Balance Dynamics and Gravity Waves in Four-Dimensional Data Assimilation

by

Lisa Neef

A thesis submitted in conformity with the requirements
for the degree of Doctor of Philosophy
Graduate Department of Physics
University of Toronto

Abstract

Balance Dynamics and Gravity Waves in Four-Dimensional Data Assimilation

Lisa Neef

Doctor of Philosophy

Graduate Department of Physics

University of Toronto

2007

This thesis examines the application of three methods of so-called four-dimensional data assimilation to dynamical models where there exists a timescale separation between vortical motion and (relatively fast) inertia-gravity waves. Using a highly simplified dynamical model which admits one nonlinear vortical mode and one inertia-gravity wave, we evaluate the relative strengths and weaknesses of linearization-based and ensemble-based sequential assimilation (i.e. two varieties of the Kalman filter), and four-dimensional variational assimilation (4DVAR).

The first part of this study is concerned with balanced flow, or flow where vortical motion dominates and inertial/gravitational motion is “slaved” to the dominant flow. The goal of assimilation in this context is to recover the true balanced state, without the excitation of spurious inertia-gravity waves. It is shown that the excitation of spurious waves becomes more difficult to control as the nonlinearity of the assimilation system is increased, for example by decreasing observation frequency. If not enough components of the true state are observed or observations are infrequent relative to the nonlinearity of the model, the explicit evolution of error covariances using a tangent-linear model can easily become quite inaccurate, which results in a highly unstable assimilation cycle wherein spurious waves are excited and not controlled. Both ensemble-based and implicit variational covariance models offer improvements, but these are themselves limited by error due to sampling and non-Gaussianity of the ensemble and by the tendency to settle into local minima of a non-quadratic cost function.

The analysis is then extended to dynamical regimes where the inertia-gravity wave becomes more important to the evolution of the system as a whole, either by increasing its magnitude, decreasing the timescale separation, or increasing the coupling between fast and slow modes. It is found that recovery of either mode from observations that contain both timescales benefits from the four-dimensional estimation of error statistics. The ability to extract both modes from observations which contain both timescales of motion depends both on the estimated fast-slow covariances, as well as the estimated error variance ascribed to the gravity wave. Recovery of a non-negligible inertia-gravity wave is found to be possible with the Kalman filter, and more so if an ensemble is used to estimate covariances, but extremely difficult for variational assimilation.

It is also found that accuracy of the assimilation for the different regimes of balance/imbalance can be weakened considerably as systematic model error is added and increased. Some typical modifications designed to counter systematic error are shown to alleviate some of these problems, but also increase the excitation of spurious imbalance.

Dedication

This thesis is dedicated to my dad, who got me into this mess in the first place.

Acknowledgements

I owe a great deal of thanks to my thesis supervisors and teachers, Drs. Ted Shepherd and Saroja Polavarapu, for advice and help, and endless patience in the face of my ridiculously optimistic promises regarding deadlines, and for sending me all over the world. My work at the University of Toronto was supported by a UT Open Fellowship and a Blythe Fellowship. Now follows a long list of people who helped make this thesis possible in some way.

Drs. Gary Morris, Lynn Sparling, and Jen McKinnon gave me much advice along the way and implicitly helped shape this thesis. Marianne Khurana and Krystyna Biel helped me to handle the (very confusing) logistical acrobatics associated with being a grad student. Dr. Chris Snyder offered very helpful feedback on my thesis in the final revision stage. For pep talks, shenanigans, and the occasional research discussion I thank Brian Ancell, Amit Ghosh, Rebekah Martin, Marek Stastna, Aldona Wiacek, Caroline Nowlan, David Sankey, and Danielle Wain.

My Mom and Dad have supported me every second, from my first day of school to the last, pushing me when I needed it, and (more importantly) giving me a loving and warm place to come home to when I needed it. Thank you for being such great examples, and also for driving me to and from O'Hare International Airport about 10,000 times. For commiseration and hilarity I thank my brothers and fellow reluctant scientists, Tobias and Daniel. My high school physics teacher, Bob Grimm, introduced me to physics in the first place, and Don Koetke and the physics faculty at Valparaiso University laid down the basics.

Grad school would have killed me if it hadn't been for the love, support, and hair-brained ideas of The Dining Room: Leah Hunter, Trent Hunter, Dylan Royal, Adrienne Caesar, Becky Hubble, Paul Hubble, Jen McGhee, Kate Hodgert, Rob Whillans and Karl Richter. Carolin Engel made me snacks and encouraged me during the final two months of writing. Much love and thanks also to Chester Li and Liz Ivkovich, Sten Witzel, Sue

Erickson, Team India, and the communities at Freedomize Toronto and Wine Before Breakfast, for friendship and faith, and doing small things with great love. This thesis was written largely to the music of Neko Case, Dar Williams, and Sufjan Stevens. All the rest comes down to Zechariah 4:6.

Contents

List of Figures

Chapter 1

Introduction

1.1 Balance, Imbalance, and the Data Assimilation Problem

1.1.1 Balance Dynamics and the Slow Manifold

Atmospheric and oceanic data assimilation is the problem of estimating the most likely state of a very complex system, given physical theory on the one hand and a set of observations on the other. Physical theory exists in the form of models, which give estimates of the state in space and time. Observations can be ground-based, space-based, or in-situ, and also involve a variety of spatial and temporal scales and levels of complexity. Both the state estimate and observations are typically large fields, and involve many dissimilar variables. While the abundance of observations of the atmosphere and oceans is increasing, and models are improving, the significant improvements made in both weather forecasts and climatological analyses in recent years are in fact largely due to improved methods of data assimilation [Rabier, 2005]. Data assimilation has two general purposes: (1) to "fill in the blanks" between observations, with the objective of finding a more complete climatological picture, and (2) to improve forecasts of processes which have finite predictability, by providing more accurate initial conditions. More than just inserting measurements into models, data assimilation algorithms are designed to use observed and modeled information of a given variable at a certain location and time to inform our estimates of other variables at other points in space and time, based on the physical relationships which connect them. At the heart of the data assimilation problem lies the exploitation of physical relationships, which are interpreted statistically in the form of multi-dimensional probability distribution functions.

Dynamical relationships are modeled in terms of motions or waves, which correspond to different physical restoring mechanisms [e.g., Kalnay, 2003, chapter 2]. Although these different wave types are not completely independent, they can be approximated as such in many meteorological and climatological applications, thereby making the problem more tractable. We can thus think of flow in terms of sound waves, which result from compression of the air; gravitational oscillations, which result from the gravitational restoring force; inertial oscillations, which result from the Coriolis force; slower vortical modes, which correspond to the conservation of potential vorticity; and Rossby modes, which result from the variation of the Coriolis parameter (f) with latitude.

Sound waves have very fast phase speeds relative to, and tend to have a negligible effect on, the motion of climatological and meteorological interest in the atmosphere and ocean. External gravitational oscillations have phase speeds $c_{\mathrm{gw}} \sim \sqrt{gH}$, where H is a typical fluid depth (in the ocean) or density scale height (in the atmosphere) and g the acceleration due to gravity. Internal gravitational oscillations (resulting from stratification) have phase speeds $c_{\mathrm{gw}} \sim NH$, where H is a typical vertical length scale and N the buoyancy or Brunt-Väisälä frequency, associated with the reduced gravity of a stratified fluid. For $H \sim 10$ km and $N \sim 10^{-2}$ s^{-1}, the phase speed of external gravity waves is on the order of about 300 m/s and that of internal waves is about 100 m/s. Inertial oscillations have phase speeds $c_{\mathrm{i}} \sim fL$, where L represents a typical horizontal lengthscale. For horizontal lengthscales $L \sim 1000$ km and at midlatitudes (where $f \sim 10^{-4}\mathrm{s}^{-1}$), the inertial phase speed is on the order of 10^2 m/s.

Now consider large-scale flow with characteristic horizontal velocity U and horizontal lengthscale $L \ll a$, where a is the radius of the Earth. Since the scale of these motions is small relative to the curvature of the earth, we can, for the sake of argument, approximate f as constant, in which case Rossby waves are filtered from the physical model, leaving inertial and gravitational oscillations, as well as largely-nondivergent, vortical flow that is approximately parallel to pressure surfaces. This vortical motion actually results from the nonlinear advection of velocity in the governing equations, and can be understood in terms of the conservation of a quantity called the potential vorticity, which can be interpreted as the circulation per volume, or the absolute vorticity of a fluid volume (that is, its intrinsic vorticity plus the Coriolis vorticity, f) divided by its depth.

Inertial and gravitational oscillations are well-separated in timescale from such vortical modes if the ratio of the flow speed U relative to the phase speeds of inertial and gravitational waves, or

$$\epsilon \equiv \frac{U}{\left(c_{\mathrm{gw}}^2 + c_{\mathrm{i}}^2\right)^{1/2}} = \frac{U}{\left(gH + f^2L^2\right)^{1/2}}, \tag{1.1}$$

is small. ϵ can also be written as

$$\epsilon \equiv \frac{\mathrm{Ro}\ \mathrm{B}}{\sqrt{1+\mathrm{B}^2}}, \tag{1.2}$$

where $\mathrm{Ro} \equiv U/fL$ is the Rossby number and $\mathrm{B} \equiv fL/\sqrt{gH}$ is the rotational Froude number. Ro measures the vorticity of the flow relative to the earth's rotation, or the ratio between the inertial timescale ($\tau_1 = f^{-1}$) and the advective timescale ($\tau_2 = L/U$). Likewise, B measures the importance of the Coriolis force relative to the gravitational restoring force. The smallness of these numbers indicates a timescale separation between different normal modes [Saujani and Shepherd, 2006]. In regimes where $\mathrm{B} \sim 1$ and $\mathrm{Ro} \ll 1$, for example, inertial and gravitational oscillations are mixed (and we talk about inertia-gravity waves), but separated in timescale from large-scale advective flow.

The coexistence of scale-separated and therefore approximately independent modes of motion leads to the concept of balance dynamics. Balanced models are simplifications of the governing equations which filter out modes or degrees of freedom which are assumed to have a negligible impact on the temporal and spatial scales of interest. A simple example of a balanced model is incompressible flow, which filters out sound waves.

In meteorological and climatological applications, "balance" generally refers to the dominance of vortical motion over inertia-gravity waves. This is because ϵ is small in the extratropical atmosphere, where the gravity and inertial timescales are a few hours or less, while the vortical or advective timescale is on the order of days. The midlatitude troposphere and lower stratosphere are "balanced" in the sense that the flow prefers slow, vortical motion over inertia-gravity waves, which carry comparatively little energy. Weather systems and large-scale climate variations consist of Rossby and vortical modes, varying on timescales of days to weeks. Inertia-gravity waves, excited by topography and frontal systems, have timescales of hours and propagate away with little influence on the large-scale flow, eventually breaking in the upper stratosphere and mesosphere. Figure 1.1 shows a sample weather map, generated from Environment Canada's Global Environment Multi-scale (GEM) Model. The solid lines are contours of geopotential height of the 500 hPa pressure surface, and the colour contours are absolute vorticity. The flagged arrows indicate wind strength and direction, and it is evident that wind is primarily parallel to the large-scale contours of height of the pressure surface. Note also that absolute vorticity maxima coincide with "troughs" in the pressure surface. Both features are indicative of balanced motion.

This preponderance of slow motion in the midlatitude troposphere and stratosphere allows us to simplify our understanding of the governing dynamics, by approximating flow on a quasi-invariant submanifold in phase space, on which gravity waves are entirely "slaved" to rotational motion [Leith, 1979]. The so-called slow manifold is defined by a system's slow variables, while the fast variables are functions of the slow variables, given

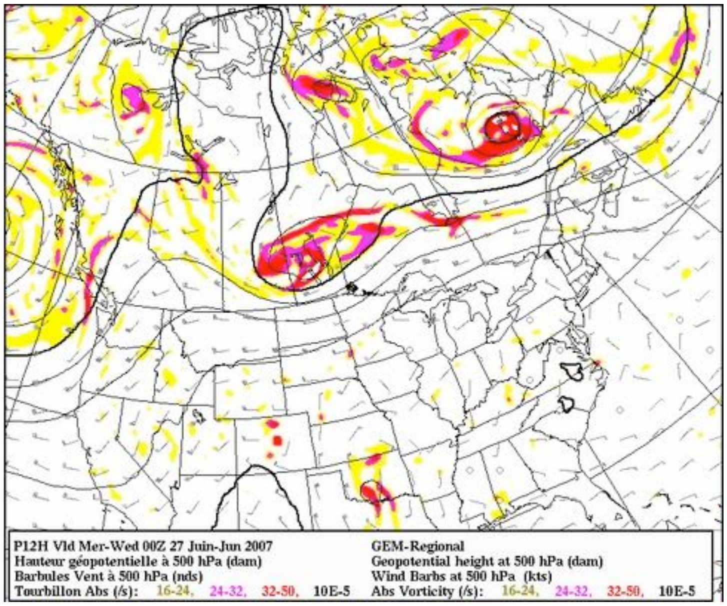
P12H Vld Mer-Wed 00Z 27 Juin-Jun 2007
Hauteur géopotentielle à 500 hPa (dam)
Barbules Vent à 500 hPa (nds)
Tourbillon Abs (/s): 16-24, 24-32, 32-50, 10E-5
GEM-Regional
Geopotential height at 500 hPa (dam)
Wind Barbs at 500 hPa (kts)
Abs Vorticity (/s): 16-24, 24-32, 32-50, 10E-5

Figure 1.1: Example of pressure and wind over North America on 27 June, 2007, from Environment Canada's GEM model. The solid lines are contours of geopotential height of the 500 hPa pressure surface, the colour shading represents absolute vorticity, and arrows represent horizontal wind. Published by Environment Canada, and available from the institution's website, *http://www.weatheroffice.gc.ca*.

by a set of relations of the form

$$\mathbf{f} = U\left(\mathbf{s};\epsilon\right), \tag{1.3}$$

where $\mathbf{f}$ represents the fast variables and $\mathbf{s}$ the slow variables. For example, under so-called geostrophic balance, the Coriolis force is assumed to dominate all other acceleration terms, such that the Coriolis and pressure forces balance and horizontal flow aligns with lines of constant pressure, p. Geostrophic wind is given by

$$u = -\frac{1}{\rho f}\frac{\partial p}{\partial y} \tag{1.4}$$

$$v = \frac{1}{\rho f}\frac{\partial p}{\partial x} \tag{1.5}$$

and can thus be thought of as *slaved* to pressure. Subsequently, wind can be inferred from pressure contours in weather maps.

There are two ways to model flow on the slow manifold. One way is to prognostically model only the "master" variables which evolve on the slow timescale, and find the "slaved" fast variables diagnostically via the slaving relations (1.3). Alternatively, it is possible to integrate the full equations, but initialize the state on the approximate slow manifold, such that the flow evolves without inertia-gravity waves. The latter has been shown to be preferable, because the truncated equations yield inaccurate forecasts. The slow manifold has since been shown to be asymptotic instead of invariant [Lorenz, 1986, Warn, 1997, Wirosoetisno and Shepherd, 2000], meaning that initialization cannot actually converge to an exact manifold, and states initialized on the slow manifold will eventually develop unbalanced motion.

The picture changes in the mesosphere, where the timescale separation remains, but the motion is primarily unbalanced, i.e. gravity waves are prevalent [Koshyk et al., 1999]. These gravity waves are generated in the troposphere and propagate upward, increasing in amplitude in the rarified air, eventually breaking and depositing their energy and momentum in the stratosphere, mesosphere, and thermosphere. This is illustrated in figure 1.2, which shows the modeled zonal velocity field at three pressure levels, from the SKYHI GCM, after Koshyk et al. [1999]. At 9.22 mb (the middle stratosphere), u is dominated by large scales. Small-scale structure increases as we move into the upper stratosphere (1.50 mb), with small scales dominating at 0.13 mb (the lower mesosphere). From model studies it is estimated that roughly 50% of total variance in the mesosphere comes from gravity waves [Polavarapu et al., 2005]. It is important to account for their presence, as the breaking of these gravity waves plays an important role in the overall dynamical picture of the middle atmosphere [Shepherd, 2000].

The balance picture also changes in the tropics, where the Coriolis parameter goes to zero, and the timescale separation between inertia-gravity waves and vortical modes accordingly decreases. The tropics also give rise to special equatorial waves, including mixed Rossby-gravity waves, as well as equatorially-trapped Kelvin waves.

1.1.2 Balance Dynamics in Data Assimilation

In the language of data assimilation, physical motion becomes statistical, which is necessary because both forecasts and observations have errors. The information given by observations and models is formulated in terms of probability distribution functions (pdfs) and relationships between variables and gridpoints are encompassed in joint distributions. If these pdfs are Gaussian, they can be fully characterized by their first and second moments, or the mean and variance. Likewise, multi-dimensional Gaussian pdfs can be characterized by covariances.

The probability of variable x at point A having a certain value, for example, is not independent of the probability of the variable y at point B having a certain value. The covariance between these points is $c_{x_A y_B} = \langle x_A y_B \rangle = \rho_{x_A y_B} \langle (x_A - \langle x_A \rangle)^2 \rangle^{1/2} \langle (y_B - \langle y_B \rangle)^2 \rangle^{1/2}$, where $\rho_{x_A y_B}$ is the statistical correlation between the two points and/or variables, and $\langle \cdot \rangle$ represents the expected value. In the same way, the errors associated with different variables and gridpoints can be related to one another in terms of error covariances. An

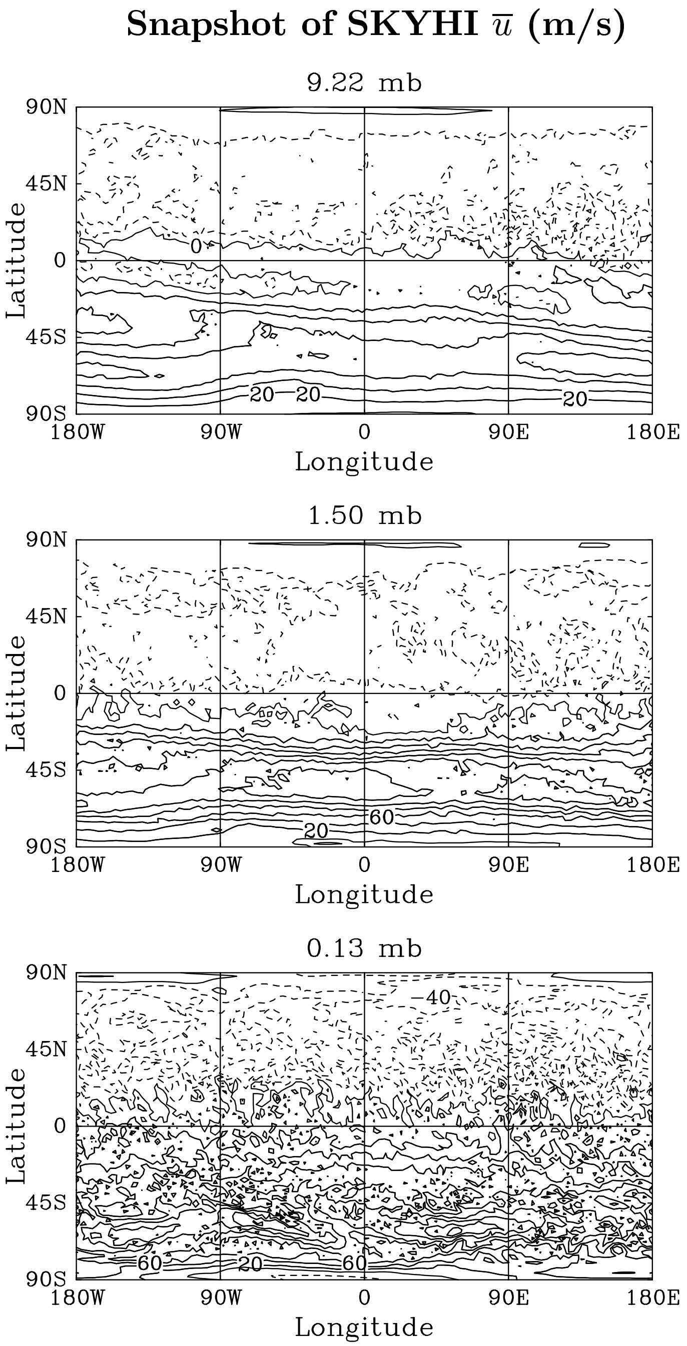

Figure 1.2: Snapshots of the zonal velocity field from the SKYHI model, at three different model levels. After Koshyk et al. [1999].

assimilation algorithm uses an *error covariance model* in order to spread information in time and space. This is where balances become important. For example, if the motion connecting two variables is an inertia-gravity wave, the covariance between the two points (and its evolution in time) will be different than if the motion is vortical. If the true state is balanced, information about one quantity (e.g. wind) can be extracted from an observation of another quantity (e.g. temperature) if the slaving relationship which defines the balance (1.3) is captured within the covariance model.

Thus knowledge of physical motion implies knowledge of joint probabilities, which is used for the interpretation of observations. Likewise, information from observations corrects physical models, and hence the estimation of dynamical relationships. Four-dimensional (4D) data assimilation takes advantage of this duality by cycling information between observations and models, to produce estimates of multivariate, spatial, and temporal error statistics which are consistent with both. As the name implies, 4D data assimilation spreads observational information in both space and time, whereas (more traditional) 3D assimilation spreads information only in space.

1.2 Challenges

It has already been shown that the problem of deriving multivariate error statistics is not entirely solved by the four-dimensionalization of data assimilation, and while several studies in recent years have demonstrated the advantages of 4D over 3D data assimilation [e.g. Rabier, 2005, Bishop et al., 2001, Lorenc and Rawlins, 2005], it has also been shown that the performance of 4D schemes in realistic settings is not always better than 3D assimilation [e.g. Houtekamer et al. 2005]. It is thus imperative to understand specific reasons why 4D schemes might either flourish or only offer minimal improvement in specific physical contexts, and to understand how multivariate covariance models are developed, given practical limitations such as nonlinearity and model error. In this

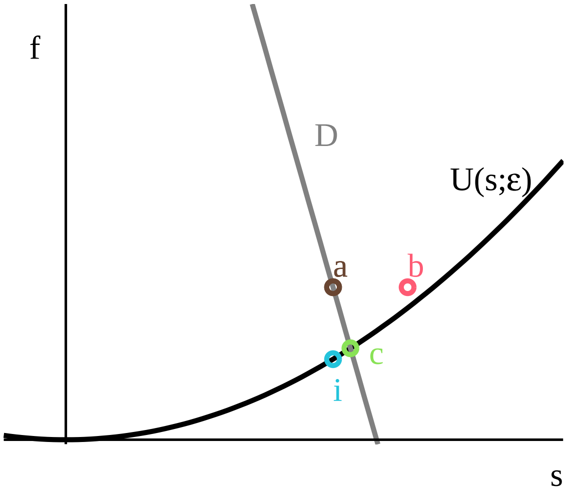

Figure 1.3: Schematic diagram of data assimilation on the slow manifold. The black line indicates the slaving relationship which defines the hypothetical manifold, and line D the data manifold. Point b represents the background state estimate, prior to assimilation. Point a represents the fit between observations and the background estimate. Point c represents the balanced state which is also part of the data manifold. Point i represents a hypothetical "initialized" state , wherein gravity waves are removed (using a balance approximation), following the insertion of the observation.

study, we address the development of multivariate error covariances from the perspective of balance dynamics and different regimes of timescale overlap.

1.2.1 Data Assimilation for Balanced States

For data assimilation in the midlatitude troposphere, where vortical motions dominate and gravity waves can be considered negligible, a problem arises when observations are inserted into models. Because observation errors project onto all degrees of freedom, including the fastest ones, the insertion of observations can destroy the dynamical balance between mass and velocity fields, causing the excitation of spurious, unrealistic inertia-gravity waves in the forecast. This is illustrated in figure 1.3, a schematic diagram of the slow manifold adapted from Leith [1979] and Daley and Puri [1980]. The nonlinear balance relationship which defines the hypothetical slow manifold (1.3) is shown by the black curve. Suppose that a perfect observation of a single variable (say, height of a pressure surface) is made. The line D represents the data manifold, or all points in the model space along which the observed variable is invariant. Let point b be the initial background state (prior to assimilation), which is close to the slow manifold. The background state is adjusted to fit the observation, say, by allowing an adjustment only on the slow variables. The resulting *analysis* state, denoted by the point a, fits the observation but is clearly farther from the slow manifold than the prior estimate. In practical terms, this means that spurious inertia-gravity waves are excited by the insertion of the observation.

Spurious inertia-gravity waves are undesirable in numerical forecasting, for several reasons. They greatly increase the distance between the estimated state, and subsequent forecasts, and the truth. In fact, the world's first numerical weather forecast, made by hand by L. F. Richardson in 1922 [Kalnay, 2003, chapter 1; Richardson, 1922] had huge errors, not because of Richardson's calculation, but because the initial conditions lead to contamination of the forecast by large spurious inertia-gravity waves. The problem is illustrated in figure 1.4, from Williamson and Temperton [1981]. The solid line shows the evolution of surface pressure in a model integration, starting from an initial state which is an assimilated analysis state, such as point a in figure 1.3. The huge variation in the pressure forecast is due to the unphysically large, spurious inertia-gravity wave which ensues from the initial conditions. Such spurious gravity waves also increase the distance between the forecast and subsequent observations (of the balanced true state); hence, subsequent observations can be rejected by the assimilation algorithm. Moreover, if the modeled state adjusts back to a balanced state (either by the dissipation of gravity waves or because the gravity waves propagate away), the observational information which projected onto the slaved fast variables is lost [e.g. Daley 1991, chapter 6].

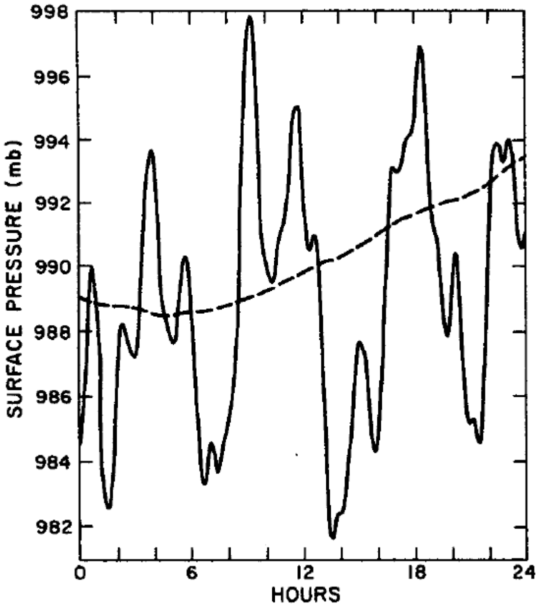

Figure 1.4: Evolution of surface pressure over a 24-hour forecast period, starting from unbalanced initial conditions (solid line) and balanced initial conditions (dashed line). After Williamson and Temperton [1981].

Instead, balance-aware data assimilation should compute a state which is a fit between the background estimate and the observations, *and* which is balanced. For the perfect-observation example in figure 1.3, this state would be given by the point c, the intersection of $U(s;\epsilon)$ and D. Traditionally, the optimal balanced state is approximated with a so-called initialization step, wherein some approximation is made to remove spurious fast motion and project the assimilated state onto the slow manifold, following the insertion of observations. The dashed line in figure 1.4 shows the evolution of surface pressure if the initial state is initialized with a balance approximation; now, surface pressure shows the slow, realistic evolution of the state. Note also that, even though the ultimate change in surface pressure over the 24-hour period is similar for the two curves, instantaneous rates of change are quite different throughout. Thus the small change in the initial conditions which happens during the initialzation step greatly improves the accuracy and physicality of the ensuing forecast. However, depending on how well balance is estimated, this approximate projection of the analysis onto the slow manifold could also result in an analysis that is farther from both the data manifold and the original background estimate, say, point i on the diagram in figure 1.3. The problem thus becomes one of incorporating the balance requirement into the computation of the optimal fit between the observations and the background estimate.

Dynamical balances are, in fact, frequently incorporated into the formulation of the error covariance models used in data assimilation, both in order to prevent the excitation of spurious inertia-gravity waves and (since balance relationships imply error correlations) to make the data assimilation multivariate. It has long been recognized that observations should, in theory, contain information about balance, which can then be incorporated into the assimilation in the form of covariance fields [Bergman, 1979, Daley and Puri, 1980]. In principle, the flow-dependent forecast error statistics developed within 4D assimilation should do exactly this: observed information about balance between variable fields informs the covariance model, which in turn causes observations to be assimilated in a balanced way. However, as we shall see in subsequent chapters, this is not always the case in the presence of realistic limitations, including nonlinearity of the model and the balance relationship.

1.2.2 Data Assimilation for Unbalanced States

The extension of models and observations (and, therefore, data assimilation) into the middle atmosphere and tropics gives rise to two new but related issues: capturing error statistics for states where the motion is predominantly unbalanced (such as the mesosphere), and/or where there is unclear timescale separation between vortical motion and inertia-gravity waves. In the middle atmosphere, where observations are sparse, the ability to extract one field from observations of another is very important for accurate simulation of features which impact climate but are difficult to simulate [e.g. Polavarapu et al. 2005]. Recovery of one field from observations of another is also difficult in the tropics, where the weakness of the Coriolis force reduces statistical coupling between mass and wind fields [Žagar et al., 2004a]. For example, if a balance relation of the form (1.3) is employed to recover the wind field from density/height observations, the process of assimilation will smooth over the (non-negligble) inertia-gravity waves present in the true state [Burgers et al., 2002]. In both contexts, the data assimilation problem is about the interpretation of observational information in terms of the correct type of motion when different motions are possible.

1.3 Overview

In this thesis we use a simplified dynamical model to address the derivation of multivariate error covariances from the perspective of balance dynamics, in three contexts: a balanced truth, an unbalanced truth, and a truth with no clear balance relation. Our model of choice has a chaotic slow mode coupled to an oscillatory fast mode, and can be initialized such that evolution of the fast variables is slaved to that of the slow variables. Analytical reasoning and numerical experiments will approach each of these problems from three angles, corresponding to the three most basic 4D methods: the Extended Kalman Filter (EKF), which uses a tangent-linear model to estimate the evolution of error covariances; the Ensemble Kalman Filter (EnKF), which uses an ensemble of states to estimate the same; and four-dimensional variational assimilation (4DVAR), which minimizes a cost function to optimize the fit between the modeled and observed states.

In comparing these three basic methods to 3D data assimilation, we seek to answer the following questions: (1) Can 4D data assimilation develop statistical models which are representative of the true dynamics? Specifically, (2) can 4D methods capture *slaving* of fast variables to slow variables? Alternatively, (3) can 4D methods capture the coexistence of vortical motion and a comparatively fast wave? (4) How sensitive is this result to the parameters which govern balance / imbalance and the parameters which govern each assimilation scheme? Finally, (5) what are the caveats and suggestions, given by idealized experiments, for realistic assimilation?

Simplified models, which isolate certain physical phenomena, are a useful tool for understanding how a given data assimilation algorithm works in a specific physical context. Models with only a few variables make the data assimilation system easy to understand analytically, and have enabled much insight into the details of specific algorithms. For example, Miller et al. [1994] used the well-known 3-component Lorenz [1963] model to show how 4D data assimilation becomes more difficult in the context of highly nonlinear dynamics, and Evensen [1994] used the same model to show that ensemble-based covariance modeling can potentially alleviate this problem. Anderson and Anderson [1999] used the same model and the related Lorenz [1980] model to show how a non-Gaussian ensemble filter (appendix B) can potentially alleviate new, ensemble-related issues, as well as the balance problem (though their results with regards to balance were largely speculative and limited to a few examples). In a simple model there is also no commitment to operational constraints, nor is there any investment in the results coming out in favour of one particular assimilation method.

The main focus of this work is on the EKF and EnKF, which are used as a proxy for the more general practical issue of linearization-based versus ensemble-based covariance modeling. Linearization-based dynamic covariance models lead to an unstable assimilation system in nonlinear systems, but this instability can be controlled if observations are frequent and accurate enough [Miller et al., 1994]. Because the EnKF preserves the effect of higher-order moments on the estimated error covariance evolution, Verlaan and Heemink [2001] proposed that the difference between the EnKF and EKF analyses gives a measure of the overall nonlinearity of the entire data assimilation system. Here we propose that the existence of a nonlinear balance relationship of the form (1.3) further increases the nonlinearity of the assimilation system, in the sense that it makes assumptions of linearity and Gaussianity, on which all three of the basic algorithms are built, more precarious. It can be shown that assimilation for both balanced and unbalanced states requires accurate representation of the balance relationship within the covariance model. This adds an additional degree of nonlinearity to the assimilation problem which can then be assessed in terms of the difference between the EKF and EnKF. Of course, this paradigm is not completely airtight, since the EnKF incurs sampling error at finite ensemble sizes and also (as will be shown in chapter 2) assumes Gaussian error statistics and hence linear error evolution. We shall also see that, in terms of balance and imbalance, the EnKF incurs new problems related to its Monte Carlo nature.

The EKF/EnKF comparison is then extended to 4DVAR, which has more or less become the practical implementation of linearization-based covariance modeling, but is also a fundamentally different approach to the same. As we shall see, the problem of balanced data assimilation changes in the context of 4DVAR. Comparisons of the Kalman filters to 4DVAR are therefore added for completeness and to build a more solid connection to practical data assimilation. To give 4DVAR as much attention as the Kalman filters would, however, be beyond the scope of the present study, and we thus restrict the analysis to simple numerical comparisons.

The model and assimilation schemes are outlined in chapter 2. In chapter 3 the basic algorithms are compared for a single chaotic timescale. In chapter 4, the comparison is extended to cases where there exists a separation of timescales between relatively fast and slow motions. In chapter 5 we consider the twin problems of cases where there is significant energy in the free fast wave, and cases where the timescale separation between slow and fast modes is unclear. The majority of the analysis of chapter 4 (but without 4DVAR) can be found in Neef et al. [2006], and the analysis of chapter 5 (again without 4DVAR) has been submitted for publication [Neef et al., 2007].

The guiding principle in the numerical experiments is to understand how a given assimilation scheme develops multivariate error covariances which reflect the true dynamics, and where these are improvements over 3D assimilation. It will be shown that the three assimilation methods return very different results, despite the fact that all are designed to compute the best fit between observations and a background estimate. Analytical reasoning and numerical results together verify the expectation that, since observations tell us about balances and balances help us to understand observations, the four-dimensionalization of covariance models is extremely useful if done right.

Several challenges will also be brought to light, and it will be shown that, ultimately, the shortcomings of all three schemes are due to the breakdown of assumptions of linearity and Gaussianity. It is arguable that the effects of nonlinearity in data assimilation may simply be viewed as model error, which can plausibly be accounted for with extra terms. This, and the effects of model error, will be investigated in chapter 6.

Chapter 2

Methodology

Data assimilation is the interaction of three components: a model, a set of observations, and the assimilation algorithm. We will refer to the combination of these as the *data assimilation system*. All three components can range from very simple to very complex, and the success or failure of the full assimilation system depends on the behavior of each component relative to the other two. It is not surprising —and will be shown again in this study— that the outcome of a data assimilation experiment tend s to be more complicated than the sum of its parts. This is why data assimilation studies often begin with low-order models and experiments where the "true state" is simulated, and hence known. Such simulations are attractive because they allow us to isolate specific dynamical phenomena, clearly understand the behavior of a given algorithm in that context, and test the sensitivity of the results to individual assimilation parameters.

To study data assimilation in the context of balanced dynamics, imbalance, and the loss of a balance relationship, we would like a model that admits both a slow mode and a gravity wave mode, allows for initialization on a slow manifold, and is nevertheless simple enough that the assimilated analysis can be easily interpreted in terms of the balanced and unbalanced components of the system. The model should also be physical in the sense that the timescale separation and slow manifold initialization are related to the real governing equations of geophysical fluid dynamics. The model should also be nonlinear, since it is nonlinearity that causes different assimilation methods to produce different results. All of these properties are offered in the model of Lorenz [1986], as modified by Wirosoetisno and Shepherd [2000]. Hereafter we will refer to this model as the "extended" Lorenz 1986 model, or exL86. The model equations and its physical properties are outlined in §2.1, and its derivation is reviewed in appendix A.

There is a wide range of data assimilation algorithms that fall under the heading of "four-dimensional," and the relative virtues and pitfalls of each proposed method are often difficult to disentangle from the details of the specific study within which each is proposed. Instead of evaluating one specific algorithm relative to another, we rather want to understand how a given *type* of assimilation method handles each of the physical contexts described in the introduction, so as to provide a framework for understanding the results of more complex studies. Because the EKF, EnKF and 4DVAR comprise the most basic implementations of 4D assimilation, they will be the focus of this study. Some extension to other algorithms is offered in appendix B.

In order to isolate the effects of assimilation parameters and balance/gravity-wave parameters on the analysis, we begin with so-called identical twin experiments, where the truth and forecast states are evolved with the same "perfect" model, which is also used to generate observations. This has two benefits: (1) the true error statistics for both the observations and the forecasts are known exactly, and (2) the forecast error at observation times comes entirely from accumulated analysis error, allowing us to isolate the relative (dis)advantages of each method. Model error is briefly considered in chapter 6.

2.1 The Extended (1986) Lorenz Model

2.1.1 Basic Equations

The exL86 model is described by the following four equations:

$$\begin{aligned} \frac{d\phi}{dt} &= w' + bz' && (2.1) \\ \frac{dw'}{dt} &= -\frac{C}{2}\sin 2(\phi + \epsilon bx) - \frac{b}{\epsilon\,(1+b^2)}x && (2.2) \\ \frac{dx}{dt} &= \frac{bw' - z'}{\epsilon} && (2.3) \\ \frac{dz'}{dt} &= \frac{x}{\epsilon\,(1+b^2)}. && (2.4) \end{aligned}$$

This system is derived by expansion of the nondimensionalized shallow water equations into a resonant wave triad [Lorenz, 1980], which yields a system that admits three wave number components of three degrees of freedom each, for a total of nine degrees of freedom. This is followed by a truncation of two of the three components, in which the gravity wave solutions are eliminated by setting the amplitude coefficients associated with divergence and geostrophic imbalance, and their time derivatives, to zero [Lorenz, 1986]. This yields a five-equation system consisting of two geostrophic components and one component which exhibits both vortical motion and an inertia-gravity wave. Following the extension of Wirosoetisno and Shepherd [2000, below], this yields the above system, which (when C is time dependent) has two normal modes: a slow vortical mode, and a nearly-linear fast mode with frequency ϵ^{-1}, where $\epsilon \ll 1$. The full derivation of the model in the form [(2.1)-(2.4)] spans four papers [Lorenz 1980, Lorenz 1986, Bokhove and Shepherd 1996, Wirosoetisno and Shepherd 2000], each with different notation. A summary derivation, using the notation of this thesis, is given in appendix A.

The four variables in (2.1)-(2.4) represent spectral coefficients. ϕ is related to the phase between the two coefficients of vorticity for the two truncated (or geostrophic) components in the original triad expansion. w' represents the amplitude of vorticity, z' the geopotential height, and x the divergence of the third (nongeostrophic) component.

The parameter b, which couples the fast and slow normal modes, corresponds to the rotational Froude number of this third triad component, and ϵ, in terms of the parameters of the system, is related to b and the Rossby number of this triad component by (1.2). The fast wave is thus indeed an inertia-gravity wave, with a timescale relative to the vortical mode proportional to the smallness of ϵ. For brevity, we will henceforth simply refer to it as a gravity wave.

The extension of Wirosoetisno and Shepherd [2000] was to give the parameter C an artificial time dependence $C = a_0 + a_1 \cos(\gamma t)$, in order to mimic the presence of additional vortical modes and to ensure that the model's slow mode is chaotic. Except where denoted otherwise (in chapter 6), we will set $a_0 = 1$, $a_1 = 0.8$, and $\gamma = 0.92$, as in Wirosoetisno and Shepherd [2000]. This results in a leading Lyapunov exponent of about $\lambda_1 \simeq 0.15$.

The system can be transformed into normal modes by defining $w \equiv w' + bz'$ and $z \equiv z' - bw'$, which physically correspond to potential vorticity and geostrophic imbalance, respectively. This is done to make the interpretation of the data assimilation experiments more straightforward. The resulting system in normal-mode form is given by

$$\frac{d\phi}{dt} = w \tag{2.5}$$

$$\frac{dw}{dt} = -\frac{C}{2}\sin\left(2\phi + 2\epsilon b x\right) \tag{2.6}$$

$$\frac{dx}{dt} = -\frac{z}{\epsilon} \tag{2.7}$$

$$\frac{dz}{dt} = \frac{x}{\epsilon} + \frac{bC}{2}\sin\left(2\phi + 2\epsilon b x\right), \tag{2.8}$$

and will be used throughout this thesis. Now (2.5)-(2.6) define the dynamics of the slow vortical mode and (2.7)-(2.8) that of the fast mode, and their coupling via ϵ and b becomes clear. This system is qualitatively similar to an elastic pendulum, or the "swinging spring," system studied by Lynch [2002]; (2.5)-(2.6) can be thought of as the slow evolution of the pendulum position, and (2.7)-(2.8) as the elastic spring oscillation. It can be shown that a swinging spring can be initialized such that the extension of the spring is entirely a function of the pendulum position. Likewise, the exL86 model can be initialized such that the fast variables, x and z, are slaved to the evolution of the slow variables, ϕ and w.

2.1.2 Slow Manifold Initialization

The lowest order approximation to a slow manifold in the exL86 system is found by setting $x = z = 0$ and evolving only ϕ and w. For $\epsilon = 0$ or $b = 0$, this manifold is exact, and results in the single-timescale system

$$\frac{d\phi}{dt} = w \tag{2.9}$$

$$\frac{dw}{dt} = -\frac{C}{2}\sin 2\phi. \tag{2.10}$$

Keeping the time dependence of C, this system is analogous to a chaotic pendulum, and corresponds physically to the quasigeostrophic equations [Cushman-Roisin, 1994, chapter 15] in that the fast solution is filtered out of the equations. The system also has some qualitative similarity to the Lorenz [1963] model. To connect this model to similar low-order model studies and to establish how nonlinearity of the slow mode affects data assimilation, experiments with this model are performed in chapter 3.

Higher-order balances can be derived by assuming that x and z are slaved to the slow variables according to slaving relations of the form (1.3), expanding each of the slaving relations in ϵ, and substituting them into the system [(2.5)-(2.8)]. To second order ($n = 2$), the resulting slaving relations are given by

$$x = U_x(\phi;\epsilon) = -\frac{\epsilon}{2}Cb\sin 2\phi + O(\epsilon^3) \tag{2.11}$$

$$z = U_z(\phi, w;\epsilon) = \epsilon^2(Cbw\cos 2\phi + \frac{C'}{2}b\sin 2\phi) + O(\epsilon^3), \tag{2.12}$$

where C' is the time derivative of C [cf. appendix A]. If a state is initialized using slaving relations to some order in ϵ, it will evolve with a free gravity wave of amplitude $\epsilon^{\nu+1}$, where ν is the order of initialization. In this study we will keep $\nu = 2$ throughout. For

$\epsilon = 10^{-1}$, for example, a balanced state will then have a residual free gravity wave of magnitude $\mathcal{O}\left(10^{-3}\right)$.

If the fast variables also contain a free gravity wave (in addition to the slaved components), the full state $\mathbf{x} = (\phi, w, x, z)^{\mathrm{T}}$ can be written as

$$\mathbf{x} = \boldsymbol{f}(\mathbf{y}) + \mathbf{g}, \tag{2.13}$$

where

$$\boldsymbol{f}(\mathbf{y}) = \begin{pmatrix} \phi \\ w \\ U_x(\phi;\epsilon) \\ U_z(\phi, w;\epsilon) \end{pmatrix} \tag{2.14}$$

is the nonlinear mapping from the slow manifold state $\mathbf{y} = (\phi, w)^{\mathrm{T}}$ to the full model state, and

$$\mathbf{g} = \begin{pmatrix} 0 \\ 0 \\ \tilde{x} \\ \tilde{z} \end{pmatrix} \tag{2.15}$$

represents the free gravity wave, with

$$\tilde{x} = x - U_x(\phi;\epsilon) \tag{2.16}$$

$$\tilde{z} = z - U_z(\phi, w;\epsilon). \tag{2.17}$$

The full fast mode (including its slaved component) can also be written in terms of action-angle variables $[I, \theta]$, where

$$I = \sqrt{x^2 + z^2} \tag{2.18}$$

$$\theta = \tan^{-1}\left(\frac{x}{z}\right). \tag{2.19}$$

Referring to (2.11)-(2.12) as the balanced components of the fast mode, we define "imbalance" as the magnitude of the free gravity wave $\tilde{I} \equiv \sqrt{(x - U_x)^2 + (z - U_z)^2} = \sqrt{\tilde{x}^2 + \tilde{z}^2}$. Wirosoetisno and Shepherd [2000] showed that the rate at which $\tilde{I}^2$ grows is approximately given by $\exp(-2.2/\epsilon)$, and so is exponentially slow as $\epsilon \to 0$. The corresponding second Lyapunov exponent, which measures chaotic drift of the fast mode, also decays with decreasing ϵ [Wirosoetisno and Shepherd, 2000].

An example illustrating the slow manifold initialization of the exL86 model is shown in figure 2.1, which shows the four model variables for two example trajectories, both run from the same initial slow-variable state, $(\phi, w)^{\mathrm{T}} = (-6.617, -0.449)^{\mathrm{T}}$, with $\epsilon = 10^{-1}$ and $b = 0.71$. Both states are initialized with (2.11)-(2.12), and a free gravity wave of magnitude $\tilde{I} = 1.5$ is then added to one run (shown in gray). In ϕ and w, the two states are indistinguishable. This shows that the coupling between the balanced and unbalanced modes is quite weak. It can be seen that the slow mode has a characteristic timescale of about 6 time units. For the balance-initialized solution (in black), the unbalanced components of x and z have magnitudes ϵ^3, and are thus not visible in the figure. For the unbalanced solution, the slaved components of x and z are overshadowed by the free gravity wave.

Returning again to the balance transformation (2.14), it is important to note that $\boldsymbol{f}(\mathbf{y})$ contains the slaving relations defined to some order in ϵ, and is thus not invertible. Chapters 4 and 5 will show that the nonlinearity of $\boldsymbol{f}$ is important for data assimilation, because of linearity approximations made in modeling the evolution of error covariances. The relative importance of the nonlinear terms in $\boldsymbol{f}$ depends on the size of ϵ and b. The slaving relationships become more nonlinear for increasing ϵ, where the separation between slow and fast variables becomes concomitantly less well defined. For increasing b (corresponding to motion which is more inertial and less gravitational), the projection of the slow mode onto the fast mode increases, and the slaving relations become more nonlinear while the timescale separation between the two modes stays the same.

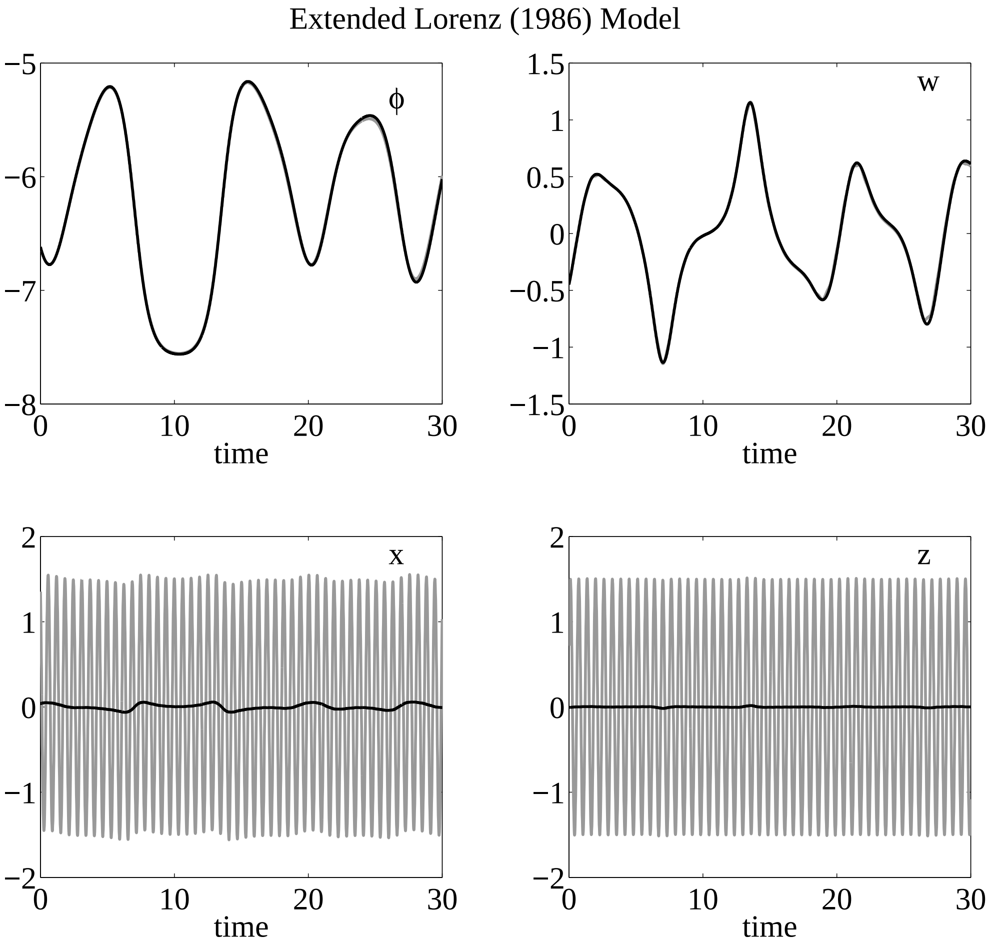

Figure 2.1: Two sample trajectories for the exL86 model, resulting from the same initial slow manifold state. The black curves in each figure represent a run initialized using [(2.11)-(2.12)], and the gray curves a run initialized with $\tilde{I} = 1.5$. In the top panels, the gray curves are essentially coincident with the black curves.

It is important to mention that a free gravity wave in this model neither propagates away nor interacts with other fast waves. This study therefore does not address the effects of adjustment from an unbalanced to a balanced state, such as would happen when gravity waves, excited either by physical mechanisms or data insertion, propagate out of the assimilation region [Daley, 1991, Žagar et al., 2004b]. Instead, our focus is on (1) the excitation of spurious gravity waves due to the assimilation cycle, and (2) the effect of true gravity waves on the recovery of the full model state from observations of the partial state.

The fact that the exL86 model is conservative does not pose a great difficulty, since the intention here is to use it to study assimilation algorithms in the context of the free atmosphere gravity wave / initialization problem, rather than dissipative processes. As pointed out by Lorenz [1986] and Wirosoetisno and Shepherd [2000], dissipation of gravity waves is not the cause of the existence of a slow manifold (rather, it is a matter of the stability of the slow mode). Therefore models such as this one can be quite representative of realistic balance dynamics. Another reason for using a conservative model is that representing dissipation —especially in a low-order model— requires some form of parameterization, which further complicates the assimilation process and yet would be completely arbitrary.

2.2 4D Data Assimilation

2.2.1 The Data Assimilation Problem

In data assimilation, we seek a state which is the best fit between a background or forecast estimate $\mathbf{x}^{\mathrm{b}}$, and a set of observations, given the probability distribution functions (pdfs) of the background estimate and observations, and employing basic estimation theory.

The background or *prior* pdf of the state, $p_{\mathbf{B}}(\mathbf{x})$, represents the distribution of possible values of the state, given all prior modeled and observed information, and has an expected value $\langle p_{\mathbf{B}}(\mathbf{x})\rangle = \mathbf{x}^{\mathrm{b}}$. At some point in time, the background estimate is informed by a vector of observations $\mathbf{z}$, which is associated with a pdf $p_{\mathbf{R}}(\mathbf{z}|\mathbf{x})$, or the conditional pdf of the observations, given the state estimate. Bayes' theorem states that the conditional posterior pdf of the state, given the new observations and the prior pdf, is given by

$$p_{\mathrm{a}}(\mathbf{x}|\mathbf{z}) = \frac{p_{\mathbf{R}}(\mathbf{z}|\mathbf{x})\, p_{\mathbf{B}}(\mathbf{x})}{p(\mathbf{z})}. \tag{2.20}$$

$p(\mathbf{z})$ is the prior pdf of the observation vector, which acts as a normalizing factor and guarantees that the probability of all possible states $\mathbf{x}$ is unity. It is given by

$$p(\mathbf{z}) = \int p_{\mathbf{R}}(\mathbf{z}|\mathbf{x}')\, p_{\mathbf{B}}(\mathbf{x}')\, d\mathbf{x}' \tag{2.21}$$

[e.g. Anderson and Anderson, 1999, Kalnay, 2003].

The simplest way to model the background and observation pdfs is as Gaussian (or normal) distributions,

$$\begin{aligned} p_{\mathbf{B}}(\mathbf{x}) &= \frac{1}{(2\pi)^{n/2}|\mathbf{B}|^{1/2}}\exp\left[-\frac{1}{2}\left(\mathbf{x}^{\mathrm{b}}-\mathbf{x}\right)^{\mathrm{T}}\mathbf{B}^{-1}\left(\mathbf{x}^{\mathrm{b}}-\mathbf{x}\right)\right] && (2.22)\\ p_{\mathbf{R}}(\mathbf{z}|\mathbf{x}) &= \frac{1}{(2\pi)^{m/2}|\mathbf{R}|^{1/2}}\exp\left[-\frac{1}{2}\left[\mathbf{z}-H(\mathbf{x})\right]^{\mathrm{T}}\mathbf{R}^{-1}\left[\mathbf{z}-H(\mathbf{x})\right]\right], && (2.23)\end{aligned}$$

where $\mathbf{R}$ and $\mathbf{B}$ are called the observation and background error covariance matricies (respectively), n and m are the dimensions of the state and observation vector (respectively), and H represents a function which maps the state to the space of observations. The error covariance matricies are defined as

$$\begin{aligned} \mathbf{B} &= \left\langle\left(\mathbf{x}-\mathbf{x}^{\mathrm{b}}\right)\left(\mathbf{x}-\mathbf{x}^{\mathrm{b}}\right)^{\mathrm{T}}\right\rangle && (2.24)\\ \mathbf{R} &= \left\langle\left[\mathbf{z}-H(\mathbf{x})\right]\left[\mathbf{z}-H(\mathbf{x})\right]^{\mathrm{T}}\right\rangle. && (2.25)\end{aligned}$$

Recall from the introduction that error covariances represent not just error amplitudes, but also correlations.

