}^b_k) \\ {}^b_{k+1} &= M_k\ {}^a_k + \ _k \end{aligned} \tag{2.23}$$

where ${}^a_k = \tilde{x}_k^a - x_k^a$ and ${}^b_k = \tilde{x}_k^b - x_k^b$. Then, from the system (2.23) the covariance matrices can be formed as

$$\begin{aligned} \overline{{}^k_a({}^k_a)^T} &= (I - K_k H_k)\,\overline{{}^k_b({}^k_b)^T}(I - K_k H_k)^T + K_k R_k K_k^T \\ \overline{{}^{k+1}_b({}^{k+1}_b)^T} &= M_k\,\overline{{}^k_a({}^k_a)^T}\,M_k^T + Q_k \end{aligned} \tag{2.24}$$

It can be seen from the comparison of the systems (2.21) and (2.24), if $\overline{{}^{k}_{b}({}^{k}_{b})^{T}} = P^{b}_{k}$ for some k, then $\overline{{}^{m}_{a}({}^{m}_{a})^{T}} = P^{a}_{m}$ and $\overline{{}^{m}_{b}({}^{m}_{b})^{T}} = P^{b}_{m}$ for all m k. This means that the analyses and backgrounds perturbations have equal covariances to the corresponding analysis and background error covariances for all subsequent analysis cycle. In this paper, they also provide another perturbed system by applying identical perturbations to observations and model error. They only choose the initial perturbation to be di erent.

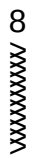
Chapter 3

Models

3.1 The Molteni coupled model

As the toy model, we use the coupled atmosphere-ocean model described by Molteni et al. (1993) which couples the chaotic Lorenz system (1963) with the linear oscillatory system. The Lorenz system has three variables X, Y and Z which represent the atmosphere, while the linear part has two variables W and V which are the representation of the ocean. The system is coupled by an arbitrary coupling parameter . It is applied through the X, Y, W and V variables. Molteni et al. (1993) have used the coupled model to examine the interaction of tropical-midlatitute. They have conducted some experiments by setting tropical forcing term W variously. The coupled system is also used by Dubois and Yiou (1999). Additionally, they have coupled the chaotic Lorenz system with the chaotic Rossler system. Molteni et al. (1993) system

is the following:



system

$$y^0(t) = f(t; y(t)) \tag{3.2}$$

with an initial condition $y(0) = y_0$. By doing Taylor expansion, the RK2 method can be derived as

$$y_{n+1} = y_n + \frac{1}{2}h(k_1 + k_2) \tag{3.3}$$

with

$$k_1 = f(t_n; y_n) \tag{3.4}$$

$$k_2 = f(t_n + h; y_n + hk_1) \tag{3.5}$$

where n represent the time step and h is the step size. The toy model is discretized by using RK2. Lawless (2006) discretize the Lorenz (1963) model with the method RK2.

Chapter 4

Experimental Setup

and R respectively (B is the background error covariance matrix while R is the observation error covariance matrix). This method gives an ensemble of analysis states at the initial time of the assimilation window and we forecast each analysis state and nd the forecast error covariance matrix calculated by the idea of ensemble Kalman Filter (EnKF) described in Section (2.1.3).

We choose Molteni et al. (1993) coupled model described in Section (3.1.1) as our toy model. Therefore, the size of state vectors x_k are 5 consisting of model variables $X; Y; Z; W$ and V. Subscript k shows the time step over the assimilation window. We transfer the continuous model into a discrete model by using the RK2 method. This numerical method which we use to generate the background, the truth and the analysis trajectories through the assimilation window, can be chosen arbitrarily. For example, Dubois and Yiou (1999) use a fourth order Runge-Kutta method (RK4) on coupled atmosphere ocean models. In this project, the Polack-Ribiere avour of conjugate gradient method is used as the iteration method to minimize the cost function, J. For this method, we set up the tolerance as 0.001 for the stopping criteria as follows

$$krJ^{i}(x_0)k$$

values of each variable of the model. We only estimate the state not the coupling parameter so it is fixed as 1. We set up the number of time steps as 20 and the step length as 0.05. Thus, the length of assimilation window becomes 1.

4.1 Generating the truth

The truth state vectors are denoted as x_k for $k = 0; :::; 20$. At the beginning, we set up x_0 as the column vector of $[1; 1; 1; 1:1; 1:1]$ to determine the initial truth state where each component of x_0 represents the initial values of model variables $X; Y; Z; W$ and V respectively. This truth state is identical in all experiments. Then, we run the RK2 method from x_0 to generate the truth trajectory over the assimilation window.

4.2 The 4DVar Algorithm

4.2.1 Generating the background

The background state vectors are denoted as x_k^b for $k = 0; :::; 20$. The initial background state x_0^b has errors and we assume that errors are uncorrelated so the background error covariance matrix B becomes diagonal and its form is

$$B = \begin{pmatrix} (\frac{X}{b})^2 & & & & 0 \\ & (\frac{Y}{b})^2 & & & \\ & & (\frac{Z}{b})^2 & & \\ & & & (\frac{W}{b})^2 & \\ 0 & & & & (\frac{V}{b})^2 \end{pmatrix} \tag{4.2}$$

where $(\ ^{X}_{b})^2; (\ ^{Y}_{b})^2; (\ ^{Z}_{b})^2; (\ ^{W}_{b})^2$ and $(\ ^{V}_{b})^2$ are the background variances of the model variables $X; Y; Z; W$ and V respectively.

Then, from Section (2.1.1) the initial background state x_0^b is generated by adding random noise to the truth state x_0 at time t_0. Di erent random noises are applied on each element of x_0. Then, Equation (2.1) becomes

$$x_0^b = x_0 + \ _b \tag{4.3}$$

where is a vector of random unbiased Gaussian errors and $_b$ is the 5 5

$$R = \begin{pmatrix} (\sigma_o^X)^2 & & & & 0 \\ & (\sigma_o^Y)^2 & & & \\ & & (\sigma_o^Z)^2 & & \\ & & & (\sigma_o^W)^2 & \\ 0 & & & & (\sigma_o^V)^2 \end{pmatrix} \tag{4.4}$$

where $(\sigma_o^X)^2; (\sigma_o^Y)^2; (\sigma_o^Z)^2; (\sigma_o^W)^2$ and $(\sigma_o^V)^2$ are the observation variances of the model variables $X; Y; Z; W$ and V respectively.

Then, from Section (2.1.1) the observations y_k at time t_k are generated by adding random noises to the truth state x_k at the same time. Then, Equation (2.2) becomes

$$y_k = x_k + \sigma_o \epsilon_k \tag{4.5}$$

where ϵ_k is the vector of random unbiased Gaussian errors at time t_k and σ_o is the 5×5 diagonal matrix whose diagonal terms are the standard deviations which are the square roots of observation variances of $X; Y; Z; W$ and V. The matrix σ_o is identical for each time step.

The states of observations depend on the observation frequency and accordingly the number of time step over the assimilation window. For example, assume the observation frequency is a vector of $[2; 2; 2; 4; 4]$ where each element represents observation frequencies of the model variables $X; Y; Z; W$ and V respectively, and assume there are 20 time steps on the assimilation window. This means that we have observations every two time steps for each variable of the atmosphere $(X; Y; Z)$ and every four time steps for each variable of the ocean $(W; V)$. Hence, this example determines $X; Y$ and Z

have 10 observations separately, in total 30 observations in the atmosphere, and W and V have 5 observations separately, in total 10 observations in the ocean. In our case, we assume that there is no observation at the initial time.

4.2.3 Determining the initial analysis state

We perform the assimilation by using the method 4DVar to nd the best analysis trajectory whose initial state x_0^a

error covariance matrix B. We use different vectors of errors denoted ϵ^i to generate each background ensemble member $x_0^{b;i}$ where i denotes the ensemble members from 1 to m. Thus, we get certain number of background states at the initial time. It is important to note that each analysis of ensemble members uses the corresponding background state $x_0^{b;i}$.

