Tellus (2008), 60A, 336–347
Printed in Singapore. All rights reserved

TELLUS

A Bayesian approach to Lagrangian data assimilation

By A. APTE[1]*, C. K. R. T. JONES[2] and A. M. STUART[3], [1]*Centre for Applicable Mathematics, Tata Institute of Fundamental Research, Bangalore, India;* [2]*Department of Mathematics, University of North Carolina, Chapel Hill, NC 27599, USA;* [3]*Mathematics Institute, University of Warwick, Coventry CV4 7AL, UK*

Manuscript received 13 June 2007; in final form 13 November 2007

ABSTRACT

Lagrangian data arise from instruments that are carried by the flow in a fluid field. Assimilation of such data into ocean models presents a challenge due to the potential complexity of Lagrangian trajectories in relatively simple flow fields. We adopt a Bayesian perspective on this problem and thereby take account of the fully non-linear features of the underlying model.

In the perfect model scenario, the posterior distribution for the initial state of the system contains all the information that can be extracted from a given realization of observations and the model dynamics. We work in the smoothing context in which the posterior on the initial conditions is determined by future observations. This posterior distribution gives the optimal ensemble to be used in data assimilation. The issue then is sampling this distribution. We develop, implement, and test sampling methods, based on Markov-chain Monte Carlo (MCMC), which are particularly well suited to the low-dimensional, but highly non-linear, nature of Lagrangian data. We compare these methods to the well-established ensemble Kalman filter (EnKF) approach. It is seen that the MCMC based methods correctly sample the desired posterior distribution whereas the EnKF may fail due to infrequent observations or non-linear structures in the underlying flow.

1. Introduction

Lagrangian instruments such as drifters and floats that are carried by the flow provide good measurements of the fluid motion and of its transport properties. The density of such Lagrangian data has increased significantly over the past few years and is expected to increase further. As a consequence, the problem of assimilating these Lagrangian data into ocean general circulation models for improving the state estimates of the ocean has rightfully received a lot of attention in recent years (Carter, 1989; Ide et al., 2002; Kuznetsov et al., 2003; Molcard et al., 2003; Özgökmen et al., 2003). These studies have used assimilation methods based on the Kalman filter or optimal interpolation, which are appropriate for linear or approximately linear systems with Gaussian errors in the data and in the prior state estimates. These assumptions are generally not valid for the problem of Lagrangian data assimilation, even in relatively simple flow fields, because of the non-linearities of the Lagrangian particle trajectories. An example of the failure of the Kalman filter based methods because of the violation of these assumptions is provided by the saddle issue (Kuznetsov et al., 2003), that occurs when an observed trajectory passes close to a saddle point of the flow.

*Corresponding author.
e-mail: apte@math.tifrbng.res.in
DOI: 10.1111/j.1600-0870.2007.00295.x

The main motivation for this study is to understand the effects of the highly non-linear nature of the Lagrangian data on the assimilation procedure. This is accomplished by adopting a Bayesian viewpoint on the Lagrangian data assimilation problem. The noisy observations of a system along with the probability density function (PDF) of errors contained in them, a dynamic model of the system, and the prior information about the initial state, given in terms of a prior PDF, can be combined, via Bayes' theorem, to give the posterior PDF of the initial state of the system (Apte et al., 2007). We emphasize that such a posterior PDF on the initial state uses 'future' observations as well—a smoother rather than a filter. This Bayes, or exact, posterior, as we will call it throughout the paper, is in general non-Gaussian even if the prior is Gaussian and the observational noise is Gaussian. This is because the transformation from initial data into the state at later times, where observations are made, is non-linear. Furthermore, the posterior on the initial state can also be pushed forward, under the dynamics, to any later time. This transformation can also introduce non-Gaussian behaviour. We present examples of such non-Gaussian posterior PDFs in perfect model, identical twin experiments in the context of Lagrangian data assimilation for the linearized shallow water model.

We also compare this exact posterior with the posterior implied by the samples from the ensemble Kalman filter (EnKF). This helps us relate the saddle issue with the non-Gaussianity of the posterior PDF by showing that the cause of the failure of the

EnKF is the non-Gaussian posterior distribution arising when the time interval between the observations is long. It will be demonstrated that the presence of centres, that is, elliptic fixed points of the flow, also leads to non-Gaussian posterior distributions and to failure of the EnKF.

In order to study the posterior distribution, we develop three sampling methods, based on the Langevin equation and the Metropolis–Hastings algorithm, that allow us to obtain an ensemble which faithfully represents this exact posterior. Ensemble based methods are attractive since averages over the ensemble allow calculation of any statistical quantity to be obtained from the PDF. We demonstrate the use of these exact sampling techniques in the linearized shallow water model.

The paper is organized as follows. In the remainder of this section, we present the application of Bayes theorem to a general data assimilation problem with a deterministic dynamic model. The linearized shallow water model that we use in our numerical experiments is described in Section 2. In Section 3, we discuss the structure of the posterior distribution and how it is affected by the choice of the prior, the observational frequency, and the dynamics of the drifters. In Section 4, we compare this exact posterior with perhaps the most commonly used, ensemble based, data assimilation method: the EnKF. The different methods used for sampling the posterior are presented and compared in Section 5. A discussion of results and some directions for future work are contained in Section 6.

In order to discuss the Bayesian framework, we write a general deterministic model for an n-dimensional state vector $x(t) \in \mathbb{R}^n$ as

$$\frac{\mathrm{d}x}{\mathrm{d}t} = f(x), \quad x(0) = x_0 \sim \zeta. \tag{1}$$

Thus, before making any observations, the initial conditions are supposed to be drawn from a prior with probability density function $p_\zeta(x_0)$. We denote the solution operator for the dynamics by Φ:

$$x(t) = \Phi(x_0; t). \tag{2}$$

We assume that the observations depend only on the state of the system at a particular time, but are subject to noise. Thus, the observation $y_k \in \mathbb{R}^m$ at time t_k can be written as

$$y_k = h[x(t_k)] + \eta_k = h[\Phi(x_0; t_k)] + \eta_k,$$

where $h : \mathbb{R}^n \to \mathbb{R}^m$. The second equality in the above equation emphasizes the fact that we consider the observations to be functions (often highly non-linear numerical functions, typically only available through numerical simulation) of the initial conditions. If we have observations at various times $t_1, \ldots, t_K$, we can write the total observational vector $y^T = ({y^T}_1, \ldots, y_K^T)$ as a function of initial condition subject to noise:

$$y = H(x_0) + \eta, \tag{3}$$

where $H(x_0)^T = ([h(x(t_1)]^T, \ldots, h[x(t_K)]^T)$ and $\eta^T = (\eta_1^T, \ldots, \eta_K^T)$. If the probability density function of the random vector η is $p_\eta : \mathbb{R}^{mK} \to \mathbb{R}$, then the conditional probability of the observations y given the initial data x_0 is $p(y \mid x_0) = p_\eta[y - H(x_0)]$. Given the prior distribution p_ζ of initial conditions and *a realization* of the observations $\hat{y}$, the posterior probability for the state vector is obtained from Bayes' theorem:

$$p(x_0|\hat{y}) = \frac{p_\eta[\hat{y} - H(x_0)]p_\zeta(x_0)}{p(\hat{y})} \propto p_\eta[\hat{y} - H(x_0)]p_\zeta(x_0), \tag{4}$$

where

$$p(y) = \int p_\eta[y - H(x_0)]p_\zeta(x_0)\,\mathrm{d}x_0$$

is a function of the observations y alone, and hence $p(\hat{y})$ is a constant for a given realization $\hat{y}$. In particular, the constant of proportionality in eq. (4) does not depend upon x_0.