The state which we seek, called the **analysis**, is the most probable state of the joint pdf (2.20). Within the assumption of Gaussian distributions, there are two ways to compute this state. The first way is to compute the estimate $\mathbf{x}$ which minimizes the mean of the square error, called the least-squares or variance-minimizing estimator. This state can be written as a linear combination of the background state and the observation vector, or rather

$$\mathbf{x}^{\mathrm{a}} = \mathbf{x}^{\mathrm{f}} + \mathbf{K}\mathbf{d}, \tag{2.26}$$

where $\mathbf{d} = \mathbf{z} - H(\mathbf{x}^{\mathrm{f}})$ is the observation increment, also called the innovation. The observation increment is multiplied by a matrix of weights $\mathbf{K}$, and to solve the problem we seek the optimum weights which give the analysis $\mathbf{x}^{\mathrm{a}}$ that has the smallest possible posterior error variance. It can be shown [e.g. Kalnay, 2003, chapter 8], that for mutually-uncorrelated observation- and background errors with zero mean values, the gain matrix

$$\mathbf{K} = \mathbf{B}\mathbf{H}^{\mathrm{T}}(\mathbf{H}\mathbf{B}\mathbf{H}^{\mathrm{T}} + \mathbf{R})^{-1} \tag{2.27}$$

minimizes $\langle (\mathbf{x} - \mathbf{x}^{\mathrm{a}})^{\mathrm{T}} (\mathbf{x} - \mathbf{x}^{\mathrm{a}}) \rangle$, such that $\mathbf{x}^{\mathrm{a}}$ is the best linear unbiased estimator (BLUE). For Gaussian prior pdfs, $\mathbf{x}^{\mathrm{a}}$ given by (2.26) is then the conditional mean of the posterior pdf (2.20). Note that $\mathbf{K}$ includes the linearized observation operator $\mathbf{H}$, which maps the innovation from the observation space to the model space.

The interpretation of $\mathbf{K}$ is easiest if one imagines a one-dimensional state with background variance σ^2, and a single observation of the same variable (such that $\mathbf{H} = 1$) with observation error variance σ^2_{obs}. In that case,

$$\mathbf{K} = \frac{\sigma^2}{\sigma^2 + \sigma^2_{\mathrm{obs}}}. \tag{2.28}$$

It can be seen that $\mathbf{K}$ weights the increments by the background error variance, normalized by the total error variance. If $\sigma^2 \gg \sigma^2_{\mathrm{obs}}$, observations are very accurate relative to the background and are hence given maximum weight, with $\mathbf{K} \rightarrow 1$. If the opposite is true, $\mathbf{K} \rightarrow 0$ and the observation is given no weight.

An alternative way of computing the analysis is to seek the state which has the maximum probability according to the posterior pdf (2.20), or its mode. For a Gaussian distribution, this state is of course equal to the variance-minimizing estimate, which is the mean of (2.20). Assuming Gaussian distributions, the log likelihood of the state, given (2.20), is

$$L = -\frac{1}{2}\left(\mathbf{x}-\mathbf{x}^{\mathrm{b}}\right)^{\mathrm{T}}\mathbf{B}^{-1}\left(\mathbf{x}-\mathbf{x}^{\mathrm{b}}\right) - \frac{1}{2}\left(\mathbf{z}-H\left(\mathbf{x}\right)\right)^{\mathrm{T}}\mathbf{R}^{-1}\left(\mathbf{z}-H\left(\mathbf{x}\right)\right). \tag{2.29}$$

We seek the state that maximizes L, which is akin to minimizing the scalar cost function

$$J\left(\mathbf{x}\right) = \frac{1}{2}\left(\mathbf{x}-\mathbf{x}^{\mathrm{b}}\right)^{\mathrm{T}}\mathbf{B}^{-1}\left(\mathbf{x}-\mathbf{x}^{\mathrm{b}}\right) + \frac{1}{2}\left[\mathbf{z}-H\left(\mathbf{x}\right)\right]^{\mathrm{T}}\mathbf{R}^{-1}\left[\mathbf{z}-H\left(\mathbf{x}\right)\right]. \tag{2.30}$$

J measures the misfit between the analysis and the background state, normalized by the background error covariance matrix, and the analysis and the observation vector, normalized by the observation error covariance matrix. Thus the maximum likelihood estimate is the state $\mathbf{x}$ which minimizes J.

The two solutions given by the minimum variance estimate and the maximum likelihood estimate lead to two types of data assimilation. The solution resulting from the minimization of J over three-dimensional space is called 3D variational data assimilation, or 3DVAR. The minimum variance estimate given by (2.26) with the optimal gain matrix (2.32) is called Optimal Interpolation, or OI [Bergman 1979]. As just shown, 3DVAR and OI return the same estimate, given Gaussian pdfs and given the same specification of covariance matrices, but differ in general. The historical differentiation between the two methods is related to approximations used to solve for $\mathbf{x}^{\mathrm{a}}$ using (2.26)-(2.32) versus using (2.29), for practical applications where $\mathbf{z} \sim \mathcal{O}\left(10^5\right)$ and $\mathbf{x} \sim \mathcal{O}\left(10^8\right)$.

Given a series of observations made in time, **forecast** states $\mathbf{x}^{\mathrm{f}}$ can now be generated by integrating a model forward until the time at which an observation is made. The model evolution can be written as

$$\mathbf{x}^{\mathrm{f}}_{k+1} = \mathcal{M}_k\left(\mathbf{x}^{\mathrm{a}}_k\right) + \mathbf{q}_k, \tag{2.31}$$

where $\mathcal{M}_k$ represents the forward evolution of the truth, $\mathbf{q}_k$ represents model error, $\mathbf{x}^{\mathrm{a}}_k$ represents the analysis made at time k, and $\mathbf{x}^{\mathrm{f}}_{k+1}$ is the forecast based on that analysis.

This forecast then becomes the background estimate when an observation is made, and can be updated according to either 3DVAR or OI. (Henceforth we will use the term "forecast" to mean the background state when there is a cycling of analysis step and model integrations, and "background" otherwise.) This update step is called the analysis step, for obvious reasons. Following the analysis step, the analysis state $\mathbf{x}_k^{\mathrm{a}}$ is evolved forward to the next observation time, using (2.31).

The analysis is thus cycled with the forward model evolution (2.31) to produce a series of forecasts and analyses that are intended to sequentially come closer to the true state. OI and 3DVAR are called three-dimensional assimilation algorithms because the error covariance matrix $\mathbf{B}$ is defined in space (spectral or physical) and over the model variables, but is static in time.

Different methods exist for estimating this matrix, a common one being the "NMC method" [Parrish and Derber, 1992], wherein error magnitudes are estimated from differences between forecasts made from different initial times. Error magnitudes are, of course, only one piece of the puzzle; a full covariance model also requires some assumption about covariance structures, or correlations. Spatial covariances in OI are typically assumed to be homogeneous, or dependent only on the relative displacement between two points (rather than the locations of the points themselves), and spatial correlations are typically assumed to be spatially isotropic functions which decay with distance (e.g. Daley 1991).

In 4D data assimilation the estimation of covariance structure and amplitudes is extended to the time dimension, by modeling not just the evolution of the state in time, but also the evolution of the pdf (2.22). If (2.22) is Gaussian, only the evolution of its mean and variance needs to be modeled. There are two ways to do this: sequentially, by evolving the covariance matrix explicitly and updating it at analysis times, or variationally, by minimizing the cost function over a time window.

2.2.2 Extended Kalman Filter

The Kalman filter can be thought of as the four-dimensional extension of OI. In the full Kalman filter [Kalman, 1960, Kalman and Bucy, 1961, Miller et al., 1994], $\mathbf{B}$ is replaced with the time dependent *forecast error covariance matrix* $\mathbf{P}_k^{\mathrm{f}}$, which is developed by evolving error covariances forward in time using the forecast model (2.31). At analysis times, $\mathbf{P}_k^{\mathrm{f}}$ is updated to reflect the reduction of error, and the information gleaned about physical correlations, by the observations. For nonlinear models (and/or nonlinear observation operators), both of these steps depend on higher order moments of the forecast error pdf, and require a closure approximation to be made. The most basic ways to do this for nonlinear models are the Extended Kalman Filter (EKF) and Ensemble Kalman Filter (EnKF).

Algorithm

The EKF algorithm [Ghil et al., 1981, Cohn and Parrish, 1999, Miller et al., 1994, Nerger et al., 2005] is as follows. The minimum-variance analysis step at timestep k is

$$\mathbf{K}_k = \mathbf{P}_k^{\mathrm{f}}\mathbf{H}_k^{\mathrm{T}}(\mathbf{H}_k\mathbf{P}_k^{\mathrm{f}}\mathbf{H}_k^{\mathrm{T}} + \mathbf{R}_k)^{-1}. \tag{2.32}$$

$$\mathbf{x}_k^{\mathrm{a}} = \mathbf{x}_k^{\mathrm{f}} + \mathbf{K}_k\left[\mathbf{z}_k - H_k(\mathbf{x}_k^{\mathrm{f}})\right]. \tag{2.33}$$

The resulting *analysis error covariance matrix* $\mathbf{P}_k^{\mathrm{a}}$ is then computed, and follows from application of the expectation operator to the analysis (2.33):

$$\begin{aligned} \mathbf{P}_k^{\mathrm{a}} &= \langle(\mathbf{x}_k^{\mathrm{a}} - \langle\mathbf{x}_k^{\mathrm{a}}\rangle)\,(\mathbf{x}_k^{\mathrm{a}} - \langle\mathbf{x}_k^{\mathrm{a}}\rangle)^{\mathrm{T}}\rangle & (2.34)\\ &= (\mathbf{I}_n - \mathbf{K}_k\mathbf{H}_k)\mathbf{P}_k^{\mathrm{f}}(\mathbf{I}_n - \mathbf{K}_k\mathbf{H}_k)^{\mathrm{T}} + \mathbf{K}_k\mathbf{R}_k\mathbf{K}_k^{\mathrm{T}}, & (2.35)\end{aligned}$$

where $\mathbf{I}_n$ is the $n \times n$ identity matrix. For the optimal gain matrix (2.32), (2.35) reduces to

$$\mathbf{P}_k^{\mathrm{a}} = (\mathbf{I} - \mathbf{K}_k\mathbf{H}_k)\mathbf{P}_k^{\mathrm{f}} \tag{2.36}$$

[e.g. Daley, 1991, chapter 4, Kalnay, 2003, chapter 5].

This error covariance estimate is then evolved forward in time by linearizing the forecast model about the forecast state at each time. Defining the error vector at time step $k+1$ as

$$\mathbf{e}^{\mathrm{f}}_{k+1} \equiv \mathbf{x}^{\mathrm{f}}_{k+1} - \mathbf{x}^{\mathrm{t}}_{k+1} \tag{2.37}$$

(where $\mathbf{x}^{\mathrm{t}}_{k+1}$ is the truth at time step $k+1$), then substituting (2.31), the forecast error at time step k can be approximated as a Taylor expansion of the model about the previous time step:

$$\begin{aligned} \mathbf{e}^{\mathrm{f}}_{k+1} &= \mathcal{M}_k(\mathbf{x}^{\mathrm{a}}_k) - \mathcal{M}_k(\mathbf{x}^{\mathrm{t}}_k) + \mathbf{q}_k & (2.38) \\ &\simeq \mathbf{M}_k \mathbf{e}^{\mathrm{a}}_k + \mathbf{q}_k. & (2.39) \end{aligned}$$

Here $\mathbf{M}_k = \partial\mathcal{M}_k(\mathbf{x}^{\mathrm{a}}_k)/\partial\mathbf{x}$ is called the tangent-linear model (TLM) about the analysis state at time step k. The forecast error covariance matrix is then found by multiplying (2.39) by its transpose and computing the expectation value, which gives

$$\mathbf{P}^{\mathrm{f}}_{k+1} = \mathbf{M}_k \mathbf{P}^{\mathrm{a}}_k \mathbf{M}^{\mathrm{T}}_k + \mathbf{Q}_k, \tag{2.40}$$

where $\langle \mathbf{q}_k \mathbf{q}^{\mathrm{T}}_k \rangle \equiv \mathbf{Q}_k$ is the estimated covariance matrix of model error. Thus, both correlations and error magnitudes are evolved in the EKF by cycling information between the model [via (2.31)] and observations [via (2.33)]. We refer to this cycling as the EKF covariance model.

The above algorithm still lacks an initial-guess error covariance matrix $\mathbf{P}^{\mathrm{f}}_0$. This is a somewhat arbitrary component of the algorithm. Since $\mathbf{P}^{\mathrm{f}}_k$ will be adjusted from observational information, it is usually assumed that the EKF will become insensitive to the initial covariance matrix formulation after a few analysis cycles. However, we will see in chapters 3-5 that this is not always the case.

Limitations of the EKF

Several practical difficulties make the EKF infeasible for practical implementation [Miller et al., 1994]. First, the EKF is extremely computationally expensive for realistic systems, because it requires one forward evolution of the TLM for each model coordinate, and also because it involves a matrix inversion at analysis times (2.32). For contemporary practical systems with state dimension $n \sim 10^8$, these operations are prohibitively expensive. Formulation of the initial covariance matrix presents a second difficulty, because it requires some sort of initial dynamical assumption about error relationships.

Third, forward evolution of $\mathbf{P}^{\mathrm{f}}_k$ by the TLM (2.40) neglects third- and higher order moments in the forecast error covariance field. The ability of the TLM to estimate the correct forecast error pdf over the interval of time between two observations depends both on the accumulated analysis error preceding an observation, and on the nonlinearity of the model evolution between observations. The repeated addition of observations and forward evolution of the state estimate can make (2.40) a more viable assumption, but can also amplify errors in estimated covariances (relative to true error statistics). The effects of errors in estimated covariances are discussed more explicitly in chapter 3. Note that underestimation of error variance means that the adjustment of covariances in (2.36) will also be underestimated, which will result in the eventual rejection of accurate observations. This phenomenon is referred to as filter divergence. It can be shown that both underestimation of variances and overestimation of correlations will eventually lead to filter divergence [Houtekamer and Mitchell, 1998, Hamill et al., 2001]. Fourth, as will be shown in chapters 4 and 5, the EKF has serious limitations in terms of balance and gravity waves.

Despite these shortcomings and the unlikelihood of its realistic implementation, the EKF is still worth examining, because of its theoretical similarity to the EnKF and 4DVAR in the limit of a perfect assimilation system. A second reason to study the EKF is that it is possible to formulate simplified, more efficient "error subspace" versions of the EKF [Dee, 1991, Nerger et al., 2005, Hoteit et al., 2005], which could make implementation of the EKF more feasible in the future.

2.2.3 Ensemble Kalman Filter

The EnKF was proposed by Evensen [1994] as a solution to the first three of the four problems listed above for the EKF. In the EnKF, the TLM-estimated evolution of error covariances (2.40) and the covariance analysis step (2.36) are replaced with a Monte Carlo estimate, using an ensemble of model states. By abandoning the TLM, the EnKF is able to preserve some nonlinearity in the evolution of error statistics. The EnKF, while often referred to as an approximation to the full Kalman filter, is therefore also a nonlinear extension of the Kalman filter.

Algorithm

Several variants of the standard EnKF exist in the literature [Evensen, 2003]. We will focus on the basic, perturbed-observation EnKF, as introduced by Evensen [1994] and modified by Burgers et al. [1998] and Houtekamer and Mitchell [1998], as a first step to illustrating the EnKF's balance properties.

In the EnKF, $\mathbf{P}_k^{\mathrm{f}}$ is computed from an ensemble of perturbations about a central forecast, and its evolution is approximated by the evolving ensemble statistics as

$$\begin{aligned}\mathbf{P}_{k+1}^{\mathrm{f}} &= \left\langle \left(\mathbf{x}_{i,k+1}^{\mathrm{f}} - \langle \mathbf{x}_{i,k+1}^{\mathrm{f}} \rangle\right) \left(\mathbf{x}_{i,k+1}^{\mathrm{f}} - \langle \mathbf{x}_{i,k+1}^{\mathrm{f}} \rangle\right)^{\mathrm{T}} \right\rangle && (2.41)\\ &= \frac{1}{N-1} \sum_{i=1}^{N} \left(\mathbf{x}_{i,k+1}^{\mathrm{f}} - \langle \mathbf{x}_{i,k+1}^{\mathrm{f}} \rangle\right) \left(\mathbf{x}_{i,k+1}^{\mathrm{f}} - \langle \mathbf{x}_{i,k+1}^{\mathrm{f}} \rangle\right)^{\mathrm{T}}, && (2.42)\end{aligned}$$

where $\langle \cdot \rangle$ now denotes the ensemble average. For the covariance matrix, the normalization factor in the ensemble average operation is $N-1$, to account for the fact that the ensemble mean is not independent of the ensemble members [Bevington and Robinson, 1992]. The i^{th} member in the ensemble is given by

$$\mathbf{x}_{i,k+1}^{\mathrm{f}} = \mathcal{M}_k \left(\mathbf{x}_{i,k}^{\mathrm{a}}\right) + \mathbf{q}_{i,k}, \tag{2.43}$$

where $\mathbf{q}_{i,k}$ represents the model error of the perturbed state.

At observation times, each ensemble member is adjusted according to

$$\mathbf{x}^{\mathrm{a}}_{i,k} = \mathbf{x}^{\mathrm{f}}_{i,k} + \mathbf{K}_k \mathbf{d}_{i,k} \tag{2.44}$$

$$\mathbf{d}_{i,k} = \mathbf{z}_k - H_k(\mathbf{x}^{\mathrm{f}}_k) \tag{2.45}$$

$$\mathbf{z}_{i,k} = \mathbf{z}_k + \mathbf{r}_{i,k}. \tag{2.46}$$

The random perturbations $\mathbf{r}_{i,k}$ represent observation error, and are chosen from a normal distribution with estimated observation error variance σ^2_{obs}. It is important to randomly perturb the observations, as pointed out by Burgers et al. [1998] and Houtekamer and Mitchell [1998], who show that bias is otherwise introduced into the assimilation (note that papers which discuss the EnKF prior to the studies of Burgers et al. [1998] and Houtekamer and Mitchell [1998] actually present a faulty algorithm).

The EnKF analysis is then simply the ensemble mean:

$$\mathbf{x}^{\mathrm{a}}_k = \langle \mathbf{x}^{\mathrm{a}}_{i,k} \rangle, \tag{2.47}$$

with analysis error covariance matrix

$$\mathbf{P}^{\mathrm{a}}_k = \frac{1}{N-1} \sum_{i=1}^{N} \left(\mathbf{x}^{\mathrm{a}}_{i,k} - \langle \mathbf{x}^{\mathrm{a}}_{i,k} \rangle \right) \left(\mathbf{x}^{\mathrm{a}}_{i,k} - \langle \mathbf{x}^{\mathrm{a}}_{i,k} \rangle \right)^{\mathrm{T}}. \tag{2.48}$$

Between observation times, we choose the analysis state to be the forecast based on the mean analysis at the last observation time —as opposed to the mean forecast, which tends to lack the full variability of the nonlinear model evolution. The distinction isn't always made clear in published descriptions of the EnKF algorithm, since the two would be equal for a linear model.

Advantages of the EnKF

By evolving the ensemble, the EnKF preserves nonlinearity in the evolution of forecast error statistics, though it still retains the assumption that error pdfs are characterizable by their mean and variance, and that the linear update (2.47) is optimal. Note that as it is conceivable that the BLUE may be an effective estimate even for non-Gaussian pdfs. We shall see below that this is indeed often the case. The EnKF also eliminates the cost of developing a TLM or (in contrast to 4DVAR assimilation) an adjoint model. Instead, each integration of an N-member EnKF requires N model integrations. Another advantage of the EnKF is that it outputs an analysis ensemble, which can subsequently be used for ensemble forecasts.

Limitations of the EnKF

The accuracy of the EnKF rests on the assumption that a finite-size forecast ensemble, and its forward evolution, capture the true error statistics between observation times and following the analysis step. Since the EnKF analysis is a weakly-nonlinear combination of model states, it is in theory constrained to the space spanned by the N-member ensemble. Smallness of the ensemble also limits the amount of information that can be brought into the assimilation before the problem becomes overdetermined [Lorenc, 2003b]. [If covariances are localized, as is commonly done (e.g. Hamill et al. 2001), this is no longer the case. In this case, however, the constraint of the assimilating model is weakened.]

For operational models, a computationally feasible EnKF typically has an ensemble size which is very small relative to the state dimension, which can be as large as $n \sim \mathcal{O}(10^8)$. Sampling error in the EnKF scales as $N^{-1/2}$. However, as pointed out by Houtekamer and Mitchell [2005], other error sources (such as non-Gaussianity of the ensemble, observation error, error in the formulation of observation error statistics, etc.) become more important as ensemble sampling error decreases. Hence, there is typically a “saturation” ensemble size, beyond which analysis error cannot be reduced by adding more ensemble members. This is illustrated briefly in chapter 3. Houtekamer and Mitchell [2005] show that, for a typical NWP global model, errors stop decreasing for more than about $N = 50$ ensemble members.

Because (2.36) is a nonlinear step (nonlinear because $\mathbf{K}_k$ contains $\mathbf{P}_k^{\mathrm{f}}$), ensemble sampling error leads to a biased ensemble distribution of error variances. Hamill et al. [2001] point out that the probability of underestimating variances is larger than that of overestimating variances. Ensemble sampling error also gives rise to spurious correlations, especially where error variances are small, leading to filter divergence. Another weakness of the EnKF is that forecast ensembles can become non-Gaussian [Lorenc, 2003b], in which case the mean and covariances estimated from the ensemble may not be good representations of the system statistics.

The accuracy of a finite-size ensemble, given these limitations, can also be justified if sufficient information from observations is brought into the data assimilation system. Other, practical, limitations and complications concerning the EnKF are discussed by Lorenc [2003b].

2.2.4 Four-dimensional Variational Assimilation (4DVAR)

Algorithm

4DVAR is the time-dependent extension of 3DVAR. In 4DVAR, the scalar cost function is not evaluated sequentially but rather is minimized over a time window $\Delta T = [t_0, t_0 + T]$, and has the form

$$J\left(\mathbf{x}_0\right) = \frac{1}{2}\left(\mathbf{x}_0 - \mathbf{x}_0^{\mathrm{b}}\right)^{\mathrm{T}} \mathbf{B}^{-1}\left(\mathbf{x}_0 - \mathbf{x}_0^{\mathrm{b}}\right) + \frac{1}{2}\sum_{i=0}^{N}\left[\mathbf{z}_i - H\left(\mathbf{x}_i\right)\right]^{\mathrm{T}} \mathbf{R}^{-1}\left[\mathbf{z}_i - H\left(\mathbf{x}_i\right)\right]. \quad (2.49)$$

The control vector in the minimization of J is now the state estimate at the beginning of the assimilation window, $\mathbf{x}_0 = \mathbf{x}(t_0)$. The first term of $J\left(\mathbf{x}_0\right)$ describes the misfit between $\mathbf{x}_0$ and a background estimate of the state at the initial time, $\mathbf{x}_0^{\mathrm{b}}$ (here we use the term *background* instead of *forecast* to reflect the fact that the update is not sequential in time). The second term describes the misfit between a series of observation vectors $\mathbf{z}_i$ made over the assimilation window, and the state estimate at observation times, which is given by $\mathbf{x}_k = \prod_{j=1}^{k} \mathcal{M}_j\left(\mathbf{x}_0\right)$. In effect, 4DVAR solves for the initial condition that results in an integrated state which best fits the observations over the time window, subject to $\mathbf{x}_0^{\mathrm{b}}$ and the model. The background matrix $\mathbf{B}$ corresponds to $\mathbf{P}_k^{\mathrm{f}}$ in the Kalman filter equations, though we again talk about background error covariances, since $\mathbf{B}$ is not explicitly evolved in time. The control vector could also be a transformed vector for which the corresponding covariance matrix is simpler to formulate or easier to invert [e.g. Lorenc 2003a].

Different methods of minimizing $J(\mathbf{x}_0)$ are possible. Minimization requires computation of the gradient of $J(\mathbf{x}_0)$ with respect to $\mathbf{x}_0$. The gradient is given by

$$\left[\frac{\partial J}{\partial \mathbf{x}_0}\right] = \mathbf{B}^{-1}\left(\mathbf{x}_0 - \mathbf{x}_0^b\right) - \sum_{i=0}^{N} \mathbf{M}_0^* \mathbf{M}_1^* ... \mathbf{M}_i^* \mathbf{H}_i^{\mathrm{T}} \mathbf{R}_i^{-1} \left[\mathbf{z}_i - H(\mathbf{x}_i)\right]. \tag{2.50}$$

$\mathbf{M}^*$ is the adjoint of the TLM, and is usually referred to simply as the adjoint model. The adjoint model is defined as

$$\langle \mathbf{M}_k \mathbf{a}, \mathbf{b} \rangle = \langle \mathbf{a}, \mathbf{M}_k^* \mathbf{b} \rangle \tag{2.51}$$

and relates the gradient of J with respect to the state at observation times to the gradient with respect to the initial state [Errico, 1997]. The minimization of J is eventually cut off at some threshold (see the comment regarding this in §2.3.3), and the resulting estimate is then the analysis trajectory over the time window. Here we use the conjugate-gradient method [Shewchuk, 1994] to perform the minimization.

Advantages of 4DVAR and Connection to the Kalman Filter

4DVAR differs from the EKF for two main reasons. First, 4DVAR integrates the full model state at each iteration of the cost function minimization. This means that 4DVAR preserves some nonlinearity in the estimation of flow-dependent covariances, and is thus similar to the EnKF, even though the details of how the fitting is done are of course quite different.

A second difference between 4DVAR and both Kalman filters is that the Kalman filter analysis at each time uses only information from analyses at previous times, whereas a

4DVAR analysis uses information from both past and future analyses. It is possible to extend the standard Kalman filter to use both past and future observations, in the so-called Kalman Smoother [e.g. Ménard and Daley, 1996], by combining a forward Kalman filter with a backwards-running Kalman filter, then smoothing the combined analysis. In the limit of linear dynamics and Gaussian errors, the final-time analysis produced by 4DVAR is equivalent to the final-time analysis of a Kalman Smoother which has been running indefinitely [Ménard and Daley, 1996, Lorenc, 2003a, Fisher et al., 2005], and Kalman Smoother experiments can be done in order to better understand the implicit modeling of covariances in 4DVAR [e.g. Lorenc, 2003a, Fisher et al., 2005]. Since the Kalman Smoother is even more expensive to implement and more complicated than the EKF, in this study we will limit the analysis to the EKF and EnKF.

Limitations of 4DVAR

The formulation of (2.49) makes the model a strong constraint on the minimization, meaning that the final minimization iterate has to be a solution admitted by the model. As in the EKF and EnKF, model error makes the 4DVAR estimate suboptimal. If the model is not perfect, the minimization of J will not really return the best estimate, i.e. the maximum of the posterior distribution.

By using the adjoint of the TLM to relate the gradient of the cost function at observation times to the gradient at the initial time, 4DVAR also assumes linear evolution of errors. This will be problematic if the assimilation period is longer than the time of validity of the TLM. A solution is to divide the total assimilation time into windows over which the TLM and its adjoint are valid approximations, and perform a separate 4DVAR analysis over each window. This means that each minimization only uses a subset of the total observation set. Consequently, there is typically an optimal assimilation window length, where the positive effect of more observations balances out the negative effect of model nonlinearity and chaotic dynamics. The optimal assimilation window depends on the timescales of motion being modeled. This is discussed further in chapters 4 and 5.

2.2.5 Comparison of the Algorithms, Effective Nonlinearity, and Practical Considerations

All three of the algorithms presented above seek the most likely state of the conditional pdf (2.20), assuming Gaussian and mutually uncorrelated background and observation error distributions. In the EKF and EnKF, evolution of the covariance model is explicit, whereas in 4DVAR, evolution of the covariance model is implicit. The assumption of Gaussianity implies an assumption of linearity. For a linear model and Guassian errors, the three methods are equivalent.

In the face of model nonlinearity and non-Gaussianity of error statistics, the results produced by each method can be quite different. One simple reason for the difference is that linearizations (of either the model or the observation operator) are made around different intermediate states. Moreover, the cost function minimization in 4DVAR is performed over finite time windows, and hence uses different sets of observations. The EnKF differs fundamentally from the EKF and 4DVAR in that it uses a Monte Carlo approach to approximate the evolution of covariances, thus trading linearization error for ensemble sampling error. Lorenc [2003b] therefore argues that, in practice, the EnKF represents the error distribution less precisely than the TLM-based 4DVAR and EKF (because of sampling error), but over longer timescales (because of nonlinearity preserved in the error evolution).

Errors in covariance estimates, even in the case of a perfect model, come from error in the TLM approximation (EKF and 4DVAR), ensemble sampling error (EnKF), inadequate formulation of the initial covariance field (EKF and 4DVAR), inadequate formulation of the initial ensemble (EnKF), and, in all three algorithms, observation sampling error (meaning the difference between actual and assumed observation error variance, over a finite number of observations), misspecification of observation error statistics, and nonlinearity of the observation operator. Hence, the nonlinearity of the estimated covariance model is actually a complex result of model nonlinearity and all other assimilation parameters which control the validity of TLM-approximations in the EKF and 4DVAR, and Gaussianity of errors in all three algorithms.

Since validity of the TLM approximation and Gaussianity of the ensemble both depend on the distance between truth and the background / forecast states, and the likelihood of the EKF and EnKF to diverge or 4DVAR to converge to a local (non-absolute) cost function minimum, all increase as the time between observations increases. It is possible to test the time of validity of the TLM, by comparing the difference between two integrations of the nonlinear model (a reference state and a perturbed state) to the TLM integration of the perturbation over the time window. If the relative error between the nonlinearly-evolved perturbation and the tangent-linearly-evolved perturbation is small over some interval, i.e. if for a perturbation α,

$$R = \frac{||\mathcal{M}\left(\mathbf{x}+\alpha\right) - \mathcal{M}\left(\mathbf{x}\right) - \mathbf{M}\alpha||}{||\mathbf{M}\alpha||} \ll 1, \tag{2.52}$$

then the TLM is valid over that interval. Figure 2.2 shows the ratio R in time for a range of initial perturbations, averaged over 15 realizations of the single-timescale model [(2.9)-(2.10)]. It can be seen that, for perturbations on the order of the initial perturbations used in our experiments ($\delta\mathbf{x} \sim 10^{-0.3}$) the relative error R ceases to be small by about 3 or 4 time units, and possibly much faster for errors larger than that.

Another source of assimilation error, which will not be addressed here, involves the time interpolation of observations, which is done when all the observations made in a certain time window are assumed to be valid at the analysis time [e.g. Houtekamer and Mitchell, 2005].

Note that, even for nonlinear/non-Gaussian error evolution, the linearity and Gaussianity assumptions made in each algorithm can be justified if the observational information is dense enough (in time and space) that small-error approximations are valid.

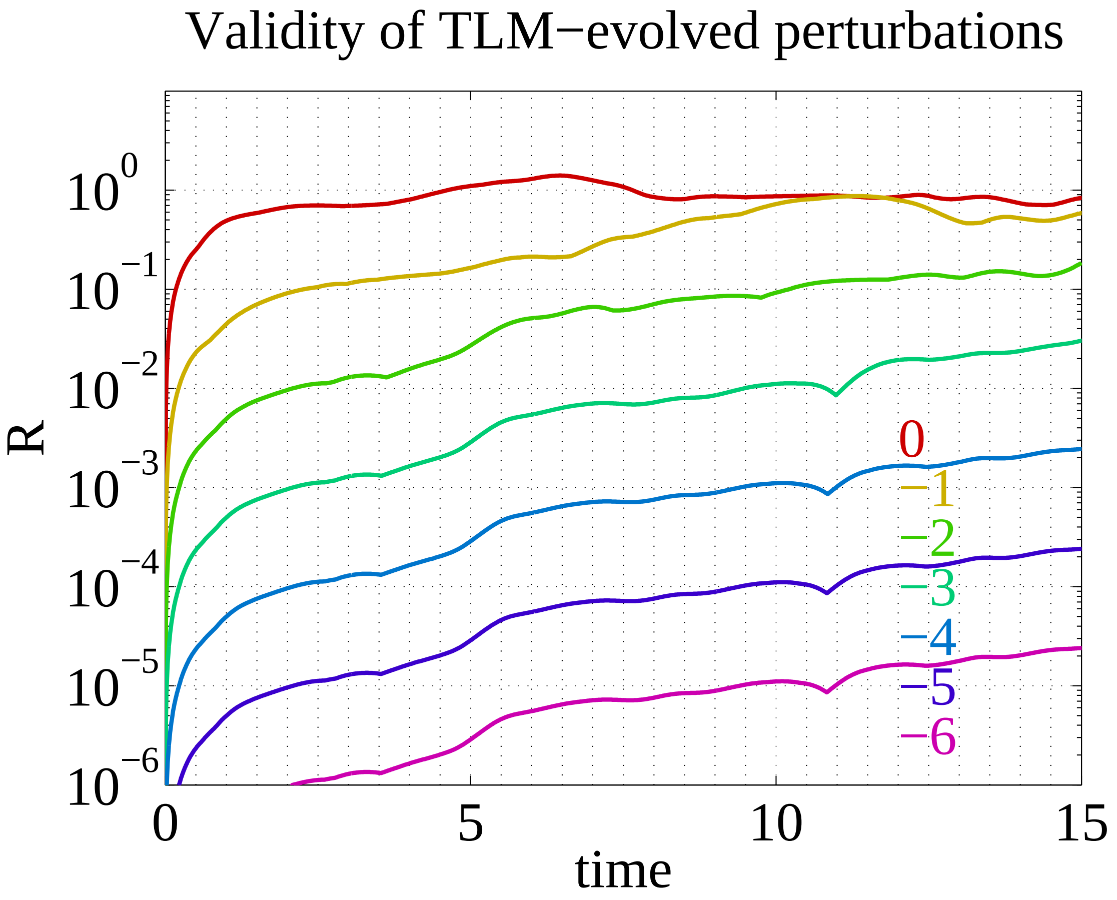

Figure 2.2: The TLM-validity ratio R (2.52) in time, for 6 different initial perturbations, the base-10 logarithm of which is indicated in the figure. Each curve represents an average over 15 random realizations of the single-timescale model [(2.9)-(2.10)].

Following Verlaan and Heemink [2001], we use the phrase "nonlinearity of the assimilation system" to encompass all the factors which control the validity of the linearity/Gaussianity assumptions in each method: frequency and density of observations, the minimization window (in 4DVAR), and of course the nonlinearity of the model. In chapter 4, it will be shown that nonlinearity of the balance relationship (or curvature of the balance manifold, as in fig. 1.3) further increases this effective nonlinearity.

The three assimilation methods also differ from one another in computational efficiency. The computational cost of both Kalman filters increases rapidly with the dimension of the problem, in the EKF because each degree of freedom requires an iteration of the TLM, and in the EnKF because more ensemble members are required to make the problem well-posed. Since 4DVAR is iterative, its cost is related mostly to the complexity of the cost function, which does not necessarily increase with state dimension. Thus, the EKF and EnKF tend to be cheaper for low-order models, but become comparable to 4DVAR for more complex models. Since the exL86 model has only 4 degrees of freedom, the EKF turns out to be the most efficient algorithm, with a 10-member EnKF requiring about twice as much computational time. Due to the high nonlinearity of the exL86 model, 4DVAR is roughly 4 times slower, and statistical results shown for 4DVAR in the next three chapters will therefore typically be averages over smaller numbers of experiments than EKF/EnKF experiments.

4DVAR has been implemented operationally at ECMWF, Météo France, the UK Met Office, Japan, and at the Canadian Meteorological Centre [Rabier, 2005]. Use of the EnKF is still largely in a discussion and testing phase [Houtekamer and Mitchell, 2005, Lorenc, 2003b], though it was implemented operationally at the Canadian Meteorological Centre in 2005 [Houtekamer and Mitchell, 2005]. The EnKF and 4DVAR are currently considered roughly similar in accuracy and ease of implementation [Lorenc 2003b, Houtekamer and Mitchell 2005]. Some practical advantages and disadvantages of the EnKF and 4DVAR are discussed in more detail by Lorenc [2003b].

2.3 Experiments

2.3.1 Experimental Outline

For reasons stated above, we begin with so-called identical twin experiments, where the forecast and truth are evolved with the same "perfect" model, and extend the analysis to the consideration of model error in chapter 6. Twin experiments address only the unrealizable case where the dynamics are completely understood, but this isolates the effects of errors in the data assimilation system.

The procedure for each assimilation experiments is as follows. A "truth" is generated by choosing initial values for the slow variables randomly from $0 \leq \phi^{\mathrm{t}} \leq 2\pi$ (a uniform distribution) and $w^{\mathrm{t}} \sim \mathcal{N}(0, 0.5^2)$ (a Gaussian distribution with variance 0.25). The full state $\mathbf{x}^t$ is then computed using (2.13), and the free gravity wave component $\mathbf{g}^{\mathrm{t}}$ which is added to the truth is given by

$$\tilde{x}^{\mathrm{t}} = \tilde{I}^{\mathrm{t}} \cos\theta^{\mathrm{t}} \tag{2.53}$$

$$\tilde{z}^{\mathrm{t}} = \tilde{I}^{\mathrm{t}} \sin\theta^{\mathrm{t}}, \tag{2.54}$$

with a prescribed free gravity wave magnitude $\tilde{I}^{\mathrm{t}}$, and θ^{t} chosen randomly from the unit circle. In chapter 4, the truth will be generated with $\tilde{I}^{\mathrm{t}} = 0$. In chapter 5, the truth will be generated with $\tilde{I}^{\mathrm{t}} > 0$, with the reference value being $\tilde{I}^{\mathrm{t}} = 1.5$. In all experiments shown in this study, the system [(2.5)-(2.8)] is integrated using a 4th-order Runge-Kutta solver with timestep $\Delta t = 0.01$. Unless noted otherwise, the reference values of the timescale separation and coupling parameters will be $\epsilon = 10^{-1}$ and $b = 0.71$.

The initial forecast in each experiment is generated by perturbing the slow component of the truth:

$$\mathbf{y}_0^{\mathrm{f}} = \mathbf{y}_0^{\mathrm{t}} + \delta\mathbf{y}_0, \tag{2.55}$$

where $\delta\mathbf{y}_0$ represents random vectors chosen from normal distributions $\mathcal{N}(0, \sigma_0^2 \mathbf{I}_2)$, where $\mathbf{I}_2$ represents the 2×2 identity matrix. The full forecast state $\mathbf{x}_0^{\mathrm{f}}$ is then computed using

(2.13), with a free gravity wave with magnitude $\tilde{I}^{\mathrm{f}}$. In most cases (unless indicated otherwise) we set $\tilde{I}^{\mathrm{f}} = 0$. In the EKF and EnKF, $\mathbf{x}_0^{\mathrm{f}}$ is integrated forward to observation times. In 4DVAR experiments, $\mathbf{x}_0^{\mathrm{f}}$ is integrated to the end of the assimilation period to form $\mathbf{x}^{\mathrm{b}}(t)$, the initial-guess background state.

2.3.2 Observations

Observations are generated at time intervals Δt^{obs}, as

$$\mathbf{z}_k = \mathbf{H}\mathbf{x}_k^t + \nu_k, \tag{2.56}$$

with observation error realizations chosen as random vectors from normal distributions, $\nu_k \sim \mathcal{N}(0, \mathbf{R})$, where $\mathbf{R} = \sigma_{\mathrm{obs}}^2 \mathbf{I}_m$ and $\sigma_{\mathrm{obs}}^2 = 0.25$ is the prescribed observation error variance. m represents the number of observed variables. All experiments are run with the correct estimation of observation error statistics, that is, with $\mathbf{R}$ in the assimilation equations equal to the true observation error covariance matrix. We drop the time subscript on $\mathbf{R}$, since the observation error covariances do not change in time in these experiments.

Generally two types of observations are compared: observations of the entire slow state

$$\mathbf{z}_{\mathrm{FIL}} = \begin{pmatrix} \phi^{\mathrm{obs}} \\ w^{\mathrm{obs}} \end{pmatrix}, \tag{2.57}$$

and observations where the slow w is exchanged for the mixed variable $w' \equiv (w - bz) \,/\, (1 + b^2)$:

$$\mathbf{z}_{\mathrm{MIX}} = \begin{pmatrix} \phi^{\mathrm{obs}} \\ w'^{\mathrm{obs}} \end{pmatrix}. \tag{2.58}$$

The former can be thought of as analogous to perfectly-filtered observations which contain no gravity wave signal, and will hence be abbreviated as FIL (though, of course, ϕ and w simply represent slow normal modes, not filtered states). The latter will be referred to as “mixed-timescale” observations, and abbreviated as MIX. The variable ϕ is observed in both cases to avoid a problem where the EnKF has a tendency to become multimodal, with groups of ensemble members clustered around harmonics of ϕ^t, if the analysis is unconstrained by observations of ϕ (see chapter 3).

In either observation case, the observation operator H is linear, and will hereafter be written as $\mathbf{H}$. Here we also drop the subscript, since the observation operator also does not change in time in our experiments. Even though the Kalman filter and 4DVAR equations can easily be genaralized to nonlinear observation operators, the assumptions of Gaussian errors throughout imply linear observation operators.

2.3.3 Comments on the Implementation of Each Algorithm

Model Error in Kalman Filter Experiments

In all EKF experiments, we set $\mathbf{Q}_k = 0$. Likewise, in all EnKF experiments outside of chapter 6 we set $\mathbf{q}_{i,k} = 0$. This makes the EKF and EnKF analogous to the strong constraint 4DVAR assimilation outlined in §2.2.4. We note in passing that (2.40) assumes that $\mathbf{e}_k^{\mathrm{a}}$ and $\mathbf{q}_k$ are uncorrelated, which implies that the $\mathbf{q}_k$ themselves are sequentially uncorrelated. This can be an extreme assumption, since systematic model errors are practically unavoidable and result not just from model shortcomings but also from errors in the data assimilation system [Dee, 2005]. Simulation of these errors will be discussed in more detail in chapter 6.

Time Windows in 4DVAR

In 4DVAR experiments, separate cost function minimizations are performed for each window, which will result in discontinuities in the analyses at the boundaries of neighboring windows. It is possible to smooth over such discontinuities [Daley, 1991, chapter 13], perhaps by adding continuity as an extra constraint in the minimization. In favour of simplicity, however, this extra constraint was not investigated.

It is desirable to use the analysis resulting from the minimization of one time window as the initial background iterate for the subsequent minimization window, which would increase continuity and improve the 4DVAR analysis as the assimilation progresses. However, in this study we do not reinitialize the forecast at each time window. Since the background error covariance matrix $\mathbf{B}$ is based on the background error field at $t = 0$, assimilation windows after ΔT_1 will have misestimated background error statistics if the forecasts are not reinitialized, and 4DVAR experiments will therefore have a deterioration of the covariance model in time. As we shall see in subsequent chapters, letting the background error covariance model deteriorate in time tells us more about what happens to 4DVAR, in terms of balance and gravity waves, when the assimilation is made increasingly flawed.

Cost Function Minimization in 4DVAR

To make 4DVAR computationally feasible for experiments over many realizations, the conjugate-gradient minimization is forced to cut off when the RMS w-error over the window drops below some threshold, or when 5 iterations have been performed. Unless stated otherwise, we place the cutoff threshold at $\langle e_w^2 \rangle_{\Delta T}^{1/2} = 0.2$, where $\langle \cdot \rangle_{\Delta T}$ signifies the average over the time window. In individual 4DVAR examples, the conjugate-gradient minimization is run out to as many iterations as are needed to illustrate the given point.

2.3.4 Measures of Assimilation Accuracy

RMS errors by components

The accuracy of each experiment is measured by the RMS true error (truth minus analysis), divided into the three components:

$$\begin{aligned} e_{\mathbf{s},\mathrm{rms}} &= \langle (\mathbf{s}^{\mathrm{t}} - \mathbf{s}^{\mathrm{a}})^{\mathrm{T}} (\mathbf{s}^{\mathrm{t}} - \mathbf{s}^{\mathrm{a}}) \rangle_T^{1/2} && (2.59) \\ e_{I,\mathrm{rms}} &= \langle (I^{\mathrm{t}} - I^{\mathrm{a}})^2 \rangle_T^{1/2} && (2.60) \\ e_{\theta,\mathrm{rms}} &= \langle (\theta^{\mathrm{t}} - \theta^{\mathrm{a}})^2 \rangle_T^{1/2}. && (2.61) \end{aligned}$$

The brackets $\langle \cdot \rangle_T$ indicate the average over all timesteps of an assimilation period of T time units. Unless noted otherwise, we set $T = 30$.

$e_{\mathbf{s},\mathrm{rms}}$ represents RMS error in the slow mode, which is defined as $\mathbf{s} = (u, v, w)^{\mathrm{T}}$. $u = \sqrt{C} \sin \phi$ and $v = \sqrt{C} \cos \phi$ are variables from the original Lorenz [1986] model formulation which are of the same scale as w (A). $e_{I,\mathrm{rms}}$ is RMS error in the magnitude of the full fast mode, including both the slaved and free gravity wave components. $e_{\theta,\mathrm{rms}}$ is RMS error in the phase of the gravity wave, and will become relevant in chapter 5, where unbalanced true states are considered.

Numerical results shown, excepting single examples, are averages of the RMS errors [(2.59)-(2.61)] over N assimilation runs, denoted by the subscripted brackets $\langle \cdot \rangle_N$. For EKF and EnKF experiments, true values of the error measures given above will also be compared to the following approximate quantities which are estimated within the assimilation cycle:

$$\begin{aligned} \sigma_{\mathbf{s}}^2 &= C \sigma_\phi^2 + \sigma_w^2 && (2.62) \\ \sigma_I^2 &= \sigma_x^2 + \sigma_z^2. && (2.63) \end{aligned}$$

Note that $\sigma_{\mathbf{s}}^2$ and σ_I^2 to not correspond exactly to the respective error variances, $\langle (\mathbf{s}^{\mathrm{t}} - \mathbf{s}^{\mathrm{a}})^2 \rangle$ and $\langle (I^{\mathrm{t}} - I)^2 \rangle$, which are actually complex nonlinear quantities. It was realized after the experiments were completed that a more accurate comparison between estimated and actual errors would be to compare, say, $\langle e_x^2 \rangle$ to σ_x^2, or $\langle e_w^2 \rangle$ to σ_w^2. We found, however, that the differences between the quantities and [(2.62)-(2.63)] and [(2.59)-(2.60)] were qualitatively similar to the differences between individual estimated and true errors, such that the results shown in the figures (specifically, figs. 4.3, 4.4, 4.10, 4.11, 5.1, 5.2, 5.8, and 5.9) do not change significantly.

Chapter 3

4D Data Assimilation for a Single Nonlinear Timescale

Recovery of the full state from partial observations is the crux of the overall data assimilation problem, and will form the basis of the more specific problems of data assimilation for balanced and unbalanced states in subsequent chapters. In the previous chapter, we saw that the solution to this problem lies in the development of an accurate covariance model, which is developed sequentially in the Kalman filter and implicitly in 4DVAR. All three of the methods presented in the previous chapter assume Gaussianity and therefore linearity, and can hence return poor estimates if nonlinearity of the data assimilation system (which is a function of observational coverage, background/forecast error, model nonlinearity, and, to be neglected until chapter 6, systematic model error) is too high.

To separate the issues of balance and gravity waves from that of dynamical nonlinearity and chaos, we first examine the three 4D algorithms and the 3D method OI, for the single-timescale [(2.9)-(2.10)] system. With only two variables, this system presents a simple environment in which to establish how each method responds as nonlinearity in the assimilation system increases, and how the three methods differ in various assimilation regimes.

3.1 Examples

Using the notation established in chapter 2, the forecast/background error covariance matrix for the single-timescale model can be written as

$$\mathbf{P}^{\mathrm{f}}, \mathbf{B} = \begin{pmatrix} \sigma_\phi^2 & c_{\phi w} \\ c_{\phi w} & \sigma_w^2 \end{pmatrix}. \tag{3.1}$$

If only one variable is observed, recovery of the full state $(\phi, w)^{\mathrm{T}}$ depends upon how well each scheme can capture the covariance $c_{\phi w}$. Three example assimilation experiments in figure 3.1 illustrate how this is done in the EKF, EnKF, and 4DVAR, for an experiment where observations of w are assimilated into the single-timescale system every $\Delta t^{\mathrm{obs}} = 6$ time units. This means that observations are made roughly once during a typical cycle of the slow mode, and beyond the time of validity of the TLM (fig. 2.2). The top two rows of plots compare the true state $\mathbf{y}^{\mathrm{t}} = (\phi, w)^{\mathrm{t}}$ to the analysis state $\mathbf{y}^{\mathrm{a}} = (\phi, w)^{\mathrm{a}}$, for the EKF [column (A)], EnKF [column (B)], and 4DVAR [column (C)]. Analyses of ϕ are shown in the first row of plots and analyses of w in the second row of plots. A 10-member ensemble is used for the EnKF; the ensemble is shown for each variable along with the ensemble-mean analysis. For the EKF and 4DVAR examples, the initial forecast error covariance matrix and background error covariance matrix, respectively, are estimated by a diagonal matrix,

$$\mathbf{P}_0^{\mathrm{f}} = \mathbf{B} = \sigma_0^2 \mathbf{I}_2, \tag{3.2}$$

thus assuming no initial correlation between ϕ and w. In 4DVAR, the minimization is performed over windows of $\Delta T = 10$ time units. Figure 3.1 depicts a case where the assimilation is successful for all three schemes, though the final analysis is different across the three cases.

For the EKF and EnKF, good analysis increments in the unobserved variable ϕ are made when growth in $c_{\phi w}$ reflects the rate at which the forecast is diverging from the

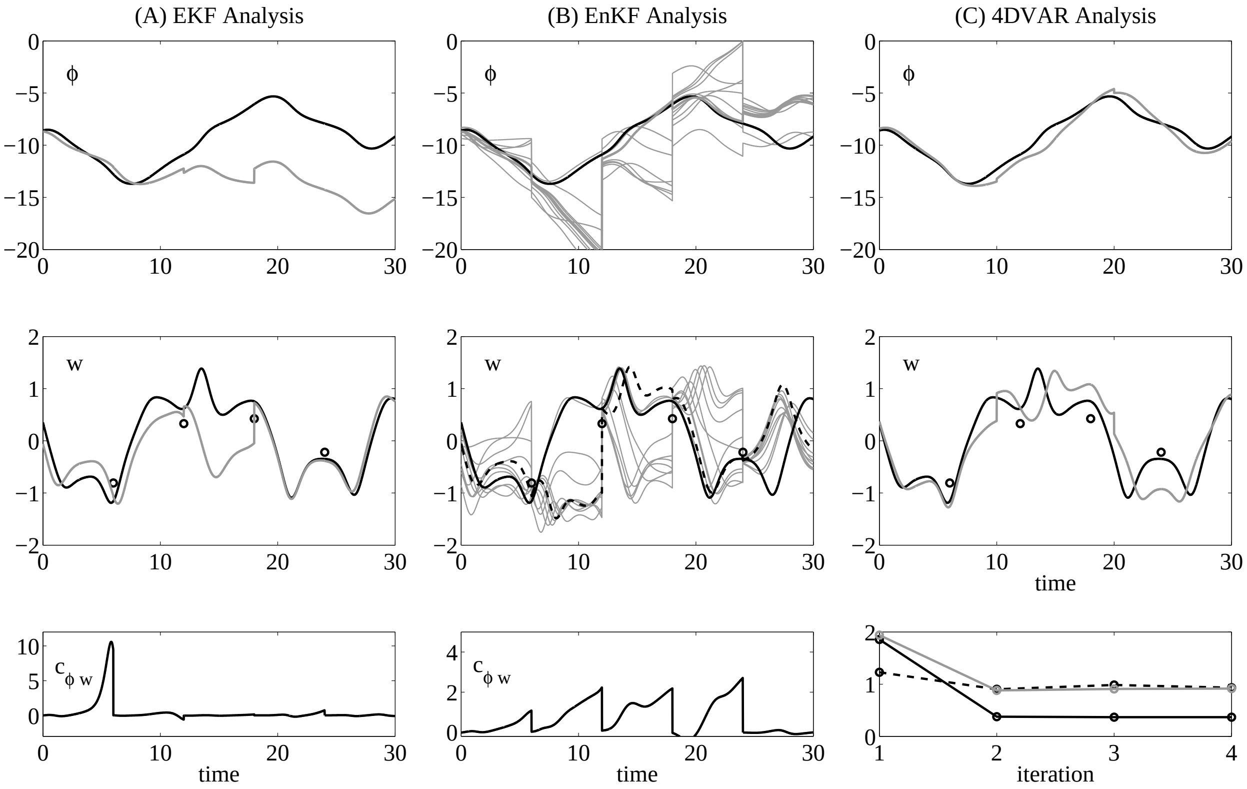

Figure 3.1: Three assimilation experiments, with the same true state and comparing the three basic algorithms: (A) EKF, (B) a 10-member EnKF, and (C) 4DVAR. Each column shows the truth (solid) and analysis (gray) for ϕ (top) and w (center). Plots of w also show the observations (circles). For the EnKF (B), the ensemble is shown in gray and the mean by a black dashed line. The bottommost plots in columns (A) and (B) also shows the estimated covariance between the two model variables, $c_{\phi w}$. The bottom plot for 4DVAR (C) shows the reduction of slow mode error (2.59) as a function of minimization iteration, for the three time windows (black solid: ΔT_1, black dashed: ΔT_2, gray solid: ΔT_3).

true state, as well as the correlation between errors in ϕ and w. The bottommost plots in columns (A) and (B) thus also show the estimated forecast error covariance $c_{\phi w}$. In column (A) it can be seen that the EKF analysis of w (which is observed) is pushed towards the truth at observation times, and the analysis of ϕ is only offset by 2π. Preceding the first observation at $t = 6$, $c_{\phi w}$ indeed grows as the forecast and true state diverge, then decreases sharply when an observation is made. Column (B) shows the analyses of ϕ and w if the EnKF is instead used to assimilate the same set of observations. The analyses of both variables are similar in quality, and growth in $c_{\phi w}$ reflects the commensurate spread in the ensemble. Note, however, that the increasing non-Gaussinaity of the ensemble in time indicates that the assimilation system is becoming increasingly nonlinear as time progresses, with the forecast of the mean analysis becoming quite different from the mean forecast.

