DAO Office Note 97-01

Office Note Series on Global Modeling and Data Assimilation

Richard B. Rood, Head
Data Assimilation Office
Goddard Space Flight Center
Greenbelt, Maryland

An Introduction to Estimation Theory *

Stephen E. Cohn

Data Assimilation Office, Goddard Laboratory for Atmospheres
Goddard Space Flight Center, Greenbelt, Maryland

Goddard Space Flight Center
Greenbelt, Maryland 20771
May 28 1997

Dedicated to the memory of Dr. James W. Pfaendtner, admired friend and esteemed colleague, who passed away on April 23, 1995.

Abstract

Despite the explosive growth of activity in the field of Earth System data assimilation over the past decade or so, there remains a substantial gap between theory and practice. The present article attempts to bridge this gap by exposing some of the central concepts of estimation theory and connecting them with current and future data assimilation approaches. Estimation theory provides a broad and natural mathematical foundation for data assimilation science.

Stochastic–dynamic modeling and stochastic observation modeling are described first. Optimality criteria for linear and nonlinear state estimation problems are then explored, leading to conditional–mean estimation procedures such as the Kalman filter and some of its generalizations, and to conditional–mode estimation procedures such as variational methods. A detailed derivation of the Kalman filter is given to illustrate the role of key probabilistic concepts and assumptions. Extensions of the Kalman filter to nonlinear observation operators and to non–Gaussian errors are then described. In a simple illustrative example, rigorous treatment of representativeness error and model error is highlighted in finite–dimensional estimation procedures for continuum dynamics and observations of the continuum state.

Contents

1 Introduction

The field of data assimilation for Earth System Science has witnessed an explosion of activity in recent years. Just a decade ago, data assimilation was regarded primarily as a means of providing initial conditions for numerical weather prediction (NWP) models. Increasingly it is being recognized that through the constant confrontation of theory (in the form of more general Earth System models) with reality (as provided by Earth System data) represented by the data assimilation process, major advances can be expected in our scientific understanding of the dynamics, variability, and interactions of all components of the Earth System over a broad range of time and space scales (Bengtsson and Shukla 1988; NAS 1991). The *desire* to carry out data assimilation is being driven largely by the rapidly increasing amount of observational data becoming available, much of it from space–borne platforms, as well as by pressing scientific and societal needs to understand the behavior of the Earth System as a whole. At the same time, the *ability* to pursue data assimilation in a physically and mathematically sound fashion is being enabled by the increasing sophistication of Earth System models and by rapid advances in computing technology.

Despite this unprecedented level of activity, Earth System data assimilation remains a young discipline, with far more open questions than solved ones. For instance, it is only intuitive that by constantly confronting Earth System models with data throughout the assimilation process, one should be able to estimate model biases and to tune free parameters, thereby offering a rigorous, data–driven means of improving our Earth System modeling capabilities. Concerted efforts along these lines have not yet begun. Even for the *static* data assimilation methods already employed operationally in NWP (e.g., Parrish and Derber 1992) there remain many open questions in covariance modeling, for example, revolving around issues such as dynamical balance, state dependence, characterization of observation errors, and identifiability of covariance parameters. Recent progress toward more *dynamic* data assimilation approaches (e.g., Andersson *et al.* 1994) raises a host of additional difficult issues, ranging from the soundness of proposed methodologies and their assumptions, to cost/benefit tradeoffs, to characterization of dynamical model errors, to long–term stability of the data assimilation process itself, and to observability of the geophysical phenomena under investigation.

The mathematical framework of *estimation theory* provides many of the tools needed to understand and approach a broad range of data assimilation problems. Estimation theory traces its origins to the efforts of astronomers some 200 years ago to understand and predict the motion of our solar system's planets, moons and asteroids. The field began to mature only in the 1960's and 1970's, along with the theory of stochastic differential equations and the development of digital computers. An excellent historical survey, along with a collection of seminal articles on various aspects and applications of estimation theory, can be found in the volume of Sorenson (1985). Applications to numerous engineering disciplines are by now standard, and textbooks in the engineering and mathematics literatures are plentiful. Among these, roughly in order of increasing level of difficulty, are Gelb (1974), Anderson and Moore (1979), Maybeck (1979), Catlin (1989), Jazwinski (1970), Casti (1985), and Omatu and Seinfeld (1989). In the more specialized Earth Sciences arena, recent texts on

data assimilation include Menke (1984), Tarantola (1987), Daley (1991), Bennett (1992), and Wunsch (1996).

In the present article, we will see that estimation theory constitutes a natural mathematical foundation for the emerging science of data assimilation. First and foremost, estimation theory offers a precise and effective way of thinking about data assimilation science. Estimation theory provides a comprehensive language, a broad conceptual framework, and a number of algorithms and approaches for addressing the open questions and scientific goals of the data assimilation enterprise.

While engineering applications of estimation theory are often small–scale and sometimes linear, Earth System data assimilation problems usually involve complex, nonlinear, large–scale models. For this reason alone, there is a substantial gap between estimation theory and Earth System data assimilation practice. The *purpose* of this article is to help bridge this gap by exposing some of the basic concepts of estimation theory to the broad Earth System data assimilation community. The focus will be on the guiding principles of estimation theory, rather than on actual computational algorithms suggested by the theory. Indeed, the sheer complexity of Earth System data assimilation problems necessitates sensible approximation: along with estimation theory, a great deal of physical reasoning and a number of techniques from modern computational mathematics and statistics must also play a large role in actually solving the fundamental problems of data assimilation in the Earth Sciences.

Estimation theory is by now a vast field, and a survey given in a single article would necessarily be perfunctory. The central ideas of estimation theory are few in number, however, so we have decided to treat just these, and to do so fairly thoroughly and rigorously, in a self–contained fashion. Further, to keep the mathematics accessible, most of the treatment will be carried out in a discrete setting. This article is therefore not a literature review; in fact, much of the material here can be located in the standard textbooks cited above. Our hope is to provide a reasonable starting point for newcomers to the fields of estimation theory and data assimilation. Readers who are already familiar with estimation theory may want to skip to § 6, where we discuss some continuum aspects of estimation theory, or perhaps to § 5.3, where nonlinearity due to non–Gaussian errors is described.

This article is organized as follows. In § 2 we introduce a generic discrete stochastic–dynamic model of the Earth System component(s) under consideration, along with a generic discrete stochastic model of the observations of the system. These two models, or variants thereof, and the probabilistic assumptions made in them, lie at the core of all estimation–theoretic approaches to data assimilation problems. Here the discrete dynamics are assumed given, and the difference between the discrete dynamics and the governing continuum dynamics is accounted for by *model error*, represented by stochastic forcing. Since the system state is assumed discrete, whereas it is the continuum state that is observed, the observation model includes a *representativeness error* term as well as a measurement error term. While in § 2 we are able to *define* model error and representativeness error rather precisely, it is not until § 6 that we show with any rigor how these error terms can actually be *treated* in data assimilation problems.

In § 3 and Appendix A we address the question of what constitutes an optimal estimate of the system state. It is shown that the *conditional mean* estimate is always unique, and has the intuitively appealing property of being also the *minimum variance* estimate for nonlinear problems as well as linear ones, independently of the nature of the probability densities of the various error sources. The *conditional mode* (*maximum a posteriori*) estimate, on the other hand, is generally not unique, but the set of conditional modes may yield a great deal more physical information about the system state than the conditional mean estimate alone. In § 3 the three basic categories of state estimation problems are also introduced: the filtering, prediction and smoothing problems. It is shown that the *four–dimensional variational* methods now under active study at several institutions represent approximate conditional mode estimation algorithms for the fixed–interval smoothing problem.

Section 4 gives a detailed derivation of the *Kalman filter*, which provides the conditional mean estimate for linear filtering problems with Gaussian–distributed errors. The purpose of this derivation is to illustrate the role of the various assumptions on which the Kalman filter is based. For most Earth System data assimilation problems, the standard Kalman filter can only be considered as a prototype algorithm in view of the many assumptions involved and in view of its computational requirements; it is certainly not an end unto itself. To progress toward less restrictive estimation procedures, it is important to understand the Kalman filter first. Following the derivation of the Kalman filter, we discuss a number of its important properties and some of its simple generalizations.

Two nonlinear generalizations of the Kalman filter are described in § 5. In the first, it is assumed that the observations are related *nonlinearly* to the state variables. After describing the practical difficulties associated with obtaining the conditional mean estimate in this case, we develop a standard approximate method of treating observation nonlinearity, known as the *locally iterated extended Kalman filter* (EKF). In the locally iterated EKF, the conditional mean is approximated by a conditional mode at observation times. It is shown that the resulting algorithm for processing the observations is algebraically equivalent to nonlinear *three–dimensional variational* analysis algorithms, which are therefore approximate conditional mode estimation algorithms. While three– and four–dimensional variational algorithms are often derived without reference to probabilistic concepts, we see that, like the (extended) Kalman filter, they rely on assumptions of Gaussian–distributed errors with mean zero.

In the second generalization, we describe by example an exact, but nonlinear and non–iterative, conditional mean estimation procedure in case the measurement errors are not Gaussian–distributed: lognormally–distributed errors are considered instead. Such errors may arise from measurements and dynamical models of nonnegative quantities, such as the mixing ratio of atmospheric trace constituents. Relationships between the multivariate Gaussian (normal) and lognormal probability densities appear in Appendix B.

In § 6 we give a simple example illustrating a number of continuum aspects of estimation theory. The governing continuum dynamics consists of the unforced scalar linear advection equation with an unknown Gaussian and statistically homogeneous initial state. Observations taken at various instants of time are arbitrary bounded linear functionals of

the continuum state, with Gaussian state–independent measurement error. It is shown first that while an exact conditional–mean filter algorithm can be written down for this problem, it cannot be implemented exactly on a computer. Following this demonstration, rather than assuming that a discretization of the continuum dynamics is already given, as in § 2, we first *define* a finite–dimensional conditional–mean filtering problem to be solved for the continuum system, and only then employ this definition to develop an appropriate discretization and an overall filter algorithm. By proceeding in this manner, it turns out for this simple example that model error does not arise, and that representativeness error can be treated exactly. The complete filter algorithm is also exact and can be implemented exactly on a computer. Supporting results, and for completeness a treatment of nonlinear dynamics, are contained in Appendix C.

By drawing upon the salient features of this simple example, we discuss finally some of the problems and prospects in accounting for model error and representativeness error in more realistic Earth System data assimilation problems. In particular it is seen that climatology, defined appropriately, may play an important role in accounting for representativeness error.

2 Discrete stochastic–dynamic and stochastic observation models

2.1 Discrete stochastic–dynamic model

The Earth System component (or components) of interest will generally be described by a system of nonlinear partial differential equations (PDEs). Denote by $\mathbf{w}_k$ the vector of prognostic (state) variables of the PDE system at time t_k. The elements of the vector $\mathbf{w}_k$ are functions, and $\mathbf{w}_k$ is assumed to belong to some function space $\mathcal{B}$. Assuming the governing PDEs to be well–posed in the sense of Hadamard (e.g., Courant and Hilbert 1962, Chapter III, § 6), there is a unique solution operator, or *propagator* $\mathbf{g}$ that yields the solution $\mathbf{w}_k$ given the solution $\mathbf{w}_{k-1}$ at an earlier time t_{k-1}:

$$\mathbf{w}_k \;=\; \mathbf{g}(\mathbf{w}_{k-1}) \;, \tag{2.1}$$

for $k = 1, 2, 3, \ldots$. While this system could be stochastically forced, for example through uncertain boundary conditions, here we neglect stochastic forcing to simplify the presentation. The system could also be internally forced by stochastic free parameters to be estimated during the course of data assimilation, for example in physical parameterizations. While a large body of literature is concerned with parameter estimation (e.g., Maybeck 1979, Chapter 10; Sorenson 1980; Caglayan and Lancraft 1983; Daley 1995; Ghil 1997), to confine the discussion we do not consider this important realm here. Parameters and forcing are considered fixed, so that the propagator $\mathbf{g}$ is deterministic. Explicit time dependence of $\mathbf{g}$ is also suppressed for notational convenience, as is dependence upon the time interval $t_k - t_{k-1}$. Although the system of PDEs itself from which we have started may be only approximate (e.g., Phillips 1966), we take (2.1) to be a correct and complete representation

of reality. This assumption, along with the assumption that $\mathbf{g}$ is deterministic, is revisited in § 6.4.

We know neither the continuum state $\mathbf{w}_k$, in particular the initial condition $\mathbf{w}_0$, nor even the operator $\mathbf{g}$. However, residing on our computer will be an approximate, discretized version of (2.1):

$$\mathbf{w}_k^d \;=\; \mathbf{f}\big(\mathbf{w}_{k-1}^d\big)\;, \tag{2.2}$$

for $k=1,\,2,\,3,\,\ldots$, the superscript d denoting " discrete". Here $\mathbf{w}_k^d$ is an n–vector approximating, with some error, the continuum state $\mathbf{w}_k$ whose evolution is given by (2.1), and $\mathbf{f}$ is the *discrete propagator*.

Next we need to define the *true state* to be estimated on the basis of observations available at times $t_1,\,t_2,\,t_3,\,\ldots$, along with an evolution equation for it. Although the continuum state given by (2.1) would appear to be the most natural candidate, we will define the true state to be a representation of the continuum state on a discrete space, so that it can be compared directly with the approximate discrete state given by (2.2); this approach will be elaborated upon and exploited more fully in § 6. To this end, define a linear operator $\mathbf{\Pi}$ from $\mathcal{B}$ to an n–dimensional function space $\mathcal{B}^n$ in a manner appropriate for the discretization (2.2). For example, if the elements of the n–vector $\mathbf{w}_k^d$ are supposed to represent averages over grid volumes, then $\mathcal{B}^n$ would consist of piecewise constant functions and the n–vector $\mathbf{\Pi}\mathbf{w}_k$ would consist of the averages of $\mathbf{w}_k$ over grid volumes. In fact, we define the (discrete) *true state* $\mathbf{w}_k^t$ as

$$\mathbf{w}_k^t \;\equiv\; \mathbf{\Pi}\mathbf{w}_k\;, \tag{2.3}$$

for $k=0,\,1,\,2,\,\ldots$; this is the representation on $\mathcal{B}^n$ of the continuum state $\mathbf{w}_k$ (cf. Cohn and Dee 1988, § 2). We remark that $\mathbf{\Pi}$ should be a projection operator, that is, $\mathbf{\Pi}^2=\mathbf{\Pi}$, but we do not require $\mathbf{\Pi}$ to be a projection operator until § 6.

The true state $\mathbf{w}_k^t$ is still unknown, since $\mathbf{w}_k$ is unknown. However, operating with $\mathbf{\Pi}$ on both sides of (2.1) gives a discrete evolution equation for $\mathbf{w}_k^t$,

$$\mathbf{w}_k^t \;=\; \mathbf{f}\big(\mathbf{w}_{k-1}^t\big) + \boldsymbol{\delta}_{k-1}^t\;, \tag{2.4}$$

where

$$\boldsymbol{\delta}_k^t \;\equiv\; \mathbf{\Pi}\mathbf{g}(\mathbf{w}_k) - \mathbf{f}(\mathbf{\Pi}\mathbf{w}_k)\;. \tag{2.5}$$

The operator $\mathbf{f}$ in (2.4) is the discrete propagator, to which we have access. The forcing term $\boldsymbol{\delta}_{k-1}^t$ is the *model error* from time t_{k-1} to time t_k. Observe that the model error defined in (2.5) is generally (continuum) state–dependent, even if the operators $\mathbf{f}$ and $\mathbf{g}$ are linear. This state dependence, as well as the dependence upon the unknown continuum propagator $\mathbf{g}$, renders the model error both unknown and unknowable from a deterministic viewpoint. However, it should be in some sense small, provided that $\mathbf{f}$ approximates $\mathbf{g}$ well. For these reasons it is appropriate, and in any case one has little choice other than, to represent this model error as a stochastic perturbation to (2.4). Here we shall simply

assume the existence of such a representation with known bias and covariance (cf. Derber 1989; Leith 1990; Daley 1992c; Bennett *et al.* 1993; Dee 1995; Dee and da Silva 1997). This will allow us to expose the probabilistic underpinnings of estimation theory in §§ 3 to 5 in a relatively simple manner. Further discussion of stochastic representation of model error is deferred to § 6.

Thus we write

$$\boldsymbol{\delta}_k^t \;=\; \mathbf{G}_k\big(\mathbf{w}_k^t\big)\,\boldsymbol{\varepsilon}_k^t\,, \tag{2.6}$$

where $\mathbf{G}_k$ is an $n \times m$ matrix depending on $\mathbf{w}_k^t$, reflecting the state dependence of model error, and $\boldsymbol{\varepsilon}_k^t$ is an m–vector stochastic process with mean

$$\widehat{\boldsymbol{\varepsilon}}_k^t \;\equiv\; \big\langle \boldsymbol{\varepsilon}_k^t \big\rangle \tag{2.7}$$

and covariance matrix

$$\mathbf{Q}_k \;\equiv\; \Big\langle \big(\boldsymbol{\varepsilon}_k^t - \widehat{\boldsymbol{\varepsilon}}_k^t\big)\big(\boldsymbol{\varepsilon}_k^t - \widehat{\boldsymbol{\varepsilon}}_k^t\big)^T \Big\rangle\,, \tag{2.8}$$

$\langle\ \rangle$ denoting the expectation operator (see Appendix A). Here $\mathbf{G}_k$, $\widehat{\boldsymbol{\varepsilon}}_k^t$ and $\mathbf{Q}_k$ are all assumed known. Substituting (2.6) into (2.4) gives

$$\mathbf{w}_k^t \;=\; \mathbf{f}\big(\mathbf{w}_{k-1}^t\big) + \mathbf{G}_{k-1}\big(\mathbf{w}_{k-1}^t\big)\,\boldsymbol{\varepsilon}_{k-1}^t\,, \tag{2.9}$$

our stochastic–dynamic model for the evolution of the discrete true state $\mathbf{w}_k^t$. Since $\boldsymbol{\varepsilon}_k^t$ is a stochastic process, so is $\mathbf{w}_k^t$. The initial condition $\mathbf{w}_0^t$ for (2.9) may also be stochastic.

2.2 Discrete stochastic observation model

It remains to formulate a stochastic model of the observed data, on the basis of which the true state $\mathbf{w}_k^t$ is to be estimated. Suppose that at times t_k, $k=1, 2, 3, \ldots$, a number p_k of observations are available and placed into a p_k–vector $\mathbf{w}_k^o$. Since these are observations of the continuum state $\mathbf{w}_k$, contaminated by some error, we write

$$\mathbf{w}_k^o \;=\; \mathbf{h}_k^c(\mathbf{w}_k) + \boldsymbol{\varepsilon}_k^m\,, \tag{2.10}$$

where $\mathbf{h}_k^c$ is the *continuum forward observation operator* from $\mathcal{B}$ to $\mathbb{R}^{p_k}$ and $\boldsymbol{\varepsilon}_k^m$ is the *measurement error*. The latter is considered stochastic, and independent of $\mathbf{w}_k$, with known mean $\widehat{\boldsymbol{\varepsilon}}_k^m$,

$$\widehat{\boldsymbol{\varepsilon}}_k^m \;\equiv\; \big\langle \boldsymbol{\varepsilon}_k^m \big\rangle\,, \tag{2.11}$$

which is the measurement error *bias*, and known covariance matrix $\mathbf{R}_k$,

$$\mathbf{R}_k \;\equiv\; \Big\langle (\boldsymbol{\varepsilon}_k^m - \widehat{\boldsymbol{\varepsilon}}_k^m)(\boldsymbol{\varepsilon}_k^m - \widehat{\boldsymbol{\varepsilon}}_k^m)^T \Big\rangle\,. \tag{2.12}$$

While *additive* measurement error is assumed in (2.10), *multiplicative* error is considered in § 5.3.

The observation operator is *linear* for state variables that are observed directly, for example by radiosondes. For most remotely–sensed data, the raw observations (such as radiances or radar backscatter) are *nonlinear* functions of the continuum state, and the corresponding elements of $\mathbf{h}_k^c(\mathbf{w}_k)$ would involve radiative transfer calculations, for example, which are integrals of nonlinear functions of $\mathbf{w}_k$. In either case, the observation operator will usually depend on a number of parameters. As was the case for the propagator, such parameters can be estimated along with the state, although we do not consider this possibility here. Instead we consider parameters to be fixed and therefore $\mathbf{h}_k^c$ to be deterministic, the so–called perfect forward model assumption.

Now the stochastic–dynamic model (2.9) was formulated in terms of the discrete true state $\mathbf{w}_k^t = \mathbf{\Pi} \mathbf{w}_k$, so we will need to introduce a *discrete forward observation operator* $\mathbf{h}_k$ that acts on $\mathbf{w}_k^t$ rather than $\mathbf{w}_k$, and rewrite the observation model (2.10) as

$$\mathbf{w}_k^o \;=\; \mathbf{h}_k\left(\mathbf{w}_k^t\right) + \boldsymbol{\varepsilon}_k^o \;, \tag{2.13}$$

where

$$\boldsymbol{\varepsilon}_k^o \;\equiv\; \boldsymbol{\varepsilon}_k^r + \boldsymbol{\varepsilon}_k^m \tag{2.14}$$

is the total *observation error*, and

$$\boldsymbol{\varepsilon}_k^r \;=\; \boldsymbol{\varepsilon}_k^r(\mathbf{w}_k) \;\equiv\; \mathbf{h}_k^c(\mathbf{w}_k) - \mathbf{h}_k(\mathbf{\Pi}\mathbf{w}_k) \tag{2.15}$$

is the *error of representativeness* (cf. Lorenc 1986), which is (continuum) state–dependent, even for *linear* forward observation operators. Note that the representativeness error (2.15) has the same form as the model error (2.5), so the problem of stochastic modeling of the representativeness error is in a sense equivalent to that of modeling the model error in the present discrete formulation.

To understand better the nature of representativeness error, let us write (2.15) as the sum of two terms,

$$\boldsymbol{\varepsilon}_k^r \;=\; \boldsymbol{\varepsilon}_k^{'} + \boldsymbol{\varepsilon}_k^{''} \;, \tag{2.16}$$

where

$$\boldsymbol{\varepsilon}_k^{'} \;\equiv\; \mathbf{h}_k^c(\mathbf{w}_k) - \mathbf{h}_k^c(\mathbf{\Pi}\mathbf{w}_k) \;, \tag{2.17}$$

$$\boldsymbol{\varepsilon}_k^{''} \;\equiv\; \mathbf{h}_k^c(\mathbf{\Pi}\mathbf{w}_k) - \mathbf{h}_k(\mathbf{\Pi}\mathbf{w}_k) \;, \tag{2.18}$$

and we assume that $\mathcal{B}^n$ is a subspace of $\mathcal{B}$, so that the expression $\mathbf{h}_k^c(\mathbf{\Pi}\mathbf{w}_k)$ is well–defined.* Now $\boldsymbol{\varepsilon}_k^{''}$ can be made as small as one pleases by employing high–order accurate integration and interpolation formulas in $\mathbf{h}_k$ to approximate $\mathbf{h}_k^c$, at least in principle, since both operators in (2.18) act on the same *discrete* true state $\mathbf{w}_k^t = \mathbf{\Pi}\mathbf{w}_k$. On the other hand, $\boldsymbol{\varepsilon}_k^{'}$ depends on the small–scale variability of $\mathbf{w}_k$. For instance, to a linear approximation (2.17) may be written as

$$\boldsymbol{\varepsilon}_k^{'} \;=\; \mathbf{H}_k^c(\mathbf{I} - \mathbf{\Pi})\mathbf{w}_k \;, \tag{2.19}$$

*A similar decomposition of the model error (2.5) has been discussed by Ménard (1994, Appendix A).

where

$$\mathbf{H}_k^c \equiv \left.\frac{\partial \mathbf{h}_k^c(\mathbf{w})}{\partial \mathbf{w}}\right|_{\mathbf{w}=\mathbf{\Pi}\mathbf{w}_k} ; \tag{2.20}$$

$\mathbf{H}_k^c$ in (2.19) operates on the *unresolved* portion $(\mathbf{I}-\mathbf{\Pi})\mathbf{w}_k$ of the continuum state $\mathbf{w}_k$. For highly variable fields such as winds, moisture and trace constituents, it is apparent that this portion of the representativeness error could easily dominate the measurement error itself.

A complete discrete theory should therefore include an adequate (stochastic) model of the representativeness error. Such a model is developed in § 6 for a specific example. Rather than attempting any degree of generality here, in §§ 3 to 5 we shall simply ignore representativeness error: our observation model is (2.13) with $\boldsymbol{\varepsilon}_k^r = \mathbf{0}$ in (2.14) and the first two moments of $\boldsymbol{\varepsilon}_k^o = \boldsymbol{\varepsilon}_k^m$ given in (2.11) and (2.12). While this is not at all realistic (and not recommended for real problems!), it shall serve our purpose of introducing the fundamentals of estimation theory in a fairly simple context.

3 Conditional mean and conditional mode estimation

3.1 Introduction

Since the state $\mathbf{w}_k^t$ that we would like to estimate is given by the stochastic–dynamic model (2.9), it has a probability distribution function. We will assume all distribution functions encountered here to be differentiable, so in fact $\mathbf{w}_k^t$ has a probability *density* function $p(\mathbf{w}_k^t)$. This is a function of n variables.

Suppose for the moment that no observations are available. If $\boldsymbol{\varepsilon}_{k-1}^t$ in (2.9) were Gaussian (see § 4 for the definition of the multivariate Gaussian density) and white in time, then the evolution of $p(\mathbf{w}_k^t)$, had we discretized only space and not time in the formulation of (2.9), would be governed by the Fokker–Planck (forward Kolmogorov) equation (cf. Epstein 1969).† This is a PDE in n independent ("spatial") variables plus time, with initial condition $p(\mathbf{w}_0^t)$. In the absence of model error, this equation simplifies to the Liouville equation, which has been studied by Ehrendorfer (1994a, b). For the size n typical of Earth Science applications, for example $n \cong 10^6 - 10^7$ in numerical weather prediction, these equations cannot be solved directly in general. If they could, however, from the resulting knowledge of $p(\mathbf{w}_k^t)$ one could then in principle calculate directly such important statistics as the ensemble mean $\langle \mathbf{w}_k^t \rangle$ and the ensemble covariance matrix, by explicit integration in n dimensions. Since this appears not to be possible, Monte Carlo and related methods have recently been explored for ensemble mean prediction (e.g., Toth and Kalnay 1993, and references therein). In practice, furthermore, generally one does not know the initial density $p(\mathbf{w}_0^t)$, nor the function $\mathbf{G}_k(\mathbf{w}_k^t)$ in (2.9), nor the density of the error process $\boldsymbol{\varepsilon}_k^t$.

In any case we see that, though unavailable, it is the entire probability density function $p(\mathbf{w}_k^t)$ that constitutes the "complete solution" of the prediction problem. In the same way,

†An analogous equation exists for the discrete–time case; for example see Jazwinski (1970, § 6.6).

it is the *conditional* probability density $p\left(\mathbf{w}_k^t \middle| \mathcal{W}_\ell^o\right)$ that constitutes the complete solution for various data assimilation problems. (See Appendix A for a brief discussion of conditional densities and expectations). Here by $\mathcal{W}_\ell^o$ we denote the set of *realizations* of all observations (2.13) available up to some time t_ℓ:

$$\mathcal{W}_\ell^o \equiv \{\mathbf{w}_1^o, \mathbf{w}_2^o, \ldots, \mathbf{w}_\ell^o\} . \tag{3.1}$$

The density $p\left(\mathbf{w}_k^t \middle| \mathcal{W}_k^o\right)$ yields the solution of the *filtering* problem at times t_k, $k=1, 2, \ldots$, while the density $p\left(\mathbf{w}_k^t \middle| \mathcal{W}_{k+L}^o\right)$ with L fixed yields the solution of the *fixed–lag smoothing* problem at times t_k (e.g., Cohn *et al.* 1994, Ménard and Daley 1996). The density $p\left(\mathbf{w}_k^t \middle| \mathcal{W}_\ell^o\right)$ for ℓ fixed and $k=\ell+1, \ell+2, \ldots$, gives the solution of the *prediction* problem when there are observed data available from time t_1 to time t_ℓ. See the texts on estimation theory cited in the Introduction for further discussion of the filtering, smoothing, and prediction problems, and Ghil (1997) for a review of applications to Earth System data assimilation.

