THE UNIVERSITY OF READING

Information Content of Observations in Variational Data Assimilation

Christine Johnson

A thesis submitted for the degree of Doctor of Philosophy

Department of Meteorology

September 2003

Declaration

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

Christine Johnson

Abstract

Data assimilation is needed to generate an analysis, which is used as the initial conditions for numerical weather prediction. Four-dimensional variational data assimilation (4D-Var) is the most advanced data assimilation algorithm to be used operationally; it uses observations that are distributed in time through the use of the model equations.

The aim of this thesis is to understand the extent to which 4D-Var can develop the structures needed for the growth and decay of baroclinic systems. Such mid-latitude storms can cause severe damage and play a key role in the evolution of the atmospheric flow. The approach taken isolates the important mechanisms in 4D-Var by considering a simple model of baroclinic instability.

Idealized case studies using the 2D Eady model consider the use of a time-sequence of observations to reconstruct the state in unobserved regions. A novel technique using the singular value decomposition of the 4D-Var observability matrix is developed, based on methods that are commonly used in satellite retrieval studies. It is used here to provide a new and useful understanding of the information content of observations in 4D-Var.

It is shown that the information that is propagated to the unobserved regions is strongly penalized by the background state and is also extremely sensitive to observational noise. This is understood by establishing a link with the literature on Tikhonov Regularization. An analysis increment will result in growth if the observations are given at only the end of the window, or if a relatively large weight is given to the background state. This may result in a poor forecast if the required analysis increments lead to decay. Two ways to maximize the benefits of 4D-Var are identified: the initial and final observations should be far apart in time, and appropriate values for the regularization parameters should be specified.

Acknowledgements

I first thank my academic supervisors, Prof. Brian Hoskins and Prof. Nancy Nichols. I am especially grateful for their enthusiasm about research in Meteorology and Mathematics and for their guidance and confidence in my work. I also thank my Met Office supervisor Sue Ballard, whose knowledge about data assimilation has been invaluable. I am grateful to Amos Lawless, for his willingness to be involved with my work and for his time and patience in helping me to understand 4D-Var.

I thank the members of staff, researchers and students in both the Meteorology and Mathematics departments, who I have had many useful discussions with. I particularly thank my thesis committee members: Ian James, Mike Pedder, and John Thuburn and members of the Dynamics group and the Data Assimilation group.

Both past and present PhD students have helped to make my time in Reading fun and I am extremely grateful for the prayers and support from members of the Christians in Meteorology group and Wycliffe Baptist Church. I thank my family for their love and care, and especially my husband Ben who has shared with me the joys and pains of PhD student life. I thank God for the privilege of begin able to study at the University of Reading and for creating an interesting world to research.

Finally, I acknowledge the financial support of the Natural Environment Research Council (NERC) and the Met Office, without whom this work would not have been possible.

Contents

‘I don’t understand you,’ said Alice. ‘It’s dreadfully confusing!’

‘That’s the effect of living backwards,’ the Queen said kindly: ‘it always makes one a little giddy at first—-’

‘Living backwards!’ Alice repeated in great astonishment. ‘I never heard of such a thing!’

‘–but there’s one great advantage in it, that one’s memory works both ways.’

– Lewis Carroll, “Through the Looking Glass”

Chapter 1

Introduction

1.1 Motivation

Data assimilation (DA) uses observational data to generate an analysis - the best estimate of the present state of the atmosphere. The analysis is used as the initial conditions for a numerical weather forecast. The chaotic nature of the atmosphere means that small errors in the initial conditions may amplify rapidly (Lorenz, 1993), and therefore the analysis needs to be as accurate as possible. For this reason, DA is one of the most important parts of numerical weather prediction (NWP).

The DA algorithm must be able to assimilate observations that are nonlinearly related to the analysis variables, for example, satellite data. The observations have errors and a sparse spatial distribution, so the DA algorithm must filter the noise and interpolate between the observation points, whilst ensuring that the fields are meteorologically realistic. The DA algorithm must also be able to solve large-dimensional problems, as in an operational context there are approximately 10^7 unknowns. Further, the algorithm must compute the analysis quickly, so that the forecast can be generated.

Variational DA finds the analysis by minimizing a cost function that contains two terms. One term penalizes the squared distance from the background state (usually a forecast that is valid at the same time as the analysis) and the second term penalizes the squared distance from the observations. The variational formulation is suitable for such large-dimensional problems, and also allows the use of observations that are (weakly) nonlinearly related to the analysis unknowns (Lorenc et al., 2000).

In three-dimensional variational DA (3D-Var), observations collected over a certain time period are assimilated by assuming that they are all taken at the same time. This means that there may be up to a three hour difference between the time of an observation and the time at which the background state is valid. 3D-FGAT (First Guess at the Appropriate Time) (Rabier et al., 1998) extends 3D-Var so that the background state is evolved to the time of the observations.

In four-dimensional variational DA (4D-Var), a whole time sequence of observations are assimilated, by linking them together with the numerical forecast model equations. The ability of 4D-Var to combine information from observations with the knowledge of the evolution of the atmosphere means that it is one of the most advanced algorithms to have been used in operational NWP (Rabier et al., 2000). 4D-Var is much more computationally expensive than 3D-Var and it is therefore important to assess whether the advantages of 4D-Var can justify the expense.

It has been shown that 4D-Var produces better results than 3D-Var in situations of baroclinic growth (e.g. Rabier and Courtier, 1992, Rabier et al., 1998, Desroziers et al., 1999, Rabier et al., 2000). Baroclinic instability is one of the dominant mechanisms for the generation of mid-latitude depressions. These depressions can cause severe damage and it is important that NWP centres are able to forecast them well. Thus, it is crucial that the DA algorithm correctly develops the initial conditions needed for such storms. The capability of 4D-Var to develop the correct vertical structures needed for the growth and decay of baroclinic systems is the subject of this thesis.

This chapter begins by briefly summarizing a history of data assimilation methods, so that we can understand why 4D-Var is known as an advanced data assimilation algorithm. Then, the current knowledge regarding 4D-Var in the presence of baroclinic growth is summarized. The chapter finishes by discussing issues that have not yet been researched, stating the key questions that are addressed in this thesis, and the layout of the rest of the thesis.

1.2 History of Data Assimilation

A brief overview the history of DA and a description of current DA algorithms are now given. More detailed overviews may be found in, for example, Daley (1991), Ghil and Malanotte-Rizzoli (1991), Wunsch (1996) and Bouttier and Courtier (2003).

In as early as 1850, the very first synoptic charts were created. Observations were plotted on geographical maps, and isobars and isotherms were drawn on by hand. There were very few observations, so the analyst would use his knowledge of a previous synoptic chart together with his knowledge of how weather systems evolve, to infer the position of the isolines in the data void regions. Knowledge about the relationships between different variables could also be applied. For example, the geostrophic balance relationship states that the wind direction is approximately parallel to the isobars, and that the isobars are closer together in regions where the wind speed is stronger. These synoptic charts were the very first type of subjective analysis. The method relied heavily upon the subjective judgement of the analyst, and as the charts were produced by hand, the method would take a great deal of time. When numerical weather models were created, the synoptic chart had to be transferred into a gridded data set to use as the initial conditions. It was realized that computers would be able to generate much better objective analyses in a shorter time. However, the underlying concepts of subjective analysis, such as using extra knowledge about the evolution of the atmosphere and atmospheric balance, are still used in present day objective analysis algorithms.

The first objective analysis algorithms fitted polynomials to the observations by minimizing the squared difference between the analysis and the observations (Gilchrist and Cressman, 1954). To achieve a good analysis, the spatial distance between the observations needed to be small in comparison to the size of the weather systems analysed, but in reality there are many data sparse areas. Therefore, it was suggested (Bergthórsson and Döös, 1955) that a background state is used as a first guess. The background state should be the best available approximation to the present state, before the use of the observational data. This could be, for example, a climatology or a forecast that is valid at the same time as the observations. This suggestion led to the iterative technique known as the Successive Corrections Method or Cressman Analysis, where the background state was modified by the observations to produce an analysis. To ensure that the analysis was smooth, the information from an observation was also used to update the surrounding grid points.

The Cressman analysis is a weighted average of the background state and the observations; however there is no direct way to specify the optimal weights. The specification of the optimal weights is important so that, for example, a good quality background state is not deteriorated by poor quality observations, and so that the information from the observations is optimally spread to the surrounding grid points. Statistical techniques are necessary to determine the expression for the optimal analysis. Such optimal estimation forms the basis of most data assimilation algorithms that are commonly used at the present time. These data assimilation algorithms are summarized in Table 1.2. Sequential algorithms such as Optimal Interpolation and the

	Use observations at the same time (Simple)	Use a time sequence of observations with the model (Advanced)
Sequential	Optimal Interpolation	Kalman Filter
Variational	3D-Var	4D-Var

Table 1.1: *A comparison of statistically optimal data assimilation algorithms, based on Ghil and Malanotte-Rizzoli (1991).*

Kalman Filter use the optimal estimation equations to compute the analysis explicitly, whilst the variational algorithms such as 3D-Var and 4D-Var compute the analysis by minimizing a cost function. The simple algorithms only use observations taken at one time level, whilst the more advanced algorithms use an entire time-sequence of observations through the use of the model dynamics. These algorithms are now discussed in detail. Only linear observation operators and models are considered, although the algorithms can be extended to consider nonlinear models. The notation in this section and in the rest of the thesis follows that advised in Ide et al. (1997).

1.2.1 3D-Var and Optimal Interpolation

Suppose that the true state of the atmosphere is represented by a vector $\mathbf{x}^t$ of dimension n, and that the first guess or background state is given by $\mathbf{x}^b$. Suppose that m observations are given in a vector $\mathbf{y}$, and are related to the true state variables through the linear observation operator $\mathbf{H}$. The background state and observations have errors $\boldsymbol{\varepsilon}^b$ and $\boldsymbol{\varepsilon}^o$, such that

$$\mathbf{x}^b = \mathbf{x}^t + \boldsymbol{\varepsilon}^b \tag{1.1}$$

$$\mathbf{y} = \mathbf{H}\mathbf{x}^t + \boldsymbol{\varepsilon}^o. \tag{1.2}$$

The errors are assumed to be unbiased ($E(\boldsymbol{\varepsilon}^b) = E(\boldsymbol{\varepsilon}^o) = 0$), where $E(x)$ denotes the expectation of x, and also to have covariances $\mathbf{B} = E(\boldsymbol{\varepsilon}^b \boldsymbol{\varepsilon}^{b^T})$, $\mathbf{R} = E(\boldsymbol{\varepsilon}^o \boldsymbol{\varepsilon}^{o^T})$.

The aim of the data assimilation algorithm is to combine the background state $\mathbf{x}^b$ and the observations $\mathbf{y}$ such that the analysis $\mathbf{x}^a$ is as 'close' to the true state as possible. In current data assimilation methods, the analysis is defined as the maximum likelihood estimate or as the minimum variance estimate. In fact both the maximum likelihood and minimum variance estimates result in the same analyses provided that the observations and background probability distribution functions (pdfs) are Gaussian. This is discussed by Lorenc (1986) and Bouttier and Courtier (2003).

The maximum likelihood estimate (or more precisely the maximum a priori estimate) approach is based on Bayesian statistics. By assuming that the background state and observations are independent, then from Bayes' theorem, it can be shown that the analysis pdf $P_a(\mathbf{x})$ can be written as:

$$P_a(\mathbf{x}) \propto P_b(\mathbf{x})P_o(\mathbf{x}) \tag{1.3}$$

where $P_b(\mathbf{x})$ and $P_o(\mathbf{x})$ are the background and observation pdfs. The maximum likelihood estimate is then given by the state which maximizes the analysis probability P_a. This can be simplified by assuming that the pdfs are Gaussian. In particular, by letting:

$$P_b(\mathbf{x}) = c_1 \exp\left[\frac{1}{2}(\mathbf{x}-\mathbf{x}^b)^T\mathbf{B}^{-1}(\mathbf{x}-\mathbf{x}^b)\right] \tag{1.4}$$

$$P_o(\mathbf{x}) = c_2 \exp\left[\frac{1}{2}(\mathbf{y}-\mathbf{H}\mathbf{x})^T\mathbf{R}^{-1}(\mathbf{y}-\mathbf{H}\mathbf{x})\right] \tag{1.5}$$

and taking the log of (1.3) then the maximum likelihood estimate $\mathbf{x}^a$ is given by the state which minimizes the cost function:

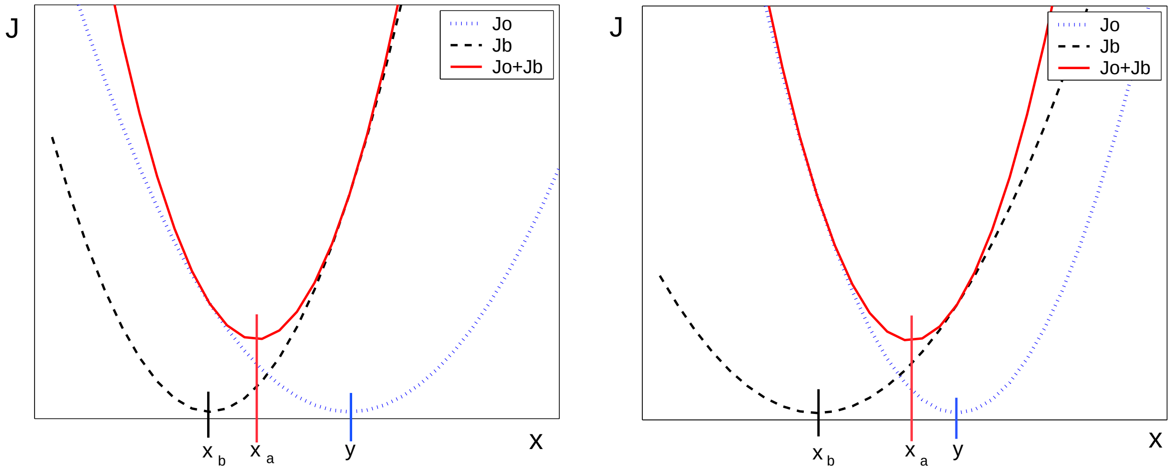
$$J(\mathbf{x}) = \frac{1}{2}\left\{(\mathbf{x}-\mathbf{x}^b)^T\mathbf{B}^{-1}(\mathbf{x}-\mathbf{x}^b) + (\mathbf{y}-\mathbf{H}\mathbf{x})^T\mathbf{R}^{-1}(\mathbf{y}-\mathbf{H}\mathbf{x})\right\}. \tag{1.6}$$

The minimum variance estimate, also known as the Best Linear Unbiased Estimate (BLUE) or the Gauss-Markov Theorem, assumes that the analysis is of the form of a linear combination of the background state and observations and uses the weights that minimize the trace of the analysis error covariance matrix. This gives the equations:

$$\mathbf{x}^a = \mathbf{x}^b + \mathbf{K}(\mathbf{y}-\mathbf{H}\mathbf{x}^b) \tag{1.7}$$

$$\mathbf{K} = \mathbf{B}\mathbf{H}^T(\mathbf{H}\mathbf{B}\mathbf{H}^T+\mathbf{R})^{-1} = (\mathbf{B}^{-1}+\mathbf{H}^T\mathbf{R}^{-1}\mathbf{H})^{-1}\mathbf{H}^T\mathbf{R}^{-1}. \tag{1.8}$$

By setting the gradient of the cost function to zero, it can be shown that the analysis that is found by minimizing the cost function (1.6) is the same as that found by solving the BLUE equations (1.7 and 1.8) directly. The weight matrix $\mathbf{K}$ specifies the optimal weights that were not specified in the Cressman method. The analysis is given by a weighted average of the background state and the observations. If the background errors are small compared to the



(a) $\mathbf{x}^b$ has small errors in comparison to $\mathbf{y}$

(b) $\mathbf{x}^b$ has large errors in comparison to $\mathbf{y}$

Figure 1.1: *Schematic diagrams of the J^o and J^b parts of the 3D-Var cost function for a state vector with only one variable. $\mathbf{x}^b$ is the background state, $\mathbf{y}$ is the observation, and $\mathbf{x}^a$ is the analysis, which is given by the minimum of $J^b + J^o$.*

observations, then the analysis would be close to the background state. If the background state errors are large compared to the observations, the analysis would be close to the observations.

The method known as Optimal Interpolation (OI) is a sequential method that finds the analysis by using the BLUE equations (1.7 and 1.8) directly. For the global data assimilation problem, the covariance matrices are very large, so are difficult to store and invert. Therefore, the optimal interpolation scheme is implemented on sub-domains of the globe (Lorenc, 1981). Further, the equations can only deal with linear observation operators. This means that satellite radiances can not be directly incorporated as the radiative transfer models are nonlinear.

The method known as Three-Dimensional Variational Data Assimilation (3D-Var) finds the analysis by minimizing the cost function (1.6) directly. In practice, the covariance matrices do not need to be inverted, so it is possible to solve the problem globally. It is also possible to extend the cost function to use nonlinear observation operators so that satellite data can be included directly. The minimization of the cost function is also a weighted, tapered, least squares method, provided that the weight matrices are given by the inverse error covariances Wunsch (1996). The analysis is given by the state which minimizes the noise vectors in the L_2 norm. The vector which the cost function is minimized with respect to is known as the control vector. The control vector may not be the same as the state vector if preconditioning is applied.

The 3D-Var cost function (1.6) has two terms. The first term is known as the J^b or background term and the second is known as the J^o or observation term. These two terms are

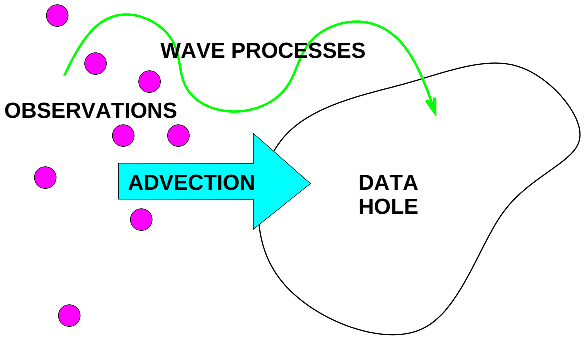


Figure 1.2: *Schematic diagram of the propagation of information into a data hole, based on Fig. 1 from Thompson (1961).*

illustrated in Fig. 1.1. If it is assumed that the background state has small errors (Fig. 1.1(a)) then the J^b term is narrow in comparison to the J^o term, so that the minimum of $J^b + J^o$ is close to the background state. However, if it is assumed that the background state has relatively large errors (Fig. 1.1(b)), then the minimum of $J^b + J^o$ is close to the observations.

1.2.2 4D-Var and the Kalman Filter

There many regions in the atmosphere that are data sparse. For example, there is little in-situ data over the oceans and also at upper levels. However, it is possible to combine a time sequence of observations with a numerical forecast model to reconstruct the atmospheric state in these data holes. This method was first proposed by Thompson (1961), and a schematic diagram based on this is shown in Fig. 1.2. At the first time level, observations are taken, and there are no observations in the data hole. These observations are then used in the initial conditions for the model, and the model is integrated. The information is then propagated into the data hole by advection or wave processes. Observations are then taken again, so that the entire state at that time has been observed. Similarly, the data hole could be positioned so that the state over the data hole is first advected and then observed.

Sasaki (1970) developed a variational method to combine the information from a time-sequence of observations with a numerical model. The numerical model was added as a constraint to the minimization, through the use of Lagrange multipliers. The control variables for the minimization were given by the model state at every time level. Such an algorithm is too large to be solved operationally, but can be simplified using the ‘reduction of the control variable’, as introduced by Le Dimet and Talagrand (1986). By formulating the problem as an optimal control problem, it is possible to use only the initial conditions as the control variables. The resulting minimization problem can be efficiently solved by using an adjoint model (Errico, 1997) to calculate the gradient of the cost function. The method that was developed is now known as 4D-Var as the time-sequence of observations provide a fourth dimension.

Full NWP models are nonlinear and so it is possible for the minimum of the cost function to be non-unique. Courtier et al. (1994) extended 4D-Var to give incremental 4D-Var, where the NWP model is linearized and a series of linear assimilation problems are solved instead. The cost may be further reduced by running the linear models at a lower resolution or with simplified physics. The incremental 4D-Var was first applied to an operational model by Rabier et al. (1998). In this thesis, only linear models are considered, and so we define the cost function for this problem only.

In general, the cost function J contains a background term J^b and an observational term J^o,

$$J(\mathbf{x}_0) = J^b + J^o \tag{1.9}$$

where $\mathbf{x}_0$ denotes the model state at time t_0 (the control-variable). The background term is the same as for 3D-Var:

$$J^b(\mathbf{x}_0) = \frac{1}{2}(\mathbf{x}_0 - \mathbf{x}^b)^T \mathbf{B}^{-1}(\mathbf{x}_0 - \mathbf{x}^b) \tag{1.10}$$

where $\mathbf{x}^b$ is the background state at the initial time, and $\mathbf{B}$ is the background error covariance matrix. The observational term is

$$J^o(\mathbf{x}_0) = \frac{1}{2}\sum_{i=1}^{N}(\mathbf{H}_i\mathbf{x}_i - \mathbf{y}_i)^T \mathbf{R}_i^{-1}(\mathbf{H}_i\mathbf{x}_i - \mathbf{y}_i) \tag{1.11}$$

where $\mathbf{y}_i$ is a vector of observations at time t_i, $\mathbf{H}_i$ is the observation operator, which converts from model space to observations space and $\mathbf{R}_i$ is the observational error covariance matrix.

The 4D-Var problem is then to minimize $J(\mathbf{x}_0)$ subject to the strong constraint that the sequence of model states must also be a solution of the model equations $\mathbf{x}_{i+1} = \mathbf{M}\mathbf{x}_i$. A strong constraint means that the model is assumed to be perfect. However it is possible to formulate the problem with a weak constraint so that the perfect model assumption can be relaxed. The 4D-Var method is illustrated in Figure 1.3. This is a constrained optimization problem. However, by using an adjoint model, the problem can be transformed into an unconstrained minimization through the use of Lagrange multipliers. This will be described in detail in Chapter 2.

It was demonstrated that the analysis for 3D-Var can also be written as a solution of the

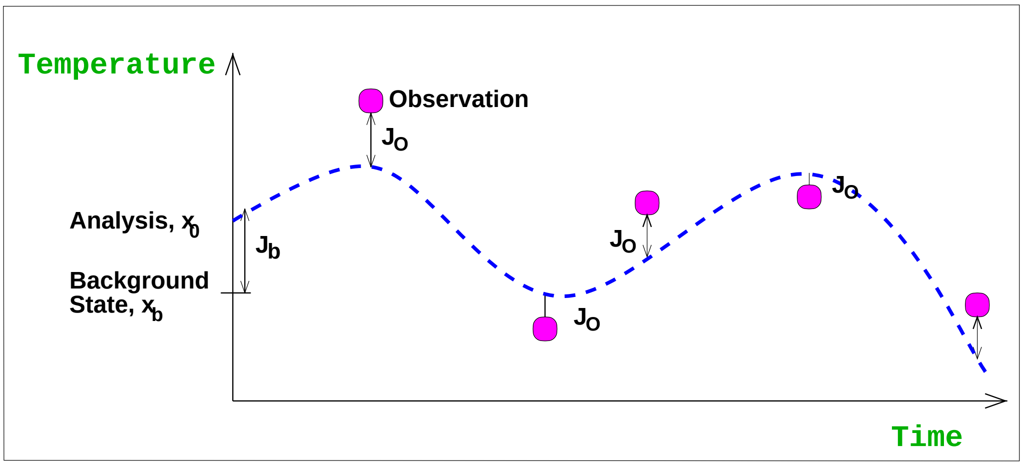


Figure 1.3: *Schematic diagram of the Four Dimensional Variational data assimilation method: minimize the squared distance between the analysis and the background state at the beginning of the assimilation window, and the squared distance between the observations and the forecast state throughout the assimilation window.*

BLUE equations that are used in optimal interpolation. Similarly, the analysis for 4D-Var can be written explicitly in the form of a sequential algorithm known as the Kalman Filter. This algorithm will be discussed below. The equivalence between 4D-Var and the Kalman Filter gives a key result that the analysis at the end of a 4D-Var assimilation window is the same as that obtained by the Kalman Filter, if the same background error covariance matrix is specified at the beginning of the window, and the models are perfect and linear. This can be proved by showing that 4D-Var and the Kalman Filter both solve the Riccati equation, as discussed by Jazwinski (1970), Ghil and Malanotte-Rizzoli (1991) and Wunsch (1996), but is perhaps more easily proved by considering a sequence of 3D-Var analyses and propagating the background error covariance matrix using the Kalman Filter (e.g. Lorenc, 1986, Li and Navon, 2001).

The Kalman Filter (Kalman, 1960) is a sequential assimilation algorithm, like the BLUE equations, but the background error covariance matrix is propagated explicitly in time. The Kalman Filter can be viewed in two stages: forecast and analysis. We denote the state covariance matrix at time t_i as $\mathbf{P}_i$, so that the error covariance at the beginning of the 4D-Var window is $\mathbf{P}_0 = \mathbf{B}$.

The Forecast step is:

$$\begin{aligned} \mathbf{x}_i^f &= \mathbf{M}\mathbf{x}_{i-1}^a \\ \mathbf{P}_i^f &= \mathbf{M}\mathbf{P}_{i-1}^a\mathbf{M}^T. \end{aligned} \tag{1.12}$$

The Analysis step is:

$$
\begin{aligned}
&\mathbf{K}_i = \mathbf{P}_i^f \mathbf{H}_i^T [\mathbf{H}_i \mathbf{P}_i^f \mathbf{H}_i^T + \mathbf{R}_i]^{-1} = [\mathbf{P}_i^{f^{-1}} + \mathbf{H}_i^T \mathbf{R}_i^{-1} \mathbf{H}_i]^{-1} \mathbf{H}_i^T \mathbf{R}_i^{-1} \\
&\mathbf{x}_i^a = \mathbf{x}_i^f + \mathbf{K}_i [\mathbf{y}_i - \mathbf{H}_i \mathbf{x}_i^f] \qquad (1.13) \\
&\mathbf{P}_i^a = [\mathbf{I} - \mathbf{K}_i \mathbf{H}_i] \mathbf{P}_i^f .
\end{aligned}
$$

$\mathbf{x}_i^a$ gives the analysis at time t_i, with associated error covariance $\mathbf{P}_i^a$. It is not possible to use the full Kalman Filter for the assimilation of the atmosphere as the size of the state vector is too large. For example, the background error covariance matrix would be $(10^7 \times 10^7)$ and therefore inverting and storing such a matrix is not possible both now and in the future. As both computer power and memory increase in the future, it is likely that the state vector will become even larger and so the covariance matrix will also become larger.

The equivalence between 4D-Var and the Kalman Filter allows an understanding of the 4D-Var algorithm. The Kalman Filter propagates the background error covariance matrix explicitly through the assimilation window, and therefore the 4D-Var algorithm also implicitly propagates the covariance matrix.

1.3 Baroclinic Instability

This thesis is concerned with the ability of 4D-Var to generate the correct analysis in regions of baroclinic instability. We therefore now describe baroclinic instability, before discussing the previous literature concerning 4D-Var and baroclinic instability.

Baroclinic instability plays a key role in the development of mid-latitude cyclones that are seen in the atmosphere. It is an instability associated with a zonal wind shear with height, which through thermal wind balance, depends on the meridional temperature gradient. The instabilities grow by converting the available potential energy in the temperature gradient of the basic state, into eddy potential and eddy kinetic energy and are an important part of the global energy cycle (Holton, 1992). There are two approaches to examining baroclinic instability. The first approach is to consider an eigenvalue problem. The eigenvectors of the model, also known as normal modes, grow exponentially without changing their spatial structure. According to linear theory, the eigenvectors with the largest eigenvalues eventually become the dominant structures in a forecast from random initial conditions. The second approach is to consider an initial-value problem. That is, to consider the non-modal (or not normal mode) growth. It is possible for some initial perturbations to grow for limited periods at faster rates than normal modes, and these perturbations change their spatial structure with time. Singular vectors of the model give such rapid linear growth over a finite time. Both the eigenvalue and initial-value approaches examine the linear growth of perturbations to a basic state, and both are highly relevant for interpreting the growth of cyclones in the atmosphere. In the rest of this thesis, we refer to normal modes of the model as modal, and any other structure as being non-modal.

1.3.1 Quasi-Geostrophic Potential Vorticity (QGPV)

Before the mechanisms for modal growth and non-modal growth are described further, we first describe a quantity known as quasi-geostrophic potential vorticity (QGPV). It is important to understand this quantity, as a model based on QGPV, known as the Eady model, is used in this thesis.

QGPV, q, also known as pseudo potential vorticity (Hoskins et al., 1985, Hoskins, 1997) may be defined in terms of the geostrophic streamfunction ψ as:

$$q = f + \nabla_h^2\psi + \frac{\partial}{\partial z}\left(\frac{f_o^2}{N^2}\frac{\partial\psi}{\partial z}\right) \tag{1.14}$$

where $f = f_0 + \beta(y)$ is the Coriolis parameter, $N^2 = \frac{g}{\theta}\frac{d\theta_o}{dz}$ is the Static Stability, where θ_o is the potential temperature of a hydrostatically balanced reference state such that $\theta = \theta_o(z) + \theta'$, and $\nabla_h^2 = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}$. Equation (1.14) is an elliptic equation and is central to what is known as the ‘Action at a Distance’ principle (Davies and Bishop, 1994). Given the QGPV and suitable boundary conditions, it is possible to infer the corresponding streamfunction field, which can be considered as a smoothed version of the QGPV. A change in the QGPV at a single point would not only give a significant change in the corresponding streamfunction at the same point but would also give a significant change in the streamfunction field in a surrounding region. This surrounding region is stretched in the vertical due to the coefficient $f_o^2/N^2 \sim 10^{-4}$.

QGPV can be considered as a dynamical tracer as it is conserved following adiabatic, geostrophic flow and is important as it combines both dynamical and thermodynamical information. This can be illustrated by considering important balance relationships in the atmosphere. Geostrophic balance states that the geostrophic wind u_g, v_g and geostrophic relative

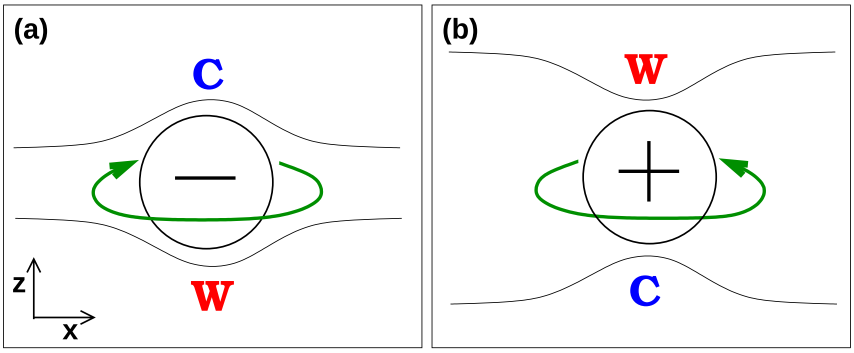


Figure 1.4: *Schematic diagrams of potential vorticity anomalies and their associated circulations (green arrows) and temperature anomalies (W and C). (a) a negative PV anomaly and (b) a positive PV anomaly. W and C represent Warm and Cold anomalies respectively. The thin lines represent isentropes (constant θ), such that the potential temperature increases with height. Based on Fig. 6 from Hoskins (1997).*

vorticity ξ_g are related to the streamfunction by:

$$(u_g, v_g) = \left(-\frac{\partial\psi}{\partial y}, \frac{\partial\psi}{\partial x}\right) \qquad\qquad \xi_g = \frac{\partial v_g}{\partial x} - \frac{\partial u_g}{\partial y} = \nabla_h^2\psi. \tag{1.15}$$

Hydrostatic balance states that the potential temperature θ' and buoyancy b are related to the streamfunction ψ by:

$$b = \frac{g}{\theta_o}\theta' = f_o\frac{\partial\psi}{\partial z}. \tag{1.16}$$

These relationships may be combined to give the Thermal Wind Balance relationship, which states that the vertical shear of the horizontal wind is related to the horizontal temperature gradient by:

$$\left(\frac{\partial u_g}{\partial z}, \frac{\partial v_g}{\partial z}\right) = \left(-f_o\frac{\partial b}{\partial y}, f_o\frac{\partial b}{\partial x}\right). \tag{1.17}$$

Using the geostrophic balance and hydrostatic balance relationships, the QGPV can also be written as:

$$q = f + \xi + f_o\frac{\partial\theta'/\partial z}{d\theta_o/dz}. \tag{1.18}$$

This equation illustrates that a positive QGPV anomaly is associated with a cyclonic circulation with a warm anomaly above and a cold anomaly below, and similarly a negative QGPV anomaly is associated with an anticyclonic circulation with a cold anomaly above and a warm anomaly below. These relationships are also illustrated by the diagrams in Fig. 1.4.

Boundary temperature anomalies can also be described in terms of QGPV. The boundary may be a lower boundary such as the ground or an upper boundary such as the tropopause. The

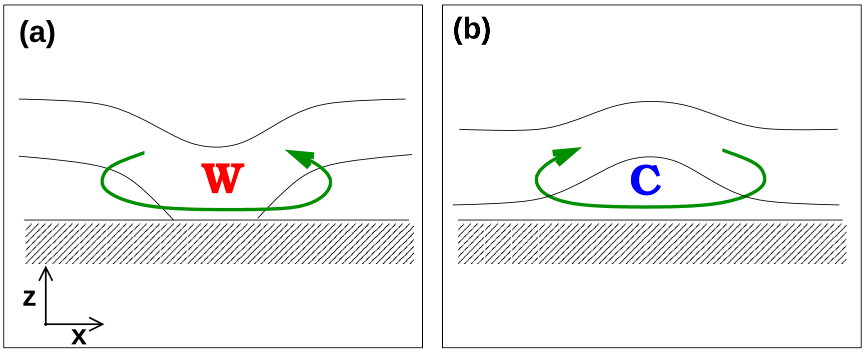


Figure 1.5: *Schematic Diagram of lower boundary temperature anomalies and their associated circulations. (a) a warm anomaly and (b) a cold anomaly. Based on Fig. 6 from Hoskins (1997).*

tropopause is the interface between the troposphere, with relatively low static stability, and the stratosphere, with relatively high static stability. The high static stability means that the tropopause may be modelled as a rigid boundary. Bretherton (1966) showed that a temperature distribution on a boundary is equivalent to a distribution of QGPV on a thin sheet just within the fluid. On the lower boundary, a warm anomaly is equivalent to a positive QGPV anomaly, and a cold anomaly is equivalent to a negative QGPV anomaly. Conversely, on the upper boundary, a warm anomaly is equivalent to a negative QGPV anomaly, and a cold anomaly is equivalent to a positive QGPV anomaly. Thus, boundary temperature anomalies have associated circulations. These are illustrated in Fig. 1.5. Consider a warm anomaly at ground. The amplitude of the anomaly decays upwards, and therefore, from thermal wind balance, the warm anomaly is associated with a cyclonic circulation. Similarly, a cold anomaly is associated with an anticyclonic circulation. Temperature anomalies at the tropopause decay downwards, and therefore they have circulations of the opposite sign. The change in sign between the boundaries is important for baroclinic instability in the Eady model. The Charney-Stern Instability Criterion (Charney and Stern, 1962, Holton, 1992) states that for baroclinic instability, there must be a change in the sign of the meridional QGPV gradient in the domain. This reversal in sign may be associated with a change in the meridional basic state QGPV gradient in the interior or with the boundary temperature gradients which are associated with boundary QGPV gradients.

1.3.2 Modal Growth

Charney (1947) and Eady (1949) formulated mathematical models for baroclinic instability. In this section, the baroclinic mechanism, based on the results by Charney and Eady, is described

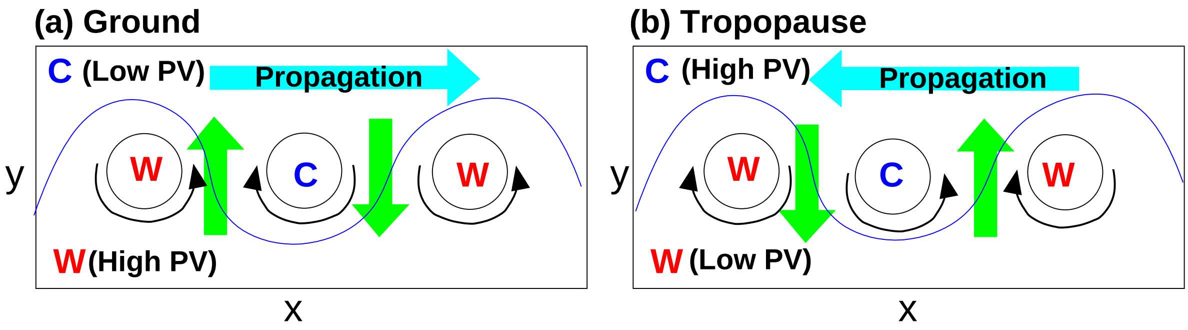


Figure 1.6: *Schematic diagrams illustrating the direction of propagation of Rossby waves at the tropopause and at the ground. Each diagram is an x-y cross section of the atmosphere at mid-latitudes. It is warm in the south and cold at the north. The circles represent temperature anomalies on the basic state meridional temperature distribution. At the tropopause, the warm anomaly (W) is associated with an anticyclonic circulation, and at the ground, the warm anomaly is associated with a cyclonic circulation. The meridional advection by the induced velocities means that the wave at the tropopause propagates westwards relative to the flow, and the wave at the ground propagates eastwards relative to the flow.*

non-mathematically. In the Charney formulation, the beta effect (y-dependence of the Coriolis parameter) and density decay with height were included and there was no rigid lid at the top. However, in the Eady formulation, beta was assumed to be zero, the density was uniform, and a lid was added to simulate the tropopause. Nevertheless, the results from the two different formulations provide similar results. This section begins by describing edge-wave propagation in the atmosphere and then describing the coupling between upper and lower edge-waves, which gives baroclinic instability.

Consider an atmosphere with a constant Coriolis parameter and with a meridional temperature gradient. This gradient is associated with a zonal wind shear with height, through thermal wind balance (1.17). The ground and the tropopause can be considered as horizontal boundaries at which the vertical velocity is very small. If it is assumed that there is no vertical motion at the boundaries, the basic state can support Rossby edge-waves, as illustrated by Fig. 1.6 and discussed in Gill (1982) for example. These waves have a maximum amplitude at the boundary, and then decay exponentially with the distance from the boundary.

The propagation of the Rossby edge-waves is now described. We first consider an edge-wave at the ground. The amplitude of the wave decays upwards, and therefore, from thermal wind balance, a warm anomaly is associated with a cyclonic circulation, whilst a cold anomaly is associated with an anticyclonic circulation as illustrated in Fig. 1.6 (a). The circulations act to move colder air to the east of the cold anomaly and warmer air to the east of the warm anomaly, so that the entire wave propagates eastwards, relative to the flow. In contrast, consider

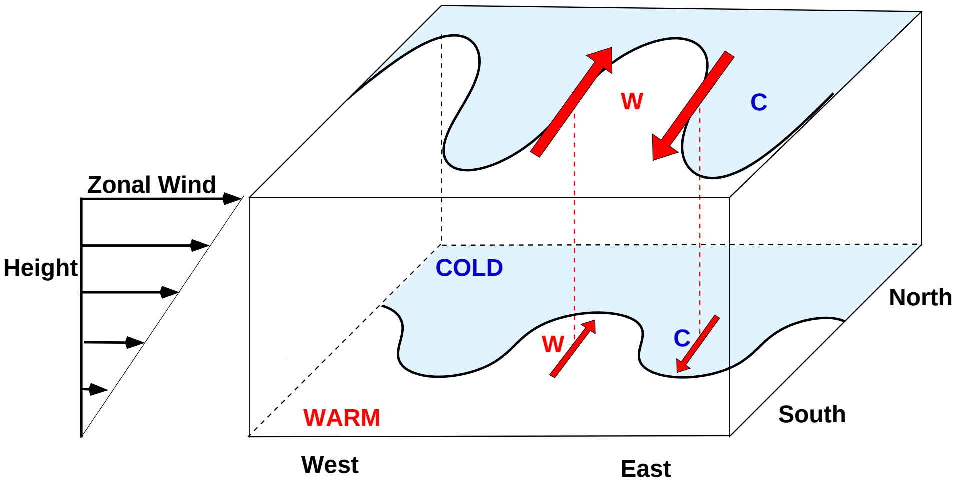


Figure 1.7: *Schematic Diagram of baroclinic instability. The meridional temperature gradient is associated with a zonal wind shear with height through thermal wind balance. There is a wave on the upper boundary (tropopause). The warm anomaly at the tropopause is associated with an anticyclonic circulation as marked by the red arrows. This circulation extends down to the ground and induces a wave on the lower boundary.*

an edge-wave at the tropopause, as illustrated in Fig. 1.6 (b). The amplitude of the wave decays downwards, and so from thermal wind balance, a warm anomaly is associated with an anticyclonic circulation and a cold anomaly is associated with a cyclonic circulation. These circulations act so that the entire wave propagates westwards, relative to the flow.

This edge-wave propagation is in fact the same as Rossby wave propagation, as the meridional temperature gradient acts in the same manner as a positive meridional PV gradient at the ground and a negative meridional PV gradient at the tropopause, as indicated also in Fig. 1.6. If a parcel of air moves from a region of high PV to a region of low PV, it must generate positive relative vorticity so that the PV of the air parcel is conserved. The circulations associated with the relative vorticity again act to move the wave so that the wave propagates westwards at the tropopause and eastwards at the ground. In the real atmosphere, the Coriolis parameter varies with latitude due to the curvature of the earth. This also affects the meridional PV gradient of the basic state, and hence also gives rise to Rossby wave behaviour.

We now consider how the upper and lower edge waves interact, as described by Davies and Bishop (1994). Consider the situation where there is an edge wave at the tropopause as shown in Fig. 1.7. The circulation associated with the warm temperature anomaly extends down to the ground. This circulation induces a wave on the lower boundary. The wave on the lower boundary also has an associated circulation (but with the opposite sign to that at the upper boundary), which then extends to the upper boundary and intensifies the upper level wave. Thus, the upper and lower level edge waves are coupled together via the meridional velocity field. This process is referred to as ‘self-development’.

From thermal wind balance, the meridional temperature gradient is also associated with a vertical shear of the zonal wind, so that the wave at the tropopause is advected eastwards faster than the wave at the ground. Since this is counter to the relative Rossby wave propagation speeds, it is possible for the edge waves to become phase-locked together, so that they move at the same speed. Thus, the two edge waves become an entire coupled wave that travels downstream. The edge waves interact with each other so that the amplitude grows in time whilst the spatial structure or shape is preserved. This is normal mode or eigenvector growth. If a model is integrated from random conditions, the normal modes with large eigenvalues will dominate the structure of the solution. The self-development mechanism relies on the fact that temperature anomalies on the upper and lower boundaries are associated with circulations of the opposite sign. This is necessary so that the Rossby waves propagate in the opposite directions, and so that they can become phase-locked together.

The spatial structure of the normal mode is vital for the growth or decay of the mode. For the growing mode, as shown in Fig. 1.8 (a), the pressure (streamfunction) field tilts westwards with height, so that the upper level ridge is close to being directly over the maximum meridional winds at the lower level. The effects of meridional advection and wave propagation mean that the maximum temperature anomalies lie just to the east of the surface low and just to the west of the surface high so that the temperature field tilts eastwards with height. For the decaying mode, as shown in Fig. 1.8 (b), the pressure field tilts eastwards with height, and the temperature field tilts westwards with height. Baroclinic growth is associated with the vertical coupling between upper and lower waves. However, this is not possible for waves at all wavelengths. Edge-waves with short wavelengths have smaller vertical scales and do not exert a large influence on the opposite boundary, so that baroclinic growth can not occur (Davies and Bishop, 1994). Thus, short waves are neutral (neither grow nor decay), whilst longer waves are baroclinically unstable. Using the Eady model, it can be shown that the wavelength of the most unstable mode is around 4000km (James, 1994). This is similar to the cyclones that are seen in real life, as discussed by Carlson (1994). Baroclinic instability is associated with a strong meridional temperature gradient, or baroclinicity, of the atmosphere. This strong gradient is found at mid-latitudes, and is more intense in the winter; hence the maximum in storm tracks are found at mid-latitudes during the winter season.

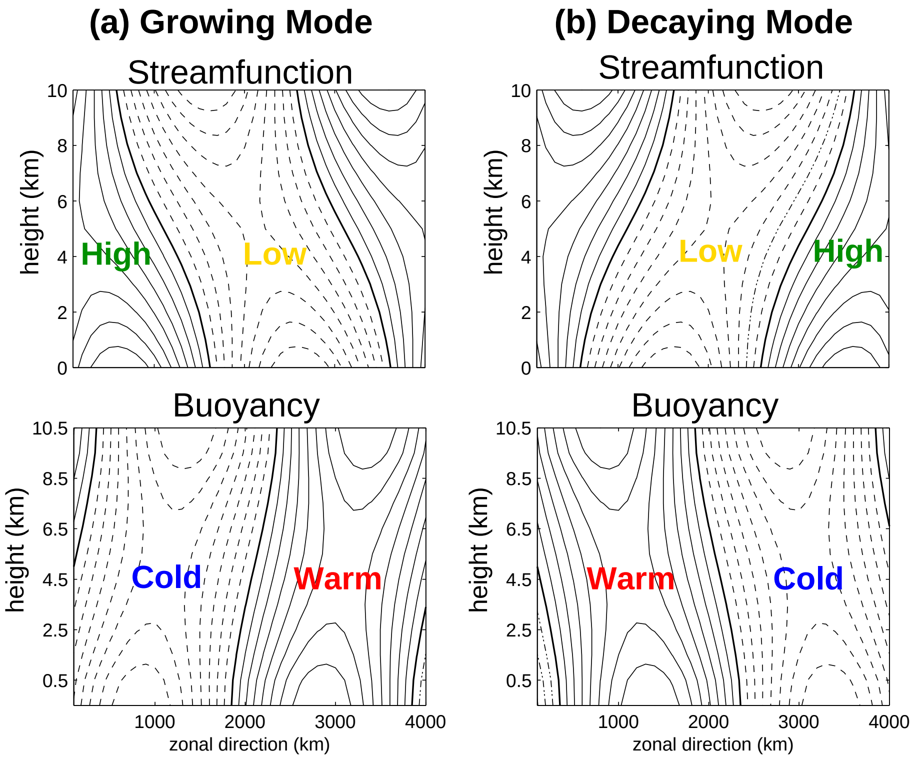
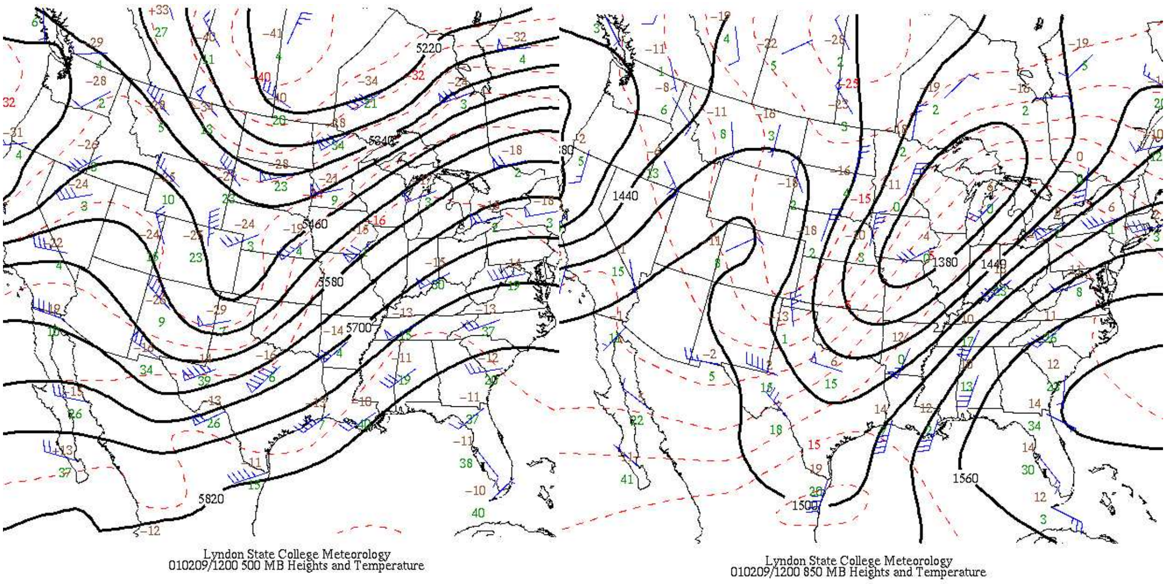


Figure 1.8: *The most rapidly (a) growing and (b) decaying Eady waves. The top panels show the streamfunction fields, with high and low regions marked, and the bottom panels show the buoyancy fields with the warm and cold regions marked.*

A real example of baroclinic growth in the atmosphere is shown in Fig. 1.9. This example was chosen over the United States as there is a large amount of data, compared with regions over the oceans. The low pressure system has developed in a region with a large temperature gradient. The low level (850 mb) trough is located to the east of the upper level (500 mb) trough, illustrating the westward tilt with height that is vital for the growth of the system.

1.3.3 Non-Modal Growth

The linear models of Charney and Eady are non-normal, which means that the discrete normal modes do not form a complete orthogonal basis. Pedlosky (1964) showed that a continuous spectrum of waves must also be included, so that any initial perturbation can be represented. In fact, the continuous spectrum involving delta functions in PV play an important role in the rapid development of perturbations. This has been shown by Farrell (1982, 1984), where the growth of perturbations is examined in the form of an initial value problem. Instead of analytically finding the eigenvectors, the equations were integrated using different initial conditions. This approach emphasized that it is possible for the growth rate of a perturbation to exceed the


Lyndon State College Meteorology
010209/1200 500 MB Heights and Temperature
Lyndon State College Meteorology
010209/1200 850 MB Heights and Temperature

(a) 500 mb

(b) 850 mb

Figure 1.9: *500mb and 850 mb level heights and temperature on the 12 Feb 2001 at 00Z. Plotted station data shows the wind, temperature and dew point temperature. The solid lines represent contours of the Geopotential Height (m), and the dashed lines represent the temperature (Celcius). Note that these contours have been determined using a data display package (McIDAS-X), and not a data assimilation algorithm. Taken from http://apollo.lsc.vsc.edu/*

exponential growth of the most unstable normal mode, over a limited period of time. Farrell (1989) extended this work, to calculate the 'optimal perturbations' which give the maximum linear growth in a finite time interval. Optimal perturbations were first calculated for a full primitive equation model by obtaining the dominant singular vectors (Buizza and Palmer, 1995, Buizza, 1997), and are now routinely calculated for use as the initial perturbations for the ECMWF ensemble prediction system (e.g. Buizza et al., 2000). In the rest of this thesis, we will use 'Optimal Perturbations' to mean the singular vectors of the model, to reduce confusion with other singular vectors that will be introduced later on.

The structures of the optimal perturbations at the initial time are characterized by the superposition of the interior PV delta functions, with a small vertical scale. For example, Fig. 1.10 shows the QGPV and buoyancy fields for a typical perturbation which gives rapid finite-time growth. Such a perturbation will be used in some of the experiments in this thesis, and the details of the calculations may be found in Section 6.3 and in Appendix A. The associated streamfunction and meridional wind fields are also shown. The initial QGPV anomalies (at T+0) are advected eastwards and westwards by the zonal shear flow. This 'unshields' the QGPV located near to the middle of the domain. The meridional winds associated with the

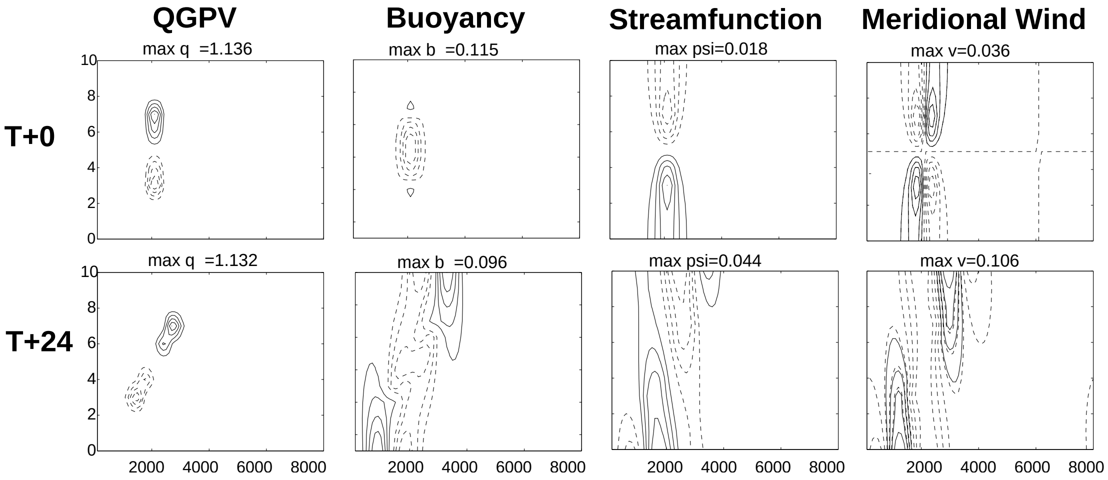


Figure 1.10: *The evolution of a perturbation which gives rapid finite-time non-modal growth. Each panel shows a z-x cross section, with the horizontal distance and height in km. The top panels show the initial perturbation and the bottom panels show the perturbation 24 hours later. The QGPV, q, Buoyancy, b, Streamfunction, ψ and Meridional Wind v fields are shown at both times. The vertical axis is the height (km) and the horizontal axis is the distance in the zonal direction (km). The basic state flow is such that the zonal wind is zero in the middle of the domain.*

interior PV produce boundary thermal anomalies, so that the perturbation evolves into a structure that resembles that of a growing normal mode. The streamfunction field has grown in amplitude, has a westward tilt with height, and new systems are emerging to the east at upper levels and to the west at low levels. This growth mechanism is also described by Badger and Hoskins (2001) and Morgan (2001), Morgan and Chien (2002).

If the true state has a small-scale interior structure, but this is not captured in the analysis, it is possible that the forecast error will grow rapidly. Such small scale errors may be due to a poor observational network or imperfect parameterization scheme (Beare et al., 2003). Sensitivity tests have shown (Rabier et al., 1996), that large forecast errors can be traced back to small analysis errors, with a strongly tilted structure. Rabier et al. (1996) state that the assimilation system must be able to “deal with structures that are both strongly tilted and small scale (in the horizontal and the vertical)”. The sensitivity tests used a linear assumption, but Beare et al. (2003) used a method termed as ’PV-sensitivity mapping’, to understand the effect of nonlinear processes. It was found that baroclinic regions near the steering level are particularly sensitive to localized PV perturbations. Targeted observations, such as dropsondes, may be added into these sensitive regions. Hence, it is important that the DA algorithm can use these extra observations to correct, or add localized sharply tilted vertical structures.

It is difficult to obtain reliable estimates of the true background state errors in an operational setting, as the true state is unknown. However, it is likely that these error structures are important as they grow fast and may trigger large forecast errors. We should therefore ensure that DA algorithms are able to capture these structures (Swanson et al., 2000, Hollingsworth, 2000).

In summary, the vertical structure of both modal and non-modal structures is very important for the growth and decay of cyclones, and hence it is vital that a DA algorithm analyses this structure correctly.

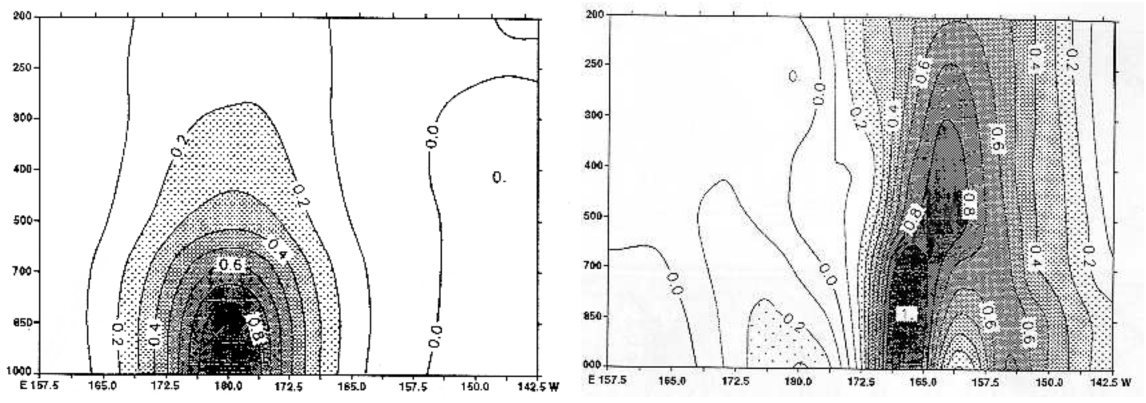
1.3.4 4D-Var in the Presence of Baroclinic Instability

We now review some of the previous studies that have shown 4D-Var to perform well in the presence of baroclinic instability.

The ability of 4D-Var to perform well in cases of baroclinic instability was first demonstrated by Courtier and Talagrand (1987). A total number of 5479 observations, during a 24 hour period, were assimilated by 4D-Var with no background term, using a spectral model of the vorticity equation with 231 degrees of freedom. The observations were mostly over the land, and in particular did not cover a region containing the Aleutian depression, yet the 4D-Var algorithm was able to combine the information from the observations with the dynamics, to reconstruct the depression. They noted that unrealistic noise was also generated in data poor areas, but that this could be reduced by adding a smoothing term to the cost function.

The reconstructive ability of 4D-Var was further demonstrated in experiments by Thépaut and Courtier (1991) where the mass field was observed and 4D-Var was used to reconstruct the vorticity field. Rabier and Courtier (1992) examined a simple baroclinic model where the small scales were observed, and the large scales were reconstructed. Tanguay et al. (1995) showed that if only the large scales are observed, the assimilation window length must be large enough so that information can be transferred to the small scales. Experiments by Thépaut et al. (1993b) where data was excluded over an area with a strong baroclinic development show that dynamics are able to infer the correct information in the unobserved regions. However, the analysed systems in the excluded data areas were slightly less intense.

To understand how 4D-Var uses information from observations, Thépaut et al. (1993a) showed that by assimilating a single observation, the analysis increment is proportional to a column of the Kalman Filter covariance matrix that is implicitly propagated in 4D-Var. The single observation experiments are therefore useful in understanding how 4D-Var spreads the



(a) 3D-Var

(b) 4D-Var

Figure 1.11: *Comparison of x-z cross sections of 3D-Var and 4D-Var structure functions for geopotential height, from single observation experiments. In both cases, the observation is placed at 1000 hPa. Note that East is on the left hand side and West is on the right. (Figures 3 and 12 from Thépaut et al. (1996)).*

information from an observation to the surrounding grid points. An example of these flow-dependent structure functions is shown in Thépaut et al. (1996). For clarity, Figs. 3 and 12 from Thépaut et al. (1996) are also shown here in Fig. 1.11. The 3D-Var structure function (background error correlation between the observation point and surrounding grid points) is isotropic with an equivalent barotropic structure, having no tilt with height. The equivalent 4D-Var structure function is anisotropic and exhibits a westward tilt with height with the upper levels showing more correlation with the lower level observation than for 3D-Var. Thépaut et al. (1996) also showed that it is important to have a long assimilation window (within the validity of the tangent linear model) to ensure that the dynamical structure functions are fully developed. Westward tilting structure functions and analysis increments can also be found in the ECMWF operational 4D-Var, as shown by Rabier et al. (1998, 2000). This westward tilt is vital for baroclinic growth

Thus, with observations at the end of the window, the analysis increments have a vertical structure that is required for baroclinic growth. The relationship between analysis increments and unstable modes has been shown by Rabier et al. (1996), where the gradient of a cost function is shown to be dominated by the most unstable components of the initial analysis error. Pires et al. (1996) showed theoretically that future observations in a 4D-Var algorithm provide accuracy of the unstable modes and experiments by Thépaut et al. (1996) demonstrated a strong link between singular vectors of the tangent linear model with 4D-Var analysis increments.

In summary, 4D-Var is able to combine information from observations with the model dynamics. This gives two main advantages in comparison to 3D-Var. First, 4D-Var is able to generate analysis increments with a westward tilt with height, as required for baroclinic growth. Second, 4D-Var is able to reconstruct or infer the state in unobserved regions. Although, these advantages have been demonstrated, there are still many questions concerning 4D-Var that remain. These are discussed in the following section.

1.4 Current Issues

We now consider some of the important issues concerning 4D-Var at the present time. In the future, there will be a greater emphasis on the use of satellite data and adaptive observations rather than in-situ data. We therefore need to assess how 4D-Var will use these observations, and whether we can expect to see a large improvement in forecast accuracy. There are also questions concerning the specification of the assumed errors. For example, there has been much research involved with the specification of the initial background error covariance matrix, $\mathbf{B}$. Ideally, the covariance should be flow-dependent and propagated from the previous assimilation window. However this is not possible and so approximations are needed. In the previous sections, we have presented 4D-Var using the model as a strong constraint. This assumes that the model is perfect, and although some methods have been suggested to account for the model error, it is still not clear how this should be dealt with. A related problem is to consider the nature of the background error. The background state is given by a forecast valid at the same time as the analysis. Therefore, the background state may contain phase errors and amplitude errors. It is therefore important to consider how 4D-Var will behave in such a case, and whether there are alternative methods that will deal with phase errors in a better way. These issues are now discussed further.

1.4.1 The Global Observing System and Adaptive Observations

The Global Observing System (GOS) (WMO, 2003) is a core component of the World Weather Watch (WWW) - the international meteorological observing system that is directed by the World Meteorological Organisation (WMO). GOS consists of the surface and radiosonde networks, and the aircraft and satellite systems that are operated by member countries of the WMO.

At the present time, there are about 13 000 land stations that provide surface observations every 3 hours. There are also about 1000 ships and 3000 moored or drifting buoys that provide surface observations over the ocean. It is important that observations are not only taken at the ground, but that the vertical profile of the atmosphere is observed. Therefore, there are approximately 600 upper-air stations that provide radiosonde observations twice a day. Aircraft data also provide an important addition at upper levels, mainly near the tropopause. Thus, the in-situ observations are mostly over the land and at the surface, giving many regions with sparse observations. For example, the southern hemisphere is particularly data sparse as it is mostly covered by the ocean. In these regions, remote-sensing instruments on satellites make a vital contribution. Observing system experiments by Andersson et al. (1991), Bouttier and Kelly (2001) showed that in the southern hemisphere satellite data does indeed have a large impact on forecast accuracy.

There are two main types of satellites: polar orbiters and geostationary satellites. Polar orbiters, or low earth orbiting satellites, orbit at an altitude of 600-1000km from pole to pole. The instruments, for example Advanced TIROS Operational Vertical Sounder (ATOVS), scan sideways to give bands or swathes of observations. The instruments on these satellites provide vertical profiles of temperature and humidity in cloud free areas. The vertical resolution of the instrument is determined by the number of channels, or wavelengths that are measured. Geostationary satellites, for example Meteosat, orbit around the equator and with the same rate of rotation as the earth so that the same part of the Earth is continuously monitored. These satellites are at a high altitude of about 36 000km, and therefore the instruments do not give such a fine resolution as the polar-orbiting satellites. The instruments on the geostationary satellites are often used to measure wind velocities in the tropics by tracking clouds and water vapour.

In the future, it is expected that there will be fewer, but more evenly distributed radiosonde and surface stations and satellite data is expected to take a greater role. The future polar orbiting satellites will carry instruments, for example the Infrared Atmospheric Sounding Interferometer (IASI), which measure the emitted radiation at a vast number of channels, giving a much increased vertical resolution (1-4km compared with the present 4-10km). The new geostationary satellites, for example Meteosat Second Generation (MSG), will also give an enhanced horizontal resolution (1 km) of derived winds at more vertical levels.

It is expected that the increased vertical resolution in satellite data will particularly benefit the analysis of regions of baroclinic instability where the vertical structure is important. It is important to assess whether these observations will be capable of identifying the correct vertical structures, and in particular, to understand how these observations are used in a 4D-Var algorithm.

In the future, adaptive or targeted observations should also provide a large contribution to GOS. These observations can be added to particular regions of the atmosphere which are thought to be important to be observed. For example, dropsondes may be released from aircraft into regions which are particularly sensitive to error growth or that need to be analysed accurately (Desroziers et al., 1999). Observations are expensive, and therefore there are restraints on the total number of observations. It is therefore important to place both the fixed observing system observations and the extra adaptive observations in the optimal positions in space and time so that the forecast errors are as small as possible. The optimal positions are non-trivial as they depend on the true state of the atmosphere, the data assimilation scheme, the forecast model and the definition of 'optimal' (Snyder, 1996, Berliner et al., 1999). Much research has focussed on choosing the optimal positions for targeted observations using the concept of sensitive regions. These regions can be identified using singular vector or sensitivity vector techniques. There has been little study on the question of where observations should be placed from the perspective of the data assimilation scheme.

1.4.2 Understanding 4D-Var

4D-Var is an expensive method in comparison to methods such as 3D-Var and FGAT (First-Guess at the Appropriate Time, described in Rabier et al. (1998), where the background state is evolved to the correct time). It is therefore important to understand the advantages of 4D-Var so that the cost can be justified and to maximize the benefits.

Many of the studies that have been used to understand 4D-Var have used single observation experiments. These have provided an understanding of the flow-dependent structure functions and the equivalence with the Kalman Filter. However, they have not provided insight into how the information from more than one observation interacts in a 4D-Var system. In particular, they cannot be used to understand how a time sequence of observations is used.

The 1D-Var equations have been used for many satellite retrieval studies. These are the same as 3D-Var, except that the state vector represents a column of the atmosphere and not the full three dimensions. Because the dimension of this problem is much smaller than the dimension of atmospheric data assimilation, mathematical techniques have been applied to give an understanding of how the information from the satellite observations are used in the 1D-

Var algorithm, for example Mateer (1965). These studies are currently useful for determining the optimal subset of satellite channels to be used in a retrieval (Rabier et al., 2002). An interesting question is whether some of the techniques that have been used in this context can be applied to 4D-Var with a view to understanding the 4D-Var algorithm further and to choosing an optimal set of observations. For example, Fisher (2003) has developed a method to evaluate the degrees of freedom for signal and entropy reduction in a 4D-Var system. These numbers give an indication of the amount of useful information contained in the observations when used in a 4D-Var system.

1.4.3 Background Error Covariance Specification

The background error covariance is generally considered to be one of the most important parts of a data assimilation algorithm as this matrix is responsible for the directions in which data is spread. For example, in a region of dense noisy observations, the background error correlations are needed to ensure that the analysis is smooth. In a region with only one observation, the background error correlations are needed to spread the information from the observations to the surrounding grid points. The background error covariance is also necessary to specify the correlations between different variables. For example, if only the pressure field is observed, and if geostrophic balance is incorporated into the covariance matrix, the algorithm will be able to infer the correct wind field.

The correlations may be specified from observation minus background statistics (Hollingsworth and Lönnberg, 1986), or from differences between forecasts and analyses verifying at the same time, for example the NMC method, (Parrish and Derber, 1992). These methods assume that the covariance can be separated into horizontal and vertical parts. It is possible to define the covariance so that the vertical correlations vary with the wave number (Rabier et al., 1998), but this still does not give flow-dependent covariances.

In theory it is possible to fully transfer information from the previous window by propagating the covariance matrix with the Kalman Filter (Li and Navon, 2001). Practically, it is not possible to evolve the whole matrix and approximations are required. The results by Thépaut et al. (1996) showed that there is a strong link between singular vectors and the structure functions, and therefore it has been suggested that it may be possible to approximate the covariance using singular vectors (e.g. Ehrendorfer and Tribbia, 1997). Algorithms known as ‘simplified’ or ‘Reduced Rank’ Kalman Filters (RRKF), (e.g. Fisher and Andersson, 2001, Beck, 2003) have been proposed as feasible methods to gain some of the benefits of the Kalman Filter. In such algorithms, only a subset of the covariance matrix (such as the parts corresponding to the optimal perturbations of the model) is propagated in time. Methods such as the geostrophic coordinate transform (Semple, 2001) have also been suggested to generate flow-dependent error structures. At the present time, it is not clear how the covariance matrix should be approximated and this area requires further research.

1.4.4 Model Error

In this chapter, both 4D-Var and the Kalman Filter have been presented with the model as a strong constraint. This assumes that the model is perfect, but this is clearly not the case. It is possible to add the model as a weak constraint, by adding extra terms to the control variable to give an extended or augmented assimilation. These extra terms could either represent the model forcing or the model parameters. By applying control theory or parameter estimation, it is then possible to use a minimization algorithm to find the optimal model variables (e.g. Griffith, 1997, Wergen, 1992, Zou et al., 1992b, Lu and Hsieh, 1997, Navon, 1997). However, the problem is that the number of control variables needs to be relatively small and so it is not possible to add a different model forcing at every time step. Also, it is not known what the covariance matrix for the model errors should be. Thus, how best to incorporate model error into 4D-Var remains an open research problem.