For 4DVAR, accurate recovery of ϕ depends on how well the time series of observations and the adjoint model constrain the cost function minimization. Column (C) shows the 4DVAR analysis for the same state and observations. The bottommost plot in this column shows the RMS slow mode error (2.59) as a function of the minimization iteration, for each of the three time windows. The analysis is very close to the truth for the first time window ($\Delta T_1 = [0, 10]$), and for this window the cost function minimum is found within the first iteration. For the other two time windows (where the initial background state is a very poor estimate of the truth, relative to specified background error covariances), the minimization settles into states which are farther from the truth, but still comparable to the corresponding Kalman filter estimates.

Figure 3.1 shows an important property, which will be illustrated in the context of the full model in subsequent chapters: even though a perfect model and the same random number realizations were used in each case, the three algorithms give three rather different analyses. As the assimilation problem is made more difficult, the assumptions made within each algorithm, and which cause the differences in the analyses, can be made to break down. Experiments in the rest of this chapter will test this in a statistical sense for each method.

3.2 Divergence of the Kalman Filter

In the Kalman filter, the underestimation of analysis errors relative to true errors, and the consequent rejection of future observations [e.g. column (A) of figure 3.1], is the phenomenon referred to as filter divergence. Figure 3.1 shows how filter divergence can result specifically from the four-dimensionalization of the assimilation: because of error in the estimated covariance evolution, covariances are overreduced in the covariance analysis steps (2.36) and (2.48), and subsequent estimated error growth is too small.

Hamill et al. [2001] give a nice discussion of how filter divergence happens in the EnKF, and show that either underestimation of forecast error variances or overestimation of correlations will lead to underestimation of analysis error variances. Overestimation of variances means that more weight is given to observations, which results in a greater adjustment of the covariance matrix towards the correct form in (2.36). If error variances are underestimated, observations are not given enough weight, resulting in analysis increments that are too weak in all directions, as well as insufficient adjustment of the forecast error covariances. Subsequent forecast error variances will then only grow to the size of the real innovation if the forecast happens to pass through a region of large error growth [Miller et al., 1994]. If forecast error variances are estimated correctly but *correlations* are overestimated, unobserved components of the state vector will be overadjusted during the analysis step, resulting in a possible breakdown of the TLM approximation.

Why does the Kalman filter tend to underestimate errors, instead of overestimating them? The reason for this can be understood most simply by considering the expression for the analysis error covariance matrix (2.36). $\mathbf{P}_k^{\mathrm{a}}$ estimated in this way gives the minimum analysis error variances possible, given the forecast and observation error variances.

It is only correct, however, if $\mathbf{K}_k$ is truly the optimal gain. If $\mathbf{K}_k$ is not the optimal gain, the posterior error variance is not really minimized, and hence is larger than what is estimated by $\mathbf{P}_k^{\mathrm{a}}$ [cf. Daley 1991, chapter 4].

A quantitative comparison of the EKF and EnKF is shown in figure 3.2, which shows average true analysis errors computed immediately following the insertion of each observation, for the EKF (left panel) and the EnKF (right panel). Average errors are computed for $\Delta t^{\mathrm{obs}} = 1, 2, 3, 4, 5, 6$ and 7, and over 600 experiments for each case, and the observed variable in all experiments is w. For the EnKF, errors are shown for ensembles of 4 members (the size of the state vector), and 10 members (2.5 times the size of the state vector). Comparison of the two panels reveals that the performance of the EnKF is overall similar to the EKF if the forecast ensemble consists of 4 forecasts, though error is lower in the EKF for $\Delta t^{\mathrm{obs}} = 1$ and 2. Here the TLM approximation is valid, with EKF errors decreasing in time for $\Delta t^{\mathrm{obs}} = 1$. Average analysis error in the EnKF is reduced when the ensemble size is increased to 10 forecasts, with the greatest reduction for more frequent observations, where error variances are small and estimated correlations therefore more critical.

Note that analysis errors grow in time for all cases shown, except for $\Delta t^{\mathrm{obs}} = 1$ for the EKF, and $\Delta t^{\mathrm{obs}} = 1$ and 2 for the 10-member EnKF, indicating that divergence of the Kalman filters will eventually happen for this model for all but the most ideal cases. Since there is no model error in these experiments and initial error statistics are estimated correctly, this divergence is entirely due to nonlinearity and sampling error.

Nonlinearity of the assimilation system does not just depend on the observation frequency, but also the configuration of observation variables. This is examined in figures 3.3 for the EKF, and figures 3.4-3.5 for the EnKF. In each figure, three observation scenarios are compared: observations of ϕ only, observations of w only, and observations of both ϕ and w, each panel comparing rms true (2.59) and estimated errors (2.62) across a range of observation intervals. The errors are also compared to average analysis errors

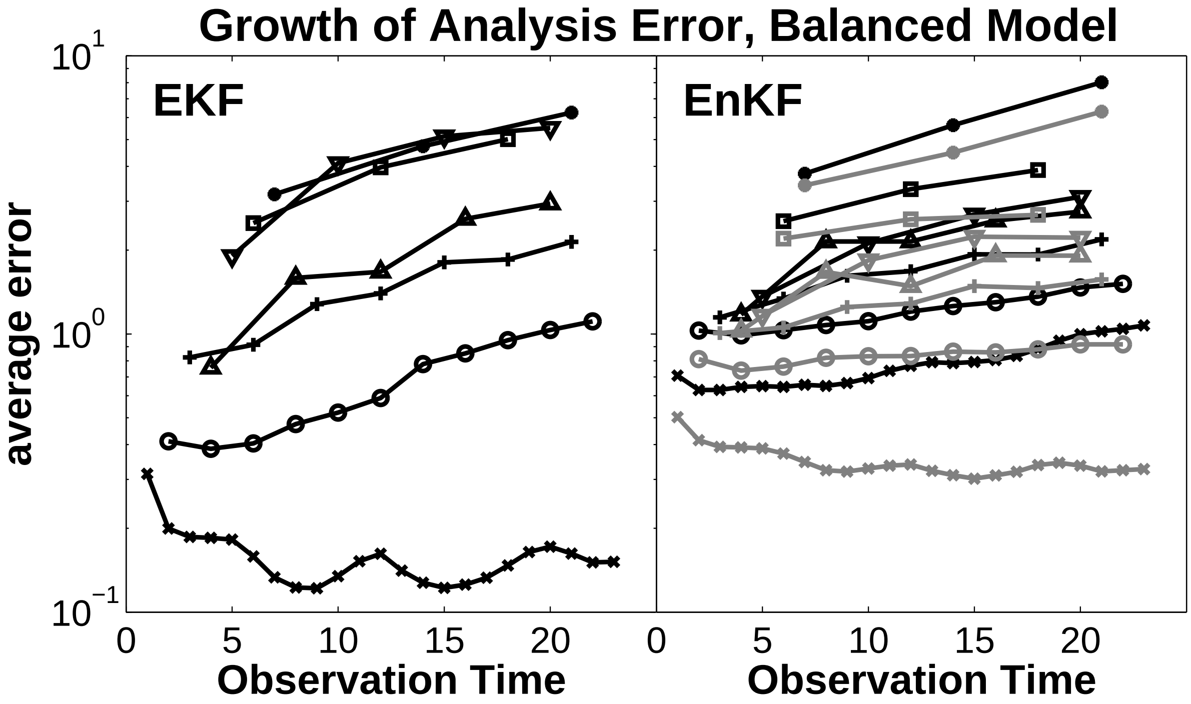

Figure 3.2: Comparison of the average state error (over 600 runs) immediately following the analysis steps, comparing the EKF (left) and EnKF (right), for the single-timescale model. For the EnKF, we also compare ensemble sizes $N = 4$ (black) and $N = 10$ (gray). Observation frequencies shown are: $\Delta t^{\text{obs}} = 1$ (x's), 2 (circles), 3 (+'s), 4 (triangles), 5 (inverted triangles), 6 (squares), and 7 (dots).

from similar experiments performed using OI, with the diagonal covariance matrix (3.2).

Let us define no-skill levels for each error component as the average difference that we would find between two randomly chosen states with no data assimilation. For the initial conditions chosen in these experiments, w oscillates chaotically between about ± 1.3 and has a climatological variance of about $\sigma_w^2 \simeq 0.45$. u and v oscillate chaotically between $\pm\sqrt{C_{\max}}$ and have climatological variances of about $\sigma_u^2 = \sigma_v^2 = 0.6$. Therefore the RMS no-skill error level for the nonlinear slow mode (the average error that we would have for no observations) is roughly

$$e_{\mathbf{s},\mathrm{NS}} \;=\; \left(\sigma_u^2 + \sigma_v^2 + \sigma_w^2\right)^{1/2} \simeq 1.3. \tag{3.3}$$

For observations of ϕ (a), EKF true errors reach the no-skill level for $\Delta t^{\mathrm{obs}} \simeq 4$, and *exceed* no-skill errors for observation intervals longer than that. The EKF also underestimates analysis errors for $\Delta t^{\mathrm{obs}} \gtrsim 2$, and exceeds corresponding OI errors. This indicates that the validity of the TLM-evolution of forecast errors in the EKF breaks down around $\Delta t^{\mathrm{obs}} = 2$ in this observation case, leading to overreduction of error variances. What actually happens in this observation case is this: as error in the estimated covariance $c_{\phi w}$ increases, it becomes more likely for the observation of ϕ to push w^a very far from the truth, sometimes into orbits where w oscillates around a value that is outside of the initialized range (which is why average errors exceed the no-skill level), even if ϕ^{a} is pushed closer to the truth. The entire forecast state then deviates much more from the truth between observations, and the linearization of the TLM around this forecast state becomes very inaccurate.

If instead only w is observed (b), the average true errors change significantly, remaining lower than corresponding OI errors for all Δt^{obs}, and is actually less than the predicted errors for $\Delta t^{\mathrm{obs}} > 4$. It is to be expected that average slow error should increase in this observation case, since observations of w contain less information about the slow mode than observations of ϕ. The reason why errors nevertheless *decrease* in the EKF is because ϕ is a phase, and slow error saturates: if w is well-observed and has negligible

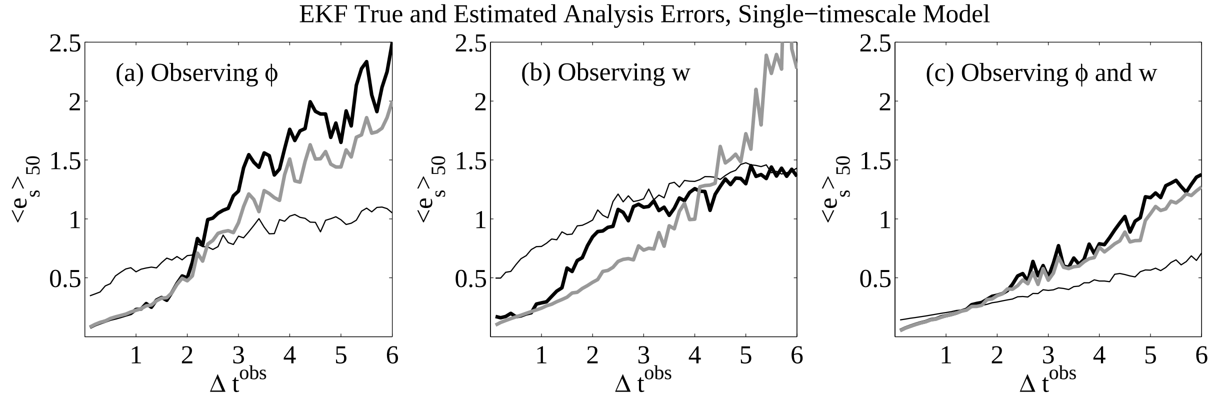

Figure 3.3: Comparison of average true errors (thick, black) compared to average estimated errors (thick, gray) for the EKF as a function of the observation interval. Three observation configurations are compared: (a) observing only ϕ, (b) observing only w, and (c) observing both variables. Each case is also compared to corresponding average true errors for a set of OI experiments (thin, black).

error while ϕ is not observed, slow mode error saturates around $e_{\mathbf{s},\mathrm{rms}} \simeq 1.3$.

Finally, if the full state is observed (c), the EKF improves overall, but actually becomes considerably worse relative OI for $\Delta t^{\mathrm{obs}} \gtrsim 2$. In this case, observational information content is high, and the recovery of either variable depends mostly on the variances σ_w^2 and σ_ϕ^2. Consequently, the neglect of covariance terms in the OI covariance matrix only makes a small difference and because variances are small, the difference between the EKF and OI then becomes more pronounced. This indicates that the relative impact of noise in estimated correlations is greater if variances are smaller, as argued by Hamill et al. [2001]. Thus, linearization around the background state remains a problem even when the information content is high.

Figure 3.4 shows the same set of experiments as figure 3.3, but for the EnKF with a (large) 15-member ensemble. In figure 3.5, EnKF analysis errors are shown as a function of ensemble size, while keeping a constant $\Delta t^{\mathrm{obs}} = 4$. For observations of ϕ [fig. 3.4 (a)], the EnKF tends to *over*estimate errors for $\Delta t^{\mathrm{obs}} \gtrsim 2$, in contrast to the corresponding

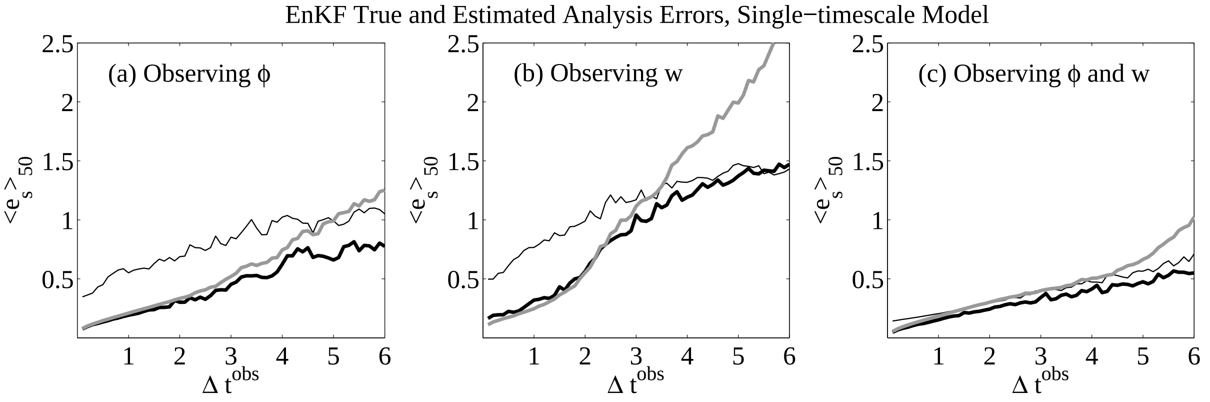

Figure 3.4: As in figure 3.3, but for the EnKF with 15 ensemble members. For comparison, average errors are again compared to average OI errors (thin, black).

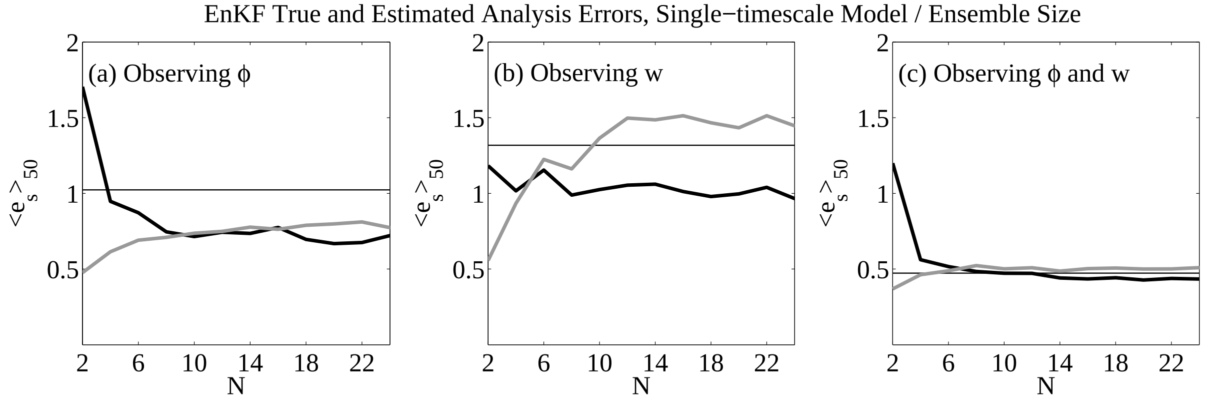

Figure 3.5: Comparison of true average EnKF analysis errors (thick, black) and estimated average analysis errors (thick, gray) as a function of ensemble size. The observation configurations being compared are as in previous figures. The thin line in each plot indicates the corresponding average OI errors.

EKF case [fig. 3.3 (a)]. Here, ensemble averaging tends to keep the mean analysis of w closer to the truth than it is in the EKF analysis, and even when ensemble members are harshly forced to a very far orbit, comparison to the rest of the ensemble in the analysis step [(2.44)-(2.47)] usually brings a far outlier back at the next observation time. Analysis increments are consequently more stable and the EnKF yields lower analysis errors than OI, even for a 3-member ensemble [fig. 3.5 (a)].

For observations of w (b), true errors now increase relative to OI. The reason again has to do with the fact that the unobserved variable is a phase. Since observations of w constrain only $\cos\phi$, it can happen that some ensemble members take on harmonics of the true value of ϕ, but remain a few π-multiples removed from ϕ^{t}. In these cases, $\cos\phi^{\mathrm{a}}$ may be quite close to $\cos\phi^{\mathrm{t}}$ — in which case error in w will grow more slowly than error in ϕ. Since w^{f} may be very close to observations, analysis increments will then be small, and the analysis increments for ϕ will underestimate actual error in ϕ. The forecast ensemble can then become multimodal, with groups of ensemble members clustered around harmonics of the true ϕ value, and unconstrained by observations. Consequently, ensemble-estimated statistics become less accurate, leading to average slow mode error which is actually similar to that in the EKF. Though ϕ in this model physically corresponds to potential vorticity, this result has implications for other phases that might be computed in an assimilation, such as wind direction; if the phase variable is not observed, the ensemble can easily become multimodal. Comparison to figure 3.5 (b) shows that errors in this case cannot be greatly improved by increasing ensemble size. Finally, for observations of both variables (c), the EnKF is consistently —but only slightly— better than OI, and remains so even for very large ensembles [fig. 3.5 (c)].

Thus, the validity of the TLM approximation, or the representativeness of the ensemble, depends not only on how much of the state is observed, but also on the nature of the observed variables. The relative value of different observation types also differs for each particular assimilation method. If the unobserved component is not terribly stable (as in the case of ϕ observations), the EKF can become pathological if Δt^{obs} is longer than the period of validity of the TLM approximation. On the other hand, if the unobserved component of the state is a phase, the EnKF can become multimodal. If very much of the state is observed (in this case, both variables), then a static covariance model may be sufficient, or even preferable, to a 4D covariance model.

3.3 Comparison of the Kalman Filters to 4DVAR

The previous section showed that TLM-derived covariances of the EKF are unreliable at large observation intervals, but still more useful than static covariances as long as instability (in this case, the state being forced into a far-away orbit in w) can be controlled. 4DVAR can be viewed as one way to control this instability, because the state estimate is updated using the full model dynamics at each minimization iteration, and because observational information is carried both forward and backward in time.

Figure 3.6 compares performances of the EKF, 15-member EnKF, OI, and 4DVAR as a function of observation interval, for observations of ϕ only (a), w only (b), and both variables (c). In these experiments the minimization iterations cut off at 5 iterations, with no cutoff threshold applied. The most striking thing about this figure is that 4DVAR has much lower assimilation errors than both Kalman filters, for all three observation cases, at large Δt^{obs}. Thus the two added components of 4DVAR —that information is carried in both directions in time and that the model is integrated at each minimization iterate— lend stability in this nonlinear model experiment. The stability of 4DVAR relative to the EKF, for increasing observation intervals, has also been noted in a low-order model context by Fisher et al. [2005]. The accuracy of 4DVAR relative to the EnKF when only w is observed (b) suggests that 4DVAR also overcomes the non-Gaussianity problems found in the EnKF at large Δt^{obs}.

On the other hand, 4DVAR has larger assimilation errors for $\Delta t^{\text{obs}} \lesssim 1$, especially in

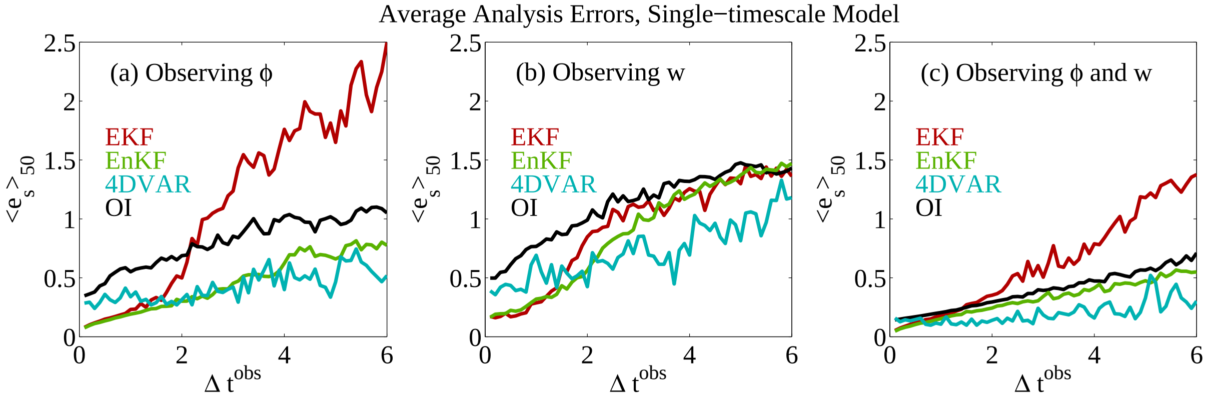

Figure 3.6: Average true errors (2.59) as a function of observation interval, comparing the EKF (red), 15-member EnKF (green), 4DVAR (blue) and OI (black). Three observation configurations are again compared: (a) observing ϕ only, (b) observing w only, and (c) observing both variables.

the two cases where only the partial state is observed [(a) and (b)]. This is a consequence of the minimization cutting off at the maximum 5 iterations. For $\Delta t^{\mathrm{obs}} \lesssim 1$, the cost function is more constrained by observations, and the minimization requires more iterations to reach the minimum. Thus, to be computationally feasible in these cases, 4DVAR is unable to perform as well as the Kalman filters if observations are very frequent.

3.4 Summary

The single-timescale experiments in this chapter highlighted the relative strengths and weaknesses for each algorithm, and their differences when nonlinearity in the assimilation system is increased.

Divergence and instability of the EKF (fig. 3.3) at observation intervals which exceed the timescales over which the TLM is a valid approximation, was also pointed out by Miller et al. [1994] and Nerger et al. [2005], and will be revisited in more detail in subsequent chapters. Note that the TLM test (2.52) shows that the validity time for the TLM is around 5 or 6 time units. Validity time for the TLM is much less in the EKF because the EKF also adjusts covariances in (2.34), which exacerbates filter divergence. In the subsequent two chapters, the TLM will be used to approximate not just nonlinear dynamics, but also the nonlinear balance relationship, and we shall see that this presents an additional source of filter instability.

The EnKF suffers from sampling error, but was found to be stable relative to OI even for small ensemble sizes. However, the ensemble can easily become non-Gaussian if, for example, the phase variable ϕ is not observed. Similar examples of ensemble non-Gaussianity at large Δt^{obs} are also shown by Anderson and Anderson [1999] in the context of the Lorenz [1963] 3-component model. In that study, a possible remedy for the error caused by non-Gaussianity is proposed, in the form of a so-called kernel or particle filter. This method is visited briefly in appendix B. If another timescale is introduced into the model, the assumption of Gaussianity must hold for both timescales. Moreover, the nonlinear balance relationship will have to be captured by the finite ensemble.

4DVAR in these experiments improved upon both problems: it is stable at large Δt^{obs}, and does not require an ensemble. On the other hand, its accuracy when observations are frequent is limited by the minimization iteration cutoff, which is in place to make the algorithm computationally feasible; we shall see that this property also affects the assimilation of balanced and unbalanced states.

Chapter 4

Balance and Excitation of Spurious Gravity Waves

4.1 Balance in the Assimilation Problem

In this chapter the classic problem of spurious excitation of gravity waves is cast into the context of 4D assimilation. The question we ask is how well the three basic 4D methods, by naturally evolving error covariances, are able to recover a balanced true state, relative to 3D assimilation and to each other. It was shown in chapters 2 and 3 that the EKF, EnKF, and 4DVAR differ in practice, for reasons which are all linked to violation of assumptions of linearity and Gaussianity. It was found that these assumptions are upheld best when the observational information content is high, but that this can also create a regime where 3D assimilation is sufficient. Weakening linearity assumptions, such as by increasing the time between observations, can significantly weaken the advantages of 4D data assimilation.

Now consider the presence of two timescales of motion, which are connected by a nonlinear balance relationship. Early studies of 4D assimilation [e.g. Cohn and Parrish, 1999] assumed that balance would be a natural side-effect of flow-dependent covariance models.

However, later studies show that this is not always the case: while 4D algorithms alleviate the initialization problem to some extent, they nonetheless develop unphysical correlations which can cause the excitation of spurious unbalanced motion [e.g. Polavarapu et al., 2000, Houtekamer and Mitchell, 2005, Lea et al., 2006]. It has also been recognized that assumptions of linearity and Gaussianity may not be justified when motions of different timescales are possible [Lorenc, 2003a].

For the EnKF, for example, Houtekamer and Mitchell [2005] report reasonably balanced analyses in experiments with an operational weather prediction model, but also find that the final analysis could be improved further by explicitly balancing it. In the context of 4DVAR, Courtier and Talagrand [1990] used a shallow water model to show that, since the algorithm uses all degrees of freedom of the problem to minimize the cost function, it generates as many gravity waves as needed in order to best fit the observations, unless the cost function minimization is somehow constrained. Though it has already been shown that balance is not perfectly preserved in a 4D analysis, it remains to be clarified how well the basic types of 4D assimilation perform relative to one another (all other factors being equal), and what it takes to retain balance. Since it is the nonlinearity of the assimilation system that makes the three methods return different results in practice, the ability of each method to capture a nonlinear balance is also likely to be different.

In this chapter, the exL86 model is used to illustrate and clarify how each of the three methods capture, or don't capture, dynamical balance. For the Kalman filters, the question is whether, and under what conditions, forecast error covariances are evolved that reflect the dynamical balance in the true state, and to what extent this implies a balanced state. For 4DVAR, the question is how the adjoint sensitivities computed in (2.50) constrain the cost function minimization to exclude gravity waves.

Experiments with a Balanced Truth

Numerical experiments in this section are performed as outlined in §2.3, with the true state in each experiment generated with zero free gravity wave ($\tilde{I}^{\mathrm{t}} = 0$), and the forecast, unless noted otherwise, also initialized with $\tilde{I}^{\mathrm{f}} = 0$. Since the truth is balanced, the two observation types (2.58) and (2.57) represent a change of variable, but both are observations of the slow mode.

To measure the degree of imbalance induced by the assimilation system, the unbalanced component of the analysis fast mode, $\tilde{I}^{\mathrm{a}}$, is shown in individual example cases. For experiments which sweep over a given assimilation parameter, we use the total error in the fast mode (2.60) as a measure of imbalance, for continuity with the experiments of chapter 5. For a balanced true state, the fast-mode error is

$$e_{I,\mathrm{rms}} = \langle (I^{\mathrm{t}} - I^{\mathrm{a}})^2 \rangle_T^{1/2} = \langle (I^{\mathrm{t}}_{\mathrm{slav}} - \tilde{I}^{\mathrm{a}} - I^{\mathrm{a}}_{\mathrm{slav}})^2 \rangle_T^{1/2}, \tag{4.1}$$

where $I^{\mathrm{t,a}}_{\mathrm{slav}}$ represents the slaved components of the fast mode for the truth and analysis states. If the unbalanced component of the fast analysis is large compared to the slaved component, e_I becomes a measure of the spurious imbalance induced by the data assimilation.

Balance in the exL86 Covariance Matrix

In the exL86 model, the forecast error covariance matrix can be written as

$$\mathbf{P} = \begin{pmatrix} \sigma_\phi^2 & c_{\phi w} & c_{\phi x} & c_{\phi z} \\ c_{\phi w} & \sigma_w^2 & c_{wx} & c_{wz} \\ c_{\phi x} & c_{wx} & \sigma_x^2 & c_{xz} \\ c_{\phi z} & c_{wz} & c_{xz} & \sigma_z^2 \end{pmatrix}. \tag{4.2}$$

Covariances involving x and z can be written in terms of two components: a balanced component resulting from the slaving relationships [(2.11) - (2.12)], and a free gravity wave component. Writing fast-variable errors as

$$e_x = e_{U_x} + e_{\tilde{x}} \tag{4.3}$$

$$e_z = e_{U_z} + e_{\tilde{z}}, \tag{4.4}$$

and assuming that the balanced and unbalanced components of the errors are uncorrelated, the fast-variable variances become

$$\sigma_x^2 = \sigma_{U_x}^2 + \sigma_{\tilde{x}}^2 = \sigma_{U_x}^2 + \frac{\sigma_{\tilde{I}}^2}{2} \tag{4.5}$$

$$\sigma_z^2 = \sigma_{U_z}^2 + \sigma_{\tilde{z}}^2 = \sigma_{U_z}^2 + \frac{\sigma_{\tilde{I}}^2}{2}. \tag{4.6}$$

where we have defined the error variance in the magnitude of the free gravity wave as $\sigma_{\tilde{I}}^2 = \sigma_{\tilde{x}}^2 + \sigma_{\tilde{z}}^2$, with $\sigma_{\tilde{x}}^2 = \sigma_{\tilde{z}}^2 = \sigma_{\tilde{I}}^2/2$.

Since the slow mode and the free gravity wave are presumed independent, covariances between fast and slow variables are functions of the slaving relationship only, for example,

$$\begin{aligned} c_{wz} &= \langle e_w e_{U_z} \rangle + \langle e_w e_{\tilde{z}} \rangle \\ &= \langle e_w e_{U_z} \rangle , \end{aligned} \tag{4.7}$$

and likewise for $c_{\phi x}$, $c_{\phi z}$, and c_{wx}. Since $\tilde{x}$ and $\tilde{z}$ define a linear wave, $c_{\tilde{x}\tilde{z}} = \langle e_{\tilde{x}} e_{\tilde{z}} \rangle = 0$, and $c_{xz} = \langle e_{U_x} e_{U_z} \rangle$.

Covariance components involving the slaving relations can be estimated by linearizing the balance transformation (2.14) about the slow component of the model state. Defining $\mathbf{e}_{\mathbf{y}}^{\mathrm{f}} = \mathbf{y}^{\mathrm{f}} - \mathbf{y}^{\mathrm{t}}$ as the error vector in terms of the slow variables, the forecast (or background) error for the full model state can be approximated with a Taylor series expansion of (2.14):

$$\mathbf{e}_{\mathbf{x}}^{\mathrm{f}} = \mathbf{x}^{\mathrm{f}} - \mathbf{x}^{\mathrm{t}} = \mathbf{L}\mathbf{e}_{\mathbf{y}}^{\mathrm{f}} + \text{nonlinear terms} \tag{4.8}$$

where $\mathbf{L} = \partial \boldsymbol{f}(\mathbf{y})/\partial \mathbf{y} \;|_{\mathbf{y}^{\mathrm{f}}}$ is the first derivative of the balance relationship, evaluated about the slow manifold state at some point in time.

Let us define a *balanced* error covariance matrix as one where the errors in the fast variables result from the slaving to the slow variables. Balanced error covariances are then found by multiplying (4.8) by its transpose, and computing the expectation value. Truncating (4.8) at the linear term, the forecast error covariance matrix in terms of the full model state can then be approximated as

$$\begin{aligned} \mathbf{B}^{\mathrm{f}}_{\mathrm{bal}}, \mathbf{P}^{\mathrm{f}}_{\mathrm{bal}} &= \langle \mathbf{e}^{\mathrm{f}}_{\mathbf{x}} (\mathbf{e}^{\mathrm{f}}_{\mathbf{x}})^{\mathrm{T}} \rangle \\ &\approx \langle (\mathbf{L}\mathbf{e}^{\mathrm{f}}_{\mathbf{y}})(\mathbf{L}\mathbf{e}^{\mathrm{f}}_{\mathbf{y}})^{\mathrm{T}} \rangle = \mathbf{L} \langle \mathbf{e}^{\mathrm{f}}_{\mathbf{y}} (\mathbf{e}^{\mathrm{f}}_{\mathbf{y}})^{\mathrm{T}} \rangle \mathbf{L}^{\mathrm{T}} = \mathbf{L}\mathbf{P}^{\mathrm{f}}_{\mathbf{y}} \mathbf{L}^{\mathrm{T}}, \qquad (4.9) \end{aligned}$$

where

$$\mathbf{B}_{\mathbf{y}}, \mathbf{P}^{\mathrm{f}}_{\mathbf{y}} = \langle \mathbf{e}^{\mathrm{f}}_{\mathbf{y}} (\mathbf{e}^{\mathrm{f}}_{\mathbf{y}})^{\mathrm{T}} \rangle \qquad (4.10)$$

is the forecast error covariance matrix in terms of the slow variables. Since (4.9) is a tangent-linear operation, a covariance matrix that is formulated in this way can be thought of as tangent to the slow manifold. This approximation will hereafter be referred to as tangent-linear balance, or TLB, in analogy to the TLM (2.40). As in the TLM, this approximation neglects higher-order statistical moments in the forecast error distribution.

The balanced covariance matrix implies that the analysis is performed only on the balanced component of the flow. To see this, consider Kalman filter analysis increments for the slow variable w and the fast variable z, for two types of observations: (slow) w, and (mixed-timescale) w'.

For an observation of w, the analysis increments are

$$\delta w^{\mathrm{a}} = w^{\mathrm{a}} - w^{\mathrm{f}} = k_{ww} \delta w^{\mathrm{obs}} \qquad (4.11)$$

$$\delta z^{\mathrm{a}} = z^{\mathrm{a}} - z^{\mathrm{f}} = k_{zw} \delta w^{\mathrm{obs}}, \qquad (4.12)$$

where $\delta w^{\mathrm{obs}} = w^{\mathrm{obs}} - w^{\mathrm{f}}$ is the observation increment. The weights

$$k_{ww} = \frac{\sigma_w^2}{\sigma_w^2 + \sigma_{\mathrm{obs}}^2} \qquad (4.13)$$

$$k_{zw} = \frac{c_{wz}}{\sigma_w^2 + \sigma_{\mathrm{obs}}^2}, \qquad (4.14)$$

are entries in the gain matrix $\mathbf{K}$, computed from (2.32). The observation of w contains information about the slaved fast mode as well as the slow mode. Therefore the fraction of the observation increment added to z in (4.12) is proportional to the balanced term c_{wz}. If c_{wz} truly captures the covariance between w and z which results from the slaving relationship, z^{f} is adjusted along a line that is tangent to the slow manifold. However, this also means that errors in estimated fast-slow covariance terms such as c_{wz} can cause a misadjustment of fast variables, resulting in the spurious excitation of a free gravity wave.

The problem changes slightly if the observed variable is the mixed-timescale quantity, w'. In this case the analysis increments become

$$\begin{aligned} \delta w^{\mathrm{a}} &= k_{ww'} \delta w'^{\mathrm{obs}} && (4.15) \\ \delta z^{\mathrm{a}} &= k_{zw'} \delta w'^{\mathrm{obs}}, && (4.16) \end{aligned}$$

with weights

$$\begin{aligned} k_{ww'} &= \frac{c_{ww'}}{\sigma_{w'}^2 + \sigma_{\mathrm{obs}}^2} && (4.17) \\ k_{zw'} &= \frac{c_{zw'}}{\sigma_{w'}^2 + \sigma_{\mathrm{obs}}^2}, && (4.18) \end{aligned}$$

where

$$\begin{aligned} c_{ww'} &= \frac{1}{1+b^2} \left\langle \left(e_w - b e_z\right) e_w \right\rangle = \frac{1}{1+b^2} \left(\sigma_w^2 - b c_{wz}\right) && (4.19) \\ c_{zw'} &= \frac{1}{1+b^2} \left\langle \left(e_w - b e_z\right) e_z \right\rangle = \frac{1}{1+b^2} \left(c_{wz} - b \sigma_z^2\right). && (4.20) \end{aligned}$$

Now, fast-slow covariances such as c_{wz} are needed not just to update the fast variables tangent to the slow manifold, but also to correctly recover information about the slow mode from observations of the mixed-timescale variable. These terms, if estimated correctly, will be a small correction to the slow-mode analysis. If fast-slow covariances are overestimated, however, they could seriously harm the slow-mode analysis. The accuracy of fast-variable analysis increments such as (4.16) now also depends on estimated fast-variable error variances (σ_z^2 in the above equations, but also σ_x^2), which have both slaved and free gravity wave components.

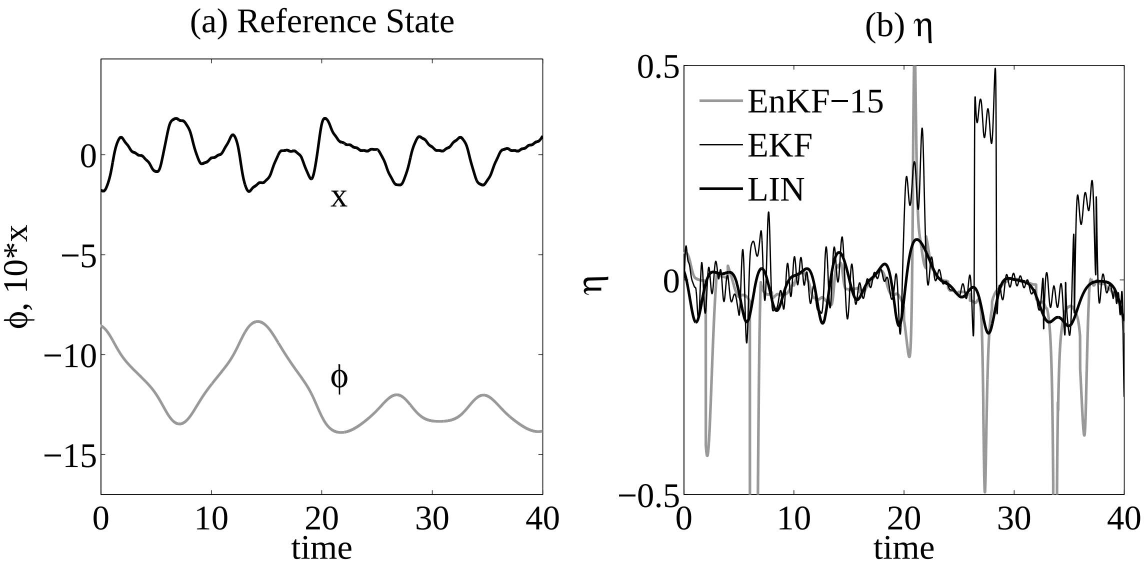

Figure 4.1: (a) Trajectories of ϕ (gray) and x (black) for a balanced example model state, where x has been multiplied by 10 in order to make its variability more visible. (b) The term η which governs the evolution of the correlation between ϕ and x for the same state (thick, black) compared to the value which is computed by a 15-member EnKF (thick gray), and an EKF (thin black).

The balanced covariance formulation (4.9) is often referred to as a strong balance constraint [Cohn and Parrish 1999, Lorenc 2003a]. However, since the covariances are evolved and updated, this constraint only holds insofar as both the balance relationship and the model dynamics can be considered as linear. The nature of balanced covariance components, and how these are explicitly evolved in the EKF and EnKF, is illustrated in Figure 4.1. Consider the correlation between errors in ϕ and x, $\rho_{\phi x} = c_{\phi x}/\sigma_\phi \sigma_x$. To examine the correlation independently of the errors themselves, we can approximate the covariance term $c_{\phi x}$ from the slaving relation (2.11) as

$$\begin{aligned} c_{\phi x} &= \langle e_\phi \frac{\partial U_x}{\partial \phi} e_\phi \rangle && (4.21) \\ &= -\epsilon C b \cos 2\phi \langle e_\phi^2 \rangle = -\epsilon C b \cos 2\phi (\sigma_\phi^2). && (4.22) \end{aligned}$$

The correlation then becomes

$$\rho_{\phi x} = -\epsilon C b \cos 2\phi \frac{\sigma_\phi}{\sigma_x} \equiv \eta_{\mathrm{LIN}} \left(\phi(t)\right) \frac{\sigma_\phi}{\sigma_x}, \tag{4.23}$$

where

$$\eta_{\mathrm{LIN}}(t) = -\epsilon C b \cos 2\phi. \tag{4.24}$$

We can evaluate the correlations estimated by each filter by comparing the quantity $\eta = \rho_{\phi x} \left(\sigma_x/\sigma_\phi\right)$ for each filter to the linearized estimate (4.24) based on the true state.

This is done for an example case in figure 4.1. Panel (a) shows $\phi(t)$ and $x(t)$ for a reference state with no free gravity wave (and with $\epsilon = 10^{-1}$ and $b = 0.71$), while (b) shows η_{LIN} for this state, compared to the values of η which are computed by a 15-member EnKF, and an EKF, both run with $\Delta t^{\mathrm{obs}} = 2$. Comparison of (a) and (b) shows that the correlation is strongly state-dependent, meaning that a dynamic covariance model is more useful than a static one. It can be seen, however, that while the EKF and EnKF both capture the overall variability of η, they can often produce large estimation errors. The ramifications of this estimation error on assimilation in different regimes are investigated numerically in the next section.

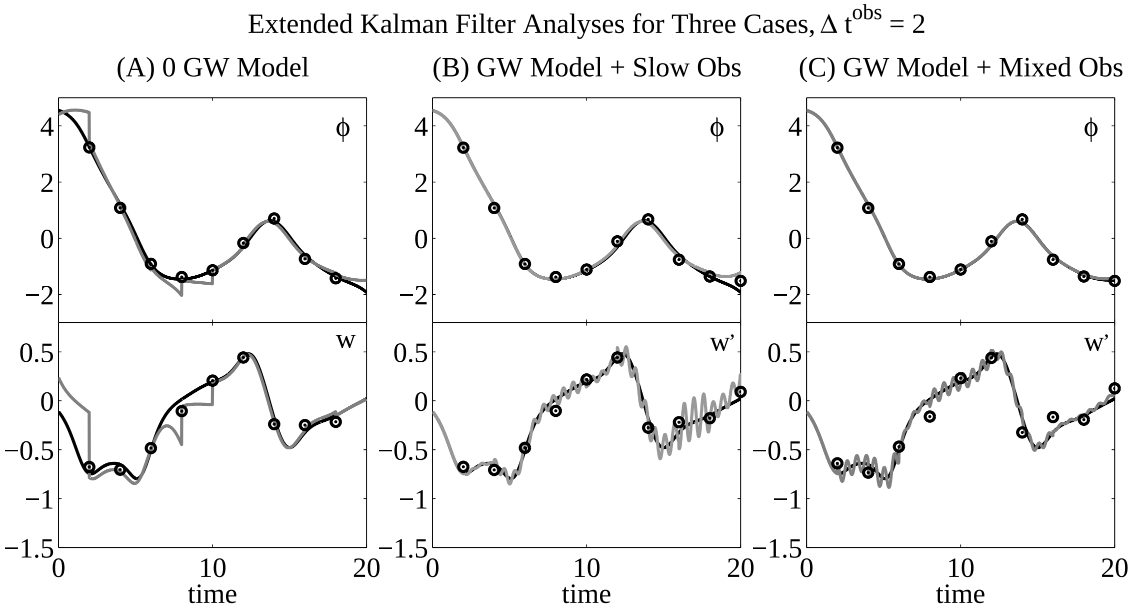

Figure 4.2: Three sample assimilation experiments with the EKF: (A) for the single-timescale model [(2.9)-(2.10)] with observations of ϕ and w, (B) for the full model [(2.1)-(2.3)] with observations of ϕ and w, and (C) for the full model with observations of ϕ and w'. The top row of plots shows ϕ for each case, and the bottom row, w' (or w for (A)). Each panel compares the true state (black) to the EKF analysis (gray) and observations (circles). In each experiment, the observation interval is $\Delta t^{\text{obs}} = 2$, and initial conditions and random error realizations are the same.

4.2 Balance in the Extended Kalman Filter

It was shown in chapter 2 that the EKF tends to misestimate correlations, and consequently overreduce variances, resulting in the rejection of observations. In this section, the ability of the EKF to preserve balance in the analysis of a balanced true state will be examined analytically and numerically.

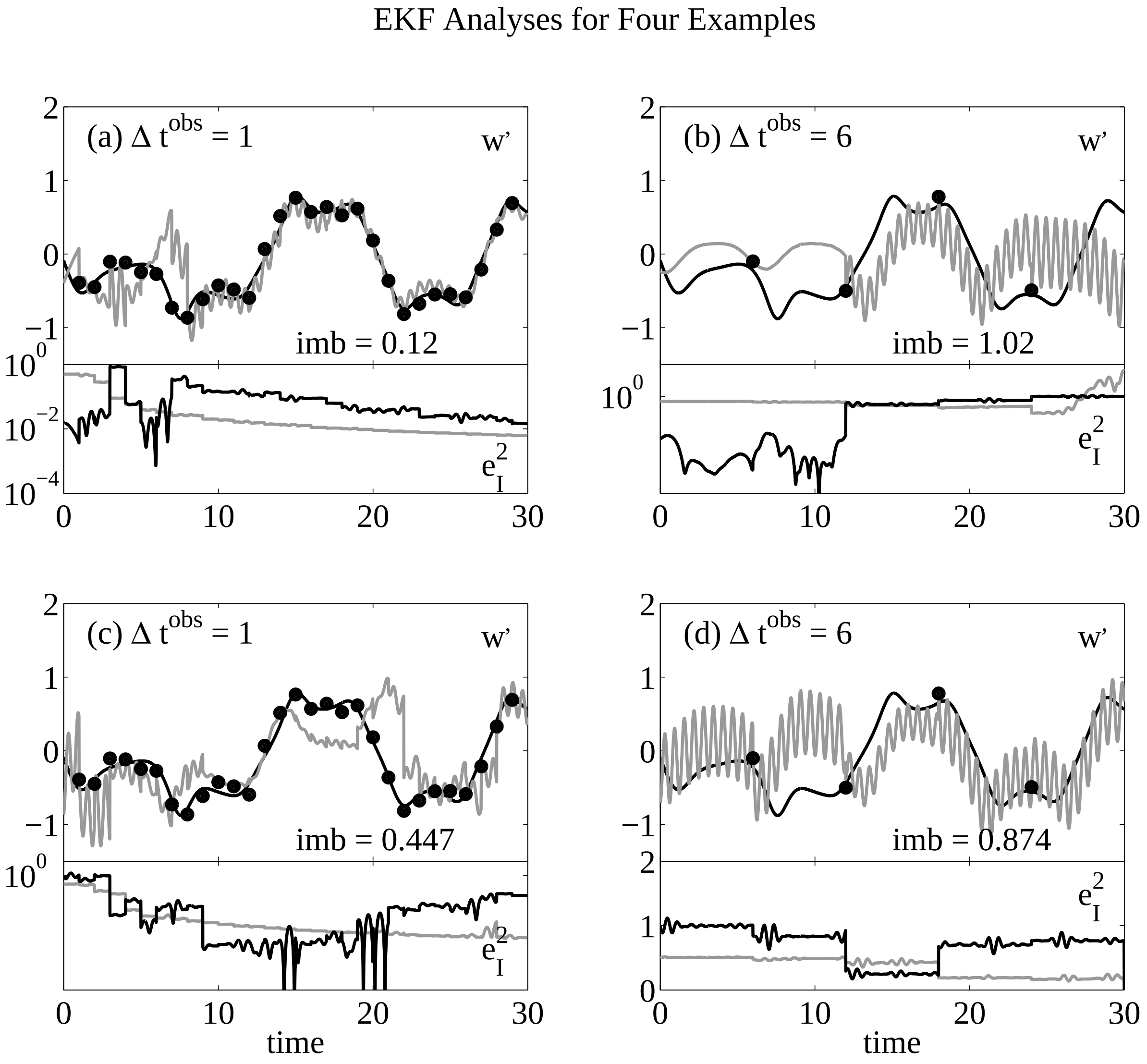

Figure 4.3: Four example EKF analyses of w' for: (a) a balanced initial forecast with observations taken every $\Delta t^{\text{obs}} = 1$ time units, (b) a balanced initial forecast with $\Delta t^{\text{obs}} = 6$, (c) an unbalanced initial forecast with $\Delta t^{\text{obs}} = 1$, and (d) an unbalanced initial forecast with $\Delta t^{\text{obs}} = 6$. The final-time values of the analysis state's free gravity wave ("imb" in the figure) are also shown. The bottom third of each panel shows the predicted fast-mode error variance (σ_I^2, gray) compared to the true square fast-mode error (e_I^2, black) in time.

4.2.1 Examples: Balance in the EKF

Figure 4.2 shows analyses of ϕ and w' for three sample EKF assimilation experiments, where the truth is a balanced state. All three experiments start from the same initial conditions and sample the same set of random numbers. In (A), the single-timescale model [(2.9)-(2.10)] is used (in which case $w' = w$). In (B), the full model is used, and completely slow observations [FIL, (2.57)] are assimilated. In (C), the full model is used, but mixed-timescale observations [MIX, (2.58)] are assimilated. The forecasts in (B) and (C) are balanced at the initial time using (2.11)-(2.12), thus reflecting a case where there is prior knowledge of the absence of gravity waves. In all four cases shown, the initial forecast error covariance matrix is estimated by a diagonal matrix,

$$\mathbf{P}_0^{\mathrm{f}} = \sigma_0^2 \mathbf{I}_4, \tag{4.25}$$

with the expectation that the series of analysis steps will adjust off-diagonal terms to more physical values as the assimilation progresses (the criticality of this assumption, in terms of balance, is explored more thoroughly below).

The slow-mode analyses (that is, of ϕ and the slow component of w') are more or less similar in each column, and largely indistinguishable from the truth. This is relatively unsurprising, since figure 3.3 showed that $\Delta t^{\mathrm{obs}} = 2$ is an adequate observation interval for observations of both ϕ and w (at least in the single-timescale model). However, the experiments in (B) and (C) differ from the single-timescale case in that the assimilation cycle also induces a spurious gravity wave in w' in both experiments, indicating projection of the analysis error onto the fast mode. This spurious projection happens even if observations are entirely slow (B). Thus we can say that the EKF analysis, even within the linear regime, does *not* necessarily return a balanced analysis.

Figure 4.3 shows another set of EKF analyses, now illustrating the effect of observation frequency. All four panels share the same balanced true state and initial perturbation, with observations now of only the mixed-timescale variable w'. In (a) and (b), the forecasts are balanced at the initial time, ($\tilde{I}^{\mathrm{f}}_0 = 0$), while in (c) and (d), the forecasts have an initial imbalance $\tilde{I}^{\mathrm{f}}_0 = 1.0$. Observations are made frequently ($\Delta t^{\mathrm{obs}} = 1$) in (a) and (c), and infrequently ($\Delta t^{\mathrm{obs}} = 6$) in (b) and (d). The predicted fast-mode error variance σ^2_I is also shown in the lower third of each panel, compared to the true square error e^2_I in the fast mode. The magnitude of the analysis free gravity wave at the end of the assimilation period is also given in each panel ("imb").

It can be seen that very frequent observations tend to retain balance in a balanced initial forecast (a), and can also reduce imbalance in an unbalanced initial forecast (c), though residual imbalance in both cases exceeds the magnitude of the slaved fast mode in the true state [which is $\mathcal{O}(\epsilon)$] Nonetheless, balance is partially restored in (c), even though no balance information is contained in the initial forecast error covariance matrix. If the time between observations is extended beyond the linear regime ($\Delta t^{\mathrm{obs}} = 6$), balance in the initially-balanced forecast deteriorates after two observations (b), and imbalance is barely reduced for the initially-unbalanced forecast (d). Furthermore, even though imbalance is not reduced in these cases, estimated fast-mode errors are reduced, and can often be orders of magnitude larger *or* smaller than actual error. Estimated fast error itself also takes on a small fast oscillation in all four cases, which is not fully removed by the addition of observations.

Though the examples shown in figures 4.2 and 4.3 represent particular realizations, they suggest three important factors for balance in the EKF. The first is the observation variable: because the covariance model is dynamic and imperfect, even FIL observations can project onto the fast manifold. In fact, it will be shown below that the spurious projection of observational information onto the free gravity wave has greater consequences when slow observations are assimilated, because these only control the slaved component of the fast variables [as shown by (4.12)].

The second factor is the strong dependence on observation frequency. Divergence of the EKF at large observation intervals (demonstrated in chapter 3, e.g. fig. 3.3) here has special consequences for balance: First, increasing Δt^{obs} increases error in the TLM approximation, and therefore estimated fast-slow covariances. Moreover, if Δt^{obs} is increased, the drift of the forecast away from the truth is greater, meaning that larger analysis increments are necessary to return the forecast to the observations. Large analysis increments amplify errors in estimated error covariances, including the balance relationships captured therein, and hence exacerbate loss of balance.

Third, estimated errors in fig. 4.3 (b) and (d) clearly do not adjust to reflect the balance relationship as observational information is added. Increasing Δt^{obs} thus doesn't harm just the state estimate, but also the estimated error covariances. It is worth investigating to what extent the covariance model may be improved by providing it with an initial knowledge of balance, as in (4.9), and the extent to which this information is retained as the assimilation progresses.

4.2.2 Slow Versus Mixed-Timescale Observations

Figure 4.4 shows EKF true errors at observation times for $\Delta t^{\mathrm{obs}} = 2$, comparing the two observation configurations, FIL (2.57) and MIX (2.58). Errors are divided into slow-mode error (2.59) and fast-mode error (2.60). The actual errors (solid lines) are compared to those estimated by the EKF cycle (dashed lines).

In the slow mode, the EKF on average diverges for both observation types: estimated errors decrease as the assimilation progresses, while true errors increase. The MIX case has larger true errors, indicating that a smaller component of the slow mode is observed. In order to recover the entire slow state, the balanced component of the observed w' must be correctly mapped to the slow variables. That information is lost in the MIX case because of the inaccuracy of the EKF covariance model at this observation frequency.