The conditional densities $p\left(\mathbf{w}_k^t \middle| \mathcal{W}_\ell^o\right)$, like the unconditional ones $p\left(\mathbf{w}_k^t\right)$, are functions of a large number of variables. Unlike unconditional densities, conditional densities are *random* functions, because they depend on the observations. For both reasons, in large Earth Science data assimilation problems, it is not possible to calculate these densities explicitly. On the other hand, algorithms for calculating the evolution of certain statistics of these densities are available, as discussed in §§ 4–6, under a number of simplifying assumptions. The first question, though, is which statistics do we want to evolve?

Two possibilities suggest themselves immediately: the conditional mean $\left\langle \mathbf{w}_k^t \middle| \mathcal{W}_\ell^o \right\rangle$, and the conditional mode, which we denote by $\mathbf{m}\left(\mathbf{w}_k^t \middle| \mathcal{W}_\ell^o\right)$. Both have obvious intuitive appeal. Both are also random n–vectors, since they depend upon realizations of the observations. The conditional mean is the "average" value of the conditional density (see Appendix A). It also has the important theoretically and intuitively appealing property of being the *minimum variance* estimate in most data assimilation problems, including the filtering, smoothing and prediction problems described above. We now explain this property, then return to the subject of conditional mode estimation.

3.2 Conditional mean estimation

Let $\mathbf{w}_k^e$ be an n–vector which is an *estimate* of the n–vector true state $\mathbf{w}_k^t$. Assume that $\mathbf{w}_k^e$ is a function of the available observed data $\mathcal{W}_\ell^o$, and define the *estimation error*

$$\boldsymbol{\varepsilon}_k = \mathbf{w}_k^t - \mathbf{w}_k^e . \tag{3.2}$$

Now let $\mathbf{S}$ be an arbitrary (but deterministic) $n \times n$ symmetric positive definite matrix, and define the quadratic *loss function* $L(\boldsymbol{\varepsilon}_k)$,

$$L(\boldsymbol{\varepsilon}_k) \equiv \boldsymbol{\varepsilon}_k^T \mathbf{S}\, \boldsymbol{\varepsilon}_k . \tag{3.3}$$

Note that L is a scalar, and that it is also a random variable, since $\boldsymbol{\varepsilon}_k$ is stochastic. For an appropriate choice of the dependent variables of the numerical model (2.2), the matrix $\mathbf{S}$

might be defined in such a way that L represents a discrete version of the total energy of the estimation error, for example.

An appealing way to define the estimate $\mathbf{w}_k^e$, then, would be to define it to be the vector that minimizes the *expected* loss $\langle L(\boldsymbol{\varepsilon}_k)\rangle$, for example the expected total energy, or more generally the *total variance* of the estimation error. It turns out that the minimizer is none other than the conditional mean: the minimum of $\langle L(\boldsymbol{\varepsilon}_k)\rangle$ with respect to $\mathbf{w}_k^e$ is attained, uniquely in fact, by taking $\mathbf{w}_k^e$ to be the *conditional mean*,

$$\mathbf{w}_k^e \;=\; \left\langle \mathbf{w}_k^t \middle|\, \mathcal{W}_\ell^o \right\rangle . \tag{3.4}$$

Thus the conditional mean estimate is also called the *minimum variance* estimate.

This result (cf. Jazwinski 1970, Theorem 5.3), proven in Appendix A, is extremely general. For instance, notice that the minimizer is *independent* of the particular choice of the positive definite matrix $\mathbf{S}$: the conditional mean simultaneously minimizes *all* quadratic functionals of the estimation error. In other words, in conditional mean estimation one never has to make an artificial choice of a particular quadratic functional to minimize. Furthermore, no assumptions about the nature of the probability densities of the stochastic forcing in (2.9) and (2.13) are required for this equivalence between conditional mean estimation and minimum variance estimation, and consequently no assumptions about the conditional density $p\left(\mathbf{w}_k^t \middle|\, \mathcal{W}_\ell^o\right)$ are required. However, under an assumption that $p\left(\mathbf{w}_k^t \middle|\, \mathcal{W}_\ell^o\right)$ is symmetric about its mean, and is unimodal (has only one local maximum), it turns out that the conditional mean minimizes the expected value of a much larger class of loss functions than just quadratic ones (see Jazwinski 1970, Theorem 5.2); this is the case, for instance, if $p\left(\mathbf{w}_k^t \middle|\, \mathcal{W}_\ell^o\right)$ is Gaussian. Another important property of the conditional mean is that it provides an *unbiased* estimate:

$$\langle \boldsymbol{\varepsilon}_k \rangle \;=\; \left\langle \mathbf{w}_k^t - \left\langle \mathbf{w}_k^t \middle| \mathcal{W}_\ell^o \right\rangle \right\rangle \;=\; \left\langle \mathbf{w}_k^t \right\rangle - \left\langle\left\langle \mathbf{w}_k^t \middle| \mathcal{W}_\ell^o \right\rangle\right\rangle \;=\; 0\,; \tag{3.5}$$

see equation (A.16). For all these reasons, the goal of conditional mean estimation, that is, of defining the state estimate in data assimilation problems by (3.4), is particularly compelling, and much literature in estimation theory is concerned with this goal. In §§ 4–6 we develop evolution equations for the conditional mean $\left\langle \mathbf{w}_k^t \middle|\, \mathcal{W}_k^o \right\rangle$ for the filtering problem.

A potential drawback of conditional mean estimation occurs when the conditional density is multimodal (has several local maxima), as may arise in nonlinear problems with multiple equilibria or multiple attractor basins; see Evensen (1994), Ghil (1997), and references therein. As a simple example, if the conditional density were *bimodal* and symmetric, then the conditional mean would lie at a *minimum* of the conditional density function, which would represent an *unstable* equilibrium point. The number of modes may increase with the dimensionality n of the problem under consideration. However, it is intuitive that the availability of dense observed data also counteracts this tendency: plentiful data serve to define the attractor basin in which the state lies.

In large–scale Earth System data assimilation problems, it is not known whether (or when) the conditional densities have multiple modes. However, Li (1991) has demonstrated the existence of multiple modes for a simple system of three quadratically interacting Rossby

modes. Bennett *et al.* (1993) found unique modes in all but one of ten cases of tropical cyclones. Bürger and Cane (1994) have devised an estimation technique which attempts to account for multimodality.

3.3 Conditional mode estimation

We conclude this section with a brief discussion of conditional mode estimation, also known as maximum *a posteriori* estimation (Maybeck 1979, § 5.5; Sorenson 1980, Chapter 5). Jazwinski (1970, §§ 6.5, 6.10) develops equations for the evolution of the conditional mode, assuming its uniqueness. A simpler approach is known in control theory literature as Pontryagin minimization (Ménard and Daley 1996) and in oceanographic literature as the generalized inverse (Bennett 1992; Bennett *et al.* 1993). These approaches, like the Kalman filter, are based on a number of simplifying probabilistic assumptions.

Specifically, suppose that $\mathbf{G}$ in (2.9) is state–independent,

$$\mathbf{G}_{k-1}\left(\mathbf{w}_{k-1}^t\right) \;=\; \mathbf{G}_{k-1}\,, \tag{3.6}$$

that $\boldsymbol{\varepsilon}_k^t$ in (2.9) is Gaussian with zero mean and white in time,

$$\boldsymbol{\varepsilon}_k^t \;\sim\; N(\mathbf{0},\,\mathbf{Q}_k), \quad \left\langle \boldsymbol{\varepsilon}_k^t \left(\boldsymbol{\varepsilon}_\ell^t\right)^T \right\rangle \;=\; \mathbf{0} \quad \text{for } k \;\neq\; \ell\,, \tag{3.7a,b}$$

that $\boldsymbol{\varepsilon}_k^o$ in (2.13) is Gaussian with zero mean and white in time,

$$\boldsymbol{\varepsilon}_k^o \;\sim\; N(\mathbf{0},\,\mathbf{R}_k), \quad \left\langle \boldsymbol{\varepsilon}_k^o \left(\boldsymbol{\varepsilon}_\ell^o\right)^T \right\rangle \;=\; \mathbf{0} \quad \text{for } k \;\neq\; \ell\,, \tag{3.8a,b}$$

that $\mathbf{w}_0^t$ is Gaussian with mean $\widehat{\mathbf{w}}_0^t$ and covariance $\mathbf{P}_0$,

$$\mathbf{w}_0^t \;\sim\; N\left(\widehat{\mathbf{w}}_0^t,\,\mathbf{P}_0\right)\,, \tag{3.9}$$

and that $\boldsymbol{\varepsilon}_k^t$, $\boldsymbol{\varepsilon}_k^o$ and $\mathbf{w}_0^t$ are mutually uncorrelated,

$$\left\langle \boldsymbol{\varepsilon}_k^t \left(\mathbf{w}_0^t\right)^T \right\rangle \;=\; \left\langle \boldsymbol{\varepsilon}_k^o \left(\mathbf{w}_0^t\right)^T \right\rangle \;=\; \left\langle \boldsymbol{\varepsilon}_k^t (\boldsymbol{\varepsilon}_k^o)^T \right\rangle \;=\; \mathbf{0}\,. \tag{3.10a,b,c}$$

Suppose also that the covariance matrices $\mathbf{P}_0$, $\mathbf{Q}_k$ and $\mathbf{R}_k$ are all nonsingular, hence positive definite. Then it can be shown (cf. § 5 below; Jazwinski 1970, pp. 151–154; Lorenc 1986) that the conditional density $p\left(\mathbf{w}_0^t,\,\mathbf{w}_1^t,\,\ldots,\,\mathbf{w}_N^t \middle|\, \mathcal{W}_N^o\right)$ is proportional to $\exp\left(-J_N\right)$, where J_N is defined as

$$J_N \;=\; \frac{1}{2}\left(\mathbf{w}_0^t - \widehat{\mathbf{w}}_0^t\right)^T \mathbf{P}_0^{-1} \left(\mathbf{w}_0^t - \widehat{\mathbf{w}}_0^t\right) \;+\; \frac{1}{2}\sum_{k=1}^{N} \left(\boldsymbol{\varepsilon}_k^o\right)^T \mathbf{R}_k^{-1} \boldsymbol{\varepsilon}_k^o \;+\; \frac{1}{2}\sum_{k=1}^{N} \left(\boldsymbol{\varepsilon}_{k-1}^t\right)^T \mathbf{Q}_{k-1}^{-1} \boldsymbol{\varepsilon}_{k-1}^t\,. \tag{3.11}$$

The assumption that $\left\langle \boldsymbol{\varepsilon}_k^t (\boldsymbol{\varepsilon}_k^o)^T \right\rangle = \mathbf{0}$ may be removed by including a cross–covariance term in (3.11). The assumptions that $\widehat{\boldsymbol{\varepsilon}}_k^o = \mathbf{0}$ and $\widehat{\boldsymbol{\varepsilon}}_k^t = \mathbf{0}$ may also be removed by subtracting $\widehat{\boldsymbol{\varepsilon}}_k^o$ and $\widehat{\boldsymbol{\varepsilon}}_k^t$ from $\boldsymbol{\varepsilon}_k^o$ and $\boldsymbol{\varepsilon}_k^t$, respectively, in (3.11).

Upon substituting (2.9) and (2.13) into (3.11), assuming $\mathbf{G}_{k-1}$ is invertible, it is seen that J_N depends only upon the free parameters (n–vectors) $\mathbf{w}_0^t$, $\mathbf{w}_1^t$, ..., $\mathbf{w}_N^t$. The conditional mode(s), or maxima of the conditional density, can therefore be found by minimizing J_N with respect to $\mathbf{w}_0^t$, $\mathbf{w}_1^t$, ..., $\mathbf{w}_N^t$. A set of minimizing states $\left\{\mathbf{w}_0^t, \mathbf{w}_1^t, \ldots, \mathbf{w}_N^t\right\}$ is called a *modal trajectory*. For fixed N, this represents a solution of the *fixed–interval smoothing* problem.

For problems of the size encountered in the Earth Sciences, minimization of (3.11) is a computationally imposing task, since the minimization is with respect to nN scalar parameters. The *representer method* (Bennett *et al.* 1993) reduces the size of this task by reducing the effective number of degrees of freedom to the number of observations available over the time interval $[t_0, t_N]$. A different simplification of the computational effort can be made by introducing one additional assumption, namely, to suppose that $\boldsymbol{\varepsilon}_k^t = \mathbf{0}$ for all k. Under this *perfect dynamical model assumption*, the final summation in (3.11) disappears. Upon substituting (2.13) into (3.11) and imposing (2.9) with all $\boldsymbol{\varepsilon}_k^t = \mathbf{0}$ as a constraint, J_N now depends on, and can be minimized with respect to, only one free n–vector: $\mathbf{w}_0^t$, for instance, or $\mathbf{w}_N^t$. This is the assumption made in current four–dimensional variational techniques (4D–VAR; e.g., Andersson *et al.* 1994), and has been studied by Ménard and Daley (1996). Thus, under a number of assumptions, these techniques attempt to calculate the modal trajectory. Nonuniqueness of the minimizing $\mathbf{w}_0^t$ would reflect either inappropriate assumptions or genuine multimodality of the corresponding conditional probability density. Simulated annealing algorithms (e.g., Tarantola 1987) can be used to locate the global minimum in this case. Courtier *et al.* (1994) have introduced a number of additional approximations that turn the 4D–VAR problem into a *quadratic* minimization problem, in which case the minimizer is always unique.

One can also view the minimization of J_N as a purely deterministic problem of minimizing errors, in which case $\mathbf{P}_0$ and $\mathbf{R}_k$ do not have a probabilistic interpretation and the minimization of J_N (under the perfect dynamical model assumption) can be regarded as least–squares curve–fitting of a deterministic model trajectory to the observed data. However, the framework of estimation theory makes clear the probabilistic interpretation of the goal of variational methods, that of conditional mode estimation, which in general is distinct from conditional mean estimation. Jazwinski (1970, p. 172) gives an interesting example of this difference.

The difference between conditional mode estimation and conditional mean estimation is due primarily to nonlinearity: it is well–known that the two are identical for linear problems with known, Gaussian statistics (see § 5 below). For nonlinear problems, in principle one would like to know both the mean and the mode, in fact all of the modes along with their probabilities in case of more than one mode. Accomplishing this for large–scale Earth Science applications, with a minimum number of simplifying assumptions, would appear to be a very challenging task.

4 The discrete Kalman filter

4.1 Introduction

We give here a complete derivation of the discrete Kalman filter, which is the minimum variance (that is, conditional mean) solution of the filtering problem under the probabilistic assumptions (3.6)–(3.10), in case the propagator and observation operator are *linear*. This derivation is *direct* in the sense that it is shown explicitly how to evolve the conditional density $p\left(\mathbf{w}_k^t \middle| \mathcal{W}_k^o\right)$. Numerous other, generally simpler derivations appear in the literature; see Talagrand (1997) for instance, and Examples (7.1)–(7.4) in Jazwinski (1970) for four distinct alternative derivations. While the present derivation is not the simplest one, it demonstrates clearly the roles of the various assumptions and thereby both indicates the obstacles and provides necessary tools to begin relaxing some of them, as we show in the following two sections.

Now let $\mathbf{f}_k$ and $\mathbf{h}_k$ be *linear* in their arguments (we introduce explicit time dependence in the propagator $\mathbf{f}$ here), so that $\mathbf{F}_k \equiv \partial \mathbf{f}_k(\mathbf{w})/\partial \mathbf{w}$ and $\mathbf{H}_k \equiv \partial \mathbf{h}_k(\mathbf{w})/\partial \mathbf{w}$ are *constant* matrices, of dimension $n \times n$ and $p_k \times n$, respectively. The stochastic–dynamic and stochastic observation models (2.9) and (2.13) then become

$$\mathbf{w}_k^t = \mathbf{F}_{k-1}\mathbf{w}_{k-1}^t + \mathbf{G}_{k-1}\boldsymbol{\varepsilon}_{k-1}^t \,, \tag{4.1}$$

$$\mathbf{w}_k^o = \mathbf{H}_k\mathbf{w}_k^t + \boldsymbol{\varepsilon}_k^o \,. \tag{4.2}$$

Suppose we are given the conditional density $p\left(\mathbf{w}_{k-1}^t \middle| \mathcal{W}_{k-1}^o\right)$. The object, then, is to calculate $p\left(\mathbf{w}_k^t \middle| \mathcal{W}_k^o\right)$: the Kalman filter is *recursive*. Denote the mean and covariance matrix, respectively, of the density $p\left(\mathbf{w}_{k-1}^t \middle| \mathcal{W}_{k-1}^o\right)$ by

$$\mathbf{w}_{k-1}^a \equiv \left\langle \mathbf{w}_{k-1}^t \middle| \mathcal{W}_{k-1}^o \right\rangle \,, \tag{4.3}$$

$$\mathbf{P}_{k-1}^a \equiv \left\langle (\mathbf{w}_{k-1}^t - \mathbf{w}_{k-1}^a)(\mathbf{w}_{k-1}^t - \mathbf{w}_{k-1}^a)^T \middle| \mathcal{W}_{k-1}^o \right\rangle \,; \tag{4.4}$$

$\mathbf{w}_{k-1}^a$ is the *analysis* at time t_{k-1}, an n–vector, the expected value of the true state $\mathbf{w}_{k-1}^t$ conditioned on all observations available up to and including that time, while $\mathbf{P}_{k-1}^a$ is the *analysis error covariance matrix*, an $n \times n$ matrix, at time t_{k-1}. At time t_0 there are no observations, so from (3.9) it follows that $p\left(\mathbf{w}_0^t\right)$ is a Gaussian density with mean $\mathbf{w}_0^a \equiv \widehat{\mathbf{w}}_0^t$ and covariance matrix $\mathbf{P}_0^a \equiv \mathbf{P}_0$. We will see that if $p\left(\mathbf{w}_{k-1}^t \middle| \mathcal{W}_{k-1}^o\right)$ is Gaussian then so is $p\left(\mathbf{w}_k^t \middle| \mathcal{W}_k^o\right)$, so by induction it will follow that all the densities $p\left(\mathbf{w}_1^t \middle| \mathcal{W}_1^o\right)$, $p\left(\mathbf{w}_2^t \middle| \mathcal{W}_2^o\right)$, ..., are in fact Gaussian. There are two main steps to this demonstration, and to the Kalman filter algorithm itself.

4.2 The forecast step

First denote the mean and covariance matrix, respectively, of the density $p\big(\mathbf{w}_k^t \big| \mathcal{W}_{k-1}^o\big)$ by

$$\mathbf{w}_k^f \;\equiv\; \big\langle \mathbf{w}_k^t \big| \mathcal{W}_{k-1}^o \big\rangle \;, \tag{4.5}$$

$$\mathbf{P}_k^f \;\equiv\; \left\langle \Big(\mathbf{w}_k^t - \mathbf{w}_k^f\Big)\Big(\mathbf{w}_k^t - \mathbf{w}_k^f\Big)^T \,\middle|\, \mathcal{W}_{k-1}^o \right\rangle ; \tag{4.6}$$

$\mathbf{w}_k^f$ is the *forecast* at the new time t_k, an n–vector, the expected value of the true state $\mathbf{w}_k^t$ conditioned on all observations up to the previous time t_{k-1}, while $\mathbf{P}_k^f$ is the *forecast error covariance matrix*, an $n \times n$ matrix, at time t_k. Substituting (4.1) into (4.5) gives

$$\mathbf{w}_k^f \;=\; \mathbf{F}_{k-1}\big\langle \mathbf{w}_{k-1}^t \big| \mathcal{W}_{k-1}^o \big\rangle \;+\; \mathbf{G}_{k-1}\big\langle \boldsymbol{\varepsilon}_{k-1}^t \big| \mathcal{W}_{k-1}^o \big\rangle \;, \tag{4.7}$$

since $\mathbf{F}_{k-1}$ and $\mathbf{G}_{k-1}$ are constant (that is, deterministic) matrices; cf. (3.6). The first expectation here is $\mathbf{w}_{k-1}^a$ according to definition (4.3). The second one is the unconditional expectation $\big\langle \boldsymbol{\varepsilon}_{k-1}^t \big\rangle$ according to the whiteness assumptions (3.7*b*) and (3.8*b*), along with (3.10), and therefore vanishes under assumption (3.7*a*). Thus we have

$$\mathbf{w}_k^f \;=\; \mathbf{F}_{k-1}\mathbf{w}_{k-1}^a \;, \tag{4.8}$$

which is indeed a forecast to time t_k from the analysis at time t_{k-1} via the linear propagator $\mathbf{F}_{k-1}$.

Substituting (4.1) and (4.8) into (4.6) gives

$$\mathbf{P}_k^f \;=\; \left\langle \Big[\mathbf{F}_{k-1}\big(\mathbf{w}_{k-1}^t - \mathbf{w}_{k-1}^a\big) + \mathbf{G}_{k-1}\boldsymbol{\varepsilon}_{k-1}^t\Big]\Big[\mathbf{F}_{k-1}\big(\mathbf{w}_{k-1}^t - \mathbf{w}_{k-1}^a\big) + \mathbf{G}_{k-1}\boldsymbol{\varepsilon}_{k-1}^t\Big]^T \,\middle|\, \mathcal{W}_{k-1}^o \right\rangle . \tag{4.9}$$

The cross–terms here vanish by (3.7*b*), (3.8*b*) and (3.10), leaving

$$\mathbf{P}_k^f \;=\; \mathbf{F}_{k-1}\mathbf{P}_{k-1}^a\mathbf{F}_{k-1}^T \;+\; \mathbf{G}_{k-1}\mathbf{Q}_{k-1}\mathbf{G}_{k-1}^T \;, \tag{4.10}$$

where again we have used (3.6) and (3.7*a*), along with the definition (4.4) of $\mathbf{P}_{k-1}^a$. Equation (4.10), which gives the evolution of $\mathbf{P}_k^f$ starting from $\mathbf{P}_{k-1}^a$, is often the most computationally demanding portion of the Kalman filter, since it involves large matrix multiplications. For reviews of efforts to ameliorate this computational burden, see Todling and Cohn (1994) and Ghil (1997). Of related interest are the recent articles by Cohn and Todling (1996), Dee (1995) and Verlaan and Heemink (1995).

Equations (4.8) and (4.10) constitute the *forecast step* of the discrete Kalman filter. Note that no Gaussian assumptions were used in their derivation. However, if $p\big(\mathbf{w}_{k-1}^t \big| \mathcal{W}_{k-1}^o\big)$ is Gaussian, then so must be $p\big(\mathbf{w}_k^t \big| \mathcal{W}_{k-1}^o\big)$ since (4.1) is a linear combination of Gaussian random vectors. In fact, we have just derived the mean $\mathbf{w}_k^f$ and covariance matrix $\mathbf{P}_k^f$ of this conditional density.

4.3 The analysis step

Here is where the Gaussian assumptions are needed quite explicitly. First, however, we develop the update equation for $p\big(\mathbf{w}_k^t \big| \mathcal{W}_k^o\big)$ given $p\big(\mathbf{w}_k^t \big| \mathcal{W}_{k-1}^o\big)$ in the absence of *any* assumptions about the nature of these densities, for use both here and in § 5. By repeatedly applying the definition (A.9) of conditional probability densities, we have

$$\begin{aligned} p\big(\mathbf{w}_k^t \big| \mathcal{W}_k^o\big) &= p\big(\mathbf{w}_k^t \big| \mathbf{w}_k^o, \mathcal{W}_{k-1}^o\big) \\ &= \frac{p\big(\mathbf{w}_k^t, \mathbf{w}_k^o, \mathcal{W}_{k-1}^o\big)}{p\big(\mathbf{w}_k^o, \mathcal{W}_{k-1}^o\big)} \\ &= \frac{p\big(\mathbf{w}_k^o \big| \mathbf{w}_k^t, \mathcal{W}_{k-1}^o\big)\, p\big(\mathbf{w}_k^t, \mathcal{W}_{k-1}^o\big)}{p\big(\mathbf{w}_k^o, \mathcal{W}_{k-1}^o\big)} \\ &= \frac{p\big(\mathbf{w}_k^o \big| \mathbf{w}_k^t, \mathcal{W}_{k-1}^o\big)\, p\big(\mathbf{w}_k^t \big| \mathcal{W}_{k-1}^o\big)\, p\big(\mathcal{W}_{k-1}^o\big)}{p\big(\mathbf{w}_k^o \big| \mathcal{W}_{k-1}^o\big)\, p\big(\mathcal{W}_{k-1}^o\big)} \\ &= \frac{p\big(\mathbf{w}_k^o \big| \mathbf{w}_k^t, \mathcal{W}_{k-1}^o\big)\, p\big(\mathbf{w}_k^t \big| \mathcal{W}_{k-1}^o\big)}{p\big(\mathbf{w}_k^o \big| \mathcal{W}_{k-1}^o\big)} \,. \end{aligned} \tag{4.11}$$

This is a quite general result, and holds for the nonlinear equations (2.9), (2.13) with no assumptions on the indicated densities (other than their existence), as well as for the linear equations (4.1) and (4.2). It is a version of Bayes' rule.

An important simplification follows, however, from the whiteness assumption on $\{\boldsymbol{\varepsilon}_k^o\}$. Referring now to the *nonlinear* observation model (2.13), observe that

$$p\big(\mathbf{w}_k^o \big| \mathbf{w}_k^t, \mathcal{W}_{k-1}^o\big) = p\big(\mathbf{w}_k^o \big| \mathbf{w}_k^t\big) \,, \tag{4.12}$$

since *given* $\mathbf{w}_k^t$, $\mathbf{w}_k^o$ depends *only* on $\boldsymbol{\varepsilon}_k^o$, which in turn is *independent* of $\mathcal{W}_{k-1}^o = \{\mathbf{w}_1^o, \mathbf{w}_2^o, \ldots, \mathbf{w}_{k-1}^o\}$ under assumption (3.8*b*).[‡] Thus (4.11) becomes

$$p\big(\mathbf{w}_k^t \big| \mathcal{W}_k^o\big) = \frac{p\big(\mathbf{w}_k^o \big| \mathbf{w}_k^t\big)\, p\big(\mathbf{w}_k^t \big| \mathcal{W}_{k-1}^o\big)}{p\big(\mathbf{w}_k^o \big| \mathcal{W}_{k-1}^o\big)} \,. \tag{4.13}$$

It remains to evaluate each of the three densities on the right side of (4.13). We already have $p\big(\mathbf{w}_k^t \big| \mathcal{W}_{k-1}^o\big)$: it is Gaussian with mean $\mathbf{w}_k^f$, given by (4.8), and covariance matrix $\mathbf{P}_k^f$, given by (4.10). From (4.2) we have

$$\big\langle \mathbf{w}_k^o \big| \mathbf{w}_k^t \big\rangle = \big\langle \mathbf{H}_k \mathbf{w}_k^t + \boldsymbol{\varepsilon}_k^o \big| \mathbf{w}_k^t \big\rangle = \mathbf{H}_k \mathbf{w}_k^t \,, \tag{4.14}$$

since $\boldsymbol{\varepsilon}_k^o$ is state–independent and has mean zero according to (3.8*a*). Therefore

$$\left\langle \Big(\mathbf{w}_k^o - \big\langle \mathbf{w}_k^o \big| \mathbf{w}_k^t \big\rangle\Big)\Big(\mathbf{w}_k^o - \big\langle \mathbf{w}_k^o \big| \mathbf{w}_k^t \big\rangle\Big)^T \,\middle|\, \mathbf{w}_k^t \right\rangle = \left\langle \boldsymbol{\varepsilon}_k^o (\boldsymbol{\varepsilon}_k^o)^T \,\middle|\, \mathbf{w}_k^t \right\rangle = \mathbf{R}_k \,, \tag{4.15}$$

[‡]Here we have used the fact that *uncorrelated* Gaussian random vectors are *independent*, but rather than assuming that the vectors $\boldsymbol{\varepsilon}_k^o$ are Gaussian and mutually uncorrelated we could have assumed instead that $\boldsymbol{\varepsilon}_k^o$ is *independent* of $\boldsymbol{\varepsilon}_\ell^o$ for $\ell < k$ without any Gaussian assumption, leading still to the conclusion that, given $\mathbf{w}_k^t$, $\boldsymbol{\varepsilon}_k^o$ is independent of $\mathcal{W}_{k-1}^o$.

according to (3.8a) and (4.2). Since $\boldsymbol{\varepsilon}_k^o$ was assumed Gaussian, $p\big(\mathbf{w}_k^o \big| \mathbf{w}_k^t\big)$ is then Gaussian with mean $\mathbf{H}_k\mathbf{w}_k^t$ and covariance matrix $\mathbf{R}_k$.

