INTERNATIONAL JOURNAL FOR NUMERICAL METHODS IN FLUIDS
Int. J. Numer. Meth. Fluids 2008; **56**:1033–1046
Published online 27 December 2007 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/fld.1698

Data assimilation: Mathematical and statistical perspectives

A. Apte[1], C. K. R. T. Jones[1], A. M. Stuart[2,*,†] and J. Voss[2]

[1]*University of North Carolina, Chapel Hill, NC, U.S.A.*
[2]*University of Warwick, Mathematics Institute, Coventry CV4 7AL, U.K.*

SUMMARY

The bulk of this paper contains a concise mathematical overview of the subject of data assimilation, highlighting three primary ideas: (i) the standard optimization approaches of 3DVAR, 4DVAR and weak constraint 4DVAR are described and their interrelations explained; (ii) statistical analogues of these approaches are then introduced, leading to filtering (generalizing 3DVAR) and a form of smoothing (generalizing 4DVAR and weak constraint 4DVAR) and the optimization methods are shown to be maximum *a posteriori* estimators for the probability distributions implied by these statistical approaches; and (iii) by taking a general dynamical systems perspective on the subject it is shown that the incorporation of Lagrangian data can be handled by a straightforward extension of the preceding concepts.

We argue that the smoothing approach to data assimilation, based on statistical analogues of 4DVAR and weak constraint 4DVAR, provides the optimal solution to the assimilation of space–time distributed data into a model. The optimal solution obtained is a probability distribution on the relevant class of functions (initial conditions or time-dependent solutions). The approach is a useful one in the first instance because it clarifies the notion of what is the optimal solution, thereby providing a benchmark against which existing approaches can be evaluated. In the longer term it also provides the potential for new methods to create ensembles of solutions to the model, incorporating the available data in an optimal fashion.

Two examples are given illustrating this approach to data assimilation, both in the context of Lagrangian data, one based on statistical 4DVAR and the other on weak constraint statistical 4DVAR. The former is compared with the ensemble Kalman filter, which is thereby shown to be inaccurate in a variety of scenarios. Copyright © 2007 John Wiley & Sons, Ltd.

Received 25 April 2007; Revised 24 October 2007; Accepted 25 October 2007

KEY WORDS: data assimilation; Bayesian statistics; 3DVAR; 4DVAR; filtering; smoothing; stochastic PDEs

*Correspondence to: A. M. Stuart, University of Warwick, Mathematics Institute, Coventry CV4 7AL, U.K.
†E-mail: a.m.stuart@warwick.ac.uk

Contract/grant sponsor: ONR; contract/grant number: N00014-05-1-0791

1. INTRODUCTION

Data assimilation is concerned with the incorporation of observational data into mathematical models. It is essential to do so in any fields that are data rich and for which well-founded predictive mathematical models exist. Geophysical applications [1], the atmospheric sciences [2] and oceanography [3] provide important application areas of this type. Here, we adopt a Bayesian view of data assimilation in which prior information (background velocity field and model error) is combined with data to provide a posterior distribution [4].

We study time-dependent problems in which the desired unknown is either the initial condition (a function of space alone) or a time-dependent function (a function of both space and time) [5]. The desired posterior probability measure is formulated on function space, without resorting to discretization in space or time. On the assumption that observational and model error statistics are known, this posterior distribution provides the optimal solution to the assimilation of space–time distributed data into a model. The approach is statistical and the optimal solution obtained is a probability distribution on the relevant class of functions (initial conditions or time-dependent solutions). Sampling from this probability distribution thus yields a representative *ensemble* of solutions. The approach introduced is a useful one for three main reasons. First it clarifies the notion of what is the optimal solution, thereby providing a benchmark against which existing approaches can be evaluated. Secondly, it provides a framework for the development of new methods for the creation of ensembles of solutions to the model, incorporating the available data in an optimal fashion; for problems where the posterior distribution is far from Gaussian, such new methods are very much required. Thirdly, by formulating the problem in function space, before discretization, a clear mathematical view of the subject is obtained, and the flexibility of using different discretization techniques for different parts of any sampling algorithm allows for optimal algorithm design.

In Section 2 we outline the general framework in which we will discuss data assimilation. Section 3 describes the optimization approaches of 3DVAR, 4DVAR and weak constraint 4DVAR. Statistical analogues of these approaches are introduced in Section 4 and related to the notions of filtering and smoothing from the signal processing literature. The optimization approaches are shown to give rise to maximum *a posteriori* estimators for these statistical approaches (the analogue of maximum likelihood estimators when a prior distribution is incorporated [4].) We show how Lagrangian data can be viewed in a general framework, subsuming both Eulerian and Lagrangian data assimilation, in Section 5. Section 6 contains two examples, based on the statistical analogues of 4DVAR and weak constraint 4DVAR. We summarize in Section 7. The majority of the material in Sections 2–5 constitutes a review, setting the context for our recent research, which is overviewed in Section 6, and where references to relevant publications are given.

