Journal of Computational Physics 231 (2012) 6401–6418

Contents lists available at SciVerse ScienceDirect

Journal of Computational Physics

journal homepage: www.elsevier.com/locate/jcp

Sequential data assimilation with multiple models

Akil Narayan [a], Youssef Marzouk [b], Dongbin Xiu [a,*]

[a] Department of Mathematics, Purdue University, West Lafayette, IN 47907, USA
[b] Department of Aeronautics and Astronautics, Massachusetts Institute of Technology, Cambridge, MA 02139, USA

ARTICLE INFO

Article history:
Received 18 January 2012
Received in revised form 6 May 2012
Accepted 3 June 2012
Available online 21 June 2012

Keywords:
Uncertainty quantification
Data assimilation
Kalman filter
Model averaging

ABSTRACT

Data assimilation is an essential tool for predicting the behavior of real physical systems given approximate simulation models and limited observations. For many complex systems, there may exist several models, each with different properties and predictive capabilities. It is desirable to incorporate multiple models into the assimilation procedure in order to obtain a more accurate prediction of the physics than any model alone can provide. In this paper, we propose a framework for conducting sequential data assimilation with multiple models and sources of data. The assimilated solution is a linear combination of all model predictions and data. One notable feature is that the combination takes the most general form with matrix weights. By doing so the method can readily utilize different weights in different sections of the solution state vectors, allow the models and data to have different dimensions, and deal with the case of a singular state covariance. We prove that the proposed assimilation method, termed direct assimilation, minimizes a variational functional, a generalized version of the one used in the classical Kalman filter. We also propose an efficient iterative assimilation method that assimilates two models at a time until all models and data are assimilated. The mathematical equivalence of the iterative method and the direct method is established. Numerical examples are presented to demonstrate the effectiveness of the new method.

1. Introduction

Numerical simulations of mathematical models are essential tools for predicting the behavior of physical systems. Myriad numerical techniques and approximations are used to simulate physical phenomena in fluid dynamics, electromagnetics, chemical systems, astrophysics, and more. Since all of these simulations involve approximations, uncertainty and error are inevitably present in their predictions. To complicate matters, a single physical process may be described by *multiple* mathematical models and numerical approximations. In addition, one may have access to empirical observations of the system—noisy and limited in number, scope, and resolution. A natural question to ask is how to combine the models and the observational data to predict the physical state with greater fidelity than can be obtained with any of the models individually?

Various techniques for model averaging and data assimilation, all of which attempt to implement such a combination of models and data, have received attention in recent years. In the case of a single dynamical model with a stream of noisy observations, the Kalman filter [15,14] is both simple and remarkably effective. The assimilation step of the Kalman filter updates the state by weighing the model prediction and the data in order to minimize a quadratic objective. This operation can be interpreted in many different ways, for instance, as a minimum variance estimator or as a Bayesian update. The

* Corresponding author. Tel.: +1 765 496 2846.
E-mail addresses: akil@purdue.edu (A. Narayan), ymarz@mit.edu (Y. Marzouk), dxiu@purdue.edu (D. Xiu).

0021-9991/$ - see front matter © 2012 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.jcp.2012.06.002

original Kalman filter was designed for linear systems, but derivate methods for filtering nonlinear systems are plentiful, e.g. the extended Kalman filter [9,11], the ensemble Kalman filter [7,8] and its variants [1,2,5,21,24], and is the subject of extensive ongoing work.

The use of multiple models for filtering, on the other hand, has seen less development. With static data sets, Bayesian model averaging (BMA) is a well-established technique for statistical prediction in the presence of model uncertainty [10]. BMA writes the predictive distribution of any quantity of interest as a weighted average of the posterior predictions due to each model; the data-dependent scalar weights are the posterior model probabilities [19]. Alternatively, one might consider non-Bayesian approaches to model averaging with static data; these are generally focused on providing point predictions rather than predictive distributions [10,6]. Dynamic model averaging (DMA) [20] generalizes BMA to the dynamical setting, where data arrive sequentially and one would like to make online predictions about the state of a system. DMA updates the state conditional on a discrete model indicator, but some Markovian dynamics (e.g. a matrix of transition probabilities or a forgetting factor) for this model indicator must be specified. As in BMA, the predictive distribution is a mixture with one component for each model. Other methods for sequential estimation with multiple models include the Interacting Multiple Model filter [4] and the generalized pseudo-Bayes framework [23]. Both of these techniques introduce dynamics for the "switching" behavior between models and update state probabilities based on innovations from a Kalman filter.

Like all of the aforementioned methods, we propose an assimilation technique that relies on a weighted arithmetic mean of the models. A limitation of the previous methods, however, is that the weights used in model averaging are scalar-valued, even when the state is vector-valued. Therefore, the assimilation process cannot employ different model weights in different sections of the state vector (corresponding to regions of space where one model might be superior to another, for instance). In this paper we consider more a general assimilation technique: if $\mathbf{u}_1, \ldots, \mathbf{u}_M$ are state vectors for M different models and $\mathbf{d}$ is the data vector, we attempt to find an assimilated state $\mathbf{w}$ of the form

$$\mathbf{w} = \sum_{m=1}^{M} \mathbf{A}_m \mathbf{u}_m + \mathbf{B}\mathbf{d}, \tag{1.1}$$

where $\mathbf{A}_m$ and $\mathbf{B}$ are matrices. Therefore, we allow our assimilation to be *the most general linear functional* of all the models and data. This form also allows the model states and data to have different dimensions. Also it is possible to have different, independent sources of data $\mathbf{d}_1, \ldots, \mathbf{d}_N$, measuring different quantities-of-interest. Mathematically they can be concatenated into a single vector $\mathbf{d}$, allowing us to use (1.1) as a general form. (We will show that assimilating data is philosophically the same as simply taking the data to be extra model states in (1.1). In other words, our method treats model predictions and data as identical mathematical objects.)

We derive our assimilation technique by minimizing a variational functional similar to, and a generalization of, the one used in the classical Kalman filter. With only a single model and data, the assimilation is identical to the Kalman filter. When two or more models are present, the direct assimilation in the form (1.1) can be employed, where the explicit forms of the matrices $\mathbf{A}_m$ and $\mathbf{B}$ are derived. For practical systems with large dimensions, we also propose a more efficient iterative assimilation method, which seeks to perform a series of two-model assimilations until all models and data are assimilated. We then establish the mathematical conditions under which the iterative assimilation is equivalent to the direct assimilation, and more importantly, invariant to the permutation of the model states. Our proposed iterative assimilation approach can also handle singular model and data covariance matrices. Like all Kalman filtering methods, the present methodology requires the prescription and propagation of the model covariances. However, the method is rather a framework and not tied to any specific algorithm for covariance propagation. Therefore, the method can be readily combined with most variants of the Kalman filter, such as the ensemble Kalman filter.

It is worth pointing out that the explicit inclusion of the data in our assimilation method (1.1) represents a subtle, and yet important, difference from the BMA. In BMA, the role of data presents itself implicitly in the form of the averaging weights, which is determined by the posterior probability. As a result, the BMA is an "average" of all models. When all models are consistently biased towards one side of the prediction, the BMA result is, by construction, guaranteed to be no better than the best available model. The form of (1.1) effectively prevents this from happening.

In Section 2 we formalize notation and present the assimilation problem. This section also reviews existing assimilation methods that are relevant to our discussion. Section 3 introduces our algorithm and presents its mathematical justification. In doing so, we compare a simultaneous assimilation procedure with a sequential assimilation procedure, and make the argument that the sequential procedure is more robust. A Bayesian interpretation of both procedures is also provided. Section 4 provides examples of our assimilation method in practice, and Section 5 follows with closing remarks.

2. Problem setup

In this section we present the overall problem of data assimilation with multiple models, formalizing the assumptions and notation to be used in subsequent analysis. We also review the Kalman filter (KF), which is widely used in data assimilation with a single model and closely related to our proposed method for multiple model assimilation.

2.1. Data assimilation with multiple models

Following usual practice in data assimilation, we present our models in the form of dynamical systems, emphasizing the temporal nature of the problem. Let $\mathbf{u}^t \in \mathbb{R}^{N_t}$, $N_t \geqslant 1$, denote the "true state" of a physical phenomenon. Typically $\mathbf{u}^t$ is a finite representation of the spatially distributed state of the physical system, after some spatial discretization has been applied. More generally, $\mathbf{u}^t$ is a quantity of interest whose evolution we wish to track. This evolution is governed by certain physical laws and processes that are not completely available to us. Nonetheless, we suppose that there exists some operator that captures the exact evolution of the system under consideration. Considering only discrete values of time t_k, we can write

$$\mathbf{u}^t(t_{k+1}) = \mathcal{L}(t_k, \mathbf{u}^t(t_k)), \tag{2.1}$$

where the operator $\mathcal{L}$ is unknown to us, but represents the evolution of the truth state $\mathbf{u}^t$ from one point in time to the next.

Instead of $\mathcal{L}$, we have a set of models indexed by $m = 1 \ldots M$, that approximate the evolution of $\mathbf{u}^t$ in different ways. In particular, these models specify the temporal evolution of distinct state vectors $\mathbf{u}_m \in \mathbb{R}^{N_m}$ via an operator g_m:

$$\begin{aligned} \frac{d\mathbf{u}_m}{dt} &= g_m(t, \mathbf{u}_m), \quad t \in (0, T], \\ \mathbf{u}_m(0) &= \mathbf{u}_{m,0}, \end{aligned} \tag{2.2}$$

with $T > 0$. Solutions of the differential equations above are the "forecast" states $\mathbf{u}_m(t)$. We define the discrete solution operator for each system, given by

$$\mathbf{u}_m(t_{k+1}) = G_m(t_k, \mathbf{u}_m(t_k)), \tag{2.3}$$

where the operator G_m is the discretized version of g_m, and may be nonlinear in $\mathbf{u}_m$. Note a strict first-order Markovian setting has been used throughout this section. This is done for mere notational convenience and does not affect our discussion below.

The forecast states are useful because they carry information about the true state. We will assume that, at any given time t_k, the forecast state variables $\mathbf{u}_m$ are linearly related to the true state $\mathbf{u}^t$, with the addition of a stochastic discrepancy. In other words, we have

$$\mathbf{u}_m = \mathbf{H}_m \mathbf{u}^t + \boldsymbol{\epsilon}_m, \quad m = 1, \ldots, M, \tag{2.4}$$

where $\mathbf{H}_m \in \mathbb{R}^{N_m \times N_t}$, and $\boldsymbol{\epsilon}_m$, satisfying $\mathbb{E}[\boldsymbol{\epsilon}_m] = \mathbf{0}$, are random variables that capture the discrepancy between the transformed true state and $\mathbf{u}_m$. Note that the $\boldsymbol{\epsilon}_m$ can be time-dependent.

In practice the relation between the forecast states and the true states could be nonlinear:

$$\mathbf{u}_m = \mathcal{H}_m(\mathbf{u}^t) + \boldsymbol{\epsilon}_m, \quad m = 1, \ldots, M, \tag{2.5}$$

for some nonlinear operators $\mathcal{H}_m$. Our technique uses a filtering procedure to assimilate different models, and in cases when the measurement operators $\mathcal{H}_m$ are not linear, then nonlinear filtering techniques such as the Unscented Kalman Filter [13,12] or particle filters [18,22] are appropriate. Here we restrict ourselves to the linear case, with the understanding that one could replace our use of the Kalman filter with any appropriate nonlinear assimilation technique. One could also consider (2.4) to be an approximation or linearization of the nonlinear operator $\mathcal{H}_m$, as often done in practice.

We use $\mathbf{d} \in \mathbb{R}^{N_d}$, $N_d \geqslant 1$, to denote a set of measurements

$$\mathbf{d} = \mathbf{H}\mathbf{u}^t + \boldsymbol{\epsilon}, \tag{2.6}$$

where $\mathbf{H} \in \mathbb{R}^{N_d \times N_t}$ is the measurement matrix and $\boldsymbol{\epsilon} \in \mathbb{R}^{N_d}$ is the measurement error satisfying $\mathbb{E}[\boldsymbol{\epsilon}] = \mathbf{0}$. Note that the relationship between the measurement and the true solution could in general be nonlinear, but we focus here on the linear case. (Again the measurement matrix $\mathbf{H}$ could be considered as a linearization of a nonlinear measurement operator.) Our use of the discrepancy terms $\boldsymbol{\epsilon}$ and $\boldsymbol{\epsilon}_m$ encompasses many sources of uncertainty including, but not limited to, model discrepancy, temporal/spatial numerical discretization error, numerical roundoff error, and measurement error. Also note that (2.6) easily generalizes to the case of multiple measurements that are conditionally independent given $\mathbf{u}^t$. For our purposes it is notationally convenient to collect all these observations into one vector $\mathbf{d}$, but the algorithm we present below can immediately be applied to assimilation problems where different measurements are treated as separate vectors.