4.3.2 Perturbed observations

The observations are generated from Equation (4.5) by choosing different ϵ_k^i for each observation y_k^i for $i = 1; :::; m$. Isaksen et al. (2010) have done this process by adding perturbations to the previous observation. Then, we have certain observations for each background state. Note that each ensemble member has the same number of observations with identical observation frequency.

4.3.3 Generating ensemble of analysis states

We have the initial backgrounds with corresponding observations. Finally, we run the 4DVar for each ensemble member to generate their initial analysis states. Thus, we have m analysis states at the initial time. Then, we generate their trajectories with the RK2 method over the assimilation window. Figure (4.1) is just an example of the evolutions of 900 analysis ensemble members with the truth state on the time length 4. This figure is plotted by our program used in the experiments.

4.3.4 The ensemble based error covariance matrix

The analysis error covariance matrix P_a at the initial time and the forecast error covariance matrix P_f at the final time of the assimilation window can

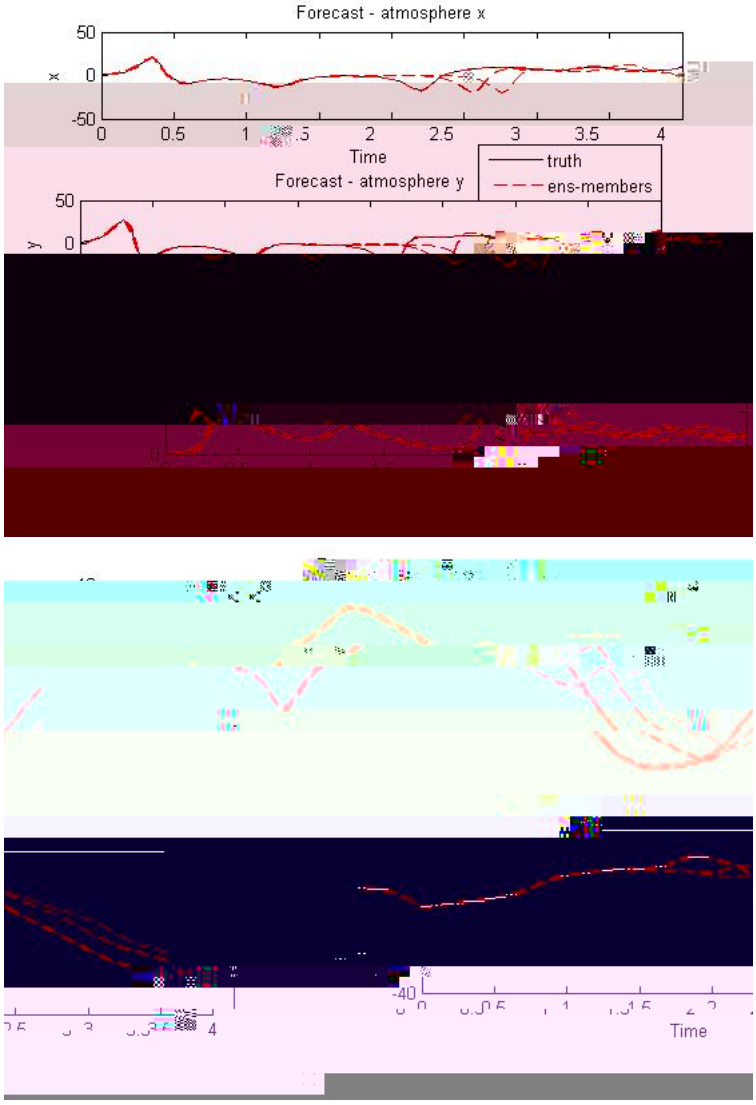


Figure 4.1: The evolutions of 900 ensemble members for each model variable

and C is calculated directly from

$$C = A^{-1} P A^{-1} \tag{4.9}$$

where A is the square root of D. The value of each component of C is in the range of $[-1; 1]$. The value 1 shows variables are fully correlated (while one variable is increasing the other also increases, and vice versa), while in the case of -1 the values are fully anti-correlated (while one variable is increasing, the other decreases, and vice versa). The closer the coefficient is to either -1 or 1, the stronger the correlation between the variables. When the value is getting close to 0, the variables are starting to become uncorrelated. The value 0 shows that the variables are independent and so they are totally uncorrelated.

Chapter 5

Assimilation experiments

In the following three experiments, we try to investigate our research questions which are described in Section (1.1). Therefore, we try to estimate the ensemble based forecast error covariance matrix P_f with different ensemble sizes. To be able to do this, we look at the convergence of P_f components separately as the ensemble size gets larger. Then, we investigate how they are affected by the error correlations of model variables and, the accuracy and the number of observations.

The stationary parameters are defined at the beginning of Chapter 4. In addition, the background error variances of the model variables $X; Y; Z; W$ and V are also fixed as 10^{-2}. Thus, the backgrounds of model variables have same accuracy (the distance between background and truth state can show small differences because of the random noise, see Equation (4.2)). We generate our ensembles from size 6 to 1800. We define the minimum ensemble size experimentally. We get the P_f full rank ($\text{rank}(P_f) = 5$) with the minimum ensemble size 6. Hence, all experiments start with 6 ensemble members. Then, the next ensemble size is set up as 100 and it carries on as

consecutive multiples of 100 up to 1800 ensemble members. Thus, we use 19 di erent ensemble sizes in each experiment. Then, we compare the values of P_f components after each ensemble. To be able to estimate the matrix P_f, it is expected that each component of the matrix converges to a number. Therefore, we try to get understanding of how many ensemble members make the P_f matrix convergence. This can give approximate relationships between error spreads in the model. Choosing a su cient and appropriate convergence condition provides a basis for our experiments.

Assume M and N are subsequent ensemble sizes where $M < N$. We de ne the relative convergence condition as

$$\frac{|e_M - e_N|}{|e_M|} < \qquad (5.1)$$

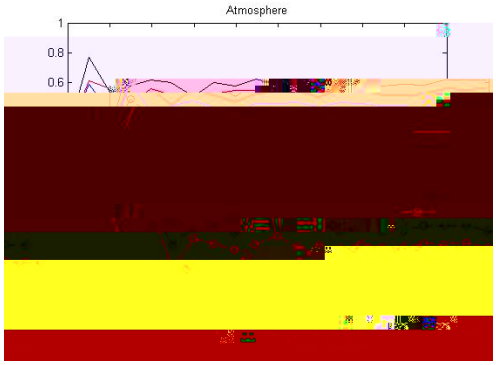
where e denotes each component of P_f and called convergence condition limit is an arbitrary number discussed in the experiments. This relative condition seems appropriate for a convergence condition because some components of P_f can have values with di erent orders of magnitude. Their approximate values are given in Tables (5.2) and (5.5). In these tables

From now on when we say convergence behaviour of any component, it represents a broken line which combines each value calculated by Equation (5.1) from the ensemble size 6 to 1800. The closeness to the convergence condition limit and the smoothness of this line will be compared in our experiments.

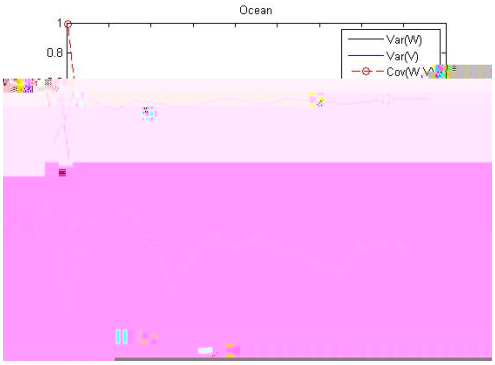
5.1 Experiment-1

In this experiment, we try to investigate our first and second research questions (see Section (1.1)). We set up the observation error variances of $X; Y; Z$ as 10^{-4} and $W; V$ as 10^{-2}. In other words, we use more accurate observations in the atmosphere than in the ocean according to the ranges of their error spreads. Additionally, the observation frequency is $[5; 5; 5; 5; 5]$ so we have observations in every five time steps for each model variable. We set up the number of time steps as 20. Hence, we have 4 observations for each variable $X; Y; Z; W$ and V at times t_5, t_{10}, t_{15} and t_{20} (we assume there is no observation at the initial time). This means that there are total 20 observations in each assimilation period that 12 of them are in the atmosphere while 8 of them are in the ocean.