In the following we use $|\cdot|$ to denote the standard finite dimensional Euclidean norm, and $|\cdot|_A = |A^{-1/2}\cdot|$ for any symmetric positive-definite matrix A. Likewise we use $\|\cdot\|$ to denote the standard L^2-norm on functions, and $\|\cdot\|_A = \|A^{-1/2}\cdot\|$ for any symmetric positive-definite operator A. We will mainly use these weighted norms with A being a covariance matrix or operator, and we will largely follow the notational conventions for such covariance matrices established in [6]. We will also use other conventions from that paper, such as the use of h (and H, $\mathscr{H}$) for observation functions and y for observations. We use the letter v to denote a velocity field, the letter z to denote passive tracer positions and the subscript 0 to denote initial conditions.

DOI: 10.1002/fld

2. DATA ASSIMILATION

2.1. The model

In the context of models from fluid mechanics we consider the problem of finding, given observations, the velocity field $v(x,t)$ for a partial differential equation of the form[‡]

$$\frac{\partial v}{\partial t} = F(v) + \eta$$
$$v(x,0) = v_0(x)$$

where η is some noise process. We start by considering the perfect model scenario where there is no noise, and the objective is to find the optimal initial velocity field $v_0(x)$ in the model

$$\begin{aligned} \frac{\partial v}{\partial t} &= F(v) \\ v(x,0) &= v_0(x) \end{aligned} \tag{1}$$

We return to the noisy case later in the paper.

2.2. The observations

We assume that we are given data in the form of observations (direct or indirect) of the velocity field $v(x,t)$. The objective of data assimilation is to find an optimal trade-off between the information available in the data and in the model. We say that the observations are *Eulerian* if they are of the velocity field itself and *Lagrangian* if they are of particles transported by the velocity field. In both cases the observations are at times $t_k \in [0,T]$, $k=1,\ldots,K$.

In the Eulerian case the observations are

$$\{y_{j,k} = h(v(x_j,t_k)) + \text{noise}\}, \quad j=1,\ldots,J \quad \text{and} \quad k=1,\ldots,K$$

The noise model can have various forms, but is assumed to be known. In the Gaussian case, assuming that correlations across space and time are known, noting that v_0 determines v uniquely, and concatenating the data, we may express

$$y = H(v_0) + \sqrt{R}\xi \tag{2}$$

where ξ is a standard Gaussian vector and R the covariance matrix. When model error is present it will be useful to view the observations as a function of v and to express

$$y = H(v) + \sqrt{R}\xi \tag{3}$$

in place of (2).

In the Lagrangian case we have

$$\{y_{j,k} = z_j(t_k) + \text{noise}\}, \quad j=1,\ldots,J \quad \text{and} \quad k=1,\ldots,K$$

[‡]We are rather loose here, and the notation is meant to incorporate a range of problems including the incompressible Navier–Stokes equation, shallow-water models, atmospheric models or ocean models.

DOI: 10.1002/fld

where

$$\frac{\mathrm{d}z_j}{\mathrm{d}t}(t) = v(z_j, t), \quad z_j(0) = z_{j,0}$$

The noise model can have various forms, but is assumed to be known. Note that the initial conditions for z_j, together with the initial condition for v, uniquely determines z_j at later times. Thus, in the Gaussian case, assuming correlations across space and time are known, and concatenating the data, we may express

$$y = \mathscr{H}(v_0, z_0) + \sqrt{R}\xi \tag{4}$$

where ξ is a standard Gaussian vector, $z_0 = (z_{1,0}, \ldots, z_{J,0})$ and R the covariance matrix.

3. 3DVAR *VERSUS* 4DVAR

Here we describe the various variants of 3DVAR and 4DVAR, which underlie the statistical approaches to data assimilation outlined in the next section. In both this and the next section we confine our attention to the Eulerian case. The Lagrangian situation will be considered thereafter.

3.1. 3DVAR

This method simply incorporates observations of a velocity field at time $t = \tau$ into a current estimated (or background) state $v^\star(x, \tau)$ at time $t = \tau$ [7]. It thus corresponds to the special case $t_k \equiv \tau$ for all k. Define

$$\begin{aligned} J_3(v) &= \sum_{j=1}^{J} \frac{1}{2r_j} |h(v(x_j, \tau)) - y_j|^2 + \frac{1}{2} \|v(x, \tau) - v^\star(x, \tau)\|_B^2 \\ &= \frac{1}{2} |H(v) - y|_R^2 + \frac{1}{2} \|v(x, \tau) - v^\star(x, \tau)\|_B^2 \end{aligned} \tag{5}$$

where R and B are the covariance matrix/operator for the observations and background state. (In the first line we have assumed a diagonal form for R but this is not necessary.)