This loss of accuracy can also be seen for FIL, where true fast errors increase in time while estimated errors are left unchanged, indicating that spurious imbalance induced by the assimilation cycle is not controlled by the observations. In contrast, when a

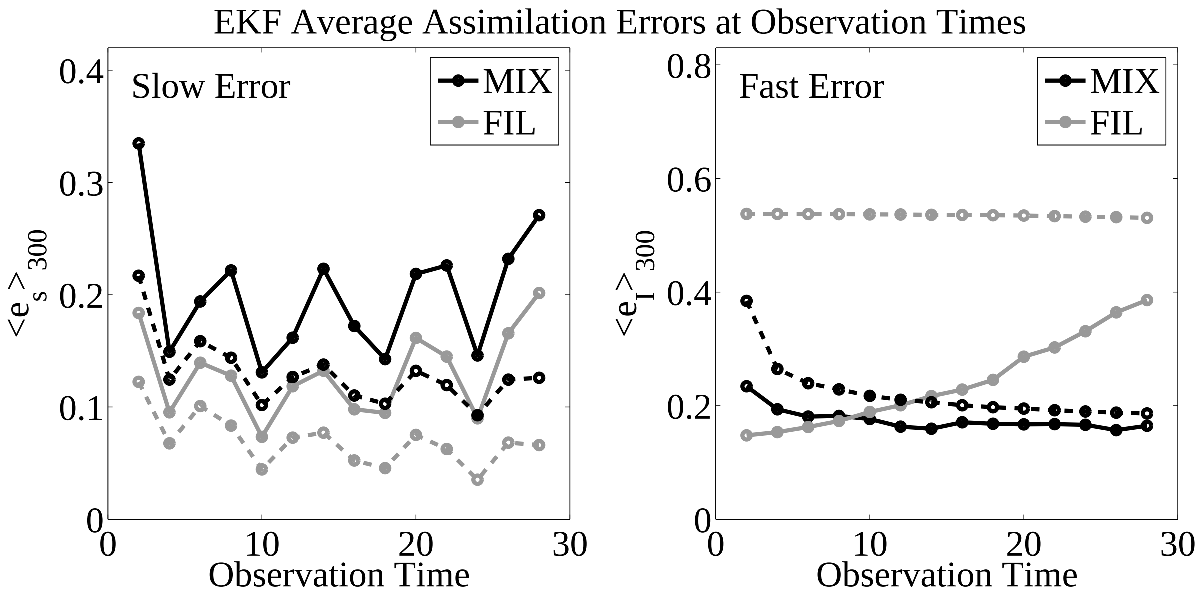

Figure 4.4: Average analysis errors for the EKF, over two types of experiments: observing the slow state (gray, FIL), and observing the mixed-timescale state (black, MIX). Errors are split into slow [(2.59), left] and fast [(2.60), right] components. For each case, the corresponding estimated errors are shown by dashed lines. Observation times are indicated by circles.

component of the fast wave is observed (MIX), fast error is reduced in time. Convergence of the EKF in the fast mode reflects the linearity of the gravity wave, in contrast to the nonlinear slow mode, where true analysis errors increase in time. Thus, though a nonlinear balance is difficult to capture in the EKF, initial imbalance can be controlled, on average, as long as observations are infrequent and include mixed-timescale variables.

4.2.3 Capturing Balance in the EKF Covariance Model: Theory

The development of accurate multivariate covariances in the EKF cycle can be seen as a problem with three components: the initial formulation of the covariance matrix, the TLM evolution of covariances (2.40), and the update step (2.36) following the insertion of observations. It is possible to examine where and how balance is lost in the EKF covariance model, using the balanced-error formulation derived in §4.1 as a guide.

1) Initial-time Covariance Matrix

Instead of specifying an initial-time error covariance matrix using (4.25), i.e. with no correlations between variables, one might instead model $\mathbf{P}_0^{\mathrm{f}}$ using knowledge of the balance relationship, as in the TLB transformation (4.9), or some approximation to it. The accuracy of such an initialization depends on how well the balance relationship is known, as well as the linearity of the balance relationship. Linearity of the balance relationship depends on the location of the state in phase space, and the size of the errors themselves; hence the effectiveness of balance-initializing $\mathbf{P}_0^{\mathrm{f}}$ will differ from case to case. Linearity of the balance relationship is also controlled by ϵ and b, becoming more nonlinear as either parameter is increased.

2) TLM Evolution of the Covariance Matrix

TLM-evolution of the covariance matrix can further impair the effect of a balance-initialized covariance matrix, because a covariance matrix initialized using the TLB approximation does not necessarily remain tangent to the slow manifold as it is evolved in the TLM (2.40), if either the model or the balance relationship are nonlinear. This can be shown using the EKF equations. If both the analysis and the true state are balanced at timestep k, such that

$$\mathbf{x}_k^{\mathrm{t}} = \boldsymbol{f}(\mathbf{y}_k^{\mathrm{t}}) \tag{4.26}$$

$$\mathbf{x}_k^{\mathrm{a}} = \boldsymbol{f}(\mathbf{y}_k^{\mathrm{a}}), \tag{4.27}$$

and if model error is zero, then the true error at the next timestep will be given by

$$\mathbf{e}_{\mathbf{x},k+1}^{\mathrm{f}} = \mathcal{M}\left[\boldsymbol{f}(\mathbf{y}_k^{\mathrm{a}})\right] - \mathcal{M}\left[\boldsymbol{f}(\mathbf{y}_k^{\mathrm{t}})\right]. \tag{4.28}$$

If the forecast model evolves a balanced state to produce a balanced state, then

$$\mathbf{e}_{\mathbf{x},k+1}^{\mathrm{f}} = \boldsymbol{f}(\mathbf{y}_{k+1}^{\mathrm{f}}) - \boldsymbol{f}(\mathbf{y}_{k+1}^{\mathrm{t}}). \tag{4.29}$$

Therefore the true error will stay tangent to the slow manifold if the model evolution is balanced and model error is zero. In contrast, if forecast errors at timestep k are balanced according to (4.9), then the forecast errors at the next timestep are given by

$$\mathbf{e}_{\mathbf{x},k+1}^{\mathrm{f}} = \mathbf{M}_k \mathbf{L}_k \mathbf{e}_{\mathbf{y},k}^{\mathrm{a}}. \tag{4.30}$$

They remain tangent to the slow manifold only to the extent that

$$\mathbf{M}_k \mathbf{L}_k \mathbf{e}_{\mathbf{y},k}^{\mathrm{a}} = \mathbf{L}_{k+1} \mathbf{e}_{\mathbf{y},k+1}^{\mathrm{f}}, \tag{4.31}$$

that is, only if *both* $\mathbf{M}_k$ and $\mathbf{L}_k$ are valid approximations to the full time evolution and balance relationships.

3) Analysis Step

For the EKF, use of the TLM is justified if the information brought in from observations in (2.36) is able to keep the evolving covariance model close to the true error statistics. Likewise, it is possible that the assimilation of observations can potentially improve the accuracy of balances represented in the covariance model. On the other hand, too much error in the computation of the optimal gain can cause the computation of the analysis error covariance matrix (2.36) to make the covariances less accurate than the estimated covariances preceding the observation. Thus, the insertion of observations can destroy the tangent-linearity of estimated covariances.

If the forecast error covariance matrix is balanced, i.e. if $\mathbf{P}^{\mathrm{f}}_k = \mathbf{L}_k\mathbf{P}^{\mathrm{f}}_{\mathbf{y},k}\mathbf{L}^{\mathrm{T}}_k$, then the gain matrix becomes

$$\begin{aligned}
\mathbf{K}_k &= \mathbf{P}^{\mathrm{f}}_k\mathbf{H}^{\mathrm{T}}\left(\mathbf{H}\mathbf{P}^{\mathrm{f}}_k\mathbf{H}^{\mathrm{T}}+\mathbf{R}\right)^{-1} && (4.32)\\
&= \mathbf{L}_k\mathbf{P}^{\mathrm{f}}_{\mathbf{y},k}\mathbf{L}^{\mathrm{T}}_k\mathbf{H}^{\mathrm{T}}\left(\mathbf{H}\mathbf{L}_k\mathbf{P}^{\mathrm{f}}_{\mathbf{y},k}\mathbf{L}^{\mathrm{T}}_k\mathbf{H}^{\mathrm{T}}+\mathbf{R}\right)^{-1} && (4.33)\\
&= \mathbf{L}_k\mathbf{P}^{\mathrm{f}}_{\mathbf{y},k}\mathbf{G}^{\mathrm{T}}_k\left(\mathbf{G}_k\mathbf{P}^{\mathrm{f}}_{\mathbf{y},k}\mathbf{G}^{\mathrm{T}}_k+\mathbf{R}\right)^{-1} && (4.34)\\
&\equiv \mathbf{L}_k\mathbf{K}_{\mathbf{y},k}, && (4.35)
\end{aligned}$$

where we have defined $\mathbf{K}_{\mathbf{y},k}$ as the gain matrix in terms of the slow variables, and $\mathbf{G}_k = \mathbf{H}\mathbf{L}_k$ as a generalized observation operator, which selects only the slow-manifold projection of the observed variable. Thus if $\mathbf{P}^{\mathrm{f}}_k$ is tangent to the slow manifold, $\mathbf{K}_k$ includes the TLB approximation.

The resulting estimated analysis error covariance matrix is then given by

$$\begin{aligned}
\mathbf{P}^{\mathrm{a}}_k &= (\mathbf{I}-\mathbf{K}_k\mathbf{H})\mathbf{L}_k\mathbf{P}^{\mathrm{f}}_{\mathbf{y},k}\mathbf{L}^{\mathrm{T}}_k && (4.36)\\
&= \mathbf{L}_k(\mathbf{P}^{\mathrm{f}}_{\mathbf{y},k}-\mathbf{L}^{-1}_k\mathbf{K}_k\mathbf{H}\mathbf{L}_k\mathbf{P}^{\mathrm{f}}_{\mathbf{y},k})\mathbf{L}^{\mathrm{T}}_k && (4.37)\\
&= \mathbf{L}_k(\mathbf{I}-\mathbf{K}_{\mathbf{y},k}\mathbf{G}_k)\mathbf{P}^{\mathrm{f}}_{\mathbf{y},k}\mathbf{L}^{\mathrm{T}}_k && (4.38)\\
&= \mathbf{L}_k\mathbf{P}^{\mathrm{a}}_{\mathbf{y},k}\mathbf{L}^{\mathrm{T}}_k. && (4.39)
\end{aligned}$$

Since $\mathbf{P}^{\mathrm{a}}_k$ can be written as $\mathbf{L}_k\mathbf{P}^{\mathrm{a}}_{\mathbf{y},k}\mathbf{L}^{\mathrm{T}}_k$, it is still tangent to the slow manifold. Thus, (2.36) remains tangent to the slow manifold to the extent that $\mathbf{L}_k$ is a valid approximation for analysis errors, which means that analysis errors must be small enough for the linearization to be valid, and $\mathbf{P}^{\mathrm{a}}_{\mathbf{y},k}$ must be accurate.

For a nonlinear model, six conditions must be met in order for balanced errors to stay balanced:

1. The TLB approximation must be valid at the initial time. Assuming that initial-time errors are small, this approximation will be valid for small enough ϵ and b, but will lose its validity as these parameters increase.

2. The forecast state must be balanced at analysis times. If observations unbalance the forecast, as in figure 4.3 (a)-(b), this will no longer be a valid assumption.

3. Evolution of the model between observations must be balanced. In the exL86 model, this is true to the order in ϵ to which the model was initialized.

4. Model error must be zero, or at least project only onto the slow manifold. Since we are restricting ourselves to perfect model experiments in this chapter, this condition will be addressed in chapter 6.

5. $\mathbf{M}_k$ and $\mathbf{L}_k$ must both be valid approximations to the model evolution and balance relationships at analysis times. This assumption will break down if analysis errors are too large.

6. $\mathbf{P}^{\mathrm{a}}_{\mathbf{y},k}$ must be an accurate estimate of slow error variances and covariances at analysis time. Chapter 3 already showed that the slow-mode covariance estimate becomes inaccurate for Δt^{obs} longer than about 3, however, just from the slow dynamics alone.

It is therefore unlikely, for a nonlinear model with a nonlinear balance relationship, that initialization of $\mathbf{P}^{\mathrm{f}}_0$ tangent to the slow manifold will ensure a balanced analysis.

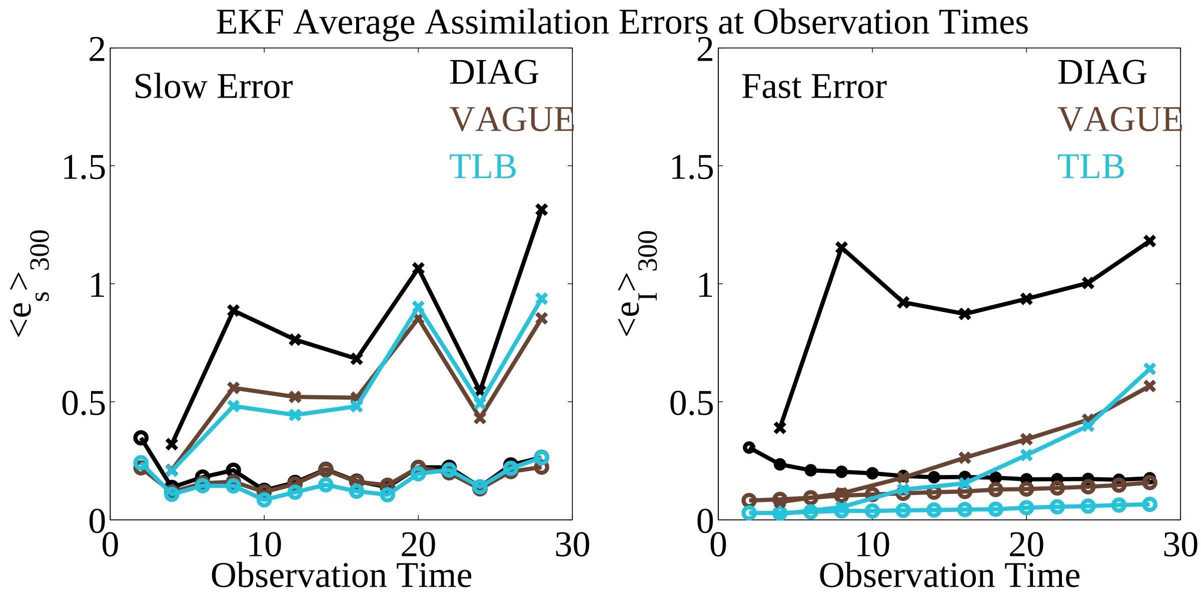

Figure 4.5: Average EKF assimilation errors at observation times, for $\Delta t^{\mathrm{obs}} = 2$ (circles) and $\Delta t^{\mathrm{obs}} = 4$ (x's), with observations of the mixed state. Each plot compares average slow [(2.59), left] and fast [(2.60), right] errors for three formulations of $\mathbf{P}_0^{\mathrm{f}}$: using the TLB transformation [(4.41), blue], using the VAGUE approximation [(4.40), brown], and using the diagonal matrix [(4.25), black].

Increasing the nonlinearity of the problem —by increasing time between observations, changing the observation variable configuration, and changing ϵ—will affect the extent to which the initialization of $\mathbf{P}_0^{\mathrm{f}}$ can balance-constrain the analysis. This is now examined numerically.

4.2.4 Capturing Balance in the EKF Covariance Model: Experiments

Figure 4.5 shows average true analysis errors at observation times, separated into slow-mode and fast-mode errors, and comparing six sets of experiments, corresponding to three initializations of $\mathbf{P}_0^{\mathrm{f}}$ and two observation intervals ($\Delta t^{\mathrm{obs}} = 2$ and 4). In the first set of experiments (DIAG), $\mathbf{P}_0^{\mathrm{f}}$ is chosen as the diagonal matrix (4.25). This is compared to a set of experiments, VAGUE, where it is assumed that the balance relations are not known, but are guessed to be functions that are proportional to ϵ, such that the error covariance matrix is approximated as

$$\mathbf{P}_0^{\mathrm{f}} = \sigma_0^2 \begin{pmatrix} 1 & 0 & \epsilon & \epsilon^2 \\ 0 & 1 & 0 & \epsilon^2 \\ \epsilon & 0 & 0 & 0 \\ \epsilon^2 & \epsilon^2 & 0 & 0 \end{pmatrix}. \tag{4.40}$$

In the third set of experiments (TLB), $\mathbf{P}_0^{\mathrm{f}}$ is initialized as

$$\mathbf{P}_0^{\mathrm{f}} = \mathbf{L}_0 \mathbf{P}_{\mathbf{y},0}^{\mathrm{f}} \mathbf{L}_0^{\mathrm{T}}, \tag{4.41}$$

where $\mathbf{P}_{\mathbf{y},0}^{\mathrm{f}} = \sigma_0^2 \mathbf{I}_2$.

The difference between the three covariance matrix initializations is greatest for fast-mode errors, where average errors in the TLB case remain lower than VAGUE and DIAG over the assimilation period for $\Delta t^{\mathrm{obs}} = 2$, and both TLB and VAGUE errors are significantly lower than DIAG errors for $\Delta t^{\mathrm{obs}} = 4$. When $\mathbf{P}_0^{\mathrm{f}}$ is initialized without correlations between variables (DIAG), average fast error immediately exceeds the magnitude of the slaved fast component of the true state [(2.11)-(2.12)], which is $\mathcal{O}(\epsilon)$. Though average fast error decreases in time for $\Delta t^{\mathrm{obs}} = 2$, it grows in time for $\Delta t^{\mathrm{obs}} = 4$, indicating the excitation of spurious gravity waves. For TLB and VAGUE, average fast error stays below the magnitude of the slaved fast mode initially, but also grows in time for $\Delta t^{\mathrm{obs}} = 4$. For slow errors, the difference only becomes clear at $\Delta t^{\mathrm{obs}} = 4$, where less information is brought from observations into the estimation of $\mathbf{P}_k^{\mathrm{f}}$. We thus see that balance eventually deteriorates in the analysis even if the initial covariance field is balance-constrained. However, this balancing can be preserved in the first few analysis steps, and even an approximation to the balance relationship, such as (4.40), is an improvement.

The effect of increasing the observation interval is examined in figure 4.6, which compares average true slow and fast errors for the three $\mathbf{P}_0^{\mathrm{f}}$ initializations over a range of

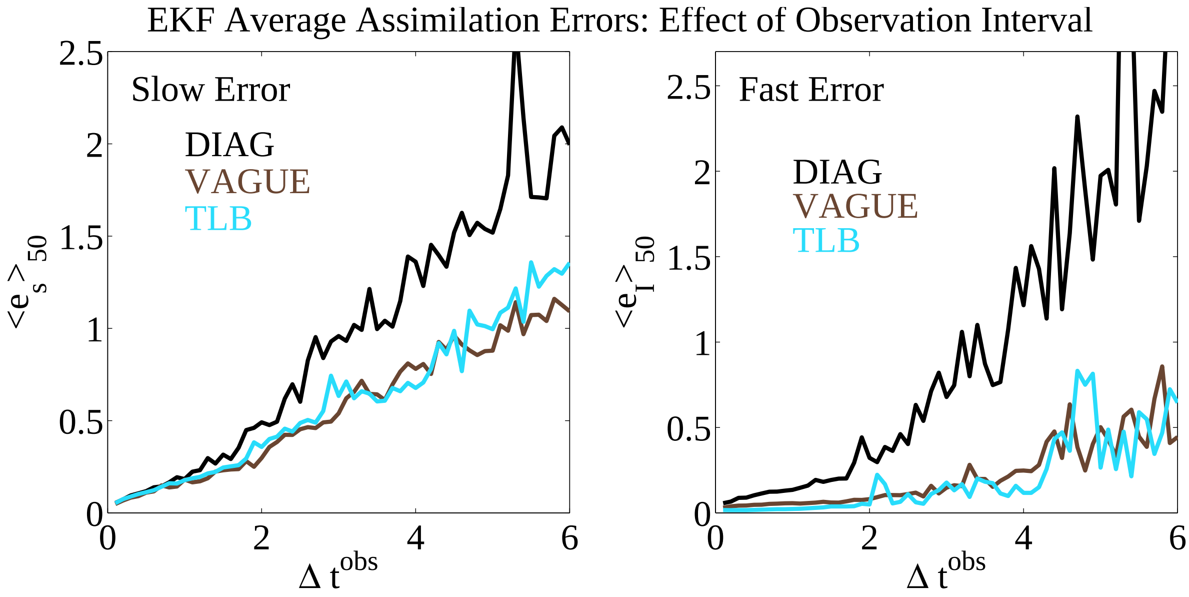

Figure 4.6: Average EKF assimilation errors over a range of observation intervals, with MIX observations. Each plot compares average slow [(2.59), left] and fast [(2.60), right] errors for three formulations of $\mathbf{P}_0^{\mathrm{f}}$: using the TLB transformation [(4.41), blue], using the VAGUE approximation [(4.40), brown], and using the diagonal matrix [(4.25), black].

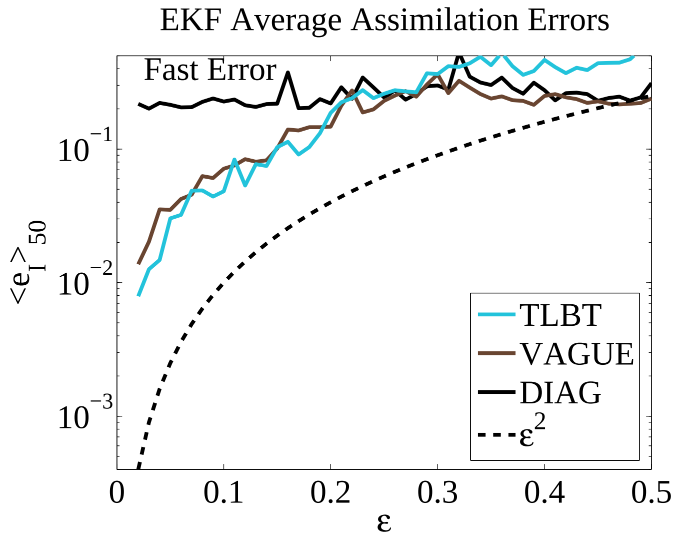

Figure 4.7: Average true analysis error in the fast mode, comparing, as in figure 4.6, three initializations of $\mathbf{P}_0^{\mathrm{f}}$: using the TLB transformation [(4.41), blue], using the VAGUE approximation [(4.40), brown], and using the diagonal matrix [(4.25), black].

observation intervals, with MIX observations (2.58). The benefit of balance-initializing the covariance matrix can be clearly seen for both error measures. Average errors grow far more with increasing Δt^{obs} for DIAG than for TLB and VAGUE. At large Δt^{obs}, where analysis increments are large, error in estimated covariances is amplified and the excitation of spurious gravity waves consequently becomes worse. This problem is alleviated by the balance-initialization of the covariance matrix in TLB and VAGUE, where fast errors are lower. Recovery of the slow state also becomes considerably more difficult with increasing Δt^{obs} for DIAG, indicating that the slow component of the observed w' is not correctly mapped to the slow manifold. Though the constraint contained in balance-initializing $\mathbf{P}_0^{\mathrm{f}}$ is most effective at large Δt^{obs}, it might be limited in realistic implementations, because memory of the EKF will decrease if there is realistic model error [Nerger et al., 2005].

Figure 4.7 examines the effectiveness of balance-initializing the covariance matrix as a function of ϵ, i.e. as the degree of timescale separation changes, again comparing the three initializations of $\mathbf{P}^{\mathrm{f}}_0$, (4.41), (4.40), and (4.25). The plot has a logarithmic y-axis, and a curve corresponding to ϵ^2 is also added, in order to emphasize the asymptotic nature of the balance relationship. VAGUE and TLB yield lower average fast-mode errors for small ϵ, but the initial formulation of $\mathbf{P}^{\mathrm{f}}_0$ no longer makes a difference for $\epsilon \gtrsim 0.3$, where the separation of "fast" and "slow" modes becomes asymptotically less well defined. Thus it becomes harder to capture slaved fast variables as slaving becomes more fuzzy. This echoes the result of Žagar et al. [2004a], who found that it is difficult to reproduce the wind field from height observations in the tropics, because the lack of a clear timescale separation between different types of waves reduced the covariance between these fields.

4.2.5 Balance Constraint in the EKF Analysis

Instead of balance-constraining only the initial covariance matrix, a balance constraint can be incorporated into the EKF cycle by performing the analysis step (2.33) in terms of the state's projection onto the slow variables only. This is done by projecting the forecast state, covariance matrix, and gain matrix onto the slow manifold, updating only the slow variables, and then mapping the analysis back to the full state.

Starting from a mixed-timescale forecast $\mathbf{x}^{\mathrm{f}}_k$, the modified algorithm is as follows:

$$\mathbf{y}^{\mathrm{f}}_k = \mathbf{F}\mathbf{x}^{\mathrm{f}}_k \tag{4.42}$$

$$\mathbf{P}^{\mathrm{f}}_{\mathbf{y},k} = \mathbf{F}\mathbf{P}^{\mathrm{f}}_k\mathbf{F}^{\mathrm{T}} \tag{4.43}$$

$$\mathbf{K}_{\mathbf{y},k} = \mathbf{P}^{\mathrm{f}}_{\mathbf{y},k}\mathbf{G}^{\mathrm{T}}_k\left(\mathbf{G}_k\mathbf{P}^{\mathrm{f}}_{\mathbf{y},k}\mathbf{G}^{\mathrm{T}}_k + \mathbf{R}\right)^{-1} \tag{4.44}$$

$$\mathbf{y}^{\mathrm{a}}_k = \mathbf{y}^{\mathrm{f}}_k + \mathbf{K}_{\mathbf{y},k}[\mathbf{z}_k - \mathbf{H}\boldsymbol{f}(\mathbf{y}^{\mathrm{f}}_k)] \tag{4.45}$$

$$\mathbf{x}^{\mathrm{a}}_k = \mathbf{f}\left(\mathbf{y}^{\mathrm{a}}_k\right). \tag{4.46}$$

Here

$$\mathbf{F} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & b \end{pmatrix} \tag{4.47}$$

is a linear mapping which projects both the model state and the forecast error covariance matrix onto the slow manifold. The gain matrix $\mathbf{K}_{\mathbf{y},k}$ is then computed for the slow variables only; note that (4.44) uses the TLB transformation to select the slow manifold projection of the observation error. The analysis is performed on the slow variables (4.45), and the analysis in terms of the full model state is computed in (4.46) by mapping the slow state back to a balanced mixed-variable state, using the slaving relations. Since the slaving relations are used in (4.46), this yields an analysis state which is necessarily balanced.

Alternatively, one could map only the covariance and gain matrices to the slow manifold (4.43) and (4.44), then update the mixed-timescale state as

$$\mathbf{x}_k^{\mathrm{a}} = \mathbf{x}_k^{\mathrm{f}} + \mathbf{L}_k \mathbf{K}_{\mathbf{y},k}[\mathbf{z}_k - \mathbf{H}\mathbf{x}_k^{\mathrm{f}}]. \tag{4.48}$$

If (4.48) is used instead of (4.46), the modified algorithm becomes similar to the simplified Kalman filter proposed by Dee [1991]. Dee [1991] showed that the Kalman filter can be made computationally cheaper by evolving only the balanced covariance matrix (4.43) forward, which requires only as many iterations of the TLM as there are balanced variables. It is pointed out that such an approximation could be minor in comparison to other approximations made in the EKF (such as linearity), in which case the modified algorithm might actually yield a more optimal covariance field, at lower cost. In other words, this simplification could be “optimal” in the sense that we neglect to explicitly model information (in this case, slaving) which is difficult to capture by the assimilation method.

Since Dee [1991] used a linear model with a linear balance relationship, the forward evolution of balanced error covariances yielded similarly balanced error covariances, and the simplified algorithm returned acceptable results. For a nonlinear model, however, (4.48) requires an additional use of the TLB transformation, and will thus be less accurate than (4.46), even without the simplification added by Dee [1991]. To avoid the additional complication of evolving only the slow component of covariances, and because the exL86 model is computationally cheap, we evolve the full covariance matrix $\mathbf{P}^{\mathrm{f}}_{\mathbf{x}}$, and solve (4.42)-(4.46) at each observation time.

A comparison between these two modified analyses [(4.45)-(4.46) versus (4.48)], is shown in figure 4.8, which compares average errors at observation times, for the two balance-constrained filters proposed above: the direct balancing given by (4.45)-(4.46), or DIR, and the indirect balance update given by (4.48), or IND. These modified analyses are also compared to the EKF initialized with the TLB transformation (denoted TLB in the figure), and results for each filter are shown for observation frequencies $\Delta t^{\mathrm{obs}} = 2$ and 4.

Both modified schemes, on average, offer an improvement over the TLB-initialized EKF in terms of fast error, especially for $\Delta t^{\mathrm{obs}} = 4$ and as the assimilation progresses in time. The two balance-constrained experiments also show slightly lower slow errors for $\Delta t^{\mathrm{obs}} = 2$. Since directly balancing the analysis [(4.45)-(4.46)] sets fast error to zero immediately following insertion of an observation, it yields nearly-zero fast error. This happens without any noticeable effect on the recovery of the slow mode. Since the indirect balancing (4.48) uses an additional approximation, it results in a less balanced state than the direct balancing. However, figure 4.8 suggests that the indirect balancing (which corresponds to the modified algorithm of Dee [1991]) may be sufficient if observations are frequent enough, indicating again that the TLB approximation is reasonable as long as errors are small.

4.3 Balance in the Ensemble Kalman Filter

The single-timescale example in chapter 3 showed that the EnKF is more stable than the EKF at large observation intervals, in the sense that the state estimate is not pushed to a far orbit where subsequent observations are rejected. Just as it does not require a TLM, the EnKF also does not require a TLB approximation to estimate balanced error

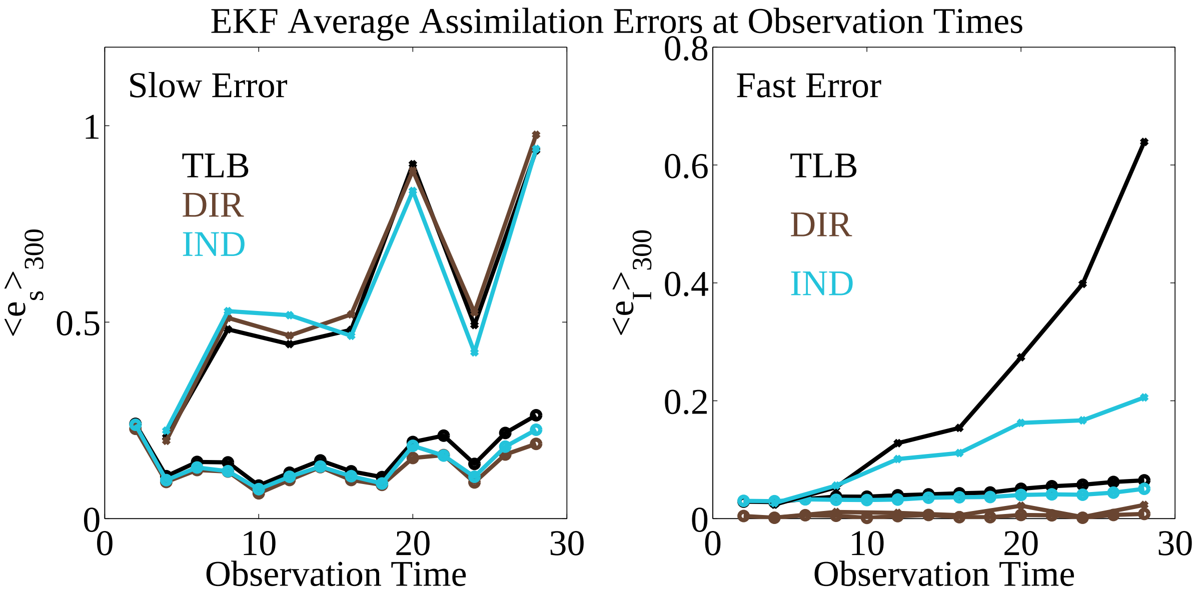

Figure 4.8: As in figure 4.5, but now comparing the TLB-initialized EKF (TLB, black) to two balance-constrained modifications: either by mapping the analysis increment with a balanced gain matrix as in (4.48) (IND, blue), or by directly balancing the analysis via [(4.45)-(4.46)] (DIR, brown). As in figure 4.5, observation frequencies shown are $\Delta t^{\text{obs}} = 2$ (circles) and 4 (x's).

covariances. The fact that the EnKF is a combination of model states further suggests that the EnKF analysis might naturally be more balanced, by virtue of the fact that there is a more physical basis for the covariance model. The averaging nature of the analysis step in itself also implies a kind of balancing: gravity wave amplitudes should average to zero if gravity wave phases in the ensemble are random. Nonetheless, as long as gravity waves are permitted in the ensemble members, a balanced analysis state is not guaranteed.

4.3.1 Balance in a Monte-Carlo Covariance Model

In the EnKF, covariances between fast and slow variables are approximated by a finite-size ensemble. While it is sometimes asserted that an ensemble of balanced forecasts will yield a balanced analysis [e.g. Szunyogh et al. 2005], $\mathbf{P}_k^{\mathrm{f}}$ will only be balanced if the balance is linear, and if $\mathbf{P}_k^{\mathrm{f}}$ is computed from an ensemble of balanced forecasts (while a realistic ensemble will converge to the correct covariance field with an error proportional to $N^{-1/2}$).

Whether or not the forward evolution and sequential updating of the ensemble indeed yields a balanced analysis depends on whether the initial N-member forecast ensemble sufficiently represents the full statistics of the true system, including balance; whether the evolution and spreading of the ensemble does not destroy its accuracy; and whether the ensemble analysis step (2.44) doesn't unbalance individual ensemble members.

1) Ensemble Representation of Balance

It can be argued that the existence of a balance relation might simplify the EnKF problem, since a balanced model state has fewer degrees of freedom than an unbalanced state, and fewer ensemble members will therefore be required to represent the error statistics of a balanced state—if the forecasts in the ensemble are all balanced. However, the nonlinear balance relationship itself must be captured.

Balanced perturbations for the exL86 model can be generated by transforming the central forecast to normal modes, adding random perturbations to ϕ^{f} and w^{f}, and then transforming back to mixed variables using (2.14). In realistic applications, this step is a bit more complicated, but similar: for example, one can randomly perturb streamfunction, and then derive wind, temperature, and pressure perturbations following some balance assumption [Mitchell et al., 2002]. In lieu of an explicit slow manifold initialization, one might also integrate the ensemble forward while filtering out fast waves with, say, a digital filter [Evensen, 1997].

2) Evolution of Ensemble Statistics Between Observation Times

As the forecast ensemble evolves between observation times, imbalance in error covariances will only grow as much as the mean imbalance in the ensemble (as opposed to the unbounded growth that would happen with the TLM evolution of errors). Eventually, the ensemble spread will become saturated in the slow mode. However, even then it could still contain information about the (slaved) fast mode.

3) Ensemble Analysis Step

If individual ensemble members become unbalanced in the analysis step (2.44), the amount of imbalance remaining in the mean state depends on the magnitude and relative phases of the fast motion in individual analyses. If the ensemble is too small or there is not enough phase mixing, some net imbalance will remain in the analysis. Moreover, if individual ensemble members include spurious imbalance, then the subsequent analysis can also be expected include spurious imbalance.

4.3.2 Examples: Balance in the EnKF

Figure 4.9 shows EnKF analyses for the same three states and observation sets shown in figure 4.2: (A) the single-timescale model with observations of w and ϕ, (B) the full

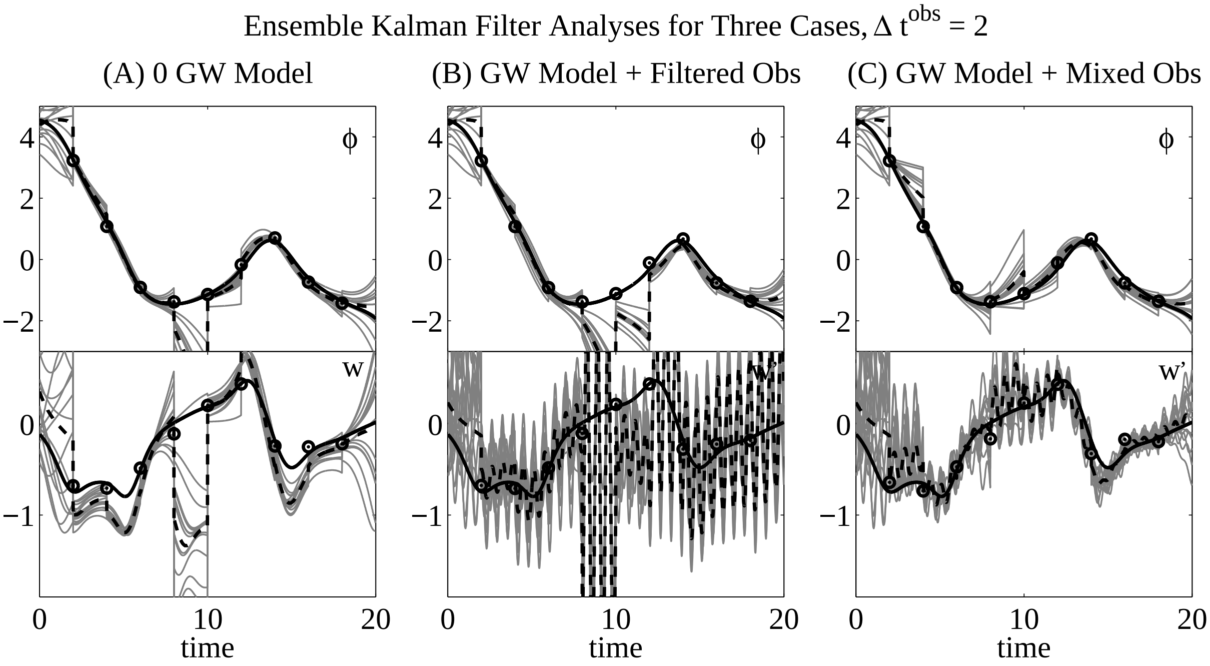

Figure 4.9: As in figure 4.2, but for a 10-member EnKF.

model with observations of w and ϕ, and (C) the full model with observations of w' and ϕ. In these experiments a 10-member ensemble is used, and all ensemble members are initially balanced. As in the EKF example (fig. 4.2), assimilation causes the analysis to become unbalanced for both observation types, even though observations are frequent enough to capture the slow mode. In (B) and (C), the ensemble mean has a smaller gravity wave than individual ensemble members, showing that at least some balancing can be achieved by averaging. However, ensemble members do phase-lock to some extent around a spurious gravity wave, leaving a substantially unbalanced ensemble mean state in both cases.

Figure 4.10 shows four more EnKF assimilation experiments, now examining the effect of observation interval, with observations of w', and all other assimilation parameters as in the corresponding EKF example (fig. 4.3). Again, the ensemble consists of 10 members in all four cases. In (a) and (b), the ensemble is balanced initially, while in (c) and (d), the initial ensemble has random free gravity wave magnitudes and phases (distributed

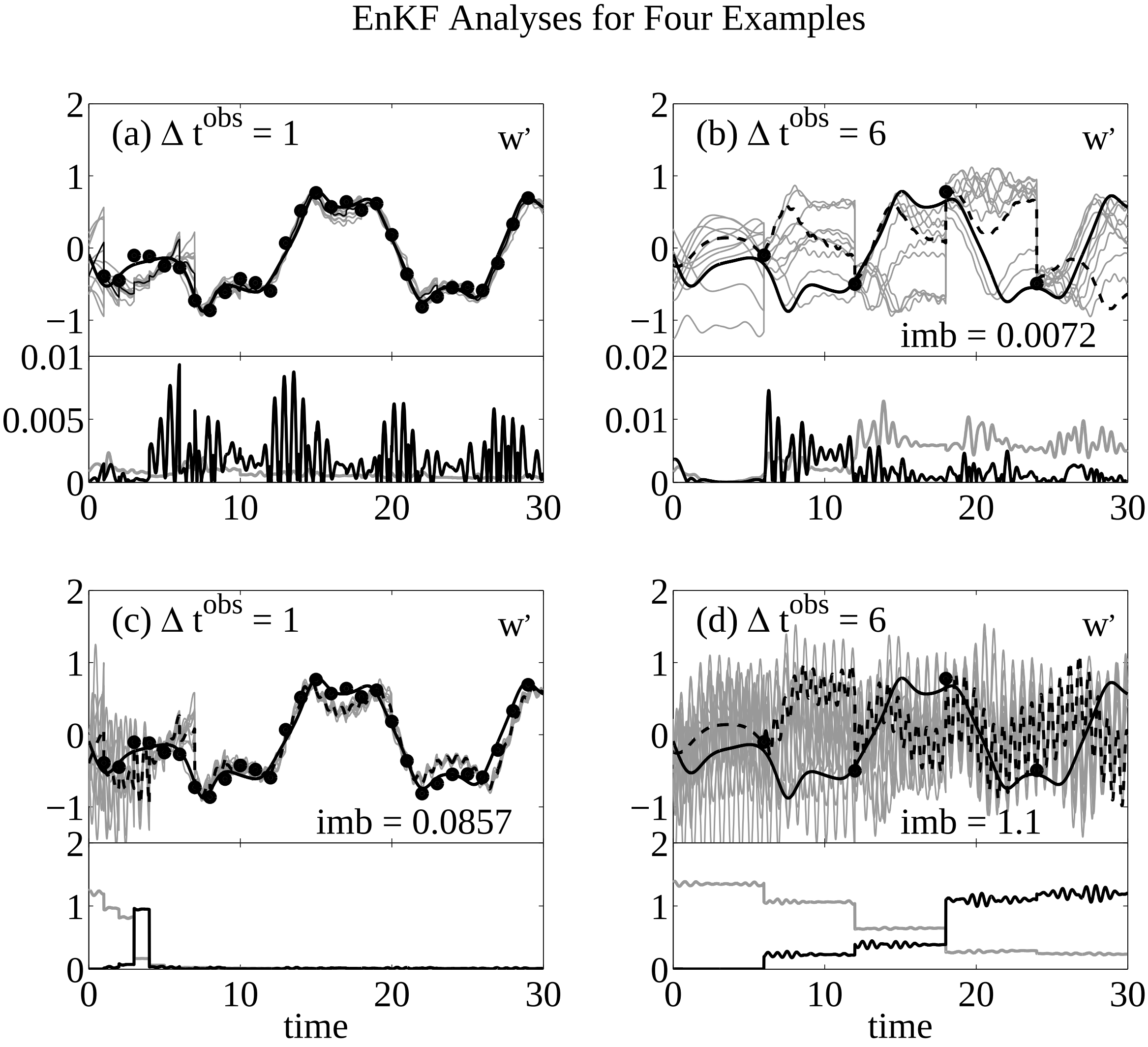

Figure 4.10: As in figure 4.3, but for the 10-member EnKF.

such that the free gravity wave in the ensemble mean is $\tilde{I}^{\mathrm{f}} = 1.0$). As in the corresponding EKF example (fig. 4.3), panels (a) and (c) show observations assimilated with $\Delta t^{\mathrm{obs}} = 1$, and panels (b) and (d) show observations assimilated every $\Delta t^{\mathrm{obs}} = 6$ time units. (Note that a linear scale is now shown in all figures.)

The initially-balanced ensemble becomes tightly crowded around the true state, with very small net imbalance, for frequent observations (a). Ensemble-estimated forecast error covariances also contain this imbalance. As in the EKF, the initially-unbalanced ensemble (d) becomes balanced as observations are brought in, although it can also be seen (by comparing the ensemble to the true state) that spread in the ensemble becomes less than the true error towards the end of the assimilation period.

Increasing the observation interval, the initially-balanced ensemble (b) diverges significantly between observations, becoming clearly non-Gaussian, though forecast error is large enough that the mean state is still brought close to the truth at observation times. As in the EKF case, the large analysis increments which occur for $\Delta t^{\mathrm{obs}} = 6$ induce imbalance in individual ensemble members, though there is enough phase-mixing that ensemble averaging reduces net imbalance substantially. Again, estimated fast errors are of the same order of magnitude as true fast errors. If observation frequency is reduced to $\Delta t^{\mathrm{obs}} = 6$ for an initially-unbalanced ensemble (d), the ensemble does not become balanced, but instead begins to phase-lock. Here, estimated fast errors begin to significantly underestimate true errors.

We thus see that the EnKF does have the advantage of ensemble averaging to keep a balanced state, but that this only works to an extent, because the assimilation of observations also causes phase-locking in an unbalanced ensemble. Moreover, such an ensemble implies that the true state is unbalanced, with the degree of imbalance in the truth unknown. We also see that the EnKF's stability at large Δt^{obs} has consequences for balance: because forecast error covariances only reflect as much imbalance as is present in the ensemble, the analysis does not cause as much imbalance at large Δt^{obs} as the

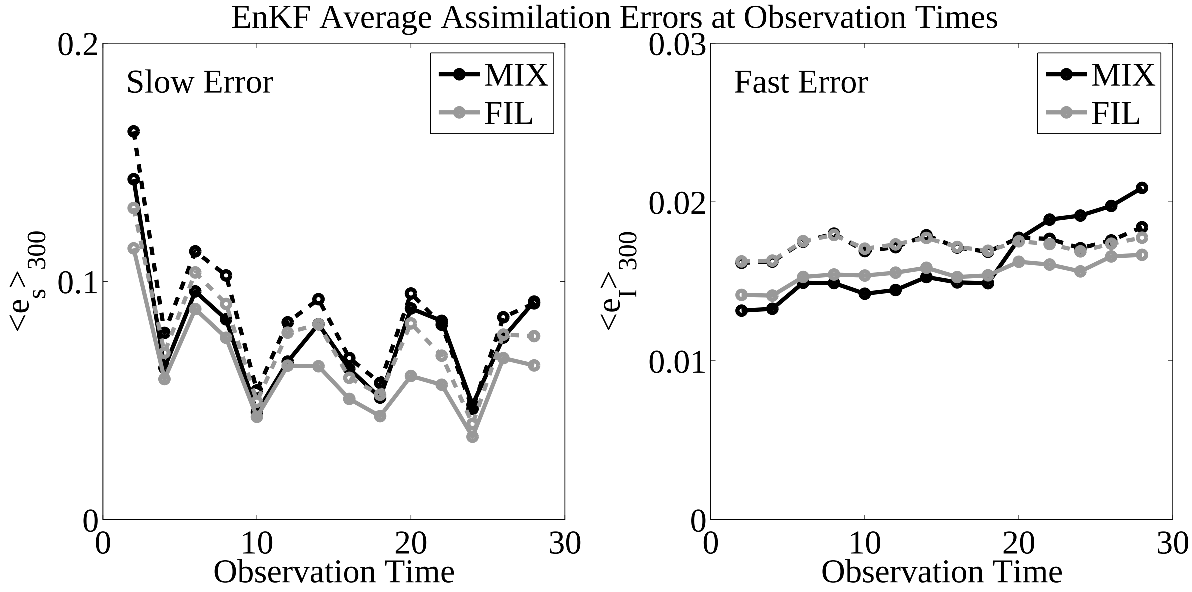

Figure 4.11: Average analysis errors for the EnKF with 50 ensemble members, over two types of experiments: observing the slow state (gray, FIL), and observing the mixed-timescale state (black, MIX). Errors are split into slow [(2.59), left] and fast [(2.60), right] components. For each case, the corresponding estimated errors are shown by dashed lines. Observation times are indicated by circles.

EKF analysis does, as long as the ensemble itself is balanced. Nonetheless, despite the very large ensemble size used here, the EnKF can become significantly unbalanced.

4.3.3 Slow Versus Mixed-Timescale Observations

Figure 4.11 shows EnKF average slow and fast-mode errors at observation times, again comparing actual to estimated errors, and MIX to FIL observations (as in figure 4.4). To minimize the effect of ensemble sampling error, ensembles of 50 forecasts are used here, with $\Delta t^{\text{obs}} = 2$. Comparison of figures 4.4 and 4.11 shows that the EnKF has lower errors than the EKF at this observation frequency, in both modes. Moreover, slow analysis errors decrease in time, with estimated errors similar to actual errors. Fast errors do grow in time, however, indicating a gradual loss of balance. Note that this

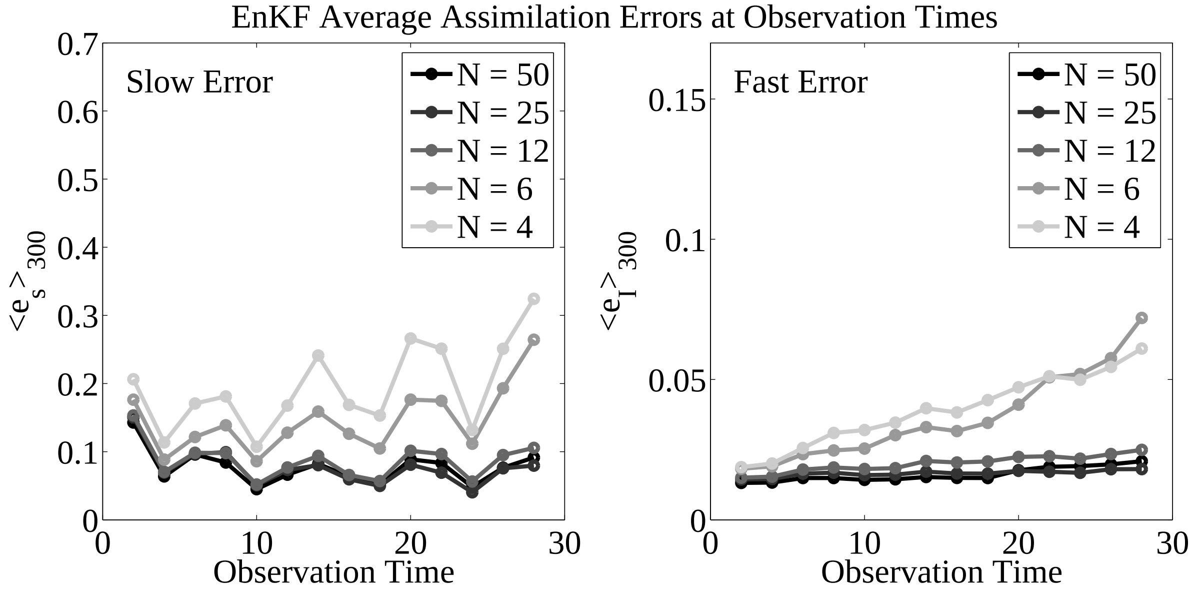

Figure 4.12: Average EnKF assimilation errors at observation times, for $\Delta t^{\mathrm{obs}} = 2$, with MIX observations. Each plot compares average slow [(2.59), left] and fast [(2.60), right] errors for different ensemble sizes.

happens for both observation types, whereas fast errors in the EKF decreased in time for MIX. The similarity of average error for the two observation types indicates that balanced covariances are captured well enough by the ensemble that the slow component is extracted correctly from mixed-timescale observations, while minimal spurious gravity wave noise is excited for slow observations. While balance is lost gradually in the EnKF even for mixed-timescale observations, the loss of balance at the end of the assimilation period is still less than after the first EKF analysis increment.

4.3.4 Ensemble Size and Observation Interval

Though the EnKF is stable over long Δt^{obs}, it is limited by ensemble size. The effect of decreasing the ensemble size is examined in figure 4.12, which compares average fast and slow analysis errors at observation times for $\Delta t^{\mathrm{obs}} = 2$, at five different ensemble sizes. Both error measures are reduced by increasing N, but do not decrease significantly for

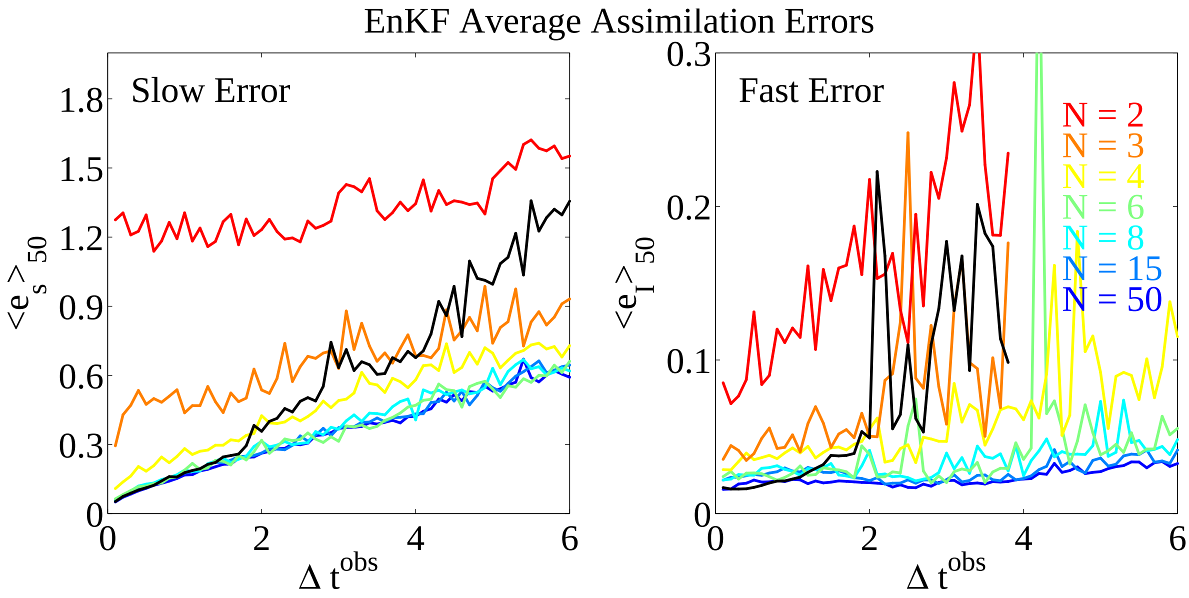

Figure 4.13: Average true analysis errors for the EnKF over a range of observation intervals, comparing average slow [(2.59), left] and fast [(2.60), right] errors, with MIX observations. Seven ensemble sizes (denoted by colors) are compared to the TLB-initialized EKF (in black).

ensemble sizes beyond 12 members, i.e. three times the dimension of the model. Average fast errors grow in time for $N = 4$ and 6, for which the EnKF is computationally similar to the EKF. However, even at these ensemble sizes the average fast error is lower than the corresponding EKF error, while average slow-mode errors are still comparable to EKF slow-mode errors (e.g. MIX in fig. 4.4). Thus the assertion of Lorenc [2003b], that ensemble-derived error statistics are limited by practically-feasible ensemble sizes but have the advantage that they are more stable in time than TLM-derived error statistics, can be extended to the issue of balance: even though the small ensemble has growing imbalance as the assimilation progresses (indicating filter divergence), growth of imbalance is relatively small over the assimilation period.

