Aspects of the Ensemble Kalman Filter

David Livings

October 3, 2005

Abstract

The Ensemble Kalman Filter (EnKF) is a data assimilation method designed to provide estimates of the state of a system by blending information from a model of the system with observations. It maintains an ensemble of state estimates from which a single best state estimate and an assessment of estimation error may be calculated. Compared to more established methods it offers advantages of reduced computational cost, better handling of nonlinearity, and greater ease of implementation.

This dissertation starts by reviewing different formulations of the EnKF, covering stochastic and semi-deterministic variants. Two formulations are selected for implementation, and the adaptation of their algorithms for better numerical behaviour is described. Next, as a subject for experiments, a simple mechanical system is described that is of interest to meteorologists as an illustration of the problem of initialisation. Experimental results are presented that show some unexpected features of the implemented filters, including ensemble statistics that are inconsistent with the actual error. Explanations of these features are provided and point to a potential flaw in the general framework for semi-deterministic formulations of the EnKF, affecting some but not all such formulations. This flaw appears to have been overlooked in the literature.

Contents

List of Figures

List of Tables

List of Symbols

Latin

$\mathbf{A}$	Ensemble adjustment matrix in EAKF.
a (superscript)	Analysis value.
e (subscript)	Ensemble value.
$\mathbf{F}$	Orthogonal matrix.
f	Frequency.
f (superscript)	Forecast value.
$\mathbf{G}$	Diagonal matrix of square roots of eigenvalues.
g	Acceleration due to gravity.
H	Nonlinear observation operator; Hamiltonian.
$\mathbf{H}$	Linear observation operator; tangent linear operator of nonlinear observation operator.
$\mathbf{I}$	Identity matrix.
i (subscript)	Ensemble member.
$\mathcal{J}$	Cost function.
$\mathbf{K}$	Kalman gain matrix.
k	Elasticity of swinging spring.
ℓ	Equilibrium length of swinging spring.
ℓ_0	Unstretched length of swinging spring.
M	Nonlinear dynamical model.
$\mathbf{M}$	Linear dynamical model; tangent linear operator of nonlinear dynamical model.

m	Dimension of observation space; mass of bob of swinging spring.
N	Ensemble size.
n	Dimension of state space.
$\mathbf{P}$	State error covariance matrix.
p	Number of nonzero singular values in SVD; generalised momentum.
$\mathbf{Q}$	Model noise covariance matrix.
$\mathbf{R}$	Observation noise covariance matrix.
r	Radial polar coordinate.
r'	Radial displacement of swinging spring relative to equilibrium length.
$\mathbf{S}$	Ensemble version of innovations covariance matrix.
T	Period of oscillation.
$\mathbf{T}$	Post-multiplier in deterministic formulations of EnKF.
t (superscript)	True value; time.
t_k	Discrete times at which observations are available and forecasts and analyses are required.
$\mathbf{U}$, $\mathbf{V}$, $\mathbf{W}$	Orthogonal matrices.
$\mathbf{X}$	Ensemble matrix.
$\mathbf{x}$	State vector.
$\mathbf{Y}$	Observation ensemble matrix.
$\mathbf{y}$	Observation vector.
$\mathbf{Z}$	*Ad hoc* matrix.
$\mathbf{z}$	*Ad hoc* vector.

Greek

β	Coefficient in serial method formulation of EnKF.
ε	Ratio of frequencies of swinging spring.
$\boldsymbol{\varepsilon}$	Observation noise vector.
$\boldsymbol{\eta}$	Model noise vector.

θ	Angular polar coordinate.
$\mathbf{\Lambda}$	Diagonal matrix of eigenvalues.
$\mathbf{\Sigma}$	Matrix of singular values in SVD.
ω	Angular frequency.

Miscellaneous

Some of the vector operators below may be applied to scalars and matrices as well.

$\mathbf{1}_N$	$N \times N$ matrix in which every element is $1/N$.
$\langle \mathbf{z} \rangle$	Population mean of $\mathbf{z}$.
$\hat{\mathbf{z}}$	Version of $\mathbf{z}$ in augmented state space; $\mathbf{z}$ scaled by inverse square root of covariance matrix; Fourier transform of $\mathbf{z}$.
$\overline{\mathbf{z}}$	Ensemble mean of $\mathbf{z}$.
$\mathbf{Z}'$	Perturbation matrix of ensemble matrix $\mathbf{Z}$.
$\widetilde{\mathbf{z}}$	Analogue in EAKF of $\mathbf{z}$ or $\widehat{\mathbf{z}}$ in ETKF.

List of Abbreviations

EAKF	Ensemble Adjustment Kalman Filter
EKF	Extended Kalman Filter
EnKF	Ensemble Kalman Filter
ETKF	Ensemble Transform Kalman Filter
KF	Kalman Filter
NWP	Numerical Weather Prediction
SVD	Singular Value Decomposition

Chapter 1

Introduction

1.1 Background

Data assimilation addresses the problem of incorporating observations into a model of some system. For example, the system could be the atmosphere of the Earth, the model could be a weather forecasting model, and the observations could be measurements made by surface stations, radiosondes, weather radars, and satellites. In this case the problem is to combine the state of the atmosphere as predicted by an earlier forecast with recent observational data to produce an updated estimate of the state of the atmosphere (known as the analysis) that can be used as the starting point for a new forecast. For a detailed overview of data assimilation in a meteorological context see Kalnay [15] or Swinbank *et al* [22].

The data assimilation techniques of 3D-Var and 4D-Var are currently popular at national meteorology centres. These are variational techniques that use numerical methods to minimise a cost function that is a weighted measure of the distances from the analysis to the forecast and the observations. The weightings in the cost function are intended to reflect the relative uncertainties in different components of the forecast and observations. The resulting analysis thus represents a combination of the twin information sources of forecast and observation, with greater weight being given to more certain information. The difference between 3D-Var and 4D-Var is that 3D-Var treats all observations as having occurred at the forecast time, whilst 4D-Var takes some account of the evolution of the system between the forecast time and the observation time. A detailed comparison of 4D-Var and the Ensemble Kalman Filter methods that are the subject of this dissertation may be found in Lorenc [18]. A major part of the effort in implementing 3D-Var or 4D-Var lies in modelling the forecast uncertainty for use in the cost function. This uncertainty is usually modelled the same for all time, which is less than ideal because the forecast uncertainty will vary depending on both the quality of the observations contributing to the analysis on which it is based, and the way in which the system evolves between the analysis and the forecast.

An evolving forecast uncertainty is provided by the Extended Kalman Filter (EKF), which is a nonlinear generalisation of the linear Kalman Filter (KF). As well as an estimate of the state of the system, these filters maintain an error covariance matrix that acts as a measure of the uncertainty in the estimate. This covariance matrix is updated with every analysis to reflect the new information provided by the observations, and is evolved along with the state estimate from the time of an analysis to the time of the next forecast. The KF and EKF are described in more detail later in this dissertation; for a comprehensive treatment see Gelb [9] or Jazwinski [14]. The EKF rose to prominence in aerospace applications where the dimension of the state space for the model is relatively small, typically nine or less. Directly extending the filter to the sort of systems encountered in numerical weather prediction (NWP), where the state space dimension may be 10^7, is beyond the capabilities of current computer technology. The EKF also shares with 4D-Var the need to implement tangent linear operators (Jacobians) for the nonlinear forecast model and the model of how observations are related to the state of the system. Doing this for a large, complicated system such as an NWP model is labour-intensive.

The Ensemble Kalman Filter (EnKF) is an attempt to overcome the drawbacks of the EKF. Two of its key ideas are to use an ensemble (statistical sample) of state estimates instead of a single state estimate and to calculate the error covariance matrix from this ensemble instead of maintaining a separate covariance matrix. If we take an ensemble size that is small but not so small that it is statistically unrepresentative, then the extra work needed to maintain an ensemble of state estimates is more that offset by the work saved through not maintaining a separate covariance matrix. The EnKF also does not use tangent linear operators, which eases implementation and may lead to a better handling of nonlinearity.

The EnKF was originally presented in Evensen [7]. An important subsequent development was the recognition by Burgers *et al* [4] (and independently by Houtekamer and Mitchell [11]) of the need to use an ensemble of pseudo-random observation perturbations to obtain the right statistics from the analysis ensemble. Deterministic methods for forming an analysis ensemble with the right statistics have also been presented. The former approach to the EnKF is comprehensively reviewed in Evensen [8], whilst variants of the latter approach are placed in a uniform framework by Tippett *et al* [23]. These variants include the Ensemble Transform Kalman Filter (ETKF) of Bishop *et al* [3] and the Ensemble Adjustment Kalman Filter (EAKF) of Anderson [2].

1.2 Goals

The goals of this dissertation are: to review the principal formulations of the EnKF; to select one or more formulations for implementation and to implement them; to perform experiments with the implemented filters using a simple mechanical system (see below) as a test case and searching for interesting phenomena; and to interpret the experimental results and draw any important conclusions.

The mechanical system to be used as a test case in the experiments is the two-dimensional swinging spring. This simple system is of interest to meteorologists because it possesses interacting motions with two distinct timescales, analogous to the Rossby and gravity waves of the atmosphere. It may be used as an illustration of the problem of initialisation (see Chapter 4).

1.3 Principal Results

The ETKF and EAKF are selected for implementation in Chapter 3. It is found to be advantageous to reformulate the raw algorithms reviewed in Chapter 2 to give algorithms that are analytically equivalent but numerically better behaved.

Experiments with the ETKF and the swinging spring in Chapter 5 show a collapse in the number of distinct ensemble members after assimilating each observation. This collapse is explained in Section 6.2 and points to a limited usefulness of the ETKF for low-dimensional systems such as the swinging spring, although the high-dimensional systems typical of NWP are unaffected.

The most important result of the dissertation is that there is a potential flaw in the general framework for semi-deterministic formulations of the EnKF as presented in Tippett *et al* [23]. This flaw may lead some formulations of the EnKF to produce analysis ensembles with statistics that are inconsistent with the actual error, being both biased (mean tending to be in the wrong place) and overconfident (standard deviations of coordinates too small). This is demonstrated experimentally in Chapter 5 and explained theoretically in Section 6.1. The ETKF is affected by this flaw (at least in some circumstances), but the EAKF is not.

1.4 Outline

Several alternative formulations of the EnKF have been published since the original paper of Evensen [7]. The principal alternatives are reviewed in Chapter 2. Formulations may be classified as stochastic or semi-deterministic depending on the degree to which they rely upon pseudo-random numbers to represent uncertainty in the system model and the observations.

The ETKF and EAKF are selected for implementation in Chapter 3. This chapter also describes the reformulation of the raw algorithms of Chapter 2 to give algorithms that are analytically equivalent but numerically better behaved.

In Chapter 4, as a subject for experiments, the two-dimensional swinging spring is introduced. This simple mechanical system is of interest to meteorologists as an illustration of the problem of initialisation. The chapter discusses the concept of initialisation and its importance for NWP.

Chapter 5 presents the results of experiments using the filter implementations described in Chapter 3 and the swinging spring system of Chapter 4. The experiments reveal some unexpected features in the ETKF, including ensemble statistics that are inconsistent with the actual error.

Chapter 6 provides explanations of the features observed in Chapter 5. The explanation of the inconsistent statistics points to a potential flaw in the general framework for semi-deterministic formulations of the EnKF, affecting some but not all such formulations. The ETKF is affected (at least in some circumstances), but the EAKF is not. This flaw appears to have been overlooked in the literature.

The dissertation concludes in Chapter 7 with a review of the preceding chapters and some suggestions for further work.

Chapter 2

Formulations of the Ensemble Kalman Filter

This chapter presents a unified exposition of various formulations of the Ensemble Kalman Filter. It encompasses the stochastic formulation reviewed in Evensen [8] and the semi-deterministic formulations reviewed in Tippett *et al* [23]. We start by looking at the problem in data assimilation that the filter is designed to solve and the standard data assimilation techniques of the Kalman and Extended Kalman Filters.

2.1 Sequential Data Assimilation

Data assimilation seeks to solve the following problem: given a noisy discrete model of the dynamics of a system and noisy observations of the system, find estimates of the state of the system. Sequential data assimilation techniques such as the Kalman Filter and its variants break this problem into a cycle of alternating forecast and analysis steps. In the forecast step the system dynamical model is used to evolve an earlier state estimate forward in time, giving a forecast state at the time of the latest observations. In the analysis step the observations are used to update the forecast state, giving an improved state estimate called the analysis. This analysis is used as the starting point for the next forecast.

We shall denote the dimension of the state space of the system by n and vectors in this space by $\mathbf{x}$, usually with various arguments, subscripts, and superscripts. In particular the true state of the system at time t_k will be denoted by $\mathbf{x}^t(t_k)$ and the forecast and analysis at this time by $\mathbf{x}^f(t_k)$ and $\mathbf{x}^a(t_k)$ respectively. We shall denote the dimension of observation space by m and the observation vector at time t_k by $\mathbf{y}(t_k)$. The argument t_k will be dropped from state and observation vectors when all quantities being discussed occur at the same time.

We shall be using random variables to model errors in the system dynamical model and in the observations. These errors lead to random errors in the forecasts and analyses. We seek forecasts and analyses that are unbiased in the sense that

$$\begin{aligned}\langle \mathbf{x}^f - \mathbf{x}^t \rangle &= 0 \\ \langle \mathbf{x}^a - \mathbf{x}^t \rangle &= 0\end{aligned}$$

where angle brackets denote the expectation operator. Information about the size and correlation of the error components is given by the error covariance matrices

$$\begin{aligned}\mathbf{P}^f &= \langle (\mathbf{x}^f - \mathbf{x}^t)(\mathbf{x}^f - \mathbf{x}^t)^T \rangle \\ \mathbf{P}^a &= \langle (\mathbf{x}^a - \mathbf{x}^t)(\mathbf{x}^a - \mathbf{x}^t)^T \rangle.\end{aligned}$$

We seek methods that calculate covariance matrices $\mathbf{P}^f$ and $\mathbf{P}^a$ as well as state estimates $\mathbf{x}^f$ and $\mathbf{x}^a$.

2.2 The Kalman and Extended Kalman Filters

We now briefly review two established data assimilation techniques. For a detailed treatment see Gelb [9] or the mathematically more sophisticated Jazwinski [14]. The Kalman and Extended Kalman Filters originally rose to prominence in aerospace applications and are popular for uses such as tracking airborne objects with radar. For these applications the state space dimension is relatively small, typically nine or less. As we shall see, there are problems extending the techniques to the much larger state space dimensions frequently encountered in geoscience applications.

2.2.1 The Kalman Filter

The Kalman Filter (KF) is a sequential data assimilation technique for use with linear models of the system dynamics and observations. It assumes that the system dynamics can be modelled by

$$\mathbf{x}^t(t_k) = \mathbf{M}\mathbf{x}^t(t_{k-1}) + \boldsymbol{\eta}(t_{k-1}) \tag{2.1}$$

where $\mathbf{M}$ is a known matrix and $\boldsymbol{\eta}(t_{k-1})$ is random model noise with known covariance matrix $\mathbf{Q}$. It is assumed that $\boldsymbol{\eta}(t_{k-1})$ is unbiased and that its values at different times are uncorrelated. The observations are assumed to be modelled by

$$\mathbf{y}(t_k) = \mathbf{H}\mathbf{x}^t(t_k) + \boldsymbol{\varepsilon}(t_k) \tag{2.2}$$

where $\mathbf{H}$ is a known matrix and $\boldsymbol{\varepsilon}(t_k)$ is random observation noise with known covariance matrix $\mathbf{R}$. It is assumed that $\boldsymbol{\varepsilon}(t_k)$ is unbiased and that its values at different times are uncorrelated. It is also assumed that there is no correlation between model noise and observation noise at any times (the same or different). It is not difficult to extend these models to the case where $\mathbf{M}$, $\mathbf{Q}$, $\mathbf{H}$, and $\mathbf{R}$ vary with time, but for notational simplicity this case is not explicitly considered here.

In addition to the models of the system dynamics and the observations, an initial state estimate $\mathbf{x}^a(t_0)$ is required and so is its error covariance matrix $\mathbf{P}^a(t_0)$. It is assumed that the error in this estimate is unbiased and uncorrelated with model noise and observation noise at any time. If observations are available at time t_0, the initial estimate may be treated as a forecast rather than an analysis.

The forecast step of the Kalman Filter evolves the analysis and analysis error covariance matrix at time t_{k-1} forward to time t_k using the equations

$$\begin{aligned}
\mathbf{x}^f(t_k) &= \mathbf{M}\mathbf{x}^a(t_{k-1}) && (2.3)\\
\mathbf{P}^f(t_k) &= \mathbf{M}\mathbf{P}^a(t_{k-1})\mathbf{M}^T + \mathbf{Q}. && (2.4)
\end{aligned}$$

The analysis step at time t_k starts by calculating the Kalman gain matrix

$$\mathbf{K}(t_k) = \mathbf{P}^f(t_k)\mathbf{H}^T(\mathbf{H}\mathbf{P}^f(t_k)\mathbf{H}^T + \mathbf{R})^{-1}. \tag{2.5}$$

The observation $\mathbf{y}(t_k)$ is then assimilated using

$$\begin{aligned}
\mathbf{x}^a(t_k) &= \mathbf{x}^f(t_k) + \mathbf{K}(t_k)(\mathbf{y}(t_k) - \mathbf{H}\mathbf{x}^f(t_k)) && (2.6)\\
\mathbf{P}^a(t_k) &= (\mathbf{I} - \mathbf{K}(t_k)\mathbf{H})\mathbf{P}^f(t_k). && (2.7)
\end{aligned}$$

Good points about the Kalman Filter include that the state update equations (2.3) and (2.6) preserve unbiasedness; that the error covariance update equations (2.4) and (2.7) are exact; and that the filter is optimal in the sense that the analysis defined by (2.6) minimises the cost function

$$\mathcal{J}(\mathbf{x}) = (\mathbf{x} - \mathbf{x}^f)^T(\mathbf{P}^f)^{-1}(\mathbf{x} - \mathbf{x}^f) + (\mathbf{y} - \mathbf{H}\mathbf{x})^T\mathbf{R}^{-1}(\mathbf{y} - \mathbf{H}\mathbf{x}). \tag{2.8}$$

This function is a weighted measure of the distances from the state $\mathbf{x}$ to the forecast $\mathbf{x}^f$ and the observation $\mathbf{y}$. Thus the analysis represents a combination of the twin information sources of forecast and observation, with greater weight being given to the more certain components.

Bad points about the Kalman Filter are that it only works for linear systems and that in applications such as numerical weather prediction (NWP) the covariance matrices are huge. A national meteorological service may use a forecasting model with 10^7 state variables and currently have to assimilate 10^5–10^6 observations per assimilation period. This gives state error covariance matrices of size $10^7 \times 10^7$ and potentially requiring hundreds of terabytes of storage. Also, the calculation of the Kalman gain (2.5) involves the inversion of a matrix of size $10^5 \times 10^5$ or larger. In the near future with more satellite data being assimilated the number of observations is expected to become comparable with the number of state variables, which will make matters worse.

2.2.2 The Extended Kalman Filter

The Extended Kalman Filter (EKF) is an attempt to extend the Kalman Filter to nonlinear dynamical systems and nonlinear observations. The linear models of dynamics (2.1) and observations (2.2) are replaced by the nonlinear variants

$$\begin{aligned}\mathbf{x}^t(t_k) &= M(\mathbf{x}^t(t_{k-1})) + \boldsymbol{\eta}(t_{k-1}) &\quad (2.9)\\ \mathbf{y}(t_k) &= H(\mathbf{x}^t(t_k)) + \boldsymbol{\varepsilon}(t_k) &\quad (2.10)\end{aligned}$$

where the matrices $\mathbf{M}$ and $\mathbf{H}$ have been replaced by potentially nonlinear functions M and H. Note the convention of using bold upright type for linear operators and standard italic type for corresponding nonlinear operators. Again, it is not difficult to extend these models to the case where M and H vary with time.

The nonlinear functions are used in the state update equations of the forecast and analysis steps:

$$\begin{aligned}\mathbf{x}^f(t_k) &= M(\mathbf{x}^a(t_{k-1}))\\ \mathbf{x}^a(t_k) &= \mathbf{x}^f(t_k) + \mathbf{K}(t_k)(\mathbf{y}(t_k) - H(\mathbf{x}^f_k)).\end{aligned}$$

The error covariance update equations (2.4) and (2.7) and the calculation of the Kalman gain (2.5) remain the same with the proviso that $\mathbf{M}$ and $\mathbf{H}$ are now the tangent linear operators (Jacobians) of M and H evaluated at $\mathbf{x}^a(t_{k-1})$ and $\mathbf{x}^f(t_k)$ respectively.

The EKF can work well, especially if the system is only weakly nonlinear. However, it relies on linear approximations to nonlinear functions and does not strictly possess several of the good features of the Kalman Filter. In particular the state update equations can no longer be guaranteed to preserve unbiasedness; the error covariance update equations are no longer exact; and the filter is no longer optimal in the sense of minimising the cost function

(2.8) (with $\mathbf{Hx}$ replaced by $H(\mathbf{x})$). The EKF also does nothing to address the problem of huge covariance matrices. In addition it is labour-intensive to implement because of the need to derive and implement tangent linear models of the dynamics and observations.

2.3 The Ensemble Kalman Filter

The EKF represents nonlinearity using derivatives that only take into account behaviour in an infinitesimal neighbourhood of a point. The Ensemble Kalman Filter (EnKF) is an attempt to represent nonlinearity by using something more spread out. The details of this approach will be discussed in the following sections, but the key ideas are to use an ensemble (statistical sample) of state estimates instead of a single state estimate; to calculate the error covariance matrix from this ensemble instead of maintaining a separate covariance matrix; and to use this calculated covariance matrix to calculate a common Kalman gain that is used to update each ensemble member in the analysis step. The hope is that the use of an ensemble will provide a better representation of nonlinearity than is achieved by the EKF.

At first sight the need to maintain a whole ensemble of state estimates makes the EnKF look computationally much more expensive that the EKF. However, the EnKF may be the cheaper of the two filters. One saving comes from the absence of a separate covariance matrix to be evolved and updated. In the case of high-dimensional systems another saving comes if we use ensemble sizes that are small compared to the number of observations (as long as they are not so small that they are statistically unrepresentative). Such ensembles lead to covariance matrices with reduced rank, and this can be exploited in the analysis step to lower the computational cost. The use of a common Kalman gain for all ensemble members also reduces the overhead of the additional ensemble members.

Another benefit of the EnKF is that it does not require tangent linear models to be implemented. This makes it especially attractive to individual researchers or small groups who wish to use data assimilation to solve problems in meteorology or oceanography without having the manpower resources of a large organisation at their disposal.

The original presentation of the EnKF was in Evensen [7]. This section follows in essence the review article Evensen [8] which incorporates later advances including the important recognition by Burgers *et al* [4] (and independently by Houtekamer and Mitchell [11]) of the need to use an ensemble of pseudo-random observation perturbations in the analysis step.

2.3.1 Notation

We shall use the subscript i to denote the individual members of an ensemble and N to denote the size of an ensemble. Thus the members of an ensemble in state space will be denoted by $\mathbf{x}_i$ $(i = 1, \ldots, N)$. A superscript f or a will frequently be added to denote a forecast or analysis ensemble.

In situations where we must give a single best state estimate from an ensemble, we shall use the ensemble mean

$$\overline{\mathbf{x}} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{x}_i.$$

When we require the error covariance matrix we shall use the ensemble covariance matrix

$$\mathbf{P}_e = \frac{1}{N-1} \sum_{i=1}^{N} (\mathbf{x}_i - \overline{\mathbf{x}})(\mathbf{x}_i - \overline{\mathbf{x}})^T.$$

Note the division by $N - 1$ rather than N. This ensures than $\mathbf{P}_e$ is an unbiased estimate of the population covariance matrix $\mathbf{P}$.

It is convenient to introduce the ensemble matrix

$$\mathbf{X} = \frac{1}{\sqrt{N-1}} \begin{pmatrix} \mathbf{x}_1 & \mathbf{x}_2 & \ldots & \mathbf{x}_N \end{pmatrix} \tag{2.11}$$

and the ensemble perturbation matrix

$$\mathbf{X}' = \frac{1}{\sqrt{N-1}} \begin{pmatrix} \mathbf{x}_1 - \overline{\mathbf{x}} & \mathbf{x}_2 - \overline{\mathbf{x}} & \ldots & \mathbf{x}_N - \overline{\mathbf{x}} \end{pmatrix}. \tag{2.12}$$

The ensemble covariance matrix may then be expressed as

$$\mathbf{P}_e = \mathbf{X}'\mathbf{X}'^T. \tag{2.13}$$

2.3.2 The Forecast Step

At its simplest the EnKF assumes the same underlying nonlinear stochastic system model (2.9) as the EKF. The forecast step evolves each ensemble member forward in time using this model:

$$\mathbf{x}_i^f(t_k) = M(\mathbf{x}_i^a(t_{k-1})) + \boldsymbol{\eta}_i(t_{k-1})$$

where $\boldsymbol{\eta}_i(t_{k-1})$ is pseudo-random noise to be drawn from a distribution with mean zero and covariance $\mathbf{Q}$. There is no covariance matrix to evolve and this offsets the extra work needed to evolve an ensemble of states. Note that the evolution of each ensemble member is independent of all other ensemble members, thus making the forecast step well-suited for implementation on a parallel computer.

In the EnKF there is no need to make the assumptions about lack of autocorrelation of model noise and lack of correlation between model noise and the initial analysis error that were made in Section 2.2.1. These assumptions are required in the KF and EKF so that $\boldsymbol{\eta}(t_{k-1})$ is uncorrelated with the analysis error $\mathbf{x}^a(t_{k-1}) - \mathbf{x}^t(t_{k-1})$, thus justifying the error covariance update equation (2.4). In the EnKF this equation is no longer used and the assumptions are not necessary. The simulation of time-correlated model noise is discussed in Evensen [8, Section 4.2.1].