1.4.5 Phase Errors

The BLUE equations optimally blend together observations and a background state. However, this may not give a good analysis in the case where there is a sharp gradient such as a front. For example, Bennett (2002) describes the case of ocean temperatures near the Gulf stream. If the background state is in the wrong place, then when observations are added, the sharp front may become smeared out. For example, Lorenc (1981) considers a front that has been mispositioned in the background state. Instead of moving the front, as a human subjective analysis would, the optimal interpolation algorithm has smeared the front out into an extended region. This is because the error covariances were not representative of the correct error structures.

We therefore need to consider new data assimilation methods which may be able to blend together the information from observations and the background state in a better way, especially in cases where the background state contains displacement errors.

Hoffman et al. (1995) suggested a technique where the cost function is split into separate terms to account for displacement errors explicitly. A variational method is then used to find a displacement vector, which may be constant throughout the domain, or defined spectrally as in Hoffman and Grassotti (1996). A similar method has also been implemented by Brewster (2002a,b), where a displacement vector is found by dividing the grid into subvolumes and using a ‘brute-force search method’. The translation is achieved by adding a pseudo wind to the model equations. A second possible technique is to apply the Monge-Kantorovitch optimal mass transfer problem (Benamou et al., 2002). In this case, a velocity field is applied to the state so that it is rearranged to be closer to the observations, whilst the size of the velocity vector is constrained to be small. The application of the Monge-Kantorovitch problem to data assimilation has been briefly discussed by Douglas (2000). A third possible technique is to adjust the PV field so that it is more consistent with water vapour imagery. This could be achieved manually (e.g. Swarbrick, 2001, Rϕsting et al., 2001, Carroll, 1997), or through the use of digital image warping (Alexander et al., 1998), which uses tie points that are defined manually. This allows the model fields to be distorted in a way that preserves the dynamical balance.

1.5 Key Questions Addressed

Section 1.4 discussed some of the important issues that need to be researched. We now focus on a subset of questions that are addressed in this thesis. The overall focus of this thesis is:

To understand the extent to which 4D-Var can develop the structures needed for the growth and decay of baroclinic systems.

The key questions are:

1. *How are observations used in 4D-Var?*

   We aim to understand how 4D-Var uses the model dynamics to spread information from observations to surrounding grid points. This is important in assessing how much better 4D-Var is compared to 3D-Var. It is known that 4D-Var is able to link together information from a time sequence of observations with the model dynamics to reconstruct the state in unobserved regions, but this process is not well understood. A particular aspect of this is to investigate whether some of the information content techniques that are used in 1D-Var satellite retrievals can be extended to 4D-Var.

2. *Why has 4D-Var been shown to perform well in regions of baroclinic instability?*

   Many studies have shown 4D-Var to develop structures that tilt westwards with height, necessary for baroclinic growth, and that there is a strong link between the analysis increments and singular vectors. We wish to develop this further by considering situations in which modal growth or decay dominates, and situations where non-modal growth dominates. With the increase in satellite data, it is important to ask whether 4D-Var will be able to capture the correct aspects of the vertical structure needed for baroclinic growth and decay.

3. *How can the benefits of 4D-Var be maximized?*

   Given that 4D-Var does give some benefits in comparison to 3D-Var, it is important to understand how these benefits can be maximized. In particular, to understand how 4D-Var can be designed so that it always performs well in regions of baroclinic instability. This includes considering what the optimal observing system would be from the perspective of the data assimilation scheme. We aim to understand where observations should be placed so that the maximum amount of useful information can be extracted. In particular, we consider where the observations should be placed in the 4D-Var assimilation time window. Many studies have shown that the 4D-Var window should be as long as possible (within the validity of the tangent linear model), and that observations at the end of the window produce flow-dependent analysis increments. However, these studies have only considered single observations. We therefore consider the best positions for observations at more than one time level.

1.6 Thesis Outline

To answer these questions, we consider idealized 4D-Var experiments with the Eady model. This is the most simple model of baroclinic instability and therefore allows the important mechanisms to be isolated.

Chapter 2 describes the development of a 4D-Var algorithm using the Eady model. The solution of the 4D-Var minimization is derived using linear algebra and Lagrange multipliers, and then the Eady model and adjoint model are described. A suitable minimization algorithm is chosen by comparing four different algorithms, and the convergence criteria are defined. A simple background error correlation model based on Laplace smoothing is also developed.

Chapter 3 discusses the results from idealized 4D-Var identical twin experiments. The experiments consider the case where the true state is given by either normal mode growth or decay, the lower level wave is observed and 4D-Var is used to reconstruct the upper level wave. These experiments are used to examine the ability of 4D-Var to reconstruct the state in unobserved regions and also to infer the vertical structure needed for the growth or decay of the state. The impacts of the background state, observational noise, background error correlations and temporal position and weights of the observations are investigated.

Chapter 4 introduces the use of the singular value decomposition (SVD) to understand 4D-Var. We first describe how the SVD has previously been used to understand the information content of observations in satellite retrieval studies. This is then extended to consider the information content of observations in 4D-Var. We will show that the 4D-Var analysis increments can usefully be written as a linear combination of the right singular vectors (RSVs) of the observability matrix, and discuss the similarities between these RSVs and optimal perturbations. The chapter finishes by discussing the computation of the SVD of the observability matrix for the Eady model.

Chapter 5 discusses the SVD of the observability matrix that was implicitly used by the experiments in Chapter 3. This is used to give a new interpretation of the 4D-Var analyses and to give a further understanding of how observations are used in 4D-Var. Finally, it is shown that 4D-Var may be formulated as a method known as Tikhonov Regularization which is often used to solve discrete ill-posed problems.

Chapter 6 considers more realistic experiments. The impact of background error correlations, in both dense and sparse data regions, is considered from an SVD perspective by examining the SVD of the normalized observability matrix. The ability of 4D-Var to generate analysis increments for non-modal growth is then considered. The final experiments consider the information content of different observing systems such as two horizontal lines and vertical lines.

Chapter 7 concludes the work of this thesis. We return to the key questions that have been addressed, and discuss the extent to which they have been answered. All the experimental work in the thesis has dealt with highly idealized case studies. We therefore discuss the implication of this thesis to operational DA. The chapter finishes by discussing the future work which follows from this thesis.

Chapter 2

Development of a 4D-Var algorithm using the Eady model

The focus of this thesis is an understanding of the extent to which 4D-Var is able to develop the structures necessary for baroclinic growth and decay. This will be addressed using 4D-Var identical twin experiments with the Eady model - a simple model of baroclinic instability.

Identical twin experiments involve two stages. First, the numerical model is integrated to provide the ‘true’ atmospheric state. Then, synthetic observations of the true state are used by the data assimilation algorithm to find the analysis, from which the forecast is generated. In identical twin experiments, the model can be assumed to be perfect and known errors may be added to the background state and observations. Thus, the identical twin experiments isolate the behaviour of the data assimilation algorithm; this is not possible using real data and real models.

The 4D-Var algorithm requires a forward model to link the observations together. In this thesis, the 2D Eady model is used so that the behaviour of 4D-Var in the presence of baroclinic instability can be addressed. The Eady model is linear, although the quasi-geostrophic equations have not been linearized. Hence we only consider linear models in the derivation of the solution to the 4D-Var minimization. The minimum of the cost function is found using a minimization algorithm which uses values of both the cost function and its gradient. The gradient of the cost function with respect to the initial state is found using an adjoint model. In this chapter, the adjoint model is described and a minimization algorithm is selected. The background error covariance matrix plays an important role in data assimilation, so the chapter finishes by describing the development of a simple correlation model. We begin by deriving the equations used to minimize the constrained 4D-Var cost function. Further details of the work in this chapter may be found in Johnson et al. (2002).

2.1 The 4D-Var algorithm

The 4D-Var algorithm was introduced in the first chapter, and can be summarized by the following.

The 4D-Var analysis $\mathbf{x}^a$ *is given by the initial state* $\mathbf{x}_0$ *which minimizes the cost function:*

$$J(\mathbf{x}_0) = J^b + J^o \tag{2.1}$$

$$= \frac{1}{2}(\mathbf{x}_0 - \mathbf{x}^b)^T \mathbf{B}^{-1}(\mathbf{x}_0 - \mathbf{x}^b) + \frac{1}{2}\sum_{i=0}^{N}(\mathbf{y}_i - \mathbf{H}\mathbf{x}_i)^T \mathbf{R}_i^{-1}(\mathbf{y}_i - \mathbf{H}\mathbf{x}_i) \tag{2.2}$$

subject to the strong constraint that $\mathbf{x}_i$ *also satisfies the linear model equations* $\mathbf{x}_{i+1} = \mathbf{M}\mathbf{x}_i$*, given* $\mathbf{x}_0$*, where* $\mathbf{x}_i = \mathbf{x}(t_i)$ *is the state vector at time* t_i*,* $\mathbf{x}^b$ *is the background state, and* $\mathbf{y}_i = \mathbf{y}(t_i)$ *is the vector of observations at time* t_i *such that the observations are given in an assimilation window of time length* $[t_0, t_N]$*.* $\mathbf{H}$ *is the observation operator which converts from state space to observation space, and* $\mathbf{B}$ *and* $\mathbf{R}$ *are the background and observation error covariance matrices.*

This is a constrained minimization, however, it can be transformed to an unconstrained minimization. To find the minimum, the gradient of the cost function with respect to the initial state is required. The equations that are used to calculate the gradient can be derived by applying linear algebra to the discrete case, or perhaps more elegantly by applying the method of Lagrange to the continuous case. Both derivations are now applied to the observation term J^o. It is not necessary to consider the background term J^b at this stage as this can simply be added to the equations for the observation term. The 4D-Var algorithm will be used for the Eady model, which is a linear model. Therefore, only linear models are considered.

2.1.1 Derivation using Linear Algebra

The derivation using linear algebra is first given. This follows Bouttier and Courtier (2003) and Lagarde et al. (2001). The observation term can be written as:

$$J^o = \sum_{i=0}^{N} J_i^o \tag{2.3}$$

where the index i is the observation time and

$$\begin{aligned} J_i^o &= \tfrac{1}{2}(\mathbf{y}_i - \mathbf{H}\mathbf{x}_i)^T\mathbf{R}_i^{-1}(\mathbf{y}_i - \mathbf{H}\mathbf{x}_i) \\ &= \tfrac{1}{2}(\mathbf{y}_i - \mathbf{H}\mathbf{M}\ldots\mathbf{M}\mathbf{x}_0)^T\mathbf{R}_i^{-1}(\mathbf{y}_i - \mathbf{H}\mathbf{M}\ldots\mathbf{M}\mathbf{x}_0). \end{aligned} \tag{2.4}$$

Then the gradient of J_i^o with respect to the initial conditions $\mathbf{x}_0$ is given by:

$$\begin{aligned} \nabla_{\mathbf{x_0}} J_i^o &= -(\mathbf{H}\mathbf{M}\ldots\mathbf{M})^T\mathbf{R}_i^{-1}(\mathbf{y}_i - \mathbf{H}\mathbf{M}\ldots\mathbf{M}\mathbf{x}_0) \\ &= -\mathbf{M}^T\ldots\mathbf{M}^T\mathbf{H}^T\mathbf{R}_i^{-1}\mathbf{d}_i \end{aligned} \tag{2.5}$$

where $\mathbf{d}_i = (\mathbf{y}_i - \mathbf{H}\mathbf{x}_i)$ denotes the innovation vector at time t_i and where the adjoint model $\mathbf{M}^T$ satisfies:

$$< \mathbf{M}\mathbf{x}, \mathbf{y} > = < \mathbf{x}, \mathbf{M}^T\mathbf{y} > \tag{2.6}$$

where $<, >$ is an inner product. Hence the gradient of the observation term is given by:

$$\nabla_{\mathbf{x}_0} J^o = -\left\{\mathbf{H}^T\mathbf{R}_0^{-1}\mathbf{d}_0 + \mathbf{M}^T(\mathbf{H}^T\mathbf{R}_1^{-1}\mathbf{d}_1 + \mathbf{M}^T(\mathbf{H}^T\mathbf{R}_2^{-1}\mathbf{d}_2 + \ldots + \mathbf{M}^T\mathbf{H}^T\mathbf{R}_N^{-1}\mathbf{d}_N)\ldots)\right\}. \tag{2.7}$$

In 4D-Var, the initial data is used as the control variables, so the adjoint model is used to propagate the gradient vector backwards in time (see for example, Lewis and Derber (1985) and Errico (1997)). It is important to note that the adjoint model is in general not the same as the inverse model $\mathbf{M}^{-1}$ (i.e. running the forward model backwards in time).

2.1.2 Derivation using Lagrange Multipliers

The derivation using Lagrange multipliers and the calculus of variations (see for example, Gelfand and Fomin (1963) and Forray (1968)) is now illustrated by considering the continuous multivariable case. Similar derivations are also given by Le Dimet and Talagrand (1986), Griffith and Nichols (1994), Griffith (1997) and Wlasak (1997).

The continuous 4D-Var problem can be stated as:

Minimize the functional $\int_{t_0}^{t_N} F(\mathbf{x}, t)dt$*, defined over an assimilation window* $[t_0, t_N],\ t_N > t_0 > 0$*, subject to the (strong) model constraint* $\dot{\mathbf{x}} = \frac{\partial \mathbf{x}}{\partial t} = \mathbf{m}(\mathbf{x}, t)$*, where* $\mathbf{x}$ *is an n-dimensional state vector and the time* $t \in [t_0, t_N]$ *is a scalar.* F *and* $\mathbf{m}$ *are scalar and vector functions* *respectively. All variables are real and are assumed to be sufficiently smooth and continuous.*

Using the method of Lagrange, the Lagrangian functional $\mathcal{L}$ can be constructed as:

$$\mathcal{L} = \int_{t_0}^{t_N} \left\{ F(\mathbf{x}, t) + \boldsymbol{\lambda}^T \left(\dot{\mathbf{x}} - \mathbf{m}(\mathbf{x}, t) \right) \right\} dt \tag{2.8}$$

$$= \int_{t_0}^{t_N} G(\mathbf{x}, \dot{\mathbf{x}}, \boldsymbol{\lambda}, \dot{\boldsymbol{\lambda}}, t) dt \tag{2.9}$$

where $\boldsymbol{\lambda}$ is an n-dimensional vector of Lagrange multipliers and $G = F + \boldsymbol{\lambda}^T(\dot{\mathbf{x}} - \mathbf{m})$ is a scalar function. Using a Taylor series expansion and integration by parts, necessary conditions for the first variation of $\mathcal{L}$ to be zero, $\delta\mathcal{L} = \mathcal{L}(\mathbf{x} + \delta\mathbf{x}, \boldsymbol{\lambda} + \delta\boldsymbol{\lambda}, t) - \mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}, t) = 0$ are given by Eulers equations:

$$\nabla_{\boldsymbol{\lambda}} G - \frac{d}{dt} \nabla_{\dot{\boldsymbol{\lambda}}} G = 0 \tag{2.10}$$

$$\nabla_{\mathbf{x}} G - \frac{d}{dt} \nabla_{\dot{\mathbf{x}}} G = 0 \tag{2.11}$$

and the transversality condition:

$$\left[\delta\mathbf{x}^T \nabla_{\dot{\mathbf{x}}} G \right]_{t_0}^{t_N} = 0. \tag{2.12}$$

Equation (2.10) gives the model constraint, and equation (2.11) gives what is known as the adjoint equation:

$$-\dot{\lambda}_j = \left(\frac{\partial \mathbf{m}}{\partial x_j} \right)^T \boldsymbol{\lambda} - \frac{\partial F}{\partial x_j} \text{ for } j = 1, \ldots, n. \tag{2.13}$$

The transversality condition (2.12) gives the final conditions $\boldsymbol{\lambda}(t_N) = 0$ and also implies that $\nabla_{\mathbf{x}(t_0)} \mathcal{L} = -\boldsymbol{\lambda}(t_0)$. That is, the gradient of $\mathcal{L}$ with respect to the initial conditions is found from the adjoint variable also at the beginning of the window. This gradient can be used by a minimization algorithm to find the minimum.

This theory can be extended to the multivariate discrete case by constructing the Lagrangian functional:

$$\mathcal{L} = J^o + \sum_{i=0}^{N} \boldsymbol{\lambda}_{i+1}^T (\mathbf{x}_{i+1} - \mathbf{M}\mathbf{x}_i) \tag{2.14}$$

and deriving the adjoint equations (Griffith, 1997):

$$\begin{aligned} &\boldsymbol{\lambda}_{N+1} = 0 \\ &\boldsymbol{\lambda}_i = \mathbf{M}^T \boldsymbol{\lambda}_{i+1} - \nabla_{\mathbf{x}_i} J_i^o \;\; i = N, \ldots, 0. \end{aligned} \tag{2.15}$$

Then the gradient of J^o at the initial time is given by

$$\nabla_{\mathbf{x}_0} J^o = -\boldsymbol{\lambda}_0. \tag{2.16}$$

If M is the forward linear model, then M^T is the **adjoint model**, $\boldsymbol{\lambda}$ is the vector of **adjoint variables** and $\nabla_{\mathbf{x}_i} J_i^o = -\mathbf{H}^T\mathbf{R}_i^{-1}(\mathbf{y}_i - \mathbf{H}\mathbf{x}_i)$ is known as the **adjoint forcing**. Thus, the result using linear algebra (2.7) gives the same result as the derivation with Lagrange multipliers (2.16).

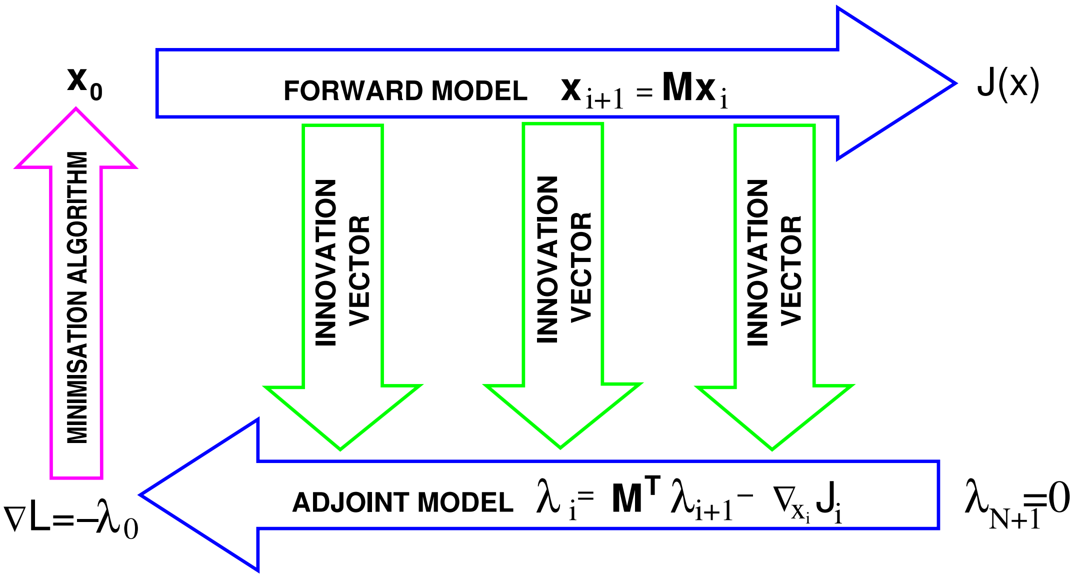


Figure 2.1: *A schematic diagram illustrating the calculation of the cost function J and the gradient of the cost function ∇J. These are both used by a descent (or minimization) algorithm to find the minimum.*

2.1.3 Summary

To summarize, the 4D-Var algorithm is solved using a minimization algorithm. On every iteration, the minimization algorithm computes the value of the cost function and the gradient of the cost function using the following steps:

1. Integrate the forward model for $i = 0, \ldots, N-1$, given the initial data $\mathbf{x}_0$ $$\mathbf{x}_{i+1} = \mathbf{M}\mathbf{x}_i$$
2. Calculate the innovation vectors for $i = 0, \ldots, N$ $$\mathbf{d}_i = \mathbf{y}_i - \mathbf{H}\mathbf{x}_i$$
3. Calculate the value of J^o $$J^o = \frac{1}{2}\sum_{i=0}^{N} \mathbf{d}_i^T \mathbf{R}_i^{-1} \mathbf{d}_i$$
4. Calculate and add the value of J^b $$J^b = \frac{1}{2}(\mathbf{x}_0 - \mathbf{x}^b)^T \mathbf{B}^{-1} (\mathbf{x}_0 - \mathbf{x}^b)$$
5. Set the adjoint variables at the final time to zero $$\boldsymbol{\lambda}_{N+1} = 0$$
6. Integrate the adjoint model backwards in time, for $i = N, \ldots, 0$, $$\boldsymbol{\lambda}_i = \mathbf{M}^T \boldsymbol{\lambda}_{i+1} - \nabla_{\mathbf{x}_i} J_i^o$$ using the adjoint forcings $\nabla_{\mathbf{x}_i} J_i^o$ $$\nabla_{\mathbf{x}_i} J_i^o = -\mathbf{H}^T \mathbf{R}_i^{-1} \mathbf{d}_i$$
7. The gradient is then given by the negative of the adjoint variables at the initial time $$\nabla_{\mathbf{x}_0} J^o = -\boldsymbol{\lambda}_0.$$
8. Calculate and add the gradient of J^b $$\nabla J^b = \mathbf{B}^{-1}(\mathbf{x}_0 - \mathbf{x}^b).$$

Some of these steps are also illustrated in the schematic diagram in Fig. 2.1.

2.2 The 2D Eady model

The 2D Eady model (Eady 1949) is a simple linear quasi-geostrophic (QG) model of baroclinic instability, and will form the basis for the experiments in this thesis. The qualitative mechanisms for baroclinic instability were described in Chapter 1. In this section, the non-dimensional equations for the Eady model are introduced. These are derived from the quasi-geostrophic equations, as given in Appendix A. The quasi-geostrophic equations are an approximation of the primitive equations for synoptic scales, which assume that the Rossby number is small and that the Burger number is unity.

The Eady model contains rigid surfaces at the ground and at the tropopause. The basic state is given by a zonal wind shear with height, that is associated with a uniform meridional temperature gradient. The density, static stability and Coriolis parameter are all taken to be constants. The Eady model equations describe the linear evolution of the perturbations to this basic state. Although the model is linear, the equations have not been linearized.

The non-dimensional Eady model equations are now described. The domain is infinite in the North-South direction y, periodic in the West-East direction x, and is between $z = -\frac{1}{2}$ and $\frac{1}{2}$. The initial state is given by the interior quasi-geostrophic potential vorticity (QGPV) perturbation, q and the perturbation buoyancy on the boundaries, b,

$$q(x, z, 0) = q_0(x, z) \qquad \text{in } z\epsilon \left[-\frac{1}{2}, \frac{1}{2}\right], \ x\epsilon[0, X] \tag{2.17}$$

$$b(x, z, 0) = b_0(x, z) \qquad \text{on } z = \pm\frac{1}{2}, \ x\epsilon[0, X]. \tag{2.18}$$

The perturbation QGPV is defined ((1.14) and (A.35) in Appendix A) as:

$$q = \frac{\partial^2\psi}{\partial x^2} + \frac{\partial^2\psi}{\partial z^2} \qquad \text{in } z\epsilon \left[-\frac{1}{2}, \frac{1}{2}\right], \ x\epsilon[0, X], \tag{2.19}$$

where ψ is the perturbation geostrophic streamfunction. The boundary conditions are periodic in the horizontal:

$$\psi(0, z, t) = \psi(X, z, t) \qquad \text{in } z\epsilon \left[-\frac{1}{2}, \frac{1}{2}\right], \ x\epsilon[0, X] \tag{2.20}$$

and, through Hydrostatic balance (1.16), the buoyancy field on the upper and lower boundaries provides the vertical boundary conditions:

$$\frac{\partial\psi}{\partial z} = b \qquad \text{on } z = \pm\frac{1}{2}, \ x\epsilon[0, X]. \tag{2.21}$$

Due to the periodic boundary conditions, an extra equation is needed to ensure that the problem for calculating ψ is well posed. Thus, the mean of the streamfunction field is arbitrarily set to zero:

$$\iint \psi dxdz = 0 \qquad \text{in } z\epsilon \left[-\frac{1}{2}, \frac{1}{2}\right], \ x\epsilon[0, X]. \tag{2.22}$$

Perturbations to the basic state are advected zonally by the basic state flow. The QGPV conservation equation ((A.34) in Appendix A),

$$\left(\frac{\partial}{\partial t} + z\frac{\partial}{\partial x}\right) q = 0 \qquad\qquad \text{in } z\epsilon \left[-\frac{1}{2}, \frac{1}{2}\right],\ x\epsilon[0, X] \qquad (2.23)$$

is derived from the QG thermodynamic equation and the QG vorticity equation. This states that QGPV is conserved following the horizontal, geostrophic, adiabatic, frictionless flow (Hoskins, 1997).

The QG thermodynamic equation reduces to:

$$\left(\frac{\partial}{\partial t} + z\frac{\partial}{\partial x}\right) \frac{\partial\psi}{\partial z} = \frac{\partial\psi}{\partial x} \qquad\qquad \text{on } z = \pm\frac{1}{2},\ x\epsilon[0, X] \qquad (2.24)$$

by assuming that there is no vertical motion at the boundaries ((A.33) in Appendix A). This equation describes the evolution of the Rossby-edge waves on the upper and lower boundaries. Notice that it is the meridional wind that provides the crucial coupling between the upper and lower waves, although this is still a 2D model.

The Eady model is discretized using 11 vertical levels for streamfunction and QGPV, with the upper and lower level buoyancy defined on levels 1 and 11. There are 40 grid points in the horizontal, giving 520 degrees of freedom. The advection equations are discretized using a Leapfrog advection scheme, and the NAG routine nag_gen_lin_sys (NAG) is used to perform an LU factorization to solve the Laplace equation. The discrete model is described in more detail in Appendix A and has previously been used, for example, by Badger (1997), Badger and Hoskins (2001) to investigate the nature of optimal perturbations, and also by Fletcher (1999) to investigate advection schemes.

The simple dynamics of this model should allow a clear understanding of the mechanisms in 4D-Var. Further, as the model is computationally cheap and has only 520 variables, it is straight-forward to explicitly compute the singular vector calculations required later in the thesis.

2.2.1 The Adjoint model

Section 2.1 showed that the adjoint model is used to calculate the gradient of the cost function J with respect to the initial state $\mathbf{x}_0$. The adjoint model for the Eady model is now described.

There are two methods to create the numerical adjoint model of a linear model: find the adjoint of the continuous equations and then discretize or find the adjoint of the discrete forward equations (Sirkes and Tziperman (1997) and Lawless (2001)). Both approaches have been taken to derive the adjoint for the Eady model, and are briefly described in Appendix B.

The adjoint of the continuous equations can be found using the Lagrange multiplier approach, as described for a general model in the previous section. Such an approach has previously been used for other simple models (e.g. Birkett and Nichols, 1983, Birkett, 1986, Xu and Nichols, 1991, Griffith and Nichols, 1994, Griffith, 1997, Wlasak, 1997, Le Dimet et al., 2002). All the equations must be considered simultaneously, and hence such an approach is not suitable for the case of large meteorological models. However, it is useful for the Eady model as it allows an understanding of the dynamics of the adjoint model. Using the notation $q_x = \frac{\partial q}{\partial x}$, the continuous forward equations are summarized by:

$$q_t + zq_x = 0 \qquad \nabla^2\psi = q, \iint \psi dxdz = 0 \qquad \text{in } z\epsilon\left[-\frac{1}{2}, \frac{1}{2}\right] \tag{2.25}$$

$$b_t + zb_x = \psi_x \qquad \psi_z = b \qquad \text{on } z = -\frac{1}{2} \tag{2.26}$$

$$b_t + zb_x = \psi_x \qquad \psi_z = b \qquad \text{on } z = +\frac{1}{2} \tag{2.27}$$

where the initial conditions $q(t = t_0)$ and $b(t = t_0)$ are given. The continuous adjoint equations are summarized by:

$$\hat{q}_\tau - z\hat{q}_x = +\hat{\psi} \qquad \nabla^2\hat{\psi} = 0, \iint \hat{\psi} dxdz = 0 \qquad \text{in } z\epsilon\left[-\frac{1}{2}, \frac{1}{2}\right] \tag{2.28}$$

$$\hat{b}_\tau - z\hat{b}_x = +\hat{\psi} \qquad \hat{\psi}_z = -\hat{b}_x \qquad \text{on } z = -\frac{1}{2} \tag{2.29}$$

$$\hat{b}_\tau - z\hat{b}_x = -\hat{\psi} \qquad \hat{\psi}_z = +\hat{b}_x \qquad \text{on } z = +\frac{1}{2} \tag{2.30}$$

where the final conditions $\hat{q}(t = t_n)$ and $\hat{b}(t = t_n)$ are given and the equations are integrated backwards in time. The time co-ordinate τ, where $\hat{q}_\tau = -\hat{q}_t$, has been introduced to make the backwards time integration explicit. The mean values of the forward and adjoint streamfunction fields are also set to zero. Comparing the adjoint equations with the forward equations, it can be seen that the direction of propagation has been reversed in the adjoint equations, the derivative boundary conditions are given by the horizontal derivative of the buoyancy field and the streamfunction is used to force both the buoyancy and the QGPV fields.

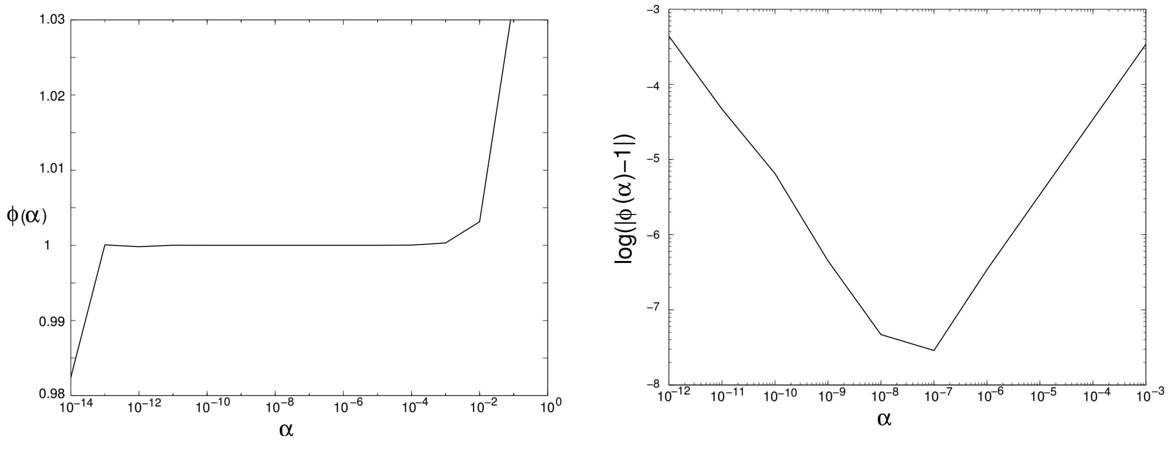
The adjoint of the discrete equations can be found by considering the linear model as a sequence of linear operators. The adjoint of each operator can be found and then these are linked together in the reverse order. Such an approach has been used to find the adjoint models of full meteorological models (e.g. Chao and Chang, 1992, Navon et al., 1992, Rosmond, 1997, Marotzke et al., 1999) and in automatic differentiation compilers such as Giering and Kaminski (1996). Special care needs to be taken if the forward model contains switches (Thuburn and Haine, 2001), although this is not relevant for the Eady model and is therefore not considered in this thesis.

To check that the adjoint model is coded correctly, the ‘Norm Test’ and the ‘Gradient Test’ are used. These are already estabilished methods to test adjoint models, for example, Navon et al. (1992), Li et al. (1994) and Rosmond (1997).

From the definition of the adjoint model (2.6), the adjoint model $\mathbf{M}^T$ should satisfy:

$$(\mathbf{M}(t_N, t_0)\mathbf{x}_0)^T(\mathbf{M}(t_N, t_0)\mathbf{x}_0) = \mathbf{x}_0^T(\mathbf{M}^T(t_0, t_N)\mathbf{M}(t_N, t_0)\mathbf{x}_0) \tag{2.31}$$

where $\mathbf{M}(t_N, t_0)$ is the linear model which is integrated from t_0 to t_N, and $\mathbf{M}^T(t_0, t_N)$ is the adjoint model which is integrated from t_N to t_0. The **Norm test** uses random initial conditions for $\mathbf{x}_0$ to check whether this relation is satisfied to the accuracy of machine precision and was used to check the adjoint for the Eady model. The initial conditions had random data with unit norm, and the model was integrated for 6 hours. The difference between $(\mathbf{M}\mathbf{x}_0)^T(\mathbf{M}\mathbf{x}_0)$ and $\mathbf{x}_0^T(\mathbf{M}^T\mathbf{M}\mathbf{x}_0)$ was zero to 16 decimal places.



(a) (b)

Figure 2.2: *Verification of the gradient calculation for the Eady model: (a) variation of ϕ with respect to α, (b) variation of $\log|\phi(\alpha) - 1|$ with respect to α.*

The **Gradient test** is used to test that both the cost function and adjoint model code are correct. From a Taylor series expansion of the cost function J,

$$J(\mathbf{x}_0 + \alpha\delta\mathbf{x}_0) = J(\mathbf{x}_0) + \alpha\delta\mathbf{x}_0^T \nabla J(\mathbf{x}_0) + \mathcal{O}(\alpha^2) \tag{2.32}$$

which can be rearranged to give

$$\phi(\alpha) = \frac{J(\mathbf{x}_0 + \alpha\delta\mathbf{x}_0) - J(\mathbf{x}_0)}{\alpha\delta\mathbf{x}_0^T \nabla J(x_0)} = 1 + \mathcal{O}(\alpha), \tag{2.33}$$

so that $\phi(\alpha) \rightarrow 1$ as $\alpha \rightarrow 0$. However, this does not hold when α is close to machine zero. Thus, $\phi(\alpha) - 1$ should be close to zero for values of α which are small but not too close to machine zero. Results of this test for the Eady model are shown in Fig. 2.2, where again, the forward model was integrated for 6 hours and random data with unit norm was used for the vector $\delta\mathbf{x}$. These figures are almost identical to those in Navon et al. (1992) and Li et al. (1994), and verify that the adjoint model, cost function and gradient of the cost function have been coded correctly.

2.3 Choice of the Minimization Algorithm and Termination Criteria

The 4D-Var analysis is given by the state which minimizes a quadratic cost function. Therefore, a suitable minimization algorithm is required. On each iteration, the forward model is used to calculate J and the adjoint model is used to calculate ∇J. This is computationally expensive, so the algorithm should ideally converge in as few iterations as possible. To give fast convergence, the algorithm needs to make good use of the the gradient information. The choice of the minimization algorithm will not affect the results, providing it has converged correctly. However, the choice is important so that the 4D-Var algorithm takes a short time to converge to the minimum.

In this section, we compare steepest descents, conjugate gradient, quasi-Newton and memoryless algorithms. Before the minimization algorithms are described, we begin by understanding the properties of the Hessian matrix.

Consider the 3D-Var cost function:

$$J(\mathbf{x}) = \frac{1}{2}(\mathbf{x} - \mathbf{x}^b)^T \mathbf{B}^{-1}(\mathbf{x} - \mathbf{x}^b) + \frac{1}{2}(\mathbf{y} - \mathbf{H}\mathbf{x})^T \mathbf{R}^{-1}(\mathbf{y} - \mathbf{H}\mathbf{x}). \tag{2.34}$$

The Hessian matrix of J is defined as the second derivative of the cost function:

$$\mathbf{A} = \nabla\nabla J = \mathbf{B}^{-1} + \mathbf{H}^T\mathbf{R}^{-1}\mathbf{H}. \tag{2.35}$$

and is constant because J is quadratic (Gill et al., 1981). By diagonalizing the Hessian matrix (see for example, Jordan and Smith (1997)), it can be shown that the isocontours of the cost function are ellipsoids whose principle axes are the eigenvectors of the Hessian with lengths proportional to the reciprocals of the square roots of the corresponding eigenvalues (Gill et al., 1981).

For example, consider the two variable quadratic function $F(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T\mathbf{A}\mathbf{x} - \mathbf{b}^T\mathbf{x}$, where $\mathbf{A}$ is a symmetric positive definite matrix:

$$\mathbf{A} = \begin{bmatrix} 5.5. & 4.5 \\ 4.5 & 5.5 \end{bmatrix} \qquad \mathbf{b} = \begin{bmatrix} 22.5 \\ 27.5 \end{bmatrix}. \tag{2.36}$$

At a minimum, $\mathbf{A}\mathbf{x} = \mathbf{b}$, and $\mathbf{x} = (0, 5)$. The eigenvectors are $(0.7, -0.7)$ and $(0.7, 0.7)$, with corresponding eigenvalues $\lambda_1 = 1$ and $\lambda_2 = 10$. The isocontours of F and the eigenvectors and eigenvalues of $\mathbf{A}$ are illustrated in Fig. 2.3.

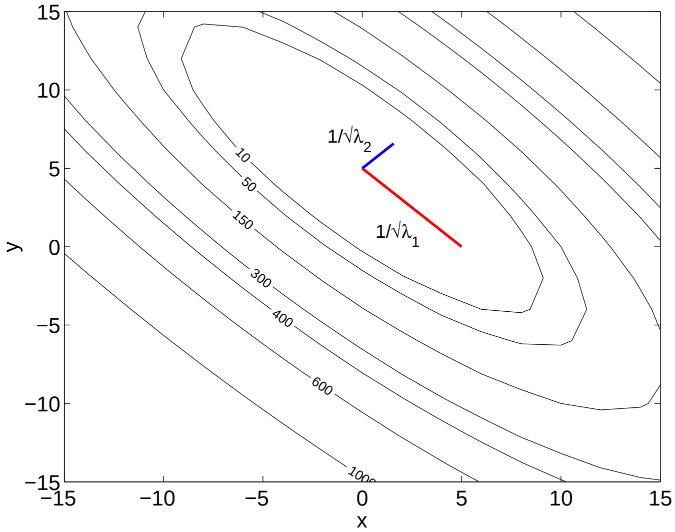


Figure 2.3: *Isocontours of the quadratic function $F = \frac{1}{2}\mathbf{x}^T\mathbf{A}\mathbf{x} - \mathbf{b}^T\mathbf{x}$. The blue and red lines show the directions of the eigenvectors of $\mathbf{A}$, with the corresponding lengths written at the side of each eigenvector, where the eigenvalues of $\mathbf{A}$ are $\lambda_1 = 1$ and $\lambda_2 = 10$.*

As the ratio between the maximum and the minimum eigenvalues increases, the isocontours become more elliptical. If one eigenvalue becomes zero, the isocontours become parallel and the minimum is then non-unique. Thus, a minimum of the cost function can only exist if all the eigenvalues are real and positive, which is satisfied if and only if the Hessian matrix is positive definite (e.g. Atkinson, 1989).

The condition number is defined as the ratio between the maximum and the minimum eigenvalue of the Hessian matrix. If this is large, then the problem is poorly conditioned, and a minimization algorithm will take a long time to reach the minimum. Some minimization algorithms, however, implicitly use information about the Hessian matrix to speed up the rate of convergence.

The minimization algorithms that are commonly used in data assimilation approximate Newtons' method (Navon and Legler, 1987), which is now described. Letting $\mathbf{x}_{k+1} = \mathbf{x}_k + \delta\mathbf{x}_k$, where k is the iteration number, a truncated Taylor series expansion of J gives:

$$J(\mathbf{x}_{k+1}) = J(\mathbf{x}_k) + (\mathbf{x}_{k+1} - \mathbf{x}_k)^T \nabla J(\mathbf{x}_k) + \frac{1}{2}(\mathbf{x}_{k+1} - \mathbf{x}_k)^T \nabla\nabla J(\mathbf{x}_{k+1} - \mathbf{x}_k). \qquad (2.37)$$

Setting the Jacobian of J with respect to $\mathbf{x}_{k+1}$ to zero, then **Newtons' method** is:

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \mathbf{A}^{-1}\nabla J_k \qquad (2.38)$$

where $\mathbf{A} = \nabla\nabla J$ is the Hessian of J, and ∇J_k is the Jacobian of J with respect to $\mathbf{x}_k$. There is no guarantee that $J(\mathbf{x}_{k+1}) < J(\mathbf{x}_k)$ and therefore it is better to modify the method (Beale, 1988), so that on each iteration a line search is performed to find a scalar $\alpha_k > 0$ which minimizes $J(\mathbf{x}_{k+1})$ such that

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \alpha_k \mathbf{A}^{-1}\nabla J_k. \qquad (2.39)$$

In 4D-Var the Hessian matrix is unknown as only J and ∇J are calculated. Therefore the Hessian matrix must be approximated.

The **steepest descent** method approximates the Hessian with an identity matrix, which results in an equation such that a step is made in the 'downhill' direction on each iteration. However, if the condition number of the Hessian matrix is large, the method is slow to converge. This is sometimes known as the 'narrow valley effect' and the algorithm is seen to zig-zag into the minimum.

There are two main types of algorithm which give much better rates of convergence than the steepest descents method by using information about the Hessian matrix: the conjugate gradient method and the quasi-Newton method.

The **conjugate gradient method** constructs a set of conjugate search directions from the set of ∇J. A pair of conjugate directions $\mathbf{d}_i$ and $\mathbf{d}_j$ are $\mathbf{A}$-orthogonal, ($\mathbf{d}_i^T \mathbf{A} \mathbf{d}_j = 0$ for $i \neq j$). This means that if a co-ordinate transformation is applied so that the isocontours are spherical, the search directions become orthogonal (Shewchuk, 1994). The use of conjugate directions means that the conditioning of the Hessian matrix is taken into account and the directions are conjugate to each other, so the algorithm is effectively not stepping in the same direction twice. By using the ∇J to construct the conjugate directions, it is possible to construct directions which are conjugate to all previous directions even though only the previous search direction needs to be stored, giving a storage of $\mathcal{O}(3n)$ where n is the dimension of the state vector (Beale, 1972, 1988). The main problem is that with inexact line searches and rounding errors, the directions may lose their conjugacy giving slower convergence and the algorithm may even need to be restarted.

The second type of algorithm is the **Quasi-Newton method**, also known as the variable metric method. This uses the ∇J to successively update an approximation to the Hessian matrix. The approximation $\mathbf{A}_k$ must satisfy the quasi-Newton condition:

$$\mathbf{A}_k(\mathbf{x}_{k+1} - \mathbf{x}_k) = \nabla J_{k+1} - \nabla J_k \tag{2.40}$$

which can be derived from (2.38) (Press et al., 1992). This method gives an estimate for the optimal step size so that line searches are not necessary. However, it requires a storage of $\mathcal{O}(n^2)$. This is not feasible for operational data assimilation as n is large, but is suitable for 4D-Var with the Eady model.

It is possible to combine the conjugate gradient and quasi-Newton methods to give **limited memory or memoryless methods**. In the conjugate gradient method, search directions are generated using the set of ∇J, but in a limited memory method, the set of $\mathbf{A}^{-1}\nabla J$ are used where $\mathbf{A}^{-1}$ is approximated using a limited number of quasi-Newton updates. As $\mathbf{A}^{-1}$ does not need to be stored, these methods require only small storage ($\mathcal{O}(7n)$).

2.3.1 Minimization Algorithm Comparison Experiments

We now use 4D-Var experiments with the Eady model to compare four different minimization algorithms: steepest descent, conjugate gradient, quasi-Newton and memoryless quasi-

Newton. The steepest descent algorithm uses ∇J_k as the search direction and uses the largest possible step size α_k such that $J_{k+1} < J_k$. The conjugate gradient algorithm, known as A22CGM (Nash, 1990, 2003), uses a linear search to bracket a minimum to find α_k. The quasi-Newton algorithms use a BFGS (Broyden-Fletcher-Goldfarb-Shanno) update and are known as CONMIN or algorithm 500 from TOMS (Shanno and Phua (1976, 1980) and Shanno and Phua (2003)). The memoryless version uses two vectors to build the current approximation of the Hessian matrix and uses Beale restarts with an inexact line search (Davidons' cubic interpolation). The algorithms have previously been used by Chao and Chang (1992) and Navon and Legler (1987).

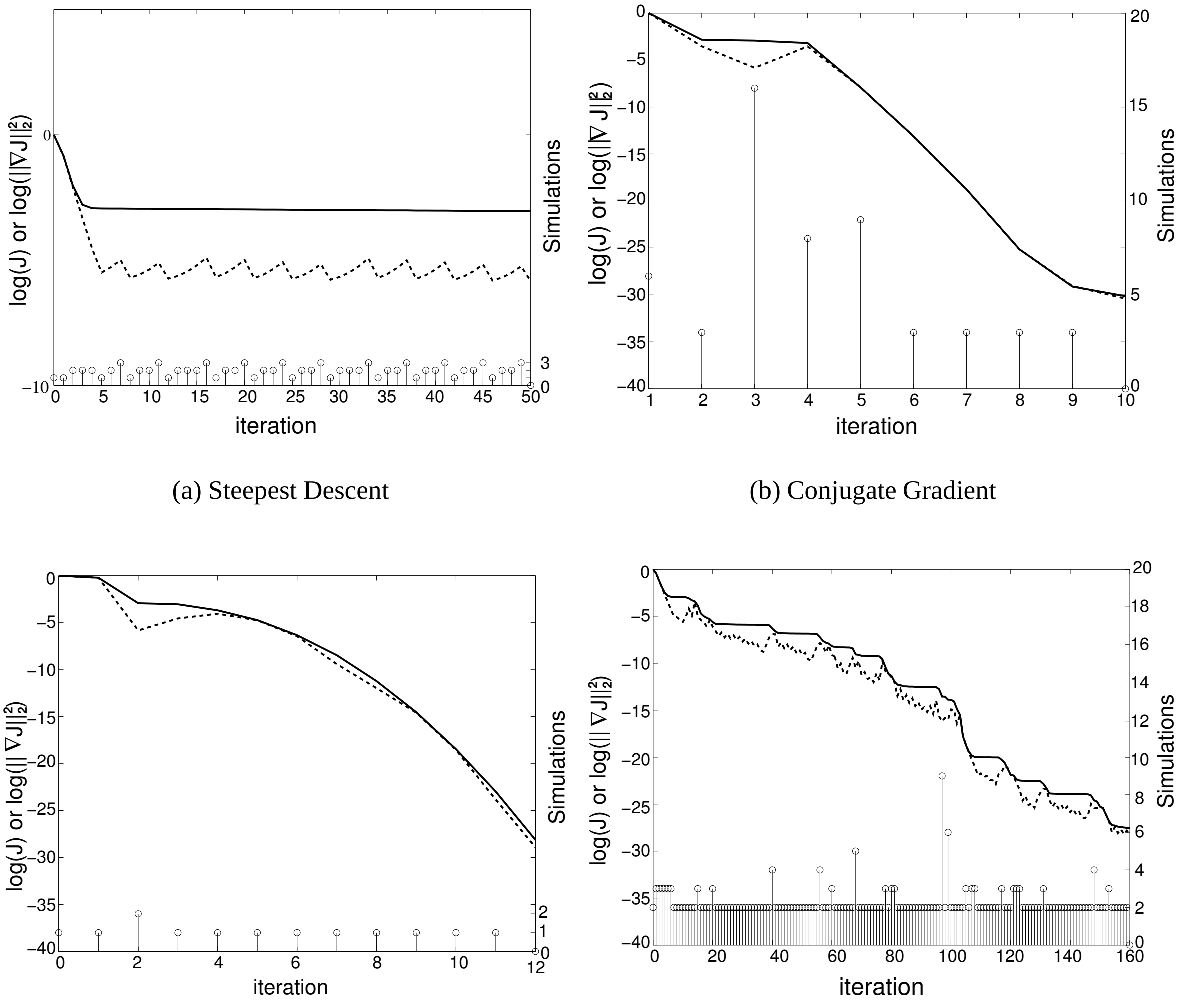
The 4D-Var experiment considered uses no background state, the true state is given by the most unstable growing Eady wave (A.38) and perfect observations of both the buoyancy on the lower boundary and the interior QGPV are provided at T+0 and T+6. This problem is well-posed without a background state. All the minimization algorithms, except for the steepest descent algorithm, are terminated when $\|\nabla J\|_2^2 < 5 \times 10^{-28}$.

The comparison of the minimization algorithms is shown in Fig. 2.4. The steepest descent algorithm shows an extremely slow rate of convergence. The magnitude of the gradient oscillates as the algorithm zig-zags into the minimum. This is due to the poor conditioning of the Hessian matrix. The experiment has been run until 200 iterations, with no further change in the rate of descent.

The conjugate gradient algorithm gives a much faster rate of convergence than the steepest descents method, and reaches the minimum in 10 iterations. However, on each iteration, many function evaluations are required. This is because the line search is found by bracketing a minimum.

The quasi-Newton method also gives a much faster rate of convergence than the steepest descents method, reaching the minimum in 10 iterations. Further, on each iteration, there are only a few function evaluations as an estimate of the optimal step size is provided. For this reason, the quasi-Newton algorithm performs better than the conjugate gradient algorithm.

The memoryless algorithm gives a slower rate of convergence than both the conjugate gradient method and the quasi-Newton method, requiring 160 iterations to give the same accuracy. This is because the Hessian is approximated with only a rank-2 matrix. It is surprising that the combined conjugate-gradient, quasi-Newton (memoryless) algorithm has a worse performance than the conjugate gradient method. This is perhaps due to the differences in the line search methods or the restarts.



(a) Steepest Descent

(b) Conjugate Gradient

(c) Quasi-Newton

(d) Memoryless

Figure 2.4: *Behaviour of the cost function J^o using (a) Steepest Descent Algorithm (b) Conjugate Gradient Algorithm (c) Quasi-Newton Algorithm (d) Memoryless Algorithm, with increasing iterations (gradient evaluations). The solid line corresponds to the cost function J, and the dotted line corresponds to the squared Euclidean norm of the gradient $||\nabla \mathbf{J}||^2$ with the magnitude on the left hand axes. The circles show the number of simulations (function evaluations) used to calculate the next step, with the magnitude on the right hand axes. This minimization is for the case with no observations on the top boundary. Note that the axes have different scales.*

Based on this comparison, the CONMIN quasi-Newton algorithm is chosen for all further 4D-Var experiments.

It is important to terminate the minimization algorithm once it has converged. To terminate before convergence is reached would not give the optimal analysis, but to terminate after it has converged would waste computer time. There are three basic ways to define the termination or convergence criteria:

1. At a minimum, $\nabla J = 0$.
2. At a minimum, there is negligible change in J.
3. At a minimum, there is negligible change in $\mathbf{x}$.

The size of these quantities can be defined using different measures and norms. For example, the absolute error and relative error of the gradient of J are defined as:

$$abs(\nabla J) = \|\nabla J\| \qquad\qquad rel(\nabla J) = \frac{\|\nabla J\|}{J} \tag{2.41}$$

where $\| \ .\|$ is a specified norm. It is hard to define a tolerance using the absolute error as it does not take into account the value of J, which may vary for different minimizations. It is also difficult to use the relative error, as this is undefined when J is zero. Therefore the combination error (Gill et al., 1981), defined as:

$$comb(\nabla J) = \frac{\|\nabla J\|}{1+J}. \tag{2.42}$$

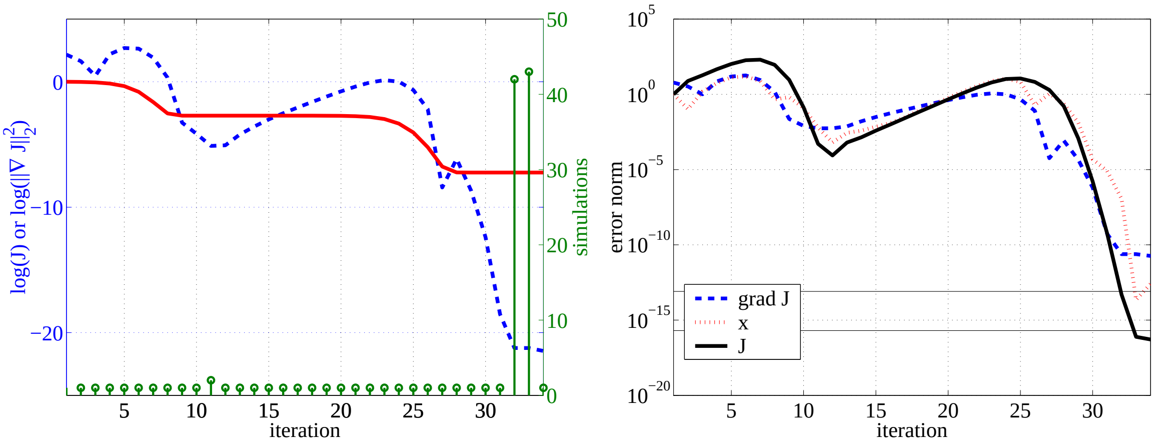
is chosen instead. This is a combination of the absolute and the relative errors. When J is zero, it gives the absolute error, but when J is large it gives a value similar to the relative error. The maximum norm (∞-norm), defined by $\|\mathbf{x}\|_\infty = \overset{max}{i} \ \ |x_i|$ is chosen to measure the size of the vectors, as it gives an indication of the extreme values.

For the experiments in the rest of this thesis, the quasi-Newton minimization algorithm is terminated if any one of the following convergence criteria are satisfied:

$$\frac{\|\nabla J\|_\infty}{1+J} \leq \tau_1 \tag{2.43}$$

$$\frac{\|\mathbf{x}_k - \mathbf{x}_{k-1}\|_\infty}{1+\|\mathbf{x}_k\|_\infty} \leq \tau_2 \tag{2.44}$$

$$\frac{\sqrt{J_k} - \sqrt{J_{k-1}}}{1+\sqrt{J_k}} \leq \tau_3 \tag{2.45}$$



(a) Behaviour of the Cost Function

(b) Termination Criteria

Figure 2.5: *The behaviour of the cost function $J = J^b + J^o$ with increasing iterations (a) cost function (solid line), squared Euclidean norm of the gradient vector (dashed line), simulations or function evaluations (circles). (See Fig. 2.4 for further details). (b) The set of termination criteria (2.43,2.44,2.45). The specified tolerances are shown by the thin solid lines.*

where τ_i are specified tolerances. J is similar to a squared Euclidean norm, therefore $\sqrt{J}$ is chosen so that this is similar to the Euclidean norm and should have a similar magnitude to the other termination criteria. The cost function J contains a multiplicative constant; therefore, to ensure that the termination criteria are robust, the cost function J and the gradient $\|\nabla J\|$ are first scaled so that they both have a value of unity at the beginning of the minimization.

The experiments in the rest of this thesis use a background state, therefore the convergence criteria for a minimization using a background state are examined. The 4D-Var experiment considered uses the true state given by the most unstable Eady wave, with a background state which has a phase error. Observations of the lower level wave are given at the beginning and the end of a 6 hour window. The weight given to the observation term is $\sigma_o^{-2} = 1$, and the weight given to the background term is $\sigma_b^{-2} = 10^{-6}$. This experiment is described in further detail in Chapter 3. The behaviour of the cost function and the different convergence criteria are shown as a function of iterations in Fig. 2.5. The tolerances are chosen so that the minimization algorithm will terminate at the point where machine precision (10^{-16}) is reached. That is, we would ideally wish to terminate at the point at which on the next iteration, there is a dramatic increase in the number of function evaluations. In Fig. 2.5(a), this occurs at iteration number 32 and therefore the tolerances should be chosen so that the minimization algorithm terminates on iteration 31. The values of the termination criteria shown in Fig. 2.5(b) show that the measure of J reaches a lower value than the other termination criteria. Therefore, this is set to have a lower tolerance. Based on Fig. 2.5(b) and plots for similar experiments (not shown), the tolerances are chosen as: $\tau_1 = 8 \times 10^{-14}, \tau_2 = 8 \times 10^{-14}, \tau_3 = 2 \times 10^{-16}$. If any of the termination criteria are satisfied, the minimization algorithm is stopped. The tolerances have been specified as relatively low values to ensure that the minimization algorithm has converged, and so it may occasionally be the case that a large amount of unnecessary simulations (function evalutions) are computed.

2.4 Background Error Correlations

The background error covariance matrix contains the variances and the correlations of the background state errors. The correlations may be between grid points corresponding to a particular model variable (auto-correlations), or may be between grid points corresponding to different model variables (cross-correlations) (Weaver and Courtier, 2001). In the Eady model, the QGPV and buoyancy errors are assumed to be uncorrelated and therefore there are no cross-correlations. In this section, we describe the specification of the background error auto-correlations that will be used in further experiments.

The auto-correlations play an important role in data assimilation as they ensure that the analysis is smooth and spread information from an observation to the surrounding grid points. That is, the analysis algorithm filters the observational noise and then interpolates the filtered data to the grid points. Daley (1985) and Hollingsworth (1987, 2003) illustrated the filtering and interpolating properties of the background error covariance using an eigenvector decomposition of the covariance matrix. The structures with small eigenvalues are damped, and the experiments by Daley (1985) showed that these are the structures with small wavelengths. As the small-scale structures are damped, the analysis is comparatively smooth.

The background error covariance matrix $\mathbf{B}$ can be written explicitly in terms of the assumed background error variances and correlations as:

$$\mathbf{B} = \boldsymbol{\sigma}\boldsymbol{\rho}\boldsymbol{\sigma} \tag{2.46}$$

(see for example, Weaver and Courtier (2001) and Kalnay (2003)), where $\boldsymbol{\sigma}$ is a diagonal matrix of background error standard deviations σ, and $\boldsymbol{\rho}$ is a symmetric matrix of background error correlation coefficients, ρ, such that the correlation coefficients $-1 < \rho < 1$ are related to the covariances by:

$$\rho(x,y) = \frac{cov(x,y)}{\sigma(x)\sigma(y)} \text{ where } -1 \leq \rho \leq 1. \tag{2.47}$$

The first data assimilation methods modelled the background error covariances using correlation functions (see for example, Julian and Thiébaux (1975), Thiébaux (1975), Daley (1991) and Kalnay (2003)). For example, a Gaussian exponential function:

$$\rho_{\text{Gaussian}}(x) = e^{-\frac{x^2}{2l^2}} \tag{2.48}$$

was used in optimal interpolation schemes (using subdomains of the globe), where the $\mathbf{B}$ matrices were inverted explicitly. Here, x is the distance between the correlated grid points and l is the length scale. This function was used for example, in the ECMWF optimal interpolation scheme (Lorenc, 1981).

It is expensive to invert the background error covariance matrix, and therefore it is better to define the inverse matrix. For example, in the the ECMWF and Met Office variational data assimilation schemes (Courtier et al. (1998), Rabier et al. (1998), Derber and Bouttier (1999) and Lorenc et al. (2000)), the J^b term is defined in spectral space using a a spherical-harmonic expansion with the correlation spectra defined such that the small wavelength modes are penalized.

A further alternative is to use a Laplace based method which is defined in grid space. For example, Derber and Rosati (1989) used an iterative Laplacian grid point filter, Schröter et al. (1993) used a second derivative smoothness constraint, and Weaver and Courtier (2001) developed a correlation method based on the diffusion equation. An advantage of these methods is that they are particularly suitable for domains with fixed boundary conditions, such as the ocean.

A simple technique to model the horizontal error correlations for the 4D-Var algorithm using the Eady model is now developed. As matrices are computationally expensive to invert, we define the inverse covariance matrix. The technique is also a Laplace based method, and we illustrate how this method is in fact very similar to defining a Gaussian correlation function. The relationship between Laplace-based correlation functions and Gaussian correlation functions was also briefly described by Rodgers (2000) and Bennett (2002). The Eady model has periodic boundary conditions in the horizontal, so the correlation model also requires periodic boundary conditions.

We first consider the Gaussian function $\rho_{\text{Gaussian}}(x)$, with spectral response (Fourier transform) $\hat{\rho}_{\text{Gaussian}}(k)$:

$$\rho_{\text{Gaussian}}(x) = e^{-\frac{x^2}{2l^2}} \qquad\qquad \hat{\rho}_{\text{Gaussian}}(k) = le^{-\frac{k^2l^2}{2}} \tag{2.49}$$

where l is the length scale and k is the wavenumber, as shown by for example Riley et al. (1998). We then define an inverse covariance matrix based on Laplace smoothing, to give a similar spectral response (or Transfer function) to the Gaussian correlation function. Following the work by Lea (2001) and Bennett (2002) the inverse correlation matrix $\boldsymbol{\rho}_{\text{Laplace}}^{-1}$ with spectral response $\hat{\rho}_{\text{Laplace}}(k)$ is defined as:

$$\boldsymbol{\rho}_{\text{Laplace}}^{-1} = w_0\mathbf{I} + w_1(\mathbf{L}_{xx})^2 \qquad\qquad \hat{\rho}_{\text{Laplace}}(k) = \frac{1}{w_0 + w_1k^4} \tag{2.50}$$

where $\mathbf{L}_{xx}$ is a second derivative matrix with periodic boundary conditions, and w_0 and w_1 are constant scalar coefficients. Choosing $w_0 = \frac{1}{l}$, and $w_1 = w_0\frac{l^4}{2}$, then $\hat{\rho}_{\text{Laplace}}$ has a similar spectral response to $\hat{\rho}_{\text{Gaussian}}$, as shown in Fig. 2.6. In the following, the coefficients w_0 and w_1 have indeed been chosen so that the Laplace-based correlation matrix has a similar spectral response to a Gaussian correlation function.

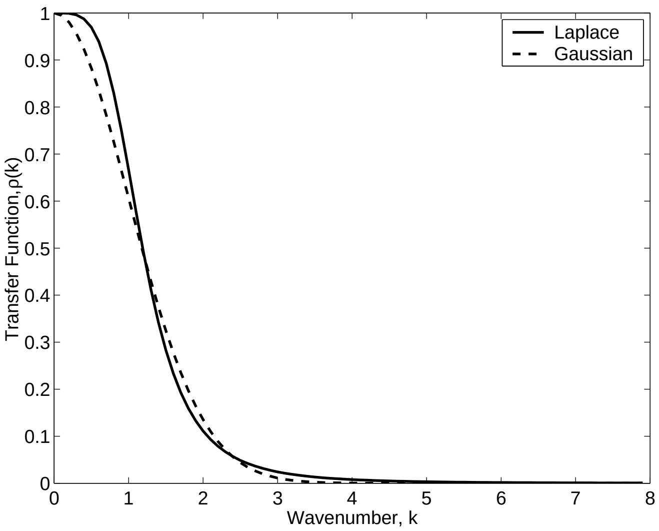


Figure 2.6: *A comparison of power spectra with $l = 1$. The dashed line represents a Gaussian correlation function with the spectral response $\hat{\boldsymbol{\rho}}_{Gaussian}(k) = le^{\frac{-k^2l^2}{2}}$, and the solid line represents a Laplace smoother with spectral response $\hat{\boldsymbol{\rho}}_{Laplace}(k) = l/(1+\frac{k^4l^4}{2})$.*

The similarity between Laplace-based correlations and Gaussian correlations can be made more clear by comparing the functions in physical grid point space rather than spectral space. We consider a domain of length $X = 200$, with $n = 80$ grid points, and grid spacing $\Delta x = X/(n-1)$ and compare the correlation functions associated with an observation at the 10^{th} grid point.

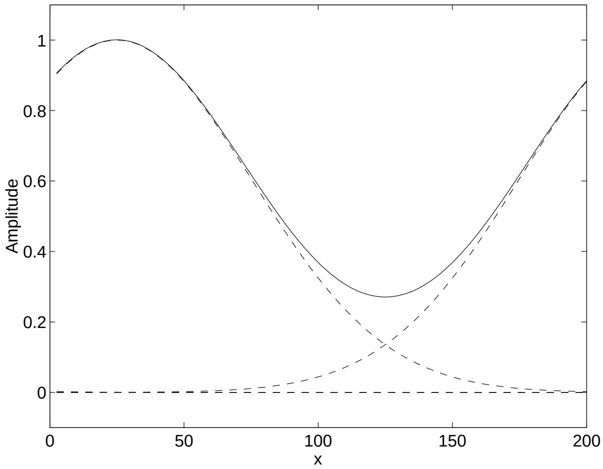


Figure 2.7: *A Gaussian correlation function with periodic boundary conditions and length scale $l = 20\Delta x$. The dashed curves represent equations f and g from equation (2.51). The solid curve represents the sum of f and g.*

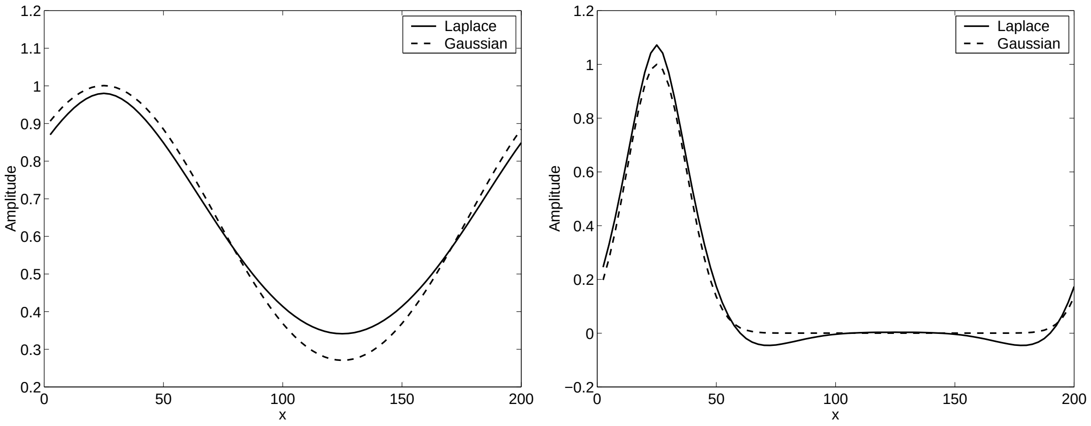
The Gaussian correlation function does not account for periodic boundary conditions. Therefore, we consider a sum of Gaussian functions:

$$f(i) = exp\left[-\left\{\frac{(i-10)\Delta x}{l}\right\}^2\right] \qquad g(i) = exp\left[-\left\{\frac{(i-(80+10))\Delta x}{l}\right\}^2\right] \tag{2.51}$$

where i is the grid point number. These curves are shown in Fig. 2.7 with $l = 20\Delta x$. The individual functions are not periodic, however the sum of the functions is periodic.

We now compare the Gaussian correlation function with the Laplace-based correlation. The $\boldsymbol{\rho}_{\text{Laplace}}$ matrix is inverted using MATLAB, and the 10th column is compared with the sum of the two correlation functions. Comparisons with length scales $l = 20\Delta x$ and $l = 5\Delta x$ are shown in Fig. 2.8 (a) and (b) respectively. There are some slight differences in the curves. For example, with $l = 20\Delta x$, ρ_{Laplace} has a smaller amplitude than ρ_{Gaussian}, but with $l = 5\Delta x$, ρ_{Laplace} has a larger amplitude. Also, for $l = 5\Delta x$, ρ_{Laplace} is sometimes negative, whilst ρ_{Gaussian} is always positive. Nevertheless, the two approaches do give very similar results.

In this section, a horizontal correlation model has been developed. The inverse background error covariance matrix is modelled using a second derivative matrix. This approach



(a) $l = 20\Delta x$

(b) $l = 5\Delta x$

Figure 2.8: *Comparison of a column of* $\boldsymbol{\rho}_{Laplace}$ *(solid curves) and a sum of* $\rho_{Gaussian}$ *functions (dashed curves), for (a)* $l = 20\Delta x$ *and (b)* $l = 5\Delta x$.

means that the $\mathbf{B}$ matrix does not need to be inverted, and periodic boundary conditions are easily incorporated. Further, the method is extremely simple to code and apply to the Eady model experiments. Whilst this technique is not a central part to this thesis, representing inverse covariances with differential operators is an interesting technique which may be suitable for operational data assimilation. Such a technique is currently being developed by Qin Xu (personal communication).

2.5 Summary

The 4D-Var algorithm, that is to be used in the identical twin experiments in the rest of this thesis, has been described in detail. The equations that are used to minimize the constrained 4D-Var cost function have been derived using two approaches: linear algebra and Lagrange multipliers. In both cases, the gradient of J is found by integrating the adjoint model backwards in time from final conditions of zero and adding the adjoint forcing at each timestep.

The 2D Eady model, used for the 4D-Var experiments has also been described. The model is one of the most simple linear models of baroclinic instability. Although it contains many approximations, this model is highly suitable as it has a small dimension and should isolate the important mechanisms in 4D-Var. The model is derived from the quasi-geostrophic equations and uses quasi-geostrophic potential vorticity (QGPV) and buoyancy as the model variables. The QGPV and buoyancy are advected by the basic state zonal wind and are linked together via an elliptic equation.