Our goal is to construct an "analyzed solution," denoted by $\mathbf{w} \in \mathbb{R}^{N_t}$ and of the form (1.1), using the forecast solutions $\{\mathbf{u}_m\}_{m=1}^M$ and the measurement $\mathbf{d}$ to provide a more accurate prediction of the true solution $\mathbf{u}^t$.

2.2. Kalman filter

The Kalman filter (KF) is widely used for data assimilation with a single model; here we only list relevant properties, and more in-depth discussion can be found in [8]. Following our setup in Section 2.1, we consider the case of $M = 1$ and $\mathbf{H}_1 = \mathbf{I}$, where $\mathbf{I}$ is the identity matrix. That is, there exists only one forecast state $\mathbf{u}_1$ and it is a direct prediction of the true state $\mathbf{u}^t$. Let $\mathbf{U}_1 \in \mathbb{R}^{N_1 \times N_1}$ be the covariance matrix of the forecast solution $\mathbf{u}_1$ and suppress the subscript 1 hereafter only in this section. The analyzed solution **w** obtained by the standard KF is a linear combination of the forecast solution **u** and the measurement **d** in the following manner,

$$\mathbf{w} = \mathbf{u} + \mathbf{K}(\mathbf{d} - \mathbf{H}\mathbf{u}), \tag{2.7}$$

where **K** is the so-called *Kalman gain matrix* defined as

$$\mathbf{K} = \mathbf{U}\mathbf{H}^T(\mathbf{H}\mathbf{U}\mathbf{H}^T + \mathbf{D})^{-1}. \tag{2.8}$$

Here the superscript T denotes the matrix transpose, and $\mathbf{D} \in \mathbb{R}^{N_d \times N_d}$ is the covariance of the measurement error $\boldsymbol{\epsilon}$. The covariance matrix of the analyzed state **w**, $\mathbf{W} \in \mathbb{R}^{N_t \times N_t}$, is then obtained by the update

$$\mathbf{W} = (\mathbf{I} - \mathbf{K}\mathbf{H})\mathbf{U}(\mathbf{I} - \mathbf{K}\mathbf{H})^T + \mathbf{K}\mathbf{D}\mathbf{K}^T = (\mathbf{I} - \mathbf{K}\mathbf{H})\mathbf{U}. \tag{2.9}$$

This assimilation process is repeated at every instance of time when data is available. The analyzed state **w** is the solution that minimizes the following variational functional:

$$\mathcal{J}[\mathbf{w}] = (\mathbf{w} - \mathbf{u})^T\mathbf{U}^{-1}(\mathbf{w} - \mathbf{u}) + (\mathbf{H}\mathbf{w} - \mathbf{d})^T\mathbf{D}^{-1}(\mathbf{H}\mathbf{w} - \mathbf{d}). \tag{2.10}$$

There are many ways to rationalize the desire to minimize the objective (2.10). The functional $\mathcal{J}[\mathbf{w}]$ is the sum of Mahalanobis distances between the known states $\mathbf{V}u$ and $\mathbf{V}d$ and we want to find a state $\mathbf{V}w$ lying as close as possible to both of them. Alternatively, we may pose the problem in the Bayesian framework: we suppose that **u** and its covariance specify a prior Gaussian distribution on state space; assuming likewise that **d** is obtained as a Gaussian perturbation from linear measurements the truth, the likelihood of observing the data can be computed. We then take the analyzed state **w** to be the mode of the posterior; equivalently, **w** minimizes the negative log-likelihood of the posterior. In this setup, the negative log-likelihood is given by $\mathcal{J}$, cf. (3.14) and Section 3.7.

3. Assimilation of multiple models

We present our algorithm for multiple model assimilation in this section. Section 3.1 begins by standardizing our notation. Our algorithm is implicitly related to the problem of computing harmonic means of covariance matrices, and so Section 3.2 introduces the concept of harmonic means on positive semi-definite matrices. An unavoidable complication that may arise in any assimilation procedure is the existence of model and/or measurement states that have competing values that cannot be clearly resolved. We acknowledge this reality in Section 3.3 and present sufficient conditions under which differing values can be unambiguously assimilated. Section 3.4 presents a version of our algorithm that simultaneously assimilates all models and measurements, and Section 3.5 follows up with our proposed sequential algorithm. We then provide a simple example to demonstrate why sequential assimilation is preferred. Section 3.7 concludes by providing a Bayesian interpretation of the multi-model assimilation scheme.

3.1. Preliminaries

We work on a complete probability space $(\Omega, \mathcal{F}, \mu)$ with Ω the collection of events, $\mathcal{F}$ a σ-algebra on sets of Ω, and μ a probability measure on $\mathcal{F}$. All random variables considered here are in the L^2 stochastic space: $u \in L^2_\mu \Rightarrow \int_\Omega \|u(\omega)\|^2 \mathrm{d}\mu(\omega) = \|u(\omega)\|^2_\mu \equiv \mathbb{E}\|u\|^2 < \infty$, where $\|u\|$ is the standard Euclidean norm when u is vector-valued; this is sufficient to ensure the existence of the mean and variance of u. When talking about limits of random variables, equality is in the L^2_μ sense: $\lim_{\varepsilon \to 0} u_\varepsilon = u \Rightarrow \|u_\varepsilon - u\|_\mu \to 0$.

Throughout this paper we will use lowercase boldface letters (e.g. **u**) to denote vectors and uppercase boldface letters (e.g. **A**) to denote matrices. $\mathbf{A}^T$ and $\mathbf{A}^\dagger$ denote the matrix transpose, and the Moore–Penrose pseudoinverse of **A**, respectively. For random vectors we will use the same letter but with uppercase to denote their corresponding covariance matrices. For example, let $\mathbf{v} \in \mathbb{R}^N$ be a random vector with zero mean, then $\mathbf{V} = \mathbb{E}[\mathbf{v}\mathbf{v}^T] \in \mathbb{R}^{N \times N}$ denotes its covariance matrix. A square matrix **A** is positive definite if $\mathbf{v}^T\mathbf{A}\mathbf{v} > 0$ for all non-trivial **v**, and is positive semi-definite (or 'semi-positive') if $\mathbf{v}^T\mathbf{A}\mathbf{v} \geqslant 0$. For any fixed size N, the space of all $N \times N$ positive definite matrices is denoted by $\mathcal{H}$, and the space of all positive semi-definite matrices by $\mathcal{H}_0$.

3.2. Matrix harmonic means

The standard way to define a harmonic mean for positive definite matrices $\mathbf{A}_1, \mathbf{A}_2 \in \mathcal{H}$ is

$$\mathbf{F}_2 = 2\left(\mathbf{A}_1^{-1} + \mathbf{A}_2^{-1}\right)^{-1}$$

and it is not difficult to imagine generalizing this to a sequence of matrices $\mathbf{A}_m \in \mathcal{H},\ m = 1, \ldots, M$:

$$\mathbf{F}_M = M\left(\sum_{m=1}^{M}\mathbf{A}_m^{-1}\right)^{-1}. \tag{3.1}$$

It is clear that this cannot be used for (non-invertible) semi-positive matrices. However, one can proceed by taking a different route.

Definition 1 (*Matrix harmonic mean*). Let $\mathbf{A}_1, \ldots, \mathbf{A}_M \in \mathcal{H}_0$, and define $\mathbf{F}_1 = \mathbf{W}_1 = \mathbf{A}_1$. For $m = 2, \ldots, M$, define:

$$\mathbf{W}_m = \mathbf{W}_{m-1}(\mathbf{W}_{m-1} + \mathbf{A}_m)^{\dagger}\mathbf{A}_m, \tag{3.2a}$$

$$\mathbf{F}_m = m\mathbf{W}_m. \tag{3.2b}$$

$\mathbf{F}_M$ is called the harmonic mean of the matrices $\mathbf{A}_1, \ldots, \mathbf{A}_M$.

This definition coincides with the traditional definition of harmonic means of $M \geqslant 2$ positive-definite matrices. More importantly, this definition is a generalization to semi-positive matrices.

Theorem 1. *For a sequence of matrices $\mathbf{A}_m, m = 1, \ldots, M$, on $\mathcal{H}$, the harmonic mean of Definition 1 coincides with the formula (3.1). For $\mathbf{A}_m$ on $\mathcal{H}_0, \mathbf{F}_M$ is*

1. *closed on $\mathcal{H}_0 : \mathbf{F}_M \in \mathcal{H}_0$;*
2. *continuous: if $\mathbf{F}_M^{\varepsilon}$ is the harmonic mean of $\mathbf{A}_1^{\varepsilon}, \ldots, \mathbf{A}_M^{\varepsilon} \in \mathcal{H}_0$ where $\mathbf{A}_m^{\varepsilon} \to \mathbf{A}_m$ then $\mathbf{F}_M^{\varepsilon} \to \mathbf{F}_M$;*
3. *consistent: $\mathbf{A}_1 = \mathbf{A}_2 = \cdots = \mathbf{A}_M \Rightarrow \mathbf{F}_M = \mathbf{A}_1$;*
4. *symmetric: For $m_1, m_2, \ldots, m_M$ any permutation of $1, 2, \ldots, M$, the harmonic mean of $\mathbf{A}_{m_1}, \ldots, \mathbf{A}_{m_M}$ is the same as that of $\mathbf{A}_1, \mathbf{A}_2, \ldots, \mathbf{A}_M$;*
5. *decreasing: $\mathbf{F}_M \leqslant \mathbf{A}_m$ for all m;*
6. *monotone: If $\mathbf{A}_m \leqslant \mathbf{B}_m$ for all m, then $\mathbf{F}_M \leqslant \mathbf{J}_M$, where $\mathbf{F}_M$ is the harmonic mean of the $\mathbf{A}_m$ and $\mathbf{J}_m$ is the harmonic mean of the $\mathbf{B}_m$.*

The above properties justify calling $\mathbf{F}_M$ a matrix mean. The Appendix contains the proof, along with mathematical discussions that are not directly related to our present task of multiple model data assimilation.

We note here that using properties of pseudoinverses one may rewrite the iterative procedure (3.2a(a)) for updating $\mathbf{W}_m$:

$$\mathbf{W}_m = \left(\mathbf{I} - \mathbf{W}_{m-1}(\mathbf{W}_{m-1} + \mathbf{A}_m)^{\dagger}\right)\mathbf{W}_{m-1} \triangleq (\mathbf{I} - \mathbf{K})\mathbf{W}_{m-1}$$

Comparing this with the Kalman filter covariance update from (2.8) and (2.9) when $\mathbf{H} = \mathbf{I}$, we see that the iterative procedure for computing harmonic means is almost identical to a standard Kalman filter procedure.

Our algorithm for model assimilation produces an assimilated model state whose covariance is proportional to the harmonic mean of the input covariances. (If the input covariances are $\mathbf{A}_m$, then the matrix $\mathbf{W}_M$ from (3.2b) is the assimilated covariance.) We will see later that the algorithm itself can be implemented by iterating a standard Kalman filter, and thus implicitly uses the iterative definition of the matrix harmonic mean (3.2b) to compute the assimilated state.

3.3. Consistent random variables

Any assimilation procedure must preclude the situation where there is no logical way to reconcile a quantitative difference between two models. For example, if we have two models of a scalar quantity of interest with $u_1(\omega) = 1$ and $u_2(\omega) = 2$ almost surely, then no assimilation procedure can produce a good synthesis. In our setting, we exclude such situations by insisting that the random variables corresponding to our model state vectors are *consistent*.