The EnKF is not limited to the simple dynamical model (2.9). Any stochastic difference or differential equation capable of numerical integration may be used. This issue is discussed in Evensen [8, Section 3.4.2].

An initial ensemble is required at time t_0. Assuming we have an initial best-guess estimate and some idea of the error in this estimate expressed through a covariance matrix, we may generate an initial ensemble by taking the best-guess estimate and adding random perturbations from a distribution determined by the covariance matrix. Because the initial estimate and covariance matrix may be crude, Evensen [8, Section 4.1] recommends integrating the initial ensemble over a time interval containing a few characteristic time scales of the dynamical system to ensure that the system is in dynamical balance and that proper multivariate correlations have developed. This is fine as long as we are not required to assimilate any observations during this interval. Other, perhaps more sophisticated, methods of ensemble generation have been devised in the context of ensemble prediction systems; see, for example, Kalnay [15, Section 6.5].

2.3.3 The Analysis Step

There are various formulations of the analysis step of the EnKF. In this section we consider the stochastic formulation of Evensen [8]. Alternative, deterministic formulations are the subject of Section 2.4. We start by considering a linear observation operator, postponing the case of nonlinear observations to Section 2.3.4. In the linear case the KF update equations (2.5) to (2.7) are exact and optimal. Therefore we wish to mimic these in the ensemble version of the analysis step.

We may define an ensemble version of the Kalman gain by

$$\mathbf{K}_e = \mathbf{P}_e^f \mathbf{H}^T (\mathbf{H}\mathbf{P}_e^f \mathbf{H}^T + \mathbf{R})^{-1}.$$

We could try updating each ensemble member using the KF equation with this gain:

$$\mathbf{x}_i^a = \mathbf{x}_i^f + \mathbf{K}_e(\mathbf{y} - \mathbf{H}\mathbf{x}_i^f).$$

This yields an update of the ensemble mean update that is like the KF:

$$\overline{\mathbf{x}^a} = \overline{\mathbf{x}^f} + \mathbf{K}_e(\mathbf{y} - \mathbf{H}\overline{\mathbf{x}^f})$$

but the ensemble perturbations satisfy

$$\mathbf{X}'^a = (\mathbf{I} - \mathbf{K}_e\mathbf{H})\mathbf{X}'^f$$

which implies that the ensemble covariance updates as

$$\mathbf{P}_e^a = (\mathbf{I} - \mathbf{K}_e\mathbf{H})\mathbf{P}_e^f(\mathbf{I} - \mathbf{K}_e\mathbf{H})^T.$$

Compared to the KF covariance update (2.7) this is too small by a factor of $(\mathbf{I} - \mathbf{K}_e\mathbf{H})^T$.

To obtain the desired statistics from the analysis ensemble we define an observation ensemble

$$\mathbf{y}_i = \mathbf{y} + \boldsymbol{\varepsilon}_i \tag{2.14}$$

where $\boldsymbol{\varepsilon}_i$ is pseudo-random noise drawn from a population with mean zero and covariance $\mathbf{R}$. We may define an observation ensemble matrix $\mathbf{Y}$, an observation ensemble perturbation matrix $\mathbf{Y}'$, and an observation ensemble covariance matrix $\mathbf{R}_e$ directly analogous to the matrices $\mathbf{X}$, $\mathbf{X}'$, and $\mathbf{P}_e$ associated with a state ensemble by (2.11) to (2.13). We define the ensemble Kalman gain with $\mathbf{R}_e$ in place of $\mathbf{R}$:

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

and update each ensemble member using this gain and the members of the observation ensemble:

$$\mathbf{x}_i^a = \mathbf{x}_i^f + \mathbf{K}_e(\mathbf{y}_i - \mathbf{H}\mathbf{x}_i^f). \tag{2.16}$$

The ensemble mean updates as

$$\overline{\mathbf{x}^a} = \overline{\mathbf{x}^f} + \mathbf{K}_e(\overline{\mathbf{y}} - \mathbf{H}\overline{\mathbf{x}^f})$$

which is like the KF but with the mean of the synthetic observation ensemble $\overline{\mathbf{y}}$ in place of the actual observation $\mathbf{y}$. We can either accept this result as it is, noting that $\overline{\mathbf{y}}$ tends to $\mathbf{y}$ as the ensemble size increases, or following Evensen [8] we may impose the constraint $\overline{\boldsymbol{\varepsilon}} = 0$ on our random vectors to ensure that $\overline{\mathbf{y}} = \mathbf{y}$ unconditionally. The ensemble perturbation matrix updates as

$$\mathbf{X}'^a = (\mathbf{I} - \mathbf{K}_e\mathbf{H})\mathbf{X}'^f + \mathbf{K}_e\mathbf{Y}'$$

which implies

$$\mathbf{P}_e^a = (\mathbf{I}-\mathbf{K}_e\mathbf{H})\mathbf{P}_e^f + (\mathbf{I}-\mathbf{K}_e\mathbf{H})\mathbf{X}'^f\mathbf{Y}'^T\mathbf{K}_e^T + \mathbf{K}_e\mathbf{Y}'(\mathbf{X}'^f)^T(\mathbf{I}-\mathbf{K}_e\mathbf{H})^T.$$

The first term on the right is the desired expression for $\mathbf{P}_e^f$. As long as our random vectors $\boldsymbol{\varepsilon}_i$ are independent of the forecast ensemble perturbations $\mathbf{x}_i^f - \overline{\mathbf{x}^f}$, the product $\mathbf{X}'^f\mathbf{Y}'^T$ and hence the second and third terms on the right hand side will tend to zero as the ensemble size increases. Once again we may either accept this result as it is or follow Evensen [8] and impose the additional constraint $\overline{(\mathbf{x}^f - \overline{\mathbf{x}^f})\boldsymbol{\varepsilon}^T} = 0$ on our random vectors $\boldsymbol{\varepsilon}_i$ to ensure the desired covariance update unconditionally.

As presented so far the algorithm for the analysis step consists of generating an observation ensemble using (2.14), calculating the ensemble Kalman gain using (2.15), and updating the ensemble members using (2.16). This does little to address the problems with huge covariance matrices that were mentioned at the end of Section 2.2.1. The calculation of $\mathbf{K}_e$ involves the calculation and storage of the $n \times n$ covariance matrix $\mathbf{P}_e^f$ and the inversion of the $m \times m$ matrix $\mathbf{H}\mathbf{P}_e^f\mathbf{H}^T + \mathbf{R}_e$. We now consider how the calculation may be arranged to avoid these problems.

We start by rewriting (2.15) in terms of ensemble matrices:

$$\mathbf{K}_e = \mathbf{X}'^f(\mathbf{X}'^f)^T\mathbf{H}^T(\mathbf{H}\mathbf{X}'^f(\mathbf{X}'^f)^T\mathbf{H}^T + \mathbf{Y}'\mathbf{Y}'^T)^{-1}.$$

The structure of the right hand side is revealed more clearly if we introduce an ensemble of forecast observations

$$\mathbf{y}_i^f = \mathbf{H}\mathbf{x}_i^f.$$

Thus $\mathbf{y}_i^f$ is what the observation would be if the true state of the system was $\mathbf{x}_i^f$ and there was no observation noise. We define matrices $\mathbf{Y}^f$ and $\mathbf{Y}'^f$ for this ensemble in the same way as $\mathbf{Y}$ and $\mathbf{Y}'$ are defined for the ensemble $\mathbf{y}_i$. In terms of these matrices

$$\mathbf{K}_e = \mathbf{X}'^f(\mathbf{Y}'^f)^T(\mathbf{Y}'^f(\mathbf{Y}'^f)^T + \mathbf{Y}'\mathbf{Y}'^T)^{-1}. \tag{2.17}$$

We now note that if we constrain the random vectors $\boldsymbol{\varepsilon}_i$ in the way discussed above so that $\mathbf{X}'^f\mathbf{Y}'^T = 0$, then it follows (on multiplying by $\mathbf{H}$) that $\mathbf{Y}'^f\mathbf{Y}'^T = 0$ and hence that the matrix we must invert can be written as

$$\mathbf{Y}'^f(\mathbf{Y}'^f)^T + \mathbf{Y}'\mathbf{Y}'^T = (\mathbf{Y}'^f + \mathbf{Y}')(\mathbf{Y}'^f + \mathbf{Y}')^T. \tag{2.18}$$

We can make this substitution in the formula for the $\mathbf{K}_e$ even if we are not constraining the $\boldsymbol{\varepsilon}_i$, justifying it on the grounds that as long as the $\boldsymbol{\varepsilon}_i$ are independent of the forecast observation perturbations $\mathbf{y}_i^f - \overline{\mathbf{y}^f}$, the product $\mathbf{Y}'^f\mathbf{Y}'^T$ tends to zero as the ensemble size increases and hence the new formula for $\mathbf{K}_e$ is as good an approximation to the true Kalman gain $\mathbf{K}$ as the old one. Now we take the singular value decomposition (SVD) of the $m \times N$ matrix that is multiplied by its transpose on the right hand side of (2.18):

$$\mathbf{Y}'^f + \mathbf{Y}' = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^T$$

where $\boldsymbol{\Sigma}$ is the $p \times p$ diagonal matrix of nonzero singular values (p being the rank of the matrix) and $\mathbf{U}$ and $\mathbf{V}$ are column-orthogonal matrices of sizes $m \times p$ and $N \times p$ respectively. From this decomposition we can find the eigenvalue decomposition

$$\mathbf{Y}'^f(\mathbf{Y}'^f)^T + \mathbf{Y}'\mathbf{Y}'^T = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^T$$

were $\boldsymbol{\Lambda} = \boldsymbol{\Sigma}\boldsymbol{\Sigma}^T$ is the diagonal matrix of nonzero eigenvalues, and the inverse follows as

$$(\mathbf{Y}'^f(\mathbf{Y}'^f)^T + \mathbf{Y}'\mathbf{Y}'^T)^{-1} = \mathbf{U}\boldsymbol{\Lambda}^{-1}\mathbf{U}^T.$$

If $p < m$ then the matrix to be inverted is singular and this inverse is a pseudo-inverse. If the number of observations m is large as in the NWP example and the ensemble size N is small compared to m, we have reduced an expensive inversion of an $m \times m$ matrix requiring $O(m^3)$ floating point operations to a cheaper SVD of an $m \times N$ matrix requiring $O(m^2N + N^3) = O(m^2N)$ operations (see Golub and Van Loan [10, Section 5.4.5])[1].

[1] The $O(mN)$ used in Evensen [8, Section 4.3.2] appears to be an error.

With the inverse in the calculation of $\mathbf{K}_e$ taken care of, we consider how the analysis step may be completed without having to compute and store excessively large matrices. We confine ourselves to systems sized as in the NWP example in which $p \leq N \ll m \leq n$. We assume that the ensemble and ensemble perturbation matrices have been computed and stored for the forecast, observation, and forecast observation ensembles. None of these matrices is larger than $n \times N$ and the majority are $m \times N$. We also assume that the above SVD has been computed and $\mathbf{U}$ and $\mathbf{\Lambda}^{-1}$ stored. These matrices are smaller still at $m \times p$ and $p \times p$ respectively. The ensemble update equation (2.16) can be written in matrix form as

$$\begin{aligned} \mathbf{X}^a &= \mathbf{X}^f + \mathbf{K}_e(\mathbf{Y} - \mathbf{Y}^f) \qquad (2.19) \\ &= \mathbf{X}^f + \mathbf{X}'^f(\mathbf{Y}'^f)^T\mathbf{U}\mathbf{\Lambda}^{-1}\mathbf{U}^T(\mathbf{Y} - \mathbf{Y}^f). \end{aligned}$$

The computations are best performed in the order set out in the following table, which also shows the size of the resulting matrix and the number of floating point operations required to compute it.

Computation	Size	Operations
$\mathbf{Z}_1 = \mathbf{\Lambda}^{-1}\mathbf{U}^T$	$p \times m$	mp
$\mathbf{Z}_2 = \mathbf{Z}_1(\mathbf{Y} - \mathbf{Y}^f)$	$p \times N$	mNp
$\mathbf{Z}_3 = \mathbf{U}\mathbf{Z}_2$	$m \times N$	mNp
$\mathbf{Z}_4 = (\mathbf{Y}'^f)^T\mathbf{Z}_3$	$N \times N$	mN^2
$\mathbf{X}^a = \mathbf{X}^f + \mathbf{X}'^f\mathbf{Z}_4$	$n \times N$	nN^2

The largest matrix is the final analysis ensemble matrix $\mathbf{X}^a$ which has size $n \times N$. The most expensive computation is also the final one with $O(nN^2)$ operations.

Assuming that multiplication by $\mathbf{H}$ and generating the random vectors $\boldsymbol{\varepsilon}_i$ are relatively cheap, the total cost of an efficient implementation of the analysis step is therefore $O(m^2N + nN^2)$ and the largest matrix that has to be stored is of size $n \times N$. This compares to $O(m^2n + n^2N)$ and $n \times n$ for a naive implementation, and $O(n^3)$ and $n \times n$ for the analysis step of the KF (see Appendix A). However, it should be added that floating point operation counts are not the whole story, especially with modern computer architectures. Memory access may be the main bottleneck, which is why the size of matrices stored is important. On parallel machines the minimisation of communication between processors will be the dominant consideration.

2.3.4 Nonlinear Observation Operators

Evensen [8, Section 4.5] presents the following technique for extending the preceding formulation of the analysis step to nonlinear observation operators of the type in (2.10). We augment the state vector with a diagnostic variable that is the predicted observation vector:

$$\hat{\mathbf{x}} = \begin{pmatrix} \mathbf{x} \\ H(\mathbf{x}) \end{pmatrix}$$

and define a linear observation operator on augmented state space by

$$\widehat{\mathbf{H}} \begin{pmatrix} \mathbf{x} \\ \mathbf{y} \end{pmatrix} = \mathbf{y}.$$

We then carry out the analysis step in augmented state space using $\hat{\mathbf{x}}$ and $\widehat{\mathbf{H}}$ in place of $\mathbf{x}$ and $\mathbf{H}$. Superficially, this technique appears to reduce the nonlinear problem to the previously-solved linear one. However, the linear problem created is not quite the same as the problem solved in Section 2.3.3, because the valid states of our system only occupy a submanifold of augmented state space instead of the whole space. Thus, whilst this is a reasonable way of formulating the EnKF for nonlinear observation operators, it is not as well-founded as the linear case, which can be justified as an approximation to the exact and optimal KF.

We now translate the augmented state space formulation of the analysis step to a formulation that does not explicitly refer to that space. First note that the observation ensemble $\mathbf{y}_i$ is independent of state space, so it and the matrices $\mathbf{Y}$ and $\mathbf{Y}'$ are the same as before. The forecast observation ensemble is

$$\mathbf{y}_i^f = \widehat{\mathbf{H}}\hat{\mathbf{x}}_i^f = H(\mathbf{x}_i^f).$$

The mean of this ensemble is

$$\overline{\mathbf{y}^f} = \overline{H(\mathbf{x}^f)}$$

and thus the forecast observation ensemble and ensemble perturbation matrices are

$$\begin{aligned} \mathbf{Y}^f &= \frac{1}{\sqrt{N-1}}\left(\begin{array}{ccc} H(\mathbf{x}_1^f) & \dots & H(\mathbf{x}_N^f) \end{array} \right) && (2.20) \\ \mathbf{Y}'^f &= \frac{1}{\sqrt{N-1}}\left(\begin{array}{ccc} H(\mathbf{x}_1^f) - \overline{H(\mathbf{x}^f)} & \dots & H(\mathbf{x}_N^f) - \overline{H(\mathbf{x}^f)} \end{array} \right). && (2.21) \end{aligned}$$

We can now use (2.17) and (2.19) to write the ensemble update in augmented state space as

$$\begin{aligned} \widehat{\mathbf{K}}_e &= \widehat{\mathbf{X}}'^f(\mathbf{Y}'^f)^T(\mathbf{Y}'^f(\mathbf{Y}'^f)^T + \mathbf{Y}'\mathbf{Y}'^T)^{-1} \\ \widehat{\mathbf{X}}^a &= \widehat{\mathbf{X}}^f + \widehat{\mathbf{K}}_e(\mathbf{Y} - \mathbf{Y}^f). \end{aligned}$$

Taking the first n rows of these equations we obtain

$$\begin{aligned} \mathbf{K}_e &= \mathbf{X}'^f(\mathbf{Y}'^f)^T(\mathbf{Y}'^f(\mathbf{Y}'^f)^T + \mathbf{Y}'\mathbf{Y}'^T)^{-1} && (2.22) \\ \mathbf{X}^a &= \mathbf{X}^f + \mathbf{K}_e(\mathbf{Y} - \mathbf{Y}^f) && (2.23) \end{aligned}$$

which is the same ensemble update as in the linear case except that $\mathbf{Y}^f$ and $\mathbf{Y}'^f$ are given in terms of the nonlinear observation operator H by (2.20) and (2.21).

2.4 Deterministic Formulations of the Analysis Step

The analysis step of the EnKF as presented in Section 2.3.3 uses a synthetic observation ensemble created by adding random vectors to the real observation. This stochastic element exposes the method to sampling errors. We now consider the alternative formulations reviewed in Tippett *et al* [23]. These are deterministic formulations that do not require a synthetic observation ensemble. However, it should be pointed out that there is a potential flaw in some of these methods that appears to have been overlooked in the literature. This is illustrated experimentally in Chapter 5 and demonstrated theoretically in Section 6.1. We shall ignore this flaw for now and give an exposition that follows in essence Tippett *et al* [23].

Unless otherwise stated we assume a nonlinear observation operator as in Section 2.3.4.

2.4.1 The General Framework

All the methods to be described in this section split the ensemble update into two parts: first the analysis ensemble mean is calculated, then the analysis ensemble perturbations. We start by examining what the method of Section 2.3.3 does to the ensemble mean. Taking the mean of both sides of (2.23) gives

$$\overline{\mathbf{x}^a} = \overline{\mathbf{x}^f} + \mathbf{K}_e(\overline{\mathbf{y}} - \overline{\mathbf{y}^f}).$$

Recall from Section 2.3.3 that we have $\overline{\mathbf{y}} = \mathbf{y}$, either exactly if we impose the right constraint on our random vectors or in the limit of large ensembles. Making this substitution gives us an update equation for the ensemble mean in which the observation ensemble has been averaged out[2]:

$$\begin{aligned} \overline{\mathbf{x}^a} &= \overline{\mathbf{x}^f} + \mathbf{K}_e(\mathbf{y} - \overline{\mathbf{y}^f}) \qquad (2.24) \\ &= \overline{\mathbf{x}^f} + \mathbf{K}_e(\mathbf{y} - \overline{H(\mathbf{x}^f)}). \end{aligned}$$

We shall use this equation in all our deterministic formulations of the analysis step. However, we cannot use $\mathbf{K}_e$ as defined by (2.22) because we no longer have an observation ensemble perturbation matrix $\mathbf{Y}'$. Instead we must

[2] Lorenc [18, Equation (4)] uses $H(\overline{\mathbf{x}^f})$ instead of $\overline{H(\mathbf{x}^f)}$. The version given above has the advantage that it arises naturally out of averaging the stochastic formulation as shown. It also takes more account of the nonlinearity of H. For linear observation operators $H(\overline{\mathbf{x}^f}) = \overline{H(\mathbf{x}^f)}$.

go back to using the population version of the observation error covariance matrix $\mathbf{R}$ instead of the ensemble version $\mathbf{R}_e = \mathbf{Y}'\mathbf{Y}'^T$. Thus we take

$$\mathbf{K}_e = \mathbf{X}'^f(\mathbf{Y}'^f)^T(\mathbf{Y}'^f(\mathbf{Y}'^f)^T + \mathbf{R})^{-1}.$$

The matrix inverted in this expression is the ensemble version of the innovations covariance matrix. It appears frequently in what follows, so we introduce a special notation for it:

$$\mathbf{S} = \mathbf{Y}'^f(\mathbf{Y}'^f)^T + \mathbf{R}. \tag{2.25}$$

We now consider the update of the ensemble perturbations. In the case of a linear observation operator we would like the ensemble covariance matrix to update like the KF covariance update (2.7). Thus in this case we require

$$\begin{aligned} \mathbf{X}'^a(\mathbf{X}'^a)^T &= \mathbf{P}_e^a \\ &= (\mathbf{I} - \mathbf{K}_e\mathbf{H})\mathbf{P}_e^f \\ &= (\mathbf{I} - \mathbf{X}'^f(\mathbf{Y}'^f)^T\mathbf{S}^{-1}\mathbf{H})\mathbf{X}'^f(\mathbf{X}'^f)^T \\ &= \mathbf{X}'^f(\mathbf{I} - (\mathbf{Y}'^f)^T\mathbf{S}^{-1}\mathbf{Y}'^f)(\mathbf{X}'^f)^T. \end{aligned}$$

The first and last terms in this chain of equations make no mention of the linear operator $\mathbf{H}$, so we impose their equality as a condition in the case of nonlinear observation operators as well. The equality will be satisfied if

$$\mathbf{X}'^a = \mathbf{X}'^f\mathbf{T} \tag{2.26}$$

where $\mathbf{T}$ is an $N \times N$ matrix that is a matrix square root of $\mathbf{I} - (\mathbf{Y}'^f)^T\mathbf{S}^{-1}\mathbf{Y}'^f$ in the sense that

$$\mathbf{T}\mathbf{T}^T = \mathbf{I} - (\mathbf{Y}'^f)^T\mathbf{S}^{-1}\mathbf{Y}'^f. \tag{2.27}$$

Note that $\mathbf{T}$ satisfying (2.27) is not unique and may be replaced by $\mathbf{TU}$ where $\mathbf{U}$ is an arbitrary $N \times N$ orthogonal matrix. The methods that we shall now consider differ in their choice of $\mathbf{T}$.

2.4.2 The Direct Method

Tippett *et al* [23] introduce a so-called direct method which, as its name suggests, takes a direct approach to finding a $\mathbf{T}$ satisfying (2.27). The first step is to solve the linear system

$$\mathbf{S}\mathbf{Z} = \mathbf{Y}'^f \tag{2.28}$$

for the $m \times N$ matrix $\mathbf{Z}$. In the case where $N \ll m$ (as in the NWP example) Tippett *et al* [23] suggest exploiting the identity

$$\mathbf{S}^{-1} = \mathbf{R}^{-1} - \mathbf{R}^{-1}\mathbf{Y}'^f(\mathbf{I} + (\mathbf{Y}'^f)^T\mathbf{R}^{-1}\mathbf{Y}'^f)^{-1}(\mathbf{Y}'^f)^T\mathbf{R}^{-1}$$

which may be verified by multiplying by $\mathbf{S}$ and using the definition (2.25) of $\mathbf{S}$. Computing $\mathbf{R}^{-1}$ for use in this identity is usually much easier than computing $\mathbf{S}^{-1}$ because $\mathbf{R}$ has a simple structure, typically diagonal. The other matrix that needs to be inverted is the $N \times N$ matrix $\mathbf{I}+(\mathbf{Y}'^f)^T\mathbf{R}^{-1}\mathbf{Y}'^f$. Thus the identity reduces the inversion of an $m \times m$ matrix to the inversion of an $N \times N$ one (although this is not the full story in the reduction of work needed to solve (2.28) because the equation can be solved without finding $\mathbf{S}^{-1}$ explicitly).

With $\mathbf{Z}$ found, the next step is to form

$$\mathbf{I} - (\mathbf{Y}'^f)^T\mathbf{S}^{-1}\mathbf{Y}'^f = \mathbf{I} - (\mathbf{Y}'^f)^T\mathbf{Z}$$

and find its matrix square root $\mathbf{T}$. Tippett *et al* [23] do not enjoin a particular method for finding the matrix square root. Since $\mathbf{I} - (\mathbf{Y}'^f)^T\mathbf{S}^{-1}\mathbf{Y}'^f$ is positive definite (as follows from (2.31) below) one possibility is to use a Cholesky factorisation algorithm such as those given in Golub and Van Loan [10, Section 4.2].

2.4.3 The Serial Method

If the observations represented by the individual components of the observation vector $\mathbf{y}$ have uncorrelated errors, they may be assimilated one at a time instead of all at once. This is a standard technique of Kalman filtering and has the advantage that it reduces the inversion of a large matrix to the inversion of a sequence of scalars. The procedure is justified because it is in effect a sequence of standard assimilation cycles with zero-length forecast steps. What is not obvious is that the result is the same as processing all observations at once. For a proof in the context of the standard KF see Dance [5, Appendix A].

The assumption of uncorrelated observation error components is the basis of the serial method of Tippett *et al* [23], which repeatedly applies an analysis scheme designed for an observation space of dimension one. In such an observation space $\mathbf{R}$ and $\mathbf{S}$ are scalars and $\mathbf{Y}'^f$ is a row vector. The method uses a closed form expression for $\mathbf{T}$ obtained by substituting

$$\mathbf{T} = \mathbf{I} - \beta(\mathbf{Y}'^f)^T\mathbf{Y}'^f \tag{2.29}$$

into (2.27) and solving for β, giving

$$\beta = \frac{1}{\mathbf{S} \pm \sqrt{\mathbf{RS}}}. \tag{2.30}$$

The serial method avoids expensive matrix inversions, but this has to be traded off against the need to apply the method multiple times at each analysis step. See Tippett *et al* [23] for a more detailed account of the issues.

Other instances of the use of serial processing with the EnKF include Houtekamer and Mitchell [13], where it is used with a stochastic formulation of the analysis step; Bishop *et al* [3], which discusses its use with the Ensemble Transform Kalman Filter to be described shortly; and Whitaker and Hamill [25], where it is used in a deterministic formulation of the analysis step that Tippett *et al* [23] show to be equivalent to (2.29) with the plus sign taken in (2.30).