The continuous adjoint equations were described, with the derivation of the adjoint of the discrete equations and the adjoint of the continuous equations outlined in Appendix B. The discrete adjoint model has been validated using both the norm test and the gradient test.

Four minimization algorithms have been compared: steepest descent, conjugate gradient, quasi-Newton and a memoryless combined quasi-Newton conjugate method. The quasi-Newton minimization algorithm has been selected as the best method for 4D-Var with the Eady model, and convergence criteria based on the value of J, ∇J and $\mathbf{x}$ have been specified.

The inverse background error covariance matrix has been modelled using Laplace smoothing and it has been shown that this is similar to using a Gaussian correlation function.

In the following chapter, the 4D-Var algorithm will be used to tackle the questions that were posed in the first chapter.

Chapter 3

4D-Var Results

Previous studies have shown 4D-Var to perform well in regions of baroclinic instability in comparison with 3D-Var, as discussed in Chapter 1. In particular, it has been shown that 4D-Var is able to reconstruct parts of the atmospheric state that are unobserved (e.g. Courtier and Talagrand, 1987, Thépaut and Courtier, 1991, Rabier and Courtier, 1992, Tanguay et al., 1995). It has also been shown that 4D-Var is able to generate westward tilting analysis increments that are necessary for baroclinic growth (e.g. Thépaut et al., 1996, Rabier et al., 1998, 2000).

Although these two properties have been demonstrated, they are not well understood. The purpose of this chapter is to investigate these properties using simple identical twin experiments with the Eady model. In the experiments, the true state is given by the most rapidly growing or decaying Eady wave. These modes grow or decay through the interaction of boundary temperature waves. The background state contains only a displacement error and observations are provided of the lower boundary wave only. There are no observations of the interior QGPV. Thus, 4D-Var must use the observations of the lower boundary wave to reconstruct or infer the correct position of the upper level wave. Due to the symmetry of the Eady model, this is equivalent to providing observations of the upper level wave and inferring the position of the lower level wave.

The chapter begins by describing the experimental design for the experiments in this chapter. Section 3.2 investigates the ability of 4D-Var to reconstruct the upper level wave and Section 3.3 investigates the ability to generate the correct vertical structure, for both growing and decaying modes. The chapter ends with a concluding discussion. For simplicity, this chapter considers the ability to reconstruct, and the ability to generate the correct vertical structures separately. It is important to note, however, that these two properties are strongly linked in the following experiments.

3.1 Experiment Design

The experiments in this chapter are now described. The **true state** initial conditions $\mathbf{x}^t$ are given by the most rapidly growing or decaying Eady wave. From equation (A.38) in Appendix A, the values of the non-dimensional buoyancy anomalies on the boundaries are defined as:

$$\text{Growing Mode: } \frac{\partial \psi'}{\partial z} = \sinh(kz) cos(kx) - \alpha \cosh(kz) \sin(kx) \text{ on } z = \pm\frac{1}{2} \tag{3.1}$$

$$\text{Decaying Mode: } \frac{\partial \psi'}{\partial z} = \sinh(kz) cos(kx) + \alpha \cosh(kz) \sin(kx) \text{ on } z = \pm\frac{1}{2} \tag{3.2}$$

where the non-dimensional wave number is $k = 1.6$, $\alpha \approx 1.5$ and the interior QGPV anomalies q are zero. The true state is then evolved using a 6 hour integration of the Eady model $\mathbf{M}(t_N, t_0)$ to give an assimilation window length of 6 hours. The notation $T + 0$ and $T + 6$ define the beginning and the end of the assimilation window respectively. It is important to use a numerical integration of the true state rather than the analytical evolution so that model error can be neglected.

Synthetic **observations** $\mathbf{y}_0$ and $\mathbf{y}_N$ of the entire lower level buoyancy field are taken from the evolved true state at the beginning and the end of the 6 hour window. For example, suppose that $\mathbf{x}_q$ is a vector of the interior QGPV of dimension 440 and $\mathbf{x}_U$ and $\mathbf{x}_L$ are vectors of the upper and lower buoyancy values respectively with dimension 40. Then, the state vector and observation operator can be defined using:

$$\mathbf{x} = \begin{bmatrix} \mathbf{x}_q \\ \mathbf{x}_U \\ \mathbf{x}_L \end{bmatrix} \qquad \mathbf{H} = \begin{bmatrix} \mathbf{0}_{(40\times440)} & \mathbf{0}_{(40\times40)} & \mathbf{I}_{(40\times40)} \end{bmatrix}. \tag{3.3}$$

Random noise ε_0 and ε_N is added to the observations so that $\mathbf{y}_0 = \mathbf{H}\mathbf{x}_0^t + \varepsilon_0$ and $\mathbf{y}_N = \mathbf{H}\mathbf{M}\mathbf{x}_0^t + \varepsilon_N$. The noise is defined to have a Gaussian distribution with standard deviation σ, using an algorithm based on Press et al. (1992).

The **background state** $\mathbf{x}^b$ is given by the same wave as the true state but with a displacement error of $\frac{1}{4}$ wavelength, which is $10\triangle x = 1000km$.

The 4D-Var algorithm, described in the previous chapter, is then used to find the analysis $\mathbf{x}^a$ that minimizes the cost function:

$$\begin{aligned} J(\mathbf{x}_0) &= (\mathbf{x}_0 - \mathbf{x}^b)^T \mathbf{B}^{-1} (\mathbf{x}_0 - \mathbf{x}^b) + (\mathbf{y}_0 - \mathbf{H}\mathbf{x}_0)^T \mathbf{R}^{-1} (\mathbf{y}_0 - \mathbf{H}\mathbf{x}_0) \\ &+ (\mathbf{y}_N - \mathbf{H}\mathbf{x}_N)^T \mathbf{R}^{-1} (\mathbf{y}_N - \mathbf{H}\mathbf{x}_N) \end{aligned} \tag{3.4}$$

subject to the constraint $\mathbf{x}_N = \mathbf{M}(t_N, t_0)\mathbf{x}_0$.

Thus, the 4D-Var algorithm must use the observations of the lower level wave, given at the beginning and end of a 6 hour window, to move the unobserved upper level wave.

These experiments are not unrealistic, although they are extremely simple. For example, in operational data assimilation, there are many surface observations and only a few upper air observations. This is particularly the case over land. Further, due to the symmetry of the model, these experiments are equivalent to the case where the upper level wave is observed and the lower level wave is reconstructed. This may be the case with satellite data or aircraft data where there are many observations of the upper troposphere but only a few observations near the surface. For example, the infra-red sounding radiometers such as HIRS and IASI can only obtain atmospheric profiles above cloud (Eyre, 2000), and current microwave sounding radiometers such as AMSU are mainly sensitive to the upper tropospheric temperature (Bouttier and Kelly, 2001). Also, satellite derived winds are mostly in the upper troposphere (M. Forsythe, Personal Communication).

3.2 Reconstruction

This section investigates how 4D-Var uses the time sequence of observations on the lower boundary to reconstruct the upper level wave. The true state is given by the growing Eady wave and so 4D-Var must use the evolution information to infer the upper level wave.

We begin by first considering perfect observations, so that the 4D-Var interpolating properties may be isolated from the filtering properties. The behaviour of the minimization algorithm is examined to gain some insight into how the information is propagated to the unobserved regions. The effect of varying the weight given to the background state is then examined. Many of the previous studies did not include a background state when examining the reconstructive properties, for example, Rabier and Courtier (1992), Tanguay et al. (1995), Thépaut and Courtier (1991). Therefore, it is important to investigate the effect of the J^b term on the reconstructive properties. The effect of observational noise is finally considered. Courtier and Talagrand (1987) found that a smoothing term was needed to reduce unrealistic small scale noise. We therefore investigate the effect of incorporating correlations (or smoothing) into the background error covariance matrix.

3.2.1 Behaviour of the Minimization

This section aims to understand how the upper level wave is reconstructed during the minimization process, to gain some insight into how the information is propagated from the lower boundary to the upper boundary.

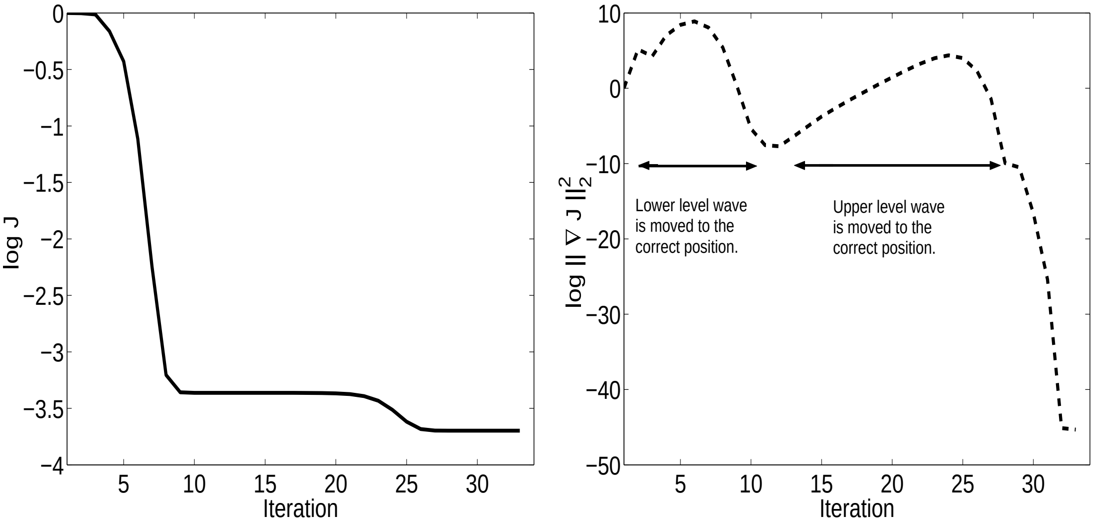
We consider the case with perfect observations ($\varepsilon_0 = \varepsilon_N = 0$) and assume that the background state and observation errors are uncorrelated so that the error covariance matrices are diagonal. We also assume that the error variances σ_b^2 and σ_o^2 are the same for all grid points and observations respectively. Both the true state and the background state QGPV fields are zero, and therefore, the background state errors for QGPV are zero. Thus, we assume that a relatively small analysis increment is added to the QGPV field. Following these assumptions, the error covariances are defined as:

$$\mathbf{R}^{-1} = \sigma_o^{-2}\mathbf{I} \qquad\qquad \mathbf{B}^{-1} = \begin{cases} \sigma_b^{-2}\mathbf{I} \text{ for Buoyancy on } \hat{z} = \pm\dfrac{1}{2} \\ 10^5\mathbf{I} \text{ for QGPV in } -\dfrac{1}{2} < \hat{z} < \dfrac{1}{2} \end{cases} \tag{3.5}$$

where the number 10^5 is chosen as an arbitrary large number so that only small amplitude analysis increments are added to the QGPV field. As only small changes are made to the QGPV, the QGPV fields are not shown in the figures that follow.

The values of J and $\|\nabla J\|_2^2$, during a minimization using $\sigma_b^{-2} = 0.04$ and $\sigma_o^{-2} = 1$, are shown in Fig. 3.1. There are two stages to the minimization. In the first stage, the lower boundary is moved to the correct position. This was established by examining the state vector on each iteration, (Johnson et al., 2002) but is not shown here. The lower boundary is observed, so this causes a dramatic decrease in the value of the cost function. In the second stage, the upper boundary is moved to the correct position. This has only a small effect on the value of J, but eventually results in a dramatic reduction in the value of $\|\nabla J\|_2^2$, as the minimum is reached.

On the 10th iteration, it may seem that the minimization algorithm has converged, as the



(a) $\log J$ (b) $\log \|\nabla J\|_2^2$

Figure 3.1: *The values of (a)* $\log J$ *and (b)* $\log \|\nabla J\|_2^2$ *during the minimization process for perfect observations with* $\sigma_o^{-2} = 1$ *and* $\sigma_b^{-2} = 0.04$. *Note that the graphs have different vertical scales.*

value of J does not change significantly. However, if the minimization algorithm was terminated after only ten iterations, the upper boundary wave would not be moved, and the reconstructive benefits of 4D-Var would be lost. Thus, it is clear that it is vital that the minimization algorithm is not terminated until it has satisfied the convergence criteria.

This experiment suggests that the minimization algorithm initially builds up the information in the observed regions before inferring the state in the unobserved regions. This behaviour suggests that the effect can be related to the conditioning of the Hessian matrix of the cost function. The minimization algorithm will initially take steps in directions such that the gradient is large. These directions correspond to eigenvectors of the Hessian, with large eigenvalues. This is especially the case for the first few iterations as the Hessian matrix is initially approximated by the identity matrix. Then the minimization algorithm will take steps in directions which correspond to the eigenvectors with small eigenvalues, before reaching the minimum. This argument is qualitative. However, it suggests that there is a relationship between the reconstruction of the upper boundary and the eigenvalues of the Hessian matrix. This relationship will be made precise in the next chapter.

3.2.2 Effect of the Background State

With diagonal covariance matrices, the cost function can be considered to contain two weights: σ_b^{-2} and σ_o^{-2}. If σ_o^{-2} is relatively large, then a large weight is given to the observations but if σ_b^{-2} is relatively large, then a large weight is given to the background state. The impact of different specifications of σ_b^{-2} on the reconstruction of the upper level wave is now examined.

Figures 3.2(a)-(b) show a 4D-Var analysis at the end of the assimilation window where a relatively small weight is given to the background state. The 4D-Var algorithm has propagated the information from the observations (circles) of the lower level wave with the model dynamics to move the unobserved upper level wave from the background state (dashed) to the correct position (dotted). It is crucial that the upper level wave is moved so that the forecast is close to the truth. A method such as 3D-Var (with no vertical correlations in the background error covariance matrix) would only move the lower level wave so that in a 3D-Var analysis the buoyancy field would tilt westwards instead of eastwards. This would produce a forecast with decay instead of growth. This simple experiment therefore illustrates one advantage of 4D-Var in comparison with 3D-Var.

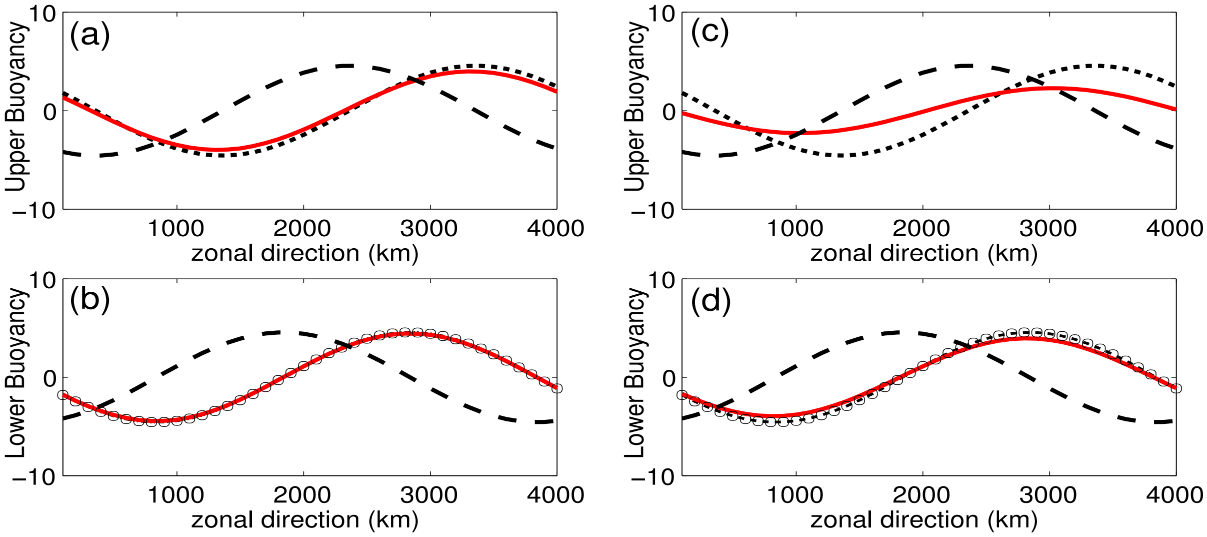


Figure 3.2: *4D-Var analyses shown at the final time of a 6 hour assimilation window, with (a)-(b) $\sigma_b^{-2} = 0.01$ and (c)-(d) $\sigma_b^{-2} = 0.1$. In both cases, perfect observations, shown by the circles, are given at the beginning and the end of the window, and are given the weight $\sigma_o^{-2} = 1$. The true state, shown by the dotted lines, is a growing Eady wave and the background state, shown by the dashed lines, has a displacement error. The upper panels show the buoyancy on the upper boundary and the lower panels show the buoyancy on the lower boundary.*

We now consider the same experiment, but with more weight given to the background state. The analysis is shown in Fig. 3.2(c)-(d). The 4D-Var algorithm has moved the upper level wave closer to the true state, but now the analysis has a phase error and an amplitude error. As the upper boundary wave is now in the wrong position and has a smaller amplitude than the true wave, the analysis gives a smaller growth rate than the true solution. To compensate for this, the amplitude of the lower boundary wave is too large at T+0 and too small at T+6.

Thus, the effect of the background state is to penalize the information propagated to the unobserved regions. Although 4D-Var is still able to reconstruct the unobserved part of the flow when the J^b term is included, if too much weight is given to the background state, these reconstructive properties are lost. Many of the experiments in previous literature did not include the J^b term in the cost function. Thus, we can perhaps infer that if a background term were included in the previous studies in the literature, the results may not have shown 4D-Var to have been so advantageous.

3.2.3 Effect of Noise on the Observations and Background Error Correlations

The effect of adding noise to the observations is now investigated. When the observations are noisy, the 4D-Var algorithm must filter the information from the observations, as well as reconstructing the upper level wave. The background error correlations play a key role in the filtering of the observational noise. Therefore, the effect of smoothing as applied through the background error correlations is also examined in this section.

The following experiments are identical to those in the previous section, except that random noise is added to the observations and correlations are incorporated into the background error covariance matrix. The random noise has a Gaussian distribution with a standard deviation of one. The background error covariance with correlations is defined as:

$$\mathbf{B}^{-1} = \frac{\sigma_b^{-2}}{l}\left(\mathbf{I} + \frac{l^4}{2}(\mathbf{L}_{xx})^2\right) \text{ for Buoyancy on } \hat{z} = -\frac{1}{2} \tag{3.6}$$

and $\mathbf{B}^{-1}$ is as before elsewhere. $\mathbf{L}_{xx}$ is a finite difference second derivative matrix in the x direction and l is the correlation length scale. This form of correlation matrix was described in detail in Chapter 2.

We first examine the effect of background error correlations with noisy observations. An analysis with noisy observations but no correlations is shown in Fig. 3.3 (a)-(b) with weights $\sigma_b^{-2} = 0.04$ and $\sigma_o^{-2} = 1$. These weights are chosen to clearly illustrate the effect of the

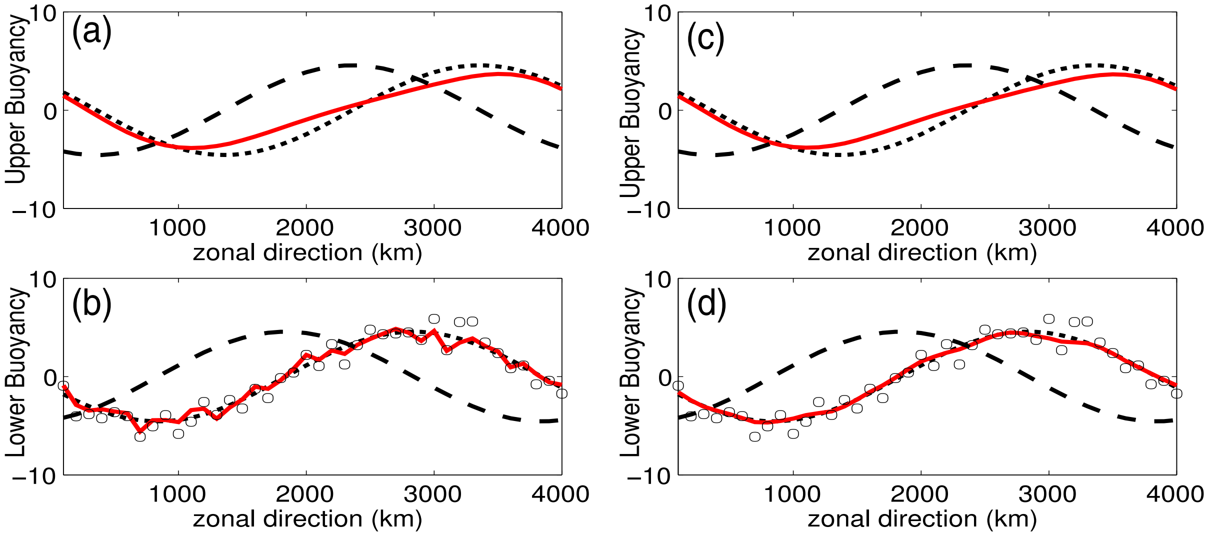


Figure 3.3: *As for Fig. 3.2 but the observations now have noise added with a Gaussian distribution of standard deviation $\sigma = 1$. The assumed background state errors have (a)-(b) no correlations and (c)-(d) correlations with length scale $l = 10\Delta x$ on the lower boundary only. In both cases, the weight given to the observations is $\sigma_o^{-2} = 1$ and the weight given to the background state is $\sigma_b^{-2} = 0.04$.*

correlations. In previous experiments, a smooth analysis could be obtained from the perfect observations. However, with noisy observations the analysis is now noisy. The same experiment but with correlations included is shown in Fig. 3.3 (c)-(d). The length scale used is $l = 10\Delta x = 1000km$. The analysis is very similar to that without correlations, except that the lower boundary is now smooth. This simple comparison has therefore shown that the correlation model, developed in the previous chapter, provides the smoothing that is needed when the observations are noisy. Thus, it is vital that background error correlations are included when noisy observations are used.

In Section 3.2.2, it was shown that if the weight σ_b^{-2} given to the background state is too large, 4D-Var is unable to reconstruct the upper level wave correctly. These experiments are now repeated but with both noisy observations and correlations on the lower boundary.

Variational data assimilation is based on minimum variance estimates and Bayesian probabilistic arguments. From these statistical derivations, we know that the a priori weights should reflect the assumed size of the errors of the background state and the observations. More specifically, the weight σ_b^{-2} should be the the reciprocal of the background state error variance, and similarly for the observation errors.

The error variance for the observations is trivial to estimate, as the errors added to the observations have a Gaussian distribution with a standard deviation of one. In this case, the statistically correct weight given to the observations is $\sigma_o^{-2} = 1$. The error variance σ_b^2 for the background state is not so obvious. The background state errors are not taken from a Gaussian distribution as we assume that the background state has a phase error. However, it is still possible to calculate a 'globally averaged' variance of the lower boundary errors using:

$$\text{variance, } \sigma_b^2 = \frac{1}{40}(\mathbf{x}^b - \mathbf{x}^t)^T(\mathbf{x}^b - \mathbf{x}^t) \tag{3.7}$$

where $\mathbf{x}^b$ and $\mathbf{x}^t$ are the background state and true state lower buoyancy fields with 40 grid points. This assumes that the error characteristics of the background state are the same for all grid points. This is a reasonable assumption since we can imagine that the wave will be advected over all the grid points. For these experiments, where $\mathbf{x}^b$ has a phase error, this gives a variance of 12.48 and hence the statistically correct weight given to the background state is $\sigma_b^{-2} = 0.08$.

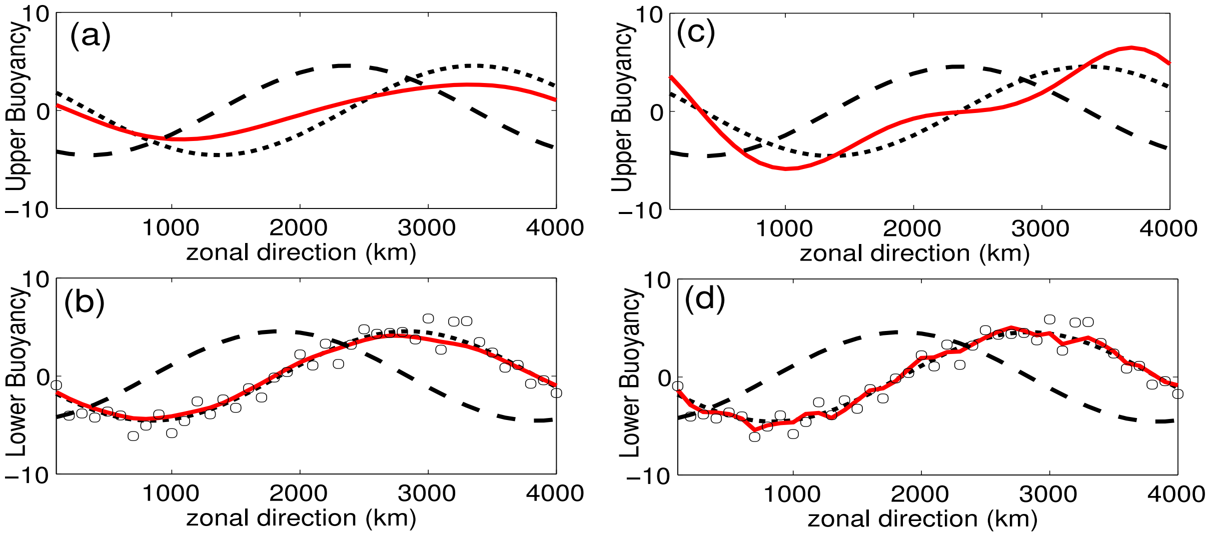


Figure 3.4: *As for Fig. 3.3 (c)-(d), but with (a)-(b) $\sigma_b^{-2} = 0.08$ and (c)-(d)$\sigma_b^{-2} = 0.01$. In both cases, there are correlations on the lower boundary with $l = 10\Delta x$, the noise has a standard deviation of $\sigma = 1$, and $\sigma_o^{-2} = 1$.*

The analysis using the statistically correct weights is shown in Fig. 3.4(a)-(b). The upper boundary wave has been moved closer to the true position, however there is still an amplitude error. This is due to the effect of the weight given to the background state. The experiments in Section 3.2.2, concerning the weight given to the background state, showed that without noise on the observations, the upper level wave can be reconstructed if less weight is given to the background state. However, this is not the case when noise is added to the observations. Figure 3.4(c)-(d) shows the same experiment as Fig. 3.4(a)-(b), but with less weight given to the background state. A non-physical wave has been generated on the upper boundary due to the presence of the noise on the observations.

Fig. 3.5 illustrates further the sensitivity of the unobserved regions to noise on the observations. A small weight is now given to the background state ($\sigma_b^{-2} = 0.004$), creating an unphysical wave on the upper boundary. If the noise on the observations is generated using a different random seed than the noise in Fig. 3.5 (a)-(b), but with the same variance, the upper level wave has a different structure, as shown in Fig. 3.5 (c)-(d). Thus, the unobserved regions are sensitive to the noise on the observations. This result was also found by Courtier and Talagrand (1987) and was briefly discussed by Bennett and Miller (1991) who interpreted the result in terms of the ill-conditioning of the underlying inverse problem.

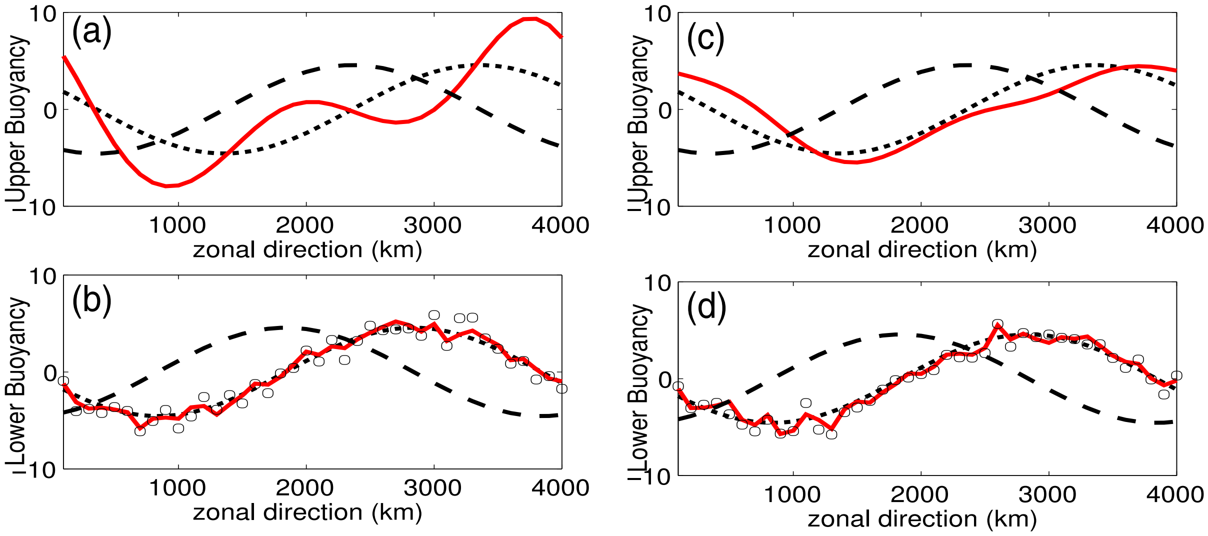


Figure 3.5: *The random noise added to the observations is generated using a different random seed (idum): (a)-(b) idum=-3, (c)-(d) idum=-4. In both cases,* $\sigma_b^{-2} = 0.004$, $\sigma_o^{-2} = 1$, *and the noise has a standard deviation of* $\sigma = 1$. *The details are as for Fig. 3.4.*

In these experiments, we have only considered correlations, or smoothing, on the lower boundary and not the upper boundary. The unobserved upper boundary has been shown to be sensitive to noise on the observations and so it seems sensible to suggest that the unphysical wave on the upper boundary may perhaps be removed by applying correlations also to the upper boundary.

The experiment shown in Fig. 3.4(c)-(d) used correlations on the lower boundary with $l = 10\Delta x$. If this is repeated, but with correlations applied to both the upper and the lower boundary, the results look very similar to those in Fig. 3.4 (c)-(d), and are therefore not shown. Thus, upper level correlations with a length scale of $l = 10\Delta x$ make little difference to the analysis. This is because the unphysical upper level wave has a wavelength which is long in comparison with the correlation length scale.

We now consider the effect of adding correlations with a longer correlation length scale. The experiment shown in Fig. 3.4(c)-(d) is repeated but with correlations using $l = 20\Delta x = 2000km$. When the correlations are only applied to the lower boundary, as in Fig. 3.6 (a)-(b), the lower boundary wave is much more smooth, but an unphysical wave has still been generated on the upper boundary. When correlations are applied to both the upper and the lower boundary, as in Fig. 3.6 (c)-(d), both the upper and the lower level wave are smooth and are close to the truth.

Thus, it is possible to create a smooth wave by increasing the length scale of the correlations. However, the length scale in these last experiments is extremely long, and would be unrealistic for operational data assimilation, where it is important to be able to resolve smaller scale structures. Thus, we may conclude that it is important to give enough weight to the background state, so that unphysical solutions are not generated in the unobserved regions.

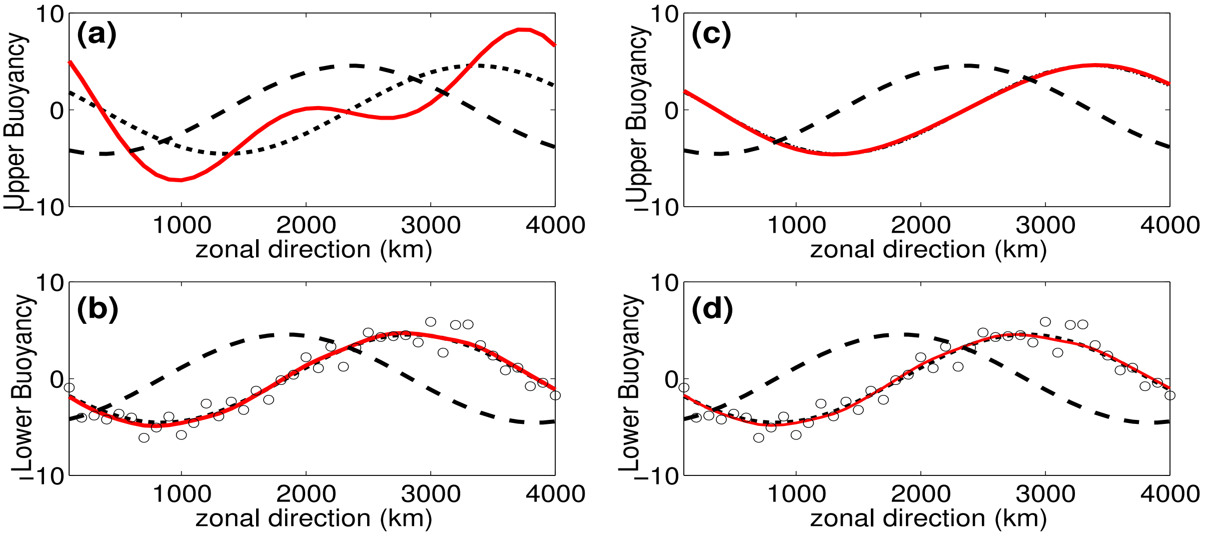


Figure 3.6: *As for Fig. 3.4, with (a)-(b) correlations are only applied to the lower boundary wave, (c)-(d), correlations are applied to both the lower boundary and the upper boundary. In both cases, observations have noise with a Gaussian distribution of standard deviation $\sigma = 1$. The weights given to the observations and the background state are $\sigma_o^{-2} = 1$ and $\sigma_b^{-2} = 0.01$. The correlations have a length scale of $l = 20\triangle x$.*

3.2.4 Signal-to-Noise Ratio

To summarize, the experiments in this section have demonstrated that it is important to specify the weights σ_b^{-2} and σ_o^{-2} correctly. That is, σ_b^2 should represent the size of the background state error variance and σ_o^2 should represent the size of the observation error variance so that the maximum amount of information may be extracted from the observations, without contaminating the analysis with noise. It will be shown in a subsequent chapter that in fact, these weights may be computed without knowing the true state.

The experiments with perfect observations showed that the background state strongly penalizes the information needed to reconstruct the state in unobserved regions. The experiments with noisy observations showed that the background state is needed to penalize the generation of unphysical waves in the unobserved regions. This can be summarized by considering what is known as the signal-to-noise ratio.

The signal-to-noise ratio is defined as the ratio of the magnitude of the signal to that of the interference or noise. The signal may be for example, an electrical current, radio wave, or a light ray. In our case, the signal is the true state, and the noise is the errors on the observations. The purpose of data assimilation is to extract the signal $\mathbf{x}^t$ from the noisy observations $\mathbf{y}$. The signal-to-noise ratio gives an indication of how much of the signal can be extracted from the noisy observations, or equivalently, how much weight should be given to the observations relative to the background state (Eyre, 2000). If an observation has a large error in comparison with the background state, then we would wish to give the observation a small weight. But, if an observation has a relatively small error, then we would wish to give a large weight to the observation. Thus, the ratio $\mu = \frac{\sigma_b}{\sigma_o}$ gives an indication of the signal-to-noise ratio.

If the specified value μ is too large and too much weight is given to the observations, then the analysis will be noisy. However, if the specified value of μ is too small and too much weight is given to the background state, the maximum amount of available information in the observations will not be extracted. Thus, the specification of the relative weight given to the observations, μ, is critical in extracting the maximum amount of useful information from the observations.

3.3 Vertical Structure and Growth Rates

The results in the previous section showed that 4D-Var is able to use a time-sequence of observations on the lower boundary to reconstruct the unobserved upper level wave. However, we did not consider the impact on the vertical structure of the system and the growth rate of the subsequent forecast. This section investigates the ability of 4D-Var to generate the correct vertical structures needed for baroclinic growth and decay. We first examine the effect of the background state when a decaying mode is observed, by repeating the experiments in

Section 3.2.2, but with the true state given by the most rapidly decaying Eady wave. The study is then extended by investigating the effect of the assimilation window length, the temporal position of the observations in the assimilation window and the temporal weights given to the observations.

3.3.1 Analysis of Decaying Modes

Many of the studies in previous literature have shown 4D-Var to perform well in cases of baroclinic growth, and the experiments in the previous section have also shown that 4D-Var is able to move the upper level wave so that the growing Eady wave has the correct vertical structure. However, there has been very little research on the behaviour of 4D-Var in the presence of baroclinic decay. It is important to understand how 4D-Var behaves when the true state is decaying, to fully assess the advantages and disadvantages of the data assimilation algorithm.

The experiments in the previous section are now extended to compare the behaviour of 4D-Var when the true state is given by either a decaying mode or a growing mode. As in Section 3.2.2, perfect observations of the lower level buoyancy are given at T+0 and T+6 and the background state has a phase error of $1000km$. When a growing mode is observed, the analyses are exactly the same as those in Section 3.2.2.

If the analysis has the correct vertical structure, the forecast will have the correct growth rate. Therefore, we choose to measure the analysis accuracy by examining the Euclidean norm of the streamfunction during the following forecast.

Fig. 3.7 gives the evolution of the norm of the streamfunction during the 6 hour assimilation window and also the following 30 hour forecast.

Fig. 3.7 (a) gives the norm evolution when a growing mode is observed. If σ_b^{-2} is small, the forecast from the analysis (dotted line) is close to the truth (solid line). If σ_b^{-2} is large, the forecast from the analysis (dashed line) has a smaller growth rate. The forecast has the wrong growth rate because the upper level wave in the analysis has the wrong position and amplitude, as shown in Fig. 3.2(c), in the previous section.

Fig. 3.7 (b) shows the norm evolution when a decaying mode is observed. If σ_b^{-2} is small, the forecast from the analysis (dotted line) is again close to the truth (solid line). However, if σ_b^{-2} is large, the forecast from the analysis grows instead of decaying. Thus, the background state has a greater detrimental effect when a decaying mode is observed.

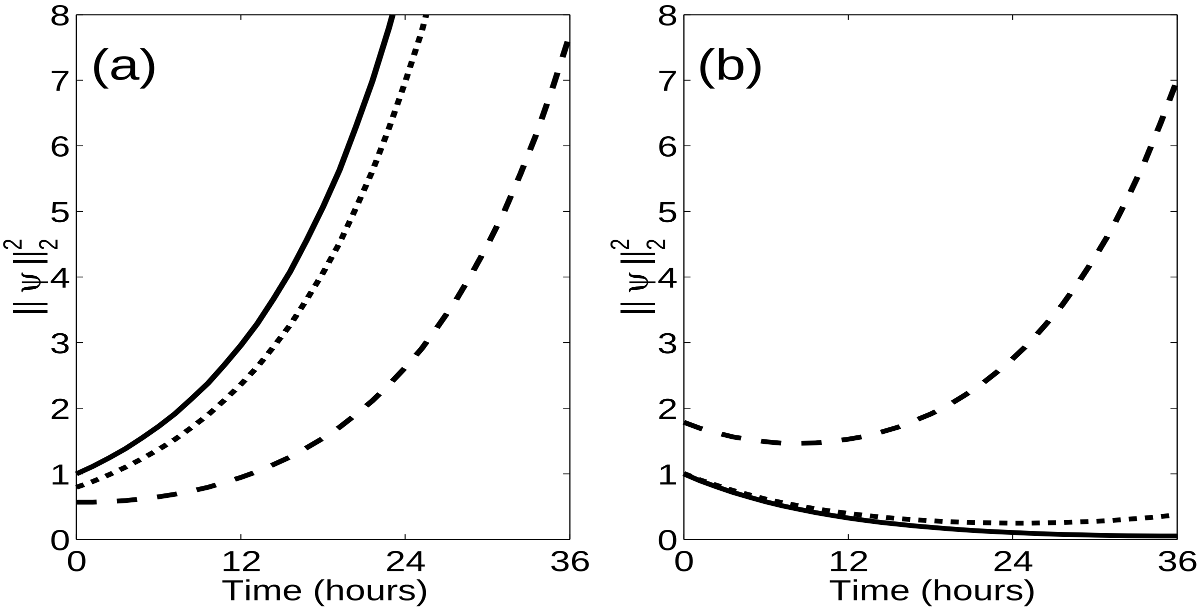


Figure 3.7: *The squared Euclidean norm of the streamfunction for forecasts from 4D-Var analyses of (a) Growing and (b) Decaying modes. In both cases, perfect observations are given at T+0 and T+6, with weights* $\sigma_o^{-2} = 1$. *The solid lines show the forecasts from the true state, the dashed lines show the forecasts from analyses with* $\sigma_b^{-2} = 0.1$, *and the dotted lines show the forecasts from analyses with* $\sigma_b^{-2} = 0.01$.

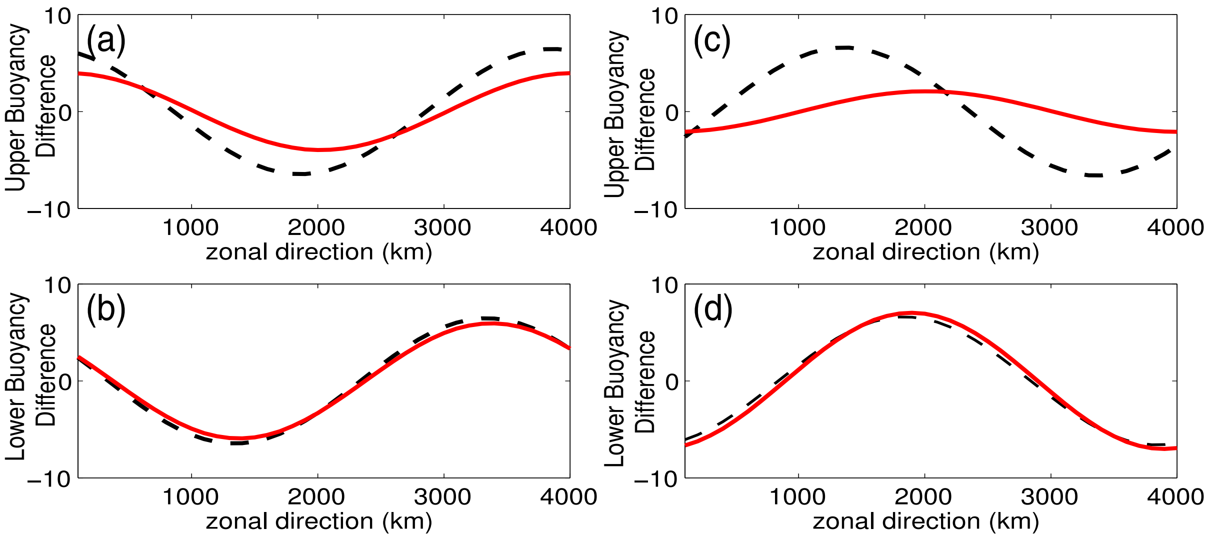


Figure 3.8: *The true analysis increment* ($\mathbf{x}_t - \mathbf{x}_b$) *(black, dashed) and the actual analysis increment* ($\mathbf{x}_a - \mathbf{x}_b$) *(red, solid) from the analysis of (a)-(b) Growing and (c)-(d) Decaying modes. In both cases, perfect observations are given at T+0 and T+6 with weights* $\sigma_o^{-2} = 1$, $\sigma_b^{-2} = 0.1$.

To compare the analysis of a decaying mode with the analysis of a growing mode, it is useful to examine the actual analysis increments ($\mathbf{x}^a - \mathbf{x}^b$) compared with the required analysis increments ($\mathbf{x}^t - \mathbf{x}^b$), as shown in Fig. 3.8. If the true solution is the growing mode, the true buoyancy field increment tilts eastwards, and if the true solution is the decaying mode then the true buoyancy field increment tilts westwards. When a growing mode is observed, the actual analysis increment is close to the required analysis increment. However, when a decaying mode is observed, the analysis increments are very different. The required increment tilts westwards, yet the actual increment tilts eastwards. So, a growing analysis increment has been added to the background state instead of a decaying analysis increment. This produces a forecast with growth instead of decay.

Again, the signal-to noise ratio $\mu = \frac{\sigma_b}{\sigma_o}$, is an important aspect in obtaining structures with the correct growth rates. This can be understood further by considering the schematic diagram in Fig. 3.9. When the observations have large errors, as shown in Fig. 3.9(a), there is a large difference between the possible growth rates of the analysis. At one extreme, the state decays and at the other extreme, the state grows. When the observations have small errors, as shown in Fig. 3.9 (b), both possible analysis extremes are growing and there is a small difference between them. Thus, it is easier to infer the growth rate when there is a small amount of noise on the observations. In the case of the Eady model, if the growth rate can be inferred, then the upper boundary wave can be corrected so that the analysis gives the correct growth during the assimilation window.

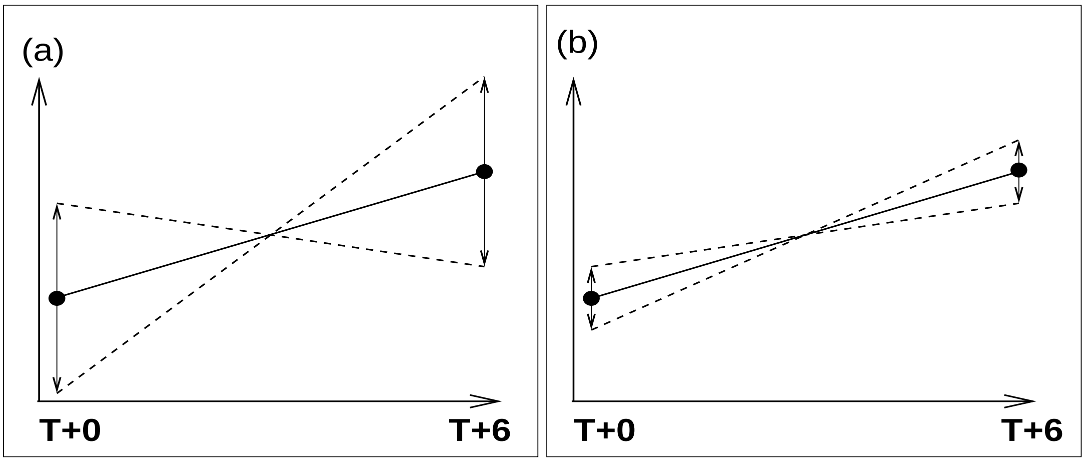


Figure 3.9: *Schematic diagram illustrating the effect of noise on the observations. The solid line represents the growth of the true state and there are observations (circles) at T+0 and T+6. The arrows represent the error bars on the observations, and the dashed lines show the forecasts from possible analyses with the most extreme growth rates. The observations have (a) relatively large errors, and (b) relatively small errors.*

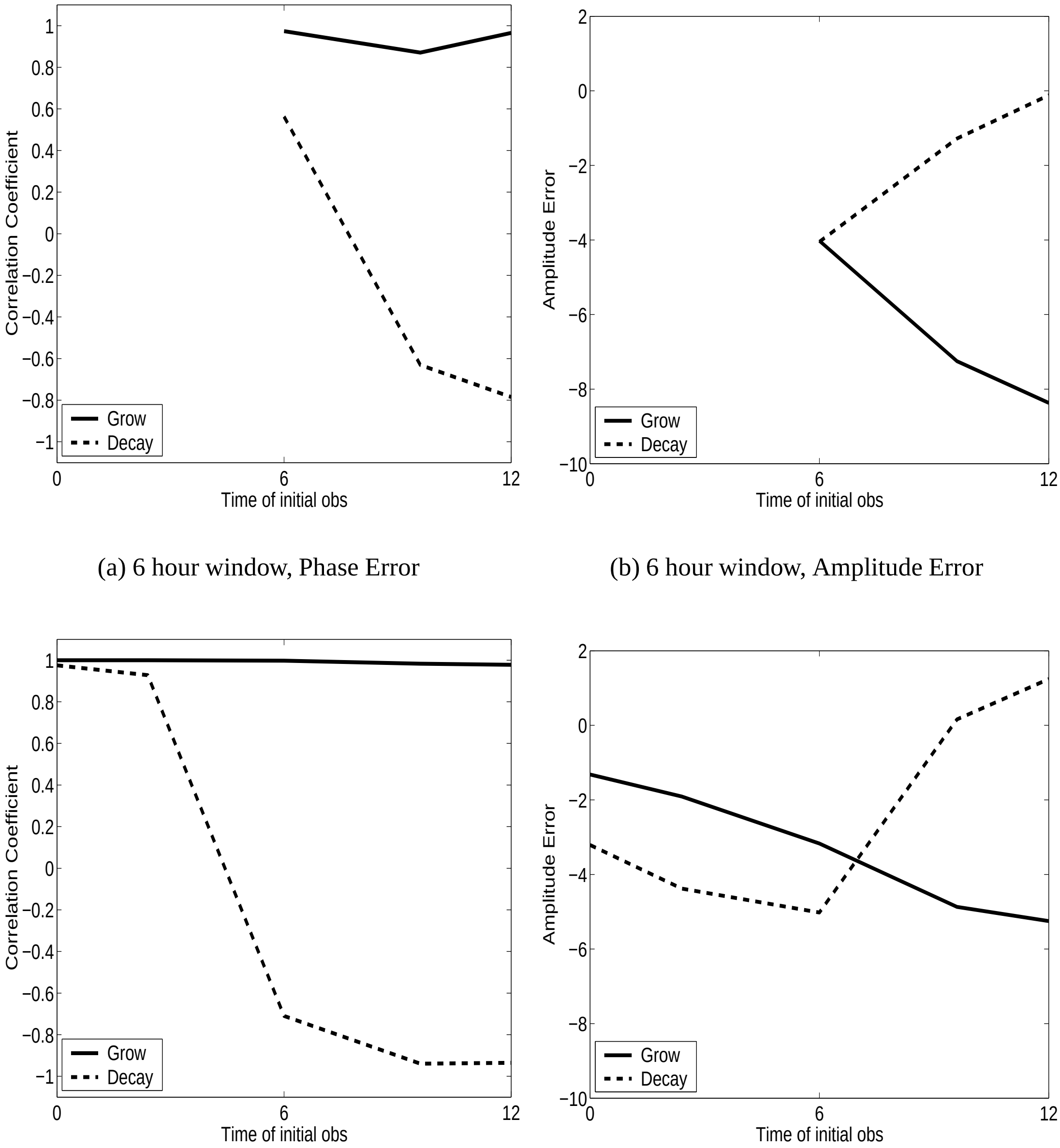
Thus, 4D-Var is able to generate the vertical structure needed for baroclinic growth, but is not always able to generate the structure needed for baroclinic decay. If the observations have large errors, then a relatively large weight must be given to the background state. This penalizes the decaying part of the analysis increment, so that in fact, a growing analysis increment may be added instead of a decaying analysis increment.

3.3.2 Temporal Position of the Observations

We now examine where the observations should be placed in the assimilation time window, so that 4D-Var can generate the best vertical structures. The equivalence between the Kalman Filter and 4D-Var means that 4D-Var implicitly evolves the background error covariance matrix through the time window. It is difficult to specify the covariance at the beginning of the window, and so it is often approximated using isotropic correlation functions which are not flow dependent. As the covariance matrix evolves through the window, it becomes flow dependent. Hence, it would seem that it is better to place the observations near to the end of the assimilation window, so that the information from the observations is spread to the surrounding grid points in a flow-dependent way. The single observation experiments by Thépaut et al. (1996) showed that it is indeed important to have as long an assimilation window as possible, to ensure that the baroclinic structures are fully developed. High-resolution experiments with the ECMWF 4D-Var system have also shown a consistent improvement in analyses from a 12-hour assimilation window compared with analyses from a 6-hour assimilation window, (Bouttier, 2001).

With this in mind, we now investigate the impact of the length of the assimilation window and the temporal position of the initial observations. The important differences between the experiments in this section and those of Thépaut et al. (1996) and Bouttier (2001) is that in these experiments we consider sets of observations at two time levels rather than at one time level, and the true state is known.

We consider a series of 4D-Var experiments with different temporal positions of the observations, and different assimilation window lengths. The experiments use either a 6 hour or 12 hour assimilation window. For a fair comparison between the 6 hour and 12 hour windows, the true state is first evolved for 6 hours if a 6 hour window is used. The true state is given by either the most rapidly growing or decaying Eady wave and the background state has a phase error of 1000 km. Observations are provided of the lower level buoyancy at two time levels.



(a) 6 hour window, Phase Error

(b) 6 hour window, Amplitude Error

(c) 12 hour window, Phase Error

(d) 12 hour window, Amplitude Error

Figure 3.10: *Correlation coefficients C and Amplitude errors A, plotted against the time of the initial observations (hours). The experiments use either a 6 or 12 hour window, with two sets of observations. The initial observations are either at the beginning of the window or somewhere in the middle of the window. The final observations are always at the end of the window. The observations are of the lower level buoyancy wave, and have noise with a Gaussian distribution, with a standard deviation of $\sigma = 0.1$. The weights are $\sigma_o^{-2} = 1$, $\sigma_b^{-2} = 0.04$ and smoothing is applied to the lower boundary through the background error correlations. The experiments with the initial observations at the end of the window, only have one set of observations at the end of the window. The solid lines represent the experiments where the true state is given by the most rapidly growing Eady wave and the dashed lines represent the experiments where the true state is given by the most rapidly decaying Eady wave.*

The final set of observations are always at the end of the window and the initial set are either at the beginning of the window or at a specified time in the middle of the window. Random noise with a Gaussian distribution with $\sigma = 0.1$ is added to the observations and smoothing is applied to the lower level through the $\mathbf{B}$ matrix. The weights given to the observations are $\sigma_o^{-2} = 1$ and $\sigma_b^{-2} = 0.04$.

Note that the weight σ_o^{-2} given to the observations is smaller than the statistically optimal value. This is to ensure that the analysis is not noisy. Also note that when the initial observations are at the end of the window, there is only one set of observations and the weight given to these is not doubled.

The correct position of the upper boundary wave is vital in determining whether the forecast from the analysis will grow or decay. Therefore, to assess the analysis accuracy, we choose to separate the phase and amplitude errors. The phase error is measured using the correlation coefficient, C, as discussed by Lawless (2001).

The correlation coefficient, C , is defined as:

$$C = \frac{cov(\mathbf{x}^a, \mathbf{x}^t)}{\sigma(\mathbf{x}^a)\sigma(\mathbf{x}^t)} \tag{3.8}$$

where $-1 \leq C \leq 1$. If $C = 1$, the analysed wave has the correct phase and if $C = -1$, the analysed wave is completely out of phase with the true wave.

The amplitude error A, is defined as:

$$A = (max(\mathbf{x}^a) - min(\mathbf{x}^a)) - (max(\mathbf{x}^t) - min(\mathbf{x}^t)). \tag{3.9}$$

If $A = 0$, the analysed wave has the correct amplitude, if $A > 0$, the analysed wave has an amplitude that is too large and if $A < 0$ the amplitude is too small.

Figure 3.10 shows a summary of the experiments with different window lengths, temporal position of the initial observations and growing and decaying true states. For example, Fig. 3.10 (a) and (b) show the correlation coefficient C and amplitude error A of the analysed upper level wave for assimilation experiments using a 6 hour window. The final observations are at T+12 and the initial observations are at T+6 , T+9 or T+12. The analysis window is defined from T+6 to T+12.

These plots show that when the growing mode is observed (solid lines), the phase error is always small, and the amplitude error increases as the initial observations are moved to the end of the window. Thus, for the growing mode, the analysed upper level wave is always in the correct position and the amplitude is the best when the initial observations are at the beginning of the window.

Figure 3.10 (a) and (b) also show that for the decaying mode (dashed lines), the phase error increases as the initial observations are moved to the end of the window. However, the amplitude error decreases. Thus, for the decaying mode, with observations at the beginning and the end of the window, the phase of the upper level wave is good although the amplitude is poor. When the observations are only at the end of the window, the analysed upper level wave has the correct amplitude. However it is completely out of phase with the true wave. This is because a growing analysis increment has been added to the background state instead of a decaying analysis increment, as discussed in Section 3.3.1.

These results are now compared with those of the 12 hour window experiments, by examining Fig. 3.10 (c) and (d). With observations at the beginning and the end of the 12 hour window, both the phase error and the amplitude errors are smaller than the errors when observations are at the beginning and the end of a 6 hour window. Thus, it seems that a longer window does give better results. However, this is not always the case when the observations are only at the end of the window.

For a growing mode, with observations at only the end of the window, the experiments with a 12 hour window give better results than those with a 6 hour window. In both cases, the upper level wave has the correct phase, but the amplitude is better for the 12 hour window than the 6 hour window. For a decaying mode, with observations at only the end of the window, the experiments with a 12 hour window give worse results than those with a 6 hour window. The phase error is larger and the wave now has too large an amplitude. That is, a large analysis increment has been added to the upper level, but it is in completely the wrong position.

To summarize, these experiments have shown that it is best to have the observations as far apart as possible in time. This means that the observations should be placed at the beginning and the end of a long assimilation window (within the validity of the tangent linear model). They have also shown that with observations at only the end of the window, a growing analysis increment is added to the background state. This gives the correct vertical structure if a growing mode is observed, but not if a decaying mode is observed. In other words, if observations are only at the end of the window, a longer assimilation window will give better results if a growing mode is observed, but worse results if a decaying mode is observed.

To illustrate further the difference between the analysis of growing and decaying modes,

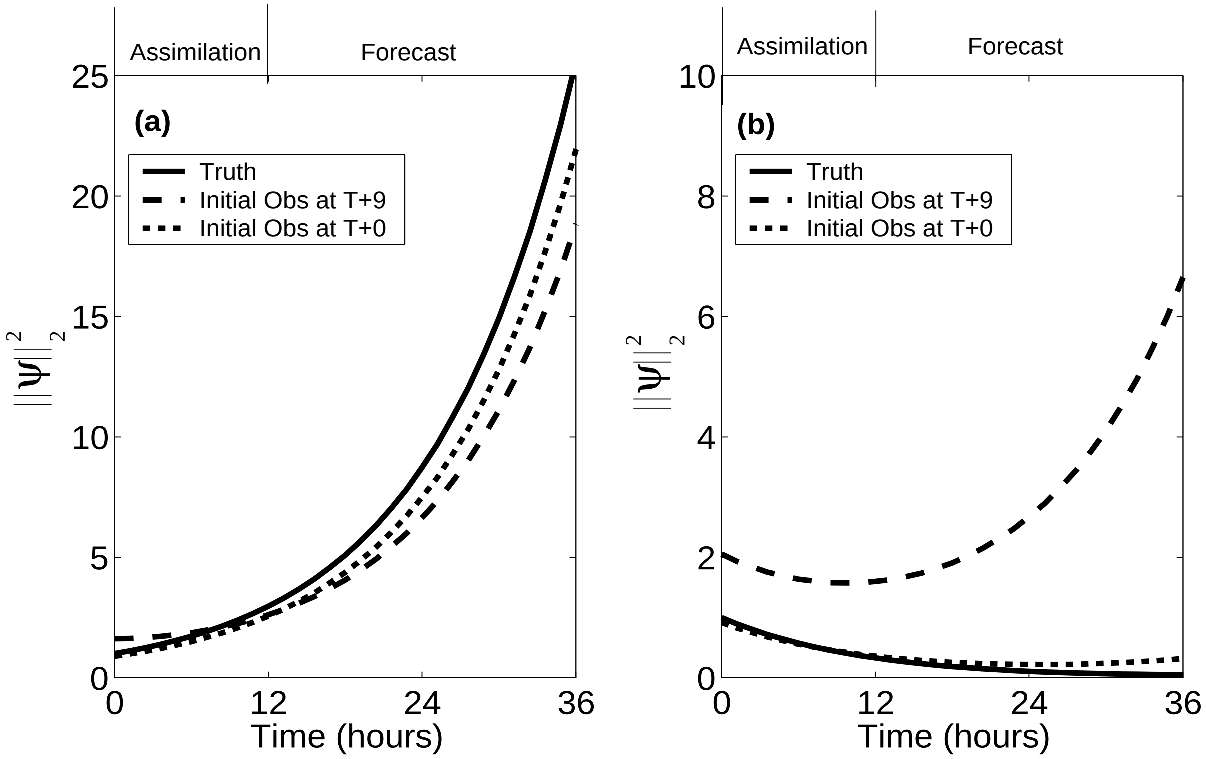


Figure 3.11: *The streamfunction norm for forecasts from 4D Var analyses of (a) Growing and (b) Decaying modes. Observations are given at either T+0 (dotted) or T+9 (dashed) and also at the end of a 12 hour window. The observations have noise with a standard deviation of* $\sigma = 0.1$*, and the weights are* $\sigma_o^{-2} = 1$*,* $\sigma_b^{-2} = 0.04$*. Smoothing is applied to the lower boundary.*

and the importance of specifying the observations as far apart as possible in time, we examine the growth rate of the following forecast, as measured by the streamfunction norm. We consider the experiments with a 12 hour assimilation window, and compare analyses with the initial observations at T+0 with analyses with the initial observations at T+9. The streamfunction norm during the 12 hour assimilation window and the following 24 hour forecast is shown in Fig. 3.11.

When the growing mode is observed (Fig. 3.11 (a)), the forecast from the analysis is closer to the true state when the initial observations are at the beginning of the window (T+0) rather than near to the end of the window (T+9). When the decaying mode is observed (Fig. 3.11 (b)), the forecast from the analysis is close the true state when the initial observations are at the beginning of the window. However, when the initial observations are near to the end of the window, a growing analysis increment is added to the background state instead of a decaying analysis increment. This produces a forecast with growth instead of decay. Thus, to give the correct decaying analysis increment, it is vital that the observations are at the beginning and the end of the window rather than both near to the end of the window.

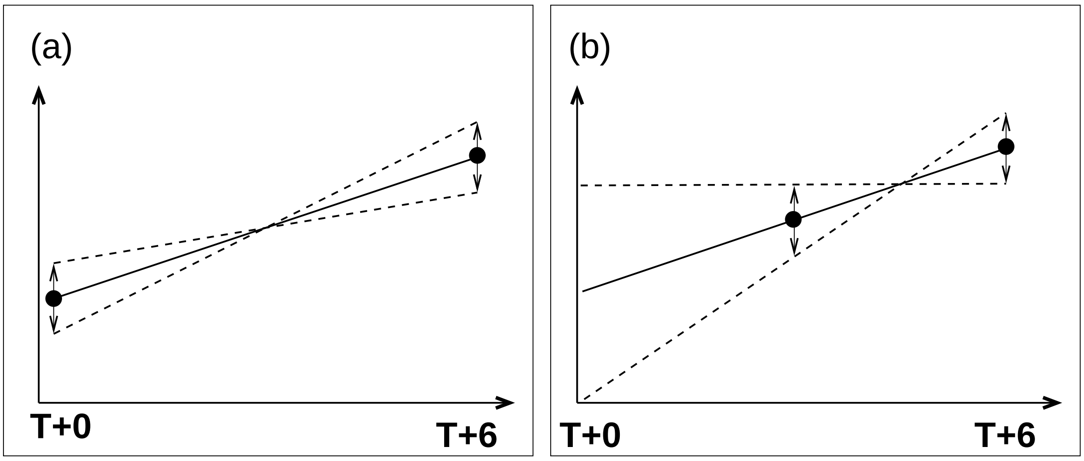


Figure 3.12: *Schematic diagram illustrating the effect of the temporal position of the observations. The solid line represents the growth of the true state and there are observations (circles) at two time levels. The arrows represent the error bars on the observations, and the dashed lines show the possible forecasts from analyses with the most extreme growth rates. The observations are at (a) T+0 and T+6, and (b) T+3 and T+6.*

This can be understood by considering the schematic diagram in Fig. 3.12. The diagram illustrates a one-variable problem with an observation (circles) at two times with associated errors as shown by the error bars. When the observations are far apart in time, the solution has grown a large amount during the time between the observations, relative to the noise on the observations. Therefore, there is a small difference between the possible extreme analyses. When the observations are close together in time, the solution has only grown a small amount relative to the noise on the observations. Therefore, it is difficult to infer the growth rate accurately. That is, there is a large difference between the possible extreme analyses. At one extreme, the forecast from the analysis will not grow, and at the other extreme, the forecast from analysis will grow rapidly. This diagram shows clearly that it is important that the observations are far apart in time, so that the growth rate can be inferred accurately. In the context of the Eady model, the correct vertical structure of the analysis can only be obtained if the growth rate is correct.

3.3.3 Temporal Weights Given to the Observations

The situation where the temporal position of the observations is fixed but the weight given to the observations can be varied is now addressed. It can be considered that twice the number of observations is equivalent to doubling the weight given to the observations. This can be justified by considering a super-observation (Lorenc, 1981) that is a linear combination of two observations that are close together. Given that there is typically twice as much radiosonde data at 00Z and 12Z than at 06Z and 18Z, the issue concerning the weight given to the observations is important in defining a 4D-Var assimilation window.

The previous experiments showed that the analysis of a decaying wave is worse when the observations are provided only near to the end of the window than when observations are near to the beginning of the window. Thus, it might be expected that it may be better to give more weight to the initial observations that the final observations.

The following experiments use a 12 hour window, with observations at T+0 and T+12. The true state is given by either the most rapidly growing or decaying Eady wave and the background state has a phase error as in previous experiments. The weight given to the background state is chosen as $\sigma_b^{-2} = 2$ and the weight given to the observations are chosen as either $\sigma_o^{-2} = 1$ given to the initial observations and $\sigma_o^{-2} = 20$ given to the final observations or $\sigma_o^{-2} = 20$ given to the initial observations and $\sigma_o^{-2} = 1$ given to the final observations. These weights are chosen so that the background state is given a relatively large weight and so that there is a large difference between the weight given to the initial and final observations.

To assess the performance of 4D-Var in the different cases, the streamfunction norm of the forecast from the analysis is examined. The case where the true state is given by the most rapidly growing Eady wave is shown in Fig. 3.13(a). It can be seen that a better forecast is achieved when more weight is given to the final time observations rather than giving more weight to the initial time observations. When a growing wave is observed, if a large weight is given to the initial observations, the analysis is required to be close to the initial observations but not to the final observations. Hence, a large analysis increment is added but it does not need to grow. If a large weight is given to the final observations, a small analysis increment can be added to the background state so that it grows to fit the final time observations. Therefore, the analyses are better if more weight is given to the final observations.

The case where the true state is given by the most rapidly decaying Eady wave is shown in Fig. 3.13(b). It can be seen that it is also the case that the analysis is better when more weight is given to the final time observations. This is opposite to the expected result and the reasons for this are not clear. The state at the final time has a smaller amplitude than at the initial time as the state is decaying. Therefore, it is perhaps the case that an analysis increment with a smaller amplitude is added when more weight is given to the final time observations. This will be answered more fully in a subsequent chapter (Section 5.6).

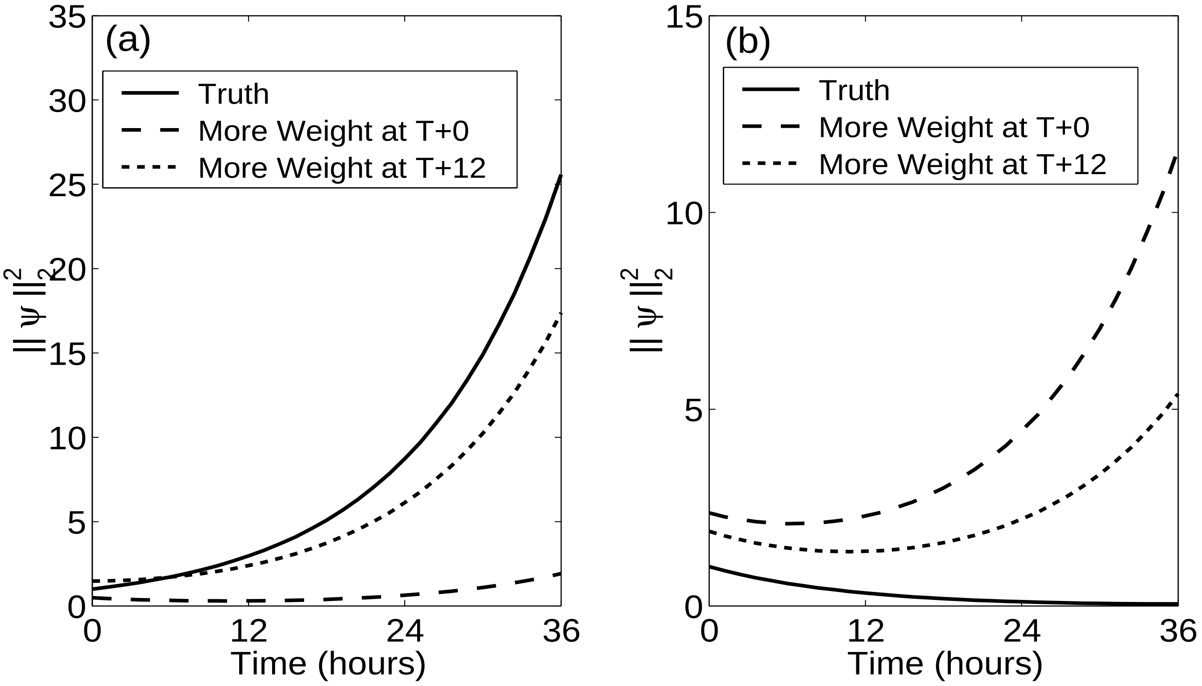


Figure 3.13: *The streamfunction norm for forecasts from 4D Var analyses of (a) Growing and (b) Decaying modes. The background state has a phase error and perfect observations of the lower boundary wave are given at T+0 and T+12. The weights are specified as* $\sigma_b^{-2} = 2$*, and* $\sigma_o^{-2} = 20$ *at T+0 and* $\sigma_o^{-2} = 1$ *at T+12 (dashed), and* $\sigma_o^{-2} = 1$ *at T+0 and* $\sigma_o^{-2} = 20$ *at T+12 (dotted). The solid line represents the true state.*

3.4 Conclusions

This chapter has focussed on two properties of 4D-Var: inferring the state in unobserved regions and generating the necessary vertical structures. These properties were investigated using simple experiments where the lower level wave was observed and 4D-Var was used to infer the position of the upper level wave. The main results are now summarized.

