ELSEVIER

A Bayesian tutorial for data assimilation

Christopher K. Wikle [a,*], L. Mark Berliner [b]

[a] *Department of Statistics, University of Missouri, 146 Middlebush Hall, Columbia, MO 65211, United States*
[b] *Department of Statistics, The Ohio State University, United States*

Abstract

Data assimilation is the process by which observational data are fused with scientific information. The Bayesian paradigm provides a coherent probabilistic approach for combining information, and thus is an appropriate framework for data assimilation. Viewing data assimilation as a problem in Bayesian statistics is not new. However, the field of Bayesian statistics is rapidly evolving and new approaches for model construction and sampling have been utilized recently in a wide variety of disciplines to combine information. This article includes a brief introduction to Bayesian methods. Paying particular attention to data assimilation, we review linkages to optimal interpolation, kriging, Kalman filtering, smoothing, and variational analysis. Discussion is provided concerning Monte Carlo methods for implementing Bayesian analysis, including importance sampling, particle filtering, ensemble Kalman filtering, and Markov chain Monte Carlo sampling. Finally, hierarchical Bayesian modeling is reviewed. We indicate how this approach can be used to incorporate significant physically based prior information into statistical models, thereby accounting for uncertainty. The approach is illustrated in a simplified advection–diffusion model.
© 2006 Elsevier B.V. All rights reserved.

Keywords: Bayes; Ensemble Kalman filter; Importance sampling; Kriging; Markov chain Monte Carlo; Particle filter

1. Introduction

What is data assimilation (DA)? In some sense, the definition is a "work in progress" depending on the application and background of those who define it. For example, Bennett [3, p. xvi] states that data assimilation involves "...interpolating fields at one time, for subsequent use as initial data in a model integration which may even be a genuine forecast". Kalnay [29, p. 136] states that DA is a "...statistical combination of observations and short-range forecasts". Talagrand [37, p. 191] writes that DA is "...the process through which all the available information is used to estimate as accurately as possible the state of the atmospheric or oceanic flow". These definitions suggest our working definition, from a statistical perspective, that *DA is an approach for fusing data (observations) with prior knowledge (e.g., mathematical representations of physical laws; model output) to obtain an estimate of the distribution of the true state of a process.* From this perspective, one needs

the following components to perform DA: a statistical model for observations (i.e., a data or measurement model), and an *a priori* statistical model for the state process (i.e., a state or process model). Issues related to DA involve choices of these two components as well as the choice for how such information is combined.

Several monographs and review papers describe various methods and approaches to DA (e.g., [39,21,12,37,3,29]). We explore the topic from a Bayesian perspective. Epstein [14] was one of the first to seriously consider Bayesian methods in the atmospheric sciences. Lorenc [30] and Tarantola [38] were among the first to write about this perspective on DA. As Lorenc showed, one of the advantages of thinking about DA from this perspective is that it provides a common methodology that links many of the seemingly disparate approaches to the subject. One might ask what is the point of our paper given that the Bayesian approach to DA is well-established. The short answer is that the field of applied Bayesian statistics in general, and the use of Monte Carlo methods specifically, has seen dramatic developments since the early 1990s. Many of these developments have found their way into the DA

* Corresponding author. Tel.: +1 573 8829659.
  E-mail address: wiklec@missouri.edu (C.K. Wikle).

literature (e.g., [1]), and some have not. Our purpose is to give an overview of the Bayesian perspective and then discuss current approaches from this perspective, as well as potential extensions based on recent developments in statistics.

2. Bayesian inference

Gelman et al. [20, p. 2] define Bayesian inference as

"...the process of fitting a probability model to a set of data and summarizing the result by a probability distribution on the parameters of the model and on unobserved quantities such as predictions for new observations".

Bayesian inference consists of three steps. In the first, one formulates a "full probability model". This is simply the joint probability distribution of all observable and unobservable components of interest (e.g., data, process, and parameters). The formulation of such models should be consistent with the knowledge of the underlying scientific processes as well as how the data were collected. The next step in this process is to find the conditional distribution of the unobservable quantities of interest given the observed data. This is formally accomplished by application of Bayes' Theorem, though, in many cases, there is no analytical solution for this step. Finally, as with all modeling, one should evaluate the fit of the model and its ability to adequately characterize the processes of interest. For more detail, see [20], and the classic texts by Berger [4] and Bernardo and Smith [9]; also see [15].

For illustration, let X denote unobservable quantities of interest and Y our data. The full probability model can always be factored into components: $p(x, y) = p(y|x)p(x) = p(x|y)p(y)$. Applying *Bayes' Rule*, we obtain

$$p(x|y) = \frac{p(y|x)p(x)}{p(y)},$$

provided $0 < p(y) < \infty$. It is illustrative to examine each component of Bayes' rule separately.

Before we begin, a few words are in order regarding notation. Random variables will be denoted by capital letters, and fixed or observed quantities will be denoted by lower-case letters. Greek letters will also refer to random variables, typically parameters. In addition, we will use $p(\ )$ to refer to a probability density function, and will use lower-case arguments in this context. Finally, bold quantities will refer to vectors or matrices.

Data distribution, $p(y|x)$: Statisticians often refer to this as a "sampling distribution" or "measurement model". It is simply the distribution of the data, given the unobservables. When viewed as a function of X for fixed y, it is known as a likelihood function, $L(x|y)$, as in classical maximum likelihood estimation. A key is that one thinks of the data *conditioned* on x. For example, if Y represents imperfect observations of temperature, and X the true (unobservable) temperature, then $p(y|x)$ quantifies the distribution of measurement errors in observing temperature, reflecting possible biases as well as instrument error.

Prior distribution, $p(x)$: This distribution quantifies our *a priori* understanding of the unobservable quantities of interest. For example, if X corresponds to temperature, then one might base this prior distribution on historical information (climatology) or perhaps from a forecast model. In general, prior distributions can be informative or non-informative, "subjective" or "objective". The choice of such distributions is an integral part of Bayesian inference.

Marginal distribution, $p(y) = \int p(y|x)p(x)\mathrm{d}x$: We assume continuous X but note that there are analogous forms (sums) for discrete X. This distribution is also known as the prior predictive distribution. Alternatively, for the observations Y, $p(y)$ can be thought of as the "normalizing constant" in Bayes' rule. Unfortunately, it is only for very specific choices of the data and prior distribution that we can solve this integral analytically.

Posterior distribution, $p(x|y)$: This distribution of the unobservables given the data is our primary interest for inference. It is proportional to the product of the data model and the prior. The posterior is the update of our prior knowledge about X as summarized in $p(x)$ given the actual observations y. In this sense, the Bayesian approach is inherently "scientific" in that it is analogous to the scientific method: one has prior belief (information), collects data, and then updates that belief given the new data (information).

2.1. Simple univariate example: Normal data, normal prior

Say we are interested in the univariate state variable, X (e.g., u-component of wind at some location). Assume we have the prior distribution (e.g., from a forecast model): $X \sim N(\mu, \tau^2)$, where "$\sim$" is read "is distributed as" and $N(a, b)$ refers to a normal (or Gaussian) distribution with mean a and variance b. Conditioned on the true value of the state process, $X = x$, assume we have n independent (given x) but noisy observations $\mathbf{Y} = (Y_1, \ldots, Y_n)'$ and thus the data model: $Y_i|X = x \sim N(x, \sigma^2)$. Then,

$$p(\mathbf{y}|x) = \prod_{i=1}^{n}(1/\sqrt{2\pi\sigma^2})\exp\{-0.5(y_i - x)^2/\sigma^2\}$$
$$\propto \exp\left\{-0.5\sum_{i=1}^{n}(y_i - x)^2/\sigma^2\right\},$$

and Bayes' rule gives

$$p(x|\mathbf{y}) \propto \exp\left\{-0.5\left[\sum_{i=1}^{n}(y_i - x)^2/\sigma^2 + (x - \mu)^2/\tau^2\right]\right\}$$
$$\propto \exp\left\{-0.5\left[x^2(n/\sigma^2 + 1/\tau^2) - 2\left(\sum y_i/\sigma^2 + \mu/\tau^2\right)x\right]\right\}.$$

We note that this is just the product of two Gaussian distributions. It can be shown (by completing the square) that the normalized product is also Gaussian with the following mean and variance:

$$X|\mathbf{y} \sim N\left((n/\sigma^2 + 1/\tau^2)^{-1}\left(\sum y_i/\sigma^2 + \mu/\tau^2\right),\right.$$
$$\left. (n/\sigma^2 + 1/\tau^2)^{-1}\right).$$

We can write the posterior mean as

$$E(X|\mathbf{y}) = \frac{\sigma^2 \tau^2}{\sigma^2 + n\tau^2}(n\bar{y}/\sigma^2 + \mu/\tau^2) \tag{1}$$

$$= w_y \bar{y} + w_\mu \mu, \tag{2}$$

where $\bar{y} = \sum_i y_i/n$, $w_y = n\tau^2/(n\tau^2 + \sigma^2)$, and $w_\mu = \sigma^2/(n\tau^2 + \sigma^2)$. Note that $w_y + w_\mu = 1$. That is, the posterior mean is a weighted average of the prior mean (μ) and the natural, data based estimate of X, $\bar{y}$. Note that as $\tau^2 \to \infty$, the data model overwhelms the prior and $p(x|\mathbf{y}) \to N(\bar{y}, \sigma^2/n)$. Alternatively, for fixed τ^2, but very large amounts of data (i.e., $n \to \infty$) the data model again dominates the posterior density. On the other hand, if τ^2 is very small, the prior is critical for comparatively small n. Though these properties are shown for the normal data, normal prior case, it is generally true that for very large datasets, the data model is the major controller of the posterior.

For purposes of DA, we note from (2) that the posterior mean can also be written as

$$E(X|\mathbf{y}) = \mu + \left(\frac{n\tau^2}{\sigma^2 + n\tau^2}\right)(\bar{y} - \mu)$$

$$= \mu + K(\bar{y} - \mu). \tag{3}$$

That is, the prior mean (μ) is adjusted toward the sample estimate ($\bar{y}$) according to the "gain", $K = (n\tau^2)/(\sigma^2 + n\tau^2)$. Analogously, the posterior variance can be rewritten:

$$\mathrm{var}(X|\mathbf{y}) = (1 - K)\tau^2, \tag{4}$$

where the posterior variance is updated from the prior variance according to the gain, K. Eqs. (3) and (4) are critical for understanding data assimilation as will be seen throughout this overview.

2.1.1. Numerical examples

Assume the prior distribution is $X \sim N(20, 3)$, and the data model is $Y_i|x \sim N(x, 1)$. We have two observations $\mathbf{y} = (19, 23)'$. The posterior mean is $20 + (6/7)(21 - 1)$ = 20.86 and the posterior variance is $(1 - 6/7)3 = 0.43$. Fig. 1 shows these distributions graphically. Since the data are relatively precise compared to the prior, we see that the posterior distribution is "closer" to the likelihood than the prior. Another way to look at this is that the gain ($K = 6/7$) is close to one, so that the data model is weighted more than the prior.

Next, assume the same observations and prior distribution, but change the data model to $Y_i|x \sim N(x, 10)$. The gain is $K = 6/16$ and the posterior distribution is $X|\mathbf{y} \sim N(20.375, 1.875)$. This is illustrated in Fig. 2. In this case, the gain is closer to zero (since the measurement error variance is relatively large compared to the prior variance) and thus the prior is given more weight.

2.1.2. Posterior predictive distribution

Often, one is interested in obtaining the distribution of a new observation $\tilde{Y}$ based on the observed data $\mathbf{y}$. Such a distribution is known as the *posterior predictive distribution* and is given
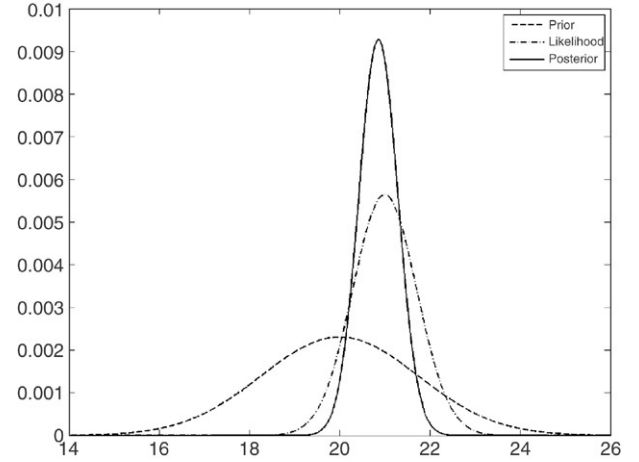

Fig. 1. Posterior distribution with normal prior and normal likelihood; relatively precise data.
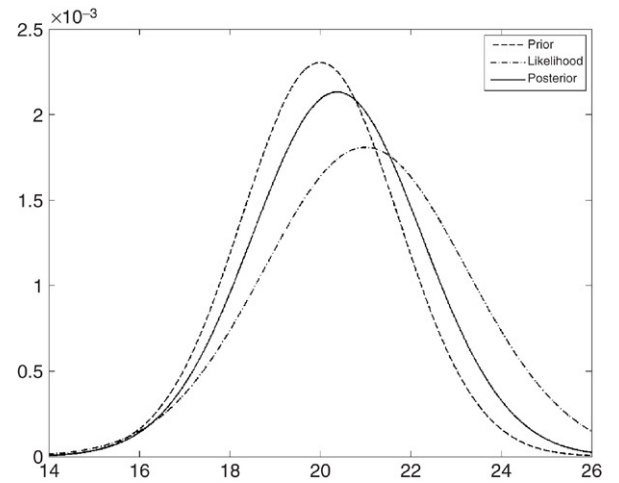

Fig. 2. Posterior distribution with normal prior and normal likelihood; relatively uncertain data.

by:

$$p(\tilde{y}|\mathbf{y}) = \int p(\tilde{y}|x)p(x|\mathbf{y})\mathrm{d}x$$

and in the normal case:

$$\propto \int \exp\{-0.5(\tilde{y} - x)^2/\sigma^2\} \exp\{-0.5(x - x_a)^2/\tau_a^2\},$$

where x_a and τ_a^2 are the posterior mean and variance, respectively (where we have made use of the conditional independence assumption that $p(\tilde{y}|x, \mathbf{y}) = p(\tilde{y}|x)$). In this case we have $\tilde{Y}|\mathbf{y} \sim N(x_a, \sigma^2 + \tau_a^2)$. Thus, the predictive mean is the posterior mean, and the predictive distribution is necessarily less precise than the posterior distribution.