2.4.4 The Ensemble Transform Kalman Filter

The Ensemble Transform Kalman Filter (ETKF) was originally introduced in Bishop *et al* [3], which describes its use to make rapid assessment of the future effect on error covariance of alternative strategies for deploying observational resources. The ETKF exploits the identity

$$\mathbf{I} - (\mathbf{Y}'^f)^T\mathbf{S}^{-1}\mathbf{Y}'^f = (\mathbf{I} + (\mathbf{Y}'^f)^T\mathbf{R}^{-1}\mathbf{Y}'^f)^{-1} \tag{2.31}$$

which may be verified by multiplying $\mathbf{I} - (\mathbf{Y}'^f)^T\mathbf{S}^{-1}\mathbf{Y}'^f$ by $\mathbf{I} + (\mathbf{Y}'^f)^T\mathbf{R}^{-1}\mathbf{Y}'^f$ and using the definition (2.25) of $\mathbf{S}$. The method starts by computing the $N \times N$ matrix $(\mathbf{Y}'^f)^T\mathbf{R}^{-1}\mathbf{Y}'^f$. This is usually much easier than computing $(\mathbf{Y}'^f)^T\mathbf{S}^{-1}\mathbf{Y}'^f$ because $\mathbf{R}$ usually has a simple structure. We then compute the eigenvalue decomposition

$$(\mathbf{Y}'^f)^T\mathbf{R}^{-1}\mathbf{Y}'^f = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T$$

where $\mathbf{U}$ is orthogonal and $\mathbf{\Lambda}$ is diagonal. It follows from (2.31) that

$$\mathbf{I} - (\mathbf{Y}'^f)^T\mathbf{S}^{-1}\mathbf{Y}'^f = \mathbf{U}(\mathbf{I} + \mathbf{\Lambda})^{-1}\mathbf{U}^T$$

and hence that a square root of the type we seek is

$$\mathbf{T} = \mathbf{U}(\mathbf{I} + \mathbf{\Lambda})^{-\frac{1}{2}}.$$

Note that $\mathbf{I} + \mathbf{\Lambda}$ is diagonal, so raising it to a power is easy. Using this $\mathbf{T}$ in (2.26) gives the ETKF.

2.4.5 The Ensemble Adjustment Kalman Filter

The Ensemble Adjustment Kalman Filter (EAKF) was originally introduced in Anderson [2]. The EAKF differs from the deterministic methods discussed so far in that it may be written in the form

$$\mathbf{X}'^a = \mathbf{A}\mathbf{X}'^f \tag{2.32}$$

although it may be written in the post-multiplier form (2.26) as well. We here assume that the observation operator is linear. The method may be extended to nonlinear observation operators by using the augmented state space of

Section 2.3.4. The first stage in finding $\mathbf{A}$ is to compute the eigenvalue decomposition

$$\mathbf{P}_e^f = \mathbf{F}\mathbf{G}^2\mathbf{F}^T \tag{2.33}$$

where $\mathbf{G}$ is the $p \times p$ diagonal matrix of positive square roots of nonzero eigenvalues and $\mathbf{F}$ is an $n \times p$ column-orthogonal matrix. We next perform the eigenvalue decomposition

$$(\mathbf{HFG})^T\mathbf{R}^{-1}\mathbf{HFG} = \widetilde{\mathbf{U}}\widetilde{\mathbf{\Lambda}}\widetilde{\mathbf{U}}^T$$

where $\widetilde{\mathbf{\Lambda}}$ is $p \times p$ diagonal and $\widetilde{\mathbf{U}}$ is $p \times p$ orthogonal. We then define

$$\mathbf{A} = \mathbf{FG}\widetilde{\mathbf{U}}(\mathbf{I} + \widetilde{\mathbf{\Lambda}})^{-\frac{1}{2}}\mathbf{G}^{-1}\mathbf{F}^T. \tag{2.34}$$

To see that the above gives the right analysis ensemble statistics note that we can find $\mathbf{F}$ and $\mathbf{G}$ in (2.33) from the SVD

$$\mathbf{X}'^f = \mathbf{FGW}^T \tag{2.35}$$

where $\mathbf{W}$ is an $N \times p$ column-orthogonal matrix. Substituting (2.34) and (2.35) into (2.32) gives

$$\mathbf{X}'^a = \mathbf{FG}\widetilde{\mathbf{U}}(\mathbf{I} + \widetilde{\mathbf{\Lambda}})^{-\frac{1}{2}}\mathbf{W}^T \tag{2.36}$$

and hence

$$\begin{aligned}
\mathbf{P}_e^a &= \mathbf{X}'^a(\mathbf{X}'^a)^T \\
&= \mathbf{FG}\widetilde{\mathbf{U}}(\mathbf{I} + \widetilde{\mathbf{\Lambda}})^{-1}\widetilde{\mathbf{U}}^T\mathbf{GF}^T \\
&= \mathbf{FG}(\mathbf{I} + (\mathbf{HFG})^T\mathbf{R}^{-1}\mathbf{HFG})^{-1}\mathbf{GF}^T.
\end{aligned}$$

Now $\mathbf{HFG}$ is a matrix with the property

$$\begin{aligned}
\mathbf{HFG}(\mathbf{HFG})^T &= \mathbf{HFG}^2\mathbf{F}^T\mathbf{H}^T \\
&= \mathbf{HP}_e^f\mathbf{H}^T \\
&= \mathbf{S} - \mathbf{R}
\end{aligned}$$

which is precisely the property of $\mathbf{Y}'^f$ that was used to establish (2.31). Therefore we may apply this identity to $\mathbf{HFG}$ to obtain

$$\begin{aligned}\mathbf{P}_e^a &= \mathbf{FG}(\mathbf{I} - (\mathbf{HFG})^T\mathbf{S}^{-1}\mathbf{HFG})\mathbf{GF}^T \\ &= \mathbf{P}_e^f - \mathbf{P}_e^f\mathbf{H}^T\mathbf{S}^{-1}\mathbf{HP}_e^f \\ &= (\mathbf{I} - \mathbf{K}_e\mathbf{H})\mathbf{P}_e^f\end{aligned}$$

which is the required ensemble covariance update.

To write the EAKF in the post-multiplier form (2.26) note that (2.35) and (2.36) imply

$$\mathbf{X}'^a = \mathbf{X}'^f\mathbf{W}\widetilde{\mathbf{U}}(\mathbf{I} + \widetilde{\mathbf{\Lambda}})^{-\frac{1}{2}}\mathbf{W}^T.$$

Note, however, that the EAKF does not quite fit into the framework of Section 2.4.1, which consists of not just the post-multiplier equation (2.26) but the square root condition (2.27) as well; see Appendix B for details.

Another deterministic formulation of the analysis step capable of being written in the pre-multiplier form (2.32) is given in Whitaker and Hamill [25].

2.5 Summary

Following the introductory material in Sections 2.1 and 2.2, this chapter has presented five formulations of the EnKF. All formulations differ from the KF and EKF in using an ensemble of state estimates instead of a single state estimate and not maintaining a separate error covariance matrix. They all offer advantages over the standard filters of reduced computational cost, better handling of nonlinearity, and greater ease of implementation through not using tangent linear models.

All formulations share the forecast step described in Section 2.3.2 and the technique for dealing with nonlinear observation operators described in Section 2.3.4. Where the formulations differ is in the analysis step. Section 2.3.3 described a stochastic formulation that makes use of an ensemble of pseudo-random observation perturbations. The other formulations are deterministic formulations that fit into a general framework described in Section 2.4.1.

This framework encompasses the direct method (Section 2.4.2, not as precisely defined as the other methods), the serial method (Section 2.4.3, limited to uncorrelated observations), the ETKF (Section 2.4.4), and the EAKF (Section 2.4.5). It was pointed out that there is a potential flaw in some of these methods; this is a major topic of Chapters 5 and 6.

The next chapter discusses the selection of an EnKF algorithm for implementation. It also describes the problems encountered in implementing the raw algorithms as presented in this chapter and how they may be reformulated to give algorithms that are analytically equivalent but numerically better behaved.

Chapter 3

Implementing an Ensemble Kalman Filter

This chapter is about the implementation of an EnKF. It describes some problems that were encountered, the solutions that were adopted, and some further improvements to the algorithms of Chapter 2. The EnKF is intended for experiments with the low-dimensional mechanical system described in Chapter 4, although it is capable of being used with other systems as well. Of the formulations of the EnKF in Chapter 2, it was initially decided to implement the ETKF. A deterministic formulation of the analysis step has the advantage over a stochastic formulation of eliminating one source of sampling error. Compared to the other deterministic formulations in Section 2.4, the ETKF has the advantage over the direct method of a more clearly defined algorithm, the advantage over the serial method of not requiring uncorrelated observations, and the advantage over the EAKF of avoiding one of the eigenvalue decompositions. However, there are some unexpected problems with the ETKF that are illustrated experimentally in Chapter 5 and explained theoretically in Chapter 6. Once these were discovered, the EAKF was implemented as well for comparison. The implementation of both filters is described in this chapter in order to keep similar material together.

3.1 Implementing the ETKF

The initial implementation of the ETKF closely followed the algorithm of Section 2.4.4. This created a problem with the eigenvalue decomposition

$$(\mathbf{Y}'^f)^T\mathbf{R}^{-1}\mathbf{Y}'^f = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^T \tag{3.1}$$

which produced eigenvalues and eigenvectors having significant imaginary parts. The reason for this is the lack of associativity in machine multiplication, leading to the ostensibly symmetric matrix $(\mathbf{Y}'^f)^T\mathbf{R}^{-1}\mathbf{Y}'^f$ becoming asymmetric when evaluated as $((\mathbf{Y}'^f)^T\mathbf{R}^{-1})\mathbf{Y}'^f$. This may be avoided by introducing the scaled forecast observation ensemble perturbation matrix

$$\widehat{\mathbf{Y}}^f = \mathbf{R}^{-\frac{1}{2}}\mathbf{Y}'^f \tag{3.2}$$

and writing

$$(\mathbf{Y}'^f)^T\mathbf{R}^{-1}\mathbf{Y}'^f = (\widehat{\mathbf{Y}}^f)^T\widehat{\mathbf{Y}}^f. \tag{3.3}$$

As long as machine multiplication is commutative this way of evaluating $(\mathbf{Y}'^f)^T\mathbf{R}^{-1}\mathbf{Y}'^f$ leads to a symmetric matrix with real eigenvalues and eigenvectors. Note that finding $\mathbf{R}^{-\frac{1}{2}}$ is easy in the common case of diagonal $\mathbf{R}$. Indeed, it is often $\mathbf{R}^{\frac{1}{2}}$ (the diagonal matrix of observation error standard deviations) that is the primary given quantity rather than $\mathbf{R}$, which makes evaluating $\mathbf{R}^{-\frac{1}{2}}$ easier still.

Regardless of the matter of symmetry, it is in any case advantageous to scale observation space quantities such as $\mathbf{Y}'^f$ by $\mathbf{R}^{-\frac{1}{2}}$ before processing them further. Such scaling has the effect of normalising observations that are possibly of disparate physical quantities with different error standard deviations so that they are dimensionless with standard deviation one. This is useful because it prevents information becoming lost due (say) to rounding errors. The advisability of such a scaling in the context of the stochastic formulation of the EnKF is mentioned in Evensen [8, Section 4.3.2]. A scaled observation operator is also part of the original presentation of the ETKF in Bishop *et al* [3] (although there it is not explicitly exploited to ensure symmetry as above).

With $\widehat{\mathbf{Y}}^f$ available, further improvements to the ETKF algorithm become possible. There is no need to perform the multiplication in (3.3) with consequent loss of accuracy and then perform the eigenvalue decomposition (3.1). Instead, we may start with the SVD

$$(\widehat{\mathbf{Y}}^f)^T = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^T \tag{3.4}$$

where $\mathbf{U}$ is $N \times N$, $\boldsymbol{\Sigma}$ is $N \times m$, and $\mathbf{V}$ is $m \times m$. The matrix $\mathbf{U}$ is the same as the matrix of eigenvectors in (3.1). The eigenvalues may be found from

$$\boldsymbol{\Lambda} = \boldsymbol{\Sigma}\boldsymbol{\Sigma}^T.$$

The ensemble perturbation matrix is then updated by

$$\begin{aligned}\mathbf{X}'^a &= \mathbf{X}'^f\mathbf{T} \\ &= \mathbf{X}'^f\mathbf{U}(\mathbf{I} + \boldsymbol{\Lambda})^{-\frac{1}{2}}.\end{aligned} \tag{3.5}$$

We shall see shortly that it is advantageous not to evaluate $\mathbf{T}$, but instead to evaluate $\mathbf{X}'^a$ by building up the product (3.5) from left to right.

The SVD (3.4) may also be exploited in the update of the ensemble mean. The ensemble Kalman gain may be written as

$$\begin{aligned}\mathbf{K}_e &= \mathbf{X}'^f(\mathbf{Y}'^f)^T(\mathbf{Y}'^f(\mathbf{Y}'^f)^T + \mathbf{R})^{-1} \\ &= \mathbf{X}'^f(\widehat{\mathbf{Y}}^f)^T(\widehat{\mathbf{Y}}^f(\widehat{\mathbf{Y}}^f)^T + \mathbf{I})^{-1}\mathbf{R}^{-\frac{1}{2}} \\ &= \mathbf{X}'^f\mathbf{U}\boldsymbol{\Sigma}(\boldsymbol{\Sigma}^T\boldsymbol{\Sigma} + \mathbf{I})^{-1}\mathbf{V}^T\mathbf{R}^{-\frac{1}{2}}.\end{aligned}$$

Note that the expensive inversion of $\mathbf{Y}'^f(\mathbf{Y}'^f)^T + \mathbf{R}$ has been reduced to the inversion of the diagonal matrix $\boldsymbol{\Sigma}^T\boldsymbol{\Sigma} + \mathbf{I}$. Instead of computing $\mathbf{K}_e$ and then computing the ensemble mean update using

$$\overline{\mathbf{x}^a} = \overline{\mathbf{x}^f} + \mathbf{K}_e(\mathbf{y} - \overline{\mathbf{y}^f})$$

it is better to first build up the product

$$\mathbf{z} = \boldsymbol{\Sigma}(\boldsymbol{\Sigma}^T\boldsymbol{\Sigma} + \mathbf{I})^{-1}\mathbf{V}^T\mathbf{R}^{-\frac{1}{2}}(\mathbf{y} - \overline{\mathbf{y}^f})$$

from right to left and then update the ensemble mean using

$$\overline{\mathbf{x}^a} = \overline{\mathbf{x}^f} + \mathbf{X}'^f\mathbf{U}\mathbf{z}.$$

This procedure avoids storing the matrix $\mathbf{K}_e$ and only needs to store a vector at each stage of building up $\mathbf{z}$. Note that the observation space quantity $\mathbf{y} - \overline{\mathbf{y}^f}$ is once again scaled by $\mathbf{R}^{-\frac{1}{2}}$ before being processed further. The matrix product $\mathbf{X}'^f\mathbf{U}$ that is used in the final step is already available from building up $\mathbf{X}'^a$ using (3.5).

The experiments presented in this dissertation use an ETKF implemented as described above in MATLAB. The SVD (3.4) is performed using the standard MATLAB `svd` function which uses the LAPACK [1] routine `DGESVD`.

3.2 Implementing the EAKF

This section takes the techniques that were used to improve the implementation of the ETKF in Section 3.1 and applies them to the EAKF. We start, however, by revisiting the SVD that is already present in the description of the EAKF in Section 2.4.5.

In was noted in Section 2.4.5 that the matrices $\mathbf{F}$ and $\mathbf{G}$ from the eigenvalue decomposition $\mathbf{P}_e^f = \mathbf{F}\mathbf{G}^2\mathbf{F}^T$ that forms the first stage of the EAKF algorithm may be found from the SVD

$$\mathbf{X}'^f = \mathbf{F}\mathbf{G}\mathbf{W}^T \tag{3.6}$$

where $\mathbf{G}$ is a $p \times p$ diagonal matrix of nonzero singular values and $\mathbf{F}$ and $\mathbf{W}$ are column-orthogonal matrices of sizes $n \times p$ and $N \times p$ respectively. The reason for requiring $\mathbf{G}$ to contain only nonzero singular values is because $\mathbf{G}^{-1}$ is required in the formula (2.34) for the ensemble adjustment matrix $\mathbf{A}$.

However, if we compute the EAKF analysis ensemble perturbation matrix as

$$\mathbf{X}'^a = \mathbf{F}\mathbf{G}\widetilde{\mathbf{U}}(\mathbf{I} + \widetilde{\mathbf{\Lambda}})^{-\frac{1}{2}}\mathbf{W}^T \tag{3.7}$$

then there is no need to evaluate $\mathbf{A}$. It may be verified that $\mathbf{X}'^a$ so calculated is unchanged if we allow $\mathbf{G}$ to include zero singular values whilst remaining square, in which case p is an upper bound for the rank of $\mathbf{X}'^f$ instead of being equal to it as before. In implementing the EAKF we use the SVD in preference to the eigenvalue decomposition to avoid the potential loss of accuracy in forming $\mathbf{P}_e^f$, and we allow the diagonal elements of $\mathbf{G}$ to be zero. A benefit of the latter relaxation is that if we have an SVD routine that is not guaranteed to eliminate all zero singular values from $\mathbf{G}$, then we may still use it in the EAKF if the consequent ease of implementation is judged to be sufficient trade-off for the loss of computational efficiency that comes from not keeping matrices as small as possible.

Turning now to the application of the techniques of Section 3.1 to the EAKF, the analogue of the observation space scaling (3.2) is the $m \times p$ matrix

$$\widetilde{\mathbf{Y}}^f = \mathbf{R}^{-\frac{1}{2}}\mathbf{H}\mathbf{F}\mathbf{G}.$$

This reduces the second matrix for which an eigenvalue decomposition is required to the machine symmetric form

$$(\mathbf{H}\mathbf{F}\mathbf{G})^T\mathbf{R}^{-1}\mathbf{H}\mathbf{F}\mathbf{G} = (\widetilde{\mathbf{Y}}^f)^T\widetilde{\mathbf{Y}}^f.$$

Once again we do not directly perform an eigenvalue decomposition of this matrix but instead perform the SVD

$$(\widetilde{\mathbf{Y}}^f)^T = \widetilde{\mathbf{U}}\widetilde{\mathbf{\Sigma}}\widetilde{\mathbf{V}}^T \tag{3.8}$$

where $\widetilde{\mathbf{U}}$ is $p \times p$, $\widetilde{\mathbf{\Sigma}}$ is $p \times m$, and $\widetilde{\mathbf{V}}$ is $m \times m$. The eigenvalue matrix is

$$\widetilde{\mathbf{\Lambda}} \;=\; \widetilde{\mathbf{\Sigma}}\widetilde{\mathbf{\Sigma}}^T.$$

We may now evaluate $\mathbf{X}'^a$ by building up the product (3.7) from left to right.

For the update of the ensemble mean we may write the ensemble Kalman gain as

$$\begin{aligned}
\mathbf{K}_e &= \mathbf{P}_e^f\mathbf{H}^T(\mathbf{H}\mathbf{P}_e^f\mathbf{H}^T + \mathbf{R})^{-1} \\
&= \mathbf{F}\mathbf{G}^2\mathbf{F}^T\mathbf{H}^T(\mathbf{H}\mathbf{F}\mathbf{G}^2\mathbf{F}^T\mathbf{H}^T + \mathbf{R})^{-1} \\
&= \mathbf{F}\mathbf{G}(\widetilde{\mathbf{Y}}^f)^T(\widetilde{\mathbf{Y}}^f(\widetilde{\mathbf{Y}}^f)^T + \mathbf{I})^{-1}\mathbf{R}^{-\frac{1}{2}} \\
&= \mathbf{F}\mathbf{G}\widetilde{\mathbf{U}}\widetilde{\mathbf{\Sigma}}(\widetilde{\mathbf{\Sigma}}^T\widetilde{\mathbf{\Sigma}} + \mathbf{I})^{-1}\widetilde{\mathbf{V}}^T\mathbf{R}^{-\frac{1}{2}}
\end{aligned}$$

where the inverted matrix is again diagonal. As in the ETKF we do not store $\mathbf{K}_e$. Instead we first build up the product

$$\mathbf{z} = \widetilde{\mathbf{\Sigma}}(\widetilde{\mathbf{\Sigma}}^T\widetilde{\mathbf{\Sigma}} + \mathbf{I})^{-1}\widetilde{\mathbf{V}}^T\mathbf{R}^{-\frac{1}{2}}(\mathbf{y} - \overline{\mathbf{y}^f})$$

from right to left. Once again this scales the observation space quantity $\mathbf{y} - \overline{\mathbf{y}^f}$ by $\mathbf{R}^{-\frac{1}{2}}$ at the earliest opportunity and only requires storing a vector at each stage. We then update the ensemble mean by evaluating

$$\overline{\mathbf{x}^a} = \overline{\mathbf{x}^f} + \mathbf{F}\mathbf{G}\widetilde{\mathbf{U}}\mathbf{z}$$

where $\mathbf{F}\mathbf{G}\widetilde{\mathbf{U}}$ is already available from building up $\mathbf{X}'^a$ using (3.7).

The experiments presented in this dissertation use an EAKF implemented as described above in MATLAB. The SVDs (3.6) and (3.8) are performed using the standard MATLAB `svd` function which uses the LAPACK [1] routine `DGESVD`. The first SVD uses the `'econ'` variant of `svd`, which is equivalent to taking $p = \min(n, N)$.

Chapter 4

The Swinging Spring and Initialisation

Chapter 5 presents the results of experiments that illustrate features of the filters discussed in Chapters 2 and 3. The system used in those experiments is the two-dimensional swinging spring introduced in this chapter. This simple mechanical system is of interest to meteorologists as an illustration of the problem of initialisation. The current chapter accordingly begins with a section introducing the concept of initialisation in the context of NWP. There follows a section on the swinging spring and a section on the numerical method used to integrate the swinging spring equations to produce the results presented in this chapter and the next. The chapter concludes with a section showing the use of the swinging spring to illustrate the techniques of linear and nonlinear normal mode initialisation. With the exception of the section on the numerical method, this chapter largely follows parts of Lynch [20], but with additions and expansions.

Investigation of the initialisation characteristics of the EnKF is an active area of research, although one that time has not permitted to be followed very far in this dissertation. A recent study using a different low-dimensional dynamical system and a purely stochastic rather than a semi-deterministic EnKF may be found in Neef *et al* [21].

4.1 Initialisation

In synoptic weather forecasting we are interested in atmospheric motions having a timescale greater than a day. These phenomena are typified by the Rossby wave solutions to the equations of motion used in NWP. These same equations also permit faster gravity wave solutions having a shorter timescale. These solutions may be of significance locally (in the lee of a steep mountain, for example) but generally they are not of interest to the forecaster and may be regarded as noise. It is advantageous to prevent the development of these unwanted high speed solutions because doing so permits longer time steps to be used in numerical integration methods without encountering instability.

It is quite possible for observational errors to create inaccurate initial conditions that trigger spurious gravity waves in NWP forecasts. These waves can lead to problems even if they do not cause numerical instability. For example, the quality control component of the forecasting system may check new observations for reasonableness against a forecast. If the forecast is inaccurate, good observations may be rejected and bad ones accepted. Another problem occurs in precipitation forecasting. A forecast with excessive gravity wave noise may have an unrealistically large vertical velocity that interacts with the humidity field to give inaccurate rainfall patterns.

A data assimilation scheme must therefore take precautions to ensure that the initial conditions for the forecast step as produced by the preceding analysis step to not inadvertently allow large gravity waves to develop. The process of taking these precautions is known as initialisation. The precautions may take the form of constraints applied within the analysis step or a post-analysis adjustment of the initial conditions.

Instead of using a complex and computationally intensive NWP model for the experiments described in this dissertation, a model of a far simpler system is used, yet one that still possesses the key property of having motions with two distinct timescales. We now define this system.

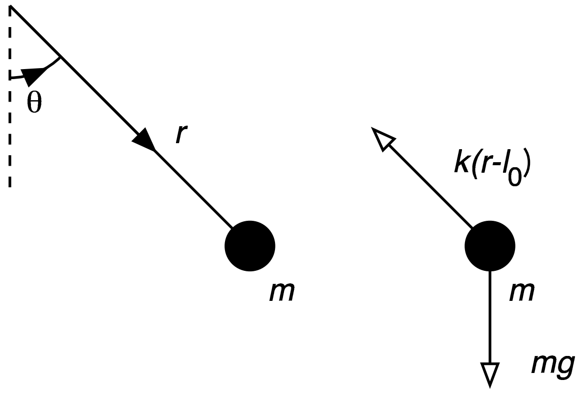

Figure 4.1: Coordinates and forces for the swinging spring. Coordinates are radius r and angle θ. Bob has mass m. Gravitational force is mg; elastic force is $k(r-\ell_0)$ where k is spring elasticity and ℓ_0 its unstretched length.

4.2 The Swinging Spring

Consider a heavy bob of mass m suspended from a fixed point in a uniform gravitational field of acceleration g by a light spring of unstretched length ℓ_0 and elasticity k. The bob is constrained to move in a vertical plane. The spring may stretch along its length but is unable to bend. (See Figure 4.1.)