We note four key points about the above formulation of the DA problem.

(1) The posterior in eq. (4) is the conditional distribution of the state at time $t = 0$ given observations over the time period $[0, t_K]$. This posterior can also be pushed forward to get the conditional distribution of the state at any time $t \in [0, t_K]$. Thus, the data assimilation problem is stated as a smoother, that is, the estimate of the state at some time instant uses observations over a time interval including the future. In contrast, a sequential filter only uses observations in the past. A smoother is natural in many physical contexts, such as the ocean state estimation, re-analysis, etc. (Cohn et al., 1994; Evensen and van Leeuwen, 2000).

(2) The methods we present for sampling the posterior $p(x_0|\hat{y})$ only need a functional form for p_η, which could be non-Gaussian and could also include correlations between observational errors η_k at different times.

(3) We have stated the problem in the perfect model scenario, which can occur only in identical twin experiments but not in any applications. Thus, for the identical twin experiments reported in Sections 3 and 4, we generated the observations using the same model, dynamic as well as observational, as the one that was used in the sampling of the posterior and EnKF. Imperfect model scenarios have been discussed in the context of various existing DA methods (Hansen, 2002; Judd and Smith, 2004) and we are exploring application of the sampling methods we present in this paper to imperfect model problems.

(4) The above formulation of the data assimilation problem does not explicitly mention the 'true' state of the system, $x^{(t)}$. In the data assimilation literature (Cohn, 1997), it is customary to state the observational model in the following form:

$$y = H[x^{(t)}] + \eta.$$

The crucial difference between this statement and the model eq. (3) is that the observation function H in eq. (3) is not a function of the 'true' state $x^{(t)}$ which is unknown in any application.

Thus, while comparing different data assimilation schemes, we do *not* measure the error with respect to the 'true' state. Instead, we take the viewpoint that for a given model [the dynamic model of eq. (1) *along with* the observational model of eq. (3)], the posterior distribution $p(x_0|\hat{y})$ contains all the information available from the model and the given realization $\hat{y}$ of the observations. The success of a DA scheme (either deterministic or statistical) should be measured by comparing its output with this posterior distribution, when the posterior is available. We also note that for comparison with sequential filtering methods, it will be necessary to push $p(x_0|\hat{y})$ forward to the final time, using the Liouville equation associated with eq. (1), to obtain $p(x(t_K)|\hat{y})$.

In applications, the usefulness of this posterior distribution depends on the context. For example, it can be used to generate state estimates of the past (re-analysis) to be compared with past observations, for example, those left out of DA for the purpose of such comparisons; or to perform parameter estimation; or to generate predictions to be then compared with future observations. In all these cases, the structure of the posterior distribution is of interest, since it contains the information from the model and the observations.

We emphasize that the probabilistic nature of the errors in the data necessitates a probabilistic estimate of the state of the system in all the above problems. We cannot a priori ask for a single best estimate of the state. The structure of the posterior will determine, a posteriori, the appropriateness of providing a single best estimate. For instance, if the posterior is unimodal and locally Gaussian-like, then the mode and the covariance can be used as a 'best' estimate of the state and its uncertainty. If the posterior is bimodal, and sharply peaked at each mode, then the location of the two modes, and their relative probabilities, constitute a useful summary of the likely states of the system, and the uncertainty in them.

(5) We have presented the posterior distribution for the case when the model dynamics is deterministic. This is usually termed as a 'strong constrained formulation.' In the case of the so-called 'weak constrained formulation,' the posterior distribution function to be considered is not just on the initial conditions, but rather on the space of paths of the model dynamics. Such a formulation is presented in (Apte et al., 2007) and its relation to various data assimilation techniques, such as 4DVAR, is presented in (Apte et al., 2008). In specific problems of interest, for example, those in oceanography, the strong constrained formulation might be too restrictive and the weak constrained formulation might be preferred.

2. Linearized shallow water model

The idealized ocean model we consider is given by the inviscid linearized shallow water equations, which have the following non-dimensional form (page 68 Pedlosky, 1986).

$$\begin{aligned}\frac{\partial u}{\partial t} &= v - \frac{\partial h}{\partial x},\\ \frac{\partial v}{\partial t} &= -u - \frac{\partial h}{\partial y},\\ \frac{\partial h}{\partial t} &= -\frac{\partial u}{\partial x} - \frac{\partial v}{\partial y},\end{aligned} \tag{5}$$

where $(x, y) \in \mathbb{R}^2$ and $t \in [0, \infty)$. The scalar fields $u(x, y, t)$ and $v(x, y, t)$ are the two components of the velocity field, and $h(x, y, t)$ the variation of the free surface height measured from the mean level. For this linear flow model, it is natural to consider the decomposition of the fields into Fourier modes. In the numerical experiments, we consider two modes:

$$\begin{aligned}u(x, y, t) &= -2\pi l \sin(2\pi kx)\cos(2\pi ly)u_0 + \cos(2\pi my)u_1(t),\\ v(x, y, t) &= 2\pi k \cos(2\pi kx)\sin(2\pi ly)u_0 + \cos(2\pi my)v_1(t),\\ h(x, y, t) &= \sin(2\pi kx)\sin(2\pi ly)u_0 + \sin(2\pi my)h_1(t),\end{aligned} \tag{6}$$

where the first term is a time-independent geostrophic mode with amplitude u_0 and the latter is a time-periodic inertial-gravity mode. On substituting eqs. (6) into eqs. (5), we get the following dynamic equations for the amplitudes.

$$\begin{aligned}\dot{u_0} &= 0,\\ \dot{u_1} &= v_1,\\ \dot{v_1} &= -u_1 - 2\pi m h_1,\\ \dot{h_1} &= 2\pi m v_1,\end{aligned} \tag{7}$$

with initial conditions $[u_0(0),\ u_1(0),\ v_1(0),\ h_1(0)]$. The geostrophic mode exhibits a cellular flow field with hyperbolic fixed points at $(x, y) = (i/2k,\ j/2l),\ i, j \in \mathbb{Z}$, joined by separatrices which prevent mixing between different cells. The time dependent inertial-gravity mode perturbs this cellular structure and can lead to mixing. We chose $k = l = m = 1$ for the numerical experiments. A typical flow field for these modes is shown in Fig. 1.

The observations are the positions $(x_i, y_i), i = 1, \ldots, M$ of M Lagrangian drifters in the above flow. For assimilation of these observations, we use an approach first proposed in Ide et al. (2002), Kuznetsov et al. (2003), augmenting the model with the equations for the drifters.

$$\dot{x_i}(t) = u[x_i(t), y_i(t), t], \qquad \dot{y_i}(t) = v[x_i(t), y_i(t), t], \qquad i = 1, \ldots, M, \tag{8}$$

with initial conditions $x_i(0)$, $y_i(0)$ and the functions u, v given in eq. (6). The drifter observations are made at discrete times $t_k = k\delta$ for $k = 1, \ldots, N$ and contain errors that are assumed to be Gaussian, uncorrelated in time, and independent of each other. Then the observational model can be written as

$$x_i^o(t_k) = x_i(t_k) + \eta_{ik}, \qquad y_i^o(t_k) = y_i(t_k) + \xi_{ik}, \tag{9}$$