2.2. Prior distributions

To many, Bayesian analysis is predicated on a belief in *subjective probability*. That is, the quantification of beliefs (however vague) about X before the data are considered. The choice of prior distributions has been the subject of much

debate and study. In fact, subjective probability may or may not be consistent with the classical long-run frequency notion of probability, which for some, is philosophically unappealing. Thus, a major historical impediment to the use of Bayesian techniques was the specification of the appropriate form of the prior. It should be noted that the choice of likelihood function in classical statistics is in many ways also subjective.

As outlined in [4] and [9], there are a variety of techniques for developing priors in practice. Priors can be obtained from past studies, from subject area "expert opinion" or scientific first principles, though compromises are typically made for mathematical simplicity. For example, conjugate prior and data model pairs of distributions lead to posterior distributions of the same functional form as the prior, as with the normal data, normal prior examples above. In other words, in such cases, one knows the normalizing constant in Bayes' rule.

There is also a significant literature on the notion of *non-informative* or *objective* priors. There are typically a variety of non-informative priors in a given problem, indicating that they do not represent total ignorance about the problem at hand. However, they serve as reference or default priors, to use when prior information is lacking or very vague. They also often yield results which match or nearly match those of traditional non-Bayesian approaches, thereby providing them with Bayesian interpretations.

2.2.1. Mixture priors

A very flexible class of priors can be constructed by forming *mixtures* of tractable (e.g., conjugate) priors. This has implications for some modern approaches in DA. Suppose we have data model $p(y_i|x)$ and $p_1(x)$ and $p_2(x)$ are both densities that give rise to the posteriors $p_1(x|\mathbf{y})$ and $p_2(x|\mathbf{y})$, respectively. Let w_1 and w_2 be any non-negative real numbers such that $w_1 + w_2 = 1$, and write the mixture prior

$$p(x) = w_1 p_1(x) + w_2 p_2(x).$$

Then, it is easy to show that the posterior distribution corresponding to $p(x)$ is:

$$p(x|\mathbf{y}) = w_1^* p_1(x|\mathbf{y}) + w_2^* p_2(x|\mathbf{y}),$$

where

$$w_i^* \propto w_i \int p(\mathbf{y}|x) p_i(x) \mathrm{d}x, \quad i = 1, 2,$$

and $w_1^* + w_2^* = 1$. Mixtures of two or more priors can be used to approximate a variety of prior information while maintaining mathematical tractability.

2.3. Multivariate normal–normal case

Assume we are interested in an $n \times 1$ vector process $\mathbf{X}$ (e.g., u-wind components at several locations), that has prior distribution, $\mathbf{X} \sim N(\boldsymbol{\mu}, \mathbf{P})$, where for now we assume that the mean $\boldsymbol{\mu}$ and covariance matrix $\mathbf{P}$ are known. In addition, we observe the $p \times 1$ data vector $\mathbf{Y}$ and assume the following data model, $\mathbf{Y}|\mathbf{x} \sim N(\mathbf{Hx}, \mathbf{R})$, where the $p \times n$ observation matrix

$\mathbf{H}$ and the observation error covariance matrix, $\mathbf{R}$, are assumed to be known.

The posterior distribution of $\mathbf{X}|\mathbf{y}$ is given by $p(\mathbf{x}|\mathbf{y}) \propto p(\mathbf{y}|\mathbf{x}) p(\mathbf{x})$. As with the univariate case, the posterior distribution is also Gaussian:

$$\mathbf{X}|\mathbf{y} \sim N((\mathbf{H}'\mathbf{R}^{-1}\mathbf{H} + \mathbf{P}^{-1})^{-1}(\mathbf{H}'\mathbf{R}^{-1}\mathbf{y} + \mathbf{P}^{-1}\boldsymbol{\mu}),$$
$$(\mathbf{H}'\mathbf{R}^{-1}\mathbf{H} + \mathbf{P}^{-1})^{-1}). \tag{5}$$

Applying some basic linear algebra, we can rewrite the posterior mean as

$$E(\mathbf{X}|\mathbf{y}) = \boldsymbol{\mu} + \mathbf{K}(\mathbf{y} - \mathbf{H}\boldsymbol{\mu}), \tag{6}$$

where $\mathbf{K} = \mathbf{PH}'(\mathbf{R} + \mathbf{HPH}')^{-1}$ is the "gain" matrix. Similarly, the posterior covariance matrix can be written as

$$\mathrm{var}(\mathbf{X}|\mathbf{y}) = (\mathbf{I} - \mathbf{KH})\mathbf{P}. \tag{7}$$

Formulas (6) and (7) are the core of the so-called *analysis step* of DA based on linear and Gaussian assumptions. Typically, in a DA problem, $\boldsymbol{\mu}$ is the forecast from some deterministic model or long term averages based on previously collected data (e.g., climatology) and $\mathbf{P}$ is the forecast error covariance matrix. One updates the prior (forecast) mean $\boldsymbol{\mu}$ based on deviations from the observations according to the "gain" $\mathbf{K}$, which is a function of the prior and data error covariance matrices. Similarly, the prior (forecast) covariance matrix is updated according to the gain, although notably, this update is not a function of the observations, but only their location and covariance structure. This latter property relies heavily on the linear, Gaussian assumptions.

2.3.1. Relationship to kriging/optimal interpolation

We consider the relationship to Kriging (geostatistics)/Optimal Interpolation (meteorology, oceanography) by a simple example. Assume $\mathbf{X} = [x(s_1), x(s_2), x(s_3)]'$ at spatial locations $s_i, i = 1, 2, 3$. Also, assume we have observations at s_2 and s_3 but not s_1: $\mathbf{y} = [y(s_2), y(s_3)]'$ and thus $\mathbf{H}$ is defined as:

$$\mathbf{H} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

Assume the prior covariance matrix that describes the (forecast) error covariance matrix is given by:

$$\mathbf{P} = \begin{bmatrix} c_{11} & c_{12} & c_{13} \\ c_{21} & c_{22} & c_{23} \\ c_{31} & c_{32} & c_{33} \end{bmatrix}.$$

Note that even though we only have observations for locations 2 and 3, it is critical that we have the covariance information between all state locations of interest (e.g., 1, 2 and 3). In this case, the "gain" is given by:

$$\mathbf{K} = \mathbf{PH}'(\mathbf{R} + \mathbf{HPH}')^{-1} = \begin{pmatrix} c_{12} & c_{13} \\ c_{22} & c_{23} \\ c_{32} & c_{33} \end{pmatrix}$$
$$\times \left(\mathbf{R} + \begin{pmatrix} c_{22} & c_{23} \\ c_{32} & c_{33} \end{pmatrix} \right)^{-1}.$$

For simplicity, assume $\mathbf{R} = \sigma^2 \mathbf{I}$ (i.e., independent measurement errors). Then, the posterior mean of $x(s_1)$ is given by:

$$E(x(s_1)|\mathbf{y}) = \mu(s_1) + w_{12}(y(s_2) - \mu(s_2))$$
$$+ w_{13}(y(s_3) - \mu(s_3)),$$

where the interpolation weights, $\mathbf{w}_1 = [w_{12}, w_{13}]'$, are given by elements from the gain matrix:

$$\mathbf{w}_1' = \begin{pmatrix} c_{12} & c_{13} \end{pmatrix} \begin{pmatrix} c_{22} + \sigma^2 & c_{23} \\ c_{32} & c_{33} + \sigma^2 \end{pmatrix}^{-1}.$$

Thus, the prior mean is adjusted by a weighted combination of the anomalies (difference between observation and prior mean) at each data location.

The mean-squared prediction error (posterior variance) at $x(s_1)$ is given by

$$\text{var}(x(s_1)|\mathbf{y}) = c_{11}$$
$$- \begin{pmatrix} c_{12} & c_{13} \end{pmatrix} \begin{pmatrix} c_{22} + \sigma^2 & c_{23} \\ c_{32} & c_{33} + \sigma^2 \end{pmatrix}^{-1} \begin{pmatrix} c_{12} \\ c_{13} \end{pmatrix}.$$

Such spatial prediction (interpolation) is the optimal (best linear unbiased) prediction assuming the parameters, $\mathbf{R}$, $\mathbf{P}$ are known. In spatial statistics this is known as *simple kriging* [31,11] and in atmospheric/oceanographic science this is known as *optimal interpolation* [18]. See [11] for details.

It is relatively simple to accommodate more complicated, unknown prior means (*ordinary kriging* if mean is constant but unknown; *universal kriging* if mean is a linear function of covariates; e.g., see [11]). These methods are easily expressed in the framework of *linear mixed models* in statistics or as variational (optimization) problems (e.g., [10]).

Numerical example: Assume we have two observations $y_2 = 16$, $y_3 = 23$ and we are interested predicting the true process x_i at these locations and a third location x_1. Our prior mean is $\mu_i = 18$, $i = 1, 2, 3$ and our prior covariance matrix is:

$$\mathbf{P} = \begin{pmatrix} 1 & 0.61 & 0.22 \\ 0.61 & 1 & 0.37 \\ 0.22 & 0.37 & 1 \end{pmatrix}.$$

Our measurement error covariance matrix is $\mathbf{R} = 0.5\mathbf{I}$. In this case, our gain (interpolation weights) is (are):

$$\mathbf{K} = \begin{pmatrix} 0.3914 & 0.0528 \\ 0.6453 & 0.0870 \\ 0.0870 & 0.6453 \end{pmatrix}$$

and the posterior mean is:

$$E(\mathbf{X}|\mathbf{y}) = \begin{pmatrix} 17.4810 \\ 17.1442 \\ 21.0527 \end{pmatrix}$$

with posterior covariance:

$$\text{var}(\mathbf{X}|\mathbf{y}) = \begin{pmatrix} 0.7508 & 0.1957 & 0.0264 \\ 0.1957 & 0.3227 & 0.0435 \\ 0.0264 & 0.0435 & 0.3227 \end{pmatrix}.$$

Thus, the optimal prediction at location 1 gives more weight to the observation at location 2 than the observation at location 3 (since it is more highly correlated (i.e., closer)). In addition, the prediction variance at location 1 is greater than location 2 and 3 since there are no data for location 1.

Now, if we let $\mathbf{P} = \mathbf{I}$, so that there is no correlation *a priori*, then the posterior mean and variance are:

$$E(\mathbf{X}|\mathbf{y}) = \begin{pmatrix} 18.0000 \\ 16.6667 \\ 21.3333 \end{pmatrix}$$

and

$$\text{var}(\mathbf{X}|\mathbf{y}) = \begin{pmatrix} 1.0000 & 0 & 0 \\ 0 & 0.3333 & 0 \\ 0 & 0 & 0.3333 \end{pmatrix}.$$

Thus, the posterior mean and variance at location 1 (no data) are just the prior mean and variance in this case.

2.4. Connections to variational approaches

It is well-known (e.g., [30,37]) that the optimal interpolation problem can be posed equivalently as a variational problem. In particular, the posterior mode (and mean) corresponding to the multivariate normal data and prior model is also found by minimizing the objective function:

$$J(\mathbf{x}) = (\mathbf{y} - \mathbf{Hx})'\mathbf{R}^{-1}(\mathbf{y} - \mathbf{Hx}) + (\mathbf{x} - \boldsymbol{\mu})'\mathbf{P}^{-1}(\mathbf{x} - \boldsymbol{\mu}). \quad (8)$$

That is, (8) is proportional to the negative of the logarithm of the posterior density given in (5). The interpretation of the solution to this objective function as a Bayesian posterior mode applies more generally to cases in which the prior on $\mathbf{X}$ is not Gaussian.

Although formally equivalent to the Bayes formulation, for high-dimensional processes it is often more computationally efficient to approach the problem from this variational perspective. Furthermore, in non-linear/non-Gaussian settings, the objective function may have multiple modes so a single solution need not summarize the posterior distribution adequately. It is also typically difficult to provide uncertainty measures for the state estimates.