The behaviour of the minimization algorithm showed that the observed regions are first corrected, and then the unobserved regions are corrected. This hints that the information needed to reconstruct the unobserved regions corresponds to eigenvectors of the Hessian matrix with small eigenvalues.

The experiments with perfect observations showed that the background state strongly penalizes the information needed to reconstruct the unobserved wave. The experiments with noisy observations showed that the reconstructed regions are sensitive to noise. Correlations may be used to smooth the noise, however, this may not be able to remove the unphysical waves in the unobserved regions if the correlation length scale is too small. Thus, it is important to give a large weight to the background state in the case of noisy observations. Together, the experiments with and without noise show that it is vital to specify the weight given to the observations relative to the weight given to the background state, so that the maximal amount of information may be extracted from the observations.

By comparing the results with growing and decaying true states, it has been shown that 4D-Var is able to generate analysis increments with vertical structures necessary for both baroclinic growth and decay. However, it is sometimes the case that a growing analysis increment is added instead of a decaying analysis increment. This occurs if a large weight is given to the background state or if all the observations are near to the end of the assimilation window. Thus, it is important that the observations are both accurate and as far apart as possible in time, so that 4D-Var is able to analyse the correct growth rate. Giving more weight to the initial observations does not improve the analysis of decaying modes.

The experiments in this chapter have led to a number of interesting results. However, they have not provided a full understanding of how observations are used in 4D-Var. In the next two chapters, we aim to provide a new understanding of the 4D-Var analyses in this chapter, using an approach based on information content concepts, that is commonly used to solve inverse problems such as satellite retrievals.

Chapter 4

Qualitative Information Content of Observations in 4D-Var

Information theory or Communication theory is concerned with what is known as the information content of a message, which is the amount of useful information contained within a message. Information theory was first used by electrical engineers to design better telecommunications systems, but now has a wide variety of applications. In particular, concepts from information theory have been applied to 1D-Var satellite retrieval studies (e.g. Mateer, 1965, Eyre, 1990, Prunet et al., 1998, Rodgers, 2000, Rabier et al., 2002). There are many different methods to evaluate or measure the information content of the observations. For example, in satellite retrieval studies, it is useful to obtain a single number as a quantitative measure of the information content. However, there are many other techniques which are useful in understanding the information content. The singular value decomposition (SVD) is one particular technique that can be used, as first described by Mateer (1965).

The SVD has previously been used to evaluate the information content of observations in 1D-Var retrievals. In this chapter, the method is extended to the temporal dimension to evaluate the information content of observations in 4D-Var. This technique should allow a new understanding of how the information from observations is combined with the model dynamics.

The chapter begins by formulating the 1D-Var/3D-Var data assimilation algorithm as an inverse problem. We then give a review of the SVD and its use in understanding the information content of observations in 3D-Var. The technique is then extended to consider the information content of observations in 4D-Var. The technique involves the right and left singular vectors of what is known as the observability matrix and so we go on to describe how these relate to the singular vectors or optimal perturbations that are more commonly used in meteorology. We also discuss how the SVD can give information about the conditioning of the problem and hence the expected rate of convergence of the minimization. The chapter finishes by describing the computational aspects of the SVD technique.

4.1 Data Assimilation as an Inverse Problem

Consider a simple example where the true state of the atmosphere is represented by a vector $\mathbf{x}^t$ of dimension n, and that m observations are given in the vector $\mathbf{y}$. Suppose that the observations are not of the atmospheric variables, but can be related to them through a set of linear equations so that:

$$\mathbf{y} = \mathbf{H}\mathbf{x}^t + \boldsymbol{\varepsilon}^o. \tag{4.1}$$

where $\boldsymbol{\varepsilon}^o$ is the observational error and $\mathbf{H}$ is the forward model. Then the best estimate $\mathbf{x}^a$ of the true state $\mathbf{x}^t$ must be found, such that a measure of $\mathbf{y} - \mathbf{H}\mathbf{x}^a$ is small. This is known as an inverse problem (Wunsch, 1996) and can be solved by formulating it as a least-squares problem. This finds the state $\mathbf{x}$ which minimizes the cost function:

$$J^o(\mathbf{x}) = \frac{1}{2}(\mathbf{H}\mathbf{x} - \mathbf{y})^T\mathbf{R}^{-1}(\mathbf{H}\mathbf{x} - \mathbf{y}). \tag{4.2}$$

J. Hadamard defined a problem to be well-posed if three requirements are met: a solution exists, the solution is unique and the solution depends continuously on the data. If any one of these three conditions are not satisfied, the problem is ill-posed (see e.g. Kalnay, 2003). In general, equation (4.2) is ill-posed. That is, there are often more unknowns than observations so the solution is non-unique. Further, even if there are as many observations as unknowns, the observations are noisy and therefore the analysis can be extremely sensitive to the noise.

For this reason more information must be provided to make the problem well-posed. For example, in 1D-Var/3D-Var a background term is added to the cost function so that the analysis is given by the state which minimizes the cost function:

$$J(\mathbf{x}) = \frac{1}{2}\left\{(\mathbf{x} - \mathbf{x}^b)^T\mathbf{B}^{-1}(\mathbf{x} - \mathbf{x}^b) + (\mathbf{H}\mathbf{x} - \mathbf{y})^T\mathbf{R}^{-1}(\mathbf{H}\mathbf{x} - \mathbf{y})\right\}. \tag{4.3}$$

Then the analysis is given by the background state in regions where there are no observations and the analysis is smooth in regions of dense noisy observations. In the rest of this chapter, we consider diagonal covariance matrices with constant variances. However, it is possible to extend the technique to include correlations, as will be shown in a subsequent Chapter.

As before, we let $\mathbf{B} = \sigma_b^2\mathbf{I}$ and $\mathbf{R} = \sigma_o^2\mathbf{I}$. Then the cost function becomes:

$$J(\mathbf{x}) = \frac{1}{2}\left\{\sigma_b^{-2}(\mathbf{x}-\mathbf{x}^b)^T(\mathbf{x}-\mathbf{x}^b) + \sigma_o^{-2}(\mathbf{H}\mathbf{x}-\mathbf{y})^T(\mathbf{H}\mathbf{x}-\mathbf{y})\right\}. \tag{4.4}$$

Multiplying the cost function by $2\sigma_o^2$, then, 3D-Var minimizes the cost function:

$$J_2(\mathbf{x}) = \mu^2\left\{(\mathbf{x}-\mathbf{x}^b)^T(\mathbf{x}-\mathbf{x}^b)\right\} + (\mathbf{H}\mathbf{x}-\mathbf{y})^T(\mathbf{H}\mathbf{x}-\mathbf{y}) \tag{4.5}$$

where $\mu^2 = \frac{\sigma_o^2}{\sigma_b^2}$ is a parameter that determines the relative weight given to the background state in comparison to the observations.

4.2 Information Content and the Singular Value Decomposition

The least squares equations have also been used for many years to solve other inverse problems such as deducing unknown constants in dynamic oceanography (Wunsch, 1977) and determining the vertical distribution of ozone in remote sensing (Mateer, 1965). The least squares equations continue to be used to solve the satellite retrieval inverse problem. For example, vertical temperature and humidity profiles are linked to observed radiances through the radiative transfer equation.

It is important to understand the information content of remotely sensed observations. That is, to understand how different observations contribute to a retrieval. By observing more radiances at different wavelengths (channels), the vertical resolution of the profile can be improved. However, there may be a point where adding further observations has a negligible effect on the retrieved profile. Further, with the vast increase in satellite data in the future, it will not be possible to include all the available observations in retrievals. Therefore, it is necessary to select an optimal subset of the observations such that the important information is retained (Rodgers (1996) and Collard (2000)).

In the case of satellite retrievals, there is a complicated relationship between the observed variables and the retrieved variables. However, many techniques have been developed to assess the information content of the observations. The information content determines how many linearly independent pieces of information are contained in a set of observations. This not only depends on the observations, but on the algorithm in which they are used, for example on the radiative transfer model and on the errors in the observations.

One of the methods used to examine the information content is the singular value decomposition (SVD). This is a matrix factorization that can be applied to any matrix, even if it is rectangular, and has many uses such as finding the rank of a matrix, reducing the storage space of a matrix (commonly used in image reconstruction and signal processing), extracting a signal from noisy observations using a truncated SVD and also defining the optimal perturbations which exhibit large finite-time growth, used to generate an ensemble of forecasts in meteorology. In this chapter we develop a technique which uses the SVD of the so-called observability matrix in 4D-Var. But first, we discuss how the SVD may be used to examine the structure and identify the important parts of the observation operator $\mathbf{H}$.

Following Golub and Van Loan (1996) and Strang (1986), the SVD of an $(m \times n)$ matrix $\mathbf{H}$ (m rows and n columns) with rank r, can be written as the product of three matrices:

$$\mathbf{H} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^{\mathbf{T}} \tag{4.6}$$

where $\mathbf{\Lambda}$ is a diagonal $(m \times n)$ matrix, with r positive singular values λ_j on the diagonal. The singular values are ordered such that $\lambda_1 > \lambda_2 > \ldots > \lambda_r > 0$. The m columns $\mathbf{u}_j$ of $\mathbf{U}$ $(m \times m)$ are also known as the left singular vectors (LSVs) and are also the eigenvectors of $\mathbf{H}\mathbf{H}^T$. The n columns $\mathbf{v}_j$ of $\mathbf{V}(n \times n)$ are known as the right singular vectors (RSVs) and are also the eigenvectors of $\mathbf{H}^T\mathbf{H}$.

The SVD can be used to identify the four fundamental subspaces known as the column space, left null space, row space and null space. The first r RSVs form a basis for the row space, whilst the remaining RSVs form a basis for the null space. This distinction is important, as all the vectors in the null space satisfy $\mathbf{H}\mathbf{x} = 0$. Similarly, the first r LSVs form a basis for the column space and the remaining LSVs form a basis for the left null space. These subspaces are illustrated in Fig. 4.1.

Both the RSVs and LSVs form orthonormal bases, so $\mathbf{U}^T\mathbf{U} = \mathbf{I}_m$ and $\mathbf{V}^T\mathbf{V} = \mathbf{I}_n$. Using this with equation (4.6) then we find that each right singular vector is mapped onto the corresponding left singular vector and the magnification is given by the corresponding singular value:

$$\mathbf{H}\mathbf{v}_j = \lambda_j\mathbf{u}_j. \tag{4.7}$$

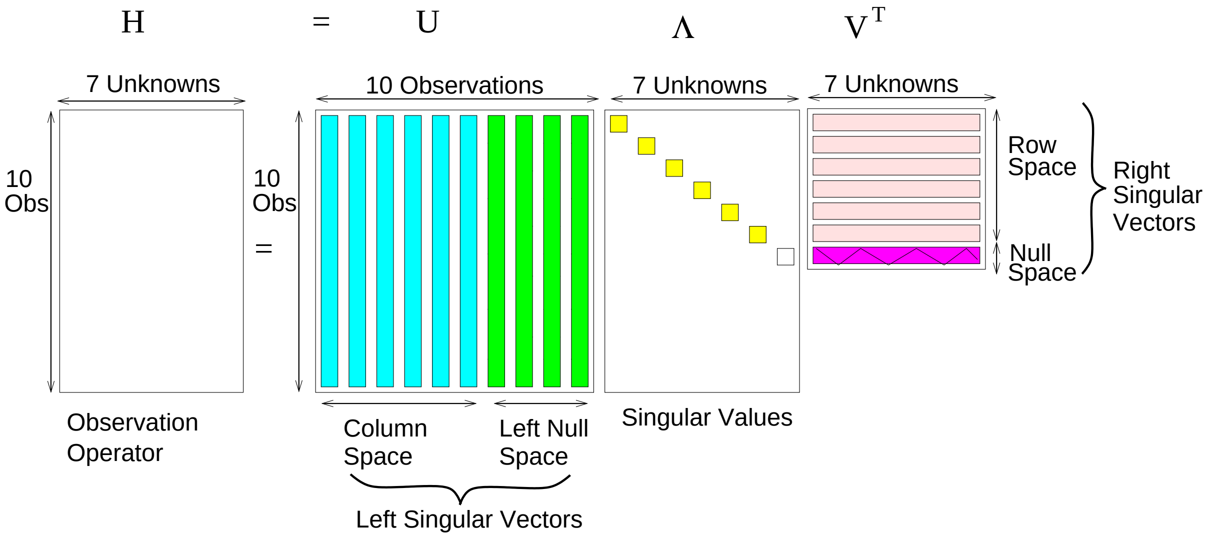


Figure 4.1: *Schematic Diagram illustrating the SVD of the observation operator* $\mathbf{H}$*. There are* $m = 10$ *observations and hence the length of the left singular vectors is 10. There are* $n = 7$ *unknown variables and hence the length of the right singular vectors is 7. As there are more observations than there are unknowns, a left-null space exists. In this case the observation operator does not have as many linearly independent equations as unknowns. Therefore, one of the singular values is zero, and so a null space also exists.*

This means that the vectors in the row space (in state space) are mapped onto vectors in the column space (in observation space). However, as the vectors in the null space have a corresponding zero singular value, the null space vectors are mapped onto zero and not onto the left-null space vectors.

It is sometimes the case that there are singular values that are non-zero but are extremely small. In this case, an effective or numerical rank may be defined, where there is a sharp decrease in the singular values (Golub et al., 1976). The RSVs that correspond to the small non-zero singular values are infact important in 4D-Var, as will be demonstrated in the next chapter.

To illustrate how the SVD can be used to identify the important part of a matrix, matlab has been used to produce an SVD analysis of a (320×200) digital image of a clown. Using the SVD, the image matrix $\mathbf{H}$ can be approximated by a rank-k matrix:

$$\mathbf{H}_k = \sum_{j=1}^{k} \lambda_j \mathbf{u}_j \mathbf{v}_j^T. \tag{4.8}$$

Some of the approximations are shown in Fig. 4.2. The rank-5 approximation shows the basic large scale structure, but it appears very blurred. More detail is added with more singular vectors, as shown by the rank-20 approximation. To the eye, the rank-60 approximation is very similar to the true image. Thus, a truncated SVD can be used to retain the dominant features of the image but discard the unnecessary small scale structures.

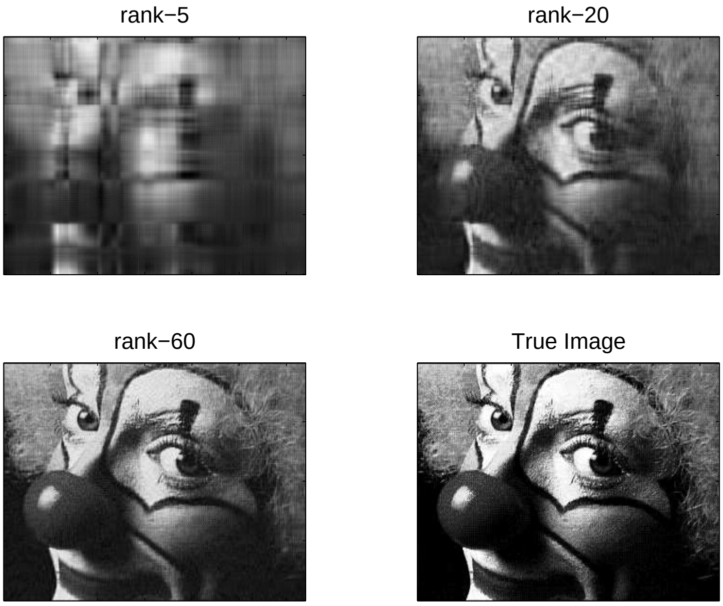


Figure 4.2: *The 'true' image of a clown, and rank-5, rank-20 and rank-60 approximations. The true image has rank-200.*

When the SVD of the observation operator is found, the RSVs lie in state space. This means that the RSVs are the same dimension as the state vector and the variables in the RSVs correspond to the variables in the state vector. Similarly, the LSVs lie in observation space. This means that the SVD can be used to identify the structure in state space that will map onto a particular structure that is observed.

Thus, the SVD can be used to identify which variables in the state space can be determined by the observations. If a retrieved state has components which lie in the null space of the observation operator, their values can not have been obtained from the measurements (Rodgers, 2000). It is obvious that a null space will exist if there are fewer observations than unknown state variables. Even if there are more observations than unknowns, it is still possible for a null space to exist; for example, if there are two observations of the same variable. A null space may also exist if the equations linking the observations with the unknowns are not all linearly independent, but this may not be obvious at first sight. In such a case, the SVD can be used to identify the rank of the matrix and also the state variables that can obtain information from the observations.

To summarize, the SVD is a useful tool which can be used to identify the dominant or important part of the observation operator. This allows the variables which can be determined from the observations to be identified. We now show this in more detail, by considering the SVD in a 1D-Var/3D-Var algorithm.

4.3 Application to 3D-Var

The solution to minimising the least-squares problem (4.5) is now written in the form of a singular vector decomposition. Setting the gradient of the cost function (4.5) to zero, gives the BLUE analysis equation:

$$\mathbf{x}^a = \mathbf{x}^b + (\mu^2\mathbf{I} + \mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T\mathbf{d} \tag{4.9}$$

where $\mathbf{d} = \mathbf{y} - \mathbf{H}\mathbf{x}^b$ is the ‘innovation vector’. Substituting an SVD of the observation operator $\mathbf{H} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^\mathbf{T}$ into the BLUE equation gives:

$$\mathbf{x}^a = \mathbf{x}^b + (\mu^2\mathbf{I} + (\mathbf{U}\mathbf{\Lambda}\mathbf{V}^T)^T(\mathbf{U}\mathbf{\Lambda}\mathbf{V}^T))^{-1}(\mathbf{U}\mathbf{\Lambda}\mathbf{V}^T)^T\mathbf{d}. \tag{4.10}$$

Letting $\mathbf{z} = \mathbf{V}^T\mathbf{x}$ and using the orthonormal property of the RSVs, $\mathbf{V}\mathbf{V}^T = \mathbf{I}$, and of the LSVs, $\mathbf{U}^T\mathbf{U} = \mathbf{I}$, then

$$\left(\mu^2\mathbf{V}\mathbf{V}^T + \mathbf{V}\mathbf{\Lambda}^2\mathbf{V}^T\right)\mathbf{V}(\mathbf{z}^a - \mathbf{z}^b) = \mathbf{V}\mathbf{\Lambda}\mathbf{U}^T\mathbf{d}, \tag{4.11}$$

and again using $\mathbf{V}^T\mathbf{V} = \mathbf{I}$, then

$$\mathbf{x}^a - \mathbf{x}^b = \mathbf{V}\left(\mu^2 + \mathbf{\Lambda}^2\right)^{-1}\mathbf{\Lambda}\mathbf{U}^T\mathbf{d}. \tag{4.12}$$

Thus, the analysis increments can be written as a linear combination of the right singular vectors of the observation operator $\mathbf{H}$,

$$\mathbf{x}^a = \mathbf{x}^b + \sum_{j=1}^{r}\frac{\lambda_j(\mathbf{u}_j^T\mathbf{d})}{\mu^2 + \lambda_j^2}\mathbf{v}_j \tag{4.13}$$

where $\mathbf{u}$,$\mathbf{v}$, λ and r are the LSVs, RSVs, singular values and rank of $\mathbf{H}$. With a similar notation to Hansen (2001), this can be written as:

$$\mathbf{x}^a - \mathbf{x}^b = \sum_{j=1}^{r} \frac{\lambda_j^2}{\mu^2 + \lambda_j^2} \frac{\mathbf{u}_j^T \mathbf{d}}{\lambda_j} \mathbf{v}_j. \tag{4.14}$$

The weight given to the RSVs is partly determined by the term $\frac{\mathbf{u}_j^T \mathbf{d}}{\lambda_j}$. If the the innovation vector $\mathbf{d}$ is similar to the LSV $\mathbf{u}_j$, then the corresponding RSV is given a large weight. For example, if they are exactly the same then $\mathbf{u}_j^T \mathbf{d} = 1$. However, if the vectors are completely orthogonal, then the RSV is given zero weight. Usually, the magnitude of $\mathbf{u}_j^T \mathbf{d}$ has a similar magnitude to λ_j, so that the magnitude of $\frac{\mathbf{u}_j^T \mathbf{d}}{\lambda_j}$ is similar for different values of λ_j.

The weight given to the RSVs is also determined by the term $f_j = \frac{\lambda_j^2}{\mu^2+\lambda_j^2}$, which are known as the Tikhonov Filter Factors (Hansen, 2001). These weights damp all the contributions to the analysis increment which have small singular values λ_j, as:

$$f_j = \begin{cases} 1 & \lambda_j >> \mu \\ \frac{1}{2} & \lambda_j = \mu \\ \frac{\lambda_j^2}{\mu^2} & \lambda_j << \mu \end{cases} \tag{4.15}$$

These weights are illustrated in Fig. 4.3. For $\mu = 0.1$, the RSVs with $\lambda > 0.1$ are given a significantly large weight, whilst the RSVs with $\lambda < 0.1$ are given much less weight. Thus, the damping of the RSVs occurs for $\lambda < \mu$. This is also discussed by Rodgers (2000).

If the observability matrix contains a null space, for example, if there are more unknowns than observations, then some singular values will be zero. In this case there would be some RSVs (in the null space) that do not have corresponding LSVs. If there is no background term, then it would not matter how much weight was given to the corresponding RSVs, as they would have no impact on the value of the cost function. Thus, the solution would be non-unique. Thus, the background state is essential to ensure that the solution is unique, by providing extra information where there is no available information from the observations.

Even if the problem has full rank, there may be some singular values that are very small. The background state also damps the RSVs that have small singular values. It will be shown that this is important as many of these RSVs contain small scale structures corresponding to noise. However, some of these RSVs also contain the important information needed in 4D-Var.

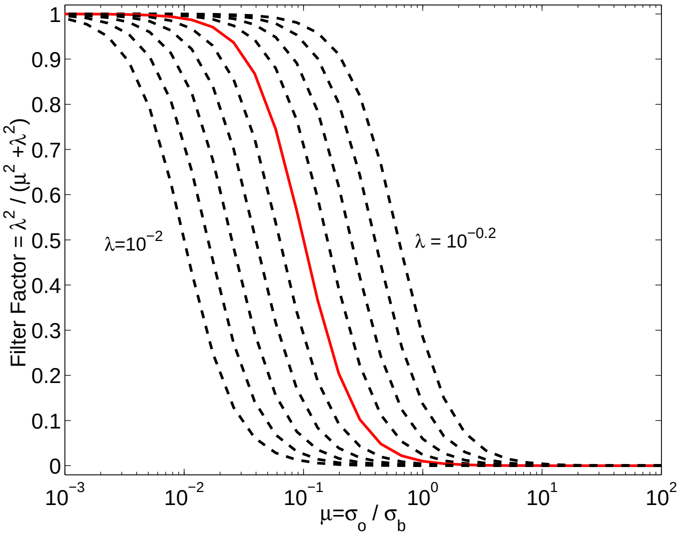


Figure 4.3: *The Tikhonov Filter Factors f as a function of the singular value λ and the relative weight parameter μ, for singular values $\lambda = 10^{-0.2i}$ for $i = 1, 10$. The red solid line indicates the values for $\lambda = 10^{-1}$.*

The SVD analysis of the 3D-Var scheme is well-known, and provides a useful technique to examine the information content of, for example, observations in satellite retrievals. This technique is now extended so that the temporal dimension in 4D-Var is included. To the best of the authors knowledge, such an extension has not been considered in previous literature. Despite being a simple extension, this will allow a new understanding of the information content of observations in 4D-Var.

4.4 Extension to 4D-Var

The 4D-Var cost function is similar to the 3D-Var cost function, except that the observations are distributed in time and linked together by the model equations. Mathematically, 4D-Var finds the analysis $\mathbf{x}^a$ which minimizes the cost function:

$$J(\mathbf{x}_0) = \sigma_b^{-2}(\mathbf{x}_0 - \mathbf{x}^b)^T(\mathbf{x}_0 - \mathbf{x}^b) + \sum_{i=0}^{N} \sigma_o^{-2}(\mathbf{y}_i - \mathbf{H}\mathbf{x}_i)^T(\mathbf{y}_i - \mathbf{H}\mathbf{x}_i) \qquad (4.16)$$

subject to the strong constraint $\mathbf{x}_{i+1} = \mathbf{M}\mathbf{x}_i$. This can be rewritten in the form of an unconstrained minimization as:

$$\begin{aligned} J(\mathbf{x}_0) &= \sigma_b^{-2}(\mathbf{x}_0 - \mathbf{x}^b)^T(\mathbf{x}_0 - \mathbf{x}^b) \\ &+ \sigma_o^{-2}(\mathbf{y}_0 - \mathbf{H}\mathbf{x}_0)^T(\mathbf{y}_0 - \mathbf{H}\mathbf{x}_0) + \sigma_o^{-2}(\mathbf{y}_1 - \mathbf{H}\mathbf{M}\mathbf{x}_0)^T(\mathbf{y}_1 - \mathbf{H}\mathbf{M}\mathbf{x}_0) \\ &+ \cdots + \sigma_o^{-2}(\mathbf{y}_N - \mathbf{H}\mathbf{M}^N\mathbf{x}_0)^T(\mathbf{y}_N - \mathbf{H}\mathbf{M}^N\mathbf{x}_0) \end{aligned} \tag{4.17}$$

which, with a simple rearrangement, gives:

$$\begin{aligned} J(\mathbf{x}_0) &= \sigma_b^{-2}(\mathbf{x}_0 - \mathbf{x}^b)^T(\mathbf{x}_0 - \mathbf{x}^b) \\ &+ \sigma_o^{-2}\begin{bmatrix} (\mathbf{y}_0 - \mathbf{H}\mathbf{x}_0)^T & \dots & (\mathbf{y}_N - \mathbf{H}\mathbf{M}^N\mathbf{x}_0)^T \end{bmatrix}\begin{bmatrix} (\mathbf{y}_0 - \mathbf{H}\mathbf{x}_0) \\ \vdots \\ (\mathbf{y}_N - \mathbf{H}\mathbf{M}^N\mathbf{x}_0) \end{bmatrix} \end{aligned} \tag{4.18}$$

or equivalently the analysis $\mathbf{x}^a$ minimizes the cost function:

$$J_2(\mathbf{x}_0) = \mu^2\|\mathbf{x}_0 - \mathbf{x}^b\|_2^2 + \|\hat{\mathbf{y}} - \hat{\mathbf{H}}\mathbf{x}_0\|_2^2. \tag{4.19}$$

The cost function is now written in the same form as for 3D-Var, except that the vector of observations now includes observations distributed in time and there is a ‘new’ observation operator $\hat{\mathbf{H}}$ and vector of observations $\hat{\mathbf{y}}$ which can be written in block matrix form as:

$$\hat{\mathbf{y}} = \begin{pmatrix} \mathbf{y}_0 \\ \mathbf{y}_1 \\ \vdots \\ \mathbf{y}_N \end{pmatrix} \qquad \hat{\mathbf{H}} = \begin{bmatrix} \mathbf{H} \\ \mathbf{H}\mathbf{M} \\ \vdots \\ \mathbf{H}\mathbf{M}^N \end{bmatrix}. \tag{4.20}$$

The matrix $\hat{\mathbf{H}}$ is known as the ‘observability matrix’ in optimal control theory, and so we will continue to use this term in the rest of this thesis. The observability matrix is also derived in Appendix B of Zou et al. (1992a) to show that if the rank of the observability matrix is equal to the number of unknowns, then it is possible to obtain a unique analysis with no background state.

The observability matrix can be thought of as an effective or generalized observation operator for 4D-Var, as it acts in a similar way to the observation operator in 3D-Var. Thus, it is possible to apply the SVD to this matrix. During the rest of this thesis, it will be shown, using the SVD, that the observability matrix plays a key role in the understanding of the mechanisms in 4D-Var.

Following a similar procedure as for 3D-Var, then the 4D-Var analysis increments can be written as

$$\mathbf{x}^a - \mathbf{x}^b = \sum_{j=1}^{r} \frac{\lambda_j^2}{\mu^2 + \lambda_j^2} \frac{\mathbf{u}_j^T \hat{\mathbf{d}}}{\lambda_j} \mathbf{v}_j \tag{4.21}$$

where $\mathbf{u}_j$, $\mathbf{v}_j$, λ_j and r are the LSVs, RSVs, singular values and rank of the 4D-Var observability matrix $\hat{\mathbf{H}}$, and $\mu^2 = \frac{\sigma_o^2}{\sigma_b^2}$ is the weight given to the background state relative to the weight given to the observations, and $\hat{\mathbf{d}} = \hat{\mathbf{y}} - \hat{\mathbf{H}}\mathbf{x}^b$ is the 4D-Var generalized innovation vector.

This is a key result of this thesis. The technique allows us to identify which components of the state vector can be identified from the observations. That is, the RSVs indicate which components of the analysis can be updated by the observations, for a particular model and observing system.

The technique also allows us to assess the influence of the background state. The relative weight given to the observations, μ^2, has been separated from the RSVs. This allows easy inference of how the analysis would change if the weights were changed. That is, we can infer which RSVs (and therefore, which components of the state vector), would be penalized if the weight given to the background state was increased.

4.5 Relationship between 4D-Var and Optimal Perturbations

The singular value decomposition is commonly used in meteorology to define a set of optimal perturbations that maximize the growth, defined by suitable metrics, in a finite-time interval (Buizza and Palmer, 1995). They are used to locate sensitive regions where errors are likely to grow and therefore used to generate a set of perturbations to use in ensemble prediction. Optimal perturbations are the right singular vectors of the linear forecast model and the growth rate is proportional to the corresponding singular value. The term ‘optimal perturbation’ is used to denote the RSVs of the model to avoid confusion with the RSVs of the observability matrix.

It is useful to understand how the RSVs of the observability matrix relate to the optimal perturbations. There are clear similarities between optimal perturbations and 4D-Var when a full set of observations are given at the end of the assimilation window as previously shown by Rabier et al. (1996) and Thépaut et al. (1996).

Optimal perturbations $\mathbf{x}_0$ are defined to maximize the ratio

$$\text{growth} = \frac{\|\mathbf{P}\mathbf{x}_N\|_{\mathbf{E}}}{\|\mathbf{x}_0\|_{\mathbf{C}}} \tag{4.22}$$

where $\mathbf{P}$ is an operator which maps from a space of dimension n to a smaller space (for example, so that the perturbations identify the maximum growth within a specified region of the atmosphere) and $\mathbf{C}$ and $\mathbf{E}$ are the initial and final time norms respectively (Barkmeijer et al., 1998). Commonly, these norms are taken to be the total energy norm. To illustrate the relationship between optimal perturbations and the RSVs defined above, we choose the norms to be the observation and background error inverse covariance matrices, and that the operator $\mathbf{P}$ is the observation operator $\mathbf{H}$, so that:

$$\text{growth} = \frac{< \mathbf{H}\mathbf{M}\mathbf{x}_0; \mathbf{R}^{-1}\mathbf{H}\mathbf{M}\mathbf{x}_0 >}{< \mathbf{x}_0; \mathbf{B}^{-1}\mathbf{x}_0 >}. \tag{4.23}$$

Using $\mathbf{R} = \sigma_o^2\mathbf{I}$ and $\mathbf{B} = \sigma_b^2\mathbf{I}$, then

$$\text{growth} = \frac{1}{\mu^2}\frac{< \mathbf{x}_0; \hat{\mathbf{H}}^T\hat{\mathbf{H}}\mathbf{x}_0 >}{< \mathbf{x}_0; \mathbf{x}_0 >} \tag{4.24}$$

where $\hat{\mathbf{H}} = \mathbf{H}\mathbf{M}(t_N, t_0)$, and $\mu = \frac{\sigma_o}{\sigma_b}$. Therefore, in 4D-Var, with observations at only the end of the assimilation window, the RSVs of the observability matrix are the same as optimal perturbations with the observation operator $\mathbf{H}$ acting as the operator $\mathbf{P}$. If a large weight is given to the background state (μ is large), then the RSVs have a large growth rate. This is intuitive, as we can imagine that the analysis increment is such that the analysis is close to the background state at the initial time and close to the observations at the final time.

There are also many differences between optimal perturbations and RSVs of the observability matrix. First, optimal perturbations are RSVs that are defined at the initial time and evolve into the LSVs which are defined at the final time. However, the RSVs of $\hat{\mathbf{H}}$ are defined in state space and evolve into LSVs which are defined in observation space. Second, by using a time sequence of observations, the analysis increments are no longer projected onto the optimal perturbations; decaying structures are also included. However, we will demonstrate in Chapter 5 that the decaying modes are penalized by the background state. Third, optimal perturbations are usually defined with respect to metrics such as total energy, enstrophy or the analysis error covariance, whilst the appropriate metrics for the RSVs of $\hat{\mathbf{H}}$ are the error covariance matrices $\mathbf{B}^{-1}$ and $\mathbf{R}^{-1}$. This last fact will be made more clear in Chapter 6, when spatial correlationsof the background state error are also included.

4.6 Rate of Convergence of the Minimization Algorithm

In the second chapter (Section 2.3), we discussed how the conditioning of the Hessian matrix is central to understanding the convergence of the minimization algorithm. We now discuss how the SVD of the observability matrix can be used to determine whether the Hessian matrix is well-conditioned.

The Hessian of the cost function (4.19), is:

$$\nabla\nabla J_2 = \mu^2\mathbf{I} + \hat{\mathbf{H}}^T\hat{\mathbf{H}}. \tag{4.25}$$

Suppose the observability matrix $\hat{\mathbf{H}}$ has RSVs $\mathbf{v}$ and associated singular values λ, then these also satisfy the eigenvector relationship:

$$\hat{\mathbf{H}}^T\hat{\mathbf{H}}\mathbf{v} = \lambda^2\mathbf{v} \tag{4.26}$$

and therefore, these are related to the Hessian matrix by:

$$(\mu^2\mathbf{I} + \hat{\mathbf{H}}^T\hat{\mathbf{H}})\mathbf{v} = (\mu^2 + \lambda^2)\mathbf{v}. \tag{4.27}$$

Thus the right singular vectors of the observability matrix are also the eigenvectors of the Hessian of the cost function. This has two important consequences.

The first consequence is that the RSVs with the smallest singular values are also the directions of the Hessian ellipsoid axes in which the isocontours are stretched the most. Thus, the minimization algorithm has difficulties in finding the directions with the smallest singular values. The minimization algorithm will first identify the RSVs with large singular values and will then identify the RSVs with small singular values later on.

The second consequence is that the conditioning of the Hessian matrix is related to the maximum and minimum singular values. That is, the condition number is defined as:

$$cond(\nabla\nabla J) = \frac{\lambda_{\text{max}}^2 + \mu^2}{\lambda_{\text{min}}^2 + \mu^2}. \tag{4.28}$$

where λ_{max} and λ_{min} are the maximum and minimum singular values of $\hat{\mathbf{H}}$. Consider the case with a zero singular value. Then, if there was no background state ($\mu = 0$), this would give an infinite condition number. That is, the analysis would be non-unique. This is expected, as a zero singular value corresponds to the observability matrix not having full-rank. If, however, a large weight is given to the background state (μ is large), then the condition number would no longer be infinite. This illustrates why the background state is needed to ensure that the problem is well-posed. Even if there were no zero singular values, adding a background state ($\mu^2 > 0$) reduces the condition number and hence improves the conditioning of the problem. Therefore, the background state also improves the convergence rate of the minimization.

4.7 Calculating the SVD for the Eady model

In the next chapter, the SVD of the observability matrix for the 4D-Var with the Eady model is discussed; the method used to find the SVD is now discussed. The details are described further in Appendix A.

There are three strategies to compute the SVD of a linear operator (Toumazou, 2001). The SVD strategy computes all the singular values and vectors of $\hat{\mathbf{H}}$. The QR strategy computes all the eigenvectors and eigenvalues of $\hat{\mathbf{H}}^T\hat{\mathbf{H}}$, and the Lanczos method is an iterative eigensolver which computes only the k largest singular values and associated singular vectors.

The Lanczos strategy is particularly useful for large problems as the algorithm does not need the linear model to be in matrix form. Therefore, the Lanczos approach is commonly taken when computing optimal perturbations of NWP models (Buizza, 1997). However, the algorithm is not able to accurately and efficiently compute all the RSVs (P. Haas, A. Beck, Personal Communication).

In this thesis, all the singular vectors and singular values are required. Therefore, the NAG routine nag_gen_svd (NAG), based on the SVD algorithm described by Golub and Van Loan (1996), is used. This algorithm requires that the linear operator is in matrix form.

Considering the case with a complete set of observations of the lower level buoyancy at the beginning and the end of the window, we compute the SVD of:

$$\hat{\mathbf{H}} = \begin{bmatrix} \mathbf{H} \\ \\ \mathbf{HM}(t_N, t_0) \end{bmatrix}. \tag{4.29}$$

To generate $\hat{\mathbf{H}}$ in matrix form, the Eady model also needs to be in matrix form. This is found by applying the discrete Eady model equations to successive columns of the identity matrix. As these initial conditions are discontinuous fields, the Lax-Wendroff numerical scheme is used, as discussed in Appendix A.

4.8 Conclusions

The 4D-Var algorithm has been considered as an inverse problem that is similar in form to 3D-Var. The observation operator in the 3D-Var cost function is replaced by the 4D-Var observability matrix, which contains both the observation operator and the linear forecast model. Writing the 4D-Var cost function in this form allows the 4D-Var analysis increments to be given by a linear combination of the right singular vectors (RSVs) of the observability matrix. The weight given to the RSVs is partly determined by the Tikhonov Filter Factors. These factors penalize the RSVs with small singular values in comparison to the relative weight given to the background state.

The SVD formulation of 4D-Var has provided a number of interesting results concerning the need for the background state, the similarities between 4D-Var and optimal perturbations and an understanding of the behaviour of the minimization algorithm.

The background state is needed to ensure that the 4D-Var inverse problem is well-posed. First, the background state ensures that the analysis is unique. This is vital if the observability matrix does not have full rank and hence a null space exists. Second, the background state is needed to penalize the RSVs with small, but non-zero, singular values. It will be shown in the next chapter that these RSVs have small scale spatial structures and hence correspond to noise. Thus, the background state is needed to ensure that the analysis does not include unphysical structures corresponding to the observational noise.

It was previously known that there are similarities between 4D-Var and optimal perturbations. The SVD formulation has provided a more precise relationship which has highlighted both the similarities and differences between the two. The main differences occur when the observations are at more than one time level and also in the metrics that are used to define the optimal perturbations.

By considering the Hessian of the cost function, it was shown that the background state acts to improve the conditioning of the problem and that the RSVs of the observability matrix are also the eigenvectors of the Hessian. This implies that the minimization algorithm will correct the directions of the RSVs with large singular values during the first few iterations.

An important consequence of the SVD formulation of 4D-Var is that it provides a useful tool to understand the information content of observations in 4D-Var. This will be demonstrated in the next chapter by discussing the SVD of the observability matrix for the Eady model.

Chapter 5

SVD Results

Results from simple 4D-Var identical twin experiments with the Eady model were presented in Chapter 3. The experiments investigated the ability of 4D-Var to reconstruct the state in unobserved regions, and to generate analysis increments with the vertical structure necessary for baroclinic growth or decay. Five main results were found: the background state penalizes the information needed to reconstruct the state in unobserved regions; the unobserved regions are particularly sensitive to observational noise; the background state penalizes the decaying analysis increments; analyses are improved when the observations are moved further apart in time; and giving more weight to the initial observations does not improve the analysis of a decaying mode. These results play a key role in assessing the advantages of 4D-Var, and towards maximizing these benefits. However, the reasons for these results are not well understood. In this chapter, the singular value decomposition (SVD) is used to provide a new understanding of 4D-Var.

It was demonstrated in Chapter 4, that the SVD provides a useful interpretation of the information content of observations in 4D-Var. This chapter begins with a brief review of the SVD technique, and then the SVD computations of the 4D-Var observability matrix are shown. These computations are used as a basis for understanding the five main results found in Chapter 3. The structure of the matrix of right singular vectors and the singular values are used to understand why the background state penalizes the information needed to reconstruct the unobserved regions, the weights given to the RSVs are used to understand why the unobserved regions are sensitive to noise, and the spatial structures of the RSVs are examined to understand why the background state penalizes the decaying modes. The SVD computations are modified so that the effect of the temporal position and weights of the observations can also be examined.

The chapter concludes by establishing a link between 4D-Var and a method known as Tikhonov regularization. This provides a more complete understanding of the 4D-Var algorithm.

5.1 Singular Vector Technique

Following the experiments in Chapter 3, observations $\mathbf{y}_0$ and $\mathbf{y}_N$ of only the lower buoyancy, are provided at the beginning and the end of an assimilation window $[t_0, t_N]$. It is assumed that there are no background error correlations so that the error covariances are diagonal with constant variances, $\mathbf{B} = \sigma_b^2\mathbf{I}$ and $\mathbf{R} = \sigma_o^2\mathbf{I}$.

Then, the 4D-Var cost function can be written as:

$$J(\mathbf{x}_0) = \frac{1}{2}\left\{\sigma_b^{-2}\|\mathbf{x}_0 - \mathbf{x}^b\|_2^2 + \sigma_o^{-2}\|\hat{\mathbf{y}} - \hat{\mathbf{H}}\mathbf{x}_0\|_2^2\right\} \tag{5.1}$$

where the generalized observation vector $\hat{\mathbf{y}}$ and the observability matrix $\hat{\mathbf{H}}$ are given by:

$$\hat{\mathbf{y}} = \begin{bmatrix} \mathbf{y}_0 \\ \\ \mathbf{y}_N \end{bmatrix} \qquad \hat{\mathbf{H}} = \begin{bmatrix} \mathbf{H} \\ \\ \mathbf{H}\mathbf{M}(t_N, t_0) \end{bmatrix} \tag{5.2}$$

where $\mathbf{x}_i$ is the state vector at time t_i, $\mathbf{y}_i$ is the vector of observations at time t_i, and $\mathbf{M}(t_N, t_0)$ is the linear Eady model such that $\mathbf{x}_N = \mathbf{M}\mathbf{x}_0$.

Setting the gradient of J with respect to $\mathbf{x}_0$ to zero, and using the singular value decomposition (SVD) of the observability matrix, $\hat{\mathbf{H}} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^T$, then the 4D-Var analysis increments can be written as:

$$\mathbf{x}^a - \mathbf{x}^b = \sum_{j=1}^{r} \frac{\lambda_j^2}{\mu^2 + \lambda_j^2}\frac{\mathbf{u}_j^T\hat{\mathbf{d}}}{\lambda_j}\mathbf{v}_j \tag{5.3}$$

where $\mathbf{u}, \mathbf{v}, \lambda$ and r are the left singular vectors (LSVs), right singular vectors (RSVs), singular values and rank of the observability matrix, j is the singular vector index, $\hat{\mathbf{d}} = \hat{\mathbf{y}} - \hat{\mathbf{H}}\mathbf{x}^b$ is the generalized innovation vector and $\mu^2 = \sigma_o^2/\sigma_b^2$ is the relative weight given to the background state in comparison to the observations.

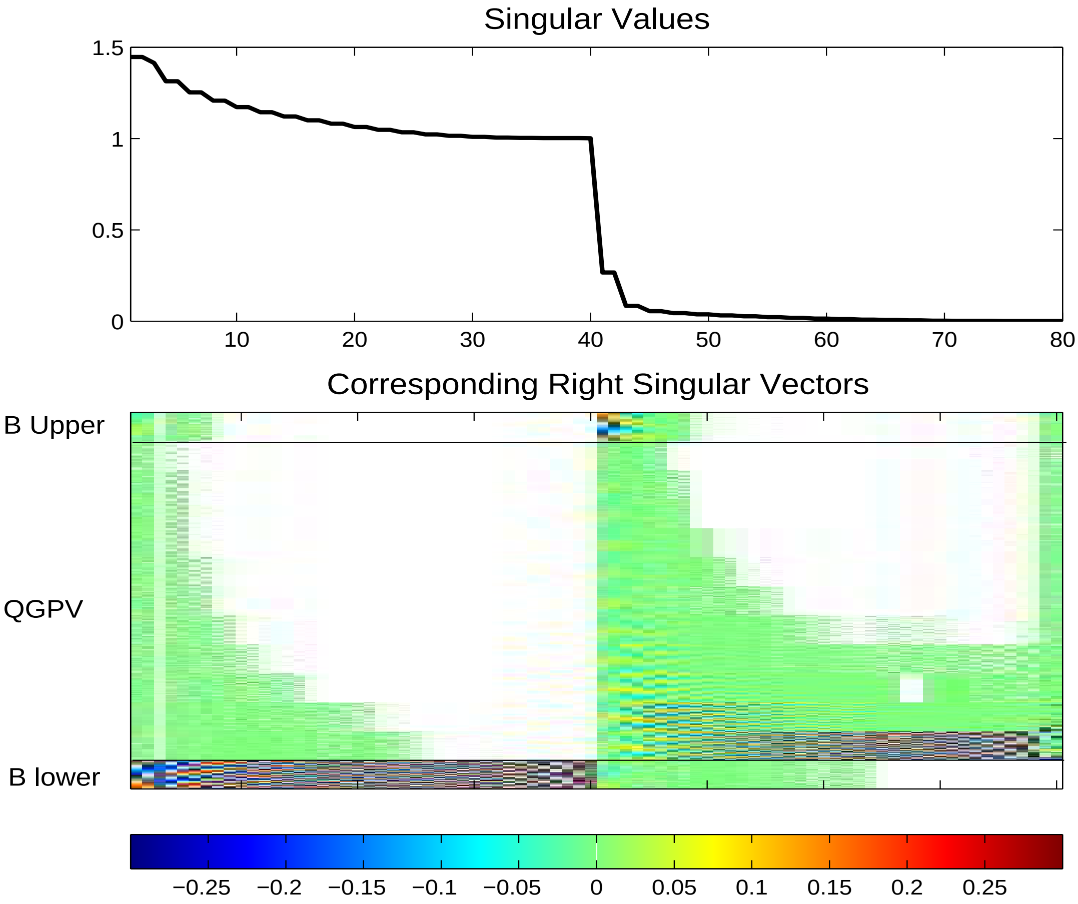


Figure 5.1: *Singular values λ and $\mathbf{V}$ matrix of the observability matrix $\hat{\mathbf{H}}$ as a function of the singular vector index, when there are observations of the lower level buoyancy at the beginning and the end of a 6 hour assimilation window. The $\mathbf{V}$ matrix is shown as an image so that each element corresponds to a colour, as shown by the colour bar. The very small values have been shaded white. Each column of $\mathbf{V}$ gives the RSV that corresponds to the singular values above, and the RSVs are such that the first 40 elements contain the upper level buoyancy (B upper), the last 40 elements contain the lower level buoyancy (B lower) and the elements in the middle correspond to the QGPV on the 11 vertical levels.*

5.2 The Background State Penalizes Important Information

In Chapter 3, it was shown that if a large weight is given to the background state, the information needed to reconstruct the state in unobserved regions is strongly penalized. The SVD of the observability matrix is now used to illustrate why the information needed to reconstruct the state is particularly sensitive to the relative weight μ^2 given to the background state.

The singular values λ and the $\mathbf{V}$ matrix of the observability matrix $\hat{\mathbf{H}}$ are shown in Fig. 5.1. There are 40 observations at two time levels, giving a total of 80 observations and hence only 80 singular values. Although there are 80 non-zero singular values, there is a sharp-drop in the singular value spectrum, suggesting an effective rank of 40.

The $\mathbf{V}$ matrix shows two distinct regions: RSVs corresponding to large singular values and RSVs corresponding to small singular values. The largest values of the first 40 RSVs are on the lower boundary (B lower), whereas the largest values of the RSVs from 40 onwards are on the upper boundary (B upper), and also on the lower boundary interior QGPV.

In both regions, there tends to be an increase in the number of oscillations in the horizontal in the RSVs, with increasing singular vector j. That is, the horizontal wavelength tends to decrease as the singular values decrease; although there are some regions where the horizontal wavelength increases with decreasing singular values, for example, $j = 30, \ldots, 40$ & $j = 70, \ldots, 80$.

The structure of the $\mathbf{V}$ matrix shows that the information needed to reconstruct the (observed) lower level wave is mainly contained in the RSVs corresponding to large singular values. In contrast, the information needed to reconstruct the (unobserved) upper level wave is mainly contained in the RSVs corresponding to small singular values.

The background state penalizes the RSVs with small singular values (from equation (5.3)). However, these RSVs contain the information needed to reconstruct the state in the unobserved regions. It is therefore evident that this is the reason why the background state penalizes the important information that is propagated from the observed regions to the unobserved regions.

We noted in Section 3.2.1 that the minimization algorithm first corrected the observed lower boundary and then corrected the unobserved upper boundary. The relationship between the RSVs of the observability matrix and the behaviour of the minimization was discussed in Section 4.6. The minimization algorithm first updates the directions with large singular values. From Fig. 5.1 the RSVs with large singular values contain the information needed to reconstruct the lower boundary, whereas the RSVs with small singular values contain the

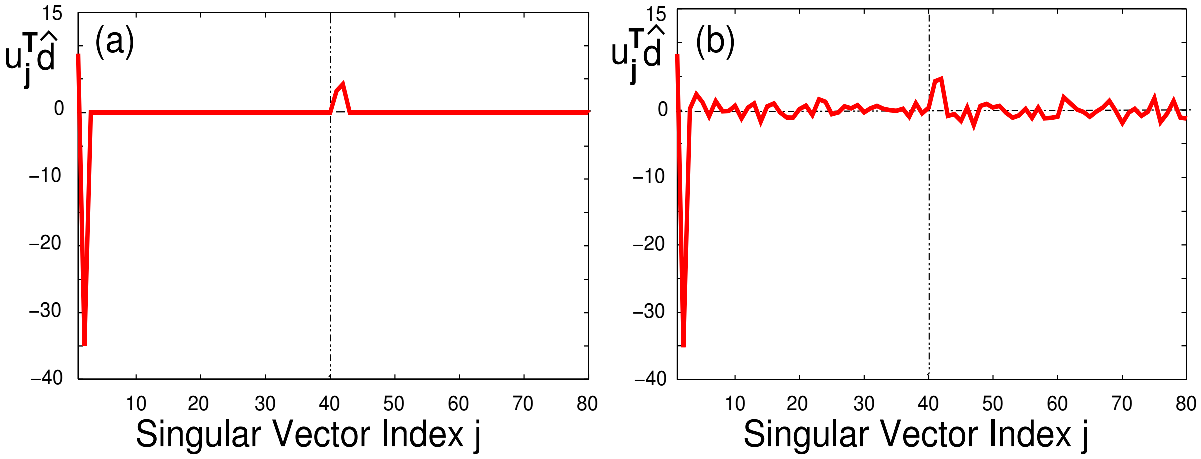


Figure 5.2: *Values of* $\mathbf{u}_j^T\hat{\mathbf{d}}$ *where* $\mathbf{u}_j$ *are the left singular vectors and* $\hat{\mathbf{d}}$ *is the generalized innovation vector. The true state is given by the most unstable Eady wave and the background state has a phase error of* $10\Delta x$*. Observations have (a) no noise and (b) noise with a Gaussian distribution with* $\sigma = 1$.

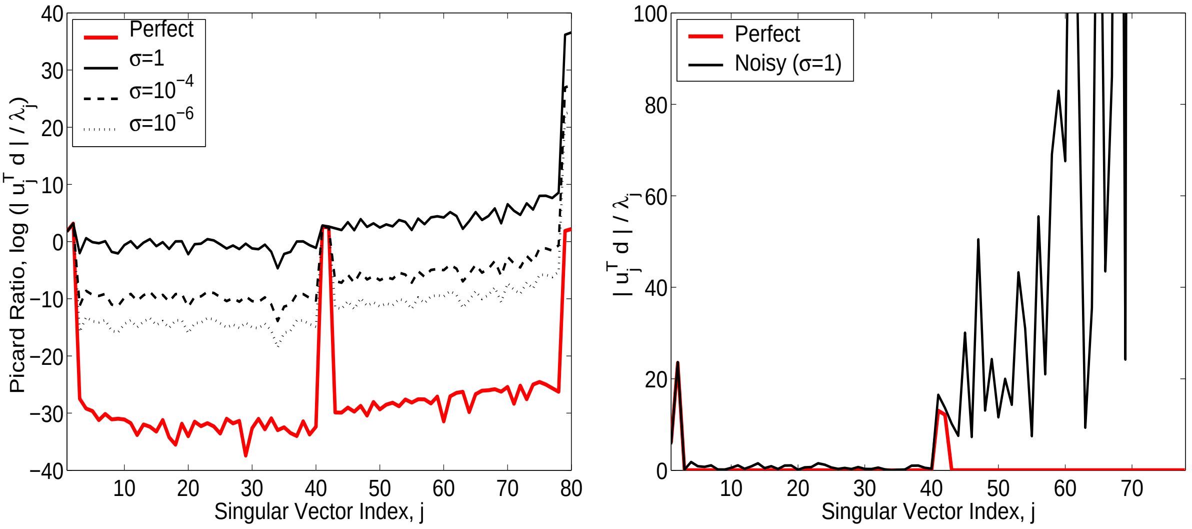
information needed to reconstruct the upper boundary. It is for this reason that the minimization algorithm first corrects the lower boundary wave and then corrects the upper boundary wave.

5.3 Unobserved Regions are Sensitive to Noise

It was shown in Chapter 3 that the unobserved regions are sensitive to the noise on the observations. If a relatively large weight is given to the observations, then an unphysical wave may be generated in the unobserved regions. The reasons for this are now demonstrated by examining the values of $\mathbf{u}_j^T\hat{\mathbf{d}}$ and what are known as the Picard ratio values.

The SVD of $\hat{\mathbf{H}}$ is only dependent on the position of the observations ($\mathbf{H}$) and the model ($\mathbf{M}$). It is independent of the true state that is observed and the observational noise. Thus, although the structure of the $\mathbf{V}$ matrix has been examined, this does not determine which particular RSVs give a large contribution to the analysis increment. The values of $\mathbf{u}_j^T\hat{\mathbf{d}}$ are now used to understand how the observational data is projected onto the particular RSVs.

For comparison with the experiments in Chapter 3, we consider the true state given by the most unstable Eady wave, and the background state with a phase error of $10\Delta x$. When there are perfect observations, there are four values of $\mathbf{u}_j^T\hat{\mathbf{d}}$ which are relatively large, arising in two pairs, as shown in Fig. 5.2(a). This means that RSVs $1\&2$ and $41\&42$ give a large contribution to the analysis increment. When noise is added to the observations, many of the other values of $\mathbf{u}_j^T\hat{\mathbf{d}}$ become almost comparable so that more RSVs are included in the analysis increment. These RSVs have smaller wavelengths, as shown by the $\mathbf{V}$ matrix in Fig. 5.1, so that the



(a) (b)

Figure 5.3: *Values of (a) the Picard ratios and (b) the ratio between $|\mathbf{u}_j^T \hat{\mathbf{d}}|$ and λ_j. In both cases the thick red solid line represents the values for perfect observations, and the thin solid line represents the values for noisy observations with standard deviation $\sigma = 1$. In (a), the case for noisy observations with standard deviation $\sigma = 10^{-4}$ (dashed line), $\sigma = 10^{-6}$ (dotted line) are also shown.*

analysis is closer to the noisy observations.

The values of $\mathbf{u}_j^T \hat{\mathbf{d}}$ do not directly show why the unobserved regions are the most sensitive to the noise. Instead, it is necessary to examine the ratios $\frac{\mathbf{u}_j^T \hat{\mathbf{d}}}{\lambda_j}$, which are the weight coefficients in equation (5.3). Since the ratios $\frac{\mathbf{u}_j^T \hat{\mathbf{d}}}{\lambda_j}$ span several orders of magnitude, it is more convenient to examine what is known as the Picard ratio, (Winkler, 1997),

$$\log\left(\frac{|\mathbf{u}_j^T \hat{\mathbf{d}}|}{\lambda_j}\right). \tag{5.4}$$

The Picard ratios for perfect observations and noisy observations are shown in Fig. 5.3 (a). In all cases, there is a sharp increase in the Picard ratio near j=80. This is due to rounding errors in the SVD computations, and can be ignored. For perfect observations (thick, red line), there are two large spikes in the value of the Picard ratio, similar to the values of $\mathbf{u}_j^T \hat{\mathbf{d}}$. This clearly highlights that only four RSVs are needed to make the correct analysis increment. When noise is added to the observations (thin, black lines) there is a dramatic increase in the Picard ratios corresponding to the other RSVs, and these values increase further when the standard deviation of the noise is increased.

The Picard ratio values illustrate why the unobserved regions are sensitive to noise. When there is no noise on the observations, the Picard ratio values are generally small, apart from the two spikes. Thus, the RSVs with small scale structures are given a small weight. When noise is added to the observations, all the Picard ratio values become large. Importantly, the values increase with the singular vector index j. As a log scale is used, this increase is significant and is perhaps more easily illustrated in Fig. 5.3 (b). It implies that with no background state, an extremely large weight is given to the RSVs with extremely small singular values ($j > 42$). From the $\mathbf{V}$ matrix, these RSVs have small scale structures and large amplitudes on the upper boundary. Thus, the analysis increment is dominated by these RSVs causing an unphysical wave to be generated on the unobserved upper boundary.

To summarize, the SVD analysis has shown that if a small weight is given to the background state, the RSVs with small singular values dominate the analysis increment. These RSVs contain small scale structures in the unobserved regions. If a large weight is given to the background state, the RSVs are strongly damped. It is for this reason that the unobserved regions are particularly sensitive to noise.

5.4 The Background State Penalizes the Decaying Modes

In Chapter 3, it was shown that the background state penalizes the decaying part of the analysis increment, so that a growing analysis increment may be added instead of the required decaying analysis increment. This is now understood by comparing the RSVs that are needed for growing and decaying analysis increments and by examining the vertical structure of the RSVs.

The previous section showed that with perfect observations, two pairs of RSVs are required to create the analysis increment. RSV $1\&2$ have the same singular value, as do RSV $41\&42$. This is due to the symmetry of the Eady model, as discussed by Preisendorfer (1988). The streamfunction fields for RSVs $1\&2$ are shown in Fig. 5.4. The RSVs have exactly the same structure except for a difference in phase. Thus, each RSV pair has a cosine and a sine component, so that the correct horizontal position of the wave can be obtained.

Given that the RSVs form sine and cosine pairs, it is useful to assess the combined weight given to an RSV pair. This is achieved by writing the sum of the weighted RSVs as a linear combination ψ of sine and cosine waves, where the weights are given by the values of $\mathbf{u}_j^T\hat{\mathbf{d}}$,

$$\psi(x,z) = \mathbf{u}_1^T\hat{\mathbf{d}}\sin(x+\phi(z)) + \mathbf{u}_2^T\hat{\mathbf{d}}\cos(x+\phi(z)). \tag{5.5}$$

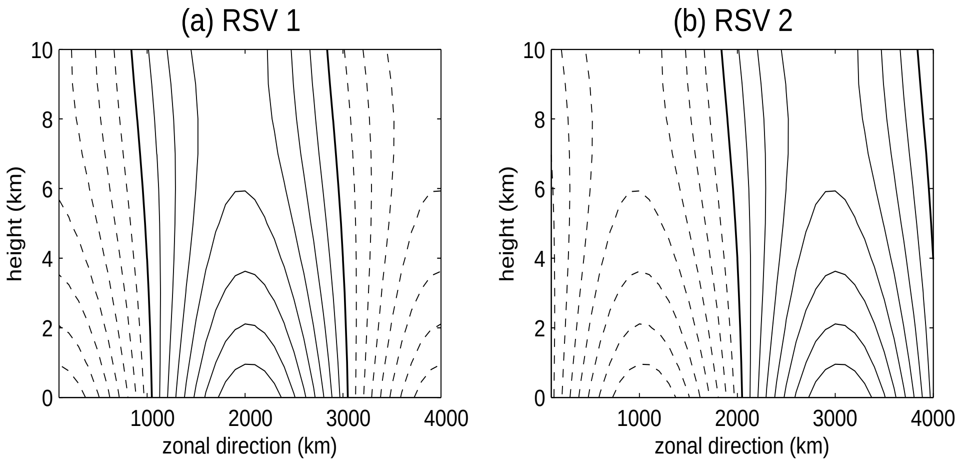


Figure 5.4: *Streamfunction fields of an RSV pair: (a) RSV 1 and (b) RSV 2. Both plots use the same contour interval.*

Setting the derivative of ψ with respect to x to zero, then the maximum or minimum value of ψ is at a distance:

$$x_{max} = \tan^{-1}\left(\frac{\mathbf{u}_1^T\hat{\mathbf{d}}}{\mathbf{u}_2^T\hat{\mathbf{d}}}\right) - \phi(z). \tag{5.6}$$

This can then be used to find the maximum amplitude of ψ:

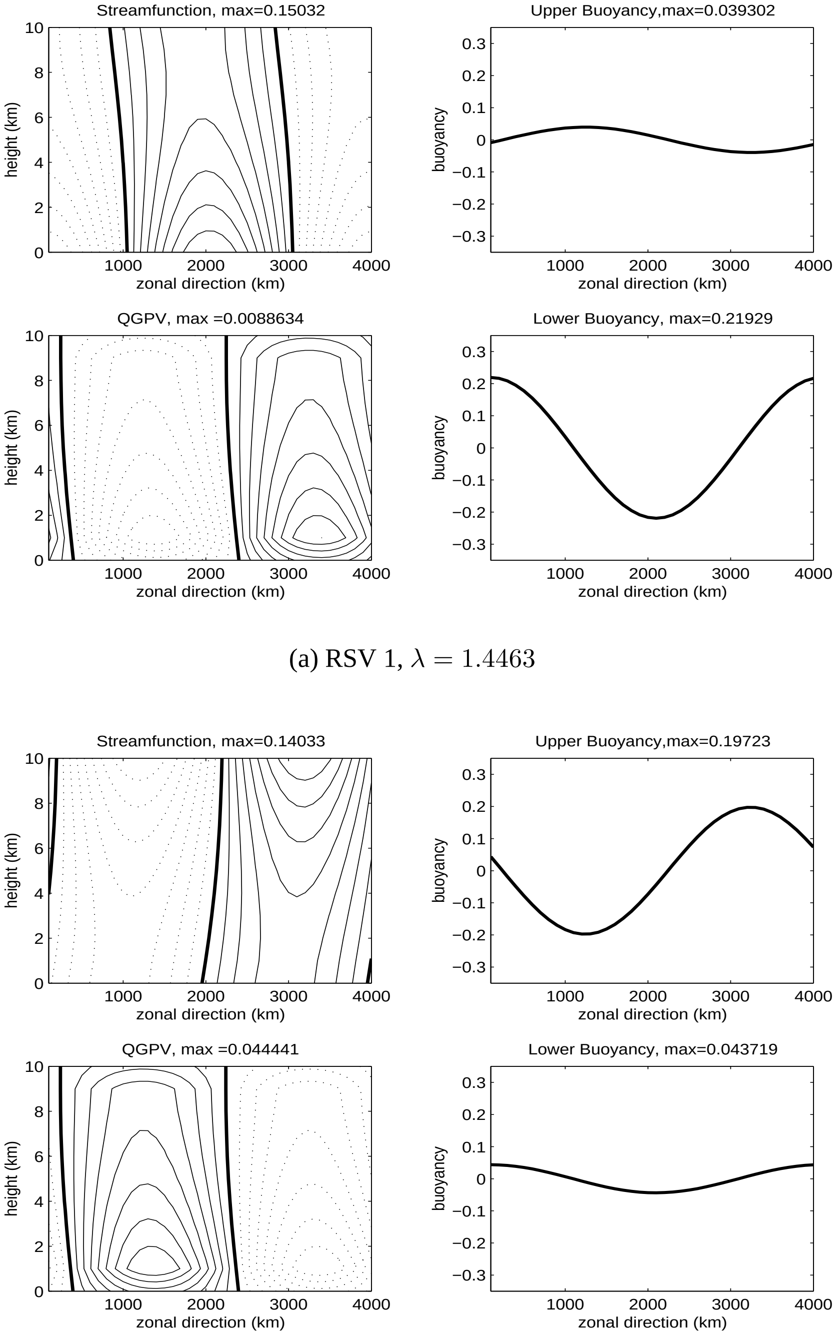
$$\psi_{max} = |\psi(x_{max})|. \tag{5.7}$$

It should be noted that this formula is used to find only the magnitude and not the sign of the combined weight.

The weight given to an RSV pair is then determined by the Picard ratio for an RSV pair, ψ_{max}/λ_j. This value allows an easier interpretation than the weights given to the separate RSV pair components. Table 5.1 shows the values of ψ_{max}/λ_j for the RSV pairs when either a growing or a decaying mode is observed, over either a 6 hour or 12 hour assimilation window. RSVs $41\&42$ have a significantly smaller singular value than RSVs $1\&2$, but they also have smaller values of ψ_{max}. Thus, the values of ψ_{max}/λ_j for the various RSV pairs are of the same order of magnitude, allowing a fair comparison.

For a 6 hour window, when a growing mode is observed, a larger weight is given to RSVs $1\&2$ than RSVs $41\&42$. However, when a decaying mode is observed, more weight is given to RSVs $41\&42$ than RSVs $1\&2$. Thus, RSVs $41\&42$ are more important when a decaying mode is observed. These vectors have very small singular values and hence they are penalized strongly when a large weight is given to the background state.

The difference between the weights for growing and decaying modes is more evident for



(a) RSV 1, $\lambda = 1.4463$

(b) RSV 41, $\lambda = 0.2660$

Figure 5.5: *RSVs shown at the initial time T+0. The RSV is defined by the buoyancy and the QGPV, but streamfunction is calculated from these fields. These RSVs give a large contribution to the analysis increment when either the most unstable growing or decaying mode is observed, and observations are taken of the lower boundary at the beginning and the end of a 6 hour assimilation window. The values at the top right of each plot give the maximum magnitudes of the fields.*

Index j	Singular Value λ_j	Growing mode is observed	Decaying mode is observed
6 hour window			
$1 \& 2$	1.4463	24.99	19.15
$41 \& 42$	0.2660	20.11	21.77
12 hour window			
$1 \& 2$	1.6168	26.47	14.47
$41 \& 42$	0.4669	19.34	25.38

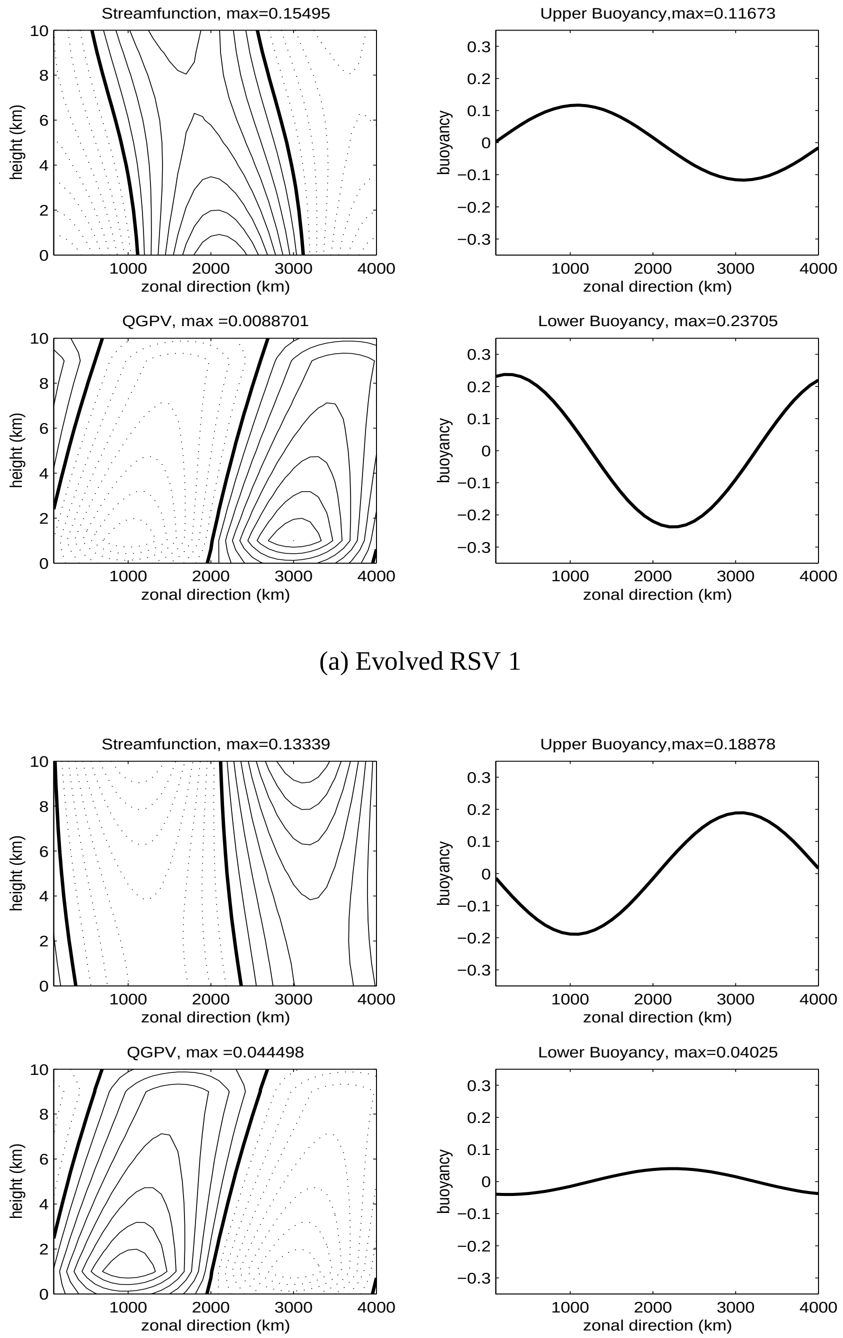
Table 5.1: *Values of ψ_{max}/λ_j for the RSV pairs. Perfect observations of the lower level buoyancy are given at the beginning and the end of either a 6 hour or 12 hour window, the true state is given by either the most rapidly growing or decaying Eady wave and the background state has a phase error. Note that by definition, $\psi_{max}/\lambda_j \geq 0$.*

a longer (12 hour) assimilation window. Thus, for a longer assimilation time, there is a more distinct splitting between the RSVs needed to produce a growing analysis increment (RSVs $1\&2$) and the RSVs needed to produce a decaying analysis increment (RSVs $41\&42$).