We locate the bob using polar coordinates (r, θ) where r is measured from the point of suspension and θ is measured from the downward vertical. The corresponding generalised momenta are the radial momentum $p_r = m\dot{r}$ and the angular momentum $p_\theta = mr^2\dot{\theta}$. The Hamiltonian of the system is the sum of the kinetic and potential energies:

$$H = \frac{1}{2m}\left(p_r^2 + \frac{p_\theta^2}{r^2}\right) + \frac{1}{2}k(r-\ell_0)^2 - mgr\cos\theta.$$

From this we may derive the equations of motion

$$\dot{\theta} = \frac{p_\theta}{mr^2} \quad (4.1)$$

$$\dot{p_\theta} = -mgr\sin\theta \quad (4.2)$$

$$\dot{r} = \frac{p_r}{m} \quad (4.3)$$

$$\dot{p_r} \quad = \quad \frac{p_\theta^2}{mr^3} - k(r - \ell_0) + mg\cos\theta. \tag{4.4}$$

This dynamical system has two stationary points: a stable equilibrium with the spring stretched and the bob at rest vertically below the suspension point, and an unstable equilibrium with the spring compressed and the bob at rest vertically above the suspension point. We shall not consider the unstable equilibrium further. The coordinates of the stable equilibrium are $(\theta, p_\theta, r, p_r) = (0, 0, \ell, 0)$ where the equilibrium length ℓ satisfies $k(\ell - \ell_0) = mg$. Linearising the system about this point we obtain

$$\begin{aligned} \dot{\theta} &= \frac{p_\theta}{m\ell^2} && (4.5)\\ \dot{p_\theta} &= -mg\ell\theta && (4.6)\\ \dot{r} &= \frac{p_r}{m} && (4.7)\\ \dot{p_r} &= -k(r - \ell). && (4.8) \end{aligned}$$

Thus we see that for small oscillations the system splits into two independent systems in the variables (θ, p_θ) and (r, p_r). The angular or rotational motion satisfies the second order differential equation

$$\ddot{\theta} + \frac{g}{\ell}\theta = 0$$

whilst the radial or elastic motion satisfies

$$\ddot{r'} + \frac{k}{m}r' = 0$$

where $r' = r - \ell$. These are equations of simple harmonic motion with angular frequencies

$$\omega_\theta = \sqrt{\frac{g}{\ell}}$$

and

$$\omega_r = \sqrt{\frac{k}{m}}$$

respectively. The ratio of frequencies is

$$\varepsilon = \frac{\omega_\theta}{\omega_r} = \sqrt{\frac{mg}{k\ell}} = \sqrt{\frac{\ell - \ell_0}{\ell}} < 1.$$

We shall be using parameter values such that $\epsilon \ll 1$. The rotational motion is a low frequency mode analogous to Rossby waves; the associated variables (θ, p_θ) are called the slow variables. The elastic motion is a high frequency mode analogous to gravity waves; the associated variables (r, p_r) are called the fast variables. For finite oscillations the motions will not be independent but will interact.

Figure 4.2 shows coordinates against time for a sample trajectory of the swinging spring obtained using the numerical method to be described in Section 4.3. It also shows the modulus of the Fourier transform of the coordinates. Following Lynch [20] the parameter values are $m = 1$, $g = \pi^2$, $k = 100\pi^2$, and $\ell = 1$. These give motions with cyclic frequencies $f_\theta = \omega_\theta/2\pi = 0.5$ and $f_r = \omega_r/2\pi = 5$, and hence a frequency ratio $\varepsilon = 0.1$. The initial conditions are $(\theta, p_\theta, r, p_r) = (1, 0, 1.05, 0)$. It can be seen from the Fourier transforms that most of the energy of the slow variables is concentrated around $f = f_\theta$ and most of the energy of the fast variables is concentrated around $f = f_r$. However, there is sign of interaction between the variables in the minor peak in the Fourier transforms of r and p_r around $f = 1 = 2f_\theta$.

4.3 Numerical Method of Integration

Note: Most of the notation in this section is not used elsewhere in the dissertation and is not included in the *List of Symbols*.

In the experiments presented in this dissertation, the swinging spring equations are integrated using the standard MATLAB `ode45` function. This function uses an explicit Runge-Kutta (4,5) pair with an adaptive step size. That is to say, an explicit fifth-order Runge-Kutta method is used to integrate the equations whilst the difference between this and a related fourth-order method is used to estimate the truncation error. The step size is adjusted

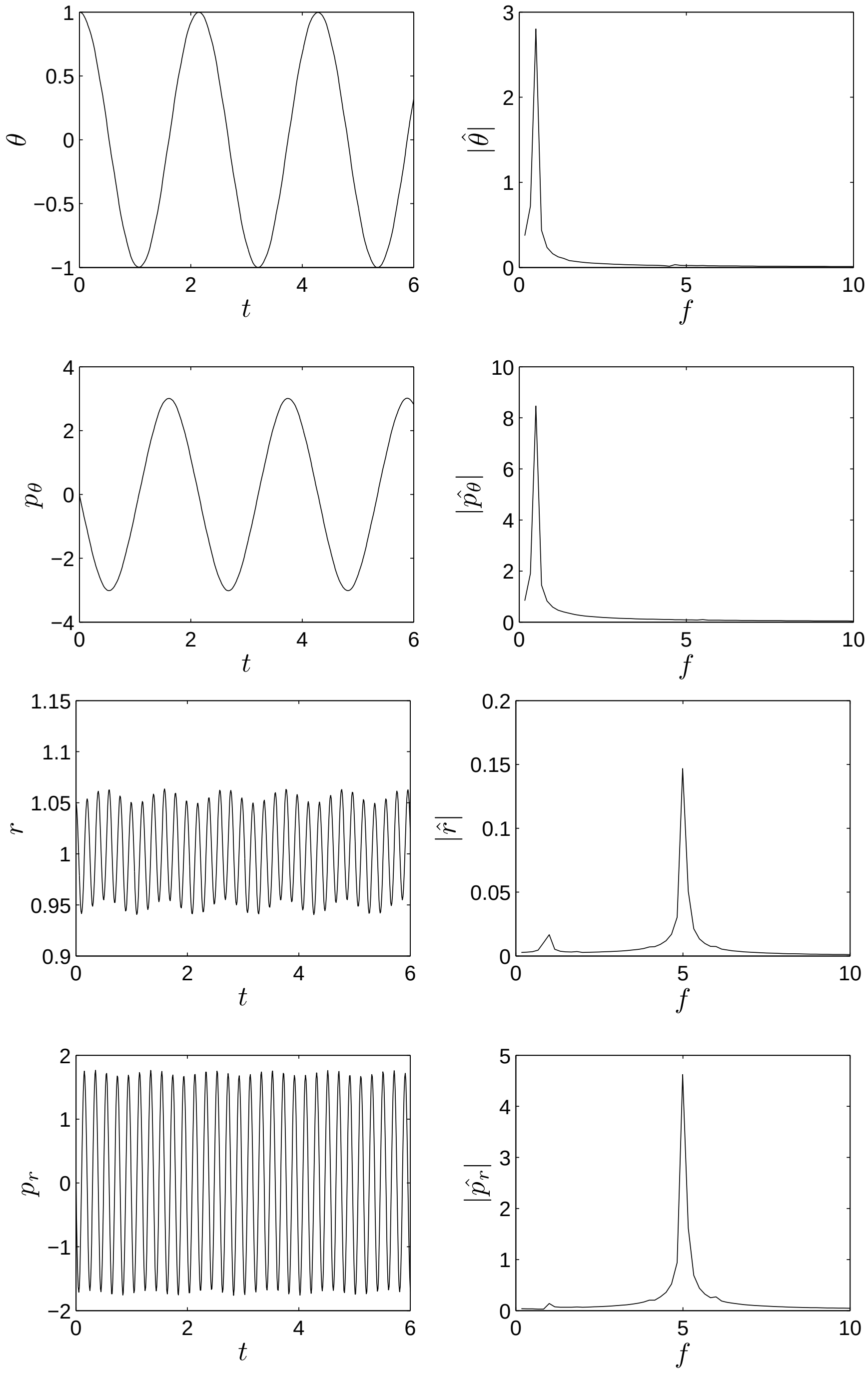

Figure 4.2: Coordinates and their Fourier transforms for an uninitialised swinging spring. Parameter values are such that rotational and elastic frequencies are $f_\theta = 0.5$ and $f_r = 5$ respectively. Initial conditions are $(\theta, p_\theta, r, p_r) = (1, 0, 1.05, 0)$.

to keep the error estimate within a user-specified tolerance. The original reference for the Runge-Kutta pair used in `ode45` is Dormand and Prince [6]; it is also discussed under the name DOPRI(5,4) in Lambert [17, Section 5.10].

The tolerance determining the step size is specified in terms of two parameters: a scalar `RelTol` specifying a relative tolerance and a vector `AbsTol` specifying an absolute tolerance for each state space coordinate. If $y(j)$ is the jth component of the solution vector at some time and $e(j)$ is the corresponding component of the error estimate from the Runge-Kutta pair, then the step size is adjusted so that

$$|e(j)| \leq \max(\texttt{RelTol} \times |y(j)|, \texttt{AbsTol}(j)). \tag{4.9}$$

The experiments in Chapter 5 use observations of state space coordinates with error standard deviations at least one-tenth the amplitude of the oscillations in the truth, where the truth is as in Figure 4.5. If we take as our primary aim in selecting values for the tolerance parameters the need to keep the truncation error small compared to the observational error, this aim may be achieved with a comfortable margin by using the default values, which are 10^{-3} for `RelTol` and 10^{-6} for every component of `AbsTol`.

There is also a parameter `MaxStep` that acts as an upper bound on the step size. This may be used to ensure stability of the numerical method. Stability analysis of a nonlinear system is a complex problem, so we confine ourselves to analysing the linearised system (4.5)–(4.8). This system decouples into two independent simple harmonic oscillators, so we consider the single oscillator that may be written in complex form as

$$\dot{y} = iy$$

and relate it to the complete linear system using scaling arguments afterwards. Lambert [17, Section 5.12] shows that a Runge-Kutta method applied to the general linear scalar system

$$\dot{y} = \lambda y$$

leads to a difference equation of the form

$$y_{k+1} = R(\hat{h})y_k$$

where $\hat{h} = \lambda h$. R is called the *stability function* of the method. Lambert gives the stability condition

$$|R(\hat{h})| < 1.$$

However, it should be borne in mind that this is in the context of solving systems with $\mathrm{Re}(\lambda) < 0$ and for which the exact solution satisfies $y \to 0$ as $t \to 0$. Our harmonic oscillator has $\mathrm{Re}(\lambda) = 0$ and $|y| =$ constant. Since the exact solution satisfies the difference equation

$$y(t_{k+1}) = e^{ih}y(t_k)$$

it is better for us to seek $\hat{h}$ such that $R(\hat{h})$ has modulus close to one and argument close to h.

For DOPRI(5,4) Lambert [17, Section 5.12] gives

$$R(\hat{h}) = 1 + \hat{h} + \frac{\hat{h}^2}{2} + \frac{\hat{h}^3}{6} + \frac{\hat{h}^4}{24} + \frac{\hat{h}^5}{120} + \frac{\hat{h}^6}{600}.$$

If we plot $|R(\hat{h})| - 1$ and $\arg(R(\hat{h})) - h$ against h, we obtain the curves in Figure 4.3. We see that both quantities are very close to zero for $h \leq 0.3$. Since the period T of our harmonic oscillator is 2π, we may conveniently take the upper bound on h to be $T/20$, which enables us to generalise immediately to an arbitrary simple harmonic oscillator. The swinging spring with the parameter values given in Section 4.2 has periods $T_\theta = 2$ and $T_r = 0.2$. Therefore we take `MaxStep` $= 0.01$. Using this value there has been no evidence of instability in the experiments.

It should be pointed out that the above is not the last word on the analysis of the numerical method or indeed on the selection of a numerical method to integrate the equations of motion. Some remarks on taking the matter further may be found in Section 7.2.1.

One must take care in using a numerical method that has not been fully

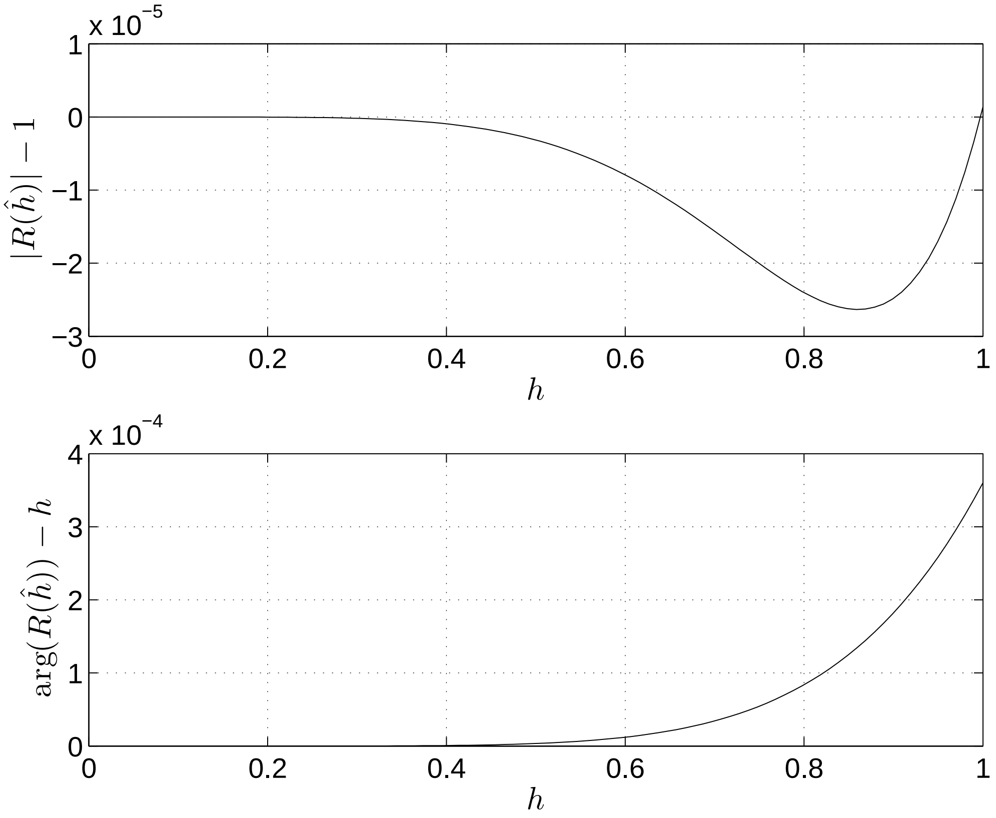

Figure 4.3: Stability function for `ode45` and the simple harmonic oscillator $y' = iy$. Stability function $R(\hat{h})$ (where $\hat{h} = ih$) is plotted as modulus relative to one and argument relative to h. We want both to be close to zero.

analysed. It is always possible that the discrete system may not have the same behaviour as the continuous system of which it is a model. Thus the results of experiments with the discrete system cannot legitimately be used to draw conclusions about the continuous system. However, this is not what we shall be doing here. For the purposes of this dissertation we may regard the discrete system as primary. We are interested in it because, as we shall verify experimentally in Section 4.4, it has properties that illustrate the problems and techniques of initialisation. The role of the continuous system then becomes that of a theory about how the system of primary interest might be working. This theory may be used to suggest what experiments it would be useful to perform and what the results might be, but we must always verify its predictions by experiment.

Similarly, when we come to use the swinging spring in our experiments with the EnKF, it will be the discrete system that we use to generate our truth and use in the forecast step. There will be no direct reference to the continuous system of which the discrete system is supposed to be a model.

4.4 Normal Mode Initialisation

A comprehensive account of the possible motions of the swinging spring may be found in Lynch [19]. Here we confine ourselves to exhibiting a few sample trajectories. We supplement the trajectory of Figure 4.2 with a couple of related trajectories illustrating the techniques of linear and nonlinear normal mode initialisation.

Suppose that we know that high frequency oscillations are absent from the motion of the swinging spring, yet due to observational errors we have predicted the motion shown in Figure 4.2. How can we adjust the initial conditions to get rid of the high frequency noise?

The technique of linear normal mode initialisation attempts to suppress high frequency oscillations by setting their initial amplitude to zero. In the case of the swinging spring this entails setting $r(0) = \ell$ and $p_r(0) = 0$. In the linear system (4.5)–(4.8) this would suppress oscillations in the fast variables for all time. However, in the full system (4.1)–(4.4) the nonlinear interac-

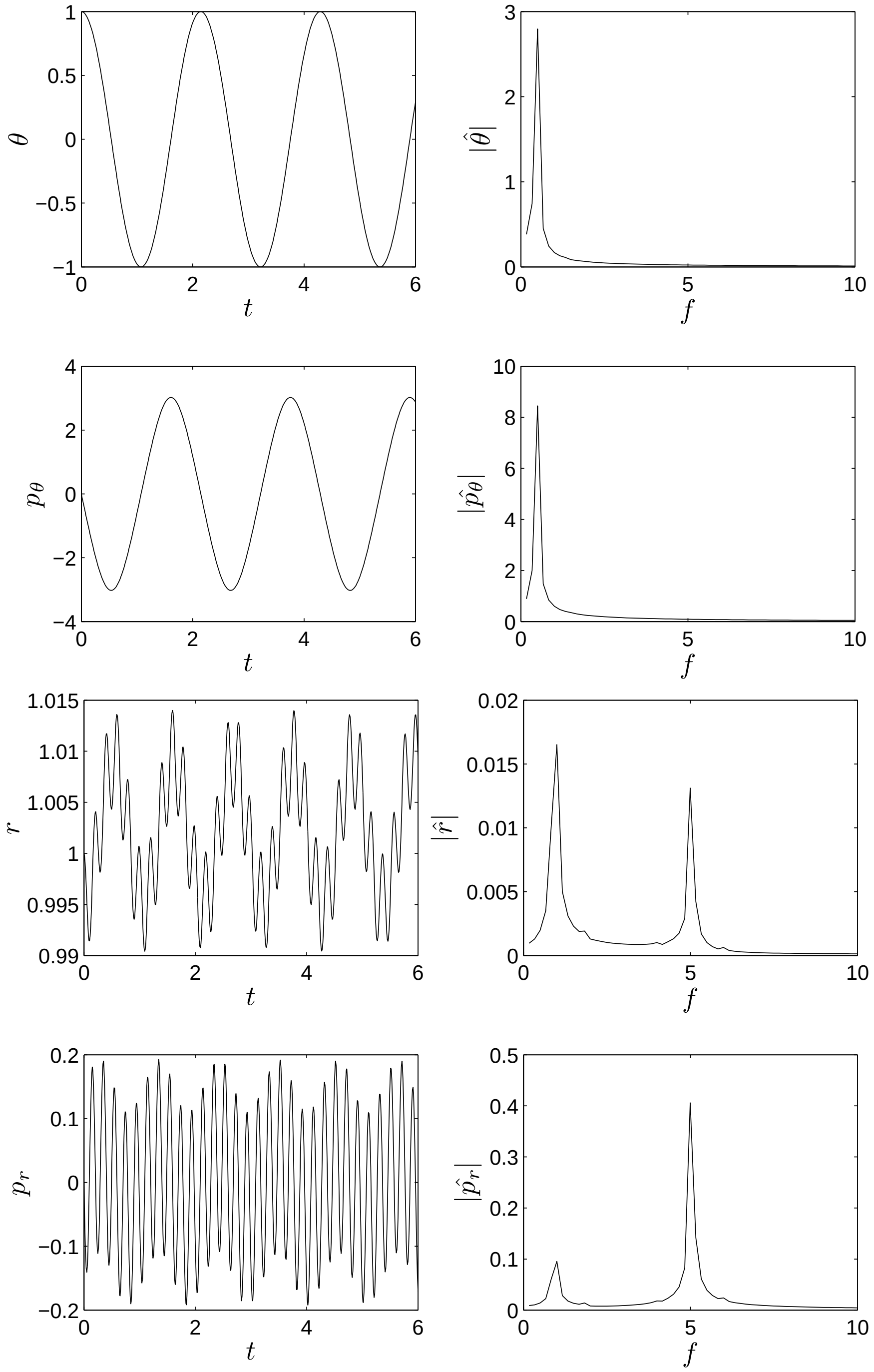

Figure 4.4: Coordinates and their Fourier transforms for a swinging spring with linear normal mode initialisation. Parameter values as in Figure 4.2 but with initial conditions $(\theta, p_\theta, r, p_r) = (1, 0, 1, 0)$. Note that r and p_r scales are one-tenth those in Figure 4.2

tion between the variables leads to the motion in the slow variables exciting high frequency oscillations in the fast variables as shown in Figure 4.4. Nevertheless, there is an improvement compared to Figure 4.2: the amplitude of the high frequency oscillations is much reduced and an underlying slow oscillation in r of frequency $f = 1 = 2f_\theta$ is clearly emerging.

The technique of nonlinear normal mode initialisation sets the initial rates of change of the fast variables to zero, the hope being that this will prevent large amplitude high frequency oscillations from developing. In the case of the swinging spring we must adjust the initial conditions so that $\dot{r}(0) = 0$ and $\dot{p_r}(0) = 0$. To achieve the first of these we simply use (4.3) and set $p_r(0) = 0$. To achieve $\dot{p_r}(0) = 0$ we calculate $\dot{\theta}(0)$ from (4.1), substitute into (4.4), and rearrange to obtain the adjusted initial value of r:

$$r(0) = \frac{\ell(1 - \varepsilon^2[1 - \cos\theta(0)])}{1 - (\dot{\theta}(0)/\omega_r)^2}. \tag{4.10}$$

One further adjustment is needed.[1] In order to ensure that the value of $\dot{\theta}(0)$ used in (4.10) is consistent with (4.1) and the new value of $r(0)$, we must set

$$p_\theta(0) = mr(0)^2\dot{\theta}(0).$$

The results of nonlinear normal mode initialisation are shown in Figure 4.5. The high frequency oscillation has been largely suppressed, with just a small residual in the Fourier transforms of the fast variables at $f = f_r$. Radial variation does not vanish, however. The spring is stretched twice per angular cycle at the bottom of the swing where the speed of the bob and the centrifugal effect are at their greatest. This leads to the observed oscillation in r and p_r of frequency $f = 1 = 2f_\theta$. The radial motion is said to be slaved to the angular motion and the phenomenon is sometimes called ‘balanced fast motion’.

[1] Lynch [20] does not mention this step, presumably because in cases where $p_\theta(0) = \dot{\theta}(0) = 0$ it makes no difference.

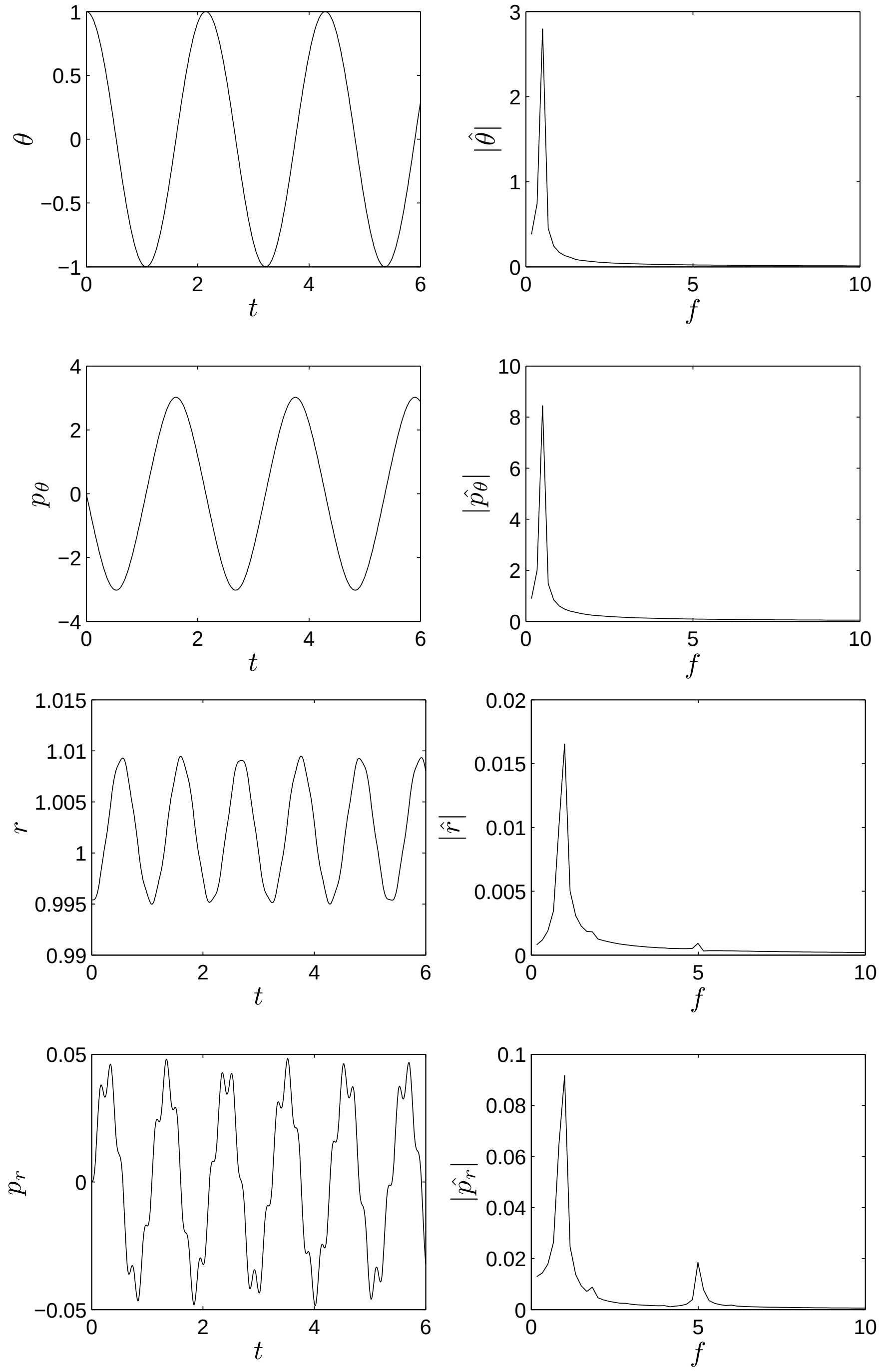

Figure 4.5: Coordinates and their Fourier transforms for a swinging spring with nonlinear normal mode initialisation. Parameter values as in Figures 4.2 and 4.4 but with initial conditions $(\theta, p_\theta, r, p_r) = (1, 0, 0.99540, 0)$. Note that p_r scale is one-quarter that in Figure 4.4

Chapter 5

Experimental Results

This chapter presents the results of experiments with an ETKF and an EAKF implemented as described in Chapter 3 using observations of the swinging spring system described in Chapter 4. The true trajectory in all experiments is the nonlinear normal mode initialised trajectory of Figure 4.5. The model used in the forecast step is the same as the model used to generate the true trajectory. The model noise is taken be zero. The experiments differ in the analysis step, the ensemble size, and whether the observations are perfect (noise-free) or imperfect.