3. Sequential approaches

We define the following notation. Let, $\mathbf{Y}_{1:t} \equiv \{\mathbf{Y}_1, \ldots, \mathbf{Y}_t\}$ and $\mathbf{X}_{0:t} \equiv \{\mathbf{X}_0, \ldots, \mathbf{X}_t\}$, and similarly for the non-random equivalents $\mathbf{y}_{1:t}$ and $\mathbf{x}_{0:t}$. The posterior distribution of the states $\mathbf{X}_{0:t}$ given the observed data $\mathbf{y}_{1:t}$ is then

$$p(\mathbf{x}_{0:t}|\mathbf{y}_{1:t}) \propto p(\mathbf{y}_{1:t}|\mathbf{x}_{0:t}) p(\mathbf{x}_{0:t}) \quad (9)$$

where $p(\mathbf{x}_{0:t})$ represents our prior knowledge of the state process, and $p(\mathbf{y}_{1:t}|\mathbf{x}_{0:t})$ represents our data or measurement distribution. Typically, a Markov assumption is applied to the prior, so that the state at time t, when conditioned on all previous states only depends on the state at time $t - 1$:

$$p(\mathbf{x}_{0:T}) = p(\mathbf{x}_0) \prod_{t=1}^{T} p(\mathbf{x}_t|\mathbf{x}_{t-1}), \quad (10)$$

where $p(\mathbf{x}_t|\mathbf{x}_{t-1})$ is the evolution distribution, $p(\mathbf{x}_0)$ is the distribution for the "initial state", and T indicates the length of the analysis period of interest. Another critical assumption is that the observations are independent given that one knows the true state. That is,

$$p(\mathbf{y}_{1:T}|\mathbf{x}_{0:T}) = \prod_{t=1}^{T} p(\mathbf{y}_t|\mathbf{x}_t). \tag{11}$$

The Markovian assumption for the evolution model (10) and the conditional independence assumption in the data distribution (11) allows one to write Bayes' rule (9) as

$$p(\mathbf{x}_{0:T}|\mathbf{y}_{1:T}) \propto p(\mathbf{x}_0) \prod_{t=1}^{T} p(\mathbf{y}_t|\mathbf{x}_t) p(\mathbf{x}_t|\mathbf{x}_{t-1}). \tag{12}$$

This form suggests that as new data becomes available, one could update the previous (optimal) estimate of the state process without having to start calculations from scratch. Such sequential updating is the focus of this section.

Filtering

In *filtering*, we assume the density $p(\mathbf{x}_{t-1}|\mathbf{y}_{1:t-1})$ is available and use it to find (i) the forecast distribution, $p(\mathbf{x}_t|\mathbf{y}_{1:t-1})$, and (ii) the analysis distribution, $p(\mathbf{x}_t|\mathbf{y}_{1:t})$. Due to the Markov assumption, the forecast or posterior predictive distribution is readily obtained from

$$p(\mathbf{x}_t|\mathbf{y}_{1:t-1}) = \int p(\mathbf{x}_t|\mathbf{x}_{t-1}) p(\mathbf{x}_{t-1}|\mathbf{y}_{1:t-1}) \mathrm{d}\mathbf{x}_{t-1}. \tag{13}$$

We then obtain the analysis distribution by Bayes' rule:

$$
\begin{aligned}
p(\mathbf{x}_t|\mathbf{y}_{1:t}) &= p(\mathbf{x}_t|\mathbf{y}_t, \mathbf{y}_{1:t-1}) \\
&\propto p(\mathbf{y}_t|\mathbf{x}_t, \mathbf{y}_{1:t-1}) p(\mathbf{x}_t|\mathbf{y}_{1:t-1}) \\
&= p(\mathbf{y}_t|\mathbf{x}_t) p(\mathbf{x}_t|\mathbf{y}_{1:t-1}).
\end{aligned}
\tag{14}
$$

Iterating between the forecast and analysis steps as new data becomes available yields pairs of distributions $p(\mathbf{x}_1|\mathbf{y}_1)$; $p(\mathbf{x}_2|\mathbf{y}_1)$, $p(\mathbf{x}_2|\mathbf{y}_{1:2})$; ...; $p(\mathbf{x}_T|\mathbf{y}_{1:T-1})$, $p(\mathbf{x}_T|\mathbf{y}_{1:T})$.

Smoothing

The term *smoothing* refers to obtaining $p(\mathbf{x}_t|\mathbf{y}_{1:T})$, the distribution of the state at some time t, given all relevant data, even data collected after time t. For $t = T$, the smoothing distribution is just the final analysis (filter) distribution. We seek to obtain the component marginal distributions of the posterior $p(\mathbf{x}_{0:T}|\mathbf{y}_{1:T})$ in a sequential fashion.

We can write the smoothing distribution as

$$p(\mathbf{x}_t|\mathbf{y}_{1:T}) = \int p(\mathbf{x}_t|\mathbf{x}_{t+1}, \mathbf{y}_{1:T}) p(\mathbf{x}_{t+1}|\mathbf{y}_{1:T}) \mathrm{d}\mathbf{x}_{t+1}. \tag{15}$$

Note that

$$p(\mathbf{x}_t|\mathbf{x}_{t+1}, \mathbf{y}_{1:T}) = p(\mathbf{x}_t|\mathbf{x}_{t+1}, \mathbf{y}_{1:t}), \tag{16}$$

since $\{\mathbf{y}_{t+1}, \ldots, \mathbf{y}_T\}$ are assumed to be independent of $\mathbf{x}_t$ given $\mathbf{x}_{t+1}$. Next, Bayes' rule gives

$$
\begin{aligned}
p(\mathbf{x}_t|\mathbf{x}_{t+1}, \mathbf{y}_{1:t}) &\propto p(\mathbf{x}_{t+1}|\mathbf{x}_t, \mathbf{y}_{1:t}) p(\mathbf{x}_t|\mathbf{y}_{1:t}) \\
&= p(\mathbf{x}_{t+1}|\mathbf{x}_t) p(\mathbf{x}_t|\mathbf{y}_{1:t}),
\end{aligned}
\tag{17}
$$

where $p(\mathbf{x}_t|\mathbf{y}_{1:t})$ is the analysis (filter) distribution for time t. Thus, given the analysis distributions, a *forward filtering–backward smoothing* algorithm can be constructed as follows:

for $t = T - 1$ to 1

- obtain $p(\mathbf{x}_t|\mathbf{x}_{t+1}, \mathbf{y}_{1:t})$ via (17), making use of the analysis (filter) distribution $p(\mathbf{x}_t|\mathbf{y}_{1:t})$ and $p(\mathbf{x}_{t+1}|\mathbf{x}_t)$
- obtain the smoothing distribution $p(\mathbf{x}_t|\mathbf{y}_{1:T})$ from (15), making use of the smoothing distribution for time $t + 1$, $p(\mathbf{x}_{t+1}|\mathbf{y}_{1:T})$, obtained at the previous iteration.

In general, one may not be able to obtain analytical representations for the forecast, analysis and smoothing distributions (13)–(15). However, in the case of normal (Gaussian) distributions and linear observation and evolution operators, one can obtain these distributions explicitly. This gives the well-known *Kalman filter* and *Kalman smoother* recursions.

3.1. Kalman filter

The *Kalman filter* is an ideal framework for sequential updating with linear model operators and Gaussian error distributions [28,27,22,21,42]. It can be derived from many different perspectives. Here, we utilize the Bayesian formulas presented previously (e.g., [32,42]).

Define the conditional expectations for "analysis" (filter) and "forecast" as $\mathbf{x}_{t|t} \equiv E(\mathbf{X}_t|\mathbf{y}_{1:t})$ and $\mathbf{x}_{t|t-1} \equiv E(\mathbf{X}_t|\mathbf{y}_{1:t-1})$, respectively. Similarly, define the conditional error covariance matrices for analysis and forecast, respectively:

$$\mathbf{P}_{t|t} = E((\mathbf{X}_t - \mathbf{x}_{t|t})(\mathbf{X}_t - \mathbf{x}_{t|t})'|\mathbf{y}_{1:t}),$$
$$\mathbf{P}_{t|t-1} = E((\mathbf{X}_t - \mathbf{x}_{t|t-1})(\mathbf{X}_t - \mathbf{x}_{t|t-1})'|\mathbf{y}_{1:t-1}).$$

Consider the measurement (data) distribution $p(\mathbf{y}_t|\mathbf{x}_t)$ given by the model:

$$\mathbf{Y}_t = \mathbf{H}_t\mathbf{x}_t + \boldsymbol{\epsilon}_t, \quad \boldsymbol{\epsilon}_t \sim N(\mathbf{0}, \mathbf{R}_t), \tag{18}$$

where $\mathbf{H}_t$ is the observation operator that maps the process to the observations, and $\mathbf{R}_t$ is the (potentially) time-varying observation (measurement) error covariance matrix.

Also, consider the evolution (or process) distribution $p(\mathbf{x}_t|\mathbf{x}_{t-1})$ given by the model:

$$\mathbf{X}_t = \mathbf{M}_t\mathbf{x}_{t-1} + \boldsymbol{\eta}_t, \quad \boldsymbol{\eta}_t \sim N(\mathbf{0}, \mathbf{Q}_t), \tag{19}$$

where $\mathbf{M}_t$ is the (linear) model operator or propagator that maps the evolution of the process in time, and $\mathbf{Q}_t$ is a noise covariance matrix perhaps representing stochastic forcing or features not resolved by the model. Typically, it is assumed that the measurement and model noise processes are independent. We have also assumed noise processes have zero mean, although this need not be the case in general.

Using conditional expectation and conditional variance arguments, the forecast distribution $\mathbf{x}_t|\mathbf{y}_{1:t-1} \sim N(\mathbf{x}_{t|t-1}, \mathbf{P}_{t|t-1})$ with mean and variance are given, respectively, by:

$$\mathbf{x}_{t|t-1} = E(\mathbf{X}_t|\mathbf{y}_{1:t-1}) = E(E(\mathbf{X}_t|\mathbf{x}_{t-1})|\mathbf{y}_{1:t-1})$$
$$= E(\mathbf{M}_t\mathbf{X}_{t-1}|\mathbf{y}_{1:t-1}) = \mathbf{M}_t\mathbf{x}_{t-1|t-1}, \quad (20)$$

and

$$\mathbf{P}_{t|t-1} = \text{var}(\mathbf{X}_t|\mathbf{y}_{1:t-1}) = E(\text{var}(\mathbf{X}_t|\mathbf{x}_{t-1})|\mathbf{y}_{1:t-1})$$
$$+ \text{var}(E(\mathbf{X}_t|\mathbf{x}_{t-1})|\mathbf{y}_{1:t-1})$$
$$= E(\mathbf{Q}_t|\mathbf{y}_{1:t-1}) + \text{var}(\mathbf{M}_t\mathbf{X}_{t-1}|\mathbf{y}_{1:t-1})$$
$$= \mathbf{Q}_t + \mathbf{M}_t\mathbf{P}_{t-1|t-1}\mathbf{M}_t'. \quad (21)$$

In addition, as with the derivation of the multivariate posterior with normal prior and normal data model, the analysis distribution is given by:

$$\mathbf{X}_t|\mathbf{y}_{1:t} \sim N((\mathbf{H}_t'\mathbf{R}_t^{-1}\mathbf{H}_t + \mathbf{P}_{t|t-1}^{-1})^{-1}$$
$$\times \quad (\mathbf{H}_t'\mathbf{R}_t^{-1}\mathbf{y}_t + \mathbf{P}_{t|t-1}^{-1}\mathbf{x}_{t|t-1}), (\mathbf{H}_t'\mathbf{R}_t^{-1}\mathbf{H}_t + \mathbf{P}_{t|t-1}^{-1})^{-1}). \quad (22)$$

Using the same matrix derivation as for the non-sequential case, we can write equivalently the mean and variance of (22):

$$\mathbf{x}_{t|t} = \mathbf{x}_{t|t-1} + \mathbf{K}_t(\mathbf{y}_t - \mathbf{H}_t\mathbf{x}_{t|t-1}) \quad (23)$$
$$\mathbf{P}_{t|t} = (\mathbf{I} - \mathbf{K}_t\mathbf{H}_t)\mathbf{P}_{t|t-1}, \quad (24)$$

where the Kalman gain is given by

$$\mathbf{K}_t = \mathbf{P}_{t|t-1}\mathbf{H}_t'(\mathbf{H}_t'\mathbf{P}_{t|t-1}\mathbf{H}_t + \mathbf{R}_t)^{-1}. \quad (25)$$

Given the parameter matrices $\mathbf{H}_t$, $\mathbf{M}_t$, $\mathbf{Q}_t$, $\mathbf{R}_t$ for $t = 1, \ldots, T$ and initial conditions (or background state) $\mathbf{x}_{0|0} \equiv \mathbf{x}^b$, $\mathbf{P}_{0|0} \equiv \mathbf{P}^b$, one can use the following Kalman filter algorithm to obtain sequential estimates of the state and associated covariance matrices:

for $t = 1$ to T

- get forecasts $\mathbf{x}_{t|t-1}$ and $\mathbf{P}_{t|t-1}$ from (20) and (21), respectively
- get gain $\mathbf{K}_t$, and analysis $\mathbf{x}_{t|t}$, and $\mathbf{P}_{t|t}$ from (25), (23) and (24), respectively.

Note, for time periods in which there are no observations, one simply skips the analysis steps and lets $\mathbf{x}_{t|t} = \mathbf{x}_{t|t-1}$, and $\mathbf{P}_{t|t} = \mathbf{P}_{t|t-1}$.

