Available online at www.sciencedirect.com

ScienceDirect

ELSEVIER

Physica D 230 (2007) 50–64

PHYSICA D

www.elsevier.com/locate/physd

Sampling the posterior: An approach to non-Gaussian data assimilation

A. Apte[a], M. Hairer[b], A.M. Stuart[b,*], J. Voss[b]

[a] Department of Mathematics, University of North Carolina, CB 3250 Phillips Hall Chapel Hill, NC 27599-3250, USA
[b] Mathematics Institute, University of Warwick, Coventry CV4 7AL, UK

Available online 8 August 2006

Abstract

The viewpoint taken in this paper is that data assimilation is fundamentally a statistical problem and that this problem should be cast in a Bayesian framework. In the absence of model error, the correct solution to the data assimilation problem is to find the posterior distribution implied by this Bayesian setting. Methods for dealing with data assimilation should then be judged by their ability to probe this distribution. In this paper we propose a range of techniques for probing the posterior distribution, based around the Langevin equation; and we compare these new techniques with existing methods.

When the underlying dynamics is deterministic, the posterior distribution is on the space of initial conditions leading to a sampling problem over this space. When the underlying dynamics is stochastic the posterior distribution is on the space of continuous time paths. By writing down a density, and conditioning on observations, it is possible to define a range of Markov Chain Monte Carlo (MCMC) methods which sample from the desired posterior distribution, and thereby solve the data assimilation problem. The basic building-blocks for the MCMC methods that we concentrate on in this paper are Langevin equations which are ergodic and whose invariant measures give the desired distribution; in the case of path space sampling these are stochastic partial differential equations (SPDEs).

Two examples are given to show how data assimilation can be formulated in a Bayesian fashion. The first is weather prediction, and the second is Lagrangian data assimilation for oceanic velocity fields. Furthermore the relationship between the Bayesian approach outlined here and the commonly used Kalman filter based techniques, prevalent in practice, is discussed. Two simple pedagogical examples are studied to illustrate the application of Bayesian sampling to data assimilation concretely. Finally a range of open mathematical and computational issues, arising from the Bayesian approach, are outlined.

Keywords: Data assimilation; Bayesian filtering; Stochastic PDEs; Langevin equation

1. Introduction

In this paper we describe a Bayesian approach to data assimilation. The approach is based on sampling from the posterior distribution on the model, after data is assimilated. We believe that this viewpoint may be useful for two primary reasons: firstly the Bayesian approach gives, in some sense, the correct theoretical answer to the data assimilation problem and other approaches which have been adopted, such as ensemble Kalman filtering, should be evaluated by their ability to approximate the posterior distribution in the Bayesian approach [26]; secondly, for any data assimilation problems which are bimodal or multimodal, Kalman based methods will necessarily fail (see [21,35]) and it will be necessary to use a Bayesian approach, such as the one described here.

From a mathematical viewpoint the main interest in this paper stems from the fact that we formulate Bayesian data assimilation in the case where the underlying model dynamics is stochastic. The basic object to sample is then a continuous time *path* (time-dependent solution of a differential equation). In this context the key concept which needs elucidation is that of a probability density in the space of paths. Once this density is defined, and a conditional density is written down which incorporates observations, the complete Bayesian framework can be employed to sample in the space of continuous time paths.

The paper is organized as follows. In Section 2 we formulate a number of variants of the data assimilation problem abstractly in the language of stochastic differential equations (SDEs). We

* Corresponding author. Tel.: +44 24 7652 2685.
E-mail address: stuart@maths.warwick.ac.uk (A.M. Stuart).

0167-2789/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.physd.2006.06.009

give two concrete examples, arising in oceanic and atmospheric science, to motivate the abstract setting. Section 3 introduces the Bayesian approach to data assimilation in the context of deterministic dynamics, where the posterior distribution that we wish to sample is on the initial data; we introduce the Langevin equation to probe this distribution, and discuss related MCMC methods. Section 4 carries out a similar program in the case where the underlying dynamics is stochastic and the posterior distribution is on the space of paths; we introduces the central idea of probability density in path space. In Section 4.2 we describe a generalization of the Langevin equation to path space, leading to nonlinear parabolic stochastic PDEs (SPDEs) which, when statistically stationary, sample from the distribution which solves the data assimilation problem; we also look at a second order Langevin equation, leading to a nonlinear damped stochastic wave equation. Section 4.3 describes another sampling strategy that might be used to sample path space, namely a Hybrid Monte Carlo technique. In Section 5 we discuss MCMC methods in path space in general terms, discussing how Metropolis–Hastings ideas might be used to improve the Langevin and Hybrid methods from the previous section, and more generally to explore a wide range of sampling techniques. In Section 6 we relate the Bayesian approach adopted here to other commonly used methods of data assimilation. Section 7 contains a pedagogical example in the case where the underlying model is deterministic; comparisons are made between the Langevin approach and various Kalman based filters. Section 8 contains a pedagogical example of Lagrangian data assimilation, based on a Gaussian random field model of a velocity field, included to illustrate the Bayesian methodology in the context of path sampling. Section 9 concludes with a description of a number of open mathematical and computational questions arising from adopting our Bayesian viewpoint on data assimilation.

The SPDE based approach to sampling continuous time paths was introduced in [38] and is subsequently analyzed in [15,16], building on analysis in [40]. (For paths conditioned only on knowing the value of the path at two points in time – *bridges* – the SPDE based approach was simultaneously written down in [31].) The SPDE approach generalizes the Langevin equation to sampling in infinite dimensions. The Langevin approach to sampling in finite dimensions is outlined in the book [32] where it is shown how to use a discretization of the Langevin equation, in conjunction with a Metropolis–Hastings accept–reject criterion, to create a Markov Chain Monte Carlo (MCMC) method. The infinite dimensional version of this MCMC method, arising when sampling the space of paths, is studied in [4]. Hybrid Monte Carlo methods, which are widely used in molecular dynamics, were generalized to sample in path space in [1], as were Langevin based methods; however that paper proceeded by discretizing the evolution equations to be sampled and then applying a finite dimensional sampling method. It is our view that it is conceptually and algorithmically preferable to formulate the sampling problem in infinite dimensions (the space of paths). It is conceptually important to know that the infinite dimensional problem makes sense mathematically. Once this infinite dimensional problem is defined, it is algorithmically important to find an efficient way of approximating it by discretization. Discretizing first, so that the sampling problem is never written down in continuous time, and then sampling, may lead to a non-optimal approximation of the desired infinite dimensional problem; see the end of Section 4.

The subject of Brownian motion and stochastic calculus is described in [19], whilst texts on SDEs include [11,29]. The subject of SPDEs is covered in the text [8].

2. The framework

In this section we write down a precise mathematical framework into which a variety of data assimilation problems can be cast. We show how Lagrangian data assimilation can be expressed as a special case of the general framework and we also discuss the issue of model error. We then give two motivational examples, and express them precisely in the language of the chosen mathematical framework. We conclude with some technical assumptions and notation that will be used in the remainder of the paper.

2.1. Mathematical setting

Data assimilation may be viewed as a form of signal processing. The *signal* that we wish to determine, and into which we wish to assimilate observational data, is assumed to satisfy the SDE

$$\frac{\mathrm{d}x}{\mathrm{d}t} = f(x) + \gamma \frac{\mathrm{d}W_x}{\mathrm{d}t}, \tag{2.1}$$

where f determines the systematic part of the evolution, and $\mathrm{d}W_x/\mathrm{d}t$ is Gaussian white noise perturbing it. In the following we will distinguish between $\gamma = 0$ (ODE) and $\gamma \neq 0$ (SDE). In the former case the Bayesian framework requires sampling in the space of initial conditions $x(0)$ only; in the latter it requires sampling in the (infinite dimensional) space of paths $\{x(t)\}$. The model equation (2.1) may be viewed as a prior distribution on the space of paths. We assume that $x(0)$ has prior distribution with density ζ.

In Bayesian data assimilation the ultimate objective is to probe the posterior probability distribution on $x(0)$ (when $\gamma = 0$) or on $\{x(t)\}$ (when $\gamma \neq 0$), conditional on some form of *observation*. If the observation is in continuous time then we denote it by $y(t)$ and assume that it too satisfies an SDE. This has the form

$$\frac{\mathrm{d}y}{\mathrm{d}t} = g(x, y) + \sigma \frac{\mathrm{d}W_y}{\mathrm{d}t}, \tag{2.2}$$

where g determines the systematic evolution of the observation, which depends on the signal x, and $\mathrm{d}W_y/\mathrm{d}t$ is a standard Gaussian white noise perturbing it, independent of the white noise $\mathrm{d}W_x/\mathrm{d}t$.

If the observation is in discrete time then we assume that we observe $y = (y_1, \dots, y_K)$ satisfying

$$y_k = h_k(x(t_k)) + \sigma_k \xi_k, \quad k = 1, \dots, K. \tag{2.3}$$

Here h_k determines which function of the signal x is observed, the ξ_k are standard i.i.d. unit Gaussian random variables $\mathcal{N}(0, I)$ and the σ_k determine their covariances; both the h_k and σ_k are indexed by k because the nature of the observations may differ at different times. We assume that the ξ_k are independent of the white noise driving (2.1), and of $x(0)$. The times $\{t_k\}$ are ordered and assumed to satisfy

$$0 < t_1 < t_2 < \cdots < t_K \le T.$$

Any observation at $t = 0$ is incorporated into ζ.

2.2. Lagrangian data assimilation

Lagrangian data assimilation arises frequently in the oceanic sciences where observations about a fluid velocity field are frequently given in terms of particles advected by the field: Lagrangian information. This may be formulated as a special case of the preceding framework, as we now show; this approach to Lagrangian data assimilation, showing that it is a special case of the general set-up, first appears in the literature in [17,21]. There are subtle differences between the cases where the observations are in continuous and discrete time and whether $\gamma = 0$ or not.

