Nonlin. Processes Geophys., 20, 1047–1060, 2013
www.nonlin-processes-geophys.net/20/1047/2013/
doi:10.5194/npg-20-1047-2013

Nonlinear Processes in Geophysics
Open Access

A potential implicit particle method for high-dimensional systems

B. Weir, R. N. Miller, and Y. H. Spitz

College of Earth, Ocean, and Atmospheric Sciences, Oregon State University, Corvallis, OR 97331, USA

Correspondence to: B. Weir (bweir@oce.orst.edu)

Received: 9 May 2013 – Revised: 9 September 2013 – Accepted: 21 October 2013 – Published: 28 November 2013

Abstract. This paper presents a particle method designed for high-dimensional state estimation. Instead of weighing random forecasts by their distance to given observations, the method samples an ensemble of particles around an optimal solution based on the observations (i.e., it is implicit). It differs from other implicit methods because it includes the state at the previous assimilation time as part of the optimal solution (i.e., it is a lag-1 smoother). This is accomplished through the use of a mixture model for the background distribution of the previous state. In a high-dimensional, linear, Gaussian example, the mixture-based implicit particle smoother does not collapse. Furthermore, using only a small number of particles, the implicit approach is able to detect transitions in two nonlinear, multi-dimensional generalizations of a double-well. Adding a step that trains the sampled distribution to the target distribution prevents collapse during the transitions, which are strongly nonlinear events. To produce similar estimates, other approaches require many more particles.

1 Introduction

Most particle filters perform poorly in very high dimensions. Their ensembles collapse onto a single particle unless the ensemble size grows exponentially with the system dimension. This is a problem of sample impoverishment, and is a manifestation of what Bellman (1957) calls “the curse of dimensionality”.

The bootstrap particle filter (BPF; Gordon et al., 1993) is a straightforward method that weighs random solutions of the dynamical model based on their proximity to observations. Even if the model and observation functions are linear and have Gaussian errors, BPF suffers from ensemble collapse as the system dimension increases (Bengtsson et al., 2008; Bickel et al., 2008; Snyder et al., 2008). There is also evidence (Snyder, 2012) that this result is more generally applicable and all particle filters suffer a similar fate.

One approach that has improved the performance of particle methods is to use mixture models rather than discrete approximations of probability distributions. This idea began with the work of Alspach and Sorenson (1972). Since then, Anderson and Anderson (1999), Chen and Liu (2000), Bengtsson et al. (2003), Kotecha and Djurić (2003), Smith (2007), Hoteit et al. (2008), Dovera and Rossa (2011), Stordal et al. (2011), Reich (2012), Frei and Künsch (2013), Sondergaard and Lermusiaux (2013a, b) and many others have developed similar approaches. In particular, all of these techniques followed from adaptations of BPF or the ensemble Kalman filter (EnKF; Evensen, 1994, 2009).

Both BPF and EnKF begin by generating an ensemble of random model forecasts that are independent of the observations. The resulting estimates are linear combinations of the forecasts, where the coefficients depend on the likelihood that the forecast produced the observations. This paper refers to such methods as explicit, in analogy with the terminology from the numerical solution of differential equations.

Explicit data assimilation methods are prone to errors when the forecast distribution is nearly singular with the distribution conditioned on the observations, also called the target or posterior. For example, this occurs when the model has multiple isolated attracting states and none of the forecasts are in the basin of attraction of the true state (Miller et al., 1999; Evensen and van Leeuwen, 2000). The singularity can be significantly reduced, however, if the stochastic model has an invariant measure (a climatology). This point is the basis of the mean field filter, maximum entropy filter, and related techniques (Eyink and Restrepo, 2000; Kim et al., 2003; Eyink et al., 2004; Eyink and Kim, 2006), which form a parametrized transformation of the model climatology with the same mean, and possibly covariance, as the forecast samples. Nevertheless, there are many stochastic processes that do not have invariant measures, including the ubiquitous Wiener process and a related example considered later in this paper.

Implicit methods, unlike explicit ones, skip the construction of a forecast distribution and work directly with the target. Their goal is to sample an "optimal" importance distribution whose difference with the target is minimal (Doucet et al., 2000). This approach has a strong theoretical basis and is effective in a variety of contexts, particularly in low to moderate dimensional problems from the geosciences (Chorin and Tu, 2009; Chorin et al., 2010; Morzfeld et al., 2012; Morzfeld and Chorin, 2012; Weir et al., 2013; Atkins et al., 2013). In these applications, the implicit methods require a factor of $O(10)$ to $O(100)$ fewer particles than BPF and EnKF to compute estimates of comparable accuracy (Morzfeld and Chorin, 2012; Weir et al., 2013).