Also from (4.2) we get

$$\big\langle \mathbf{w}_k^o \big| \mathcal{W}_{k-1}^o \big\rangle \;=\; \big\langle \mathbf{H}_k\mathbf{w}_k^t + \boldsymbol{\varepsilon}_k^o \big| \mathcal{W}_{k-1}^o \big\rangle = \mathbf{H}_k\mathbf{w}_k^f \;, \tag{4.16}$$

from definition (4.5) and assumptions (3.8b), (3.10c), along with the assumption (3.8a) that $\langle \boldsymbol{\varepsilon}_k^o \rangle = \mathbf{0}$. We then have

$$\begin{aligned} &\left\langle \Big(\mathbf{w}_k^o - \big\langle \mathbf{w}_k^o \big| \mathcal{W}_{k-1}^o \big\rangle \Big) \Big(\mathbf{w}_k^o - \big\langle \mathbf{w}_k^o \big| \mathcal{W}_{k-1}^o \big\rangle \Big)^T \,\middle|\, \mathcal{W}_{k-1}^o \right\rangle \\ &\qquad = \left\langle \Big[\mathbf{H}_k \Big(\mathbf{w}_k^t - \mathbf{w}_k^f \Big) + \boldsymbol{\varepsilon}_k^o \Big] \Big[\mathbf{H}_k \Big(\mathbf{w}_k^t - \mathbf{w}_k^f \Big) + \boldsymbol{\varepsilon}_k^o \Big]^T \,\middle|\, \mathcal{W}_{k-1}^o \right\rangle . \end{aligned} \tag{4.17}$$

Again the cross–terms vanish as in (4.9), leaving

$$\left\langle \Big(\mathbf{w}_k^o - \big\langle \mathbf{w}_k^o \big| \mathcal{W}_{k-1}^o \big\rangle \Big) \Big(\mathbf{w}_k^o - \big\langle \mathbf{w}_k^o \big| \mathcal{W}_{k-1}^o \big\rangle \Big)^T \,\middle|\, \mathcal{W}_{k-1}^o \right\rangle \;=\; \mathbf{H}_k\mathbf{P}_k^f\mathbf{H}_k^T + \mathbf{R}_k \;, \tag{4.18}$$

from (3.8a) and (4.16). The density $p\big(\mathbf{w}_k^o \big| \mathcal{W}_{k-1}^o\big)$ is Gaussian because $\boldsymbol{\varepsilon}_k^o$ was assumed to be Gaussian–distributed and $p\big(\mathbf{w}_k^t \big| \mathcal{W}_{k-1}^o\big)$ is Gaussian.

Substituting these results into (4.13) and using the definition of the multivariate Gaussian density, we have finally

$$p\big(\mathbf{w}_k^t \big| \mathcal{W}_k^o\big) \;=\; \frac{p_1\, p_2}{p_3} \;, \tag{4.19}$$

where

$$p_1 \;=\; (2\pi)^{-\frac{p}{2}} |\mathbf{R}|^{-\frac{1}{2}} \exp\Big[-\tfrac{1}{2} \big(\mathbf{w}^o - \mathbf{H}\mathbf{w}^t \big)^T \mathbf{R}^{-1} \big(\mathbf{w}^o - \mathbf{H}\mathbf{w}^t \big) \Big] \;, \tag{4.20}$$

$$p_2 \;=\; (2\pi)^{-\frac{n}{2}} \big| \mathbf{P}^f \big|^{-\frac{1}{2}} \exp\Big[-\tfrac{1}{2} \big(\mathbf{w}^t - \mathbf{w}^f \big)^T \big(\mathbf{P}^f \big)^{-1} \big(\mathbf{w}^t - \mathbf{w}^f \big) \Big] \;, \tag{4.21}$$

$$p_3 \;=\; (2\pi)^{-\frac{p}{2}} \big| \mathbf{H}\mathbf{P}^f\mathbf{H}^T + \mathbf{R} \big|^{-\frac{1}{2}} \exp\Big[-\tfrac{1}{2} \big(\mathbf{w}^o - \mathbf{H}\mathbf{w}^f \big)^T \big(\mathbf{H}\mathbf{P}^f\mathbf{H}^T + \mathbf{R} \big)^{-1} \big(\mathbf{w}^o - \mathbf{H}\mathbf{w}^f \big) \Big] \;, \tag{4.22}$$

where the symbol $|\;\;|$ denotes the matrix determinant and for notational convenience we have omitted the time index k which should appear on all vectors and matrices in (4.20) to (4.22). Thus,

$$p\big(\mathbf{w}_k^t \big| \mathcal{W}_k^o\big) \;=\; c\, \exp\big(-\tfrac{1}{2} J\big) \;, \tag{4.23}$$

where

$$c \;=\; (2\pi)^{-\frac{n}{2}} |\mathbf{R}|^{-\frac{1}{2}} \big| \mathbf{P}^f \big|^{-\frac{1}{2}} \big| \mathbf{H}\mathbf{P}^f\mathbf{H}^T + \mathbf{R} \big|^{\frac{1}{2}} \;, \tag{4.24}$$

and

$$\begin{aligned} J \;=\; & \big(\mathbf{w}^o - \mathbf{H}\mathbf{w}^t \big)^T \mathbf{R}^{-1} \big(\mathbf{w}^o - \mathbf{H}\mathbf{w}^t \big) \;+\; \big(\mathbf{w}^t - \mathbf{w}^f \big)^T \big(\mathbf{P}^f \big)^{-1} \big(\mathbf{w}^t - \mathbf{w}^f \big) \\ & - \big(\mathbf{w}^o - \mathbf{H}\mathbf{w}^f \big)^T \big(\mathbf{H}\mathbf{P}^f\mathbf{H}^T + \mathbf{R} \big)^{-1} \big(\mathbf{w}^o - \mathbf{H}\mathbf{w}^f \big) \;. \end{aligned} \tag{4.25}$$

With considerable hindsight, define $\mathbf{P}^a$ by

$$(\mathbf{P}^a)^{-1} \;=\; \left(\mathbf{P}^f\right)^{-1} + \mathbf{H}^T\mathbf{R}^{-1}\mathbf{H}\,; \tag{4.26}$$

$\mathbf{P}^a$ will be shown to be the analysis error covariance matrix (4.4) at time t_k. From the Sherman–Morrison–Woodbury formula (e.g., Golub and Van Loan 1983), (4.26) may be rewritten as

$$\mathbf{P}^a \;=\; (\mathbf{I}-\mathbf{K}\mathbf{H})\mathbf{P}^f\,, \tag{4.27}$$

where

$$\mathbf{K} \;=\; \mathbf{P}^f\mathbf{H}^T\left(\mathbf{H}\mathbf{P}^f\mathbf{H}^T+\mathbf{R}\right)^{-1} \tag{4.28}$$

is the *Kalman gain matrix*. Substituting (4.27) into the expression $\mathbf{P}^a\mathbf{H}^T\mathbf{R}^{-1}$ and using (4.28) yields

$$\mathbf{P}^a\mathbf{H}^T\mathbf{R}^{-1} \;=\; \mathbf{K}\,. \tag{4.29}$$

From (4.29) it follows that

$$\begin{aligned}
&\left(\mathbf{w}^o-\mathbf{H}\mathbf{w}^t\right)^T\mathbf{R}^{-1}\left(\mathbf{w}^o-\mathbf{H}\mathbf{w}^t\right)\\
&\qquad=\;\left[\left(\mathbf{w}^o-\mathbf{H}\mathbf{w}^f\right)-\mathbf{H}\left(\mathbf{w}^t-\mathbf{w}^f\right)\right]^T\mathbf{R}^{-1}\left[\left(\mathbf{w}^o-\mathbf{H}\mathbf{w}^f\right)-\mathbf{H}\left(\mathbf{w}^t-\mathbf{w}^f\right)\right]\\
&\qquad=\;\left(\mathbf{w}^o-\mathbf{H}\mathbf{w}^f\right)^T\mathbf{R}^{-1}\left(\mathbf{w}^o-\mathbf{H}\mathbf{w}^f\right)\;+\;\left(\mathbf{w}^t-\mathbf{w}^f\right)^T\mathbf{H}^T\mathbf{R}^{-1}\mathbf{H}\left(\mathbf{w}^t-\mathbf{w}^f\right)\\
&\qquad\quad-\left(\mathbf{w}^t-\mathbf{w}^f\right)^T\left(\mathbf{P}^a\right)^{-1}\left[\mathbf{K}\left(\mathbf{w}^o-\mathbf{H}\mathbf{w}^f\right)\right]\\
&\qquad\quad-\left[\mathbf{K}\left(\mathbf{w}^o-\mathbf{H}\mathbf{w}^f\right)\right]^T\left(\mathbf{P}^a\right)^{-1}\left(\mathbf{w}^t-\mathbf{w}^f\right)\,.
\end{aligned} \tag{4.30}$$

Substituting this result into (4.25) and using (4.26) yields the expression

$$\begin{aligned}
J \;=\;&\left(\mathbf{w}^t-\mathbf{w}^f\right)^T(\mathbf{P}^a)^{-1}\left(\mathbf{w}^t-\mathbf{w}^f\right)\\
&+\left(\mathbf{w}^o-\mathbf{H}\mathbf{w}^f\right)^T\left[\mathbf{R}^{-1}-\left(\mathbf{H}\mathbf{P}^f\mathbf{H}^T+\mathbf{R}\right)^{-1}\right]\left(\mathbf{w}^o-\mathbf{H}\mathbf{w}^f\right)\\
&-\left(\mathbf{w}^t-\mathbf{w}^f\right)^T(\mathbf{P}^a)^{-1}\left[\mathbf{K}\left(\mathbf{w}^o-\mathbf{H}\mathbf{w}^f\right)\right]\;-\;\left[\mathbf{K}\left(\mathbf{w}^o-\mathbf{H}\mathbf{w}^f\right)\right]^T(\mathbf{P}^a)^{-1}\left(\mathbf{w}^t-\mathbf{w}^f\right)\,.
\end{aligned} \tag{4.31}$$

But from (4.28) and (4.29) it follows that

$$\begin{aligned}
\mathbf{R}^{-1}-\left(\mathbf{H}\mathbf{P}^f\mathbf{H}^T+\mathbf{R}\right)^{-1} \;&=\; \left(\mathbf{H}\mathbf{P}^f\mathbf{H}^T+\mathbf{R}\right)^{-1}\left[\left(\mathbf{H}\mathbf{P}^f\mathbf{H}^T+\mathbf{R}\right)-\mathbf{R}\right]\mathbf{R}^{-1}\\
&=\; \mathbf{K}^T\mathbf{H}^T\mathbf{R}^{-1}=\mathbf{K}^T(\mathbf{P}^a)^{-1}\mathbf{K}\,,
\end{aligned} \tag{4.32}$$

so that we have finally

$$J \;=\; \left[\left(\mathbf{w}^t-\mathbf{w}^f\right)-\mathbf{K}\left(\mathbf{w}^o-\mathbf{H}\mathbf{w}^f\right)\right]^T\left(\mathbf{P}^a\right)^{-1}\left[\left(\mathbf{w}^t-\mathbf{w}^f\right)-\mathbf{K}\left(\mathbf{w}^o-\mathbf{H}\mathbf{w}^f\right)\right]\,. \tag{4.33}$$

Now, to simplify the expression for the constant c in (4.24), let

$$\mathbf{M} \;\equiv\; \mathbf{H}\,\mathbf{P}^f\mathbf{H}^T+\mathbf{R}\,, \tag{4.34}$$

and define the partitioned matrix

$$\mathbf{A} \equiv \begin{bmatrix} \mathbf{P}^f & \mathbf{P}^f\mathbf{H}^T \\ \mathbf{H}\mathbf{P}^f & \mathbf{M} \end{bmatrix} = \begin{bmatrix} \mathbf{I}_n & \mathbf{0} \\ \mathbf{H} & \mathbf{I}_p \end{bmatrix} \begin{bmatrix} \mathbf{P}^f & \mathbf{0} \\ \mathbf{0} & \mathbf{R} \end{bmatrix} \begin{bmatrix} \mathbf{I}_n & \mathbf{H}^T \\ \mathbf{0} & \mathbf{I}_p \end{bmatrix}, \tag{4.35}$$

where $\mathbf{I}_n$ and $\mathbf{I}_p$ are the $n \times n$ and $p \times p$ identity matrices. It follows that

$$|\mathbf{A}| = \left|\mathbf{P}^f\right||\mathbf{R}| ; \tag{4.36}$$

cf. Householder (1964, p. 17). The matrix $\mathbf{A}$ may also be factored as

$$\mathbf{A} = \begin{bmatrix} \mathbf{I}_n & \mathbf{P}^f\mathbf{H}^T\mathbf{M}^{-1} \\ \mathbf{0} & \mathbf{I}_p \end{bmatrix} \begin{bmatrix} \mathbf{P}^a & \mathbf{0} \\ \mathbf{0} & \mathbf{M} \end{bmatrix} \begin{bmatrix} \mathbf{I}_n & \mathbf{0} \\ \mathbf{M}^{-1}\mathbf{H}\mathbf{P}^f & \mathbf{I}_p \end{bmatrix}, \tag{4.37}$$

according to (4.27), (4.28) and (4.34), from which it follows that

$$|\mathbf{A}| = |\mathbf{P}^a||\mathbf{M}| . \tag{4.38}$$

Therefore

$$|\mathbf{P}^a| = \left|\mathbf{P}^f\right||\mathbf{R}||\mathbf{M}|^{-1} , \tag{4.39}$$

and (4.24) becomes

$$c = (2\pi)^{-\frac{n}{2}}|\mathbf{P}^a|^{-\frac{1}{2}} . \tag{4.40}$$

Equations (4.33) and (4.40) show that the density $p\left(\mathbf{w}_k^t \middle| \mathcal{W}_k^o\right)$ in (4.23) is Gaussian with mean

$$\left\langle \mathbf{w}_k^t \middle| \mathcal{W}_k^o \right\rangle = \mathbf{w}_k^f + \mathbf{K}_k\Big(\mathbf{w}_k^o - \mathbf{H}_k\mathbf{w}_k^f\Big) , \tag{4.41}$$

and covariance matrix

$$\left\langle \Big(\mathbf{w}_k^t - \left\langle \mathbf{w}_k^t \middle| \mathcal{W}_k^o \right\rangle\Big)\Big(\mathbf{w}_k^t - \left\langle \mathbf{w}_k^t \middle| \mathcal{W}_k^o \right\rangle\Big)^T \middle| \mathcal{W}_k^o \right\rangle = \mathbf{P}_k^a , \tag{4.42}$$

where the time index has been re–introduced. Referring back to (4.3), (4.4), (4.27) and (4.28), we finally have the *analysis update* equations for the Kalman filter:

$$\mathbf{w}_k^a = \mathbf{w}_k^f + \mathbf{K}_k\Big(\mathbf{w}_k^o - \mathbf{H}_k\mathbf{w}_k^f\Big) , \tag{4.43}$$

$$\mathbf{K}_k = \mathbf{P}_k^f\mathbf{H}_k^T\Big(\mathbf{H}_k\mathbf{P}_k^f\mathbf{H}_k^T + \mathbf{R}_k\Big)^{-1} , \tag{4.44}$$

$$\mathbf{P}_k^a = (\mathbf{I} - \mathbf{K}_k\mathbf{H}_k)\mathbf{P}_k^f . \tag{4.45}$$

The complete Kalman filter algorithm thus consists of these three equations, along with the forecast equations (4.8) and (4.10). This section concludes with a brief discussion of some of the properties and generalizations of the Kalman filter.

4.4 Properties

Observe first that the covariance evolution equations (4.10), (4.45), (4.44) do not depend on the observations, so that the conditional covariances $\mathbf{P}_k^f$, $\mathbf{P}_k^a$ are actually unconditional covariances. This property is a result of the linearity of (4.1) and (4.2), as well as the assumption that all of the error statistics are Gaussian. For small enough problems the covariance evolution, and in particular the sequence of Kalman gain matrices (4.44) needed in (4.43), may therefore be *precomputed* under this rather restrictive scenario, before the observations are actually taken. This property, along with the particular form of (4.8) and (4.43), leads also to a number of algebraically equivalent forms of the Kalman filter algorithm stated here, which can reduce computational costs and/or enhance computational stability. See Maybeck (1979, Chapter 7) for a fairly complete discussion of alternative Kalman filter algorithms.

The Kalman filter indeed acts as a (low–pass) filter, removing unwanted noise from the observations $\mathcal{W}_k^o$. See Daley and Ménard (1993) for a discussion of the filtering properties of the Kalman filter.

There exists a large body of literature on stability properties of the Kalman filter and its sensitivity to parameters, beginning with Kalman's (1960) seminal article, and covered well in most texts on estimation theory. The stability results depend heavily, in turn, on the observability and controllability properties of the system (4.1), (4.2) under consideration, also discussed in most texts; see also Cohn and Dee (1988), Ghil (1997).

An error sensitivity property of considerable practical importance is the following. In practice we seldom know either the observation error covariance matrix $\mathbf{R}_k$ or the model error covariance matrix $\mathbf{Q}_k$. Suppose, however, that we have access to *conservative* estimates of each, $\widetilde{\mathbf{R}}_k$ and $\widetilde{\mathbf{Q}}_k$, that is, $\widetilde{\mathbf{R}}_k \geqslant \mathbf{R}_k$ and $\widetilde{\mathbf{Q}}_k \geqslant \mathbf{Q}_k$ for each t_k.[¶] Then if we calculate a Kalman filter using $\widetilde{\mathbf{R}}_k$ and $\widetilde{\mathbf{Q}}_k$, the resulting approximate forecast and analysis error covariances, denoted by $\widetilde{\mathbf{P}}_k^f$ and $\widetilde{\mathbf{P}}_k^a$, have the property that $\widetilde{\mathbf{P}}_k^f \geqslant \mathbf{P}_k^f$ and $\widetilde{\mathbf{P}}_k^a \geqslant \mathbf{P}_k^a$, for each t_k, where $\mathbf{P}_k^f$ and $\mathbf{P}_k^a$ now denote the *actual* covariances resulting from our (suboptimal) filter calculations involving $\widetilde{\mathbf{R}}_k$ and $\widetilde{\mathbf{Q}}_k$ (Jazwinski 1970, Theorem 7.6), assuming a conservative estimate $\widetilde{\mathbf{P}}_0^a \geqslant \mathbf{P}_0^a$ also. Thus we *know* that we have computed conservative estimates of $\mathbf{P}_k^f$ and $\mathbf{P}_k^a$, and in particular we have conservative estimates of the forecast and analysis error *variances* (diagonal elements of $\mathbf{P}_k^f$ and $\mathbf{P}_k^a$) as well. More generally, the *performance evaluation equations* can be implemented to study and interpret the results of numerous suboptimal filter schemes, in particular those that approximate the dynamics of $\mathbf{P}_k^f$ in (4.10), (4.45); cf. Todling and Cohn (1994), Cohn and Todling (1996).

The effect of estimating $\mathbf{P}_k^f$ conservatively, that is, of *overestimating* the forecast error covariance matrix, is to assign more weight to the observed data than one would otherwise, resulting in "noisier" analyses. This avoids the problem of *filter divergence* (cf. Jazwinski 1970, §§ 8.8–8.12), however, which occurs when the filter "thinks" it is doing better than

[¶]By $\mathbf{A} \geqslant \mathbf{B}$ for symmetric matrices $\mathbf{A}$ and $\mathbf{B}$ of like dimension, we mean that the difference $\mathbf{A} - \mathbf{B}$ is positive semidefinite.

it actually is, that is, when $\mathbf{P}_k^f$ is *underestimated* and observations thereby receive too little weight: in the extreme case, if observations are neglected then data assimilation accomplishes nothing, and the estimated state drifts away according to the (incorrect) model dynamics. Thus a standard rule–of–thumb is to estimate unknown error covariances conservatively.

A fundamental property of the Kalman filter is expressed by the *Innovations Theorem* (Kailath 1968, Theorem 2), which says that the sequence of *innovations* $\mathbf{w}_k^o - \mathbf{H}_k \mathbf{w}_k^f$ is Gaussian and white in time, i.e.,

$$\left\langle \left(\mathbf{w}_k^o - \mathbf{H}_k \mathbf{w}_k^f \right) \left(\mathbf{w}_\ell^o - \mathbf{H}_\ell \mathbf{w}_\ell^f \right)^T \right\rangle \; = \; \mathbf{0} \quad \text{for } k \; \neq \; \ell \, ; \tag{4.46}$$

this can be verified directly from the Kalman filter equations and assumptions. In fact a nonlinear version of this result is also true (Frost and Kailath 1971, Theorem 2): for *nonlinear* observation models in continuous time with white, Gaussian, state–independent observation error, and essentially [see their equations (3) and (4)] no assumptions on the nature of the probability density of the true state, hence on that of the density of the model error, the (nonlinear) innovation process is still white and Gaussian for the optimal (conditional mean, i.e., minimum variance) nonlinear filter. While optimal filters for *nonlinear* problems cannot generally be expressed in closed form because the first and second moments of $p\left(\mathbf{w}_k^t \middle| \mathcal{W}_k^o\right)$ become coupled with higher–order moments (see Appendix C),[‡] this is still a very powerful result: by routinely *monitoring* the observed–minus–forecast residuals for whiteness, we can rationally assess the proximity of a given suboptimal filtering algorithm to optimality, and also assess putative improvements to the algorithm. Daley (1992b) has shown how such monitoring can even be useful in diagnosing weaknesses in an operational atmospheric data assimilation system.

In fact, in numerical weather prediction and other large–scale data assimilation problems, one does *not* generally know the form of the probability densities appearing in (4.13) from which the analysis equations (4.43)–(4.45) were derived. For the various assumptions in this section we have seen them to be Gaussian, but in practice one has at best only limited knowledge of the ingredients $\mathbf{w}_k^f$, $\mathbf{P}_k^f$, $\mathbf{H}_k$ and $\mathbf{R}_k$ of the analysis equations, let alone any assurance that the densities are Gaussian. Yet these equations in fact are simply the matrix formulation of the usual "optimal interpolation" analysis equations (e.g., Bergman 1979; Lorenc 1981) without data selection (da Silva *et al.* 1995). One expects, then, that these equations can be derived from much simpler assumptions than those we have stated, and indeed they can, as follows.

Given only the observation model (4.2), along with $\langle \boldsymbol{\varepsilon}_k^o \rangle = \mathbf{0}$ and $\left\langle \boldsymbol{\varepsilon}_k^o (\boldsymbol{\varepsilon}_k^o)^T \right\rangle = \mathbf{R}_k$, and no assumptions on the densities of $\mathbf{w}_k^t$ and $\boldsymbol{\varepsilon}_k^o$ themselves, then of all estimators of the *linear* form (4.43), the choice of $\mathbf{K}_k$ that minimizes the scalar $\left\langle \left(\mathbf{w}_k^a - \mathbf{w}_k^t \right)^T \mathbf{S} \left(\mathbf{w}_k^a - \mathbf{w}_k^t \right) \right\rangle$ for all positive definite $n \times n$ matrices $\mathbf{S}$ is none other than the Kalman gain (4.44); cf. Jazwinski (1970, Example 7.4).[‖] This is the *best linear unbiased estimate* (BLUE) property of the

[‡]See Casti (1985, pp. 173–175) for an example of a nonlinear filtering problem that does have a closed–form solution.

[‖]See also § 3.2 of the present article. Observe also that for $\mathbf{S} = \mathbf{I}$, this scalar is identical to *trace* $\mathbf{P}_k^a$, where

Kalman filter (more accurately, the best affine minimum variance estimate property; see Catlin 1989, § 5.3). The drawback, however, is that if the densities in (4.13) are *not* Gaussian, then the analysis equations (4.43)–(4.45) do *not* give the conditional mean analysis, which is the minimum variance analysis. Non–Gaussian densities lead to nonlinear analysis equations for the conditional mean, even for linear observation operators. We return to this subject in § 5.

4.5 Generalizations

To bring the ideas of Kalman filtering into the realm of actual Earth Science data assimilation problems, perhaps the most obviously necessary generalization is to nonlinear observation operators and nonlinear dynamics, which are discussed in § 5 and Appendix C, respectively. Also, in § 2 we have raised the necessity of careful treatment of continuum aspects of the theory, which are discussed in § 6 and Appendix C. Generalization to smoothing problems of various types have been mentioned already in § 3; see Maybeck (1979, Chapter 8) and Anderson and Moore (1979, Chapter 7) for fairly extensive discussions.

Here for completeness we mention just a few simple generalizations of the discrete linear theory presented already. First, in the presence of *known* model error bias $\mathbf{G}_k \widehat{\boldsymbol{\varepsilon}}_k^t = \mathbf{G}_k \langle \boldsymbol{\varepsilon}_k^t \rangle$, cf. (2.6) and (2.7), and/or *known* measurement error bias (2.11), it is straightforward to generalize the standard Kalman filter equations. The forecast equation (4.8) becomes

$$\mathbf{w}_k^f \;=\; \mathbf{F}_{k-1}\mathbf{w}_{k-1}^a + \mathbf{G}_{k-1}\widehat{\boldsymbol{\varepsilon}}_{k-1}^t \;, \tag{4.47}$$

and follows readily from (4.7). The analysis equation (4.43) becomes

$$\mathbf{w}_k^a \;=\; \mathbf{w}_k^f + \mathbf{K}_k\Big(\mathbf{w}_k^o - \mathbf{H}_k\mathbf{w}_k^f - \widehat{\boldsymbol{\varepsilon}}_k^o\Big) \;, \tag{4.48}$$

which arises, for example, by considering $\mathbf{w}_k^o - \widehat{\boldsymbol{\varepsilon}}_k^o$ as an unbiased "pseudo–observation" vector. The covariance equations (4.10), (4.45), as well as the equation for the Kalman gain (4.44) are seen to remain unchanged by following their derivation closely.

Model and observation biases, however, like their error covariances, are seldom actually known. Rather, they either need to be estimated along with the state itself, or else their presence should at least be accounted for in the filtering procedure. The former approach can be accommodated by *state augmentation*, that is, by simply including the bias parameters as additional state variables to be estimated. This introduces additional computational expense, and one must assume either that the biases are constant in time, or else have access to a dynamical model for the biases (Jazwinski 1970, § 8.4). In case the biases are indeed constant in time, the latter approach, of accounting for biases but not estimating them directly, is accomplished with less expense than the former, through a generalization known as the Schmidt–Kalman filter (Jazwinski 1970, § 8.4); see also Caglayan and Lancraft (1983) for a nonlinear treatment. Dee and da Silva (1997) have developed a model error bias

$\mathbf{P}_k^a$ was defined in (4.4), the minimization of which is often used as a heuristic criterion for deriving the Kalman gain matrix.

estimation scheme suitable for large–scale models. Derber (1989) has introduced a model error bias estimation scheme in a variational context.