Figure 4.13 examines the effect of observation interval. This figure is analogous to figure 4.6 for the EKF, but compares ensembles of $N = 2,\ 3,\ 4,\ 6,\ 8,\ 15$ and 50 forecasts to the EKF with $\mathbf{P}_0^{\mathrm{f}}$ initialized using the TLB approximation. First, note that fast error in the EnKF is controlled more by ensemble size than observation interval. For ensembles larger than 3 members, the advantage of the EnKF becomes clear for $\Delta t^{\mathrm{obs}} \geq 2$. Even for 4 ensemble members, where it is difficult for unbalanced motion to average out completely, far less net imbalance is induced in individual ensemble members, relative to the EKF. This implies that a substantial fraction of the EnKF's success comes from the retention of nonlinearity in the covariance evolution. Nonetheless, effect of ensemble averaging can be seen in the clear improvement from $N = 6$ to $N = 12$ in figure 4.12 (b). The combined effect is that fast error in the EnKF is controlled more by ensemble size than by observation frequency.

For 3 ensemble members, where the slow mode is difficult to capture, less imbalance is still induced in the EnKF, relative to the EKF, for $\Delta t^{\mathrm{obs}} \gtrsim 2$. However, imbalance is still quite large for $N = 2$ and 3 (ensembles smaller than the state dimension). For $\Delta t^{\mathrm{obs}} \lesssim 2$ and $N \leq n$, the EKF returns both the best slow mode analysis and the most balanced analysis.

It is also interesting to note that EnKF fast errors are slightly larger for $\Delta t^{\mathrm{obs}} \lesssim 2$ than for $2 \lesssim \Delta t^{\mathrm{obs}} \lesssim 4$. This is a robust result for different ensemble sizes, and is explainable: more frequent analysis of the ensemble means that estimated forecast errors (corresponding to the spread of the ensemble) are small, while errors in estimated correlations (due to the finite ensemble size) have a larger impact. Consequently, individual ensemble members and the ensemble-mean become more unbalanced. There thus seems to be a range of observation intervals which are optimal in terms of computing a balanced analysis. Lorenc [2003b] points out that the tendency of the EnKF to diverge for a large observation density could pose serious problems as the abundance of observations increases. Figure 4.13 suggests that one aspect of this problem could be spurious imbalance: as observations become more frequent, the tendency of the ensemble to lock onto a spurious gravity wave increases. An important caveat here is that the model only admits one gravity wave frequency, and that this frequency is the same for the truth and forecast. In chapter 6 it will be shown that ensemble phase-locking becomes much less likely if the truth and forecast are evolved with models which admit different gravity wave frequencies.

In summary, we find that the EnKF has one major balance-related advantage over the EKF, which is stability of estimated covariances over long observation intervals. While the EnKF also benefits from averaging, this requires a large ensemble and sufficient phase mixing of the ensemble.

4.4 Balance in 4DVAR

4.4.1 Balance in Variational Assimilation

We now extend the analysis briefly to 4DVAR. We have seen (chapter 3) that 4DVAR is more stable than the EKF at large Δt^{obs}, while also bypassing sampling error issues in the EnKF. Since the practical difference between the EKF and 4DVAR results largely from nonlinearity of the assimilation system, it can be expected that loss of balance will also happen differently in 4DVAR.

4.4.2 Examples: Balance in 4DVAR

Figure 4.14 shows three 4DVAR example assimilation experiments, which correspond to the experiments shown in figures 4.2 and 4.9 and use the same truth, initial forecast, and observations. As in those figures, the single-timescale model is used in (A), the gravity wave model with FIL observations in (B), and the gravity wave model with MIX observations in (C). In analogy to the corresponding EKF experiment, $\mathbf{B}$ is estimated as the diagonal matrix (4.25) in all three cases. The minimization window length is $\Delta T = 10$, which means that two iterative minimizations of the cost function are performed in

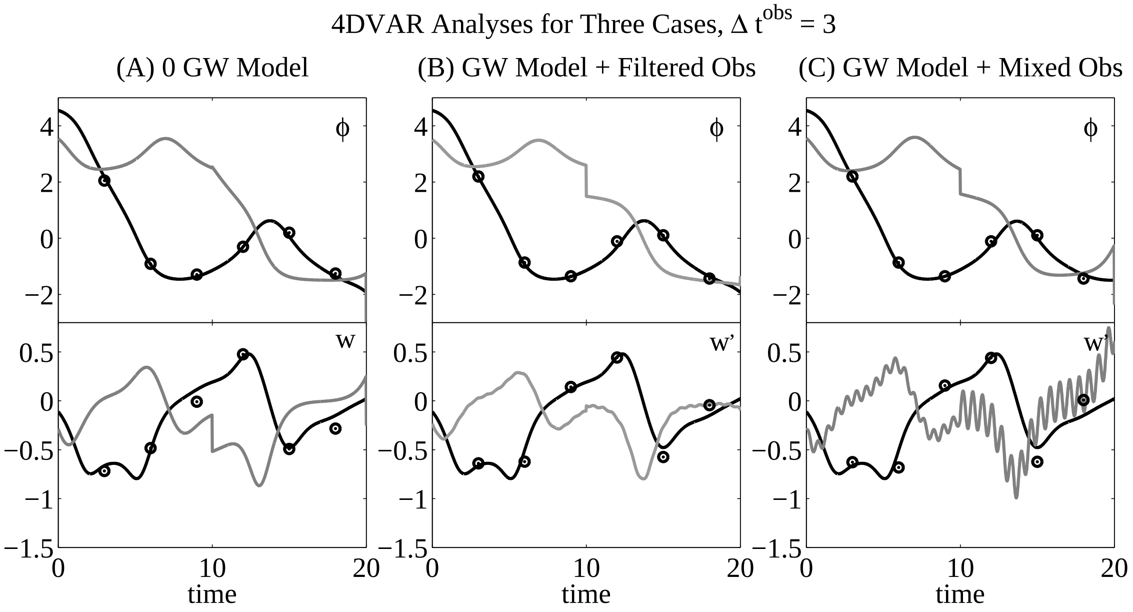

Figure 4.14: As in figures 4.2 and 4.9, but for 4DVAR, with $\Delta T = 10$ and 5 minimization iterations.

each example, over $\Delta T_1 = [0, 10]$ and $\Delta T_2 = [10, 20]$. The minimizations in this example are run out to five iterations, with no cutoff threshold.

For ΔT_1 in this example, the state falls into a local minimum of the cost function at the second minimization iteration, yielding a slow-mode analysis in (A)-(C) which is considerably worse than in the corresponding EKF (fig. 4.2) and EnKF (fig. 4.9) analyses. The difference between the three methods illustrates the point made in chapter 3 that the overall nonlinearity of each experiment also depends on the assimilation method used. It was shown in chapter 3 that $\Delta t^{\text{obs}} = 3$ is something of a boundary between the linear and nonlinear assimilation regimes. In this particular case, $\Delta t^{\text{obs}} = 3$ presents a very nonlinear problem for 4DVAR (the cost function has local minima in which the analysis becomes "trapped") but not for the EKF. There is also a clear difference between the two minimization windows: practically no fit to observations is achieved in the first window, while a somewhat better fit is achieved in the second window, even though the initial background error over ΔT_2 is much larger.

Despite the failure of 4DVAR to capture the slow mode, spurious imbalance in this example is only induced in (C), with a spurious gravity wave generated in the second window that is roughly twice as large as the gravity wave induced in the first window. Interestingly, greater imbalance happens to be induced in the time window where the slow mode is captured more accurately. This is not entirely an accident: we will see below that a stronger adjustment of the initial background guess towards observations frequently results in a greater loss of balance.

The effect of observation frequency and the ability to keep balance in a balanced background state (or restore balance in an unbalanced background state) is examined in figures 4.15 and 4.16, which show four experiments that are analogous to figures 4.3 and 4.10, again using the same random number realizations. Figure 4.15 shows the w' truth and analyses for (a) a balanced initial background state with frequent observations, (b) a balanced initial background state with infrequent observations, (c) an unbalanced initial background state with frequent observations, and (d) an unbalanced initial background state with infrequent observations. Figure 4.16 shows the RMS error in the fast mode (2.60) for each of the three assimilation windows and for each case, as a function of minimization iteration number.

For frequent observations and a balanced initial background (a), 4DVAR yields a state which is practically indistinguishable from the truth, and almost perfectly balanced. Fast error in each window remains at $\mathcal{O}\left(\epsilon^2\right)$ throughout the minimization. Extending the observation interval to $\Delta t^{\text{obs}} = 6$ (b) yields a state which is balanced for the first two time windows (ΔT_1 and ΔT_2), but significantly unbalanced for ΔT_3. Note that imbalance grows with increasing minimization iterations [fig. 4.16 (b)], which means that the net imbalance in the analysis state would be less if the minimization were cut off earlier. In contrast, for frequent observations and an unbalanced background state (c), continuing iterations restore balance in the second and third time windows. Over ΔT_1, where the background estimate is closest to the truth, there is no clear reduction of imbalance.

Increasing the observation interval to $\Delta t^{\mathrm{obs}} = 6$ (d), the unbalanced forecast remains unbalanced for all three time windows.

As in the EKF and EnKF, spurious imbalance can grow in 4DVAR if observations are infrequent (b), and can be reduced if observations are frequent (c). However, figure 4.16 points out an interesting difference between 4DVAR and the Kalman filters: spurious imbalance grows or is decreased, not in time, but with iterations of the minimization [as in the third window in (b) and (c)]. This is examined more closely in figure 4.17, which shows another 4DVAR assimilation example, for a special case where the minimization is allowed to run out to nine iterations. In this example, $\Delta t^{\mathrm{obs}} = 2$ and $\Delta T = 10$, with an assimilation period of $T = 30$, such that there are five observations in each of the three windows, and one observation which is shared between neighbouring windows. The analyses of ϕ and x are shown in panels (a) and (b), and the RMS slow and fast analysis errors, for each window, are shown as a function of iteration in panels (c) and (d). In this example, the background estimate of ϕ is so far from the truth that practically no fit to the truth is achieved in $\Delta T_2 = [10, 20]$ and $\Delta T_3 = [20, 30]$, even within nine iterations (c). However, in terms of x (b), imbalance is actually *lowest* over ΔT_2, because the minimization settles on a slightly more balanced state at the eighth iteration —even though there is only a very small change in the slow state over these nine iterations. For $\Delta T_1 = [0, 10]$, where the slow mode fit is best, imbalance starts out low but increases with subsequent iterations. Note that, for the first few iterations, imbalance increases most quickly (iteration-wise) in those windows (ΔT_2 and ΔT_3) where the prior estimate is far from the truth and iteration steps in the slow mode are therefore larger.

This example illustrates an interesting property of 4DVAR: even after the slow-mode analysis has more or less settled into an analysis state near the minimum of the cost function, small iterations can increase or decrease imbalance in the analysis. Greater imbalance is also induced by larger minimization increments. This means that the net spurious imbalance generated by the 4DVAR analysis depends both on how many iter-

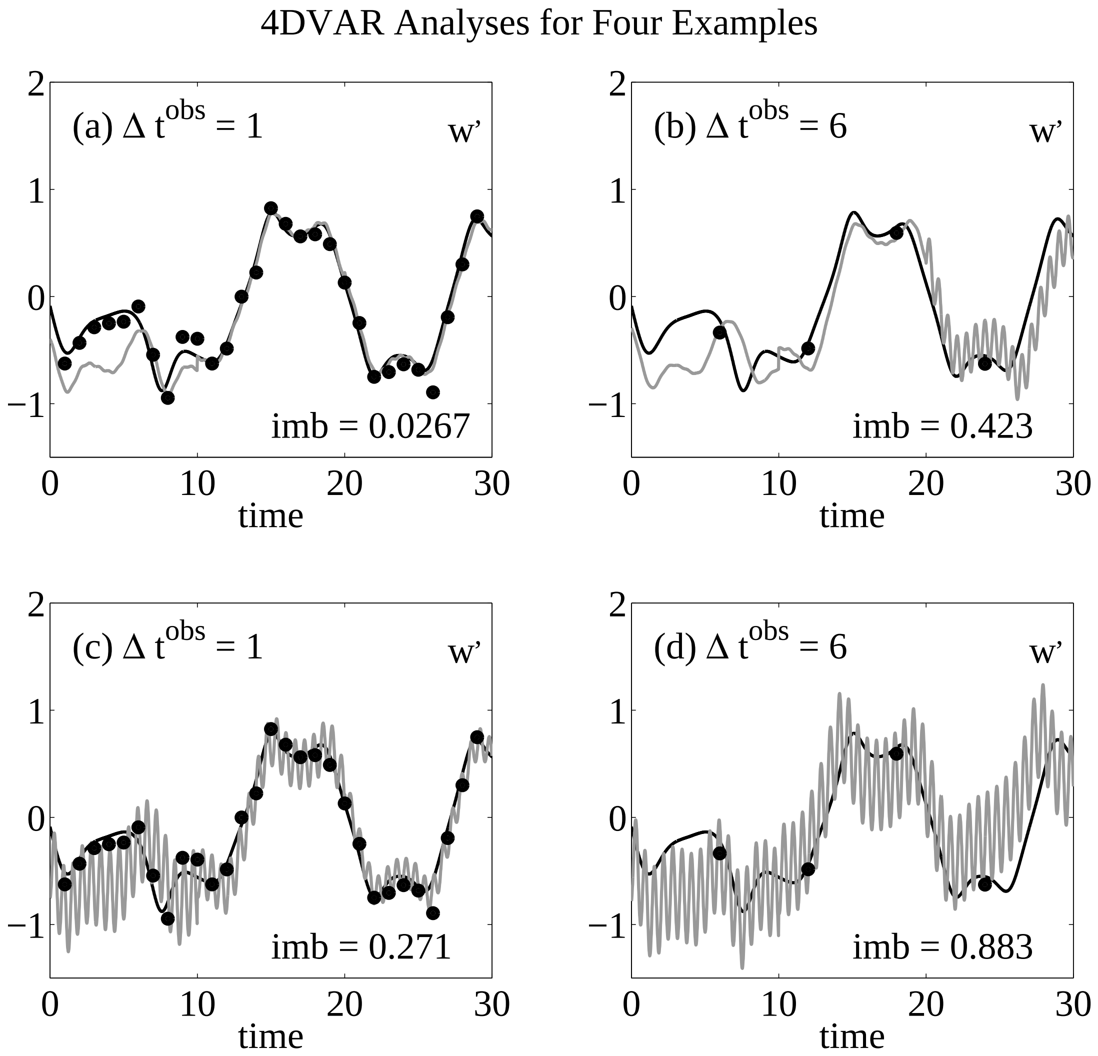

Figure 4.15: Four example 4DVAR analyses of w' for: (a) a balanced initial forecast with observations taken every $\Delta t^{\text{obs}} = 1$ time units, (b) a balanced initial forecast with $\Delta t^{\text{obs}} = 6$, (c) an unbalanced initial forecast with $\Delta t^{\text{obs}} = 1$, and (d) an unbalanced initial forecast with $\Delta t^{\text{obs}} = 6$. The final-time values of the analysis state's free gravity wave ("imb" in the figure) are also shown.

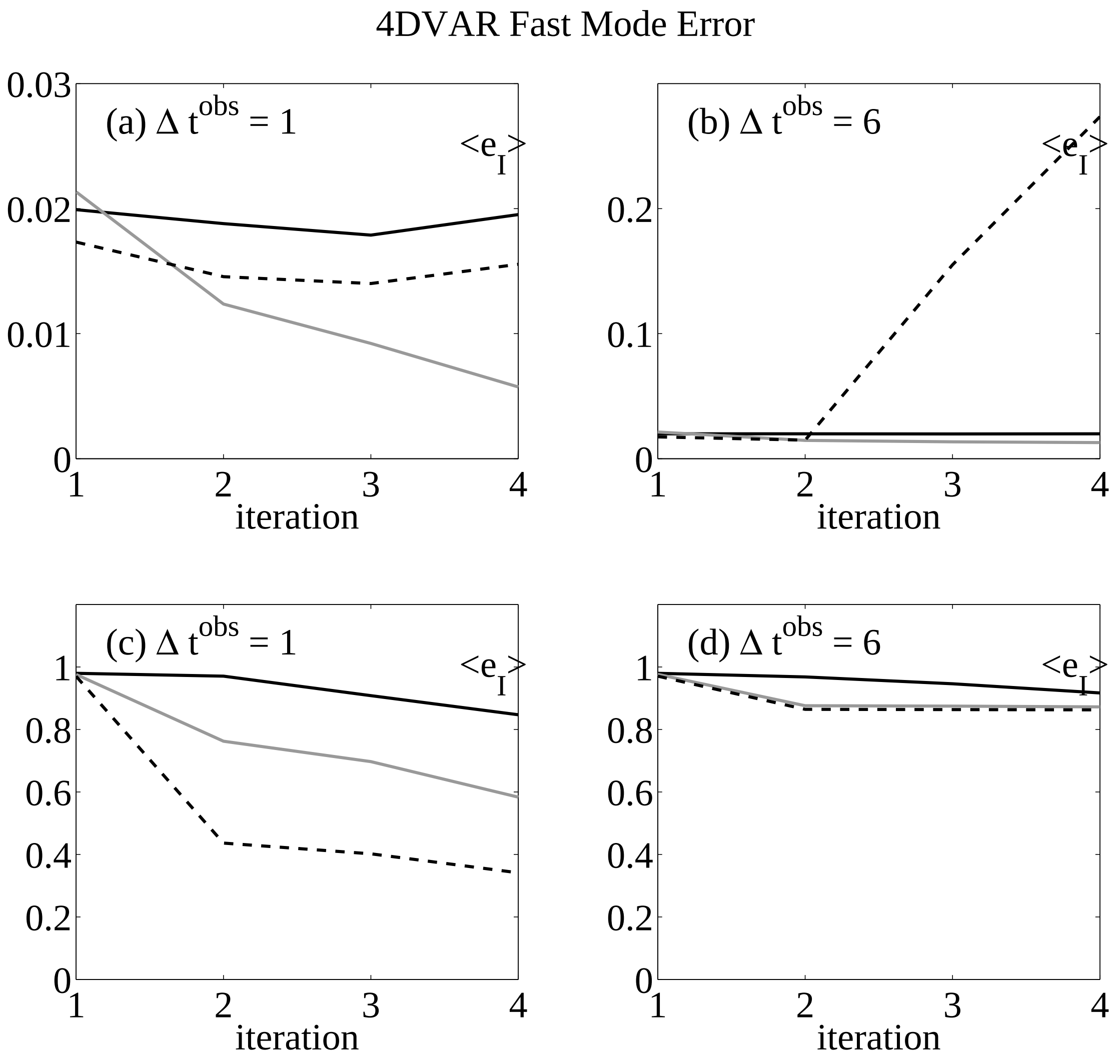

Figure 4.16: Fast-mode error (2.60) as a function of conjugate-gradient iteration, for the three assimilation windows in each example shown in figure 4.15: $\Delta T_1 = [0, 10]$ (black, solid), $\Delta T_2 = [10, 20]$ (gray, solid) and $\Delta T_3 = [20, 30]$ (black, dashed).

ations are performed, and the distance between the background estimate and the cost function minimum.

4.4.3 Slow Versus Mixed-Timescale Observations

Having established that recovery of a balanced state is a very different process in 4DVAR than in the Kalman filters, we now want to see what this difference means for various assimilation parameters. The difference between slow and mixed-timescale observations is examined in figure 4.18, which shows average slow and fast analysis errors at observation times, and is analogous to figures 4.4 and 4.11. The two observation types, FIL (2.57) and MIX (2.58), are compared for experiments with $\Delta t^{\text{obs}} = 2$ and $\Delta T = 10$ (as in fig. 4.17). In these experiments, the minimization was cut off when the threshold $e_{w,\text{rms}} = 0.2$ was reached for each time window.

Note first that average 4DVAR fast errors are much lower than in the EKF, and almost as low as those in the EnKF, which benefits from ensemble averaging. Thus 4DVAR adds not just stability (as in chapter 3), but also helps to preserve balance. 4DVAR behaves quite differently from the EKF and EnKF for fast errors, showing smaller average fast errors for FIL observations, and more spurious imbalance for MIX observations, whereas the situation was reversed in the Kalman filters. This implies that balance is more or less preserved by assimilating slow observations in 4DVAR. For MIX, 4DVAR generates as much imbalance as needed to fit the mixed-timescale observations, with imbalance increasing depending on how far the initial forecast iterate lies from the cost function minimum.

While 4DVAR shows a large difference between observation types in fast-mode errors (like the EKF), errors for MIX and FIL are more or less similar for slow-mode errors (like the EnKF). Average slow-mode errors in 4DVAR are also similar in magnitude to EnKF errors (fig. 4.11). This suggests another advantage of the implicit covariance model of 4DVAR, which is that loss of balance does not have as great an impact on the slow-mode

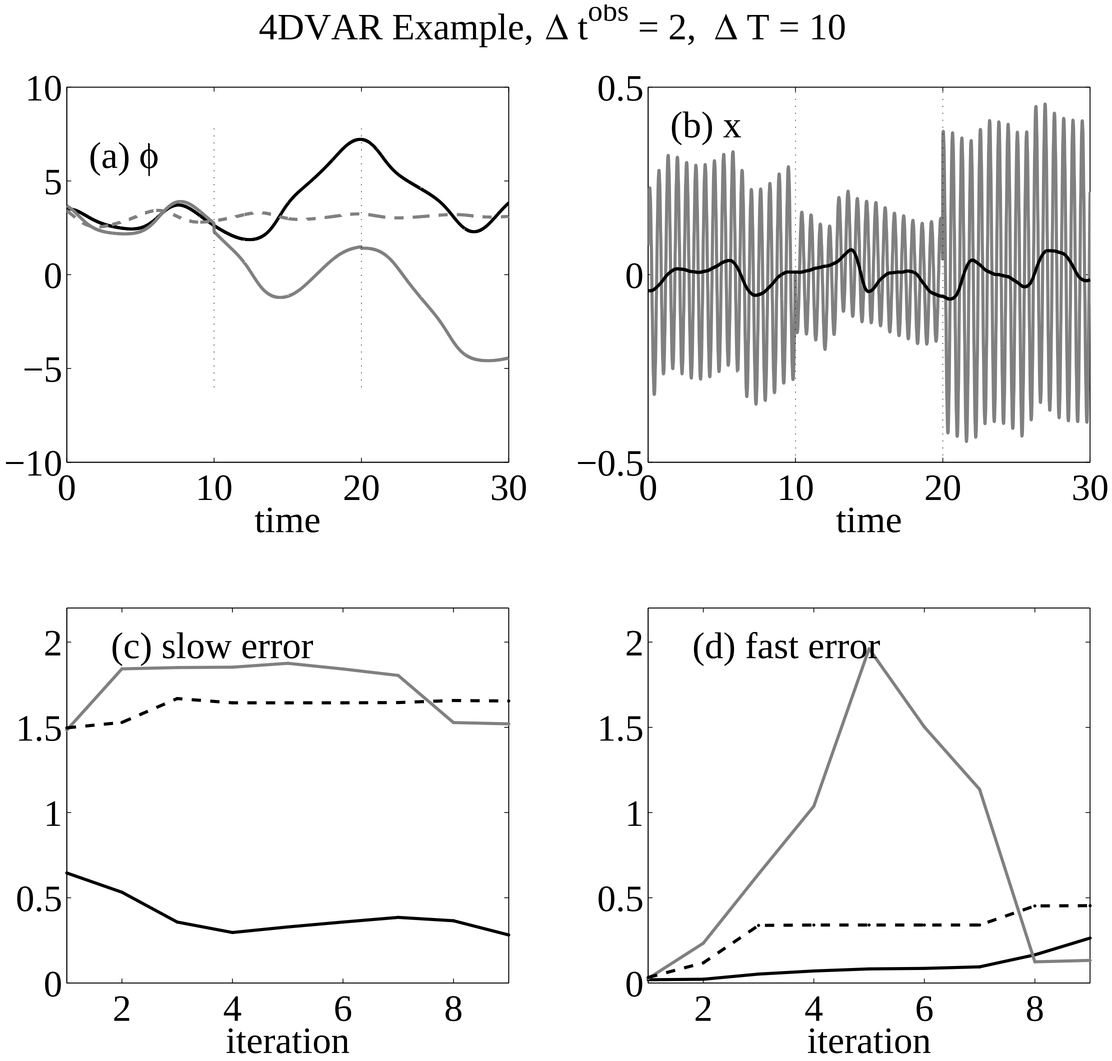

Figure 4.17: A 4DVAR assimilation experiment, with $\Delta t^{\text{obs}} = 2$ and $\Delta T = 10$. The truth (black) and analysis (gray, solid) are shown for ϕ in (a) and for x in (b). In (a), the initial background guess for ϕ is also shown (gray, dashed). (c) shows the slow-mode error over successive minimization iterations, for windows ΔT_1 (black, solid), ΔT_2 (gray, solid), and ΔT_3 (black, dashed). (d) shows the same, but for fast-mode error.

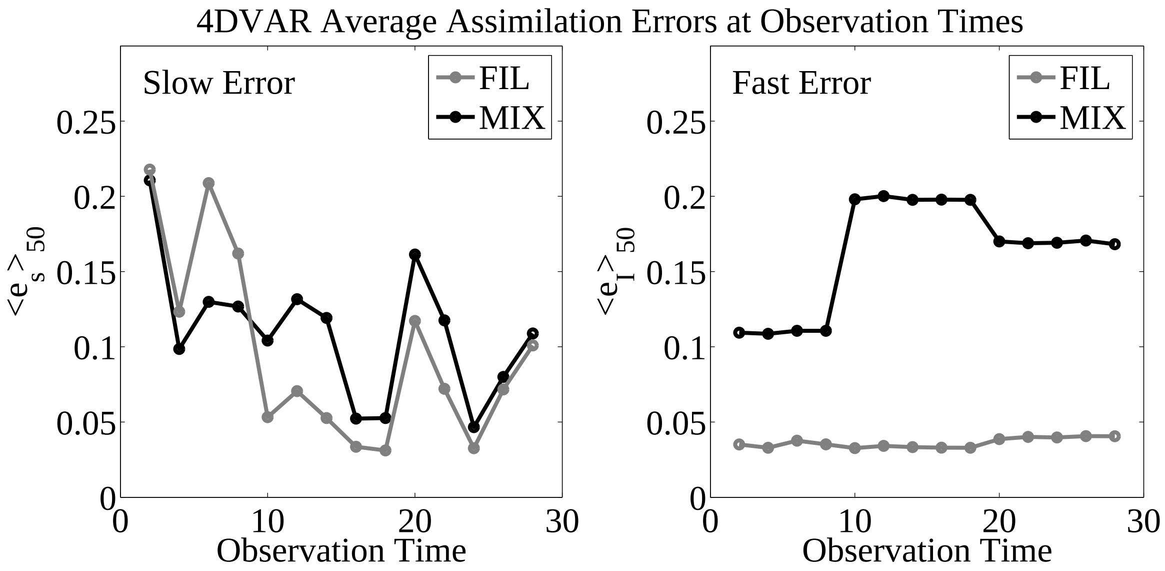

Figure 4.18: Average analysis errors for 4DVAR, over two types of experiments: observing the slow state ϕ and w (gray, FIL) and observing the mixed-timescale state ϕ and w' (black, MIX). Errors are split into slow (left) and fast (right) components. Observation times are indicated by circles.

analysis in 4DVAR as it does in the EKF.

4.4.4 Balance Constraint in the Background Error Covariance Matrix

In the experiments shown thus far, **B** was initialized as a diagonal matrix (4.25) and hence without any knowledge of balance. Despite this, assimilation experiments in figures 4.14 and 4.18 both showed that 4DVAR induces less imbalance, on average, than the EKF does with a balance-initialized covariance matrix.

Of course, there is no reason why the background error covariance matrix in 4DVAR cannot also be constrained according to the linearized balance relationship (4.9). In addition to providing a balance constraint, balancing of **B** contracts the search space in the cost function minimization [Lorenc, 2003a], thereby speeding up convergence of the minimization (in realistic models). The same constraint can also be imposed by using the balance relationship to transform the model variables into mutually uncorrelated components, such that the background error covariance matrix becomes the identity matrix [Weaver et al., 2005]. Because $\mathbf{B}$ constrains the space of available analyses, this is considered a strong constraint. Balance can also be made a weak constraint by defining $\mathbf{B}$ with nonzero variance in the unbalanced components of fast variables.

However, even a balanced covariance matrix really only constrains the minimization insofar as the TLM and adjoint models are accurate. In the EKF, nonlinearity meant that $\mathbf{P}^{\mathrm{f}}_k$ did not stay tangent to the slow manifold. Analogously, we can expect that a balanced $\mathbf{B}$ will not guarantee a balanced analysis when both the model and the balance relationship are nonlinear. In fact, it has already been shown [Courtier and Talagrand, 1990, Polavarapu et al., 2000] that 4DVAR can yield an unbalanced analysis even when balance is enforced as a strong constraint, as long as observations have errors which project onto inertia-gravity waves. The simplicity of the exL86 model allows us to illustrate how this happens.

Figure 4.19 examines the impact of balancing $\mathbf{B}$ in terms of average errors. It is similar to figure 4.5, and compares 4DVAR slow and fast average errors at observation times for three formulations of $\mathbf{B}$: the TLB transformation (4.41), the approximate balance (4.40), and the diagonal matrix (4.25). In all experiments in this figure, $\Delta t^{\mathrm{obs}} = 2$ and $\Delta T = 10$. Comparison of fast errors for the three cases shows that the different formulations of $\mathbf{B}$ have only a slight effect, with TLB and VAGUE yielding slightly lower average fast errors, while there is no discernible difference between the slow-mode errors.

It was shown for the EKF that the relative benefit of initializing the covariance matrix increases as the observation interval, and hence the nonlinearity of the assimilation system, increases. The same is examined for 4DVAR in figure 4.20, which shows average fast and slow errors as a function of Δt^{obs}, again comparing the three formulations of $\mathbf{B}$. As in figure 4.19, there is no clear difference in slow-mode errors for different for-

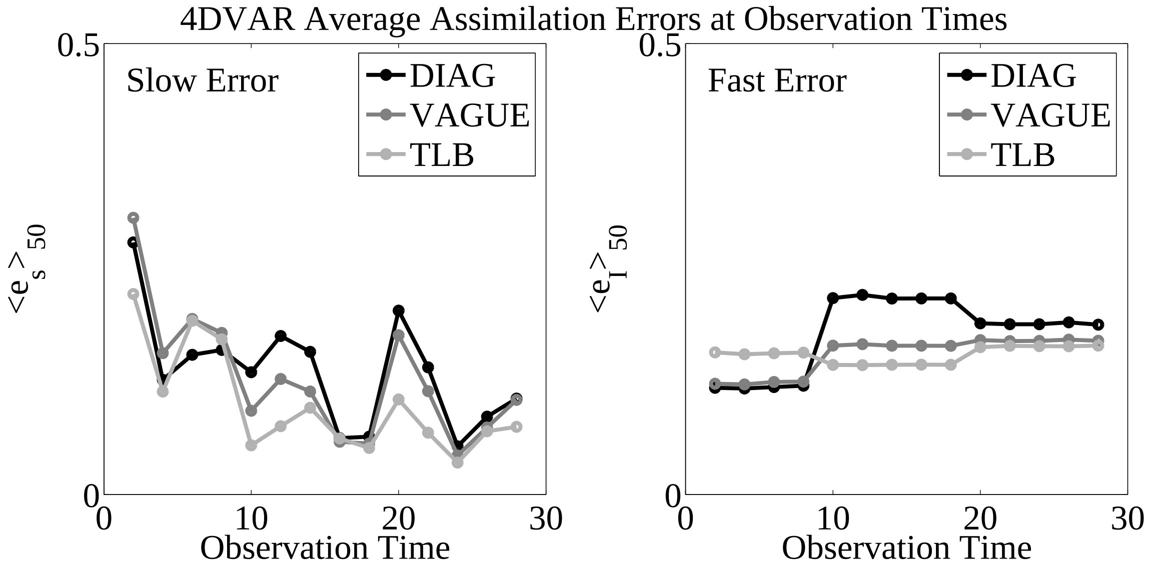

Figure 4.19: Average 4DVAR assimilation errors at observation times, for $\Delta t^{\mathrm{obs}} = 2$, for MIX observations. Each plot compares average slow [(2.59), left] and fast [(2.60), right] errors for three formulations of $\mathbf{P}_0^{\mathrm{f}}$: using the TLB transformation [(4.41), light gray], using the VAGUE approximation [(4.40), medium gray], and using the diagonal matrix [(4.25), black].

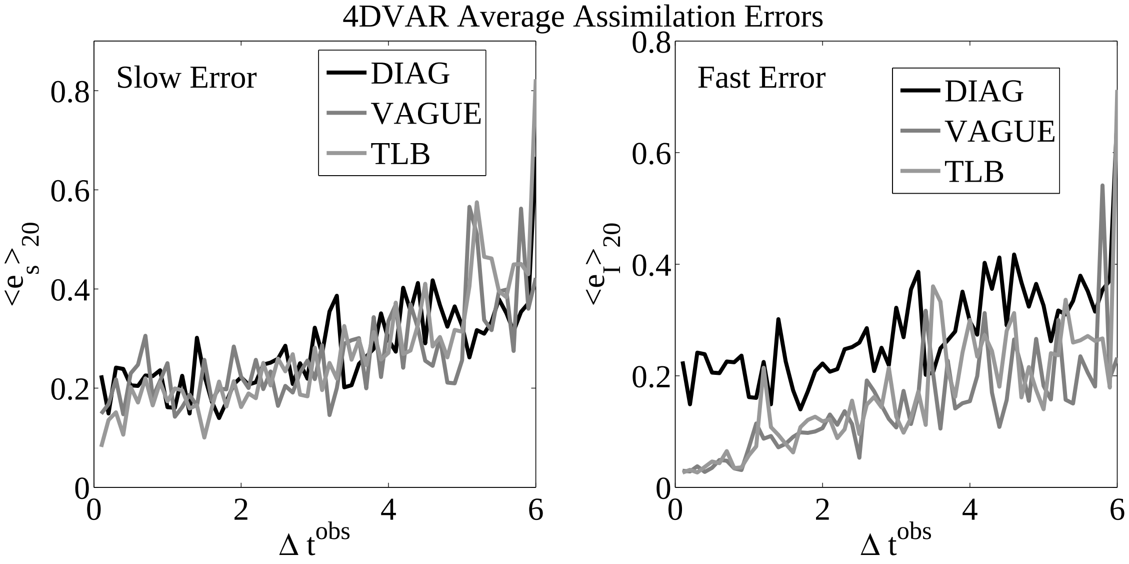

Figure 4.20: Average 4DVAR assimilation errors over a range of observation intervals, with MIX observations. Each plot compares average slow [(2.59), left] and fast [(2.60), right] errors for three formulations of $\mathbf{P}_0^{\mathrm{f}}$: using the TLB transformation [(4.41), light gray], using the VAGUE approximation [(4.40), medium gray], and using the diagonal matrix [(4.25), black].

mulations of $\mathbf{B}$. In terms of fast errors, the benefit of balance-initializing $\mathbf{B}$ decreases with increasing Δt^{obs}. This is the opposite of what happens in the EKF (fig. 4.6), where balance-initialization of $\mathbf{P}_0^{\mathrm{f}}$ is most effective at large Δt^{obs}. The difference is this: for the EKF, increasing Δt^{obs} is more likely to cause instability in the fast analysis, which can be alleviated if covariances are balance-initialized. In 4DVAR, increasing Δt^{obs} means that the minimization is harder to constrain, but 4DVAR does not share the EKF's gravity wave instability, because the covariance model is not explicitly evolved and updated. Even though the adjoint of the TLM is used to find the cost function minimum (2.50), 4DVAR integrates the full model at each minimization iteration. Therefore 4DVAR, like the EnKF, uses (in part) the actual model dynamics to infer balance. Thus even at $\Delta t^{\mathrm{obs}} = 6$, 4DVAR fast errors are on average lower than corresponding EKF errors at $\Delta t^{\mathrm{obs}} = 4$ (fig. 4.6).

4.4.5 Assimilation Window

Nonlinearity can also be increased in 4DVAR by increasing the size of the observation window. If ΔT is too large, the TLM and its adjoint become poor approximations. If ΔT is too small, the state is not fit to enough observations at the same time. Figure 4.21 (A) examines average slow and fast errors as a function of ΔT, for experiments with $\Delta t^{\mathrm{obs}} = 2$, mixed-timescale observations, and $\mathbf{B}$ formulated with the TLB approximation (4.41). In these experiments, the minimization cutoff threshold is at $e_{w,\mathrm{rms}} = 0.2$. Here we find that there is a trade-off between capturing the slow mode and capturing balance: slow-mode error is lowest for $\Delta T \sim 6$ or 7, while fast error simply decreases with ΔT.

The optimal time window for the slow mode makes sense, since it corresponds to the time of validity of the TLM. For windows longer than that, nonlinearity begins to affect the minimization, and the cost function will cease to be quadratic. This is illustrated in figure 4.21 (B), which shows slices of the cost function along ϕ, for a balanced example state where the truth is given by the initial slow state $(\phi_0, w_0)^{\mathrm{T}} = (0.19, 0.21)^{\mathrm{T}}$. The cost

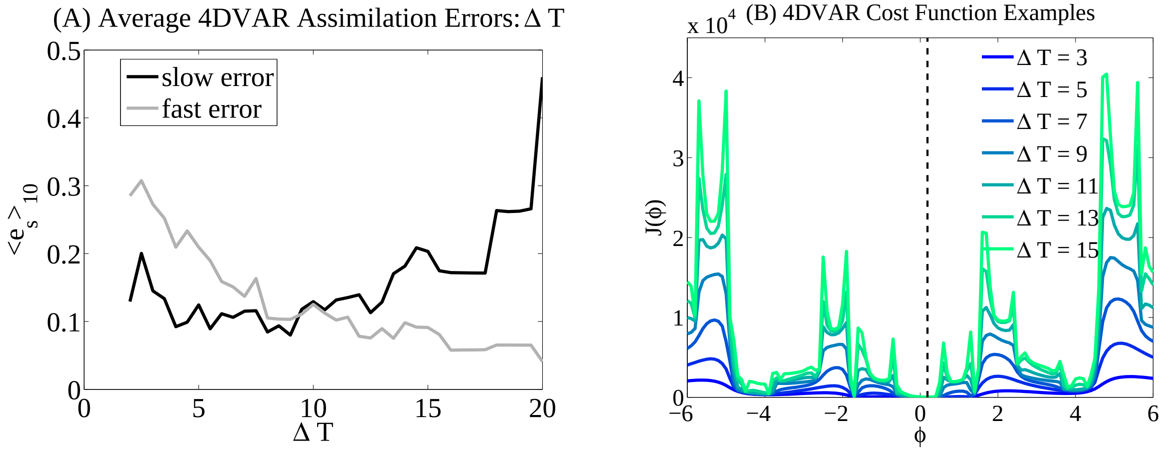

Figure 4.21: (A) Average 4DVAR assimilation errors, as a function of the assimilation window, comparing average slow (black) and fast errors (gray), for experiments with $\Delta t^{\text{obs}} = 2$, MIX observations, and $\mathbf{B}$ computed with (4.41). (B) Slices of the 4DVAR cost function along ϕ, for an example case with $\Delta t^{\text{obs}} = 2$, MIX observations, and a variety of minimization windows ΔT. The true initial state, $\phi_0 = 0.1865$, is indicated by a dashed line.

function displays local minima for all ΔT shown, becoming increasingly jagged as ΔT is increased. If the estimate settles into a local minimum of the cost function, subsequent iterations will not bring it closer to the observation. Since less adjustment of the slow mode means that less spurious imbalance is induced in the fast mode, analyses with long ΔT are more balanced.

4.4.6 Other Balance Constraints in 4DVAR

There are other ways to add a balance constraint to the minimization, such as by adding a slow manifold projection to the TLM and its adjoint. This was done by Courtier and Talagrand [1990] using nonlinear normal mode initialization and by Polavarapu et al. [2000] using a digital filter, and is analogous to the balance constraints added to the EKF in §4.2.5, though discussing the fine details of this similarity is beyond the scope of this study. Another way to add a balance constraint is suggested by Courtier and

Talagrand [1990], who propose adding a penalty term to the cost function minimization, and then reinitializing each iterate. This issue has been further studied by Thépaut and Courtier [1991], Polavarapu et al. [2000], and Gauthier and Thépaut [2001], and we refer to those studies, while noting that even without an extra balance constraint, 4DVAR in our experiments is still preferable to both the EKF and EnKF in recovering a balanced state.

4.5 Summary and Discussion

In this chapter, experiments with the exL86 model showed that a perfectly-balanced state is generally difficult to recover in 4D data assimilation, outside of the ideal case of very frequent observations.

While 4DVAR and the EnKF will lose balance if too much error accrues in the assimilation, both methods turn out to be significantly more stable than the EKF, in regards to balance, with 4DVAR yielding the most balanced analyses, especially as observations become less frequent. The advantages of the EnKF and 4DVAR come from use of the nonlinear model to infer (some) covariance evolution. Even when pushed to their limits (i.e. by decreasing ensemble size or observation frequency, or increasing the number of minimization iterations in 4DVAR) both methods by-and-large are improvements over the EKF, though the EnKF becomes somewhat worse than the EKF when the ensemble is smaller than the state dimension.

A few more comments must be made to connect these results to more realistic problems. The generation of unconstrained imbalance from observations of entirely slow variables, which is then not controlled because the fast variables are not observed, echoes a similar result by Lea et al. [2006], who found that spurious inertial oscillations in an ocean model were induced by the assimilation because the observations (sea surface height, temperature, and salinity) contained almost no information about that type of motion.

For the EKF and 4DVAR in our experiments, balance was improved somewhat by balance-constraining the initial error covariance matrix, but this generally had no impact on recovery of the slow mode. Interestingly, Weaver et al. [2005] (again in an ocean model) found that balance-constraining the initial background error covariance matrix in 4DVAR significantly decreased errors relative to 3DVAR for all fields, and suggested that constraining the cost function in a balanced way somehow helps to unleash the full power of 4DVAR. That study did not discuss the effect of model nonlinearity, however; our results suggest that, for highly nonlinear contexts, the constraint imposed by a balanced covariance matrix is small.

Lorenc [2003b] has argued that 4DVAR is practically preferable to the EnKF in terms of imbalance, because it is possible to balance-constrain the cost function minimization. The experiments of this chapter, however, showed that nonlinearity limits the effectiveness of balance constraints in 4DVAR, and that there may be regions in the space of assimilation parameters where the EnKF returns a more balanced analysis than 4DVAR (e.g. $\Delta t^{\mathrm{obs}} < 1$), without any external constraints.

The improvement of the EKF when a balance constraint was added to the analysis suggests that a simplified algorithm such as the one suggested by Dee [1991], where balanced covariances are incorporated manually into the assimilation cycle, can yield better results than the full EKF. This idea has similarity to the use of so-called perturbation forecast models (PFMs) in place of the TLM, albeit in the context of 4DVAR [Lorenc, 2003a, Rabier, 2005]. PFMs are designed to model the evolution of an approximately Gaussian pdf for motion at a certain range of scales, while purposely excluding scales about which little is known. Clever use of PFMs and additional balance constraints could make it possible to better preserve balance in realistic applications of 4DVAR. The fact that 4DVAR was an improvement over the EKF even without these additional constraints is encouraging.

It was also found that the EnKF is susceptible to a form of filter divergence in terms of balance, wherein the forecast ensemble is close to the truth, but tightly clustered around a spurious gravity wave, because the ensemble members have locked phases [e.g. figure 4.9 (B)]. It can be argued that the EnKF will likely be more balanced in practice, where more than one gravity wave frequency is admitted, and gravity waves can propagate away or dissipate between observations [e.g. Szunyogh et al. 2005]. However, the results of Houtekamer and Mitchell [1998] suggest that spurious imbalance in the ensemble mean analysis is difficult to avoid for realistic ensemble sizes, and for observation intervals at or near realistic gravity wave timescales, where gravity waves have less time to propagate away. This leaves the somewhat unfortunate result that more frequent observations could cause greater imbalance and ensemble phase-locking, as in figure 4.12, though it remains to be seen whether this is the case for more complex models.

Overall, the experiments of this chapter achieved two purposes: (1) they illustrated the initialization problem, and its interpretation in sequential/variational assimilation, in a straightforward and simple environment (finding that 4DVAR preserved balance more, overall) and (2) they serve as preparation for the next chapter, which deals with physical cases where the true state is not balanced, and about which much less is understood from a data assimilation perspective.

Chapter 5

Gravity Waves in the Truth

The experiments in the previous chapter showed that the problems associated with misestimated fast-slow correlations can be alleviated to some extent by adding a balance constraint to the assimilation, either by balance-constraining the initial covariance matrix, or by adding a slow manifold projection to the analysis step. The EKF in particular could be made far more useful by adding a balance constraint to the analysis. Such constraints will no longer be useful, however, if the true state is not balanced. This is the case in the mesosphere and upper stratosphere, where flows are dominated by gravity waves [Koshyk et al., 1999], and in the tropics, where the timescale separation between vortical modes, gravity waves, and equatorial waves becomes unclear [Žagar et al., 2004a].

The issues of an unbalanced truth and/or unclear timescale separation therefore raise the question of how well 4D assimilation schemes can develop the right covariance representation for cases where traditional balance constraints do not represent the full system. Szunyogh et al. [2005] report a case of an ensemble-based 4D method capturing a real gravity wave, and suggest that well-formulated 4D covariance models have the potential to capture gravity waves which are present in the truth but not in the model. For tropical data assimilation, Žagar et al. [2004a] showed that recovery of the full state from partial observations requires accurate representation of special tropical wave solutions in the background covariance model. However, the development in time of accurate covariance models in either case is not yet well understood.

A related question concerns observations: when the slaving relationship does not represent the full dynamics, and since different observations project differently onto vortical and inertia-gravity wave modes, it is not clear which observation types are most useful for recovery of the full state. Observation frequency is another complicating factor, since inertia-gravity waves have timescales (hours to the inertial period) which are similar to or faster than typical data assimilation intervals (usually 6 or 12 hours). It is easy to see that, if observations are assimilated roughly once in a fast period, it will be difficult to glean the wave's magnitude and frequency (unless there is sufficient background knowledge about the true gravity wave).

Unbalanced Truth Experiments

In the experiments of this chapter, the truth is chosen to contain a free gravity wave with magnitude $\tilde{I}^{\mathrm{t}} = \sqrt{\tilde{x}^2 + \tilde{z}^2}$ and frequency ϵ, while the forecast is initialized with $\tilde{I}^{\mathrm{f}} = 0$. In the EKF and 4DVAR experiments, $\mathbf{P}_0^{\mathrm{f}}$ and $\mathbf{B}$ are estimated as follows:

$$\mathbf{P}_0^{\mathrm{f}} = \mathbf{L}\mathbf{P}_{\mathbf{y}}^{\mathrm{f}}\mathbf{L}^{\mathrm{T}} + \mathbf{P}_{\mathrm{GW}}. \tag{5.1}$$

The first term is the TLB approximation derived in chapter 4, while the second term,

$$\mathbf{P}_{\mathrm{GW}} = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & \sigma_{\tilde{x}}^2 & 0 \\ 0 & 0 & 0 & \sigma_{\tilde{z}}^2 \end{pmatrix} \tag{5.2}$$

represents the free gravity wave component of initial fast variable variances, with

$$\sigma_{\tilde{x}}^2 = \sigma_{\tilde{z}}^2 = \frac{\sigma_{\tilde{I}}^2}{2}. \tag{5.3}$$

In the EnKF, ensemble members are initialized with free gravity wave magnitudes chosen randomly from $\tilde{I}_i \sim \mathcal{N}\left(0, \sigma_{\tilde{I},0}^2\right)$, and with gravity wave phases θ_i chosen randomly from a uniform distribution between $[0, \pi]$, such that the initial ensemble mean gravity wave has zero amplitude. Since initial forecast states are balanced in all experiments, the actual initial free gravity wave error is equal to the true free gravity wave magnitude, and we therefore set $\sigma^2_{\tilde{I},0} = \left(\tilde{I}^{\mathrm{t}}\right)^2$ in both filters. The sensitivity of assimilation errors to this formulation is examined below.

The relevant error measures in these cases are the slow mode error (2.59), the fast error (2.60), and the error in the gravity wave phase (2.61). The no-skill error for the slow mode in these experiments remains ~ 1.6, as in (3.3). For fast error, no-skill means that no gravity wave is forced in the initially-balanced analysis, in which case the average fast error is

$$e_I = \langle \left(\tilde{I}^{\mathrm{t}} + I^{\mathrm{t}}_{\mathrm{slav}} - \tilde{I}^{\mathrm{a}} + I^{\mathrm{a}}_{\mathrm{slav}} \right)^2 \rangle^{1/2} \simeq \tilde{I}^{\mathrm{t}}, \tag{5.4}$$

where the slaved components of the fast mode in the analysis and truth, $I^{\mathrm{t}}_{\mathrm{slav}}$ and $I^{\mathrm{a}}_{\mathrm{slav}}$, are small compared to $\tilde{I}^{\mathrm{t}}$. For gravity wave phase, the no-skill level is $\pi/2 \simeq 1.5$, the RMS of random gravity wave phase errors distributed normally between 0 and π.

Fast and Slow Analysis Increments

In chapter 4, it was shown that the fast-slow covariances which result from the balance relationship are needed in order to correctly recover information about the slow mode from observations of the mixed-timescale variables, and to prevent the excitation of a spurious wave in the fast variables. Since only the balanced components of the fast variables can be gleaned from slow variable observations, the fast mode can now be only partially recovered in the FIL case. Fast increments in this case have the form (4.12), which shows that the fast-slow covariances must be captured in order to correctly update the slaved component of the fast mode.

To capture the free gravity wave from mixed-timescale observations [where fast analysis increments have the form (4.16)] the unbalanced components of fast-variable error variances (σ_z^2 and σ_x^2) must be large enough. For the unbalanced truth case, the free gravity wave variance components (which obey approximately linear dynamics) will likely dominate, so that fast variables should, in principle, be easier to capture. In the presence of a gravity wave, recovery of the slow mode from mixed-timescale observations involves analysis increments of the form (4.15) and thus again requires accurate modeling of fast-slow covariances.

5.1 Gravity Waves in the Extended Kalman Filter

5.1.1 Example

Figure 5.1 shows the EKF analysis for an example state with gravity wave parameters $\tilde{I}^{\mathrm{t}} = 1.5$, $\epsilon^{-1} = 10$, and $b = 0.71$. Observations are assimilated every $\Delta t^{\mathrm{obs}} = 3$ time units, comparing MIX observations (2.58) in column (A) and FIL observations (2.57) in column (B). The truth and analysis for ϕ and x are shown in the top panels, and the true square fast-mode error (e_I^2, black) is compared to the filter-estimated fast error variance (σ_I^2, gray) in the bottom panel.

In (A) the slow mode and gravity wave phase are more or less captured by the end of the assimilation period. However, the x analysis is generally overforced to a gravity wave magnitude which is too large, and then not sufficiently corrected by subsequent observations. Because the gravity wave is linear, σ_I^2 (lower left panel) matches the true square error over the interval of time prior to the first observation. Nonetheless, the EKF overreduces estimated fast error variance at observation times, and eventually diverges. Since fast mode variance was estimated correctly prior to the first observation, the overreduction of fast mode variance and the associated filter divergence must (as in the previous chapter) result from the error in EKF-estimated covariances. Covariance estimation error can also be seen in the corresponding ϕ analysis: though ϕ^{a} is pushed towards observations at every observation time, it deviates more and more from the truth

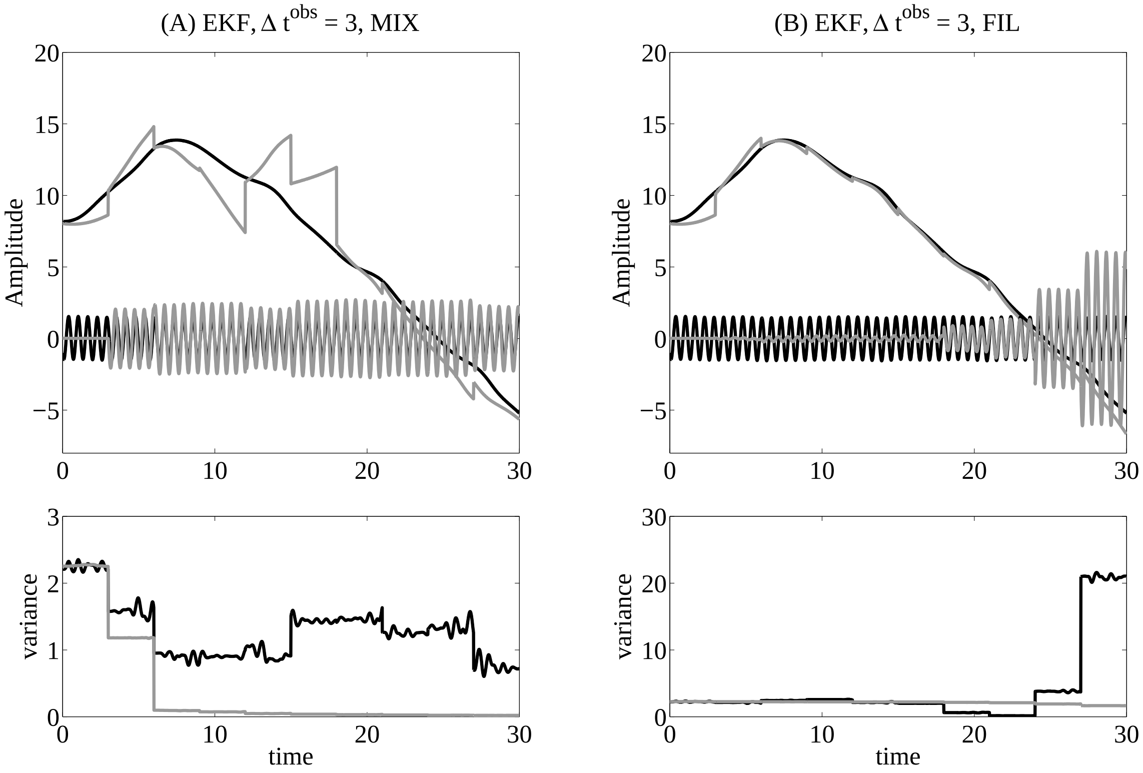

Figure 5.1: The true state (black) and EKF analyses (gray) of $x(t)$ and $\phi(t)$, for two experiments with a true gravity wave magnitude of $\tilde{I}^{\mathrm{t}} = 1.5$. Mixed-timescale observations are assimilated in (A) and slow observations in (B). The lower panel in each column shows the true square fast error e_I^2 (black) and the associated error variance estimate σ_I^2 (gray).

between observations —indicating that the tendency of the entire state is estimated less accurately in the EKF.