The statistics literature focuses on the formal treatment of these models when parameters are unknown (especially, $\mathbf{M}_t$, $\mathbf{Q}_t$ and $\mathbf{R}_t$). In cases where these are not time-varying, and the dimensionality n is relatively low, one can use the E–M algorithm [34] or numerical maximum likelihood methods [25] to obtain estimates. In addition, fully-Bayesian (i.e., hierarchical Bayesian) methods can be used for estimating parameters as described in Section 5, as well as [42], and [35].

3.1.1. Kalman filter example

To illustrate the Kalman filter, we consider a simple simulation experiment. Assume we have the univariate measurement model,

$$Y_t = x_t + \epsilon_t, \quad \epsilon_t \sim N(0, 0.1), \quad (26)$$
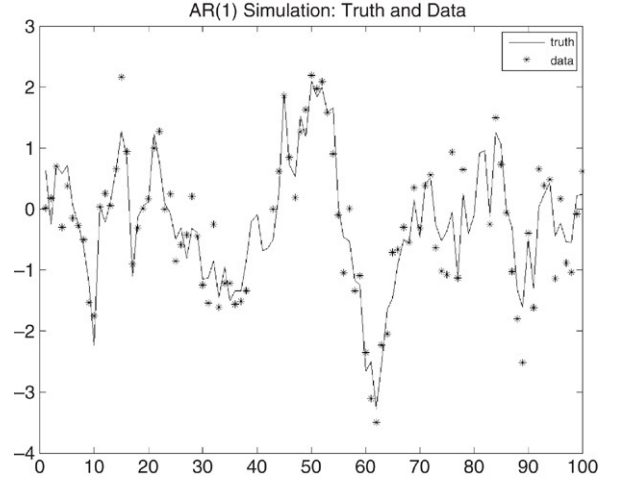

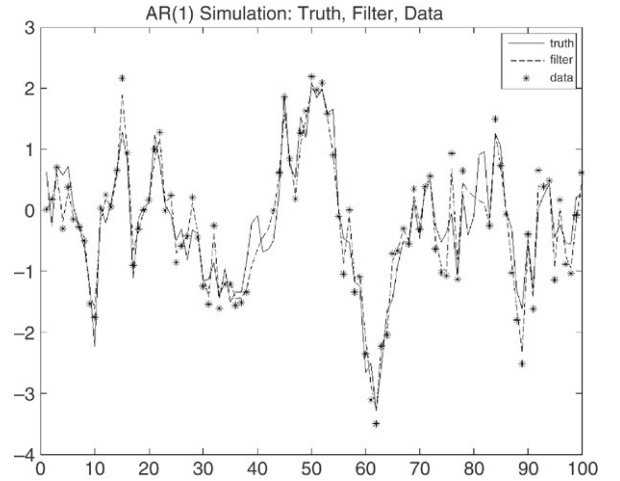Fig. 3. First-order autoregressive process simulation and data.

Fig. 4. First-order autoregressive process simulation Kalman filter state estimates, truth and data.

where clearly in this case $R = 0.1$ and $H = 1$ for all t. Also, assume we have a forecast (or process) model that follows a simple first-order autoregressive (AR(1)) process,

$$X_t = 0.7x_{t-1} + \eta_t, \quad \eta_t \sim N(0, 0.5), \quad (27)$$

where in this case $Q = 0.5$ and $M = 0.7$ for all t. Given an initial condition $x_0 \sim N(0, 1)$ we simulate x_t and y_t for $t = 1, \ldots, 100$. In addition, we let the data at times $t = 40, \ldots, 43$ and $t = 80, \ldots, 83$ be missing. Our objective is to obtain the filtered state estimate $\hat{x}_{t|t}$ and associated variances $\hat{P}_{t|t}$ for all times, given the data y_t, $t = 1, \ldots, 39, 44, \ldots, 79, 84, \ldots, 100$.

Fig. 3 shows the truth x_t and data y_t for the simulation. The Kalman filter state estimates $\hat{x}_{t|t}$ are shown along with the data and true (simulated) state process in Fig. 4. Note that the filter estimates are typically between the noisy observations and the true state, as expected. Also note that the filtered values are not particularly close to the truth in areas of missing data. This is illustrated more clearly in Fig. 5 which shows the estimate of the filter variance $\hat{P}_{t|t}$ for each time. There is a clear increase in
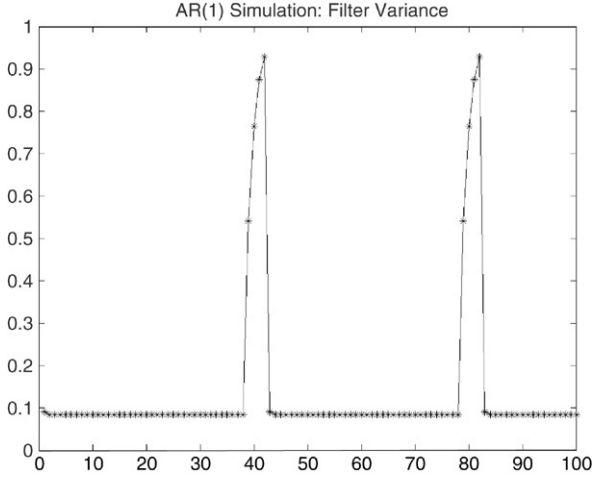
Fig. 5. First-order autoregressive process simulation Kalman filter variance estimates.

the variance as the number of consecutive missing data points increases.

3.2. Kalman smoother

As mentioned previously, in situations where we are interested in the distribution $p(\mathbf{x}_t|\mathbf{y}_{1:T})$, $t = 1, \ldots, T$, then we are said to be interested in the *smoothing* distribution of $\mathbf{X}_t$. That is, the estimate of $\mathbf{X}_t$ at time t given all the data (both before and after time t). This would be useful for a retrospective analysis (i.e., re-analysis). As suggested by the general algorithm given previously, one can utilize the Kalman filter and associated backwards recursion formulas to obtain the smoothing distributions. This procedure is sometimes known as the Kalman Smoother. This result (and other types of smoothers) can be derived from various perspectives. For extensive development, we refer the reader to one of the many excellent texts that discuss filtering and smoothing in detail (e.g., see [27,2,35]). We take a Bayesian approach here (e.g., see [42]).

As with the Kalman filter derivation, let $\mathbf{x}_{t|T} \equiv E(\mathbf{X}_t|\mathbf{y}_{1:T})$ and $\mathbf{P}_{t|T} \equiv \text{var}(\mathbf{X}_t|\mathbf{y}_{1:T})$. We seek sequential formulas for $p(\mathbf{x}_t|\mathbf{y}_{1:T})$ and assume that we have access to the analysis (filter) distributions $\mathbf{X}_t|\mathbf{y}_{1:t} \sim N(\mathbf{x}_{t|t}, \mathbf{P}_{t|t})$ and the forecast distributions $\mathbf{X}_{t+1}|\mathbf{y}_{1:t} \sim N(\mathbf{x}_{t+1|t}, \mathbf{P}_{t+1|t})$ for $t = 1, \ldots, T$. These are available from the Kalman filter algorithm.

First, note that we can obtain $p(\mathbf{x}_t|\mathbf{x}_{t+1}, \mathbf{y}_{1:t})$ from Bayes' rule as suggested by (17). This distribution is then proportional to the product of two Gaussian distributions. By completing the square and using some matrix algebra, this distribution is $N(\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t)$ where

$$\boldsymbol{\mu}_t = \mathbf{x}_{t|t} + \mathbf{J}_t(\mathbf{x}_{t+1} - \mathbf{M}_{t+1}\mathbf{x}_{t|t}) \tag{28}$$

and

$$\boldsymbol{\Sigma}_t = \mathbf{P}_{t|t} - \mathbf{J}_t\mathbf{M}_{t+1}\mathbf{P}_{t|t}, \tag{29}$$

where

$$\mathbf{J}_t \equiv \mathbf{P}_{t|t}\mathbf{M}'_{t+1}(\mathbf{M}_{t+1}\mathbf{P}_{t|t}\mathbf{M}'_{t+1} + \mathbf{Q}_{t+1})^{-1}$$
$$= \mathbf{P}_{t|t}\mathbf{M}'_{t+1}\mathbf{P}^{-1}_{t+1|t}. \tag{30}$$

Next, one obtains the smoother distribution from (15). Specifically, $\mathbf{X}_t|\mathbf{y}_{1:T} \sim N(\mathbf{x}_{t|T}, \mathbf{P}_{t|T})$, where

$$\mathbf{x}_{t|T} = E(\mathbf{X}_t|\mathbf{y}_{1:T}) = E(E(\mathbf{X}_t|\mathbf{x}_{t+1}, \mathbf{y}_{1:t})|\mathbf{y}_{1:T}) = E(\boldsymbol{\mu}_t|\mathbf{y}_{1:T})$$
$$= \mathbf{x}_{t|t} + \mathbf{J}_t(\mathbf{x}_{t+1|T} - \mathbf{x}_{t+1|t}) \tag{31}$$

and

$$\mathbf{P}_{t|T} = \text{var}(\mathbf{X}_t|\mathbf{y}_{1:T}) = E(\text{var}(\mathbf{X}_t|\mathbf{x}_{t+1}, \mathbf{y}_{1:t})|\mathbf{y}_{1:T})$$
$$+ \text{var}(E(\mathbf{X}_t|\mathbf{x}_{t+1}, \mathbf{y}_{1:t})|\mathbf{y}_{1:T})$$
$$= E(\boldsymbol{\Sigma}_t|\mathbf{y}_{1:T}) + \text{var}(\boldsymbol{\mu}_t|\mathbf{y}_{1:T})$$
$$= \mathbf{P}_{t|t} + \mathbf{J}_t(\mathbf{P}_{t+1|T} - \mathbf{P}_{t+1|t})\mathbf{J}'_t. \tag{32}$$

Thus, given that one has the forecast and analysis distributions $\mathbf{x}_{t|t}$, $\mathbf{P}_{t|t}$, $\mathbf{x}_{t+1|t}$, and $\mathbf{P}_{t+1|t}$ for $t = 1, \ldots, T$, the smoother algorithm can be implemented as follows:

For $t = T - 1$ to 1

- obtain $\mathbf{J}_t$ from (30)
- obtain $\mathbf{x}_{t|T}$ from (31)
- obtain $\mathbf{P}_{t|T}$ from (32).

Note that this type of smoothing algorithm is known as a *fixed interval smoother*. Other types of smoothing algorithms (e.g., fixed-point and fixed-lag) can be implemented in "real time" (e.g., see [24]) but are not considered here.

4. Monte Carlo sampling and data assimilation

The forecast and analysis distributions given in (13) and (14), respectively, cannot be obtained explicitly for non-Gaussian models and/or nonlinear dynamic operators. We also assume that the dimensionalities of the relevant vectors prohibit direct numerical integration. A traditional approach to handling nonlinear observation and/or evolution models is local (tangent linear) linearization of the model and evolution operators. In the sequential case, this is known as *extended Kalman filtering* (e.g., [24]). Additionally, some types of non-Gaussian error structures can be accommodated by allowing error structures to be convex mixtures of Gaussian distributions. Alternatively, we may rely on Monte Carlo (MC) methods. This section briefly considers several different MC approaches. For details and theoretical discussion of MC methods, see [33].

4.1. Basic Monte Carlo sampling

Historically, the primary use of Monte Carlo (MC) is for the estimation of integrals (or expectations for probability models). As before, we let the data be represented by $\mathbf{Y}_{1:t} \equiv \{\mathbf{Y}_1, \ldots, \mathbf{Y}_t\}$ and use analogous notation for the state process over time. Let f be a function of the state process and assume a Bayesian context in which we observe data $\mathbf{y}_{1:t}$ and are interested in the conditional expectation of $f(\mathbf{X}_{0:t})$ given the data:

$$E(f(\mathbf{X}_{0:t})|\mathbf{y}_{1:t}) = \int f(\mathbf{x}_{0:t})p(\mathbf{x}_{0:t}|\mathbf{y}_{1:t})d\mathbf{x}_{0:t}$$
$$= \frac{\int f(\mathbf{x}_{0:t})p(\mathbf{y}_{1:t}|\mathbf{x}_{0:t})p(\mathbf{x}_{0:t})d\mathbf{x}_{0:t}}{\int p(\mathbf{y}_{1:t}|\mathbf{x}_{0:t})p(\mathbf{x}_{0:t})d\mathbf{x}_{0:t}},$$

(assuming the integrals exist). A MC estimate can be obtained as follows:

1. generate m pseudo-random realizations, $\mathbf{x}_{0:t}^i$ from $p(\mathbf{x}_{0:t}|\mathbf{y}_{1:t})$, $i = 1, \ldots, m$
2. evaluate f for each realization and compute the arithmetic average of the results, $\hat{\mathrm{E}}(f(\mathbf{X}_{0:t})|\mathbf{y}_{1:t}) = (1/m)\sum_{i=1}^m f(\mathbf{x}_{0:t}^i)$.