We conclude this section by mentioning that straightforward generalizations of the Kalman filter exist in case the model error and observation error are correlated with each other (Jazwinski 1970, Examples 7.5 and 7.6) and in case the observation error is correlated (rather than white) in time, provided it can be modeled as a Markov process (Jazwinski 1970, Example 7.7). The case of time–correlated model error can also be treated, and has been studied by Daley (1992a).

5 Nonlinear analysis updates

5.1 Introduction

Here we discuss the nonlinear analysis equations that arise when either the observation operator $\mathbf{h}$ is nonlinear or the observation error density is not Gaussian. In the former case it will be seen that in practice one can usually only obtain an approximate formula for the minimum variance (conditional mean) analysis vector $\mathbf{w}^a$. The approximation described here leads to the analysis step of the *locally iterated extended Kalman filter*, which is seen to be algebraically equivalent to global variational analysis algorithms (e.g., Parrish and Derber 1992; Heckley *et al.* 1992). In the latter case, an exact formula is obtained when the observation error density is related to the Gaussian density in a known way.

5.2 Nonlinear observation operators

In order to describe the nature of the approximation usually made in this case, first we rewrite the general result (4.11). By the definition (A.9) of conditional probability densities, we may write

$$p\big(\mathbf{w}_k^t \big|\, \mathcal{W}_{k-1}^o\big) \;=\; \frac{p\big(\mathbf{w}_k^t,\, \mathcal{W}_{k-1}^o\big)}{p\big(\mathcal{W}_{k-1}^o\big)}\,, \tag{5.1}$$

and also

$$p\big(\mathbf{w}_k^o \big|\, \mathbf{w}_k^t,\, \mathcal{W}_{k-1}^o\big)\, p\big(\mathbf{w}_k^t,\, \mathcal{W}_{k-1}^o\big) \;=\; p\big(\mathbf{w}_k^o,\, \mathbf{w}_k^t,\, \mathcal{W}_{k-1}^o\big)\,. \tag{5.2}$$

Therefore

$$\begin{aligned}
\int p\big(\mathbf{w}_k^o \big|\, \mathbf{w}_k^t,\, \mathcal{W}_{k-1}^o\big)\, p\big(\mathbf{w}_k^t \big|\, \mathcal{W}_{k-1}^o\big)\, d\mathbf{w}_k^t \;&=\; \frac{1}{p\big(\mathcal{W}_{k-1}^o\big)} \int p\big(\mathbf{w}_k^o,\, \mathbf{w}_k^t,\, \mathcal{W}_{k-1}^o\big)\, d\mathbf{w}_k^t \\
&=\; \frac{p\big(\mathbf{w}_k^o,\, \mathcal{W}_{k-1}^o\big)}{p\big(\mathcal{W}_{k-1}^o\big)} \\
&=\; p\big(\mathbf{w}_k^o \big|\, \mathcal{W}_{k-1}^o\big)\,,
\end{aligned} \tag{5.3}$$

where the second equality follows from definition (A.5) and the third from (A.9). Equation (5.3) allows (4.11) to be written as

$$p\big(\mathbf{w}_k^t \big|\, \mathcal{W}_k^o\big) \;=\; \frac{p\big(\mathbf{w}_k^o \big|\, \mathbf{w}_k^t,\, \mathcal{W}_{k-1}^o\big)\, p\big(\mathbf{w}_k^t \big|\, \mathcal{W}_{k-1}^o\big)}{\displaystyle\int p\big(\mathbf{w}_k^o \big|\, \mathbf{w}_k^t,\, \mathcal{W}_{k-1}^o\big)\, p\big(\mathbf{w}_k^t \big|\, \mathcal{W}_{k-1}^o\big) d\mathbf{w}_k^t}\;; \tag{5.4}$$

here the denominator is just the integral of the numerator with respect to $\mathbf{w}_k^t$. From (5.4) it follows that $\int p\big(\mathbf{w}_k^t \big|\, \mathcal{W}_k^o\big) d\mathbf{w}_k^t = 1$ as expected.

Now consider the nonlinear observation model

$$\mathbf{w}^o \;=\; \mathbf{h}\big(\mathbf{w}^t\big) + \boldsymbol{\varepsilon}^o\,, \tag{5.5}$$

with

$$\boldsymbol{\varepsilon}^o \;\sim\; N(\mathbf{0},\, \mathbf{R})\,, \tag{5.6}$$

and the time subscript k is omitted for notational convenience. From (4.12) and from arguments identical to those leading to (4.20), we have

$$p\big(\mathbf{w}_k^o \big|\, \mathbf{w}_k^t\,, \mathcal{W}_{k-1}^o\big) \;=\; (2\pi)^{-\frac{p}{2}} |\mathbf{R}|^{-\frac{1}{2}} \exp\Big\{ -\tfrac{1}{2} \big[\mathbf{w}^o - \mathbf{h}\big(\mathbf{w}^t\big)\big]^T \mathbf{R}^{-1} \big[\mathbf{w}^o - \mathbf{h}\big(\mathbf{w}^t\big)\big] \Big\}\,. \tag{5.7}$$

We will *assume* that the prior density $p\big(\mathbf{w}_k^t \big|\, \mathcal{W}_{k-1}^o\big)$ in (5.4) is Gaussian, and therefore given by (4.21), although it should be kept in mind that this is already an approximation since the state $\mathbf{w}_k^t$ conditioned on observations obtained nonlinearly from past states will not generally be Gaussian.

From (5.7) and (4.21) it follows that the numerator N in (5.4) can be written as

$$N \;=\; c\, \exp(-J)\,, \tag{5.8}$$

where

$$c \;=\; (2\pi)^{-\frac{p}{2}} (2\pi)^{-\frac{n}{2}} |\mathbf{R}|^{-\frac{1}{2}} \big|\mathbf{P}^f\big|^{-\frac{1}{2}}, \tag{5.9}$$

and

$$J = J\big(\mathbf{w}^t\big) \;\equiv\; \frac{1}{2} \big(\mathbf{w}^t - \mathbf{w}^f\big)^T \big(\mathbf{P}^f\big)^{-1} \big(\mathbf{w}^t - \mathbf{w}^f\big) \;+\; \frac{1}{2} \big[\mathbf{w}^o - \mathbf{h}\big(\mathbf{w}^t\big)\big]^T \mathbf{R}^{-1} \big[\mathbf{w}^o - \mathbf{h}\big(\mathbf{w}^t\big)\big]\,. \tag{5.10}$$

Thus we may write (5.4) as

$$p\big(\mathbf{w}_k^t \big|\, \mathcal{W}_k^o\big) \;=\; \frac{\exp\big[-J\big(\mathbf{w}_k^t\big)\big]}{\displaystyle\int \exp\big[-J\big(\mathbf{w}_k^t\big)\big] d\mathbf{w}_k^t}\,. \tag{5.11}$$

The difficulty in evaluating $p\big(\mathbf{w}_k^t \big|\, \mathcal{W}_k^o\big)$, and therefore its mean and covariance, lies first in evaluating the integral in (5.11). While we have already seen that this is straightforward

in the linear case, it is not trivial in case the observation operator $\mathbf{h}$ is nonlinear. It may be possible to evaluate this integral analytically for some simple nonlinearities. For realistic Earth Science applications, though, only methods such as Monte Carlo integration would appear to be feasible, although still expensive, since this is an n–dimensional integral. Additional integrations would be required to calculate the first two moments of $p\left(\mathbf{w}_k^t \middle| \mathcal{W}_k^o\right)$. However, one approximation has already been made, namely that $p\left(\mathbf{w}_k^t \middle| \mathcal{W}_{k-1}^o\right)$ is Gaussian, so further approximations not "worse" than this one should be useful.

One approach would be to approximate just the mean and covariance of the density $p\left(\mathbf{w}_k^o \middle| \mathcal{W}_{k-1}^o\right)$ that the integral in (5.4) represents. For example, from (5.5) we have to first order that

$$\left\langle \mathbf{w}_k^o \middle| \mathcal{W}_{k-1}^o \right\rangle \;=\; \left\langle \mathbf{h}\left(\mathbf{w}_k^t\right) \middle| \mathcal{W}_{k-1}^o \right\rangle \cong \mathbf{h}\left(\left\langle \mathbf{w}_k^t \middle| \mathcal{W}_{k-1}^o \right\rangle \right) = \mathbf{h}\left(\mathbf{w}_k^f\right) . \tag{5.12}$$

In carrying out such an approach, one would still have to arrive at a density for $p\left(\mathbf{w}_k^o \middle| \mathcal{W}_{k-1}^o\right)$, which would not be Gaussian, such that the quotient in (5.4) would integrate to unity. A different approach, the most common one, is described next.

Since the denominator of (5.11) is simply a normalizing constant, independent of $\mathbf{w}_k^t$, and the numerator is readily available in (5.10), it is straightforward to calculate the *maximum* of (5.11) with respect to $\mathbf{w}_k^t$, that is, the *mode* of the conditional density $p\left(\mathbf{w}_k^t \middle| \mathcal{W}_k^o\right)$. This is the analysis step of the locally iterated extended Kalman filter (Jazwinski 1970, § 9.7): to approximate the mean of $p\left(\mathbf{w}_k^t \middle| \mathcal{W}_k^o\right)$ by its mode,

$$\mathbf{w}_k^a \;\cong\; \mathbf{m}\left(\mathbf{w}_k^t \middle| \mathcal{W}_k^o\right) ; \tag{5.13}$$

see the discussion in the last paragraph of § 3.1 for notation. This analysis is biased, unless by chance it happens that $\mathbf{m}\left(\mathbf{w}_k^t \middle| \mathcal{W}_k^o\right) = \left\langle \mathbf{w}_k^t \middle| \mathcal{W}_k^o \right\rangle \equiv \mathbf{w}_k^a$; see (3.4), (3.5), and definition (4.3). The mode may not be unique.

From (5.11) it follows that maxima of $p\left(\mathbf{w}_k^t \middle| \mathcal{W}_k^o\right)$ coincide with minima of $J\left(\mathbf{w}_k^t\right)$ defined in (5.10). The gradient vector $\partial J(\mathbf{w})/\partial \mathbf{w}$ is obtained by differentiating (5.10):

$$\frac{\partial J(\mathbf{w})}{\partial \mathbf{w}} \;=\; \left(\mathbf{P}^f\right)^{-1}\left(\mathbf{w} - \mathbf{w}^f\right) \;+\; \mathbf{H}^T(\mathbf{w})\,\mathbf{R}^{-1}\left[\mathbf{h}(\mathbf{w}) - \mathbf{w}^o\right] , \tag{5.14}$$

where

$$\mathbf{H}(\mathbf{w}) \;\equiv\; \frac{\partial \mathbf{h}(\mathbf{w})}{\partial \mathbf{w}} \tag{5.15}$$

is the *tangent linear forward observation operator*. The variational methods now gaining widespread use (e.g., Parrish and Derber 1992; Heckley *et al.* 1992) generally solve for a minimum (denoted by $\mathbf{w}^a$ hereafter) of $J(\mathbf{w})$, namely a vector $\mathbf{w}^a$ such that $\partial J(\mathbf{w})/\partial \mathbf{w} = \mathbf{0}$ at $\mathbf{w} = \mathbf{w}^a$, by employing (5.14) in a gradient descent method (e.g., Navon and Legler 1987). The method most commonly seen in the estimation theory literature is the following quasi–Newton method, which relies explicitly on the form of $J(\mathbf{w})$ given by (5.10). This method circumvents the need for choosing a step size, which is sometimes a source of difficulty in descent methods.

By differentiating (5.14) one finds that the $(i,\,j)^{\text{th}}$ element of the *Hessian matrix* $\partial^2 J(\mathbf{w})/\partial \mathbf{w}^2$ is given by

$$\left[\frac{\partial^2 J}{\partial \mathbf{w}^2}\right]_{ij} \;=\; \left[\left(\mathbf{P}^f\right)^{-1} + \mathbf{H}^T(\mathbf{w})\,\mathbf{R}^{-1}\mathbf{H}(\mathbf{w})\right]_{ij} + \left\{\mathbf{H}^{jT}(\mathbf{w})\,\mathbf{R}^{-1}\left[\mathbf{h}(\mathbf{w}) - \mathbf{w}^o\right]\right\}_i , \tag{5.16}$$

where the matrices $\mathbf{H}^j(\mathbf{w})$, $j=1,\,2,\,\ldots\,,\,n$ are defined by

$$\mathbf{H}^j(\mathbf{w}) \;\equiv\; \frac{\partial \mathbf{H}(\mathbf{w})}{\partial \mathbf{w}_j}\,; \tag{5.17}$$

these latter matrices vanish in case $\mathbf{h}(\mathbf{w})$ is linear or affine, since then $\mathbf{H}$ defined in (5.15) is independent of $\mathbf{w}$. Newton's method for minimizing $J(\mathbf{w})$ is the iteration

$$\mathbf{w}_{\ell+1} \;=\; \mathbf{w}_\ell \;-\; \left[\frac{\partial^2 J(\mathbf{w}_\ell)}{\partial \mathbf{w}_\ell^2}\right]^{-1} \frac{\partial J(\mathbf{w}_\ell)}{\partial \mathbf{w}_\ell}\,, \quad \ell = 0,\,1,\,2,\,\ldots\,, \tag{5.18}$$

with $\mathbf{w}_0 = \mathbf{w}^f$ for instance. This iteration converges quadratically to a (local) minimum if the Hessian matrix is positive definite. There may be multiple minima of course, which would be the case if $p\left(\mathbf{w}_k^t \middle|\, \mathcal{W}_k^o\right)$ is multimodal.

The quasi–Newton method is obtained by neglecting the second term in (5.16), which we have seen to arise only from nonlinearity, while retaining the first term, which is present even for linear observation operators. This leads to a significant computational simplification as well as an easily verified convergence criterion. Thus we write

$$\frac{\partial^2 J}{\partial \mathbf{w}^2} \;\cong\; \left(\mathbf{P}^f\right)^{-1} + \mathbf{H}^T(\mathbf{w})\,\mathbf{R}^{-1}\,\mathbf{H}(\mathbf{w})\,, \tag{5.19}$$

and substitute this expression into the iteration (5.18). This approximate Hessian is positive definite if, for example, $\left(\mathbf{P}^f\right)^{-1}$ is positive definite, so convergence is easy to guarantee, although it may be less than quadratic since the Hessian has been approximated.

Now for notational convenience, define

$$\mathbf{H}_\ell \;\equiv\; \mathbf{H}(\mathbf{w}_\ell)\,, \tag{5.20}$$

$$\mathbf{K}_\ell \;\equiv\; \mathbf{P}^f\mathbf{H}_\ell^T\left(\mathbf{H}_\ell\mathbf{P}^f\mathbf{H}_\ell^T + \mathbf{R}\right)^{-1}. \tag{5.21}$$

Then from (4.26)–(4.28) and (5.19) we have

$$\frac{\partial^2 J}{\partial \mathbf{w}^2} \;\cong\; \left[(\mathbf{I} - \mathbf{K}_\ell\mathbf{H}_\ell)\mathbf{P}^f\right]^{-1}. \tag{5.22}$$

Substituting this result into (5.18) and using (5.14) gives the iteration

$$\mathbf{w}_{\ell+1} \;=\; \mathbf{w}_\ell \;-\; (\mathbf{I} - \mathbf{K}_\ell\mathbf{H}_\ell)\left\{\mathbf{w}_\ell - \mathbf{w}^f + \mathbf{P}^f\mathbf{H}_\ell^T\mathbf{R}^{-1}\left[\mathbf{h}(\mathbf{w}_\ell) - \mathbf{w}^o\right]\right\}, \tag{5.23}$$

or, in view of (4.27) and (4.29),

$$\mathbf{w}_{\ell+1} \;=\; \mathbf{w}_\ell \;+\; (\mathbf{I} - \mathbf{K}_\ell\mathbf{H}_\ell)(\mathbf{w}^f - \mathbf{w}_\ell) \;+\; \mathbf{K}_\ell\left[\mathbf{w}^o - \mathbf{h}(\mathbf{w}_\ell)\right]. \tag{5.24}$$

Since this requires two linear system solves per iteration, that is, two operations with the gain matrix $\mathbf{K}_\ell$ defined in (5.21), we regroup terms in (5.24) to obtain finally the quasi–Newton iteration

$$\mathbf{w}_{\ell+1} \;=\; \mathbf{w}^f \,+\, \mathbf{K}_\ell \big[\mathbf{w}^o - \mathbf{h}(\mathbf{w}_\ell) + \mathbf{H}_\ell \big(\mathbf{w}_\ell - \mathbf{w}^f\big)\big] \;, \tag{5.25}$$

requiring only one linear system solve per iteration. That is, according to (5.21) one first solves

$$\big(\mathbf{H}_\ell \mathbf{P}^f \mathbf{H}_\ell^T + \mathbf{R}\big)\mathbf{x}_\ell \;=\; \mathbf{w}^o - \mathbf{h}(\mathbf{w}_\ell) + \mathbf{H}_\ell \big(\mathbf{w}_\ell - \mathbf{w}^f\big) \;, \tag{5.26}$$

and then sets

$$\mathbf{w}_{\ell+1} \;=\; \mathbf{w}^f \,+\, \mathbf{P}^f \mathbf{H}_\ell^T \mathbf{x}_\ell \;; \tag{5.27}$$

at convergence ($\mathbf{w}_{\ell+1} \cong \mathbf{w}_\ell \equiv \mathbf{w}_\infty$), one sets

$$\mathbf{w}^a \;=\; \mathbf{w}_\infty \;. \tag{5.28}$$

Observe that for linear observation operators, (5.25) reduces to the (non–iterative) Kalman filter analysis update equation (4.43) upon re–introducing the time index k. Equations (5.26), (5.27) represent a simple nonlinear extension of the Physical–space Statistical Analysis System under development at the NASA/Goddard Data Assimilation Office (da Silva *et al.* 1995).

We note that (5.25) may also be viewed as a simple Picard iteration for minimizing J, i.e., for solving the nonlinear equation

$$\frac{\partial J}{\partial \mathbf{w}} \;=\; \mathbf{0} \;. \tag{5.29}$$

According to (5.14), this equation can be written as

$$\Big[\big(\mathbf{P}^f\big)^{-1} + \mathbf{H}^T(\mathbf{w})\,\mathbf{R}^{-1}\,\mathbf{H}(\mathbf{w})\Big]\mathbf{w} \;=\; \big(\mathbf{P}^f\big)^{-1}\mathbf{w}^f \,+\, \mathbf{H}^T(\mathbf{w})\,\mathbf{R}^{-1}\big[\mathbf{H}(\mathbf{w})\mathbf{w} - \mathbf{h}(\mathbf{w}) + \mathbf{w}^o\big] \;, \tag{5.30}$$

leading to the iteration

$$\Big[\big(\mathbf{P}^f\big)^{-1} + \mathbf{H}_\ell^T \mathbf{R}^{-1} \mathbf{H}_\ell\Big]\mathbf{w}_{\ell+1} \;=\; \big(\mathbf{P}^f\big)^{-1}\mathbf{w}^f \,+\, \mathbf{H}_\ell^T \mathbf{R}^{-1}\big[\mathbf{H}_\ell \mathbf{w}_\ell - \mathbf{h}(\mathbf{w}_\ell) + \mathbf{w}^o\big] \;. \tag{5.31}$$

Defining

$$\mathbf{P}_\ell^{-1} \;\equiv\; \big(\mathbf{P}^f\big)^{-1} + \mathbf{H}_\ell^T \mathbf{R}^{-1} \mathbf{H}_\ell \;, \tag{5.32}$$

this iteration may be written as

$$\mathbf{w}_{\ell+1} \;=\; \mathbf{P}_\ell \big(\mathbf{P}^f\big)^{-1}\mathbf{w}^f \,+\, \mathbf{P}_\ell \mathbf{H}_\ell^T \mathbf{R}^{-1}\big[\mathbf{H}_\ell \mathbf{w}_\ell - \mathbf{h}(\mathbf{w}_\ell) + \mathbf{w}^o\big] \;. \tag{5.33}$$

From (4.26), (4.27) and (5.21), equation (5.32) may be written as

$$\mathbf{P}_\ell \;=\; (\mathbf{I} - \mathbf{K}_\ell \mathbf{H}_\ell)\mathbf{P}^f \;, \tag{5.34}$$

and with (4.29) allows (5.33) to be written as

$$\mathbf{w}_{\ell+1} \;=\; (\mathbf{I}-\mathbf{K}_\ell\mathbf{H}_\ell)\mathbf{w}^f \;+\; \mathbf{K}_\ell\big[\mathbf{H}_\ell\mathbf{w}_\ell - \mathbf{h}(\mathbf{w}_\ell) + \mathbf{w}^o\big]\;, \tag{5.35}$$

from which the iteration (5.25) follows.

It remains to find an approximate expression for the analysis error covariance matrix $\mathbf{P}^a$. By analogy with the linear case, it is tempting to consider $\mathbf{P}_\ell$ defined in (5.34) as an approximate error covariance matrix for the ℓ^{th} iterate $\mathbf{w}_\ell$,** and therefore to consider

$$\mathbf{P}_\infty \;\equiv\; (\mathbf{I}-\mathbf{K}_\infty\mathbf{H}_\infty)\mathbf{P}^f \tag{5.36}$$

as an approximate analysis error covariance matrix at convergence. Note from (5.32) and (5.34) that this $\mathbf{P}_\infty$ is identical to the inverse of the approximate Hessian matrix of (5.19) at convergence. We will see that $\mathbf{P}_\infty$ defined in (5.36) does indeed give $\mathbf{P}^a$ to within a tangent linear approximation; see also Rabier and Courtier (1992, Appendix B).

To see this, linearize the observation model (5.5) about the ℓ^{th} iterate $\mathbf{w}_\ell$:

$$\mathbf{w}^o \;\cong\; \mathbf{h}(\mathbf{w}_\ell) + \mathbf{H}_\ell\big(\mathbf{w}^t - \mathbf{w}_\ell\big) + \boldsymbol{\varepsilon}^o\;, \tag{5.37}$$

where $\mathbf{H}_\ell$ was defined in (5.20) and (5.15). Regrouping terms, rewrite (5.37) as

$$\mathbf{w}^o \;\cong\; \mathbf{H}_\ell\mathbf{w}^t + \big[\mathbf{h}(\mathbf{w}_\ell) - \mathbf{H}_\ell\mathbf{w}_\ell + \boldsymbol{\varepsilon}^o\big]\;. \tag{5.38}$$

Comparing (5.25) with (4.48), it is seen that $\mathbf{w}_{\ell+1}$ in (5.25) is precisely the analysis vector one would obtain from the (linear) Kalman filter upon considering the term $\mathbf{h}(\mathbf{w}_\ell)-\mathbf{H}_\ell\mathbf{w}_\ell$ in (5.38) as an observation error bias, assumed uncorrelated with $\mathbf{w}^t$. The matrix $\mathbf{P}_\ell$ defined in (5.34), by comparison with (4.45), is therefore indeed the error covariance matrix

$$\mathbf{P}^a_{k,\ell} \;\equiv\; \Big\langle \big(\mathbf{w}^t_k - \mathbf{w}_{k,\ell}\big)\big(\mathbf{w}^t_k - \mathbf{w}_{k,\ell}\big)^T \,\Big|\, \mathcal{W}^o_k \Big\rangle \tag{5.39}$$

for the linearized observation model (5.38), and $\mathbf{P}_\infty$ defined in (5.36) is the corresponding error covariance matrix for the "converged" linearized observation model

$$\begin{aligned}\mathbf{w}^o \;&\cong\; \mathbf{H}_\infty\mathbf{w}^t + \big[\mathbf{h}(\mathbf{w}_\infty) - \mathbf{H}_\infty\mathbf{w}_\infty + \boldsymbol{\varepsilon}^o\big] \\ &=\; \mathbf{H}(\mathbf{w}^a)\mathbf{w}^t \;+\; \big[\mathbf{h}(\mathbf{w}^a) - \mathbf{H}(\mathbf{w}^a)\mathbf{w}^a + \boldsymbol{\varepsilon}^o\big]\;.\end{aligned} \tag{5.40}$$

To summarize, the locally iterated extended Kalman filter proceeds at observation times t_k as follows. First, assume the prior density $p\big(\mathbf{w}^t_k \big|\, \mathcal{W}^o_{k-1}\big)$ is Gaussian. Then calculate the mode of the density $p\big(\mathbf{w}^t_k \big|\, \mathcal{W}^o_k\big)$ by carrying (5.25) or an equivalent iteration to convergence, denoting the result by $\mathbf{w}_{k,\infty}$. Approximate the conditional mean $\mathbf{w}^a_k$ by setting $\mathbf{w}^a_k = \mathbf{w}_{k,\infty}$. Finally, approximate the observation model by

$$\mathbf{w}^o_k \;=\; \mathbf{H}_k(\mathbf{w}^a_k)\mathbf{w}^t_k \;+\; \big[\mathbf{h}_k(\mathbf{w}^a_k) - \mathbf{H}_k(\mathbf{w}^a_k)\mathbf{w}^a_k + \boldsymbol{\varepsilon}^o\big]\;, \tag{5.41}$$

**This $\mathbf{P}_\ell$ need not be calculated since it does not appear in (5.25), nor in (5.26)–(5.27).

assume that $\mathbf{h}_k(\mathbf{w}_k^a) - \mathbf{H}_k(\mathbf{w}_k^a)\mathbf{w}_k^a$ is not correlated with $\mathbf{w}_k^t$, and thereby calculate $\mathbf{P}_k^a$ defined in (4.4) by the standard update equation (4.45), in which $\mathbf{H}_k = \mathbf{H}_k(\mathbf{w}_k^a)$ is now defined in (5.15) and $\mathbf{K}_k = \mathbf{K}_k(\mathbf{w}_k^a)$ is now defined by

$$\mathbf{K}_k \;=\; \mathbf{P}_k^f\,\mathbf{H}_k^T(\mathbf{w}_k^a)\Big[\mathbf{H}_k(\mathbf{w}_k^a)\,\mathbf{P}_k^f\,\mathbf{H}_k^T(\mathbf{w}_k^a) \;+\; \mathbf{R}_k\Big]^{-1}. \tag{5.42}$$

Note that, unlike the linear case, in the nonlinear case $\mathbf{P}_k^a$ depends upon the observations themselves, through dependence upon $\mathbf{w}_k^a$. The forecast step (4.8), (4.10) proceeds as usual.

The approximate analysis update described here has not yet been tested fully for large–scale Earth Science data assimilation problems, and the extent to which the approximations involved are reasonable is not yet known. From a theoretical point of view, approximation of the conditional mean by a conditional mode seems less than ideal. Furthermore, nonlinear observation operators arise mainly from remote–sensing devices, whose observations (e.g., radiances) often contain much redundant information (Joiner and da Silva 1997). Thus from a practical viewpoint also, the computational expense of three–dimensional analysis iterations may sometimes not be warranted by the data themselves. For both of these reasons, research directed toward the assimilation of *retrieved* remotely–sensed data products which are related linearly to state variables has begun recently (Joiner and da Silva 1997), following the retrieval error analysis of Rodgers (1990).

5.3 Lognormal observation errors

We have seen that in the estimation–theoretic approach to data assimilation problems, a stochastic–dynamic model, a stochastic observation model, and explicit assumptions on the nature of the probability densities involved in these models form the essential ingredients of the problem statement. Actual algorithms, often necessarily approximate, only ensue once the stochastic problem is formulated completely.

So far we have only considered Gaussian errors, and cited the BLUE property of the Kalman filter in case the only knowledge at one's disposal is the first two moments of the errors. Additional information is sometimes available, however, and if so it can and should be used. For example, R. Ménard (personal communication) has recently obtained evidence that measurement (retrieval) errors for the mixing ratio of several atmospheric trace constituents observed from limb sounders on board the Upper Atmosphere Research Satellite (UARS) tend to be *lognormally* distributed, rather than normally (Gaussian) distributed.[††] In addition there is evidence that forecast mixing ratio errors from a transport model designed to assimilate trace constituent data (Ménard *et al.* 1995; Lyster *et al.* 1997) also tend to be lognormally distributed. In fact the lognormal distribution arises quite naturally for the concentration of trace constituents themselves, according to the theory of successive random dilutions (Ott 1995, Chapters 8, 9). Here we show very briefly how the standard Kalman filter analysis equations can be modified to accommodate lognormally–distributed errors. Similar arguments can be used to develop analysis equations for other densities that

[††]See Appendix B for relationships between the first two moments of the normal and lognormal densities.

are related to the Gaussian density. For more general densities, one may revert back to the general result (4.11).