To understand the difference between RSVs $1\&2$ and RSVs $41\&42$, the vertical structures of the RSVs are now examined. As the RSVs are in pairs, only one RSV from each pair is shown. Figure 5.5 shows RSV 1 and RSV 41 at the beginning of the time window. The streamfunction field for RSV 1 is the same as that in Fig. 5.4(a). For RSV1, the magnitude of the buoyancy on the lower boundary is much larger than that on the upper boundary. The streamfunction tilts westwards with height and the buoyancy field tilts eastwards with height; these are both characteristics of growing normal modes and will therefore result in growth. In contrast, for RSV 41, the magnitude of the buoyancy is larger on the upper boundary than the lower boundary. The streamfunction tilts eastwards with height and the buoyancy field tilts westwards with height; these are both characteristics of decaying normal modes. The QGPV has relatively small but non-zero values and therefore gives little contribution to the growth and decay in comparison to the upper and lower temperature waves.

The RSVs are defined at the beginning of the assimilation window. However, they can be evolved to the end of the window, by integrating them with the Eady model, to give evolved RSVs, shown in Fig. 5.6. As the RSVs are not the singular vectors of the model, the evolved RSVs are not the same as the LSVs.

For RSV 1, there is a large increase in amplitude of the upper level wave, and the maximum



(a) Evolved RSV 1

(b) Evolved RSV 41

Figure 5.6: *As for Fig.5.5, but now the RSVs are shown at the final time T+6.*

amplitude of the streamfunction field has also increased. The QGPV field has been sheared by the basic state flow. The upper level wave has moved slightly westwards and the lower level has moved slightly eastwards so that at the final time, the structure is more similar to a growing normal mode structure.

For RSV 41, there is a reversal in the sign of the lower buoyancy wave, and again, the upper level wave has moved slightly westwards and the QGPV field has been sheared by the basic state flow. In contrast to RSV 1, the maximum amplitude of the streamfunction field has decreased.

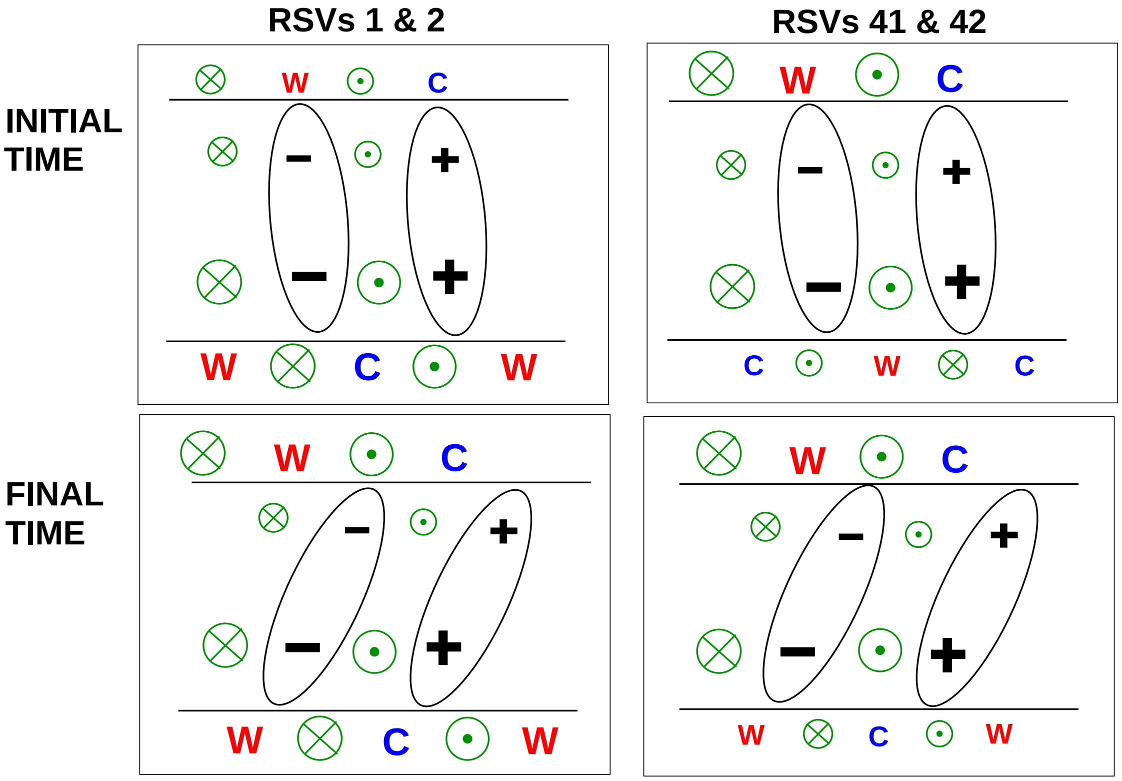


Figure 5.7: *Schematic Diagram showing the evolution of the right singular vectors. The panels on the left show the evolution of RSV 1 (large singular value) and the panels on the right show the evolution of RSV 41 (small singular value), both at the beginning and the end of the assimilation window. The boundary temperature anomalies are indicated by W (warm) and C (cold), and the interior QGPV anomalies are indicated by +(positive) and -(negative). The circles indicate the direction and magnitude of the meridional wind associated with the QGPV anomalies in the interior and the buoyancy anomalies on the boundaries.*

The schematic diagram in Fig. 5.7 summarizes the evolution mechanisms of the RSVs. For RSVs 1&2, the maximum meridional wind (horizontal derivative of streamfunction) lies directly underneath the maximum temperature anomaly on the upper boundary. Thus the lower level circulation intensifies the amplitude of the upper level wave. As the lower level wave has

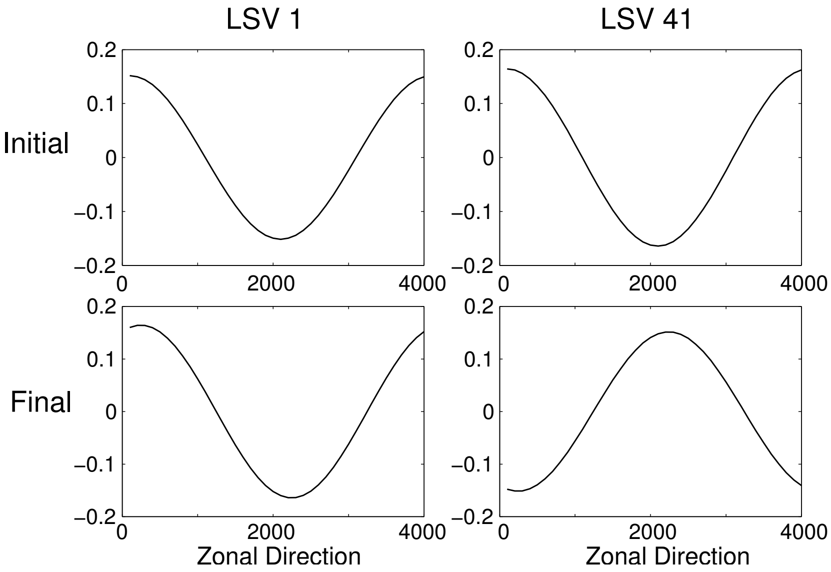


Figure 5.8: *LSVs 1 and 41. The panels on the left show LSV 1 at T+0 and T+6, and the panels on the right show LSV 41 at T+0 and T+6. Note that the LSVs are defined in observation space, and hence are given by the lower level buoyancy field at the initial and the final time.*

a relatively large amplitude this is a large effect. The circulation from the upper level wave also acts to slightly intensify the amplitude of the lower level wave so that the streamfunction field grows. The basic state zonal wind acts on the QGPV field so that it tilts eastwards with height at the final time. Rossby wave propagation on the upper and lower boundaries acts to move the upper level wave westwards and the lower level wave eastwards. Thus, at the final time the structure is similar to a growing normal mode structure. There is a smaller difference between the amplitude of the upper and lower buoyancy waves and the streamfunction field has a larger westward tilt with height than at the beginning of the window.

For RSVs $41\&42$, the position of the upper level wave acts to weaken the lower level wave. As the upper level wave has a relatively large amplitude, this is a large effect. Further, the circulation associated with the QGPV field reinforces this effect so that the wave actually becomes zero and then starts to grow again in the opposite direction. The circulation from the lower level wave also acts to slightly weaken the upper level wave so that the streamfunction field decays. Again, the structure at the final time is similar to a growing normal mode.

To understand how the observational data is projected onto the RSVs, it is of interest to also examine the structure of the LSVs. If an LSV $\mathbf{u}_j$ is in the same direction as the generalized innovation vector $\hat{\mathbf{d}}$, then the corresponding RSV $\mathbf{v}_j$ is given a large weight. The LSVs lie in the observation space which is given by buoyancy on the lower boundary at both the initial and the final time. LSVs $1\&41$ are shown in Fig. 5.8. LSV 1 has a similar structure at T+6 to that at T+0, with a small change in the amplitude. This implies that the weights given to RSVs $1\&2$ are determined by the general shape and position of the observed wave. In contrast, LSV 41 at T+6 has the opposite sign to the wave at T+0. This implies that the weights given to RSVs $41\&42$ are determined by the change in magnitude of the observed wave between T+0 and T+6. If LSV 41 is added to LSV 1, a decaying wave results, and if LSV 41 is subtracted from LSV 1, a growing wave results. Thus, the LSVs with small singular values are detecting the growth or decay of the system.

To summarize, RSVs $1\&2$ contain the information needed to reconstruct the state in the observed regions and to give a growing analysis increment. RSVs $41\&42$ contain the information needed to reconstruct the state in unobserved regions and also to give a decaying analysis increment. In particular, they are needed to detect the growth. The growth may be small in comparison to the amplitude of the wave and hence is harder to detect and therefore a large weight should be given to the background state if the observations are relatively noisy. It is for this reason that RSVs $41\&42$ have a small singular value and thus RSVs $41\&42$ are penalized so that the position of the upper level wave cannot be determined.

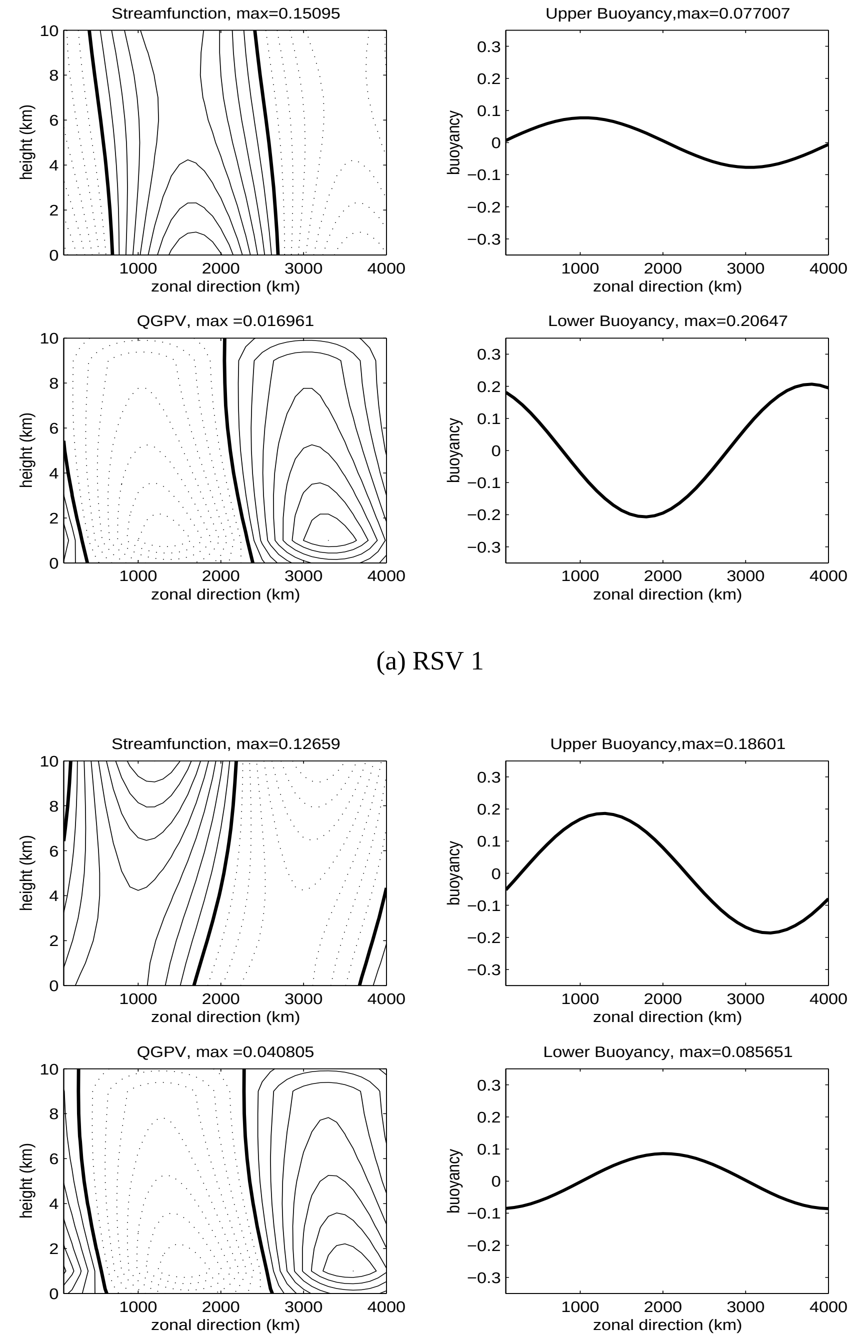
When the assimilation window is longer, there is a more distinct splitting between the weight given to the RSV pairs when growing and decaying modes are observed and the singular value of the RSVs $41\&42$ is larger. These aspects are investigated in the following section.

5.5 Observations Should be Placed Far Apart in Time

In Chapter 3, it was shown that it is best to place the observations as far apart as possible in time. It was also shown that if observations are only at the end of the window, a longer assimilation window will give better results if a growing mode is observed, but worse results if a decaying mode is observed.

The reasons for this are now investigated by repeating the SVD computations but with different temporal observing systems. We first examine the RSVs when a 12 hour assimilation window is used with observations at T+0 and T+12, and we then examine the singular values when the temporal position of the initial observations is changed.

Figure 5.9 shows the structures of the RSVs for a 12 hour assimilation window, with observations of the lower level buoyancy at T+0 and T+12. The structures are very similar to those for a 6 hour window (Fig. 5.5), but there are some important differences. The QGPV field exhibits a greater tilt with height, as the longer window provides a longer time to untilt.



(a) RSV 1

(b) RSV 41

Figure 5.9: *As for Fig. 5.5, but for a 12 hour window.*

Also, there is a smaller difference between the amplitude of the wave on the upper boundary and that on the lower boundary.

The upper level wave in RSV 1 and the lower level wave in RSV 41 are larger for a longer assimilation window. Thus, the RSVs exhibit tilted structures that are more intense. That is, for a longer time window, RSV 1 exhibits a more pronounced vertical structure for growth whilst RSV 41 exhibits a more pronounced vertical structure for decay. Thus, there is a more distinct difference in the vertical structure for the two RSV pairs.

The difference in the amplitude of the buoyancy waves on the upper and lower levels is important for the reconstruction of the wave with the correct vertical structure. If a large weight is given to the background state, a small weight will be given to RSVs $41\&42$ and the analysis increment will be dominated by RSVs $1\&2$. The amplitude of the upper level wave is larger for a longer time window. Thus, a larger analysis increment will be added to the upper boundary, in a position to give growth in the following forecast. Therefore, if a growing mode is observed at the beginning and the end of the window, a longer assimilation window gives a better analysis. However, if a decaying mode is observed at the beginning and the end of the window, a longer assimilation window may give a worse analysis because although a large analysis increment is added, the position of the upper level wave relative to the lower level wave is incorrect.

The temporal position, t_I, of the initial observations is now considered by calculating the SVD of the observability matrix:

$$\hat{\mathbf{H}} = \begin{bmatrix} \mathbf{HM}(t_I, t_0) \\ \\ \mathbf{HM}(t_N, t_0) \end{bmatrix} \tag{5.8}$$

where $\mathbf{M}(t_I, t_0)$ represents the integration of the linear model from the beginning of the assimilation window t_0 to the time of the initial observations t_I, and $\mathbf{M}(t_N, t_0)$ represents the integration of the linear model from t_0 to the end of the assimilation window t_N.

For a fixed assimilation window length, the position of the initial observations has little impact on the structure of the RSVs and so their structures are not shown. There is, however, a dramatic change in the size of the singular values of the second pair of RSVs, as shown in Fig. 5.10.

For both a 6 hour and 12 hour window, the singular value of the second pair of RSVs decreases as the initial set of observations is moved towards the end of the window. If the

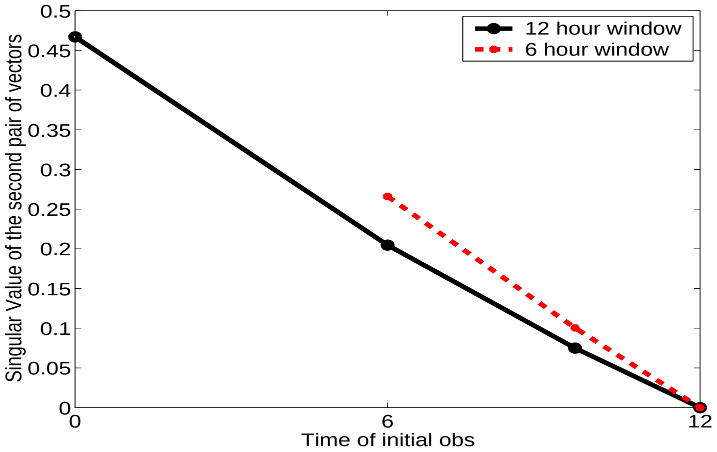


Figure 5.10: *The singular values of the second pair of RSVs (not necessarily* 41&42 *) plotted against the time of the initial observations. The solid line represents the values for a 12 hour window, and the dashed line represents the values for a 6 hour window. In both cases, the final observations are given at T+12.*

initial observations are at the end of the window (T+12), then the singular value is zero. In other words, when there are only observations at the end of the window, there is only one pair of RSVs that contribute to the analysis increment (RSVs 1&2).

This diagram explains why it is better to place the observations as far apart as possible in time. When a large weight is given to the background state, the RSVs with small singular values are penalized. Thus, the weight given to the second pair of RSVs decreases as the initial observations are moved to the end of the assimilation window. Alternatively, as the initial observations are moved to the end of the window, more weight should be given to the observations; this can only occur if the observations become more accurate. With observations at only the end of the window, the analysis increment is always a growing solution. Even if the weight given to the background state is zero, it is not possible to add an analysis increment with a decaying structure.

To summarize, a longer assimilation window produces RSVs with more vertical structure. RSVs 1&2 have a larger amplitude on the upper boundary and RSVs 41&42 have a larger amplitude on the lower boundary. With observations at only the end of the window, RSVs 1&2 alone contribute to the analysis increment. Thus, if a growing analysis increment is required, a longer assimilation window gives better results. This is consistent with Thépaut et al. (1996), where it was shown that the length of the assimilation period is crucial to ensure fully developed dynamical structure functions. However, it is for the same reason that a longer assimilation window will give worse results if the observations are only at the end of the window and a decaying analysis increment is required.

If the observations are further apart in time, the singular value of the second pair of RSVs is larger, so that the weight given to the RSVs is also larger (when considering a significant weight given to the background state). This is consistent with the fact that it is easier to detect the growth of the true state over a longer assimilation window: a larger singular value of the second pair of RSVs indicates that the observations contain more useful information.

5.6 Temporal Weights Given to the Observations

In Chapter 3, it was shown that the analyses with more weight given to the final observations were better than the analyses with more weight given to the initial observations. In particular, it was shown that the analysis of the decaying mode is not improved by giving more weight to the initial observations.

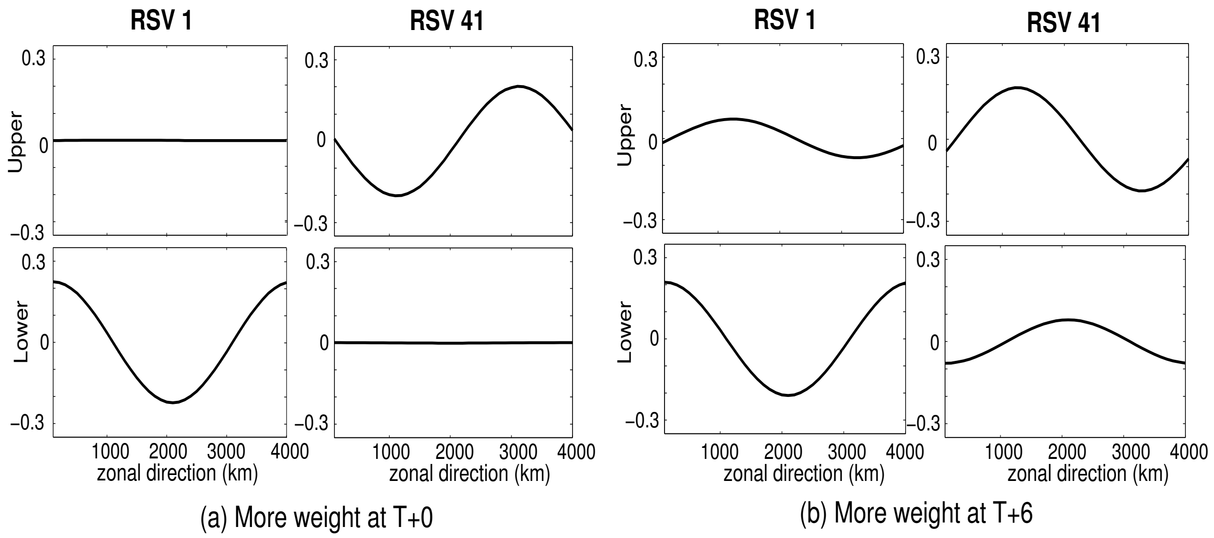
The SVD is now used to understand the effect of the temporal weights given to the observations. We consider a 6 hour window with observations at T+0 and T+6 and with 10 times more weight given to either the initial of the final observations. This is achieved by calculating the SVD of:

$$\begin{bmatrix} w_1\mathbf{H} \\ \\ w_2\mathbf{HM}(t_N, t_0) \end{bmatrix} \tag{5.9}$$

with the scalar weights w_1 and w_2 defined by $w_1 = 1, w_2 = 10$ to give more weight to the final observations and $w_1 = 10, w_2 = 1$ to give more weight to the initial observations.

The structures of the RSVs of the observability matrix that give a large contribution to the analysis increment when the most rapidly growing or decaying Eady wave is observed are now examined. The RSVs for the case when more weight is given to the initial observations are shown in Fig. 5.11(a) and the RSVs for the case when more weight is given to the final observations are shown in Fig. 5.11(b). When more weight is given to the initial observations, there is a large difference between the amplitudes of the upper and lower waves so that the information needed to reconstruct the lower level wave and the information needed to reconstruct the upper level wave is clearly separated. When more weight is given to the final observations, there is a small difference between the amplitudes of the upper and lower waves and thus there is a stronger vertical tilt. These structures are similar to those for a 12 hour window.

If we consider the case where a relatively large weight is given to the background state,



(a) More weight at T+0

(b) More weight at T+6

Figure 5.11: *The upper and lower buoyancy field for RSVs 1 and 41 of the observability matrix for the case of a 6 hour window with observations of the lower boundary buoyancy at T+0 and T+6. (a) The weight given to the initial observations is 10 times greater than that given to the final observations (*$w_1 = 10$*,* $w_2 = 1$*), and (b) the weight given to the final observations is 10 times greater than that given to the initial observations (*$w_1 = 1$*,* $w_2 = 10$*).*

then because RSV $41\&42$ have small singular values, they are strongly penalized. Therefore, the analysis increment is dominated by RSVs $1\&2$. If an eastward tilting analysis increment (growing) is required, then it is better to give more weight to the final time observations. If a westward tilting analysis increment (decaying) is required, then it makes little difference as to whether more weight is given to the initial or the final observations. In both cases, a growing analysis increment will be added to the background state. If more weight is given to the initial observations, a large analysis increment must be added to the background state so that the analysis increment is close to the observations at the beginning of the window. If more weight is given to the final observations, a small amplitude analysis increment may be added to the background state. This increment has a strong vertical tilt so that it grows rapidly during the window so that the analysis is close to the final observations.

If a decaying analysis increment is required, giving more weight to the initial observations does not improve the analysis. The only way to obtain an analysis increment with the required growth rate is to give a relatively small weight to the background state so that the RSVs with small singular values (RSVs $41\&42$) can be included in the analysis increment.

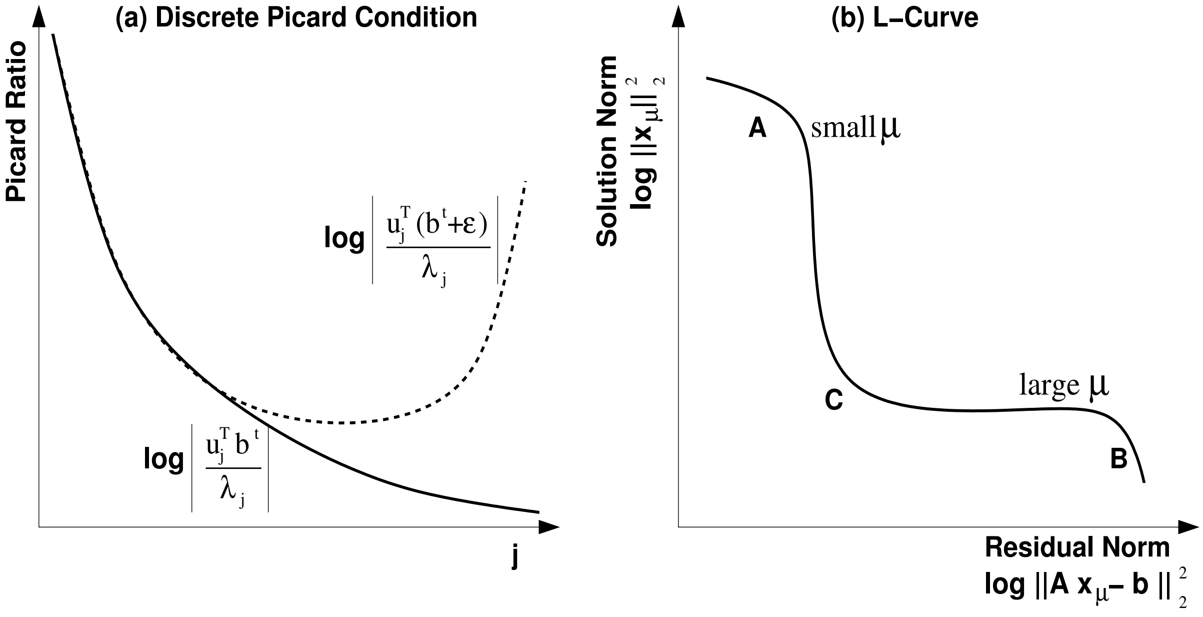

(a) Discrete Picard Condition
Picard Ratio
$\log \left| \frac{u_j^T (b^t+\varepsilon)}{\lambda_j} \right|$
$\log \left| \frac{u_j^T b^t}{\lambda_j} \right|$
j
(b) L-Curve
Solution Norm
$\log ||x_\mu||_2^2$
A
small μ
C
large μ
B
Residual Norm
$\log ||A x_\mu - b||_2^2$

Figure 5.12: *Schematic Diagram illustrating the (a) Discrete Picard condition (b) L-Curve*

5.7 Tikhonov Regularization

In Chapter 4, the SVD and information content concepts that are used in 1D-Var satellite retrievals were extended to 4D-Var. This technique has been used in this chapter to understand how observations are combined with the model dynamics in 4D-Var. One of the results from this study is that the background state is needed to filter the noisy components of the solution. This aspect gives strong links to a method that is widely known in the mathematical literature as Tikhonov Regularization (Aleksandrov, 1976). The method of Tikhonov Regularization is now introduced and then a link between 4D-Var and Tikhonov Regularization is established.

Following Winkler (1997) and Hansen (2001), suppose that the true state $\mathbf{x}^t$ satisfies the matrix equation:

$$\mathbf{A}\mathbf{x}^t = \mathbf{b}^t \tag{5.10}$$

and that the given data $\mathbf{b}$ which is used to infer the true state $\mathbf{x}^t$ has errors ε,

$$\mathbf{b} = \mathbf{b}^t + \varepsilon. \tag{5.11}$$

Then, the inverse problem is to find the state $\mathbf{x}^a$ which minimizes:

$$||\mathbf{A}\mathbf{x} - \mathbf{b}||_2^2. \tag{5.12}$$

This is a discrete ill-posed problem if the solution $\mathbf{x}^a$ is sensitive to the data $\mathbf{b}$. Strictly speaking, an ill-posed problem must have an infinite dimension, however certain finite dimensional problems have similar properties to those of ill-posed problems and are known as discrete ill-posed problems (Hansen, 1992).

Extra information about the desired solution is necessary to make the problem well conditioned. In Tikhonov Regularization, this is achieved by also requiring that the 2-norm of the solution is small. The standard form for Tikhonov Regularization is to minimize:

$$||\mathbf{Ax} - \mathbf{b}||_2^2 + \mu^2 ||\mathbf{x}||_2^2 \tag{5.13}$$

where μ is the regularization parameter which controls the amount of smoothing. The solution $\mathbf{x}_\mu$ is known to be a good approximation to the true solution provided that the exact solution is dominated by the large singular values. That is, if the exact solution satisfies the discrete Picard condition:

The exact coefficients $\mathbf{u}_j^T \mathbf{b}^t$ decay, on average, faster than the singular values λ_j.

Thus, the ratio $\frac{\mathbf{u}_j^T \mathbf{b}^t}{\lambda_j}$ must decay, where $\mathbf{u}_j$ and λ_j are the LSVs and singular values of the matrix $\mathbf{A}$. A schematic diagram illustrating the typical behaviour of the values of the Picard ratios is shown in Fig. 5.12(a). The solid line represents the exact coefficients for the exact data, $\frac{\mathbf{u}_j^T \mathbf{b}^t}{\lambda_j}$. They decrease, so the discrete Picard condition is satisfied. The dashed line represents the coefficients for noisy data, $\frac{\mathbf{u}_j^T (\mathbf{b}^t + \boldsymbol{\varepsilon})}{\lambda_j}$. These values decrease to a minimum and then increase. Thus, if the regularization parameter μ is small, the solution is dominated by the RSVs with very small singular values. The regularization parameter should be chosen so that these components are damped, and so that the approximation is close to the true solution.

The parameter μ controls the amount of smoothing. If too much regularization is imposed, the solution will not fit the data well and $||\mathbf{Ax}_\mu - \mathbf{b}||_2^2$ will be large; but if too little regularization is imposed, the solution will be dominated by the data errors and $||\mathbf{x}_\mu||_2^2$ will be large. Thus, it is important to specify μ well.

The L-Curve is one method that may be used to choose μ, and may also be used to illustrate the fundamental foundations for Tikhonov Regularization. The L-Curve is a parametric plot of $\log ||\mathbf{x}_\mu||_2$ against $\log ||\mathbf{Ax}_\mu - \mathbf{b}||_2$ and is known as an L-Curve due to the L shape of the curve. The log-log scale is used so that the corner of the L-Curve is emphasized. It has been proved (see for example, Winkler (1997) and Hansen (2001)), that if the unperturbed data $\mathbf{b}^t$ satisfies the discrete Picard condition, the noise $\boldsymbol{\varepsilon}$ is an unbiased random vector with a diagonal covariance matrix and $||\boldsymbol{\varepsilon}|| << ||\mathbf{b}^t||$, the L-Curve assumes the shape shown by the schematic diagram in Fig. 5.12 (b). The L-Curve is always concave at the ends near the axes and a corner exists at C. When μ is small (region A), the solution is dominated by the effect of noise, and hence $\log \|\mathbf{A}\mathbf{x}_\mu - \mathbf{b}\|_2^2$ is small and $\log \|\mathbf{x}_\mu\|_2^2$ is large. When μ is large (region B), the solution has been over-smoothed and hence $\log \|\mathbf{A}\mathbf{x}_\mu - \mathbf{b}\|_2^2$ is large and $\log \|\mathbf{x}_\mu\|_2^2$ is small. The appropriate value for μ therefore lies at the corner of the L shape (point C). This is known as the L-Curve criterion.

The 4D-Var algorithm is now explicitly formulated as Tikhonov Regularization. The 4D-Var cost function can be written as (see Section 4.4):

$$J(\mathbf{x}_0) = (\mathbf{x}_0 - \mathbf{x}^b)^T \mathbf{B}^{-1} (\mathbf{x}_0 - \mathbf{x}^b) + (\hat{\mathbf{y}} - \hat{\mathbf{H}}\mathbf{x}_0)^T \mathbf{R}^{-1} (\hat{\mathbf{y}} - \hat{\mathbf{H}}\mathbf{x}_0). \tag{5.14}$$

To account for correlations in the error covariances, let $\mathbf{B} = \sigma_b^2 \boldsymbol{\rho}_B$ and $\mathbf{R} = \sigma_o^2 \boldsymbol{\rho}_R$, then the cost function can be rewritten as:

$$J(\mathbf{x}_0) = \sigma_b^{-2} \|\boldsymbol{\rho}_B^{-\frac{1}{2}} (\mathbf{x}_0 - \mathbf{x}^b)\|_2^2 + \sigma_o^{-2} \|\boldsymbol{\rho}_R^{-\frac{1}{2}} (\hat{\mathbf{H}}\mathbf{x}_0 - \hat{\mathbf{y}})\|_2^2. \tag{5.15}$$

Letting $\boldsymbol{\chi} = \boldsymbol{\rho}_B^{-\frac{1}{2}} (\mathbf{x}_0 - \mathbf{x}^b)$ and $\hat{\mathbf{d}} = \hat{\mathbf{y}} - \hat{\mathbf{H}}\mathbf{x}^b$, then

$$\sigma_o^2 J(\boldsymbol{\chi}) = \mu^2 \|\boldsymbol{\chi}\|_2^2 + \|\boldsymbol{\rho}_R^{-\frac{1}{2}} \hat{\mathbf{H}} \boldsymbol{\rho}_B^{\frac{1}{2}} \boldsymbol{\chi} - \boldsymbol{\rho}_R^{-\frac{1}{2}} \hat{\mathbf{d}}\|_2^2 \tag{5.16}$$

where again, $\mu^2 = \frac{\sigma_o^2}{\sigma_b^2}$.

This is in the standard form for Tikhonov Regularization (5.13) where $\mathbf{A} = \boldsymbol{\rho}_R^{-\frac{1}{2}} \hat{\mathbf{H}} \boldsymbol{\rho}_B^{\frac{1}{2}}$ and $\mathbf{b} = \boldsymbol{\rho}_R^{-\frac{1}{2}} \hat{\mathbf{d}}$. Therefore, it is now possible to apply our understanding of the discrete Picard condition and the L-Curve to 4D-Var.

The 4D-Var experiments in Section 3.2.3 (for example, Fig. 3.4) considered the true state given by the most unstable Eady wave and a background state with a phase error. Observations of the lower level buoyancy were given at T+0 and T+6 with noise with a Gaussian distribution with $\sigma = 1$. Background error correlations were applied to the lower level buoyancy with $l = 10\Delta x$. Using the known true state, the statistically optimal inverse variances were $\sigma_o^{-2} = 1$ and $\sigma_b^{-2} = 0.08$. An L-Curve for this experiment is now found, to determine whether the ratio $\mu^2 = \sigma_o^2/\sigma_b^2$ can be found without the knowledge of the true state. The curve is found by repeating the 4D-Var analysis with different values for σ_b^{-2} and with $\sigma_o^{-2} = 1$ fixed.

The L-Curve, shown in Fig. 5.13 for this case is a parametric plot of $\log \|\boldsymbol{\rho}_B^{-\frac{1}{2}} (\mathbf{x}^a(\mu) - \mathbf{x}^b)\|_2$ and $\log \|\hat{\mathbf{y}} - \hat{\mathbf{H}}\mathbf{x}^a(\mu)\|_2$, or equivalently of $\log \sqrt{(\sigma_b^2 J_{min}^b)}$ and $\log \sqrt{(\sigma_o^2 J_{min}^o)}$. The

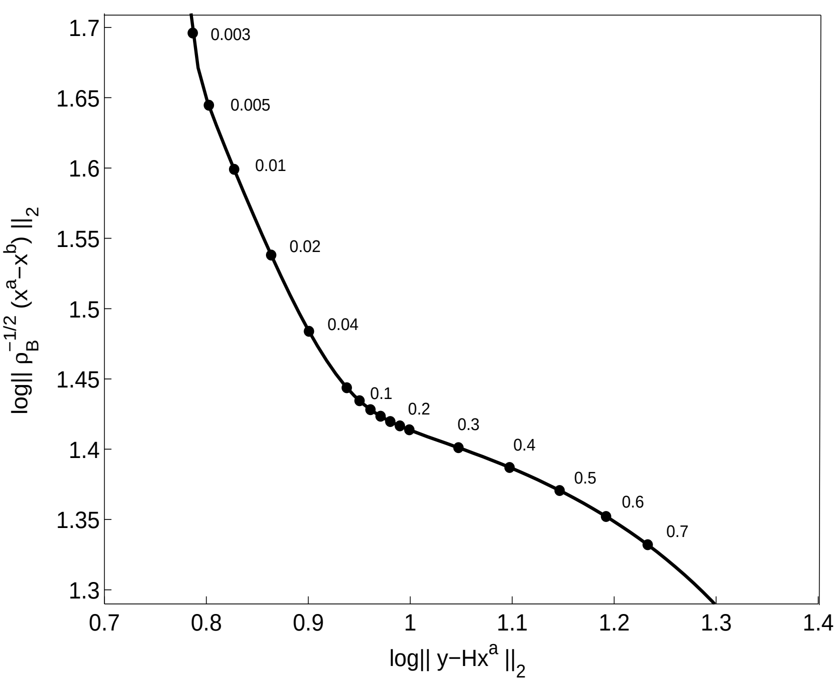


Figure 5.13: *The L-Curve: a parametric plot of* $\log\sqrt{\sigma_b^2 J^b_{min}} = \log\|\boldsymbol{\rho}_B^{-\frac{1}{2}}(\mathbf{x}^a_\mu - \mathbf{x}^b)\|_2$ *and* $\log\sqrt{\sigma_o^2 J^o_{min}} = \log\|\hat{\mathbf{H}}\mathbf{x}^a_\mu - \hat{\mathbf{y}}\|_2$. *The values of* $\sigma_b^{-2} = \mu^2$ *are written by the side of each point and* σ_o^{-2} *is fixed.*

value of μ^2 at the corner of the curve is 0.14; this is slightly larger than the statistically optimal value of 0.08. Thus, the L-Curve does indeed find a parameter in the correct vicinity of the optimal value, but is perhaps slightly too large. This oversmoothing is consistent with the results found by Hansen (2001).

The values of the Picard ratios for 4D-Var with no correlations were shown in Fig. 5.3. The values for perfect observations show two spikes and do not decay on average with the singular vector index. That is, the underlying true solution does not satisfy the discrete Picard condition. It is for this reason that 4D-Var is not always able to reconstruct the unobserved regions and that the analysis is a poor approximation to the true state. When the observations are noisy, the RSVs with small singular values must be damped. However, in the process of damping the unwanted RSVs, the background state also damps the important information.

The SVD computations shown in this chapter have not considered correlations; these will be considered in the next chapter. It will be shown that the effect of the correlations is to re-order the RSVs so that the discrete Picard condition can be satisfied.

5.8 Conclusions

In this chapter we have used the singular value decomposition (SVD) to understand how 4D-Var combines the information from observations with the model dynamics. For comparison with the experiments in Chapter 3, we have examined the case where only the lower boundary is observed and either the most rapidly growing or decaying Eady wave is observed.

The SVD of the observability matrix showed that for this case, only two pairs of RSVs are needed to form the correct analysis increment. The first pair have a large singular value and are needed to reconstruct the state in the observed regions and also to give a growing analysis increment. The second pair have a small singular value and are needed to reconstruct the state in the unobserved regions and to give a decaying analysis increment.

The Picard ratio values for perfect observations do not decay with the singular vector index. This means that when noise is added to the observations, and other RSVs are given a large weight, it is impossible to damp the unwanted noisy RSVs whilst retaining the second pair of RSVs. It is for this reason that the background state strongly penalizes the information needed to reconstruct the state in the unobserved regions and also to give a decaying analysis increment.

The Picard ratio values for noisy observations increase with the singular vector index. Thus, the RSVs with extremely small singular values dominate the analysis increment if there is no background state. However, it is these RSVs which are strongly damped by the background state. These RSVs contain small scale structures with a large amplitude in the unobserved region. It is for this reason that the unobserved regions are particularly sensitive to noise on the observations and also to the weight given to the background state.

We have shown that the choice of the regularization parameter μ is essential in generating an analysis that has extracted the maximum amount of available information but that does not contain unphysical structures due to noise. By relating 4D-Var to Tikhonov Regularization, we have shown that it is possible to use the data (the observations and the background state), to find the appropriate choice for the regularization parameter. This has been demonstrated using the L-Curve.

If the assimilation window length is increased, the RSVs develop more vertical structure. When the initial and final observations are close together in time of the background state is given a large weight, the growing RSVs dominate the analysis increment and therefore a longer assimilation window gives better results if a growing analysis increment is required, but worse results if a decaying analysis increment is required.

If the initial and final observations are moved further apart in time, the singular value of the second pair of RSVs increases. This indicates that the observations provide more useful information when they are far apart in time.

The SVD computations have provided a new understanding of 4D-Var, and in particular the extent to which 4D-Var can reconstruct the state unobserved regions and generate the correct vertical structures, and also how these benefits can be maximized. However, only very simple experiments have been examined and it is important to ask how these results might differ for more realistic data assimilation cases.

The section on Tikhonov Regularization (Section 5.7) showed that correlations may also be considered by formulating the problem with a change of variable. The background error correlations play a key role in spreading the information from observations in both dense and sparse data regions. In all the experiments so far we have only considered the case where the true state is given by the most rapidly growing or decaying Eady wave and where only the lower boundary buoyancy is observed. It is therefore important to investigate 4D-Var for different true states and observing systems. In particular, it is important to consider the case where the true state exhibits non-modal growth, and to consider also vertical lines of observations, for example given by radiosondes. These issues will be investigated in the next chapter.

Chapter 6

Extension to More Realistic Cases

The experiments in previous chapters considered the assimilation of a full horizontal line of observations of the lower level buoyancy at two time levels, and where the true state was given by either the most rapidly growing or decaying Eady wave. This has been useful in understanding how 4D-Var can reconstruct the upper level wave, and does simulate a real data case where only surface observations (or upper air) observations are used. Both the 4D-Var analyses and the SVD computations have illustrated how 4D-Var combines the information from observations with the model dynamics. However, the experiments were highly idealized. The purpose of this chapter is to understand whether the previous results can be applied to an operational 4D-Var algorithm. This is investigated by extending the experiments to more realistic, although still idealized cases.

In operational data assimilation, the observational data contains many vertical temperature profiles, for example, from radiosondes or satellites. Therefore, it is important to understand how 4D-Var uses the information from vertical temperature profiles and how this compares and contrasts with the assimilation of horizontal lines of observations. Since a vertical profile of observations samples the vertical structure of the atmosphere, it is to be expected that 4D-Var will be able to use this data to generate analyses with good vertical structures.

All the previous experiments in this thesis have considered the true states given by the most rapidly growing or decaying Eady wave. The vertical structure of these normal modes is vital for the vertical coupling between the upper and lower boundary waves and hence for the growth or decay of the waves. Chapter 1 also discussed perturbations which can grow faster than the exponential growth of normal modes. The vertical structure of these perturbations is very different to that of the normal modes as they are characterized by interior QGPV structures with small spatial scales. Therefore, it is important to understand whether 4D-Var is able to generate analyses with such non-modal structures. Vertical profiles sample such structures well, and therefore it is expected that 4D-Var should be able to extract the information to generate analyses with structures necessary for non-modal growth.

Vertical temperature profiles give sparse data in the horizontal. Therefore, the interpolating effect of background error correlations is important in such cases. The 4D-Var experiments in Section 3.2.3 showed that correlations play an important role in creating a smooth analysis when a full line of noisy observations is assimilated. None of the SVD experiments, however, have considered correlations or sparse observations. Thus, it is important to investigate how the SVD results differ when correlations are included and also when sparse observations are assimilated.

All the experiments in this chapter consider the assimilation of observations of the interior buoyancy field. In the first section, the technique used to handle observations of the interior buoyancy is described and the assimilation of interior buoyancy is compared with the assimilation of the lower level buoyancy. In the second section, the technique to compute the SVD incorporating correlations is described and the filtering and interpolating effects of correlations are investigated. In the third section, the assimilation of observations from a true state with non-modal growth is compared with that from modal growth. The fourth section considers the assimilation of observations from different observing systems. First, the experiments are extended to consider the assimilation of two horizontal lines of observations, and then to consider the assimilation of two vertical lines of observations.

6.1 Observing Interior Buoyancy

This section describes the assimilation of observations of the interior buoyancy. Such observations will be used for all the experiments in the rest of this chapter. The assimilation of a horizontal line of interior buoyancy is compared with that of a horizontal line of buoyancy on the lower boundary, by examining the SVD of the observability matrix:

$$\hat{\mathbf{H}} = \begin{bmatrix} \mathbf{H} \\ \\ \mathbf{HM} \end{bmatrix} \tag{6.1}$$

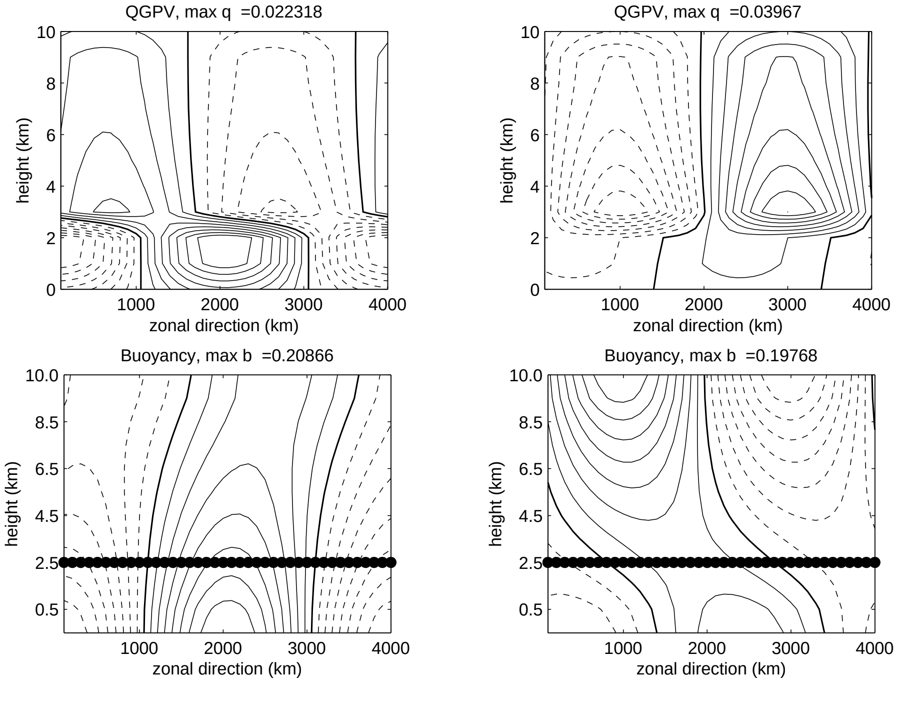
where $\mathbf{H}$ is the observation operator and $\mathbf{M}$ is the linear Eady model.

The SVD of the observability matrix gives an indication of the information that is contained in the observations when they are assimilated using 4D-Var. The background state penalizes the RSVs with small singular values. This is important in removing the noisy components from the analysis increment, but may also remove some of the important information that is contained in the observations.

In the Eady model, when the lower boundary buoyancy is observed, the observation operator $\mathbf{H}$ is simply a matrix of ones and zeros, but when the interior buoyancy is observed, the observation operator must contain diagnostic equations which link the interior buoyancy to the control variables (QGPV and buoyancy on the boundaries). The non-dimensional buoyancy is given by the vertical derivative of the non-dimensional streamfunction field. Therefore, the observation operator first uses the QGPV and buoyancy on the boundaries to find the streamfunction field, and then the interior buoyancy is given by the vertical derivative of the streamfunction field. This is described in further detail in Appendix A.

The structures of the RSVs, for the assimilation of the interior buoyancy at a height of 2.5km, are now examined. RSVs 2 and 10 are shown in Fig. 6.1. The first pair have an eastward tilting buoyancy field with a maximum in amplitude on the lower boundary. The second pair have a westward tilting buoyancy field with a maximum in amplitude on the upper boundary. Although the observations are closer to the middle of the domain than in the previous experiments, there is still a distinction between the information needed to reconstruct the boundary that is closest to the observations and the information that is needed to reconstruct the boundary that is furthest from the observations. Also, the information to give a growing analysis increment is found in the first RSV pair, whilst the information to give a decaying analysis increment is found in the second RSV pair. The difference between these RSVs and the RSVs for observations of the lower boundary is the dipole structure in the QGPV field. Such a structure is expected from the relationship between temperature and QGPV. A warm buoyancy anomaly is related to a negative QGPV anomaly above and positive QGPV anomaly below, as shown by the schematic diagrams in Fig. 1.4. It is expected that this structure will enable the reconstruction of small-scale structures in QGPV that are necessary for non-modal growth.

The effect of the height of the horizontal line of observations is now considered. The singular values of the two RSV pairs that are necessary to form the analysis increment when the most rapidly growing or decaying Eady wave is observed are shown in Fig. 6.2. As the height



(a) RSV 2 and 3, $\lambda = 0.9689$ (b) RSV 10 and 11, $\lambda = 0.2219$

Figure 6.1: *Right singular vectors of the observability matrix that give a large contribution to the analysis increment for the analysis of the most rapidly growing or decaying Eady wave. Observations of the interior buoyancy at a height of 2.5km are given at the beginning and the end of a 6 hour window.*

of the horizontal line is increased from 0.5km to 4.5km, the singular value of the first pair of RSVs decreases. The structure of the RSVs also changes as the height is increased (not shown). When the observations are at 0.5km, the first pair of RSVs has a relatively large amplitude on the lower boundary and the second pair have a relatively large amplitude on the upper boundary. When the observations are moved higher to 4.5km, the RSVs have nearly equal amplitudes on the upper and lower boundaries. As the height of the observations is increased even further, the first pair of RSVs gain a larger amplitude on the upper boundary and the second pair gain a larger amplitude on the lower boundary. Thus, the first pair of RSVs contains the information needed to reconstruct the boundary which is closest to the observations, whilst the second pair of RSVs contains the information needed to reconstruct the boundary which is furthest from the observations. At all heights, the first pair of RSVs has an eastward tilting buoyancy field and the second pair of RSVs has a westward tilting buoyancy field. Thus, the second pair of RSVs always contains the information to give a decaying analysis increment. When the observations are in the middle of the domain, there is a smaller difference between the singular values so

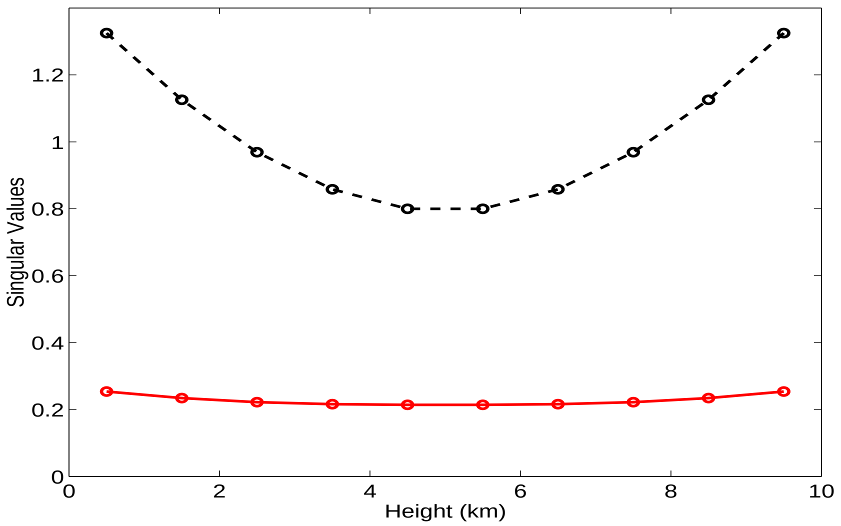


Figure 6.2: *Singular values of the two pairs of RSVs that contribute to the analysis increment when the most rapidly growing or decaying Eady wave is observed. Observations of a horizontal line of the interior buoyancy are given at T+0 and T+6 and at a specified height (the abscissa). The black dashed line shows the singular values of the first pair of RSVs and the red solid line shows the singular values of the second pair of RSVs.*

there is less distinction between the information needed to infer the growing and decaying parts of the analysis increment. The singular value of these 'decaying' RSVs is always smaller than that for the 'growing' RSVs, even when the observations are in the middle of the domain. Thus, the background state will continue to strongly penalize the decaying part of the analysis increment.

To summarize, the SVD of the observability matrix for a horizontal line of interior buoyancy observations has shown that there are still two pairs of RSVs needed to form the analysis increment. Therefore, the conclusions from the previous chapter can be applied to the assimilation of interior buoyancy. That is, the first pair of RSVs with a large singular value, is needed to reconstruct the state on the boundary closest to the observations and to give a growing analysis increment whilst the second pair of RSVs, with a small singular value, is needed to reconstruct the state on the boundary furthest from the observations and to give a decaying analysis increment. As the background state penalizes the RSVs with small singular values, this implies that the background state may penalize the information needed to reconstruct the state on the boundary furthest from the observations and also penalize the decaying part of the analysis increment. The structures of the RSVs show that if the horizontal line is moved from near the lower boundary to the middle of the domain, there is less of a distinction between the information to reconstruct the upper level wave and the information needed to reconstruct the lower level wave. However, the singular values show that there is still a distinction between the information needed to reconstruct the growing and decaying parts of the analysis increment.

6.2 Background Error Correlations

The SVD experiments are now extended to understand the filtering and interpolating effects of correlations. The section begins by rewriting the 4D-Var analysis increments in terms of the RSVs of what will be known as the normalized observability matrix.

Consider the minimization of the 4D-Var cost function:

$$J(\mathbf{x}_0) = (\mathbf{x}_0 - \mathbf{x}^b)^T \mathbf{B}^{-1}(\mathbf{x}_0 - \mathbf{x}^b) + (\hat{\mathbf{y}} - \hat{\mathbf{H}}\mathbf{x}_0)^T \mathbf{R}^{-1}(\hat{\mathbf{y}} - \hat{\mathbf{H}}\mathbf{x}_0) \tag{6.2}$$

where the generalized observation vector $\hat{\mathbf{y}}$ and the observability matrix $\hat{\mathbf{H}}$ are given by:

$$\hat{\mathbf{y}} = \begin{bmatrix} \mathbf{y}_0 \\ \\ \mathbf{y}_N \end{bmatrix} \qquad \hat{\mathbf{H}} = \begin{bmatrix} \mathbf{H} \\ \\ \mathbf{HM} \end{bmatrix}. \tag{6.3}$$

We assume that the background state and observations have uniform error variances, σ_b^2 and σ_o^2 so that the covariances may be split into the variance and correlation components:

$$\mathbf{B} = \sigma_b^2 \boldsymbol{\rho}_B \qquad \mathbf{R} = \sigma_o^2 \boldsymbol{\rho}_R \tag{6.4}$$

where $\boldsymbol{\rho}_B$ and $\boldsymbol{\rho}_R$ are the background and observation error correlations respectively. Defining the co-ordinate transformation $\boldsymbol{\chi} = \boldsymbol{\rho}_B^{-\frac{1}{2}}(\mathbf{x}_0 - \mathbf{x}^b)$, then the cost-function may be written in the standard, preconditioned form:

$$\sigma_o^2 J(\boldsymbol{\chi}) = \mu^2 \|\boldsymbol{\chi}\|_2^2 + \|\boldsymbol{\rho}_R^{-\frac{1}{2}} \hat{\mathbf{H}} \boldsymbol{\rho}_B^{\frac{1}{2}} \boldsymbol{\chi} - \boldsymbol{\rho}_R^{-\frac{1}{2}} \hat{\mathbf{d}}\|_2^2 \tag{6.5}$$

where the generalized innovation vector is $\hat{\mathbf{d}} = \hat{\mathbf{y}} - \hat{\mathbf{H}}\mathbf{x}^b$ and $\mu = \sigma_o/\sigma_b$.

In the previous chapters, the SVD of the observability matrix $\hat{\mathbf{H}}$ was examined. However, by using the co-ordinate transformation, the cost function has now been transformed such that the analysis increments may be written as a linear combination of the RSVs of $\boldsymbol{\rho}_R^{-\frac{1}{2}} \hat{\mathbf{H}} \boldsymbol{\rho}_B^{\frac{1}{2}}$ which we will call the ‘Normalized Observability matrix’. The term normalized is used to be consistent with the terms used in Rabier et al. (2002) where the SVD of a normalized Jacobian matrix (or equivalently, a normalized observation operator) is examined.

If $\mathbf{u}_j$, $\mathbf{v}_j$, λ_j and r now denote the LSVs, RSVs, singular values and rank of the normalized observability matrix, $\boldsymbol{\rho}_R^{-\frac{1}{2}}\hat{\mathbf{H}}\boldsymbol{\rho}_B^{\frac{1}{2}}$, then the analysis increments can be written as:

$$\boldsymbol{\chi}^a = \boldsymbol{\rho}_B^{-\frac{1}{2}}(\mathbf{x}_0^a - \mathbf{x}^b) = \sum_{j=1}^{r} \frac{\lambda_j^2}{\mu^2 + \lambda_j^2} \frac{\mathbf{u}_j^T \boldsymbol{\rho}_R^{-\frac{1}{2}}\hat{\mathbf{d}}}{\lambda_j}\mathbf{v}_j \tag{6.6}$$

Thus, the RSVs of the normalized observability matrix determine the structure of the analysis increments when correlations are included.

In the following, observation correlations are not considered, so $\boldsymbol{\rho}_R = \mathbf{I}$. However, it is interesting to note that the impact of temporal observation error correlations may be investigated using this technique. Such correlations are highly relevant for the assimilation of satellite data, but at the present time, it is unclear how such correlations should be treated.

The matrix $\hat{\mathbf{H}}\boldsymbol{\rho}_B^{\frac{1}{2}}$ first needs to be calculated before the SVD can be computed. In particular, a method is required to compute $\boldsymbol{\rho}_B^{\frac{1}{2}}$. In this thesis, the background error correlations are defined as:

$$\boldsymbol{\rho}_B^{-1} = w_0\mathbf{I} + w_1(\mathbf{L}_{xx})^2 \tag{6.7}$$

where w_0 and w_1 are positive constants depending on the correlation length l, and $\mathbf{L}_{xx}$ is a finite difference second derivative matrix (see equation (2.50)). There are many decompositions which satisfy $\boldsymbol{\rho}_B = \mathbf{C}\mathbf{C}^T$. Here, we choose to find the real, symmetric positive definite square root of $\boldsymbol{\rho}_B$ using the Schur decomposition, following Higham (1984) and Strang (1986). This derivation is simplified considerably by first noting the properties of $\boldsymbol{\rho}_B^{-1}$:

1. $\boldsymbol{\rho}_B^{-1}$ is a real symmetric matrix ($n \times n$),
2. the diagonal elements of $\boldsymbol{\rho}_B^{-1}$ are all positive,
3. $\boldsymbol{\rho}_B^{-1}$ is strictly diagonally dominant[1].

[1] The matrix $\mathbf{A}$ is strictly diagonally dominant if and only if

$$|a_{ii}| > \sum_{\substack{j=1 \\ j\neq 1}}^{n} |a_{ij}| \quad i = 1, \ldots, n.$$

where a_{ij} is the element of A on the ith row and jth column.

From these three properties, it can be shown that $\boldsymbol{\rho}_B^{-1}$ is also symmetric positive definite[2].

As $\boldsymbol{\rho}_B^{-1}$ is symmetric positive definite, it is invertible, and $\boldsymbol{\rho}_B$ is also symmetric positive definite (e.g. Noble and Daniel, 1988).

Any square matrix $\boldsymbol{\rho}_B$ may be written as a Schur decomposition (e.g. Atkinson, 1989):

$$\boldsymbol{\rho}_B = \mathbf{V}\mathbf{D}\mathbf{V}^T \tag{6.8}$$

where $\mathbf{V}$ is a unitary matrix and $\mathbf{D}$ is an upper triangular matrix whose diagonal elements are the eigenvalues of $\boldsymbol{\rho}_B$.

As $\boldsymbol{\rho}_B$ is symmetric, then the eigenvalues of $\boldsymbol{\rho}_B$ are real and the eigenvectors form an orthonormal basis. This means that $\mathbf{D}$ is then a diagonal matrix with the eigenvalues of $\boldsymbol{\rho}_B$ on the diagonal, and the eigenvectors of $\boldsymbol{\rho}_B$ form the orthonormal columns of $\mathbf{V}$.

As $\boldsymbol{\rho}_B$ is symmetric positive definite, then the eigenvalues of $\boldsymbol{\rho}_B$ are also all positive. Thus, the square roots of the eigenvalues are real. This means that a real square root of $\boldsymbol{\rho}_B$ can be defined as:

$$\boldsymbol{\rho}_B^{\frac{1}{2}} = \mathbf{V}\mathbf{D}^{\frac{1}{2}}\mathbf{V}^T \tag{6.9}$$

where $\mathbf{D}^{\frac{1}{2}}$ is a diagonal matrix whose diagonal elements are square roots of the eigenvalues of $\boldsymbol{\rho}_B$. This may be verified by forming $\boldsymbol{\rho}_B^{\frac{1}{2}}\boldsymbol{\rho}_B^{\frac{1}{2}} = (\mathbf{V}\mathbf{D}^{\frac{1}{2}}\mathbf{V}^T)(\mathbf{V}\mathbf{D}^{\frac{1}{2}}\mathbf{V}^T) = \mathbf{V}\mathbf{D}\mathbf{V}^T = \boldsymbol{\rho}_B$.

There exist both positive and negative square roots of the eigenvalues, so that there are 2^n possible square roots (Higham, 1984). Here, the positive square roots are chosen so that the unique, real and positive definite square root is

$$\boldsymbol{\rho}_B^{\frac{1}{2}} = \mathbf{V}\mathbf{D}^{\frac{1}{2}}\mathbf{V}^T. \tag{6.10}$$

where $\mathbf{D}^{\frac{1}{2}}$ is a diagonal matrix whose elements are the positive square roots of the eigenvalues.

The matrix $\boldsymbol{\rho}_B^{-1}$ is inverted using the NAG routine nag_sym_mat_inv, (NAG) which computes the inverse of a real symmetric matrix. The Schur decomposition of $\boldsymbol{\rho}_B$ is then found using nag_sym_eig_all, which computes all the eigenvalues and eigenvectors of a real symmetric matrix. The eigenvalues and eigenvectors are then used to compute $\boldsymbol{\rho}_B^{\frac{1}{2}}$ using (6.9), which

[2] The matrix $\mathbf{A}$ is symmetric positive definite if and only if

$$\mathbf{x}^T\mathbf{A}\mathbf{x} = \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}x_ix_j > 0 \;\; \forall \mathbf{x} \neq 0, \;\; \mathbf{x}\epsilon\mathbb{R} \text{ and } \mathbf{A} = \mathbf{A}^T.$$

is then premultiplied by $\hat{\mathbf{H}}$.

6.2.1 Relationship to Optimal Perturbations

It was briefly mentioned in Section 4.5 that the appropriate metrics for the RSVs are the background and observation error covariances, although this was not justified. The relationship between the RSVs and optimal perturbations are now reconsidered to show that these are indeed the relevant metrics.