Under independent sampling this average converges (almost surely) to $\mathrm{E}(f(\mathbf{X}_{0:t})|\mathbf{y}_{1:t})$ as m approaches infinity. This also holds if realizations are stationary and ergodic though not necessarily independent. We note that the rate of convergence is independent of the dimensionality of the integrand (for details, see [33]). We can also approximate the distribution $p(\mathbf{x}_{0:t}|\mathbf{y}_{1:t})$ by:

$$\hat{p}(\mathbf{x}_{0:t}|\mathbf{y}_{1:t}) \equiv p^m(\mathbf{x}_{0:t}|\mathbf{y}_{1:t}) = \frac{1}{m}\sum_{i=1}^m \delta_{\mathbf{x}_{0:t}^i}$$

where $\delta_{\mathbf{x}_{0:t}^i}$ is a Dirac-delta mass on the trajectory $\mathbf{x}_{0:t}^i$. In practice, one typically considers kernel density estimates of the posterior (e.g., [36,41]).

4.2. Importance sampling Monte Carlo

When direct simulation from $p(\mathbf{x}_{0:t}|\mathbf{y}_{1:t})$ is very difficult, *Importance Sampling Monte Carlo* (*ISMC*) is suggested. The idea is that we consider another distribution, say $g(\mathbf{x}_{0:t}|\mathbf{y}_{1:t})$, that is comparatively easy to sample. We generate m samples $\mathbf{x}_{0:t}^i$ from $g(\mathbf{x}_{0:t}|\mathbf{y}_{1:t})$ and evaluate f for each. A standard procedure for estimating $E(f(\mathbf{X}_{0:t})|\mathbf{y}_{1:t})$ is to weight each sample or ensemble member to adjust for the fact that they are not from the target posterior as follows:

$$\hat{\mathrm{E}}(f(\mathbf{X}_{0:t})|\mathbf{y}_{1:t}) = \sum_{i=1}^m w^i f(\mathbf{x}_{0:t}^i),$$

where the weights w^i are defined by

$$w^i = \frac{p(\mathbf{x}_{0:t}^i|\mathbf{y}_{1:t})/g(\mathbf{x}_{0:t}^i|\mathbf{y}_{1:t})}{\sum_{j=1}^m p(\mathbf{x}_{0:t}^j|\mathbf{y}_{1:t})/g(\mathbf{x}_{0:t}^j|\mathbf{y}_{1:t})}. \tag{33}$$

In this case, we can approximate the posterior distribution by

$$p^m(\mathbf{x}_{0:t}|\mathbf{y}_{1:t}) = \sum_{i=1}^m w^i \delta_{\mathbf{x}_{0:t}^i},$$

or, in practice, by a kernel-density approximation (e.g., [41]).

While efficient selection of the importance function g is the subject of much research, we only consider the following common approach. Assume that g is the prior or forward model distribution:

$$g(\mathbf{x}_{0:t}|\mathbf{y}_{1:t}) = p(\mathbf{x}_{0:t}).$$

Thus, given a sample from $p(\mathbf{x}_0)$, one obtains MC trajectories by simulating from the forward model. Applying (33) for this

g, we obtain

$$w^i \propto \frac{p(\mathbf{x}_{0:t}^i|\mathbf{y}_{1:t})}{p(\mathbf{x}_{0:t}^i)} \propto p(\mathbf{y}_{1:t}|\mathbf{x}_{0:t}^i). \tag{34}$$

Such an ensemble-based approach for obtaining samples from the posterior is sometimes referred to as an *ensemble smoother*. The crucial point to note here is that this approach can be implemented even if we cannot find the normalizing constant of the target posterior $p(\mathbf{x}_{0:t}|\mathbf{y}_{1:t})$.

4.3. Sequential Monte Carlo

In principle, a general sequential MC algorithm follows from the usual sequential update distributions (13) and (14).

As mentioned in the previous section, the choice of the prior distribution as the importance distribution gives weights proportional to the likelihood. We note that these weights can be computed iteratively as follows. Let the weight on ensemble member i at time t be denoted w_t^i. Since the data are assumed to be conditionally independent across time, it follows that

$$w_t^i \propto p(\mathbf{y}_{1:t}|\mathbf{x}_{0:t}^i) \propto p(\mathbf{y}_t|\mathbf{x}_t^i)w_{t-1}^i. \tag{35}$$

Assume we have m samples (or "particles") from the analysis (posterior) distribution at time $t-1$, $p(\mathbf{x}_{t-1}|\mathbf{y}_{1:t-1})$, denoted by $\mathbf{x}_{t-1|t-1}^i$, $i = 1, \ldots, m$. The analysis distribution at time $t-1$ can then be approximated by:

$$p^m(\mathbf{x}_{t-1}|\mathbf{y}_{1:t-1}) = \sum_{i=1}^m \delta_{\mathbf{x}_{t-1|t-1}^i} w_{t-1}^i,$$

where w_{t-1}^i are the normalized ISMC weights for time $t-1$. As with basic Monte Carlo, one often considers a kernel-density estimate of this distribution in practice. This, can then be used to estimate the forecast distribution as the following mixture:

$$p^m(\mathbf{x}_t|\mathbf{y}_{1:t-1}) = \sum_{i=1}^m p(\mathbf{x}_t|\mathbf{x}_{t-1|t-1}^i)w_{t-1}^i.$$

Using the fact that the weights can be updated as in (35), the analysis distribution at time t is then given by:

$$p^m(\mathbf{x}_t|\mathbf{y}_{1:t}) \propto p(\mathbf{y}_t|\mathbf{x}_t) \sum_{i=1}^m p(\mathbf{x}_t|\mathbf{x}_{t-1|t-1}^i)w_{t-1}^i$$

$$= \sum_{i=1}^m p(\mathbf{x}_t|\mathbf{x}_{t-1|t-1}^i)w_t^i.$$

Unfortunately, when the state process or data vectors are of relatively high dimension, the importance weights will degenerate as t increases (i.e., only a few, or one, sample gets all of the weight) and the posterior distribution is not adequately represented by the sample. Various practical solutions to this problem have been considered. We briefly (and generally) discuss a few of these in the following subsections.

4.3.1. Particle filtering

A comprehensive overview of the particle filtering literature can be found in [13]. As mentioned above, overcoming the

degeneracy problem is crucial for practical implementation of such approaches. One way to address this problem is to eliminate the particles having low importance weights and to multiply particles that have high weights (e.g., [23]). There are many approaches to dealing with this degeneracy problem and it is still an active area of research. For example, the Bootstrap filter [13] is implemented as follows:

1. Initialization, $t = 0$
   - for $i = 1, \ldots m$ sample $\mathbf{x}_{0|0}^i \sim p(\mathbf{x}_0)$ and set $t = 1$
2. Importance sampling step
   - for $i = 1, \ldots, m$ sample $\tilde{\mathbf{x}}_t^i \sim p(\mathbf{x}_t|\mathbf{x}_{t-1}^i)$ and set $\tilde{\mathbf{x}}_t^i = \{\mathbf{x}_{t-1}^i, \tilde{\mathbf{x}}_t^i\}$
   - for $i = 1, \ldots, m$ evaluate the importance weights $\tilde{w}_t^i = p(\mathbf{y}_t|\tilde{\mathbf{x}}_t^i)$ (note: in this algorithm these weights are not proportional to the weights at the previous time, $t - 1$, due to the resampling in Step 3, which induces equal weights on the resample)
   - normalize IS weights
3. Selection step
   - resample with replacement m particles $\{\mathbf{x}_t^i : i = 1, \ldots, m\}$ from the set $\{\tilde{\mathbf{x}}_t^i : i = 1, \ldots, m\}$ according to importance weights.
   - set $t = t + 1$ and go to step 2.

Although algorithms such as the Bootstrap filter help with the degeneracy issue, experience has shown that in high dimensional problems like those one would encounter in practical atmospheric/oceanic DA problems, the degeneracy issue is still problematic. A closely related idea is to use MC in the context of the Kalman filter. Such methods are typically referred to as *ensemble Kalman filters*.

4.3.2. Ensemble Kalman filter

The ensemble Kalman filter (EnKF) was originally developed by Evensen [16], Evensen and van Leeuwen [17] and Houtekamer and Mitchell [26]. The basic approach uses Monte Carlo samples to approximate the forecast distribution while, critically, still using the nonlinear forward model. In particular, one estimates the prior (forecast) means and variance/covariance matrices with the Monte Carlo samples (ensembles). These are then used in the linear KF update formulas to obtain the analysis distribution. There are many different variants of this approach and it remains a very active area of research. We will consider a very basic form here. This "classical" EnKF approach can be considered as an approximation to the sequential importance sampling algorithm.

Assume one has m independent samples from the analysis distribution at time $t - 1$, $\mathbf{x}_{t-1|t-1}^i$. Implicitly, we assume that these samples are equally weighted, so $w_{t-1}^i = 1/m$. As we have seen with ISMC, this assumption is not valid in general and will lead to biased samples. Since it is assumed that we have independent samples from the posterior at time $t - 1$, we can use basic MC to represent the forecast distribution (again, like sequential ISMC but with weights $= 1/m$):

$$p^m(\mathbf{x}_t|\mathbf{y}_{1:t-1}) = (1/m) \sum_{i=1}^m p(\mathbf{x}_t|\mathbf{x}_{t-1|t-1}^i).$$

Next, we assume that the forecast distribution can be characterized by its first two moments (or, equivalently, that it is Gaussian with mean $\mathbf{x}_{t|t-1}^i$ and (estimated) variance/covariance matrix $\hat{\mathbf{P}}_{t|t-1}$). Then, the update (analysis distribution at time t) is given by:

$$p^m(\mathbf{x}_t|\mathbf{y}_{1:t}) \propto (1/m) p(\mathbf{y}_t|\mathbf{x}_t) \sum_{i=1}^m N(\mathbf{x}_{t|t-1}^i, \hat{\mathbf{P}}_{t|t-1}).$$

Clearly, if we assume the measurement distribution is also Gaussian, then this analysis distribution can be computed analytically as usual for the Bayesian normal prior, normal data model case by using the Kalman filter update equations.

Specifically, assume a linear (or linearized) observation model $\mathbf{y}_t = \mathbf{H}_t \mathbf{x}_t + \mathbf{e}_t$, where the covariance of $\mathbf{e}_t$ is $\mathbf{R}_t$. Let $\mathbf{x}_{t|t-1}$ and $\mathbf{P}_{t|t-1}$ denote the mean and covariance matrix, respectively, of the predictive distribution $p(\mathbf{x}_t|\mathbf{y}_{1:t-1})$. We seek to obtain a viable ensemble from the analysis distribution, $p(\mathbf{x}_t|\mathbf{y}_{1:t})$, or at least ensemble based estimates of its mean $\mathbf{x}_{t|t}$ and covariance matrix $\mathbf{P}_{t|t}$. Thus, assuming one has available independent samples $\mathbf{x}_{t-1|t-1}^i, i = 1, \ldots, m$ from the analysis distribution at time $t - 1$, the following steps form the basic algorithm.

- Forecast each of the samples $\mathbf{x}_{t-1|t-1}^i$ forward using the evolution model:

$$\mathbf{x}_{t|t-1}^i = \mathcal{M}(\mathbf{x}_{t-1|t-1}^i) + \eta_t^i, \quad \eta_t^i \sim N(\mathbf{0}, \mathbf{Q}).$$

  Note that in applications in which no model error is assumed, one just evolves the sample forward using the $\mathcal{M}()$ model evolution, but no additive noise.

- Use the forecast samples to calculate a sample forecast covariance matrix, $\hat{\mathbf{P}}_{t|t-1}$:

$$\hat{\mathbf{P}}_{t|t-1} = \frac{1}{m-1} \sum_{i=1}^m (\mathbf{x}_{t|t-1}^i - \hat{\mathbf{x}}_{t|t-1})(\mathbf{x}_{t|t-1}^i - \hat{\mathbf{x}}_{t|t-1})'$$

  where $\hat{\mathbf{x}}_{t|t-1} = (1/m) \sum_{i=1}^m \mathbf{x}_{t|t-1}^i$.

- Use the Kalman Filter update equations to update each forecast sample given the sampled observations:

$$\mathbf{x}_{t|t}^i = \mathbf{x}_{t|t-1}^i + \mathbf{K}_t(\mathbf{y}_t + \mathbf{e}_t^i - \mathbf{H}_t \mathbf{x}_{t|t-1}^i)$$

  where

$$\mathbf{K}_t = \hat{\mathbf{P}}_{t|t-1} \mathbf{H}_t' (\mathbf{H}_t \hat{\mathbf{P}}_{t|t-1} \mathbf{H}_t' + \mathbf{R})^{-1}$$

  and

$$\mathbf{e}_t^i \sim N(\mathbf{0}, \mathbf{R}), \quad i = 1, \ldots, m$$

  give simulated observations $\mathbf{y}_t + \mathbf{e}_t^i$ (necessary to ensure that the analysis distribution has the appropriate spread). In addition,

$$\hat{\mathbf{P}}_{t|t} = \frac{1}{m-1} \sum_{i=1}^m (\mathbf{x}_{t|t}^i - \hat{\mathbf{x}}_{t|t})(\mathbf{x}_{t|t}^i - \hat{\mathbf{x}}_{t|t})'$$

  where $\hat{\mathbf{x}}_{t|t} = (1/m) \sum_{i=1}^m \mathbf{x}_{t|t}^i$.

Critically, we note that unless $\mathcal{M}$ is linear, $p(\mathbf{x}_t|\mathbf{y}_{1:t-1})$ cannot be Gaussian if $p(\mathbf{x}_{t-1}|\mathbf{y}_{1:t-1})$ is Gaussian. Hence, the EnKF is analogous to a "best linear-in-the observations"

Bayesian procedure (e.g., [42]). It is often considered a desirable feature of the EnKF that it yields the correct updating if all distributions are Gaussian. However, in nonlinear cases, since Gaussianity cannot hold for all time, the EnKF must yield biased samples and estimates, even for unlimited sample sizes (i.e., it corresponds to unweighted ISMC procedures).