First we remark that strictly nonnegative quantities, such as salinity, temperature, wind speed, and mixing ratio of water vapor or other atmospheric constituents, as well as remotely–sensed radiance measurements, and therefore errors in these quantities, cannot be strictly Gaussian–distributed (although they may be approximately so), since the Gaussian density assigns positive probability to negative values of these quantities. As a result, observation models such as (4.2) or (5.5) may not be appropriate under the usual assumption of Gaussian measurement error $\boldsymbol{\varepsilon}^o$. For example, if $\mathbf{w}^t$ is a vector of mixing ratios then each element of $\mathbf{w}^t$ must be nonnegative, while if $\boldsymbol{\varepsilon}^o$ is assumed Gaussian then according to the *model* (4.2) there is a nonzero probability of recording a negative observation, even though the *actual* observations are all nonnegative.

A simple example illustrates how this mismatch between model and reality can result in negative (and therefore incorrect) analyses. Consider a forecast vector $\mathbf{w}^f = \left[\mathrm{w}_1^f, \mathrm{w}_2^f \right]^T$ of dimension $n=2$, with $\mathrm{w}_1^f \geqslant 0$ and $\mathrm{w}_2^f \geqslant 0$, having error covariance matrix

$$\mathbf{P}^f \;=\; \left[\begin{array}{cc} a & \rho\sqrt{ab} \\ \rho\sqrt{ab} & b \end{array} \right] , \tag{5.43}$$

with $a>0$, $b>0$, and $|\rho|<1$. Suppose there is a direct observation w_1^o of w_1^t according to (4.2), so that $\mathbf{H}=[1\;\;0]$, and let $R=r>0$. From (4.43) and (4.44) one obtains

$$\mathrm{w}_1^a \;=\; \frac{1}{a+r}\Big(r\,\mathrm{w}_1^f \;+\; a\mathrm{w}_1^o \Big) , \tag{5.44}$$

so that $\mathrm{w}_1^a \geqslant 0$ if $\mathrm{w}_1^o \geqslant 0$, while

$$\mathrm{w}_2^a \;=\; \mathrm{w}_2^f \;+\; \frac{\rho\sqrt{ab}}{a+r}\Big(\mathrm{w}_1^o \;-\; \mathrm{w}_1^f \Big) , \tag{5.45}$$

so it can happen that $\mathrm{w}_2^a < 0$ if

$$\rho\Big(\mathrm{w}_1^o \;-\; \mathrm{w}_1^f \Big) \;<\; 0 \,. \tag{5.46}$$

If $\mathrm{w}_2^f = 0$, then the simple condition (5.46) alone results in a negative analysis $\mathrm{w}_2^a < 0$. If $\mathrm{w}_2^f > 0$, then $\mathrm{w}_2^a < 0$ is obtained if either

$$\frac{\rho}{\sqrt{\dfrac{a}{b}} + \dfrac{r}{\sqrt{ab}}} \;>\; \frac{\mathrm{w}_2^f}{\mathrm{w}_1^f \;-\; \mathrm{w}_1^o} > 0 \,, \tag{5.47}$$

or

$$\frac{\rho}{\sqrt{\dfrac{a}{b}} + \dfrac{r}{\sqrt{ab}}} \;<\; \frac{\mathrm{w}_2^f}{\mathrm{w}_1^f \;-\; \mathrm{w}_1^o} < 0 \,. \tag{5.48}$$

Thus a negative analysis occurs for sufficiently small $\left|\dfrac{\mathrm{w}_2^f}{\mathrm{w}_1^f - \mathrm{w}_1^o}\right|$, sufficiently small ratios a/b and r^2/ab, and sufficiently large $|\rho|$.

For nonnegative fields such as mixing ratios, then, in place of the observation model (4.2), let us model the observations $\mathbf{w}^o$ according to

$$\log \mathbf{w}^o \;=\; \mathbf{H} \log \mathbf{w}^t \,+\, \log \boldsymbol{\varepsilon}^o \,, \tag{5.49}$$

where the logarithm is applied componentwise to the vector elements, we assume the elements of $\mathbf{w}^t$ and $\boldsymbol{\varepsilon}^o$ are nonnegative, and we omit the time index k. This observation model assumes the observations are also nonnegative since

$$\mathbf{w}^o \;=\; \exp\bigl(\mathbf{H} \log \mathbf{w}^t \,+\, \log \boldsymbol{\varepsilon}^o\bigr) \,. \tag{5.50}$$

In fact, defining the logarithmically interpolated state $\mathbf{z}^t$ by

$$\log \mathbf{z}^t \;=\; \mathbf{H} \log \mathbf{w}^t \,, \tag{5.51}$$

(5.49) is equivalent to

$$\mathrm{w}_j^o \;=\; \epsilon_j^o \, \mathrm{z}_j^t \,, \qquad j = 1, 2, \ldots, p \,, \tag{5.52}$$

so that ϵ_j^o is just the *relative error* of the j^{th} observation.

Assuming that $\boldsymbol{\varepsilon}^o$ and $\mathbf{z}^t$ are independent and that the observations are not biased, from (5.52) it follows that

$$\Bigl\langle \epsilon_j^o \Bigr\rangle \;=\; 1 \,, \qquad j = 1, 2, \ldots, p \,. \tag{5.53}$$

Now assume that the density $p(\boldsymbol{\varepsilon}^o)$ is lognormal, so that $p(\log \boldsymbol{\varepsilon}^o)$ is normal, and denote the covariance matrix of $\boldsymbol{\varepsilon}^o$ by $\mathbf{R}$:

$$R_{ij} \;=\; \Bigl\langle (\epsilon_i^o - 1)\Bigl(\epsilon_j^o - 1\Bigr)\Bigr\rangle \,. \tag{5.54}$$

From (B.8) and (B.9) we then have

$$\log \boldsymbol{\varepsilon}^o \;\sim\; N(\mathbf{b}^o, \mathbf{B}^o) \,, \tag{5.55}$$

where

$$B_{ij}^o \;=\; \log\,(1 + R_{ij}) = \log \Bigl\langle \epsilon_i^o \epsilon_j^o \Bigr\rangle \,, \tag{5.56}$$

and

$$b_j^o \;=\; -\tfrac{1}{2} B_{jj}^o = \log \Bigl\langle \Bigl(\epsilon_j^o\Bigr)^2 \Bigr\rangle^{-\frac{1}{2}} \,. \tag{5.57}$$

Equations (5.55)–(5.57) complete the definition of the observation model (5.49).

Now take $\mathbf{v}^t \equiv \log \mathbf{w}^t$ as the state vector and assume the prior density $p\left(\mathbf{w}_k^t \middle| \mathcal{W}_{k-1}^o\right)$ to be lognormal, so that the density $p\left(\mathbf{v}_k^t \middle| \mathcal{W}_{k-1}^o\right)$ is normal with mean and covariance matrix denoted by $\mathbf{v}_k^f$ and $\mathbf{B}_k^f$ respectively. From (5.49) and (5.55) it is seen that we are now in the standard Kalman filter analysis situation, with "observations" $\log \mathbf{w}^o$ and "observation" error bias $\mathbf{b}^o$. From (4.48), (4.44) and (4.45), the optimal (minimum variance, conditional mean) analysis update equations follow:

$$\mathbf{v}^a = \mathbf{v}^f + \mathbf{K}\left(\log \mathbf{w}^o - \mathbf{H}\mathbf{v}^f - \mathbf{b}^o\right) , \tag{5.58}$$

$$\mathbf{K} = \mathbf{B}^f \mathbf{H}^T \left(\mathbf{H}\mathbf{B}^f\mathbf{H}^T + \mathbf{B}^o\right)^{-1} , \tag{5.59}$$

$$\mathbf{B}^a = (\mathbf{I} - \mathbf{K}\mathbf{H})\mathbf{B}^f . \tag{5.60}$$

The optimal analysis $\mathbf{w}^a$ of $\mathbf{w}^t$ itself can be recovered using (B.1):

$$\mathrm{w}_j^a = \exp\left(\mathrm{v}_j^a + \tfrac{1}{2}B_{jj}^a\right) ; \tag{5.61}$$

it is clear that each $\mathrm{w}_j^a \geqslant 0$. The corresponding analysis error covariance is given, if desired, by (B.2):

$$P_{ij}^a = \mathrm{w}_i^a \mathrm{w}_j^a \left(e^{B_{ij}^a} - 1\right) . \tag{5.62}$$

As was the case for nonlinear observation operators, $\mathbf{P}^a$ depends on the observations themselves, through dependence on $\mathbf{w}^a$. The forecast equations may proceed either from the pair $(\mathbf{w}^a, \mathbf{P}^a)$ or, perhaps preferably, directly from $(\mathbf{v}^a, \mathbf{B}^a)$, if the dynamics are based on the evolution equations for $\mathbf{v}^t$ rather than $\mathbf{w}^t$.

6 A simple illustrative example

6.1 Introduction

In the previous section it was seen that the probabilistic assumptions made in formulating the observation model are critical in determining the appropriate analysis update equations, and hence the analysis itself. In § 2 it was argued that the representativeness error term in the observation model may sometimes play an especially important role.

In this section we give a very simple example of an estimation problem arising from continuum dynamics, in which representativeness error can be treated exactly and the exact optimal (conditional mean) state estimate on a well–defined finite–dimensional function space $\mathcal{B}^n$ can actually be calculated. While this example is purposely contrived to make exact treatment possible, the intention is to describe a conceptual framework that may help guide the development of approximate estimation algorithms for more realistic problems. For instance, in this framework we will see that "climatology", defined appropriately, plays a central role in the treatment of representativeness error.

The point–of–view will be to define *first* the space $\mathcal{B}^n$ on which the estimation problem is to be solved, and only *after* doing so to develop an appropriate discretization of the

dynamics. Until now we have *first* assumed the existence of a discrete model, and *then* defined the discrete estimation problem; this led to model error and representativeness error. In the simple example given here we will see that by defining $\mathcal{B}^n$ first and only then discretizing, there will be no model error at all, and the representativeness error can be accounted for exactly. We begin by stating the continuum problem, including all of our assumptions. Afterwards, we will extract and solve an appropriate finite–dimensional problem.

6.2 The continuum problem

Suppose the state $\mathrm{w}=\mathrm{w}(x,t)$ is governed by the one–dimensional scalar advection equation

$$\frac{\partial \mathrm{w}}{\partial t} + u\frac{\partial \mathrm{w}}{\partial x} \;=\; 0\,, \tag{6.1}$$

with constant advection speed u (known and deterministic), periodic boundary conditions on $[0,\,2\pi]$,

$$\mathrm{w}(2\pi,t) \;=\; \mathrm{w}(0,t)\,, \quad t \geqslant t_0\,, \tag{6.2}$$

and real, *unknown* initial condition $\mathrm{w}(x,t_0)=\mathrm{w}_0(x)$ which is considered to be *random*. For notational simplicity we take $\mathrm{w}_0(x)$ to be a random field defined over all of $\mathbb{R}$, not just for $x\in[0,\,2\pi]$. Then by the solution of (6.1), (6.2), we mean the random field

$$\mathrm{w}(x,t) \;=\; \mathrm{w}_0\Big(x - u(t-t_0)\Big)\,, \quad t \geqslant t_0\,, \tag{6.3}$$

provided that

$$\mathrm{w}_0(x+2\pi) \;=\; \mathrm{w}_0(x) \quad \text{for all } x \in \mathbb{R}\,. \tag{6.4}$$

Since $\mathrm{w}_0(x)$ is a random field, condition (6.4) needs to be interpreted in a probabilistic sense, which we will do in Assumption 6.2 below. We assume first that $\mathrm{w}_0(x)$ has finite second moments, so that the mean function

$$\widehat{\mathrm{w}}_0(x) \;\equiv\; \big\langle \mathrm{w}_0(x) \big\rangle\,, \tag{6.5}$$

the covariance function

$$W_0(x_1,x_2) \;\equiv\; \Big\langle \big[\mathrm{w}_0(x_1) - \widehat{\mathrm{w}}_0(x_1)\big]\big[\mathrm{w}_0(x_2) - \widehat{\mathrm{w}}_0(x_2)\big] \Big\rangle\,, \tag{6.6}$$

and the uncentered covariance function

$$Y_0(x_1,x_2) \;\equiv\; \big\langle \mathrm{w}_0(x_1)\mathrm{w}_0(x_2) \big\rangle\,, \tag{6.7}$$

all exist. We will make the following four additional assumptions on the random initial field $\mathrm{w}_0(x)$, which in light of the simple form of the solution (6.3) will ultimately serve to define the function space $\mathcal{B}$ in which the random field $\mathrm{w}(x,t)$ is sought. The first of these is somewhat technical, and therefore will be discussed in detail.

Assumption 6.1 *There exist positive constants α and β independent of x such that*

$$\left\langle \left[\mathrm{w}_0(x+h) - \mathrm{w}_0(x) \right]^2 \right\rangle \; < \; \alpha h^2 \tag{6.8}$$

for all $h < \beta$ and all $x \in \mathbb{R}$.

This assumption can be verified directly for a given $Y_0(x_1, x_2)$ since the left side of (6.8) evaluates to $Y_0(x+h, x+h) - 2Y_0(x+h, x) + Y_0(x, x)$. Thus the assumption implies existence of the second derivative $\partial^2 Y_0(x_1, x_2)/\partial x_1 \partial x_2$ at all diagonal points $x_1 = x_2$, and therefore existence of this second derivative on all of $\mathbb{R} \times \mathbb{R}$ (cf. Jazwinski 1970, p. 63, Corollary 1), which is tacit, for instance, in the usual derivation of wind–wind covariance functions from height–height covariance functions for geostrophically–related height and wind errors on the sphere (cf. Daley 1991, § 5.3). Existence of this (deterministic) derivative in turn implies existence of the (random) derivative $\partial \mathrm{w}_0(x)/\partial x$ in the *mean–square sense* (cf. Jazwinski 1970, Theorem 3.5), which is the sense in which the derivatives in (6.1) are to be interpreted, and also implies that the operations of expectation and differentiation commute (cf. Jazwinski 1970, Theorem 3.6). The latter property is used in the derivation of the continuum Kalman filter equations (Appendix C). Assumption 6.1 is satisfied, in particular, if $\mathrm{w}_0(x)$ is *bandlimited*; cf. Papoulis (1984, p. 332).

Assumption 6.1 implies that the random field $\mathrm{w}_0(x)$ is *almost surely sample continuous* on $\mathbb{R}$ (Loève 1963, p. 520), a fact that will be important in defining observations of the state $\mathrm{w}(x, t)$ later in this subsection. This means that *all realizations* of $\mathrm{w}_0(x)$ are continuous functions on $\mathbb{R}$, except for a set of realizations of probability zero *independent of* x, a strong notion of continuity. Existence of $\partial^2 Y_0(x_1, x_2)/\partial x_1 \partial x_2$ only implies *almost sure continuity* of $\mathrm{w}_0(x)$, referred to in the literature also as continuity *with probability 1*, which means that realizations are continuous at any *given* point x with probability 1. Loève (1963, p. 501) discusses the distinction between the concept of almost sure sample continuity and the weaker concept of almost sure continuity.

In place of Assumption 6.1 one might consider a still weaker assumption such as *mean–square continuity* of $\mathrm{w}_0(x)$,

$$\lim_{h \to 0} \left\langle \left[\mathrm{w}_0(x+h) - \mathrm{w}_0(x) \right]^2 \right\rangle \; = \; 0 \quad \text{for all } x \in \mathbb{R} \,, \tag{6.9}$$

which is equivalent to continuity of $Y_0(x_1, x_2)$ at each diagonal point $x_1 = x_2$ (cf. Jazwinski 1970, Theorem 3.2; Papoulis 1984, p. 225). Realizations of mean–square continuous random fields, however, can be extremely irregular and often display fractal characteristics (cf. Tarantola 1987, Example 7.4), although they are always regular enough to be integrable over finite intervals (Loève 1963, p. 520). While such fields describe a variety of natural phenomena, we employ Assumption 6.1 instead, which, as we will see later, renders *point* observations of $\mathrm{w}(x, t)$ meaningful.

Assumption 6.2 *There exist positive constants γ and δ independent of x such that*

$$\left\langle \left[\mathrm{w}_0(x + 2\pi + h) - \mathrm{w}_0(x) \right]^2 \right\rangle \; < \; \gamma h^2 \tag{6.10}$$

for all $h < \delta$ and all $x \in \mathbb{R}$.

This assumption implies *almost sure sample* 2π*–periodicity* of the random field $\mathrm{w}_0(x)$: this is the sense in which (6.4) is meant in our example. In other words, all realizations of $\mathrm{w}_0(x)$ satisfy (6.4), except for a set of realizations of probability zero independent of x. A weaker (and more frequently encountered) sense, implied by Assumption 6.2, is *mean–square* 2π*–periodicity*:

$$\Big\langle \big[\mathrm{w}_0(x+2\pi) - \mathrm{w}_0(x)\big]^2 \Big\rangle \;=\; 0 \quad \text{for all } x \in \mathbb{R}\;; \tag{6.11}$$

cf. Papoulis (1984, p. 230).

Assumption 6.3 *The random field* $\mathrm{w}_0(x)$ *is Gaussian, that is, for each positive integer* m *and each set of points* $x_1, x_2, \ldots, x_m$ *in* $\mathbb{R}$*, the random variables* $\mathrm{w}_0(x_1)$, $\mathrm{w}_0(x_2)$, $\ldots$, $\mathrm{w}_0(x_m)$ *are jointly Gaussian.*

It follows from (6.3) that in fact the solution $\mathrm{w}(x,t)$ is Gaussian for all time $t \geqslant t_0$.

Assumption 6.4 *The random field* $\mathrm{w}_0(x)$ *is homogeneous (wide–sense stationary when viewed as a stochastic process rather than as a random field). That is, the mean function* $\widehat{\mathrm{w}}_0(x)$ *in* (6.5) *is actually a constant* c*,*

$$\widehat{\mathrm{w}}_0(x) \;\equiv\; \big\langle \mathrm{w}_0(x) \big\rangle = c\;, \tag{6.12}$$

and the covariance function $W_0(x_1,x_2)$ *in* (6.6) *is a function* C *of* $x_1 - x_2$ *alone,*

$$W_0(x_1,x_2) \;=\; C(x_1 - x_2)\;. \tag{6.13}$$

The constant c *and the function* $C(x_1 - x_2)$ *are assumed known.*

From definition (6.6) it follows that $C(x_1-x_2)=C(x_2-x_1)=C(|x_1-x_2|)$; $C(x)$ is an *even* function. In other words, $\mathrm{w}_0(x)$ is also *isotropic*, which is the case for all homogeneous random fields in one dimension.

Now, from (6.3) and (6.12) it follows that the *unconditional mean*

$$\widehat{\mathrm{w}}(x,t) \;\equiv\; \big\langle \mathrm{w}(x,t) \big\rangle \tag{6.14}$$

is a constant for all time $t \geqslant t_0$:

$$\widehat{\mathrm{w}}(x,t) \;=\; \Big\langle \mathrm{w}_0\Big(x - u(t-t_0)\Big) \Big\rangle = c\;. \tag{6.15}$$

From (6.3), (6.6) and (6.13) it follows that the *unconditional covariance function*

$$W(x_1,x_2,t) \;\equiv\; \Big\langle \big[\mathrm{w}(x_1,t) - \widehat{\mathrm{w}}(x_1,t)\big]\big[\mathrm{w}(x_2,t) - \widehat{\mathrm{w}}(x_2,t)\big] \Big\rangle \tag{6.16}$$

is also constant in time and depends only on $x_1 - x_2$:

$$\begin{aligned} W(x_1, x_2, t) &= W_0\Big(x_1 - u(t-t_0),\, x_2 - u(t-t_0)\Big) \\ &= C\Big(\big[x_1 - u(t-t_0)\big] - \big[x_2 - u(t-t_0)\big]\Big) \\ &= C(x_1 - x_2)\,. \end{aligned} \tag{6.17}$$

Therefore $\mathrm{w}(x,t)$ is a homogeneous Gaussian random field for all time $t \geqslant t_0$, with unconditional mean c and unconditional covariance function $C(x_1 - x_2)$ which are both independent of time. Since these statistics are independent of time, we will refer to c as the *climatological mean* and to $C(x_1 - x_2)$ as the *climatological covariance function*. Thus the notion of climatology in this example differs from the conventional one: our usage refers directly to ensemble statistics rather than time–average statistics, and does not require ergodicity.

Replacing x in (6.8) by $x - u(t - t_0)$, it follows from (6.3) that

$$\Big\langle \big[\mathrm{w}(x+h,t) - \mathrm{w}(x,t)\big]^2 \Big\rangle \;<\; \alpha h^2\,, \tag{6.18}$$

for all $h < \beta$, all $x \in \mathbb{R}$, and all $t \geqslant t_0$. Similarly, from (6.3) and (6.10) it follows that

$$\Big\langle \big[\mathrm{w}(x+2\pi+h,t) - \mathrm{w}(x,t)\big]^2 \Big\rangle \;<\; \gamma h^2\,, \tag{6.19}$$

for all $h < \delta$, all $x \in \mathbb{R}$, and all $t \geqslant t_0$. Thus the function space $\mathcal{B}$ is defined to consist of all homogeneous Gaussian random fields $\mathrm{w}(x,t)$ having mean c, covariance function $C(x_1 - x_2)$, and satisfying (6.18) and (6.19). These random fields are almost surely sample continuous, 2π–periodic functions on $\mathbb{R}$. In particular, $\mathcal{B}$ includes the *realizations* of such fields.

Now suppose that noisy linear observations of $\mathrm{w}(x,t)$ are available at discrete instants of time t_k for $k = 1, 2, 3, \ldots$:

$$\mathbf{w}_k^o \;=\; \mathbf{H}_k\, \mathrm{w}(\cdot\,,\, t_k) + \boldsymbol{\varepsilon}_k^m\,. \tag{6.20}$$

Here $\mathbf{w}_k^o$ is a p_k–vector as usual, $\mathbf{H}_k$ denotes a *known*, deterministic p_k–vector–valued bounded linear operator on $\mathcal{B}$ acting on the fundamental interval $[0, 2\pi]$, and we will assume that the measurement error $\boldsymbol{\varepsilon}_k^m$ is a Gaussian p_k–vector uncorrelated with $\mathrm{w}_0(x)$,

$$\Big\langle \Big(\boldsymbol{\varepsilon}_k^m - \langle \boldsymbol{\varepsilon}_k^m \rangle\Big)\, \mathrm{w}_0(x) \Big\rangle \;=\; \mathbf{0}\,, \tag{6.21}$$

for all k and for all $x \in \mathbb{R}$. From (6.3) and (6.21) it follows that $\boldsymbol{\varepsilon}_k^m$ is not correlated with the state $\mathrm{w}(x,t_k)$, nor with the *signal* $\mathbf{H}_k\, \mathrm{w}(\cdot\,,\, t_k)$. The signal is a Gaussian–distributed p_k–vector since it is a linear operation on the state, which is Gaussian (cf. Loève 1963, p. 485). We also assume that the measurement error bias,

$$\widehat{\boldsymbol{\varepsilon}}_k^m \;\equiv\; \langle \boldsymbol{\varepsilon}_k^m \rangle\,, \tag{6.22}$$

and the measurement error covariance matrix,

$$\mathbf{R}_k \;\equiv\; \left\langle (\boldsymbol{\varepsilon}_k^m - \widehat{\boldsymbol{\varepsilon}}_k^m)(\boldsymbol{\varepsilon}_k^m - \widehat{\boldsymbol{\varepsilon}}_k^m)^T \right\rangle , \tag{6.23}$$

are both known. Finally, we assume that the measurement error is white in time.

Let us write the ℓ^{th} element of the vector observation equation (6.20) as

$$(\mathbf{w}_k^o)_\ell \;=\; (\mathbf{s}_k)_\ell \,+\, (\boldsymbol{\varepsilon}_k^m)_\ell \,, \tag{6.24}$$

where $\mathbf{s}_k \equiv \mathbf{H}_k\, \mathrm{w}(\cdot\,,\, t_k)$ is the signal. What we actually observe is a *realization* of the random field $\mathrm{w}(x, t_k)$. It followed from Assumption 6.1 that almost all realizations are *continuous* functions, in a very strict sense. We define $\mathbf{H}_k$ only for the continuous realizations. We have also stipulated that $\mathbf{H}_k$ should be a *bounded* linear operator on $\mathcal{B}$ (acting on the fundamental interval $[0,\, 2\pi]$), a physically natural requirement which means for continuous functions $\mathrm{w}(x, t_k)$ on $[0,\, 2\pi]$ that

$$|(\mathbf{s}_k)_\ell| \;\leqslant\; M_{k,\ell} \max_{0 \leqslant x \leqslant 2\pi} |\mathrm{w}(x, t_k)| \,, \tag{6.25}$$

for some constants $M_{k,\ell}$ independent of $\mathrm{w}(x, t_k)$. It follows that a large class[‡‡] of signals can be represented as

$$(\mathbf{s}_k)_\ell \;=\; \int_0^{2\pi} \mathrm{w}(x, t_k) f_{k\ell}(x) dx \,, \tag{6.26}$$

where the functions $f_{k\ell}(x)$ are any integrable functions on $[0,\, 2\pi]$, since then we can take

$$M_{k,\ell} \;=\; \int_0^{2\pi} |f_{k\ell}(x)| dx \,. \tag{6.27}$$

In particular, *point* observations are permissible: in this case

$$(\mathbf{s}_k)_\ell \;=\; \mathrm{w}(x_\ell, t_k) \,, \tag{6.28}$$

x_ℓ for $\ell = 1,\, 2,\, \ldots\,,\, p_k$ denoting observation points at time t_k, which is obtained by setting $f_{k\ell}(x) = \delta(x - x_\ell)$ in (6.26), δ denoting the Dirac δ–function, for which $M_{k,\ell} = 1$ in (6.25);[*] see also Papoulis (1984, pp. 235–245, 271–283). More generally, the weighting functions $f_{k\ell}(x)$ can be considered as *aperture functions* (cf. Daley 1993) or *averaging kernels* (Backus and Gilbert 1970), and in this example are assumed known, inasmuch as $\mathbf{H}_k$ is assumed known.

[‡‡] The entire class is described by the Riesz representation theorem for continuous functions (e.g., Royden 1968, p. 310).

[*] Had we assumed only mean–square continuity (6.9) of $\mathrm{w}_0(x)$ rather than Assumption 6.1, then the most we could conclude about integrability of $\mathrm{w}(x, t_k)$ is *almost sure sample square–integrability* (Loève 1963, p. 520), i.e., that almost all realizations lie in the space $L_2[0,\, 2\pi]$ of square–integrable functions on $[0,\, 2\pi]$. The Riesz representation theorem for L_p spaces (e.g., Royden 1968, p. 121) would then imply that the functions $f_{k\ell}(x)$ must also lie in $L_2[0,\, 2\pi]$, thus *precluding* point observations since the Dirac δ–function is not square–integrable. In other words, point observations would tell us nothing about the realization of $\mathrm{w}(x, t_k)$ in this case.