Definition 2. Consider $L^2(\Omega)$ random vectors $\mathbf{u}_m \in \mathbb{R}^{N_m}$, $m = 1, \ldots, M$, each paired with a measurement matrix $\mathbf{H}_m \in \mathbb{R}^{N_m \times N_t}$, where N_t is the size of the truth state.[1] Then $\mathbf{u}_m$ are consistent if the components in each random variable corresponding to zero variance can be assimilated, almost surely without contradiction, from the other variables. More precisely, let S_m^0 be a matrix of orthogonal column vectors corresponding to the nullspace of $\mathbf{U}_m$. Then $\mathbf{u}_m$ are consistent if the linear system defined by the M vector-valued equations

$$\left(S_m^0\right)^T \mathbf{H}_m \mathbf{w} = \left(S_m^0\right)^T \mathbb{E}[\mathbf{u}_m], \quad m = 1, 2, \ldots, M \tag{3.3}$$

has at least one solution for the vector $\mathbf{w}$.

The condition (3.3) essentially states that the random variables $\mathbf{u}_m$ do not have competing zero-variance components. If all the random variables $\mathbf{u}_m$ have strictly positive covariance kernels, then they are automatically consistent. In all that follows, we will only consider collections of random variables that are consistent.

[1] Recall from (2.4) that $\mathbf{H}_m$ is only used to connect model states to the truth solution. The source of randomness, whether stemming from ϵ_m or from randomness in the truth, is here irrelevant.

3.4. Main result: direct assimilation method

We now prescribe a method for assimilating multiple models along with data. The single-model Kalman filter procedure can be identified as minimizing the variational functional (2.10). This suggests that an assimilation procedure for assimilating multiple models with data can be proposed simply by making appropriate changes to the variational functional.

Theorem 2. *Consider a set of M forecast states $\mathbf{u}_m, m = 1, \dots, M$, satisfying (2.4) and with error covariance $\mathbf{U}_m = \mathbb{E}[\boldsymbol{\epsilon}_m \boldsymbol{\epsilon}_m^T] \in \mathcal{H}$, and data vector (2.6) with error covariance $\mathbf{D} = \mathbb{E}[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T] \in \mathcal{H}$. Assume that the row spaces of $\mathbf{H}$ and $\mathbf{H}_m$ span all $\mathbb{R}^{N_t}$:*

$$\operatorname{span}\{\operatorname{ran}\mathbf{H}, \operatorname{ran}\mathbf{H}_1, \operatorname{ran}\mathbf{H}_2, \dots, \operatorname{ran}\mathbf{H}_M\}. \tag{3.4}$$

Define

$$\mathcal{J}[\mathbf{w}] = \sum_{m=1}^{M} (\mathbf{H}_m\mathbf{w} - \mathbf{u}_m)^T \mathbf{U}_m^{-1} (\mathbf{H}_m\mathbf{w} - \mathbf{u}_m) + (\mathbf{H}\mathbf{w} - \mathbf{d})^T \mathbf{D}^{-1} (\mathbf{H}\mathbf{w} - \mathbf{d}), \tag{3.5}$$

then the minimizer satisfies

$$\mathbf{w} = \mathbf{W}_M \left(\sum_{m=1}^{M} \mathbf{H}_m^T \mathbf{U}_m^{-1} \mathbf{u}_m + \mathbf{H}^T \mathbf{D}^{-1} \mathbf{d} \right), \tag{3.6}$$

where $\mathbf{W}_M$ is given by:

$$\mathbf{W}_M = \left(\sum_{m=1}^{M} \mathbf{H}_m^T \mathbf{U}_m^{-1} \mathbf{H}_m + \mathbf{H}^T \mathbf{D}^{-1} \mathbf{H} \right)^{-1}.$$

Proof. The proof can be easily obtained by straightforward calculus, where the condition (3.4) is required to assure strict positivity of $\mathcal{J}[\mathbf{w}]$. □

The analyzed state is $\mathbf{w}$ and the analyzed model states are

$$\mathbf{v}_m = \mathbf{H}_m \mathbf{w}, \quad m = 1, \dots, M. \tag{3.7}$$

The matrices $\mathbf{A}_m$ in (1.1) are thus given by

$$\mathbf{A}_m = \mathbf{W}_m \mathbf{H}_m^T \mathbf{U}_m^{-1}. \tag{3.8}$$

The technical assumption (3.4) is made[2] to ensure a unique minimizer for the quadratic form $\mathcal{J}$. While (3.6) will produce the assimilated state we propose, the assumptions $\mathbf{U}_m, \mathbf{D} \in \mathcal{H}$, i.e., strictly positive definite covariances, may be too restrictive in practice. (Note that under such an assumption all random variables in Theorem 2 are automatically consistent.) This restriction can be relaxed by using a more robust iterative assimilation method.

3.5. Main result: iterative assimilation method

We now present the iterative assimilation method that behaves in a more robust manner, in the sense that it can readily deal with semi-positive covariance matrices.

Theorem 3. *Consider a set of M forecast variables $\mathbf{u}_m, m = 1, \dots, M$, satisfying (2.4) and with error covariance $\mathbf{U}_m = \mathbb{E}[\boldsymbol{\epsilon}_m \boldsymbol{\epsilon}_m^T] \in \mathcal{H}_0$, and data vector (2.6) with error covariance $D = \mathbb{E}[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T] \in \mathcal{H}_0$. Assume[3] that $\mathbf{H}_1 = \mathbf{I}$; set $\mathbf{w}_1 = \mathbf{u}_1$ and $\mathbf{W}_1 = \mathbf{U}_1$. For $m = 2, 3, \dots, M$, let*

$$\begin{aligned}
\mathbf{K}_m &= \mathbf{W}_{m-1}\mathbf{H}_m^T \left(\mathbf{H}_m \mathbf{W}_{m-1} \mathbf{H}_m^T + \mathbf{U}_m \right)^{\dagger} \\
\mathbf{w}_m &= \mathbf{w}_{m-1} + \mathbf{K}_m(\mathbf{u}_m - \mathbf{H}_m\mathbf{w}_{m-1}) = \mathbf{w}_{m-1} + \mathbf{W}_{m-1}\mathbf{H}_m^T \left(\mathbf{H}_m \mathbf{W}_{m-1} \mathbf{H}_m^T + \mathbf{U}_m \right)^{\dagger} (\mathbf{u}_m - \mathbf{H}_m \mathbf{w}_{m-1}), \\
\mathbf{W}_m &= (\mathbf{I} - \mathbf{K}_m\mathbf{H}_m)\mathbf{W}_{m-1} = \mathbf{W}_{m-1} - \mathbf{W}_{m-1}\mathbf{H}_m^T \left(\mathbf{H}_m \mathbf{W}_{m-1} \mathbf{H}_m^T + \mathbf{U}_m \right)^{\dagger} \mathbf{H}_m \mathbf{W}_{m-1}
\end{aligned} \tag{3.9}$$

[2] In the absence of any prior information about the truth state, we also consider this condition as required to avoid infinities: if we know nothing about the truth state $\mathbf{u}^t$, and (3.4) is violated, all the models and data essentially contain insufficient information to say anything about certain components of $\mathbf{u}^t$. Nevertheless we must assign *some* value to these components. In order to communicate our lack of knowledge, prescribing infinite variance is the only quantitatively accurate assignment.

[3] This assumption is made for simplicity and is related to the concern from footnote 2. Weakening this assumption ($\mathbf{H}_1 \neq \mathbf{I}$) is possible but requires special treatment to avoid infinite variances in the early stages of the assimilation procedure. When $\mathbf{H}_1 \neq \mathbf{I}$ one can formulate a well-defined assimilation procedure so long as (3.4) holds.

Finally, assimilate the data vector:

$$\begin{aligned}
\mathbf{K} &= \mathbf{W}_M\mathbf{H}\left(\mathbf{H}\mathbf{U}_M\mathbf{H}^T + \mathbf{D}\right)^{\dagger} \\
\mathbf{w}_{M+1} &= \mathbf{w}_M + \mathbf{K}(\mathbf{d} - \mathbf{H}\mathbf{u}_M), \\
\mathbf{W}_{M+1} &= (\mathbf{I} - \mathbf{K}\mathbf{H})\mathbf{W}_M.
\end{aligned} \tag{3.10}$$

$\mathbf{W}_m$ *is the covariance of* $\mathbf{w}_m$*. When* $\mathbf{U}_m,\ \mathbf{D} \in \mathcal{H}$*, then* $\mathbf{w}_{M+1}$ *is the same as* $\mathbf{w}$ *from (3.6). If* $\mathbf{u}_1, \ldots, \mathbf{u}_M$ *are consistent random variables, then* $\mathbf{w}_M$ *is independent of the ordering in the iterative procedure (3.9).*

We give the proof in the Appendix. It is in this proof that we require the definition of consistent model states in order to guarantee that the assimilated state does not depend on the ordering of the model states.

There are many observations to make about the iterative procedure defined in Theorem 3.

- The iterative method allows the covariance matrices to be semi-positive. When the covariance matrices are positive, the iterative method is equivalent to the direct method. Therefore the iterative method has a mathematically wider range of applicability than the direct method.
- The assumption $\mathbf{H}_1 = \mathbf{I}$ is made only to make presentation of the assimilation procedure clearer. Generally we are interested in dynamical systems, so there is a model state $\mathbf{w}$ from the previous time step that can be used
- The ordering of the assimilation procedure does not matter if the random variables $\mathbf{u}_m,\ m = 1, \ldots, M$, and $\mathbf{d}$ are consistent. In fact it is also independent of the data/model ordering. Therefore, one can treat data as another set of model simulation results. Consequently, multiple and conditionally independent sources of data can be assimilated by simply performing additional data assimilation steps, independent of the ordering.
- The matrix $\mathbf{W}_M$ can be interpreted as a harmonic mean of the matrices $\mathbf{U}_m$ when paired with the measurement matrices $\mathbf{H}_m$. If the $\mathbf{H}_m$ are all identity matrices, then $\mathbf{W}_M$ is exactly $1/M$ times the harmonic mean of the $\mathbf{U}_m$, where Definition 1 is used when any of the $\mathbf{U}_m$ is semi-positive. See also the discussion at the end of Section 3.2.
- The iterative procedure outlined above is essentially a sequential application of a standard Kalman filter update. (Use of the pseudoinverse is the main difference.) Thus, assimilation of each new model may be viewed as a single-model Kalman filter update.
- The iterative scheme allows one to assimilate any subset of models and data at times when they are available. One practical use of this is the ability to assimilate only model states in the absence of data. An example of such a situation is given in Section 4.3.

The iterative assimilation method can be implemented in a straightforward manner:

- *Initialization.* For model $\mathbf{u}_1$ and data $\mathbf{d}$, perform the standard Kalman update to obtain an analyzed model state $\mathbf{w}_1$ with covariance $\mathbf{W}_1$.
- *Iteration.* For $m = 2, \ldots, M$, apply the procedure (3.9). In other words, consider the present assimilated model state $\mathbf{w}_{m-1}$ with its covariance $\mathbf{W}_{m-1}$ to be the forecast state, and conduct a Kalman update using the new model prediction $\mathbf{u}_m$ as "data" (with measurement matrix $\mathbf{H}_m$) to obtain the new analyzed state $\mathbf{w} = \mathbf{w}_{M+1}$ and the analyzed model states $\mathbf{v}_m = \mathbf{H}_m\mathbf{w}$.

3.6. Some motivating examples

The direct assimilation approach is applicable to cases when the quadratic form $\mathcal{J}$ from (3.5) is positive definite; this is satisfied, for example, under the assumptions of Theorem 2. In other cases, for example when some components have zero variance, the iterative procedure is more appropriate. However, the iterative assimilation approach is only robust if the model states are consistent. If the model states are not consistent, then one must abandon the hope of producing an assimilation that is faithful to all the models. We now present examples that showcase these various situations.