There are several modifications to the EnKF algorithm that are typically used in practice. First, because of the dimensionality of the state process in most DA applications, it is not possible to evolve the error covariance matrices according to the Kalman filter equations. Rather, the forecast samples are used to calculate a *sample* forecast covariance matrix as given above. However, typically, m is relatively small because it is too expensive to run the forward model. Thus, the covariance estimates are not stable or not of full rank when the dimension of $\mathbf{X}_t$ is larger than m. The standard solution is to consider the *Shur product* (or *Hadamard product*). This is an "element-by-element" multiplication of the ensemble estimated covariance matrix by a correlation matrix $\mathbf{S}$ that has "compact support". That is, one uses $\mathbf{P}_{t|t-1} \circ \mathbf{S}$ where $\circ$ is the Shur product and $\mathbf{S}$ is a correlation matrix. We note that the product of a covariance matrix and a correlation matrix is a covariance matrix.

In addition, it is computationally more efficient to calculate the elements of $\mathbf{K}_t$ directly, rather than $\mathbf{P}_{t|t-1}$:

$$\hat{\mathbf{P}}_{t|t-1}\mathbf{H}'_t = \frac{1}{m-1}\sum_{i=1}^{m}(\mathbf{x}^i_{t|t-1} - \mathbf{x}_{t|t-1})$$
$$\times (\mathbf{H}_t\mathbf{x}^i_{t|t-1} - \overline{\mathbf{H}_t\mathbf{x}_{t|t-1}})'$$

$$\mathbf{H}_t\hat{\mathbf{P}}_{t|t-1}\mathbf{H}'_t = \frac{1}{m-1}\sum_{i=1}^{m}(\mathbf{H}_t\mathbf{x}^i_{t|t-1} - \overline{\mathbf{H}_t\mathbf{x}_{t|t-1}})$$
$$\times (\mathbf{H}_t\mathbf{x}^i_{t|t-1} - \overline{\mathbf{H}_t\mathbf{x}_{t|t-1}})',$$

where $\overline{\mathbf{H}_t\mathbf{x}_{t|t-1}} = (1/m)\sum_{i=1}^{m}\mathbf{H}_t\mathbf{x}^i_{t|t-1}$.

As might be expected, for situations where m is small relative to the state dimension, effects from sampling variability can be problematic. Some approaches to EnKF such as the square-root filter [40] attempt to address this issue, more or less.

EnKF example: Consider again the AR(1) simulation discussed in Section 3.1.1. Given these same simulated data, we implement an EnKF with 10 ensemble members. Fig. 6 shows the filter state estimates in this case, compared to the Kalman filter estimate discussed in Section 3.1.1. In this simple case, the EnKF is remarkably close to the Kalman filter estimates even with only 10 ensemble members. However, as shown in Fig. 7, the estimated state variances for the EnKF with 10 ensembles is quite variable when compared to the Kalman filter variance. Also shown in this figure is the filter variance for a 100 member EnKF. This estimate is clearly more stable relative to the Kalman filter variance than the 10 member ensemble.

5. Hierarchical models

Relatively recent computational advancements in Bayesian statistics have led to increased use of hierarchical models for complicated problems (e.g., see [20] for a general overview). From a DA perspective, such models are ideal for retrospective
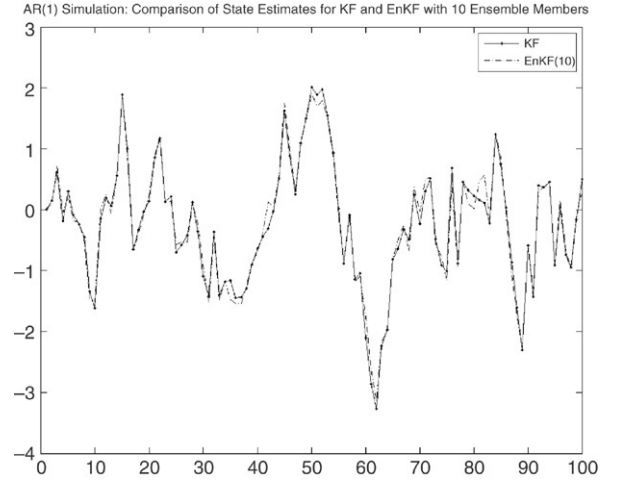

Fig. 6. First-order autoregressive process simulation Kalman filter state estimates and EnKF filter state estimates based on 10 ensemble members.
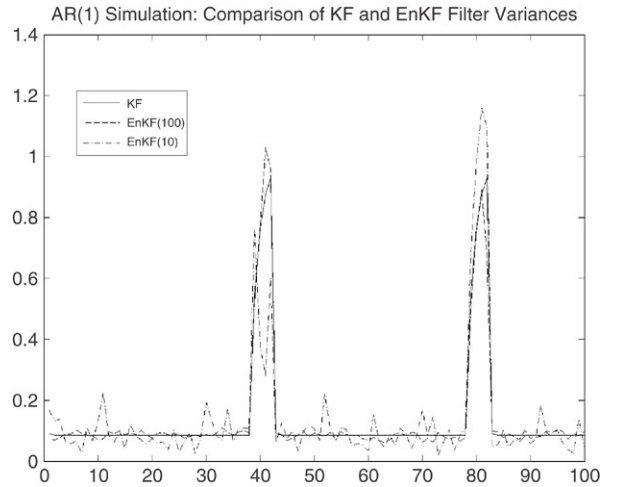

Fig. 7. First-order autoregressive process simulation Kalman filter variance estimates and EnKF variance estimates for 10 and 100 member ensembles.

analysis where there is significant uncertainty in the process and parameters that control the model. The central idea of hierarchical modeling is based on the notion of conditioning, and that by factoring a complicated joint distribution into a series of conditional distributions, one is better able to utilize scientific knowledge and characterize uncertainty. Say we have three random variables, a, b, c. Basic principles of probability always allow one to factor the joint distribution of these variables as, for example: $p(a, b, c) = p(a|b, c)p(b|c)p(c)$. It is typically the case that one can specify these component distributions (i.e., $p(a|b, c)$, $p(b|c)$, and $p(c)$) more easily than the full joint distribution. For environmental modeling, as initially described in [5], it is convenient to consider a general three-stage factorization of p(data, process, parameters). These stages are given by:

Stage 1. Data model: p(data|process, parameters)
Stage 2. Process model: p(process|parameters)
Stage 3. Parameter model: p(parameters)

Note that the first two stages in this general hierarchy are just the first two stages of the state–space formulation used

in sequential updating (measurement distribution and state evolution distribution). The difference is that here one considers that there are parameters that should be considered as random variables, and thus we must have a modeling stage for those parameters. In a modeling context, one is interested in the distribution of the process and parameters given the data. In principle, this can be obtained by Bayes' rule:

$$p(\text{process, parameters}|\text{data}) \propto p(\text{data}|\text{process, parameters})$$
$$p(\text{process}|\text{parameters}) p(\text{parameters}). \tag{36}$$

For an overview of this framework in an atmospheric setting, see [49] and [43], or for a more general overview in terms of environmental modeling see [6] and [44]. Here, we outline briefly these stages.

5.1. Component models

Although the factorization (36) seems simple, it is in fact quite powerful in that one can accommodate various sources of uncertainty in the component models. Furthermore, each of these component models can often be further factorized depending on the specific problem of interest.

5.1.1. Data models

Let Y_a be data associated with a process X and let θ_a be parameters associated with the data (measurement) model. As in the usual state–space formulation presented previously, the data model is written: $p(y_a|x, \theta_a)$. This conditional distribution is much simpler than the unconditional distribution of Y_a since most of the complicated structure in the data comes from the process X. More generally, the power of the hierarchical approach comes from the ability to accommodate multiple data sets at various resolutions and alignments. For example, given observations Y_c, Y_a for the same process, X, we often can write:

$$p(y_a, y_c|x, \theta_a, \theta_c) = p(y_a|x, \theta_a) p(y_c|x, \theta_c).$$

In other words, conditioned on the true process, the data are often independent. Clearly, it is not typically the case that the data sets would be *unconditionally* independent. In addition, this framework presents a natural way to accommodate data at differing resolutions or from distinct platforms (e.g., [46,19,44, 48]).

For multivariate processes x_a, x_c, we often can write:

$$p(y_a, y_c|x_a, x_c, \theta_a, \theta_c) = p(y_a|x_a, \theta_a) p(y_c|x_c, \theta_c),$$

where again, conditional on the true processes, it is often reasonable to assume that the data are independent.

5.1.2. Process models

As with data models, process models are also often factored into a series of conditional models. For example,

$$p(x_a, x_c|\theta_x) = p(x_a|x_c, \theta_x) p(x_c|\theta_x).$$

This is exactly the assumption that is made in the state–process model used for the sequential analysis (in that case, the "a" and "c" subscripts refer to time t and $t-1$, respectively, etc.). Such

factorizations are also important for simplifying multivariate processes as well. For example, Berliner et al. [8] consider such a conditional framework for modeling the ocean conditional on the atmosphere, which provides a unified probabilistic approach for coupling processes.

5.1.3. Parameter models

Parameter models can also be factored into subcomponents. For example, we might assume,

$$p(\theta_a, \theta_c, \theta_x) = p(\theta_a) p(\theta_c) p(\theta_x).$$

That is, we might assume parameter distributions are independent (if justified) or we might be able to use previous studies to facilitate the specification of these models. For example, measurement error parameters can often be obtained from previous studies which focus on such issues. Furthermore, process parameters often carry scientific insight such as spatially-dependent diffusion parameters [45] or turbulence characteristics as in [46]. In other cases, we don't know much about the parameters and use vague or non-informative distributions for parameters. Alternatively, we might use data-based estimates for such parameters, although the use of such estimates does not follow the traditional Bayesian paradigm.

5.2. Conceptual example

One of the strengths of the hierarchical approach is that it can accommodate the knowledge that the "deterministic" model for the process of interest is inherently inadequate to describe the real-world process. For example, it is often the case that relatively simple differential equation-based models for the process, with random parameters (and possibly complicated dependence structures) can model processes more complicated than suggested by the original deterministic differential equations.

For example, assume the true system follows the one-dimensional *Burgers Equation*

$$\frac{\partial u}{\partial t} = -u\frac{\partial u}{\partial x} + A\frac{\partial^2 u}{\partial x^2}.$$

As is the case in reality, assume our physical knowledge of the system is incomplete and that we (naively) believe that linear advection diffusion equation dynamics are appropriate for the system,

$$\frac{\partial u}{\partial t} = -\alpha\frac{\partial u}{\partial x} + A\frac{\partial^2 u}{\partial x^2}.$$

Finite differencing suggests the following difference equation representation for the u process at a location, s:

$$u_t(s) = \theta_1 u_{t-1}(s) + \theta_2 u_{t-1}(s + \delta_s) + \theta_3 u_{t-1}(s - \delta_s),$$

where θ_i are functions of δ_s (spatial discretization), δ_t (time discretization), A, and α. In vector form, for $\mathbf{u}_t = [u_t(s_1), \ldots, u_t(s_n)]'$:

$$\mathbf{u}_t = \mathbf{M}\mathbf{u}_{t-1} + \mathbf{M}^b\mathbf{u}_{t-1}^b,$$

where $\mathbf{M}$, $\mathbf{M}^b$ are functions of $\theta_1, \ldots, \theta_3$; $\mathbf{u}_t^b = [u_t(s_0) \\ u_t(s_{n+1})]'$ corresponds to the boundary process. Specifically,

$$\mathbf{M} = \begin{bmatrix} \theta_1 & \theta_2 & 0 & \cdots & 0 \\ \theta_3 & \theta_1 & \theta_2 & & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & & \theta_3 & \theta_1 & \theta_2 \\ 0 & \cdots & 0 & \theta_3 & \theta_1 \end{bmatrix} \quad \text{and} \quad \mathbf{M}^b = \begin{bmatrix} \theta_3 & 0 \\ 0 & 0 \\ \vdots & \vdots \\ 0 & 0 \\ 0 & \theta_2 \end{bmatrix}.$$

It is important to note that we do not assume this is the correct model, but rather let the state process ($\mathbf{U}_t$) be random, allow the parameters be random, and add (possibly) correlated noise in an attempt to allow the data to help modify our prior dynamical model. In this case, given observations $\mathbf{y}_t$, we can write the following hierarchical model:

- *Data model:*

  $$\mathbf{y}_t | \mathbf{u}_t, \mathbf{R}_t \sim N(\mathbf{H}_t \mathbf{u}_t, \mathbf{R}_t), \quad t = 1, \ldots, T$$

  where $\mathbf{y}_t$ is $p_t \times 1$, $\mathbf{u}_t$ is $n \times 1$, $\mathbf{H}_t$ is an $p_t \times n$ matrix that maps data to prediction locations, and $\mathbf{R}_t$ is a $p_t \times p_t$ observation error covariance matrix.