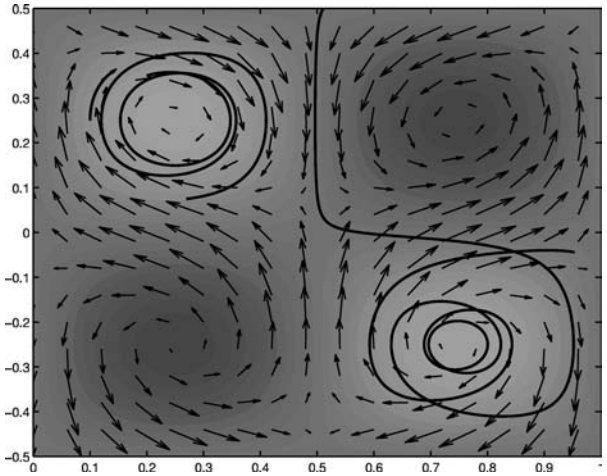

Fig. 1. A typical flow field showing the cellular structure perturbed by the inertial-gravity mode, along with some of the drifter trajectories used in the numerical experiments. The shading denotes the height field and the arrows show the velocity field.

where η_{ik} and ξ_{ik} are $\mathcal{N}(0, R)$ (Gaussian with mean 0 and covariance R) and are independent identically distributed random variables. We re-emphasize that the sampling methods we present can be readily generalized to non-Gaussian errors that are correlated in time and also dependent on each other, as long as we know, or assume, the functional form of their probability distribution function.

Following the discussion in Section 1, the above set-up with a deterministic model and noisy observations naturally reduces the data assimilation problem to that of sampling the posterior distribution of initial conditions. The full model, corresponding to eq. (1), consists of eqs. (7) and (8) for the $4 + 2M$ dimensional state vector $x \equiv (u_0, u_1, v_1, h_1, x_1, y_1, \ldots, x_M, y_M)^T$ of the flow and drifters. Also, writing the $2MN$ dimensional observation and noise vectors $y = [x^o{}_1(t_1), y^o{}_1(t_1), x^o{}_2(t_1), \ldots, y^o_M(t_N)]$ and $\eta = (\eta_{11}, \xi_{11}, \eta_{21}, \ldots, \xi_{MN})$, the observational model can be written in the form of eq. (3) with $x_0 = [u_0(0), u_1(0), v_1(0), h_1(0), x_1(0), y_1(0), \ldots, y_M(0)]$. Thus we will sample the initial conditions of drifters as well as the velocity field, given a single realization $\hat{y}$ of the drifter observations at later times.

We emphasize that we have chosen periodic boundary conditions and the Fourier decomposition eq. (6) for the velocity and height fields. Thus, by using the drifter observations over a certain spatial domain to 'infer' the Fourier components, we get information about the flow over the whole plane $\mathbb{R}^2$. In practical problems with specific boundary conditions, this certainly might not be the case and the 'propagation over space' of the information from drifter observations will raise issues that have not been addressed here.

We solved the flow and the drifter eqs. (7) and (8) using a fourth order Runge–Kutta scheme with a time-step of 10^{-4}. Over the time periods used in our experiments, the error in the flow equations was seen to be negligible by comparing the numerical solutions with the exact solutions of these linear equations. The numerical error in the drifter equations was also seen to be negligible by comparing the trajectories calculated with a much smaller time-step of 10^{-7}.

3. Structure of the posterior

We now study the posterior distributions for various identical twin experiments of the Lagrangian data assimilation problem discussed above. Specifically, we study three cases: a short length of the trajectory of the drifter, a longer trajectory that stays within a cell of the geostrophic flow, and a trajectory that traverses the cellular boundaries. For these different cases, we present the exact posterior obtained using three sampling methods: Langevin stochastic differential equation (LSDE), Metropolis adjusted Langevin algorithm (MALA) and Random walk Metropolis–Hastings (RWMH). The details of these methods and their comparisons are discussed in Section 5.

In Section 4, we compare this posterior with the distribution from the EnKF. The EnKF is implemented as a sequential filter which aims to approximate the exact posterior. Thus, it is natural to compare the exact and the EnKF distributions of the state at the time of the final observation, when the smoother and an ideal filter agree (Evensen and van Leeuwen, 2000). To this end, the posterior samples obtained using the exact sampling methods are pushed forward from the initial time to the final time using the solution operator Φ from eq. (2). This is done using the same numerical scheme as the one used in creating the samples from MALA, RWMH and EnKF. Thus, the distributions shown in this section are those at the final observation time.

We will see in Section 5 that the adaptive version of MALA is the most efficient of the different sampling methods. For all the examples presented in this section, the samples were obtained using this method. In fact, for most of the more complex examples, the non-adaptive algorithms discussed later were inefficient to the point of being almost impossible to use.

3.1. A short trajectory

This trajectory is the short arc in the upper left-hand side of Fig. 1. We observed the position of a drifter at five time instances over the period $t \in [0.005, 0.025]$. The observational error is $\sigma_{\text{obs}} = 0.005$ and the initial true state is $(u_0, u_1, v_1, h_1, x_1, y_1) = (1, 0, 0.5, 0, 0.1, 0.25)$. We use this trajectory to compare the different sampling methods and this comparison (see Fig. 10) is discussed in Section 5.4. Here, we point out that, in the marginals, the posterior is very close to being Gaussian, even though this is the posterior of the state at the final observation time.

We also study the effect on the posterior distribution of varying the prior distribution. Fig. 2 shows the posterior for different priors described in the figure legend. There are a few points to note.

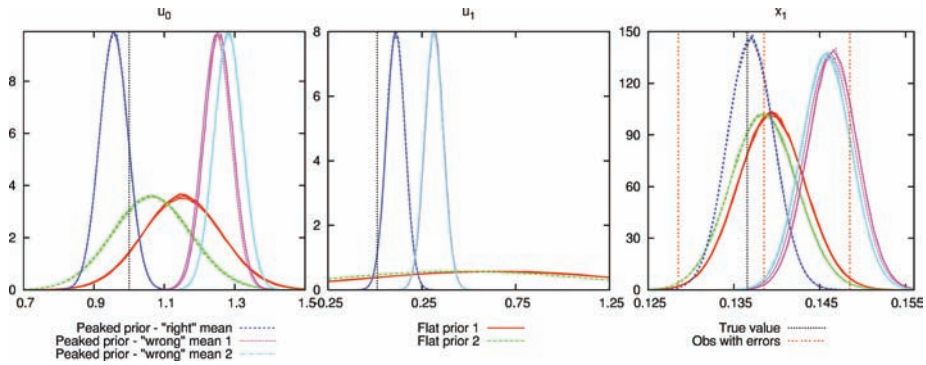

Fig. 2. The marginal posteriors for different priors for the experiment discussed in Section 3.1: the 'flat' priors have covariance of $1.0\mathcal{I}$ each but differ in their means. The 'peaked' priors have covariance of $0.0025\mathcal{I}$. The three different peaked priors differ in the prior mean—'right' mean corresponds to a prior mean close to the 'true' state, 'wrong' mean 1 is far from the 'true' state but the prior mean particle position within the same cell as the initial value, and the 'wrong' mean 2 is also far from the 'true' state with the prior mean particle position in another cell. For clarity, only part of the flat posterior of u_1 is shown. The error shown in x_1 is at the two standard deviations level. In this and other figures showing the marginal posteriors, different curves for each prior correspond to different realizations of MALA.

(1) The prior mean has significant effect on the posterior mean while the posterior variance depends on the prior variance. Of course, when the prior variance is smaller, the effect of the prior mean is stronger.

(2) The dependence of the posterior mean on the prior mean in different variables of state space is significantly different. For example, when the prior is flat, the posterior mean depends notably on the prior mean only in the (u_0, v_1) directions but not others.