Suppose we have two models $\mathbf{u}_1, \mathbf{u}_2 \in \mathbb{R}^2$ with identity observation matrices $\mathbf{H}_1 = \mathbf{H}_2 = \mathbf{I}$. The direct assimilation scheme is to form the variable

$$\mathbf{w} = (\mathbf{U}_1^{-1} + \mathbf{U}_2^{-1})^{-1}\left[\mathbf{U}_1^{-1}\mathbf{u}_1 + \mathbf{U}_2^{-1}\mathbf{u}_2\right]. \tag{3.11}$$

If $\mathbf{U}_1$ and $\mathbf{U}_2$ are positive definite, there is no issue. But let us explore the issue of continuity with respect to semi-positive definite matrices. Let us parameterize these random variables with respect to a small positive perturbation ε:

$$\begin{aligned}
\mathbf{U}_1^{\varepsilon} &= \begin{pmatrix} \varepsilon & 0 \\ 0 & 1 \end{pmatrix}, \quad \mathbf{U}_2^{\varepsilon} = \begin{pmatrix} 1 & 0 \\ 0 & \varepsilon \end{pmatrix}, \\
\mathbf{u}_1(\omega) &= \begin{pmatrix} u_{11} \\ u_{12} \end{pmatrix}, \quad \mathbf{u}_2(\omega) = \begin{pmatrix} u_{21} \\ u_{22} \end{pmatrix}.
\end{aligned}$$

The values of $\mathbf{u}_1$ and $\mathbf{u}_2$ are not important, but clearly their uncertainties depend on ε. Small values of ε correspond to the case where $\mathbf{u}_1$ has little uncertainty in its first component, and $\mathbf{u}_2$ has little uncertainty in its second component. The 'sensible' way to assimilate such states then, is to put most of the weight on u_{11} and u_{22}. Indeed the direct assimilation procedure does exactly this; for $\varepsilon > 0$ we obtain from (3.11)

$$\mathbf{w}^\varepsilon = \frac{1}{\varepsilon+1}\begin{pmatrix} u_{11}+\varepsilon u_{21} \\ \varepsilon u_{12}+u_{22}\end{pmatrix}.$$

We can define $\mathbf{w} = \lim_{\varepsilon\to 0}\mathbf{w}^\varepsilon$ and $\mathbf{U}_j = \lim_{\varepsilon\to 0}\mathbf{U}_j^\varepsilon$. Clearly, $\mathbf{w} = (u_{11}, u_{22})^T$. One can verify that either iterative procedure

$$\mathbf{w} = \mathbf{u}_1 + \mathbf{U}_1(\mathbf{U}_1+\mathbf{U}_2)^\dagger(\mathbf{u}_2-\mathbf{u}_1) \tag{3.12a}$$

$$\mathbf{w} = \mathbf{u}_2 + \mathbf{U}_2(\mathbf{U}_2+\mathbf{U}_1)^\dagger(\mathbf{u}_1-\mathbf{u}_2) \tag{3.12b}$$

produces this state $\mathbf{w}$ regardless of ordering (indeed $\mathbf{U}_1+\mathbf{U}_2 = \mathbf{I} > 0$ so the pseudoinverse is not even necessary). One may then be tempted to use the direct approach (3.11), but by naïvely replacing all direct inverses $^{-1}$ by pseudoinverses †. If we do this and then apply (3.11), we obtain a state $\mathbf{w} = (u_{21}, u_{12})^T$. This state is the opposite of what we should do: we completely ignore the components that we have the most information about. This shows that the direct assimilation approach cannot be remedied for positive semi-definite covariances.

Let us now show how the iterative scheme treats consistent random variables with semi-positive covariances. Again we take $\mathbf{u}_1, \mathbf{u}_2 \in \mathbb{R}^2$, and we now prescribe the covariances

$$\mathbf{U}_1^\varepsilon = \begin{pmatrix} \varepsilon & 0 \\ 0 & \varepsilon\end{pmatrix}, \quad \mathbf{U}_2^\varepsilon = \begin{pmatrix} 1 & 0 \\ 0 & \varepsilon\end{pmatrix}.$$

We now expect that the procedure should discard u_{21} in favor of u_{11}, and should equally weigh the values from u_{12} and u_{22}. When $\varepsilon > 0$ we can use either the direct or iterative schemes to obtain

$$\mathbf{w}^\varepsilon = \begin{pmatrix} \frac{1}{\varepsilon+1}(u_{11}+\varepsilon u_{21}) \\ \frac{1}{2}u_{12}+\frac{1}{2}u_{22}\end{pmatrix}.$$

Taking limits we obtain $\mathbf{w}^{(1)} \triangleq \lim_{\varepsilon\to 0}\mathbf{w}^\varepsilon = \left(u_{11}, \frac{1}{2}u_{12}+\frac{1}{2}u_{22}\right)^T$ as expected. Again define $\mathbf{U}_1 = \lim_{\varepsilon\to 0}\mathbf{U}_1^\varepsilon$ and similarly for $\mathbf{U}_2$. The iterative scheme (3.12a) produces the state $\mathbf{w}^{(2)} = (u_{11}, u_{12})^T$ whereas the second scheme (3.12b) produces $\mathbf{w}^{(3)} = (u_{11}, u_{22})$. Both of these are potentially different from $\mathbf{w}^{(1)}$, and from each other. However, if $\mathbf{u}_1$ and $\mathbf{u}_2$ are consistent random variables, then u_{12} and u_{22} are almost surely equal, and therefore $\mathbf{w}^{(1)} = \mathbf{w}^{(2)} = \mathbf{w}^{(3)}$ almost surely, so that the final choice among the three options is largely irrelevant.

If the random variables are not consistent, then all three $\mathbf{w}$ vectors differ with nonzero probability, but in this case there can be no remedy: we have two models with a non-vanishing discrepancy that are both entirely sure of their own accuracy.

We close this section by noting that the ability to assimilate components with vanishing variance opens up attractive possibilities. One may enforce constraints of the analyzed state vector (e.g. conservation of mass) in a post-processing step by treating the constraints as zero-variance data.

3.7. Bayesian interpretation of multi-model assimilation

To place the present scheme in context, it is instructive to consider its Bayesian interpretation. Let $\mathbf{w}^k \triangleq \mathbf{w}(t_k)$ denote the assimilated state at time t_k. From the Bayesian perspective, $\mathbf{w}^k$ is a random variable whose distribution captures the current state of knowledge about the truth state $\mathbf{u}^t(t_k)$. In other words, the distribution of $\mathbf{w}^k$ is the posterior distribution of the system state, given the data up to time t_k and the available models.

To be more specific, the present scheme provides the mean and covariance of the posterior probability density $p(\mathbf{w}^k|\mathbf{d}^{1:k}, \mathcal{M}_1, \ldots, \mathcal{M}_M)$, where $\mathbf{d}^{1:k} \triangleq (\mathbf{d}(t_k), \mathbf{d}(t_{k-1}), \ldots, \mathbf{d}(t_1))$ are the data provided up to assimilation timestep t_k and $\mathcal{M}_m$ are individual models. If we had only one model $\mathcal{M}_1$ and one source of data, the posterior density of $\mathbf{w}^k$ would be written as

$$p\left(\mathbf{w}^k|\mathbf{d}^{1:k}, \mathcal{M}_1\right) \propto p\left(\mathbf{d}^k|\mathbf{w}^k, \mathcal{M}_1\right)p\left(\mathbf{w}^k|\mathbf{d}^{1:k-1}, \mathcal{M}_1\right) = p\left(\mathbf{d}^k|\mathbf{w}^k\right)\int p(\mathbf{w}^k|\mathbf{w}^{k-1}, \mathcal{M}_1)p\left(\mathbf{w}^{k-1}|\mathbf{d}^{1:k-1}, \mathcal{M}_1\right)d\mathbf{w}^{k-1}. \tag{3.13}$$

As is typical in Kalman filtering variants, the assimilation step approximates the data likelihood $p\left(\mathbf{d}^k|\mathbf{w}^k\right)$ and the forecast distribution $p\left(\mathbf{w}^k|\mathbf{d}^{1:k-1}, \mathcal{M}_1\right)$ as Gaussian, even if propagation of uncertainty through a forecast model results in a non-Gaussian distribution. In this case, if the forecast state has mean $\mathbf{u}_1^k$, covariance $\mathbf{U}_1^k$, and $\mathbf{H}_1 = \mathbf{I}$, such that

$$\mathbf{w}^k = \mathbf{u}_1^k + \boldsymbol{\epsilon}_1^k, \quad \boldsymbol{\epsilon}_1^k \sim N(0, \mathbf{U}_1^k),$$

then the posterior mean and covariance of $\mathbf{w}^k$ are exactly given by the Kalman update specified in Section 2.2.

With more than one model, it is natural to interpret models $\mathcal{M}_m$, $m \geqslant 2$, as providing additional terms in the likelihood function, e.g. $p_m\left(\mathbf{u}_m^k|\mathbf{w}^k, \mathbf{d}^{1:k-1}, \mathcal{M}_m\right)$, where $\mathbf{u}_m^k \triangleq \mathbf{u}_m(t_k)$. These terms can be justified as follows. Beginning with the multi-model prior $p\left(\mathbf{w}^{k-1}|\mathbf{d}^{1:k-1}, \mathcal{M}_{1:M}\right)$, each model propagates this distribution forward to the next assimilation time and adds its own sources of randomness (e.g. parametric uncertainty or process noise). Each resulting model-specific forecast distribution is then described only by its mean $\mathbf{u}_m^k$ and covariance $\mathbf{U}_m^k$. One can write the forecasts as

$$\mathbf{H}_m\mathbf{w}^k = \mathbf{u}_m^k + \boldsymbol{\epsilon}_m^k,$$

where $\epsilon_m^k \sim N(0, \mathbf{U}_m^k)$. The resulting likelihood term is

$$p_m\left(\mathbf{u}_m^k|\mathbf{w}^k, \mathbf{d}^{1:k-1}, \mathcal{M}_{1:M}\right) \propto \exp\left(-\frac{1}{2}\left(\mathbf{H}_m\mathbf{w}^k - \mathbf{u}_m^k\right)^T\left(\mathbf{U}_m^k\right)^{-1}\left(\mathbf{H}_m\mathbf{w}^k - \mathbf{u}_m^k\right)\right). \tag{3.14}$$

Note that the conditioning on *all* models $\mathcal{M}_{1:M}$ reflects the fact that all of these forecasts began with the assimilated multi-model state at timestep $k-1$, but the model-specific subscript in p_m emphasizes that subsequent forecasting was performed only with the mth model.

Since the model forecasts are conditionally independent given $\mathbf{w}^k$, the likelihood terms can be combined to yield the multi-model posterior density:

$$p(\mathbf{w}^k|\mathbf{d}^{1:k}, \mathcal{M}_{1:M}) \propto p(\mathbf{d}^k|\mathbf{w}^k)\left(\prod_{m=2}^{M} p_m(\mathbf{u}_m^k|\mathbf{w}^k, \mathbf{d}^{1:k-1}, \mathcal{M}_{1:M})\right) p_1(\mathbf{w}^k|\mathbf{d}^{1:k-1}, \mathcal{M}_{1:M}). \tag{3.15}$$

Here the prior predictive $p_1(\mathbf{w}^k|\mathbf{d}^{1:k-1}, \mathcal{M}_{1:M})$ is chosen to reflect the forecast of any one of the M models, here $\mathcal{M}_1$. Assuming $\mathbf{H}_1 = \mathbf{I}$, this distribution is Gaussian with mean $\mathbf{u}_1^k$ and covariance $\mathbf{U}_1^k$.

We can motivate the form of the variational functional (3.5) for the direct assimilation method from this Bayesian analysis: note that maximizing the posterior probability given by (3.14) and (3.15) is equivalent to minimizing its negative logarithm. The negative log-posterior (modulo constants) is given by the functional $\mathcal{J}[w]$ from (3.5).

Compared to other model averaging techniques, the present scheme does not need to specify additional dynamics on the model space. Instead, the role of the covariance is paramount in determining the weight assigned to each model prediction and to the data. The covariance-based scheme thus allows matrix-valued weights and is a natural generalization of the Kalman filter.

4. Examples

In this section we provide a few simple examples that illustrate the broad range of applicability of the model assimilation approach previously detailed. Our implemented methods use the iterative procedure outlined in Theorem 3, and we assume that all models represent consistent random variables. For our examples, this is a valid assumption: all involved random variables have strictly positive covariances. The cost of the assimilation procedure does not suffer greatly from use of the pseudoinverse implementation as a safeguard.

Our first example concerns a rudimentary differential equation $y' = ay$ and uses Taylor polynomials as the models. The second example is a related stochastic differential equation (SDE) example that uses the same Taylor polynomial models, but showcases the complex interplay that can occur between the models and the data. Finally, we consider a periodic one-dimensional advection problem with non-constant wavespeed to show how the strengths of different models can be combined by using the multimodel assimilation approach.