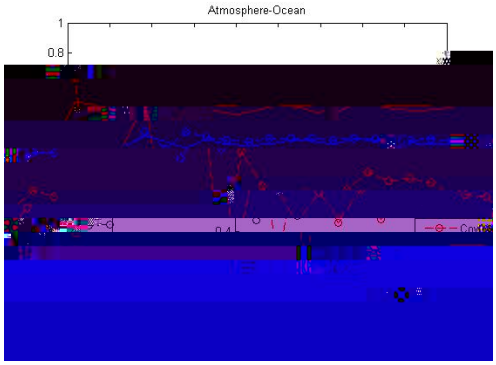
Figure (5.1) shows the relative values of P_f components with the ensemble sizes from 6 to 1800. The values of each component are plotted by multiplying with an appropriate one of the multiples of 10 to make its evolution in the range of $[-1; 1]$. For example, the values of $Var(X)$ in each ensemble are between 10^{-5} and 10^{-4}. We plot its values by multiplying 10^4. If the value range is between 10^{-5} and 10^{-3}, it will be plotted by multiplying



(a) Atmosphere part



(b) Ocean part



(c) Atmosphere-Ocean part

Figure 5.1: The relative values of each component of P_f for di erent ensemble sizes

10^3. This way is one of the clearest ways to see their behaviour in the same scale. Otherwise, it is hard to see their relative changes in di erent scales. Their approximate values can be seen in the Experiment-1 part of Table (5.2). In this table, the represented values are the average of the values of P_f

the sense of the convergence condition in Equation (5.1). To be able to get a more clear idea of their convergences, we plot their relative changes with respect to Equation (5.1) with the ensemble size getting bigger. The results are seen in Figures (5.2), (5.3) and (5.4). Some peak values cannot be seen in these gures because we xed the scale of gures between 0:5 and 1 to clarify their changes. On these gures, there are two dotted lines: red and green. We will discuss them later as a limit of convergence condition in Equation (5.1). As we expect from Figure (5.1), the remarkable jumps are seen on the relative changes of $Cov(W;V)$ (Figure (5.4)(d)) and $Cov(X;W)$ (Figure (5.3)(a)) with the ensemble size getting larger.

To have a general idea about how the error in a component is correlated with the error in another one, we calculate the correlation matrix of P_f represented in Section (4.3.4). Table (5.1) shows the average values of each di erent component of correlation matrix C. We calculate their arithmetic means from the results of each ensemble size from 6 to 1800. We do not think plotting their values at each ensemble member necessary, because the values of each component is quite close to the its mean value. From this table, we can see there is almost no correlation between the errors of X and W, and similarly between the errors of W and V. On the other hand, the atmosphere itself has highly correlated errors.

If we consider the convergence behaviour of one component with its corresponding correlation mean value, we can see that they seem to be related. For example, consider the case of $Cov(X;W)$ in Figure (5.3)(a), it has the most spurious behaviour with the ensemble sizes and then we see in Table (5.1) that the errors of W and V have the lowest correlation (0:0151). On

the other hand, the convergence behaviour of $Cov(X;V)$ in Figure (5.3)(b) has much smoother behaviour than $Cov(X;W)$. Then, we look at the error correlations of X and V in Table (5.1), they are almost fully correlated with the correlation value 0.8558. Therefore, we can conclude that the less correlation in errors the less convergence tendency in its convergence behaviour, and vice versa. From this view, we can expect to see the best convergence behaviour on the variances of $X;Y;Z;W$ and V because they are fully correlated in their own right (each diagonal term of correlation matrix C is 1). This can determine the limit of convergence condition .

Now, we can focus on the limit of convergence condition . In the gures showing relative conditions of each model variable, there are two speci ed s. The red dotted line is the line of 0.05 whereas the green one is that of 0.2. From Figure (5.2), to be able to make the variances of each component always under the green dotted line after a certain ensemble size (i.e the size of 300), we can de ne the limit of convergence condition as the line 0.2 (the green dotted line). Otherwise, on the condition of 0.05 (the red dotted line) we cannot generalize that these variances have always satis ed the convergence condition after a certain ensemble size. Thus, from now on we de ne the limit of convergence condition as 0.2.

The reason for signi cant jumps on the convergence behaviour of $Cov(X;W)$ can be because of the model system (see Equation (3.1)). There is no direct relationship between X and W variables. This may indicate that more ensemble members are needed to capture the information between their errors. From this view, we can expect similar jumps on the convergence behaviours of $Cov(Y;V)$, $Cov(Z;W)$ and $Cov(Z;V)$, because there are also no direct

Atmosphere	
Cor(X,Y)	0.7737
Cor(X,Z)	-0.6748
Cor(Y,Z)	-0.9596
Ocean	
Cor(W,V)	-0.1113

Atmosphere-Ocean	
Cor(X,W)	-0.0151
Cor(X,V)	0.8558
Cor(Y,W)	0.5304
Cor(Y,V)	0.4471
Cor(Z,W)	-0.5340
Cor(Z,V)	-0.2847

Table 5.1: Averages of each error correlation from the ensemble size 6 to 1800 in Experiment-1

relationships between Y and V, Z and W, and Z and V. However, their

numbers used in generating ensemble members. Therefore, considering the convergence behaviours of the components of P_f from the ensemble size 6 to 1800 in gures can be more logical to compare the convergences of components.

As the ensemble size getting bigger from 6 to 1800, the convergence behaviour of Cov(W;V) shows signi cant uctuations especially up to the size 800 in Figure (5.1)(b). In addition, it can be seen in Table (5.1) that there is a small correlation in the errors of W and V (Cor(W,V)=-0.1113). These results can be from the parameters we set up. In this experiment, we have less accurate observations in the ocean (each variance of W and V is 10^{2}) than in the atmosphere (each variance of X;Y and Z is 10^{4}), and the number of observations in the ocean (8 observations) less than in the atmosphere (12 observations). In the following two experiments, Experiment-2 and Experiment-3 we will examine the e ects of the accuracy of ocean observations and the number of ocean observations on the convergence of P_f respectively. The setup parameters used in this experiment will provide a basis for the parameters used in other two. In the gures, the results of each following experiment will be plotted with the results of this experiment to notice their di erences clearly.

5.2 Experiment-2

In this experiment, we examine the e ects of the accuracy of ocean observations on the convergence of P_f, which is a part of our third research question (see Section (1.1)). Thus, we do similar experiments as in Experiment-1 with more accurate ocean observations. The only di erence in the parameters set

up in the previous experiment is that the observation error variance of W and V becomes 10^{-4} from 10^{-2}. By doing so, we make the observation variances of all variables ($X; Y; Z; W$ and V) the same. Table (5.2) shows the approximate value of each component of P_f. This table is created by the mean of the values of each P_f component as the ensemble size gets bigger from 6 to 1800. We compare the results of Experiment-1 and 2 and see that more accurate ocean observations lead the errors to spread less. Remarkable decreases are seen in the ocean covariances and the cross-covariances between the atmosphere and ocean. However, the convergence property becomes more erratic in general that we can also expect from Table (5.3) according to our hypothesis. In other words, when we get more accurate ocean observations there are more fluctuations on the convergence behaviour of each component of P_f. In this table, except the atmosphere part, the approximate error correlations in the ocean and in the atmosphere-ocean part are almost zero so we can expect significant fluctuations on their convergence behaviours. Then to see whether or not our expectations are satisfied, the relative changes of P_f components are plotted in Figures (5.5),(5.6) and (5.7). The figures also include the results of the first experiment to see the differences between them clearly. More accurate ocean observations cause significant jumps in convergence behaviours of most P_f components.