Under the stated assumptions, the conditional mean forecasts and analyses at times t_k, and their corresponding conditional covariance functions, are given (exactly) by the continuum version of the Kalman filter (see Appendix C). These are defined by

$$\mathrm{w}_k^f(x) \equiv \left\langle \mathrm{w}(x,t_k) \middle| \mathcal{W}_{k-1}^o \right\rangle , \tag{6.29}$$

$$\mathrm{w}_k^a(x) \equiv \left\langle \mathrm{w}(x,t_k) \middle| \mathcal{W}_{k}^o \right\rangle , \tag{6.30}$$

$$P_k^f(x_1,x_2) \equiv \left\langle \Big[\mathrm{w}(x_1,t_k) - \mathrm{w}_k^f(x_1)\Big]\Big[\mathrm{w}(x_2,t_k) - \mathrm{w}_k^f(x_2)\Big] \middle| \mathcal{W}_{k-1}^o \right\rangle , \tag{6.31}$$

$$P_k^a(x_1,x_2) \equiv \left\langle \Big[\mathrm{w}(x_1,t_k) - \mathrm{w}_k^a(x_1)\Big]\Big[\mathrm{w}(x_2,t_k) - \mathrm{w}_k^a(x_2)\Big] \middle| \mathcal{W}_{k}^o \right\rangle . \tag{6.32}$$

At time t_0 we have

$$\mathrm{w}_0^a(x) = \big\langle \mathrm{w}_0(x) \big\rangle = c , \tag{6.33}$$

$$P_0^a(x_1,x_2) = W_0(x_1,x_2) = C(x_1-x_2) . \tag{6.34}$$

At times $t_k > t_0$, $\mathrm{w}_k^f(x)$ and $P_k^f(x_1,x_2)$ are given by the solutions $\widehat{\mathrm{w}}(x,t_k)$ and $P(x_1,x_2,t_k)$, respectively, of the differential equations

$$\frac{\partial \widehat{\mathrm{w}}}{\partial t} + u\frac{\partial \widehat{\mathrm{w}}}{\partial x} = 0 , \tag{6.35}$$

$$\frac{\partial P}{\partial t} + u\frac{\partial P}{\partial x_1} + u\frac{\partial P}{\partial x_2} = 0 , \tag{6.36}$$

with initial conditions

$$\widehat{\mathrm{w}}(x,t_{k-1}) = \mathrm{w}_{k-1}^a(x) , \tag{6.37}$$

$$P(x_1,x_2,t_{k-1}) = P_{k-1}^a(x_1,x_2) ; \tag{6.38}$$

thus we have simply

$$\mathrm{w}_k^f(x) = \mathrm{w}_{k-1}^a\Big(x - u(t_k - t_{k-1})\Big) , \tag{6.39}$$

$$P_k^f(x_1,x_2) = P_{k-1}^a\Big(x_1 - u(t_k - t_{k-1}),\, x_2 - u(t_k - t_{k-1})\Big) . \tag{6.40}$$

For the analysis update we have

$$\mathrm{w}_k^a(x) = \mathrm{w}_k^f(x) + \mathbf{K}_k(x)\Big(\mathbf{w}_k^o - \mathbf{H}_k\,\mathrm{w}_k^f - \widehat{\boldsymbol{\varepsilon}}_k^m\Big) . \tag{6.41}$$

Here the innovation vector $\mathbf{w}_k^o - \mathbf{H}_k\mathrm{w}_k^f - \widehat{\boldsymbol{\varepsilon}}_k^m$ is a (column) p_k–vector as usual, and the gain $\mathbf{K}_k(x)$ is the (row) p_k–vector function of x defined by

$$\mathbf{K}_k(x) = \Big[\mathbf{H}_{2k}P_k^f(x\ ,\cdot)\Big]^T\Big\{\mathbf{H}_{1k}\Big[\mathbf{H}_{2k}P_k^f(\cdot\ ,\cdot)\Big]^T + \mathbf{R}_k\Big\}^{-1} , \tag{6.42}$$

where $\mathbf{H}_{2k}P_k^f(\cdot\ ,\cdot)$ denotes the (column) p_k–vector function of x_1 obtained by acting with $\mathbf{H}_k$ on the x_2 variable of $P_k^f(x_1,x_2)$ and $\mathbf{H}_{1k}$ denotes the action of $\mathbf{H}_k$ on the x_1 variable;

the innovation covariance (in braces) to be inverted is a $p_k \times p_k$ matrix as usual, and the x–dependence of $\mathbf{K}_k(x)$ arises solely from the factor $\Big[\mathbf{H}_{2k} P_k^f(x\ ,\ \cdot)\Big]^T$. Finally, the analysis error covariance function P_k^a is given by

$$P_k^a(x_1, x_2) \ = \ P_k^f(x_1, x_2) - \mathbf{K}_k(x_1)\mathbf{H}_{1k} P_k^f(\cdot\ , x_2)\ . \tag{6.43}$$

The update equations (6.41), (6.42), (6.43) appear in Appendix C in a slightly different notation as equations (C.26), (C.27), (C.35) for the more general case of a *vector* state in *several* space variables.

Equations (6.39)–(6.43), with initial conditions (6.33), (6.34), constitute the complete solution of the continuum filtering problem we have posed. Even for this very simple problem, however, for general $\mathbf{H}_k$ they cannot be solved exactly on a computer, because the analysis update equations (6.41)–(6.43) require access to P_k^f *as a function.* Rather than propose an approximate means of solving these filter equations, we now extract from the continuum filtering problem one which can in fact be solved exactly on a computer. We remark here that the solution of the continuum filtering problem did not require the homogeneity Assumption 6.4; the initial conditions (6.33) and (6.34) need only be the (unconditional) mean $\big\langle \mathrm{w}_0(x) \big\rangle$ and covariance $W_0(x_1, x_2)$ of $\mathrm{w}(x, t_0)$, whatever they may be. The new filtering problem will make explicit use of the homogeneity assumption on the initial field.

6.3 A finite-dimensional problem

By Assumption 6.1, the random field $\mathrm{w}_0(x)$ is mean–square differentiable, therefore mean–square continuous, and therefore mean–square integrable; cf. Jazwinski (1970, § 3.4). It follows that we can define the *random Fourier coefficients*

$$a_j \ \equiv \ \frac{1}{2\pi} \int_0^{2\pi} \Big[\mathrm{w}_0(x) - \big\langle \mathrm{w}_0(x) \big\rangle \Big] e^{-\mathrm{i} j x} dx\ . \tag{6.44}$$

These are Gaussian random variables in light of Assumption 6.3 (cf. Loève 1963, p. 485); a_0 is real and $a_{-j} = \bar{a}_j$ since $\mathrm{w}_0(x)$ is real. Since the operations of expectation and mean–square integration commute (cf. Jazwinski 1970, Theorem 3.8), it follows from (6.44) that

$$\langle a_j \rangle \ = \ 0 \quad \text{for all } j\ . \tag{6.45}$$

Further, the homogeneity Assumption 6.4 implies that

$$\langle a_i \bar{a}_j \rangle \ = \ 0 \quad \text{for } i \neq j\ , \tag{6.46}$$

while

$$\big\langle |a_j|^2 \big\rangle \ = \ \frac{1}{2\pi} \int_0^{2\pi} C(x) e^{-\mathrm{i} j x} dx \quad \text{for all } j\ . \tag{6.47}$$

In other words, the variances $\langle |a_j|^2 \rangle$ are just the Fourier coefficients of the function $C(x)$ defined by (6.6) and (6.13):

$$C(x) \;=\; \sum_{j=-\infty}^{\infty} \langle |a_j|^2 \rangle e^{\mathrm{i}jx} \,, \tag{6.48}$$

where the convergence is pointwise and uniform in x.† We saw earlier that $C(x)$ is an even function, $C(-x)=C(x)$, which follows also from (6.48) since $a_{-j}=\bar{a}_j$. Thus the expansion (6.48) can be rewritten as

$$C(x) \;=\; \langle (a_0)^2 \rangle + 2 \sum_{j=1}^{\infty} \langle |a_j|^2 \rangle \, \cos(jx) \,. \tag{6.49}$$

Similarly, according to (6.12) and (6.44), the Gaussian random variables a_j themselves are the Fourier coefficients of the random field $\mathrm{w}_0(x) - \langle \mathrm{w}_0(x) \rangle = \mathrm{w}_0(x) - c$:

$$\mathrm{w}_0(x) \;=\; c + \sum_{j=-\infty}^{\infty} a_j e^{\mathrm{i}jx} \,, \tag{6.50}$$

where the convergence is both mean–square and almost sure (cf. Loève 1963, pp. 485–486; Papoulis 1984, pp. 301–302; Yaglom 1987, § 2.7, Example 3). In view of (6.3) we have also

$$\mathrm{w}(x,t) \;=\; c + \sum_{j=-\infty}^{\infty} a_j e^{\mathrm{i}j[x \,-\, u(t \,-\, t_0)]} \,. \tag{6.51}$$

This simple expansion gives rise naturally to a finite–dimensional filtering problem, as follows.

Suppose we are interested only in a "large–scale" analysis of $\mathrm{w}(x,t)$, which we will define to be the conditional mean of the first $N \geqslant 0$ waves in the expansion (6.51) of $\mathrm{w}(x,t)$. Thus let Π denote the operator from $\mathcal{B}$ to $\mathcal{B}^n$, $n=2N+1$, defined for random fields $\mathrm{w}_0(x)$ satisfying Assumptions 6.1 to 6.4 by the truncated random Fourier series

$$\Pi \mathrm{w}_0(x) \;=\; c + \sum_{j=-N}^{N} a_j e^{\mathrm{i}jx} \,; \tag{6.52}$$

Π is a projection operator,

$$\Pi^2 \;=\; \Pi \,. \tag{6.53}$$

Then the *large–scale field* $\mathrm{w}^\ell(x,t)$ is defined to be

$$\mathrm{w}^\ell(x,t) \;\equiv\; \Pi \mathrm{w}(x,t) = c + \sum_{j=-N}^{N} a_j e^{\mathrm{i}j[x \,-\, u(t \,-\, t_0)]} \,, \tag{6.54}$$

†$C(x)$ is twice–differentiable according to Assumption 6.1, in particular the first derivative $dC(x)/dx$ is continuous, and it follows from (6.11) and (6.12) that $C(x)$ is 2π–periodic. Standard results from Fourier analysis imply pointwise and uniform convergence in this case.

the latter equality following from (6.51); $\mathrm{w}^\ell = \mathrm{w}^t$ in the notation of the previous sections. The *small–scale field* $\mathrm{w}^s(x,t)$ is then

$$\mathrm{w}^s(x,t) \;\equiv\; (\mathrm{I}-\Pi)\mathrm{w}(x,t) = \sum_{|j|>N} a_j e^{\mathrm{i}j[x-u(t-t_0)]} \,, \tag{6.55}$$

so that

$$\mathrm{w}(x,t) \;=\; \mathrm{w}^\ell(x,t) + \mathrm{w}^s(x,t) \,. \tag{6.56}$$

The estimation problem will be to calculate the *large–scale* forecasts and analyses,

$$\mathrm{w}^{\ell f}_k(x) \;\equiv\; \left\langle \mathrm{w}^\ell(x,t_k) \middle| \mathcal{W}^o_{k-1} \right\rangle \,, \tag{6.57}$$

$$\mathrm{w}^{\ell a}_k(x) \;\equiv\; \left\langle \mathrm{w}^\ell(x,t_k) \middle| \mathcal{W}^o_{k} \right\rangle \,, \tag{6.58}$$

along with the corresponding error covariances

$$P^{\ell f}_k(x_1,x_2) \;\equiv\; \left\langle \left[\mathrm{w}^\ell(x_1,t_k) - \mathrm{w}^{\ell f}_k(x_1)\right]\left[\mathrm{w}^\ell(x_2,t_k) - \mathrm{w}^{\ell f}_k(x_2)\right] \middle| \mathcal{W}^o_{k-1} \right\rangle \,, \tag{6.59}$$

$$P^{\ell a}_k(x_1,x_2) \;\equiv\; \left\langle \left[\mathrm{w}^\ell(x_1,t_k) - \mathrm{w}^{\ell a}_k(x_1)\right]\left[\mathrm{w}^\ell(x_2,t_k) - \mathrm{w}^{\ell a}_k(x_2)\right] \middle| \mathcal{W}^o_{k} \right\rangle \,. \tag{6.60}$$

Before proceeding, observe first that

$$\mathrm{w}^a_k(x) - \mathrm{w}^{\ell a}_k(x) \;=\; \left\langle \mathrm{w}^s(x,t_k) \middle| \mathcal{W}^o_k \right\rangle \,, \tag{6.61}$$

according to (6.30), (6.56) and (6.58); a similar relation holds for $\mathrm{w}^f_k(x) - \mathrm{w}^{\ell f}_k(x)$. The conditional expectation on the right side of (6.61) does not vanish in general: writing a single aperture function $f_{k\ell}(x)$ as a Fourier series

$$f_{k\ell}(x) \;=\; \sum_{m=-\infty}^{\infty} b_m e^{\mathrm{i}mx} \,, \tag{6.62}$$

the integral in (6.26) evaluates to

$$\int_0^{2\pi} \mathrm{w}(x,t_k) f_{k\ell}(x) dx \;=\; 2\pi \sum_{j=-\infty}^{\infty} a_j b_{-j} e^{-\mathrm{i}ju(t_k-t_0)} \,. \tag{6.63}$$

Thus the right side of (6.61) vanishes only if for all the aperture functions $b_j = 0$ for all $|j|>N$, assuming the original field $\mathrm{w}_0(x)$ has power at all small scales ($\langle |a_j|^2 \rangle \neq 0$ for all $|j|>N$); cf. (6.55). In fact, the closer the observations are to point observations, the closer the spectrum $\{b_j\}$ is to being flat ($|b_j| = $ constant). However, the *unconditional* expectation $\langle \mathrm{w}^s(x,t_k) \rangle$ vanishes according to (6.45) and (6.55):

$$\left\langle \mathrm{w}^s(x,t_k) \right\rangle \;=\; 0 \,. \tag{6.64}$$

Small–scale information therefore arises solely from the observations in this simple example. From (6.61) and (6.64) we see that the large–scale analysis $\mathrm{w}_k^{\ell a}(x)$ amounts to discarding the small–scale information contained in all current and past observations.

To solve the filtering problem (6.57)–(6.60), first write the observation equation (6.20) as

$$\mathbf{w}_k^o \;=\; \mathbf{H}_k\,\mathrm{w}^\ell(\cdot\,,t_k) + \boldsymbol{\varepsilon}_k^r + \boldsymbol{\varepsilon}_k^m\;, \tag{6.65}$$

where the *representativeness error* $\boldsymbol{\varepsilon}_k^r$ is given by

$$\boldsymbol{\varepsilon}_k^r \;=\; \mathbf{H}_k\,\mathrm{w}^s(\cdot\,,t_k)\;, \tag{6.66}$$

and has mean zero according to (6.64). Since the coefficients a_j are Gaussian–distributed, as is the measurement error $\boldsymbol{\varepsilon}_k^m$, and since $\mathbf{H}_k$ is a linear operator, the *large–scale signal* $\mathbf{H}_k\,\mathrm{w}^\ell(\cdot\,,t_k)$ is still a Gaussian random vector and so is the *observation error* $\boldsymbol{\varepsilon}_k^o \equiv \boldsymbol{\varepsilon}_k^r + \boldsymbol{\varepsilon}_k^m$; cf. (2.14). The large–scale signal, representativeness error and measurement error are also mutually uncorrelated in view of (6.21), (6.46), (6.54) and (6.55). Therefore the analysis update equations (6.41)–(6.43) still hold if we replace w_k^f, w_k^a, P_k^f and P_k^a by $\mathrm{w}_k^{\ell f}$, $\mathrm{w}_k^{\ell a}$, $P_k^{\ell f}$ and $P_k^{\ell a}$, and if we replace the measurement error covariance matrix $\mathbf{R}_k$ by the observation error covariance matrix. Let us now calculate this matrix.

According to (6.16), (6.17), (6.49) and (6.55),

$$\begin{aligned}\Big\langle \mathrm{w}^s(x_1,t)\mathrm{w}^s(x_2,t)\Big\rangle \;&=\; 2\sum_{j>N}\big\langle |a_j|^2\big\rangle\,\cos\big[j(x_1-x_2)\big] \\ &\equiv\; C^s(x_1-x_2) \\ &=\; C(x_1-x_2) - C^\ell(x_1-x_2)\;;\end{aligned} \tag{6.67}$$

$C^s(x_1-x_2)$ is the *small–scale climatological covariance* and

$$\begin{aligned}C^\ell(x_1-x_2) \;&\equiv\; \Big\langle \Big[\mathrm{w}^\ell(x_1,t)-c\Big]\Big[\mathrm{w}^\ell(x_2,t)-c\Big]\Big\rangle \\ &=\; \big\langle (a_0)^2\big\rangle + 2\sum_{j=1}^{N}\big\langle |a_j|^2\big\rangle\,\cos\big[j(x_1-x_2)\big]\end{aligned} \tag{6.68}$$

is the *large–scale climatological covariance*. Note that since we have assumed the function $C(x_1-x_2)$ to be known, for example as any of the traditional isotropic covariance models (cf. Daley 1991), then $C^s(x_1-x_2)$ can be calculated by use of the latter equalities in (6.67) and (6.68). Thus the observation error covariance matrix is

$$\begin{aligned}\Big\langle \Big(\boldsymbol{\varepsilon}_k^r + \boldsymbol{\varepsilon}_k^m - \big\langle \boldsymbol{\varepsilon}_k^m\big\rangle\Big)\Big(\boldsymbol{\varepsilon}_k^r + \boldsymbol{\varepsilon}_k^m - \big\langle \boldsymbol{\varepsilon}_k^m\big\rangle\Big)^T\Big\rangle \;&=\; \big\langle \boldsymbol{\varepsilon}_k^r(\boldsymbol{\varepsilon}_k^r)^T\big\rangle + \mathbf{R}_k \\ &=\; \mathbf{H}_{1k}\big[\mathbf{H}_{2k}C^s(\cdot-\cdot)\big]^T + \mathbf{R}_k\;,\end{aligned} \tag{6.69}$$

which is the sum of the representativeness error covariance matrix and the measurement error covariance matrix. Here the notation is as in (6.42).

The analysis update equations are therefore given by

$$\mathrm{w}_k^{\ell a}(x) \;=\; \mathrm{w}_k^{\ell f}(x) + \mathbf{K}_k^{\ell}(x)\Big(\mathbf{w}_k^{o} - \mathbf{H}_k \mathrm{w}_k^{\ell f} - \widehat{\boldsymbol{\varepsilon}}_k^{m}\Big)\;, \tag{6.70}$$

$$\mathbf{K}_k^{\ell}(x) \;=\; \Big[\mathbf{H}_{2k} P_k^{\ell f}(x\;,\,\cdot)\Big]^T \left(\mathbf{H}_{1k}\left\{\mathbf{H}_{2k}\Big[P_k^{\ell f}(\cdot\,,\,\cdot) + C^s(\cdot - \cdot)\Big]\right\}^T + \mathbf{R}_k\right)^{-1}\;, \tag{6.71}$$

$$P_k^{\ell a}(x_1, x_2) \;=\; P_k^{\ell f}(x_1, x_2) - \mathbf{K}_k^{\ell}(x_1)\mathbf{H}_{1k} P_k^{\ell f}(\cdot\,,\,x_2)\;. \tag{6.72}$$

Except for those giving rise to the representativeness error covariance matrix, each of the integrals represented by $\mathbf{H}_k$, $\mathbf{H}_{1k}$, and $\mathbf{H}_{2k}$ in these equations can be evaluated exactly: they act on functions in the finite–dimensional space $\mathcal{B}^n$, or $\mathcal{B}^n \times \mathcal{B}^n$, and by orthogonality of the basis (sinusoidal) functions, the aperture functions $f_{k\ell}(x)$ in these integrals may be truncated to their projections $\Pi f_{k\ell}(x)$ in evaluating the integrals; cf. (6.20), (6.24), (6.26) and (6.63). The integrals involved in calculating the representativeness error covariance matrix would have to be calculated numerically, to some specified degree of accuracy. The initial analysis equations, as in (6.33) and (6.34), are given by

$$\mathrm{w}_0^{\ell a}(x) \;\equiv\; \Big\langle \mathrm{w}^{\ell}(x, t_0)\Big\rangle = c\;, \tag{6.73}$$

$$P_0^{\ell a}(x_1, x_2) \;\equiv\; \left\langle \Big[\mathrm{w}^{\ell}(x_1, t_0) - c\Big]\Big[\mathrm{w}^{\ell}(x_2, t_0) - c\Big]\right\rangle = C^{\ell}(x_1 - x_2)\;; \tag{6.74}$$

cf. (6.68).

It remains to calculate the evolution of the conditional mean and covariance between observation times. Since the projection operator Π commutes with $u\partial/\partial x$ and $\partial/\partial t$, from (6.1) and (6.54) we find that

$$\frac{\partial \mathrm{w}^{\ell}}{\partial t} + u\frac{\partial \mathrm{w}^{\ell}}{\partial x} \;=\; 0\;. \tag{6.75}$$

Taking the conditional expectation with respect to $\mathcal{W}_{k-1}^{o}$ in (6.75), it follows that $\mathrm{w}_k^{\ell f}(x)$ defined in (6.57) is given by the solution $\widehat{\mathrm{w}}^{\ell}(x, t_k)$ of the equation

$$\frac{\partial \widehat{\mathrm{w}}^{\ell}}{\partial t} + u\frac{\partial \widehat{\mathrm{w}}^{\ell}}{\partial x} \;=\; 0\;, \tag{6.76}$$

starting from initial condition $\widehat{\mathrm{w}}^{\ell}(x, t_{k-1}) = \mathrm{w}_{k-1}^{\ell a}(x)$. Similarly, $P_k^{\ell f}(x_1, x_2)$ is given by the solution $P^{\ell}(x_1, x_2, t_k)$ of the equation

$$\frac{\partial P^{\ell}}{\partial t} + u\frac{\partial P^{\ell}}{\partial x_1} + u\frac{\partial P^{\ell}}{\partial x_2} \;=\; 0\;, \tag{6.77}$$

starting from initial condition $P^{\ell}(x_1, x_2, t_{k-1}) = P_{k-1}^{\ell a}(x_1, x_2)$. Since the initial conditions for (6.76) and (6.77) lie in the finite–dimensional spaces $\mathcal{B}^n$ and $\mathcal{B}^n \times \mathcal{B}^n$ respectively, these equations may also be solved exactly, for example by evolving the spectral coefficients directly. Thus the filtering problem (6.57)–(6.60) is solved.

We note that in this simple example there is no "aliasing" of the small–scale information contained in the observations onto the large–scale analyses and forecasts: *by definition* [see

(6.54), (6.57), (6.58)], these extract only the large–scale information contained in the observations, the small–scale information being simply discarded [see (6.61) and the discussion thereafter]. Further, since $\mathrm{w}^s(x,t)$ lies in $\mathcal{B}\backslash\mathcal{B}^n$ (the complement of $\mathcal{B}^n$ in $\mathcal{B}$), so does the conditional expectation on the right side of (6.61), so that

$$\Pi\left\langle \mathrm{w}^s(x,t_k)\middle|\mathcal{W}_k^o\right\rangle \;=\; 0\;, \tag{6.78}$$

which follows from (6.53) and (6.55). Therefore, since $\Pi\mathrm{w}^{\ell a}(x,t)=\mathrm{w}^{\ell a}(x,t)$, operating on (6.61) with Π gives the relation

$$\mathrm{w}_k^{\ell a}(x) \;=\; \Pi\mathrm{w}_k^a(x)\;; \tag{6.79}$$

that is, the analyses resulting from the finite–dimensional algorithm (6.70)–(6.72), (6.76), (6.77) are just the projections onto $\mathcal{B}^n$ of those resulting from the infinite–dimensional algorithm (6.39)–(6.43). Similar relations hold for $\mathrm{w}_k^{\ell f}$, $P_k^{\ell a}$ and $P_k^{\ell f}$.

6.4 Concluding remarks

Let us now summarize the results of this very simple example, and use them to provide a way of thinking about realistic geophysical data assimilation problems. First of all, we have seen that by defining $\mathcal{B}^n$ first, and only then developing a discretization, an exact, implementable filter algorithm has been designed. This filter algorithm involves, perhaps surprisingly, no model error term. Had a different discretization of the dynamics been chosen, i.e., one incompatible with the discrete estimation problem imposed by the definition of $\mathcal{B}^n$, then model error would have arisen. Such model error in this simple example could perhaps be modeled stochastically by considering the leading–order terms in the truncation error expansion of the chosen discretization. While in principle it appears best to define $\mathcal{B}^n$ first, then to define the finite–dimensional estimation problem to be solved, and only as the final step to develop an appropriate discretization, in the real world this will be a practical impossibility at least for some time, since large–scale geophysical models take many years to develop: currently we are usually *given* a discrete model, then asked to develop a data assimilation algorithm. For this reason alone, model error is inevitable.

There are many other sources of model error, however. In our simple example, the absence of model error was due to the invariance of the continuum dynamics under the action of the projection operator Π from $\mathcal{B}$ to $\mathcal{B}^n$. If the advection speed had not been constant, this invariance would no longer have held. In this case, a different choice of $\mathcal{B}^n$ could perhaps ameliorate model error. For most *nonlinear* problems, it is unlikely that for *any* choice of $\mathcal{B}^n$ one could develop a projection operator under which the dynamics would be invariant. For instance, energy– and enstrophy–cascade processes (cf. Gauthier *et al.* 1995; Tanguay *et al.* 1995) would likely lead to model error; this error could possibly be modeled stochastically (cf. Leith 1990). Assumptions made from the outset in the governing continuum dynamics, such as the hydrostatic assumption and the traditional shallowness approximations (Phillips 1966) in the atmospheric primitive equations also lead to model error. Finally, stochastic forcing arises from uncertain parameters in physical parameterizations and boundary conditions. Errors from all these sources will ultimately have to

be modeled, and the models tuned by adaptive procedures such as that suggested by Dee (1995).

Concerning representativeness error, it should be noted that the exact treatment in our example was enabled by the homogeneity of the random field $\mathrm{w}(x,t)$: this property led to the absence of correlation between the large–scale part of the signal $\mathbf{H}_k\mathrm{w}^\ell$ and the representativeness error $\mathbf{H}_k\mathrm{w}^s$, and also to state–independence of the mean and covariance matrix of the representativeness error. While geophysical fields generally do not have this property, it is sometimes possible to introduce a change of coordinates such that homogeneity or isotropy holds approximately (e.g., Desroziers and Lafore 1993; Carton and Hackert 1990; Derber and Rosati 1989; Vanmarcke 1983, p. 81). The "kernel" $P^{\ell f} + C^s$ in (6.71) upon which the observation operator $\mathbf{H}_k$ acts generally contains power at all spatial scales. Current–generation global analysis systems for numerical weather prediction also involve such a kernel (Parrish and Derber 1992; Heckley *et al.* 1992), but truncate it at finite spectral resolution and instead lump the representativeness error covariance matrix together with the measurement error covariance matrix. Equation (6.71) suggests that it may be more natural to sum the forecast error covariance model $\left(\sim P^{\ell f}\right)$ together with a small–scale climatological covariance model $(\sim C^s)$ in accounting for representativeness error, resulting in a covariance model with power at all scales. In principle this is possible by modeling the sum directly as a covariance *function* (with power at all scales), rather than as a truncated spectral expansion. The Physical–space Statistical Analysis System (da Silva *et al.* 1995) is one effort being developed along these lines. By the analogy between model error and representativeness error drawn in § 2, it appears that the stochastic forcing $\boldsymbol{\varepsilon}_k^t$ in (2.9) must also contain power at all scales in general, and therefore should be considered as a random field rather than as a random vector.

In the example it was also seen to be important to evaluate the action of the integrals represented by the observation operators $\mathbf{H}_k$ *as actual integrals*, as opposed to, say, the simple interpolations carried out in conventional optimal interpolation schemes (e.g., McPherson *et al.* 1979; Lorenc 1981). This may be possible in operational practice, but only if the discrete function space $\mathcal{B}^n$ is defined precisely, for instance, only if we know precisely what is meant by the grid–point values of a numerical prediction model. Precise definition of $\mathcal{B}^n$ and implementation of integral observation operators is likely to be important for properly assimilating satellite radiances or retrieved products representing averages over regions of the spatial domain, and even for "point" observations such as those obtained from most *in situ* measurements. Of course, we do not know the aperture functions or averaging kernels precisely. Ultimately it may be necessary to parameterize them and to estimate the parameters during the data assimilation process.