4.1. An exponential model

Consider a system whose truth is given by

$$\frac{\mathrm{d}u^t}{\mathrm{d}t} = au^t, \quad (u^t)_0 = u_0,$$

for some u_0. The observations are scalar and the measurement matrix H is the identity (here, the scalar 1). The measurement noise ε is assumed to be $\mathcal{N}(0, \sigma^2)$. For any $m \geqslant 1$, the forecast model propagators G_m from (2.3) are given by

$$G_m(v) = \left(\sum_{k=0}^{m-1}\frac{(a\Delta t)^k}{k!}\right)v,$$

which is a degree-$(m-1)$ Taylor approximation of the true solution. We first let $\Delta t = 0.05,\ a = 1,\ u_0 = 0.1$, and $\sigma = 0.05$. We use $M = 2$ models (constant and linear approximations). Having information about the models, we assume that the standard deviation of the propagation error $(u^t)_n - G_m((u^t)_{n-1})$ is given by $0.1\Delta t^{m-1}$: i.e., the constant model is assumed to have error with standard deviation $0.1\Delta t$, while the linear model's error has standard deviation $0.1\Delta t^2$. Naturally these are not entirely accurate and we can make far better assumptions about the error, but this will serve our purposes. Furthermore, we assume that data is only available every 5 timesteps—in the absence of data, the propagators G_m are used to obtained analyzed states without any assimilation.

Note that by construction, all the forecast models are consistently biased below the true solution and data. If one adopts Bayesian model averaging (BMA), then the analyzed state is guaranteed to be less accurate than the best available forecast

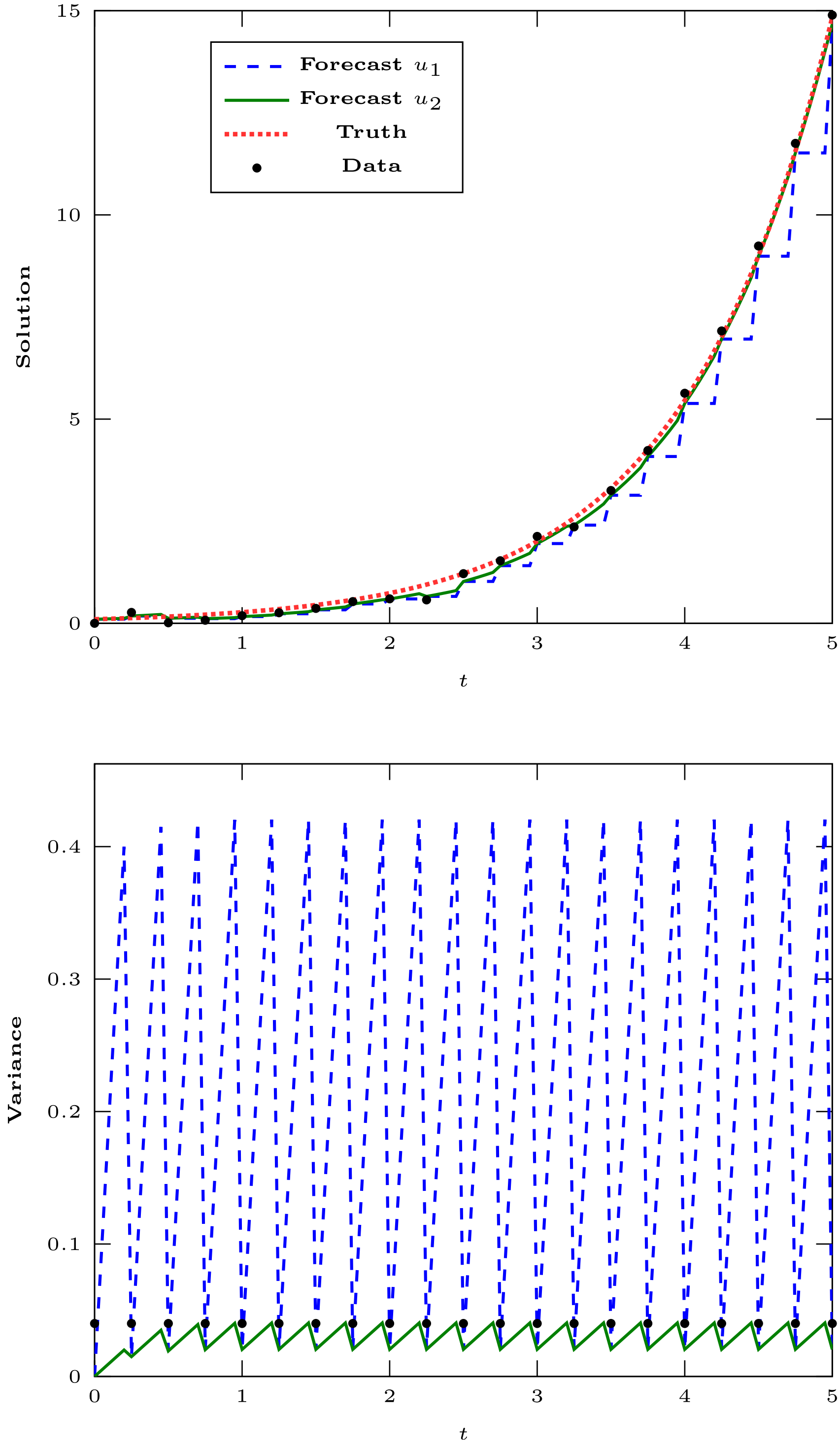

Fig. 4.1. Plot of the evolution of the forecast states along with indication of analyzed states and data with $M = 2$ forecast models.

model G_M. This is because in the approaches such as BMA, data only manifest itself in the averaging weights via the posterior probability. In our assimilation method, this shortcoming is naturally avoided.

We show the results of the data assimilation simulation in Fig. 4.1. We note in particular that while each of the two models undershoots the solution, the weight on the data is enough to bring the analyzed state closer to the data point. The variances of the forecasts are also plotted and the results are not surprising: the variances increase according to our coarse assumption about the numerical discretization error, and at assimilation times the variance is reduced dramatically.

In Fig. 4.2 we show the evolution of these assimilation weights. In this case, the filter considers the computed variances of the $m = 1$ and $m = 2$ model; compared to the measurement error variance, the data is slightly more accurate than the models, so the filter takes more information from the data. We run another test by (a) decreasing σ^2 to 4×10^{-3}, (b) increasing our assumed numerical discretization error to Δt^m, and (c) increasing the number of models to $M = 4$, incorporating both quadratic and cubic Taylor approximations. We show only the assimilation weights in the right-hand plot of Fig. 4.2. Now the variances of the quadratic and cubic models are comparable to the data measurement error, so the filter places more weight on the two accurate models. Of course, by decreasing the measurement noise σ to a low enough level, we could obtain data weights that are much larger than any of the polynomial forecasts. The figure shows that a steady state is achieved quickly in terms of the assimilation weights that are placed on the data and models.

4.2. An SDE with multiplicative white noise

Let $W_t = W(t,\omega)$ be a standard Wiener process. We consider in this example the stochastic differential equation given by

$$\mathrm{d}u^t = au^t\mathrm{d}t + bu^t\mathrm{d}W_t, \quad u^t(0) = u_0,$$

which has an exact solution in terms of an Ito integral [16]:

$$u^t = u_0 \exp\left(\left(a - \frac{1}{2}b^2\right)t + bW_t\right). \tag{4.1}$$

We use the same models as the last example—Taylor polynomial approximations based upon the current state. We must augment our assumption about the propagation error. We let the truth solution be a realization of (4.1). Brownian motion is simulated by solving $du = dW_t$ with an Euler method of stepsize 10^{-4}. By making the crude assumption that $\exp(W_t) \sim 1 + W_t + \frac{1}{2}(W_t)^2$, we obtain that the standard deviation in a Δt step of propagation is $u_0 b\sqrt{\Delta t}\exp((a - \frac{1}{2}b^2)\Delta t)$. Every Δt time units, variances of the previous amount, as well as that corresponding to the discretization error of $0.1\Delta t^{m-1}$, are added to the current U_m covariance. Since we have multiplicative noise, this noise is larger when u_m is large, and smaller when u_m is small. Therefore we expect the assimilation procedure to place more weight on the models when $|u^t|$ is small, and less weight on the models when $|u^t|$ is large.

We take $a = b = 1$ with $u_0 = 0.5$. We assimilate data, polluted by independent $\mathcal{N}(0, 2.5 \times 10^{-3})$ observational errors, every 50 propagation steps and integrate up to $t = 3$. The results of the evolution are shown in Fig. 4.3. The left-hand plot shows that our assimilation method follows the truth solution with some relative degree of accuracy; we note that this happens despite the data, which sometimes is very inaccurate. The assimilation weights plotted on the right-hand side show that the model puts weight on the data when the noise is small compared with the expected noise of the models. In other cases it prefers the models over the data.

4.3. An advection example

Finally we present a few variations on the following example: our truth solution obeys the one-dimensional wave equation

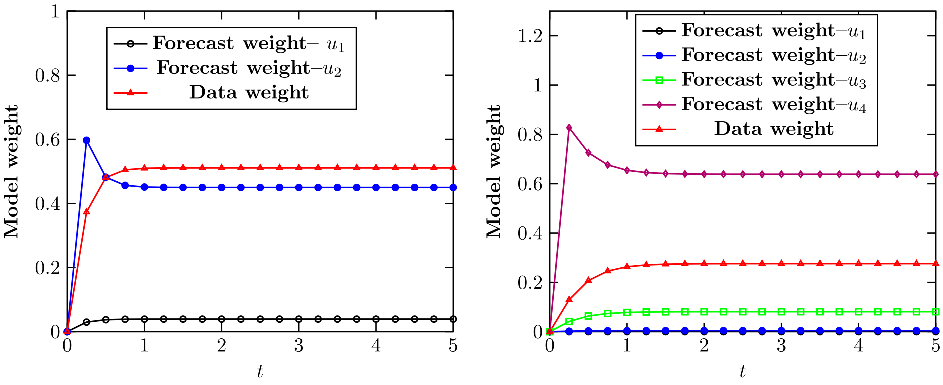

Fig. 4.2. Plot of the evolution of the assimilation weights for each of the forecast states along with the data. Left: $M = 2$ forecast models. Right: $M = 4$ forecast models.

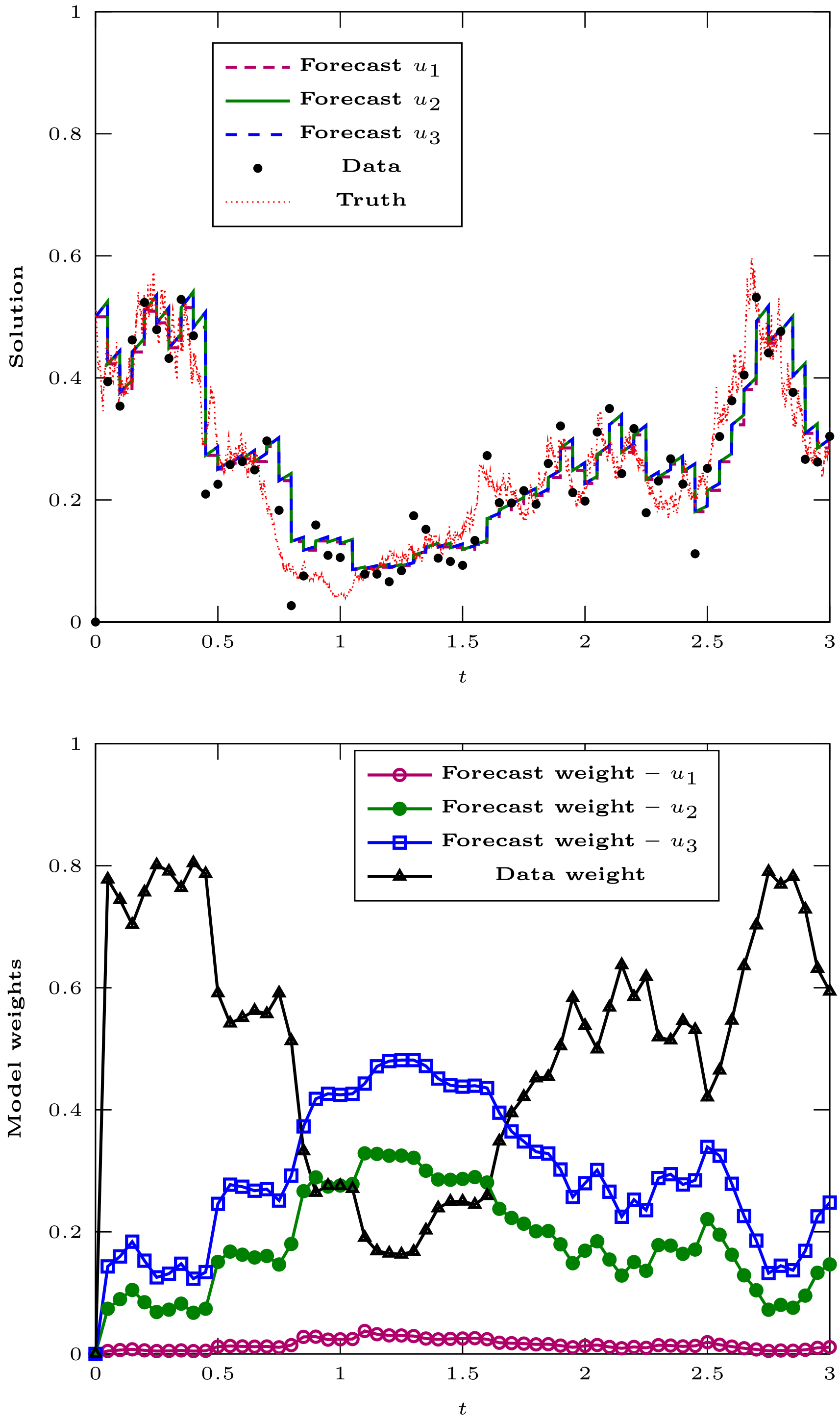

Fig. 4.3. Plot of the evolution of the forecast states (top) along with the truth solution and provided data with $M = 3$ forecast models for the SDE problem. At this scale, models 2 and 3 are visually indistinguishable. Also plotted are the assimilation weights (bottom) for the models and the data. Since models 2 and 3 are piecewise linear and quadratic, respectively, they are indistinguishable in this plot due to the small stepsize and frequent assimilation.