Now choose $\widehat{v}$ to minimize $J_3(v)$:

$$\widehat{v} = \operatorname*{argmin}_v J_3(v) \tag{6}$$

This constitutes 3DVAR and produces an improved state $\widehat{v}(x, \tau)$ at $t = \tau$, which incorporates observations into the current estimate from the model. Note that J_3 is quadratic if the observation operators h are linear.

3.2. 4DVAR

This method aims to incorporate data, concerning the velocity field, which is distributed in time on the interval $[0, T]$ [8]. This is used to improve the current estimate of the initial velocity field

DOI: 10.1002/fld

$v_0(x) = v(x,0)$. Specifically, given a background initial state $v_0^\star(x)$, we define

$$J_4(v_0) = \sum_{j=1,k=1}^{J,K} \frac{1}{2r_{j,k}} |h(v(x_j,t_k)) - y_{j,k}|^2 + \frac{1}{2}\|v_0(x) - v_0^\star(x)\|_B^2$$

$$= \frac{1}{2}|H(v_0) - y|_R^2 + \frac{1}{2}\|v_0(x) - v_0^\star(x)\|_B^2 \tag{7}$$

(In the first line we have again assumed a diagonal form for R but this is not necessary.) Here $v(x,t)$ is the velocity field with initial state $v_0(x)$; that is, the solution of (1). This is sometimes termed a *hard constraint*: it is assumed that the model dynamics are obeyed exactly. Note that, even if the observation operator h is linear, the functional J_4 is not quadratic unless the dynamics of (1) are also linear.

We choose $\widehat{v}_0$ to minimize $J_4(v_0)$:

$$\widehat{v}_0 = \operatorname*{argmin}_{v_0} J_4(v_0) \tag{8}$$

This method can be varied so that, for example, the background information consists not only of the initial velocity field $v_0^\star$ but also $v^\star(x,t)$, the velocity field at later times t.

Note that 4DVAR is considerably more complex than 3DVAR because the function H depends on v_0 through the solution v of (1). For this reason it is hard to use 4DVAR on systems that are sensitive to initial conditions (for example chaotic) and over long time intervals compared with the typical separation time of trajectories. In this situation weak constraint 4DVAR is more natural.

3.3. 4DVAR (weak constraint)

This method is similar to 4DVAR except that the model dynamics of (1) is now no longer incorporated as a hard constraint. Instead satisfaction of the model dynamics is imposed weakly through an additional term in the cost function to be minimized [9]. Specifically we define

$$J_w(v) = \sum_{j=1,k=1}^{J,K} \frac{1}{2r_{j,k}} |h(v(x_j,t_k)) - y_{j,k}| + \frac{1}{2}\|v_0(x) - v_0^\star(x)\|_B^2$$

$$+ \frac{1}{2}\int_0^T \left\| \frac{\partial v}{\partial t}(x,t) - F(v(x,t)) \right\|_Q^2 \mathrm{d}t$$

$$= \frac{1}{2}|H(v) - y|_R^2 + \frac{1}{2}\|v_0(x) - v_0^\star(x)\|_B^2$$

$$+ \frac{1}{2}\int_0^T \left\| \frac{\partial v}{\partial t}(x,t) - F(v(x,t)) \right\|_Q^2 \mathrm{d}t$$

Here Q is a covariance matrix that quantifies the level of confidence in the model equations.

We choose $\widehat{v}$ to minimize $J_w(v)$:

$$\widehat{v} = \operatorname*{argmin}_{v} J_w(v) \tag{9}$$

DOI: 10.1002/fld

This minimization is now more complex than for 4DVAR as it involves finding an entire approximate trajectory $\{v(x,t)\}_{t\in[0,T]}$ of (1), not just an initial condition. (In other words a function of space–time, not just of space.) However, for reasons detailed at the end of the last subsection, it is desirable to impose the weak constraint when the dynamics is sensitive to initial conditions and long time intervals, and the additional complexity is thus sometimes necessary.

Various variants are possible concerning the manner in which the weak constraint is imposed. For instance the cost function above corresponds to an error model that is uncorrelated in time; it is possible (and indeed sometimes natural) to add time–correlation information. Furthermore, the background state may be distributed in time, not just on the initial conditions.

4. STATISTICAL PERSPECTIVE

The perspective in the previous section is to pose data assimilation as an optimization problem to estimate the best possible velocity field. Instead, since the observations, and possibly the model itself, are subject to noise, our statements about the velocity field also have a natural probabilistic interpretation [4]. This leads to a Bayesian perspective on data assimilation in which observations are used to convert a prior distribution on velocity fields into a posterior. See [10, 11], for example, in the context of the atmospheric sciences and, in the context of applications to oil reservoir simulation, see [1].

4.1. 3DVAR and filtering

We may take a probabilistic view of the problem by sampling from the probability density function (pdf) for $v(x,\tau)$ proportional to

$$\exp(-J_3(v)) \tag{10}$$

The background $v^\star(x,\tau)$ is the mean of a prior Gaussian distribution with covariance B. The posterior probability density given by (10) is found by applying Bayes rule and incorporating the observations, assuming that the error in them is Gaussian with covariance R. The posterior is Gaussian only if H is linear.