- *Process model:*

  $$\mathbf{u}_t | \mathbf{u}_{t-1}, \boldsymbol{\theta}, \mathbf{Q}(\gamma) \sim N(\mathbf{M}(\boldsymbol{\theta}) \mathbf{u}_{t-1} + \mathbf{M}^b(\boldsymbol{\theta}) \mathbf{u}_{t-1}^b, \mathbf{Q}(\gamma)),$$

  for $t = 1, \ldots, T$ where we have used notation to indicate that the propagator matrices depend on the parameters $\boldsymbol{\theta}$. In addition, we note that it is typically the case that the error covariance matrix $\mathbf{Q}$ will be parameterized in the hierarchical setting, and we denote those parameters by γ (where these parameters are sometimes given distributions at a lower stage of the model hierarchy). In some cases, one might assume the boundary process $\mathbf{u}_t^b$ is known. However, this is often not realistic and one should model this process as well. A significant advantage of the hierarchical approach is that this is fairly easy to do by specifying a distribution for this process at a lower stage of the model hierarchy (e.g., [47]). That is, in the primary process stage, one is modeling the interior process *conditional* on this boundary process (assuming it is known), whereas the variability of the boundary process is accounted for at the lower stage. We also must specify the prior distribution for the initial condition, $\mathbf{u}_0 \sim N(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$, where we give $\boldsymbol{\mu}_0$ and $\boldsymbol{\Sigma}_0$ similarly to the specification of background means and covariances in traditional DA.

- *Parameter models:* Priors for θ_i are also needed. That is, we must specify the joint distribution of θ_1, θ_2, and θ_3. This may be simplified if we take into consideration the relationships suggested by the finite difference representation:

  $$\tilde{\theta}_1 = 1 - \frac{2\delta_t A}{\delta_s^2},$$

  $$\tilde{\theta}_2 = \frac{\delta_t A}{\delta_s^2} - \frac{\delta_t \alpha}{\delta_s},$$

  $$\tilde{\theta}_3 = \frac{\delta_t A}{\delta_s^2} + \frac{\delta_t \alpha}{\delta_s},$$

  where δ_s and δ_t are the spatial and temporal discretization intervals. It might be reasonable to assume conditional

independence in this case, $\theta_i \sim N(\tilde{\theta}_i, \sigma_i^2)$, where σ_i^2 are specified to reflect ones prior belief in how closely the system should follow the advection–diffusion dynamics. Since the dependence of these parameters is most likely due to the common parameters (e.g., A and α), if we condition on these, then it might be a reasonable prior model to assume independence of the θ_i. In that case, one must then specify a prior distribution for A and α. In such a prior, one might allow these parameters to vary spatially (e.g., [45]). In that case, spatial dependence is built into the distributions of these parameters to allow them to vary relatively smoothly over the spatial domain, and to borrow strength from data rich areas to data poor areas. Typically, since we are not trying to solve the PDE explicitly, we do not specify values for δ_s and δ_t. Rather, these are accounted for implicitly in the random parameters $\tilde{\theta}_i$ or absorbed in the distributions of A and α.

Importantly, one may need to constrain these distributions to ensure non-explosive (stable) growth. Depending on the specific parameterization chosen, one might be able to work out the appropriate analytical conditions on θ_i to guarantee stability (i.e., so that the eigenvalues of $\mathbf{M}(\boldsymbol{\theta})$ are less than one in modulus). Alternatively, one might have to check these conditions numerically as the θ_i parameters are sampled in the estimation algorithm (e.g., MCMC as discussed in the next section). However, experience has shown that such stability considerations (e.g., CFL conditions) are not as important in Bayesian hierarchical models as in traditional numerical solutions to PDEs because data is available to naturally constrain the solution if needed. It is important to remember that we are not seeking numerical solutions to the PDE here, but rather are using the discretized PDE to suggest the form of the statistical model. This is a promising area of research in hierarchical Bayesian physical/statistical modeling.

5.3. Bayesian computation for hierarchical models, MCMC

How does one do the Bayesian computation for hierarchical models? In relatively low-dimensional problems, importance sampling can be used (e.g., [8]). In higher dimensional problems, one typically uses some form of *Markov Chain Monte Carlo (MCMC)* algorithm. In MCMC simulation, one constructs a Markov chain to have a stationary, ergodic distribution which coincides with the posterior distribution of interest. Simulations from the chain then converge to realizations from the posterior. Asymptotically, the samples for a given variable sampled in this way are identically distributed, but dependent. For a more formal discussion of MCMC see, for example, [33].

As it turns out, there are several relatively simple algorithms that accomplish this. These include the Gibbs, Metropolis-Hastings, and slice samplers. For illustration, consider the Gibbs sampler. First, we must define the necessary distributions. Assume that our data correspond to $\mathbf{y}$ and our variables of interest (including process and parameters) are denoted by $\boldsymbol{\theta}_i, i = 1, \ldots, p$. Then, our target posterior, for which we would like to draw samples

is given by: $p(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_p|\mathbf{y})$. To implement the sampler, for each θ_i one needs the so-called *full conditional distribution*: $p(\boldsymbol{\theta}_i|\text{all other } \boldsymbol{\theta}_j, \mathbf{y})$; we abbreviate this $p(\boldsymbol{\theta}_i|\cdot)$. Then, the Gibbs sampling algorithm is given by:

- Initialize: $\boldsymbol{\theta}_1^{(0)}, \ldots, \boldsymbol{\theta}_p^{(0)}$
- Iterate: given $\boldsymbol{\theta}_1^{(i)}, \ldots, \boldsymbol{\theta}_p^{(i)}$, generate $\boldsymbol{\theta}_1^{(i+1)}, \ldots, \boldsymbol{\theta}_p^{(i+1)}$ from the following sequence of full-conditionals:

$$p(\boldsymbol{\theta}_1^{(i+1)}|\boldsymbol{\theta}_2^{(i)}, \ldots, \boldsymbol{\theta}_p^{(i)}, \mathbf{y})$$
$$p(\boldsymbol{\theta}_2^{(i+1)}|\boldsymbol{\theta}_1^{(i+1)}, \boldsymbol{\theta}_3^{(i)}, \ldots, \boldsymbol{\theta}_p^{(i)}, \mathbf{y})$$
$$p(\boldsymbol{\theta}_3^{(i+1)}|\boldsymbol{\theta}_1^{(i+1)}, \boldsymbol{\theta}_2^{(i+1)}, \boldsymbol{\theta}_4^{(i)}, \ldots, \boldsymbol{\theta}_p^{(i)}, \mathbf{y})$$
$$\vdots$$
$$p(\boldsymbol{\theta}_p^{(i+1)}|\boldsymbol{\theta}_1^{(i+1)}, \ldots, \boldsymbol{\theta}_{p-1}^{(i+1)}, \mathbf{y}).$$

This algorithm is easy to implement if the full conditionals can be derived and are relatively easy to simulate from. Otherwise, additional sampling methods within the sampler must be utilized (see [33]). One must also decide on initial conditions to start the Markov chain, how long to let the chain "burn-in", how to account for the dependency in the samples, and how to establish convergence. These issues are beyond the scope of this article, but can be found in the literature (e.g., [33]). In addition, even if the full conditional distributions are conceptually easy, implementation may be difficult or impossible for problems with high-dimensional data, process, or parameters. Furthermore, implementation can be problematic in cases with non-linear operators and/or distributions for which one cannot analytically derive the full-conditional distributions.

5.3.1. Simple Gibbs sampling example

To illustrate the MCMC Gibbs sampler approach to DA, we again consider the simple AR(1) process with measurement error example discussed in Section 3.1.1. In this case, our data model is given by (26) and our process model given by (27). We also have an initial condition prior $X_0 \sim N(0, 1)$ as stated in Section 3.1.1. Assume for illustration that we "know" the measurement error variance $R = 0.1$ but do not know the evolution operator M nor the process variance σ_η^2. Thus, we specify prior distributions for these, $M \sim U(-1, 1)$, $\sigma_\eta^2 \sim \text{IG}(2, 1)$, where $U(w_1, w_2)$ corresponds to a continuous uniform distribution between w_1 and w_2, and $\text{IG}(a, b)$ corresponds to an inverse gamma distribution with shape parameter a and scale parameter b.

In this case, the posterior distribution of interest is:

$$p(M, \sigma_\eta^2, x_0, \ldots, x_T|y_1, \ldots, y_T)$$
$$\propto \prod_{t=1}^{T} p(y_t|x_t)p(x_t|x_{t-1}, M, \sigma_\eta^2)p(x_0)p(M)p(\sigma_\eta^2). \quad (37)$$

Given the uncertainty in the parameters, we cannot find the normalizing constant for this posterior distribution analytically. We can, however, use MCMC (specifically, a Gibbs sampler) to obtain samples from this distribution. To do so, we need the following full-conditional distributions:

$$p(x_0|\cdot) \propto p(x_1|x_0, M, \sigma_\eta^2)p(x_0) = N((M^2/\sigma_\eta^2 + 1)^{-1}$$
$$\times (Mx_1/\sigma_\eta^2), (M^2/\sigma_\eta^2 + 1)^{-1}) \quad (38)$$

$$\begin{aligned}
p(x_t|\cdot) &\propto p(x_{t+1}|x_t, M, \sigma_\eta^2)p(x_t|x_{t-1}, M, \sigma_\eta^2)p(y_t|x_t)\\
&= N((M^2/\sigma_\eta^2 + 1/\sigma_\eta^2 + 1/\sigma_\epsilon^2)^{-1}\\
&\quad \times (Mx_{t+1}/\sigma_\eta^2 + Mx_{t-1}/\sigma_\eta^2 + y_t/\sigma_\epsilon^2),\\
&\quad (M^2/\sigma_\eta^2 + 1/\sigma_\eta^2 + 1/\sigma_\epsilon^2)^{-1}),\\
&\quad \text{for } t = 1, \ldots, T-1.
\end{aligned} \quad (39)$$

$$\begin{aligned}
p(x_T|\cdot) &\propto p(x_T|x_{T-1}, M, \sigma_\eta^2)p(y_T|x_T)\\
&= N((1/\sigma_\eta^2 + 1/\sigma_\epsilon^2)^{-1}\\
&\quad \times (Mx_{T-1}/\sigma_\eta^2 + y_T/\sigma_\epsilon^2), (1/\sigma_\eta^2 + 1/\sigma_\epsilon^2)^{-1}). \quad (40)
\end{aligned}$$

$$\begin{aligned}
p(M|\cdot) &\propto \prod_{t=1}^{T} p(x_t|x_{t-1}, M, \sigma_\eta^2)p(M)\\
&= N_{[-1,1]}\left(\left(\sum_{t=1}^{T} x_{t-1}^2/\sigma_\eta^2\right)^{-1}\right.\\
&\quad \left.\times \left(\sum_{t=1}^{T} x_t x_{t-1}/\sigma_\eta^2\right), \left(\sum_{t=1}^{T} x_{t-1}^2/\sigma_\eta^2\right)^{-1}\right), \quad (41)
\end{aligned}$$

where $N_{[-1,1]}()$ indicates a normal distribution truncated between -1 and 1. Finally,

$$\begin{aligned}
p(\sigma_\eta^2|\cdot) &\propto \prod_{t=1}^{T} p(x_t|x_{t-1}, M, \sigma_\eta^2)p(\sigma_\eta^2)\\
&= \text{IG}\left(T/2 + a, \left(1/b + 0.5\sum_{t=1}^{T}(x_t - Mx_{t-1})^2\right)^{-1}\right). \quad (42)
\end{aligned}$$

Note that the normal full-conditional distributions can be derived by completing the square and the inverse gamma distribution can be derived by recognizing that the IG prior on σ_η^2 from a normal distribution is conjugate (e.g., see [20]). We also note that the full-conditionals for x_t, $t = 1, \ldots, T$ assume that there are no missing data. The corresponding full-conditional for the case of a missing y_t is also normal but without the terms corresponding to y_t and σ_ϵ^2. These are not presented here for space considerations.