We start with discrete time observations and consider the situation where $\gamma = 0$. Assume that the Lagrangian information about x is carried in z, where

$$\frac{\mathrm{d}z}{\mathrm{d}t} = g(x, z).$$

The observations are

$$y_k = h_k(z(t_k)) + \sigma_k \xi_k, \quad k = 1, \ldots, K.$$

By re-defining $x \mapsto (x, z)$, $f \mapsto (f, g)$ and the h_k we can formulate this as in the previous subsection for discrete time observations and $\gamma = 0$. An important point to notice is that part of the data assimilation process may involve sampling the initial data for the Lagrangian variables as well as for x. Hence the reason why the vector x is extended to incorporate z as well as x.

We now consider the case $\gamma \neq 0$ and again study discrete time observations. Assume that the Lagrangian information about x is carried in z, where

$$\frac{\mathrm{d}z}{\mathrm{d}t} = g(x, z) + \eta \frac{\mathrm{d}W_z}{\mathrm{d}t}.$$

The observations are

$$y_k = h(z(t_k)) + \sigma_k \xi_k, \quad k = 1, \ldots, K.$$

Again, by re-defining $x \mapsto (x, z)$, $f \to (f, g)$, $W_x \mapsto (W_x, W_z)$, $\gamma \mapsto (\gamma, \eta)$ and the h_k we can formulate this as in the previous subsection, now for discrete time observations and $\gamma \neq 0$. Since the Lagrangian data is in discrete time, but the Lagrangian variables evolve stochastically in continuous time, part of the data assimilation process involves sampling paths of the Lagrangian variables between the observations. This is the reason why the vector x is extended to incorporate z as well as x.

In the case where the Lagrangian information is carried in y, and y is a continuous time path satisfying Eq. (2.2), the observation is in continuous time. Hence this may be directly formulated as in the case of continuous time observations in the previous subsection, for both $\gamma = 0$ and $\gamma \neq 0$.

2.3. Model error

If the model error can be represented as Gaussian white noise in time then it is already clearly representable in the mathematical framework given by (2.1). However, typically, the precise nature of the model error would not be known; more precisely γ would be unknown. In this context the methods described in this paper would need to be extended to include sampling from the distribution on γ, given some prior information on it. This falls into the realm of parameter estimation, and is a natural extension of the Bayesian framework given here.

Of course model error may not be Gaussian white in time: it may include systematic non-random contributions, as well as noise which is time correlated. However, the framework given can be extended to cover such situations, and would require the estimation of parameters representing the form of the systematic model error, as well as the memory kernel for the noise; the latter will be most easily estimated if it is assumed to be exponentially decaying, since the model can then still be expressed in Markovian form.

2.4. Motivational examples

When discretized in space, a typical model for numerical weather prediction is an ODE system with dimension of order 10^8. In the absence of model error and external forcing, an equation of the form (2.1) is obtained, with $\gamma = 0$. In this context the state x represents the nodal values of the unknown quantities such as velocity, temperature, pressure and so forth. The observations which we wish to assimilate are then various projections of the state x, possibly different at different times, and may be viewed as subject to independent Gaussian white noises. We thus obtain observations y of the form (2.3).

A second motivational example is that of Lagrangian data assimilation in the ocean (see [21] for work in this direction). For expository purposes consider trying to make inference about a 2D velocity field governed by the noisy incompressible Navier–Stokes equations, by means of Lagrangian particle trajectories. If we assume periodicity in space then we may write the velocity field $v(y, t)$ as an (incompressible) trigonometric series

$$v(y, t) = \sum_{k \in \mathcal{K}} \mathrm{i}k^{\perp} x^k(t) \exp(\mathrm{i}k \cdot y).$$

The vector x made up of the x^k then satisfies an equation like (2.1). Now imagine a set of Lagrangian drifters, indexed by ℓ, and with positions $y_\ell(t)$ governed by

$$\frac{\mathrm{d}y_\ell}{\mathrm{d}t} = v(y_\ell, t) + \sigma_\ell \frac{\mathrm{d}W_\ell}{\mathrm{d}t}.$$

From the representation of the velocity field it is clear that

$$v(y,t)=\chi(x(t),y)$$

for some function χ linear in x and hence that the collection of Lagrangian drifters satisfy an equation of the form (2.2), with $g(x,y)$ found by concatenating the $\chi(x,y_\ell)$ over each drifter y_ℓ. If data from the drifters (obtained by GPS for example) is assumed to be essentially continuous in time then we may view (2.2) as giving the observational data y which is to be assimilated. (As mentioned above it is also possible to formulate Lagrangian data assimilation in the case where the drifters are observed only at discrete times.)

2.5. *Assumptions and notation*

In Eq. (2.1) we have $f:\mathbb{R}^d\to\mathbb{R}^d$, $\gamma\in\mathbb{R}^{d\times d}$ and W_x is standard d-dimensional Brownian motion. We assume either that $\gamma=0$, or that γ is invertible and we define $\Gamma=\gamma\gamma^{\mathrm{T}}$. In Eq. (2.2) we have $g:\mathbb{R}^d\times\mathbb{R}^m\to\mathbb{R}^m$, $\sigma\in\mathbb{R}^{m\times m}$ and W_y is standard m-dimensional Brownian motion, independent of W_x. We assume that σ is invertible and we define $\Sigma=\sigma\sigma^{\mathrm{T}}$. In Eq. (2.3) we have $h_k:\mathbb{R}^d\to\mathbb{R}^m$ and $\sigma_k\in\mathbb{R}^{m\times m}$. The ξ_k are assumed independent of W_x. We also assume that σ_k is invertible and define $\Sigma_k=\sigma_k\sigma_k^{\mathrm{T}}$.

For any positive definite $n\times n$ covariance matrix A we define the inner product on $\mathbb{R}^n$ given by

$$\langle a,b\rangle_A=a^{\mathrm{T}}A^{-1}b$$

and the induced norm $\|\cdot\|_A^2=\langle\cdot,\cdot\rangle_A$. This notation is used extensively in the following sections, with A equal to Γ, Σ or Σ_j, and also with $A=R$ where R is a covariance matrix formed by concatenating the discrete time observations into a single vector.

3. Initial data sampling and SDEs

We start by considering the case where $\gamma=0$ so that the posterior distribution to be sampled is on the initial data, and is finite dimensional.

3.1. *Density on initial conditions*

The dynamics are governed by

$$\frac{\mathrm{d}x}{\mathrm{d}t}=f(x),\quad x(0)=x_0\sim\zeta$$

and we use the solution operator for the dynamics to write

$$x(t)=\Phi(x_0;t). \tag{3.1}$$

We observe $h(x(t))$ at discrete times, subject to independent noises, and write (2.3) as

$$y_k=h(x(t_k))+\eta_k,\quad \eta_k\sim\mathcal{N}(0,\Sigma_k).$$

If we define

$$y^{\mathrm{T}}=\{y_k^{\mathrm{T}}\}_{k=1}^K,\quad \eta^{\mathrm{T}}=\{\eta_k^{\mathrm{T}}\}_{k=1}^K,$$
$$H(x_0)^{\mathrm{T}}=\{h(\Phi(x_0;t_k))^{\mathrm{T}}\}_{k=1}^K$$

and let R be the covariance matrix of the Gaussian random variable η, then

$$y=H(x_0)+\eta,\quad \eta\sim\mathcal{N}(0,R).$$

From this we find the pdf for the joint random variable (x_0,y) by first conditioning on x_0 and then multiplying by the prior on x_0. Define

$$J(x_0,y)=\frac{1}{2}\|y-H(x_0)\|_R^2.$$

Then the pdf for (x_0,y) is

$$\rho(x_0,y)\propto\zeta(x_0)\exp(-J(x_0,y)). \tag{3.2}$$

By Bayes' rule

$$\rho(x_0|\,y)\propto\rho(x_0,y)$$

with constant of proportionality depending only on y. Hence we may use the expression (3.2) as the basis for sampling x_0 given y, in any method which requires the pdf only up to a multiplicative constant. We discuss such methods in the next two subsections.

It is worth noting at this point that the commonly adopted *4DVAR* approach [24] corresponds to choosing x_0 to maximize $\rho(x_0|\,y)$. It may hence be viewed as a maximum likelihood method for determination of x_0. If the random variable $x_0|\,y$ is Gaussian and has small variance then this is natural. But if the random variable is far from Gaussian with small variance, for example if it is bimodal, 4DVAR clearly comprises an ineffective way to probe the posterior distribution on x_0 given observations y. It is thus of interest to understand the structure of the posterior distribution in order to know whether 4DVAR is a useful approach. The structure of the posterior distribution depends in a complicated way on the underlying dynamics of x, as well as the nature and number of the observations.

3.2. *Langevin SDE*

Sampling from the distribution of $x_0|\,y$ can be achieved by, amongst many possibilities, solving the *Langevin equation*. This is simply

$$\frac{\mathrm{d}x_0}{\mathrm{d}s}=\nabla_{x_0}\ln\rho(x_0|\,y)+\sqrt{2}\frac{\mathrm{d}W}{\mathrm{d}s}. \tag{3.3}$$

This equation has $\rho(x_0|\,y)$ as an invariant density and is ergodic under mild assumptions on ρ.[1] Hence the empirical measure (histogram) generated by a single solution path over a long time interval will approximate the desired posterior density. More precisely, under the assumption of ergodicity, we will have

$$\lim_{S\to\infty}\frac{1}{S}\int_0^S\phi(x_0(s))\mathrm{d}s\to\int_{\mathbb{R}^d}\phi(x)\rho(x|\,y)\mathrm{d}x, \tag{3.4}$$

for functions ϕ of the initial distribution. In practice the limit $S=\infty$ cannot be obtained, but a single numerical trajectory

[1] For the equation to be well defined the conditional density needs to be differentiable in x_0; if it is not then more care is required in defining the Langevin equation.

of (3.3) over a long time interval $s \in [0, S]$ can be used to approximate the desired target density. Notice that the time-like variable s is an artificial *algorithmic time* introduced to facilitate sampling from the desired density.