(3) Even when the prior is peaked around the 'true' state (peaked prior with 'right' mean), the posterior mean is *not* the same as the true state. But, we see that the posterior mean of the particle position is within $2\sigma_{\text{obs}}$ of the observations, even in the case when the prior is peaked but with a mean which is far from the truth (peaked prior with 'wrong' mean). We will see later that even this need not be the case since the posterior contains information not just about this last observation but about all the previous ones as well.

3.2. A trajectory that stays in a cell

This is the trajectory that circles around in the upper left 'cell' shown in Fig. 1. We used three different observational sets for this trajectory and they are shown in Fig. 3. The observations are taken over a time period, $t \in [0, 0.5]$. The difference between the three observational sets is the number of observations, which is 100, 20 and 6 for the first, second, and third set respectively. The observational error is $\sigma_{\text{obs}} = 0.005$ and the initial true state is $(u_0, u_1, v_1, h_1, x_1, y_1) = (1.0, 0.5, 0.5, 0.5, 0.2, 0.35)$. The prior distribution is assumed to be very flat with a variance of 1.0. The marginal posterior distributions for the state variables u_0, x_1, and y_1 at the final time $t = 0.5$ are shown in Fig. 4. The marginals for the remaining flow variables show very similar patterns. The main conclusions to be drawn from this comparison are the following.

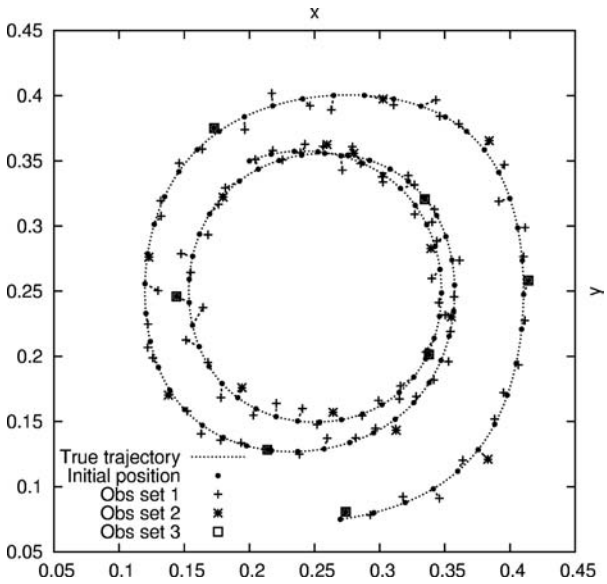

Fig. 3. Three different observational sets for the trajectory discussed in Section 3.2.

(1) As the frequency of observations increases, the posterior becomes more peaked.

(2) Even with very different number of observations, the marginal posteriors remain close to being Gaussian, in spite of the fact that over this time period, the drifter dynamics is highly non-linear.

(3) Even with 100 observations, the posterior mean is not the same as the 'truth.'

3.3. A trajectory that crosses between cells: saddle issues

This is the trajectory that crosses between the two right cells in Fig. 1, crossing close to the saddle point near $(x, y) = (0.5, 0.5)$.

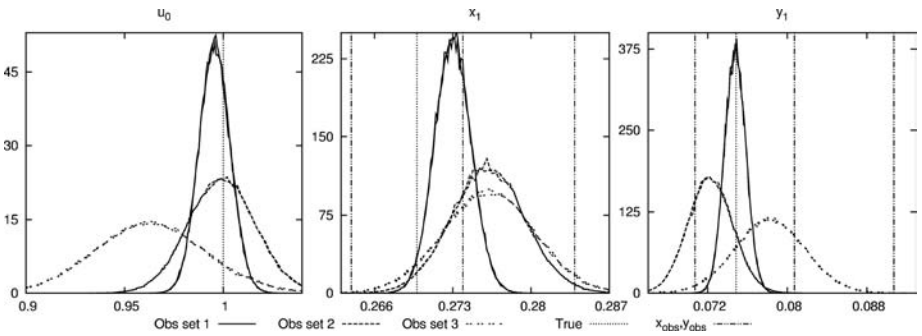

Fig. 4. The marginal posterior distributions using three different observational sets for the trajectory discussed in Section 3.2. The final observation along with an error of $2\sigma_{\text{obs}}$ and the true values are also shown.

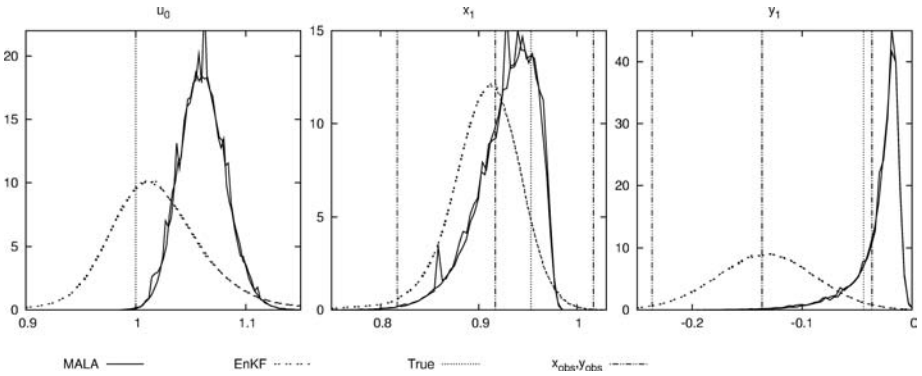

Fig. 5. Comparison of the exact marginal posteriors with those from the EnKF, for the trajectory that crosses between the cells (Section 3.3). The adaptive MALA used for sampling the posterior is not fully converged.

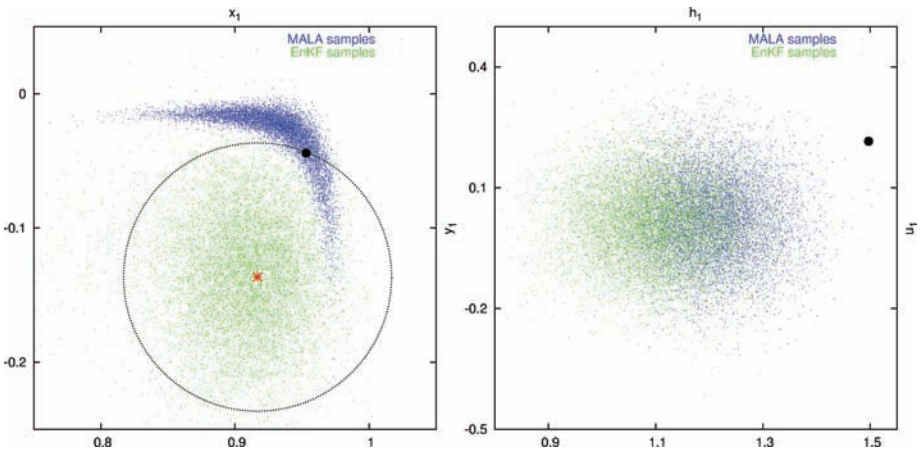

Fig. 6. Scatter plots for the position variables (left-hand panel) and for (h_1, u_1) variables along with the true values ($\bullet$) and a $2\sigma_{\text{obs}}$ ellipse around the observation ($*$) at the final time $t = 1.0$ for the trajectory that crosses between the cells (Section 3.3).