The velocity field $\widehat{v}$ found from (6) is the maximum *a posteriori* estimator. If the updated pdf of the velocity field is updated sequentially with $\tau : t_k \mapsto t_{k+1}$ then we obtain a *filter*: a method that alternates between model updates in time and incorporation of data. Assuming that the prior is Gaussian, in general, an approximation since the underlying Liouville equation that propagates the density between t_k and t_{k+1} will not preserve Gaussianity unless the dynamics is linear. Making the Gaussian approximation leads to Kalman filters and their variants such as the extended Kalman filter and the ensemble Kalman filter (EnKF). Non-Gaussian problems are typically approximated via particle filters [12].

4.2. 4DVAR and smoothing

In the context of 4DVAR we may also take a statistical perspective. We view observations as noisy and, hence, the initial condition is only known to us probabilistically. We sample from the pdf for $v_0(x)$ proportional to

$$\exp(-J_4(v_0)) \tag{11}$$

DOI: 10.1002/fld

The background $v_0^\star(x)$ is the mean of a prior Gaussian distribution with covariance B. The posterior probability density given by (11) is found by applying Bayes rule and incorporating the observations, assuming that the error in them is Gaussian with covariance R. The posterior distribution on u is non-Gaussian unless the dynamics of (1) and the observation operator are linear. The velocity field $\widehat{v}_0$ found from (8) is the maximum *a posteriori* estimator.

Unlike the previous subsection, there is no efficient sequential update available here: the posterior pdf on the initial data depends on data at all $\{t_k\}_{k=1}^K$ from $[0, T]$. This is referred to as smoothing rather than filtering. See [5] for a perspective on this version of smoothing as a form of data assimilation, and methods for sampling from the posterior distribution. If the distribution on v_0 is pushed forward to final time t_K then the resulting distribution on v at time t_K agrees with the filtering distribution calculated recursively as outlined at the end of the previous subsection. Thus, an accurate sampling of the smoothing distribution can be used to benchmark various *approximate* filters such the extended and EnKFs.

4.3. 4DVAR (weak constraint) and smoothing

We may also consider weak constraint 4DVAR as the basis for a statistical viewpoint in which we have a pdf for the solution $v(x, t)$. In this context we no longer have the model dynamics (1) but rather the *stochastic dynamics* given by

$$\begin{aligned} \frac{\partial v}{\partial t} &= F(v) + \sqrt{Q}\frac{\partial W}{\partial t} \\ v(x, 0) &= u(x) \end{aligned} \tag{12}$$

where $\partial W/\partial t$ is a space–time white noise and Q is the covariance of the noise in space.

The posterior pdf for $v(x, t)$ is now proportional to

$$\exp(-J_w(v)) \tag{13}$$

Again $v_0^\star(x)$ is the mean of a Gaussian prior on initial conditions. The model stochastic dynamics given by (12) defines a prior on the solution $v(x, t)$ trajectory. The posterior distribution is again, as for 4DVAR, non-Gaussian unless the dynamics and observations are linear.

The field $\widehat{v}(x, t)$ found from (9) is the maximum *a posteriori* estimator. To sample from the distribution (13), sophisticated sampling is required: there are boundary values in space and time. See [5] for a perspective on this version of smoothing as a form of data assimilation, and methods for sampling from the posterior distribution.

5. LAGRANGIAN DATA

Here we show how the preceding optimization and statistical perspectives can be applied to the problem of assimilating Lagrangian data into models. We achieve this by extending the Eulerian set-up to incorporate Lagrangian data. The basic idea we outline is useful because, once the viewpoint is understood, it becomes clear that, mathematically, Eulerian and Lagrangian data assimilations are both specific cases of a single framework concerning the assimilation of data into a dynamical system and, in principle, Lagrangian data assimilation may be tackled by all the methods we have already outlined for the Eulerian case. However, in practice of course, the structure of the posterior

DOI: 10.1002/fld

distributions may be affected considerably by the type of observations. The papers [13, 14] were the first to extend data assimilation to Lagrangian data in a systematic fashion.

5.1. The problem

The aim is to find $v_0(x)$ the initial velocity field and $z_{j,0}$ the initial particle positions satisfying

$$\frac{\partial v}{\partial t} = F(v)$$

$$v(x,0) = v_0(x)$$

$$\frac{\mathrm{d}z_j}{\mathrm{d}t} = v(z_j,t), \quad j=1,\ldots,J$$

$$z_j(0) = z_{j,0}, \quad j=1,\ldots,J$$

We observe

$$\{y_{j,k} = z_j(t_k) + \text{noise}\}, \quad j=1,\ldots,J \quad \text{and} \quad k=1,\ldots,K$$

5.2. Lagrangian data assimilation as standard data assimilation

We concatenate $z=(z_1,\ldots,z_J)$ and $z_0=(z_{1,0},\ldots,z_{J,0})$ and define the observation function $h(v,z)=z$. The data assimilation problem then looks identical to the Eulerian case, extended from a dynamical model for the velocity field v alone to a dynamical model for the pair v,z, and with a particular observation function that corresponds to projection onto particle positions z. Everything that we have said about Eulerian data assimilation may now be generalized to this case [13, 14].