Finally, we reiterate the role of the Gaussian assumptions made here (as well as in operational data assimilation systems), along with the independence of the measurement error from the signal. The Gaussian assumptions lead, as we have seen, to true conditional mean (minimum variance) estimation procedures if in fact they are correct. Gaussian assumptions can be checked, at least in part, by monitoring statistics of the observed–minus–forecast residuals. As demonstrated in § 5.3, for alternative densities related simply to the Gaussian density, it is straightforward to modify the estimation algorithm appropriately, by

a change of dependent variable. In the simple example given here, the dynamics are invariant under the change of variable $v = \log w$, and the analysis algorithm would be trivially modified as suggested in § 5.3.

Independence of measurement error from the signal depends on the measuring device itself. For most *in situ* measurements, the assumption of independence may be justified, provided the devices are properly calibrated. For retrieved satellite products, a method is currently being developed by Joiner and da Silva (1997) in part to ensure this independence.

Acknowledgments

The author is indebted to all of his colleagues at the Data Assimilation Office, who have exerted a great deal of influence on the content of this article. In particular, discussions with R. Ménard and A. da Silva led to the lognormal development in § 5 and to the example in § 6, respectively. The material in Appendix B was kindly provided by G. Gaspari. Thoughtful reviews of the manuscript were provided by R. Daley, D. Dee, G. Gaspari, P. Lyster, R. Ménard, K. Omidvar, A. da Silva, N. S. Sivakumaran, R. Todling, and R. Yang. The difficult text processing effort was carried out by Q. Philpot, N. S. Sivakumaran, and M. Sienkiewicz. Financial support was provided by NASA Headquarters, through Grants 578–41–52–20 and 579–21–32–20, and through the EOS Interdisciplinary Science projects led by R. Rood and M. Schoeberl.

Appendix A
Conditional Probability Densities and Expectations

Here we review only those facts about conditional probability densities and expectations that allow us to give a self–contained proof of the equivalence of conditional mean estimation and minimum variance estimation; cf. (3.4). Background material can be found in most textbooks on probability theory. We make no notational distinction here between a random variable and its realizations. All integrals defined below are assumed to exist.

If $\boldsymbol{z}$ is a random n–vector, its expected value (or mean, or first moment) is the vector $\langle \boldsymbol{z} \rangle$ whose i^{th} element is defined by

$$\langle \boldsymbol{z} \rangle_i \equiv \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \zeta_i \, p_z(\boldsymbol{\zeta}) d\zeta_1 \cdots d\zeta_n \,, \tag{A.1}$$

where p_z is the probability density function of $\boldsymbol{z}$. We abbreviate this definition of $\langle \boldsymbol{z} \rangle$ by the notation

$$\langle \boldsymbol{z} \rangle = \int \boldsymbol{\zeta} \, p_z(\boldsymbol{\zeta}) d\boldsymbol{\zeta} \,, \tag{A.2}$$

where the integration is over all of $\mathbb{R}^n$. If $f(\boldsymbol{z})$ is a deterministic function of $\boldsymbol{z}$, then we also

have

$$\langle f(\boldsymbol{z})\rangle \;=\; \int f(\boldsymbol{\zeta})p_z(\boldsymbol{\zeta})d\boldsymbol{\zeta}\;. \tag{A.3}$$

Now denote the first m elements of $\boldsymbol{z}$ by the m–vector $\boldsymbol{x}$, and the remaining $n-m$ elements by the $(n-m)$–vector $\boldsymbol{y}$, so that $\boldsymbol{z}=\left[\boldsymbol{x}^T,\boldsymbol{y}^T\right]^T$. The joint probability density function $p_{x,y}(\boldsymbol{x},\boldsymbol{y})$ is then defined as

$$p_{x,y}(\boldsymbol{x},\boldsymbol{y}) \;\equiv\; p_z(\boldsymbol{z})\;, \tag{A.4}$$

and the marginal densities $p_x(\boldsymbol{x})$ and $p_y(\boldsymbol{y})$ are defined by

$$p_x(\boldsymbol{x}) \;\equiv\; \int p_{x,y}(\boldsymbol{x},\boldsymbol{\eta})d\boldsymbol{\eta}\;, \tag{A.5}$$

$$p_y(\boldsymbol{y}) \;\equiv\; \int p_{x,y}(\boldsymbol{\xi},\boldsymbol{y})d\boldsymbol{\xi}\;, \tag{A.6}$$

the former integral being over $\mathbb{R}^{n-m}$ and the latter over $\mathbb{R}^m$. The expected values $\langle\boldsymbol{x}\rangle$ and $\langle\boldsymbol{y}\rangle$ are then given by

$$\langle\boldsymbol{x}\rangle \;=\; \int \boldsymbol{\xi}p_x(\boldsymbol{\xi})d\boldsymbol{\xi}\;, \tag{A.7}$$

$$\langle\boldsymbol{y}\rangle \;=\; \int \boldsymbol{\eta}p_y(\boldsymbol{\eta})d\boldsymbol{\eta}\;, \tag{A.8}$$

the former integral being over $\mathbb{R}^m$ and the latter over $\mathbb{R}^{n-m}$, so that $\langle\boldsymbol{z}\rangle=\left[\langle\boldsymbol{x}\rangle^T,\langle\boldsymbol{y}\rangle^T\right]^T$.

The conditional density of $\boldsymbol{x}$ given $\boldsymbol{y}$, written $p_{x|y}(\boldsymbol{x}|\boldsymbol{y})$, is defined as

$$p_{x|y}(\boldsymbol{x}|\boldsymbol{y}) \;\equiv\; \frac{p_{x,y}(\boldsymbol{x},\boldsymbol{y})}{p_y(\boldsymbol{y})}\;, \tag{A.9}$$

by analogy with the usual definition of the probability of occurrence of an event A given the occurrence of an event B,

$$Pr(A|B) \;\equiv\; \frac{Pr(A\cap B)}{Pr(B)}\;. \tag{A.10}$$

Note that if $\boldsymbol{x}$ and $\boldsymbol{y}$ are independent, that is, if $p_{x,y}(\boldsymbol{x},\boldsymbol{y})=p_x(\boldsymbol{x})p_y(\boldsymbol{y})$, then the intuitive result $p_{x|y}(\boldsymbol{x}|\boldsymbol{y})=p_x(\boldsymbol{x})$ follows from (A.9). The expected value of $\boldsymbol{x}$ given $\boldsymbol{y}$, written $\langle\boldsymbol{x}|\,\boldsymbol{y}\rangle$, is defined by

$$\langle\boldsymbol{x}|\,\boldsymbol{y}\rangle \;\equiv\; \int \boldsymbol{\xi}\,p_{x|y}(\boldsymbol{\xi}|\boldsymbol{y})d\boldsymbol{\xi}\;, \tag{A.11}$$

and is a function of the random vector $\boldsymbol{y}$. However, if $\boldsymbol{x}$ and $\boldsymbol{y}$ are independent, by comparing (A.7) and (A.11) it follows that $\langle\boldsymbol{x}|\,\boldsymbol{y}\rangle=\langle\boldsymbol{x}\rangle$.

Now, since $\langle \boldsymbol{x} | \, \boldsymbol{y} \rangle$ is a function of $\boldsymbol{y}$, upon taking the expectation with respect to $\boldsymbol{y}$ it follows from (A.3) that

$$\langle \langle \boldsymbol{x} | \, \boldsymbol{y} \rangle \rangle \;=\; \int \langle \boldsymbol{x} | \, \boldsymbol{\eta} \rangle p_y(\boldsymbol{\eta}) d\boldsymbol{\eta} \; . \tag{A.12}$$

On the other hand, from (A.7), (A.5) and (A.9) we have

$$\langle \boldsymbol{x} \rangle \;=\; \int \boldsymbol{\xi} \left[\int p_{x|y}(\boldsymbol{\xi} | \, \boldsymbol{\eta}) p_y(\boldsymbol{\eta}) d\boldsymbol{\eta} \right] d\boldsymbol{\xi} \; . \tag{A.13}$$

Interchanging the order of integration here gives

$$\langle \boldsymbol{x} \rangle \;=\; \int \left[\int \boldsymbol{\xi} \, p_{x|y}(\boldsymbol{\xi} | \, \boldsymbol{\eta}) d\boldsymbol{\xi} \right] p_y(\boldsymbol{\eta}) d\boldsymbol{\eta} \, , \tag{A.14}$$

or, from (A.11),

$$\langle \boldsymbol{x} \rangle \;=\; \int \langle \boldsymbol{x} | \, \boldsymbol{\eta} \rangle p_y(\boldsymbol{\eta}) d\boldsymbol{\eta} \; . \tag{A.15}$$

Comparing (A.12) and (A.15) shows that

$$\langle \langle \boldsymbol{x} | \, \boldsymbol{y} \rangle \rangle \;=\; \langle \boldsymbol{x} \rangle \, , \tag{A.16}$$

a fundamental identity we will need.

Another basic identity is that

$$\Big\langle \mathbf{g}^T(\boldsymbol{y}) \, \boldsymbol{x} \Big| \boldsymbol{y} \Big\rangle \;=\; \mathbf{g}^T(\boldsymbol{y}) \langle \boldsymbol{x} | \, \boldsymbol{y} \rangle \, , \tag{A.17}$$

if the vector $\mathbf{g}$ is a function of $\boldsymbol{y}$ alone. This follows directly from the definition (A.11) of conditional expectation.

Now we establish the relationship (3.4). Denote by $\boldsymbol{\mu}_k$ the conditional mean of the state $\mathbf{w}_k^t$ given the observations $\mathcal{W}_\ell^o$,

$$\boldsymbol{\mu}_k \;\equiv\; \big\langle \mathbf{w}_k^t \big| \, \mathcal{W}_\ell^o \big\rangle \; . \tag{A.18}$$

From (A.16) it follows that

$$\Big\langle L(\boldsymbol{\varepsilon}_k) \Big\rangle \;=\; \Big\langle \Big\langle L(\boldsymbol{\varepsilon}_k) \Big| \mathcal{W}_\ell^o \Big\rangle \Big\rangle \, , \tag{A.19}$$

where $L(\boldsymbol{\varepsilon}_k)$ was defined in (3.3). Substituting (3.2) and (3.3) into (A.19) yields

$$\Big\langle L(\boldsymbol{\varepsilon}_k) \Big\rangle \;=\; \Big\langle \Big\langle \big(\mathbf{w}_k^t - \mathbf{w}_k^e\big)^T \mathbf{S} \big(\mathbf{w}_k^t - \mathbf{w}_k^e\big) \Big| \mathcal{W}_\ell^o \Big\rangle \Big\rangle \; . \tag{A.20}$$

Adding and subtracting $\boldsymbol{\mu}_k$ in (A.20) gives

$$\Big\langle L(\boldsymbol{\varepsilon}_k) \Big\rangle \;=\; \Big\langle \Big\langle \Big[\big(\mathbf{w}_k^t - \boldsymbol{\mu}_k\big) + \big(\boldsymbol{\mu}_k - \mathbf{w}_k^e\big) \Big]^T \mathbf{S} \Big[\big(\mathbf{w}_k^t - \boldsymbol{\mu}_k\big) + \big(\boldsymbol{\mu}_k - \mathbf{w}_k^e\big) \Big] \Big| \mathcal{W}_\ell^o \Big\rangle \Big\rangle \; . \tag{A.21}$$

Since $\boldsymbol{\mu}_k$ is a function only of $\mathcal{W}_\ell^o$ according to (A.18), and since the estimate $\mathbf{w}_k^e$ was assumed to be a function only of the observations $\mathcal{W}_\ell^o$ also, from (A.17) we have

$$\Big\langle (\boldsymbol{\mu}_k - \mathbf{w}_k^e)^T \mathbf{S} (\mathbf{w}_k^t - \boldsymbol{\mu}_k) \Big| \mathcal{W}_\ell^o \Big\rangle \;=\; (\boldsymbol{\mu}_k - \mathbf{w}_k^e)^T \mathbf{S} \Big\langle (\mathbf{w}_k^t - \boldsymbol{\mu}_k) \Big| \mathcal{W}_\ell^o \Big\rangle \,. \tag{A.22}$$

But

$$\Big\langle (\mathbf{w}_k^t - \boldsymbol{\mu}_k) \Big| \mathcal{W}_\ell^o \Big\rangle \;=\; \mathbf{0} \,, \tag{A.23}$$

according to definition (A.18), so the expression (A.22) vanishes. Therefore we can write (A.21) as

$$\Big\langle L(\boldsymbol{\varepsilon}_k) \Big\rangle \;=\; \Big\langle \Big\langle (\mathbf{w}_k^t - \boldsymbol{\mu}_k)^T \mathbf{S} (\mathbf{w}_k^t - \boldsymbol{\mu}_k) \Big| \mathcal{W}_\ell^o \Big\rangle \Big\rangle \;+\; \Big\langle \Big\langle (\boldsymbol{\mu}_k - \mathbf{w}_k^e)^T \mathbf{S} (\boldsymbol{\mu}_k - \mathbf{w}_k^e) \Big| \mathcal{W}_\ell^o \Big\rangle \Big\rangle \,. \tag{A.24}$$

Using (A.16) and definition (3.3) again, this becomes

$$\Big\langle L(\boldsymbol{\varepsilon}_k) \Big\rangle \;=\; \Big\langle L\big(\mathbf{w}_k^t - \boldsymbol{\mu}_k\big) \Big\rangle + \Big\langle L\big(\boldsymbol{\mu}_k - \mathbf{w}_k^e\big) \Big\rangle \,. \tag{A.25}$$

The first term on the right side of (A.25) is independent of the estimate $\mathbf{w}_k^e$. The second term is minimized uniquely (since $\mathbf{S}$ was assumed positive definite) by the choice (3.4), that is, by setting

$$\mathbf{w}_k^e \;=\; \boldsymbol{\mu}_k \,, \tag{A.26}$$

in which case the second term vanishes. Thus, $\langle L(\boldsymbol{\varepsilon}_k) \rangle$ is minimized uniquely by the conditional mean $\boldsymbol{\mu}_k$, and the value of $\langle L(\boldsymbol{\varepsilon}_k) \rangle$ at the minimum is $\big\langle L\big(\mathbf{w}_k^t - \boldsymbol{\mu}_k\big) \big\rangle$.

Appendix B
The Lognormal Distribution

Here we describe the relationships between the first two moments of the multivariate normal (Gaussian) and lognormal probability densities.

Suppose $\mathbf{v} \in R^n$ is normally distributed with mean $\langle \mathbf{v} \rangle$ and covariance matrix $\mathbf{B}$, denoted $\mathbf{v} \sim N(\langle \mathbf{v} \rangle, \mathbf{B})$. If the components w_j of a vector $\mathbf{w}$ are defined by $\mathrm{w}_j = \exp(\mathrm{v}_j)$ for $j = 1, 2, \ldots, n$, then $\mathbf{w}$ is said to be lognormally distributed, written $\mathbf{w} \sim LN(\langle \mathbf{w} \rangle, \mathbf{P})$. The mean vector $\langle \mathbf{w} \rangle$ is given by

$$\langle \mathrm{w}_j \rangle \;=\; \exp\Big(\langle \mathrm{v}_j \rangle + \tfrac{1}{2} B_{jj} \Big) \,, \tag{B.1}$$

and the covariance matrix $\mathbf{P}$ by

$$P_{jk} \;=\; \langle \mathrm{w}_j \rangle \langle \mathrm{w}_k \rangle \Big(e^{B_{jk}} - 1 \Big) \,. \tag{B.2}$$

The mean vector and covariance matrix characterize the multivariate lognormal density completely, as is also the case for the normal density.

To see (B.1), recall that the characteristic function $\phi_{\rm V}(\boldsymbol{\omega}) \equiv \left\langle \exp\left(\mathrm{i}\,\boldsymbol{\omega}^T \mathbf{v}\right)\right\rangle$ of $\mathbf{v}$ is given by

$$\phi_{\rm V}(\boldsymbol{\omega}) \;=\; \exp\Big(\mathrm{i}\,\boldsymbol{\omega}^T \langle \mathbf{v}\rangle - \tfrac{1}{2}\boldsymbol{\omega}^T \mathbf{B}\boldsymbol{\omega}\Big)\;, \tag{B.3}$$

as shown in many elementary probability texts. Substituting into (B.3) the vector $\boldsymbol{\omega}=\boldsymbol{\omega}_j$ defined by $\boldsymbol{\omega}_j = -\mathrm{i}\,\mathbf{e}_j$, where $\mathrm{i}=\sqrt{-1}$ and $\mathbf{e}_j$ denotes the $j^{\rm th}$ column of the $n\times n$ identity matrix, gives (B.1) immediately. Similarly, substituting $\boldsymbol{\omega}=\boldsymbol{\omega}_{jk} \equiv \boldsymbol{\omega}_j + \boldsymbol{\omega}_k$ into (B.3) gives

$$\langle \mathrm{w}_j \mathrm{w}_k\rangle \;=\; \exp\Big[\langle \mathrm{v}_j\rangle + \langle \mathrm{v}_k\rangle + \tfrac{1}{2}(B_{jj} + B_{kk} + 2B_{jk})\Big] = \langle \mathrm{w}_j\rangle\langle \mathrm{w}_k\rangle \exp(B_{jk})\;, \tag{B.4}$$

so that

$$P_{jk} \;\equiv\; \Big\langle \Big(\mathrm{w}_j - \langle \mathrm{w}_j\rangle\Big)\Big(\mathrm{w}_k - \langle \mathrm{w}_k\rangle\Big)\Big\rangle = \langle \mathrm{w}_j \mathrm{w}_k\rangle - \langle \mathrm{w}_j\rangle\langle \mathrm{w}_k\rangle\;, \tag{B.5}$$

from which (B.2) follows.

Straightforward algebraic calculations from (B.1) and (B.2) show that, if we are given $\mathbf{w} \sim LN(\langle \mathbf{w}\rangle, \mathbf{P})$ and define $\mathbf{v} = \log \mathbf{w}$ (componentwise), then $\mathbf{v} \sim N(\langle \mathbf{v}\rangle, \mathbf{B})$, with

$$\langle \mathrm{v}_j\rangle \;=\; \log \langle \mathrm{w}_j\rangle - \tfrac{1}{2} B_{jj}\;, \tag{B.6}$$

and

$$B_{jk} \;=\; \log\left(1 + \frac{P_{jk}}{\langle \mathrm{w}_j\rangle\langle \mathrm{w}_k\rangle}\right). \tag{B.7}$$

In the special case that $\langle \mathrm{w}_j\rangle = 1$ for all j, discussed in § 5.3, one has simply

$$\langle \mathrm{v}_j\rangle \;=\; -\tfrac{1}{2} B_{jj}\;, \tag{B.8}$$

and

$$B_{jk} \;=\; \log\left(1 + P_{jk}\right); \tag{B.9}$$

in particular $\langle \mathrm{v}_j\rangle \leqslant 0$ for each j.

Appendix C
Filtering Theory on the Continuum

While the discrete theory developed in §§ 2–5 had the virtue of keeping the mathematics fairly simple, it did not allow for an adequate treatment of model error or representativeness error. In § 6 it was seen that continuum theory is required to address these issues

fully, and is therefore developed here. In this appendix we carry out no discretization, so there will be no representativeness error; instead this error is treated in § 6. We also do not consider stochastic forcing of the continuum dynamics, primarily to simplify the mathematical development; see, however, the discussion in § 6.4. The observations will be supposed linear, since the main difficulties surrounding the treatment of nonlinear observations were described already in § 5.2. The continuum dynamics will be nonlinear, however, so that we can highlight the role of closure approximations. Such approximations do not arise in the linear case.

While essentially exact filters for nonlinear dynamics can be obtained through Monte Carlo approaches (Evensen 1994; Ménard 1994), here we will invoke the *second–moment closure* (third– and higher–moment discard) *approximation*. The development will be brief and formal. See Cohn (1993) and references therein for more detailed discussion of this approximation. Curtain (1975) and Omatu and Seinfeld (1989) summarize rigorous treatments of *linear* stochastic PDEs in estimation theory, establishing the relationship between rigorous and formal approaches.

Let the m–vector state $\mathbf{w}=\mathbf{w}(\mathbf{x},t)$ satisfy a system of m nonlinear PDEs ($m=1$ for scalar, univariate dynamics):

$$\frac{\partial \mathbf{w}}{\partial t} + \boldsymbol{f}(\mathbf{w}) \;=\; \mathbf{0}\,, \tag{C.1}$$

where $\boldsymbol{f}(\mathbf{w})=\boldsymbol{f}(\mathbf{w};\partial/\partial\mathbf{x})$ denotes an m–vector partial differential operator acting on the spatial variables $\mathbf{x}$ of the state $\mathbf{w}(\mathbf{x},t)$, which is assumed to lie in some function space $\mathcal{B}$ for each time t. The (unknown) initial condition $\mathbf{w}(\mathbf{x},t_0)=\mathbf{w}_0(\mathbf{x})\in\mathcal{B}$ will be considered to be a *random field* (e.g., Yaglom 1987; Vanmarcke 1983) with *known* mean

$$\widehat{\mathbf{w}}_0(\mathbf{x}) \;\equiv\; \Big\langle \mathbf{w}_0(\mathbf{x}) \Big\rangle\,, \tag{C.2}$$

and *known* covariance function

$$\mathbf{W}_0(\mathbf{x}_1,\mathbf{x}_2) \;\equiv\; \left\langle \Big[\mathbf{w}_0(\mathbf{x}_1)-\widehat{\mathbf{w}}_0(\mathbf{x}_1)\Big] \Big[\mathbf{w}_0(\mathbf{x}_2)-\widehat{\mathbf{w}}_0(\mathbf{x}_2)\Big]^T \right\rangle\,; \tag{C.3}$$

this is an $m\times m$ matrix function of two sets of spatial variables $\mathbf{x}_1$ and $\mathbf{x}_2$, and by definition satisfies the symmetry property

$$\mathbf{W}_0^T(\mathbf{x}_1,\mathbf{x}_2) \;=\; \mathbf{W}_0(\mathbf{x}_2,\mathbf{x}_1)\,. \tag{C.4}$$

Now suppose a p_k–vector of observations $\mathbf{w}_k^o$ taken at discrete instants of time t_k, $k=1,\,2,\,3,\,\ldots$, is related linearly to the state variables and corrupted by additive noise:

$$\mathbf{w}_k^o \;=\; \mathbf{H}_k\,\mathbf{w}(\cdot\,,t_k) + \boldsymbol{\varepsilon}_k^m\,, \tag{C.5}$$

where $\mathbf{H}_k$ is a linear operator on the function space $\mathcal{B}$. The p_k–vector measurement error $\boldsymbol{\varepsilon}_k^m$ is assumed to be Gaussian, white in time, and independent of the state $\mathbf{w}(\mathbf{x},t_k)$ and the initial field $\mathbf{w}_0(\mathbf{x})$. The measurement error bias

$$\widehat{\boldsymbol{\varepsilon}}_k^m \;\equiv\; \langle \boldsymbol{\varepsilon}_k^m \rangle\,, \tag{C.6}$$

a p_k–vector, and covariance matrix

$$\mathbf{R}_k \;\equiv\; \left\langle (\varepsilon_k^m - \widehat{\varepsilon}_k^m)(\varepsilon_k^m - \widehat{\varepsilon}_k^m)^T \right\rangle , \tag{C.7}$$

a $p_k \times p_k$ matrix, are both assumed to be known.

To solve the filtering problem, first we need to develop an equation for the evolution between times t_{k-1} and t_k of the *conditional mean*

$$\widehat{\mathbf{w}} \;=\; \widehat{\mathbf{w}}(\mathbf{x},t) \equiv \left\langle \mathbf{w}(\mathbf{x},t) \middle| \mathcal{W}_{k-1}^o \right\rangle . \tag{C.8}$$

If $\boldsymbol{f}=\boldsymbol{f}(\mathbf{w})$ were linear (or quadratic) in $\mathbf{w}$, the equation developed here would be exact. In the second–moment closure approximation, we expand $\boldsymbol{f}(\mathbf{w})$ about $\widehat{\mathbf{w}}$, assuming $\boldsymbol{f}$ is twice continuously differentiable with respect to $\mathbf{w}$, and retain terms up to second order only:

$$f_\ell(\mathbf{w}) \;=\; f_\ell(\widehat{\mathbf{w}}) + \mathbf{L}_\ell(\widehat{\mathbf{w}})\mathbf{e} + \tfrac{1}{2}\,\mathrm{tr}\left[\mathbf{F}_\ell(\widehat{\mathbf{w}})\mathbf{e}\mathbf{e}^T\right] , \tag{C.9}$$

for $\ell=1, 2, \ldots, m$. Here f_ℓ denotes the ℓ^{th} element of the vector $\boldsymbol{f}$; $\mathbf{e}$ is the m–vector function

$$\mathbf{e} \;=\; \mathbf{e}(\mathbf{x},t) \equiv \mathbf{w}(\mathbf{x},t) - \widehat{\mathbf{w}}(\mathbf{x},t) ; \tag{C.10}$$

$\mathbf{L}_\ell$ is the ℓ^{th} *row* of the *tangent linear operator* $\mathbf{L}$, the $m \times m$ matrix partial differential operator whose $(\ell,j)^{\text{th}}$ element is given by

$$L_{\ell j}(\widehat{\mathbf{w}}) \;\equiv\; \left.\frac{\partial f_\ell(\mathbf{w})}{\partial \mathrm{w}_j}\right|_{\mathbf{w}=\widehat{\mathbf{w}}} ; \tag{C.11}$$

$\mathbf{F}_\ell$ is the *Hessian operator*, an $m \times m$ matrix partial differential operator whose $(i,j)^{\text{th}}$ component is

$$\left[\mathbf{F}_\ell(\widehat{\mathbf{w}})\right]_{ij} \;\equiv\; \left.\frac{\partial^2 f_\ell(\mathbf{w})}{\partial \mathrm{w}_i \partial \mathrm{w}_j}\right|_{\mathbf{w}=\widehat{\mathbf{w}}} ; \tag{C.12}$$

and "tr" denotes the *trace* of a matrix. Taking conditional expectations in (C.9) gives

$$\left\langle f_\ell(\mathbf{w}) \middle| \mathcal{W}_{k-1}^o \right\rangle \;=\; f_\ell(\widehat{\mathbf{w}}) + \tfrac{1}{2}\,\mathrm{tr}\left[\mathbf{F}_\ell(\widehat{\mathbf{w}})\mathbf{V}\right] , \tag{C.13}$$

where

$$\mathbf{V} \;=\; \mathbf{V}(\mathbf{x},t) \equiv \left\langle \mathbf{e}(\mathbf{x},t)\, \mathbf{e}^T(\mathbf{x},t) \middle| \mathcal{W}_{k-1}^o \right\rangle \tag{C.14}$$

is the (conditional) *variance function*; $\mathbf{V}$ is a symmetric $m \times m$ matrix function whose diagonal elements are the conditional variances of the m state variables and whose off–diagonal elements are conditional cross–covariances between different state variables evaluated at a given spatial location $\mathbf{x}$.