Ten observations are taken over a time period, $t \in [0, 1]$. The observational error variance is $\sigma_{\text{obs}} = 0.05$ and the initial true state is $(u_0, u_1, v_1, h_1, x_1, y_1) = (1.0, 0.2, 1.3, 1.4, 0.51, 0.498)$. The prior distribution is taken to be very flat with a variance of 1.0. The marginal posterior distributions are shown in Fig. 5. The scatter plots for the position variables and for (h_1, u_1) using samples from the exact posterior are shown in Fig. 6. The scatter plots for other variables are qualitatively similar. (For later use in Section 4, this figure also shows the posterior distributions from the EnKF, which should be ignored for the time-being.) This posterior shows a very interesting structure that is clearly affected by the dynamics as well as the observations.

(1) The posterior is clearly non-Gaussian. The effect of the dynamics of the drifters is seen as follows. The geostrophic mode has separatrices at $x_1 = 1$ and $y_1 = 0$. But the true position of

the drifter and its observation at the final time lie in the 'cell' satisfying $x_1 < 1$ and $y_1 > 0$. Correspondingly, there are almost no samples outside that cell. This gives rise to the interesting 'boomerang' shape in the scatter plot for the position (x_1, y_1) of the drifter at the final observation time, as shown in Fig. 6. On the other hand the scatter plot of the posterior distribution for the initial position (not shown here) shows a similar behaviour but near the initial position of the drifter.

(2) We also see that the posterior is very much dependent not just on the final observation but all the earlier ones as well. In contrast to the previous cases studied, cf. Figs. 2 and 4, the marginal in the y_1 position coordinates lies almost entirely outside the $2\sigma_{\text{obs}}$ circle around the final time observation.

(3) Even though the true final position of the drifter is 'within' the marginal posterior, the true final velocity is well 'outside' the posterior. Thus, in this case, the drifter observations fail to give enough information about the flow to estimate the true flow. We re-emphasize that we do not consider this to be a failure of the data assimilation scheme—the 'truth' is only available because of the identical twin experiment set-up and in practice, comparison with 'truth' cannot be made. This leaves open the question of 'consistent comparison of the posterior' with the observations.

4. Comparison with the EnKF

We now compare the posterior implied by the samples from the EnKF, which we will call the 'EnKF posterior,' with the exact posterior for the cases presented above. We implemented the perturbed observation version of the EnKF (described in detail in, for example, Evensen, 2004). Though various different modifications (covariance inflation, deterministic EnKF, localization) could show qualitative improvements (Evensen, 2004), our point of view was that such changes do not address the main shortcoming, which is the assumption of Gaussianity of the prior distribution at each observation time.

The EnKF posterior for the trajectory that crosses a separatrix was shown in Figs. 5 and 6. We see that for this case of a strongly non-Gaussian posterior, the EnKF fails to approximate the correct posterior. The EnKF marginal in the position coordinates is centred around the observation. The prior mean and covariance in the position coordinates before assimilating the final observation were

$$(x_1, y_1)^{\text{prior}}_{\text{mean}} = (0.89313, -0.11918),$$

$$\mathbf{P}^{\text{prior}}_{xy} = \begin{bmatrix} 0.0029 & -0.0005 \\ -0.0005 & 0.0094 \end{bmatrix},$$

and the corresponding posterior values were

$$(x_1, y_1)^{\text{post}}_{\text{mean}} = (0.90613, -0.13344),$$

$$\mathbf{P}^{\text{post}}_{xy} = \begin{bmatrix} 0.0013 & -0.00005 \\ 0.00005 & 0.0020 \end{bmatrix},$$

whereas the observation was $(0.91684, -0.13655)$ with a covariance of $\mathbf{R}_{xy} = \text{diag}\,(0.0025, 0.0025)$. Thus, we see that the prior is already 'close' to the observation and with a comparable covariance. The same effect persisted even with significant covariance inflation.

At the other extreme, the EnKF posteriors for the first example, of a short trajectory with high observational frequency, are almost exactly the same as the exact posteriors shown in Fig. 2. In fact, they *are* included in that figure. Only for the 'peaked prior with wrong mean' (the magenta and cyan lines in Fig. 2), was the mean of the EnKF posterior different from that of the exact posterior by less than a tenth of its standard deviation.

In order to further understand how the EnKF approximates (or fails to approximate) the exact posterior, we now discuss the EnKF posterior for the second example in the previous section. Figs 7 and 8 show the exact and the EnKF posteriors for two of the six variables for the three different observation sets for the longer trajectory shown in Fig. 3, and also for a fourth observation set described below. We recall that the length of the trajectory is the same for the three sets but they differ in the number of observations, and consequently, in the frequency of observations. We note a few key aspects of the EnKF approximation:

(1) When the time interval between observations is small (observation set 1), the EnKF approximates the exact posterior very well.

(2) Increasing the time between observations increases the discrepancy between the EnKF and the exact posteriors, that is, the EnKF approximation becomes worse.

(3) In order to see whether keeping the time period between the observations the same but increasing the number of observations can lead to a better performance for the EnKF, we considered another observational set, called 'Obs set 4' in Figs. 7 and 8. For this set, the observational frequency was $\delta = 0.075$, the same as that in the third set, but the number of observations was $M = 20$, the same as in the second set. We see that the EnKF posterior is still not the same as or close to the posterior from MALA. Hence, the time interval between the observations is seen to be the main factor affecting the performance of the EnKF.

(4) The EnKF approximation seems to fail when the observations are infrequent even though the posterior is close to a Gaussian. This can be understood using Fig. 9 which shows the distributions for the position coordinates x_1, y_1 of the drifter, before and after assimilating the first and second observation. We note that the posterior after assimilating the first observation gives samples that show an approximately Gaussian distribution near the elliptic fixed point at $(x_1, y_1) = (0.25, 0.25)$. But, after evolving these samples up to the next observation time, they form a 'ring' seen in the upper right hand plot. This is a direct consequence of the non-linear evolution equations for the drifters. Applying the Kalman filter with this prior is well beyond the realm of validity of Kalman filter. After assimilating

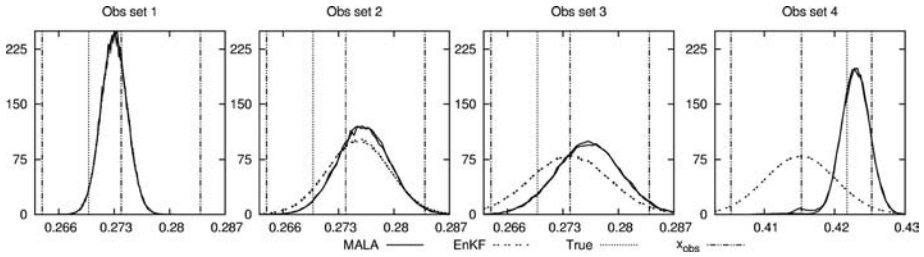

Fig. 7. The posterior for x position of the drifter using the MALA and from the EnKF for four observational sets for the trajectory shown in Fig. 3 (Section 3.2).

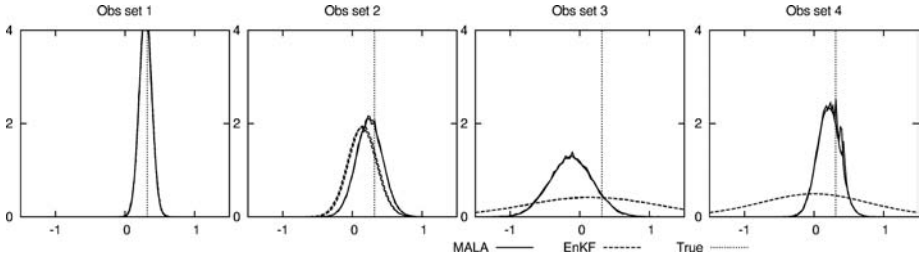

Fig. 8. Same as Fig. 7 but for u_1.