To implement the Gibbs sampler, one could use the following pseudo-code:

Assign initial values to: $x_t^{(0)}$, $t = 0, \ldots, T$, $M^{(0)}$, $\sigma_\eta^{2(0)}$
for $k = 1$ to m
    $x_0^{(k)} = $ sample from (38) given $x_1^{(k-1)}$, $M^{(k-1)}$, $\sigma_\eta^{2,(k-1)}$
    for $t = 1, \ldots, T-1$, $x_t^{(k)} = $ sample from (39) given $x_{t-1}^{(k)}$, $x_{t+1}^{(k-1)}$, $M^{(k-1)}$, $\sigma_\eta^{2,(k-1)}$
    $x_T^{(k)} = $ sample from (40) given $x_{T-1}^{(k)}$, $M^{(k-1)}$, $\sigma_\eta^{2,(k-1)}$
    $M^{(k)} = $ sample from (41) given $x_t^{(k)}$, $\sigma_\eta^{2,(k-1)}$
    $\sigma_\eta^{2,(k)} = $ sample from (42) given $x_t^{(k)}$, $M^{(k)}$
end

Given data and starting values, one could generate m samples from this algorithm. We note that for Markovian state–space process models such as presented here, there are
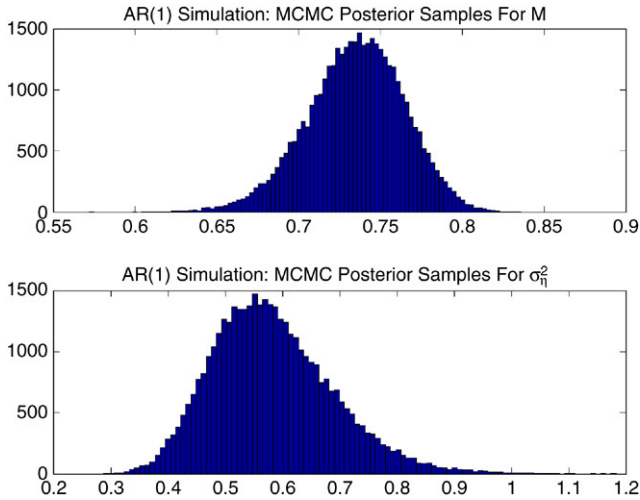
Fig. 8. MCMC Posterior samples for M (top) and σ_η^2 (bottom) based on the AR(1) simulation described in Section 3.1.1.
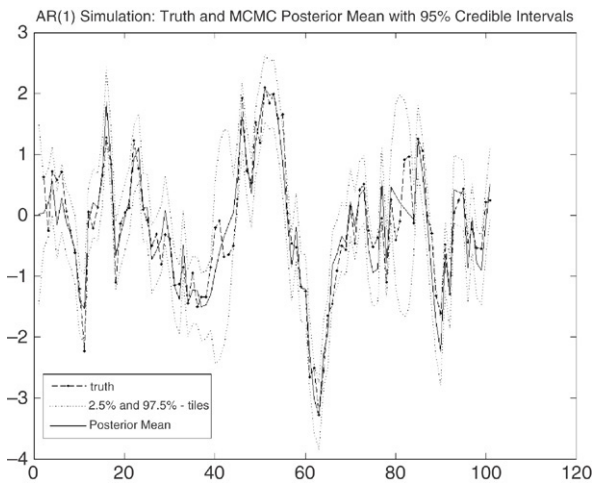

Fig. 9. MCMC posterior mean and 2.5% and 97.5%-tiles for AR(1) simulation described in Section 3.1.1.

more efficient sampling approaches than the basic approach given above. These typically make use of the Kalman filter and smoother algorithms and effectively allow one to update the full state–process at once (e.g., see [42] for an overview).

This algorithm was implemented on the simulated data discussed in Section 3.1.1 with an MCMC sample of 40,000 after a 10,000 sample burn-in. Fig. 8 shows histograms of MCMC samples from the posterior for M and σ_η^2. Clearly, the true values for these parameters (0.7 and 0.5, respectively) used in the simulation are contained in these posterior distributions. In addition, Fig. 9 shows the posterior mean of the state process (x_t), and the 95% Bayesian credible intervals (for each time) as a measure of uncertainty. Note that the posterior distribution seems to cover the true process and the uncertainty in the data free regions ($t = 40$–43, $t = 80$–83) is much larger, as expected.

5.3.2. Real world examples

Although hierarchical Bayesian methods are not yet practical for operational DA in the atmospheric sciences, they

have been used successfully in various smaller-scale contexts related to DA. For example,

- Berliner et al. [7] develop a Bayesian hierarchical model for tropical Pacific sea surface temperature. This model uses qualitative dynamics in reduced dimensional (EOF) space, conditioned on hidden processes. In this case, the dynamical propagator is time-varying, depending on the current and projected climate state.
- Wikle et al. [46] use a hierarchical Bayesian model to predict tropical surface winds given high-resolution satellite scatterometer observations and low-resolution analysis fields. In this context the data models included (low resolution) data from analysis fields and high resolution, but incomplete, satellite scatterometer observations of winds. The process model was based on equatorial shallow water equations with random evolution parameters for large-scale dynamics and a multiresolution (wavelet) process for small-scale dynamics. Parameter model distributions for the equatorial modes was based on the climatology, and priors for multiresolution modes based on observed power-law (5/3 slope) scaling behavior in tropical surface winds over the ocean.
- Berliner et al. [8] use the Bayesian hierarchical approach to couple models for the atmosphere and ocean. That is, they develop the idea of hierarchical coupling of complicated systems, where each subsystem is also modeled hierarchically. Approximate dynamics, with random parameters and noise terms are used to account for model uncertainty and unmodeled components. A hybrid importance sampler, the Gibbs sampler algorithm, is used and the model is tested in an observation, simulation system experiment.

6. Conclusion

The Bayesian framework is the ideal probabilistic framework for combining information. It follows that it provides a complete and general perspective for data assimilation. In the case of linear operators and Gaussian error distributions, well-known equations for optimal interpolation follow from the Bayesian development. Furthermore, in the sequential estimation setting, the Kalman filter and smoother can be derived from a Bayesian perspective. More importantly, the Bayesian paradigm provides the more general probabilistic DA linkage for non-Gaussian or nonlinear processes, in both the retrospective and sequential settings. However, it is typically not possible to derive analytically the posterior distributions in these cases. Thus, one must consider various (local linear) approximations and/or Monte Carlo sampling. For example, as shown here, the particle filter (sequential importance sampling) gives a useful, approximate posterior distribution through Monte Carlo sampling for nonlinear operators and non-Gaussian distributions. However, due to the curse of dimensionality, it cannot be applied to problems with high dimensional data and state spaces. The ensemble Kalman filter is arguably a viable and practical alternative to the particle filter.

In cases where there are complicated data sources, and/or uncertainty in the process and parameter models,

the Bayesian hierarchical modeling approach is a plausible extension to the usual state–space modeling approach used in sequential updating. In addition, the hierarchical approach suggests that relatively simple physical models with random parameters and/or correlated error processes can model real-world processes with complicated spatio-temporal structure. This approach can also be applied to multiple processes, and suggests a probabilistic approach for coupling models. Although such models cannot be implemented in extremely high-dimensions, such limitations will be less of a factor as computing technology improves and necessary algorithms are developed.

Acknowledgments

References

[1] J.L. Anderson, A local least squares framework for ensemble filtering, Mon. Weather Rev. 129 (2003) 2884–2903.

[2] B.D.O. Anderson, J.B. Moore, Optimal Filtering, Prentice Hall, Englewood Cliffs, New Jersey, 1979.

[3] A.F. Bennett, Inverse Modeling of the Ocean and Atmosphere, Cambridge University Press, Cambridge, 2002.

[4] J.O. Berger, Statistical Decision Theory and Bayesian Analysis, Springer-Verlag, New York, 1985.

[5] L.M. Berliner, Hierarchical Bayesian time series models, in: K. Hanson, R. Silver (Eds.), Maximum Entropy and Bayesian Methods, Kluwer Academic Publishers, 1996, pp. 15–22.

[6] L.M. Berliner, Physical–statistical modeling in geophysics, J. Geophys. Res. 108 (D24) (2003) 8776. doi:10.1029/2002JD002865.

[7] L.M. Berliner, C.K. Wikle, N. Cressie, Long-lead prediction of Pacific SST's via Bayesian dynamic modeling, J. Climate 13 (2000) 3953–3968.

[8] L.M. Berliner, R.F. Milliff, C.K. Wikle, Bayesian hierarchical modeling of air–sea interaction, J. Geophys. Res. 108 (C4) (2003) 3104. doi:10.1029/2002JC001413.

[9] J.M. Bernardo, A.F.M. Smith, Bayesian Theory, John Wiley & Sons, Inc., New York, 1994.

[10] P. Courtier, Variational methods, J. Meteorol. Soc. Jpn 75 (1997) 211–218.

[11] N.A.C. Cressie, Statistics for Spatial Data, Revised Edition, Wiley & Sons, 1993, 900 pp.

[12] R. Daley, Atmospheric Data Analysis, Cambridge University Press, London, 1991.

[13] A. Doucet, N. de Freitas, N. Gordon (Eds.), Sequential Monte Carlo Methods in Practice, Springer-Verlag, New York, 2001.

[14] E.S. Epstein, A Bayesian approach to decision making in applied meteorology, J. Appl. Meteorol. 1 (1962) 169–177.

[15] E.S. Epstein, Statistical Inference and Prediction in Climatology: A Bayesian Approach, American Meteorological Society, Boston, MA, 1985.

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

[17] G. Evensen, P.J. van Leeuwen, Assimilation of gestat altimeter data for the Agulhas current using the ensemble Kalman filter, Mon. Weather Rev. 124 (1996) 85–96.

[18] L.S. Gandin, Objective analysis of meteorological fields, in: Gidrometeorologicheskoe Izdatel'stvo (GIMIZ) (Israel Program for Scientific Translations, Jerusalem, Trans.) Leningrad 1963 (Original work published 1965).

[19] A.E. Gelfand, L. Zhu, B.P. Carlin, On the change of support problem for spatio-temporal data, Biostatistics 2 (2001) 31–45.

[20] A. Gelman, J.B. Carlin, H.S. Stern, D.B. Rubin, Bayesian Data Analysis, second edn, Chapman and Hall, CRC, Boca Raton, 2004.

[21] M. Ghil, P. Malanotte-Rizzoli, Data assimilation in meteorology and oceanography, in: Advances in Geophysics, vol. 33, Academic Press, 1991, pp. 141–266.

[22] M. Ghil, S.E. Cohn, J. Tavantzis, K. Bube, E. Isaacson, Applications of estimation theory to numerical weather prediction, in: L. Bengtsson, M. Ghil, E. Källén (Eds.), Dynamic Meteorology: Data Assimilation Methods, Springer-Verlag, 1981, pp. 139–224.

[23] N.J. Gordon, D.J. Salmond, A.F.M. Smith, Novel approaches to nonlinear/non-Gaussian Bayesian state estimation, IEE Proc. F 140 (1993) 107–113.

[24] M.S. Grewal, A.P. Andrews, Kalman Filtering Theory and Practice, Prentice Hall, Englewood Cliffs, New Jersey, 1993, p. 381.

[25] N.K. Gupta, R.K. Mehra, Computational aspects of maximum likelihood estimation and reduction in sensitivity function calculations, IEEE Trans. Automat. Contr. AC-19 (1974) 774–783.

[26] P.L. Houtekamer, H.L. Mitchell, Data assimilation using an ensemble Kalman filter technique, Mon. Weather Rev. 126 (1998) 796–811.

[27] A.H. Jazwinski, Stochastic Processes and Filtering Theory, Academic Press, San Diego, 1970.

[28] R.E. Kalman, A new approach to linear filtering and prediction problems, Trans. ASME J. Basic Eng. 82 (1960) 35–45.

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

[30] A.C. Lorenc, Analysis methods for numerical weather prediction, Q. J. Roy. Meteor. Soc. 112 (1986) 1177–1194.

[31] G. Matheron, Principles of geostatistics, Econ. Geol. 58 (1963) 1246–1266.

[32] J. Meinhold, N.D. Singpurwalla, Understanding the Kalman filter, Amer. Statistician 37 (1983) 123–127.

[33] C.P. Robert, G. Casella, Monte Carlo Statistical Methods, Springer-Verlag, New York, 2004.

[34] R.H. Shumway, D.S. Stoffer, An approach to time series smoothing and forecasting using the EM algorithm, J. Time Ser. Anal. 3 (1982) 253–264.

[35] R.H. Shumway, D.S. Stoffer, Time Series Analysis and Its Applications, Springer-Verlag, New York, 2000, 549 pp.

[36] B.W. Silverman, Density Estimation for Statistics and Data Analysis, Chapman & Hall, London, 1986.

[37] O. Talagrand, Assimilation of observations, an introduction, J. Meteorol. Soc. Jpn 75 (1997) 191–209.

[38] A. Tarantola, Inverse Problem Theory: Methods for Data Fitting and Model Parameter Estimation, Elsevier, New York, 1987.

[39] H.J. Thiebaux, M.A. Pedder, Spatial Objective Analysis with Applications in Atmospheric Science, Academic Press, 1987.

[40] M.K. Tippett, J.L. Anderson, C.H. Bishop, T.M. Hamill, J.S. Whitaker, Ensemble square-root filters, Mon. Weather Rev. 131 (2003) 649–667.

[41] M. West, Approximating posterior distributions by mixtures, J. R. Stat. Soc., Ser. B 55 (1993) 409–422.

[42] M. West, J. Harrison, Bayesian Forecasting and Dynamic Models, second edn, Springer-Verlag, New York, 1997.

[43] C.K. Wikle, L.M. Berliner, N. Cressie, Hierarchical Bayesian space-time models, J. Environ. Ecol. Stat. 5 (1998) 117–154.

[44] C.K. Wikle, Hierarchical models in environmental science, Int. Stat. Rev. 71 (2003) 181–199.

[45] C.K. Wikle, Hierarchical Bayesian models for predicting the spread of ecological processes, Ecology 84 (2003) 1382–1394.

[46] C.K. Wikle, R.F. Milliff, D. Nychka, L.M. Berliner, Spatiotemporal hierarchical Bayesian modeling: Tropical ocean surface winds, J. Amer. Stat. Assoc. 96 (2001) 382–397.

[47] C.K. Wikle, L.M. Berliner, R.F. Milliff, Hierarchical Bayesian approach to boundary value problems with stochastic boundary conditions, Mon. Weather Rev. 131 (2003) 1051–1062.

[48] C.K. Wikle, L.M. Berliner, Combining information across spatial scales, Technometrics 47 (2005) 80–91.

[49] L.M. Berliner, J.A. Royle, C.K. Wikle, R.F. Milliff, Bayesian methods in the atmospheric sciences, Bayesian Statistics 6 (1998) 83–100.