In Figure (5.6), except the case of $Cov(X; W)$ the convergence behaviour shows more fluctuations in overall when we compare the Experiment-1 result. For example, $Cov(X; V)$ does not satisfy the convergence condition after 500 ensemble members although it always satisfies the convergence condition after 300 ensemble members in Experiment-1. However, in the case of $Cov(X; W)$ the overall convergence behaviour is better than in the first

experiment. When we compare the error correlations of both experiments in Tables (5.1) and (5.3), it can be seen that while $Cor(X;W)$ increases from -0.0151 to -0.1188 (consider the magnitudes of numbers), others in the atmosphere-ocean part decrease remarkably.

In the variances figure (Figure (5.5)), all of them satisfy the converge condition first with less ensemble size than that of Experiment-1. For example, $Var(Z)$ (in Figure (5.5)(c)) satisfies the convergence condition first with the ensemble size 200 in Experiment-1 whereas it needs 100 ensemble members in Experiment-2.

As we can see in Figure (5.7), more accurate ocean observations make the atmosphere covariances $Cov(X;Y)$, $Cov(X;Z)$, $Cov(Y;Z)$ convergence with all ensemble sizes. If we look at their error correlations in Table (5.3), they are almost fully correlated with the values 0.9915, $0{:}9758$ and $0{:}9570$ respectively. However, in the case of $Cov(W;V)$ there is an almost opposite situation. While $Cov(W;V)$ converges with the 300 ensemble members in Experiment-1, here it needs much more ensemble members than in the first experiment to converge. A decrease from 9570

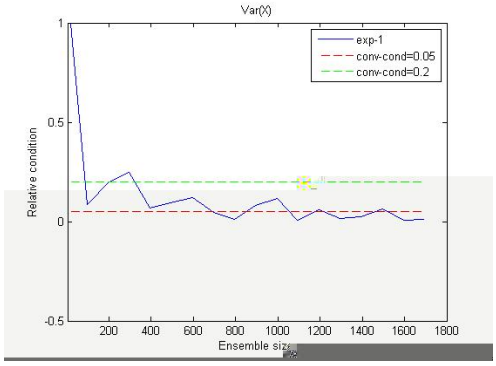
	Atmosphere	
	Experiment-1	Experiment-2
Var (X)	0.00005324	0.00001426
Var (Y)	0.00004628	0.00002096
Var (Z)	0.00006124	0.00003629
Cov (X,Y)	0.00003841	0.00001714
Cov (X,Z)	-0.00003864	-0.00002220
Cov (Y,Z)	-0.00005113	-0.00002639

	Ocean	
	Experiment-1	Experiment-2
Var (W)	0.00389633	0.00002497
Var (V)	0.00466764	0.00002583

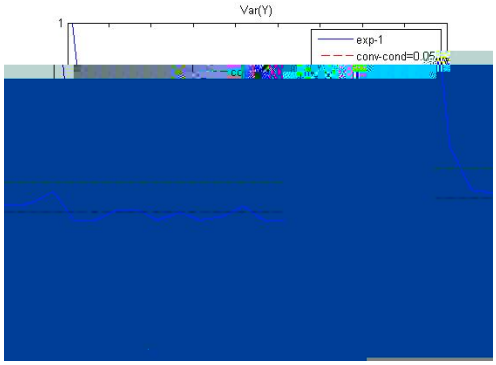
Atmosphere-Ocean	
Experiment-1	Experiment-2

the ensemble size 6 to 1800 in the gures because at each ensemble size the values can show some di erences if we do same experiment with exactly same parameters. Random numbers used generating ensemble members can cause small di erences in their values at the certain ensemble size. For the case of $Cov(X; W)$, we can compare their correlation changes between Table (5.1) and (5.4). The error correlation between X and W increase slightly in the case wd

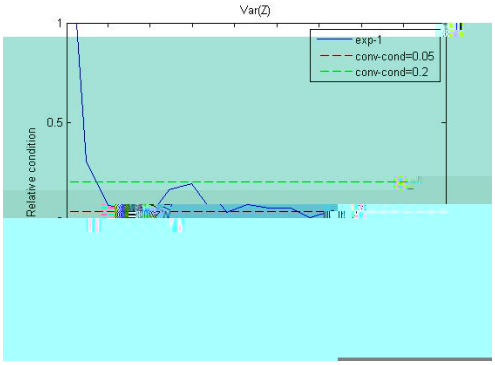
	Atmosphere	
	Experiment-1	Experiment-3
Var (X)	0.00005324	0.00003756



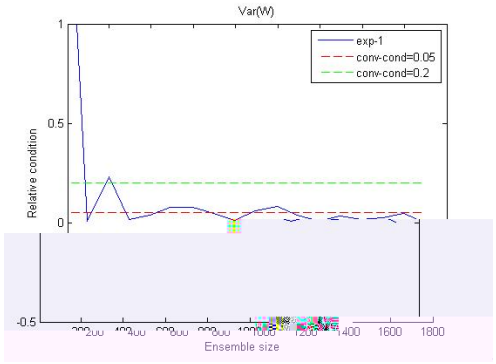
(a) Var(X)



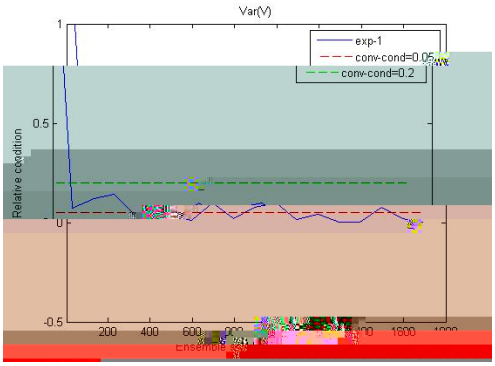
(b) Var(Y)



(c) Var(Z)

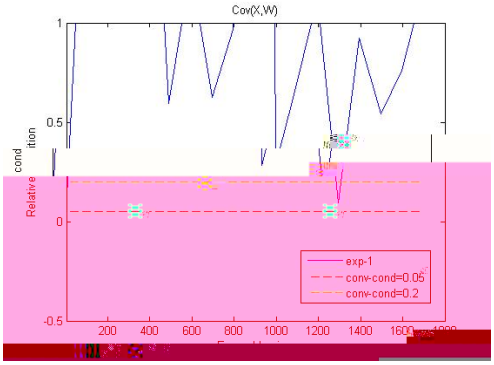


(d) Var(W)

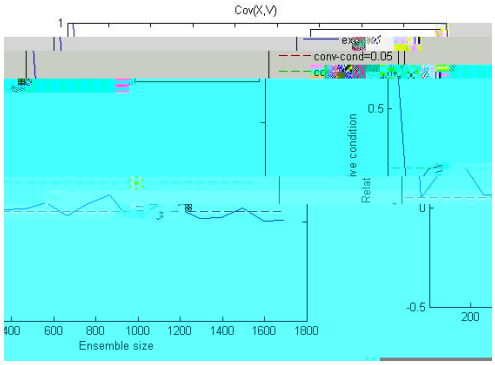


(e) Var(V)