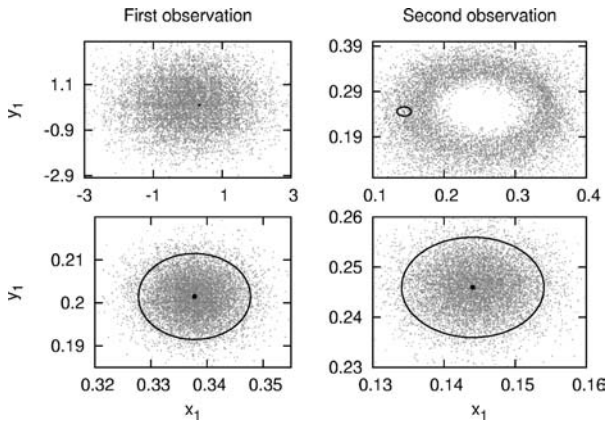

Fig. 9. The scatterplots for the position variables using the EnKF ensemble before (upper row) and after (lower row) after assimilating the first (left-hand column) and the second (right-hand column) observations from the observational set 3 of Fig. 3. The corresponding observations (black dot) and $2\sigma_{\text{obs}}$ circles are shown as well.

the second observation, the posterior is again approximately Gaussian.

We also note that the number of samples needed to get converged distributions using the EnKF is in the range of 10^5–10^6 for the examples presented above. This is comparable to the number of samples needed for convergence of other methods presented in the next section. The computational effort per sample for the EnKF is the same as that for RWMH which is much less than that for MALA. Thus, the computation effort needed to get converged distributions from the adaptive version of RWMH and the EnKF are comparable and indeed, in some cases, RWMH converges faster. It would be interesting and practically relevant to study the differences between the sampling methods (MALA and RWMH) and the EnKF when only small ensemble sizes are used.

5. Sampling methods

In this section, we describe, and compare, various methods used to generate an ensemble of samples from the posterior probability density function. Throughout this section, we will denote this density by $\pi(z)$. Thus, in the notation used in eq. (4), $z \equiv x_0$ and $\pi(z) \equiv p(x_0|\hat{y})$.

5.1. Langevin equation

Given the density $\pi(z)$, consider the LSDE (Robert and Casella, 1999),

$$\frac{\mathrm{d}z}{\mathrm{d}s} = \Lambda \nabla \ln \pi(z) + \sqrt{2\Lambda}\frac{\mathrm{d}W}{\mathrm{d}s},$$

where Λ is any positive definite matrix and W is the standard Brownian motion. The invariant density of this equation is $\pi(z)$. If it is ergodic (a conditions on the tails of π Roberts and Tweedie, 1996), then a single long trajectory of the LSDE will have an empirical density converging to $\pi(z)$. Thus a solution $z(s)$ of the Langevin equation could be used to calculate any statistical quantity that can be calculated using $\pi(z)$. In particular, averages of any function $f(z)$ with respect to π (such as the mean or covariance of π) can be calculated using time averages over the solution $z(s)$ of the Langevin equation:

$$\int f(z)\pi(z)\,\mathrm{d}z = \lim_{T\to\infty}\frac{1}{T}\int_0^T f[z(s)]\,\mathrm{d}s. \tag{10}$$

For the Lagrangian data assimilation problem of Section 2, we solved the above equation using the Euler–Maruyama discretization:

$$z_{n+1} = z_n + \Lambda \nabla \ln \pi(z_n)\delta_s + \sqrt{2\delta_s \Lambda}\ \omega_n, \tag{11}$$

where ω_n are iid $\mathcal{N}(0, I)$. Numerically, the integral in the RHS of eq. (10) is approximated by a sum:

$$\int f(z)\pi(z)\,\mathrm{d}z \approx \frac{1}{N}\sum_{n=1}^{N} f(z_n).$$

In general, for any finite δ_s, the distribution implied by the ensemble $\{z_n\}_{n=1}^N$ generated using eq. (11) does not approach $\pi(z)$ as $N \to \infty$, but rather approaches an approximation $\tilde{\pi}(z; \delta_s)$ (Talay, 1995). Thus, it is necessary to trade the finite sample size error with the approximation error to optimize the algorithm. The Metropolis–Hastings methods of the next section do not suffer from this issue since, by construction, they sample the desired density π in stationarity. Furthermore, we found that the Metropolis–Hastings methods are more efficient at sampling the distribution than the Euler-discretized Langevin equation.

5.2. Metropolis–Hastings algorithms

We use two different methods for Metropolis–Hastings sampling. In the MALA (Robert and Casella, 1999; Roberts and Rosenthal, 2001), the proposal is given by the Euler discretization of the LSDE, eq. (11):

$$z^* = z_n + \Lambda \nabla \ln \pi(z_n)\delta_s + \sqrt{2\delta_s \Lambda}\ \omega_n. \tag{12}$$

In the RWMH (Robert and Casella, 1999), the proposal is

$$z^* = z_n + \sqrt{2\delta_s \Lambda}\ \omega_n, \tag{13}$$

In both cases, $z^* \sim \mathcal{N}(\mu(z), \Sigma)$, that is, the joint PDF $q(z, z^*)$ of the proposed state z^* and the current state z is

$$q(z, z^*) \propto \exp\left[-\frac{1}{2}\|z^* - \mu(z)\|_\Sigma^2\right]. \tag{14}$$

Here, $\|z\|_\Sigma^2 = z^T \Sigma^{-1} z$, $\Sigma = 2\delta_s\ \Lambda$ and for MALA $\mu(z) = z + \Lambda \nabla \ln \pi(z)\ \delta_s$ while for RWMH $\mu(z) = z$. The standard Metropolis–Hastings criterion (Robert and Casella, 1999) is used for accepting or rejecting the proposed state: if $\alpha = \min\{[\pi(z^*)q(z^*, z_n)/\pi(z_n)q(z_n, z^*)], 1\} > u_n$, then $z_{n+1} = z^*$, otherwise $z_{n+1} = z_n$, where $u_n \sim U(0, 1)$ are iid uniform random variables. We also implemented the adaptive version of these algorithms introduced in (Atchade and Rosenthal, 2005; Atchade, 2006). In the adaptive algorithms, the 'proposal mean and covariance' μ and Λ as well as the 'time step' δ_s in eqs. (12)–(13) is adapted at each step in the following manner.

$$\begin{aligned}
\mu_{n+1} &= \mu_n + \gamma_n(z_n - \mu_n),\\
\Lambda_{n+1} &= \Lambda_n + \gamma_n[(z_n - \mu_n)(z_n - \mu_n)^T - \Lambda_n],\\
\delta_{s,n+1} &= \delta_{s,n} + \delta_{s,n}\gamma_n(\alpha - \tau).
\end{aligned}$$

Here, γ_n is a sequence, $\gamma_n = c_0/n$ with a constant $c_0 \sim O(1)$. This gives an asymptotic optimal acceptance rate τ for these Markov chains: the algorithm learns the covariance structure as the chain progresses. The detailed description can be found in (Atchade, 2006).