Ignoring the free gravity wave (B) improves the analysis of ϕ, at the cost of accuracy in x. While it is to be expected that the gravity wave cannot be fully captured in the FIL case, here the analysis induces an accidental gravity wave, which doesn't match the true gravity wave (except in frequency) and which, more importantly, is not controlled by observations and thus allowed to grow. Since information about the balanced components [(2.11)-(2.12)] of the fast variables can only be gleaned from observations of the slow state to the extent that the balance relationship is sufficiently represented by the forecast error covariance matrix, not observing the fast variables can be quite harmful if the balance relationship is *not* captured accurately (as we know is the case in the EKF). This is essentially the same instability as was identified in chapter 4, but in a different context. It is an especially salient point in the present context, because observations in physical situations where gravity waves are large and significant tend to be sparse and irregular in time [Polavarapu et al., 2005]. We thus see that (1) lack of observed fast variable information, as in chapter 4, can induce spurious unbalanced motion, and (2) even with fast variable observations, there is filter divergence in the gravity wave, despite its linear dynamics.

5.1.2 Observation Type and Interval

The result shown in figure 5.1 is examined more quantitatively in figure 5.2, which shows actual and estimated errors at observation times, comparing average quantities for FIL and MIX, divided into slow mode error (2.59), fast-mode error (2.60), and gravity wave phase error (2.61). Error due to use of the TLM can again be seen in EKF slow mode errors, which become larger than estimated errors after the first observation. For FIL, fast errors grow in time and gravity wave phase errors stay at the no-skill level. While it is to be expected that the gravity wave cannot be fully captured in the FIL case, growth

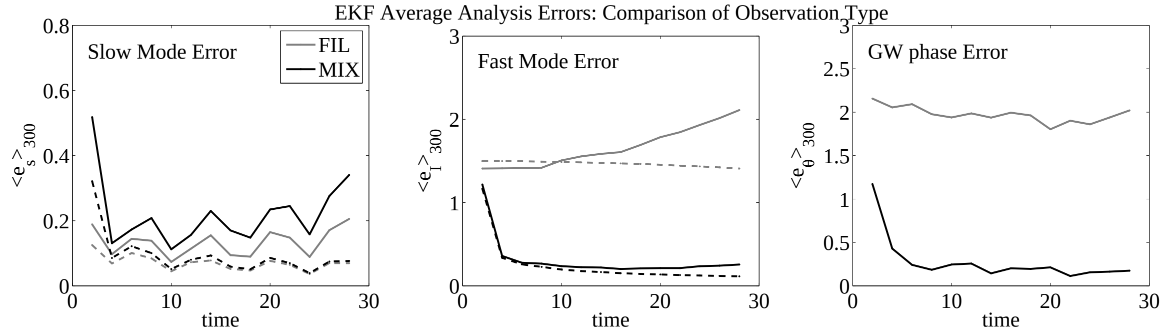

Figure 5.2: Actual EKF analysis errors (solid lines) and predicted errors (dashed lines) at observation times for $\Delta t^{\text{obs}} = 2$, with errors divided into the three components [(2.59)-(2.61)]. Each plot compares FIL observations [(2.57), gray] and MIX observations [(2.58), black].

of $\langle e_I \rangle_{300}$ in each filter indicates spurious projection onto the fast mode, as in figure 4.4 for the balanced truth case. Since estimated errors are updated only in proportion to estimated fast-slow covariances, actual errors in both filters then diverge from estimated errors. However, both $\langle e_I \rangle_{300}$ and $\langle e_\theta \rangle_{300}$ on average decrease in time for MIX, indicating that the fast wave *can* be captured, though actual fast errors still tend to be larger than estimated fast errors.

Figure 5.3 examines how the difference between the observation types changes when the assimilation system is made more nonlinear by increasing Δt^{obs}. Errors for the MIX case are also compared to a set of OI experiments (MIX-OI), with mixed-timescale observations and the same initial covariance matrix (5.1). For MIX, recovery of the slow mode becomes worse as Δt^{obs} increases. Recovery of the slow mode is improved for FIL, but here fast-mode error, due to the spurious projection of the slow mode onto the fast mode, becomes even larger for $\Delta t^{\text{obs}} > 2$. Average fast mode and gravity wave phase errors grow with Δt^{obs} for the MIX case, almost to the level of FIL, even though the gravity wave is partially observed. OI, which involves no incoming cycling from observations, has lower errors in both the slow and fast modes at large Δt^{obs}.

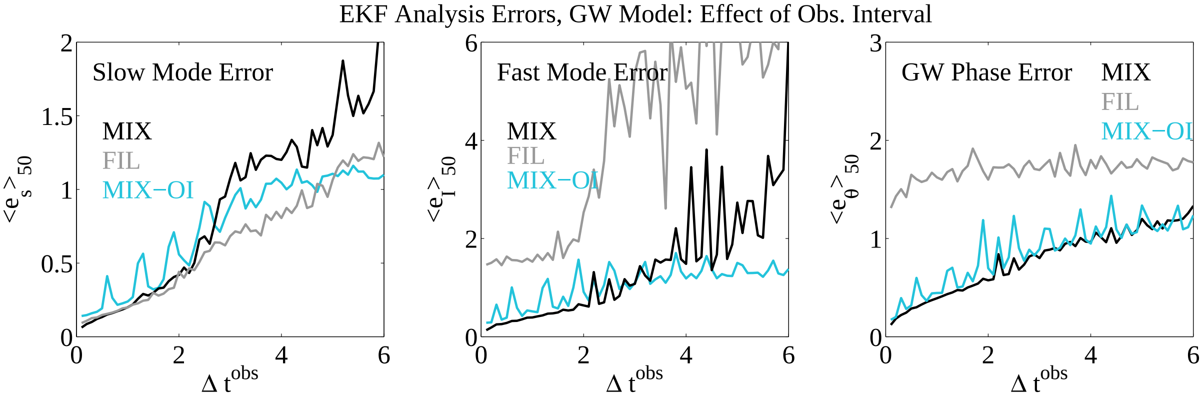

Figure 5.3: Average analysis errors for the EKF, divided into the three components [(2.59)-(2.61)] and shown for a range of observation intervals. Slow observations [(2.57), gray] are compared to mixed-timescale observations [(2.58), black]. These experiment sets are also compared to average OI errors for mixed-timescale observations (blue).

However, the EKF and OI are still on average able to achieve error reduction in the gravity wave phase for all observation intervals shown, if observations are of mixed-timescale. This is because the gravity wave is linear, and gravity wave phase is bounded. For MIX, the EKF also offers some improvement over OI for $\Delta t^{\rm obs} < 2$, particularly when $\Delta t^{\rm obs}$ is a multiple of the gravity wave period, $T_{\rm GW} = 2\pi\epsilon \simeq 0.63$. An increase in error is to be expected in that case, since observing a wave once every cycle does not give any information about the wave amplitude or phase. Since the same peaks do not appear in the average EKF errors, evolution of covariances in time thus helps overcome the bias created by poor temporal sampling of a gravity wave.

5.1.3 Formulation of the Initial Covariance Matrix

We might ask whether the problem of capturing two separate modes can be better conditioned by judiciously choosing the initial covariance matrix, as was done in chapter 4. For $\tilde{I}^{\rm t} = 1.5$, we found that the induction of the spurious gravity wave from FIL observations is reduced somewhat by the TLB-initialization of off-diagonal terms in $\mathbf{P}_0^{\rm f}$ (5.1),

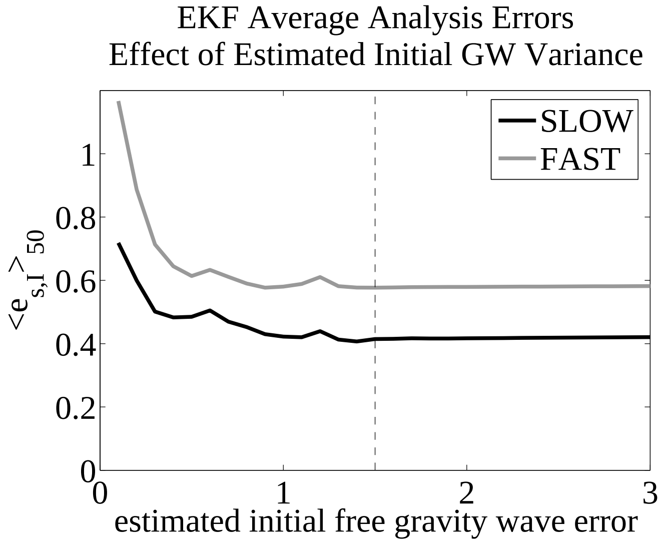

Figure 5.4: Average slow [(2.59), black] and fast [(2.60), gray] errors for the EKF, as a function of the gravity wave magnitude standard deviation ($\sigma_{\tilde{I},0}$) estimated in the initial covariance matrix. The true initial free gravity wave error is indicated by the vertical dashed line.

but this result is not plotted since it is essentially similar to the same result found for the EKF in the balanced truth case (fig. 4.6). Furthermore, the EKF's recovery of the slow state from mixed-timescale observations is found not to be sensitive to the initialization of fast-slow covariances.

It is possible, however, to improve the EKF's recovery of either mode by changing the estimated initial variance attributed to the free gravity wave in (5.2)-(5.3). Figure 5.4 shows average slow and fast errors for the EKF as a function of $\sigma_{\tilde{I},0}$. Both errors decrease as $\sigma_{\tilde{I},0}$ increases, but are essentially independent of this quantity once it is sufficiently large. Gravity wave phase errors are omitted from this figure because they look similar to the other two measures.

Comparison of estimated and actual fast variable error variances in time for these experiments shows that the EKF usually overreduces forecast error variances in the analysis step (as in the lower panel of figure 5.1). Since the gravity wave doesn't grow between

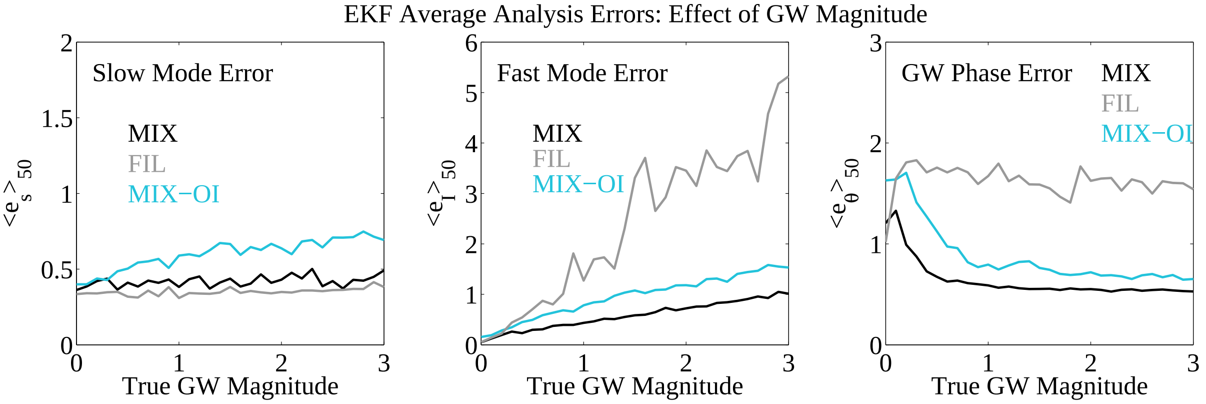

Figure 5.5: Average EKF analysis errors divided into the three components [(2.59)-(2.61)], over a range of true gravity wave magnitudes, $\tilde{I}^{\rm t}$, comparing MIX [(2.58), black] and FIL [(2.57), gray] observations, and a set of OI experiments (MIX-OI, blue) with MIX observations and the same initial covariance matrix as EKF.

observations, the overreduced fast error variances continue to underestimate true error variance, and subsequent observations are not given enough weight. Starting from an initial fast variable error variance that is much larger than the actual variance delays the point where fast variable errors become underestimated, by a few observation times. At the same time, a larger initial free gravity wave error variance means that the adjustment of this variance will be stronger during the analysis step (2.36). This in turn causes the estimated variances to be overreduced. As a result, increasing the initial estimate of $\sigma_{\tilde{I},0}$ tends to reduce errors only to an extent, so that average errors cannot really be decreased for $\sigma_{\tilde{I},0} > 0.5$. Thus, as in the balanced-truth case, the EKF is sensitive to initial formulation of the covariance model, but has a tendency to diverge as the assimilation cycle winds on.

5.1.4 Effect of Free Gravity Wave Magnitude

The free gravity wave can be made more significant for the evolution of the full system if its magnitude ($\tilde{I}^{\rm t}$) is increased. Figure 5.5 examines the effect of increasing free gravity wave magnitude for the EKF and the two observation types. The MIX case is again compared to a corresponding set of OI runs with corresponding covariance matrix and observations.

For MIX, $\langle e_I \rangle$ grows with $\tilde{I}^{\mathrm{t}}$ because the initial forecast is balanced, and it takes longer for the assimilation cycle to generate a gravity wave as large as the true gravity wave. In this case, gravity wave phase errors decrease with increasing gravity wave magnitude, simply because a larger free gravity wave magnitude means that the ratio between observation error and the gravity wave signal is decreased. Here the EKF has slightly lower errors than OI though, since the gravity wave is linear and therefore easier to capture, the improvement of $\langle e_\theta \rangle$ saturates rather quickly. Interestingly, MIX-OI errors clearly do increase with the magnitude of the true gravity wave. This indicates a benefit to the EKF's covariance model, despite its shortcomings.

For FIL, slow mode errors can be expected to increase with the magnitude of the free gravity wave, since the influence of the free gravity wave on the evolution of the slow mode is not captured in this case. However, though fast errors still generally exceed the no-skill levels for FIL, this does not affect slow mode errors, which show no clear sensitivity to $\tilde{I}^{\mathrm{t}}$. Even for $\tilde{I}^{\mathrm{t}} = 3$, a free gravity wave that is twice as large as that shown in the example (fig. 5.1), recovery of the slow mode is still better if observations of the full slow mode, rather than observations which contain the gravity wave signal, are assimilated.

5.1.5 Gravity Wave Period

The role of the gravity wave in the evolution of the system can also be changed by changing the timescale separation parameter ϵ, or the coupling parameter, b (see next section). Increasing ϵ means that the gravity wave becomes slower relative to the slow mode, while the balanced components of x and z become larger, and the balance relationships [(2.11)-(2.12)] more nonlinear.

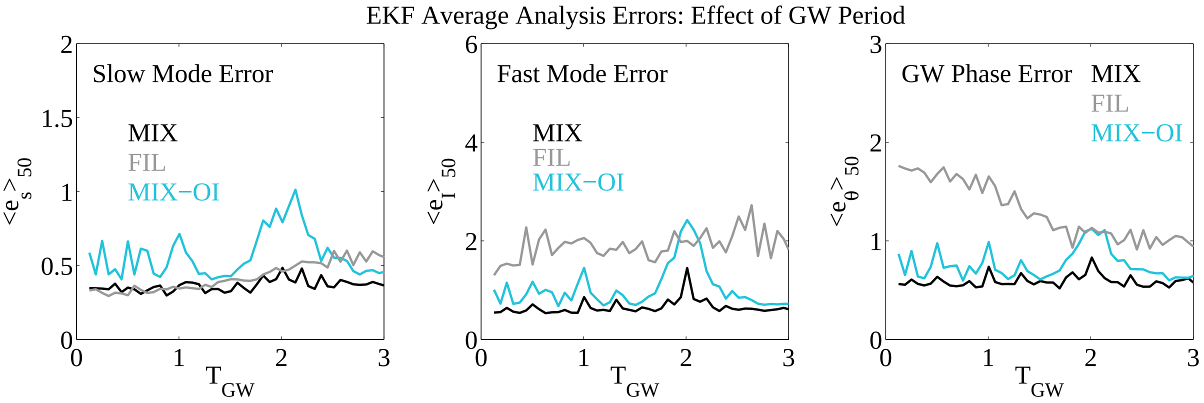

Figure 5.6: As in figure 5.5, but now showing errors as a function of the period of the true-state gravity wave.

Figure 5.6 compares average true errors for sets of experiments where ϵ is increased, comparing MIX to FIL as a function of the period of the gravity wave ($T_{\text{GW}} = 2\pi\epsilon$), and again comparing MIX to a set of OI experiments (MIX-OI). For FIL, average slow error increases with T_{GW}, with errors surpassing the MIX case for T_{GW} greater than about 1 or 1.5. This means that it is harder to ignore the gravity wave as timescales move closer together. This makes sense, since slow mode analysis increments become larger if the ignored gravity wave has a larger impact on the evolution of the slow mode. For larger analysis increments, covariance estimation error is amplified.

One might expect slow errors to decrease with increasing T_{GW} for mixed-timescale observations, since observations of w' contain more information about the slow mode at larger ϵ. Since slow mode errors for MIX don't decrease on average, it seems that most of the error in the MIX case comes from not observing the entire vortical mode, and can't be greatly improved by increasing the slaved component of the fast mode. Gravity wave phase errors, however, do decrease with increasing ϵ for FIL, reflecting the fact that a greater component of the fast variables is slaved to the slow mode, and thus informed by slow observations, as ϵ increases.

In the MIX case there are also peaks in all three error measures at $T_{\text{GW}} = 1$ and 2,

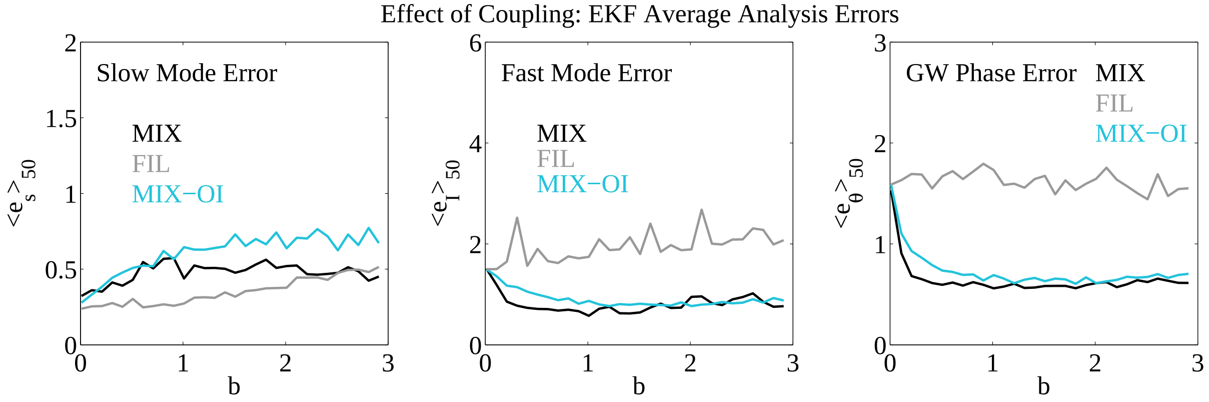

Figure 5.7: As in figure 5.6, but now showing errors as a function of the coupling parameter, b.

where the gravity wave period is commensurable with $\Delta t^{\mathrm{obs}} = 2$. These peaks correspond to the increased errors for OI in figure 5.3, and though they also appear for the EKF here, they are larger for OI. This indicates again that the evolution of the covariance matrix helps to reduce the effect of this bias, because information about the true gravity wave is taken up and spread forward in time.

5.1.6 Slow-fast Coupling

Figure 5.7 compares the effect of changing b while keeping ϵ and $\tilde{I}^{\mathrm{t}}$ fixed. b also changes the nonlinearity of the balance relationship and hence the accuracy of estimated correlations [fig. 4.1], but without changing the timescale of the gravity wave. This again means that the evolution of the slow mode between observations depends more heavily on the fast variables, while the slaved components of the fast varlables become slightly larger. b also changes the information content of observations. As b increases, w' becomes more gravity wave dominated, while z' becomes more slow mode dominated. This means that MIX observations, as defined in (2.58), contain increasingly more information about the gravity wave, and less about the slow mode.

As in the previous figure, average slow mode error increases for FIL as the gravity wave becomes more important for the evolution of the whole system. There is also, again, an advantage in the EKF relative to OI: though MIX observations contain less information about the slow mode as b increases, $\langle e_s \rangle$ only clearly increases with b for OI. It is somewhat surprising that the EKF can achieve this, even though estimated fast-slow covariances are fairly poor.

Fast mode and gravity wave phase errors both decrease with increasing b in the MIX case, an obvious result since more information about the gravity wave is contained in the observations as b increases. Even though the balanced components of fast variables are increased slightly, they are still of order ϵ, so about a tenth of the full gravity wave magnitude. Thus, gravity wave magnitude and phase error do not decrease with increasing b (as they did for increasing ϵ) in the FIL case.

5.2 Gravity Waves in the Ensemble Kalman Filter

We now turn our attention back to the EnKF, which was shown in chapter 4 to have more accurate fast-slow covariances, but which also clearly benefitted from averaging to produce a balanced analysis. For an unbalanced truth, the challenge is instead to capture both fast and slow motions with accurate amplitudes.

5.2.1 Example

Figure 5.8 shows the analyses from an 8-member EnKF, corresponding to the example shown in figure 5.1, again comparing MIX and FIL observations.

In the case of mixed-timescale observations (A), the ensemble locks onto the true gravity wave magnitude and phase within a few observations. In contrast to the corresponding EKF case [fig. 5.1 (A)], EnKF-estimated fast mode variances are also a closer approximation to the true square error in the fast mode, though this error is still underestimated somewhat. Changing to FIL observations, the slow mode analysis is improved slightly

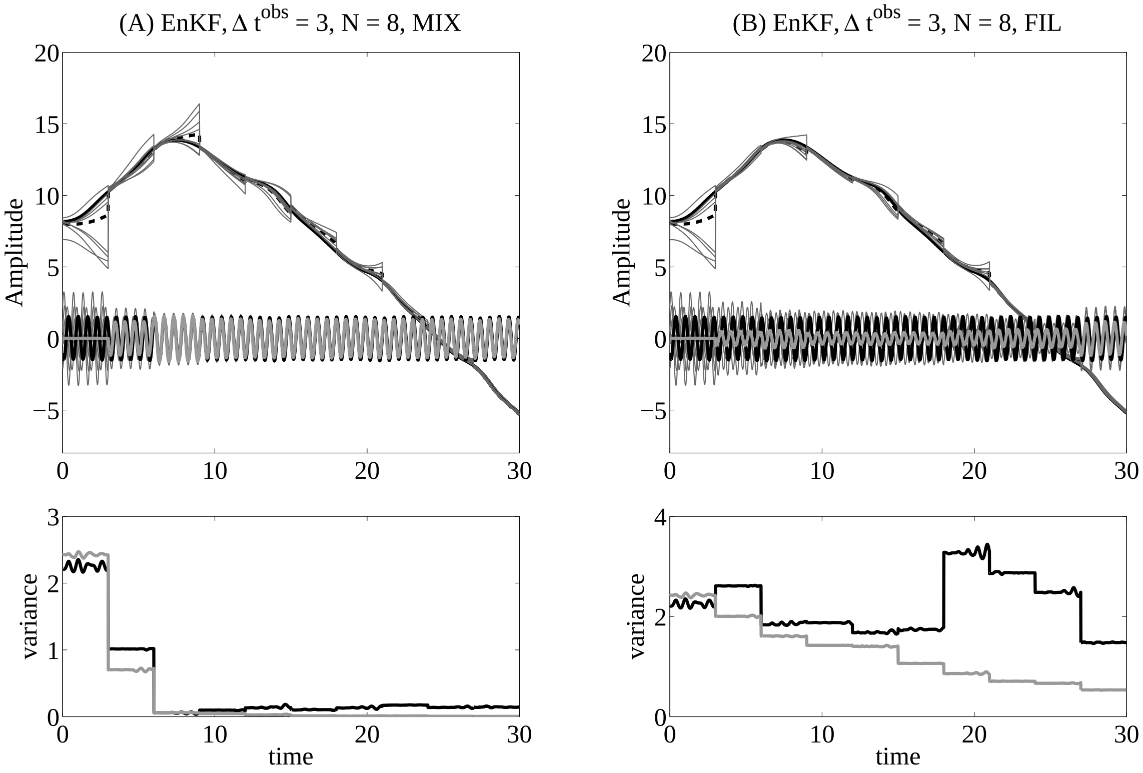

Figure 5.8: The true state (black) and EnKF analyses (dashed line) of $x(t)$ and $\phi(t)$, for two EnKF experiments with a true gravity wave magnitude of $\tilde{I}^{\mathrm{t}} = 1.5$. Mixed-timescale observations are assimilated in (A) and slow observations in (B). The ensemble in each case is shown in gray. The lower panel in each column shows the square fast error in gravity wave magnitude e_I^2 (black) and the associated error variance estimate σ_I^2 (gray).

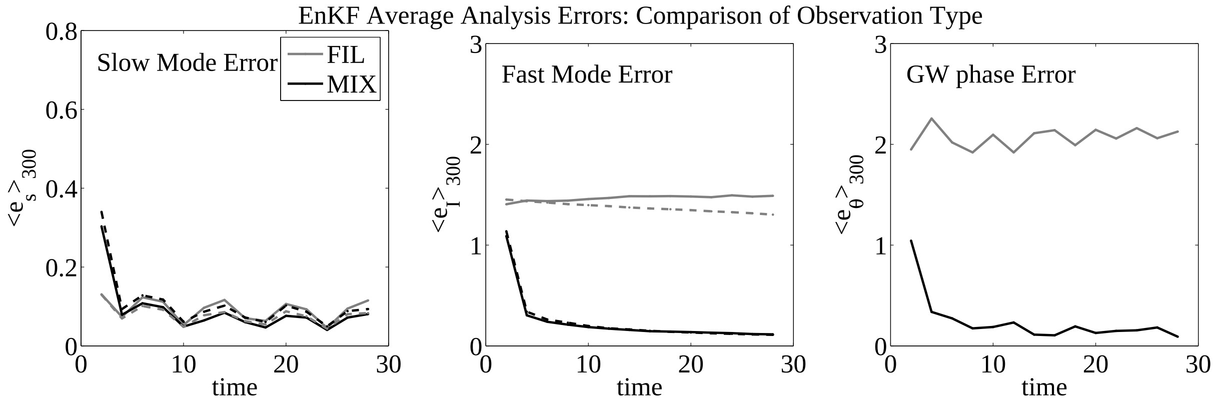

Figure 5.9: Actual EnKF analysis errors (solid lines) and predicted errors (dashed lines) at observation times for $\Delta t^{\text{obs}} = 2$, with errors divided into the three components [(2.59)-(2.61)]. Each plot compares FIL observations [(2.57), gray] and MIX observations [(2.58), black].

while the gravity wave is not captured. Assimilation reduces the initially-present gravity waves in individual ensemble members, but they do begin to phase-lock somewhat, leaving a net (spurious) gravity wave. We thus see a possible advantage of ensemble-estimated covariances, in that both modes can be captured from mixed-timescale observations, while the spurious gravity wave is reduced (though not eliminated) for slow observations.

5.2.2 Observation Type and Interval

The difference between the two observation configurations is examined in terms of average errors in figure 5.9, which shows actual and estimated errors at observation times for MIX and FIL, again divided into the three error components, as in figure 5.2. In this case, a 50-member EnKF is used, so that the covariance model is as accurate as possible. There are two clear differences from the EKF case: First, there is less difference between estimated and true fast error in FIL, indicating, again, less spurious projection onto the fast mode from slow observations. A new result is that slow mode errors show no clear difference between the two observation types, indicating that the slow mode is retrieved

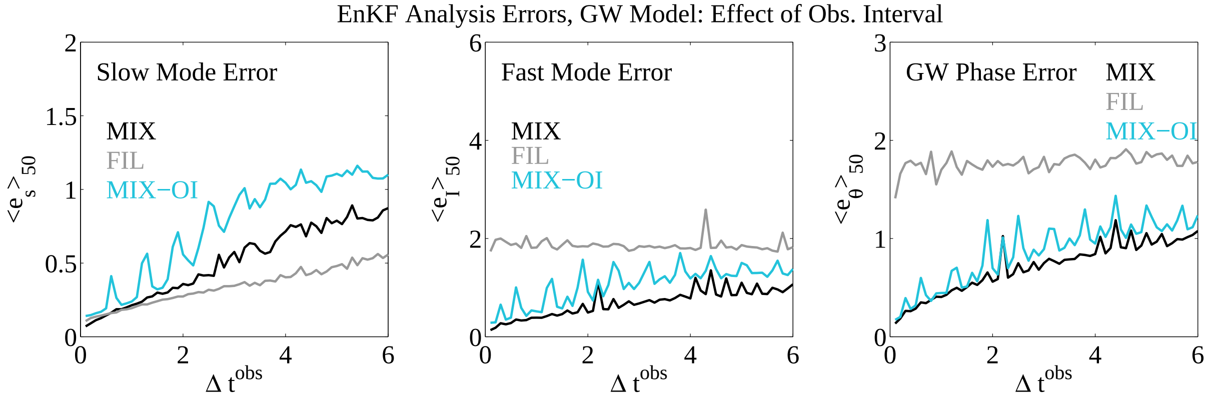

Figure 5.10: Average analysis errors for the EnKF, divided into the three components [(2.59)-(2.61)], and shown for a range of observation intervals. FIL observations [(2.57), gray] are compared to MIX observations [(2.58), black] and OI with MIX observations (blue).

from mixed-timescale observations as though the full slow mode had been observed.

Again casting this result into different regimes of nonlinearity, figure 5.10 compares the two observation configurations across increasing observation interval, for an EnKF with 50 ensemble members, again comparing MIX to FIL. For comparison with the corresponding EKF figure (fig. 5.3), the axes have been kept the same and the set of MIX-OI experiments is again included. Here the EnKF again shows the stability shown in chapters 3 and 4: at Δt^{obs} much larger than what is allowed for the TLM to be a valid approximation, the EnKF still returns errors that are low relative to OI, for both observation types. There is also, as in chapter 4, very little uncontrolled projection onto the gravity wave; fast-mode and gravity wave phase errors remain near the no-skill levels for FIL, and are significantly lower for MIX. Recovery of the slow mode in the 50-member EnKF is not as sensitive to the two observation types as in the EKF, though the difference between FIL and MIX increases with increasing Δt^{obs}. Again, we see that retrieval of the slow mode from mixed-timescale observations becomes more difficult as covariance estimation error increases.

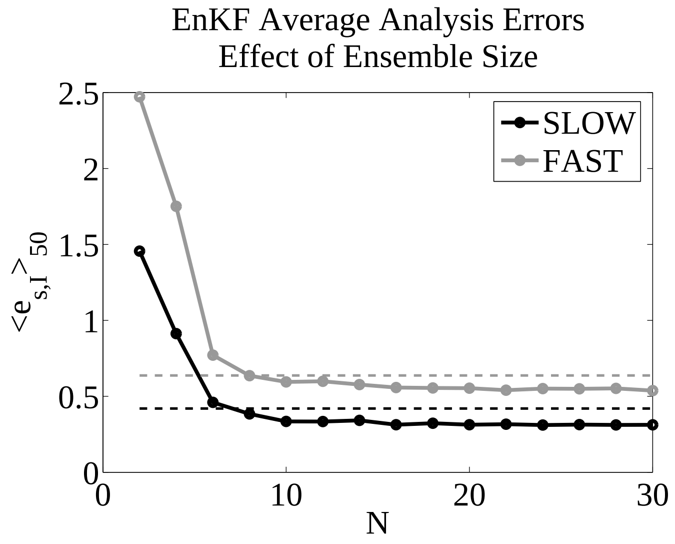

Figure 5.11: Average true slow [(2.59), black] and fast errors [(2.60), gray] for the EnKF as a function of ensemble size, N, with $\tilde{I}^{\mathrm{t}} = 1.5$. The average EKF slow and fast magnitude errors for $\Delta t^{\mathrm{obs}} = 2$ are shown by dashed lines.

5.2.3 Ensemble Size and Initial Covariance Model

The ensemble size in the experiments above is, of course, quite large. Average EnKF slow and fast errors are plotted over a range of ensemble sizes in figure 5.11, for experiments with $\Delta t^{\mathrm{obs}} = 2$ and MIX observations. For comparison, the average EKF slow and fast errors for $\Delta t^{\mathrm{obs}} = 2$ are also shown by dashed lines. Both error measures (as well as gravity wave phase error, not shown here) stop decreasing for ensembles greater than about 15 members, with little difference between 10 and 15 ensemble members. For the slow mode, the EnKF becomes comparable to the EKF at about 6 ensemble members; for the fast mode, it is about 8. The difference implies that the (almost linear) gravity wave is slightly more difficult to capture from the MIX observation configuration as defined by (2.58), and is more sensitive to the accuracy of covariance estimates.

It was also found, though not shown, that estimation error due to sampling becomes clearer when $\Delta t^{\mathrm{obs}} \lesssim 1$, with the EnKF diverging slightly more easily here than for intermediate observation intervals, as in figure 4.13. In this case, to get more accurate

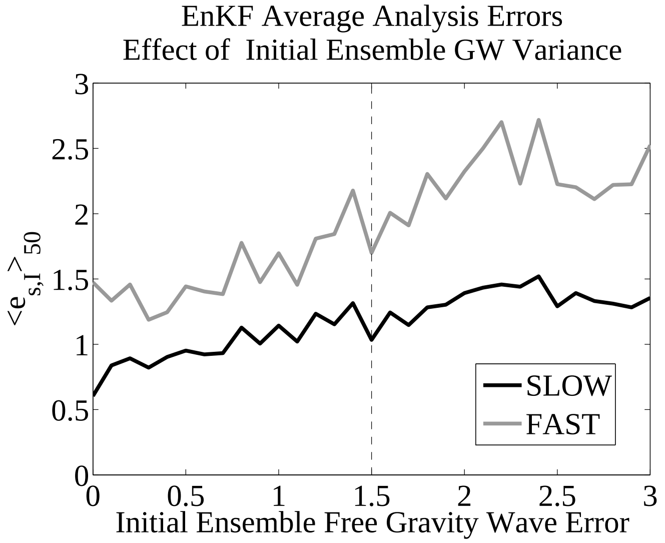

Figure 5.12: Average true slow mode error [(2.59), black] and fast error [(2.60), gray] for the EnKF as a function of the gravity wave magnitude standard deviation ($\sigma_{\tilde{I},0}$) with which the ensemble is generated. The true initial free gravity wave error is indicated by a line.

estimated error statistics, the observation interval needs to be increased, which decreases analysis quality. The EnKF has a tendency to diverge because the ensemble phase-locks around the wrong gravity wave. Whereas the ensemble spreads naturally between observation times in the chaotic slow mode, it does not in the nearly-linear gravity wave.

This leads again to the question of whether filter divergence can be alleviated by changing the initial (pre-assimilation) covariance model. Initial estimated fast variable variances in the EnKF can be controlled by changing the distribution of free gravity wave magnitudes in the initial ensemble. This is done in figure 5.12, which shows average slow and fast errors over a range of $\sigma_{\tilde{I},0}$, the standard deviation of the distribution from which the free gravity waves in the initial ensemble are generated, with $N = 50$. This figure is suprisingly different from the corresponding EKF experiment (fig. 5.4). Here, average fast errors are actually lowest for $\sigma_{\tilde{I}} \sim 0.5$, which is an underestimation of the initial error in the free gravity wave magnitude ($e_{\tilde{I}} = \tilde{I}^{\mathrm{t}} = 1.5$), and tend to increase for estimates larger than that. Figure 5.12 actually shows two competing effects: while increasing $\sigma_{\tilde{I},0}$ increases the weight applied to observations of the fast mode and thus tends to reduce gravity wave errors, it increases slow mode errors. The explanation for this is found in the examination of individual experiments, which show that a strong analysis increment in the fast variables (due to a wide distribution of free gravity wave amplitudes in the ensemble) can overreduce the spread of the ensemble in the slow mode, causing an overreduction of ensemble spread for subsequent observation intervals. This in turn harms the gravity wave analysis, causing the $\langle e_I \rangle$ curve in figure 5.12 to increase again, with an "optimal" $\sigma_{\tilde{I}}$ of about 0.5.

It must be mentioned that this result only happens on average, and that increasing $\sigma_{\tilde{I},0}$ improves the analyses of both modes in some realizations. The main point made by figures 5.4 and 5.12 is that overestimation of error variance for one mode can cause overreduction of error variances overall, and can therefore be detrimental to the recovery of the other mode. Furthermore, we find again that the reactions of two different assimilation systems to similar changes in data assimilation parameters can be quite different. The experiments of this section suggest that divergence of the EnKF may be quite difficult to prevent for small ensembles.

5.2.4 Effect of Free Gravity Wave Magnitude

It remains to examine the EnKF with respect to the parameters which define the gravity wave. The effect of free gravity wave magnitude on the EnKF analysis is examined in figure 5.13, which compares average analysis errors for MIX and FIL observations, using a 15-member EnKF.

Like the EKF, the EnKF slow mode analysis shows no sensitivity to the magnitude of the free gravity wave in slow mode errors. While fast errors stay at the no-skill level for FIL observations, they are significantly below the no-skill level for MIX, and also lower than corresponding EKF errors (fig. 5.5). This means that the (large-ensemble) EnKF

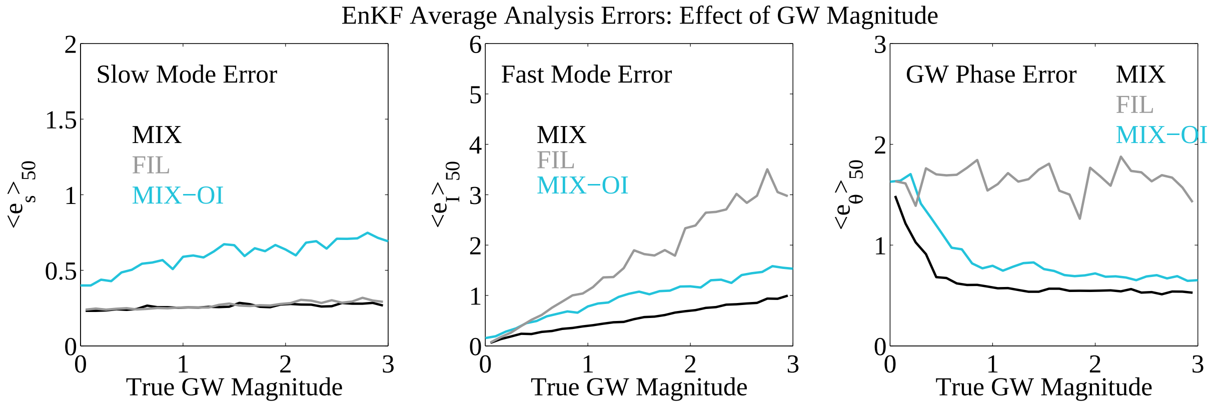

Figure 5.13: Average analysis errors for a 15-member EnKF, divided into the three components [(2.59)-(2.61)], over a range of true gravity wave magnitudes, $\tilde{I}^{\mathrm{t}}$, comparing MIX [(2.58), black] and FIL [(2.57), gray] observations, and OI with MIX observations (blue).

adjusts the ensemble-mean from a zero initial gravity wave to the correct gravity wave magnitude more easily than does the EKF. As in the EKF, gravity wave phase error decreases with $\tilde{I}^{\mathrm{t}}$ for mixed-timescale observations, again because of a presumably lower error-to-signal ratio. All in all, figure 5.13 shows that the EnKF in general is better able to capture either mode when both are observed, though, as implied by figure 5.11, this still requires a fairly large ensemble.

5.2.5 Gravity Wave Period and Slow-Fast Coupling

Figure 5.14 compares average MIX and FIL analysis errors over a range of gravity wave periods, for a 15-member EnKF. This figure is similar to the corresponding EKF figure (fig. 5.14) in that MIX gives lower slow mode errors than FIL as the impact of the gravity wave on the evolution of the slow mode is increased (by making the gravity waves slower). The difference between observation types is less pronounced in the EnKF, indicating that the larger analysis increments, which amplify covariance estimation error, are less critical here.

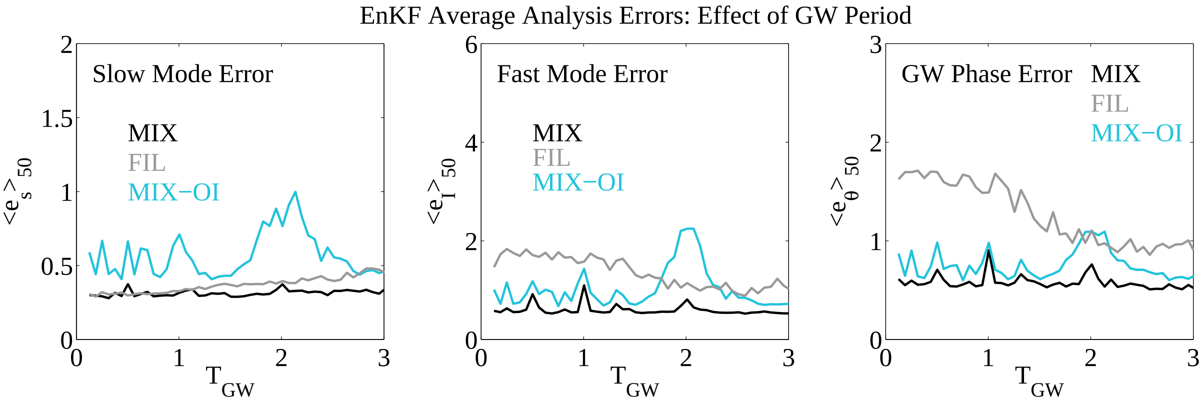

Figure 5.14: As in figure 5.13, but now showing errors as a function of the period of the true-state gravity wave.

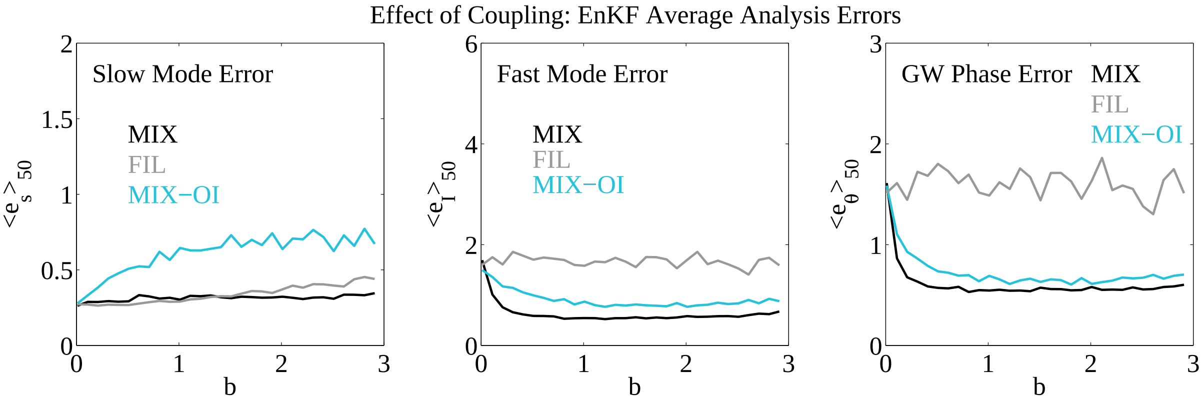

Figure 5.15: As in figure 5.14, but now showing errors as a function of the coupling parameter, b.

Figure 5.15 shows average EnKF errors as a function of the coupling parameter, b. The result is basically similar to the corresponding EKF result (fig. 5.14), with the exception that there is, again, less difference between slow and mixed-timescale observations in slow mode errors. The similarity of the EKF and EnKF errors over changing ϵ and b means that the effective nonlinearity of the assimilation system doesn't change as these parameters change in the presence of a free gravity wave in the truth —though the effectiveness of flow-dependent covariances clearly does increase.

5.3 Gravity Waves in 4DVAR

So far we have seen that the EKF and EnKF are both able to disentangle the fast and slow modes, as long as both are observed and observations are frequent, though even then the EnKF is better able to capture the free gravity wave. We now compare this result to 4DVAR. 4DVAR was shown in chapter 4 to yield more balanced states, in part because its implicit covariance model is more stable than that of the Kalman filters over large Δt^{obs}, but also because the minimization was cut off before too much spurious imbalance could be induced.

5.3.1 Example

Figure 5.16 (a) shows the 4DVAR analyses of ϕ and x for the MIX experiment shown in figures 5.1 (A) and 5.8 (A), with the assimilation period divided into three minimization windows of size $\Delta T = 10$. The same experiment is repeated in figure 5.17, but now assimilating slow observations (2.57). Also shown in each figure is the reduction of errors in the slow mode (b) and fast mode (c) over successive minimization iterations for each window. The minimization is cut off at nine iterations in this example.

In both cases, convergence of the slow mode analysis happens within three or four iterations for all three windows. For FIL (fig. 5.17), the free gravity is, of course, not recovered at all, but neither is there a spurious gravity wave. For MIX (fig. 5.16), the gravity wave magnitude is underestimated over the first time window, even in nine iterations. This is the window where the background estimate of ϕ is closest to the truth and hence requires the least adjustment. It can be seen [fig. 5.16 (c)] that adjustment of the fast mode over the iterations is consequently too small in this window. Note that in chapter 4 larger increments at the minimization iterations meant a greater loss of balance. Here, conversely, larger increments mean a greater ability to match the amplitude of the free gravity wave.

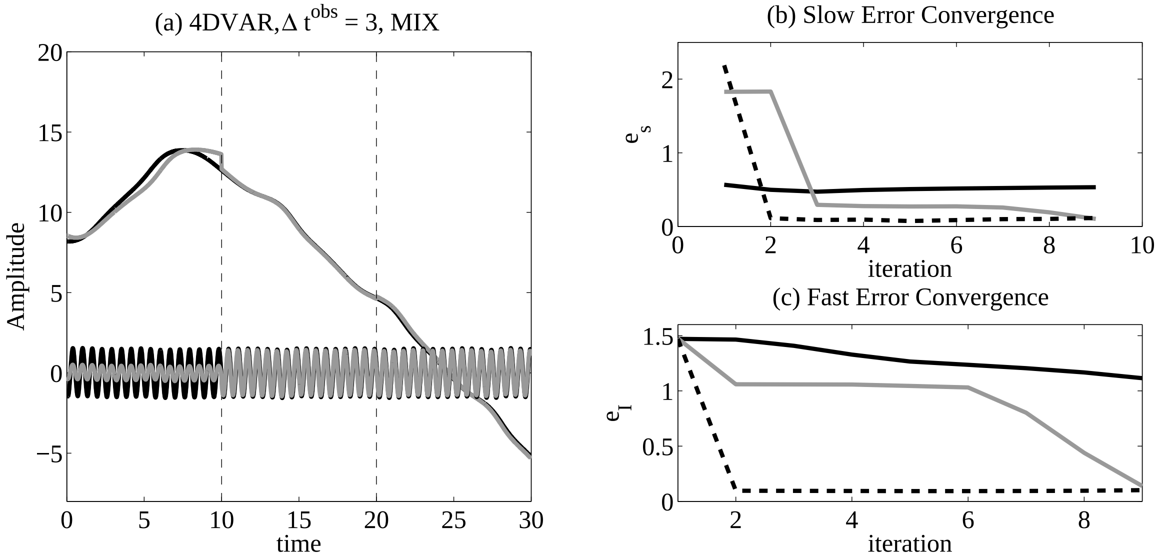

Figure 5.16: (a) The true state (black) and 4DVAR analysis (gray) of $x(t)$ and $\phi(t)$, for a sample case with a true gravity wave magnitude of $\tilde{I}_{\mathrm{t}} = 1.5$. Boundaries between the three minimization time windows ($\Delta T_1 = [0, 10]$, $\Delta T_2 = [10, 20]$, and $\Delta T_3 = [20, 30]$) are denoted by dashed lines. (b) Slow mode error as a function of minimization iteration, for the first time window (black, solid), the second time window (gray, solid), and the third time window (black, dashed). (c) As in (b), but for fast-mode error. In this case, mixed-timescale observations (2.58) are assimilated, and $\Delta t^{\mathrm{obs}} = 3$.

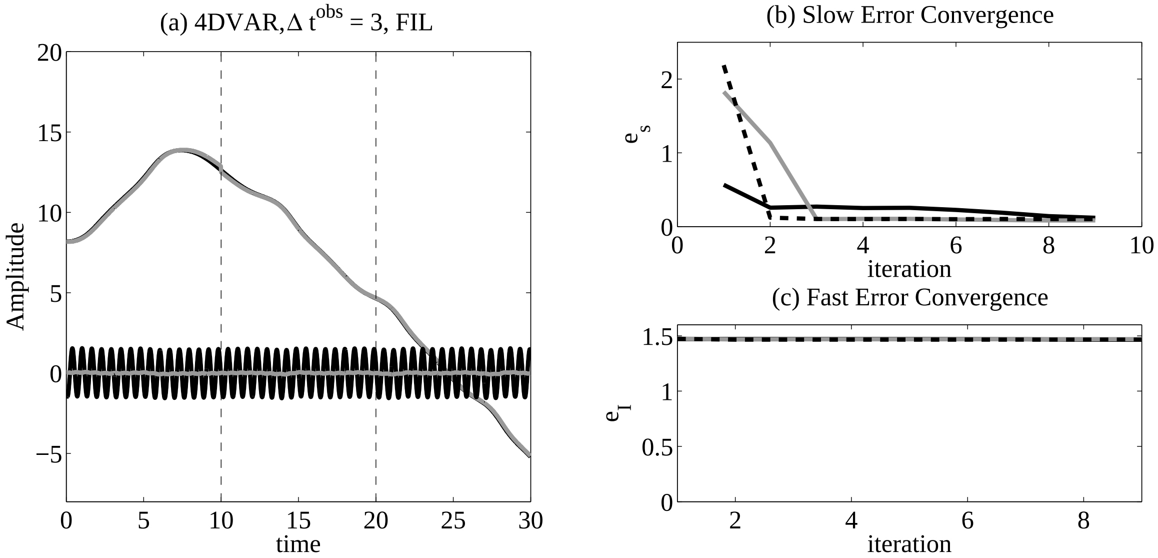

Figure 5.17: As in figure 5.16, but assimilating slow observations (2.57).

5.3.2 Observation Type

To compare the effect of observation variable, average errors at observation times for MIX and FIL are shown in figure 5.18. In these experiments, the minimization is again cut off at the threshold $e_{w,\mathrm{rms}} = 0.2$. As in the EnKF but in contrast to the EKF, there is no difference between the two observation configurations in average slow mode errors, indicating that (for this observation frequency) 4DVAR is able to recover the full slow mode even from mixed-timescale observations to the same level of accuracy as when slow observations are assimilated.

Recovery of the free gravity wave for MIX is extremely weak; this is a sign of the minimization iteration cutoff. A difference can also be seen between the three minimization windows; fast errors are decreased most in the second and third time windows, but not at all for the first time window. This is because the background estimate is closest to the truth in the first time window, and only requires one or two iterations to be brought below the slow error threshold, whereas more iterations are necessary to fit the slow mode in the second and third time windows, and these iterations also involve larger steps —thereby generating larger free gravity waves. Reduction of gravity wave phase

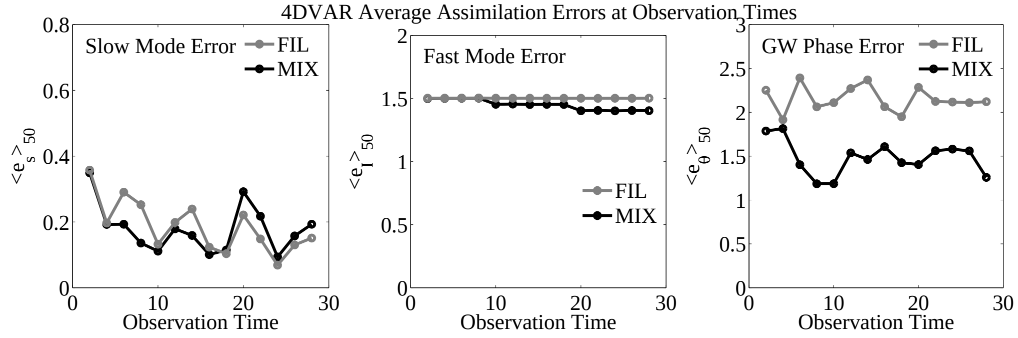

Figure 5.18: Average 4DVAR analysis errors at observation times for $\Delta t^{\text{obs}} = 2$, with errors divided into the three components [(2.59)-(2.61)]. Each plot compares FIL observations [(2.57), gray] and MIX observations [(2.58), black].

error is also much weaker in 4DVAR than in the Kalman filters.

Figure 5.19 shows average 4DVAR analysis errors as a function of observation interval, again comparing the two observation configurations to OI with mixed-timescale observations. We found that, in order to noticeably reduce fast errors below the no-skill level, 4DVAR had to be run out to at least 3 or 4 minimization iterations. Hence, the experiments in this figure are performed with 5 conjugate gradient iterations, and no cutoff threshold. For FIL, the slow mode shows no clear sensitivity to the observation interval. For MIX, recovery of the slow mode becomes slightly more difficult as Δt^{obs} increases. This tells us that the fast-slow covariances implied by the adjoint sensitivity calculation become worse estimates as nonlinearity in the assimilation increases. However, the difference between the observation types is still far smaller than in the EKF (fig. 5.19).