The experiments reveal some features of the filters that require explanation. These explanations are furnished in Chapter 6.

5.1 Experiments with the ETKF

We start by looking at the results of experiments with the ETKF and frequent, perfect observations of all four coordinates. The first observation is at time 0.1 and subsequent ones follow at intervals of 0.1. Although the actual observation errors are zero, the covariance matrix **R** passed to the filter is that of observations having uncorrelated errors with the standard deviations listed in Table 5.1. These standard deviations are close to one-tenth of the amplitude of the oscillations in the truth. The same covariance matrix is used in generating the initial ensemble, which is done as follows. An ensemble of

Coordinate	Standard deviation
θ	0.1
p_θ	0.3
r	7×10^{-4}
p_r	5×10^{-3}

Table 5.1: Observation error standard deviations passed to the filter in experiments with perfect observations. These standard deviations are also used to generate the initial ensemble.

pseudo-random vectors is drawn from a normal distribution with the given covariance matrix as its covariance matrix and the true initial state as its mean. This ensemble is then translated slightly so that the ensemble mean coincides exactly with the true initial state (this is an analogue for the initial ensemble of taking perfect observations). The translated random ensemble is used as the initial ensemble.

The result of an experiment with ensemble size $N = 10$ is shown in Figures 5.1 and 5.2. The graphs show the difference between the filter and the truth. Figure 5.1 shows the individual ensemble members, whilst Figure 5.2 shows the ensemble mean and the ensemble mean ± ensemble standard deviation. All graphs show the expected initial rapid decrease in filter error. However, after about two seconds the filter settles into an oscillation about the true trajectory with only a slight further decrease in the general level of the error. The graphs of ensemble statistics show some considerable intervals of time during which the true state of the system (represented by zero on the vertical axis) is outside the band defined by the ensemble mean ± ensemble standard deviation. This suggests that the ensemble statistics may be inconsistent with the actual error, either the mean being biased or the standard deviation being too small or both. This is confirmed by computing the fraction of analyses having an ensemble mean within one ensemble standard deviation of the truth for each coordinate. For unbiased, normally-distributed, analysis errors with standard deviation equal to the ensemble standard deviation, one would expect this fraction to be about 0.68 (from statistical tables such as Kreyszig [16, Appendix 5, Table A7]). The actual errors need not be normally-distributed, but this is still a useful guide. In the

Experiment	θ	p_θ	r	p_r
ETKF, perfect observations, $N = 10$	0.31	0.32	0.32	0.29
ETKF, perfect observations, $N = 50$	0.92	0.93	0.90	0.93
ETKF, imperfect observations, $N = 10$	0.56	0.56	0.94	0.94
ETKF, imperfect observations, $N = 50$	0.66	0.67	1.00	1.00
EAKF, perfect observations, $N = 10$	1.00	1.00	1.00	1.00
EAKF, perfect observations, $N = 50$	1.00	1.00	1.00	1.00
EAKF, imperfect observations, $N = 10$	0.67	0.68	0.95	0.96
EAKF, imperfect observations, $N = 50$	0.70	0.72	1.00	1.00

Table 5.2: Fraction of analyses with ensemble mean within one ensemble standard deviation of truth. Computed from 100 runs of the filters with different random initial conditions and, in the case of imperfect observations, different random observation errors. For normally-distributed errors this fraction should be about 0.68.

case shown in Figure 5.2 the actual fractions are 0.43, 0.43, 0.23, and 0.37 for θ, p_θ, r, and p_r respectively. Further confirmation is provided by running the filter 100 times with different random initial ensembles and computing the same fractions. The first row of Table 5.2 shows that they are all around 0.3.

The result of increasing the ensemble size to $N = 50$ is shown in Figures 5.3 and 5.4. The statistics in Figure 5.4 look more consistent than those in Figure 5.2, with the band defined by the ensemble mean ± ensemble standard deviation encompassing the truth (zero on the vertical axis) most of the time. This is confirmed by running the filter 100 times again to obtain the numbers in the second row of Table 5.2. The fraction of analyses with ensemble mean within one ensemble standard deviation of truth is 0.90 or more for each coordinate. That this is in excess of the expected 0.68 is explainable by the observations being error-free whilst the filter is operating on the assumption that they have the error standard deviations shown in Table 5.1.

Figure 5.3 reveals another feature of the ETKF that is visible in Figure 5.1 but becomes more obvious when the ensemble size is increased: after the first observation is assimilated, the number of distinct ensemble members collapses to five, some trajectories presumably being occupied by multiple members.

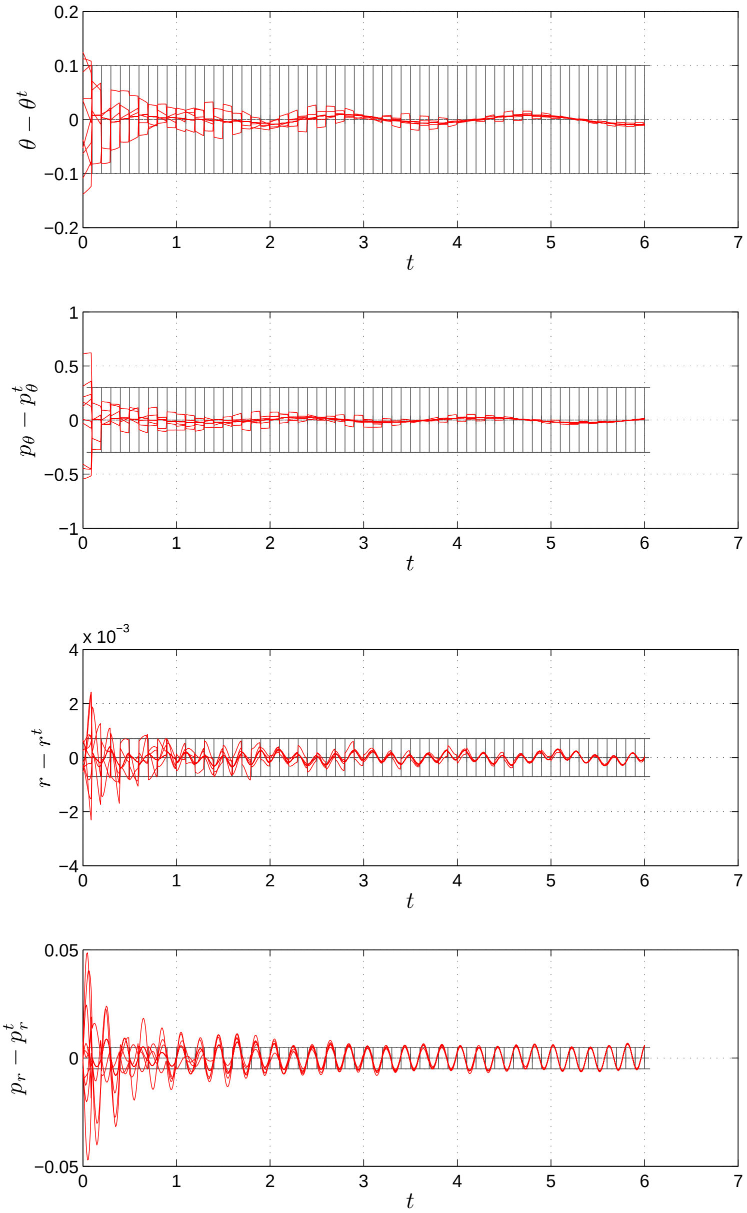

Figure 5.1: ETKF, perfect observations, $N = 10$: ensemble members. Coordinates are plotted relative to truth. Lines showing individual ensemble members are superimposed on observations plotted as error bars. Radius of error bars equals standard deviation passed to filter. Because observations in this case are frequent and perfect, error bars form a grid vertically centred on zero.

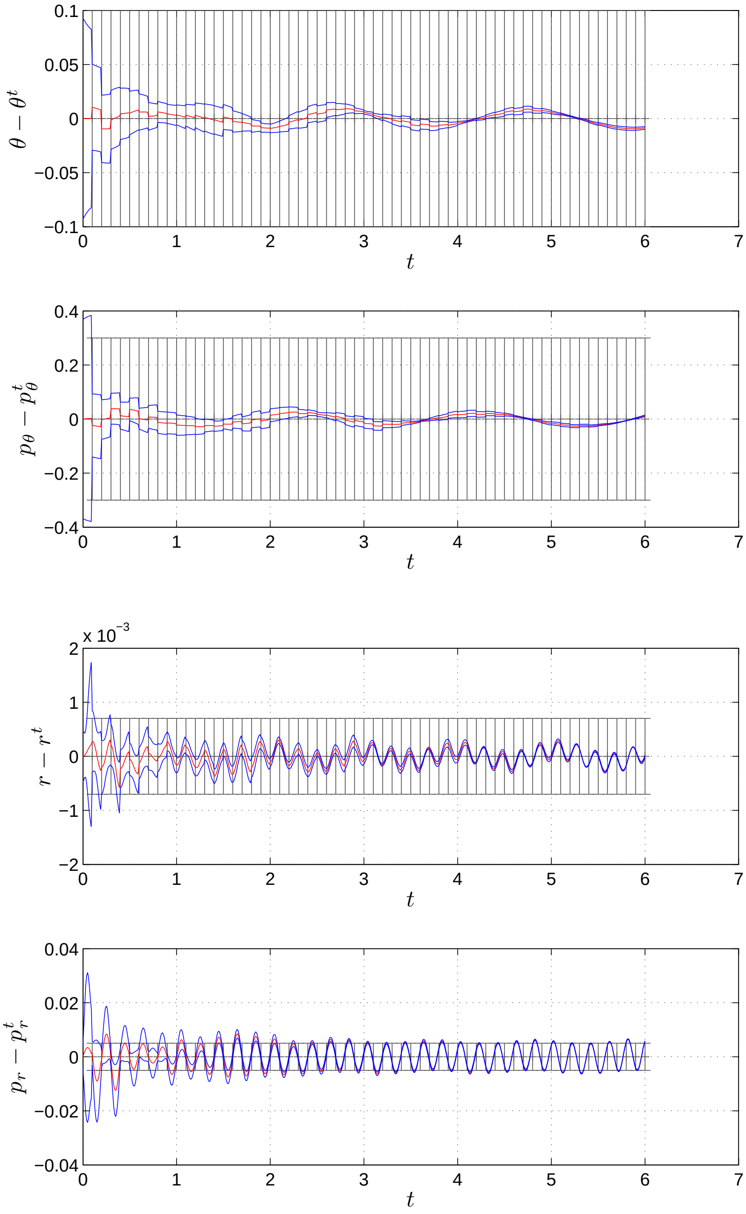

Figure 5.2: ETKF, perfect observations, $N = 10$: ensemble statistics. Coordinates are plotted relative to truth. Three lines showing ensemble mean and ensemble mean ± ensemble standard deviation are superimposed on observations plotted as error bars. Errors bar conventions are as in Figure 5.1.

Coordinate	Standard deviation
θ	0.1
p_θ	3
r	0.06
p_r	1.5

Table 5.3: Standard deviations used to generate the initial ensemble in experiments with imperfect observations.

Like the inconsistent ensemble statistics for $N = 10$ this is a feature requiring explanation.

In the experiments presented so far, the observations have been specialised in that they have been noise-free, frequent, and made of all four coordinates of the system. We now relax these assumptions to see what effect this has on the ensemble statistics and number of distinct ensemble members. For the next experiments the interval between observations is increased to 0.37. As well as being larger than the previous interval of 0.1, this interval is chosen because it is not a submultiple of the natural oscillation periods $T_\theta = 2$ and $T_r = 0.2$ of the system (thus removing another specialising assumption of previous experiments). Instead of observing all coordinates, only θ is observed. As before, the observation error standard deviation passed to the filter for θ is 0.1, but now random errors of this magnitude really are added to the observations. The initial ensemble is generated using a diagonal covariance matrix corresponding to the standard deviations listed in Table 5.3. The standard deviation for θ is the same as that used for observations. The standard deviations for the other coordinates are approximately equal to the amplitudes of the uninitialised oscillations in Figure 4.2. The intention is that the initial ensemble represents almost complete ignorance about these coordinates. The initial ensemble is generated using pseudo-random vectors as in the experiments with perfect observations except that there is no final translation to make the ensemble mean coincide exactly with the true initial state.

The result of an experiment with imperfect observations and an ensemble size $N = 10$ is shown in Figures 5.5 and 5.6. Statistics from 100 runs are shown in the third row of Table 5.2. These runs use different random obser-

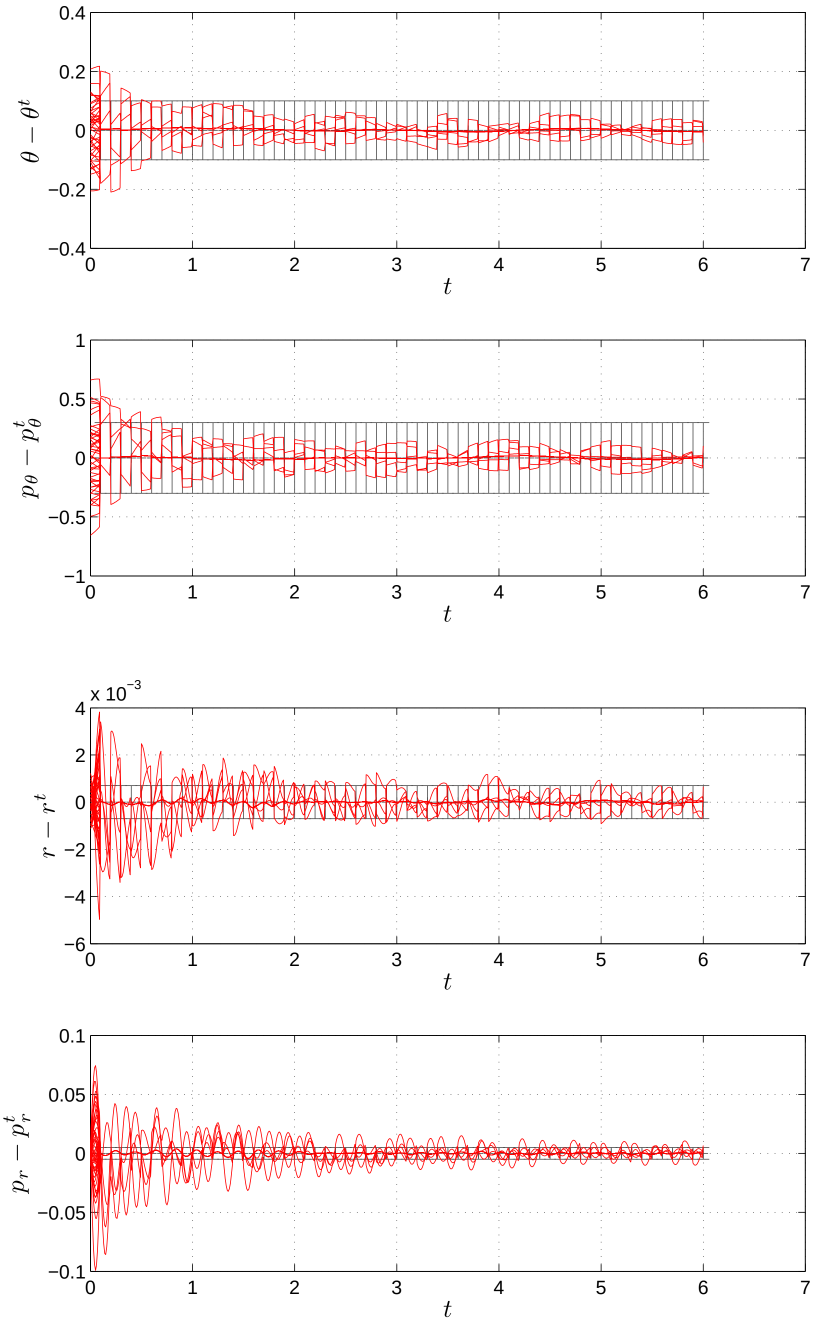

Figure 5.3: ETKF, perfect observations, $N = 50$: ensemble members. Plotting conventions as in Figure 5.1.

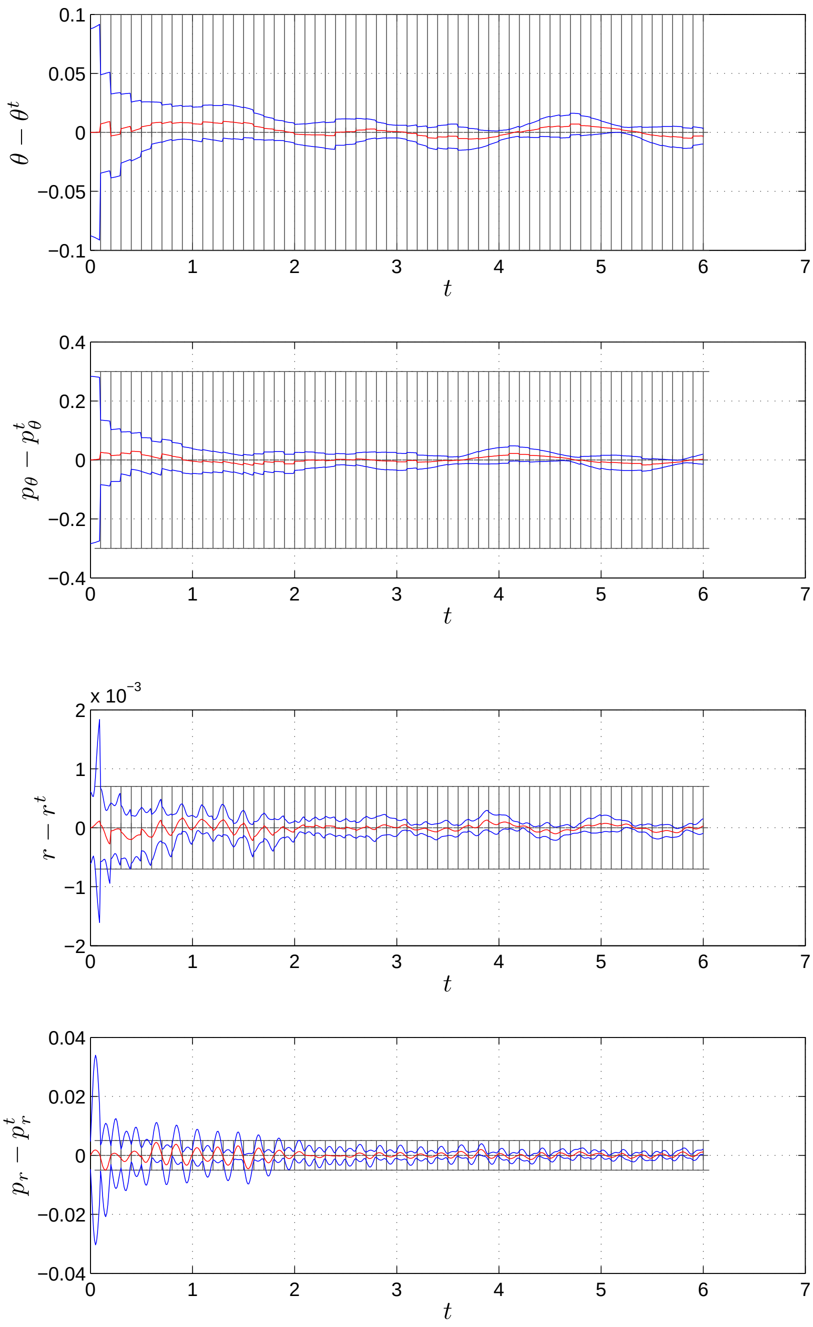

Figure 5.4: ETKF, perfect observations, $N = 50$: ensemble statistics. Plotting conventions as in Figure 5.2.

vation errors as well as different random initial ensembles. The fraction of analyses with the ensemble mean within one standard deviation of the truth is 0.56 for θ and p_θ, which is short of the expected 0.68 but an improvement on the just over 0.30 in the perfect observation case. For r and p_r the fractions are unexpectedly large at 0.94. A possible explanation is that this is due to ignorance on the part of the filter rather the accuracy. It can be seen from the lower two graphs in Figure 5.6 that there is no decrease in the general level of the ensemble standard deviation in r and p_r from its initial value representing complete ignorance of the values of these coordinates, so it is not surprising that the ensemble mean should agree with the truth to within this large margin of error. That observations of θ alone should provide little or no information about r and p_r is not surprising when it is recalled that in the linearised system (4.5)–(4.8) θ and p_θ are totally independent of r and p_r. It is also possible that the improved ensemble statistics in θ and p_θ are due to larger standard deviations rather than more accurate means, the increase in standard deviation coming from the longer time interval between observations of θ and the absence of any direct observations of p_θ. The increase in standard deviation may be seen on comparing the top two graphs of Figure 5.2 with the corresponding graphs of Figure 5.6 and noting that the vertical axis range of the latter is at least 10 times that of the former (the scales may also be compared by noting that the error bars in the top graphs have the same length in each case).

As for the collapse in the number of distinct ensemble members, there is nothing visible in Figure 5.5 in the graphs for p_θ, r, and p_r (the unobserved coordinates) but the graph for θ shows a collapse to two distinct states after the assimilation of each observation. There is a subsequent a fanning out of distinct ensemble members from one of these states (presumably due to differences in the other coordinates) whilst the other state continues as a single ensemble member on its own.

Increasing the ensemble size to $N = 50$ gives the graphs in Figures 5.7 and 5.8 and the 100-run statistics in the fourth row of Table 5.2. As in the case of perfect observations, increasing the ensemble size improves the ensemble statistics. The numbers for θ and p_θ in Table 5.2 are close the ideal

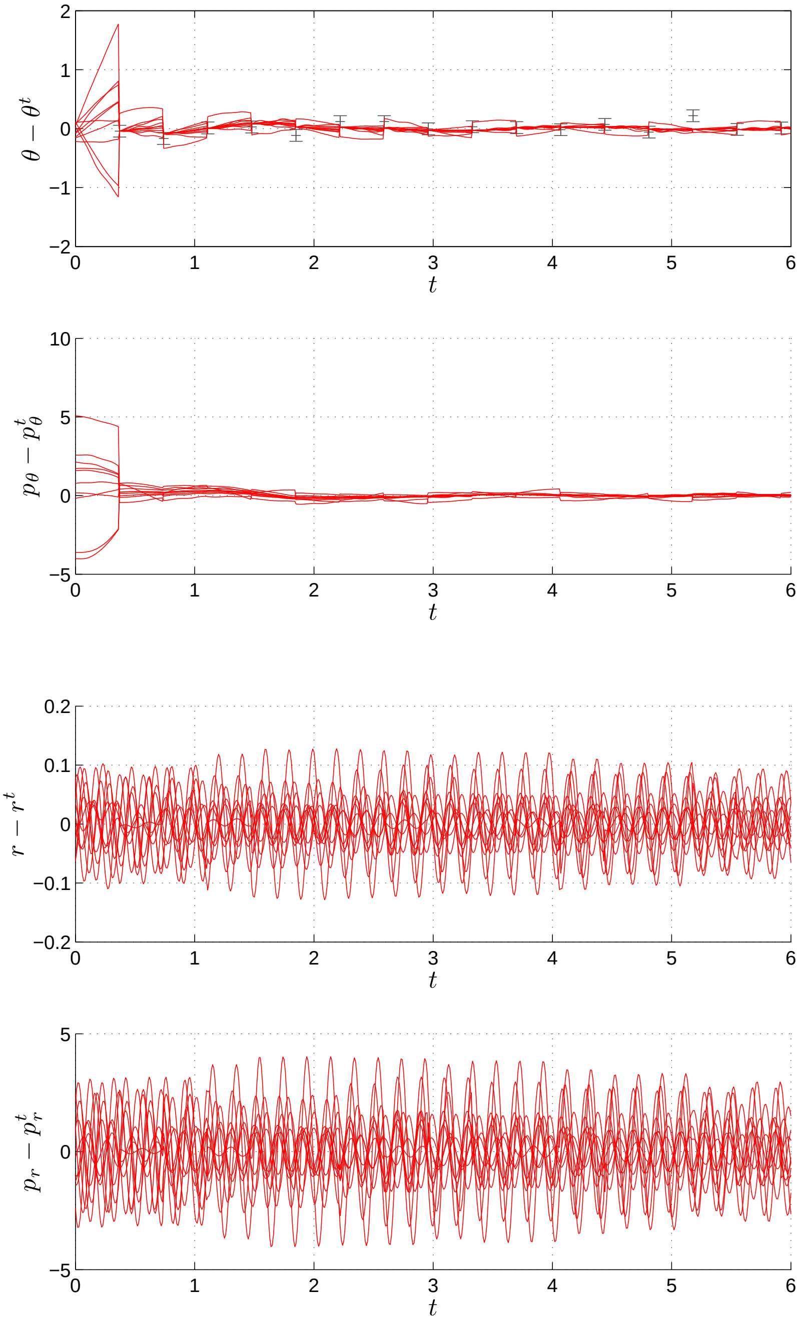

Figure 5.5: ETKF, imperfect observations, $N = 10$: ensemble members. Plotting conventions as in Figure 5.1. Only the first graph shows error bars because only the first coordinate is observed.