For expository purposes let us consider an analogue of 4DVAR for this Lagrangian problem. Note that all particle positions may be viewed as functions of the initial velocity field and the initial particle positions. Define

$$\begin{aligned} J_l(v_0,z_0) &= \sum_{j=1,k=1}^{J,K} \frac{1}{2r_{i,j}} |z_j(t_k) - y_{j,k}|^2 + \frac{1}{2}\|v_0(x) - v_0^\star(x)\|_B^2 \\ &\quad + \sum_{j=1}^{J} \frac{1}{2\omega_j} |z_{j,0} - z_{j,0}^\star|^2 \\ &= \frac{1}{2}|\mathscr{H}(v_0,z_0) - y|_R^2 + \frac{1}{2}\|v_0(x) - v_0^\star(x)\|_B^2 + \frac{1}{2}|z_0 - z_0^\star|_\Omega^2 \end{aligned} \tag{14}$$

Here both v_0 and z_0 are assumed to have background values $v_0^\star, z_0^\star$, and B (resp. Ω) quantifies the uncertainty in the former (resp. latter). Then the analogue of 4DVAR consists of solving the following minimization problem:

$$(\widehat{v}_0, \widehat{z}_0) = \operatorname*{argmin}_{v_0,z_0} J_l(v_0,z_0) \tag{15}$$

The statistical analogue is to sample from

$$\exp(-J_l(v_0,z_0)) \tag{16}$$

DOI: 10.1002/fld

Again we have assumed that the covariance in the observations is diagonal, but this may be relaxed. We may also incorporate model error into both the evolution of the velocity field and the evolution of the passive tracers.

6. APPLICATIONS

We give two examples of the statistical variant on 4DVAR, both in the Lagrangian context. The first corresponds to the perfect model scenario of (1) [15] and the second to a situation where the model is imposed as a weak constraint which, in the statistical viewpoint, corresponds to a model of the form (12), driven by noise [5].

6.1. Perfect model scenario

We are interested in finding Fourier coefficients of $(v,h)|_{t=0}$ the initial conditions for the linearized shallow-water equations:

$$\frac{\partial v}{\partial t} = Jv - \nabla h, \quad (x,t) \in \Omega \times [0,\infty)$$

$$\frac{\partial h}{\partial t} = -\nabla \cdot v, \quad (x,t) \in \Omega \times [0,\infty)$$

Here Ω is the unit square and J is a skew-symmetric matrix. We impose periodic boundary conditions on v and h.

We assume that we are given observations $y_{j,k} = z_j(t_k) + \zeta_{j,k}$ of the passive tracers

$$\frac{\mathrm{d}z_j}{\mathrm{d}u} = v(z_j, u)$$

where $z_j \in \mathbb{R}^2$. Here the noise is Gaussian with mean zero. Figure 1 shows a typical flow field and the trajectories of three passive tracers.

The posterior distribution function given in (16) was sampled using five different Monte-Carlo Markov Chain (MCMC) methods of Metropolis–Hastings (MH) type, based on Langevin (MALA) and random walk (RWMH) proposals, in one case using adaptive preconditioning based on learning the covariance structure during the course of the computation. MCMC is a methodology whereby a given target probability distribution is sampled by constructing a Markov chain for which the target is invariant. Such a Markov chain can be constructed by taking a given Markov chain that is easy to sample, and accepting or rejecting proposals from this chain according to the MH criterion [4]. Assuming that the resulting Markov chain is ergodic, the time series from it will have a histogram that converges to the desired target distribution.

The results from these sampling methods are shown in Figure 2. The first two columns show samples from the posterior of two components of the velocity field, and the third column the posterior on the first coordinate of the particle position. A detailed description and comparison of the different methods may be found in [15]. For the purposes of this short paper, it suffices to note that the final row, which uses an adaptive MCMC method, may be viewed as providing the *exact* posterior distribution, and does so in the most efficient fashion. We will now compare such exact posteriors for the smoothing distribution with the output of some frequently used Kalman-based filters, all compared at the final time t_K.