Because 4DVAR has almost no spurious projection of observational information onto the fast mode in the FIL case, average fast mode and gravity wave phase errors in this case also barely deviate from the no-skill level. For MIX, recovery of the fast mode magnitude is also more stable over Δt^{obs} than in the EKF, and is comparable to the EnKF for large Δt^{obs}. However, 4DVAR generally has difficulty finding the phase of

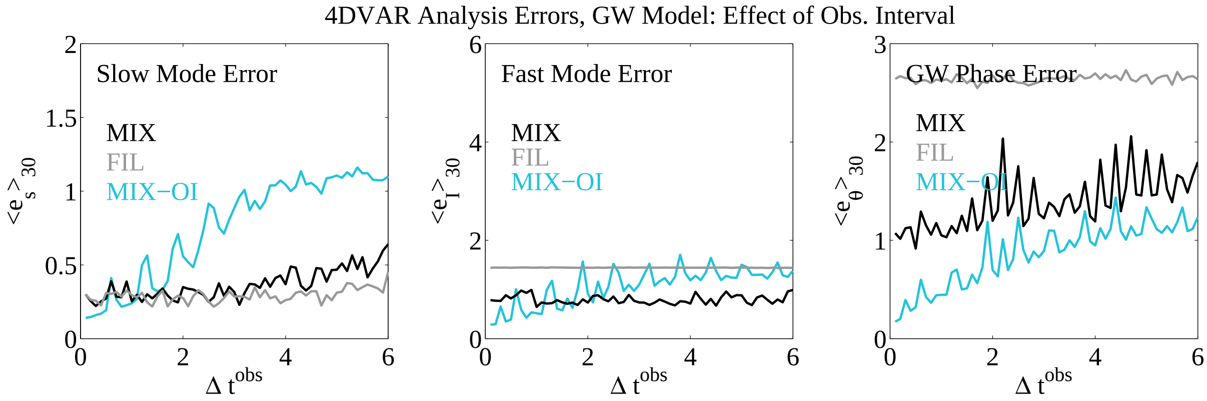

Figure 5.19: Average analysis errors for 4DVAR, divided into the three components [(2.59)-(2.61)], and shown for a range of observation intervals. FIL observations [(2.57), gray] are compared to MIX observations [(2.58), black] and OI with MIX observations (blue).

the free gravity wave, even though the minimization is run out to enough iterations to recover the fast mode magnitude. While average gravity wave phase error is less than the no-skill level for MIX, it is still much larger than in the corresponding EKF (fig. 5.3) and EnKF (fig. 5.10) cases, and is even larger than average errors for OI.

As shown in the previous two chapters, another way to increase nonlinearity of the 4DVAR analysis is to change the size of the minimization window, ΔT. In figure 5.20, the three error measures are examined as a function of ΔT. Again, the minimization is cut off at 5 iterations but with no slow error threshold. As in the balanced truth case (fig. 4.21), the optimal minimization window for the slow mode (as well as for gravity wave phase) is around $\Delta T = 7$. However, now having a longer minimization window also increases fast-mode error. Recall that larger ΔT means that the slow mode analysis is adjusted less towards the true state, because the cost function becomes nonlinear. While in chapter 4 this meant less excitation of a spurious gravity wave, here it means that there is less opportunity to generate a sufficiently large gravity wave in the analysis. As a result, the ability to capture the gravity wave is closely linked to the ability to capture

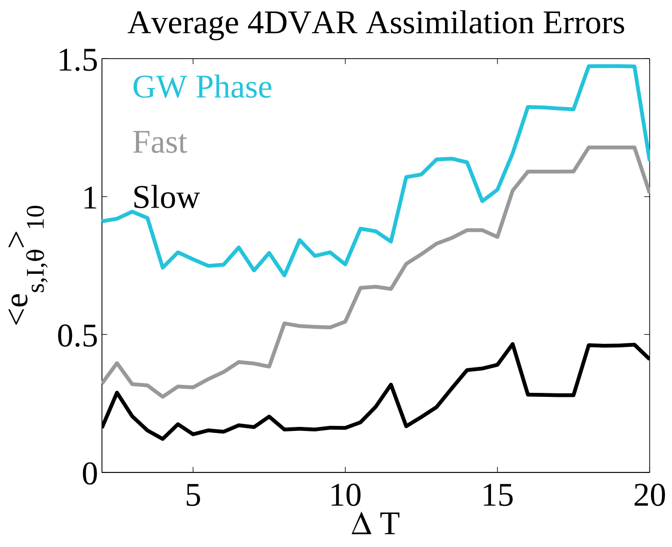

Figure 5.20: Average 4DVAR assimilation errors, as a function of the assimilation window, comparing average slow error (black), fast error (gray), and gravity wave phase error (blue), for experiments with $\Delta t^{\mathrm{obs}} = 2$, and MIX observations.

the slow mode.

5.3.3 Effect of Free Gravity Wave Magnitude

So far we have seen that 4DVAR can only capture the full magnitude of the free gravity wave if roughly twice as many minimization iterations are done as are needed to capture the slow mode. What about free gravity waves of different magnitudes? Figure 5.21 compares average 4DVAR errors as a function of $\tilde{I}^{\mathrm{t}}$, for experiments where the minimization algorithm is cut off when $e_{w,\mathrm{rms}} \leq 0.2$, again comparing MIX to FIL and MIX-OI.

As in the previous figures, the FIL experiments do not at all reduce fast error in 4DVAR, but slow error for FIL also does not increase with $\tilde{I}_t$, as it does in OI. For MIX, fast mode and gravity wave phase errors are closer to the no-skill level when the gravity wave amplitude small, but are lowered (relative to the no-skill level) as $\tilde{I}^{\mathrm{t}}$ is increased. Recall that, for larger free gravity waves, the deviation of the slow mode from the truth is greater, and minimization increments will therefore be larger. Having already seen

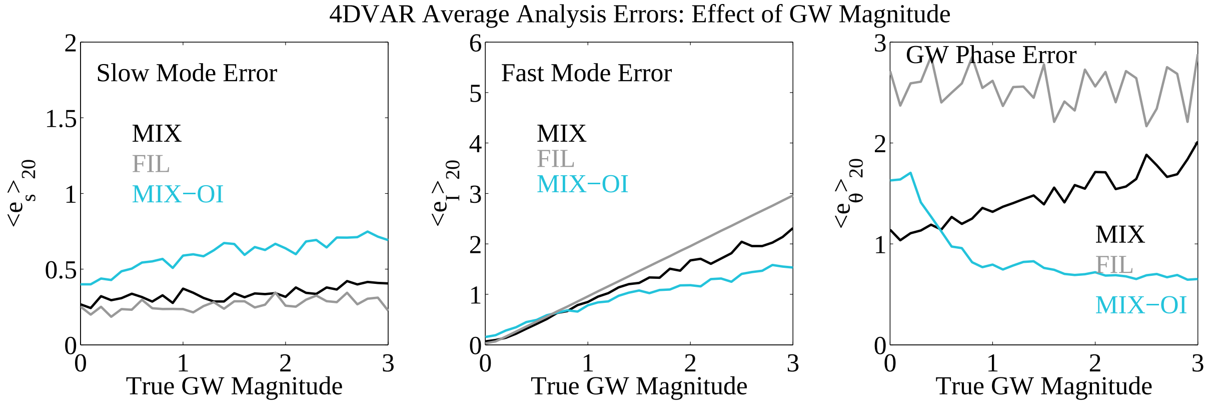

Figure 5.21: Average analysis errors for 4DVAR, divided into the three components [(2.59)-(2.61)], over a range of true gravity wave magnitudes, $\tilde{I}^{\rm t}$, comparing MIX [(2.58), black] and FIL [(2.57), gray] observations, and OI with MIX observations (blue).

that larger minimization increments lead to larger gravity wave generation (both in the balanced and unbalanced truth cases), it follows that a larger gravity wave is generated for large $\tilde{I}^{\rm t}$. Overall reduction from the no-skill level is therefore greatest when $\tilde{I}$ is large. Of course, these gravity waves are still too small because the minimization is cut off when the w-error threshold is passed, and both fast mode and gravity wave phase errors typically exceed corresponding OI errors.

$\langle e_I \rangle$ and $\langle e_\theta \rangle$ can of course be reduced by allowing the minimization to iterate longer, but these results are omitted since they don't really show anything new. As in the EKF and EnKF, recovery of the slow mode is stable in these experiments, even when the (largely unrecovered) free gravity wave is large. Experiments with increasing ϵ and b showed a similar result — namely that recovery of the fast mode is improved when the distance between the background estimate and true state is larger — and are thus also omitted.

5.4 Discussion

Since observations are far less abundant in the middle atmosphere and the tropics than they are in the midlatitude troposphere, accurate data assimilation is an even greater necessity in those regions. The experiments of this chapter show that dynamic error covariances can be quite beneficial, both in recovering the slow mode when the fast mode is unobserved, and in recovering both modes from mixed-timescale observations. The implications for realistic applications are as follows.

The very accurate covariance model of an EnKF with a large ensemble could easily retrieve both modes. Realistically, ensemble size is of course extremely limited. However, even with less accurate covariance models, both the EKF and EnKF were found to be improvements over OI, especially in the task of disentangling the slow and fast modes from mixed-timescale observations. Moreover, by developing dynamic error covariances, all three methods are able to overcome the increase in analysis error (in OI) when the observation interval is a multiple of the gravity wave period. These results suggest that 4D methods will become extremely useful for assimilation in the unbalanced regions of the middle atmosphere.

Though ϵ and b control the nonlinearity of the slaving relationship between fast and slow variables, it was found that increasing either parameter primarily changes the nonlinearity of the assimilation system by increasing the size of analysis increments for observations which ignore the gravity wave, resulting in greater amplification of error in estimated correlations. If fast variables are observed, there is little difference between the EKF and EnKF in these experiments, aside from the peaks at $\Delta t^{\mathrm{obs}} = nT_{\mathrm{GW}}$. Thus, the balance parameters don't necessarily affect the effective nonlinearity of the assimilation system. This result is encouraging, as it suggests that, as long as the assimilation system is able to separate the two modes without external balance constraints, it should also be suitable in the tropics.

4DVAR, which improves upon the EKF's TLM-based covariance model by not updating covariances explicitly (as shown in previous chapters), is comparable to the EnKF in its ability to retrieve the slow mode from mixed-timescale observations. However, it is, overall, difficult for 4DVAR to retrieve the free gravity wave, unless about twice as many iterations of the minimization are performed as are necessary to capture the slow mode. The tendency of 4DVAR to increase the analysis free gravity wave with increasing minimization iterations has similarity with the results of Tanguay et al. [1995], who found that 4DVAR recovers scales which are fast relative to the observations as the minimization continues. Though that study was concerned mainly with spatial scales, we see a similar principle here: because the model contains scales not resolved in the observations, the unobserved scales can in a sense be filled in by continuing the iterations of the minimization. Information is transferred thereby from the spatial scale of the observations to smaller scales. In our experiments, this greatly increases the cost of 4DVAR experiments relative to the Kalman filters.

Another interesting difference between 4DVAR and the Kalman filters was found for the phase of the free gravity wave. Even if enough minimization iterations are done for 4DVAR to recover the fast mode magnitude, error in the gravity wave phase is *still* reduced only slightly —whereas both the EKF and the EnKF found gravity wave phase quite easily. This could be an advantage of sequential assimilation, though it can also be argued that the EKF and EnKF are so successful at recovering the phase because the gravity wave frequency (ϵ^{-1}) is already known. This is addressed briefly in the next chapter.

Chapter 6

Model Error and Bias Correction in the EnKF

6.1 Introduction

The previous two chapters showed that dynamic error covariance models benefit the recovery of both balanced and unbalanced states, as long as the assumptions upon which each assimilation scheme is built are valid. The EnKF in particular performed quite well without additional balance or gravity wave constraints, as it showed stability and the ability to recover both modes over large Δt^{obs} (e.g. fig. 4.13), and was easily adaptable to unbalanced truth (fig. 5.10) and large-ϵ (fig. 5.14) cases.

It can be argued, of course, that at least part of the EnKF's success in these experiments can be attributed to the fact that the ensemble was generated from a perfect model. Models are always imperfect in the real world, which makes 4D covariance models which assume a perfect model fundamentally biased —even for frequent observations, clever initial covariance formulation, a large ensemble, etc. If model bias is not accounted for within the data assimilation system, the forecast/analysis errors which are estimated by the assimilation scheme will underestimate true errors, giving too much credence to the model and not enough to observations. Another caveat for the EnKF results shown here is that the ensemble sizes required to exceed the accuracy of the EKF (e.g. figs. 4.12 and 5.11) and to control the fast wave ($N \simeq 4-6$) were around or above the state dimension. For an operational model, this is a prohibitively expensive requirement.

There exists a large amount of literature suggesting modifications to and alternative formulations of the standard 4D assimilation schemes, which are intended to deal with both the small-ensemble and model error problems. Some of these modifications take the form of additional terms added to the equations, while others replace the forward evolution of covariances, or the analysis step (of the state or its covariance estimate), with an alternative formulation. In 4DVAR, it is possible to add extra terms to the cost function [Daley, 1991, chapter 13, Courtier and Talagrand, 1990, Fisher et al., 2005], or to the perturbation forecast model [Lorenc, 2003a], or to make model error parameters part of the control vector [Lorenc, 2003a]. In the Kalman filter, one can add “forgetting factors” to the evolution of covariances, which decrease the link between forecast covariances and the preceding analyses, and thus lessen the model constraint on the assimilation. It is also possible to adaptively estimate biases, using information from observations. In this chapter, we address the effect of forgetting factors and adaptive bias estimation.

It is important to note that such modifications need to account not just for errors in the model and parametrizations of unresolved physics, but also for the cumulative effects of an imperfect assimilation system, due to nonlinearity, sampling error, and practical limitations. Following Dee [2005], we refer to the sum of all these effects as the **bias**.

Because bias consists of so many, largely cumulative effects, it has been pointed out [Dee, 2005, Ehrendorfer, 2006] that modifications intended to correct bias cannot be trusted blindly based on a few successful experiments, but must be understood as fundamentally as possible. Dee [2005] also points out that the large variety of bias sources means that, while bias-aware assimilation is rather simple in principle, the formulation of bias models is difficult. In the exL86 model, for example, we have found that bias in the EnKF comes not just from model error, but also from the clustering of the ensemble around the wrong or a spurious gravity wave. If a suggested trick improves the analysis in one case, it may not necessarily accomplish the same in every other system, and could, in fact, make things worse. Dee [2005] points out that misspecified bias correction (modifications to the algorithm which treat the wrong bias source) will exacerbate filter divergence. Moreover, since modifications and alternative formulations are expensive to implement operationally, it makes sense to begin the comparison in the realm of low-order models, where individual phenomena (in our case, balance and imbalance) can be isolated and understood in a simple and transparent manner.

In this chapter, we focus on the covariance model which required the fewest external constraints —that of the EnKF— and consider what happens to the analyses of balanced and unbalanced states when systematic errors are present in the model. We then compare three modifications to the standard EnKF for both perfect and imperfect implementations of the exL86 model, and examine how the analyses of the slow and fast modes change under these additional terms.

6.2 Simulation of Model Error

The first step is to see to what extent the advantages of the EnKF's covariance model break down in the presence of model error. First, recall that the *eventual* divergence of the EnKF (meaning: ensemble spread becoming less than true analysis error), even for a perfect model, is almost inevitable (e.g. fig. 4.11). Filter divergence was found in the experiments of chapter 4, where the ensemble phase-locked around a spurious gravity wave, and in chapter 5, where the ensemble phase-locked around the wrong gravity wave. The possibility of obtaining a biased ensemble from a perfect model has also been noted by Whitaker and Hamill [2002] and Houtekamer et al. [2005]. Addition of model error to an intrinsically-flawed data assimilation system then implies that the distance between the truth and the central forecast will grow faster than the estimated forecast error, leading to further filter divergence.

6.2.1 Model Error in the Gravity Wave

In all experiments shown so far, forecasts and true states were evolved with the same model, so that the analyses were always perfect in gravity wave frequency. It can be argued that this is at least part of the reason why the EKF and EnKF converged rather easily in gravity wave phase (e.g. fig. 5.8). In the real world, observed gravity wave frequencies differ from model-resolved frequencies, and both differ from the truth. For example, fast gravity waves are less likely to be observed than slow gravity waves, simply because fast waves spend less time in an instrument's observation region [Alexander and Barnet, 2006]. Moreover, gravity-wave frequencies are altered by model numerics. To examine this issue, we consider two sets of experiments in which the perfect model restriction is removed for the gravity wave parameters, ϵ and b.

In the first set of experiments, the EnKF with 50 ensemble members is run over a range of gravity wave frequencies ϵ^{f} assumed by the forecast model, with truths generated with gravity waves of frequency $\epsilon^{\mathrm{t}} = 10^{-1}$ and magnitude $\tilde{I}^{\mathrm{t}} = 1.5$. In the second set of experiments, the parameter b^{f} in the forecast model is changed, while keeping the true value constant at $b^{\mathrm{t}} = 0.71$. Column (A) shows the three error measures as a function of the forecast model gravity wave period, $T^{\mathrm{f}}_{\mathrm{GW}} = 2\phi\epsilon^{\mathrm{f}}$, while Column (B) shows analysis errors over a range of b^{f}. The true values of gravity wave period or coupling parameter are indicated in each plot.

Column (A) shows that analysis errors associated with the gravity wave shoot to no-skill levels for any $T^{\mathrm{f}}_{\mathrm{GW}} \neq T^{\mathrm{t}}_{\mathrm{GW}}$. That fast error occasionally exceeds 1.5 indicates the excitation of spuriously large, uncontrolled gravity waves. Failure to capture the gravity wave also increases slow mode error. Since the gravity wave is not captured at all in these cases, analysis error also increases when the gravity wave period is a multiple

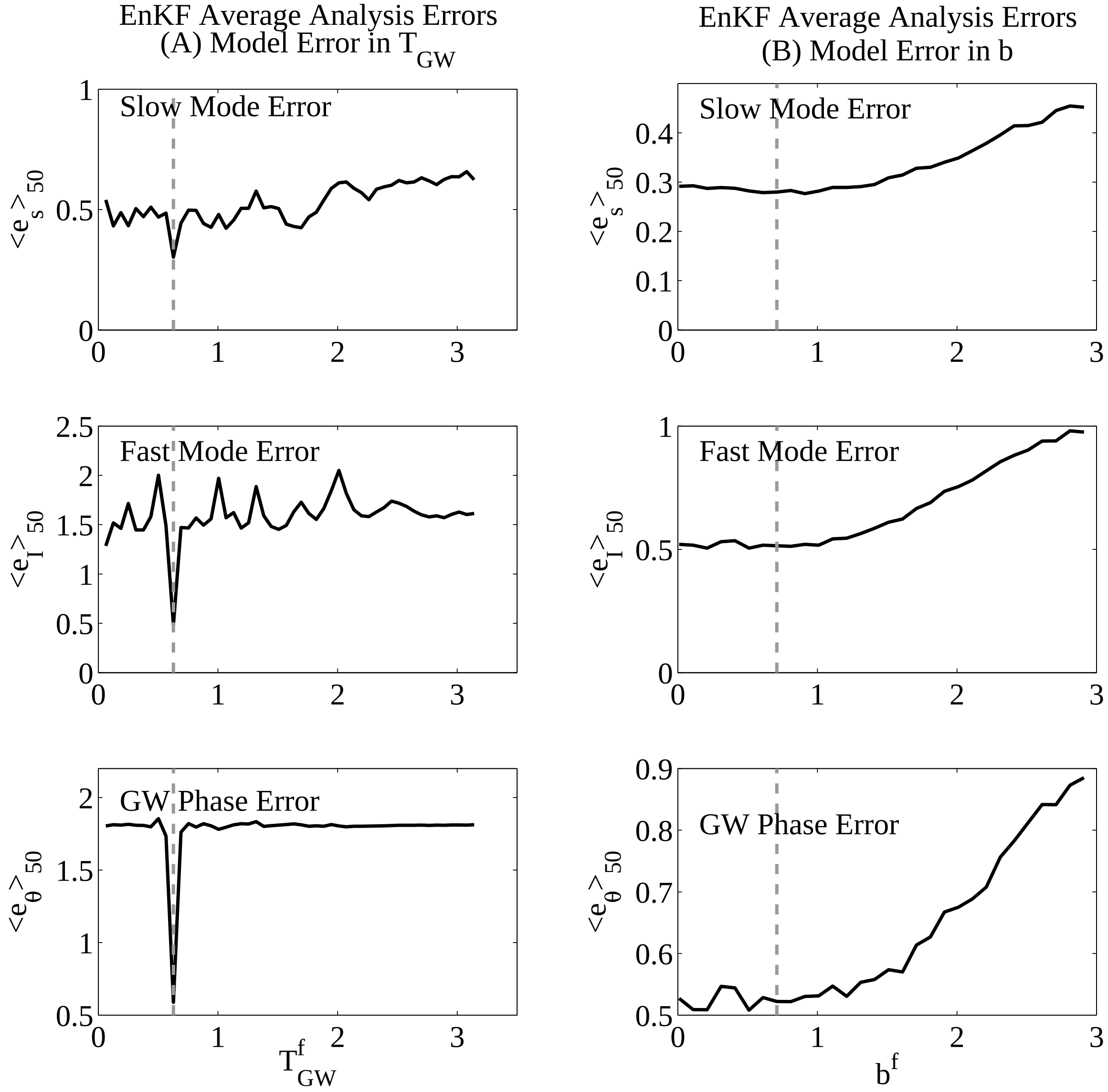

Figure 6.1: (A) Average analysis errors (as in previous figures), changing the timescale separation parameter in the forecast model, ϵ^{f}. (B) The same, but changing the parameter b^{f} which is estimated by the forecast model. In both experiments, a 50-member EnKF is used.

of the observation interval, and we see small peaks at $T_{\mathrm{GW}} = 1$ and 2, as in the OI experiments in figures 5.3 and 5.6.

A less extreme case of model error is shown in Column (B), where there is still some skill in estimating gravity wave phase, even for $b^{\mathrm{f}} = 3$. Underestimating this parameter ($b^{\mathrm{f}} < 0.71$) increases error less than when it is overestimated ($b^{\mathrm{f}} > 0.71$). This results from the fact that b changes the information content of the observation, from being purely slow to being mixed. For $b^{\mathrm{f}} > 0.71$, the assimilation is assuming that w' is more mixed than it is really, resulting in adjustment of the gravity wave away from its true amplitude and phase, and not enough adjustment of the slow variables, which suffer from the misestimated gravity wave. For $b^{\mathrm{f}} < 0.71$, the assimilation is assuming that w' is less mixed than it really is, and thus adjusts the slow mode with more information than is actually contained in the observations, while the gravity wave is not adjusted enough. This increases slow mode errors slightly, while errors for the gravity wave, which is easier to capture, stay about the same.

Though these two experiments are highly idealized (since only one type of gravity wave is admitted in each, truth and forecast), they show that the EnKF's ease in locking onto a linear gravity wave phase will be limited by any differences between the resolved frequencies in observations and model. When the ability to capture a gravity wave is lost, the analysis of the vortical mode suffers as well. Incidentally, repeating the experiments of figure 6.1 with the EKF (not shown) shows similar behavior but overall larger average errors. This suggests that accurate 4D development of forecast error covariances can nevertheless help to reduce this consequence of model error in resolved gravity wave parameters.

6.2.2 Model Error in the Slow Mode

A less extreme way to simulate systematic model error in the exL86 model is to give $C(t)$, which controls the degree of chaos in the slow mode, different time dependencies in the forecast model. In the experiments to follow we generate the true state with $C^{\mathrm{t}}(t) = a_0 + a_1 \cos(\gamma t)$, with $a_0 = 1$, $a_1 = 0.8$, and $\gamma = 0.92$, as in previous chapters. Ensembles will now be generated with $C^{\mathrm{f}}(t) = a_0^{\mathrm{f}} + a_1^{\mathrm{f}} \cos(\gamma t)$, under the following three model error scenarios:

ME0 The reference case of zero model error. Here the forecast is generated with the same parameters as the truth.

ME1 The forecast has a nonchaotic slow mode. In this case, $a_1^{\mathrm{f}} = 0$, with all other parameters as in the truth.

ME2 The forecast slow mode has greater variability. Here $a_0^{\mathrm{f}} = 4$, with all other parameters as in the truth.

Figure 6.2 illustrates the three model error scenarios for a reference experiment where the true state is balanced, and MIX observations (2.58) are assimilated every $\Delta t^{\mathrm{obs}} = 4$ time units using a 10-member ensemble. Figure 6.2 (A) shows the EnKF analysis for the perfect model scenario (ME0), while (B) and (C) show the analyses which result from model error scenarios ME1 and ME2, respectively. All other parameters and random number realizations are the same. The assimilation window in these and subsequent examples has been shortened to 15 time units in order to make the analysis of the gravity wave more visible. RMS slow and fast errors over the assimilation window are also shown for each experiment.

Some general effects of systematic model error (which have nothing to do with balance and gravity waves) are illustrated by these examples, and can be seen most clearly by comparing the ϕ analyses between the three model error cases. For ME0, the ensemble is spread more or less evenly about the central analysis, though the spread is tight compared to the actual error around $t = 5$. For ME1 and ME2, things look much worse: The ME1 ensemble (B) doesn't spread enough between observations (reflecting the lack of chaos in the forecast model) and becomes tightly clustered around a poor estimate, doubling the rms slow mode error. The ME2 ensemble spreads too much between observations, and becomes bimodal —tripling the slow mode error.

In regards to balance, the process of assimilation reduces net imbalance in the ME0 ensemble, though, as in chapter 4, the ensemble still locks onto a small gravity wave by the end of the assimilation period. Loss of balance and phase-locking effects are far more egregious for ME1: instead of reducing imbalance, the assimilation of observations mainly causes more ensemble phase-locking. Imbalance is actually lowest in the ME2 case, but we shall see below that this is a fluke for this particular realization.

These examples show again that the advantages of 4D covariance models, in regards to balance and gravity waves, can diminish in the presence of model error, as in §6.2.1. The next three subsections will consider how filter divergence due to systematic error in the slow mode, and the associated loss of balance or recovery of the true-state gravity wave, can be controlled by adding extra terms to the EnKF analysis.

6.3 Modifications to the EnKF Analysis

We now examine three simple ways of overcoming bias in the EnKF, which have varying degrees of complexity: covariance inflation, the addition of white noise to the ensemble, and an adaptive bias estimation algorithm.

6.3.1 Covariance Inflation / Deflation

Formulation

Anderson and Anderson [1999] suggest that a very simple way to prevent filter divergence is to simply inflate error covariances, by multiplying ensemble deviations from the mean by a factor $\beta > 1$. This gives the assimilation system less confidence in the forecast estimate than determined by the assimilation cycle, while preserving the forecast itself. It can be argued that some form of covariance inflation is always necessary, since the

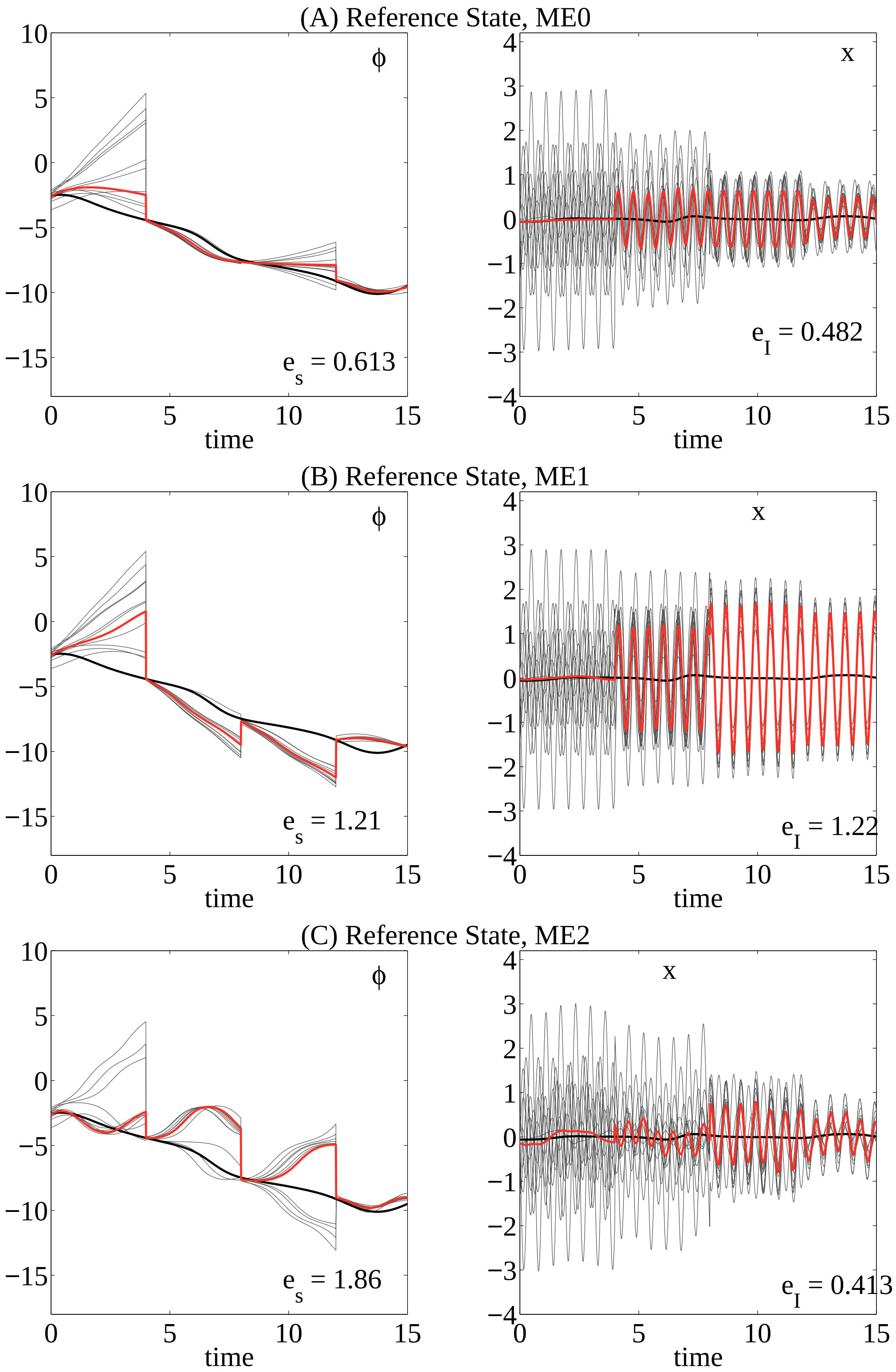

Figure 6.2: EnKF reference example, comparing zero model error [ME0, (A)], and model error scenarios ME1 (B) and ME2 (C), for MIX observations (2.58), with $\Delta t^{\mathrm{obs}} = 4$ and a 10-member ensemble. In all figures, the truth (black) is compared to the analysis ensemble (gray) and ensemble mean (red).

EnKF has a tendency to underestimate covariances even without systematic model error. Moreover, it should help to err on the side of overestimation of error variances, since overestimation can be self-correcting while underestimation of variances leads to filter divergence. Ensemble inflation has been shown by Whitaker and Hamill [2002] to improve the EnKF analysis for both a low-order model and an intermediate global circulation model, even in the absence of model error. Similarly, Hoteit et al. [2005] and Fisher et al. [2005] found that covariance inflation improved the EKF analysis in perfect model experiments, as did Anderson and Anderson [1999] with a variant of the EnKF known as a particle filter (appendix B).

In the EnKF, forecast error covariances can be inflated by multiplying the deviation of each ensemble member from the mean by β, then adding this term back to the mean:

$$\hat{\mathbf{x}}_i^{\mathrm{a}} = \langle \mathbf{x}^{\mathrm{a}} \rangle + \beta \left(\mathbf{x}_i^{\mathrm{a}} - \langle \mathbf{x}^{\mathrm{a}} \rangle \right), \tag{6.1}$$

such that the inflated error covariance matrix becomes

$$\hat{\mathbf{P}}^{\mathrm{a}} = \beta^2 \mathbf{P}^{\mathrm{a}}. \tag{6.2}$$

With ensemble inflation, the covariance between ϕ and x, for example, becomes

$$\hat{c}_{\phi x} = \beta^2 \rho_{\phi x} \sigma_\phi \sigma_x. \tag{6.3}$$

Thus the effect of EnKF-estimated multivariate error correlations is amplified if $\beta > 1$.

It is conceivable that the EnKF could develop accurate multivariate relationships while underestimating error magnitudes. Nerger et al. [2005], for example, found that spatial covariance *structures* can be well-estimated in the EnKF while *amplitudes* of covariances are underestimated. If the EnKF is able to naturally develop correct correlations even when the ensemble spread is small, inflating the ensemble while preserving correlation structure could be beneficial. On the other hand, inflating covariances can also amplify undesirable effects, such as error due to non-Gaussianity or sampling. For $\beta < 1$, the ensemble distribution is "deflated" and the effect of estimated error relationships is then de-amplified.

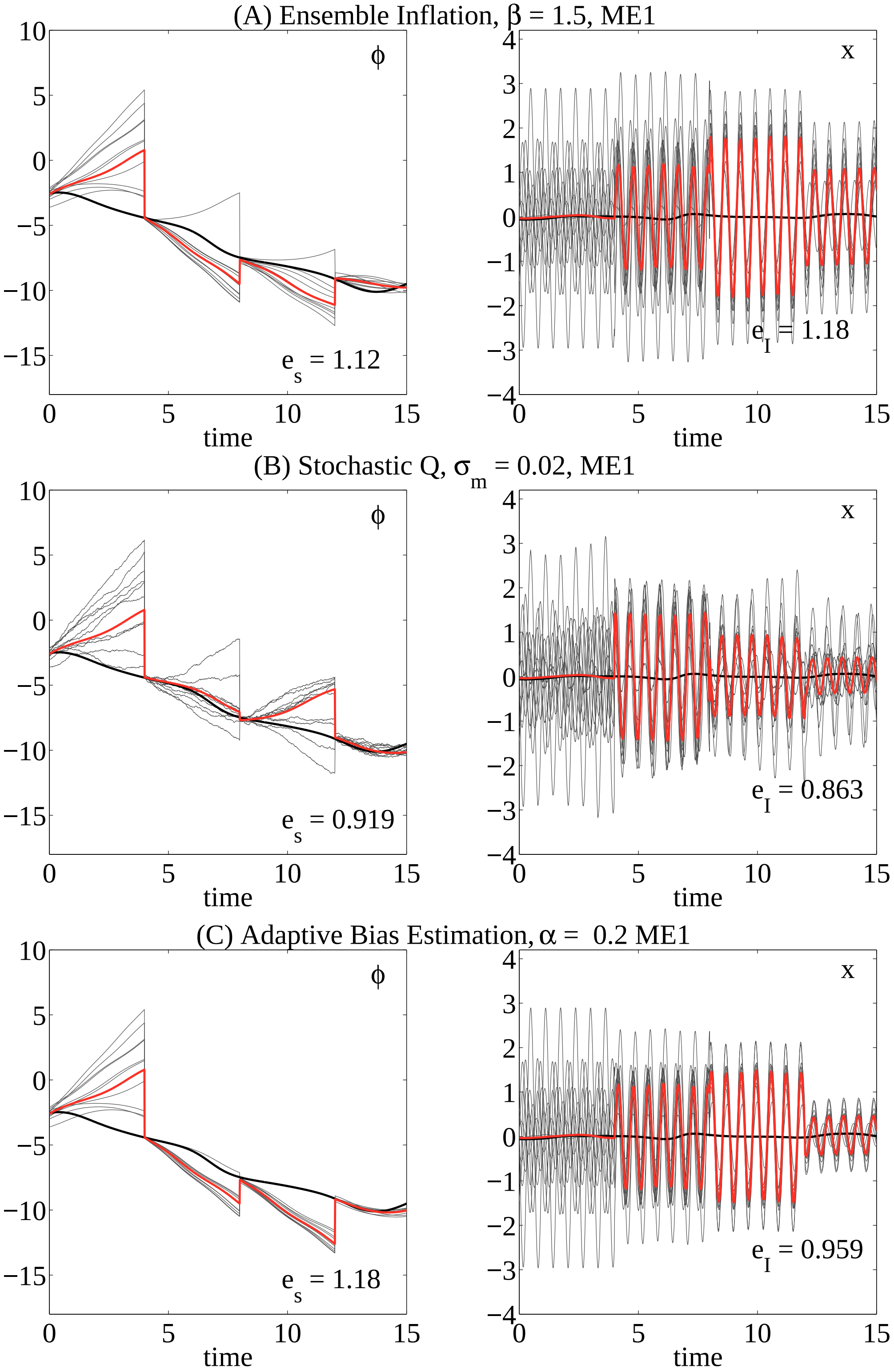

Figure 6.3: As in figure 6.2 (B), but with (A) ensemble inflation with inflation factor $\beta = 1.5$, (B) stochastic forcing with $\sigma_{\mathrm{m}} = 0.02$, and (C) application of the bias estimation algorithm [(6.4)-(6.8)] and $\alpha = 0.2$.

Example

An illustration of the effect of ensemble inflation on the recovery of a balanced state is shown in figure 6.3 (A), where we repeat the ME1 experiment of figure 6.2, but now with inflation of the ensemble by a factor $\beta = 1.5$ after each analysis step. The slow mode analysis in this case is now better than in the ME1 reference case [fig. 6.2 (B)]: error is reduced, and the ensemble does not deviate radically from the truth between observations. Though slow error is not reduced to that of the ME0 case [fig. 6.2 (A)], ensemble spread around the mean is improved relative to true error. A small reduction of imbalance can be seen in the analysis of x, where the net spurious gravity wave is slightly smaller than in figure 6.2 (B). However, the ensemble is still very much unbalanced and phase-locked.

Quantitative Comparison: Ensemble Inflation

A more quantitative examination of this result is shown in figure 6.4 (A), which shows average slow and fast analysis errors as a function of the inflation factor, with β ranging from 0.5 to 3, with $\Delta t^{\mathrm{obs}} = 2$ and 10-member ensembles. Average errors are compared for each model error scenario. The boundary between "inflation" and "deflation" has been marked with a line at $\beta = 1$.

For ME0, ensemble inflation ($\beta > 1$) largely *increases* error in both modes, though slight inflation ($\beta \sim 1.1$) was found to reduce errors in some experiments (an effect which is hidden in the average shown in the figure). This indicates that ensemble-estimated errors are usually appropriate in the perfect-model case, and giving more weight to observations only occasionally improves things. The ME0 curves look similar to figure 3 of Anderson and Anderson [1999], which showed the same statistics for the Lorenz [1963] 3-component model, and found an optimal inflation factor of about 1.1 for a perfect model.

Ensemble inflation does decrease slow mode error for ME1 and ME2, though not

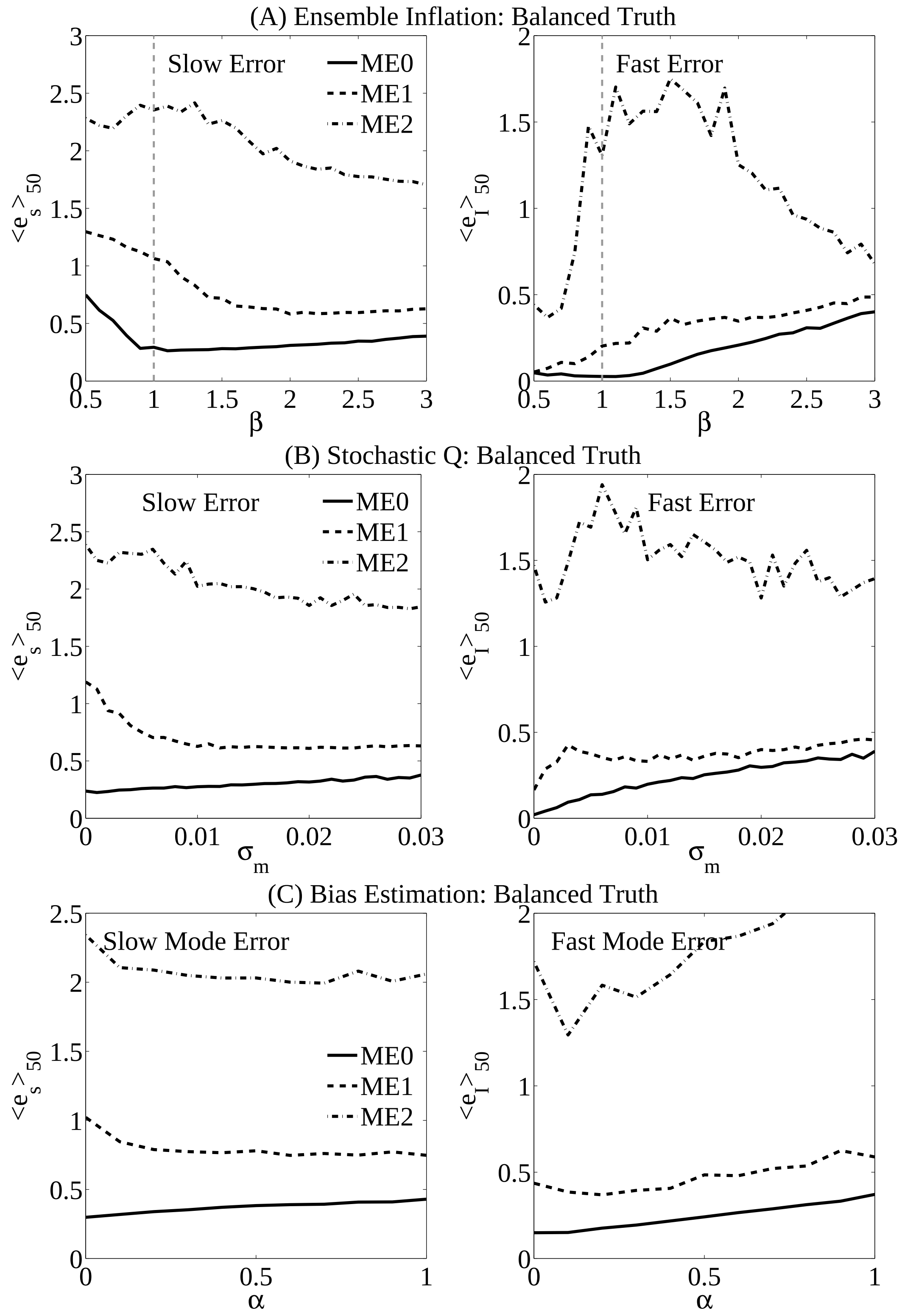

Figure 6.4: Average EnKF slow [(2.59), left] and fast [(2.60), right] errors for the three model error scenarios and a balanced truth, versus (A) inflation factor β, (B) standard deviation σ_m in the stochastic forcing matrix, and (C) the bias estimation parameter α.

to that of ME0. Fast errors are only decreased by very heavy inflation ($\beta > 2$) for ME2, and by deflation ($\beta < 1$) for ME1 and ME2. Since inflation only corrects underestimated covariance amplitudes, it cannot compensate for the systematically-flawed fast-slow correlations which come from the biased ensemble. Hence ensemble inflation primarily increases fast error for ME1, by amplifying the imbalance in individual ensemble members. However, ensemble inflation does decrease fast errors for ME2, for $\beta > 2$. Examination of individual experiments suggests that this happens because strong inflation helps to undo the extreme ensemble phase-locking that happens in the ME2 case. By and large, however, we can say that ensemble inflation, while useful for overcoming systematic error in the vortical mode, amplifies the covariance estimation errors which cause spurious imbalance.

A consequence of this is that *deflation* ($\beta < 1$) decreases fast errors for ME1 and ME2. This suggests that one solution to balance problems in the EnKF might be simply to manually reduce the variances associated with fast variables, thereby bringing the ensemble of fast variables closer to the mean state. This result may have implications for general cases where spurious correlations arise due to sampling error. Polavarapu et al. [2005], for example, found that overestimated correlations coupled to large variances can create unrealistic analysis increments in the mesosphere, from the assimilation of stratospheric data (with biases growing because of a lack of mesospheric observations). In such a case, manually deflating covariances could potentially improve the assimilation.

Figure 6.5 (A) shows the same statistics as figure 6.4, but for sets of experiments where the truth contains a free gravity wave with $\tilde{I}^{\mathrm{t}} = 1.5$, as in chapter 5. The left column in this figure looks remarkably similar to the left column of figure 6.4. Further, ensemble inflation now improves the analyses of both modes in both ME1 and ME2. Recall that slight overestimation of ensemble gravity wave variance improved the fast mode analysis (fig. 5.12). Here we see the same effect, and since estimated slow mode errors are also inflated, the slow mode analysis is not made worse by giving more credence

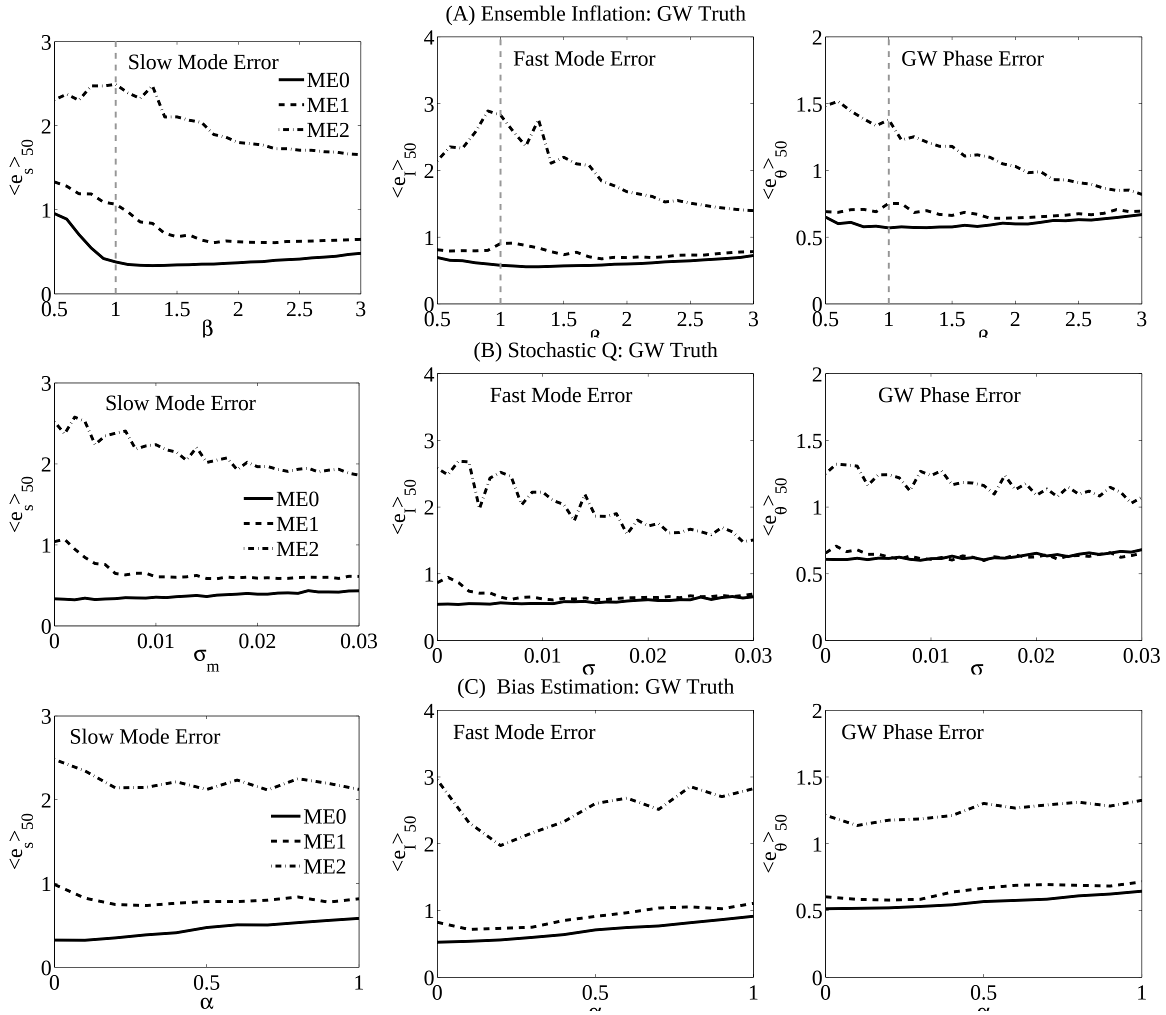

Figure 6.5: As in figure 6.4, but for a true state with free gravity wave $\tilde{I}_t = 1.5$.

to observations of the fast mode. In summary, ensemble inflation is by-and-large a good idea, especially when two timescales are present, but can also destroy balance.

6.3.2 Stochastic Model Error Forcing

Formulation

Another way to account for bias is by modeling the $\mathbf{q}_k$ term in (2.31) as a white noise process. In the EnKF, this can be done by adding random vectors $\mathbf{q}_{i,k} \sim \mathcal{N}(0, \sigma_m^2)$ to individual ensemble members, and computing $\mathbf{P}_k^{\mathrm{f}}$ from the ensemble (2.42) as before.

Unlike ensemble inflation/deflation, the addition of stochastic terms adds a random component to estimated correlation structures, which could potentially destroy misestimated correlations and thus prevent observations from either shocking the system into highly unbalanced states, as in figure 6.2 (B), or causing the phase-locking of the ensemble around a spurious gravity wave. Too much random forcing can, of course, also destroy the potentially valuable correlations developed within the assimilation cycle.

Example

In figure 6.3 (B), random terms with standard deviation $\sigma_{\mathrm{m}} = 0.02$ are added to individual ensemble members in the ME1 reference case. As in the ensemble inflation experiment [6.3 (A)], the small stochastic perturbations successfully keep the ensemble from clustering too tightly around the mean in the slow mode, resulting in a more accurate analysis. In x, it can be seen that stochastic forcing also increases imbalance in individual ensemble members. However, it also causes crucial phase mixing, resulting in a net analysis which is more balanced, though the spurious gravity wave is still not eliminated.

Quantitative Comparison: Stochastic Forcing

Figures 6.4 (B) and 6.5 (B) show average slow and fast errors over a range of σ_m, for the three model error scenarios, in the context of a balanced truth [fig. 6.4 (B)] and a free gravity wave truth with $\tilde{I}^{\mathrm{t}} = 1.5$ [fig. 6.5 (B)]. In both figures, the effect of stochastic forcing is similar to that of ensemble inflation: forcing increases errors slightly for ME0 and improves the recovery of the slow mode for ME1 and ME2, but increases fast errors for ME0 and ME1, and offers no real improvement for ME2. We thus see again that the imbalance induced by the filter modification is greater than the effect of the phase mixing it creates. For the gravity wave truth [fig. 6.5 (B)], on the other hand, stochastic forcing, like ensemble inflation, decreases all three error measures for ME1 and ME2.

Here, the additional terms evidently increase phase-mixing, and since the free gravity wave is active, it can be controlled by observations. We note in passing that the addition of a diagonal matrix $\mathbf{Q}$ (not shown) did not significantly change the results.

These results have similarity to those from the Canadian operational EnKF [Houtekamer and Mitchell, 2005], where model error realizations which have both balanced and unbalanced components are added to the ensemble. There, the unbalanced component of model error is needed to reduce bias in the analyzed temperature field, but has also been shown to cause imbalance in the ensemble.

More complex ways to account for model error stochastically may be possible. One particular shortcoming of the above formulation is that one would expect that a forecast which incorrectly represents the true dynamical balances will have estimation errors that are smoothly correlated in space and time [Houtekamer and Mitchell 2005, Lorenc 2003a]. Houtekamer and Mitchell [2005] suggest that the random vector $\mathbf{q}_{i,k}$ could also be estimated by simultaneously running different models, or the same model with different parameterizations. Model error could also be modeled by giving the random terms the same structure as is currently used to approximate background error covariances in 3D assimilation [Houtekamer and Mitchell, 2005, Houtekamer et al., 2005]. Figure 6.4 (B) suggests that more complex stochastic model error forcing should also include consideration of balance.

6.3.3 Sequential Bias Estimation

Another bias formulation, designed to be more dynamically consistent than inflation or stochastic forcing, has been suggested by Dee [2005], who proposes a sequential bias estimation algorithm wherein an estimate of systematic model error, $\mathbf{b}(t)$, is updated in proportion to innovations, in analogy to the Kalman filter equations.

Formulation

The algorithm of Dee [2005] was originally proposed for the EKF. Here we have adapted it to the EnKF in the following form:

$$\mathbf{b}_{i,k} = \mathbf{b}_{i,k-1} - \mathbf{K}_k^{\mathrm{b}} \left[\mathbf{z}_{i,k} - \mathbf{H}\left(\mathbf{x}_{k,i}^{\mathrm{f}} - \mathbf{b}_{i,k} \right) \right] \tag{6.4}$$

$$\mathbf{x}_{i,k}^{\mathrm{a}} = \left(\mathbf{x}_{i,k}^{\mathrm{b}} - \mathbf{b}_{i,k} \right) + \mathbf{K}_k \left[\mathbf{z}_{i,k} - \mathbf{H}\left(\mathbf{x}_{k,i}^{\mathrm{f}} - \mathbf{b}_{i,k} \right) \right] \tag{6.5}$$

$$\mathbf{K}_k = \mathbf{P}_k^{\mathrm{f}} \mathbf{H}^{\mathrm{T}} \left[\mathbf{H}\mathbf{P}_k^{\mathrm{f}} \mathbf{H}^{\mathrm{T}} + \mathbf{R} \right]^{-1} \tag{6.6}$$

$$\mathbf{K}_k^{\mathrm{b}} = \mathbf{P}_k^{\mathrm{b}} \mathbf{H}^{\mathrm{T}} \left[\mathbf{H}\mathbf{P}_k^{\mathrm{b}} \mathbf{H}^{\mathrm{T}} + \mathbf{H}\mathbf{P}_k^{\mathrm{f}} \mathbf{H}^{\mathrm{T}} + \mathbf{R} \right]^{-1} \tag{6.7}$$

$$\mathbf{x}_k^{\mathrm{a}} = \langle \mathbf{x}_{i,k}^{\mathrm{a}} \rangle \tag{6.8}$$

In this formulation, each ensemble member has a corresponding estimated bias vector $\mathbf{b}_{i,k}$. Starting from some initial guess, these are updated at observation times according to the observation increment (6.4), weighted by a special bias-gain matrix, $\mathbf{K}_k^{\mathrm{b}}$. This matrix is computed from the bias covariance matrix $\mathbf{P}_k^{\mathrm{b}}$, weighted by the total estimated bias, forecast error, and observation error (6.7). Individual ensemble members are computed as before, but with the estimated bias removed (6.5). The analysis state is, as in the regular EnKF, the ensemble mean (6.8).

It remains to define the initial bias estimate $\mathbf{b}_{i,k}$ and the bias covariance matrix $\mathbf{P}_k^{\mathrm{b}}$. Since the process is adaptive, we choose $\mathbf{b}_{i,k} = 0$. $\mathbf{P}_k^{\mathrm{b}}$ is difficult to formulate, since it must account for systematic error resulting from both the model and the data assimilation system. Here we follow the suggestion of Dee [2005], approximating $\mathbf{P}_k^{\mathrm{b}}$ as a fraction of the forecast error covariance matrix, $\mathbf{P}_k^{\mathrm{b}} = \alpha \mathbf{P}_k^{\mathrm{f}}$. Since $\mathbf{P}_k^{\mathrm{f}}$ contains physically meaningful error correlations to the extent that the ensemble is representative of true errors, this seems like a plausible structure.