5.3. Computational cost of different sampling methods

Note that for $z = x_0$ and $\pi(z) = p(x_0|\hat{y})$ given in eq. (4), the drift, the second term in eqs. (11) and (12), is given by

$$\nabla_{x_0} \ln p(x_0 \mid \hat{y}) = \frac{-1}{p_\eta(x_0 \mid \hat{y})}\sum_{k=1}^{M} \nabla_{x_0}\big\{p_\eta[\hat{y} - h(x_k)]\big\} + \nabla p_\xi(x_0), \tag{15}$$

where $x_k(x_0) \equiv x(t_k) = \Phi(x_0; t_k)$ and the derivative under the sum requires calculation of $\nabla_{x_0} x_k(x_0)$ and of $\nabla_{x_k} h(x_k)$. For the dynamics given by eq. (1), we see that $L(t) \equiv \nabla_{x_0} x(x_0) = \partial x(t)/\partial x_0$ satisfies the equation

$$\frac{\mathrm{d}L(t)}{\mathrm{d}t} = \nabla f[x(t)]\ L(t), \qquad L(0) = \mathcal{I}. \tag{16}$$

In our numerical implementation, we solved the above equation using a fourth order Runge-Kutta scheme with a time-step of 2×10^{-4}, giving comparable accuracy as that for eqs. (7)–(8).

Since $L(t)$ is a $n \times n$ matrix, each step of the Euler-discretized Langevin equation and each sample from MALA require integration of the $n + n^2$ coupled equations (1) and (16) over the time interval $t \in [0, t_M]$. Since the proposal density $\pi(z)$ does not require evaluation of $L(t)$, each sample of the RWMH requires integrating only the n eq. (1). Similarly for EnKF, each sample requires integrating these n coupled equations from the initial time to the final. Thus, we see that the computational effort per sample for RWMH and EnKF increases linearly with the dimension n of the state space while it scales quadratically for MALA. On the other hand, we will see in Section 5.4 that the number of samples required for convergence is significantly less for MALA than for RWMH. Hence the actual computational effort required for convergence of these algorithm depends on the dimension of state space and the time interval over which observations are taken.

It would be interesting to implement the various approximations used for 4D-VAR (Courtier et al., 1994) to compute $L(t)$ approximately and study the effect of these approximations on samples generated by MALA.

5.4. Convergence of different sampling methods

Figure 10 shows the comparison between different methods for sampling from the first of the three numerical experiments described in Section 3: a short trajectory containing five observations of a single drifter position. The figure shows the histograms for the u_0 (left), u_1 (centre), and x_1 (right) variables using the samples obtained by different methods. The top row uses samples from the Euler-discretized LSDE while the second row is from

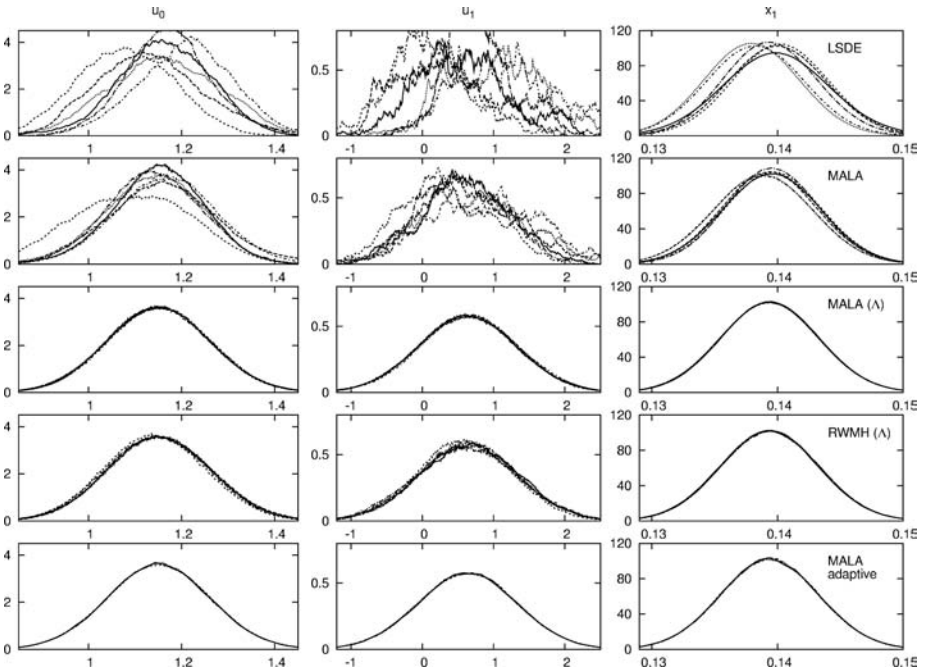

Fig. 10. The marginal posterior distributions for u_0 (left-hand panels), u_1 (centre panels), and x_1 (right-hand panels) using samples from various different methods, explained in detail in the text. The different curves in each plot are two noise realizations for three different starting values. The third and fifth row show the converged posterior distribution.

MALA, both using $\Lambda = \mathrm{Id}$. The distribution using MALA with $\Lambda = \mathrm{diag}\,\{10, 100, 100, 100, 1, 1\}$ is shown in the third row, while the fourth row uses samples from RWMH using $\Lambda = \mathrm{diag}\,\{20, 20, 20, 20, 1, 1\}$. The number of samples is 10^7 and the step sizes δ_s (adjusted to give acceptance rates of around 50% for MALA and 25% for RWMH) are 10^{-6}, 7×10^{-6}, 7×10^{-6}, and 1.5×10^{-5} respectively for these top four rows. The bottom row shows the histogram from samples obtained using the adaptive MALA. The number of samples for the adaptive case is 5×10^5, twenty time smaller than the non-adaptive cases from the top four rows.

We note several important conclusions, though these could be possibly very problem specific.

(1) The Euler-discretized LSDE with $\Lambda = \mathcal{I}$ (top row) is quite slow at sampling the distribution and it is unclear whether it is converging to the correct distribution.

(2) The MALA with $\Lambda = \mathcal{I}$ (second row), which requires almost the same computational effort as the Langevin equation, is significantly better.

(3) The MALA with a proposal covariance of $\Lambda = \mathrm{diag}\,\{10, 100, 100, 100, 1, 1\}$ (third row) shows a significant improvement in convergence compared to the above two cases. The number of samples is the same (10^7) in the first three, and also the fourth, row. But we see that only MALA with non-identity Λ shows the converged posterior distributions. With significantly more number of samples, the other methods, with the possible exception of Euler-discretized LSDE, also show converged posteriors that are the same as that shown in this row. Comparison of RWMH with different proposal covariances Λ also shows that significant improvement in convergence is achieved by use of an appropriately chosen Λ.

(4) The fourth row shows the distribution from the RWMH, with a proposal covariance of $\Lambda = \mathrm{diag}\,\{20, 20, 20, 20, 1, 1\}$. Comparing these distributions with those in the third row, we see that RWMH converges more slowly, requiring more samples to show a converged distribution, than MALA.

(5) The bottom row shows the converged posterior, which is of course the same as that from the third row, using the adaptive version of MALA, but using only 5×10^5 samples. This is the most efficient of all methods for all the different numerical experiments we performed.

The results for other numerical experiments are similar. In fact, for most of the other cases, the non-adaptive methods did not fully converge even with 10^7 samples and only the adaptive method gave a converged distribution. A 'good' proposal covariance always dramatically improved the convergence and the good guesses turned out to be close to the covariance of the distribution π to be sampled. Thus, a general conclusion seems to be that sampling of the posterior distributions in these Lagrangian data assimilation problems requires a good guess of the structure of the posterior distribution, which is of course not known a priori.

The main advantage of the adaptive algorithm for our problem is that we do not have to make any such guesses about π. The

mean and the covariance of the proposal, and the step size of the adaptive algorithm, are adjusted at each step. It was observed that the final proposal mean and covariance of the adaptive algorithm were very close to the mean and covariance of π while the final step size was usually larger than the one chosen for the 'optimal' non-adaptive version. In a sense, the adaptive algorithm learns the covariance structure of the density to be sampled.