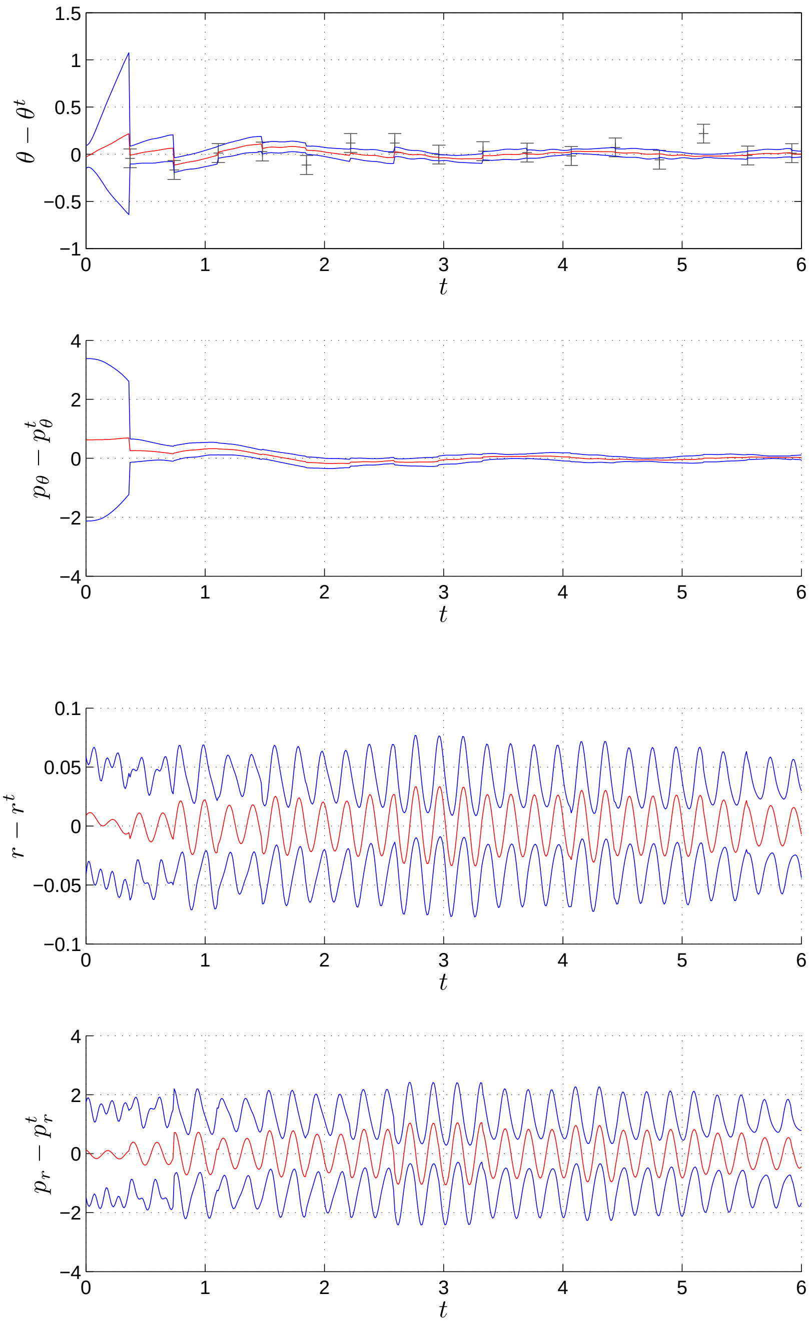

Figure 5.6: ETKF, imperfect observations, $N = 10$: ensemble statistics. Plotting conventions as in Figure 5.2. Only the first graph shows error bars because only the first coordinate is observed.

of 0.68. The numbers for r and p_r are both 1.00, quite possibly for the reason given above in the case $N = 10$: the ensemble standard deviation for these two coordinates does not decrease from its initial value representing complete ignorance (see the bottom two graphs of Figure 5.8).

The graph of θ in Figure 5.7 shows the same collapse-and-fan structure noted above in the case $N = 10$. There is now also sign of an outlier in the graph of p_θ, presumably due to the influence of the outlier in θ through the equation of motion (4.2). There may be some sign of this in Figure 5.5 too, but it is clearer in this case.

This ends the experiments with the ETKF. There is a summary of the results in Section 5.3 at the end of the chapter.

5.2 Experiments with the EAKF

Repeating the ETKF experiments with the EAKF reveals neither of the features observed for the former in Section 5.1. The 100-run EAKF statistics in the lower half of Table 5.2 show no sign of inconsistency. For the perfect observations (fifth and sixth rows) the ensemble mean is within one ensemble standard deviation of the truth for virtually all analyses, as one would expect with perfect observations. For the imperfect observations (seventh and eighth rows) the fraction is close to the expected 0.68 for θ and p_θ, and is close to 1.00 for r and p_r. The latter may be explained as in Section 5.1 by the filter having a large standard deviation that correctly reflects its almost complete ignorance of the true state of these variables.

Also absent from the EAKF results is the collapse in the number of distinct ensemble members that occurs with the ETKF. Graphs illustrating both this and the absence of inconsistent ensemble statistics may be found in Appendix C

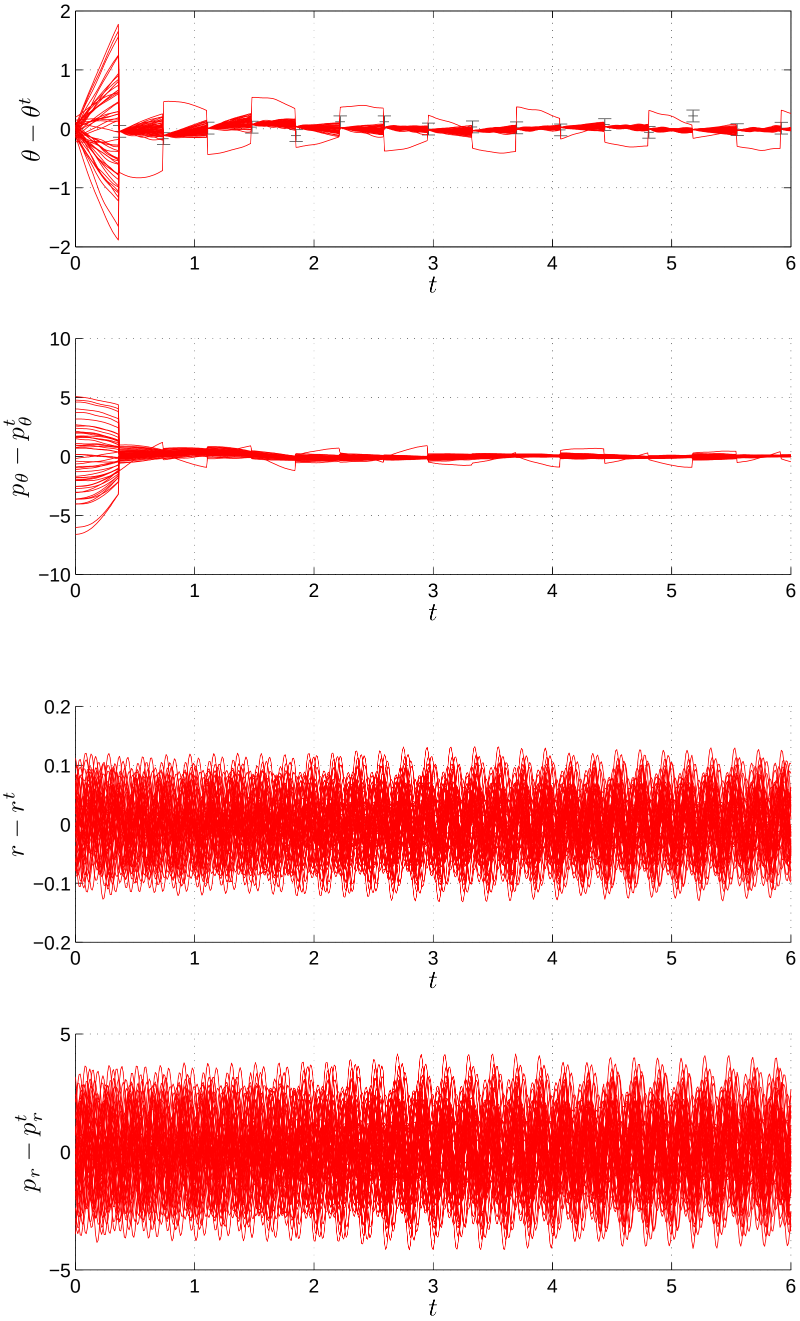

Figure 5.7: ETKF, imperfect observations, $N = 50$: ensemble members. Plotting conventions as in Figure 5.5.

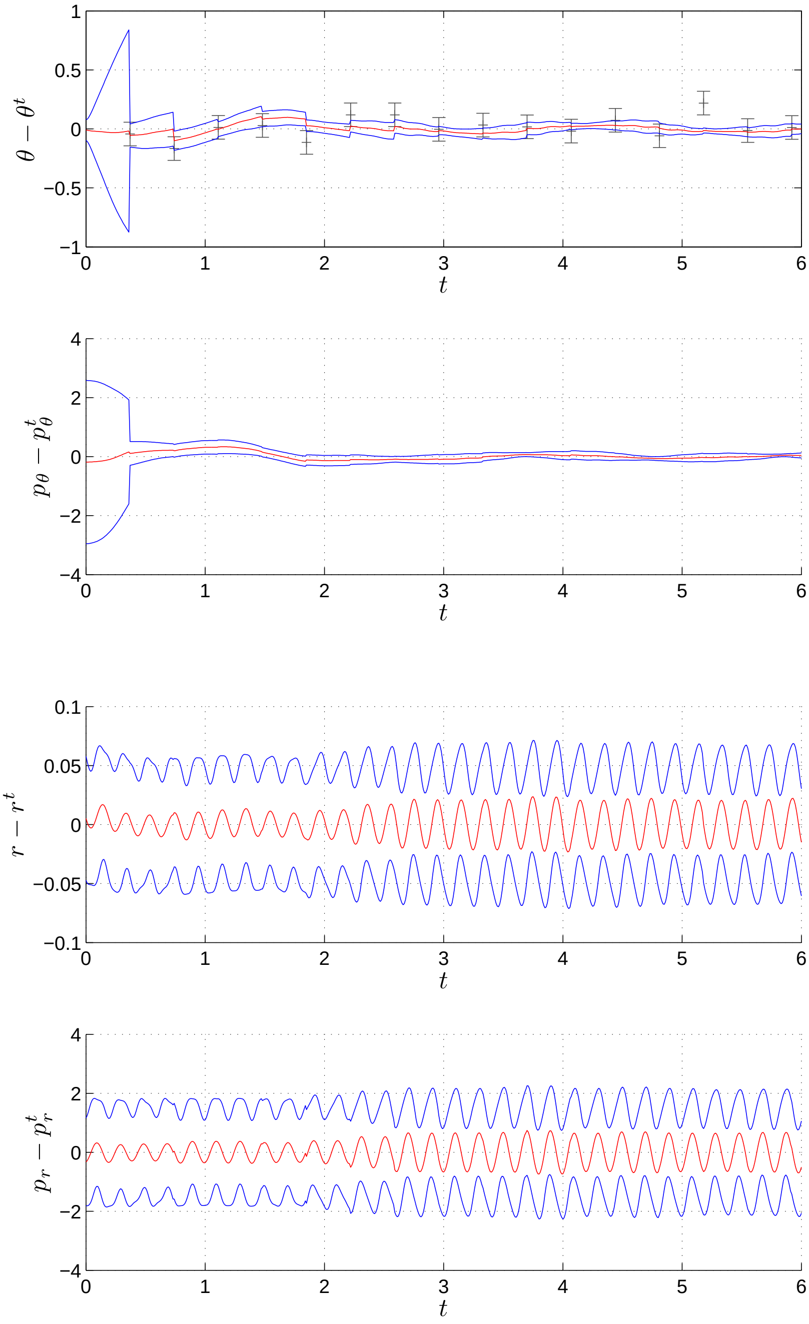

Figure 5.8: ETKF, imperfect observations, $N = 50$: ensemble statistics. Plotting conventions as in Figure 5.6.

5.3 Summary

The experiments in this chapter have revealed two features of the ETKF that require further investigation. The first is that the filter may produce analysis ensembles with statistics that are inconsistent with the actual error, either the mean being biased or the standard deviations of the coordinates being too small or both. This statistical inconsistency seems to decrease with increasing ensemble size. It may also be masked or reduced by increased filter uncertainty due to uncertain initial conditions, less frequent observations, or fewer observed coordinates.

The second feature is that each assimilation of an observation by the ETKF produces a collapse in the number of distinct values of the observed coordinates in the ensemble. It may be conjectured from the results presented that when m coordinates are observed, there is a collapse in the number of distinct values of each observed coordinate to $m + 1$ following assimilation. It may be further conjectured from the results that m of these values are occupied by single ensemble members whilst the remaining $N - m$ members occupy the remaining value. Note that such a collapse will only have an effect on ensembles with $N > m + 1$. Thus a collapse is likely to be apparent with low-dimensional systems such as the swinging spring, but not with NWP-type systems that have $N \ll m$.

The EAKF appears not to possess either of these features. An explanation for their presence in the ETKF is given in Chapter 6.

Chapter 6

Explanation of Experimental Results

Two features of the ETKF emerged from the experiments of Chapter 5. The first is that the filter may produce analysis ensembles with statistics that are inconsistent with the actual error, and the second is that each assimilation of an observation produces a collapse in the number of distinct values of the observed coordinates in the ensemble. This chapter presents explanations of both features. The explanation of the inconsistent statistics reveals a potential flaw in deterministic formulations of the analysis step of the EnKF. This flaw appears to have been overlooked in the literature and its discovery is probably the most important thing in this dissertation.

6.1 Inconsistent Analysis Ensemble Statistics

We start by reviewing the general framework for deterministic formulations of the analysis step of the EnKF. This framework was given in Section 2.4.1 and is based on Tippett *et al* [23]. The ensemble update is broken into two parts. First the analysis ensemble mean is calculated using

$$\overline{\mathbf{x}^a} = \overline{\mathbf{x}^f} + \mathbf{K}_e(\mathbf{y} - \overline{\mathbf{y}^f}). \tag{6.1}$$

Then the analysis ensemble perturbation matrix is calculated using

$$\mathbf{X}'^a = \mathbf{X}'^f\mathbf{T} \tag{6.2}$$

where $\mathbf{T}$ is an $N \times N$ matrix satisfying

$$\mathbf{T}\mathbf{T}^T = \mathbf{I} - (\mathbf{Y}'^f)^T\mathbf{S}^{-1}\mathbf{Y}'^f. \tag{6.3}$$

The analysis ensemble members are formed by adding $\overline{\mathbf{x}^a}$ to the columns of $\sqrt{N-1}\mathbf{X}'^a$ in accordance with the definition (2.12) of an ensemble perturbation matrix.

It is tacitly assumed in Tippett *et al* [23] that (6.2) yields a valid analysis ensemble perturbation matrix for any choice of $\mathbf{T}$ satisfying (6.3). However, definition (2.12) implies that the mean of the columns of an ensemble perturbation matrix must be zero, and this does not necessarily follow from (6.2) and (6.3). To see this, let $\mathbf{T}$ be a particular solution of (6.3). Then a general solution is $\mathbf{TU}$ where $\mathbf{U}$ is an arbitrary $N \times N$ orthogonal matrix. The corresponding general analysis ensemble perturbation matrix is

$$\mathbf{X}'^a = \mathbf{X}'^f\mathbf{T}\mathbf{U}.$$

Now let $\overline{\mathbf{Z}}$ denote the mean of the column vectors of the matrix $\mathbf{Z}$; that is, if

$$\mathbf{Z} = \begin{pmatrix} \mathbf{z}_1 & \mathbf{z}_2 & \dots & \mathbf{z}_N \end{pmatrix}$$

where the $\mathbf{z}_i$ are column vectors, then

$$\overline{\mathbf{Z}} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{z}_i.$$

Note that $\overline{\mathbf{Z}_1\mathbf{Z}_2} = \mathbf{Z}_1\overline{\mathbf{Z}_2}$. It follows that

$$\overline{\mathbf{X}'^a} = \mathbf{X}'^f\mathbf{T}\overline{\mathbf{U}}. \tag{6.4}$$

Thus $\overline{\mathbf{X}'^a} = 0$ if and only if $\overline{\mathbf{U}}$ lies in the null space of $\mathbf{X}'^f\mathbf{T}$. The vector $\overline{\mathbf{U}}$ has length $1/\sqrt{N}$ and can be made to point in any direction by an appropriate choice of $\mathbf{U}$. Therefore, unless $\mathbf{X}'^f\mathbf{T} = 0$ (in which case the analysis ensemble collapses to a point), there will be at least some choices of $\mathbf{U}$ that give $\overline{\mathbf{X}'^a} \neq 0$ and hence an invalid analysis ensemble perturbation matrix. We shall see that the individual methods discussed in Section 2.4 differ as to whether they yield $\overline{\mathbf{X}'^a} = 0$ unconditionally.

At first glance the length $1/\sqrt{N}$ of $\overline{\mathbf{U}}$ in (6.4) appears to offer hope of proving that $\mathbf{X}'^a$ is a valid analysis ensemble perturbation matrix in the limit of large ensembles. This hope is reinforced by the observation that $\mathbf{X}'^f\mathbf{T}$ should be bounded in some sense as $N \to \infty$ on account that $\mathbf{X}'^f\mathbf{T}(\mathbf{X}'^f\mathbf{T})^T = \mathbf{P}_e^a$, which should tend to a limit $\mathbf{P}^a$ as $N \to \infty$ or at least remain bounded. However, recall that there is a factor of $1/\sqrt{N-1}$ in the definition (2.12) of an ensemble perturbation matrix and it is in fact the mean of the columns of $\sqrt{N-1}\mathbf{X}'^a$ that we require to be zero. This factor of $\sqrt{N-1}$ cancels the length of $\overline{\mathbf{U}}$ and has so far foiled the author's attempts at a proof of a result along these lines (although one may be possible).

Equation (6.2) in conjunction with the general solution of (6.3) remains a valid way of transforming a matrix square root $\mathbf{X}'^f$ of the covariance matrix $\mathbf{P}_e^f$ into a matrix square root $\mathbf{X}'^a$ of the covariance matrix $\mathbf{P}_e^a$ with $\mathbf{P}_e^f$ and $\mathbf{P}_e^a$ related as in the Kalman Filter. However, although all ensemble perturbation matrices are matrix square roots of the corresponding ensemble covariance matrix, the converse (as we have seen above) is not true.

To see the effect of constructing an analysis ensemble from $\overline{\mathbf{x}^a}$ calculated using (6.1) and an invalid analysis ensemble perturbation matrix, introduce the notation $\mathbf{x}_i'^a$ for the columns of $\sqrt{N-1}\mathbf{X}'^a$. Then the members of the analysis ensemble will be

$$\mathbf{x}_i = \overline{\mathbf{x}^a} + \mathbf{x}_i'^a.$$

The mean of this ensemble will be

$$\overline{\mathbf{x}} = \overline{\mathbf{x}^a} + \overline{\mathbf{x}'^a}.$$

But $\overline{\mathbf{x}'^a} = \sqrt{N-1}\,\overline{\mathbf{X}'^a} \neq 0$ (because we are assuming $\mathbf{X}'^a$ is invalid), and so there is a bias in the ensemble mean. Furthermore, the ensemble covariance matrix of the ensemble $\mathbf{x}_i$ is

$$\begin{aligned}
\mathbf{P}_e &= \frac{1}{N-1}\sum_{i=1}^{N}(\mathbf{x}_i - \overline{\mathbf{x}})(\mathbf{x}_i - \overline{\mathbf{x}})^T \\
&= \frac{1}{N-1}\sum_{i=1}^{N}(\mathbf{x}_i'^a - \overline{\mathbf{x}'^a})(\mathbf{x}_i'^a - \overline{\mathbf{x}'^a})^T \\
&= \frac{1}{N-1}\left(\sum_{i=1}^{N}\mathbf{x}_i'^a(\mathbf{x}_i'^a)^T - N\overline{\mathbf{x}'^a}\,\overline{\mathbf{x}'^a}^T\right) \\
&= \mathbf{P}_e^a - \frac{N}{N-1}\overline{\mathbf{x}'^a}\,\overline{\mathbf{x}'^a}^T.
\end{aligned} \tag{6.5}$$

Thus $\mathbf{P}_e \neq \mathbf{P}_e^a$ and in particular the ensemble standard deviation will be too small for any coordinate in which there is also a bias in the mean.

Analysis ensembles with biased mean or too small standard deviations or both is exactly what was observed in the experiments with the ETKF in Chapter 5. This suggests that the ETKF is one of those methods that may yield an invalid analysis ensemble perturbation matrix in at least some circumstances. This may be confirmed algebraically, as shown in Appendix D. The experiments of Chapter 5 also suggest that the statistical inconsistency becomes less significant as the ensemble size increases or the filter becomes more uncertain due to other factors (see Sections 5.3). This is an area for further investigation; the author has not yet able to prove a result along these lines.

The EAKF always yields a valid analysis ensemble perturbation matrix because it may be written in the pre-multiplier form

$$\mathbf{X}'^a = \mathbf{A}\mathbf{X}'^f$$

which implies

$$\overline{\mathbf{X}'^a} = \mathbf{A}\overline{\mathbf{X}'^f} = 0.$$

This explains why inconsistent analysis ensemble statistics were not observed with the EAKF in Chapter 5.

The direct method of Section 2.4.2 does not specify how the matrix square root $\mathbf{T}$ in (6.3) is to be found, so no general statement can be made about whether this method yields a valid analysis ensemble perturbation matrix. The serial method of Section 2.4.3 always yields a valid matrix because by (2.29)

$$\mathbf{X}'^a = \mathbf{X}'^f\mathbf{T} = \mathbf{X}'^f - \beta\mathbf{X}'^f(\mathbf{Y}'^f)^T\mathbf{Y}'^f$$

and hence

$$\overline{\mathbf{X}'^a} = \overline{\mathbf{X}'^f} - \beta\mathbf{X}'^f(\mathbf{Y}'^f)^T\overline{\mathbf{Y}'^f} = 0.$$

In summary, there is a potential flaw in deterministic formulations of the analysis step of the EnKF, leading to analysis ensembles with statistics that are inconsistent with the actual error. The EAKF and serial method are immune, but the ETKF and direct method are not. This flaw appears to have been overlooked in Tippett *et al* [23] and elsewhere in the literature.

6.2 Collapse in Number of Distinct Ensemble Members

It was conjectured in Section 5.3 on the basis of experimental results that when the ETKF assimilates an observation of m system coordinates there is a collapse in the number of distinct values of each of these coordinates in the ensemble to $m+1$. It was further conjectured that m of these values are occupied by single ensemble members whilst the remaining $N-m$ members occupy the remaining value. This will now be demonstrated algebraically in the case of an arbitrary linear observation operator $\mathbf{H}$ and an ensemble size N greater than the observation space dimension m.

Define a matrix $\mathbf{Y}'^a$ by

$$\mathbf{Y}'^a = \mathbf{H}\mathbf{X}'^a.$$

This may interpreted as an analysis observation ensemble perturbation matrix analogous to $\mathbf{Y}'^f$, but such an interpretation is not essential to what follows. For the ETKF we have (in the notation of Section 3.1)

$$\begin{aligned}\mathbf{Y}'^a &= \mathbf{H}\mathbf{X}'^f\mathbf{U}(\mathbf{I}+\mathbf{\Lambda})^{-\frac{1}{2}}\\ &= \mathbf{Y}'^f\mathbf{U}(\mathbf{I}+\mathbf{\Lambda})^{-\frac{1}{2}}\end{aligned}$$

$$
\begin{aligned}
&= \mathbf{R}^{\frac{1}{2}}\widehat{\mathbf{Y}}^f\mathbf{U}(\mathbf{I}+\mathbf{\Lambda})^{-\frac{1}{2}} \\
&= \mathbf{R}^{\frac{1}{2}}\mathbf{V}\mathbf{\Sigma}^T(\mathbf{I}+\mathbf{\Lambda})^{-\frac{1}{2}}. \qquad (6.6)
\end{aligned}
$$

Since we are assuming $N > m$, we may assume that the $N \times m$ matrix $\mathbf{\Sigma}$ has the form

$$\mathbf{\Sigma} = \begin{pmatrix} \mathbf{\Sigma}_1 \\ 0 \end{pmatrix}$$

where $\mathbf{\Sigma}_1$ is an $m \times m$ diagonal matrix and 0 stands for an $(N-m) \times m$ matrix of zeros. It follows that

$$\mathbf{\Lambda} = \mathbf{\Sigma}\mathbf{\Sigma}^T = \begin{pmatrix} \mathbf{\Lambda}_1 & 0 \\ 0 & 0 \end{pmatrix}$$

where $\mathbf{\Lambda}_1 = \mathbf{\Sigma}_1\mathbf{\Sigma}_1^T$. It may then be shown that

$$\mathbf{\Sigma}^T(\mathbf{I}+\mathbf{\Lambda})^{-\frac{1}{2}} = \begin{pmatrix} \mathbf{\Sigma}_1^T(\mathbf{I}+\mathbf{\Lambda}_1)^{-\frac{1}{2}} & 0 \end{pmatrix}$$

where 0 stands for an $m \times (N-m)$ matrix of zeros. Substituting into (6.6) gives

$$\mathbf{Y}'^a = \begin{pmatrix} \mathbf{R}^{\frac{1}{2}}\mathbf{V}\mathbf{\Sigma}_1^T(\mathbf{I}+\mathbf{\Lambda}_1)^{-\frac{1}{2}} & 0 \end{pmatrix}.$$

Thus $\mathbf{Y}'^a$ has at most m nonzero columns and at least $N-m$ zero columns.