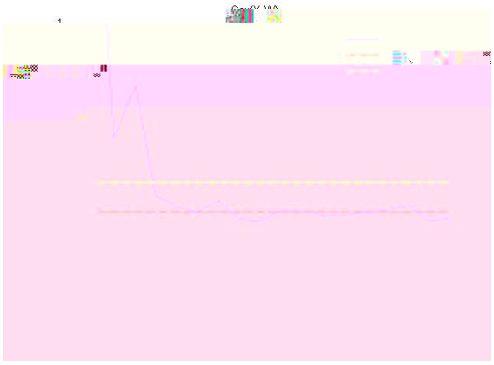
Figure 5.2: The values of convergence condition of variances for di erent ensemble sizes in Experiment-1



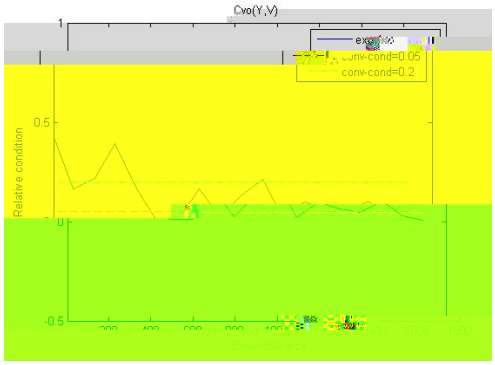
(a) Cov(X,W)



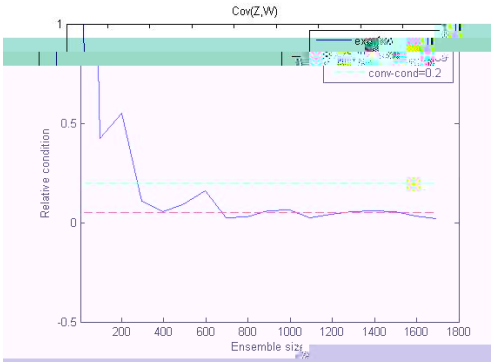
(b) Cov(X,V)



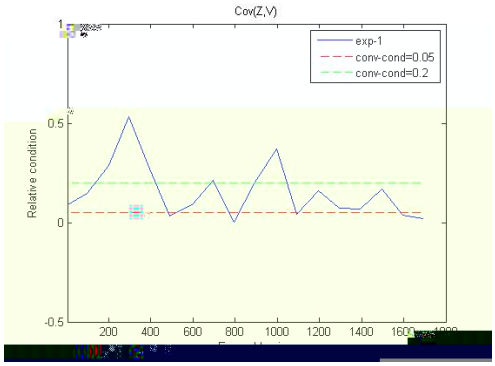
(c) Cov(Y,W)



(d) Cov(Y,V)

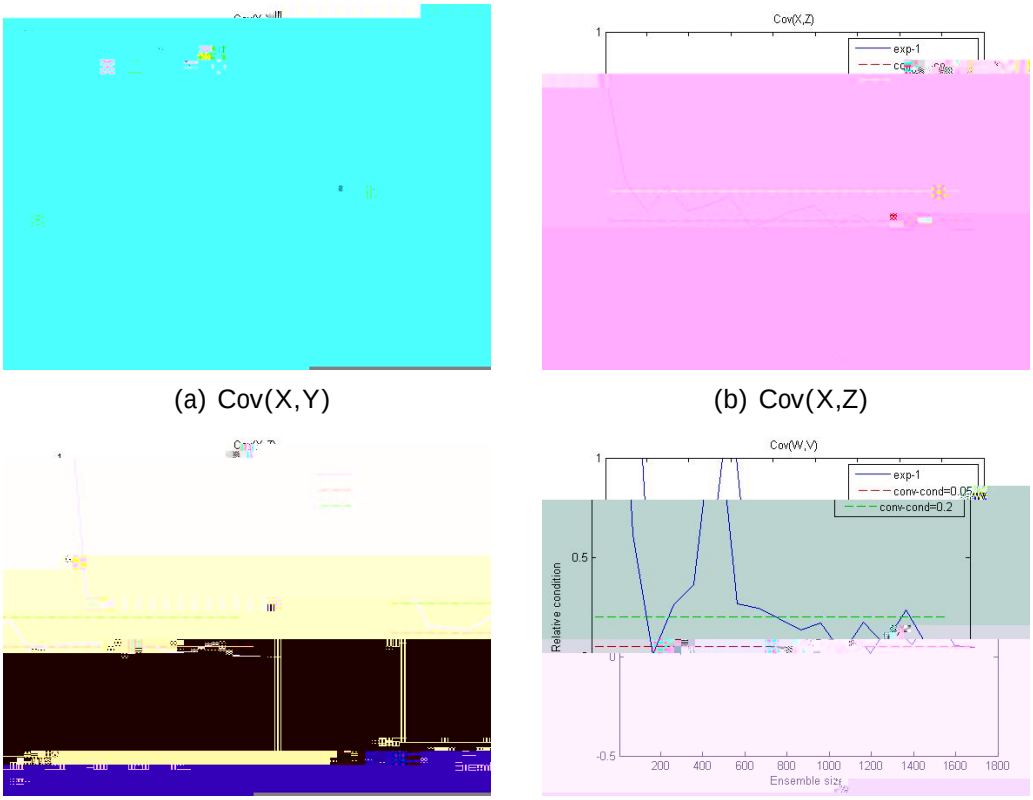


(e) Cov(Z,W)



(f) Cov(Z,V)

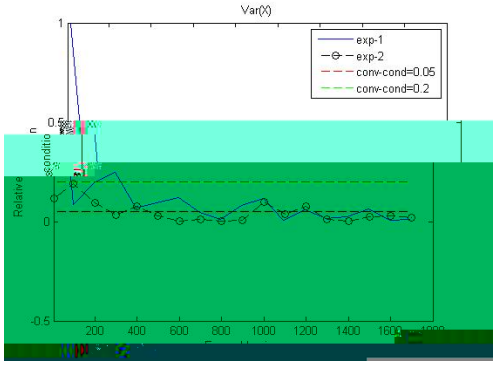
Figure 5.3: The values of convergence condition of cross-covariances for different ensemble sizes in Experiment-1



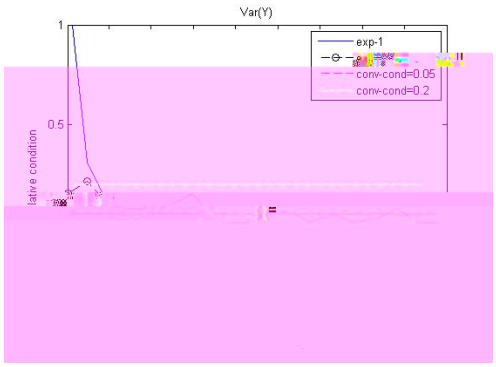
(a) Cov(X,Y) (b) Cov(X,Z)

(c) Cov(Y,Z) (d) Cov(W,V)

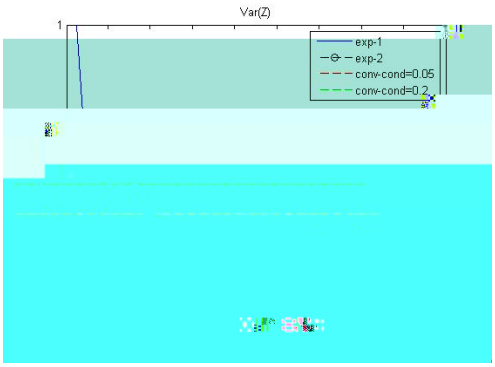
Figure 5.4: The values of convergence condition of covariances for di erent ensemble sizes in Experiment-1



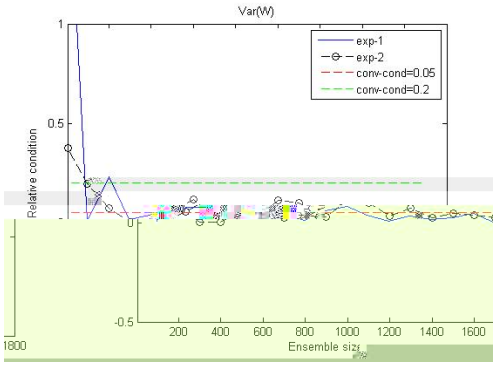
(a) Var(X)