$$\frac{\partial u^t}{\partial t} = (1 + c(x,\omega))\frac{\partial u^t}{\partial x}, \quad x \in [0, 2\pi), \tag{4.2}$$

with initial data $u^t(t=0) = u_0(x,\omega)$, meaning that u_0 is random. Before discussing the numerical scheme, we first characterize the wavespeed c and the initial data u_0. We introduce the orthonormal Fourier series functions $\phi_n(x) = \frac{1}{\sqrt{2\pi}}\exp(inx)$. We let $c(x,\omega)$ be a weakly stationary Gaussian random field with a permissible covariance function $C(r) = \mathbb{E}[c(x,\omega) - \mathbb{E}c(x)][c(x+r,\omega) - \mathbb{E}c(x+r)]$ defined by a sum of Fourier modes:

$$C(r) = \sum_{|n|\leqslant K} \widehat{C}_n \phi_n(r). \tag{4.3}$$

We also want $c > 0$ for all x with probability 1, which we enforce with high probability by imposing the mean $\mathbb{E}c(x) = 4$. The $\widehat{C}_n$ are given by $\frac{1}{1+n^2}$. We set $K = 15$ for all the following simulations. The truth solution u^t obeys the above equation for a particular realization of the random field.

A similar set of assumptions is made about the initial data u_0: it has the covariance function (4.3) with $K = 50$ and the same values for the $\widehat{C}_n$. The truth initial data again is a realization of this field. All our discretizations are finite-difference methods, so all the fields are evaluated at grid points to produce degrees of freedom.

The numerical method we use to solve (4.2) will be an upwind finite-difference method, and we use periodic boundary conditions. The vector of unknowns (point-evaluations) for model m is $\mathbf{u}_m$, and entry n in the vector is denoted $u_{m,n}$. Since $c > 0$, we have for example the following first and third order methods, assuming equidistant grids with stepsize h:

$$\frac{\partial u_{m,n}}{\partial x} \approx D_1(u_{m,n}) = \frac{1}{h}(u_{m,n+1} - u_{m,n}),$$
$$\frac{\partial u_{m,n}}{\partial x} \approx D_3(u_{m,n}) = \frac{1}{h}\left(-\frac{1}{6}u_{m,n+2} + u_{m,n+1} - \frac{1}{2}u_{m,n} - \frac{1}{3}u_{m,n-1}\right).$$

We then apply temporal discretization to the semidiscrete system

$$\frac{\mathrm{d}u_{m,n}}{\mathrm{d}t} = D_j(u_{m,n}) \quad n = 1, 2, \ldots, N,$$

where the choice of accuracy j depends on the model. For all methods we use a fourth-order strong stability preserving explicit Runge–Kutta scheme to perform the temporal evolution with $\Delta t = 10^{-3}$ for all models and simulations. The truth solution is the finite-difference solution with $N = 900$ equispaced points using the third-order upwind operator D_3.

We consider four models; the first model $\mathbf{u}_1$ is a first order method using D_1, but has a very fine equidistant mesh with $N = 900$. The second through fourth models are all third-order upwind schemes using D_3, but have different coarser meshes. For example, on the interval $[0, 2\pi/3)$, $\mathbf{u}_2$ is 3 times coarser than the grid for $\mathbf{u}_1$, and on $[2\pi/3, 2\pi)$ it is 10 times coarser. A qualitative plot of the grids for each model is shown in Fig. 4.4.

4.3.1. Assimilation of models

The first example here is to show that data-blind assimilation is also a useful method that can efficiently combine existing models to provide a better guess of the true solution, provided rough error estimates for each of the models are available. In this case we are not inserting randomness into the system: we generate one realization of c and use this as the wavespeed for all models. However, due to the spatial discretization errors, we assume that each propagator G_m augments the covariance matrix as follows:

$$U_m(t_{k+1}) = U_m(t_k) + J_m, \tag{4.4}$$

where J_m is a diagonal matrix, whose nth diagonal entry is

$$(J_m)_{n,n} = \left(\bar{h}_{m,n}\right)^{p(m)} \tag{4.5}$$

and $\bar{h}_{m,n}$ is the average of the stepsizes of the left and right of the nth node of model m, and p denotes the spatial order of accuracy of each model with $p(m) = 1$ for $m = 1$ and $p(m) = 3$ for $m = 2,\ 3,\ 4$. Although J_m is meant to be an error estimator for the models, the approximation given by (4.5) is very rough. Furthermore, the covariance update (4.4) is not accurate since it does not take into account the fact that error is advected across the domain, and assumes that the discretization error is uncorrelated on the grid. Nevertheless, we proceed to use this covariance update to define the assimilation procedure. We assimilate every 10 timesteps and integrate up to $t = 3$ (300 assimilations). Since the average wavespeed is about 4, this corresponds to about 2 full cycles of the wave across the domain. The ℓ^2 error is measured as on the grid corresponding to $\mathbf{u}_1$,

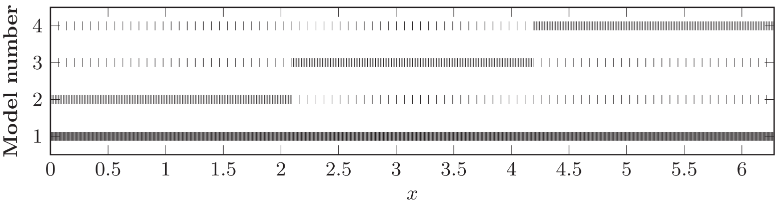

Fig. 4.4. Qualitative plots of the grid densities for each of the four models considered. Model 1 has the finest grid, but uses a low-order finite difference. Models 2–4 have coarser grids in different parts of the interval, but use higher-order finite difference approximations.

which is the finest grid. We show the evolution of this error in the top plot of Fig. 4.5, along with the errors of each of the models *without* any assimilation procedures. We see that the assimilated state indeed picks the best parts of each solution and combines them into one.

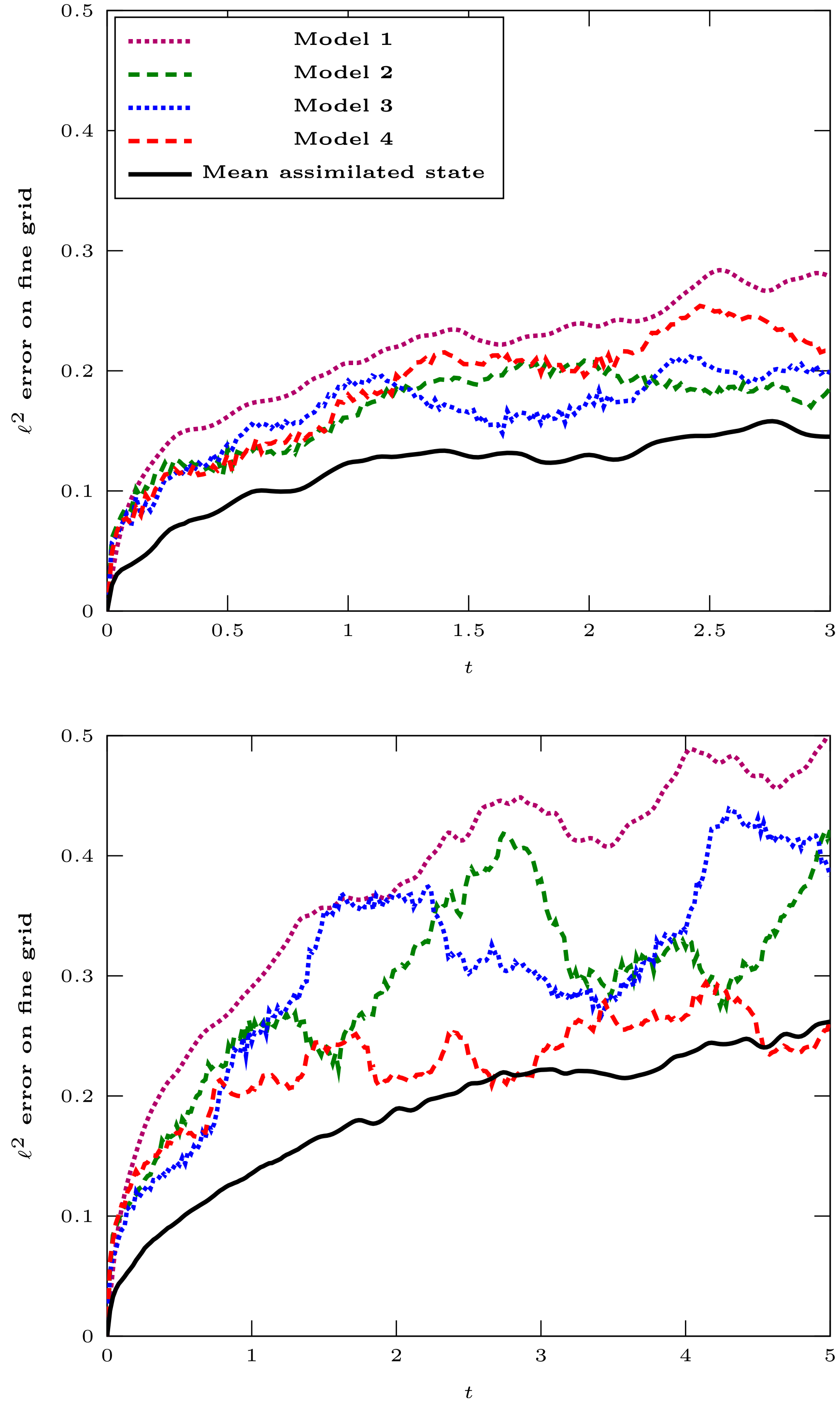

Fig. 4.5. Error plots for Section 4.3.1. Evolution of the ℓ^2 error (on the fine grid) against the truth for the four-model assimilation procedure. The top plot shows a non-random simulation with multimodel assimilation with covariance update defined by (4.4). The bottom plot shows a simulation where there is uncertainty present in the wavespeed and the covariance is computed via a 10^4-size ensemble.