Optimal perturbations maximize the ratio:

$$\text{growth} = \frac{\|\mathbf{P}\mathbf{x}_N\|_{\mathbf{E}}}{\|\mathbf{x}_0\|_{\mathbf{C}}} \tag{6.11}$$

where $\mathbf{P}$ is an operator which reduces the n-dimensional vector to a vector with a smaller dimension, and $\mathbf{C}$ and $\mathbf{E}$ are the initial and final time norms. These norms are now chosen to be the background and observation inverse error covariances, $\mathbf{B}^{-1}$ and $\mathbf{R}^{-1}$ respectively and the observation operator $\mathbf{H}$ is used as $\mathbf{P}$ so that:

$$\text{growth} = \frac{< \mathbf{HMx}_0; \mathbf{R}^{-1}\mathbf{HMx}_0 >}{< \mathbf{x}_0; \mathbf{B}^{-1}\mathbf{x}_0 >}. \tag{6.12}$$

Letting $\mathbf{R} = \sigma_o^2\boldsymbol{\rho}_R$ and $\mathbf{B} = \sigma_b^2\boldsymbol{\rho}_B$, then

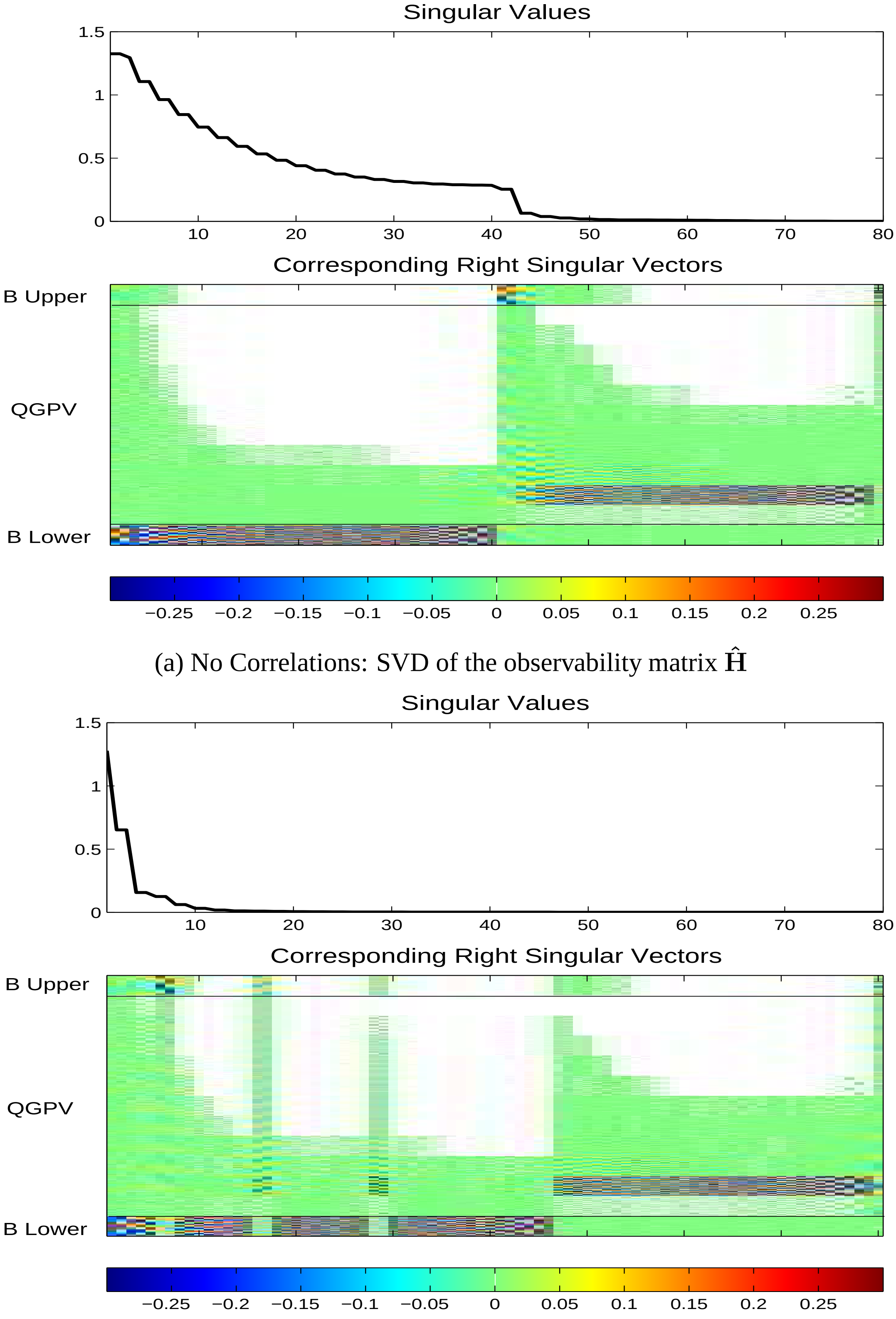
$$\text{growth} = \frac{1}{\mu^2}\frac{< \boldsymbol{\rho}_R^{-\frac{1}{2}}\mathbf{HMx}_0; \boldsymbol{\rho}_R^{-\frac{1}{2}}\mathbf{HMx}_0 >}{< \boldsymbol{\rho}_B^{-\frac{1}{2}}\mathbf{x}_0; \boldsymbol{\rho}_B^{-\frac{1}{2}}\mathbf{x}_0 >}. \tag{6.13}$$

Using the co-ordinate transformation $\boldsymbol{\chi} = \boldsymbol{\rho}_B^{-\frac{1}{2}}\mathbf{x}_0$ then

$$\text{growth} = \frac{1}{\mu^2}\frac{< \boldsymbol{\rho}_R^{-\frac{1}{2}}\mathbf{HM}\boldsymbol{\rho}_B^{\frac{1}{2}}\boldsymbol{\chi}; \boldsymbol{\rho}_R^{-\frac{1}{2}}\mathbf{HM}\boldsymbol{\rho}_B^{\frac{1}{2}}\boldsymbol{\chi} >}{< \boldsymbol{\chi}; \boldsymbol{\chi} >}. \tag{6.14}$$

so that these are also the RSVs of $\boldsymbol{\rho}_R^{-\frac{1}{2}}\hat{\mathbf{H}}\boldsymbol{\rho}_B^{\frac{1}{2}}$ where $\hat{\mathbf{H}} = \mathbf{HM}$.

Thus the optimal perturbations of the model $\mathbf{M}$ that are found using the operator $\mathbf{H}$ and with the metrics given by the background error correlations at the initial time and the observation error correlations at the final time are equivalent to the RSVs of the normalized observability matrix $\boldsymbol{\rho}_R^{-\frac{1}{2}}\hat{\mathbf{H}}\boldsymbol{\rho}_B^{\frac{1}{2}}$.



(a) No Correlations: SVD of the observability matrix $\hat{\mathbf{H}}$

(b) With Correlations: SVD of the normalized observability matrix $\hat{\mathbf{H}}\boldsymbol{\rho}_B^{\frac{1}{2}}$

Figure 6.3: *Singular values λ and $\mathbf{V}$ matrix of the (a) observability matrix and (b) normalized observability matrix, when there is a full line of observations of the interior buoyancy at a height of 0.5km, at T+0 and T+6. The correlations are applied to every level with a length scale of $l = 10\Delta x$.The details are as for Fig. 5.1.*

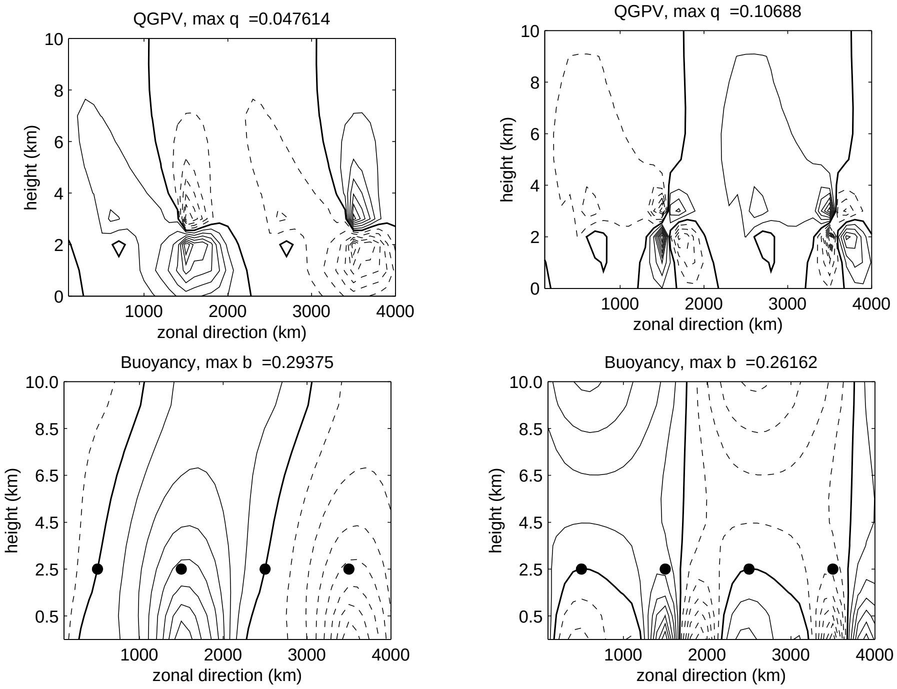
6.2.2 Impact of Correlations on the Assimilation of Dense Observations

The 4D-Var experiments in Section 3.2.3 showed that when noisy observations are assimilated, the background state must be given enough weight to penalize the noise and to give a smooth analysis. However, the background state also acts to penalize the information needed to reconstruct the state in unobserved regions. When correlations were added in the background error covariance matrix, it was possible to extract more of the information needed to reconstruct the upper level wave whilst giving a smooth analysis. Thus, background error correlations act to filter the noise from observations that have a dense spatial distribution.

The SVD computations in Section 5.3 showed that with perfect observations of the most rapidly growing or decaying Eady wave, two pairs of RSVs were required to form the analysis increment. However, with noisy observations, many of the other RSVs, with smaller-scale spatial structures were also given a large weight. The filtering effect of correlations is now understood by computing the SVD of the normalized observability matrix.

The SVDs of $\hat{\mathbf{H}}$ and $\hat{\mathbf{H}}\boldsymbol{\rho}_B^{\frac{1}{2}}$ for the case with a horizontal line of interior buoyancy observations at a height of 0.5km are shown in Fig. 6.3. If the true state is given by the most rapidly growing Eady wave and the background state has a phase error, then with no correlations, the RSVs needed to form the analysis increment are given by RSVs $1\&2$ and $41\&42$. But with correlations, the RSVs needed to form the analysis increment are given by RSVs $2\&3$ and $5\&6$. This is because the RSVs with small scale structures and a large amplitude on the lower boundary are now associated with very small singular values. Thus, the effect of the correlations is to re-order the RSVs so that the RSVs with large scales are associated with the largest singular values. Note that the second pair of RSVs still have a small singular value; this is important as there is no change in the observations, and hence there is also no change in the information that can be extracted from the observation. The re-ordering of the RSVs means that the regularization parameter μ can be chosen so that the RSVs that are needed to reconstruct the upper level wave are given a large weight, but so that the RSVs with small scale structures are strongly penalized by the background state.

The SVD has shown that the background error correlations act to bias the analysis increments towards the spatial structures that are expected. This is achieved by strongly penalizing the error structures that are not expected. In this case, the unexpected structures are the RSVs with small spatial scales. When correlations are included, these RSVs have smaller singular values and are therefore penalized more by the background state.



(a) RSV 2 and 3, $\lambda = 0.3$

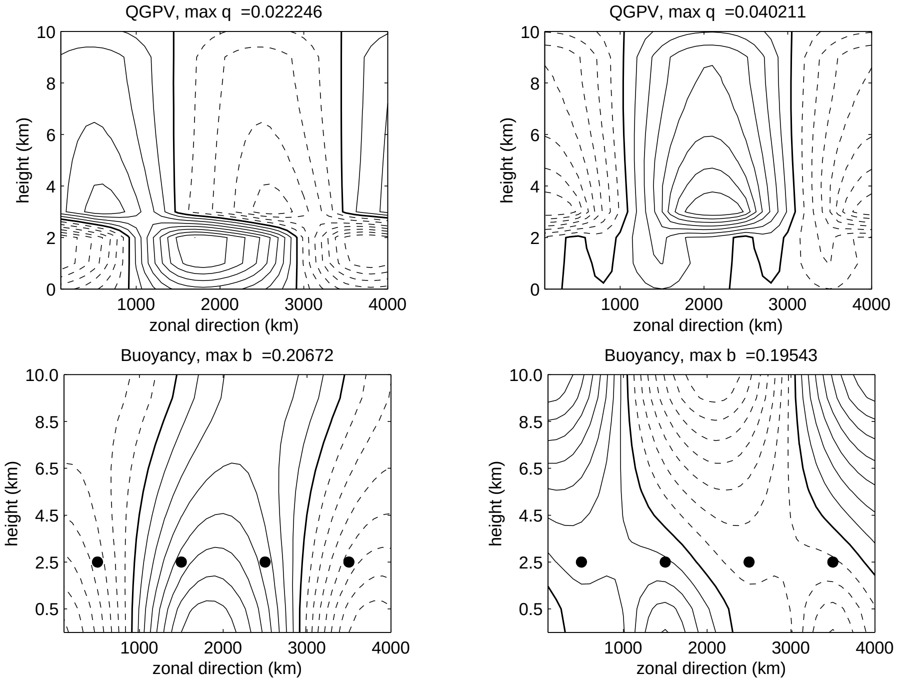
(b) RSV 5 and 6, $\lambda = 0.11$

Figure 6.4: *As for Fig. 6.1 but with four sparse observations at T+0 and T+6, and no correlations.*

6.2.3 Impact of Correlations on the Assimilation of Sparse Observations

The SVD of $\hat{\mathbf{H}}$ and $\hat{\mathbf{H}}\boldsymbol{\rho}_B^{\frac{1}{2}}$ are now used to consider the information content of sparse observations. First, the SVD of $\hat{\mathbf{H}}$ for a horizontal line of only four observations at T+0 and T+6 is found. The RSV pairs which give a large contribution to the analysis increment, if the most rapidly growing or decaying Eady wave is observed, are shown in Fig. 6.4. Again, there are two pairs of RSVs, but their structure is vastly different to that for a full horizontal line of observations. These RSVs have small horizontal scales with maxima and minima in the buoyancy field at the positions of the observations. This makes it difficult to interpret the information that is contained in the RSVs.

It is more useful to examine the RSVs of $\hat{\mathbf{H}}\boldsymbol{\rho}_B^{\frac{1}{2}}$ so that the effect of correlations can also be considered. These RSVs are shown in Fig. 6.5. In contrast to the RSVs of $\hat{\mathbf{H}}$, these RSVs now have large horizontal scales and can usefully be compared to the RSVs for a full horizontal line of observations. RSVs $2\&3$ for the sparse observations (Fig. 6.5 (a)) are almost identical to RSVs $2\&3$ for the full line of observations (Fig. 6.1(a)). RSVs $5\&6$ for the sparse observations (Fig. 6.5 (b)) are also very similar to RSVs $10\&11$ for the full line of observations (Fig. 6.1(b)),



(a) RSV 2 and 3, $\lambda = 0.15$

(b) RSV 5 and 6, $\lambda = 0.03$

Figure 6.5: *RSVs of the normalized observability matrix, with correlations with a length scale of $l = 10\Delta x$ applied to every level. Four sparse observations are given at T+0 and T+6.*

although there are still some small scales near to the lower boundary.

The similarities between the RSVs for a full line of observations and those for sparse observations mean that the results concerning the reconstruction of the unobserved regions and the vertical structure of analysis increments are still relevant for the assimilation of sparse observations; the background state penalizes the information needed to reconstruct the unobserved regions and also penalizes the decaying part of the analysis increment.

There are, however, two fundamental differences between assimilating a full line of observations and assimilating sparse observations. The first difference is that the sparse observations need to be more accurate than the full line of observations to produce the same analysis. This is illustrated by the difference in the magnitude of the corresponding singular values. The singular values for the sparse observations are 7 times smaller than those for a full set of observations because the number of observations are also reduced by a factor of 10. With smaller singular values, the regularization parameter μ also needs to be smaller so that the same amount of information can be extracted; this is equivalent to stating that the observations need to be more accurate. The second difference is that only the large scales can be successfully reconstructed using sparse data. In particular, at least four independent observations per wavelength are needed to reconstruct a wave. With only two observations, the position of the wave cannot be inferred, and the maxima and minima of the wave are simply at the positions of the observations. This was found through 4D-Var experiments (not shown) but may be deduced by considering polynomial interpolation. That is, $m + 1$ independent pieces of information are needed to determine the parameters for a polynomial of degree m (e.g. Atkinson, 1989).

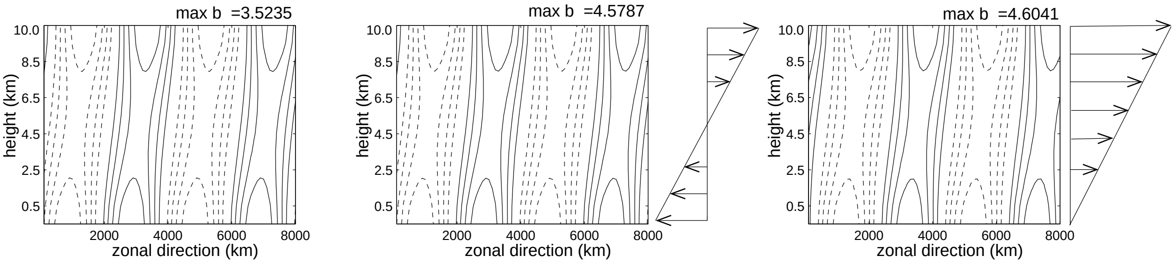
To summarize, the SVD of the normalized observability matrix $\hat{\mathbf{H}}\boldsymbol{\rho}_B^{\frac{1}{2}}$ has illustrated both the filtering and interpolating effect of background error correlations. The SVD of $\hat{\mathbf{H}}\boldsymbol{\rho}_B^{\frac{1}{2}}$ has also shown that the conclusions from previous experiments can be applied to the assimilation of sparse data, provided that only the large scales need to be reconstructed and that the sparse data is less noisy than the dense data.

6.3 Non-Modal and Modal Growth

The ability to generate the vertical structures necessary for modal and non-modal growth is now investigated. The experiments consider a background state with both zero QGPV and zero buoyancy and the true state given by either the most rapidly growing Eady wave, as shown in Fig. 6.6, or by an interior QGPV dipole as shown in Fig. 6.7, which is associated with a negative temperature anomaly which may have resulted from diabatic cooling. The equations for the non-modal initial conditions are given in Appendix A. The Eady wave results in exponential growth through the vertical coupling between the boundary waves, whereas the QGPV dipole structure results in rapid finite-time growth through the PV-unshielding mechanism, as discussed by Badger and Hoskins (2001). Figures 6.6(c) and 6.7(c) show the final state from integrations with a basic state such that the zonal wind is zero on the lower boundary. These true states will not be used until Section 6.5, but are given here for comparison.

In the following experiments, the observations are given by a full horizontal line of the interior buoyancy at a height of 4.5km at T+0 and T+6. This line passes through the centre of the buoyancy anomaly. The horizontal domain has been increased to 80 grid points to ensure that the spatial extent of the perturbation to the flow is smaller than the domain length.

It will be found that there is a significant dependence on the choice of the specified background error variances. For this reason, we choose to formulate the problem as solving an ill-posed inverse problem using Tikhonov's method of regularization, with multiple regularization parameters.



(a) Initial True State

(b) Final True State and basic state zonal wind

(c) Final True State and basic state zonal wind

Figure 6.6: *Buoyancy fields for the true state for modal growth (a)T+0, (b)T+6 such that the basic state zonal wind is zero in the centre of the domain, (c) T+6 such that the basic state zonal wind is zero on the lower boundary. The QGPV fields are all zero and are therefore not shown.*

The 4D-Var control variable $\mathbf{x}_0$ contains both the interior QGPV variables $\mathbf{x}_q$ and the boundary buoyancy variables $\mathbf{x}_T$ so that $\mathbf{x}_0$ can be written as:

$$\mathbf{x}_0 = \begin{bmatrix} \mathbf{x}_q \\ \\ \mathbf{x}_T \end{bmatrix}. \tag{6.15}$$

It is assumed that the background error covariance is diagonal, so that there are no auto-correlations or cross-correlations. However, we now assume that the QGPV and buoyancy have different background error variances. Denoting the background error variances of the QGPV and buoyancy as σ_q^2 and σ_T^2 respectively, and the observation error variance as σ_o^2 then:

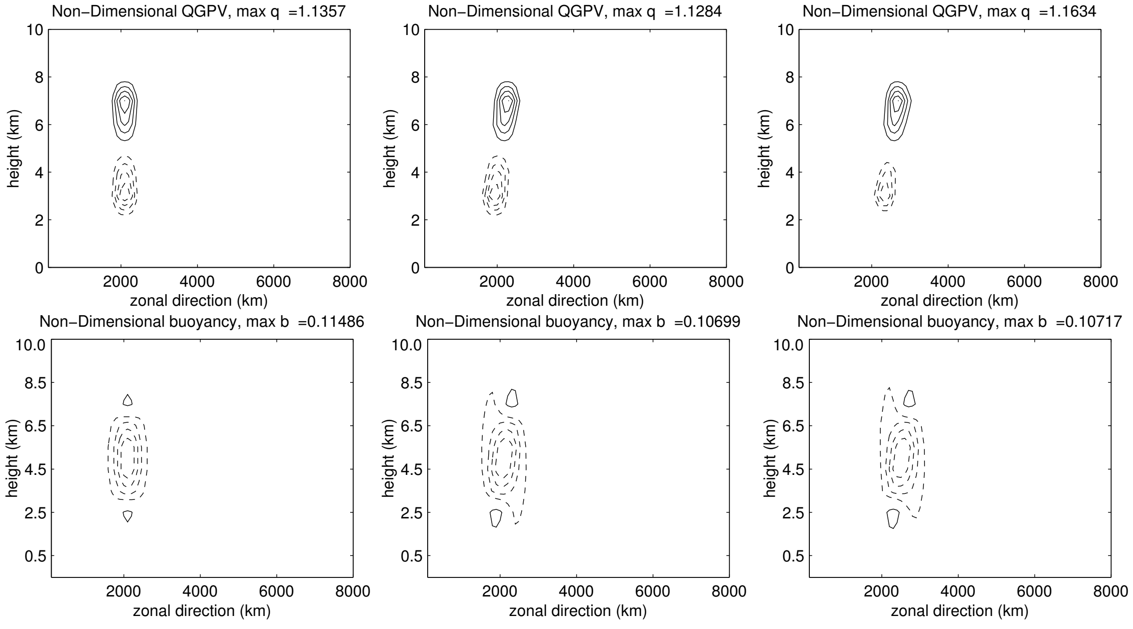
$$\mathbf{B} = \begin{pmatrix} \sigma_q^2\mathbf{I} & 0 \\ \\ 0 & \sigma_T^2\mathbf{I} \end{pmatrix} \qquad\qquad \mathbf{R} = \sigma_o^2\mathbf{I}. \tag{6.16}$$

With $\mathbf{x}^b = 0$, then the 4D-Var problem can be formulated as minimizing the cost function:

$$J(\mathbf{x}_0) = \mu_q^2\|\mathbf{x}_q\|_2^2 + \mu_T^2\|\mathbf{x}_T\|_2^2 + \|\hat{\mathbf{y}} - \hat{\mathbf{H}}\mathbf{x}_0\|_2^2 \tag{6.17}$$

where $\mu_q = \frac{\sigma_o}{\sigma_q}$, $\mu_T = \frac{\sigma_o}{\sigma_T}$ are regularization parameters. The importance of choosing the appropriate parameters μ_q and μ_T is now illustrated.

The case where the true state is given by the most unstable normal mode is first considered.

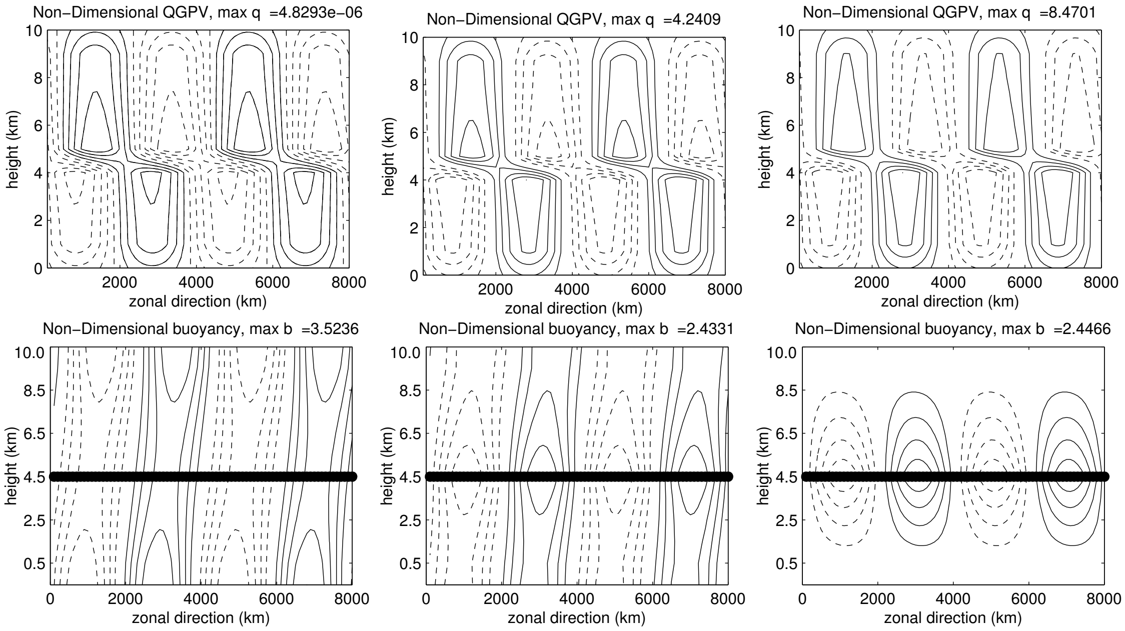


(a) Initial True State (b) Final True State (c) Final True State

Figure 6.7: *As for Fig. 6.6, but for non-modal growth.*

A comparison of three 4D-Var analyses with different values of μ_q and μ_T is shown in Fig. 6.8. If the parameters are specified as: $\mu_q^2 = 1$ and $\mu_T^2 = 10^{-5}$, then a large analysis increment is added to the buoyancy fields (amplitude is 3.5) and a small increment is added to the interior QGPV (amplitude is 10^{-6}). This gives an analysis that is close to the true state, as shown in Fig. 6.8 (a). In contrast, if the parameters are specified as $\mu_q^2 = 10^{-5}$ and $\mu_T^2 = 1$, then a large analysis increment is added to the interior QGPV (amplitude is 8.5) and a small increment is added to the buoyancy on the boundaries (amplitude is 2.4), as shown in Fig. 6.8 (c), giving a very poor analysis. The QGPV field of the analysis has a rich vertical structure, and so the structure will grow using a PV-unshielding mechanism rather than a boundary coupling mechanism.

The parameters chosen in these cases give two extremes. When the parameters represent the error variances well, the analysis is close to the truth. However, at the opposite extreme, when the parameters do not represent the error variances, the analysis is far from the the true state. The analysis shown in Fig. 6.8 (b) uses ‘climatological’ parameters which are in the middle of these two extremes. A medium sized increment has been added to both the buoyancy field and the QGPV field, giving an analysis that is also between the two extremes. Such ‘climatological’ values are representative of an operational 4D-Var algorithm, where the back-



(a) $\sigma_q^2 = 1$, $\sigma_T^2 = 10^5$

(b) $\sigma_q^2 = 2 \times 10^3$, $\sigma_T^2 = 10^2$

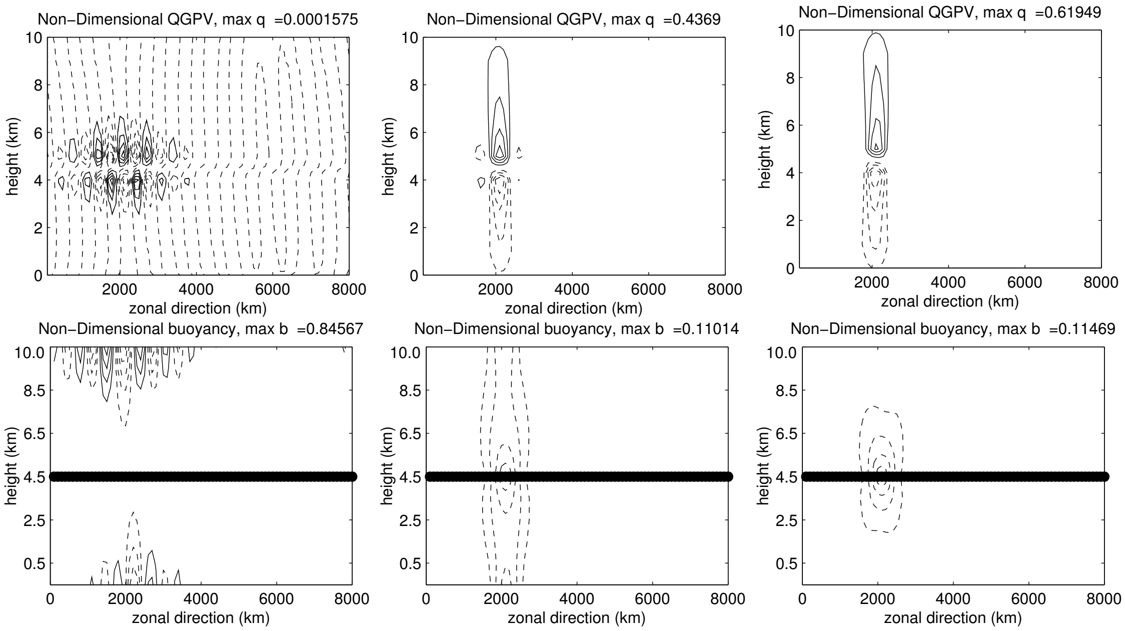
(c) $\sigma_q^2 = 10^5$, $\sigma_T^2 = 1$

Figure 6.8: *4D-Var analyses shown at T+0 with the true state given by modal growth. The upper plots show non-dimensional QGPV and the lower plots show non-dimensional buoyancy. Perfect observations of a horizontal line of buoyancy are given at T+0 and T+6. The assumed observation error variance is* $\sigma_o^2 = 1$. *The assumed background error variances are (a)*$\sigma_q^2 = 1$, $\sigma_T^2 = 10^5$, *(b)*$\sigma_q^2 = 2 \times 10^3$, $\sigma_T^2 = 10^2$ *and (c)*$\sigma_q^2 = 10^5$, $\sigma_T^2 = 1$.

ground error covariance is estimated using a climatology of error statistics.

The case where the true state is given by a QGPV dipole, characteristic of non-modal growth, is now considered. The 4D-Var analyses are repeated with the same parameters as for the normal mode case. If the parameters are specified as $\mu_q^2 = 1$ and $\mu_T^2 = 10^{-5}$, then 4D-Var tries to add a large analysis increment to the boundaries, so that the interior buoyancy in the observed regions will have the correct initial amplitude. This gives an analysis that has completely the wrong structure, as shown in Fig. 6.9 (a). The amplitude of the buoyancy field is 0.8, whereas the amplitude of the QGPV field is 1.4×10^{-4}. In contrast, if the parameters are specified as $\mu_q^2 = 10^{-5}$ and $\mu_T^2 = 1$ then a large analysis increment is correctly added to the interior QGPV and the analysis resembles the true state, as shown in Fig. 6.9 (c). An analysis using ‘climatological’ parameters in between the two extremes is shown in Fig. 6.9 (b). As expected, the analysis is better when the parameters are a good representation of the actual background error variances.

To illustrate further the detrimental effect of the ‘climatological parameters’, as used by operational 4D-Var at the present time, the growth rates of the forecasts from the analyses



(a) $\sigma_q^2 = 1$, $\sigma_T^2 = 10^5$ (b) $\sigma_q^2 = 2 \times 10^3$, $\sigma_T^2 = 10^2$ (c) $\sigma_q^2 = 10^5$, $\sigma_T^2 = 1$

Figure 6.9: *As for Fig. 6.8, except the true state is given by non-modal growth with non-zero interior QGPV.*

are examined. The growth rates are measured using the perturbation kinetic energy norm as defined in Badger and Hoskins (2001). The non-dimensional kinetic energy norm is defined as:

$$KE = \iint v'^2 dxdz \tag{6.18}$$

where v' is the perturbation meridional wind, and the KE growth rate at time t, $\sigma_{KE}(t)$ is defined as:

$$\sigma_{KE}(t) = \frac{1}{2\Delta t} \ln \left(\frac{KE_{t+\Delta t}}{KE_{t-\Delta t}} \right) \tag{6.19}$$

where Δt is the time step. This measure allows a clear view of non-modal growth and also allows a clear comparison with the work by Badger and Hoskins (2001).

The kinetic energy growth rate of the state during the 6 hour assimilation window and the following 18 hour forecast is now examined for both modal and non-modal growth.

When the true state is given by modal growth (Fig. 6.10 (a)), the true kinetic energy growth rate σ_{KE} (T) is constant with time, as expected. The forecast from the analysis with the ‘appropriate parameters’ for the modal growth case (M) ($\sigma_q^2 = 1$, $\sigma_T^2 = 10^5$, and $\sigma_o^2 = 1$, Fig. 6.8(a)) also has the same growth rate. However, the growth rate of the forecast from the analysis with ‘climatological parameters’ (C) ($\sigma_q^2 = 2 \times 10^3$, $\sigma_T^2 = 10^2$, and $\sigma_o^2 = 1$,

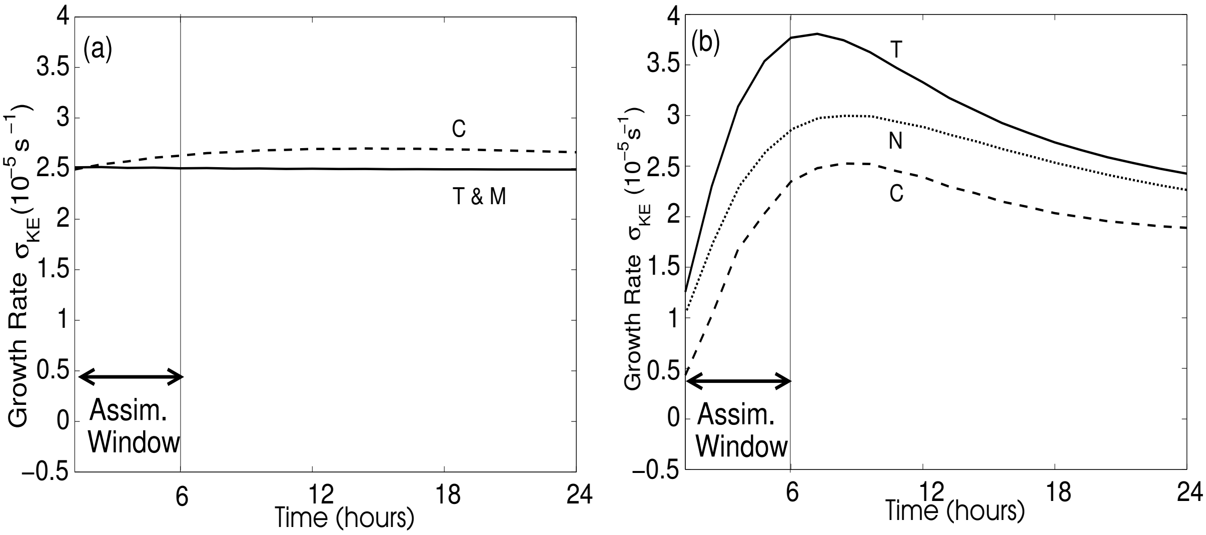


Figure 6.10: *Kinetic energy growth rates σ_{KE} $(10^{-5}s^{-1})$ for the true state (T) given by (a) the most rapidly growing normal mode and (b) a PV-dipole perturbation, during the 6 hour assimilation window and following 18 hour forecast. The growth rates from the truth (T, solid), and analyses with climatological parameters (C) (dashed) and with appropriate parameters for modal (M) and non-modal (N) (dotted) are shown. C: $\sigma_q^2 = 2 \times 10^3$, $\sigma_T^2 = 10^2$, M: $\sigma_q^2 = 1$, $\sigma_T^2 = 10^5$ N: $\sigma_q^2 = 10^5$, $\sigma_T^2 = 1$.*

Fig. 6.8(b)) is larger than that of the true state because the analysed QGPV field has a large amplitude and grows through the PV-unshielding mechanism.

When the true state is given by non-modal growth (Fig. 6.10 (b)), the true kinetic energy growth rate (T) varies with time. There is a peak at 6 hours and then it decreases to the value of the growth rate of the most unstable normal mode. The growth rate of the forecast from the analysis with the 'appropriate parameters' for the non-modal growth case (N) ($\sigma_q^2 = 10^5$, $\sigma_T^2 = 1$, and $\sigma_o^2 = 1$, Fig. 6.9(c))) also reaches a value that exceeds the rate for normal mode growth, although it is never as high as the that for the true state. The growth rate of the forecast from the analysis with 'climatological parameters' (C) ($\sigma_q^2 = 2 \times 10^3$, $\sigma_T^2 = 10^2$, and $\sigma_o^2 = 1$, Fig. 6.9(b))) is far less than that of the true state and in fact, it never reaches a value that is larger than that of the normal mode. This is because a large proportion of the analysis increment has been added to the boundaries and not to the interior QGPV. Thus, the analysis does not contain the vertical structure that is needed to give the rapid growth that is seen in the true solution.

To summarize, for normal-mode growth, the best analysis and forecast are achieved when the parameters are such that a large analysis increment is added to the boundaries; for non-modal growth the best analysis is achieved when the parameters are such that a large analysis increment is added to the interior. Although the background error covariances are diagonal, 4D-Var has been able to generate analysis increments for both modal and non-modal growth, providing that the regularization parameters are specified appropriately.

6.4 Two Horizontal Lines

In the final section (Section 6.5) in this chapter we will consider how 4D-Var uses the information from vertical temperature profiles and how this compares and contrasts with the use of horizontal lines of observations. The previous results (Chapter 3) have shown that a time-sequence of observations in the horizontal can provide information about the vertical structure. The question is whether vertical profiles can provide a significant increase in the amount of useful information about the vertical structure.

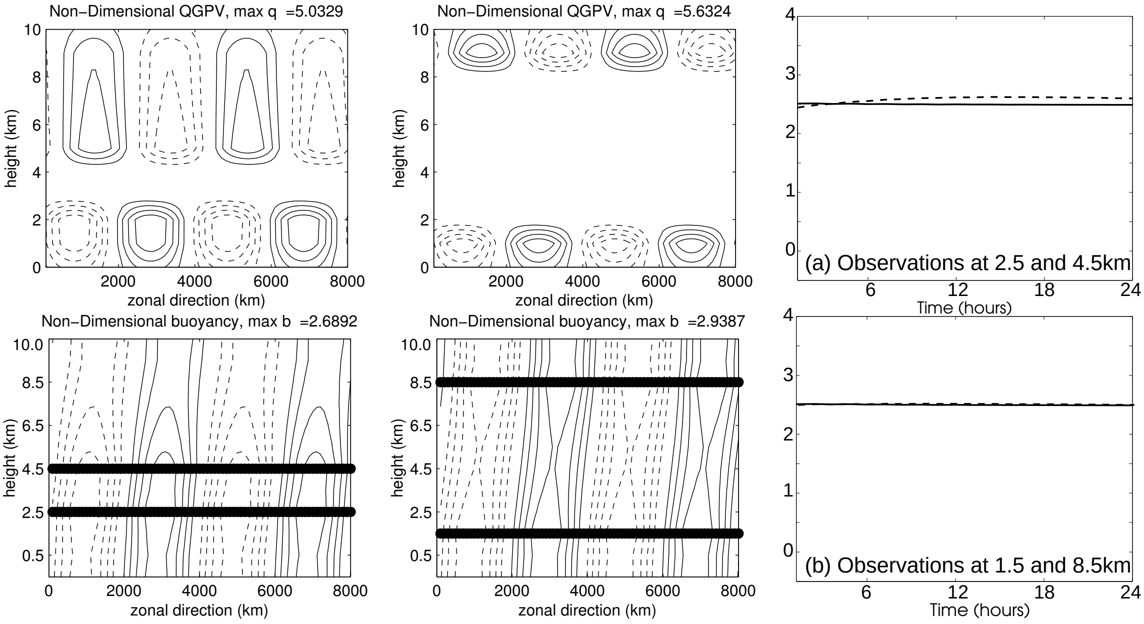
As a step towards understanding the information from vertical profiles, the assimilation of two horizontal lines at different vertical levels is considered. This should provide a strong link with the knowledge that has already been gained from the previous experiments. The assimilation of two horizontal lines of observations is considered in this section.

6.4.1 4D-Var Experiments

In the following experiments, perfect observations of horizontal lines of the interior buoyancy at two heights are given at the beginning and the end of a 6 hour window. The background state is zero and the true state is given by either the most rapidly growing Eady wave or by a QGPV dipole. For the QGPV dipole, one set of observations passes close to the centre of the buoyancy anomaly, whilst the other set of observations samples the zero values below the anomaly.

Analyses using the ‘climatological values’ of $\sigma_q^2 = 2 \times 10^3$, $\sigma_T^2 = 10^2$ and $\sigma_o^2 = 1$ are first examined. Note that these values have not been derived from a climatology of error statistics but were chosen for the experiments in the previous section (6.3) to represent the values used in operational 4D-Var.

The analyses for the true state given by modal growth are shown in Fig. 6.11. Again, as for the assimilation of a single horizontal line, the buoyancy field correctly tilts eastwards with height, but the interior QGPV is not zero, but instead has a large amplitude. However, the region between the two horizontal lines does have QGPV values that are relatively small. When the horizontal lines are moved further apart (b), the region with near zero QGPV also extends. To be able to determine whether these analyses are better than that for the assimilation



(a) Observations at 2.5 and 4.5km

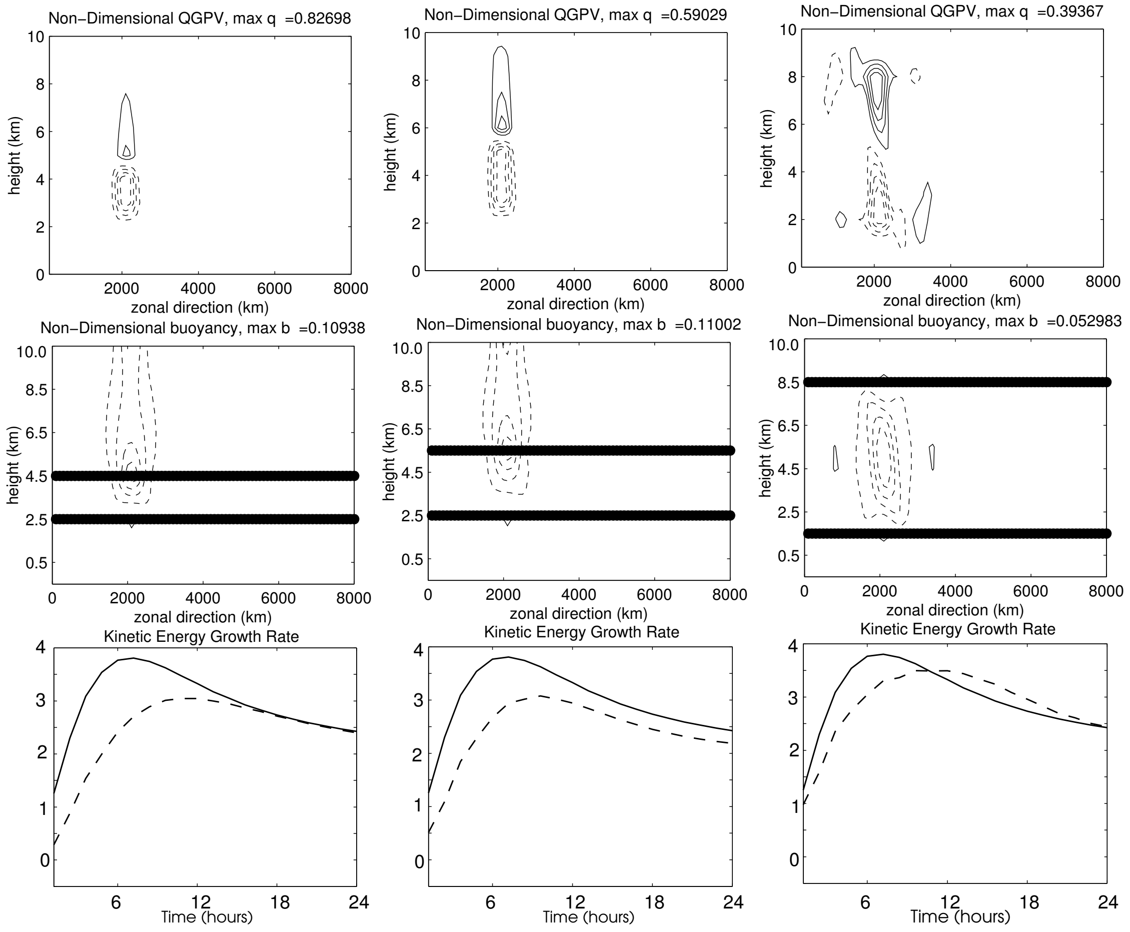
(b) Observations at 1.5 and 8.5 km

(c) Kinetic Energy Growth Rates, σ_{KE} $(10^{-5}s^{-1})$

Figure 6.11: *4D-Var analyses shown at T+0 with the true state given by modal growth. Perfect observations of two horizontal lines of buoyancy are given at T+0 and T+6 at either (a)2.5 and 4.5 km or (b) 1.5 and 8.5km. In both cases the weights are are given by:* $\sigma_q^2 = 2 \times 10^3$, $\sigma_T^2 = 10^2$, $\sigma_o^2 = 1$. *The associated kinetic energy growth rates* σ_{KE} $(10^{-5}s^{-1})$ *from the true state (solid) and from the analyses are shown in (c).*

of a single line of buoyancy, it is useful to examine the growth rate of the following forecast. The kinetic energy growth rates for the 6 hour window, and following 18 hour forecast are shown in Fig. 6.11 (c). When the two horizontal lines are close together (a), there is some improvement in the analysed growth rate in comparison to the assimilation of a single line of observations. When the two horizontal lines are further apart, there is even more improvement, and there is little distinction between the true growth rate and the analysed growth rate.

The analyses for the true state given by non-modal growth are shown in Fig. 6.12. When an extra horizontal line is given at 2.5km (a), there is some information indicating that the buoyancy is zero at that height and therefore the spatial extent of the analysed buoyancy perturbation has been cut off below 2.5km. When the upper line is moved higher in the domain (b) , the structure of the analysis looks very similar, however, the minimum amplitude of the perturbation is now higher and coincides with the position of the observations. Thus, the 4D-Var algorithm has not been able to infer that the minimum amplitude in the perturbation lies in between the two lines of observations, and instead the minimum in the analysis is found at the



(a) Observations at 2.5 and 4.5 km, weights: $\sigma_q^2 = 2 \times 10^3$, $\sigma_T^2 = 10^2$, $\sigma_o^2 = 1$.

(b) Observations at 2.5 and 5.5 km, weights: $\sigma_q^2 = 2 \times 10^3$, $\sigma_T^2 = 10^2$, $\sigma_o^2 = 1$.

(c) Observations at 1.5 and 8.5km, weights: $\sigma_q^2 = 10^5$, $\sigma_T^2 = 1$, $\sigma_o^2 = 1$.

Figure 6.12: *4D-Var analyses shown at T+0 with the true state given by non-modal growth. Perfect observations of two horizontal lines of buoyancy are given at T+0 and T+6 at (a) 2.5 and 4.5km, (b) 2.5 and 6.5km and (c) 1.5 and 8.5km. The weights are given by: (a) and (b)* $\sigma_q^2 = 2 \times 10^3$, $\sigma_T^2 = 10^2$, $\sigma_o^2 = 1$ *and (c)* $\sigma_q^2 = 10^5$, $\sigma_T^2 = 1$, $\sigma_o^2 = 1$. *The top panels show the QGPV fields, the middle panels show the buoyancy fields and the bottom panels show the associated Kinetic Energy growth rates* σ_{KE} $(10^{-5}s^{-1})$, *from the truth (solid) and from the analyses (dashed).*

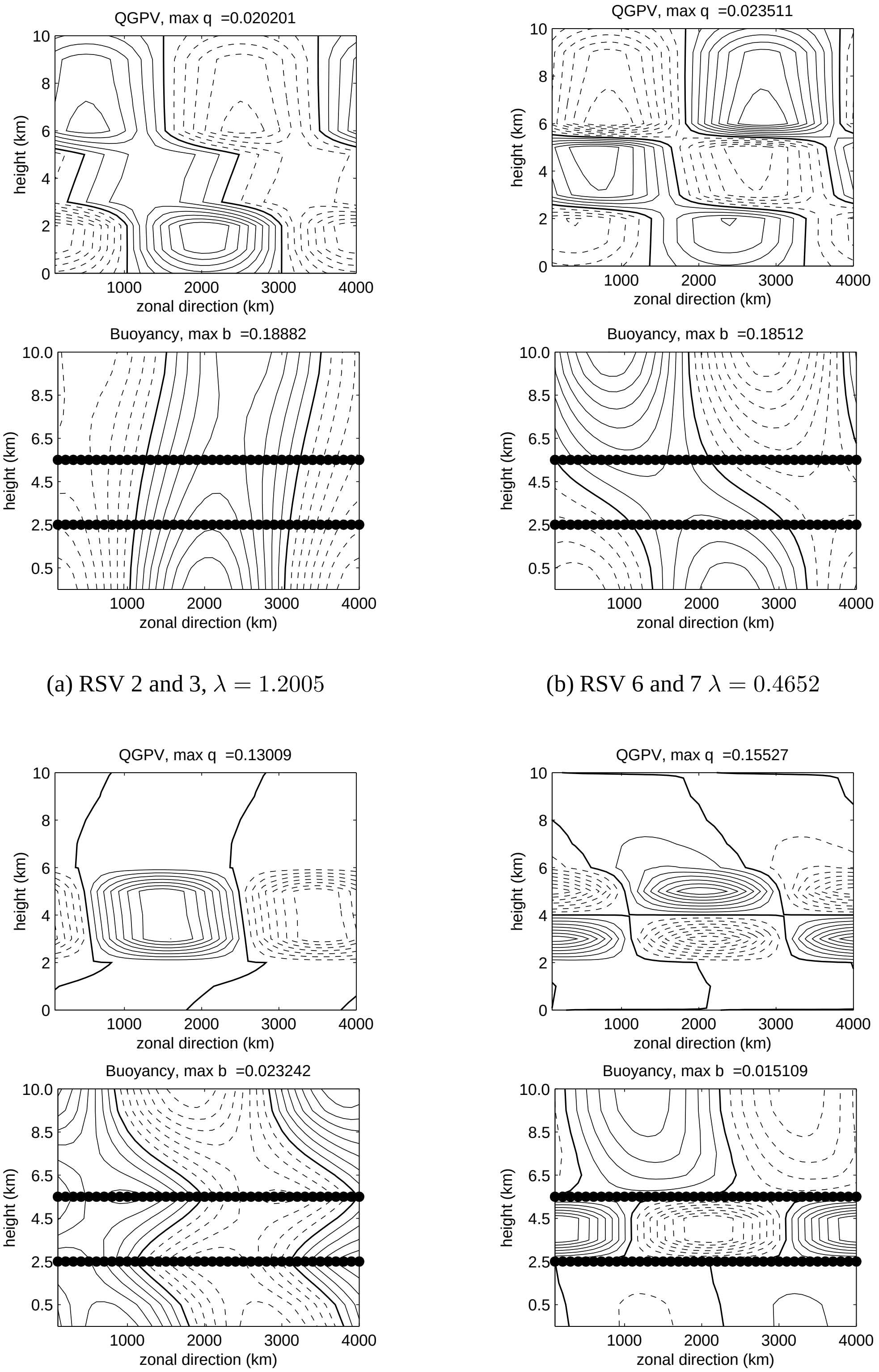
positions of the observations.

Again, it is useful to examine the growth rates for the following forecasts. These are shown in the bottom panels of Fig. 6.12. With observations at 2.5 and 4.5km, the growth rate now reaches a value of $3\times10^5\ s^{-1}$, as opposed to the value of only $2.5\times10^5\ s^{-1}$ which was attained with the assimilation of only a single line of observations. With observations at 2.5 and 5.5km, the growth rate again peaks at $3\times10^5\ s^{-1}$, but then the growth rate reduces in the last 12 hours.

In both analyses, the maximum in amplitude was found at the position of the observations. This leads on to an investigation of whether 4D-Var is indeed able to infer the perturbation if it lies in an unobserved region. The following experiment uses lines of observations at heights of 2.5 and 7.5 km. These are below and above the spatial extent of the initial perturbation, but should be able to detect the growth of the perturbation. The results from the previous chapters showed that to be able to infer the state in unobserved regions, it is important to give a relatively large weight to the observations in comparison to the background state. Therefore, the parameters are now adjusted to $\sigma_q^2 = 10^5$, $\sigma_T^2 = 1$ and $\sigma_o^2 = 10^{-5}$. These weights are chosen to maximize the use of the information from the observations. The analysis is shown in Fig. 6.12(c). The perturbation has been correctly inferred by 4D-Var, with a maxima in the buoyancy field at the correct position. The growth rate attains a maximum of $3.4\times10^5\ s^{-1}$, but the peak growth rate occurs at 12 hours rather than at 6 hours. This experiment has shown that 4D-Var is able to reconstruct the interior QGPV perturbations that are necessary for rapid finite-time growth, although a relatively large weight needs to be given to the observations.

6.4.2 SVD Experiments

To make this understanding more complete, it is necessary to examine the RSVs that are used to reconstruct the most unstable Eady wave. From the values of $\mathbf{u}_j^T\hat{\mathbf{d}}$ there are four RSV pairs that contribute to the analysis increment. The QGPV and buoyancy fields for these vectors are shown in Fig. 6.13. RSVs $2\&3$ (Fig. 6.13(a)) have an eastward tilting buoyancy field with a maximum in amplitude on the lower boundary. These are very similar to the first pair of RSVs for a single line of observations. RSVs $6\&7$ (Fig. 6.13(b)) have a westward tilting buoyancy field with a maximum in amplitude on the upper boundary. These are very similar to the second pair of RSVs for a single line of observations. Thus the information content of two horizontal lines of observations is very similar to the the information content of one horizontal line. However, with two lines of observations, there are also another two pairs of



(a) RSV 2 and 3, $\lambda = 1.2005$

(b) RSV 6 and 7 $\lambda = 0.4652$

(c) RSV 19 and 20, $\lambda = 0.1257$

(d) RSV 127 and 128 $\lambda = 0.0084$

Figure 6.13: *RSVs of the 4D-Var observability matrix with no correlations, with two horizontal lines (80 observations) at the beginning and the end of a 6 hour window, when either the most rapidly growing or decaying Eady wave is observed.*

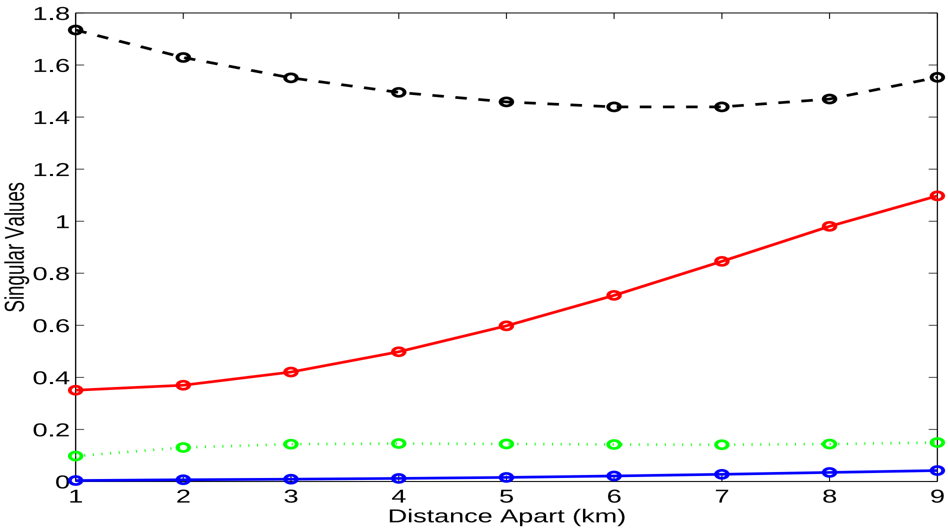


Figure 6.14: *Singular values of the four pairs of RSVs that contribute to the analysis increment when the most rapidly growing or decaying Eady wave is observed. Observations of two horizontal line of the interior buoyancy are given at T+0 and T+6. The lower line of observations is given at 0.5km and the vertical distance between the horizontal lines is given by the abscissa. The singular values of the first, second, third and fourth pairs of RSVs are represented by the black, red, green and blue lines respectively.*

RSVs that have very different structures. RSVs $19\&20$ (Fig. 6.13(c)) contain a PV monopole that is situated between the two horizontal lines, and RSVs $127\&128$ (Fig. 6.13(d)) contain a PV dipole that is situated between the two horizontal lines. These RSVs have extremely small singular values and hence an extremely small weight also needs to be given to the background state so that these RSVs may be included in the analysis increment. These RSVs are needed to contribute to the analysis increment when the most unstable Eady wave is observed. Similar RSVs are used for the analysis increment when the non-modal perturbation is observed.

It is surprising that although there is more information about the vertical structure of the wave, there is still a large distinction between the growing and decaying parts of the analysis increment. That is, there is a large difference in the singular values of RSVs $2\&3$ and RSVs $6\&7$. This is possibly because the horizontal lines are close together. Therefore, we now investigate how the singular values differ if the distance between the horizontal lines is increased. Singular values of the four pairs of RSVs that contribute to the analysis increment when the most rapidly growing or decaying Eady wave is observed are shown in Fig. 6.14. The lower line is given at 0.5km and the height of the upper line is varied. As the distance between the horizontal lines increases, the singular value of the first pair of RSVs decreases whilst the singular value of the second pair of RSVs increases. Thus, when the horizontal lines are close together, there is a large distinction between the information for the growing and decaying modes; when the horizontal lines are far apart, there is a smaller distinction between the information for the growing and decaying modes. However, even when the horizontal lines are as far apart as possible, there is still a significant difference between the singular values of the first and second pairs of RSVs. Thus, the decaying part of the analysis increment will always be penalized more than the growing part of the analysis increment, regardless of the height of the two horizontal lines.

To summarize, the 4D-Var and SVD experiments both showed that the analyses of normal modes are improved when the horizontal lines of observations are further apart. The SVD experiments also showed that the decaying part of the analysis increment is still strongly penalized by the background state even though there is more information about the vertical structure. The 4D-Var experiments showed that if a relatively large weight is given to the background state the maxima in the analysis is found at the position of the observations, although it is possible to use the time-evolution information to infer the state in an unobserved regions provided that sufficiently large weight is given to the observations (the observations must be accurate enough).

6.5 Vertical Lines

The following 3D-Var and 4D-Var experiments consider the assimilation of vertical lines of observations of the interior buoyancy. Each vertical line contains an observation at each of the twelve vertical levels. At a single instant in time, a vertical line of observations cannot provide any information about the vertical tilt of the atmosphere. To gain information about the vertical tilt, at least two vertical profiles are required. In 3D-Var, the vertical profiles need to sample the atmosphere at different points in space, but in 4D-Var the vertical profiles may sample the atmosphere at different points in time. Therefore, we now consider the assimilation of two vertical lines in both 3D-Var and 4D-Var. The 3D-Var experiments assimilate two vertical profiles given at the same time, and the 4D-Var experiments assimilate a single profile given at both the beginning and the end of the assimilation window.

The first experiments consider the ability to generate analysis increments necessary for modal growth and decay, whilst the final experiments consider the ability to generate analysis increments necessary for non-modal growth.

6.5.1 3D-Var and 4D-Var Experiments

3D-Var and 4D-Var analyses using two vertical lines are now examined. The true state is given by the most rapidly growing Eady wave, shown in Fig. 6.6(a), and the background state is zero. The purpose of these experiments is to determine whether 4D-Var is able to use the model dynamics to link together the observations distributed in time and successfully infer the vertical tilt of the state.

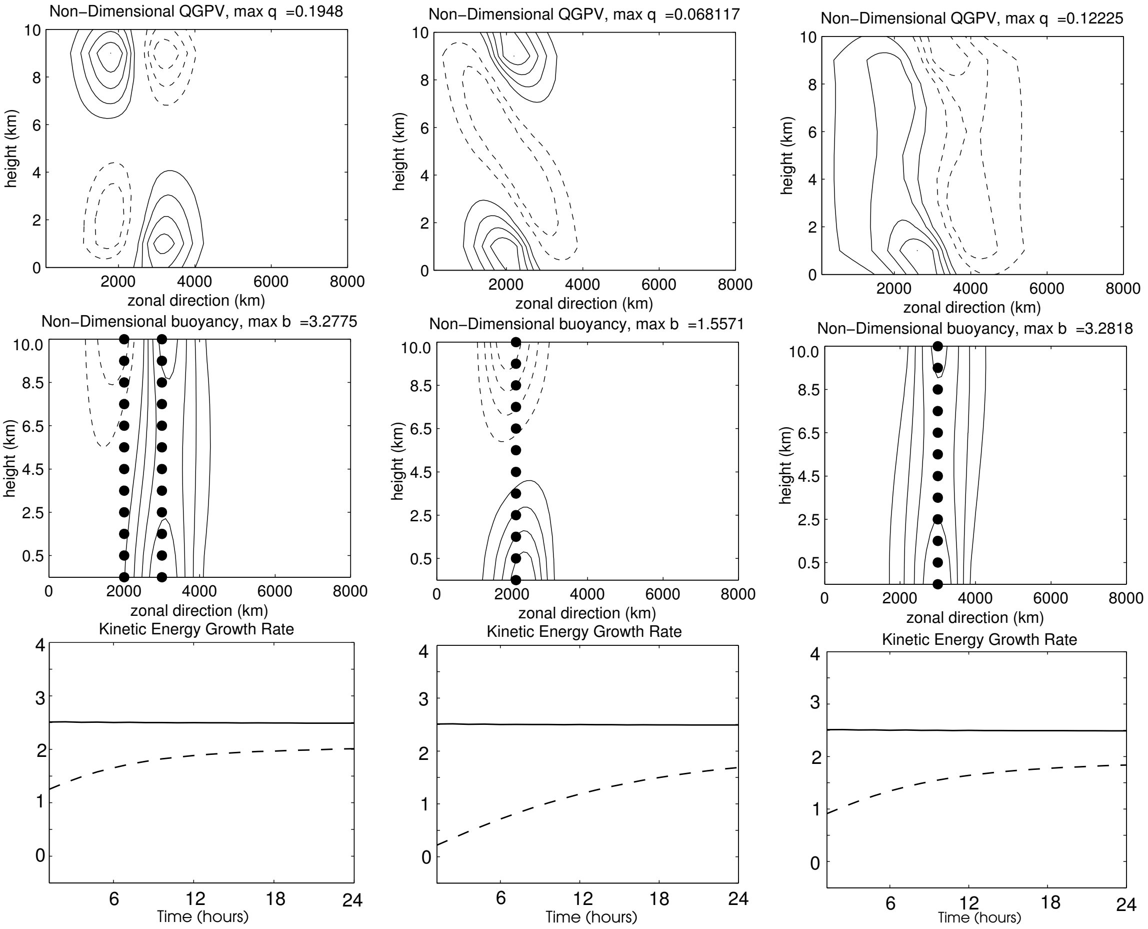
The assumed background error variances for the QGPV and temperature are now the same. The weights are chosen so that a relatively large weight is given to the observations: $\sigma_b^{-2} = 10^{-2}$ and $\sigma_o^{-2} = 1$. Horizontal correlations with a length scale of $l = 5\Delta x$ are applied to both the buoyancy and the QGPV at every vertical level, so that the information from the observations is distributed to the surrounding grid points.

A 3D-Var analysis using two vertical lines of buoyancy observations at 2000km and 3000km is shown in Fig. 6.15(a). The analysis increment has only been added to a small region near to the observations. The size of this region is determined by the correlation length scale. The buoyancy field tilts eastwards with height in the regions between the observations, as required. However, the analysis does not exhibit any tilt in the regions to the east and west of the observations because there are no further vertical profiles to link together.

The experiment is repeated but for 4D-Var with only a single vertical line of observations at both the beginning and the end of a 6 hour assimilation window. This experiment is designed to assess whether the model dynamics can provide the necessary information so that the correct vertical structure can be obtained using two vertical lines that are distributed in time instead of space.

The 4D-Var analysis for vertical lines at 2000km is shown in Fig. 6.15(b). At 2000km, the true vertical structure has a sharp gradient and a small amplitude. The analysis is similar, with a cold anomaly in the upper half of the domain and a warm anomaly in the lower half. The upper anomaly is slightly to the west of the position of the observations and the lower anomaly is slightly to the east. These anomalies are advected by the basic state wind so that the maxima are at 2000km at T+6.

The 4D-Var analysis for vertical lines at 3000km is shown in Fig. 6.15(c). At 3000km, the true vertical structure has a large amplitude and is symmetrical about the middle of the domain. It can clearly be seen that the buoyancy field of the analysis tilts eastwards with height as required.



(a) 3D-Var

(b) 4D-Var with Observations at 2000km

(c) 4D-Var with Observations at 3000km

Figure 6.15: *Analyses where the true state is given by the most rapidly growing Eady wave, shown at T+0. The assumed variances are given by: $\sigma_o^{-2} = 1$, $\sigma_b^{-2} = 10^{-2}$, and horizontal correlations are applied with a horizontal length scale $l = 5\Delta x$. (a) 3D-Var analysis with two vertical lines of buoyancy observations, (b) 4D-Var analysis using a vertical line of observations at 2000km at T+0 and T+6, (c) 4D-Var analysis using a vertical line of observations at 3000km at T+0 and T+6. The top panels show the QGPV fields, the middle panels show the buoyancy fields and the bottom panels show the associated Kinetic Energy growth rates σ_{KE} $(10^{-5}s^{-1})$, from the truth (solid) and from the analyses (dashed).*

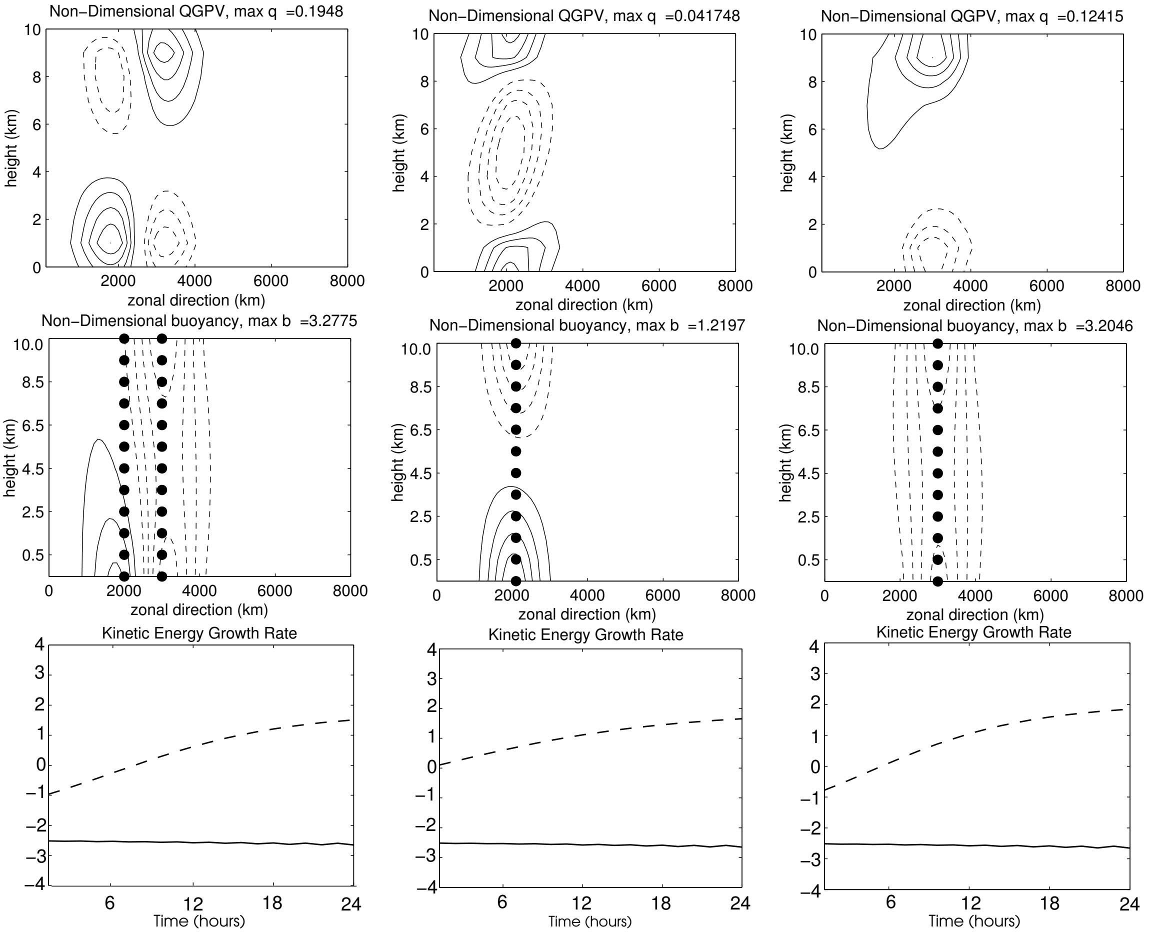
To assess the performance of 4D-Var, it is necessary to examine the growth rates for the following forecast. The associated kinetic energy growth rates are also shown in Fig. 6.15. None of the analyses give the correct growth rate because the analyses have localized spatial structures. It can clearly be seen that the 4D-Var analysis with observations at 3000km gives a growth rate that is close to that for the 3D-Var analysis. Thus, 4D-Var is able to determine the correct vertical structure from the time-sequence of observations. However, with observations at 2000km, the resulting growth rate is very small

The three experiments are repeated but with the true state given by the most rapidly decaying Eady wave; the analyses are shown in Fig. 6.16. The 3D-Var analysis (a) has a westward tilting buoyancy field in the region between the observations, as required. This results in a negative growth rate during the 6 hour assimilation window but a positive growth rate for the next 18 hours. The 4D-Var analysis with observations at 2000km has a cold anomaly in the upper half of the domain and a warm anomaly in the lower half. This results in a positive growth rate during both the assimilation window and the following forecast. The 4D-Var analysis with observations at 3000km does give a westward tilting buoyancy field as required and this results in a negative growth rate during the assimilation window, but is again positive during the forecast.

The experiments have shown that 4D-Var is able to use a time-sequence of vertical profiles to infer the correct vertical structure. This is similar to 4D-Var with horizontal lines of observations, where the time-sequence was used to infer the unobserved boundaries. The difference between the two is that with a horizontal line of observations, it is the time-evolution information that is used to infer the unobserved regions. However, with a vertical line of observations, it is not necessarily the same part of the state that is observed due to the horizontal advection. Thus, the 4D-Var with observations distributed in time is very similar to 3D-Var with observations distributed in space. If we imagine advecting the position of the initial observations, by the basic state flow, to the final time, then the vertical profiles may be thought of as being distributed in space and hence the vertical tilt may be inferred from this.