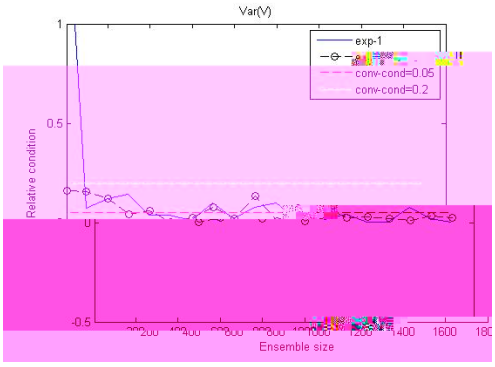
(b) Var(Y)



(c) Var(Z)

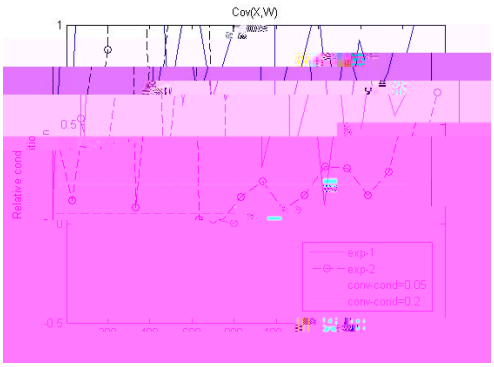


(d) Var(W)

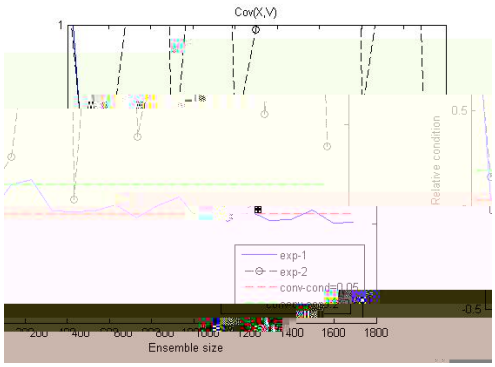


(e) Var(V)

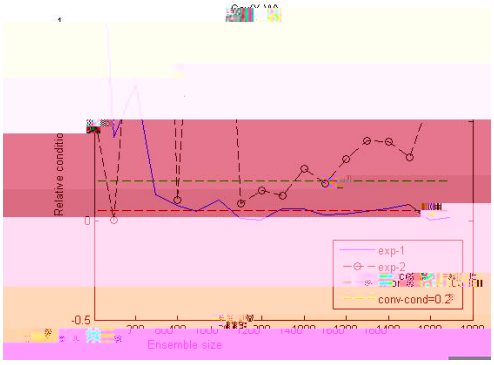
Figure 5.5: The values of convergence condition of variances for di erent ensemble sizes in Experiment-2 are compared with the results in Experiment-1



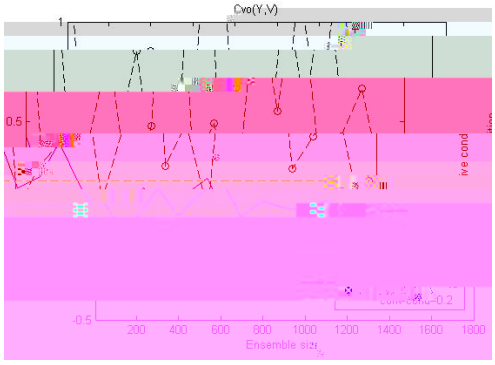
(a) Cov(X,W)



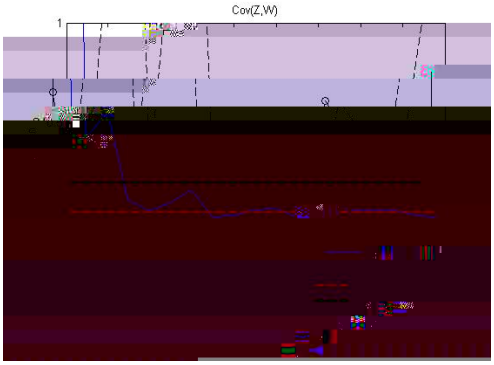
(b) Cov(X,V)



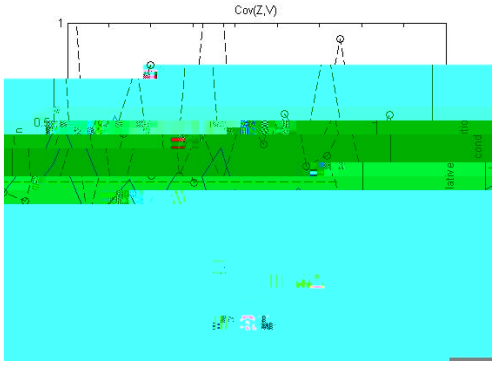
(c) Cov(Y,W)



(d) Cov(Y,V)

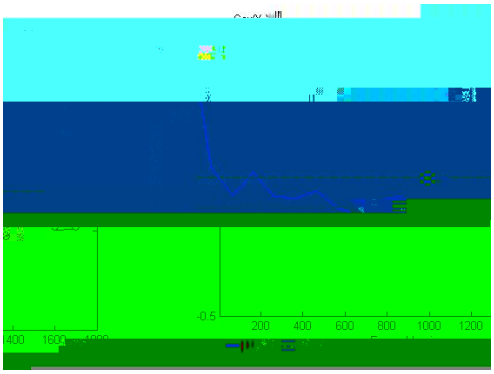


(e) Cov(Z,W)

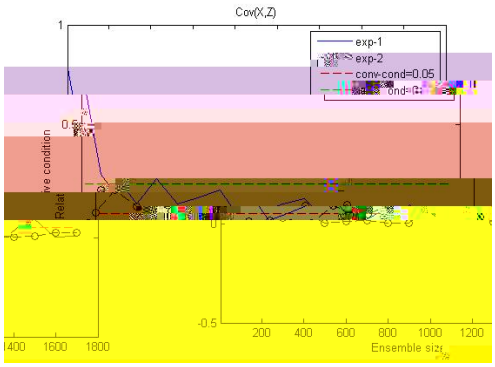


(f) Cov(Z,V)

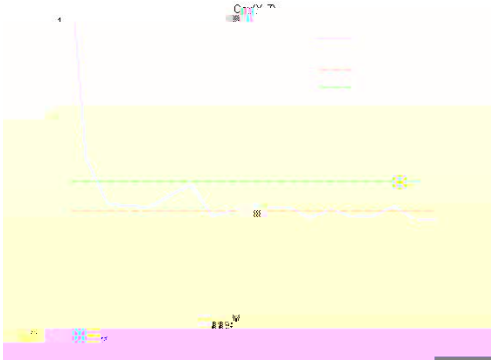
Figure 5.6: The values of convergence condition of cross-covariances for di erent ensemble sizes in Experiment-2 are compared with the results in Experiment-1



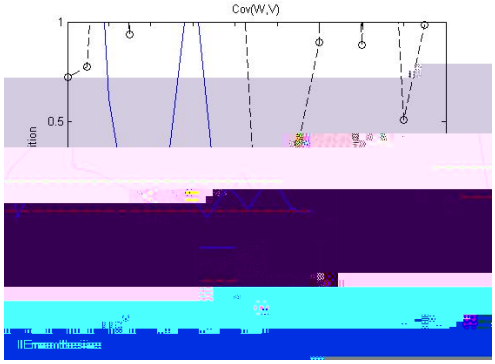
(a) Cov(X,Y)



(b) Cov(X,Z)