6. Conclusions and discussion

In this paper, we study the Lagrangian data assimilation problem from a Bayesian viewpoint. We study the posterior distribution of the state of the system given its observations, a dynamic model, and a model for the noise in the observations, in the context of the linearized shallow water velocity field. In a perfect model scenario, which we discussed in this paper, the ensemble of samples from this posterior is the optimal ensemble to use in data assimilation. Such an ensemble gives us information about the variability in our estimation of the initial conditions of the system, given the specific realization of the observations. This posterior distribution shows interesting structures that are affected by the dynamics of the model as well as the observations assimilated in that distribution.

We compared this posterior to the posterior distribution implied by the ensemble Kalman filter. The main factor affecting the performance of the EnKF is seen to be the time interval between the observations—the longer this interval, the worse is the approximation by the EnKF. In the Lagrangian data assimilation problem, the presence of the centre, that is, an elliptic fixed point of the flow, is seen to give rise to strongly non-Gaussian distributions that lead to the failure of the EnKF.

We used three different methods for sampling this posterior distribution: the LSDE; MALA and RWMH. The comparison of these methods lead us to conclude that the adaptive versions of the Metropolis–Hastings algorithms are the most efficient at sampling, at least in the Lagrangian data assimilation problems we studied.

The adaption of these sampling techniques to large complex models presents significant computational challenges. There are three distinct issues that need further consideration:

(1) The MALA method requires computation of the matrix $L(t)$ which in turn requires the linearized dynamics, cf. eq. (16). The use of approximations, such as those used in 4DVAR, would be essential to speed up the sampling and make a practical method for high-dimensional problems. It would be necessary to study the effects of these approximations on the sampling techniques.

(2) We have used a very large ensemble to get the posterior distributions. The use of smaller ensembles will of course speed up the computations. A significant challenge is to adapt these methods to get a faithful representation of the posterior even with smaller numbers of samples.

(3) The sampling techniques discussed above require a functional form of the prior distribution, cf. eq. (15). It would be useful to adapt these techniques to the cases when only samples from the prior are given. Then, the posterior samples generated using data over an earlier time period can be used as prior samples for assimilating subsequent observations.

Apart from the above computational challenges, there are two important conceptual issues that need further consideration.

(1) We have discussed the posterior distribution given observations of the system over a finite time period. A major conceptual challenge is to understand the properties of this posterior when we take into account observations over longer and longer periods of time and to study the limiting posterior distribution.

(2) Due to the sensitive dependence on initial conditions inherent in problems of this kind, it is likely that imposing the dynamics as a strong constraint will make the posterior distribution hard to sample—its support may be essentially concentrated on low dimensional structures, highly stretched along certain directions, for example. Understanding the effect of weak constraints in the amelioration of such effects is thus of great interest in this fully Bayesian context. In essence this means understanding the effect of adding noise models to the dynamic equations in the probability model, meaning that the posterior is now concentrated on time-dependent solutions, not just initial conditions. The mathematical framework for incorporation of noise in the model is outlined in Apte et al. (2007). Some preliminary numerical experiments are described in Apte et al. (2008). Furhter work to understand the relative merits of weak versus strong constraints is called for.

(3) As we pointed out in the introduction, the identical twin experiments presented in this paper assume perfect model scenario. The application of this method to a real system, when the observations are not taken from another model run, will necessarily require studying the imperfect model scenario. The interpretation of the posterior in these scenarios, and especially in real applications when the 'true' state is not available, is a major conceptual challenge. In the context of Lagrangian data assimilation, we are exploring the use of the sampling techniques in the case when the observations are taken from the dynamics of an inertial particle with small mass whereas the model used in data assimilation is that of a Lagrangian particle. The use of the above sampling techniques for assimilating data from a physical system, not from a perfect or imperfect model, will certainly provide an understanding of the relation between the information contained in the model dynamics and in the data.

7. Acknowledgments

AA and CKRTJ would like to acknowledge the support of ONR (grant number N00014-04-1-0215) and SAMSI (grant number

03-SC-NSF-1009). AS would like to acknowledge the support of ONR. The authors also like to thank Jie Yu for various discussions about linear shallow water equations.

References

Apte, A., Hairer, M., Stuart, A. and Voss, J. 2007. Sampling the posterior: an approach to non-gaussian data assimilation. *Physica D* **230**, 50–64.

Apte, A., Jones, C. K. R. T., Stuart, A. M. and Voss, J. 2008. Ensemble data assimilation. *Int. J. Numer. Methods Fluids*, in press.

Atchade, Y. 2006. An adaptive version for the metropolis adjusted langevin algorithm with a truncated drift. *Methodol. Comput. Appl. Probab.* **8**, 235–254.

Atchade, Y. and Rosenthal, J. 2005. On adaptive Markov chain Monte Carlo algorithms. *Bernoulli* **11**, 815–828.

Carter, E. 1989. Assimilation of Lagrangian data into a numerical model. *Dyn. Atmos. Oceans* **13**, 335.

Cohn, S. E. 1997. An introduction to estimation theory. *J. Met. Soc. Japan* **75**, 257–288.

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

Courtier, P., Thépaut, J.-N. and Hollingsworth, A. 1994. A strategy for operational implementation of 4D-VAR. *Quart. J. Roy. Meteo. Soc.* **120**, 1367–1387.

Evensen, G. 2004. Sampling strategies and square root analysis schemes for the enkf with correction. *Ocean Dyn.* **54**, 539–560.

Evensen, G. and van Leeuwen, P. J. 2000. An ensemble kalman smoother for nonlinear dynamics. *Mon. Wea. Rev.* **128**, 1852–1867.

Hansen, J. 2002. Accounting for model error in ensemble-based state estimation and forecasting. *Mon. Wea. Rev.* **130**, 2373–2391.

Ide, K., Kuznetsov, L. and Jones, C. 2002. Lagrangian data assimilation for point-vortex system. *J. Turbulence* **3**, 053.

Judd, K. and Smith, L. A. 2004. Indistinguishable states II: the imperfect model scenario. *Physica D* **196**, 224–242.

Kuznetsov, L., Ide, K. and Jones, C. 2003. A method for assimilation of Lagrangian data. *Mon. Wea. Rev.* **131**, 2247.

Molcard, A., Piterbarg, L., Griffa, A., Özgökmen, T. and Mariano, A. 2003. Assimilation of drifter positions for the reconstruction of the Eulerian circulation field. *J. Geophys. Res.* **108**, 3056.

Özgökmen, T., Molcard, A., Chin, T., Piterbarg, L. and Griffa, A. 2003. Assimilation of drifter positions in primitive equation models of mid-latitude ocean circulation. *J. Geophys. Res.* **108**, 3238.

Pedlosky, J. 1986. *Geophysical Fluid Dynamics*, Springer, New York, USA.

Robert, C. and Casella, G. 1999. *Monte Carlo Statistical Methods*, Springer, New York.

Roberts, G. and Rosenthal, J. 2001. Optimal scaling for various Metropolis-Hastings algorithms. *Stat. Sci.* **16**, 351–367.

Roberts, G. and Tweedie, R. 1996. Exponential convergence of langevin distributions and their discrete approximations. *Bernoulli* **2**, 341–363.

Talay, D. 1995. Simulation and numerical analysis of stochastic differential systems: a review. In: *Probabilistic Methods in Applied Physics* Vol. 451 of Lecture Notes in Physics, Chapter 3 (eds P. Kree and W. Wedig). Springer-Verlag, Berlin, Germany., 63–106.