From (3.2) we see that the Langevin equation becomes

$$\frac{\mathrm{d}x_0}{\mathrm{d}s} = \nabla_{x_0} \ln \zeta(x_0) - \nabla_{x_0} J(x_0, y) + \sqrt{2}\frac{\mathrm{d}W}{\mathrm{d}s}.$$

Here

$$\nabla_{x_0} J(x_0, y) = -\nabla_{x_0} H(x_0)^{\mathrm{T}} R^{-1}[y - H(x_0)]. \tag{3.5}$$

Notice that the operators H and $\nabla_{x_0} H$ are calculated (or approximated) for the implementation of 4DVAR. Thus this technology can be transported to numerical algorithms for the Langevin equation arising in this context.

There are various generalizations of the Langevin equation that can also be useful for sampling — including the *second order Langevin equation* and *pre-conditioning*. We restrict discussion of these methods to the (infinite dimensional) context of sampling path space, described in Section 4.

3.3. Hybrid Monte Carlo

Another method that is successful in the context of sampling certain high dimensional probability distributions is *Hybrid Monte Carlo*. The starting point is the Hamiltonian system of equations

$$\frac{\mathrm{d}^2 x_0}{\mathrm{d}s^2} = \nabla_{x_0} \ln \rho(x_0 | y). \tag{3.6}$$

This equation defines a solution operator

$$\mathcal{M}: \left(x_0(0), \frac{\mathrm{d}x_0}{\mathrm{d}s}(0)\right) \mapsto \left(x_0(S), \frac{\mathrm{d}x_0}{\mathrm{d}s}(S)\right)$$

mapping initial conditions to the solution at time S. With the notation

$$P_x : (x, y) \mapsto x$$

we construct the Markov chain

$$x^{n+1} = P_x \mathcal{M}(x^n, \xi^n)$$

where the ξ^n are chosen to be i.i.d. Gaussian random variables with distribution $\mathcal{N}(0, I)$. This Markov chain has $\rho(x_0| y)$ as an invariant density.

3.4. Continuous time observations

Now assume that the Lagrangian information is carried in y, where

$$\frac{\mathrm{d}y}{\mathrm{d}t} = g(x, y) + \sigma \frac{\mathrm{d}W_y}{\mathrm{d}t}.$$

Let

$$H(x_0, y, t) = g(\Phi(x_0; t), y)$$

and define

$$J(x_0, y) = \frac{1}{2}\int_0^{\mathrm{T}} \left\{ \left\| \frac{\mathrm{d}y}{\mathrm{d}t} - H(x_0, y, t) \right\|_{\Sigma}^2 + \nabla_y \cdot H(x_0, y, t) \right\} \mathrm{d}t.$$

It turns out (and we discuss this further in the next section) that $\exp(-J(x_0, y))$ may be thought of as a density on path space for y given x_0. Hence we may deduce that the pdf for (x_0, y) is again of the form

$$\rho(x_0, y) \propto \zeta(x_0) \exp(-J(x_0, y)), \tag{3.7}$$

as before, and that Bayes' rule gives

$$\rho(x_0| y) \propto \rho(x_0, y).$$

We may again apply any sampling method which requires knowledge about the posterior for x_0 given y only up to a multiplicative constant. In particular we may employ the Langevin SDE or Hybrid Monte Carlo. Both of these require the derivative $\nabla_{x_0} J(x_0, y)$ and this is

$$\int_0^{\mathrm{T}} \left\{ -\nabla_{x_0} H(x_0, y, t)^{\mathrm{T}} \Sigma^{-1} \left(\frac{\mathrm{d}y}{\mathrm{d}t} - H(x_0, y, t) \right) + \frac{1}{2} \nabla_{x_0} \left(\nabla_y \cdot H(x_0, y, t) \right) \right\} \mathrm{d}t. \tag{3.8}$$

4. Path space sampling and SPDEs

We now consider the case where $\gamma \neq 0$ and is invertible. Now the posterior distribution to be sampled is on the space of paths, and is hence infinite dimensional.

4.1. Density in path space

In order to develop a Bayesian approach to path sampling for $\{x(t)\}_{t \in [0,T]}$, conditional on observations, we need to define a probability density in path space. To this end we define the following functionals:

$$I(x) = \frac{1}{2} \int_0^{\mathrm{T}} \left(\left\| \frac{\mathrm{d}x}{\mathrm{d}t} - f(x) \right\|_{\Gamma}^2 + \frac{1}{2} \nabla_x \cdot f(x) \right) \mathrm{d}t,$$

$$J(x, y) = \frac{1}{2} \int_0^{\mathrm{T}} \left(\left\| \frac{\mathrm{d}y}{\mathrm{d}t} - g(x, y) \right\|_{\Sigma}^2 + \frac{1}{2} \nabla_y \cdot g(x, y) \right) \mathrm{d}t,$$

$$J_D(x, y) = \frac{1}{2} \sum_{k=1}^{K} \| y_k - h_k(x(t_k)) \|_{\Sigma_k}^2 .$$

Note that where the observation y appears in J it is a function, and where it appears in J_D it is a finite vector. Roughly speaking these three functionals are sums (or integrals) of squared independent noises. The extra divergence terms in I and J occur because all terms are interpreted in a symmetric fashion, with respect to the time-like variable. Thus all derivatives in the t-direction below should be approximated in a centred fashion. The divergence terms arise when converting Itô (non-centred) to Stratonovich (centred) stochastic integrals in I and J.

Here $I(x)$ is known as the Onsager–Machlup functional for (2.1) and the unconditional density for paths x solving (2.1) may be thought of as being proportional to (see [13])

$$Q(x) := q(x)\zeta(x(0))$$

where

$$q(x) := \exp\{-I(x)\}$$

and ζ is the density of the initial condition for $x(t)$. Similarly $I(x) + J(x, y)$ is the Onsager–Machlup functional for (2.1) and (2.2), with unconditional density for paths x, y found by exponentiating the negative of this functional. Hence, by Bayes' rule, the conditional density for paths x solving (2.1), given observation of y solving (2.2), may be thought of as being proportional to $Q(x) := q(x)\zeta(x(0))$ where

$$q(x) := \exp\{-I(x) - J(x, y)\}.$$

Similarly the conditional density for paths x solving (2.1), given observation of y from (2.3), may be thought of as being proportional to $Q(x) := q(x)\zeta(x(0))$ where

$$q(x) := \exp\{-I(x) - J_D(x, y)\}.$$

Note that, in all cases, q maps the Sobolev space of functions with square integrable first derivative $H^1([0, T])$ into the positive reals $\mathbb{R}^+$. The observations y parameterize $q(x)$.

In the following two sections we will introduce continuous and discrete time Markov chains whose invariant measure samples from densities on path space such as the functionals $Q(x)$ defined above. This will lead to SPDEs in Section 4.2 and a Markov chain constructed through a PDE with random initial data in Section 4.3. The development is analogous to that in the previous section, but is now infinite dimensional.

Defining the SPDEs will require calculation of the variational derivatives of $I(x)$, $J(x, y)$ and $J_D(x, y)$ with respect to x. We list these derivatives here. To this end it is useful to define

$$\mathcal{F}(x) = \frac{1}{2}\|f(x)\|_\Gamma^2 + \frac{1}{2}\nabla_x \cdot f(x)$$
$$\mathcal{H}(x) = \Gamma^{-1}\mathrm{d}f(x) - \mathrm{d}f(x)^{\mathrm{T}}\Gamma^{-1},$$

where $\mathrm{d}f : \mathbb{R}^d \to \mathbb{R}^{d\times d}$ is the Jacobian of f. We also use $\mathrm{d}g : \mathbb{R}^d \times \mathbb{R}^m \to \mathbb{R}^{m\times d}$ to denote the Jacobian of g with respect to x and $\mathrm{d}h_j : \mathbb{R}^d \to \mathbb{R}^{m\times d}$ to denote the Jacobian of h_j with respect to x. Then the required variational derivatives are:

$$\frac{\delta I}{\delta x} = -\Gamma^{-1}\frac{\mathrm{d}^2x}{\mathrm{d}t^2} + \mathcal{H}(x)\frac{\mathrm{d}x}{\mathrm{d}t} + \nabla_x\mathcal{F}(x)$$
$$\frac{\delta J}{\delta x} = -\mathrm{d}g(x, y)^{\mathrm{T}}\Sigma^{-1}\left[\frac{\mathrm{d}y}{\mathrm{d}t} - g(x, y)\right] + \frac{1}{2}\nabla_x\{\nabla_y \cdot g(x, y)\},$$
$$\frac{\delta J_D}{\delta x} = -\sum_{k=1}^{K}\mathrm{d}h(x(t_k))^{\mathrm{T}}\Sigma_k^{-1}[y_k - h_k(x(t_k))]\delta(t - t_k). \quad (4.1)$$

Notice that the last derivative is made up of point sources at the t_k. If $t_K = T$ then the jump induced by the delta function modifies the boundary condition at $t = T$ in the SPDEs that we write down in the next two sections. Otherwise the delta jumps are in the interior of the domain for the SPDEs.

One important observation here is that the presence of the second term in $\mathcal{F}$, namely the divergence of f, is something which has caused some controversy in the physics literature. A least squares definition of the density, based on Gaussian white noise, misses the term. Even if it is included, its magnitude — the factor $\frac{1}{2}$ — has been queried [22]. The analysis in [16, 31] and numerical experiments [38] are unequivocal that its presence is necessary and that the pre-factor of $\frac{1}{2}$ is the correct choice.

It is also because of this second term in $\mathcal{F}$ that we have concerns about sampling methods which first discretize the SDE (2.1) and then apply standard finite dimensional sampling techniques [1]. Such an approach can lead to a very indirect and numerically unsatisfactory approximation of the second term (see [38]). For this reason we strongly recommend employing the methodology outlined in this paper: namely to formulate an infinite dimensional sampling method in path space, and then approximate it.