Abbreviating (C.13) by the slightly abusive notation

$$\left\langle \boldsymbol{f}(\mathbf{w}) \middle| \mathcal{W}_{k-1}^{o} \right\rangle \;=\; \boldsymbol{f}(\widehat{\mathbf{w}}) + \tfrac{1}{2}\,\mathrm{tr}\Big[\mathbf{F}(\widehat{\mathbf{w}})\mathbf{V}\Big]\,, \tag{C.15}$$

and taking conditional expectations in (C.1) leads to the *mean equation*

$$\frac{\partial \widehat{\mathbf{w}}}{\partial t} + \boldsymbol{f}(\widehat{\mathbf{w}}) + \tfrac{1}{2}\,\mathrm{tr}\Big[\mathbf{F}(\widehat{\mathbf{w}})\mathbf{V}\Big] \;=\; \mathbf{0}\,. \tag{C.16}$$

The mean equation is similar to the original dynamical equation (C.1), which only governs the evolution of individual realizations, but incorporates a *nonlinear bias correction term* $\frac{1}{2}\mathrm{tr}\,(\mathbf{FV})$. The truncated expansion of $\boldsymbol{f}(\mathbf{w})$ in (C.9) is *exact* if $\boldsymbol{f}$ is quadratically nonlinear, which holds for advective nonlinearity for instance, and in this case (as well as the linear case) the mean equation is therefore also exact. Solving the mean equation for nonlinear $\boldsymbol{f}$ requires access to the conditional variance function $\mathbf{V}(\mathbf{x},t)$, but not to the entire conditional *covariance* function $\mathbf{P}(\mathbf{x}_1,\mathbf{x}_2,t)$, which is defined by

$$\mathbf{P}(\mathbf{x}_1,\mathbf{x}_2,t) \;\equiv\; \left\langle \mathbf{e}(\mathbf{x}_1,t)\,\mathbf{e}^T(\mathbf{x}_2,t) \middle| \mathcal{W}_{k-1}^{o} \right\rangle, \tag{C.17}$$

a function of *two* sets of spatial variables like $\mathbf{W}_0(\mathbf{x}_1,\mathbf{x}_2)$; cf. (C.3). From (C.14) and (C.17) it follows that

$$\mathbf{V}(\mathbf{x},t) \;=\; \mathbf{P}(\mathbf{x},\mathbf{x},t)\,. \tag{C.18}$$

In practice it may be possible to *model* $\mathbf{V}(\mathbf{x},t)$, or to describe its evolution through Monte Carlo methods. Approximate evolution equations for $\mathbf{V}(\mathbf{x},t)$ can be developed in some special cases (Cohn 1993).

To obtain an approximate covariance evolution equation, first use (C.1), (C.8) and (C.10) to find that

$$\frac{\partial \mathbf{e}}{\partial t} + \boldsymbol{f}(\mathbf{w}) - \left\langle \boldsymbol{f}(\mathbf{w}) \middle| \mathcal{W}_{k-1}^{o} \right\rangle \;=\; \mathbf{0}\,. \tag{C.19}$$

Substituting (C.9) and (C.15) into (C.19) gives

$$\frac{\partial \mathbf{e}}{\partial t} + \mathbf{L}(\widehat{\mathbf{w}})\mathbf{e} \;=\; \mathbf{0}\,, \tag{C.20}$$

where terms quadratic in $\mathbf{e}$ have been discarded because in the covariance equation they become cubic or quartic; along with the truncated expansion (C.9) employed to derive the mean equation, this completes the second–moment closure approximation. Under an alternative assumption that the estimation error $\mathbf{e}(\mathbf{x},t)$ is Gaussian–distributed, the cubic terms would still vanish and the quartic terms would be expressed as functions of the quadratic terms (Jazwinski 1970, §§ 9.3 and 9.4; Miller *et al.* 1994). As it stands, (C.20) is a *linear* PDE, coupled *nonlinearly* to the mean equation (C.16) through the dependence of $\mathbf{L}$ upon $\widehat{\mathbf{w}}$.

According to definition (C.17), the conditional estimation error covariance function $\mathbf{P}=\mathbf{P}(\mathbf{x}_1,\mathbf{x}_2,t)$ has the symmetry property

$$\mathbf{P}^T(\mathbf{x}_1,\mathbf{x}_2,t) \;=\; \mathbf{P}(\mathbf{x}_2,\mathbf{x}_1,t)\,, \tag{C.21}$$

and its time derivative $\frac{\partial \mathbf{P}}{\partial t}$ satisfies

$$\frac{\partial \mathbf{P}}{\partial t} \;=\; \left\langle \frac{\partial \mathbf{e}_1}{\partial t}\,\mathbf{e}_2^T \,\middle|\, \mathcal{W}^o_{k-1} \right\rangle + \left\langle \frac{\partial \mathbf{e}_2}{\partial t}\,\mathbf{e}_1^T \,\middle|\, \mathcal{W}^o_{k-1} \right\rangle^T , \tag{C.22}$$

where $\mathbf{e}_j \equiv \mathbf{e}(\mathbf{x}_j, t)$ for $j=1,\,2$. Substituting (C.20) into (C.22) yields the *covariance evolution equation*

$$\frac{\partial \mathbf{P}}{\partial t} + \mathbf{L}_1\mathbf{P} + \Big(\mathbf{L}_2\mathbf{P}^T\Big)^T \;=\; \mathbf{0}\,, \tag{C.23}$$

where $\mathbf{L}_j \equiv \mathbf{L}(\widehat{\mathbf{w}}(\mathbf{x}_j, t))$ denotes the tangent linear operator acting on the variables $\mathbf{x}_j$ of $\mathbf{P}(\mathbf{x}_1, \mathbf{x}_2, t)$ for $j=1,\,2$. Equation (C.23) is a PDE in twice the number of spatial variables as the mean equation (C.16) with which it is coupled.

Equations (C.16) and (C.23) constitute the forecast step of the second–moment closure filter for *nonlinear* continuum dynamics. In the linear case they are exact. Their solution at time t_k, starting from initial conditions

$$\mathbf{w}^a_{k-1}(\mathbf{x}) \;\equiv\; \widehat{\mathbf{w}}(\mathbf{x}, t_{k-1})\,, \tag{C.24}$$

$$\mathbf{P}^a_{k-1}(\mathbf{x}_1, \mathbf{x}_2) \;\equiv\; \mathbf{P}(\mathbf{x}_1, \mathbf{x}_2, t_{k-1})\,, \tag{C.25}$$

respectively [cf. (C.8), (C.17)], will be denoted by $\mathbf{w}^f_k(\mathbf{x})$ and $\mathbf{P}^f_k(\mathbf{x}_1, \mathbf{x}_2)$. The initial conditions for $k=1$ are given by (C.2) and (C.3), respectively.

Unlike their discrete *linear* counterparts (4.8) and (4.10), in the nonlinear case (C.16) and (C.23) are *coupled*, nonlinearly in fact: the linear equation (C.23) depends on $\mathbf{L}(\widehat{\mathbf{w}})$, while the nonlinear equation (C.16) depends on $\mathbf{V}(\mathbf{x}, t) = \mathbf{P}(\mathbf{x}, \mathbf{x}, t)$. In the *extended Kalman filter* the nonlinear bias correction term $\frac{1}{2}\mathrm{tr}\,(\mathbf{FV})$ in the conditional mean equation (C.16) is omitted, rendering the mean equation independent of the covariance equation. Omission of this term has been shown both theoretically (Cohn 1993) and numerically (Evensen 1994; Ménard 1994) to lead to spurious unbounded growth of variance for some nonlinear problems. Thus it is likely to be important to account for this term, either directly as in (C.16) or through Monte Carlo simulation of the conditional mean dynamics arising from (C.1), in data assimilation schemes of the future. Ménard (1994) has shown for the Burgers equation, however, that while the mean equation (C.16) is exact in this case since the Burgers equation is quadratically nonlinear, evaluating the nonlinear bias correction term by solving the covariance evolution equation leads to poor results because of the second–moment closure approximation in the covariance equation (C.23).

Equations for the conditional mean analysis $\mathbf{w}^a_k(\mathbf{x})$, an m–vector function of the spatial variables $\mathbf{x}$, are now developed under the assumption that $\mathbf{w}(\mathbf{x}, t_k)$ is a Gaussian random field. For linear dynamics this holds automatically if $\mathbf{w}_0(\mathbf{x}) = \mathbf{w}(\mathbf{x}, t_0)$ is Gaussian, but for nonlinear dynamics this is an approximating assumption. These analysis equations still provide the best *linear* unbiased estimate in the absence of this assumption, as discussed in § 4.4. An appropriate change of dependent variables can also be useful, as discussed in § 5.3.

Under the stated assumptions, the analysis update equation has the form

$$\mathbf{w}^a_k(\mathbf{x}) \;=\; \mathbf{w}^f_k(\mathbf{x}) + \mathbf{K}_k(\mathbf{x})\Big(\mathbf{w}^o_k - \mathbf{H}_k\mathbf{w}^f_k - \widehat{\boldsymbol{\varepsilon}}^m_k\Big)\,; \tag{C.26}$$

cf. (4.43) and (4.48). Here the gain $\mathbf{K}_k$ is an $m \times p_k$ matrix function of $\mathbf{x}$, and the innovation $\mathbf{w}_k^o - \mathbf{H}_k \mathbf{w}_k^f - \widehat{\varepsilon}_k^m$ is a p_k–vector as usual; while $\mathbf{w}_k^f$ is a function of $\mathbf{x}$, the operation $\mathbf{H}_k \mathbf{w}_k^f$ produces a p_k–vector independent of $\mathbf{x}$, as in (C.5). Thus the $\mathbf{x}$–dependence of $\mathbf{w}_k^a(\mathbf{x})$ arises solely through that of $\mathbf{w}_k^f(\mathbf{x})$ and that of the gain $\mathbf{K}_k(\mathbf{x})$.

Rather than deriving the optimal gain, here we simply write down the equation for it, by analogy with (4.44), omitting the time subscript k now for notational convenience:

$$\mathbf{K}(\mathbf{x}_1) \;=\; \left(\mathbf{H}_2 \mathbf{P}^{fT}\right)^T \Big[\mathbf{H}_1 \left(\mathbf{H}_2 \mathbf{P}^{fT}\right)^T + \mathbf{R}\Big]^{-1} . \tag{C.27}$$

Here $\mathbf{R} = \mathbf{R}_k$ is the $p_k \times p_k$ measurement error covariance matrix (C.7), while $\mathbf{H}_1$ and $\mathbf{H}_2$ denote the action of the observation operator $\mathbf{H}$ on the variables $\mathbf{x}_1$ and $\mathbf{x}_2$, respectively, of the forecast error covariance matrix function $\mathbf{P}^f = \mathbf{P}_k^f(\mathbf{x}_1, \mathbf{x}_2)$. Thus $\mathbf{H}_2 \mathbf{P}^{fT}$ is a $p_k \times m$ matrix function of $\mathbf{x}_1$, and $\mathbf{H}_1 \left(\mathbf{H}_2 \mathbf{P}^{fT}\right)^T$ is an ordinary $p_k \times p_k$ matrix; the innovation covariance matrix to be inverted in (C.27) has dimension $p_k \times p_k$.

Finally, we derive the update equation for the analysis error covariance

$$\mathbf{P}_k^a(\mathbf{x}_1, \mathbf{x}_2) \;\equiv\; \left\langle \mathbf{e}_k^a(\mathbf{x}_1) \mathbf{e}_k^{aT}(\mathbf{x}_2) \,\middle|\, \mathcal{W}_k^o \right\rangle , \tag{C.28}$$

an $m \times m$ matrix function of $\mathbf{x}_1$ and $\mathbf{x}_2$, where

$$\mathbf{e}_k^a(\mathbf{x}) \;\equiv\; \mathbf{w}(\mathbf{x}, t_k) - \mathbf{w}_k^a(\mathbf{x}) \; ; \tag{C.29}$$

cf. (C.10), (C.17), (C.25). Substituting (C.5) and (C.26) into (C.28) yields

$$\mathbf{P}^a(\mathbf{x}_1, \mathbf{x}_2) \;=\; \left\langle \left[(\mathbf{I} - \mathbf{K}_1 \mathbf{H}_1)\mathbf{e}_1^f - \mathbf{K}_1(\varepsilon^m - \widehat{\varepsilon}^m)\right] \left[(\mathbf{I} - \mathbf{K}_2 \mathbf{H}_2)\mathbf{e}_2^f - \mathbf{K}_2(\varepsilon^m - \widehat{\varepsilon}^m)\right]^T \,\middle|\, \mathcal{W}_k^o \right\rangle , \tag{C.30}$$

where the time index has been omitted, $\mathbf{I}$ denotes the $m \times m$ identity matrix, $\mathbf{K}_j \equiv \mathbf{K}_k(\mathbf{x}_j)$ for $j = 1, 2$, and

$$\mathbf{e}_{j,k}^f \;\equiv\; \mathbf{w}(\mathbf{x}_j, t_k) - \mathbf{w}_k^f(\mathbf{x}_j) \; , \tag{C.31}$$

for $j = 1, 2$. The cross–terms that appear when the bracketed terms in (C.30) are multiplied vanish because of the assumed independence of the measurement error and the state, so that $\left\langle \mathbf{w}(\mathbf{x}_j, t_k) (\varepsilon_k^m)^T \right\rangle = \mathbf{0}$, and because the assumed whiteness of the measurement error and its independence of the initial state implies that $\left\langle \mathbf{w}_k^f(\mathbf{x}_j) (\varepsilon_k^m)^T \right\rangle = \mathbf{0}$. These assumptions also allow the conditioning on $\mathcal{W}_k^o$ in the remaining two terms to be replaced by conditioning on $\mathcal{W}_{k-1}^o$, so that (C.30) becomes

$$\mathbf{P}^a(\mathbf{x}_1, \mathbf{x}_2) \;=\; (\mathbf{I} - \mathbf{K}_1 \mathbf{H}_1) \Big[(\mathbf{I} - \mathbf{K}_2 \mathbf{H}_2) \mathbf{P}^{fT}\Big]^T + \mathbf{K}_1 \mathbf{R} \mathbf{K}_2^T . \tag{C.32}$$

This is the so–called *Joseph form* (Bucy and Joseph 1968, pp. 175–176) of the analysis error covariance, which holds for arbitrary gains.

Equation (C.32) may be written as

$$\mathbf{P}^a(\mathbf{x}_1,\mathbf{x}_2) \;=\; (\mathbf{I}-\mathbf{K}_1\mathbf{H}_1)\mathbf{P}^f + \mathbf{M}\,, \tag{C.33}$$

where

$$\begin{aligned}\mathbf{M} \;&\equiv\; \mathbf{K}_1\mathbf{R}\mathbf{K}_2^T - (\mathbf{I}-\mathbf{K}_1\mathbf{H}_1)\big(\mathbf{K}_2\mathbf{H}_2\mathbf{P}^{fT}\big)^T \\ &=\; \Big[\mathbf{K}_1\mathbf{R} - (\mathbf{I}-\mathbf{K}_1\mathbf{H}_1)\big(\mathbf{H}_2\mathbf{P}^{fT}\big)^T\Big]\mathbf{K}_2^T \\ &=\; \Big\{\mathbf{K}_1\Big[\mathbf{H}_1\big(\mathbf{H}_2\mathbf{P}^{fT}\big)^T + \mathbf{R}\Big] - \big(\mathbf{H}_2\mathbf{P}^{fT}\big)^T\Big\}\mathbf{K}_2^T\,. \end{aligned} \tag{C.34}$$

Upon substituting (C.27) into (C.34) one finds that $\mathbf{M}=\mathbf{0}$, so that

$$\mathbf{P}^a(\mathbf{x}_1,\mathbf{x}_2) \;=\; (\mathbf{I}-\mathbf{K}_1\mathbf{H}_1)\mathbf{P}^f\,, \tag{C.35}$$

the analysis error covariance update equation; cf. (4.45). Equations (C.26), (C.27) and (C.35) together constitute the continuum analysis update equations.

References

Anderson, B. D. O., and J. B. Moore, 1979: *Optimal Filtering*. Prentice–Hall, 357pp.

Andersson, E., J. Pailleux, J.–N. Thépaut, J. Eyre, A. McNally, G. Kelly, and P. Courtier, 1994: Use of cloud–cleared radiances in three/four dimensional variational assimilation. *Quart. J. Roy. Meteor. Soc.*, **120**, 627–654.

Backus, G., and F. Gilbert, 1970: Uniqueness in the inversion of inaccurate gross Earth data. *Philos. Trans. Roy. Soc. London*, **266**, 123–192.

Bengtsson, L., and J. Shukla, 1988: Integration of space and in situ observations to study global climate change. *Bull. Amer. Meteor. Soc.*, **69**, 1130–1143.

Bennett, A. F., 1992: *Inverse Methods in Physical Oceanography*. Cambridge U. Press, 346pp.

Bennett, A. F., L. M. Leslie, C. R. Hagelberg, and P. E. Powers, 1993: Tropical cyclone prediction using a barotropic model initialized by a generalized inverse method. *Mon. Wea. Rev.*, **121**, 1714–1729.

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

Bucy, R. S., and P. D. Joseph, 1968: *Filtering for Stochastic Processes with Applications to Guidance*. Wiley–Interscience, 195pp.

Bürger, G., and M. A. Cane, 1994: Interactive Kalman filtering. *J. Geophys. Res. Oceans*, **99**, 8015–8031.

Caglayan, A. K., and R. E. Lancraft, 1983: A separated bias identification and state estimation algorithm for nonlinear systems. *Automatica*, **19**, 561–570.

Carton, J. A., and E. C. Hackert, 1990: Data assimilation applied to the temperature and circulation in the tropical Atlantic, 1983–84. *J. Phys. Oceanogr.*, **20**, 1150–1165.

Casti, J. L., 1985: *Nonlinear System Theory.* Academic Press, 261pp.

Catlin, D. E., 1989: *Estimation, Control, and the Discrete Kalman Filter.* Springer–Verlag, 274pp.

Cohn, S. E., 1993: Dynamics of short–term univariate forecast error covariances. *Mon. Wea. Rev.*, **121**, 3123–3149.

Cohn, S. E., and D. P. Dee, 1988: Observability of discretized partial differential equations. *SIAM J. Numer. Anal.*, **25**, 586–617.

Cohn, S. E., and R. Todling, 1996: Approximate data assimilation schemes for stable and unstable dynamics. *J. Met. Soc. Japan*, **74**, 63–75.

Cohn, S. E., N. S. Sivakumaran, and R. Todling, 1994: A fixed–lag Kalman smoother for retrospective data assimilation. *Mon. Wea. Rev.*, **122**, 2838–2867.

Courant, R., and D. Hilbert, 1962: *Methods of Mathematical Physics, Vol. II: Partial Differential Equations.* Wiley/Interscience, 830pp.

Courtier, P., J.–N. Thépaut, and A. Hollingsworth, 1994: A strategy for operational implementation of 4D–VAR, using an incremental approach. *Quart. J. Roy. Meteor. Soc.*, **120**, 1367–1387.

Curtain, R., 1975: A survey of infinite-dimensional filtering. *SIAM Review*, **17**, 395–411.

Daley, R., 1991: *Atmospheric Data Analysis.* Cambridge U. Press, 457pp.

Daley, R., 1992a: The effect of serially correlated observation and model error on atmospheric data assimilation. *Mon. Wea. Rev.*, **120**, 164–177.

Daley, R., 1992b: The lagged innovation covariance: A performance diagnostic for atmospheric data assimilation. *Mon. Wea. Rev.*, **120**, 178–196.

Daley, R., 1992c: Estimating model–error covariances for application to atmospheric data assimilation. *Mon. Wea. Rev.*, **120**, 1735–1746.

Daley, R., 1993: Estimating observation error statistics for atmospheric data assimilation. *Ann. Geophysicae*, **11**, 634–647.

Daley, R., 1995: Estimating the wind field from chemical constituent observations: Experiments with a one–dimensional extended Kalman filter. *Mon. Wea. Rev.*, **123**, 181–198.

Daley, R., and R. Ménard, 1993: Spectral characteristics of Kalman filter systems for atmospheric data assimilation. *Mon. Wea. Rev.*, **121**, 1554–1565.

da Silva, A., J. Pfaendtner, J. Guo, M. Sienkiewicz, and S. Cohn, 1995: Assessing the effects of data selection with DAO's Physical–space Statistical Analysis System. *Second International Symposium on Assimilation of Observations in Meteorology and Oceanography*, Tokyo, 13–17 March 1995, World Meteorological Organization, 273–278.

Dee, D. P., 1995: On–line estimation of error covariance parameters for atmospheric data assimilation. *Mon. Wea. Rev.*, **123**, 1128–1145.

Dee, D. P., and A. da Silva, 1997: Data assimilation in the presence of forecast bias. *Quart. J. Roy. Meteor. Soc.*, in press.

Derber, J. C., 1989: A variational continuous assimilation technique. *Mon. Wea. Rev.*, **117**, 2437–2446.

Derber, J., and A. Rosati, 1989: A global oceanic data assimilation system. *J. Phys. Oceanogr.*, **19**, 1333–1347.

Desroziers, G., and J.–P. Lafore, 1993: A coordinate transformation for objective frontal analysis. *Mon. Wea. Rev.*, **121**, 1531–1553.

Ehrendorfer, M., 1994a: The Liouville equation and its potential usefulness for the prediction of forecast skill. Part I: Theory. *Mon. Wea. Rev.*, **122**, 703–713.

Ehrendorfer, M. 1994b: The Liouville equation and its potential usefulness for the prediction of forecast skill. Part II: Applications. *Mon. Wea. Rev.*, **122**, 714–728.

Epstein, E. S., 1969: Stochastic dynamic prediction. *Tellus*, **21**, 739–759.

Evensen, G., 1994: Sequential data assimilation with a nonlinear quasi–geostrophic model using Monte Carlo methods to forecast error statistics. *J. Geophys. Res. Oceans*, **99**, 10143–10162.

Frost, P. A., and T. Kailath, 1971: An innovations approach to least–squares estimation – Part III: Nonlinear estimation in white Gaussian noise. *IEEE Trans. Automatic Control*, **AC–16**, 217–226.

Gauthier, P., M. Tanguay, and P. Bartello, 1995: Four–dimensional data assimilation with a wide range of scales. *Second International Symposium on Assimilation of Observations in Meteorology and Oceanography*, Tokyo, 13–17 March 1995, World Meteorological Organization, 123–128.

Gelb, A. (ed.), 1974: *Applied Optimal Estimation*. The MIT Press, 374pp.

Ghil, M., 1997: Advances in sequential estimation for atmospheric and oceanic flows. This volume.

Golub, G. H., and C. F. Van Loan, 1983: *Matrix Computations*. The Johns Hopkins University Press, 476pp.

Heckley, W. A., P. Courtier, J. Pailleux, and E. Andersson, 1992: The ECMWF variational analysis: General formulation and use of background information. *Proc. ECMWF Workshop on Variational Assimilation*, 9–12 Nov. 1992. ECMWF, Shinfield Park, Reading, Berkshire, RG2 9AX, England, 49–94.

Householder, A. S., 1964: *The Theory of Matrices in Numerical Analysis*. Dover, 257pp.

Jazwinski, A. H., 1970: *Stochastic Processes and Filtering Theory*. Academic Press, 376pp.

Joiner, J., and A. da Silva, 1997: Efficient methods to assimilate satellite retrievals based on information content. *Quart. J. Roy. Meteor. Soc.*, in press.

Kailath, T., 1968: An innovations approach to least–squares estimation. Part I: Linear filtering in additive white noise. *IEEE Trans. Automatic Control*, **AC–13**, 646–655.

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

Leith, C. E., 1990: Stochastic backscatter in a subgrid–scale model: Plane shear mixing layer. *Phys. Fluids*, **2**, 297–299.

Li, Y., 1991: A note on the uniqueness problem of variational adjustment approach to four–dimensional data assimilation. *J. Met. Soc. Japan*, **69**, 581–585.

Loève, M., 1963: *Probability Theory*, 3$^{\text{rd}}$ ed. D. Van Nostrand Co., 685pp.

Lorenc, A. C., 1981: A global three–dimensional multivariate statistical interpolation scheme. *Mon. Wea. Rev.*, **109**, 701–721.

Lorenc, A. C., 1986: Analysis methods for numerical weather prediction. *Quart. J. Roy. Meteor. Soc.*, **112**, 1177–1194.

Lyster, P. M., S. E. Cohn, R. Ménard, L.–P. Chang, S.–J. Lin, and R. G. Olsen, 1997: Parallel implementation of a Kalman filter for constituent data assimilation. *Mon. Wea. Rev.*, in press.

Maybeck, P. S., 1979: *Stochastic Models, Estimation, and Control*. Volume 1. Academic Press, 423pp.

McPherson, R. D., K. H. Bergman, R. E. Kistler, G. E. Rasch, and D. S. Gordon, 1979: The NMC operational global data assimilation system. *Mon. Wea. Rev.*, **107**, 1445–1461.

Ménard, R., 1994: *Kalman filtering of Burgers' equation and its application to atmospheric data assimilation*. Ph. D. Thesis, McGill University. Stormy Weather Group scientific report **NW–100**, 211pp.

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

Ménard, R., P. Lyster, L.–P. Chang, and S. E. Cohn, 1995: Middle atmosphere assimilation of UARS constituent data using Kalman filtering: preliminary results. *Second International Symposium on Assimilation of Observations in Meteorology and Oceanography*, Tokyo, 13–17 March 1995, World Meteorological Organization, 235–238.

Menke, W., 1984: *Geophysical Data Analysis: Discrete Inverse Theory.* Academic Press, 260pp.

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

National Academy of Sciences (NAS), 1991: *Four–Dimensional Model Assimilation of Data: A Strategy for the Earth System Sciences.* Report of the Panel on Model–Assimilated Data Sets for Atmospheric and Oceanic Research, National Academy Press, 78pp.

Navon, I. M., and D. M. Legler, 1987: Conjugate–gradient methods for large–scale minimization in meteorology. *Mon. Wea. Rev.*, **115**, 1479–1502.

Omatu, S., and J. H. Seinfeld, 1989: *Distributed Parameter Systems: Theory and Applications.* Oxford U. Press, 430pp.

Ott, W. R., 1995: *Environmental Statistics and Data Analysis.* Lewis Publishers, Boca Raton, Florida, 313pp.

Papoulis, A., 1984: *Probability, Random Variables, and Stochastic Processes*, 2^{nd} ed. McGraw–Hill, 576pp.

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

Phillips, N. A., 1966: The equations of motion for a shallow rotating atmosphere and the "traditional approximation". *J. Atmos. Sci.*, **23**, 626–628.

Rabier, F., and P. Courtier, 1992: Four–dimensional assimilation in the presence of baroclinic instability. *Quart. J. Roy. Meteor. Soc.*, **118**, 649–672.

Rodgers, C. D., 1990: Characterization and error analysis of profiles retrieved from remote sounding measurements. *J. Geophys. Res.*, **95**, 5587–5595.

Royden, H. L., 1968: *Real Analysis*, 2^{nd} ed. Macmillan, 349pp.

Sorenson, H. W., 1980: *Parameter Estimation: Principles and Problems.* Marcel Dekker, 382pp.

Sorenson, H. W. (ed.), 1985: *Kalman Filtering: Theory and Application.* IEEE Press, 457pp.

Talagrand, O., 1997: Assimilation of observations, an introduction. This volume.

Tanguay, M., P. Bartello, and P. Gauthier, 1995: Four–dimensional data assimilation with a wide range of scales. *Tellus*, **47A**, 974–997.

Tarantola, A., 1987: *Inverse Problem Theory: Methods for Data Fitting and Model Parameter Estimation.* Elsevier, 613pp.

Todling, R., and S. E. Cohn, 1994: Suboptimal schemes for atmospheric data assimilation based on the Kalman filter. *Mon. Wea. Rev.*, **122**, 2530–2557.

Toth, Z., and E. Kalnay, 1993: Ensemble forecasting at NMC: The generation of perturbations. *Bull. Amer. Meteor. Soc.*, **74**, 2317–2330.

Vanmarcke, E., 1983: *Random Fields: Analysis and Synthesis.* The MIT Press, 382pp.

Verlaan, M., and A. W. Heemink, 1995: Reduced rank square root filters for large scale data assimilation problems. *Second International Symposium on Assimilation of Observations in Meteorology and Oceanography*, Tokyo, 13–17 March 1995, World Meteorological Organization, 247–252.

Wunsch, C., 1996: *The Ocean Circulation Inverse Problems.* Cambridge U. Press, 442pp.

Yaglom, A. M., 1987: *Correlation Theory of Stationary and Related Random Functions. Volume I: Basic Results.* Springer–Verlag, 526pp.