DOI: 10.1002/fld

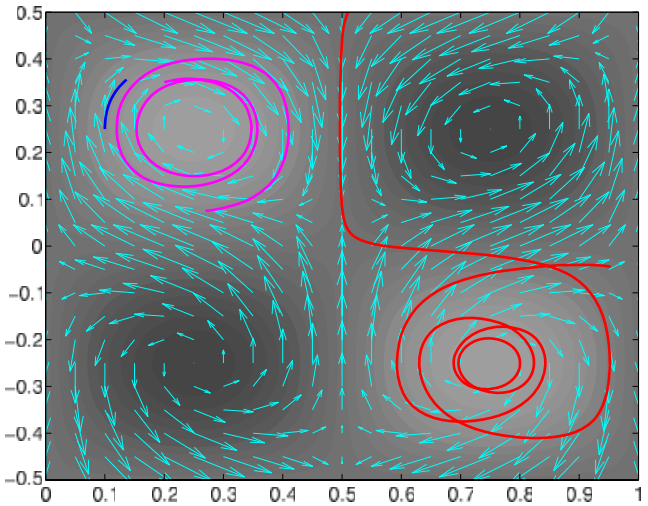

Figure 1. Snapshot of flowfield and particle trajectories.

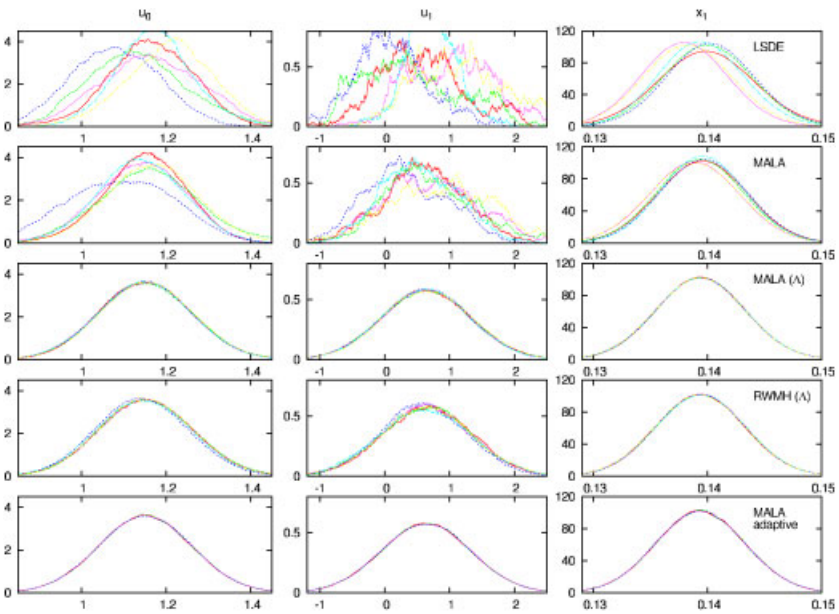

Figure 2. Five different sampling methods for the posterior. The last row may be viewed as giving the 'exact' posterior.

Figure 3 shows a comparison with the exact posterior on the coordinates of the observed particle in blue (found by using resolved samples from the adaptive MALA method) and its approximation by the EnKF algorithm [16] in green. The true solution is marked with a blue

DOI: 10.1002/fld

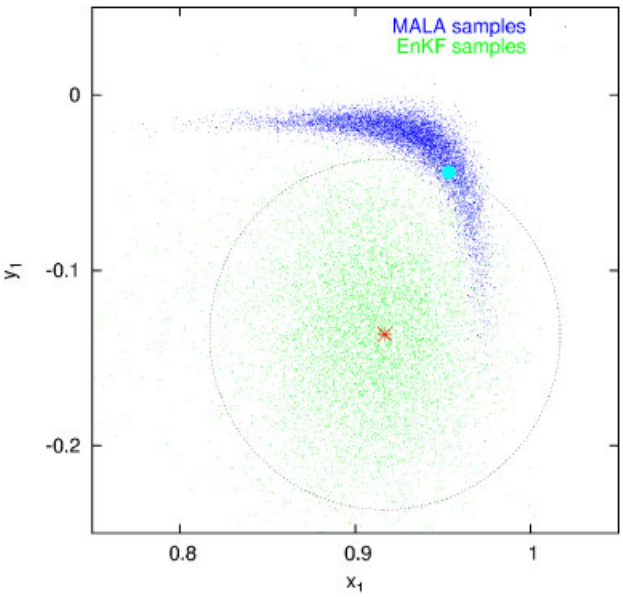

Figure 3. EnKF (green) and MALA (blue) approximations of the true posterior.

circle and the observation by a red asterix; the ellipse is of size two (observational) standard deviations around the asterix. We see that in this highly nonlinear dynamical model, the EnKF fails to accurately approximate the exact posterior distribution, due to the inappropriate Gaussian assumptions underlying it. A detailed discussion of this issue may be found in [15].