4.2. Langevin SPDEs which sample path space

As illustrated in finite dimensions, the basic idea of *Langevin methods* is to construct a potential given by the gradient of the logarithm of the target density and to consider motion in this potential, driven by noise [32,33] — see (3.3). In the path space case the desired target density is proportional to $Q(x) = q(x)\zeta(x(0))$. Ignoring the boundary conditions (i.e. ζ) for a moment, we obtain the following SPDE for $x(t, s)$:

$$\frac{\partial x}{\partial s} = \frac{\delta \ln q(x)}{\delta x} + \sqrt{2}\frac{\partial W}{\partial s}, \quad (s, t) \in (0, \infty) \times (0, T). \quad (4.2)$$

Here s is an algorithmic time introduced to facilitate sampling in the space of paths, parameterized by real time t, and $\frac{\partial W}{\partial s}$ is a white noise in (t, s). The variational derivative of $\ln q(x)$ gives a second order differential operator in t and so the PDE is of reaction–diffusion type, subject to noise. The details of the SPDE depend upon whether the sampling of x is unconditional, or subject to observations y; the latter may be in discrete or continuous time. The previous section implicitly calculates the derivative of $\ln q(x)$ in each of these three cases, through the variational derivatives of $I(x)$, $J(x)$ and $J_D(x)$.

To find boundary conditions for the SPDE we argue in the standard fashion adopted in the calculus of variations. Notice that

$$\ln Q(x + \Delta x) = \ln Q(x) + \left(\frac{\delta}{\delta x}\ln Q(x), \Delta x\right) + \mathcal{O}(\|\Delta x\|^2)$$

where $(\cdot, \cdot)$ is the $L^2([0, T])$ inner product and $\|\cdot\|$ an appropriate norm. Now

$$\left(\frac{\delta}{\delta x}\ln Q(x), \Delta x\right) = \left(\frac{\delta}{\delta x}\ln q(x), \Delta x\right)$$
$$+ \left\langle\frac{\mathrm{d}x(0)}{\mathrm{d}t} - f(x(0)) + \Gamma\nabla_x \ln\zeta(x(0), \Delta x(0))\right\rangle_\Gamma$$
$$- \left\langle\frac{\mathrm{d}x(T)}{\mathrm{d}t} - f(x(T)), \Delta x(T)\right\rangle_\Gamma .$$

The first term on the right hand side gives the contribution to the derivative of $Q(x)$ appearing in the interior of the SPDE. Equating the second and third terms to zero, for all possible variations Δx, we obtain the following boundary conditions for the SPDE:

$$\frac{\partial x}{\partial t} - f(x) + \Gamma \nabla_x \ln \zeta(x) = 0, \quad t = 0, \tag{4.3}$$

$$\frac{\partial x}{\partial t} - f(x) = 0, \quad t = T. \tag{4.4}$$

The resulting SPDE (4.2)–(4.4) then has the desired equilibrium distribution.

When the observations are in discrete time and the last observation coincides with the last point at which we wish to sample x (so that $t_K = T$) the delta function at $t = t_K$ in the variational derivative of $\ln q(x)$ does not appear in the interior $t \in (0, T)$ and instead modifies the second boundary condition to read

$$\frac{\partial x}{\partial t} - f(x) - \Gamma \mathrm{d}h_K(x)^{\mathrm{T}} \Sigma_K^{-1} [y_K - h_K(x)] = 0, \quad t = T. \tag{4.5}$$

The nonlinear boundary conditions (4.4) and (4.5) both arise from jumps in the derivative induced by the Dirac masses contained in the boundary term with $t_K = T$ in (4.1).

Note that the case $h(x) = x$ and $y_J = x^+$ gives, in the limit where $\Sigma_K \to 0$, the Dirichlet boundary condition $x = x^+$ at $t = T$. Choosing ζ to be a Gaussian centred at x^-, and taking the limit of variance to zero, will also give a Dirichlet boundary condition $x = x^-$ at $t = 0$. These Dirichlet boundary conditions arise naturally in some applications of path sampling when bridges are studied [31,38].

By generalizing the *second order Langevin method* we obtain the following SPDE for $x(t, s)$:

$$\frac{\partial^2 x}{\partial s^2} + \iota \frac{\partial x}{\partial s} = \frac{\delta \ln q(x)}{\delta x} + \sqrt{2\iota} \frac{\partial W}{\partial s}, \\ (s, t) \in (0, \infty) \times (0, T), \tag{4.6}$$

with boundary conditions (4.3) and (4.4). Here $\iota > 0$ is an arbitrary positive parameter whose value may be optimized to improve sampling. This SPDE is a damped driven wave equation which yields the desired equilibrium distribution, when marginalized to x. The equilibrium distribution gives white noise in true time direction t for the momentum variable $\frac{\partial x}{\partial s}$ and this is hence natural initial data for the momentum variable.

It is also of interest to discuss *preconditioned Langevin equations*. Let $\mathcal{G}$ denote an arbitrary positive definite self-adjoint operator on the space of paths and consider the following SPDEs derived from (4.2) and (4.6) respectively:

$$\frac{\partial x}{\partial s} = \mathcal{G} \frac{\delta \ln q(x)}{\delta x} + \sqrt{2\mathcal{G}} \frac{\partial W}{\partial s}, \quad (s, t) \in (0, \infty) \times (0, T)$$

and

$$\mathcal{G}^{-1} \frac{\partial^2 x}{\partial s^2} + \iota \frac{\partial x}{\partial s} = \mathcal{G} \frac{\delta \ln q(x)}{\delta x} + \sqrt{2\iota\mathcal{G}} \frac{\partial W}{\partial s}, \\ (s, t) \in (0, \infty) \times (0, T).$$

Some examples substantiating this idea are given in [16] and [4]; in particular they show how to incorporate inhomogeneous boundary conditions in this context. Formally both these SPDEs preserve the desired invariant measure, for any choice of $\mathcal{G}$.

The simplest way to use any of the Langevin SPDEs described above to probe the desired (conditional) distribution on path space is as follows. Given some function $\phi: C([0, T], \mathbb{R}^d) \to \mathbb{R}$ (such as the maximum value along the path, or the value of $|x(t)|^2$ at some time point $t = \tau$) solve one of the Langevin SPDEs numerically, discretizing with increment Δs in the algorithmic time direction, thereby generating a sequence $x_n(t) \approx x(t, n\Delta s)$ (in practice this will need to be discretized along the path in t as well as in s). For M sufficiently large, the collection $\{x_n(t)\}_{n \geq M}$ form approximate samples from the desired distribution in path space. Hence, as $N \to \infty$, the average

$$\frac{1}{N} \sum_{n=0}^{N-1} \phi(x_n(t)) \tag{4.7}$$

will converge, by ergodicity, to an approximation of the average of ϕ in the desired conditional distribution. This is a discrete time analogue of (3.4). (The fact that we obtain an approximation, rather than the exact stationary value, results from discretization of the SPDE in t, s — see [37,39].) The role of $\mathcal{G}$ is to accelerate convergence as $N \to \infty$ and this point is discussed in the conclusions.

4.3. Hybrid Monte Carlo methods which sample path space

By generalizing the *Hybrid Monte Carlo method* we obtain the following Markov chain $x_n(t)$. Setting $\iota = 0$ in the SPDE (4.6) gives the PDE

$$\frac{\partial^2 x}{\partial s^2} = \frac{\delta \ln q(x)}{\delta x}, \quad (s, t) \in (0, \infty) \times (0, T). \tag{4.8}$$

The boundary conditions are again (4.3) and (4.4). This equation defines a solution operator

$$\mathcal{M}: \left(x(0), \frac{\partial x}{\partial s}(0) \right) \mapsto \left(x(S), \frac{\partial x}{\partial s}(S) \right)$$

mapping initial conditions to the solution at algorithmic time $s = S$. With the notation

$$P_x: (x, y) \mapsto x$$

we construct the Markov chain

$$x^{n+1} = P_x \mathcal{M}(x^n, \xi^n) \tag{4.9}$$

where the ξ^n are chosen to be i.i.d. Gaussian white noises in the true time direction t. This yields the desired equilibrium distribution. The formula (4.7) can again be used to probe the desired conditional distribution. Each step of the Markov chain requires the solution of a nonlinear wave equation over an interval of length S in s. Because numerical approximation of the wave equation (and hence $\mathcal{M}$) can lead to errors the formula (4.7) will in practice again only give an approximation of the true ergodic limit as $N \to \infty$. Pre-conditioning can also be used in the context of the Hybrid Monte Carlo method, replacing (4.8) by

$$\frac{\partial^2 x}{\partial s^2} = \mathcal{G}^2 \frac{\delta \ln q(x)}{\delta x}, \quad (s,t) \in (0,\infty) \times (0,T).$$

Again, $\mathcal{G}$ is used to accelerate convergence to stationarity. In this case the Markov chain (4.9) is generated by Gaussian ξ with mean zero and covariance $\mathcal{G}^2$.

The Hybrid Monte Carlo method was introduced and studied for discretizations of the path sampling problem in [1] where choices for the operator $\mathcal{G}$ were also discussed.

5. Remarks on other MCMC methods

The Langevin S(P)DEs and the Hybrid Monte Carlo methods both give rise to Markov chains which, if solved exactly (which is impossible in almost all practical situations), sample exactly from the desired distribution in their stationary measure. They are all examples of MCMC methods. But there is no reason to restrict sampling methods to these particular MCMC methods and in this section we briefly outline directions which might be fruitfully pursued to get improved sampling. We restrict our discussion to the case of path sampling as this high (infinite) dimensional setting is particularly challenging.

5.1. Metropolis–Hastings

In practice the MCMC methods in the previous section require numerical approximation of an (S)PDE in (s,t). This will incur errors and hence the stationary distribution will only be sampled approximately. The errors arising from integration in s can be corrected by means of a Metropolis–Hastings accept–reject criterion (see [25,32]). Furthermore, optimizing the choice of time-step in s can improve efficiency of the algorithm — we outline this below.

To apply the Metropolis–Hastings idea in path space, first discretize the path $\{x(t)\}$ giving rise to a vector x at the grid points. In the case of discrete observations this grid should ideally be chosen to include the observation times $\{t_j\}$. The signal $\{y(t)\}$ in the case of continuous time observations should also be discretized on the same grid.

The target density $Q(x)$ can then be approximated, using finite differences on the integrals, to define a finite dimensional target density $Q_D(x)$. By discretizing the SPDEs in the previous section on the same grid of points in t, as well as discretizing in s, we obtain a *proposal distribution*. Moves according to this proposal distribution (discretized SPDE) are then accepted or rejected with the Metropolis–Hastings probability leading to a Markov chain with invariant density $Q_D(x)$. Thus the effect of error introduced by integrating in s is removed; and the error due to approximation in t is controlled by the approximation of $Q(x)$ by $Q_D(x)$.