Example

In figure 6.3 (C), the adaptive bias estimation algorithm [(6.4)-(6.8)] with $\alpha = 0.2$ is added to the ME1 reference experiment. With this modification, the spread in the slow error reduced slightly, and imbalance in the individual ensemble members is also reduced from figure 6.2 (B). However, phase-locking of the ensemble about a small spurious gravity wave remains.

Quantitative Comparison: Adaptive Bias Estimation

Figure 6.4 (C) shows the effect of adaptive bias estimation on average fast and slow analysis errors for the balanced truth case, and figure 6.5 (C) shows the same for unbalanced truths, with $\tilde{I}^{\mathrm{t}} = 1.5$. The effect on the slow mode is similar to that of ensemble inflation and stochastic forcing in both cases: slow errors are increased slightly for ME0, and reduced somewhat for ME1 and ME2. The effect of the three modifications on recovery of the slow mode in both the balanced-truth and unbalanced-truth cases is about the same. Fast errors, on the other hand, are decreased slightly for $\alpha \simeq 0.2$ in the ME1 and ME2 cases in both figures, but not for most other values of α. Thus the adaptive bias algorithm is able to decrease spurious imbalance [fig. 6.4 (C)], but the effect is slight and evidently very sensitive to the specification of the bias covariance matrix ($\mathbf{P}^{\mathrm{b}}_{k}$). Recovery of the true gravity wave [fig. 6.5 (C)] is just as difficult, and gravity wave phase errors [fig. 6.5 (C)] show no clear decrease at all.

It is telling that the most sophisticated of the three modifications shown does not offer a greater improvement in the analysis errors. It is plausible that creative changes to the adaptive algorithm could improve the results, but the fact remains that a more sophisticated algorithm does not necessarily make for a better algorithm.

6.3.4 Discussion

We have seen that the EnKF's ability to recover either a balanced state or coexistent fast and slow modes declines considerably in the presence of model error. Increasing model error, and thereby both covariance estimation error and the size of analysis increments, increases the tendency of the EnKF to diverge. Specifically, the ensemble becomes more likely to phase-lock about the wrong gravity wave magnitude.

In the presence of model error, the EnKF can benefit from additional terms which loosen the model constraint. All three of the modifications examined in this chapter improved the recovery of the slow mode, both in the context of a balanced truth (fig. 6.4) and in the presence of a true gravity wave (fig. 6.5). A nonzero free gravity wave is recoverable with ensemble inflation or with the addition of stochastic terms, and possibly with the adaptive bias estimation algorithm (though the effect in our experiments was not clear), but balance is easily lost for all three methods. Overall, the improvement offered by all three modifications was approximately similar. In realistic applications, it will be important to consider balance when formulating modifications to the analysis, perhaps by formulating a more physical bias-estimation model, or, more simply, by a combination of covariance inflation and deflation.

Chapter 7

Conclusions

7.1 Summary and Discussion

4D data assimilation methods combine observations and model integrations to approximate the statistical error relationships between modeled variables over time and space. For nonlinear models, the three most basic ways to do this are by evolving covariances between observations using a tangent-linear model (the EKF), by evolving an ensemble of states (the EnKF), or by minimizing a cost function (4DVAR). The goal of this study was to assess whether and how 4D assimilation algorithms are able to capture multivariate error relationships in the context of a dynamical model which admits two modes evolving on very different timescales.

This problem has two components: (1) so-called balanced dynamics, where the free fast motion is zero and the fast variables are slaved to the slow, and (2) the "unbalanced" case where both timescales have significant energy. The former corresponds to the midlatitude troposphere and lower stratosphere, where motion is primarily vortical and the effect of gravity waves is negligible, and the latter to the upper stratosphere and mesosphere, where inertia-gravity waves dominate. The experimental environment was also extended to regimes where the timescale separation between inertia-gravity waves and vortical modes becomes unclear, a context which corresponds roughly to assimilation in the tropics.

Though the three basic 4D methods represent three different implementations of the same principle, they are fundamentally different in their treatment of nonlinear and/or non-Gaussian dynamics, and can therefore return very different analysis states if the effective nonlinearity of the assimilation system is high. It follows that the three basic algorithms will also differ in their representation of a nonlinear balance relationship.

The primary question of whether 4D algorithms can develop covariance models which are representative of a system which admits two timescales and exhibits balance dynamics, can be answered as yes — with several important details and some unexpected consequences, to be summarized below. From the experiments in chapters 3-5, a picture emerged wherein the accuracy of the estimated covariance model depends on the degree of nonlinearity preserved in the assimilation system. An overall improvement was found in going from the EKF to the two algorithms designed to be more nonlinearity-preserving, the EnKF and 4DVAR, but the improvements also depended on specific physical contexts and assimilation parameters.

Taking average fast error (2.60) as a measure of the assimilation system's ability to capture either a balanced state or free fast motion, figure 7.1 summarizes the relative accuracy of the EKF, EnKF, and 4DVAR, relative to each other and to basic OI. In (a) effective nonlinearity is increased by increasing Δt^{obs}, with $\tilde{I}^{\mathrm{t}} = 0$ (the balanced-truth case). In (b), relevance of the fast mode is increased by increasing free gravity wave magnitude, with $\Delta t^{\mathrm{obs}} = 2$. In (c), the computational cost of the EnKF is changed by changing N, with $\Delta t^{\mathrm{obs}} = 2$ and $\tilde{I}^{\mathrm{t}} = 0$. In (d), the computational cost of 4DVAR is changed by changing the number of minimization iterations allowed, with $\Delta t^{\mathrm{obs}} = 2$ and $\tilde{I}^{\mathrm{t}} = 1.5$.

The first component of the problem is capturing slaving when the true state is balanced, summarized in figure 7.1 (a) and (c). It was shown in chapter 4 that the way in

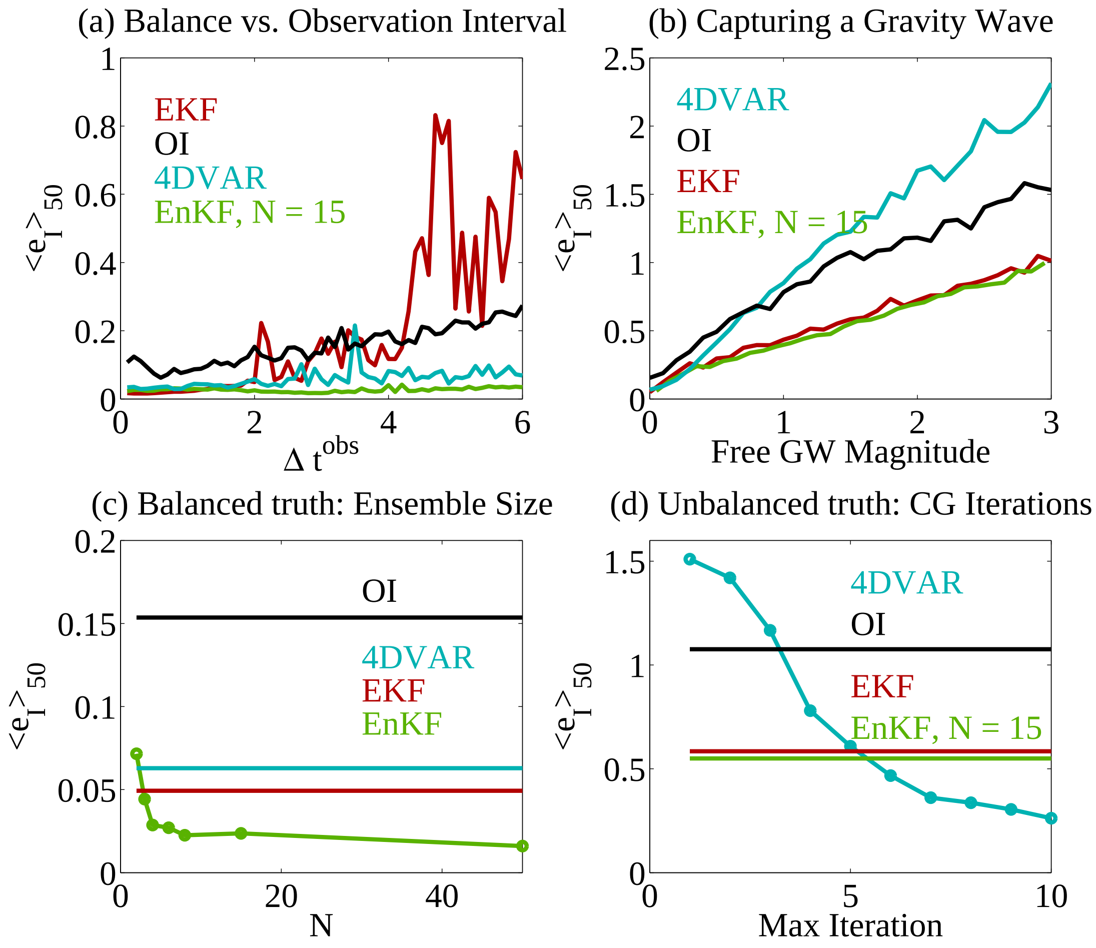

Figure 7.1: A summary comparison of the three 4D algorithms (EKF, EnKF, and 4DVAR) and OI. Average fast mode error (2.60) is compared while changing four parameters: (a) increasing the time between observations for a balanced true state, (b) increasing the magnitude of the true state's free gravity wave, for $\Delta t^{\rm obs} = 2$, (c) increasing ensemble size for a balanced true state and $\Delta t^{\rm obs} = 2$, and (d) increasing the number of iterations performed by 4DVAR, for a true state with $\tilde{I}^{\rm t} = 1.5$ and $\Delta t^{\rm obs} = 2$. In all four panels, the lines represent averages over 50 experiments, and the observations are of the mixed-timescale state MIX (2.58).

which balance is lost for each type of algorithm depends on how error in the estimation of covariances propagates through the assimilation cycle, and is therefore tied to the properties which differentiate the assimilation algorithms from one another. Slaving implies that covariances between fast and slow variables result from the (nonlinear and asymptotic) slaving relationship, and are hence difficult to capture in an assimilation cycle that relies largely on linearizations. Figure 7.1 (a) shows that the EKF becomes unstable from the increase of covariance estimation error with increasing Δt^{obs}, while the EnKF and 4DVAR both preserve balance and clearly surpass OI in accuracy. The EKF error covariance model becomes inaccurate for two reasons: (1) the linearized approximation of covariances is often not valid, and (2) the subsequent adjustment of the analysis error covariances frequently exacerbates covariance estimation error.

The EnKF has two properties which imply a more balanced analysis than that produced by the EKF cycle: (1) imbalance in the evolution of covariances is limited to the net imbalance in individual ensemble members, which does not grow during the forecast step since the evolution is nonlinear; and (2) the analysis state is an ensemble average, implying that spurious imbalance will average out as long as the ensemble is sufficiently phase-mixed. Though neither of these properties guarantees a balanced analysis, numerical experiments found that only a relatively small amount of imbalance is induced by the EnKF cycle, even when Δt^{obs} is large [fig. 7.1 (a)]. The limiting factor for the EnKF is, of course, the ensemble size. This is summarized in figure 7.1 (c), which shows average fast errors (for $\Delta t^{\mathrm{obs}} = 2$) for increasing ensemble size to corresponding fast errors produced by the other three methods. Even though slaving reduces the space of possible true states to an approximate two-dimensional submanifold, it was found that three or four ensemble members are required in order to make the EnKF more balance-preserving than the EKF and 4DVAR. However, considering that EnKF fast errors are still far lower than OI fast errors even when only two ensemble members are used (an ensemble size at which the chaotic slow mode is virtually not captured at all, fig. 4.13), the EnKF can be considered quite useful with respect to balance dynamics.

The other alternative to the EKF, 4DVAR, also infers covariance information from the full model evolution, as in the EnKF. Additionally, information in 4DVAR is spread both ways in time. As a result, far less spurious projection of analysis error onto the fast mode was found for 4DVAR, and it performs similarly overall to the EnKF [fig. 7.1 (a)]. Recall that the algorithm, as formulated in the present study, was intrinsically flawed for all but the first time window. Given this handicap, the reluctance of 4DVAR to deviate from a balanced initial background state, and its general success in capturing the slow mode (e.g. fig. 4.19), is encouraging. The degree of balance captured in the 4DVAR analysis was found to depend strongly on the number of iterations performed in the minimization, and the magnitude of the steps taken within the minimization, with more spurious imbalance generated if increments were large, and/or the minimization was allowed to iterate longer. This is something of a happy accident, since the number of minimization iterations performed translates into extra computational cost. If a balanced analysis state is desired, our results suggest that 4DVAR, run to as many iterations as are needed to fit the slow motion, will be both the cheapest and most balance-preserving algorithm.

Though it was found that balance could be retained considerably if the initial-guess covariance field involved some form of tangent-linear balance transformation, such a constraint is not exact if either the model or the balance relationship is nonlinear. Alternatively, it is also possible to balance-constrain the assimilation externally, for example by performing the analysis step on the slow manifold and inferring analysis fast variables diagnostically. Experiments with the EKF (fig. 4.8) showed that such a constraint can significantly improve the analysis in regimes where balance is otherwise lost, but relies upon accurate knowledge of the balance relationship. The usefulness of external balance constraints for 4D assimilation is discouraging, since, after all, one of the goals of 4D data assimilation is to do without external covariance model constraints. In fact it was found that recovery of the slow mode was relatively unaffected by application of balance constraints. This implies that the overall advantages of 4D data assimilation over 3D assimilation would remain even if balance is enforced externally. However, we found no results suggesting that application of an external balance constraint will further improve 4D assimilation relative to 3D.

We mention in passing that Lorenc [2003b] has argued that 4DVAR may also be further preferable to the EnKF, because the EnKF can incur spurious imbalance from the localization of error covariances [Hamill et al., 2001, Houtekamer and Mitchell, 2001, Mitchell et al., 2002, Lorenc, 2003b, Houtekamer and Mitchell, 2005, Houtekamer et al., 2005]. Localization of covariances is required when the number of ensemble members is less than the number of observations being assimilated (in which case the problem becomes ill-posed), and is typically needed to avoid the effect of spurious correlations where variances are small.

The other side of the balance problem is the recovery of both timescales when the true state consists of a vortical mode and free unbalanced motion (chapter 5). In this case, it can be shown that the accurate representation of covariances which result from the balance relationship is still important, but that the covariance model now also needs to account for variance ascribed to the free fast motion. Choice of observation variables now also becomes an issue. If fast variables are not assimilated, only the component which is slaved to the slow mode can be controlled by observations. In this case, the results from the balanced-truth case can be carried forward in that fast-slow error covariances must be modeled accurately in order not to induce spurious, uncontrolled fast motion. When the gravity wave is unobserved and only slow variables are assimilated, recovery of the slow mode is compromised and a spurious gravity wave is induced — a familiar problem for which nonlinearity-preserving algorithms are again needed. Not surprisingly, the EnKF and 4DVAR are much more able to ignore the gravity wave if no fast variables are observed [e.g. figs. 5.13 and 5.21].

If fast variables are observed, the problem becomes one of distinguishing between the two modes: recovering a chaotic mode while also finding a gravity wave's magnitude and phase based on observations which are by-and-large infrequent relative to the period of the wave. The ability of each method to capture gravity waves of different magnitudes is summarized in figure 7.1 (b), which shows average fast errors as a function of $\tilde{I}^{\mathrm{t}}$. Here we see a very different result than the one shown in panel (a): the 15-member EnKF still has the lowest errors, but it is now followed closely in accuracy by the EKF, not 4DVAR. Despite its inability to capture fast-slow error covariances, the EKF is quite adept at capturing the gravity wave, especially its phase, as long as observations are assimilated which project onto the gravity wave.

The EnKF proved to be even better than the EKF at recovering the gravity wave, and was quite flexible in various regimes of imbalance and timescale separation, and stable over large Δt^{obs}. Of course, this result also hinged on ensemble size, with the EnKF losing accuracy relative to the EKF for ensembles less than about 8 members (fig. 5.11), i.e the benefits of the EnKF required ensemble size of more than twice the state dimension of the model. A weakness of the EnKF was that the lack of chaos in the fast mode gave ensemble members a tendency to phase-lock onto the wrong gravity wave. Slight overestimation of ensemble fast variance alleviated this problem, but only to an extent.

4DVAR, on the other hand, had great difficulty recovering the gravity wave, both in magnitude and phase, unless the number of cost function minimization iterations was increased beyond what was necessary to converge in the slow mode. Figure 7.1 (d) shows the decrease in fast error as a function of minimization iteration, relative to the other three methods, for $\tilde{I}^{\mathrm{t}} = 1.5$. For the exL86 model, between two and four iterations are needed to fit the slow mode but, as can be seen in the figure, at least six iterations are needed in order to surpass the EKF and EnKF in fast mode accuracy.

Sensitivity to the physical parameters ϵ and b, which govern the balance relationship and the free gravity wave, was tested primarily in the Kalman filters. It was found that increasing either parameter had nearly the same effect on the EKF as on the EnKF. Even though both parameters control the degree of nonlinearity of the balance relationship, increasing either parameter did not directly influence the ability of each method to capture balanced covariances. This indicates that changing ϵ and b primarily changes the information content of the observations, while the intrinsic nonlinearity of the balance relationship does not seem to matter. It was also found that observations of both timescales become more important as the period of the true gravity wave, or the coupling between the two modes, increase. In these limits, the value of time-evolved error covariances was also found to increase (figs. 5.6-5.7 and 5.14-5.15). The EKF and EnKF were also found to more easily overcome the error associated with observing at or near the gravity wave period, because information is spread forward in time.

The majority of the numerical experiments were performed in a perfect model context. Since a substantial amount of error in 4D data assimilation can come from accumulated error in the assimilation process, such experiments reveal the practical differences between assimilation algorithms. Experiments with model error in the slow mode (§6.2.2) showed that systematic error further exacerbates divergence of the Kalman filter, causing the EnKF ensemble to tend more strongly towards spurious phase-locking. Experiments which introduced model error into the model-estimated gravity wave parameters suggested that, ultimately, the ability to capture the fast wave may depend on the accuracy of the modeled frequencies. Simple tests of bias-estimation modifications to the EnKF were found to be quite harmful to the analysis if the truth is balanced, but improve the recovery of both modes if it is unbalanced.

7.2 Points for Practical Implementation and Future Research

Out of the above results, it is possible to distill some overall conclusions for assimilation with operational models and real observations.

- 4DVAR does not require the generation of an ensemble and yet tends to preserve balance in the context of a balanced truth. As such, 4DVAR could be a useful and cheaper algorithm for assimilation in cases where balance dynamics dominate.

- For unbalanced true states (roughly speaking, the tropics, middle atmosphere, and meso- and convective scales), the value of time-evolved error covariances, relative to static covariances, increases. There are at least two reasons for this: (a) 4D algorithms can better distinguish between slow and fast modes because the error covariances associated with slaving are estimated following the flow, and (b) 4D algorithms can overcome the error associated with observing at or near the periods of the fast waves, because information is spread forward in time. As operational model lids are raised and observations become more diverse and abundant, it will therefore be very important to pursue the development of accurate and efficient 4D data assimilation.

- Observations which project onto the fast waves are important in unbalanced-truth applications. This is especially true as parameters which govern the influence of the estimated fast waves on the estimated slow flow (e.g. the true gravity wave magnitude, the timescale ratio ϵ, or the rotational Froude number b) increase.

- Though it is not yet known what practical ensemble sizes are needed to realize the advantages of the EnKF in the unbalanced regions of the atmosphere, it is possible that the number might be computationally infeasible. Likewise, it is plausible that the observation frequency required to make the EKF (or a lower-order approximation to it) stable is not achievable. In these cases, 4DVAR would remain as the only option, yet the ability of 4DVAR to capture a partially-observed fast wave is questionable. Thus no scheme can presently be pin-pointed as the best choice for assimilation in the middle atmosphere and tropics. It may be that the best method to use will depend on individual applications, or that some "hybrid" scheme, which combines aspects of each method, could be the best choice.

Points which require further research are as follows:

- The gravity wave in the exL86 model neither propagates away nor is dissipated between observation times. Szunyogh et al. [2005] point out that, realistically, gravity wave events may not be present in the assimilation domain throughout the entire observation interval, and thus may not be captured. To address this issue, an analysis similar to the present study could be carried out with a model that is more complex than the exL86 model, yet still simple enough to retain the transparency of this analysis. A model which admits more than one gravity wave frequency, or more spatial degrees of freedom, would make it possible to address the effects of, say, geostrophic adjustment, localization of error covariances, observability of a gravity wave over a spatial observation grid, or propagation of a true gravity wave in space.

- A rigorous comparison of the balance properties of the EnKF and 4DVAR in both the balanced-truth and unbalanced-truth contexts could be carried out in a larger model. In particular, we would like to know whether the inability of 4DVAR to capture fast waves, and its strong tendency to retain balance, remain as more degrees of freedom are added to the model. Moreover, it is important to establish, in terms of operational models, the assimilation parameters (e.g. ensemble size and minimization iteration cutoff values) at which computational cost for each method outweighs its advantages.

- A model in which the fast mode is chaotic (in which case slaving is impossible), such as the two-timescale model of Lorenz (1995), is another interesting context for examining assimilation for multiple time scales, and could address problems of mesoscale and convective-scale assimilation [e.g. Snyder and Zhang, 2003].

- Since the effects of gravity waves in the atmosphere are generally parameterized, it would be interesting to study the problem of fitting gravity wave *parameters*, rather than the fast variables, to gravity wave observations. For the exL86 model, for example, one could make ϵ and b part of the analysis vector, then solve for the optimal gain matrix or minimize the cost function with respect to these parameters.

- Another complication not discussed here might be the effect of time-interpolation of observations, wherein observations made over a time window are taken as valid at the central time. Time interpolation of the model state to the observations relies on the assumption of a balanced short-range forecast [Houtekamer and Mitchell, 2005]. For an unbalanced truth, time-interpolation will likely become more difficult.

- Finally, a more thorough extension to tropical data assimilation, where special wave solutions arise (due to the change of sign of f at the equator) and complicate the dynamical picture, is necessary. An example of such a study is that of Žagar et al. [2004b]. It would be very enlightening to compare the perfomance of the EKF and EnKF in similar experiments.

Appendix A

Derivation of the exL86 Model

The following is a summary derivation of the exL86 model, tracing the development of this simple system through four papers: Lorenz [1980], Lorenz [1986], Bokhove and Shepherd [1996], and Wirosoetisno and Shepherd [2000].

In Lorenz [1980], the f-plane shallow water equations are nondimensionalized and then simplified by expanding vorticity, divergence, and height as an interacting resonant wave triad. This yields a system of nine equations which describe the evolution of the vorticity, divergence, and height of three interacting waves. These amplitudes are then transformed into normal modes, corresponding to potential vorticity, divergence, and geostrophic imbalance. In Lorenz [1986], the latter two variables are eliminated for two of the three waves, which leaves two geostrophic vorticity modes, and a third wave which admits both vortical motion and a gravity wave. U and V are the vorticity amplitudes of the two truncated waves, and W, X, and Z are, respectively, the potential vorticity, divergence, and geostrophic imbalance of the third wave. The equations which describe their evolution are given by:

$$\frac{dU}{dT} = -VW + bVZ \tag{A.1}$$

$$\frac{dV}{dT} = UW - bUZ \tag{A.2}$$

$$\frac{dW}{dT} = -UV \tag{A.3}$$

$$\frac{dX}{dT} = -Z \tag{A.4}$$

$$\frac{dZ}{dT} = bUV + X \tag{A.5}$$

These equations describe a nonlinearly interacting vorticity triad (U, V, and W), coupled to an inertia-gravity wave (X and Z). The parameter b is the rotational Froude number of the third wave, defined as $b \equiv fL/\sqrt{gH}$, where f is the Coriolis parameter, L represents the horizontal lengthscale of the motion, g is the acceleration due to gravity, and H the depth of the fluid.

Bokhove and Shepherd [1996] emphasize the timescale separation between the nonlinear vortical mode and the gravity wave by scaling the variable amplitudes $U =: \epsilon u$, $V =: \epsilon v$, $W =: \epsilon w$, $X =: \epsilon x$, $Z =: \epsilon z$, and scaling time $T =: t/\epsilon$. The scaled system is

$$\frac{du}{dt} = -vw + bvz \tag{A.6}$$

$$\frac{dv}{dt} = uw - buz \tag{A.7}$$

$$\frac{dw}{dt} = -uv \tag{A.8}$$

$$\frac{dx}{dt} = -\frac{z}{\epsilon} \tag{A.9}$$

$$\frac{dz}{dt} = buv + \frac{x}{\epsilon} \tag{A.10}$$

For $\epsilon \ll 1$, x and z vary on a timescale that is fast compared to the evolution of u, v and w. From the dimensions of the original equations, it can be shown that ϵ, which defines the inverse of the ratio between the advective timescale (corresponding to t) and that of the inertia-gravity wave, is given by $\epsilon = Rb/\sqrt{1+b^2}$, where $R = U/fL$ is the Rossby number. b measures the importance of vorticity relative to the gravitational restoring force, while R measures the importance of rotation relative to the planet's rotation (and is thus a non-dimensional measure of amplitude). This means that ϵ is

$$\epsilon = \frac{U}{\sqrt{gH + f^2L^2}} = \frac{U}{c_{\mathrm{IGW}}}, \tag{A.11}$$

where $c_{\mathrm{IGW}} = \sqrt{gH + f^2L^2}$ is the phase speed of the inertia-gravity wave.

This system is further simplified by noting the invariant $C^2 = u^2 + v^2$, and defining $u =: \sqrt{C}\cos\phi'$, $v =: \sqrt{C}\sin\phi'$, and $\phi := \phi' - \epsilon bx$. This yields the following four-variable system:

$$\begin{aligned}
\frac{d\phi}{dt} &= w && \text{(A.12)}\\
\frac{dw}{dt} &= -\frac{C}{2}\sin\left(2\phi + 2\epsilon bx\right) && \text{(A.13)}\\
\frac{dx}{dt} &= -\frac{z}{\epsilon} && \text{(A.14)}\\
\frac{dz}{dt} &= \frac{x}{\epsilon} + \frac{bC}{2}\sin\left(2\phi + 2\epsilon bx\right). && \text{(A.15)}
\end{aligned}$$

Bokhove and Shepherd [1996] showed that the vortical mode in (A.12)-(A.15) has periodic solutions for most balanced initial conditions. In order to make the evolution of ϕ and w sufficiently chaotic, Wirosoetisno and Shepherd [2000] let C vary in time as $C(t) = k_0 + k_1\cos\gamma t$.

Since observed quantities are *not* clearly separated into "slow" or "fast" variables, it makes sense to transform w and z in the above system back to mixed variables, for the purposes of observations. This is simply done by defining $w \equiv w' + bz'$ and $z \equiv z' - bw'$. Though the normal-mode system is used in this thesis, the mixed-timescale system is given here for completeness:

$$\begin{aligned}
\frac{d\phi}{dt} &= w' + bz' && \text{(A.16)}\\
\frac{dw'}{dt} &= -\frac{C}{2}\sin 2(\phi + \epsilon bx) - \frac{\alpha^2 b}{\epsilon}x && \text{(A.17)}\\
\frac{dx}{dt} &= \frac{bw' - z'}{\epsilon} && \text{(A.18)}\\
\frac{dz'}{dt} &= \frac{\alpha^2 x}{\epsilon} && \text{(A.19)}
\end{aligned}$$

where $\alpha = (1+b^2)^{-\frac{1}{2}}$. In this system, ϕ describes the (geostrophic) vorticity of wave components 1 and 2, and w', x, and z' the (non-geostrophic) vorticity, divergence, and height, respectively, of wave component 3.

Slaving relations can be derived by postulating slaving of the form

$$\mathbf{f} = U\left(\mathbf{s}; \epsilon\right), \tag{A.20}$$

for each of the fast variables, following Warn et al. [1995]. Now we seek a functional form for $\mathbf{f}$. This can be done by expanding the slaving relations in ϵ, as

$$U_{x,z} = U_{x,z}^{(0)}\left(\phi, w\right) + \epsilon U_{x,z}^{(1)}\left(\phi, w\right) + \epsilon^2 U_{x,z}^{(2)}\left(\phi, w\right) + \ldots, \tag{A.21}$$

then substituting these expansions into [(A.19)-(A.19)]. Up to $\mathcal{O}\left(\epsilon^2\right)$, this gives the slaving relations,

$$x = U_x(\phi;\epsilon) = -\frac{\epsilon}{2}Cb\sin 2\phi + O(\epsilon^3) \tag{A.22}$$

$$z = U_z(\phi, w;\epsilon) = \epsilon^2(Cbw\cos 2\phi + \frac{C'}{2}b\sin 2\phi) + O(\epsilon^3). \tag{A.23}$$

Appendix B

Deterministic Ensemble Filters

The experiments of chapters 3-5 showed that two shortcomings of the EnKF are that (1) it incurs sampling error for small ensemble sizes, and (2) it assumes a Gaussian forecast error distribution even though the forecast ensemble may be very non-Gaussian.

One source of sampling error in the EnKF is the perturbation of observations for each ensemble member (2.46). To prevent this problem, several methods have been proposed in recent years [Anderson and Anderson, 1999, Tippett, 2002, Whitaker and Hamill, 2002, Szunyogh et al., 2005] wherein the correct analysis error distribution is computed first [according to (2.36)], and the ensemble is then generated according to this distribution. These alternative formulations can be thought of as "deterministic" ensemble filters, in contrast to the "stochastic" perturbed-observation EnKF [Leeuwenburgh et al., 2005]. The ensemble's non-Gaussianity, particularly at large Δt^{obs}, is addressed by Anderson and Anderson [1999], who suggest a possible remedy in the so-called kernel or partcle filter, which is based upon the principle that we are seeking not a state, but a pdf of states, given background and observation pdfs.

In this Appendix, we briefly consider two algorithms which fall into the category of deterministic filters: the Ensemble Square Root Filter [Whitaker and Hamill, 2002] and a particle filter [Anderson and Anderson, 1999, Pham, 2001, Xiong et al., 2006].

B.1 Ensemble Square Root Filter

In a square root filter [Tippett, 2002], one solves for the transform matrix $\mathbf{T}_k$ which satisfies

$$\mathbf{Z}_k^{\mathrm{f}}\mathbf{T}_k\mathbf{T}_k^{\mathrm{T}}{\mathbf{Z}^{\mathrm{f}}_k}^{\mathrm{T}} = \mathbf{P}_k^{\mathrm{a}} = \left(\mathbf{I}_n - \mathbf{K}_k\mathbf{H}\right)\mathbf{P}_k^{\mathrm{f}}, \tag{B.1}$$

where

$$\mathbf{P}_k^{\mathrm{f}} \equiv \mathbf{Z}_k^{\mathrm{f}}{\mathbf{Z}_k^{\mathrm{f}}}^{\mathrm{T}}. \tag{B.2}$$

The square root of the forecast error covariance matrix is computed from the ensemble as $\mathbf{Z}_k^{\mathrm{f}} = \mathbf{X}_k^{\mathrm{f}} \cdot (N-1)^{-1/2}$, where $\mathbf{X}_k^{\mathrm{f}}$ represents an $n \times N$ matrix which contains the ensemble of deviations from the mean. Because the matrix square root is nonunique, several methods of solving for $\mathbf{T}_k$ exist. If implemented in the same model, these would yield different analysis ensembles which span the same subspace. Their theoretical equivalence and practical differences are described in Tippett [2002].

A particularly simple method of computing $\mathbf{T}_k$ is given by the Ensemble Square Root Filter (EnSRF) of Whitaker and Hamill [2002]. Here, observations are assimilated one at a time, which makes the matrix square root a simple scalar operation. This also lowers computational cost for more complex systems, where the matrix square root computation is expensive. As long as observations are uncorrelated with one another (which is certainly the case for our experiments), this algorithm is equivalent to other variants of the square root method. (Further discussion on the validity of serial assimilation of observations is given in Houtekamer and Mitchell [2001] and Bishop et al. [2001].)

Following the notation of Tippett [2002], the EnSRF algorithm is implemented as follows. For each observation, the matrix $\mathbf{Z}_k^{\mathrm{f}}$ is multiplied by a transform matrix

$$\mathbf{T}_{j,k} = \mathbf{I}_n - \gamma_{k,j}\mathbf{V}_{k,j}\mathbf{V}_{k,j}^{\mathrm{T}}, \tag{B.3}$$

where

$$\begin{aligned}
\mathbf{V}_{k,j} &= \mathbf{H}_j\mathbf{Z}^{\mathrm{f}}_k && \text{(B.4)}\\
\beta_{k,j} &= \left(\mathbf{D}_{k,j} + \sqrt{\mathbf{R}_j\mathbf{D}_{k,j}}\right)^{-1} && \text{(B.5)}\\
\mathbf{D}_{k,j} &= \mathbf{H}_j\mathbf{P}^{\mathrm{f}}_k\mathbf{H}^{\mathrm{T}}_j + \mathbf{R}_j. && \text{(B.6)}
\end{aligned}$$

Here j denotes individual observations, and $\mathbf{R}_j$ and $\mathbf{H}_j$ represent the error covariance and observation operator, respectively, for individual observations. $\mathbf{R}_j$, $\mathbf{H}_j$, $\mathbf{D}_{k,j}$ and $\mathbf{V}_{k,j}$ are actually scalars when observations are serially assimilated, but are written as matricies for consistency with the general algorithm.

It can be shown that

$$\langle\mathbf{Z}^{\mathrm{a}}_k\mathbf{Z}^{\mathrm{a}}_k\rangle^{\mathrm{T}} = \left(\mathbf{I}_n - \mathbf{K}_k\mathbf{H}\right)\mathbf{P}^{\mathrm{f}}_k = \mathbf{P}^{\mathrm{a}}_k. \tag{B.7}$$

Ensemble members are then updated by adding the updated perturbations

$$\mathbf{Z}^{\mathrm{a}}_k = \mathbf{Z}^{\mathrm{f}}_k\prod_{j=1}^{m}\mathbf{T}_{j,k} \tag{B.8}$$

back to the ensemble mean. Then the analysis state is, as before, given by the ensemble mean.

B.2 Particle or Kernel Filter

In a so-called kernel filter, the joint pdf of background and observation errors, rather than the covariance matrix, is estimated at observation times. At the analysis time, new ensemble members are drawn from the conditional posterior pdf

$$p_{\mathrm{a}}\left(\mathbf{x}|\mathbf{z}\right) = \frac{1}{N}p_{\mathbf{R}}\left(\mathbf{z}|\mathbf{x}\right)p_{\mathbf{B}}\left(\mathbf{x}\right). \tag{B.9}$$

The analysis state and error covariance matrix can then be computed by resampling an ensemble out of the pointwise product (B.9). This requires a model of the observation and forecast error distribution, with the latter being estimated from the ensemble. Xiong

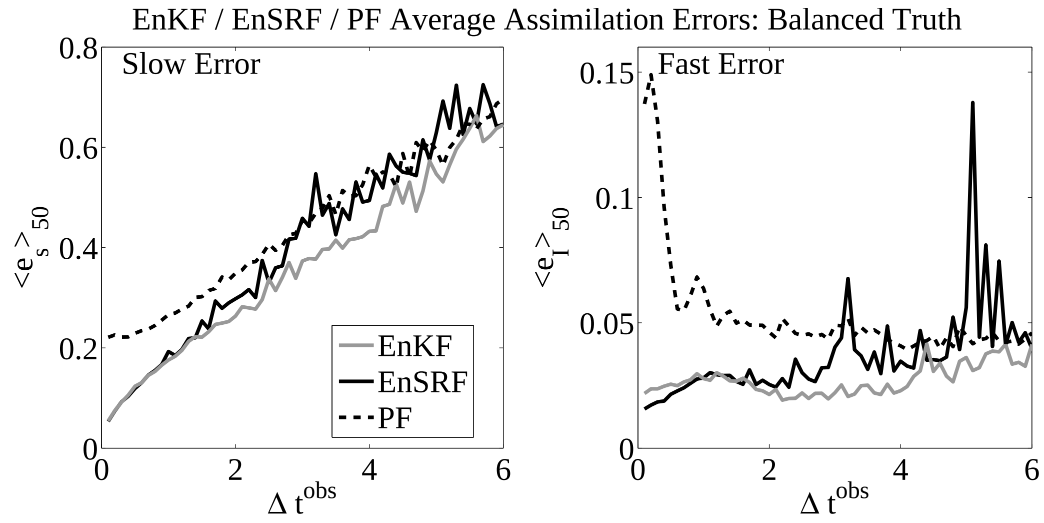

Figure B.1: Comparison of average analysis errors between the standard (perturbed-observation) EnKF, the serial EnSRF, and the PF as a function of observation interval. As in previous figures, slow errors (2.59) are shown in the left panel, and fast errors (2.60) in the right panel. Each line represents an average over 50 experiments.

et al. [2006] suggest that one way of approximating the pointwise product is to give each ensemble member a weight based on the relative probability of its innovation:

$$p_{i,k} = \frac{p_{\mathbf{R}}\left(\mathbf{d}_{i,k}\right)}{\sum_{l=1}^{N} p_{\mathbf{R}}\left(\mathbf{d}_{l,k}\right).}, \tag{B.10}$$

Ensemble members with probability below a certain level are discarded, and new ones are generated according to a Gaussian distribution which is fit to the remaining ensemble members. Methods of this type are generally referred to as kernel or particle filters; using the specific algorithm of Xiong et al. [2006], we use the term particle filter, or PF.

B.3 Quantitative Comparisons

Figure B.1 shows average slow and fast analysis errors for balanced-truth experiments and over a range of observation intervals, comparing the EnKF, EnSRF, and PF. 10-member ensembles are used for all three filters.

EnSRF errors are similar to the EnKF, but differ in two ways: (a) as observation interval increases, the EnSRF has more cases of large, spurious gravity waves, reflected in the large peaks in $\langle e_I \rangle$ and a larger overall increase in $\langle e_s \rangle$, and (b) the EnSRF returns the most balanced analyses at smaller Δt^{obs}, rather than at $\Delta t^{\mathrm{obs}} \sim 3$ (which is the optimal observation interval in the EnKF).

The PF shows a radically different behavior, in terms of balance, from the other two methods. Unlike the EnKF and EnSRF, fast errors in the PF are *largest* for short Δt^{obs}, and decrease as the time between observations increases. For $\Delta t^{\mathrm{obs}} \lesssim 2$ we find that the PF has a tendency to give most of the weight to one highly unbalanced member of the ensemble, and less weight to the rest of the ensemble, which has less spurious imbalance. The extreme loss of balance at frequent observations also causes larger slow mode errors at these observation intervals. There is thus no point on this plot where the PF yields a better analysis than the other two methods.

To expand the analysis to unbalanced truth states, figure B.2 compares average EnSRF and PF analysis errors to EnKF errors over a range of magnitudes of the true gravity wave, now also comparing average errors in gravity wave phase. Here the performance of the EnSRF and EnKF is similar overall, though the EnKF on average has lower slow mode errors. The PF shows significantly higher errors in both modes, though error is still below the no-skill levels throughout.

B.4 Discussion

The above experiments are provided as a brief exploration of how robust our results regarding the ensemble-based covariance model are when extended to other ensemble-based filters. We therefore forego a detailed interpretation of the results, keeping in mind that thorough, unified formulations of the two algorithms introduced here have not yet been developed.

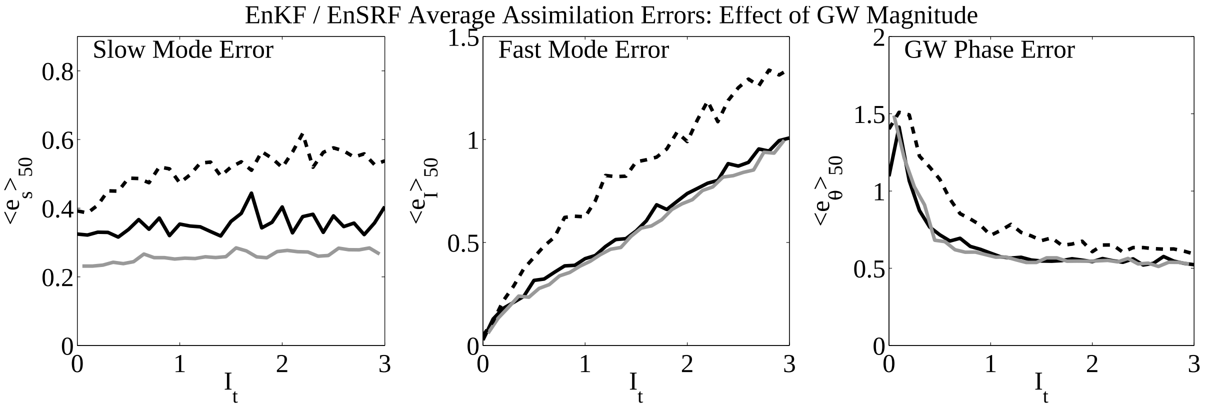

Figure B.2: Comparison of average analysis errors between the standard (perturbed-observation) EnKF, the serial EnSRF, and the PF, as a function of the true-state gravity wave magnitude. As in previous figures, slow errors (2.59) are shown in the left panel, and fast errors (2.60) in the middle and right panels. Each line represents an average over 50 experiments.

Deterministic filters have been shown in some studies to be less susceptible to filter divergence than the standard EnKF [e.g. Whitaker and Hamill, 2002, Anderson and Anderson, 1999], while other studies [e.g. Lawson and Hansen, 2004, Leeuwenburgh et al., 2005] have argued that deterministic filters break down more easily than stochastic filters as the nonlinearity of the assimilation system increases. Our results seem to suggest the latter; by forcing the ensemble to represent the posterior error distribution which is consistent with the minimum-variance estimate, some physicality in the ensemble-estimated covariance model seems to be lost. As a result, loss of balance in both the EnSRF and PF is greater, and recovery of the slow mode in the presence of an unbalanced truth becomes more difficult.

In regards to the PF, Anderson and Anderson [1999] argue that a more accurate representation of the prior pdf should also yield more balanced analyses, thereby making the PF preferable in a balance context. In particular, Anderson and Anderson [1999] perform experiments with the 9-variable Lorenz [1980] model (see appendix A) which suggest that this might indeed be the case. Our results, however, have more similarity with those of van Leeuwen [2003], who shows that in some regimes the PF ensemble collapses to a single ensemble member, because sampling error accrues too quickly in the ensemble representation of the prior pdf. The difference between our results and those of Anderson and Anderson [1999] may be due to nonlinearity and chaos. This is plausible since the 9-component Lorenz [1980] model and the exL86 model differ primarily in that the former is integrable (when balanced), while the later is chaotic.

Bibliography

Alexander, M. J., and C. Barnet, 2006: Using satellite observations to constrain parameterizations of gravity wave effects for global models. *J. Atmos. Sci.*, **63**, 2963–2977.

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

Bergman, K. H., 1979: Multivariate analysis of temperatures and winds using optimal interpolation. *Mon. Wea. Rev.*, **107**, 1423–1444.

Bevington, P. R., and D. K. Robinson, 1992: *Data Reduction and Error Analysis for the Physical Sciences*. 2nd ed., McGraw-Hill.

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

Bokhove, O., and T. G. Shepherd, 1996: On Hamiltonian balanced dynamics and the slowest invariant manifold. *J. Atmos. Sci.*, **53**, 276–297.

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

Burgers, G., M. A. Balmaseda, F. C. Vossepoel, G. J. van Oldenborgh, and P. J. van

Leeuwen, 2002: Balanced ocean-data assimilation near the equator. *J. Phys. Ocean.*, **32**, 2509–2529.

Cohn, S. E., and D. F. Parrish, 1999: The behavior of forecast error covariances for a Kalman filter in two dimensions. *Mon. Wea. Rev.*, **119**, 1757–1785.

Courtier, P., and O. Talagrand, 1990: Variational assimilation of meteorological observations with the direct and adjoint shallow-water equations. *Tellus*, **42A**, 531–549.

Cushman-Roisin, B., 1994: *Introduction to Geophysical Fluid Dynamics*. Prentice Hall.

Daley, R., 1991: *Atmospheric Data Analysis*. Cambridge University Press.

Daley, R., and K. Puri, 1980: Four-dimensional data assimilation and the slow manifold. *Mon. Wea. Rev.*, **108**, 85–99.

Dee, D. P., 1991: Simplification of the Kalman filter for meteorological data assimilation. *Quart. J. Roy. Meteorol. Soc.*, **117**, 367–384.

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

Ehrendorfer, M., 2006: . Review of issues concerning ensemble-based data assimilation techniques, Presentation at the Seventh International Workshop on Adjoint Applications in Dynamic Meteorology.

Errico, R. M., 1997: What is an adjoint model? *Bull. Am. Met. Soc.*, **78**, 2577–2591.

Evensen, G., 1994: Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast errors statistics. *J. Geophys. Res.*, **C99**, 10,143–10,162.

Evensen, G., 1997: Advanced data assimilation for strongly nonlinear dynamics. *Mon. Wea. Rev.*, **125**, 1342–1354.

Evensen, G., 2003: The ensemble Kalman filter: theoretical formulation and practical implementation. *Ocean Dynamics*, **53**, 343–367.

Fisher, M., M. Leutbecher, and G. Kelly, 2005: On the equivalence between Kalman smoothing and weak-constraint four-dimensional variational data assimilation. *Quart. J. Roy. Meteorol. Soc.*, **131**, 3235–3246.

Gauthier, P., and J.-N. Thépaut, 2001: Impact of the digital filter as a weak constraint in the preoperational 4dvar assimilation system of Metéo-France. *Mon. Wea. Rev.*, **129**, 2089–2102.

Ghil, M., S. Cohn, J. Tavantzis, K. Bube, and E. Isaacson, 1981: *Application of estimation theory to numerical weather prediction*, pp. 139–224, in *Dynamic Meteorology*, Springer-Verlag.

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

Hoteit, I., G. Korres, and G. Triantafyllou, 2005: Comparison of extended and ensemble based Kalman filters with low and high resolution primitive equation models. *Non. Proc. Geophys.*, **12**, 755–765.

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

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

Houtekamer, P., H. L. Mitchell, G. Pellerin, M. Buehner, M. Charron, L. Spacek, and B. Hansen, 2005: Atmospheric data assimilation with an ensemble Kalman filter: Results with real observations. *Mon. Wea. Rev.*, **133**, 604–620.

Houtekamer, P. L., and H. L. Mitchell, 2005: Ensemble Kalman filtering. *Quart. J. Roy. Meteorol. Soc.*, **131**, 3269–3289.

Kalman, R., 1960: A new approach to linear filtering and prediction problems. *Trans. ASME, Ser. D, J. Basic Eng.*, **82**, 35–45.

Kalman, R., and R. Bucy, 1961: New results in linear filtering and prediction theory. *Trans. ASME, Ser. D, J. Basic Eng.*, **83**, 905–108.

Kalnay, E., 2003: *Atmospheric Modeling, Data Assimilation and Predictability*. Cambridge University Press.

Koshyk, J. N., B. A. Boville, K. Hamilton, E. Manzini, and K. Shibata, 1999: Kinetic energy spectrum of horizontal motions in middle-atmosphere modles. *J. Geophys. Res.*, **104**, 27,177–27,190.

Lawson, W. G., and J. A. Hansen, 2004: Implications of stochastic and deterministic filters as ensemble-based data assimilation methods in varying regimes of error growth. *Mon. Wea. Rev.*, **132**, 1966–1981.

Lea, D. J., T. W. Haine, and R. F. Gasparovic, 2006: Observability of the Irminger sea circulation using variational data assimilation. *Quart. J. Roy. Meteorol. Soc.*, **132**, 1545–1576.

Leeuwenburgh, O., G. Evensen, and L. Bertino, 2005: The impact of ensemble filter definition on the assimilation of temperature profiles in the tropical pacific. *Quart. J. Roy. Meteorol. Soc.*, **131**, 3191–3300.

Leith, C. E., 1979: Nonlinear normal mode initialization and quasi-geostrophic theory. *J. Atmos. Sci.*, **37**, 958–968.

Lorenc, A. C., 2003a: Modelling of error covariances by 4d-var data assimilation. *Quart. J. Roy. Meteorol. Soc.*, **129**, 3167–3182.

Lorenc, A. C., 2003b: The potential of the ensemble Kalman filter for NWP — a comparison with 4d-var. *Quart. J. Roy. Meteorol. Soc.*, **129**, 3183–3204.

Lorenc, A. C., and F. Rawlins, 2005: Why does 4d-var beat 3d-var? *Quart. J. Roy. Meteorol. Soc.*, **131**, 3247–3257.

Lorenz, E. N., 1963: Deterministic non-periodic flow. *J. Atmos. Sci.*, **20**, 130–141.

Lorenz, E. N., 1980: Attractor sets and quasi-geostrophic equilibrium. *J. Atmos. Sci.*, **37**, 1685–1699.

Lorenz, E. N., 1986: On the existence of a slow manifold. *J. Atmos. Sci.*, **43**, 1547–1558.

Lynch, P., 2002: *The Swinging Spring: a Simple Model for Atmospheric Balance*, pp. 64–108, in *Large–Scale Atmosphere–Ocean Dynamics: Vol II: Geometric Methods and Models*, Cambridge University Press.

Ménard, R., and R. Daley, 1996: The application of Kalman smoother theory to the estimation of 4d-var error statistics. *Tellus*, **48A**, 221–237.

Miller, R. N., M. Ghil, and F. Gauthiez, 1994: Advanced data assimilation in strongly nonlinear dynamical systems. *J. Atmos. Sci.*, **51**, 1037–1056.

Mitchell, H. L., P. L. Houtekamer, and G. Pellerin, 2002: Ensemble size, balance, and model-error representation in an ensemble Kalman filter. *Mon. Wea. Rev.*, **130**, 2791–2808.

Neef, L. J., S. M. Polavarapu, and T. G. Shepherd, 2006: Four-dimensional data assimilation and balanced dynamics. *J. Atmos. Sci.*, **60**, 1840–1858.

Neef, L. J., S. M. Polavarapu, and T. G. Shepherd, 2007: Gravity waves in nonlinear sequential data assimilation. *J. Atmos. Sci.*, p. submitted.

Nerger, L., W. Hiller, and J. Schröter, 2005: A comparison of error subspace Kalman filters. *Tellus*, **57A**, 715–735.

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

Pham, D. T., 2001: Stochastic methods for sequential data assimilation in strongly nonlinear systems. *Mon. Wea. Rev.*, **129**, 1194–1207.

Polavarapu, S., M. Tanguay, and L. Fillion, 2000: Four-dimensional variational data assimilation with digital filter initialization. *Mon. Wea. Rev.*, **128**, 2491–2510.

Polavarapu, S., T. G. Shepherd, Y. Rochon, and S. Ren, 2005: Some challenges of middle atmosphere data assimilation. *Quart. J. Roy. Meteorol. Soc.*, **131**, 3513–3527.

Rabier, F., 2005: Overview of global data assimilation developments in numerical weather-prediction centres. *Quart. J. Roy. Meteorol. Soc.*, **131**, 3215–3233.

Richardson, L. F., 1922: *Weather prediction by numerical process*. 2nd ed., Cambridge University Press.

Saujani, S., and T. G. Shepherd, 2006: A unified theory of balanced dynamics in the extratropics. *J. Fluid Mech.*, **569**, 447–464.

Shepherd, T. G., 2000: The middle atmosphere. *J. Atmos. Solar-Terr. Phys.*, **62**, 1587–1601.

Shewchuk, J. R., 1994: An introduction to the conjugate gradient method without the agonizing pain. *Unpublished Manuscript, available at http://www.cs.berkeley.edu/ jrs/*.

Snyder, C., and F. Zhang, 2003: Assimilation of simulated radar observations with an ensemble Kalman filter. *Mon. Wea. Rev.*, **131**, 1663–1677.

Szunyogh, I., E. Kostelich, G. Gyarmati, D. Patil, B. R. Hunt, E. Kalnay, E. Ott, and J. Yorke, 2005: Assessing a local ensemble Kalman filter: Perfect model experiments with the NCEP global model. *Tellus*, **57A**, 528–545.

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 data assimilation using the adjoint of a multilevel primitive equation model. *Quart. J. Roy. Meteorol. Soc.*, **117**, 1225–1254.

Tippett, M. K., 2002: Ensemble square root filters. *Mon. Wea. Rev.*, **131**, 1485–1490.

van Leeuwen, P. J., 2003: A variance-minimizing filter for large-scale applications. *Mon. Wea. Rev.*, **131**, 2071–2084.

Verlaan, M., and A. W. Heemink, 2001: Nonlinearity in data assimilation applications: A practical method for analysis. *Mon. Wea. Rev.*, **129**, 1578–1589.

Warn, T., 1997: Nonlinear balance and quasi-geostrophic sets. *Atmos.-Ocean*, **35**, 135–145.

Warn, T., O. Bokhove, T. G. Shepherd, and G. K. Vallis, 1995: Rossby number expansions, slaving principles, and balance dynamics. *Quart. J. Roy. Meteorol. Soc.*, **121**, 723–739.

Weaver, A., C. Deltel, E. Machu, S. Ricci, and N. Daget, 2005: A multivariate balance operator for variational ocean data assimilation. *Quart. J. Roy. Meteorol. Soc.*, **131**, 3605–3625.

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

Williamson, D. L., and C. Temperton, 1981: Normal mode initialization for a multilevel grid-point model, part ii: Nonlinear aspects. *Mon. Wea. Rev.*, **109**, 744–757.

Wirosoetisno, D., and T. G. Shepherd, 2000: Averaging, slaving and balanced dynamics in a simple atmospheric model. *Physica*, **D 141**, 37–53.

Xiong, X., I. M. Navon, and B. Uzunoglu, 2006: A note on the particle filter with posterior Gaussian resampling. *Tellus*, **58A**, 456–460.

Žagar, N., N. Gustafsson, and Källén, 2004a: Variational data assimilation in the tropics: the impact of a background-error constraint. *Quart. J. Roy. Meteorol. Soc.*, **130**, 103–125.

Žagar, N., N. Gustafsson, and E. Källén, 2004b: Dynamical response of equatorial waves in four-dimensional variational data assimilation. *Tellus*, **56A**, 29–46.