6.2. *Model error*

Now consider a problem with noise: to find (v, h) solving the linearized noisy shallow-water equations:

$$\frac{\partial v}{\partial t} = Jv - \nabla h - Q\gamma v + \sqrt{\frac{2\gamma Q}{\beta}}\xi, \quad (x,t) \in \Omega \times [0,\infty)$$

$$\frac{\partial h}{\partial t} = -\nabla \cdot v, \quad (x,t) \in \Omega \times [0,\infty)$$

given continuous time observation of passive tracers:

$$\frac{\mathrm{d}z_j}{\mathrm{d}u} = v(z_j, u) + \zeta_j$$

Again Ω is a unit square and we impose periodic boundary conditions on v and h; the particles $z_j \in \mathbb{R}^2$. Here ξ is a space–time white noise and the ζ_j are time white noise. The operator Q induces spatial correlations in the noise and it is natural to choose it so that it has constants in its null space, thereby ensuring preservation of the mean velocity field under the dynamics. (However,

DOI: 10.1002/fld

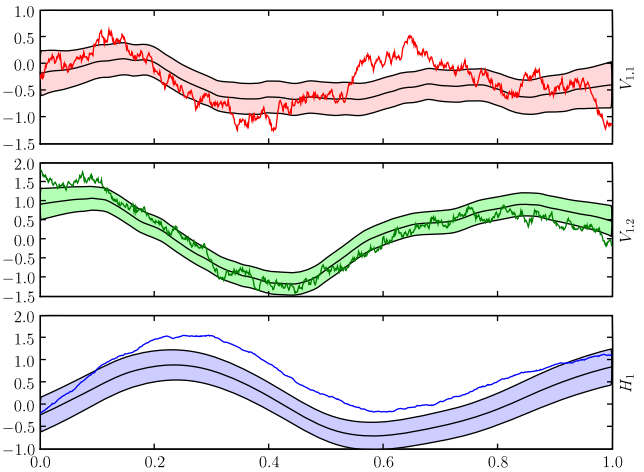

Figure 4. Reconstruction of x_i; five tracers are used.

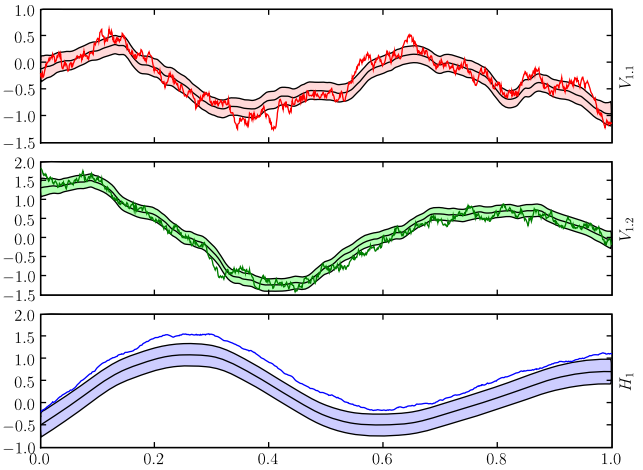

Figure 5. Reconstruction of x_i; 50 tracers are used.

in practice, we add a small amount of noise to the first Fourier coefficient and to the equation for the height field h.)

Figures 4–6 show reconstruction of the first two Fourier coefficients of v (the first two panels in each figure) and the first Fourier coefficient of the height field (the third panel in each figure). The three figures correspond to 5, 50 and 500 tracer particles, respectively. The bands represent one standard deviation about the mean of the posterior distribution, and the non-smooth curves the underlying exact signal, or 'truth'. Note that the velocity Fourier coefficients are well reconstructed for large numbers of particles; uncertainty remains in the height field, however, because the Lagrangian tracers do not probe it directly and because we include a small amount of noise in

DOI: 10.1002/fld

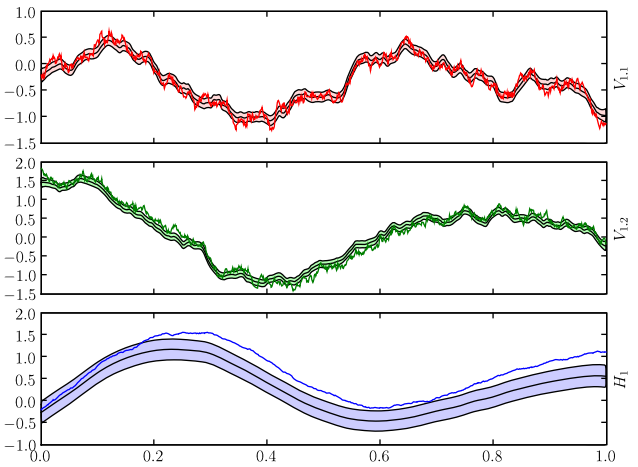

Figure 6. Reconstruction of x_i; 500 tracers are used.

the equation for its evolution. For more detailed discussion of the methods employed to find these distributions see [5].