If a small time-step is used in s then the proposal distribution is not far from the current position of the Markov chain. This is known as a local proposal and for these there is a well-developed theory of optimality for the resulting MCMC methods [33]. The variance of an estimator in a Markov chain is given by the integrated autocorrelation function. Roughly speaking, very small steps in s are undesirable because the correlation in the resulting Markov chain is high, leading to high variance in estimators, which is inefficient; on the other hand, large steps in s lead to frequent rejections, which is also inefficient, again because correlation between steps is high when rejections are included. Choosing the optimal scaling of the step in s, with respect to the number of discretization points used along the path $\{x(t)\}$, is an area of current research activity [4], building on the existing studies of MCMC methods in high dimensions [33]. In the context of Metropolis–Hastings, good choices for the pre-conditioner $\mathcal{G}$ are ones which approximately equilibrate the convergence rates in different Fourier modes of the distribution. With this in mind, an interesting choice for $\mathcal{G}$ is a Green's operator for $-\frac{\mathrm{d}^2}{\mathrm{d}t^2}$ with homogeneous boundary conditions (see [16,4,1]).

If the integration time S is small in the Hybrid Monte Carlo method, then again the proposal distribution is local in nature. However, larger S will lead to better decorrelation, and hence efficiency, if the rejection rate is not too large. Hence it is of interest to study optimal choices for S, as a function of the number of discretization points, for this problem.

5.2. Global moves

Langevin methods have a potential problem for the sampling of multimodal distributions, namely that they can get stuck in a particular mode of the distribution for long times, because of the local (in state space) nature of the proposals. The Hybrid Monte Carlo method goes some way to ameliorating this issue as it allows free vibrations in the Hamiltonian given by the logarithm of the target density, and this is known to be beneficial in many finite dimensional sampling problems. However it is undoubtedly the case that sampling in path space will frequently be accelerated if problem specific global moves are incorporated into the proposal distributions. This is an open area for investigation. In the context of bridges the paper [20] contains some ideas that might form the basis of global proposal moves; but these are not likely to extend to data assimilation directly.

6. Relationship to other approaches

The purpose of this section is to discuss the approach advocated in this paper in relation to others prevalent in practice.

The first observation is that *4DVAR* is, in general, likely to be a highly ineffectual way of probing the posterior distribution; it will only be of value when the distribution is close to Gaussian, and has small variance — see the discussion in Section 3. 4DVAR was first studied for data assimilation in [24,6,36]. More recent references include [23,18,27,28,14,2,3].

The second observation is that, in the language of signal processing, the Bayesian method proposed here is performing *smoothing*, not *filtering*. This is because we sample from

$x(t), t \in [0, T]$ given the entire set of observations on $[0, T]$, whereas filtering would sample from $x(t)$ given only observations in $[0, t]$. Filtering is appropriate in applications where the data is on-line. But for off-line data, smoothing is quite natural. Off-line situations arise when performing parameter estimation, for example. There are also applications in Lagrangian data assimilation for oceanic velocity fields where data is only available infrequently and the off-line setting is appropriate.

The third observation concerns the relationship between what we advocate here, and the standard method for performing filtering for nonlinear SDEs conditional on observations. The rest of this section is devoted to this relationship. Standard methods are based on the *Zakai equation* and its generalizations. The Zakai equation is a linear partial differential equation for the probability density of the signal, conditional on observations. It is thus in the form of a Fokker–Planck equation, driven by noise (the observation). Informally it may be derived by employing the unconditional Fokker–Planck equation for (2.1) as a prior, and incorporating the observations via Bayes' law; the Markovian structure of the signal and observations allows filtering to be performed sequentially $0 \to T$. Smoothing can then be performed by means of a backward sweep, using a similar linear SPDE, incorporating data in reverse time $T \to 0$. See [34], Chapter 6, and the bibliographical Notes on Chapter 6, for further details and references.

A significant problem with use of the Zakai equation in the context of high dimensional problems ($d \gg 1$) is that the *independent* variables are in $\mathbb{R}^d$ and it is notoriously difficult to solve PDEs in high dimensions. *Particle filters* are a good tool for approximation of the Zakai equation in moderate dimension [7], but can be difficult to use in very high dimension. Weather prediction leads to d of order 10^8 and solution of the Zakai equation by particle filters is impractical. In this context two simplifications are usually introduced. The first is to use the *extended Kalman filter* (EKF) [5] which proceeds by linearizing the system and propagating a Gaussian model for the uncertainty; it is hence necessary to update the mean in $\mathbb{R}^d$ and the covariance matrix in $\mathbb{R}^{d \times d}$ sequentially, a task which is significantly easier than solving the Zakai equation. However even this approximation is impractical for large d and further approximations, primarily to effect dimension reduction on the covariance matrix, are performed; this leads to the *ensemble Kalman filter* (EnKF) [9] and its generalizations [30].

The approach we advocate in this paper is conceptually quite different from those based on the Zakai equation, and its Gaussian approximations. Instead of trying to sample from the probability distribution of the signal, at each point in time, by sequential means, we try to sample an entire path of the signal, from a distribution on path space. This leads to a nonlinear SPDE in one space dimension (t) and one time-like dimension indexing the sampling (s). The high dimension d enters as dimension of the *dependent* variable $x(t, s)$ which solves the SPDE; in contrast the Zakai equation has dimension d in the *independent* variable. The nonlinear SPDE proposed here hence has a considerable computational advantage over methods based on the Zakai equation, at least for problems which cannot be approximated in a Gaussian fashion.

7. Pedagogical example — sampling initial data

We study an example to illustrate Langevin sampling in the initial data space — i.e. when $\gamma = 0$ in (2.1). Thus we are in the framework of Section 3. We use the Langevin equation to probe the desired probability distribution, and compare our results with both the extended Kalman filter (EKF), and the ensemble Kalman filter (EnKF). We choose an example where the posterior can be calculated exactly, and then pushed forward to the final time where it is fair to compare both filtering methods (EKF, EnKF) and smoothing methods (our posterior sampling).

In order to illustrate our comparison between numerical methods we take the following explicitly solvable example. We study the equation

$$\frac{\mathrm{d}x}{\mathrm{d}t} = x - x^3, \quad x(0) = x_0 \sim \mathcal{N}(a, \sigma_{\text{init}}^2),$$

noting that this equation, and its linearization, can both be solved exactly. The observations are in discrete time and take the form

$$y_k = x(k\delta) + \mathcal{N}(0, \sigma_{\text{obs}}^2), \quad i = 1, \ldots, K.$$

Given the observations, the posterior on the initial data can be calculated exactly, using the fact that the solution operator $\Phi(x_0; t)$ in (3.1) can be calculated analytically, and no numerical approximation is needed. The exact solution operator is also used in the Langevin sampler and particle filters. Furthermore, we also use the fact that the derivative of Φ with respect to x_0 can be calculated analytically; this enables us to find the term $\nabla_{x_0} H(x_0)$ in the Langevin equation (3.5) explicitly, without resorting to numerical approximation. For more complex problems these tasks will have to be carried out by numerical approximation, of course.

The exact posterior distribution on x_0 can be mapped forward explicitly to obtain the exact posterior at any time t, including $t = T = K\delta$. Notice that the exact posterior corresponds to solving the smoothing problem. The Langevin approach hence directly approximates the smoothing problem. Both EKF and EnKF approximate the filtering problem. Filtering and smoothing, if exact, only coincide at the final time $t = T$. Hence we compare the methods at this time, for which EKF (resp. EnKF) is identical to the smoother analogue EKS (resp. EnKS).

We now present three numerical experiments illustrating the behaviour of the Langevin sampler, in comparison with Kalman based methods. We use the perturbed observation EnKF as presented in [9]. In all of the three figures presented in this section, the solid black curve is the exact posterior. Our interest is hence in how well this is replicated by the different sampling methods. We choose $a = -0.1, \sigma_{\text{init}} = 0.2, \sigma_{\text{obs}} = 0.8$, and $K = 10$. The three figures differ only in the frequency of observations δ (which is 0.095, 0.09, and 0.3 for Figs. 1–3

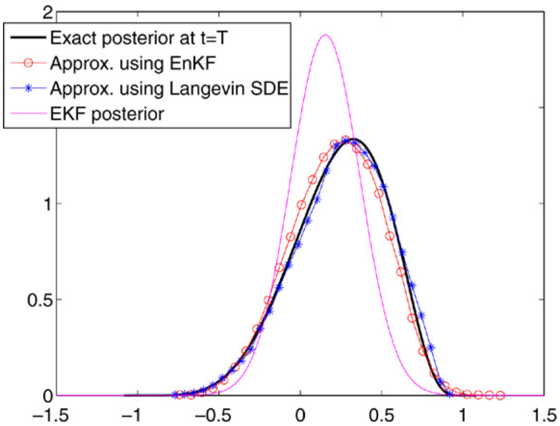

Fig. 1. Comparison of the exact posterior distribution, the Langevin approximation, and approximations by EnKF and EKF.

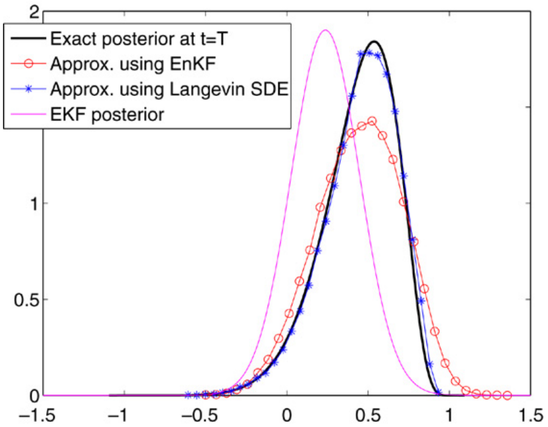

Fig. 2. Comparison of the exact posterior distribution, the Langevin approximation, and approximations by EnKF and EKF.

respectively) and the initial condition x_0 (which is 0.5 for Figs. 1 and 2 but 0.0001 for Fig. 3). Note that the actual initial condition used is not the mean of the prior distribution on $x(0)$. We chose a very large sample size (50 000) for both the EnKF and Langevin method, so that we can compare the results without dealing with sampling issues.