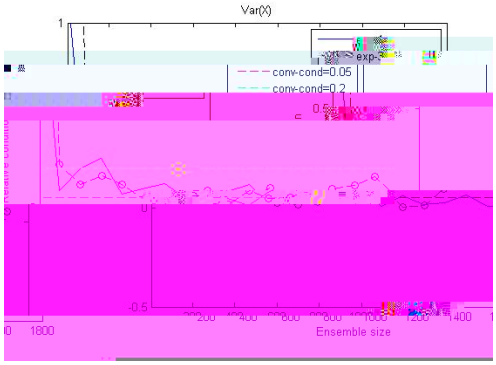


(c) Cov(Y,Z)

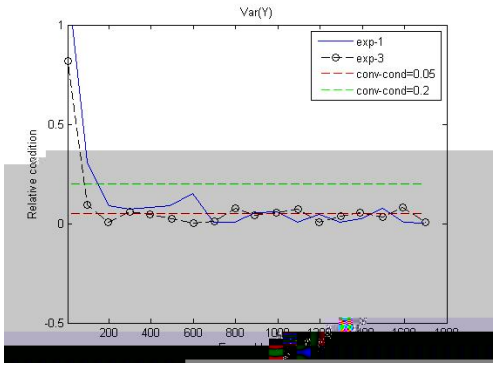


(d) Cov(W,V)

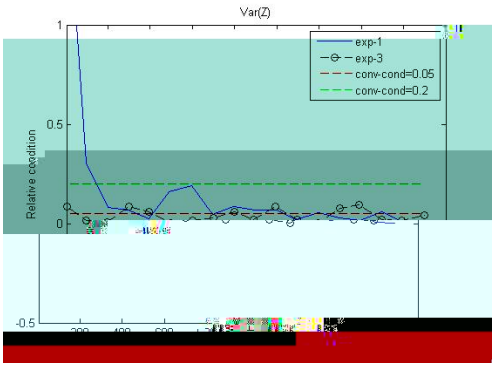
Figure 5.7: The values of convergence condition of covariances for di erent ensemble sizes in Experiment-2 are compared with the results in Experiment-1



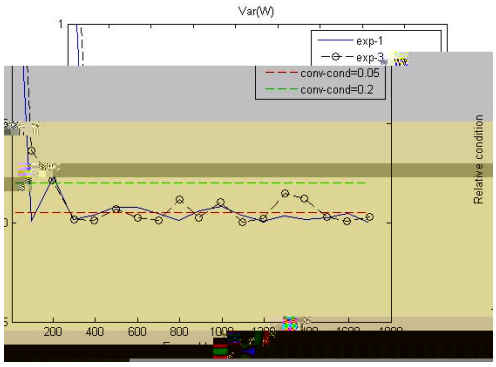
(a) Var(X)



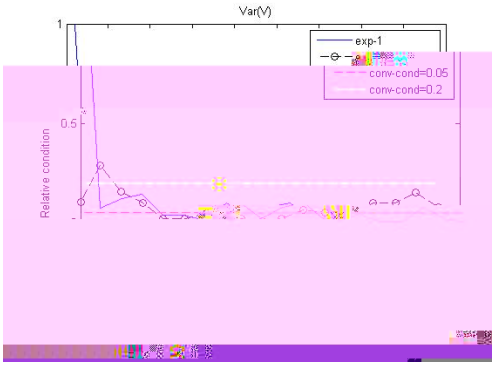
(b) Var(Y)



(c) Var(Z)

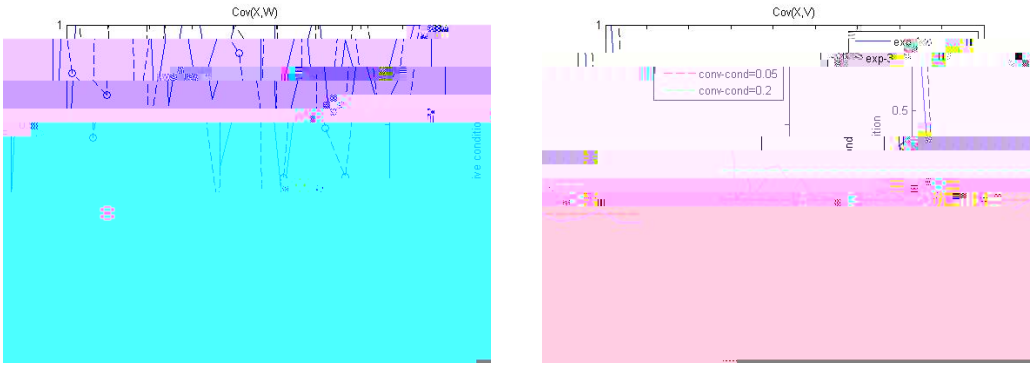


(d) Var(W)

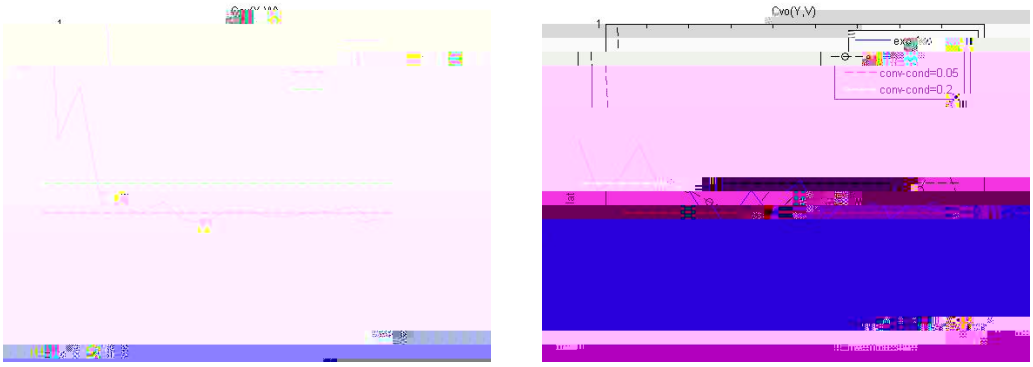


(e) Var(V)

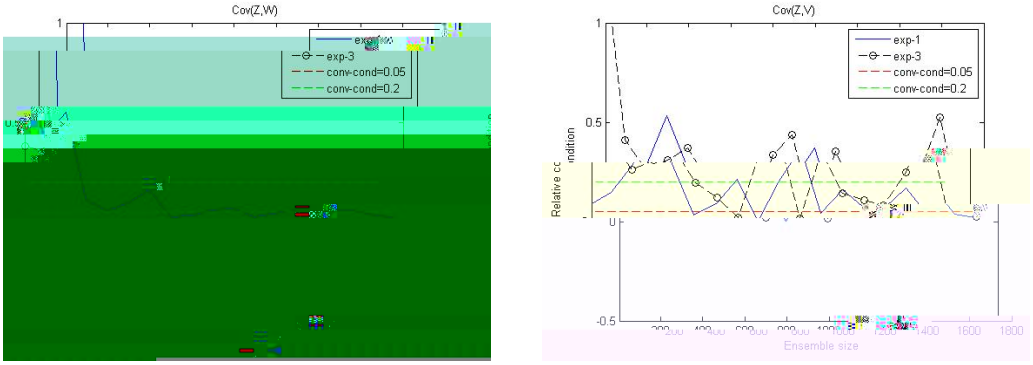
Figure 5.8: The values of convergence condition of variances for di erent ensemble sizes in Experiment-3 are compared with the results in Experiment-1



(a) Cov(X,W) (b) Cov(X,V)

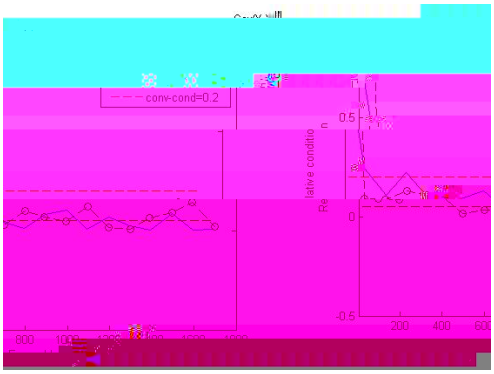


(c) Cov(Y,W) (d) Cov(Y,V)