7. CONCLUSIONS

In this paper we have highlighted the following well-known points regarding data assimilation:

- 3DVAR and 4DVAR are minimization techniques that differ through whether time-distributed data are incorporated into the cost function.
- 4DVAR and 4DVAR (weak) differ through whether the model is imposed exactly; in the latter case, error in the satisfaction of the dynamical equations is incorporated as part of the cost function.
- Adopting a Bayesian viewpoint shows that all of these variational methods have natural statistical analogues: filtering and smoothing. The variational methods find a maximum *a posteriori* estimator—the analogue of maximum likelihood estimators when a prior distribution is incorporated [4], typically as a regularizer.
- Lagrangian data assimilation can be framed as a generalization of the standard case of Eulerian data assimilation; thus, there are natural analogues of 3DVAR, 4DVAR and 4DVAR (weak) for the Lagrangian case.

The main new ideas that we have highlighted in this paper are as follows:

- It is insightful to formulate (smoothing) Bayesian data assimilation problems on function space, without discretizing in space and/or time; this allows for a clearer understanding of the mathematical structure, and allows discretizations to be optimized for the purposes of sampling, once a probability measure on function space is defined.

DOI: 10.1002/fld

- The full power of MCMC methods should be brought to bear on sampling these Bayesian (smoothing) posterior distributions arising in data assimilation. This allows for the calculation of the 'right' answer and hence for the evaluation of various approximations.
- Approximate filters, such as the EnKF, can behave poorly; we illustrated this fact on a highly non-Gaussian problem arising in Lagrangian data assimilation.

The primary challenges arising in this area are as follows:

- Sampling function space is extremely costly. (Typical discretizations of function space in weather forecasting currently involve $\mathcal{O}(10^7)$ unknowns at each instance in time.) Carrying out fully resolved MCMC simulations in this context is currently out of the question without new ideas. However, it may be possible to marry some of the current methods used to make 4DVAR efficient, such as adjoint methods, low-rank approximations and so forth, with MCMC proposals in such a fashion that useful ensemble information can be obtained efficiently and in the context of highly non-Gaussian posterior distributions. Carrying out a research program that effects this would be extremely valuable.
- An alternative to the Bayesian (smoothing) techniques that we use here and that, in principle, capture the correct posterior, is the use of particle filters. These are very effective in low dimensions, but suffer from severe computational problems in high dimensions. Understanding the relative merits of attacking the smoothing problem by MCMC methods, and the use of particle filters, provides an important research area in the study of high-(infinite) dimensional data assimilation problems.

REFERENCES

1. Farmer CL. Bayesian field theory applied to scattered data interpolation and inverse problems. In *Algorithms for Approximation*, Iske A, Leveseley J (eds). Springer: Berlin, 2007; 147–166.
2. Kalnay E. *Atmospheric Modeling*, *Data Assimilation and Predictability*. Cambridge University Press: Cambridge, 2003.
3. Bennett AF. *Inverse Modeling of the Ocean and Atmosphere*. Cambridge University Press: Cambridge, 2002.
4. Kaipio J, Somersalo E. *Statistical and Computational Inverse Problems*. Springer: Berlin, 2004.
5. Apte A, Hairer M, Stuart AM, Voss J. Sampling the posterior: an approach to non-Gaussian data assimilation. *Physica D* 2007; **230**:50–64.
6. Ide K, Courtier P, Ghil M, Lorenc AC. Unified notation for data assimilation: operational, sequential and variational. *Journal of the Meteorological Society of Japan* 1997; **75**:181–189.
7. Courtier P, Anderson E, Heckley W, Pailleux J, Vasiljevic D, Hamrud M, Hollingworth A, Rabier F, Fisher M. The ECMWF implementation of three-dimensional variational assimilation (3d-var). *Quarterly Journal of the Royal Meteorological Society* 1998; **124**:1783–1808.
8. Lorenc AC. Analysis methods for numerical weather prediction. *Quarterly Journal of the Royal Meteorological Society* 1986; **112**:1177–1194.
9. Derber JC. A variational continuous assimilation technique. *Monthly Weather Review* 1989; **117**:2437–2446.
10. Sneddon G. A statistical perspective on data assimilation in numerical models. *Studies in the Atmospheric Sciences*. *Lecture Notes in Statistics*, vol. 144. Springer: Berlin, 2000; 7–21.
11. Berliner LM. Monte Carlo based ensemble forecasting. *Statistics and Computing* 2001; **11**:269–275.
12. Doucet A, DeFreitas N, Gordon N. *Sequential Monte Carlo Methods in Practice*. Springer: Berlin, 2003.
13. Ide K, Kuznetsov L, Jones CKRT. Lagrangian data assimilation for point–vortex system. *Journal of Turbulence* 2002; **3**:053.
14. Kuznetsov L, Ide K, Jones CKRT. A method for assimilation of Lagrangian data. *Monthly Weather Review* 2003; **131**:2247–2260.
15. Apte A, Jones CKRT, Stuart AM. A Bayesian approach to Lagrangian data assimilation. *Tellus* 2007; submitted.
16. Evensen G. The ensemble Kalman filter: theoretical formulation and practical implementation. *Ocean Dynamics* 2003; **53**:343–367.

DOI: 10.1002/fld