Fig. 1 shows a situation in which both the Langevin sampling and EnKF reproduce the target posterior density very well; the EKF, performs quite poorly. That the EKF performs poorly is fairly typical for problems with any appreciable nonlinear effects and Fig. 2 again shows the EKF performing poorly. In this case the EnKF is appreciably better than the EKF, but is outperformed by the Langevin method. Finally, Fig. 3 shows a situation where the EnKF fails to produce a reasonable approximation at all, but once again the Langevin method performs excellently. (The EKF is not shown here.)

The moral of these numerical experiments is that standard techniques, widely used in practice, and based on approximations of the Kalman filter, can fail when applied to nonlinear problems which are far from Gaussian. The Langevin method, however, is very robust (although this does come at the expense of a considerable increase in computational complexity). For this reason we proceed in the next section to generalize the Langevin method to the sampling of path space, necessary whenever the basic model dynamics (2.1) is stochastic — $\gamma \neq 0$.

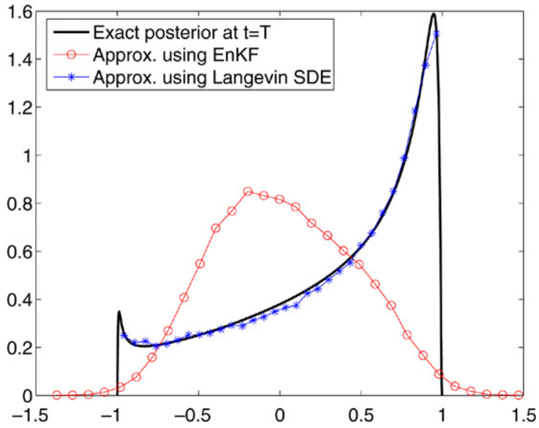

Fig. 3. Comparison of the exact posterior distribution, the Langevin approximation, and approximations by EnKF.

8. Pedagogical example — sampling path space

We discuss a simple example motivated by Lagrangian data assimilation. We use the example to illustrate the use of the (first order) Langevin SPDE for sampling conditional paths of (2.1) when $\gamma \neq 0$. In the previous example the exact posterior was available analytically so that evaluation of the methods studied was straightforward. In this path space example we choose a problem where the posterior mean can be calculated so that we may again evaluate the sampling method.

Consider a one dimensional velocity field of the form

$$v(y,t) = x_1(t) + x_2(t)\sin(y) + x_3(t)\cos(y)$$

where the $x_k(t)$ are Ornstein–Uhlenbeck processes solving

$$\frac{\mathrm{d}x_k}{\mathrm{d}t} = -\alpha x_k + \gamma \frac{\mathrm{d}W_{x,k}}{\mathrm{d}t}. \tag{8.1}$$

We assume that the particles are initially stationary and independent so that each $x_k(0)$ is distributed as $\mathcal{N}(0, \gamma^2/2\alpha)$, with density $\zeta(x) \propto \exp\{-\alpha x^2/\gamma^2\}$.

We study the question of making inference about the paths $\{x_k(t)\}$ from the observation of L drifters $\{y_\ell\}_{\ell=1}^{L}$ moving in the velocity field, and subject to random forcing idealized as white noise (e.g. molecular diffusion):

$$\frac{\mathrm{d}y_\ell}{\mathrm{d}t} = v(y_\ell, t) + \sigma \frac{\mathrm{d}W_{y,\ell}}{\mathrm{d}t}. \tag{8.2}$$

Here the $W_{x,k}$ and $W_{y,\ell}$ are independent standard Brownian motions. The initial conditions for the y_ℓ are i.i.d. random variables drawn from the distribution $\mathcal{N}(0, 2\pi)$.

Writing $y = (y_1, \dots, y_L)^{\mathrm{T}}$ and $W_y = (W_{y,1}, \dots, W_{y,L})^{\mathrm{T}}$ we obtain

$$\frac{\mathrm{d}y}{\mathrm{d}t} = h(y)x + \sigma \frac{\mathrm{d}W_y}{\mathrm{d}t}, \tag{8.3}$$

where $h: \mathbb{R}^L \to \mathbb{R}^{L\times 3}$, $\sigma \in \mathbb{R}^+$.

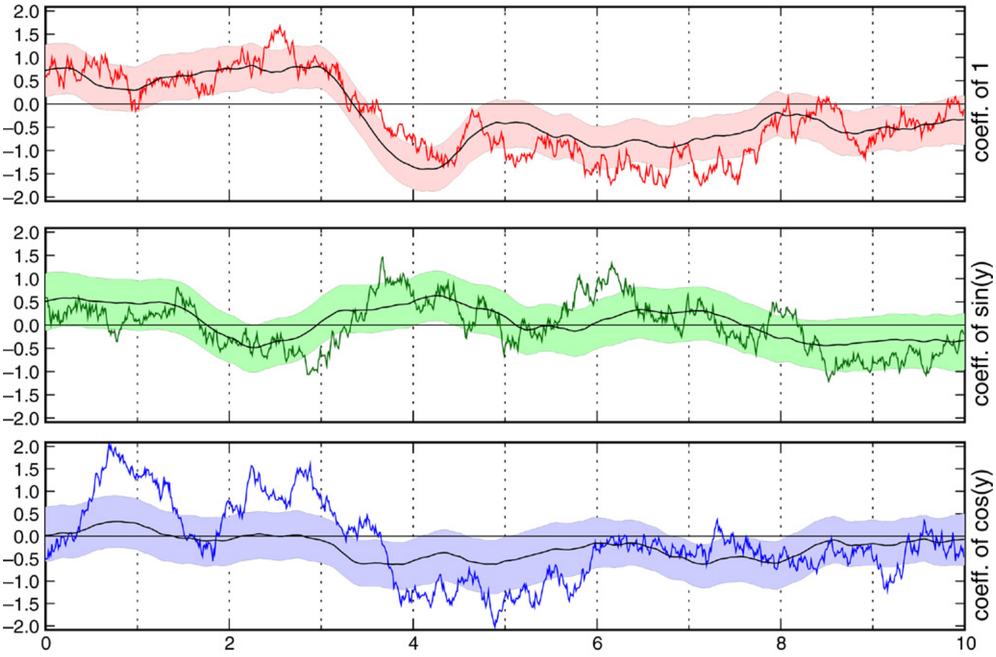

Fig. 4. Reconstruction of the x_i solving (8.1), together with one standard deviation bounds, on $s \in [0, 100]$; 5 drifters are used.

In this case the Langevin SPDE (4.2)–(4.4) is hence

$$\begin{aligned}\frac{\partial x}{\partial s} &= \frac{1}{\gamma^2}\frac{\partial^2 x}{\partial t^2} - \frac{\alpha^2}{\gamma^2}x + \frac{1}{\sigma^2}h(y)^{\mathrm{T}}\left[\frac{\mathrm{d}y}{\mathrm{d}t} - h(y)x\right] \\ &\quad - \frac{1}{2}\nabla_y \cdot h(y)^{\mathrm{T}} + \sqrt{2}\frac{\partial W}{\partial s}, \quad (s,t)\in(0,\infty)\times(0,T) \\ \frac{\partial x}{\partial t} &= +\alpha x, \quad (s,t)\in(0,\infty)\times\{0\} \\ \frac{\partial x}{\partial t} &= -\alpha x, \quad (s,t)\in(0,\infty)\times\{T\} \\ x &= x_0, \quad (s,t)\in 0\times[0,T].\end{aligned}$$

Because the SPDE is linear, the mean $\bar{x}$ in the stationary measure is found by removing the derivative in s and the noise to obtain

$$\begin{aligned}&\frac{1}{\gamma^2}\frac{\mathrm{d}^2\bar{x}}{\mathrm{d}t^2} - \frac{\alpha^2}{\gamma^2}\bar{x} - \frac{1}{\sigma^2}h(y)^{\mathrm{T}}h(y)\bar{x} \\ &\quad = -\frac{1}{\sigma^2}h(y)^{\mathrm{T}}\frac{\mathrm{d}y}{\mathrm{d}t} + \frac{1}{2}\nabla_y \cdot h(y)^{\mathrm{T}}, \quad t\in(0,T), \\ &\frac{\mathrm{d}\bar{x}}{\mathrm{d}t} = +\alpha\bar{x}, \quad t=0, \\ &\frac{\mathrm{d}\bar{x}}{\mathrm{d}t} = -\alpha\bar{x}, \quad t=T.\end{aligned}$$

Note that if $\sigma \ll \min(\gamma, 1)$ then, formally, the equation for the mean is dominated by the *normal equations*

$$h(y)^{\mathrm{T}}\left[\frac{\mathrm{d}y}{\mathrm{d}t} - h(y)\bar{x}\right] \approx 0$$

which arise from trying to solve the overdetermined Eq. (8.3) for x, when the noise is ignored. But when noise is present, however small, $\frac{\mathrm{d}y}{\mathrm{d}t}$ exists only as a distribution (it has the regularity of white noise) and so the second order differential operator in x, which incorporates prior information on x, is required to make sense of the mean.

Our numerical experiments are conducted as follows. We set $\alpha = \gamma = \sigma = 1$ and generated a single path for each x_k, $k = 1, 2, 3$ solving (8.1) on the interval $t \in [0, 10]$, using stationary initial conditions as described above. We also generated the trajectories of 500 drifters y_i moving according to (8.2), with initial conditions drawn from a Gaussian distribution as described above. We then chose L drifter paths, with $L = 5, 50$ and 500 respectively, and solved the Langevin SPDE to sample from the distribution of the x_k. We integrated over 100 algorithmic time units in s and approximated the mean of the x_k, together with one standard deviation, using (4.7). We also calculated the mean directly by solving the boundary value problem for $\bar{x}$. In all cases we used a formally second order accurate approximation in the spatial variable t, and for time integration we used a linearly implicit method with Crank–Nicolson approximation of the leading order differential operator. We emphasize that the signals x_k are not available to the Langevin SPDE or the boundary value problem: only information about the drifters y_ℓ is used to reconstruct the x_k. The signals are shown in the following figures so that the reconstruction of the signal may be judged.