In the case where $\mathbf{H}$ is the projection operator onto m of the system coordinates, the rows of $\mathbf{Y}'^a$ equal the rows of $\mathbf{X}'^a$ corresponding to the observed coordinates. It follows that at least $N-m$ of the analysis ensemble members will coincide in these coordinates with $\overline{\mathbf{x}^a}$ calculated from (6.1) whilst at most m members will differ. This is what was conjectured with the slight weakening that there may be a further collapse in the number of distinct values for observed coordinates due to zero columns in $\mathbf{R}^{\frac{1}{2}}\mathbf{V}\mathbf{\Sigma}_1^T(\mathbf{I}+\mathbf{\Lambda}_1)^{-\frac{1}{2}}$.

Unlike the discovery of Section 6.1, this result is a limitation of the ETKF rather than a potential flaw. It says that one should not apply the filter to ensembles with $N > m+1$. This limits the filter's usefulness for low-dimensional systems such as the swinging spring (or more precisely for systems with low-dimensional observation spaces). NWP-type applications with $N \ll m$ are not affected.

Chapter 7

Conclusions

7.1 Summary and Discussion

Chapter 2 introduced the EnKF and gave several alternative formulations of the algorithm. These alternatives may be classified as stochastic (reviewed in Evensen [8]) or semi-deterministic (reviewed in Tippett *et al* [23]). The difference between the formulations is in the analysis step; all share the same stochastic forecast step. The deterministic formulations of the analysis step all fit into the general framework described in Section 2.4.1.

In Chapter 3 two algorithms were selected for implementation: the ETKF (originally presented in Bishop *et al* [3]) and the EAKF (originally presented in Anderson [2]). The chapter described how the raw algorithms of Chapter 2 were reformulated to give algorithms that are analytically equivalent but numerically better behaved. The first main step in the reformulation is to scale observation space quantities by the inverse square root of the observation error covariance matrix $\mathbf{R}^{-\frac{1}{2}}$ at the earliest possible opportunity. This has the effect of normalising observations of possibly disparate physical quantities with different error standard deviations so that they are dimensionless with standard deviation one. This prevents information becoming lost due to rounding errors. The scaling also enables all matrices for which an eigenvalue decomposition is required to be written in the form $\mathbf{Z}\mathbf{Z}^T$. This is a precursor to the second step in the reformulation, which is to replace all such eigenvalue decompositions with an SVD of $\mathbf{Z}$. There is then no need to form $\mathbf{Z}\mathbf{Z}^T$ with the consequent loss of accuracy. Chapter 3 also showed how to order the computations in the ETKF and EAKF so as to minimise storage requirements and maximise reuse of intermediate results.

Chapter 4 introduced the two-dimensional swinging spring. As motivation for its study the chapter also briefly introduced the concept of initialisation that is of importance in NWP. It used the swinging spring to illustrate the techniques of linear and nonlinear normal mode initialisation. The method used to numerically integrate the equations of motion was described and approximately analysed to find method parameter values that give acceptable truncation error and guard against instability.

The results of experiments using an ETKF and an EAKF with observations of the swinging spring were presented in Chapter 5. The experiments revealed two features of the ETKF that were explained in Chapter 6. The first is that the filter may produce analysis ensembles with statistics that are inconsistent with the actual error, the mean being biased and the standard deviations of the coordinates too small. This was traced to a potential flaw in the general framework for deterministic formulations of the analysis step of the EnKF. This flaw appears to have been overlooked in the literature and its discovery is probably the most important thing in this dissertation. Put briefly, the general framework uses the transform $\mathbf{X}'^a = \mathbf{X}'^f\mathbf{T}$ for suitable $\mathbf{T}$ to convert the forecast ensemble perturbation matrix $\mathbf{X}'^f$ into an analysis ensemble perturbation matrix $\mathbf{X}'^a$. It is tacitly assumed that any $\mathbf{X}'^a$ resulting from this transform is a valid ensemble perturbation matrix. However, a valid ensemble perturbation matrix must have the mean of its column vectors equal to zero, and it is shown in Section 6.1 that the restrictions placed on $\mathbf{T}$ are not sufficient to ensure this in the general case. Whether or not $\mathbf{X}'^a$ is a valid ensemble perturbation matrix depends on the particular choice of $\mathbf{T}$. For the EAKF it is always valid, whilst for the ETKF it need not be.

The inconsistent analysis ensemble statistics thus produced are undesirable for a number of reasons beyond the simple fact that a biased mean tends to put the filter's supposed best estimate in the wrong place. Such a bias would not be too great a problem if it were accompanied by an increase in the size of the error estimate provided by the filter's covariance matrix. Users of the output would then be aware of the increased error, although they would remain unaware that part of the error is systematic rather than random. However, here we have a decrease in the size of the error estimate rather than an increase, and indeed equation (6.5) shows that the worse the bias, the worse the overconfidence of the error estimate.

A biased and overconfident analysis has the potential to create problems at later times in any Kalman-type filter. Such an analysis is likely to lead to a biased and overconfident forecast. The filter will then give more weight than it should to the forecast in the next analysis step and less to the observation. This will prevent the observation from properly correcting the bias in the forecast and the next analysis will be biased and overconfident as well. In extreme cases the filter may become increasingly overconfident until it is in effect a free-running forecast model diverging from the truth and taking no notice of observations.

Inconsistent ensemble statistics have been observed in formulations of the EnKF other than the ETKF. Houtekamer and Mitchell [11] present results showing problems with a stochastic EnKF and Anderson [2] discusses the issue in the context of the EAKF. The causes of the inconsistencies in these cases must be different to that described for the ETKF in Section 6.1. The authors attribute them to the use of small ensembles and to other approximations made in the course of deriving the filters. Various solutions to the problem have been proposed in the literature. Houtekamer and Mitchell [11] use a pair of ensembles with the covariance calculated from each ensemble being used to assimilate observations into the other. The justification for such an approach is discussed further in van Leeuwen [24] and Houtekamer and Mitchell [12]. Anderson [2] uses a tunable scalar covariance inflation factor.

The other feature of the ETKF revealed by the experiments only affects systems in which the observation space dimension m and the ensemble size N satisfy $N > m+1$ (thus it does not affect NWP-type systems with $N \ll m$). When m of the system coordinates are observed, then each assimilation of an observation is followed by a collapse in the number of distinct values of each of the observed coordinates in the ensemble to $m+1$. Of these values, m are occupied by single ensemble members and the remaining value is occupied by the remaining $N-m$ members. Unlike the first feature this is not really a flaw in the ETKF, but rather a limitation on the dimension of the systems to which it may be usefully applied. In particular, it is now seen not to be well-suited to experiments with low-dimensional systems such as the swinging spring.

7.2 Further Work

Three areas may be identified for further investigation: the numerical method used to integrate the swinging spring equations, initialisation techniques for the EnKF, and the inconsistent analysis ensemble statistics from semi-deterministic formulations of the EnKF.

7.2.1 Numerical Integration of the Swinging Spring Equations

It must be admitted that the stability analysis of Section 4.3 was rather crude. A more careful treatment would at least investigate the discrete system that results from applying the Runge-Kutta method to the full nonlinear system of ODEs rather than to the linearised system. Better knowledge of the stability properties may allow a larger value to be used for the `MaxStep` parameter to `ode45`. It was decided to err on the side of caution in choosing the value in Section 4.3 on account of possible differences between the nonlinear and linear systems; as a result it is `MaxStep` rather than the error estimate (4.9) that is the dominant factor in limiting the step size. Increasing step size and reducing run time is especially useful when one wishes to carry out many Monte Carlo runs such as those used to produce Table 5.2.

There is also the question of the choice of method used to integrate the equations. The MATLAB `ode45` function is a good general-purpose ODE solver, but may not be the best choice for the swinging spring. Given that the system possess motions with two distinct timescales and we are primarily interested in solutions in which the fast motions are suppressed, it could be argued that we should be using a solver designed for stiff systems; and indeed it is the fast motion timescale T_r that determined the size of `MaxStep` in Section 4.3 rather than the ten-times larger T_θ. However, this is not the whole story, especially if we are conducting experiments with ensembles in which some members have significant fast motion.

7.2.2 Initialisation and the Ensemble Kalman Filter

The original plan for this dissertation involved using the swinging spring system to investigate initialisation techniques for the EnKF. Unfortunately, investigation of the issues arising from the implementation and testing of the filters themselves did not leave time to pursue this line of enquiry. A recent study where it is pursued is Neef *et al* [21], which uses a different four-dimensional dynamical system (the extended Lorenz model) to investigate initialisation properties of a stochastic EnKF in comparison with a conventional EKF. It would be an interesting exercise to repeat the study using the swinging spring to see whether the same conclusions are reached. The EnKF could be a stochastic formulation or the EAKF, or both could be used and compared.

7.2.3 Inconsistent Analysis Ensemble Statistics

As previously mentioned, probably the most important thing in this dissertation is the discovery that some deterministic formulations of the EnKF produce invalid analysis ensemble perturbation matrices, leading to analysis ensembles with statistics that are inconsistent with the actual error. The possibility of this is shown experimentally with the ETKF in Section 5.1, proved algebraically in the context of the general framework for deterministic formulations of the analysis step in Section 6.1, and again proved algebraically with a specific example for the ETKF in Appendix D. It is also shown in Section 6.1 that the EAKF is immune to this problem.

There remains the question of how significant the statistical inconsistency is and what parameters control the degree of inconsistency. The experimental results of Section 5.1 suggest that the inconsistency decreases with increasing ensemble size, and an analytic proof of this conjecture is the first priority, either in the general case or in the specific case of the ETKF. Some of the factors aiding and frustrating such a proof are mentioned in Section 6.1.

It is noted in Section 5.3 that other factors appearing to affect the degree of inconsistency are the uncertainty of the initial conditions, the frequency of the observations, and the number of observed variables. The first two factors are probably only relevant as far as they affect the spread of the forecast ensemble that is passed to the analysis step. It is this spread as expressed through the ensemble covariance matrix and the relation of this matrix to the observation error covariance matrix that are likely to be key. Note that the number of observed variables—the third factor mentioned in Section 5.3—may be regarded as a property of the observation error covariance matrix. Results relating these factors to the degree of statistical inconsistency may be sought analytically or experimentally. In the latter case note that the problem is purely one of the analysis step and no dynamical model or forecast step is required. A forecast ensemble to pass to the analysis algorithm may be randomly generated from an assumed forecast covariance matrix and the true state. It is true that in practice one may have a forecast ensemble with inconsistent statistics, but by starting with a consistent forecast ensemble we may isolate the effect of the analysis step algorithm. An observation to pass to the analysis algorithm may likewise be randomly generated from the true state and an assumed observation error covariance matrix. By varying the two covariance matrices (and possibly other parameters as well) and assessing the effect on the analysis ensemble statistics, it may be possible to form conjectures for analytic proof.

Finally, if many Monte Carlo runs are to be performed for these or other experiments, it would be best to employ a little more finesse than was used to generate the statistics in Table 5.2. There the number 100 was almost arbitrarily decided upon as the number of runs to perform in each case. If the machinery of statistical hypothesis testing can be brought to bear on the design of the experiments, then the number of runs may be reduced to the minimum necessary to produce meaningful results and the number of cases that may be tested in a given time will be maximised.

Appendix A

Additional Operation Counts

This appendix is a supplement to Section 2.3.3. The following operation counts are based on $O(a^3)$ to invert an $a \times a$ matrix and $O(abc)$ to multiply an $a \times b$ matrix by a $b \times c$ matrix. Recall that we are considering an NWP system with $N \ll m \leq n$. Recall also that we are assuming that multiplication by $\mathbf{H}$ is cheap.

A.1 Analysis Step of KF

- $O(m^3)$ to form inverse of $\mathbf{H}\mathbf{P}^f\mathbf{H}^T + \mathbf{R}$ in formula (2.5) for $\mathbf{K}$.
- $O(m^2 n)$ to form $\mathbf{K}$ as product of this inverse and $\mathbf{P}^f\mathbf{H}^T$.
- $O(n^3)$ to form $\mathbf{P}^a = (\mathbf{I} - \mathbf{K}\mathbf{H})\mathbf{P}^f$.
- State update is negligible.
- Total $O(m^3 + m^2 n + n^3) = O(n^3)$.

A.2 Naive Implementation of Analysis Step of Stochastic EnKF

- $O(n^2 N)$ for multiplication $\mathbf{P}_e^f = \mathbf{X}'^f(\mathbf{X}'^f)^T$.

- $O(m^3)$ to form inverse of $\mathbf{HP}_e^f\mathbf{H}^T + \mathbf{R}_e$ in formula (2.15) for $\mathbf{K}_e$.
- $O(m^2 n)$ to form $\mathbf{K}_e$ as product of this inverse and $\mathbf{P}_e^f\mathbf{H}^T$.
- $O(mnN)$ to form $\mathbf{X}^a = \mathbf{X}^f + \mathbf{K}_e(\mathbf{Y} - \mathbf{HX}^f)$.
- Total $O(m^3 + m^2 n + mnN + n^2 N) = O(m^2 n + n^2 N)$.

Appendix B

The EAKF and the General Deterministic Framework

This appendix is a supplement to Section 2.4.5. It is shown there that the EAKF may be written in the post-multiplier form (2.26) with

$$\mathbf{T} = \mathbf{W}\widetilde{\mathbf{U}}(\mathbf{I} + \widetilde{\mathbf{\Lambda}})^{-\frac{1}{2}}\mathbf{W}^T.$$

This is one half of the general framework discussed in Section 2.4.1; the other half is the square root condition (2.27). For $\mathbf{T}$ defined as above it can be shown that

$$\mathbf{T}\mathbf{T}^T = \mathbf{W}\mathbf{W}^T - (\mathbf{Y}'^f)^T\mathbf{S}^{-1}\mathbf{Y}'^f.$$

Thus to conclude that (2.27) holds we must show that

$$\mathbf{W}\mathbf{W}^T = \mathbf{I}. \tag{B.1}$$

But $\mathbf{W}$ is an $N \times p$ column orthogonal matrix, so for the above to be true we must have $p = N$. Recall that p is the number of nonzero eigenvalues of $\mathbf{P}_e^f$ or equivalently the rank of $\mathbf{X}'^f$. But $\mathbf{X}'^f$ is an $n \times N$ matrix with at least one linear relation between the columns (they sum to zero). Therefore $p < N$ and (B.1) does not hold. Thus the EAKF only partially fits into the general framework.

There is an argument in Tippett *et al* [23, Section 3a] purporting to show that $\mathbf{T}$ for the EAKF is equal to $\mathbf{T}$ for the ETKF post-multiplied by an orthogonal matrix. This would show that the EAKF fully fits into the general framework. But the supposed orthogonal matrix is $\mathbf{G}^{-1}\mathbf{F}^T\mathbf{X}'^f$ (in the notation of this dissertation). This has size $p \times N$ and is thus not square and consequently not orthogonal.

Appendix C

Graphs of EAKF Experiments

This appendix is a supplement to Section 5.2. Figures C.1 to C.8 correspond to Figures 5.1 to 5.8 and result from repeating the ETKF experiments with the EAKF. They have been placed in an appendix because they merely confirm that the unexpected features of the ETKF observed in Section 5.1 are absent from the EAKF. The graphs of ensemble statistics (Figures C.2, C.4, C.6, and C.8) show no sign of inconsistency with the actual error, whilst the graphs of ensemble members (Figures C.1, C.3, C.5, and C.7) show no collapse in the number of distinct ensemble members in any coordinate or at any time.

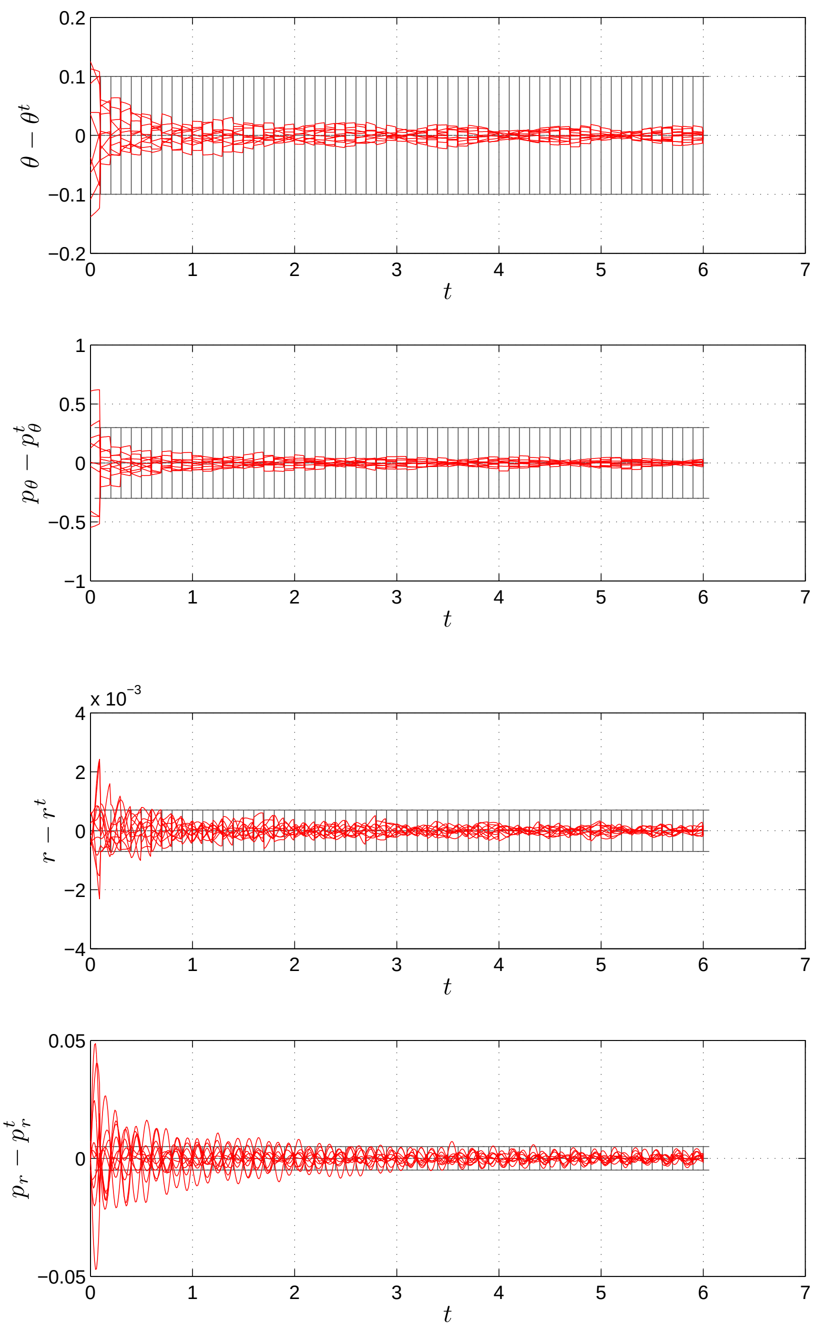

Figure C.1: EAKF, perfect observations, $N = 10$: ensemble members. This corresponds to Figure 5.1 for the ETKF.

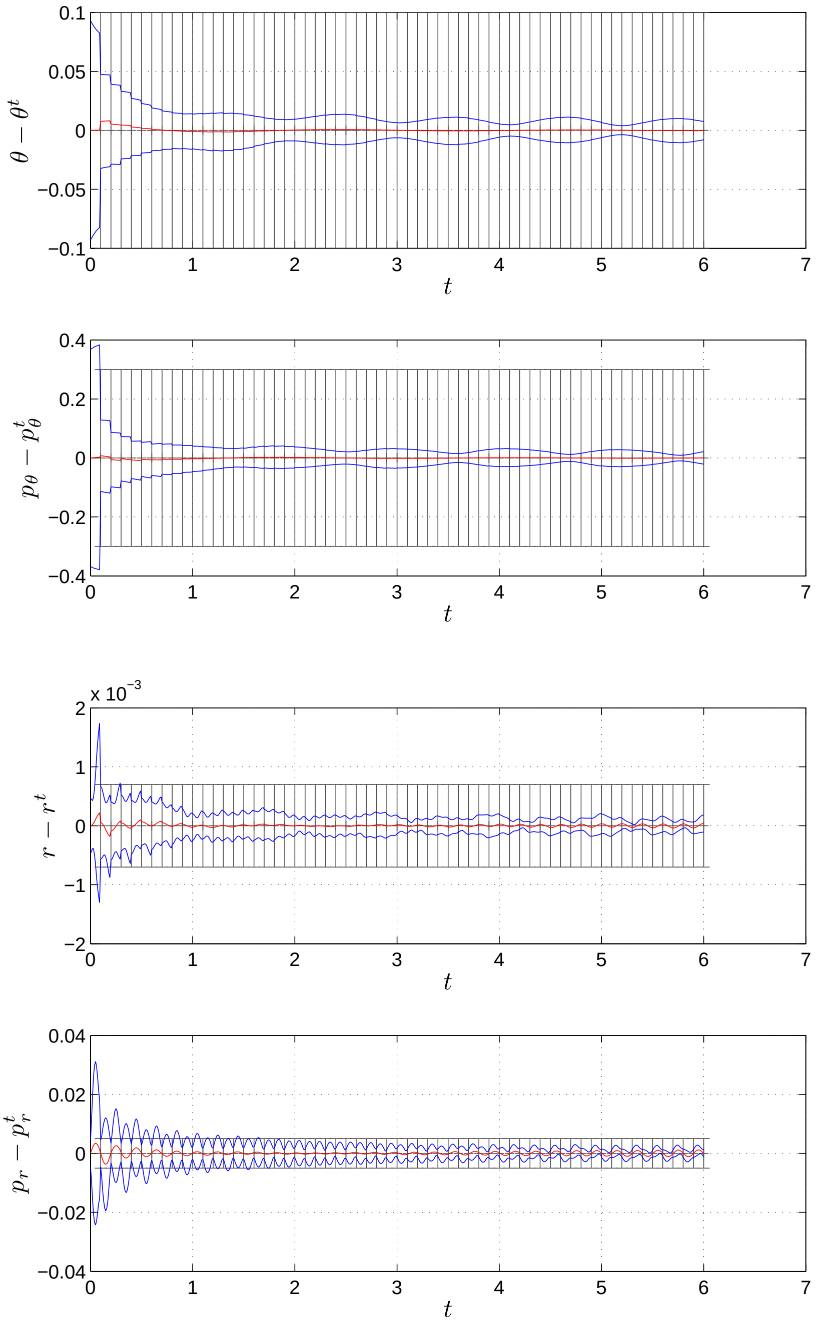

Figure C.2: EAKF, perfect observations, $N = 10$: ensemble statistics. This corresponds to Figure 5.2 for the ETKF.

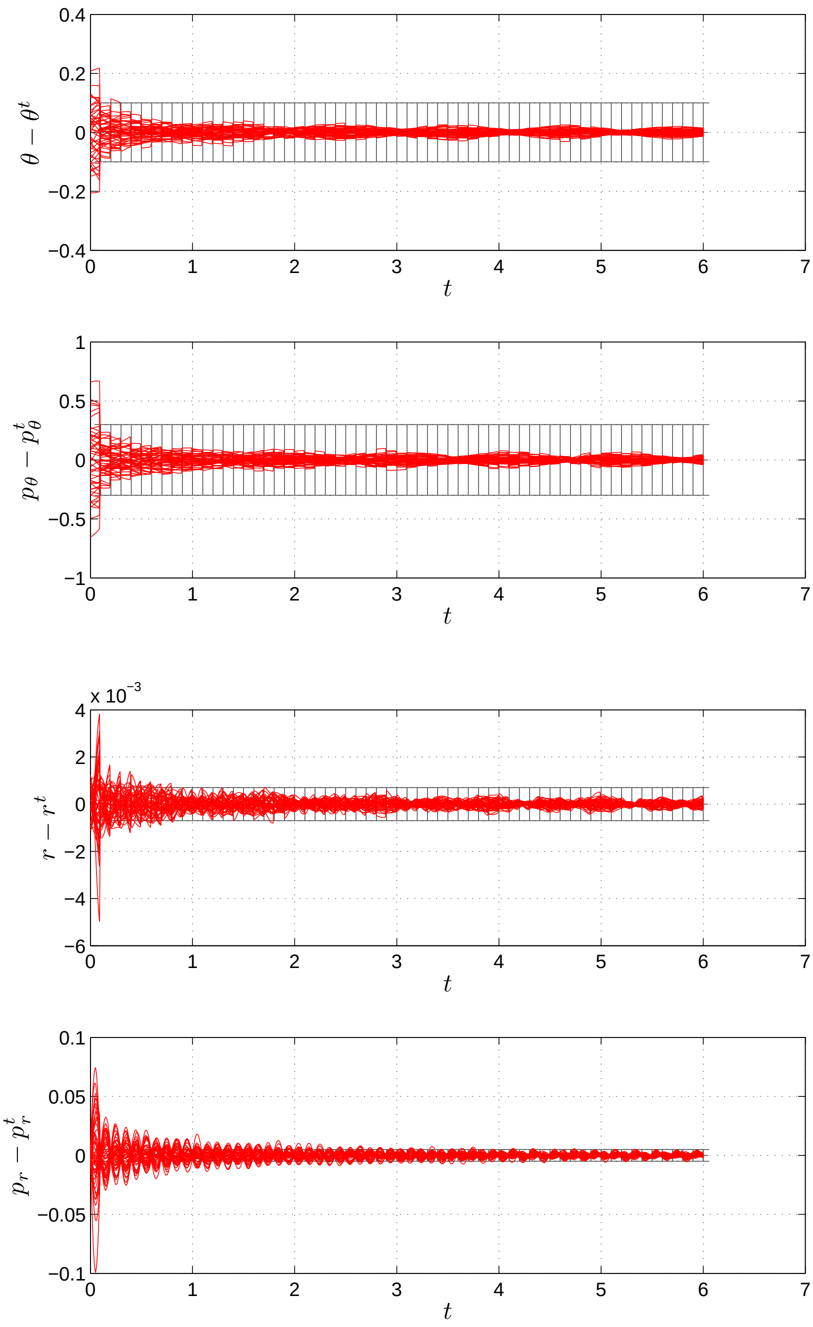

Figure C.3: EAKF, perfect observations, $N = 50$: ensemble members. This corresponds to Figure 5.3 for the ETKF.