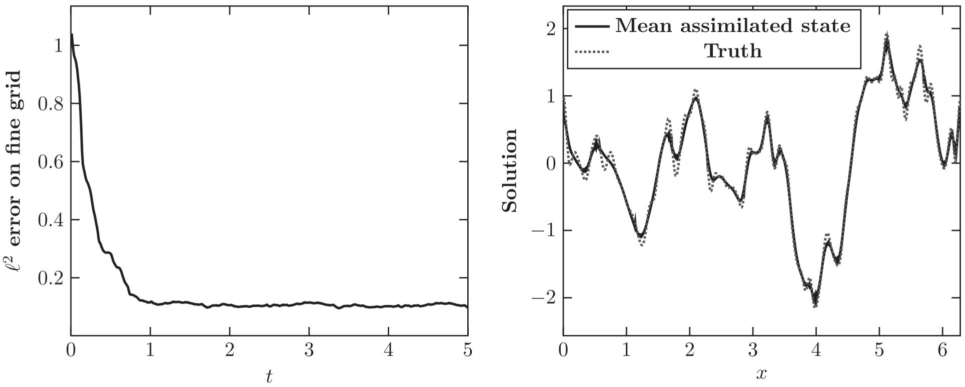

Fig. 4.6. Truth wave vs. mean assimilated state for the advection model of Section 4.3.2. The snapshot on the right is shown at $t = 5$ after 500 assimilations and about three cycles of the wave through the domain.

To further test the multimodel assimilation, we now test a more difficult and realistic situation. We take the same models again. Suppose we are uncertain about the wavespeed; we again choose some realization of $c(x,\omega)$ as the truth, and for each model m, we perturb $c(x_{m,n})$ with a mean-zero Gaussian, whose variance is proportional to $\bar{h}_{m,n}^{p(m)}$. This uncertainty is meant to mirror numerical discretization uncertainty, so now we forego the covariance update (4.4) and this time employ an ensemble with 10^4 realizations to compute the covariance. The initial data for each model is exact—the only uncertainty that is present is due to the uncertain wavespeed. We again see from the right-hand plot of Fig. 4.5 that because we have chosen the uncertainty in the wavespeed to mimic the numerical discretization error, the discrete ℓ^2 between the truth and the mean assimilated state is consistently smaller than that of any one of the models without assimilation.

4.3.2. Data assimilation

Now we introduce data into the system. We again consider the four models of the previous setup. However, now we take 11 equispaced point-value measurements on the domain. The data noise ϵ is a multivariate mean-zero Gaussian with covariance $\sigma^2\mathbf{I}$, where $\sigma^2 = 10^{-4}$ and $\mathbf{I}$ is the 11×11 identity matrix. We now also let the initial condition be uncertain: $u_0(x)$ is a random field described in the same way as $c(x,\omega)$.

Therefore, randomness is inserted into the system via three mechanisms:

- The truth initial data $u_0(\omega)$ is a realization from the covariance function (4.3), and the ensemble of the initial data is simply taken as different realizations of this same field, evaluated at the appropriate grid values.
- The truth wavespeed is a realization of $c(\omega)$, but the gridpoint evaluations $c(x_{m,n})$ (constant in time) are perturbed by zero-mean multivariate Gaussians whose covariance is J_m, defined by (4.5).
- The data vector $\mathbf{d}$ is assumed to have covariance $10^{-4}\mathbf{I}$, where $\mathbf{I}$ is the 11×11 identity matrix.

We still employ an ensemble method with size 10^4, but we also encode information about the assumed accuracy of the numerical models by performing the approximate covariance update (4.4) after computing the ensemble covariance. Therefore our covariance incorporates uncertainty in the initial data and wavespeed through the ensemble, and uncertainty in the numerical model through a post-operation at each time step of (4.4). We assimilate every 5 timesteps and integrate up to $t = 5$, which is more than 3 full rotations through the domain, and corresponds to 500 assimilations. We show the mean assimilated state at $t = 5$, compared to the truth, in Fig. 4.6. We emphasize that each of the models individually, even with the data assimilation, do not closely follow the true wave at all. Only when we assimilate all the models do we obtain the reasonable mean state shown in Fig. 4.6.

5. Conclusion

In this paper we presented a new framework for sequential data assimilation with multiple sources of models and data. The framework can be considered as a notable generalization of the Kalman filter, which is widely used for single model data assimilation. We presented the mathematical properties of the new method. More importantly, we also presented an effective iterative algorithm that essentially renders the implementation of the new method as a recursive two-model Kalman filter scheme. The mathematical conditions under which the iterative algorithm is valid are established.

We remark that there remain many important modeling and computational challenges in the context of multiple model averaging and filtering.

- The number of models considered may be extremely large, leading to significant computational expense.

- One might like to remove models that are shown to be ineffective, based on some metric. A good example is the Occam's window technique [17].
- Effective and rigorous means of specifying the error covariance for each model in the current method is a challenge. This is the same difficulty faced in the traditional Kalman filter setup. It represents the requirement to have a good prior understanding of the models. (A similar requirement in BMA or DMA exists, as one needs to specify the prior model probabilities.)
- Our method is employs a linear filtering technique to assimilate models and data. In situations when the truth state is nonlinearly related to the model states as given by (2.5) then nonlinear filtering techniques become necessary. Such extensions of our method could employ the Unscented Kalman Filter, particle filters, batch filters, or numerical solutions of the Fokker–Planck equation.

We did not attempt to address these broader challenges in the present paper, and will conduct further study in the current framework to address them.

Acknowledgments

This work is in part supported by AFOSR, DOE/NNSA, NSF.

Appendix A

Here we compile proofs of results along with some mathematically relevant discussions. Given a linear subspace $V \subset \mathbb{R}^N$, we will use the notation $\mathbf{P}_V$ to denote the Euclidean orthogonal projection operator onto V. The orthogonal complement of V in $\mathbb{R}^N$ is $V^\perp$. Our main goals are the proofs of Theorem 1 and 3: that the harmonic mean $\mathbf{F}_M$ defined in 1 satisfies properties that justify calling it a mean, and that the iterative procedure for assimilation is order-independent. We start with the harmonic mean, and since our main tool will be induction, it proves useful to consider the special case $M = 2$.

Lemma 1. $\mathbf{F}_2$ *is continuous in both arguments.*

Proof. We show that $\mathbf{W} \triangleq \mathbf{W}_2$ is continuous. First we note that if $(\ker \mathbf{A}_1) \cap (\ker \mathbf{A}_2) = \{\mathbf{0}\}$, then there is little to prove: $\mathbf{A}_1 + \mathbf{A}_2$ is invertible, and the standard inverse is continuous on the open set of invertible matrices, and so therefore $\mathbf{W}_2$ is also. Therefore, assume that $\mathbf{A}_1$ and $\mathbf{A}_2$ have kernels with a nontrivial intersection, $\mathcal{K} = (\ker \mathbf{A}_1) \cap (\ker \mathbf{A}_2)$. Let

$$\mathbf{W}^\varepsilon = \left(\mathbf{A}_1 + \frac{1}{2}\varepsilon\mathbf{I}\right)(\mathbf{A}_1 + \mathbf{A}_2 + \varepsilon\mathbf{I})^{-1}\left(\mathbf{A}_2 + \frac{1}{2}\varepsilon\mathbf{I}\right)$$

We proceed by showing that for all $\mathbf{v} \in \mathbb{C}^N$, $\mathbf{W}^\varepsilon\mathbf{v} \to \mathbf{W}\mathbf{v}$. Let $\mathbf{v} = \mathbf{v}_\mathcal{K} + \mathbf{v}_{\mathcal{K}^\perp}$, where $\mathbf{v}_\mathcal{K} = \mathbf{P}_\mathcal{K}\mathbf{v}$. Let $\mathbf{I} = \mathbf{P}_\mathcal{K} + \mathbf{P}_{\mathcal{K}^\perp} = \mathbf{I}_\mathcal{K}\mathbf{I}_\mathcal{K}^T + \mathbf{I}_{\perp\mathcal{K}}\mathbf{I}_{\perp\mathcal{K}}^T$ be the spectral resolution of the identity on $\mathcal{K}$ and $\mathcal{K}^\perp$. Using the eigen decompositions of $\mathbf{A}_j$ and $\mathbf{A}_1 + \mathbf{A}_2$, we have

$$\mathbf{A}_j + \varepsilon\mathbf{I} = \mathbf{I}_{K^\perp}\left(\mathbf{I}_{\mathcal{K}^\perp}^T\mathbf{A}_j\mathbf{I}_{\mathcal{K}^\perp} + \varepsilon\mathbf{I}\right)\mathbf{I}_{\mathcal{K}^\perp}^T + \varepsilon\mathbf{P}_\mathcal{K}$$
$$(\mathbf{A}_1 + \mathbf{A}_2 + \varepsilon\mathbf{I})^{-1} = \frac{1}{\varepsilon}\mathbf{P}_\mathcal{K} + \mathbf{I}_{\mathcal{K}^\perp}\left[\mathbf{I}_{\mathcal{K}^\perp}^T(\mathbf{A}_1 + \mathbf{A}_2)\mathbf{I}_{\mathcal{K}^\perp} + \varepsilon\mathbf{I}\right]^{-1}\mathbf{I}_{\mathcal{K}^\perp}^T.$$

This implies, for example, that $(\mathbf{A}_1 + \mathbf{A}_2 + \varepsilon\mathbf{I})^{-1}\mathbf{v}_{\mathcal{K}^\perp} = \mathbf{I}_{\mathcal{K}^\perp}\left[\mathbf{I}_{\mathcal{K}^\perp}^T(\mathbf{A}_1 + \mathbf{A}_2)\mathbf{I}_{\mathcal{K}^\perp} + \varepsilon\mathbf{I}\right]^{-1}\mathbf{I}_{\mathcal{K}^\perp}^T\mathbf{v}_{\mathcal{K}^\perp}$, so we have that

$$\mathbf{W}^\varepsilon\mathbf{v}_{\mathcal{K}^\perp} = \mathbf{I}_{\mathcal{K}^\perp}\underbrace{\left(\mathbf{I}_{\mathcal{K}^\perp}^T\mathbf{A}_2\mathbf{I}_{\mathcal{K}^\perp} + \frac{\varepsilon}{2}\mathbf{I}\right)}_{(a)}\underbrace{\left[\mathbf{I}_{\mathcal{K}^\perp}^T(\mathbf{A}_1 + \mathbf{A}_2)\mathbf{I}_{\mathcal{K}^\perp} + \varepsilon\mathbf{I}\right]^{-1}}_{(b)}\underbrace{\left(\mathbf{I}_{\mathcal{K}^\perp}^T\mathbf{A}_1\mathbf{I}_{\mathcal{K}^\perp} + \frac{\varepsilon}{2}\mathbf{I}\right)}_{(c)}\mathbf{I}_{\mathcal{K}^\perp}^T\mathbf{v}_{\mathcal{K}^\perp}.$$

Terms (a) and (c) converge to $\mathbf{I}_{\mathcal{K}^\perp}^T\mathbf{A}_j\mathbf{I}_{\mathcal{K}^\perp}$ as $\varepsilon \to 0$. Since $\left[\mathbf{I}_{\mathcal{K}^\perp}^T(\mathbf{A}_1 + \mathbf{A}_2)\mathbf{I}_{\mathcal{K}^\perp}\right]$ is invertible, then term (b) approaches the expression without the ε term. Noting that $(\mathbf{A}_1 + \mathbf{A}_2)^\dagger = \mathbf{I}_{\mathcal{K}^\perp}\left[\mathbf{I}_{\mathcal{K}^\perp}^T(\mathbf{A}_1 + \mathbf{A}_2)\mathbf{I}_{\mathcal{K}^\perp}\right]^{-1}\mathbf{I}_{\mathcal{K}^\perp}^T$, and that the operator $\mathbf{I}_{\mathcal{K}^\perp}\mathbf{I}_{\mathcal{K}^\perp} = \mathbf{P}_{\mathcal{K}^\perp}$ is equivalent to the identity on $\operatorname{ran}\mathbf{A}_j = \mathcal{K}^\perp$, we obtain

$$\mathbf{W}^\varepsilon\mathbf{v}_{\mathcal{K}^\perp} \to \mathbf{A}_1\mathbf{I}_{K^\perp}\left[\mathbf{I}_{\mathcal{K}^\perp}^T(\mathbf{A}_1 + \mathbf{A}_2)\mathbf{I}_{\mathcal{K}^\perp}\right]^{-1}\mathbf{I}_{\mathcal{K}^\perp}^T\mathbf{A}_2\mathbf{v}_{\mathcal{K}^\perp} = \mathbf{A}_1(\mathbf{A}_1 + \mathbf{A}_2)^\dagger\mathbf{A}_2\mathbf{v}_{\mathcal{K}^\perp}.$$

The same properties derived above show that $\mathbf{W}^\varepsilon\mathbf{v}_\mathcal{K} = \frac{\varepsilon}{2}\mathbf{P}_\mathcal{K}\frac{1}{\varepsilon}\mathbf{P}_\mathcal{K}\frac{\varepsilon}{2}\mathbf{P}_\mathcal{K}\mathbf{v}_\mathcal{K}$, and so $\mathbf{W}^\varepsilon\mathbf{v}_\mathcal{K} = \frac{1}{4}\varepsilon\mathbf{v}_\mathcal{K} \to \mathbf{0} = \mathbf{A}_1(\mathbf{A}_1 + \mathbf{A}_2)^\dagger\mathbf{A}_2\mathbf{v}_\mathcal{K}$. Thus $\mathbf{W}^\varepsilon\mathbf{v} \to \mathbf{A}_1(\mathbf{A}_1 + \mathbf{A}_2)^\dagger\mathbf{A}_2\mathbf{v}$. □

Induction now allows us to show that $\mathbf{F}_M$ as defined in (3.2b) satisfies standard operator mean qualities.