The results are shown in Figs. 4–6, corresponding to $L = 5, 50$ and 500 respectively. In each figure we consider x_1 in the top panel, x_2 in the middle and x_3 at the bottom. The actual signal x_k is the non-smooth curve whilst the mean of the desired conditional distribution, found by solving the equation for $\bar{x}$, is the smooth curve. The shaded bands show an estimate of one standard deviation about the mean, with both mean and standard deviation estimated by time averaging solution of the Langevin SPDE in s.

The figures illustrate two facts, one a property of the path sampling procedure we propose in this paper, the second a

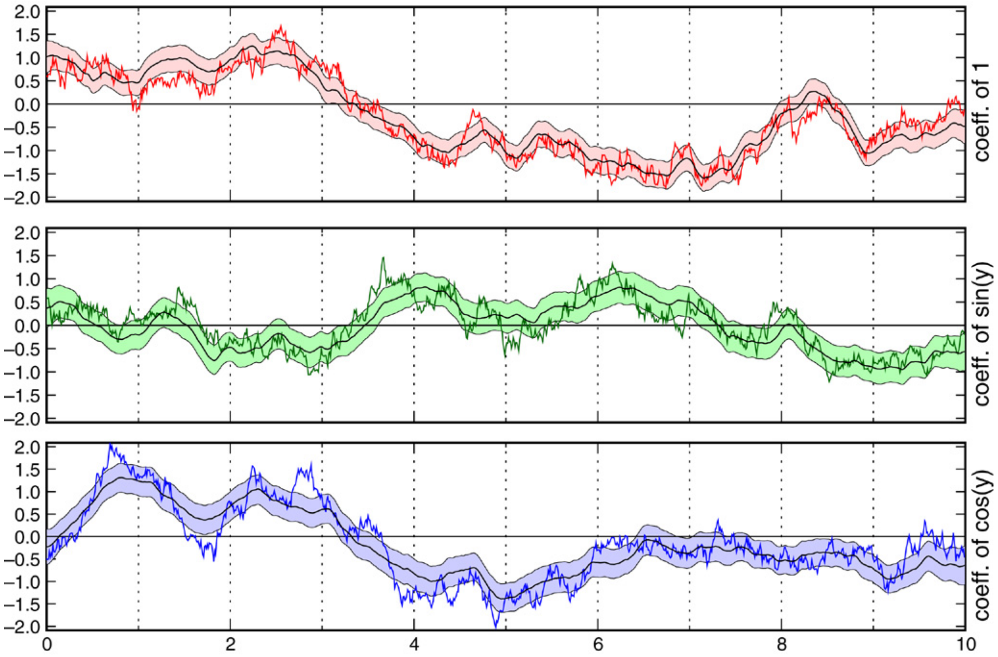

Fig. 5. Reconstruction of the x_i solving (8.1), together with one standard deviation bounds, on $s \in [0, 100]$; 50 drifters are used.

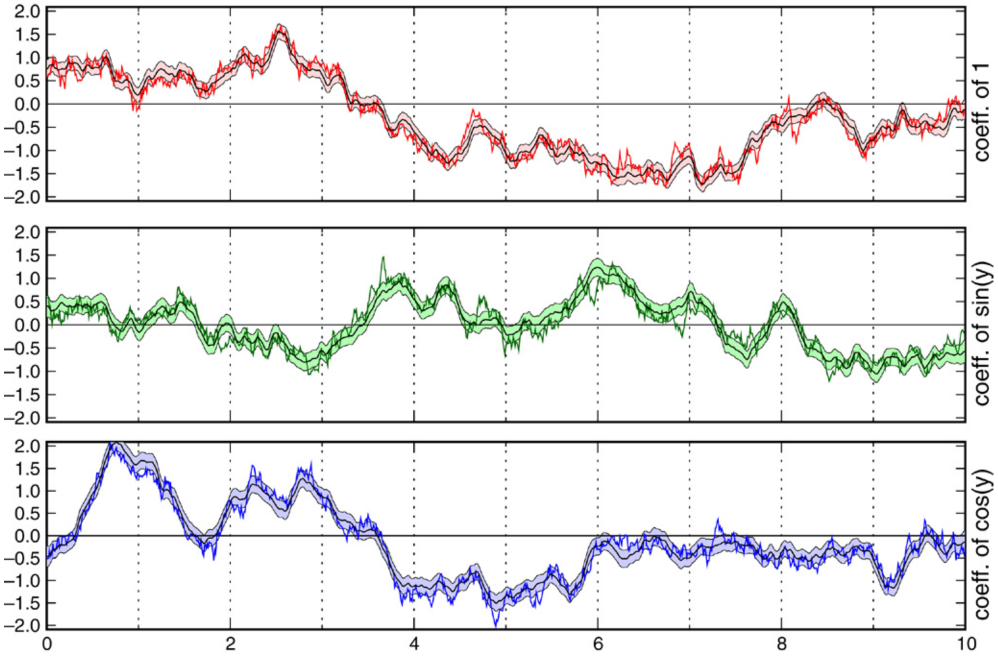

Fig. 6. Reconstruction of the x_i solving (8.1), together with one standard deviation bounds, on $s \in [0, 100]$; 500 drifters are used.

property of the desired conditional distribution for this data assimilation problem. The first fact is this: because the true mean $\bar{x}$ lies in the middle of the shaded band, it is clear that the estimate of the mean, calculated through time averaging, is accurate at $s = 100$. The second fact is this: as L is increased our ability to recover the actual signal increases; this is manifest in the fact that the mean gets closer to the signal, and the standard deviation bounds get tighter.

To give some insight into how long the Langevin SPDE has to be integrated to obtain accurate time averages, we generated data analogous to that in Fig. 4, but only integrated to algorithmic time $s = 10$. The results are shown in Fig. 7. The fact that $\bar{x}$ no longer lies in the middle of the shaded bands, at least for some parts of the paths, indicates that the time average of the path has not converged to the mean value in the stationary distribution.

9. Challenges

We have presented an approach to data assimilation that will be useful for problems which are highly non-Gaussian.

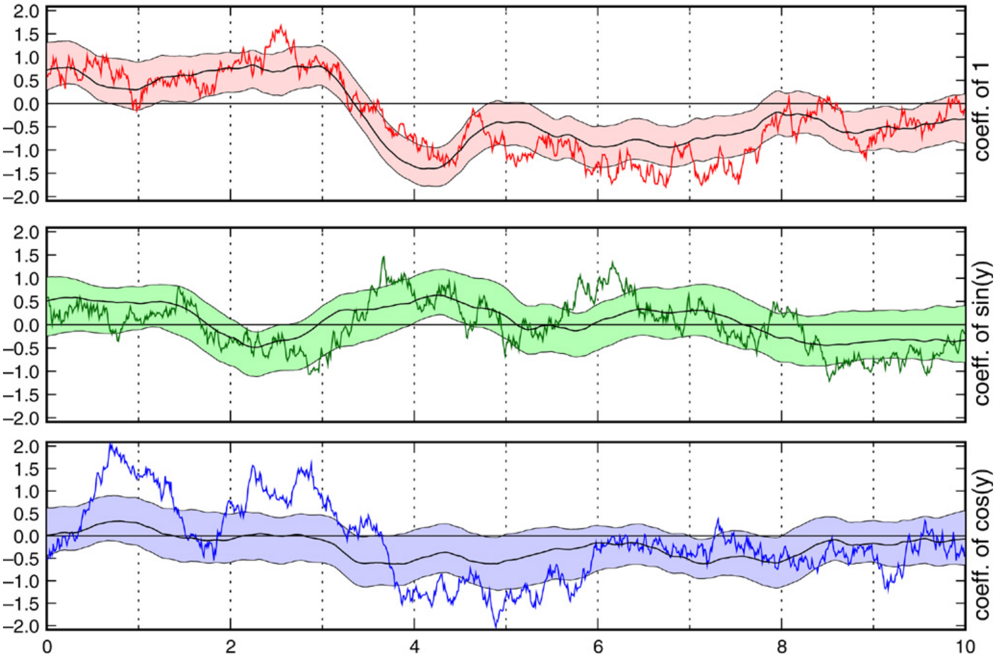

Fig. 7. Reconstruction of the x_i solving (8.1), together with one standard deviation bounds, on $s \in [0, 10]$; 5 drifters are used. Note that the estimate of the mean (the middle of the shaded bands) is not always close to the actual mean (the smooth curve). This should be contrasted with Fig. 4 which is on a longer interval in algorithmic time s.

It is a Bayesian framework based on sampling the posterior distribution by MCMC methods, especially the Langevin equation. Both deterministic and stochastic model dynamics are considered. In the former case the posterior is on the initial data; in the latter case it is on the space of paths. The approach outlined here presents a number of significant scientific challenges. We outline some of these, breaking the challenges down into three categories: applications, mathematical and computational.

9.1. Applications

- In the context of short term weather prediction, Gaussian based Kalman filter approximation often appears quite effective; it would be interesting to quantify this by comparing with the Bayesian approach described here.
- In the context of Lagrangian data assimilation for oceans, it would be of interest to use the methodology proposed here to study the multimodal problems which often arise quite naturally, and for which the extended Kalman filter diverges.
- For both weather prediction and ocean modelling it would be of interest to incorporate the methodology proposed here for the purposes of parameter estimation. In this context the paths of (2.1) are treated as missing data which are sampled to enable estimation of parameters appearing in (2.1) itself. A Gibbs sampler [32] could be used to alternate between the missing data and the parameters.
- There are many other potential applications of this methodology in chemistry, physics, electrical engineering and econometrics, for example.