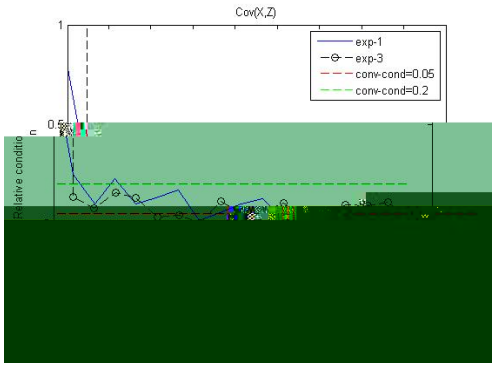


(e) Cov(Z,W) (f) Cov(Z,V)

Figure 5.9: The values of convergence condition of cross-covariances for di erent ensemble sizes in Experiment-3 are compared with the results in Experiment-1

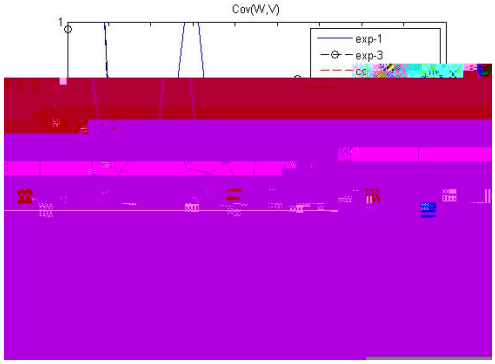


(a) Cov(X,Y)



(b) Cov(X,Z)

(c) Cov(Y,Z)



(d) Cov(W,V)

Figure 5.10: The values of convergence condition of covariances for di erent ensemble sizes in Experiment-3 are compared with the results in Experiment-1

Chapter 6

Discussion

6.1 Summary and Conclusion

The aim of this project is to investigate following questions:

(1) How many ensemble members are needed to capture the forecast covariances correctly?

(2) What are the effects of error correlations on the estimation of forecast covariances?

(3) How these are affected by observation errors and frequencies?

To do this, first we generate a method which is an ensemble of non-incremental 4DVar methods. In this method, we generate ensemble members by perturbing background and observations with different random numbers. Then, we try to estimate the forecast error covariance matrix with different ensemble sizes. We try to look at the convergence behaviour of P_f compo-

nents as the ensemble size getting larger and demonstrate the e ects of error correlations on the convergence of P_f. Then, by changing the accuracies and the number of ocean observations we try to show the di erences in the convergence behaviours of P_f components.

In our experiments, to be able to estimate P_f, we investigate the convergence behaviour of its components as the ensemble size getting bigger. Each components of P_f converges to di erent numbers which can have di erent orders of magnitude. Thus, relative convergence condition is chosen to determine their convergence. Every component of P_f is tested separately with respect to the relative condition. We mostly focus on their convergence behaviour not the numbers they converge. We can also guess their convergence behaviour by looking at the error correlations. From our experiment results, we see that there are strong relationship between convergence behaviour of one component of P_f and the error correlation of its corresponding variables. The less correlation in errors the less convergence tendency in its convergence behaviour, and vice versa. It is important to know that the experimental setups used in this dissertation are not ideal ones. We use particular parameters sets. It would be useful change the parameters and do similar experiments.

In the rst experiment, we try to estimate the matrix P_f in the system of more accurate observations with more observations in the atmosphere than in the ocean. We see that there are signi cant jumps on the convergence behaviour of $Cov(X;W)$ and $Cov(W;V)$. When we look at the error correlations of (X and W) and (W and V) separately, there are almost no correlation between them. The reason for the convergence behaviour of $Cov(X;W)$ can be because of the model system. There is no direct con-

nection between X

discretizing the model step by step. At the final time, we have an ensemble based forecast error covariance matrix P_f which we try to estimate in our experiments. After the estimation of the matrix, it should be used for the next assimilation window as a background error covariance matrix. Then, it would be great to generate new analysis ensemble members by using P_f.

The matrix P_f includes the effects of the forecast. Investigating the convergence of P_a, which is the analysis error covariance matrix at the initial time, would be salient to eliminate the forecast effects on this matrix. We would suggest to do same experiments for the estimation of P_a.

Bibliography

Anderson, J. L. and Anderson, S. L. (1999). A Monte Carlo implementation of the nonlinear filtering problem to produce ensemble assimilations and forecasts. Monthly Weather Review, 127(12), 2741-2758.

Bouttier, F. and Courtier, P. (2002). Data assimilation concepts and methods. Meteorological Training Course Lecture Series, ECMWF http://old.ecmwf.int/newsevents/training/rcourse_notes/DATA_ASSIMILATION/ASSIM_CONCEPTS/Assim_concepts2.html

Cheng, H., Jardak, M., Alexe, M., and Sandu, A. (2010). A hybrid approach to estimating error covariances in variational data assimilation. Tellus A, 62(3), 288-297.

Dimet, F. X. L. and Talagrand, O. (1986). Variational algorithms for analysis and assimilation of meteorological observations: theoretical aspects.

1-7.

Evensen, G. (1994). Sequential data assimilation with a nonlinear quasigeostrophic model using Monte Carlo methods to forecast error statistics. Journal of Geophysical Research: Oceans (19782012), 99(C5), 10143-10162.

Fairbairn, D., Pring, S. R., Lorenc, A. C., and Roulstone, I. (2014). A comparison of 4DVar with ensemble data assimilation methods. Quarterly Journal of the Royal Meteorological Society, 140(678), 281-294.

Grewal, M. S., and Andrews, A. P. (2011). Kalman ltering: theory and practice using MATLAB. John Wiley and Sons.

Hamill, T. M., Whitaker, J. S., and Snyder, C. (2001). Distance-dependent ltering of background error covariance estimates in an ensemble Kalman Filter. Monthly Weather Review, 129(11), 2776-2790.

Isaksen, L., Bonavita, M., Buizza, R., Fisher, M., Haseler, J., Leutbecher, M., and Raynaud, L. (2010). Ensemble of data assim-

Kalman, R. E. (1960). A new approach to linear filtering and prediction problems. Journal of Fluids Engineering, 82(1), 35-45.

Kalnay, E. (2003). Atmospheric modelling, data assimilation, and predictability. Cambridge university press.

Lawless, A.S. (2006). Data assimilation with the Lorenz equations. http://darc.nerc.ac.uk/, 12/08/2006, simple models

Lawless, A.S. (2013). Variational data assimilation for very large environmental problems.In Large Scale Inverse Problems: Computational Methods and Applications in the Earth Sciences(2013), Eds. Cullen, M.J.P., Freitag, M. A., Kindermann, S., Scheichl, R., Radon Series on Computational and Applied Mathematics 13. De Gruyter, pp. 55-90.

Liu, C., Xiao, Q., and Wang, B. (2008). An ensemble-based four-dimensional variational data assimilation scheme. Part I: Technical formulation and preliminary test. Monthly Weather Review, 136(9), 3363-3373.

Lorenz, E. N. (1963). Deterministic non-periodic flow. Journal of the atmospheric sciences, 20(2), 130-141.

Molteni, F., Ferranti, L., Palmer, T. N., and Viterbo, P. (1993). A dynamical interpretation of the global response to equatorial Pacific SST anomalies. Journal of climate, 6(5), 777-795.

Nichols, N. K. (2010). Mathematical concepts of data assimilation. In Data Assimilation (pp. 13-39). Springer Berlin Heidelberg.

Voss, R. and Sausen, R. (1996). Techniques for asynchronous and periodically synchronous coupling of atmosphere and ocean models. Climate Dynamics, 12(9), 605-614.