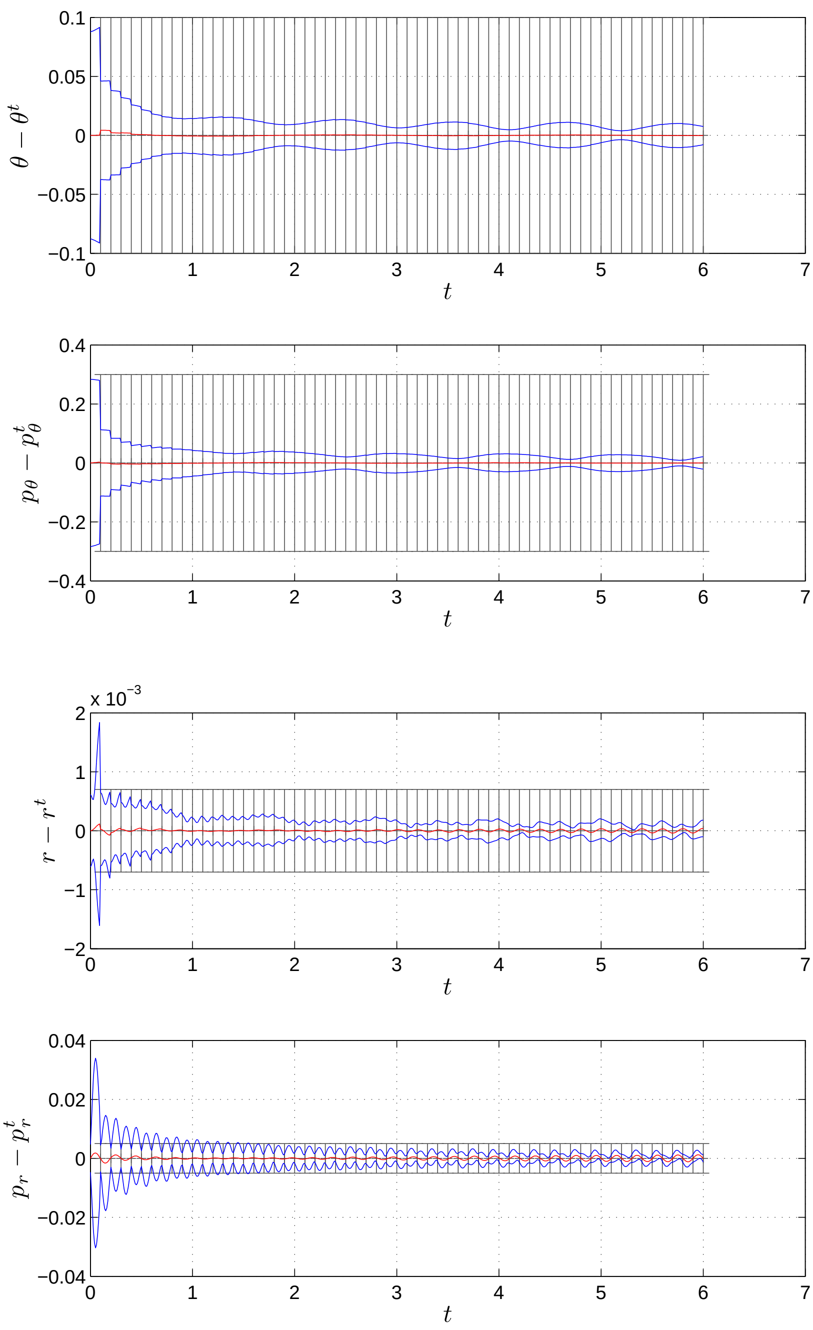

Figure C.4: EAKF, perfect observations, $N = 50$: ensemble statistics. This corresponds to Figure 5.4 for the ETKF.

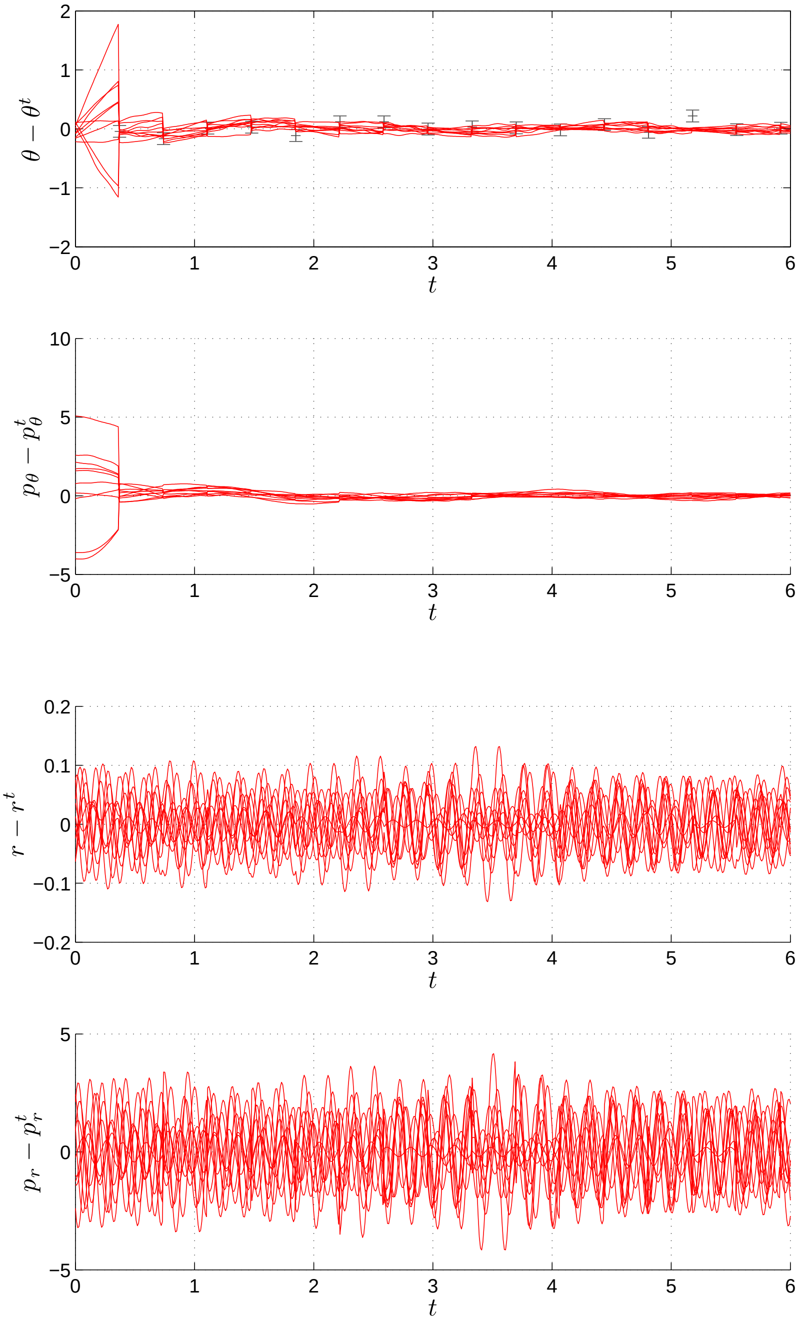

Figure C.5: EAKF, imperfect observations, $N = 10$: ensemble members. This corresponds to Figure 5.5 for the ETKF.

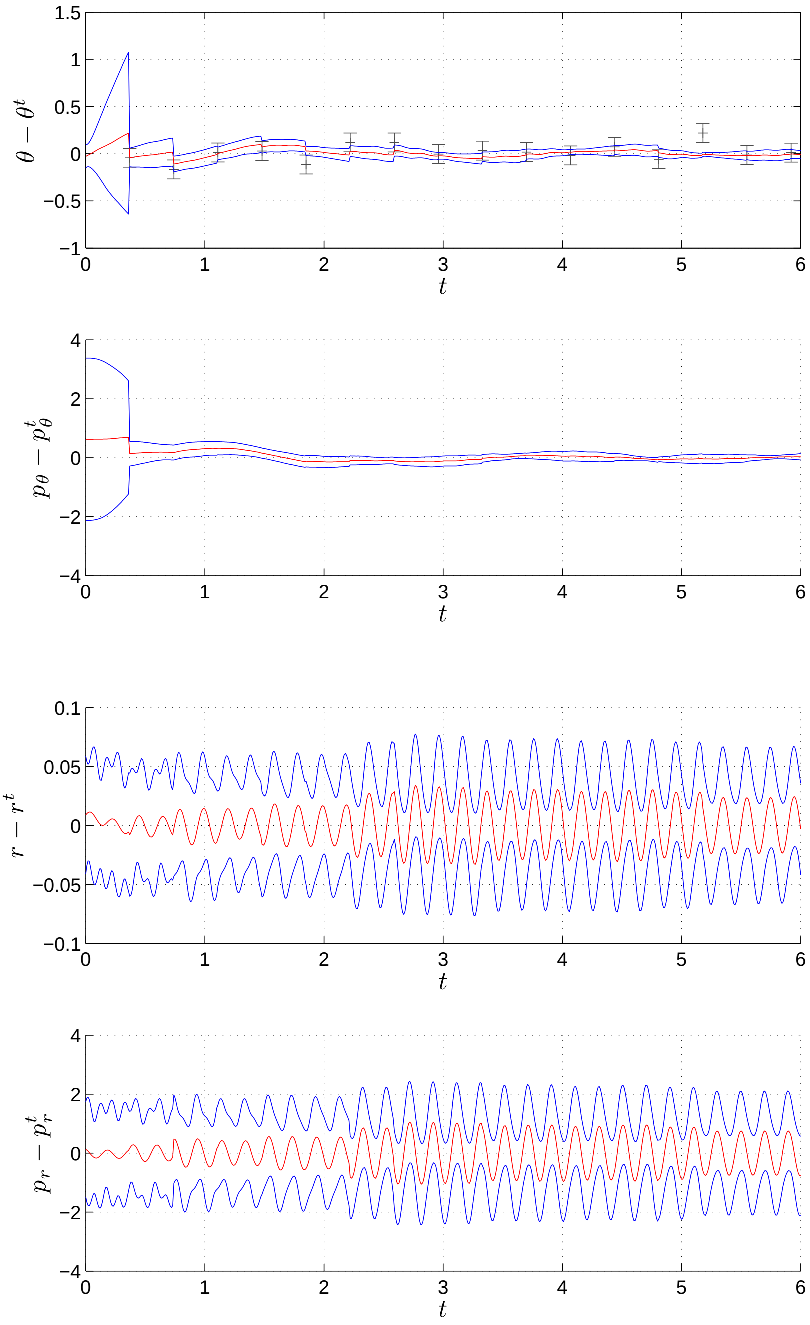

Figure C.6: EAKF, imperfect observations, $N = 10$: ensemble statistics. This corresponds to Figure 5.6 for the ETKF.

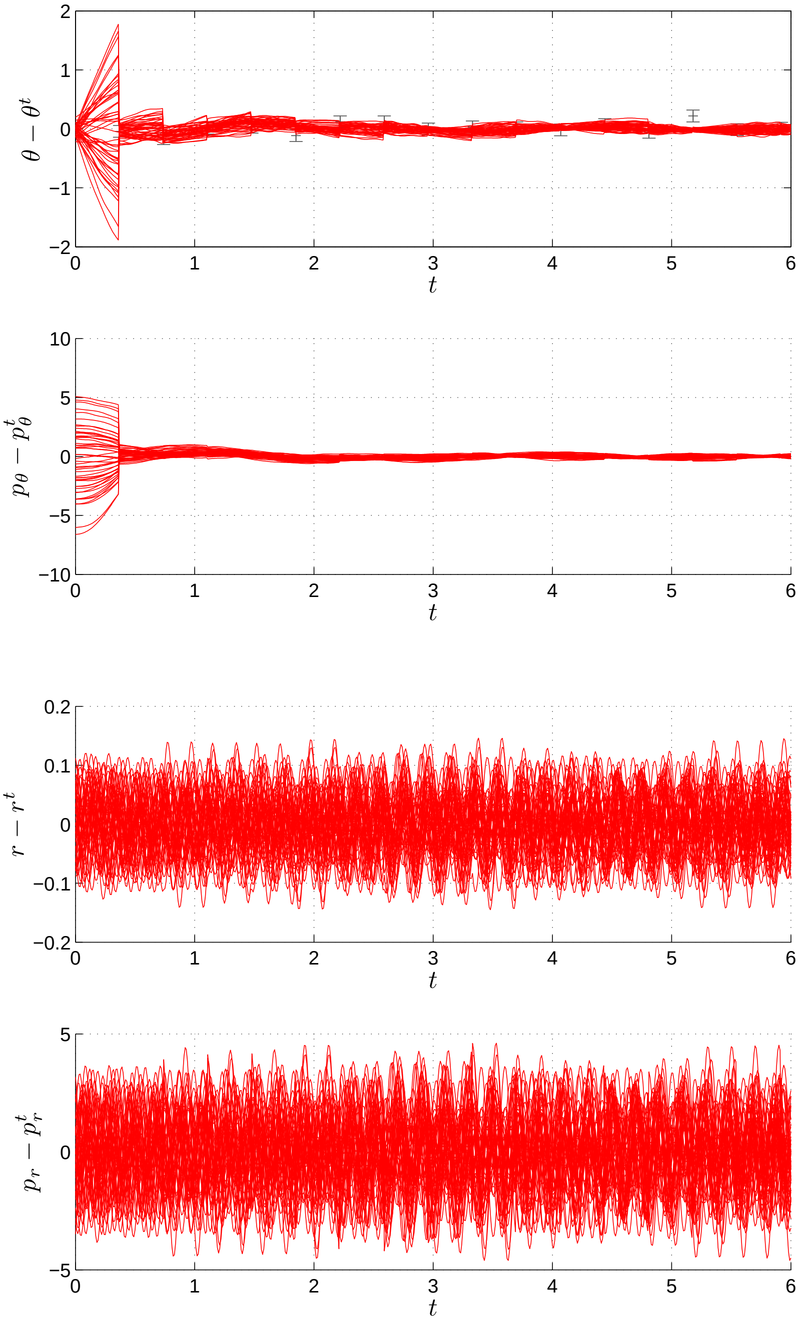

Figure C.7: EAKF, imperfect observations, $N = 50$: ensemble members. This corresponds to Figure 5.7 for the ETKF.

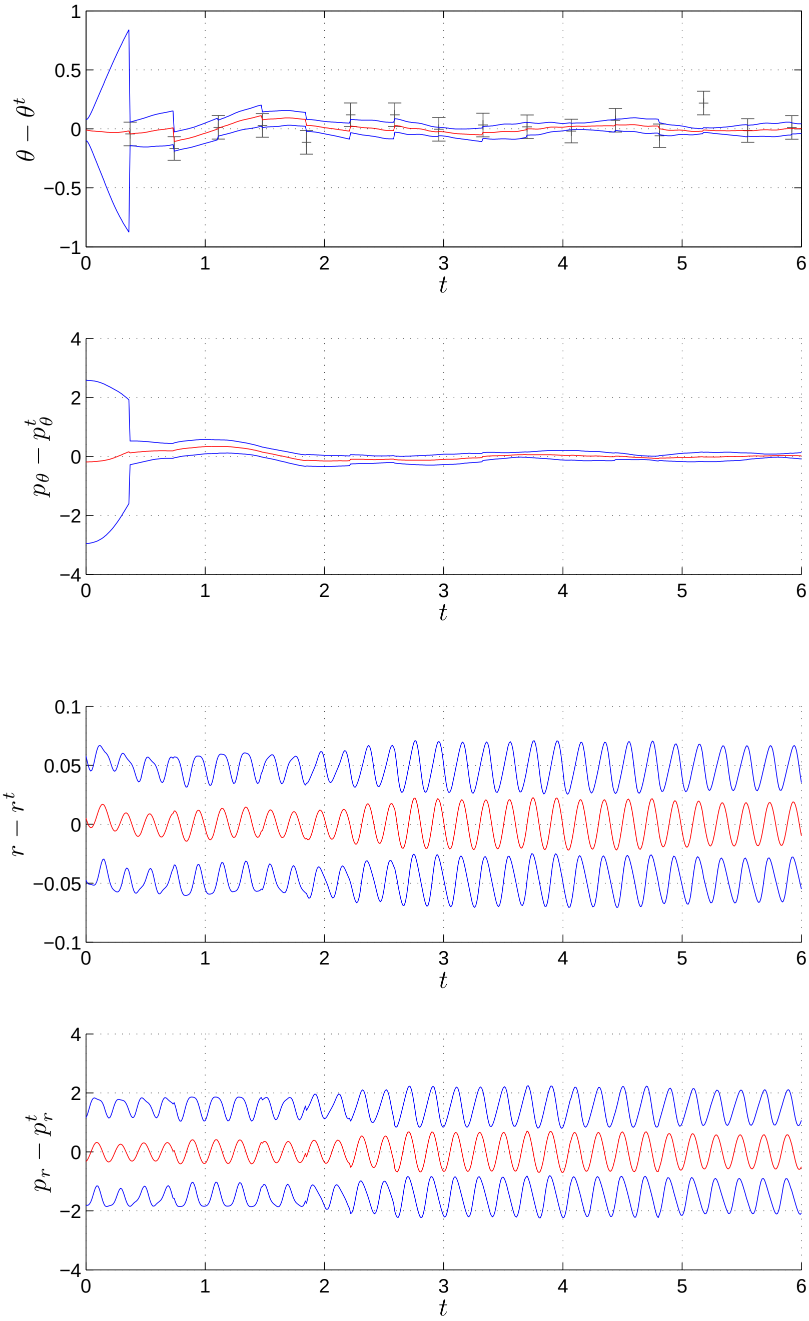

Figure C.8: EAKF, imperfect observations, $N = 50$: ensemble statistics. This corresponds to Figure 5.8 for the ETKF.

Appendix D

Example of an Invalid $\mathbf{X}'^a$ from the ETKF

This appendix is a supplement to Section 6.1. We shall construct a forecast ensemble that leads to an invalid analysis ensemble perturbation matrix (one with $\overline{\mathbf{X}'^a} \neq 0$) from the ETKF. We assume that the observation operator is linear with $\mathbf{H} = \mathbf{R} = \mathbf{I}$. Suppose that $n > N$ and that the forecast ensemble matrix has the form

$$\mathbf{X}^f = \begin{pmatrix} \mathbf{I} \\ 0 \end{pmatrix}.$$

Introduce the notation $\mathbf{1}_N$ to stand for the $N \times N$ matrix in which every element is $1/N$. Note that $\mathbf{1}_N$ and hence $\mathbf{I} - \mathbf{1}_N$ is idempotent. The forecast ensemble perturbation matrix is

$$\mathbf{X}'^f = \mathbf{X}^f - \mathbf{X}^f \mathbf{1}_N = \mathbf{X}^f(\mathbf{I} - \mathbf{1}_N) = \begin{pmatrix} \mathbf{I} - \mathbf{1}_N \\ 0 \end{pmatrix}.$$

For the ETKF we must find the eigenvalue decomposition of

$$(\mathbf{Y}'^f)^T \mathbf{R}^{-1} \mathbf{Y}'^f = (\mathbf{X}'^f)^T \mathbf{X}'^f = \mathbf{I} - \mathbf{1}_N.$$

Define column N-vectors $\mathbf{z}_i$ $(i = 1, \ldots, N-1)$ to have 1 in the first i rows, $-i$ in the next row, and 0 elsewhere. It is not difficult to show that these $\mathbf{z}_i$

are mutually orthogonal and

$$(\mathbf{I} - \mathbf{1}_N)\mathbf{z}_i = \mathbf{z}_i.$$

Define $\mathbf{z}_N$ to be the column N-vector with 1 in every row. Then $\mathbf{z}_N$ is orthogonal to the other $\mathbf{z}_i$ and

$$(\mathbf{I} - \mathbf{1}_N)\mathbf{z}_N = 0.$$

If we let $\mathbf{U}$ be the orthogonal matrix with columns equal to normalised versions of $\mathbf{z}_i$ $(i = 1, \ldots, N)$ then we have the eigenvalue decomposition

$$\mathbf{I} - \mathbf{1}_N = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T$$

where

$$\mathbf{\Lambda} = \begin{pmatrix} \mathbf{I}_{N-1} & 0 \\ 0 & 0 \end{pmatrix}$$

$\mathbf{I}_{N-1}$ being the $(N-1) \times (N-1)$ identity matrix. Let $\mathbf{U}_{N-1}$ denote the $N \times (N-1)$ matrix consisting of the first $N-1$ columns of $\mathbf{U}$. Then the ETKF update equation is

$$\begin{aligned} \mathbf{X}'^a &= \mathbf{X}'^f\mathbf{U}(\mathbf{I} + \mathbf{\Lambda})^{-\frac{1}{2}} \\ &= \begin{pmatrix} \mathbf{I} - \mathbf{1}_N \\ 0 \end{pmatrix} \mathbf{U} \begin{pmatrix} \frac{1}{\sqrt{2}}\mathbf{I}_{N-1} & 0 \\ 0 & 1 \end{pmatrix} \\ &= \begin{pmatrix} \mathbf{U}_{N-1} & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} \frac{1}{\sqrt{2}}\mathbf{I}_{N-1} & 0 \\ 0 & 1 \end{pmatrix} \\ &= \begin{pmatrix} \frac{1}{\sqrt{2}}\mathbf{U}_{N-1} & 0 \\ 0 & 0 \end{pmatrix}. \end{aligned}$$

Since the columns of $\mathbf{U}_{N-1}$ are $N-1$ orthonormal vectors, it follows that the length of $\overline{\mathbf{X}'^a}$ is $\sqrt{N-1}/\sqrt{2}N$. Therefore $\overline{\mathbf{X}'^a} \neq 0$ and $\mathbf{X}'^a$ is an invalid analysis ensemble perturbation matrix.

Bibliography

[1] E. Anderson, Z. Bai, C. Bischof, S. Blackford, J. Demmel, J. Dongarra, J. Du Croz, A. Greenbaum, S. Hammarling, A. McKenney, and D. Sorensen. *LAPACK User's Guide*. SIAM, 3rd edition, 1999. Also available as `http://www.netlib.org/lapack/lug/lapack_lug.html`.

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

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

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

[5] S. L. Dance. Issues in high resolution limited area data assimilation for quantitative precipitation forecasting. *Physica D*, 196:1–27, 2004.

[6] J. R. Dormand and P. J. Prince. A family of embedded Runge-Kutta formulae. *J. Comp. Appl. Math.*, 6:19–26, 1980.

[7] G. Evensen. Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics. *J. Geophys. Res.*, 99(C5):10143–10162, 1994.

[8] G. Evensen. The Ensemble Kalman Filter: Theoretical formulation and practical implementation. *Ocean Dynamics*, 53:343–367, 2003.

[9] A. Gelb, editor. *Applied Optimal Estimation.* The M.I.T. Press, 1974.

[10] G. H. Golub and C. F. Van Loan. *Matrix Computations.* The Johns Hopkins University Press, 3rd edition, 1996.

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

[12] P. L. Houtekamer and H. L. Mitchell. Reply. *Mon. Wea. Rev.*, 127:1378–1379, 1999.

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

[14] A. H. Jazwinski. *Stochastic Processes and Filtering Theory.* Academic Press, 1970.

[15] E. Kalnay. *Atmospheric Modeling, Data Assimilation and Predictability.* Cambridge University Press, 2003.

[16] E. Kreyszig. *Advanced Engineering Mathematics.* John Wiley & Sons, 8th edition, 1999.

[17] J. D. Lambert. *Numerical Methods for Ordinary Differential Systems.* John Wiley & Sons, 1991.

[18] A. C. Lorenc. The potential of the ensemble Kalman filter for NWP—a comparison with 4D-Var. *Q. J. R. Meteorol. Soc.*, 129:3183–3203, 2003.

[19] P. Lynch. The swinging spring: A simple model for atmospheric balance. In J. Norbury and I. Roulstone, editors, *Large-Scale Atmosphere-Ocean Dynamics II: Geometric Methods and Models*, pages 64–108. Cambridge University Press, 2002.

[20] P. Lynch. Introduction to initialization. In R. Swinbank, V. Shutyaev, and W. A. Lahoz, editors, *Data Assimilation for the Earth System*, pages 97–111. Kluwer Academic Publishers, 2003.

[21] L. Neef, S. Polavarapu, and T. Shepherd. Four-dimensional data assimilation and balanced dynamics. *J. Atmos. Sci.*, 2005. Submitted.

[22] R. Swinbank, V. Shutyaev, and W. A. Lahoz, editors. *Data Assimilation for the Earth System.* Kluwer Academic Publishers, 2003.

[23] M. K. Tippett, J. L. Andersen, C. H. Bishop, T. M. Hamil, and J. S. Whitaker. Ensemble square root filters. *Mon. Wea. Rev.*, 131:1485–1490, 2003.

[24] P. J. van Leeuwen. Comment on “Data assimilation using an ensemble Kalman filter technique”. *Mon. Wea. Rev.*, 127:1374–1377, 1999.

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

Acknowledgements

I wish to thank my supervisors Sarah Dance and Nancy Nichols for their help and for continually prodding me to make this dissertation better. I wish there had been time to incorporate all their good suggestions.

I would like to thank the Departments of Mathematics and Meteorology for creating the course and accepting me on it, Sue Davis and Peter Sweby for their help with organisational aspects, and my fellow MSc students for brandy in a pub after measuring the weather outside in the wind and rain.

Finally, I wish to thank NERC for their funding.

Declaration

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

David Livings