The ability of 4D-Var to use vertical profiles to generate the correct vertical structure for non-modal growth is now considered. A vertical profile samples the vertical structure of such perturbations well, and therefore it is expected that the assimilation of vertical profiles should give better analyses than for horizontal lines.

The analyses and growth rates shown in Fig. 6.17 illustrate that the appropriate choice of the regularization parameters is still vital to give a good analysis. In (a), a large weight has been given to the background state QGPV so that a relatively large analysis increment has been

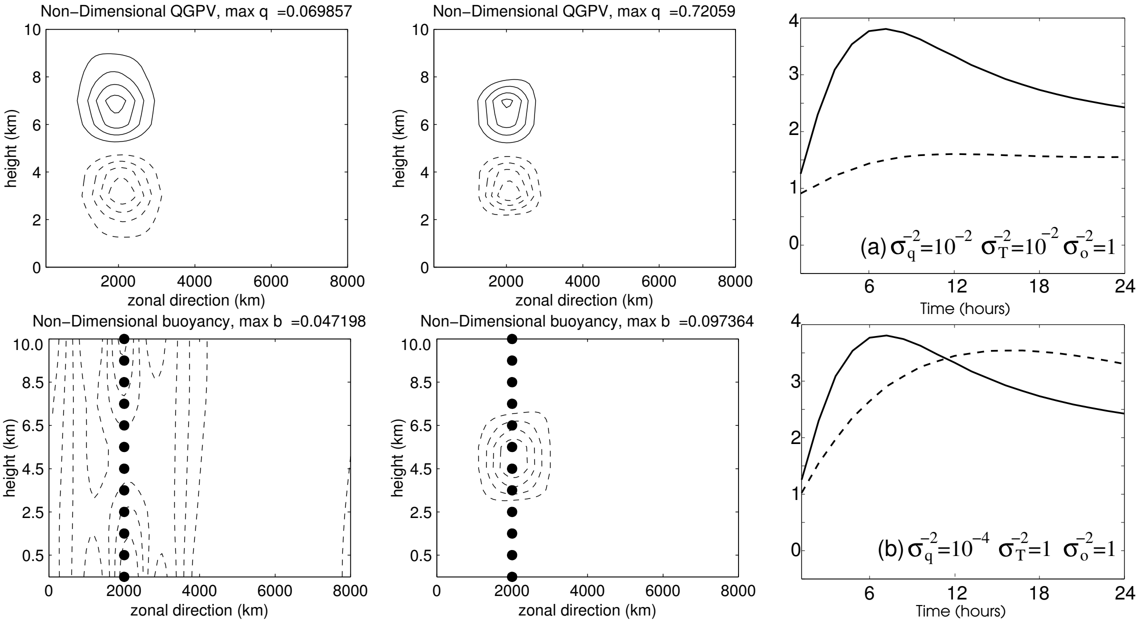


(a) 3D-Var

(b) 4D-Var with Observations at 2000km

(c) 4D-Var with Observations at 3000km

Figure 6.16: *As for Fig. 6.15 but with the true state given by the most rapidly decaying Eady wave.*



(a) $\sigma_q^{-2} = 10^{-2}, \sigma_T^{-2} = 10^{-2}, \sigma_o^{-2} = 1$

(b) $\sigma_q^{-2} = 10^{-4}$, $\sigma_T^{-2} = 1$, $\sigma_o^{-2} = 1$

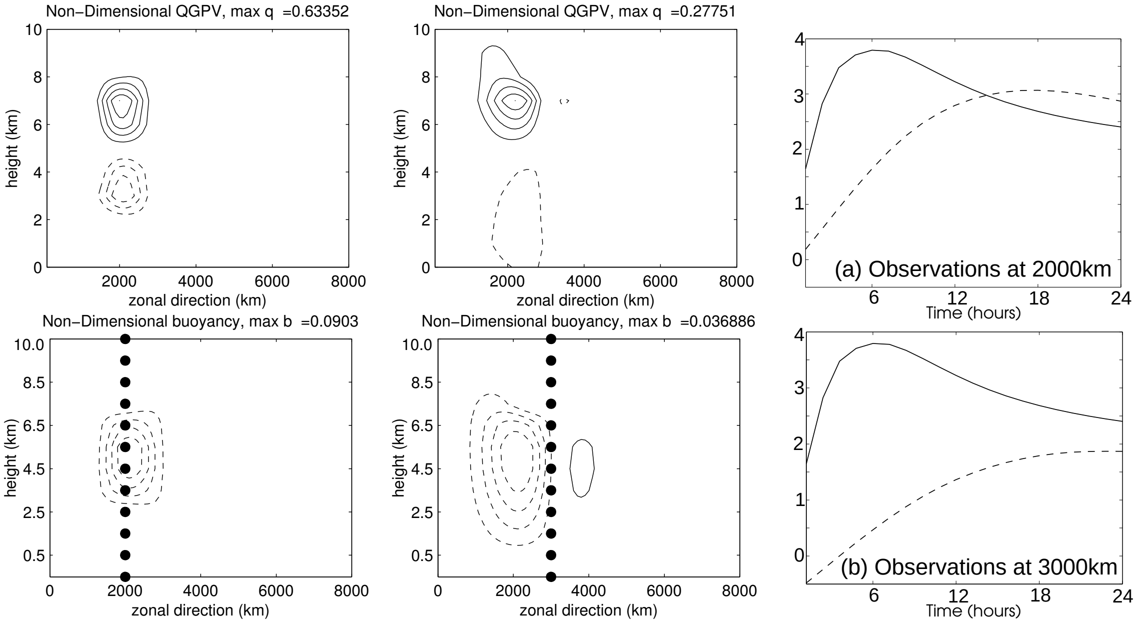
(c) Kinetic Energy Growth Rates, $\sigma_{KE}(10^{-5}s^{-1})$

Figure 6.17: *Analyses where the true state is given by non-modal growth, shown at T+0. A vertical line of buoyancy observations is given at T+0 and T+6, and horizontal correlations are applied with a length scale of $l = 5\Delta x$. The assumed variances are given by (a) $\sigma_q^{-2} = 10^{-2}$, $\sigma_T^{-2} = 10^{-2}$, $\sigma_o^{-2} = 1$ and (b) $\sigma_q^{-2} = 10^{-4}$, $\sigma_T^{-2} = 1$, $\sigma_o^{-2} = 1$. In both cases the basic state zonal wind is zero in the middle of the domain. The associated kinetic energy growth rates σ_{KE} $(10^{-5}s^{-1})$ from the true state (solid) and from the analyses are shown in (c).*

added to the boundaries. Although the QGPV field contains a dipole structure, the buoyancy field does not contain the monopole and the amplitudes are small. Thus, the associated growth rate is very small and so the finite-time growth is missed in the forecast. The spatial extent of the perturbation is too large in the horizontal due to the choice of the horizontal correlation length scale.

If the weight given to the QGPV background state is reduced then the analysis (b) has the correct structure and the growth rate achieves a maximum of nearly $3.5 \times 10^{-5}s^{-1}$. The horizontal length scale of the perturbation is now smaller as there is less weight given to the background state QGPV.

The experiments for vertical profiles of non-modal growth are now repeated but with the basic state such that the zonal wind is zero on the lower boundary. When the basic state flow was zero in the middle of the domain, the observations at 2000km sampled the centre of the buoyancy anomaly at both T+0 and T+6. When the basic state flow is zero on the lower



(a) Observations at 2000km

(b) Observations at 3000km

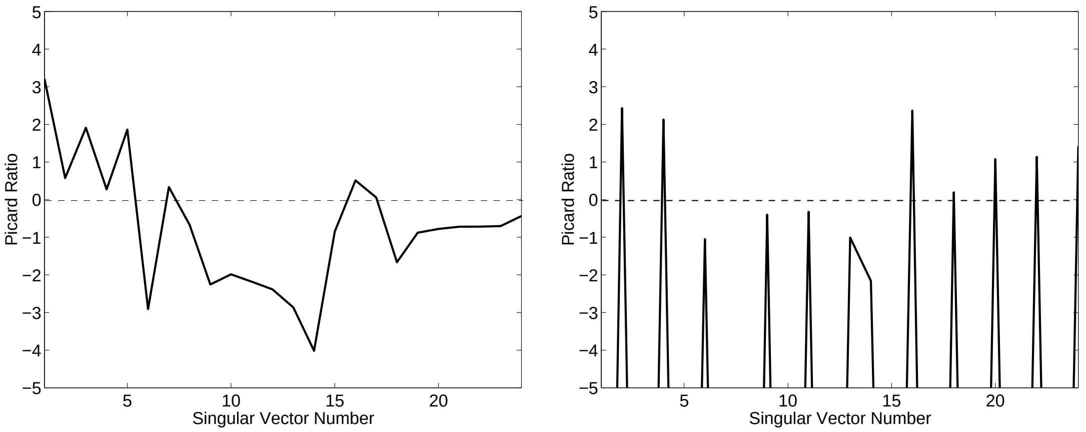
(c) Kinetic Energy Growth Rates, $\sigma_{KE}(10^{-5}s^{-1})$

Figure 6.18: *Analyses where the true state is given by non-modal growth, shown at T+0. Horizontal correlations are applied with a length scale of $l = 5\Delta x$ and the assumed variances are given by $\sigma_q^{-2} = 10^{-4}$, $\sigma_T^{-2} = 1$, $\sigma_o^{-2} = 1$. A vertical line of buoyancy observations is given at (a) 2000km at T+0 and T+6, and (b) 3000km at T+0 and T+6. In both cases the basic state zonal wind is zero on the lower boundary. The associated kinetic energy growth rates σ_{KE} $(10^{-5}s^{-1})$ from the true state (solid) and from the analyses are shown in (c).*

boundary, the buoyancy anomaly is advected eastwards. Consequently, with observations at 2000km, the buoyancy anomaly is sampled at T+0 but not at T+6, and with observations at 3000km, the buoyancy anomaly is sampled at T+6 but not at T+0. The effect of the different basic state on the analysis increments is now investigated.

The analysis for observations at 2000km is shown in Fig. 6.18 (a). The buoyancy anomaly and QGPV dipole have been reconstructed. The spatial extent of the buoyancy anomaly is larger to the east of the observations than to the west. This is because the observations at T+6 sampled the state to the west of the anomaly, where the true values are zero. The growth rate in (c) shows that the KE growth rate attains a value of nearly $3 \times 10^{-3}s^{-1}$. Thus, the perturbation does give rapid growth, although this is again 6 hours later than the rapid growth of the true state. The maximum growth rate is slightly less than that for the basic state with zero flow in the middle of the domain.

The analysis for observations at 3000km is shown in Fig. 6.18 (b). The buoyancy anomaly



(a) Observations at 3000 km

(b) Observations at 2000km

Figure 6.19: *Picard Ratio Values* $\left(\log\left|\frac{\mathbf{u}_j^T\hat{\mathbf{d}}}{\lambda_j}\right|\right)$ *with the true state given by the growing Eady wave, and a background state of zero, with a vertical line of observations at (a) 3000km and (b) 2000km at T+0 and T+6.*

has correctly been inferred to the west of the observation position. The reconstruction of the upper QGPV anomaly is better, with a larger amplitude, than that of the lower anomaly because the upper anomaly is advected further by the basic state. The surrounding regions have very small negative values so that the mean of the QGPV field is zero. This is due to the constraint that the mean of $\hat{\psi}$ is zero (2.28). The values of the KE growth rate show that the perturbation does not give rapid finite-time growth, perhaps because the lower QGPV anomaly is small in comparison to the upper QGPV anomaly and because the maximum amplitudes are very small.

In both the 2000km and 3000km cases, the growth rates are relatively small during the 6 hour assimilation window. This is because the maximum in the buoyancy anomaly was only sampled at one time, rather than at both times.

6.5.2 SVD Experiments

The SVD of the normalized observability matrix for vertical lines is now examined. The results should clearly show how the information from a vertical line of observations is combined with the model dynamics.

The Picard ratio (5.4):

$$\text{Picard Ratio} = \log\left|\frac{\mathbf{u}_j^T\hat{\mathbf{d}}}{\lambda_j}\right| \tag{6.20}$$

was introduced in Chapter 5, where $\mathbf{u}_j$ and λ_j are the LSVs and singular values of the normalized observability matrix, and $\hat{\mathbf{d}}$ is the generalized innovation vector. These values may be used to determine which RSVs contribute to the analysis increment. The Picard Ratio values, for the most rapidly growing Eady wave with a vertical line of observations at either 3000km or 2000km are shown in Fig. 6.19.

When a single horizontal line of observations at T+0 and T+6 were assimilated, two pairs of RSVs contributed to the analysis increment. For a double line of observations at T+0 and T+6, there were four pairs of RSVs. For a vertical line of observations, it can be seen that there are no longer just a few RSVs that contribute to the true analysis increment; with observations at 3000km, all 24 RSVs give a significant contribution to the analysis increment, and with observations at 2000km, 12 RSVs give a significant contribution to the analysis increment. Thus, there are two RSVs for each vertical level. The RSVs do not form pairs as there is only one observation at each vertical level.

The background state strongly penalizes the RSVs with small singular values so that the RSVs with large singular values will dominate the analysis increment when a relatively large weight is given to the background state. Therefore, we examine in detail the structures of only the first four RSVs. The QGPV, buoyancy and streamfunction fields of the first four RSVs are shown in Fig. 6.20. The amplitudes of the QGPV fields are small in comparison to the buoyancy fields, so that it is the buoyancy fields that dominate the structures. The values of the buoyancy field at the position of the observations for RSV 1 are all negative. This corresponds to a streamfunction field with negative values in the upper half and positive values in the lower half. The streamfunction field also has a westward tilt with height.

In contrast, the values of buoyancy at the position of the observations for RSV 2 are positive in the upper half and negative in the lower half. This is associated with an equivalent barotropic streamfunction field.

The buoyancy fields for RSVs 3 and 4 have much smaller scale structures, making it difficult to interpret the information that is contained in these vectors.

To understand the information that is contained in the RSVs, it is useful to examine the difference between the values of $\mathbf{u}_j^T\hat{\mathbf{d}}$ for the growing and decaying modes. These values emphasize the first few RSVs that have large singular values, whereas the Picard ratio values (in Fig. 6.19) show all the RSVs that are given a large weight. As only the first four RSVs are now being considered, the values of $\mathbf{u}_j^T\hat{\mathbf{d}}$ therefore allow a clearer comparison than the Picard ratio values.

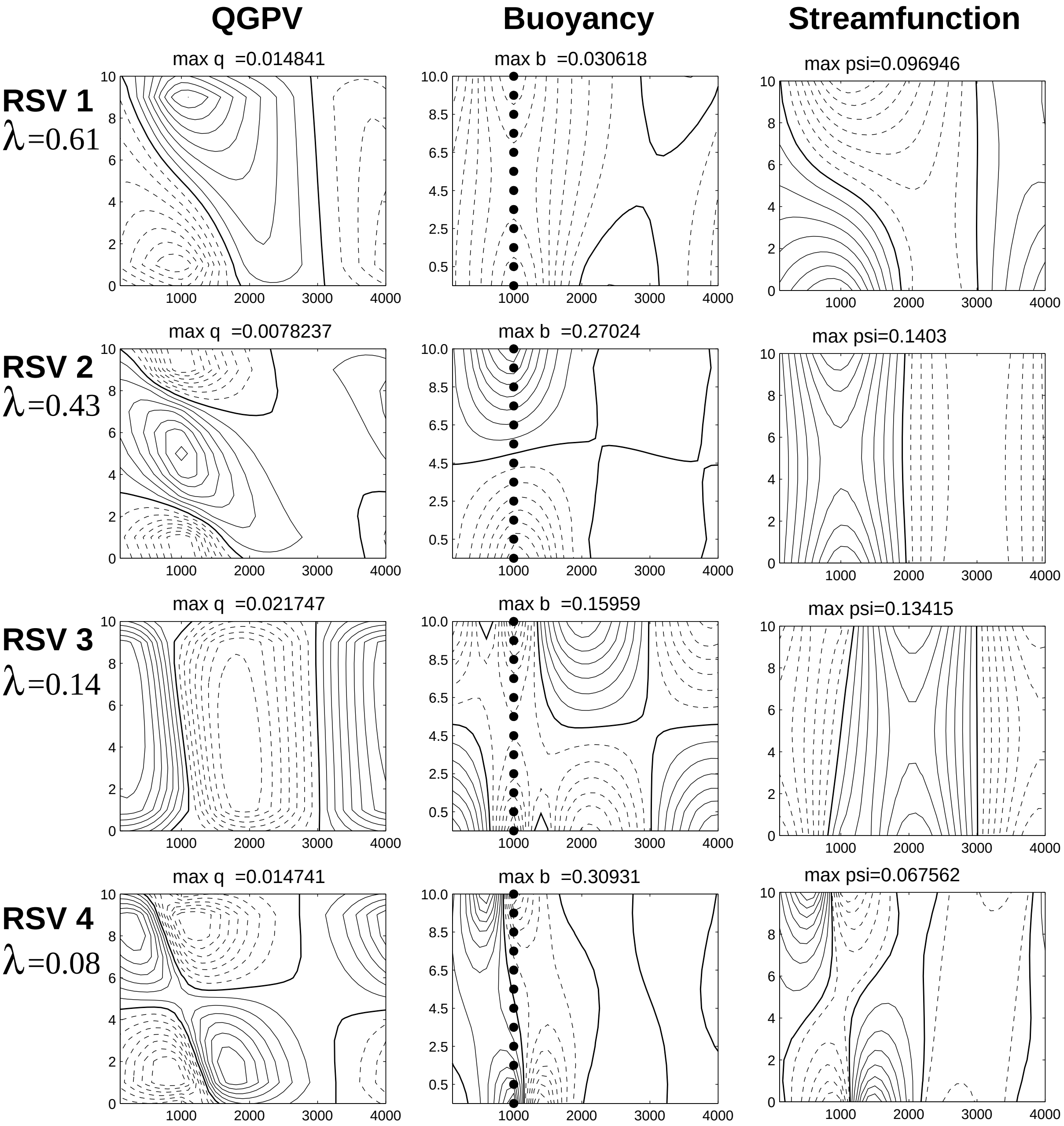


Figure 6.20: *The first four RSVs for the normalized observability matrix with a vertical line of buoyancy observations at T+0 and T+6. Horizontal correlations are applied with a length scale of $l = 5\Delta x$. Each row gives the QGPV field, Buoyancy field and streamfunction field for an RSV. The horizontal axes are the zonal directions(km) and the vertical axes are the height (km). ($\sigma_o^{-2} = 1$, $\sigma_q^{-2} = 1$, $\sigma_T^{-2} = 1$).*

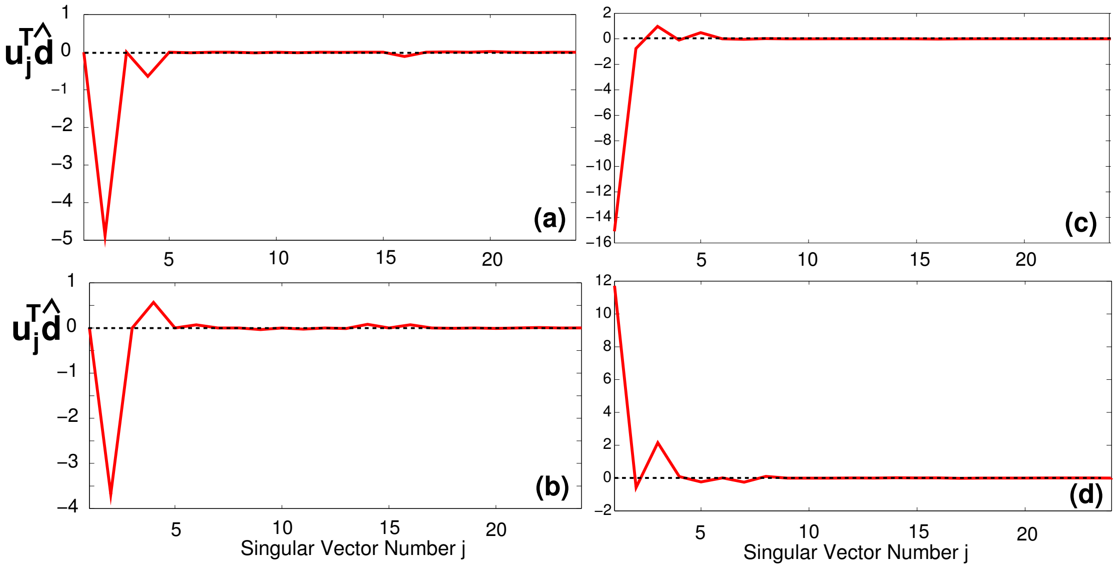


Figure 6.21: *Values of $\mathbf{u}_j^T\hat{\mathbf{d}}$, with perfect observations of a vertical line of buoyancy at T+0 and T+6 at (a)-(b) 2000km, (c)-(d) 3000km, and with the true state given by the most rapidly (a),(c) Growing and (b),(d) Decaying Eady wave. The background state is zero. The zero values are represented by the dashed line for clarity.*

Fig. 6.21(a) show the values of $\mathbf{u}_j^T\hat{\mathbf{d}}$ when the true state is given by the most rapidly growing Eady wave, with a background state of zero, and with a vertical line of buoyancy observations at 2000km. The values show that a large contribution comes from RSVs 2 and 4. Both RSVs 2 and 4 are given negative weights when the growing wave is observed. When this is repeated for the decaying Eady wave, a negative weight is given to RSV 2 and a positive weight is given to RSV 4 (Fig. 6.21(b)).

This is then repeated when the observations are at 3000km. The values of $\mathbf{u}_j^T\hat{\mathbf{d}}$ show that RSVs 1 and 3 give a large contribution to the analysis increment. When a growing mode is observed, a negative weight is given to RSV 1 and a positive weight is given to RSV 3. When a decaying mode is observed, negative weights are given to both RSV 1 and RSV 3.

The difference between the sign of the weights for the growing and decaying modes can be explained by the structures of the LSVs. Fig. 6.22 shows the first four LSVs. The LSVs are defined in observation space, so are vertical profiles at T+0 and T+6. LSVs 1 and 3 have structures that are symmetrical about the centre of the domain, whereas LSVs 2 and 4 have structures that are anti-symmetrical about the centre. It is for this reason that LSVs $1\&3$ and $2\&4$ occur together. The first two LSVs have very similar structures at the initial and the final time, whereas the second two LSVs change sign between the initial and final times. Thus, the second two LSVs are needed to determine the structure needed for growth or decay. For

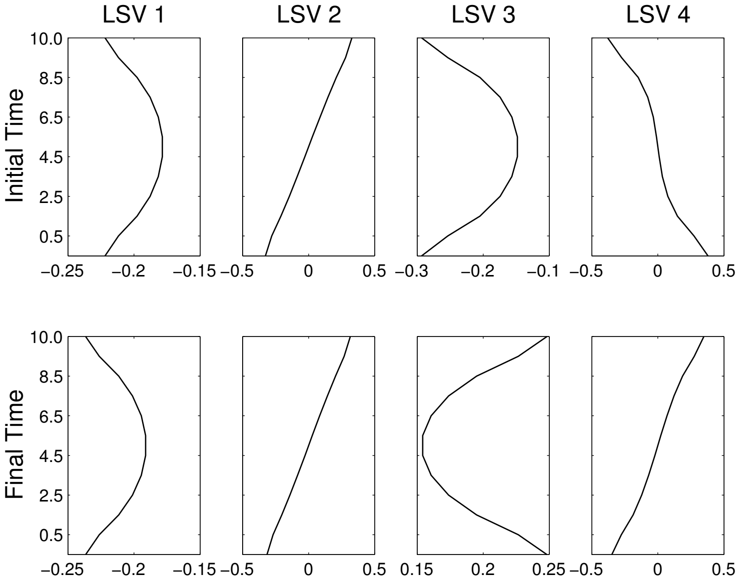


Figure 6.22: *The first four LSVs of the normalized observability matrix. The ordinate gives the height(km) and the abscissa gives the amplitude. Note that the abscissae use different scales.*

example, if LSV 1 and LSV 3 are added together, then a decaying structure is obtained, but if LSV 3 is subtracted from LSV 1, then a growing structure is obtained.

The analysis increments for non-modal growth had a very different structure to those for modal growth. This was due to the different specification of the regularization parameters. The SVD is now computed for the normalized observability matrix, $\hat{\mathbf{H}}\mathbf{B}^{\frac{1}{2}}$, where $\sigma_q^{-2} = 10^{-4}$ and $\sigma_T^{-2} = 1$ (and $\sigma_o^{-2} = 1$) correlations with a length scale of $l = 5\Delta x$ are also applied and where the basic state is such that the zonal wind is zero on the lower boundary. The first four RSVs are shown in Fig. 6.23. In Fig. 6.20, the RSVs were dominated by the buoyancy fields, whereas now the RSVs are dominated by the QGPV fields. This illustrates again how the effect of the covariance is to bias the analysis increments towards the expected structures. RSV 1 and RSV 2 have QGPV dipoles and buoyancy monopoles. For RSV1, the buoyancy is maximum to the west of the observations and therefore corresponds to the information contained in the final time observations. For RSV 2, the buoyancy is maximum to the east of the observations and therefore corresponds to the information contained in the initial time observations. RSV 3 contains three QGPV anomalies and RSV contains four QGPV anomalies. Thus, the vertical scale of the structures decrease with the singular values. This is similar to the assimilation of horizontal lines where the horizontal spatial scales decreased with increasing singular vector index.

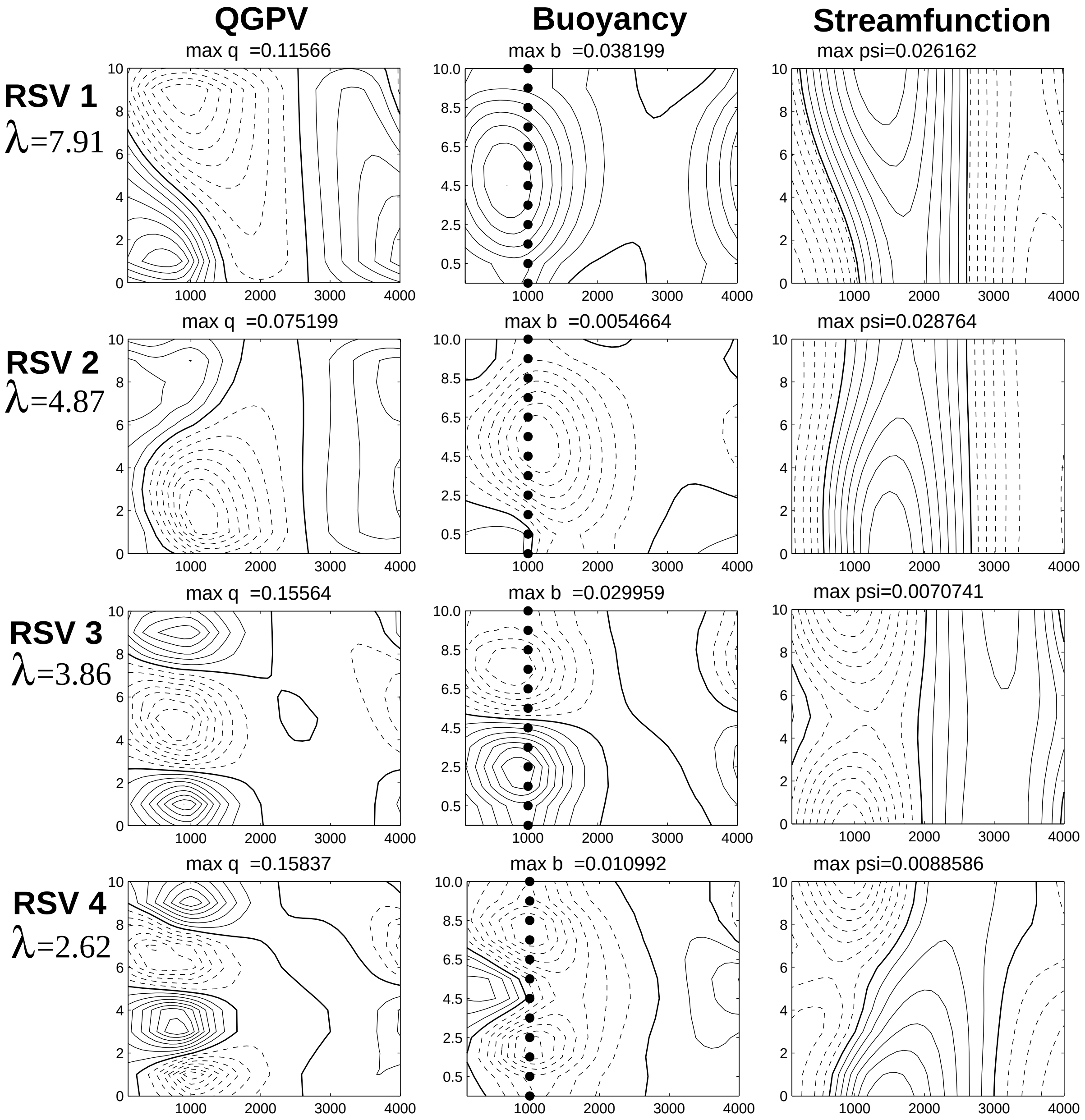


Figure 6.23: *As for Fig. 6.20, but with weights $\sigma_q^{-2} = 10^{-4}$ and $\sigma_T^{-2} = 1$, and for the basic state with zero flow on the lower boundary.*

To summarize, the 4D-Var experiments for modal growth and decay have shown that 4D-Var is able to generate analysis increments with the correct vertical structures. However, the decaying part of the increment is strongly penalized so that in the subsequent forecast, the analysis increment starts to grow instead of decaying. The 4D-Var experiments also showed that the analysis is better if the position of the observations is at the maxima and minima of the required analysis increment. The corresponding SVD experiments showed that the first two RSVs are needed to determine the general structure of the analysis increment whilst the second two RSVs are needed to determine the growth or decay of the analysis increment.

The 4D-Var experiments for non-modal growth have shown that 4D-Var is able to use vertical profiles of observations to generate the correct vertical structure provided that the regularization parameters are specified appropriately. With the basic state such that the zonal flow is zero on the lower boundary, 4D-Var is again able to infer the part of the state in the unobserved regions. This is clearly illustrated by the SVD experiments where the first RSV contains a maxima to the west of the position of the observations. This RSV corresponds to the information obtained from the final time observations, and as it has the largest singular value so that it is not strongly penalized by the background state.

6.6 Conclusions

This chapter has extended the idealized 4D-Var experiments to more realistic cases by considering the effect of background error correlations, different true states and different observing systems. In all the experiments in this chapter we have considered the assimilation of observations of the interior buoyancy field. The SVD results have shown that the results from previous chapters can be applied to the assimilation of horizontal lines of the interior buoyancy field. It has also been found that as the horizontal line is moved nearer to the middle of the domain, there is less distinction between the information needed to reconstruct the upper and lower boundaries, but still a large distinction between the growing and decaying parts.

The filtering and interpolating effect of background error auto-correlations have been understood from an SVD perspective. For dense observations, it has been found that correlations act to bias the analysis increments towards the expected spatial structures. This is achieved by penalizing the unexpected RSV structures with small spatial scales. For sparse observations, the RSVs of the normalized observability matrix have similar structures to the RSVs for a full line of observations. This means that it is possible to apply the previous conclusions to the assimilation of sparse data. However, to be able to retrieve the same amount of information from sparse observations, the observations must be a great deal more accurate than the full line of observations. Further, only the large-scale structures can be successfully reconstructed using sparse observations.

It was shown in Chapter 5 that the specification of the appropriate value of the ratio between the observation error variance and the background error variance is vital to extract the maximum amount of information. This ratio can be considered as a regularization parameter in the context of Tikhonov Regularization. This concept has now been extended to multiple regularization parameters. It has been found that the correct analysis increment structures for both modal and non-modal growth can be generated provided that the background error variances for the QGPV and buoyancy fields are specified appropriately. If 4D-Var uses 'climatological values' for these parameters, the growth rates of the subsequent forecast can be vastly different from the truth. For example, if the required analysis increment has a vertical structure which leads to rapid finite-time, non-modal growth, a relatively large analysis increment needs to be added to the interior. If, however, the background error variances are specified so that a relatively large analysis increment is added to the boundaries, the analysis will produce a forecast with a very small growth rate. Thus, the appropriate specification of the regularization parameters, on each analysis, is vital for the analysis of extreme weather events such as mid-latitude storms.

The 4D-Var experiments concerning the assimilation of two horizontal lines showed that the analyses for modal growth are improved if the lines of observations are moved further apart. The analyses for non-modal growth showed that the maximum in the analysis increment is found at the position of the observations if a large weight is given to the background state. It was also possible to infer the maximum in an unobserved region if more weight is given to the observations. The corresponding SVD experiments showed that the decaying part of the analysis increment is still strongly penalized despite the extra information about the vertical structure. The extra horizontal line is used to provide information about the state in between the two horizontal lines. The time-evolution information is used, for example, to infer the position of a temperature maximum in between the two lines of temperature observations. These RSVs have very small singular values, so a large weight must be given to the observations.

The experiments concerning the assimilation of vertical profiles showed that 4D-Var is able to use vertical profiles to generate the correct vertical structures. The analysis increments have structures that are extremely localized in the horizontal, however, instead of the large-scale structures produced by the assimilation of horizontal lines. The region surrounding the position of the observations can still be inferred from the time-sequence of observations. The 4D-Var experiments for modal-growth showed that both growing and decaying analysis increments can be generated, but the decaying part of the analysis increment is again strongly penalized so that the forecast quite rapidly gives growth instead of decay. The corresponding SVD experiments showed that the RSVs with large singular values are used to give the general structure of the analysis increment and result in growth. The RSVs with small singular values are used to give the correct sign of the growth rate. This is similar to the assimilation of horizontal lines. The 4D-Var experiments for non-modal growth showed that the appropriate choice of the regularization parameters is still vital in obtaining a good analysis, even though there is more information about the vertical structure. When the appropriate parameters are specified, an analysis increment resulting in rapid finite-time growth can be obtained. When the basic state flow is such that the perturbation is advected through the position of the observations, 4D-Var is still able to infer the position of the perturbation, although if there are fewer observations at the maximum of the perturbation, the growth rate of the analysis increment is reduced significantly. The corresponding SVD experiments showed that the RSVs with the largest singular values correspond to the information needed to infer the position of the maxima, and the RSVs with smaller singular values correspond to the information needed to reconstruct smaller scale structures. Thus, there is a great deal of useful information that can be extracted from the observations, provided that the appropriate regularization parameters are specified.

In conclusion, we have shown that the results from previous chapters can be extended to more realistic situations. It was expected that the assimilation of vertical profiles would give much better analyses of the vertical structure than the assimilation of a single horizontal line. Indeed, the vertical profiles are able to generate vertical structures for both modal and non-modal growth. If the non-modal perturbations are advected through the position of the observations by the basic state flow, 4D-Var is still able to infer the position of the maxima. With a few, sparse, vertical profiles, the SVD with correlations has shown that the time-evolution information in the horizontal can still be used to infer the vertical structure. However, we have also shown that the decaying parts of the analysis increment are still strongly penalized by the background state and that the vertical structures leading to either modal or non-modal growth can only be achieved if the appropriate regularization parameters are chosen and if the observations are sufficiently accurate.

Chapter 7

Conclusions

4D-Var is one of the most advanced data assimilation algorithms to be used in operational numerical weather prediction, as it combines the information from the observations with the knowledge of the atmospheric dynamics and physics. The aim of this thesis is to understand the extent to which 4D-Var can develop the vertical structures needed for the growth and decay of baroclinic systems.

Most of the development of 4D-Var has previously been made in the context of the full weather forecast problem; here, 4D-Var has been examined using idealized case studies with the Eady model of baroclinic instability. A 4D-Var algorithm using the Eady model has been developed. This included the development of the adjoint model, a background error correlation model based on Laplace smoothing and a comparison of minimization algorithms.

A novel technique for examining the information content of observations in 4D-Var has been developed. The technique is a straightforward temporal extension of methods that are commonly used to examine the information content of observations in satellite retrievals and is based on the singular value decomposition (SVD) of the normalized observability matrix. The technique has enabled the gaining of a new understanding of how the information from a time-sequence of observations is combined with the model dynamics in 4D-Var.

The majority of experiments in this thesis have considered how 4D-Var can use a time-sequence of the lower level wave to reconstruct the position of the upper level wave. These experiments have provided an understanding of both the reconstruction of the state in unobserved regions and the generation of the vertical structures needed for baroclinic growth or decay. These experiments were extended in the Chapter 6 to more realistic cases. In particular, the effect of correlations, different true states and different observing systems were considered.

This final chapter returns to the key questions that were posed in Chapter 1, discusses how the results from this thesis can be applied to operational NWP, and ends with a discussion of some of the possible directions for future work.

7.1 Answers to the Key Questions

The results from this thesis are now summarized by returning to the three key questions that were posed in Chapter 1.

1. How are observations used in 4D-Var?

A time-sequence of observations provides information about the atmospheric state in the region that is observed; through the use of the equations for the time-development of the system, it also provides information about the state in unobserved regions. For example, the experiments with horizontal lines of observations showed that 4D-Var is able to use the time-evolution information to infer the rate of growth or decay and is hence able to infer the vertical structure. The experiments with vertical lines of observations showed that 4D-Var is able to link the observations together through the model advection to infer the vertical tilt of the state near the observations.

It was shown in Chapter 4 that the 4D-Var analysis increments can be written as a linear combination of the RSVs of the 4D-Var observability matrix. This formulation was used in Chapters 5 and 6 to examine the information content of observations in 4D-Var. By considering the true state given by the most rapidly growing or decaying Eady wave, with either horizontal or vertical lines of perfect observations, it was shown that the RSVs with large singular values contain the information needed to infer the state in the observed regions, whilst the RSVs with small singular values contain the information needed to infer the state in the unobserved regions.

It was also shown that when the observations have errors, the RSVs with smaller spatial scales are also given significantly large weights. These RSVs completely dominate the analysis increment so that unphysical structures are generated in the unobserved regions. Thus, the analysis is extremely sensitive to the observational noise, and for this reason, 4D-Var with no J^b term can be considered as a discrete ill-posed inverse problem, even if there are enough observations to define a unique solution. Such a problem may be solved using Tikhonov regularization; in 4D-Var, this is equivalent to adding the background term to the cost function.

Thus, a link between the literature on 4D-Var and the literature on Tikhonov regularization, which is used to solve many types of ill-posed inverse problems, has been established.

The background state acts to penalize the RSVs with small singular values and small spatial scales that correspond to noise. This is necessary to create a smooth analysis. However, the background state may therefore also penalize the RSVs that contain the information needed to reconstruct the state in the unobserved regions. The weight given to the background state in comparison to the observations determines how many RSVs are penalized. This signal-to-noise ratio can be considered as a regularization parameter, and it is important to specify the appropriate value so that the maximum amount of useful information can be extracted from the observations, but that the analysis is sufficiently smooth.

The background term also provides a priori information such as auto-correlations. Correlations act to filter the observational noise in dense data regions so that the analysis is smooth. From an SVD perspective, the correlations act to bias the analysis increments towards the expected large-scale structures, by penalizing the unexpected structures. Such correlations allow more of the useful information to be extracted from the observations as it is possible to penalize the noisy structures without penalizing the information needed to reconstruct the state in the unobserved regions. Background error correlations also act to interpolate the information from sparse observations. This means that a time-sequence of sparse observations may also be used to reconstruct the state in the unobserved regions provided that only the large scales need to be reconstructed and also that the observations are sufficiently accurate.

The 4D-Var algorithm may also be considered to have multiple regularization parameters so that the background state has different error variances for each variable. Again, from an SVD perspective, this allows the analysis increments to be biased towards the expected structures and allows the information from the observations to be used in a better way.

2. Why has 4D-Var been shown to perform well in regions of baroclinic instability?

If observations are only given at the end of the window, then the analysis increment can be considered as a linear combination of optimal perturbations. These structures maximize the amount of growth during the assimilation window, so that an analysis increment with a small amplitude is added to the background state but that the analysis is close to the observations at the final time.

With a horizontal line of observations at only one time level, it is not possible to infer the growth rate. Therefore, to infer the vertical structure, observations are required at two time-levels or more. The SVD experiments showed that with observations at both the beginning and the end of the window, the analysis increments are a linear combination of both growing and decaying structures. This means that both growing and decaying analysis increments can be generated. However, the decaying vector has a small singular value and so it is strongly penalized by the background state. Therefore, even with a time-sequence of observations, 4D-Var is likely to add a growing analysis increment.

It is for this reason that 4D-Var performs well in regions of baroclinic growth. Although a growing analysis increment is added to the background state, it is not necessarily the correct flow-dependent structure. For example, if a decaying analysis increment is required, but a growing analysis increment is added, then this will be completely detrimental to the forecast.

3. How can the benefits of 4D-Var be maximized?

This work has highlighted two ways to maximize the benefits of 4D-Var: the initial and final observations should be as far apart as possible in time, and the appropriate values for the regularization parameters should be specified.

The experiments concerning the temporal position and weights given to the observations in the assimilation window showed that the best analyses are achieved if the observations are as far apart as possible in time and if more weight is given to the final time observations than the initial time observations. Thus, the assimilation window should be designed to be as long as possible (within the validity of the tangent linear assumption), and such that there are many observations at both the beginning and the end of the window. This agrees with the results by Thépaut et al. (1996), where it was shown that the structure functions are more fully developed for a longer assimilation window. However, it also implies that the observations at the beginning of the window also play a crucial role in generating vertical structures with the correct attributes.

The experiments also showed that the position of the observations should ideally be near to the maxima and minima in the required analysis increments. For example, if the required analysis increment has a structure which leads to modal growth, the observations should be placed near to the upper and lower boundaries, but if the required analysis increment has a structure which leads to non-modal growth, the observations should be placed in the interior. It is possible to design the observing system so that the observations are far apart in time. However, the true state and hence the required analysis increments are unknown. Therefore, it is not possible to know in advance the optimal spatial positions for the observations.

If the observational errors are larger than the amount of growth during the assimilation window, it is not possible to infer the growth rate. Thus, a time-sequence of noisy observations may not be able to provide any information about the growth rate. In such instances, the knowledge of the model dynamics will not be able to reconstruct the state in the unobserved regions. Therefore, to maximize the benefits of 4D-Var, the observations need to be accurate enough to infer the growth rate or vertical tilt. If the observations are not accurate enough to infer such information, it is doubtful that 4D-Var would be significantly beneficial in comparison to 3D-Var. Although flow-dependent structures would be generated, they would not necessarily have the correct attributes.

To maximize the amount of useful information that is extracted from the observations, it is important to choose the appropriate value for the regularization parameter μ^2, which is the ratio between the observation error variance and the background error variance. It may seem from the BLUE equations (1.7 and 1.8), that the error covariance matrices need to be known a priori. However, the discussion in this thesis has shown that the observations and the background state can be used to specify the appropriate value for μ. This was illustrated using a technique known as the L-Curve, which is commonly used in problems involving Tikhonov Regularization.

Not only is it important to specify the ratio between the J^b and the J^o terms, but also to specify multiple regularization parameters. For example, the background error variances may be different for different variables in different geographical regions and at different times. As the parameters differ for each analysis, they must be re-specifed on each analysis cycle. The specification of such parameters is particularly important for the analysis of extreme weather events such as mid-latitude storms and can be considered as a technique to generate flow-dependent analysis increments.

7.2 Implications for Operational NWP

The experiments in this thesis are highly idealized, in comparison to operational data assimilation. The Eady model is extremely simple in comparison to full NWP models and in particular only a linear (although not linearized), perfect model with simple true states has been considered. Nevertheless, the understanding of how the information from observations is combined with the model dynamics can be applied to understand the processes in operational 4D-Var. The benefits of 4D-Var in comparison to 3D-Var, the implications for an algorithm known as the reduced rank Kalman Filter, and techniques to maximize the benefits of 4D-Var are now discussed.

It is important to assess the benefits of 4D-Var, given that 4D-Var is significantly more expensive than 3D-Var (Rabier et al., 1998), and that 4D-Var requires a great deal of time to develop the tangent linear and adjoint models. Previous studies (Rabier et al., 2000) have shown that 4D-Var has a clear advantage over 3D-Var in regions of rapid cyclogenesis. Such benefits have been attributed to the dynamical evolution of the covariance matrix (e.g. Thépaut et al., 1993a). The work in this thesis has also shown that with observations at only the end of the window, the vertical structure of the analysis increments are more fully developed and lead to a faster growth rate. However this is not necessarily better if the required analysis increment is decaying. Observations at the beginning of the assimilation window are also needed so that 4D-Var is able to infer the growth rate. We have shown that 4D-Var is able to use a time-sequence of observations to infer the state in unobserved regions, and to infer the growth rate during the assimilation window. It is these two abilities that give 4D-Var clear advantages in comparison to 3D-Var. Thus, the dynamical evolution of the covariance matrix is not necessarily the only reason for the benefits of 4D-Var. This has important implications for the development of data assimilation algorithms such as simplified Kalman Filters and Ensemble Kalman Filters.

Experiments with the Reduced Rank Kalman Filter (RRKF) have shown that the RRKF has an entirely neutral impact on the analysis quality (e.g. Fisher and Andersson, 2001, Beck, 2003). The reasons for the neutral impact are not understood, but the work from this thesis may aid towards an understanding. 4D-Var is likely to add a growing analysis increment even if a decaying analysis increment is required, and hence 4D-Var already handles the growing structures well. This is therefore perhaps the reason why evolving the covariances corresponding to the growing structures has little impact on 4D-Var. Further research is required to investigate whether this is indeed the case.

This work has highlighted that the benefits of 4D-Var are not only due to the propagation of the covariance matrix, but to the use of time-evolution information. The time-evolution information that is contained in the observations plays an important role in ensuring that the analysis leads to a good forecast. The dramatic increase in the number of observations in the future should also give more time-evolution information. Hence, it should be expected that the benefits of 4D-Var will become more apparent in the future.

It is vital that the maximum amount of useful information contained in the observations is extracted. This thesis has identified two ways to maximize the amount of information that is extracted: to use observations that are placed far apart in time, and to choose the appropriate regularization parameters.

Previous research concerning targeted observations has mainly considered where the sensitive atmospheric regions are, but here we have considered where the observations should be placed in both space and time so that the data assimilation algorithm can extract the maximum amount of useful information. Previous studies have shown that it is important to place observations at the end of a long assimilation window. The work in this thesis has highlighted that the observations at the beginning of the assimilation window also play an important role, particularly for the accuracy of the following forecast. For operational data assimilation, it is likely that it is still the case that the initial observations are important, although the observations in the middle of the window may also be important in the case of, for example, non-modal growth. Therefore, the observing system and assimilation window should be designed with this in mind.

The specification of the multiple regularization parameters has been shown to play a vital role in extracting the information contained in the observations. This has previously been considered (e.g. Wahba and Wendelberger, 1980, Dee, 1995, Desroziers and Ivanov, 2001), but has not been implemented in an operational data assimilation scheme, although online covariance estimation is currently being developed in the HIRLAM (High Resolution Limited Area Modelling) variational data assimilation system (M. Lindskog, personal communication). The work in this thesis has illustrated that the specification of such parameters is vital to exploit the benefits of 4D-Var and therefore further research is needed to identify a robust and feasible method to calculate the appropriate values. This is discussed further in the next section.

7.3 Future Development

Having begun to answer the questions which were posed in Chapter 1, we now consider how this work may be extended.

1. How can better analyses of the decaying modes be obtained?

One of the main conclusions from this thesis is that the decaying part of the analysis increment is strongly penalized by the background state. The reason for this is that the control variables are defined at the beginning of the window. This is a result of the reduction of the size of the problem using the ‘reduction of the control variable’ (Le Dimet and Talagrand, 1986). If the control variables were at the end of the window, then the growing structures would be penalized instead (Pires et al., 1996). Thus, to eliminate the distinction between the growing and decaying modes, it would be necessary to reformulate the 4D-Var problem so that the control variables are given by the state vector at every time level, or at least by the state at both the beginning and the end of the time window. For example, the model may be added as a weak constraint and an elliptic problem (rather than a hyperbolic problem) solved (Sasaki, 1970). Such an approach is currently being considered by Juckes (2003a,b).

2. What technique should be used to calculate the regularization parameters?

This work has also shown that it is important to specify the appropriate values for the regularization parameters, which are the ratios of the error variances. In particular, the estimation of multiple parameters is particularly important for the analysis of extreme weather events. Although the true state is unknown, it is possible to obtain the appropriate values for these parameters from the data (the background state and the observations). This was demonstrated with the L-Curve for a single parameter μ in Chapter 5. It is possible to extend the L-Curve framework to an L-hypersurface to consider multiple parameters (Belge et al., 2002). This has not been addressed in this thesis, but illustrates that it is possible to obtain the appropriate values for multiple regularization parameters, from the data. Both the 4D-Var and SVD experiments showed that the specified background error correlations act to bias the analysis increments towards the expected structures. One question that follows from this is whether it is possible to use the data to specify the appropriate values for correlation length scales. Again, this has not been addressed in this thesis, but is considered by Wahba and Wendelberger (1980).

An important question that has not been addressed is how the appropriate values should be calculated. There have been a number of suggestions for the calculation of such parameters and these are now briefly outlined.

The L-Curve (Hansen, 2001), described in Chapter 5, is calculated by repeating the 4D-Var analysis many times with different parameters. This is costly for large problems, and is therefore not appropriate for operational data assimilation. Wahba and Wendelberger (1980) suggested the use of a method known as Generalized Cross-Validation (GCV). This uses the criteria that a good choice for the regularization parameters is the ability to predict the value of the field where the observational data are withheld and calculates the appropriate parameter by minimizing a GCV function. Dee (1995) suggested a method based on the assumption that the covariance matrix of the innovation vectors has a Gaussian distribution. The parameter, based on this maximum-likelihood concept, is also found by minimizing a function.

Both the GCV and the maximum-likelihood techniques require the evaluation of the trace of large matrices. Therefore, the minimization of the functions is non-trivial. Dee (1995) used a simple descent algorithm, with no gradient evaluation, to find the minimum, but a more sophisticated algorithm would be required if many parameters needed to be estimated. To reduce the computational cost, the problem would need to be divided into sub-domains. A more sophisticated approach is to estimate upper and lower bounds of the functions using randomized trace estimation, Gauss quadrature and Lanczos bidiagonalization. Golub and von Matt (1996) describe such an approach, and also illustrate how the minimum of the GCV function may be found. Fisher (2003) has also used such a method to calculate the degrees of freedom in the ECMWF 4D-Var system.

An alternative method to estimate the parameters is to use iterative tuning. Talagrand (1998) showed that if the background and observation error statistics have been specified correctly, the value of the cost function at the minimum should be equal to the number of observations (see also, Rodgers (2000)). Based on this concept, Desroziers and Ivanov (2001) suggested an iterative technique to tune covariance parameters so that the correct value of the cost function is attained. They showed that it is possible to use such a technique to tune the observation errors in a 3D-Var global analysis.

It is not clear which approach would be most suitable for the purposes of data assimilation. Therefore, future work is needed to identify a method to calculate the appropriate values that is robust, and possible to use within a global 4D-Var scheme.

The analyses using different regularization parameters had very different structures and resulted in very different forecasts. Therefore an alternative possibility is not to tune the parameters, but to repeat the analysis using different parameters to generate an ensemble of initial conditions that are consistent with the background state, the observations and the model. This is perhaps an important application as ensemble forecasting is likely to become increasingly important in the future.

3. What is the effect of model error, nonlinearity, vertical correlations, cross-correlations and temporal observation correlations?

This thesis has not considered the effect of model error, nonlinearity, vertical correlations, cross-correlations and temporal observations correlations; these are all important issues for operational data assimilation.

There are a number of issues concerning the model error. If the model is wrong but model error is not taken into account in the data assimilation algorithm, then it is likely that the state at the end of the assimilation window has larger errors than at the beginning of the window, and hence it may be the case that the initial observations are even more important than the final observations. A second issue is that the model may propagate the wrong information into the unobserved regions and hence the benefits of 4D-Var may be lost. The third issue is to consider the case when model bias or model parameters are estimated by the data assimilation algorithm as well as the initial conditions. Then, it would be interesting to extend the information content concepts to understand whether the observations contain enough information to estimate the model errors.

Only a linear model has been considered in this study, therefore it is important to extend the studies to understand the effect of nonlinear models. The 4D-Var problem with a non-linear model is solved as a sequence of linear problems in incremental 4D-Var. The singular vectors will depend on the linearization state, which may change throughout the window, and will also change on each outer loop. Therefore, the particular questions to answer are: how do the singular vectors of the observability matrix depend on the linearization state, and how is the information from the observations used to change the linearization state?

This thesis has only considered the effect of horizontal background error correlations and not vertical correlations. In the Eady model, the elliptic QGPV equation provided some vertical correlations so it was not necessary to apply vertical correlations. However it would be of particular interest to understand the effect of vertical correlations on the assimilation of growing and decaying modes, where the vertical structure is important for the growth rate.

Cross-correlations between different variables are an important part of data assimilation to ensure that analysis fields are balanced. They also allow unobserved model fields to be inferred. The SVD approach may aid the understanding of the impact of such cross-correlations in a 4D-Var algorithm.

Temporal observation correlations are becoming increasingly important, as the amount of satellite and radar data increases; however, it is not clear how to account for such correlations in 4D-Var. It is possible to included temporal error correlations in the definition of the normalized observability matrix, and such a formulation may aid an understanding of the effect of such correlations.

4. What is the information content of observations in 4D-Var with different dynamical *models?*

The singular value decomposition of the observability matrix has provided a useful understanding of how the information from observations is combined with the model dynamics, and in particular to understand how the state in unobserved regions is reconstructed. This concept could usefully be applied to other data assimilation problems such as data assimilation in the tropics and mesoscale data assimilation.

There are two main difficulties for atmospheric data assimilation in the tropics. The first is that there is a wide range of types of wave motion in the tropics (Holton, 1992), and the second is that there are very few wind observations in comparison to mass observations and therefore the mass field is needed to infer the wind field (Brzovic, 2003). It would be useful to use the singular value decomposition to examine the information that is contained in the observations. In particular, this technique could be used to determine whether the mass field can be used to successfully reconstruct the wind field, and to understand whether analysis increments corresponding to different wave structures can be generated.

An important current area of research is the assimilation of precipitation (radar) data into mesoscale models. One of the main problems in the assimilation of precipitation data is that in many cases, the impact of the assimilation of precipitation data only remains for the first several hours of the forecast. This is because the model temperature and humidity profiles are not adjusted appropriately and therefore cannot continue to give the required precipitation. Therefore, the question is whether it is possible to use observations of precipitation to infer the necessary temperature and humidity fields (Jones and Macpherson, 1997). This is a challenging question as both nonlinearity and model error are important. The SVD technique could be applied in this context to examine the information that is contained in the precipitation observations.

5. How can better analyses be obtained when the background state has a phase error?

Current data assimilation methods blend together the observations and the background state. Many of the experiments in this thesis have considered the background state to have a displacement error, and have shown that it is possible for the amplitude of the analysis to be reduced. Such an effect would be more noticable in a region with sharp gradients, such as a front. If a front is in the wrong place in the background field, then it is possible for the feature to become smeared out in the analysis. It is therefore important to consider alternative data assimilation algorithms that aim to extract the maximum amount of useful information from both the observations and the background state.

A possible algorithm would be to allow for both amplitude and phase errors, so that it is possible to shift the background state closer to the observations. Such a technique could be considered as a technique to generate flow-dependent background error covariance structures, as the analysis increments would depend explicitly on the background state.

Appendix A

Eady Model

In this thesis, the non-dimensional 2D-Eady model is used in 4D-Var identical twin experiments where the true state is given by either the most rapidly growing or decaying Eady wave or by an interior QGPV dipole perturbation that results in non-modal growth.

This appendix begins with a description of the quasi-geostrophic equations, from which the 2D Eady model equations are derived. This is followed by a description of the non-dimensional variables and a co-ordinate change. The equations used for the modal and non-modal initial conditions are then given. The appendix ends with a description of the particular discretization of the Eady model that is used in this thesis and also the details of the methods used to handle observations of the interior buoyancy and to calculate the SVD of the observability matrix.

A.1 Quasi-Geostrophic Equations

The quasi-geostrophic (QG) equations are an approximation to the primitive equations and describe the essentially geostrophic motion for mid-latitude synoptic scales. The equations are simplified by using Cartesian co-ordinates and assuming that the atmosphere is shallow and hydrostatically balanced. Frictional and diabatic effects are also neglected.

The Boussinesq approximation is used to simplify the equations. We consider the fluctuations to the static state of the atmosphere such that the basic state is only a function of height:

$$\theta = \theta_o(z) + \theta'(x, y, z, t) \tag{A.1}$$

$$p = p_o(z) + p'(x, y, z, t) \tag{A.2}$$

$$\rho = \rho_o(z) + \rho'(x, y, z, t) \tag{A.3}$$

where θ denotes the potential temperature, p denotes pressure, ρ denotes the density, and x, y, z and t are the zonal, meridional and vertical co-ordinates and time. It is assumed that the vertical motion is small in comparison to the height and that the motion is anelastic. It is also assumed that the inertial effects of the variations in the basic state density can be ignored, but the buoyancy effects cannot. The Boussinesq approximation also means that there is no variation in the height of the tropopause, H.

The static stability N^2 of the basic state (or Brunt-Väisälä frequency N) is defined as:

$$N^2(z) = g\frac{d\ ln\theta}{dz} = \frac{g}{\theta}\frac{d\theta_o}{dz} \tag{A.4}$$

and it is assumed that the perturbation stratification ($\frac{d\theta'}{dz}$) is negligible.

The quasi-geostrophic equations are derived from a scale analysis of the primitive equations, based upon typical mid-latitude synoptic scales. Typical synoptic scale lengths for the atmosphere at mid-latitudes are the horizontal velocity scale $U\sim 10ms^{-1}$, the horizontal length scale $L\sim 1000km$, the height of the tropopause $H\sim 10km$, the Coriolis parameter $f\sim 10^{-4}s^{-1}$ and the static stability $N^2\sim 10^{-4}s^{-2}$. Using such values, it can be assumed that the Rossby number, Ro is small:

$$Ro = \frac{U}{fL} \ll 1. \tag{A.5}$$

This is equivalent to assuming that the relative vorticity ξ is small in comparison to the planetary vorticity f. It can also be assumed that the stratification parameter (or Burger number), Bu is unity:

$$Bu = \left(\frac{NH}{fL}\right)^2 = \left(\frac{L_R}{L}\right)^2 = 1 \tag{A.6}$$

where L_R is the Rossby radius of deformation.

From the scale analysis (Pedlosky (1987), Holton (1992), James (1994) and Muraki et al.

(1999)), the thermodynamic equation reduces to:

$$D_g b + wN^2 = 0 \tag{A.7}$$

where $b = \frac{g}{\theta}\theta'$ is the buoyancy, g is the gravitational constant, w is the vertical velocity, and $D_g = \frac{\partial}{\partial t} + u_g \frac{\partial}{\partial x} + v_g \frac{\partial}{\partial y}$ is the geostrophic derivative where u_g and v_g are the geostrophic velocities. This QG equation states that potential temperature is conserved following dry adiabatic motion and that as air rises, it cools via adiabatic expansion.

Similarly, the vertical component of the vorticity equation reduces to:

$$D_g(f + \xi_g) = \frac{f_o}{\rho_o}\frac{\partial}{\partial z}(\rho_o w) \tag{A.8}$$

where $f = f_o + \beta y$ is the Coriolis parameter and $\xi_g = \frac{\partial v_g}{\partial x} - \frac{\partial u_g}{\partial y}$ is the geostrophic relative vorticity. This QG equation implies that with no vertical motion at the ground, mid-tropospheric ascent implies that the column of air is stretched, so that the absolute vorticity $f + \xi_g$ will increase by an amount proportional to the product of column stretching and the planetary vorticity, f_o. Note that this equation only contains a vortex stretching term and not a vortex tilting (or twisting) term (Hoskins, 1997).

The QG thermodynamic equation can be combined with the QG vorticity equation, by eliminating the vertical velocity w, to give the QG potential vorticity equation:

$$D_g q = 0 \tag{A.9}$$

where

$$q = f + \xi_g + \frac{1}{\rho_o}\frac{\partial}{\partial z}\left(\frac{\rho_o f_o b}{N^2}\right) \tag{A.10}$$

and is referred to as the QG potential vorticity (PV). This elliptic equation describes the so-called ‘invertibility principle’ as, when suitable boundary conditions are provided, this equation can be used to derive the primitive variables such as temperature, pressure and horizontal and vertical winds. Equation A.9 states that QGPV is conserved following horizontal, geostrophic, adiabatic, frictionless motion.

The QGPV equation is now rewritten in terms of the geostrophic streamfunction ψ, which is defined as:

$$\psi = \frac{p'}{\rho_o f_o}. \tag{A.11}$$

The geostrophic velocities and buoyancy can then be written in terms of streamfunction; from Geostrophic balance:

$$(u_g, v_g) = (-\frac{\partial \psi}{\partial y}, \frac{\partial \psi}{\partial x}), \tag{A.12}$$

and from Hydrostatic balance:

$$b = \frac{g}{\theta_o}\theta' = f\frac{\partial \psi}{\partial z}. \tag{A.13}$$

To summarize, the QG equations are given by:

$$D_g q = 0 \qquad \textit{The QG Potential Vorticity Equation}, \tag{A.14}$$

$$D_g b + N^2 w = 0 \qquad \textit{The QG Thermodynamic Equation}. \tag{A.15}$$

Thus, the fundamental dynamical variables are QGPV and potential temperature (buoyancy) and these are related to streamfunction by:

$$q = f + \nabla_h^2 \psi + \frac{1}{\rho_o}\frac{\partial}{\partial z}\left(\rho_o \frac{f_o^2}{N^2}\frac{\partial \psi}{\partial z}\right) \qquad \textit{Definition of QGPV}, \tag{A.16}$$

$$b = f\frac{\partial \psi}{\partial z} \qquad \textit{Definition of Buoyancy}. \tag{A.17}$$

where $\nabla_h^2 = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial z^2}$. Thus, the buoyancy provides suitable boundary conditions for the elliptic equation for streamfunction.

A.2 The 2D Eady Model

The 2D Eady model (Eady, 1949) equations are now derived from the QG equations. It is assumed that there are two rigid boundaries; one at the ground and one at a height H to represent the tropopause. The tropopause may be modelled as a rigid boundary due to the high static stability of the stratosphere. It is also assumed that there is no vertical motion at these rigid surfaces, the density ρ_o and the Static stability N are constants and the Coriolis parameter f is a constant (f-plane approximation, $\beta = 0$).

The Eady model equations describe the linear evolution of perturbations from a basic state. The basic state is assumed to be only dependent on y, whilst the perturbations are independent of y,

$$q = \bar{q}(y) + q'(x,z,t), \tag{A.18}$$

$$b = \bar{b}(y) + b'(x,z,t). \tag{A.19}$$

The basic state consists of a uniform meridional temperature gradient that is necessary for baroclinic instability. Through thermal wind balance, this is associated with a linear vertical zonal wind shear:

$$\bar{u} = Az. \tag{A.20}$$

The basic state is illustrated in the schematic diagram in Fig. A.1. The Coriolis parameter is a constant, so $\bar{q} = f$. The zonal wind perturbation is zero as $u'_g = -\frac{\partial\psi'}{\partial y}$ and ψ' is assumed to be independent of y. This allows the derivation of a linear model of perturbations which have no constraint on their size. That is, we do not need to linearize the model about the basic state as there is no need to neglect any small terms. The associated basic state and perturbation variables can be written as:

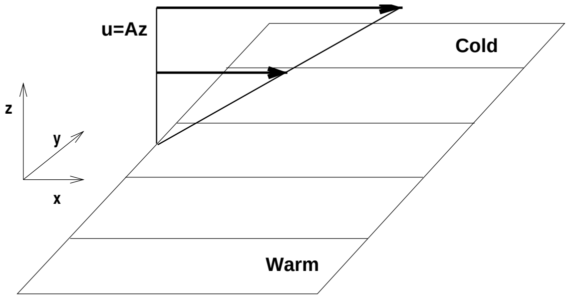


Figure A.1: *Basic state of the Eady model. The meridional temperature gradient is associated with a linear zonal wind shear with height, through thermal wind balance. The model contains rigid lids at both the ground and the tropopause.*

$$u_g = \bar{u}(z) + 0 \tag{A.21}$$

$$v_g = 0 + v'(x,z,t) \tag{A.22}$$

$$\psi = \bar{\psi}(y,z) + \psi'(x,z,t). \tag{A.23}$$

Using these definitions, the QG potential vorticity equation (A.14) becomes:

$$\left(\frac{\partial}{\partial t} + \bar{u}\frac{\partial}{\partial x} + v'\frac{\partial}{\partial y}\right)(\bar{q} + q') = 0 \tag{A.24}$$

which gives:

$$\left(\frac{\partial}{\partial t}+\bar{u}\frac{\partial}{\partial x}\right)q'=0 \text{ in } z\epsilon(0,H) \tag{A.25}$$

where from (A.16),

$$q'=\frac{\partial^2\psi'}{\partial x^2}+\frac{f^2}{N^2}\frac{\partial^2\psi'}{\partial z^2} \text{ in } z\epsilon(0,H). \tag{A.26}$$

This elliptic equation requires suitable boundary conditions. These are provided by the buoyancy from the Hydrostatic equation (A.17). Assuming that the vertical velocity w is zero on the upper and lower boundaries, then the QG Thermodynamic equation (A.15), that advects the buoyancy field, becomes:

$$\left(\frac{\partial}{\partial t}+\bar{u}\frac{\partial}{\partial x}+v'\frac{\partial}{\partial y}\right)(\bar{b}+b')=0 \text{ on } z=0,H \tag{A.27}$$

which gives:

$$\left(\frac{\partial}{\partial t}+\bar{u}\frac{\partial}{\partial x}\right)b'+v'\frac{\partial\bar{b}}{\partial y}=0 \text{ on } z=0,H. \tag{A.28}$$

Now, from (A.12, A.13, and A.20),

$$v'\frac{\partial\bar{b}}{\partial y}=\frac{\partial\psi'}{\partial x}\frac{\partial}{\partial y}\left(f\frac{\partial\bar{\psi}}{\partial z}\right)=\frac{\partial\psi'}{\partial x}\frac{\partial}{\partial z}\left(\frac{\partial\bar{\psi}}{\partial y}\right)f=-\frac{\partial\psi'}{\partial x}\frac{\partial}{\partial z}\bar{u}(z)f=-\frac{\partial\psi'}{\partial x}Af \tag{A.29}$$

so that the Thermodynamic equation becomes:

$$\left(\frac{\partial}{\partial t}+\bar{u}\frac{\partial}{\partial x}\right)\frac{\partial\psi'}{\partial z}=A\frac{\partial\psi'}{\partial x} \text{ on } z=0,H. \tag{A.30}$$

The left hand side describes the zonal advection of the temperature wave, whilst the right hand side describes the meridional advection. This meridional advection provides the crucial coupling between the upper and lower waves.

A.3 Non-Dimensional Equations

New variables and a co-ordinate change are now introduced to non-dimensionalize the Eady model. These are introduced for two reasons. The first reason is that the non-dimensional variables simplify the equations by removing the constants. The second reason is that the Eady model is used in an optimization problem, and therefore non-dimensionalizing should improve the conditioning of the problem (Gill et al., 1981).

The following co-ordinate transformation is introduced:

$$\tilde{z} = \frac{z - \frac{H}{2}}{H} \qquad \tilde{x} = \frac{x - \frac{AH}{2}t}{L_R} \qquad \tilde{t} = \frac{fA}{N}t \tag{A.31}$$

where $L_R = \frac{NH}{f}$ is the Rossby radius of deformation. The new non-dimensional variables $\tilde{\psi}'$, $\tilde{q}'$ and $\tilde{b}'$ are related to the dimensional variables ψ', q' and b' by:

$$\tilde{\psi}' = \frac{\psi'}{\psi_0} \qquad \tilde{q}' = \frac{L_R^2}{\psi_0}q' \qquad \tilde{b}' = \frac{H}{\psi_0 f}b' \tag{A.32}$$

where ψ_0 is the amplitude of ψ'.