9.2. Mathematical

- The SPDEs which arise as the formal infinite dimensional Langevin equations, and the related PDE which arises in the Hybrid Monte Carlo method, all lead to significant problems in analysis concerned with the existence, uniqueness, ergodicity and rate of convergence to stationarity. Some of these issues have been resolved for particular forms of nonlinearity in (2.1) and (2.2) (see [15,16]) primarily for vector fields f, and g in the case of continuous time observations, which are combinations of gradients and linear vector fields.
- For non-gradient vector fields the presence of the term $\mathcal{H}(x)\frac{\partial x}{\partial t}$ causes particular problems in the development of a theory for the SPDE as, when the solution operator for the linear part of the Langevin SPDE is applied to it, a definition of stochastic integral is required. Numerical evidence as well as the derivation of $I(x)$ by means of the Girsanov formula, suggests that this should be a Stratonovich-type centred definition, but the mathematical analysis remains to be developed. A related, but simpler, mathematical question arises in the interpretation of the stochastic integral with respect to y arising in (3.8).
- In some applications the underlying path to be sampled arises from an SPDE itself: i.e. Eq. (2.1) is itself an SPDE; it would be of interest to derive the relevant Langevin SPDE here, in which the variable t would appear as a spatial variable, in addition to the spatial derivatives already appearing in (2.1).
- We have assumed for simplicity that white noise affects all components of the signal and observation equations; relaxing this assumption is natural in some applications,

and it would be of interest to find the relevant SPDEs for sampling in this case; as mentioned in Section 2 this case arises when studying model error.

9.3. Computational

- Sampling the posterior distribution of the smoothing problem is, in general, costly in terms of computational time. A major challenge is to understand situations where sampling the posterior of the smoothing problem is necessary from an applied viewpoint, and then to develop efficient algorithms for doing so.
- If the dimension d is high then, since the number of dependent variables in the SPDEs proposed here will scale like d, techniques are required to reduce the dimensionality for sampling; multiscale methods are likely to be useful in this context [12]. Some interesting work in this direction, using relative entropy, may be found in [10].
- Within the context of Langevin algorithms it would be of interest to study choices of the pre-conditioner $\mathcal{G}$, and discretization method for the SPDE, which lead to efficient algorithms; efficiency in this context should be measured through the integrated autocorrelation function which quantifies the fluctuations in estimates of the form (4.7), for expectations of $\phi(x(\cdot))$ with respect to the desired conditional measure [33].
- Similar considerations apply to Hybrid Monte Carlo methods, and the choice of pre-conditioner.
- It is also of interest to compare first order and second order Langevin based methods with one another and with the Hybrid Monte Carlo method, once good pre-conditioners have been found. See [1] for a step in this direction.
- The use of other MCMC methods to sample the desired probability measures on path space should also be explored. It is common practical experience that, whilst Langevin-type methods are provably efficient within the context of methods using local (in state space) proposals [33], greater speed-ups can often be obtained by incorporating additional global moves, based on problem specific knowledge.
- The issue of how to discretize the SPDE is also non-trivial. In particular for non-gradient vector fields in (2.1) and (2.2), the term $\mathcal{H}(x)\frac{\partial x}{\partial t}$ needs to be discretized carefully (as discussed above centred differencing is necessary in our formulations of the SPDE) essentially for the same reasons that the SPDE theory is hard to develop in this case.

Acknowledgements

The authors are grateful to Greg Eyink, Kayo Ide, Chris Jones and Arthur Krener for helpful input. AMS and JV are grateful to the EPSRC (UK) and ONR (USA) for funding. AA would like to acknowledge the support of ONR (grant number N00014-04-1-0215) and SAMSI (grant number 03-SC-NSF-1009).

References

[1] F. Alexander, G. Eyink, J. Restrepo, Accelerated Monte–Carlo for optimal estimation of time series, J. Stat. Phys. 119 (2005) 1331–1345.

[2] A.F. Bennet, Inverse Methods in Physical Oceanography, University Press, Cambridge, 1999.

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

[4] A. Beskos, G.O. Roberts, A.M. Stuart, J. Voss, An MCMC method for diffusion bridges, Ann. Appl. Prob. (submitted for publication).

[5] D.E. Caitlin, Estimation, Control and the Discrete Kalman Filter, Springer, New York, 1989.

[6] P. Courtier, O. Talagrand, Variational assimilation of meteorological observations with the adjoint vorticity equation (II): Numerical results, Quart. J. R. Meteorol. Soc. 113 (1987) 1329–1368.

[7] D. Crisan, P. Del Moral, T.J. Lyons, Discrete filtering using branching and interacting particle systems, Markov Process. Related Fields 5 (1999) 293–318.

[8] G. Da Prato, J. Zabczyk, Stochastic Equations in Infinite Dimensions, in: Encyclopedia of Mathematics and its Applications, vol. 44, Cambridge University Press, 1992.

[9] G. Evensen, The ensemble Kalman filter: Theoretical formulation and practical implementation, Ocean Dynam. 53 (2003) 343–367.

[10] G.L. Eyink, S. Kim, A maximum entropy method for particle filtering, J. Stat. Phys. (2006) (in press).

[11] C.W. Gardiner, Handbook of Stochastic Methods, Springer, Berlin, 1985.

[12] D. Givon, R. Kupferman, A.M. Stuart, Extracting macroscopic dynamics: Model problems and algorithms, Nonlinearity 17 (2004) R55–R127.

[13] R. Graham, Path integral formulation of general diffusion processes, Z. Physik B 26 (1977) 281–290.

[14] A.K. Griffith, N.K. Nichols, Adjoint techniques in data assimilation for treating systematic model error, J. Flow, Turbul. Combust. 65 (2001) 469–488.

[15] M. Hairer, A.M. Stuart, J. Voss, P. Wiberg, Analysis of SPDEs arising in path sampling, part I: The Gaussian case, Commun. Math. Sci. 3 (2005) 587–603.

[16] M. Hairer, A.M. Stuart, J. Voss, Analysis of SPDEs arising in path sampling, part II: The nonlinear case, Ann. Appl. Prob. (submitted for publication).

[17] K. Ide, L. Kuznetsov, C.K.R.T. Jones, Lagrangian data assimilation for point vortex systems, J. Turbul. 3 (2002) 53.

[18] C. Johnson, N.K. Nichols, B.J. Hoskins, A singular vector perspective of 4DVar: Filtering and interpolation, Quart. J. R. Meteorol. Soc. 131 (2005) 1–20.

[19] I. Karatzas, S.E. Shreve, Brownian Motion and Stochastic Calculus, second ed., Springer, 1991.

[20] A. Krener, Reciprocal diffusions in flat space, Probab. Theory Related Fields 107 (1997) 243–281.

[21] L. Kuznetsov, K. Ide, C.K.R.T. Jones, A method for assimilation of Lagrangian data, Mon. Weather Rev. 131 (2003) 2247–2260.

[22] F. Langouche, D. Roekaerts, E. Tirapegui, Functional integral methods for stochastic fields, Physica 95A (1979) 252–274.

[23] A.S. Lawless, S. Gratton, N.K. Nichols, An investigation of incremental 4DVar using non-tangent linear models, Quart. J. R. Meteorol. Soc. 131 (2005) 459–476.

[24] F.-X. Le Dimet, O. Talagrand, Variational algorithms for analysis and assimilation of meteorological observations: Theoretical aspects, Tellus A 38 (1986) 97–110.

[25] J. Liu, Monte Carlo Strategies in Scientific Computing, Springer, New York, 2001.

[26] A. Lorenc, Analysis methods for numerical weather prediction, Quart. J. R. Meteorol. Soc. 112 (1986) 1177–1194.

[27] N. Nichols, Data assimilation: Aims and basic concepts, in: R. Swinbank, V. Shutyaev, W.A. Lahoz (Eds.), Data Assimilation for the Earth System, Kluwer Academic, 2003, pp. 9–20.

[28] N.K. Nichols, Treating model error in 3-D and 4-D data assimilation, in: R. Swinbank, V. Shutyaev, W.A. Lahoz (Eds.), Data Assimilation for the Earth System, Kluwer Academic, 2003, pp. 127–135.

[29] B. Øksendal, Stochastic Differential Equations: An Introduction with Applications, fifth ed., Springer, Berlin, 1998.

[30] E. Ott, B.R. Hunt, I. Szunyogh, A.V. Zimin, E.J. Kostelich, M. Corazza, E. Kalnay, D.J. Patil, J.A. Yorke, A local ensemble Kalman filter for atmospheric data assimilation, Tellus A 56 (2004) 415–428.

[31] M. Reznikoff, E. Vanden-Eijnden, Invariant measures of stochastic PDEs, C. R. Acad. Sci. Paris 340 (2005) 305–308.

[32] C.P. Robert, G. Casella, Monte Carlo Statistical Methods, Springer, New York, 1999.

[33] G. Roberts, J.S. Rosenthal, Optimal scaling for various Metropolis–Hastings algorithms, Statist. Sci. 16 (4) (2001) 351–367.

[34] B.L. Rozovskii, Stochastic Evolution Systems: Linear Theory and Applications to Nonlinear Filtering, Kluwer, The Netherlands, 1990.

[35] H. Salman, L. Kuznetsov, C.K.R.T. Jones, K. Ide, A method for assimilating Lagrangian data into a shallow-water equation ocean model, Mon. Weather Rev. 134 (2006) 1081–1101.

[36] Y. Sasaki, Some basic formalisms in numerical variational analysis, Mon. Weather Rev. 98 (1970) 875–883.

[37] T. Shardlow, A.M. Stuart, A perturbation theory for ergodic Markov chains with application to numerical approximation, SIAM J. Numer. Anal. 37 (2000) 1120–1137.

[38] A.M. Stuart, J. Voss, P. Wiberg, Conditional path sampling of SDEs and the Langevin MCMC method, Commun. Math. Sci. 2 (2004) 685–697.

[39] D. Talay, Second-order discretization schemes for stochastic differential systems for the computation of the invariant law, Stoch. Stoch. Rep. 29 (1990) 13–36.

[40] J. Zabczyk, Symmetric solutions of semilinear stochastic equations, in: G. Da Prato, L. Tubaro (Eds.), Stochastic Partial Differential Equations and Applications II (Proceedings, Trento 1988), in: Lecture Notes in Mathematics, vol. 1390, Springer, 1988, pp. 237–256.