Proof (*Theorem* 1). When appropriate, we show the results first for $\mathbf{W}_M$. All the results are a consequence of (i) the inductive definition and (ii) continuity for $M = 2$. First we show the results for $M = 2$: For (1), continuity implies that W_2 is Hermitian, as well as that the eigenvalues are non-negative. Continuity (2) is Lemma 1. Consistency (3) for $\mathbf{F}_2$ is a direct result of the fact that $\mathbf{A}\mathbf{A}^{\dagger}\mathbf{A} = \mathbf{A}$ for all matrices $\mathbf{A}$ ($\mathbf{W}_2$ is not consistent). Argument symmetry (4) is straightforward since $\mathbf{W}_2$ is Hermitian. (5) is an easy consequence of continuity and the same result for $\mathbf{W}_2$ on $\mathcal{H}$. We borrow monotonicity (6) from the theory of strictly positive matrices [3] – taking limits along with continuity, this implies monotonicity on the completed space $\mathcal{H}_0$. To extend the results to general M, induction is used. Since $\mathbf{F}_M = M\mathbf{W}_M$, the results (1), (2), (4)–(6) clearly hold for $\mathbf{F}_M$ as well. □

We conclude by proving the order-independence of the iterative scheme (3.9).

Proof (*Theorem* 3). We will outline the proof only for the case when $\mathbf{H}_m = \mathbf{I}$ for all m. When $\mathbf{H}_m \neq \mathbf{I}$ the ideas are essentially the same, and the operations are similar to the standard single-model Kalman filter derivation, which is described in many other texts.

If all the covariance matrices $\mathbf{U}_m$ are invertible, then the iterative procedure is identical to the direct procedure, as can be seen by manipulating standard matrix inverses with e.g. the Sherman–Morrison–Woodbury formula. It is thus independent of the ordering. Now assume that there is at least one $\mathbf{U}_m$ with vanishing eigenvalues. We will show independence of ordering for $M = 2$ and leave the rest to induction.

We first assume that $\mathbf{H}_m = \mathbf{I}$ for all m. Suppose that $(\ker \mathbf{U}_1) \cap (\ker \mathbf{U}_2)$ is the trivial subspace. Then $\mathbf{U}_1 + \mathbf{U}_2$ is invertible, so that

$$\begin{aligned}
\mathbf{w}_2 &= (\mathbf{I} - \mathbf{U}_1(\mathbf{U}_1 + \mathbf{U}_2)^{\dagger})\mathbf{u}_1 + \mathbf{U}_1(\mathbf{U}_1 + \mathbf{U}_2)^{\dagger}\mathbf{u}_2 \\
&= (\mathbf{I} - \mathbf{U}_1(\mathbf{U}_1 + \mathbf{U}_2)^{-1})\mathbf{u}_1 + \mathbf{U}_1(\mathbf{U}_1 + \mathbf{U}_2)^{-1}\mathbf{u}_2 \\
&= \mathbf{U}_2(\mathbf{U}_1 + \mathbf{U}_2)^{-1}\mathbf{u}_1 + (\mathbf{I} - \mathbf{U}_2(\mathbf{U}_1 + \mathbf{U}_2)^{-1})\mathbf{u}_2 \\
&= (\mathbf{I} - \mathbf{U}_2(\mathbf{U}_1 + \mathbf{U}_2)^{\dagger})\mathbf{u}_2 + \mathbf{U}_2(\mathbf{U}_1 + \mathbf{U}_2)^{\dagger}\mathbf{u}_1
\end{aligned}$$

and the equivalence of the first and last lines show the order-independence for $M = 2$. Thus, if the pairwise kernel intersections of $\{\mathbf{U}_m\}_{m=1}^{M}$ are all trivial, then the above shows order-independence for general M. Now suppose that $N = \ker \mathbf{U}_1 \cap \ker \mathbf{U}_2$ is nontrivial so that the pseudoinverse is distinct from the traditional inverse. Then $\mathbf{U}_1 + \mathbf{U}_2 + k\mathbf{P}_N > 0$ for any $k > 0$, and $(\mathbf{U}_1 + \mathbf{U}_2)^{\dagger}\mathbf{u} = (\mathbf{U}_1 + \mathbf{U}_2 + k\mathbf{P}_N)^{-1}\mathbf{u}$ for all $\mathbf{u} \in \operatorname{ran}(\mathbf{U}_1 + \mathbf{U}_2) = \operatorname{ran}(\mathbf{I} - \mathbf{P}_N)$. Let $\mathbf{v}_j \triangleq (\mathbf{I} - \mathbf{P}_N)\mathbf{u}_j$. Then assimilating $\mathbf{u}_2$ into $\mathbf{u}_1$ yields

$$\begin{aligned}
\mathbf{w}^{(1)} &= \mathbf{u}_1 + \mathbf{U}_1(\mathbf{U}_1 + \mathbf{U}_2)^{\dagger}(\mathbf{u}_2 - \mathbf{u}_1), \\
&= (\mathbf{I} - \mathbf{P}_N)\mathbf{u}_1 + \mathbf{P}_N\mathbf{u}_1 - \mathbf{U}_1(\mathbf{U}_1 + \mathbf{U}_2)^{\dagger}[\mathbf{P}_N(\mathbf{u}_2 - \mathbf{u}_1) + (\mathbf{I} - \mathbf{P}_N)(\mathbf{u}_2 - \mathbf{u}_1)], \\
&= \mathbf{v}_1 + \mathbf{U}_1(\mathbf{U}_1 + \mathbf{U}_2 + \mathbf{P}_N)^{-1}(\mathbf{v}_2 - \mathbf{v}_1) + \mathbf{P}_N\mathbf{u}_1.
\end{aligned}$$

Similarly, assimilating $\mathbf{u}_1$ into $\mathbf{u}_2$ yields

$$\mathbf{w}^{(2)} = \mathbf{v}_2 + \mathbf{U}_2(\mathbf{U}_1 + \mathbf{U}_2 + \mathbf{P}_N)^{-1}(\mathbf{v}_1 - \mathbf{v}_2) + \mathbf{P}_N\mathbf{u}_2.$$

Since $\mathbf{P}_N\mathbf{v}_j = \mathbf{0}$, then

$$\mathbf{w}^{(1)} - \mathbf{w}^{(2)} = \mathbf{P}_N(\mathbf{u}_1 - \mathbf{u}_2).$$

Now if $\mathbf{u}_1$ and $\mathbf{u}_2$ are consistent model states, then there exists a vector $\mathbf{y}$ such that $\mathbb{E}\mathbf{P}_N\mathbf{u}_1 = \mathbb{E}\mathbf{P}_N\mathbf{u}_2 = \mathbf{P}_N\mathbf{y}$, and so $\mathbb{E}[\mathbf{P}_N\mathbf{u}_1 - \mathbf{P}_N\mathbf{u}_2] = \mathbf{0}$. But $\mathbf{P}_N\mathbf{u}_1$ is a random variable with zero covariance so that $\mathbb{E}\mathbf{P}_N\mathbf{u}_1 = \mathbf{P}_N\mathbf{u}_1$ almost surely, and the same for $\mathbf{u}_2$. Therefore, $\mathbf{P}_N\mathbf{u}_1 = \mathbf{P}_N\mathbf{u}_2$ almost surely, and this implies that $\mathbf{w}^{(1)} = \mathbf{w}^{(2)}$ almost surely.

We have shown that ordering is independent for two consistent random variables when the observation matrices are both the identity. By appropriate permutations, any group of random variables that are consistent can be assimilated in any order. □

References

[1] J.L. Anderson, An ensemble adjustment Kalman filter for data assimilation, Monthly Weather Review 129 (2001) 2884–2903.

[2] J.L. Anderson, S.L. Anderson, A Monte Carlo implementation of the nonlinear filtering problem to produce ensemble assimilations and forecasts, Monthly Weather Review 127 (1999) 2741–2758.

[3] R. Bhatia, Positive Definite Matrices, Princeton University Press, 2006.

[4] H.A.P. Blom, An efficient filter for abruptly changing systems, The 23rd IEEE Conference on Decision and Control 23 (1984) 656–658.

[5] G. Burgers, P.V. Leeuwen, G. Evensen, Analysis scheme in the ensemble Kalman filter, Monthly Weather Review 126 (1998) 1719–1724.

[6] C.G.H. Diks, J.A. Vrugt, Comparison of point forecast accuracy of model averaging methods in hydrologic applications, Stochastic Environmental Research and Risk Assessment 24 (6) (2010) 809–820.

[7] G. Evensen, Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics, Journal of Geophysical Research 99 (1994) 10,143–10,162.

[8] G. Evensen, Data Assimilation: The Ensemble Kalman Filter, second ed., Springer, 2009.

[9] A. Gelb, Applied Optimal Estimation, MIT Press, Cambridge, 1974.

[10] J.A. Hoeting, D. Madigan, A.E. Raftery, C.T. Volinsky, Bayesian model averaging: a tutorial, Statistical Science 14 (4) (1999) 382–417.

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

[12] S. Julier, J. Uhlmann, H.F. Durrant-Whyte, A new method for the nonlinear transformation of means and covariances in filters and estimators, IEEE Transactions on Automatic Control 45 (3) (2000) 477–482.

[13] S.J. Julier, J.K. Uhlmann, New extension of the Kalman filter to nonlinear systems, Proceedings of SPIE 3 (1) (1997) 182–193.

[14] R. Kalman, R. Bucy, New results in linear prediction filter theory, Transactions of the ASME. Series D, Journal of Basic Engineering 83D (1961) 85–108.

[15] R. E Kalman, A new approach to linear filtering and prediction problems, Transactions of the ASME. Series D, Journal of Basic Engineering 82 (1960) 35–45.

[16] P.E. Kloeden, E. Platen, Numerical Solution of Stochastic Differential Equations, corrected ed., Springer, 2011.

[17] D. Madigan, A.E. Raftery, Model selection and accounting for model uncertainty in graphical models using Occam's window, Journal of the American Statistical Association 89 (428) (1994) 1535–1546.

[18] R. Van Der Merwe, A. Doucet, N. De Freitas, E. Wan, The unscented particle filter, Advances in Neural Information Processing Systems 96 (CUED/F-INFENG/TR 380) (2001) 584–590.

[19] A. Raftery, T. Gneiting, F. Balabdaoui, M. Polakowski, Using Bayesian model averaging to calibrate forecast ensembles, Monthly Weather Review 133 (2005) 1155–1174.

[20] A.E. Raftery, M. Kárný, P. Ettler, Online prediction under model uncertainty via dynamic model averaging: application to a cold rolling mill, Technometrics 52 (1) (2010) 52–66.

[21] M.K. Tippett, J.L. Anderson, C.H. Bishop, T.M. Hamill, J.S. Whitaker, Ensemble square-root filters, Monthly Weather Review 131 (2003) 1485–1490.

[22] P.J. van Leeuwen, Nonlinear data assimilation in geosciences: an extremely efficient particle filter, Quarterly Journal of the Royal Meteorological Society 136 (653) (2010) 1991–1999.

[23] K. Watanabe, S.G. Tzafestas, Generalized pseudo-Bayes estimation and detection for abruptly changing systems, Journal of Intelligent & Robotic Systems 7 (1) (1993) 95–112.

[24] J.S. Whitaker, T.M. Hamill, Ensemble data assimilation without perturbed observations, Monthly Weather Review 130 (2002) 1913–1924.