The non-dimensional equations become:

$$\left(\frac{\partial}{\partial \tilde{t}} + \tilde{z}\frac{\partial}{\partial \tilde{x}}\right)\frac{\partial \tilde{\psi}'}{\partial \tilde{z}} = \frac{\partial \tilde{\psi}'}{\partial \tilde{x}} \qquad \text{on } z = \pm\frac{1}{2} \tag{A.33}$$

$$\left(\frac{\partial}{\partial \tilde{t}} + \tilde{z}\frac{\partial}{\partial \tilde{x}}\right)\tilde{q}' = 0 \qquad \text{in } z\epsilon(-\frac{1}{2},\frac{1}{2}) \tag{A.34}$$

$$\frac{\partial^2 \tilde{\psi}'}{\partial x^2} + \frac{\partial^2 \tilde{\psi}'}{\partial z^2} = \tilde{q}' \qquad \text{in } z\epsilon(-\frac{1}{2},\frac{1}{2}). \tag{A.35}$$

Although it is only the derivatives of $\tilde{\psi}'$ that are of interest, the derivatives are found in practice by first calculating $\tilde{\psi}'$. Therefore an extra equation is needed so that the problem for $\hat{\psi}'$ is well posed. We impose that

$$\int\!\!\int \tilde{\psi}' dx dz = 0 \tag{A.36}$$

so that the mean value of the streamfunction in the domain is zero.

For simplicity, the tildes and primes will be omitted in all further equations.

A.4 Modal and Non-Modal Initial Conditions

All the identical twin experiments in this thesis use a true state with initial conditions given by either the most rapidly growing or decaying normal mode or by a perturbation which leads to non-modal rapid finite-time growth. The equations for these states are now given.

The normal mode solutions can be found analytically by assuming that $q' = 0$ and substituting a solution of the form:

$$\psi(x,z,t) = \hat{\psi}(z)e^{-ik(x-ct)} \tag{A.37}$$

into the non-dimensional Eady model equations, where k is the non-dimensional wave number and c is the non-dimensional phase speed. It can be shown (e.g. Pedlosky, 1987) that when $k > 2.4$, then c is real and so the corresponding solutions form pairs of neutral modes; and when $k < 2.4$ then c is imaginary, and the solutions form pairs of stationary growing and decaying modes. It can also be shown that the maximum non-dimensional growth or decay rate $\sigma = \pm kc_i$ of 0.31 corresponds to a non-dimensional wavenumber $k = 1.6$.

The equations for the most rapidly growing and decaying normal modes are then given by:

$$\psi = \psi_0 e^{\sigma t}\left[\cosh(kz)\cos(kx) - \alpha\sinh(kz)\sin(kx)\right] \qquad (\text{Growing Mode}) \tag{A.38}$$

$$\psi = \psi_0 e^{-\sigma t}\left[\cosh(kz)\cos(kx) + \alpha\sinh(kz)\sin(kx)\right] \qquad (\text{Decaying Mode}) \tag{A.39}$$

where $k = 1.6$, $\sigma = 0.31$ and

$$\alpha = \sqrt{\frac{1 - \frac{k}{2}\tanh\frac{k}{2}}{\frac{k}{2}\coth\frac{k}{2} - 1}} \text{ and } \sigma = |kc_i| > 0. \tag{A.40}$$

The initial state for non-modal growth is defined by an interior QGPV dipole perturbation with zero values for the buoyancy on the boundaries. The QGPV dipole is defined by:

$$q(x, z) = sf(x)g(z) \tag{A.41}$$

where

$$f(x) = \begin{cases} -\frac{L}{4} - \frac{x}{2} + \frac{\sin(2kx)}{4k} & -\frac{L}{2} < x < -\frac{L}{4} \\ -\frac{L}{8} - \frac{\cos(kx)}{k} & -\frac{L}{4} < x < \frac{L}{4} \\ -\frac{L}{4} + \frac{x}{2} - \frac{\sin(2kx)}{4k} & \frac{L}{4} < x < \frac{L}{2} \end{cases} \tag{A.42}$$

with $f(x) = 0$ for $x < -\frac{L}{2}$ and $x > \frac{L}{2}$, $L = 1000km$, and $k = \frac{2\pi}{L}$,

$$g(z) = exp\left(-\frac{1}{D^2}(z - 3000)^2\right) - exp\left(-\frac{1}{D^2}(z - 7000)^2\right) \tag{A.43}$$

for $2 < z < 8km$, and $g(z) = 0$ otherwise, and with $D = 1500km$. The perturbation is scaled to an appropriate size using $s = 10^{-6}$.

A.5 Discrete Equations

The discretization of the Eady model equations is now described. This particular discretization has previously been used, for example, by Badger(1997,2001) and by Fletcher (1999).

The continuous non-dimensional Eady model equations (A.33 to A.35) can be summarized as the advection of the buoyancy on the boundaries, and QGPV in the interior:

$$\left(\frac{\partial}{\partial t} + z\frac{\partial}{\partial x}\right) q = 0 \qquad \text{in } z\epsilon \left[-\frac{1}{2}, \frac{1}{2}\right] \tag{A.44}$$

$$\left(\frac{\partial}{\partial t} + z\frac{\partial}{\partial x}\right) b = \frac{\partial \psi}{\partial x} \qquad \text{on } z = \pm\frac{1}{2} \tag{A.45}$$

where the QGPV q and buoyancy b are related to the streamfunction ψ, by:

$$q = \nabla^2 \psi \qquad \text{in } z\epsilon \left[-\frac{1}{2}, \frac{1}{2}\right] \tag{A.46}$$

$$b = \frac{\partial \psi}{\partial z} \qquad \text{on } z = \pm\,\frac{1}{2} \tag{A.47}$$

where $\nabla^2 = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial z^2}$.

These equations are discretized on a domain with 40 grid points in the horizontal and 11 vertical levels. The value of q and ψ at the ith horizontal grid point, jth vertical level at time-level t are denoted by $q^t_{i,j}$ and $\psi^t_{i,j}$, where the horizontal grid spacing is Δx, the vertical spacing is Δz and the time-step is Δt. $j = 11$ represents the upper boundary and $j = 1$ represents the lower boundary so that the value of the buoyancy at the ith grid point on the lower boundary is denoted by $b_{i,j}$ for $j = 1$, and similarly for the upper boundary, by $b_{i,j}$ for $j = 11$. The non-dimensional grid-spacing and time-steps are chosen as $\Delta x = 0.1$, $\Delta z = 0.1$ and $\Delta t = 0.1728$. These correspond to dimensional steps $\Delta x = 100km$, $\Delta z = 1km$ and $\Delta t = 4320s$ if the constants are given by $N = 10^{-2}s^{-1}$, $f = 10^{-4}s^{-1}$, $H = 10km$ and $A = 4 \times 10^{-3}s^{-1}$.

In the majority of experiments, the basic state flow is such that the zonal wind is zero in the centre of the domain. However, there are some experiments in Chapter 6 where the basic state flow is such that the zonal wind is zero on the lower boundary. This is achieved by adding 0.5 to each value of z. Then, to satisfy the CFL (Courant-Friedrichs-Lewy) condition (e.g. Durran, 1999), the time step is halved to $\Delta t = 0.0864$.

The model can be summarized as follows. Steps 2 and 3 are repeated for every time step.

Step 1 **Initial Conditions.** The initial conditions are given by the QGPV in the interior:

$$q^t_{i,j} \text{ for } i = 1, \ldots, 40,\ j = 1, \ldots, 11,\ t = 1 \tag{A.48}$$

and the buoyancy on the boundaries:

$$b^t_{i,j} \text{ for } i = 1, \ldots, 40,\ j = 1\&11,\ t = 1. \tag{A.49}$$

Step 2 **Calculate the streamfunction.** The elliptic equation (A.46) is discretized using a 5-point star stencil:

$$q^t_{i,j} = \frac{\psi^t_{i-1,j} - 2\psi^t_{i,j} + \psi^t_{i+1,j}}{\Delta x^2} + \frac{\psi^t_{i,j-1} - 2\psi^t_{i,j} + \psi^t_{i,j+1}}{\Delta z^2}. \tag{A.50}$$

This formula is repeated for $i = 1, \ldots, 11$ and $j = 1, \ldots, 40$ and is then written in matrix form.

From the periodic boundary conditions then $\psi^t_{i-1,j} = \psi^t_{40,j}$ when $i = 1$ and similarly, $\psi^t_{i+1,j} = \psi^t_{1,j}$ when $i = 40$. When $j = 1$, $\psi^t_{i,j-1}$ does not exist and when $j = 11$, $\psi^t_{i,j+1}$ does not exist. However, they can be found by using the Hydrostatic balance equation (A.47). If this is discretized by:

$$b^t_{i,j} = \frac{\psi^t_{i,j+1} - \psi^t_{i,j-1}}{2\Delta z} \tag{A.51}$$

then $\psi^t_{i,j-1} = \psi^t_{i,j+1} - 2\Delta z b^t_{i,j}$ for $j = 1$ and $\psi^t_{i,j+1} = \psi^t_{i,j-1} + 2\Delta z b^t_{i,j}$ for $j = 11$. The values of $b_{i,j}$ for $j = 1\&11$ are then added to the right hand side of the matrix. Some of the rows are then multiplied by $\frac{1}{2}$ to make the matrix self-adjoint.

To illustrate the matrix form clearly, we consider the more simple case of a smaller domain with $i = 1, \ldots, 3$ and $j = 1, \ldots, 3$. The matrix equation can be written as:

$$\left(\begin{array}{ccc|ccc|ccc}
\frac{1}{2}f & \frac{1}{2}d & \frac{1}{2}d & e & 0 & 0 & 0 & 0 & 0 \\
\frac{1}{2}d & \frac{1}{2}f & \frac{1}{2}d & 0 & e & 0 & 0 & 0 & 0 \\
\frac{1}{2}d & \frac{1}{2}d & \frac{1}{2}f & 0 & 0 & e & 0 & 0 & 0 \\
\hline
e & 0 & 0 & f & d & d & e & 0 & 0 \\
0 & e & 0 & d & f & d & 0 & e & 0 \\
0 & 0 & e & d & d & f & 0 & 0 & e \\
\hline
0 & 0 & 0 & e & 0 & 0 & \frac{1}{2}f & \frac{1}{2}d & \frac{1}{2}d \\
0 & 0 & 0 & 0 & e & 0 & \frac{1}{2}d & \frac{1}{2}f & \frac{1}{2}d \\
0 & 0 & 0 & 0 & 0 & e & \frac{1}{2}d & \frac{1}{2}d & \frac{1}{2}f
\end{array}\right)
\begin{pmatrix} \psi_{1,1} \\ \psi_{2,1} \\ \psi_{3,1} \\ \psi_{1,2} \\ \psi_{2,2} \\ \psi_{3,2} \\ \psi_{1,3} \\ \psi_{2,3} \\ \psi_{3,3} \end{pmatrix}
=
\begin{pmatrix}
\frac{1}{2}(-deq_{1,1} + e2\Delta z b_{1,1}) \\
\frac{1}{2}(-deq_{2,1} + e2\Delta z b_{2,1}) \\
\frac{1}{2}(-deq_{3,1} + e2\Delta z b_{3,1}) \\
-deq_{1,2} \\
-deq_{2,2} \\
-deq_{3,2} \\
\frac{1}{2}(-deq_{1,3} - e2\Delta z b^{1,3}) \\
\frac{1}{2}(-deq_{2,3} - e2\Delta z b^{2,3}) \\
\frac{1}{2}(-deq_{3,3} - e2\Delta z b^{3,3})
\end{pmatrix} \tag{A.52}$$

where $d = -\Delta z^2$, $e = -\Delta x^2$ and $f = 2(\Delta z^2 + \Delta x^2)$ and the horizontal and vertical lines are added for clarity. To impose the constraint that the mean value of the streamfunction is zero,

$$\iint \psi dx dz = 0, \tag{A.53}$$

then a small constant $(= 0.1)$ is added to every element of the matrix. Without this constant, the matrix would be singular and hence ill-posed.

Thus, the initial conditions are used to form a vector $\mathbf{r}^t$, which contains the values of $r^t_{i,j}$, at time level $t = t$:

$$r^t_{i,j} = \begin{cases} \alpha q^t_{i,j} + \beta b^t_{i,j} & \text{for } j = 1 \\ 2\alpha q^t_{i,j} & \text{for } j = 2, \ldots, 10 \\ \alpha q^t_{i,j} - \beta b^t_{i,j} & \text{for } j = 11 \end{cases} \tag{A.54}$$

for $i = 1, \ldots, 40$, where $\alpha = -\frac{1}{2}\Delta z^2 \Delta x^2$ and $\beta = -\Delta x^2 \Delta z$ are constant scalars. The streamfunction variables $\psi_{i,j}^t$ are then found by solving the matrix equation:

$$\mathbf{A}\boldsymbol{\psi}^t = \mathbf{r}^t \tag{A.55}$$

where $\mathbf{A}$ is a square matrix containing the coefficients corresponding to a five point differencing scheme, and $\boldsymbol{\psi}^t$ is a vector containing the values of $\psi_{i,j}^t$ for $i = 1, \ldots, 40$, $j = 1, \ldots, 11$ and $t = t$. This matrix equation is solved using an LU factorization using the NAG routine nag_gen_lin_sys (NAG).

Step 3 **Advect the QGPV and buoyancy.** The interior QGPV, q^t and buoyancy on the upper and lower boundaries, b^t at time t are advected to the next time step, $t + 1$ using the Leapfrog (Centred Time, Centred Space) discretization (with the Forward Time, Centred Space scheme for the first time step). For the interior QGPV,

$$q_{i,j}^{t+1} = q_{i,j}^t - \frac{c_j}{2}(q_{i+1,j}^t - q_{i-1,j}^t) \qquad \text{for } t = 1 \tag{A.56}$$

$$q_{i,j}^{t+1} = q_{i,j}^{t-1} - c_j(q_{i+1,j}^t - q_{i-1,j}^t) \qquad \text{for } t = 2, \ldots, T-1 \tag{A.57}$$

where $c_j = z_j \frac{\Delta t}{\Delta x}$, for $j = 1, \ldots, 11$ and $i = 1, \ldots, 40$, again using periodic boundary conditions. For the buoyancy on the boundaries,

$$b_{i,j}^{t+1} = b_{i,j}^t - \frac{c_j}{2}(b_{i+1,j}^t - b_{i-1,j}^t) + \frac{\Delta t}{2\Delta x}(\psi_{i+1,j}^t - \psi_{i-1,j}^t) \qquad \text{for } t = 1 \tag{A.58}$$

$$b_{i,j}^{t+1} = b_{i,j}^{t-1} - c_j(b_{i+1,j}^t - b_{i-1,j}^t) + \frac{\Delta t}{\Delta x}(\psi_{i+1,j}^t - \psi_{i-1,j}^t) \qquad \text{for } t = 2, \ldots, T-1 \tag{A.59}$$

for $j = 1\&11$ and $i = 1, \ldots, 40$, using periodic boundary conditions.

A.6 Observing Interior Buoyancy

The observation operator $\mathbf{H}$ for observations $\mathbf{y}$ of the interior buoyancy field is now described. The control variables $\mathbf{x}$ are defined as the interior QGPV and buoyancy on the upper and lower boundaries, and hence the interior buoyancy is not a control variable. The 4D-Var experiments in Chapter 3 consider observations of the lower boundary buoyancy, and as this is a control variable, the observation operator is simply a matrix of ones and zeros (3.3). The 4D-Var experiments in Chapter 6 consider observations of the interior buoyancy and as this is not a control variable, the observation operator must contain dynamcial equations to link the observed variables to the control varibles. In practice, the observation operator is no longer a matrix, but a sequence of linear operations which can be summarized as:

$$\mathbf{H}: \ \mathbf{x} \xrightarrow{1} \boldsymbol{\psi} \xrightarrow{2} \mathbf{b} \xrightarrow{3} \mathbf{y}. \tag{A.60}$$

The first operator uses the QGPV and buoyancy in the control vector $\mathbf{x}$ to calculate the corresponding streamfunction field $\boldsymbol{\psi}$ using (A.55). The second operator uses the streamfunction field $\boldsymbol{\psi}$ to calculate the interior buoyancy $\mathbf{b}$ using the Hydrostatic balance relation (A.17). This is discretized as:

$$b_{i,j+\frac{1}{2}} = \frac{\psi_{i,j+1} - \psi_{i,j}}{\Delta z} \text{ for } j = 0, \ldots, 9, \ i = 1, \ldots, 40 \tag{A.61}$$

so that the buoyancy field then contains the values of buoyancy at heights $0.5, 1.5, \ldots, 9.5$ km and also at 0 and 10 km from the buoyancy on the boundaries. The third operator applies a matrix of ones and zeros to select the individual observation locations, for example to give horizontal or vertical lines, to give the vector of observations $\mathbf{y}$.

The adjoint $\mathbf{H}^T$ is simply given by the adjoints of each operator and is in reverse order:

$$\mathbf{H}^T: \ \mathbf{y} \xrightarrow{3^T} \mathbf{b} \xrightarrow{2^T} \boldsymbol{\psi} \xrightarrow{1^T} \mathbf{x}. \tag{A.62}$$

A.7 Calculating the SVD of the 4D-Var Observability Matrix

The experiments in Chapters 5 and 6 use the singular value decomposition of the 4D-Var observability matrix:

$$\hat{\mathbf{H}} = \begin{bmatrix} \mathbf{H} \\ \\ \mathbf{H}\mathbf{M}(t_N, t_0) \end{bmatrix}. \tag{A.63}$$

This is found using the SVD algorithm nag_gen_svd (NAG, Golub and Van Loan, 1996), which requires that $\hat{\mathbf{H}}$ is in matrix form.

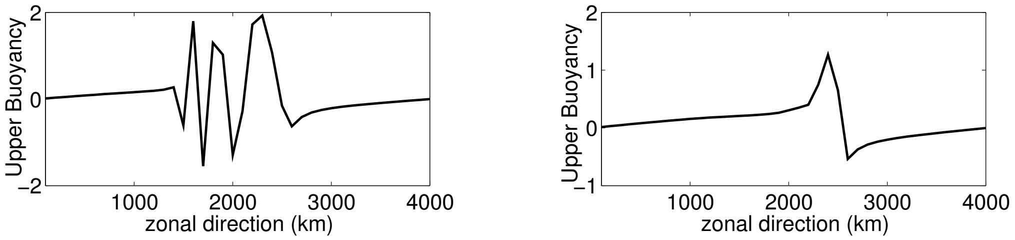
The matrix form $\mathbf{M}$ of the discrete Eady model $\mathcal{M}$ is found by applying the discrete Eady model equations to successive columns of the identity matrix:

$$\mathbf{M} = \mathbf{M}\mathbf{I} = \mathcal{M}(\mathbf{e}_1)\mathcal{M}(\mathbf{e}_2)\cdots\mathcal{M}(\mathbf{e}_{520}) \tag{A.64}$$

where the vectors $\mathbf{e}_i$ are the columns of the identity matrix:

$$\mathbf{e}_1 = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \qquad \mathbf{e}_2 = \begin{pmatrix} 0 \\ 1 \\ \vdots \\ 0 \end{pmatrix} \qquad \cdots \qquad \mathbf{e}_{520} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{pmatrix}. \tag{A.65}$$

The linear operator $\mathbf{M}$ is applied to discontinuous field, and therefore it is important to consider the choice of the numerical advection scheme. The Leapfrog scheme (A.59) is used for the discrete Eady model and although it is suitable for smoothly varying functions such as the Eady wave, it is not suitable for propagating sharp discontinuities. This is illustrated in Fig. A.2(a) where the Eady model with the Leapfrog scheme has been applied to the initial conditions given by a spike in the buoyancy field at the centre of the domain. The upper level spike is advected eastwards by the flow but there is a trail of short waves upstream of the perturbation. This is due to wavenumber dependent phase speed errors (dispersion) and the computational mode.



(a) Leapfrog

(b) Lax-Wendroff

Figure A.2: *Comparison of the upper level buoyancy fields at time T+6 (5 timesteps) with (a) Leapfrog (b) Lax-Wendroff numerical advection schemes in the Eady model. The initial conditions are given by a spike in both the upper and lower buoyancy fields at* $x = 2000km$ *with an amplitude of* 4 *on the upper boundary and* -4 *on the lower boundary.* $c = z\frac{\Delta t}{\Delta z} = 0.864$.

An alternative numerical scheme is the Lax-Wendroff scheme. This scheme is derived from a Taylor series expansion and can be written for the buoyancy advection equation as:

$$b_{i,j}^{t+1} = b_{i,j}^{t} - \frac{c_j}{2}(b_{i+1,j} - b_{i-1,j}^{t}) + \frac{c_j^2}{2}(b_{i+1,j}^{t} - b_{i,j}^{t} + b_{i-1,j}^{t}) + \frac{\Delta t}{2\Delta x}(\psi_{i+1,j}^{t} - \psi_{i-1,j}^{t}) \quad \text{(A.66)}$$

The modified equation approach (e.g. Le Veque, 1992, Durran, 1999), can be used to show that the Lax-Wendroff scheme is both dispersive and diffusive and this means that the short wavelengths are damped. Further, as the Lax-Wendroff scheme is a two-time level scheme, it does not suffer from a computational mode. For these reasons, the Lax-Wendroff scheme is more successful than the Leapfrog scheme in advecting a spike. This is shown in Fig. A.2(b).

The matrix form $\mathbf{M}$ of the Eady model is tested by integrating the models for 24 hours with the initial state $\mathbf{x}_0$ given by the most unstable Eady wave.

For the Leapfrog discretization,

$$||\mathbf{M}\mathbf{x} - \mathcal{M}(\mathbf{x})||_2 = 2.5 \times 10^{-13} \quad \text{(A.67)}$$

and for the Lax-Wendroff discretization,

$$||\mathbf{M}\mathbf{x} - \mathcal{M}(\mathbf{x})||_2 = 4.3 \times 10^{-14}. \quad \text{(A.68)}$$

Therefore, the error is slightly smaller is the Lax-Wendroff discretization is used. The Lax-Wendroff scheme is used for all the SVD computations in this thesis.

Appendix B

Adjoint Model

This appendix describes the adjoint model equations in both continuous and discrete form. The adjoint of the continuous equations can be found using Lagrange mulitpliers whilst the adjoint of the discrete equations can be found by considering the model as a sequence of linear operations. It is particularly useful to consider the adjoint of the continous equations to be able to understand the discrete adjoint equations.

B.1 Continuous Equations

The forward Eady model continuous equations can be summarized as:

$$q_t + zq_x = 0 \qquad \nabla^2\psi = q \qquad \text{in } z\epsilon\left[-\frac{1}{2}, \frac{1}{2}\right] \tag{B.1}$$

$$b_t + zb_x = \psi_x \qquad \psi_z = b \qquad \text{on } z = \pm\frac{1}{2} \tag{B.2}$$

and with periodic boundary conditions in the horizontal where $q = q(x, z, t)$ is the QGPV, $b = b(x, z, t)$ is the buoyancy, and $\psi = \psi(x, z, t)$ is the streamfunction, and using the notation $q_t = \frac{\partial}{\partial t}$, and $\nabla^2 = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial z^2}$.

The adjoint of the continuous equations can be found using the calculus of variations, following the work by, for example, Gelfand and Fomin (1963), Forray (1968), Birkett and Nichols (1983), Xu and Nichols (1991) and Griffith and Nichols (1994). The forward model contains both differential equation constraints and algebraic constraints. However, all the equations are linked together and must therefore all be considered simultaneously in the Lagrangian functional. Further, some equations are defined over the interior whilst the others are defined over the boundaries and so the Lagrangian functional must be defined accordingly. The Lagrangian functional $\mathcal{L}$ may be defined as:

$$\mathcal{L} = \int\limits_0^T\iint\limits_\Omega \hat{q}(q_t + zq_x)dxdzdt + \int\limits_0^T\oint\limits_\Gamma \hat{b}(b_t + zb_x - \psi_x)dxdt \tag{B.3}$$

$$+ \int\limits_0^T\iint\limits_\Omega \hat{\psi}(\nabla^2\psi - q)dxdzdt + \int\limits_0^T\oint\limits_\Gamma \hat{\psi}(\psi_z - b)dxdt \tag{B.4}$$

where $\hat{q} = \hat{q}(x,z,t)$, $\hat{b} = \hat{b}(x,z,t)$, $\hat{\psi} = \hat{\psi}(x,z,t)$ are Lagrange multipliers, Ω is the rectangular domain, Γ is the boundary surrounding the domain and T is the assimilation window length.

The Lagrange multiplier $\hat{\psi}$ is used for both the elliptic equation and the derivative boundary condition (last two terms) as the elliptic equation is defined over the interior of the domain whilst the derivative equation is defined on the boundaries. That is, in the third term, $\hat{\psi}$ is defined over the interior, whilst in the fourth term $\hat{\psi}$ is defined on the boundaries. The constraint that the mean of ψ is zero is omitted from the Lagrangian functional as it is only the derivatives of ψ that are required, and hence the forward equations are well-posed without this constraint.

The first variation of $\mathcal{L}$ can be written as a function of δq, δb, and $\delta\psi$ using Greens Theorem in the Plane (or integration by parts). Then, from the Fundamental Lemma of the calculus of variations, it can be shown that the continuous adjoint Eady model equations are given by:

$$\hat{q}_\tau - z\hat{q}_x = +\hat{\psi} \qquad \nabla^2\hat{\psi} = 0 \qquad \text{in } z\epsilon\left[-\frac{1}{2},\frac{1}{2}\right] \tag{B.5}$$

$$\hat{b}_\tau - z\hat{b}_x = +\hat{\psi} \qquad \hat{\psi}_z = -\hat{b}_x \qquad \text{on } z = -\frac{1}{2} \tag{B.6}$$

$$\hat{b}_\tau - z\hat{b}_x = -\hat{\psi} \qquad \hat{\psi}_z = +\hat{b}_x \qquad \text{on } z = +\frac{1}{2} \tag{B.7}$$

with periodic boundary conditions in the horizontal. The time co-ordinate τ has been introduced, such that $\hat{q}_\tau = -\hat{q}_t$ so that the adjoint equations are propagated backwards in time. The derivative boundary conditions imply that the streamfunction ψ and adjoint streamfunction $\hat{\psi}$ are only unique up to an additive constant. In the forward equations, we added the constraint that the mean of ψ was zero, and by analogy with the derivation of the adjoint equations for the discrete model (Section B.2), we also apply a similar constraint in the adjoint equations so that the mean of $\hat{\psi}$ is zero:

$$\iint\limits_{\Omega} \psi dxdz = 0 \qquad \iint\limits_{\Omega} \hat{\psi} dxdz = 0. \tag{B.8}$$

From the transversality conditions, the final state at the end of the assimilation window is zero:

$$\hat{q}(x,z,T) = 0 \qquad \text{in } z\epsilon \left[-\frac{1}{2}, \frac{1}{2}\right] \tag{B.9}$$

$$\hat{b}(x,z,T) = 0 \qquad \text{on } z = \pm\frac{1}{2} \tag{B.10}$$

and the gradients of $\mathcal{L}$ with respect to the forward variables at the beginning of the window are defined as:

$$\frac{\partial \mathcal{L}}{\partial q(x,z,0)} = -\hat{q}(x,z,0) \qquad \text{in } z\epsilon \left[-\frac{1}{2}, \frac{1}{2}\right] \tag{B.11}$$

$$\frac{\partial \mathcal{L}}{\partial b(x,z,0)} = -\hat{b}(x,z,0) \qquad \text{on } z = -\frac{1}{2} \tag{B.12}$$

$$\frac{\partial \mathcal{L}}{\partial b(x,z,0)} = -\hat{b}(x,z,0) \qquad \text{on } z = +\frac{1}{2}. \tag{B.13}$$

B.2 Discrete Equations

The discrete equations for the Eady model can be summarized as.

Step 1 **Initial Conditions.** The initial conditions are given by the QGPV in the interior:

$$q^t_{i,j} \text{ for } i = 1, \ldots, 40,\ j = 1, \ldots, 11,\ t = 1 \tag{B.14}$$

and the buoyancy on the boundaries:

$$b^t_{i,j} \text{ for } i = 1, \ldots, 40,\ j = 1\&11,\ t = 1, \tag{B.15}$$

where i is the horizontal grid point, j is the vertical level and t is the time level, with horizontal grid spacing Δx, vertical level spacing Δz and time spacing Δt.

Step 2 **Calculate the streamfunction.** The initial conditions are used to form a vector $\mathbf{r}^t$, which contains the values of $r^t_{i,j}$, at time level $t = t$:

$$r^t_{i,j} = \begin{cases} \alpha q^t_{i,j} + \beta b^t_{i,j} & \text{for } j = 1 \\ 2\alpha q^t_{i,j} & \text{for } j = 2, \ldots, 10 \\ \alpha q^t_{i,j} - \beta b^t_{i,j} & \text{for } j = 11 \end{cases} \tag{B.16}$$

for $i = 1, \ldots, 40$, where α and β are constant scalars. The streamfunction variables $\psi^t_{i,j}$ are then found by solving the matrix equation:

$$\mathbf{A}\boldsymbol{\psi}^t = \mathbf{r}^t \tag{B.17}$$

where $\mathbf{A}$ is a square matrix containing the coefficients corresponding to a five point differencing scheme, and $\boldsymbol{\psi}^t$ is a vector containing the values of $\psi^t_{i,j}$ for $i = 1, \ldots, 40$, $j = 1, \ldots, 11$ and $t = t$.

Step 3 **Advect the QGPV and buoyancy.** The interior QGPV, q^t and buoyancy on the upper and lower boundaries, b^t at time t are advected to the next time step, $t + 1$ using the Leapfrog (Centred Time, Centred Space) discretization (with the Forward Time, Centred Space scheme for the first time step). For the interior QGPV,

$$q^{t+1}_{i,j} = q^t_{i,j} - \frac{c_j}{2}(q^t_{i+1,j} - q^t_{i-1,j}) \qquad \text{for } t = 1 \tag{B.18}$$

$$q^{t+1}_{i,j} = q^{t-1}_{i,j} - c_j(q^t_{i+1,j} - q^t_{i-1,j}) \qquad \text{for } t = 2, \ldots, T-1 \tag{B.19}$$

where $c_j = z_j \frac{\Delta t}{\Delta x}$, for $j = 1, \ldots, 11$ and $i = 1, \ldots, 40$, again using periodic boundary conditions. For the buoyancy on the boundaries,

$$b^{t+1}_{i,j} = b^t_{i,j} - \frac{c_j}{2}(b^t_{i+1,j} - b^t_{i-1,j}) + \frac{\Delta t}{2\Delta x}(\psi^t_{i+1,j} - \psi^t_{i-1,j}) \quad \text{for } t = 1 \tag{B.20}$$

$$b^{t+1}_{i,j} = b^{t-1}_{i,j} - c_j(b^t_{i+1,j} - b^t_{i-1,j}) + \frac{\Delta t}{\Delta x}(\psi^t_{i+1,j} - \psi^t_{i-1,j}) \quad \text{for } t = 2, \ldots, T-1 \tag{B.21}$$

for $j = 1\&11$ and $i = 1, \ldots, 40$, using periodic boundary conditions.

The adjoint of the discrete equations may be found by considering the linear model in matrix form. For example, consider the discrete linear model in matrix form, $\mathbf{M}$ that is integrated over one time step: $\mathbf{x}_{t+1} = \mathbf{M}\mathbf{x}_t$. The adjoint of the matrix $\mathbf{M}$ is simply the complex conjugate transpose, and as all the variables are real in the following cases, this is simply the transpose of the matrix. This may be further simplified by considering the model as a sequence of linear operations $\mathbf{L}_n, \ldots, \mathbf{L}_1$:

$$\mathbf{M} = \mathbf{L}_n \ldots \mathbf{L}_2 \mathbf{L}_1. \tag{B.22}$$

From the definition of the transpose of a matrix, the adjoint model is then:

$$\mathbf{M}^T = \mathbf{L}_1^T \mathbf{L}_2^T \ldots \mathbf{L}_n^T. \tag{B.23}$$

Thus, the adjoint model may be considered as the exact reversal of the sequence of adjoint operations. Such an approach is also described by Chao and Chang (1992), Navon et al. (1992), Rosmond (1997) and Marotzke et al. (1999), where adjoints of oceanic and atmospheric models are derived and also by Giering and Kaminski (1996). This approach has been used to develop the adjoint model code for the Eady model. It is also possible to consider the following equations as the discretization of the continuous equations that were given in the previous section.

The adjoint model is integrated backwards in time, and starts with the final conditions $\hat{q}^T = 0$ and $\hat{b}^T = 0$ (and $\hat{q}^{T+1} = 0$ and $\hat{b}^{T+1} = 0$). These are then advected backwards in time to the the previous time step, using the adjoint advection equations (step 3). Then the adjoint of the streamfunction calculation is used to find the right hand side forcing for the advection equations (step 2). Steps 3 and 2 are then repeated for every time step.

Step 3 **Adjoint of the advection equations.** For $t = T, \ldots, 3$, $i = 1, \ldots, 40$ with periodic boundary conditions:

$$\hat{q}_{i,j}^{t-1} = \hat{q}_{i,j}^{t+1} + c_j(\hat{q}_{i+1,j}^t - \hat{q}_{i-1,j}^t) \qquad \text{for } j = 1, \ldots, 11 \tag{B.24}$$

$$\hat{b}_{i,j}^{t-1} = \hat{b}_{i,j}^{t+1} + c_j(\hat{b}_{i+1,j}^t - \hat{b}_{i-1,j}^t) \qquad \text{for } j = 1\&11 \tag{B.25}$$

$$\hat{\psi}_{i,j}^{t-1} = -\frac{\Delta t}{\Delta x}(\hat{b}_{i+1,j}^t - \hat{b}_{i-1,j}^t) \qquad \text{for } j = 1\&11, \tag{B.26}$$

and for $t = 2$, $i = 1, \ldots, 40$ with periodic boundary conditions:

$$\hat{q}_{i,j}^{t-1} = \hat{q}_{i,j}^{t} + \frac{c_j}{2}(\hat{q}_{i+1,j}^{t} - \hat{q}_{i-1,j}^{t}) \qquad \text{for } j = 1 \ldots 11 \qquad \text{(B.27)}$$

$$\hat{b}_{i,j}^{t-1} = \hat{b}_{i,j}^{t} + \frac{c_j}{2}(\hat{b}_{i+1,j}^{t} - \hat{b}_{i-1,j}^{t}) \qquad \text{for } j = 1\&11 \qquad \text{(B.28)}$$

$$\hat{\psi}_{i,j}^{t-1} = -2\frac{\Delta t}{\Delta x}(\hat{b}_{i+1,j}^{t} - \hat{b}_{i-1,j}^{t}) \qquad \text{for } j = 1\&11. \qquad \text{(B.29)}$$

Step 2 **Adjoint of the streamfunction calculation.** The $\hat{\psi}^{t-1}$ field is used to find the values of $\hat{r}_{i,j}^{t-1}$ by solving the matrix equation:

$$\mathbf{A}\hat{\mathbf{r}}^{t-1} = \hat{\boldsymbol{\psi}}^{t-1}. \qquad \text{(B.30)}$$

Note that this is the same equation as for the forward model. The forward matrix equation is solved using an LU decomposition. However, it is not necessary to find the adjoint of the LU decomposition; instead the adjoint equation $\mathbf{A}^T\hat{\mathbf{r}} = \hat{\boldsymbol{\psi}}$ can be solved and in this case, $\mathbf{A}$ is self-adjoint. Thus, the same LU decomposition that is used to solve the matrix equation in the forward model can be used to solve the matrix equations in the adjoint model. The $\hat{r}^{t-1}$ variables are then used to update the q^{t-1} and b^{t-1} variables using the assignment statements:

$$(\hat{q}_{i,j}^{t-1})^{l} = (\hat{q}_{i,j}^{t-1})^{l-1} + 2\alpha(\hat{r}_{i,j}^{t-1})^{l-1} \qquad \text{for } j = 2, \ldots, 10 \qquad \text{(B.31)}$$

$$(\hat{q}_{i,j}^{t-1})^{l} = (\hat{q}_{i,j}^{t-1})^{l-1} + \alpha(\hat{r}_{i,j}^{t-1})^{l-1} \qquad \text{for } j = 1\&11 \qquad \text{(B.32)}$$

$$(\hat{b}_{i,j}^{t-1})^{l} = (\hat{b}_{i,j}^{t-1})^{l-1} + \beta(\hat{r}_{i,j}^{t-1})^{l-1} \qquad \text{for } j = 1 \qquad \text{(B.33)}$$

$$(\hat{b}_{i,j}^{t-1})^{l} = (\hat{b}_{i,j}^{t-1})^{l-1} - \beta(\hat{r}_{i,j}^{t-1})^{l-1} \qquad \text{for } j = 11. \qquad \text{(B.34)}$$

where the indices l and $l-1$ denotes the vales of the variables just before and after the execution of the assignment.

From the structure of the $\mathbf{A}$ matrix (A.52) and from equations (B.26) and (B.29), then it is clear that the boundary conditions for $\hat{r}$ are given by

$$\frac{\partial \hat{r}}{\partial z} = +\frac{\partial b}{\partial x} \qquad \text{on } z = +\frac{1}{2} \qquad \text{(B.35)}$$

$$\frac{\partial \hat{r}}{\partial z} = -\frac{\partial b}{\partial x} \qquad \text{on } z = -\frac{1}{2}. \qquad \text{(B.36)}$$

Also, equations (B.31 to B.33) can be interpreted as providing the right hand side forcing for the advection equations. Thus, the discrete model is consistent with the continuous adjoint equations. Note, however, that the $\hat{\psi}$ variables in the continuous equations are equivalent to the $\hat{r}$ variables in the discrete equations.

Glossary of Symbols and Acronyms

4D-Var Notation

n	Dimension of the state vector
m	Dimension of the (generalized) observation vector
t_0	Initial time
t_N	Assimilation window length
$\mathbf{x}_i^t$	True state at time t_i
$\mathbf{x}^b$	Background state
$\mathbf{y}_i$	Observations at time t_i
$\mathbf{x}^a$	Analysis at time t_0
$\mathbf{d}_i$	Innovation vector at time t_i
$\mathcal{M}$	Linear forward model operator in equation form
$\mathbf{M}$	Linear forward model in matrix form
$\mathbf{M}^T$	Adjoint model in matrix form
$\mathbf{H}_i$	Observation operator at time t_i
$\mathbf{K}$	Kalman Gain matrix
$\mathbf{B}$	Specified background error covariance matrix
$\mathbf{R}_i$	Specified observation error covariance matrix at time t_i
$\boldsymbol{\rho}_B$	Specified background error correlation matrix
$\boldsymbol{\rho}_R$	Specified observation error correlation matrix
l	Horizontal correlation length scale
$\boldsymbol{\varepsilon}^b$	Background state errors
$\boldsymbol{\varepsilon}^o$	Observation errors
σ^2	Variance of the observational noise
σ_o^2	Specified observation error variance
σ_b^2	Specified background state error variance
μ^2	Ratio of the background and observation error variances
J	Cost function

J^b	Background cost function term
J^o	Observation cost function term
$\nabla_{\mathbf{x}_{t_i}} J$	Gradient of the cost function with respect to $\mathbf{x}_{t_i}$
$T+0$	Beginning of the assimilation window
$T+6$	6 hours into the assimilation window
t_I	Time of the initial observations

Singular Vector Decomposition Notation

$\mathbf{u}_j$	Left singular vector (LSV)
$\mathbf{v}_j$	Right singular vector (RSV)
λ_j	Singular value
j	Singular vector index
$\mathbf{U}$	Orthonormal matrix with columns given by the LSVs
$\mathbf{V}$	Orthonormal matrix with columns given by the RSVs
$\boldsymbol{\Lambda}$	Diagonal matrix with diagonal entries given by the singular values
$\hat{\mathbf{H}}$	4D-Var observability matrix
$\hat{\mathbf{y}}$	4D-Var generalized observation vector
$\hat{\mathbf{d}}$	4D-Var generalized innovation vector
$\boldsymbol{\chi}$	Pre-conditioned control variable
$\mathbf{C}$	Initial time norm
$\mathbf{E}$	Final time norm

Eady Model Notation

x	Horizontal distance in the zonal direction
y	Horizontal distance in the meridional direction
z	Height
q	Quasi-Geostrophic Potential Vorticity (QGPV)
f	Coriolis parameter
g	Gravitational acceleration
θ	Potential temperature
ψ	Geostrophic streamfunction

N^2	Static stability
$b^t_{i,j}$	Buoyancy at horizontal grid point i, vertical level j and time level t
u	Zonal wind
v	Meridional wind
$\mathbf{x}_q$	QGPV variables of the control vector
$\mathbf{x}_T$	Buoyancy variables of the control vector
σ^2_q	Specified background error variance for the QGPV
σ^2_T	Specified background error variance for the Buoyancy
σ_{KE}	Kinetic energy growth rate

Acronyms

DA	Data Assimilation
NWP	Numerical Weather Prediction
4D-Var	Four-Dimensional Variational Data Assimilation
3D-Var	Three-Dimensional Variational Data Assimilation
FGAT	First Guess at the Appropriate Time
PDF	Probability Distribution Function
RRKF	Reduced Rank Kalman Filter
QGPV	Quasi-Geostrophic Potential Vorticity
SVD	Singular Value Decomposition
RSV	Right Singular Vector
LSV	Left Singular Vector
GCV	Generalized Cross Validation

References

Aleksandrov, P. S., 1976. The principal mathematical discoveries of A. N. Tikhonov. *Russian Mathematical Surveys*, **31**:13–15.

Alexander, G., J. Weinman, and J. Schols, 1998. Use of digital warping of microwave integrated water vapor imagery to improve forecasts of marine extratropical cyclones. *Mon. Weather Rev.*, **126**:1469–1496.

Andersson, E., A. Hollingsworth, G. Kelly, P. Lönnberg, J. Pailleux, and Z. Zhang, 1991. Global observing system experiments on operational statistical retrievals of satellite sounding data. *Mon. Weather Rev.*, **119**:1851–1864.

Atkinson, K. E., 1989. *An Introduction to Numerical Analysis*. J. Wiley and Sons, second edition, pp. 693.

Badger, J., 1997. *Mechanisms for Rapid Synoptic Development*. PhD thesis, Department of Meteorology, University of Reading.

Badger, J. and B. J. Hoskins, 2001. Simple initial value problems and mechanisms for baroclinic growth. *J. Atmos. Sci.*, **58**:38–49.

Barkmeijer, J., M. V. Gijzen, and F. Bouttier, 1998. Singular vectors and estimates of the analysis-error covariance metric. *Q. J. R. Meteorol. Soc.*, **124**:1695–1713.

Beale, E. M. L., 1972. A derivation of conjugate gradients. In Lootsma, F. A., editor, *Numerical Methods for Nonlinear Optimization*, chapter 4, pp. 39–43. Academic Press.

Beale, E. M. L., 1988. Multi-dimensional optimization. In *Introduction to Optimization*, chapter 4, pp. 25–36. Wiley.

Beare, R. J., A. J. Thorpe, and A. A. White, 2003. The predictability of extratropical cyclones: Nonlinear sensitivity to localized potential-vorticity perturbations. *Q. J. R. Meteorol. Soc.*, **129**:219–237.

Beck, M. A., 2003. *Data Assimilation and Covariance Dynamics in Atmospheric Models*. PhD thesis, University of Vienna.

Belge, M., M. E. Kilmer, and E. L. Miller, 2002. Efficient determination of multiple regularization parameters in a generalized L-curve framework. *Inverse Problems.*, **18**:1161–1183.

Benamou, J. D., Y. Brenier, and K. Guittet, 2002. The Monge-Kantorovitch mass transfer and its computational fluid mechanics formulation. *International Journal for Numerical Methods in Fluids*, **40**:21–30.

Bennett, A. F., 2002. *Inverse Modelling of the Ocean and Atmosphere*. Cambridge University Press, pp. 234.

Bennett, A. F. and R. N. Miller, 1991. Weighting initial conditions in variational assimilation schemes. *Mon. Weather Rev.*, **119**:1098–1102.

Bergthórsson, P. and B. Döös, 1955. Numerical weather map analysis. *Tellus*, **7**:329–340.

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

Birkett, N. R. C., 1986. Non-linear optimal control of tidal power schemes in long estuaries. Numerical Analysis Report 9/86, Department of Mathematics, University of Reading.

Birkett, N. R. C. and N. K. Nichols, 1983. Optimal control problems in tidal power generation. Numerical Analysis Report 8/83, Department of Mathematics, University of Reading.

Bouttier, F., 2001. The development of 12-hourly 4D-Var. Tech. Memo. 348, ECMWF.

Bouttier, F. and P. Courtier, 2003. Data assimilation concepts and methods. Meteorological training course lecture series, ECMWF.
URL `http://www.ecmwf.int/`.

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

Bretherton, F. P., 1966. Critical layer instability in baroclinic flows. *Q. J. R. Meteorol. Soc.*, **92**:325–345.

Brewster, K. A., 2002. Phase-correcting data assimilation and application to storm-scale numerical weather prediction. Part I: Method description and simulation testing. *Mon. Weather Rev.*, **131**:480–492.

Brewster, K. A., 2002. Phase-correcting data assimilation and application to storm-scale numerical weather prediction. Part II: Application to a severe storm outbreak. *Mon. Weather Rev.*, **131**:493–507.

Brzovic, N., 2003. *Atmospheric Data Assimilation in the Tropics*. PhD thesis, Stockholm University. Thesis for the degree of Filosofie Licentiat.

Buizza, R., 1997. *The Singular Vector Approach to the Analysis of Perturbation Growth in the Atmosphere*. PhD thesis, University of London.

Buizza, R., J. Barkmeijer, T. N. Palmer, and D. S. Richardson, 2000. Current status and future developments of the ECMWF ensemble prediction system. *Meteorol. Appl.*, **7**:163–175.

Buizza, R. and T. Palmer, 1995. The singular-vector structure of the atmospheric global circulation. *J. Atmos. Sci.*, **52**:1434–1456.

Carlson, T. N., 1994. *Mid-Latitude Weather Systems*. Routledge, pp. 507.

Carroll, E. B., 1997. A technique for consistent alterations of NWP output fields. *Meteorol. Appl.*, **4**:171–178.

Chao, W. and L. Chang, 1992. Development of a four-dimensional variational analysis system using the adjoint method at GLA. part 1: Dynamics. *Mon. Weather Rev.*, **120**:1661–1673.

Charney, J. G., 1947. The dynamics of long waves in a baroclinic westerly current. *Journal of Meteorology*, **4**:135–162.

Charney, J. G. and M. E. Stern, 1962. On the stability of internal baroclinic jets in a rotating atmosphere. *J. Atmos. Sci.*, **19**:159–172.

Collard, A. D., 2000. Assimilation of IASI and AIRS data: Information content and quality control. In *Exploitation of the New Generation of Satellite Instruments for Numerical Weather Forecasts*, Seminar Proceedings, pp. 201–224, Reading, UK. ECMWF.

Courtier, P., E. Anderson, W. Heckley, J. Pailleux, D. Vasiljevic, M. Hamrud, A. Hollingsworth, F. Rabier, and M. Fisher, 1998. The ECMWF implementation of three-dimensional variational assimilation (3D-Var). I: Formulation. *Q. J. R. Meteorol. Soc.*, **124**:1783–1807.

Courtier, P. and O. Talagrand, 1987. Variational assimilation of meteorological observations with the adjoint vorticity equation (ii): Numerical results. *Q. J. R. Meteorol. Soc.*, **113**: 1329–1368.

Courtier, P., J.-N. Thépaut, and A. Hollingsworth, 1994. A strategy for operational implementation of 4D-Var, using an incremental approach. *Q. J. R. Meteorol. Soc.*, **120**:1367–1387.

Daley, R., 1985. The analysis of synoptic scale divergence by a statistical interpolation procedure. *Mon. Weather Rev.*, **113**:1066–1079.

Daley, R., 1991. *Atmospheric Data Analysis*. Cambridge University Press, pp. 457.

Davies, H. C. and C. H. Bishop, 1994. Eady edge waves and rapid development. *J. Atmos. Sci.*, **51**:1930–1946.

Dee, D. P., 1995. On-line estimation of error covariance parameters for atmospheric data assimilation. *Mon. Weather Rev.*, **123**:1128–1145.

Derber, J. and F. Bouttier, 1999. A reformulation of the background error covariance in the ECMWF global data assimilation system. *Tellus*, **51A**:195–221.

Derber, J. and A. Rosati, 1989. A global oceanic data assimilation system. *J. Phys. Oceanogr.*, **19**:1333–1347.

Desroziers, G. and S. Ivanov, 2001. Diagnosis and adaptive tuning of observation-error parameters in a variational assimilation. *Q. J. R. Meteorol. Soc.*, **127**:1433–1452.

Desroziers, G., B. Pouponneau, J. Thepaut, M. Janiskova, and F. Veerse, 1999. Four-dimensional variational analyses of FASTEX situations using special observations. *Q. J. R. Meteorol. Soc.*, **125**:3339–3358.

Douglas, R., 2000. Rearrangements of functions with applications to meteorology and ideal fluid flow. Internal Report 118, Joint Centre for Mesoscale Meteorology.

Durran, D. R., 1999. *Numerical methods for wave equations in geophysical fluid dynamics*. Number 32 in Texts in Applied Mathematics. Springer-Verlag, New York, pp. 465.

Eady, E. T., 1949. Long waves and cyclone waves. *Tellus*, **1**:33–52.

Ehrendorfer, M. and J. Tribbia, 1997. Optimal prediction of forecast error covariances through singular vectors. *J. Atmos. Sci.*, **54**:286–313.

Errico, F. M., 1997. What is an adjoint model? *Bulletin of the American Meteorological Society*, **78**:2577–2591.

Eyre, J. R., 1990. The information content of data from satellite sounding systems: a simulation study. *Q. J. R. Meteorol. Soc.*, **116**:401–434.

Eyre, J. R., 2000. Planet earth seen from space: basic concepts. In *Exploitation of the New Generation of Satellite Instruments for Numerical Weather Forecasts*, Seminar Proceedings, pp. 5–20, Reading, UK. ECMWF.

Farrell, B. F., 1982. The initial growth of disturbances in a baroclinic flow. *J. Atmos. Sci.*, **39**: 1663–1686.

Farrell, B. F., 1984. Modal and non-modal baroclinic waves. *J. Atmos. Sci.*, **41**:668–673.

Farrell, B. F., 1989. Optimal excitation of baroclinic waves. *J. Atmos. Sci.*, **46**:1193–1206.

Fisher, M., 2003. Estimation of entropy reduction and degrees of freedom for signal for large variational analysis systems. Tech. Memo. 397, ECMWF.

Fisher, M. and E. Andersson, 2001. Developments in 4D-Var and Kalman Filtering. Tech. Memo. 347, ECMWF.

Fletcher, S., 1999. Numerical Approximations to Buoyancy Advection in the Eady Model. Master's thesis, Department of Mathematics, University of Reading.

Forray, M. J., 1968. *Variational Calculus in Science and Engineering*. McGraw-Hill Book Company, pp. 221.

Gelfand, I. M. and S. V. Fomin, 1963. *Calculus of Variations*. Prentice-Hall, Englewood Cliffs, pp. 232.

Ghil, M. and P. Malanotte-Rizzoli, 1991. Data assimilation in meteorology and oceanography. *Advances in Geophysics*, **33**:141–266.

Giering, R. and T. Kaminski, 1996. Recipes for adjoint code construction. Report 212, Max-Planck-Institute for Meteorology.

Gilchrist, B. and G. Cressman, 1954. An experiment in objective analysis. *Tellus,* **6**:309–318.

Gill, A. E., 1982. *Atmosphere-Ocean Dynamics*. Academic Press, pp. 662.

Gill, P. E., W. Murray, and M. H. Wright, 1981. *Practical Optimization*. Academic Press, pp. 401.

Golub, G., V. Klema, and G. W. Stewart, 1976. Rank degeneracy and least squares problems. Technical report tr-456, Computer Science Department, University of Maryland.

Golub, G. H. and C. F. Van Loan, 1996. *Matrix computations*. The John Hopkins University Press, third edition, pp. 694.

Golub, G. H. and U. von Matt, 1996. Generalized cross-validation for large scale problems, revised version. Stanford SCCM Report 96-08, Stanford University.

Griffith, A. K., 1997. *Data Assimilation for Numerical Weather Prediction using Control Theory*. PhD thesis, Department of Mathematics, University of Reading.

Griffith, A. K. and N. K. Nichols, 1994. Data assimilation using optimal control theory. Numerical Analysis Report 10/94, Department of Mathematics, University of Reading.

Hansen, P. C., 1992. Regularization tools: A matlab package for analysis and solution of discrete ill-posed problems. Manual IMM-REP-98-6, Informatics and Mathematical Modelling, Technical University of Denmark.
URL `http://www.imm.dtu.dk/ pch/`.

Hansen, P. C., 2001. The L-curve and its use in the numerical treatment of inverse problems. In Johnston, P., editor, *Computational Inverse Problems in Electrocardiology,* pp. 119–142. WIT Press, Southampton.

Higham, N. J., 1984. Computing real square roots of a real matrix. Numerical Analysis Report 89, Department of Mathematics, University of Manchester.

Hoffman, R. N. and C. Grassotti, 1996. A technique for assimilating SSM/I observations of marine atmospheric storms: Tests with ECMWF analyses. *Mon. Weather Rev.,* **35**:1177–1188.

Hoffman, R. N., Z. Liu, J. Louis, and C. Grassotti, 1995. Distortion representation of forecast errors. *Mon. Weather Rev.*, **123**:2758–2770.

Hollingsworth, A., 1987. Objective analysis for numerical weather prediction. In Matsuno, T., editor, *Short- and Medium- Range Numerical Weather Prediction*, pp. 11–60. WMO/IUGG NWP Symposium. Special Volume of the J. Meteor. Soc. Japan: Collection of papers presented at the WMO/IUGG NWP symposium, Tokyo, 4-8 August 1986.

Hollingsworth, A., 2000. Numerical weather prediction: Paradigms and principles. In *Exploitation of the new Generation of Satellite Instruments for Numerical Weather Forecasts*, Seminar Proceedings, pp. 1–4, Reading, UK. ECMWF.

Hollingsworth, A., 2003. Filtering and projection of information as a consequence of the optimality principles of variational analysis - lecture 1. ECMWF Training Course Lecture Notes.
URL `http://www.ecmwf.int/`.

Hollingsworth, A. and P. Lönnberg, 1986. The statistical structure of short-range forecast errors as determined from radiosonde data. Part I: The wind field. *Tellus*, **38A**:111–136.

Holton, J., 1992. *An Introduction to Dynamic Meteorology*. Academic Press, pp. 511.

Hoskins, B. J., 1997. A potential vorticity view of synoptic development. *Meteorol. Appl.*, **4**.

Hoskins, B. J., M. E. McIntyre, and A. W. Robertson, 1985. On the use and significance of isentropic potential vorticity maps. *Q. J. R. Meteorol. Soc.*, **111**:877–946.

Ide, K., P. Courtier, and M. Ghil, 1997. Unified notation for data assimilation: Operational, sequential and variational. *J. Meteor. Soc. Japan*, **75**:181–189.

James, I. N., 1994. *Introduction to Circulating Atmospheres*. Cambridge University Press, pp. 422.

Jazwinski, A., 1970. *Stochastic Processes and Filtering Theory*. Academic Press, pp. 376.

Johnson, C., N. K. Nichols, B. J. Hoskins, S. P. Ballard, and A. S. Lawless, 2002. Four dimensional variational data assimilation in the presence of idealised rapidly growing weather systems. Numerical Analysis Report 1/02, Department of Mathematics, University of Reading.

Jones, C. D. and B. Macpherson, 1997. A latent heat nudging scheme for the assimilation of precipitation data into an operational mesoscale model. *Meteorol. Appl.*, **4**(3):269–277.

Jordan, D. W. and P. Smith, 1997. *Mathematical Techniques*. Oxford University Press, second edition, pp. 788.

Juckes, M. N., 2003. Data analysis and process models: Part i: Ordinary differential equations. Preprint for the Q. J. R. Meteorol. Soc.

Juckes, M. N., 2003. Data analysis and process models: Part ii, two dimensional linear processes. Preprint for the Q. J. R. Meteorol. Soc.

Julian, P. and H. Thiébaux, 1975. On some properties of correlation functions used in optimum interpolation schemes. *Mon. Weather Rev.*, **103**:605–616.

Kalman, R. E., 1960. A new approach to linear filtering and prediction problems. *Transaction of the ASME - Journal of Basic Engineering*, **82**:35–45.

Kalnay, E., 2003. *Atmospheric Modeling, Data Assimilation and Predictability*. Cambridge University Press, pp. 341.

Lagarde, T., A. Piacentini, and O. Thual, 2001. A new representation of data-assimilation methods: The PALM flow-charting approach. *Q. J. R. Meteorol. Soc.*, **127**:189–207.

Lawless, A., 2001. *Development of Linear Models for Data Assimilation in Numerical Weather Prediction*. PhD thesis, Department of Mathematics, University of Reading.

Le Dimet, F.X. and O. Talagrand, 1986. Variational algorithms for analysis and assimilation of meteorological observations: Theoretical aspects. *Tellus*, **38A**:97–110.

Le Dimet, F. X., I. Navon, and D. Daescu, 2002. Second order information in data assimilation. *Mon. Weather Rev.*, **130**:629–648.

Le Veque, R. J., 1992. *Numerical Methods for Conservation Laws*. Lectures in Mathematics. Birkhäuser Verlag, pp. 214.

Lea, D. J., 2001. *Joint Assimilation of Sea Surface Temperature and Sea Surface Height*. PhD thesis, University of Oxford.

Lewis, J. M. and J. C. Derber, 1985. The use of adjoint equations to solve a variational adjustment problem with advective constraints. *Tellus*, **37A**:309–322.

Li, Y., I. Navon, W. Yang, X. Zou, J. Bates, S. Moorthi, and R. Higgins, 1994. Four-dimensional variational data assimilation experiments with a multilevel semi-lagrangian semi-implicit general circulation model. *Mon. Weather Rev.*, **122**:966–983.

Li, Z. and I. M. Navon, 2001. Optimality of 4D Var and its relationship with the Kalman Filter and Smoother. *Q. J. R. Meteorol. Soc.*, **127**:661–683.

Lorenc, A., 1981. A global three-dimensional multivariate statistical interpolation scheme. *Mon. Weather Rev.*, **109**:701–721.

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

Lorenc, A. C., S. P. Ballard, R. S. Bell, N. B. Ingleby, P. L. F. Andrews, D. M. Barker, J. R. Bray, A. M. Clayton, T. Dalby, D. Li, T. J. Payne, and F. W. Saunders, 2000. The Met. Office global three-dimensional variational data assimilation scheme. *Q. J. R. Meteorol. Soc.*, **126**: 2991–3012.

Lorenz, E. N., 1993. *The Essence of Chaos*. UCL Press, pp. 227.

Lu, J. and W. Hsieh, 1997. Adjoint data assimilation in couple atmosphere-ocean models: Determining model parameters in a simple model. *Q. J. R. Meteorol. Soc.*, **123**:2115–2139.

Marotzke, J., R. Giering, K. Q. Zhang, D. Stammer, C. Hill, , and T. Lee, 1999. Construction of the adjoint MIT ocean general circulation model and application to Atlantic heat transport sensitivity. *J. Geophys. Res.*, **104**:29529–29548.

Mateer, C. L., 1965. On the information content of umkehr observations. *J. Atmos. Sci.*, **22**: 370–381.

Morgan, M., 2001. A potential vorticity and wave activity diagnosis of optimal perturbation evolution. *J. Atmos. Sci.*, **58**:2518–2544.

Morgan, M.C. and C.-C. Chien, 2002. Diagnosis of optimal perturbation evolution in the Eady model. *J. Atmos. Sci.*, **59**:169–185.

Muraki, D. J., C. Snyder, and R. Rotunno, 1999. The next-order corrections to quasigeostrophic theory. *J. Atmos. Sci.*, **56**:1547–1560.

NAG, 2003. Numerical algorithms group (nag).
URL `http://www.nag.co.uk/`.

Nash, J. C., 1990. *Compact Numerical Methods for Computers: Linear Algebra and Function Minimization*, pp. 186–206. Adam Hilger, IOP Publishing, second edition.

Nash, J. C., 2003. A22CGM code. Guide to Available Mathematical Software, National Institute of Standards and Technology.
URL `http://gams.nist.gov/serve.cgi/Module/NASHLIB/A22/11240/`.

Navon, I. M., 1997. Practical and theoretical aspects of adjoint parameter estimation and identifiability in meteorology and oceanography. *Dynamics of Atmospheres and Oceans*, **27**:55–79.

Navon, I. M. and D. M. Legler, 1987. Conjugate-gradient methods for large-scale minimization in meteorology. *Mon. Weather Rev.*, **115**:1479–1502.

Navon, I. M., X. Zou, J. Derber, and J. Sela, 1992. Variational data assimilation with an adiabatic version of the NMC spectral model. *Mon. Weather Rev.*, **120**:1433–1446.

Noble, B. and J. W. Daniel, 1988. *Applied Linear Algebra*. Prentice-Hall, Englewood Cliffs, third edition, pp. 521.

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

Pedlosky, J., 1964. An initial value problem in the theory of baroclinic instability. *Tellus*, **16**: 12–17.

Pedlosky, J., 1987. *Geophysical Fluid Dynamics*. Springer-Verlag, second edition, pp. 710.

Pires, C., R. Vautard, and O. Talagrand, 1996. On extending the limits of variational assimilation in nonlinear chaotic systems. *Tellus*, **48A**:96–121.

Preisendorfer, R. W., 1988. *Principle Component Analysis in Meteorology and Oceanography*. Number 17 in Developments in Atmospheric Science. Elsevier, pp. 425. Edited by C. D. Mobley.

Press, W., S. Teukolsky, W. Vetterling, and B. Flannery, 1992. *Numerical Recipes in Fortran 77*, volume 1. Academic Press.

Prunet, P., J.-N. Thépaut, and V. Cassé, 1998. The information content of clear sky IASI radiances and their potential for numerical weather prediction. *Q. J. R. Meteorol. Soc.*, **124**: 211–241.

Rϕsting, B., J. E. Kristjansson, and J. Sunde, 2001. Can modification of the PV in the numerical analyses improve the model simulation. Research Report 126, DNMI.

Rabier, F. and P. Courtier, 1992. Four-dimensional assimilation in the presence of baroclinic instability. *Q. J. R. Meteorol. Soc.*, **118**:649–672.

Rabier, F., N. Fourrié, D. Chafaï, and P. Prunet, 2002. Channel selection methods for infrared atmospheric sounding interferometer. *Q. J. R. Meteorol. Soc.*, **128A**:1011–1027.

Rabier, F., H. Järvinen, E. Klinker, J. Mahfouf, and A. Simmons, 2000. The ECMWF operational implementation of four-dimensional variational assimilation. I: Experimental results with simplified physics. *Q. J. R. Meteorol. Soc.*, **126**:1143–1170.

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

Rabier, F., J. Thepaut, and P. Courtier, 1998. Extended assimilation and forecast experiments with a 4D-Var assimilation system. *Q. J. R. Meteorol. Soc.*, **124**:1861–1887.

Riley, K. F., M. P. Hobson, and S. J. Bence, 1998. *Mathematical methods for Physics and Engineering*. Cambridge University Press.

Rodgers, C. D., 1996. Information content and optimisation of high spectral resolution measurements. In *Optical Spectroscopic Techniques and Instrumentation for Atmospheric and Space Research II*, volume 2830, pp. 136–147. SPIE.

Rodgers, C. D., 2000. *Inverse Methods for Atmospheric Sounding, Theory and Practice*, volume 2 of *Atmospheric, Oceanic and Planetary Physics*. World Scientific, pp. 238.

Rosmond, T. E., 1997. A technical description of the NRL adjoint modelling system. Memorandum Report NRL/MR/7532/97/7230, Naval Research Laboratory.

Sasaki, Y., 1970. Some basic formalisms in numerical variational analysis. *Monthly Weather Review*, **98**:875–883.

Schröter, J., U. Seiler, and M. Wenzel, 1993. Variational assimilation of geosat data into an eddy-resolving model of the gulf stream extension area. *J. Phys. Oceanogr.*, **23**:925–953.

Semple, A., 2001. A meteorological assessment of the geostrophic co-ordinate transform and error breeding system when used in 3D variational data assimilation. NWP Technical Report 357, Met Office.

Shanno, D. F. and K. H. Phua, 1976. Algorithm 500: Minimization of unconstrained multivariate functions. *ACM Transactions On Mathematical Software*, **2**:87–94.

Shanno, D. F. and K. H. Phua, 1980. A remark on algorithm 500:minimization of unconstrained multivariate functions. *ACM Transactions of Mathematical Software*, **6**:618–622.

Shanno, D. F. and P. K. H. Phua, 2003. CONMIN code. Journal of ACM Transactions on Mathematical Software, Guide to Available Mathematical Software, National Institute of Standards and Technology.
URL `http://gams.nist.gov/serve.cgi/Module/TOMS/500/8535/`.

Shewchuk, J. R., 1994. An introduction to the conjugate gradient method without the agonizing pain, edition 1.25. Technical report, School of Computer Science, Carnegie Mellon University, Pittsburgh.

Sirkes, Z. and E. Tziperman, 1997. Finite difference of adjoint or adjoint of finite difference? *Mon. Weather Rev.*, **125**:3373–3378.

Snyder, C., 1996. Summary of an informal workshop on adaptive observations and fastex. *Bull. Amer. Meteor. Soc.*, **77**:953–961.

Strang, G., 1986. *Linear Algebra and its applications*. Harcourt Brace Jovanovich College Publishers.

Swanson, K. L., T. N. Palmer, and R. Vautard, 2000. Observational error structures and the value of advanced assimilation techniques. *J. Atmos. Sci.*, **57**:1327–1340.

Swarbrick, S. J., 2001. Applying the relationship between potential vorticity fields and water vapour imagery to adjust initial conditions in NWP. *Meteorol. Appl.*, **8**:221–228.

Talagrand, O., 1998. A posteriori evaluation and verification of analysis and assimilation algorithms. In *Diagnosis of data assimilation systems*, Workshop Proceedings, pp. 17–28, Reading, UK. ECMWF.

Tanguay, M., P. Bartello, and P. Gauthier, 1995. Four-dimensional data assimilation with a wide range of scales. *Tellus*, **47A**:974–997.

Thépaut, J.-N. and P. Courtier, 1991. Four-dimensional variational data assimilation using the adjoint of a multilevel primitive-equation model. *Q. J. R. Meteorol. Soc.*, **117**:1225–1254.

Thépaut, J.-N., P. Courtier, G. Belaud, and G. Lemaître, 1996. Dynamical structure functions in 4D variational assimilation: A case study. *Q. J. R. Meteorol. Soc.*, **122**:535–561.

Thépaut, J.-N., R. N. Hoffman, and P. Courtier, 1993. Interactions of dynamics and observations in a 4D variational assimilation. *Mon. Weather Rev.*, **121**:3393–3414.

Thépaut, J-N., D. Vasiljevic, and P. Courtier, 1993. Variational assimilation of conventional meteorological observations with a multilevel primitive-equation model. *Q. J. R. Meteorol. Soc.*, **119**:153–186.

Thiébaux, H. J., 1975. Experiments with correlation representations for objective analysis. *Mon. Weather Rev.*, **103**:617–627.

Thompson, P. D., 1961. A dynamical method of analysing meteorological data. *Tellus*, **13**: 334–349.

Thuburn, J. and T. W. N. Haine, 2001. Adjoints of nonoscillatory advection schemes. *Journal of Computational Physics*, **171**:616–631.

Toumazou, V., 2001. Using a Lanczos eigensolver in the computation of empirical orthogonal functions. *Mon. Weather Rev.*, **129**:1243–1250.

Wahba, G. and J. Wendelberger, 1980. Some new mathematical methods for variational objective analysis using splines and cross validation. *Mon. Weather Rev.*, **108**:1122–1143.

Weaver, A. and P. Courtier, 2001. Correlation modelling on the sphere using a generalized diffusion equation. *Q. J. R. Meteorol. Soc.*, **127**:1815–1846.

Wergen, W., 1992. The effect of model errors in variational assimilation. *Tellus*, **44A**:297–313.

Winkler, J. R., 1997. Tikhonov regularisation in standard form for polynomial basis conversion. *Appl. Math. Modelling*, **21**:651–662.

Wlasak, M., 1997. Variational Data Assimilation: A Study. Master's thesis, Department of Mathematics, University of Reading.

WMO, 2003. Global Observing System.
URL `http://www.wmo.ch/web/www/OSY/GOS.html`.

Wunsch, C., 1977. Determining the general circulation of the oceans: A preliminary discussion. *Science*, **196**:871–875.

Wunsch, C., 1996. *The Ocean Circulation Inverse Problem*. Cambridge University Press, pp. 442.

Xu, Z. and N. Nichols, 1991. Hydrodynamic modelling and optimal control of tidal power schemes in long estuaries. Numerical Analysis Report 6/91, Department of Mathematics, University of Reading.

Zou, X., I. M. Navon, and F. X. Le Dimet, 1992. Incomplete observations and control of gravity waves in variational data assimilation. *Tellus*, **44A**:273–296.

Zou, X., I. M. Navon, and F. X. Le Dimet, 1992. An optimal nudging data assimilation scheme using parameter estimation. *Q. J. R. Meteorol. Soc.*, **118**:1163–1186.