The implicit particle method introduced in this paper avoids ensemble collapse in high dimensions in three ways. First, it forms a mixture model approximation of the background distribution of the state at the previous assimilation time. Second, it uses numerical optimization to find the most probable model solution given the observations and samples around that solution. Third, if the target/posterior distribution is strongly non-Gaussian, it further improves the results by refining the sampled importance distribution to better approximate the target.

While it is possible to improve the estimates of BPF and EnKF significantly, e.g., using Markov chain Monte Carlo resampling methods (Weare, 2009), running-in-place (Kalnay and Yang, 2010), the finite-size EnKF (Bocquet, 2011), and iterative EnKF (Bocquet and Sakov, 2012), only their simplest forms are considered here. Some particle filters (van Leeuwen, 2010; van Leeuwen, 2011; Ades and van Leeuwen, 2013) do, in fact, perform well in high dimensions. Yet it remains unclear if these approaches satisfy the tail decay properties necessary for convergence (Geweke, 1989). Surveys of many other assimilation techniques can be found in the reviews of van Leeuwen (2009) and Bocquet et al. (2010).

The remainder of the paper proceeds as follows. The state estimation problem is introduced next. After that, Sect. 3 describes the mixture-based implicit particle smoother (MIPS) in a general form. This method is applied to a high-dimensional example with a linear model and Gaussian statistics in Sect. 4 and to multi-dimensional generalizations of the double-well problem in Sect. 5. Both sections include comparisons with BPF, EnKF, and the implicit particle filter. The final section summarizes the results and conclusions from these examples. Throughout, vectors are written in bold italics, matrices in regular bold, random variables in capital letters and their realizations in lowercase letters.

2 State estimation

The system state is a stochastic process $\{\boldsymbol{X}_m : m = 0, 1, \ldots\}$ in N_x dimensions that satisfies the equation

$$\boldsymbol{X}_{m+1} = \mathcal{M}(\boldsymbol{X}_m, \boldsymbol{X}_{m+1}) + \sqrt{\mathbf{Q}}\boldsymbol{E}_{m+1}, \tag{1}$$

where $\mathcal{M}$ is a discrete-time dynamical model, $\{\boldsymbol{E}_m\}$ is a (dimensionless) standard normal/Gaussian process, which represents model error, $\mathbf{Q}$ is the (dimensional) error covariance matrix and $\sqrt{\mathbf{Q}}$ is any square root of $\mathbf{Q}$, i.e., $\sqrt{\mathbf{Q}}\sqrt{\mathbf{Q}}^T = \mathbf{Q}$. The model dependence on the new state $\boldsymbol{X}_{m+1}$ is included to account for implicit numerical time discretizations (Kloeden and Platen, 1999). In general, the initial condition is imprecisely known, and the value of its probability density function (pdf) at a realization $\boldsymbol{x}_0$ is denoted $p(\boldsymbol{x}_0)$.

The stochastic model is supplemented with noisy observations at a subsequence $\{t_{\mathrm{m}(n)} : n = 1, 2, \ldots\}$ of the model times such that

$$\boldsymbol{Y}_n = \mathcal{H}\left(\boldsymbol{X}_{\mathrm{m}(n)}\right) + \sqrt{\mathbf{R}}\boldsymbol{D}_n \tag{2}$$

for a given function $\mathcal{H}$, (dimensionless) observation error process $\{\boldsymbol{D}_n\}$ and (dimensional) covariance matrix $\mathbf{R}$. The goal of data assimilation is to efficiently sample from the distribution of model solutions conditioned on a sequence of realizations $\{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_k\}$ of the observations (2) at successive times. The pdf of this stochastic process is denoted $p\left(\boldsymbol{x}_{0:\mathrm{m}(k)} \,|\, \boldsymbol{y}_{1:k}\right)$, which uses the shorthand $z_{i:j}$ for a given sequence $\{z_i, \ldots, z_j\}$.

It is possible to assimilate each new observation and discard it afterward because the target pdf satisfies the recursion relationship

$$p(\boldsymbol{x}_{0:\mathrm{m}(k+1)} \,|\, \boldsymbol{y}_{1:k+1}) \propto p\left(\boldsymbol{x}_{0:\mathrm{m}(k)} \,|\, \boldsymbol{y}_{1:k}\right) \cdot p\left(\boldsymbol{x}_{\mathrm{m}(k)+1:\mathrm{m}(k+1)} \,|\, \boldsymbol{x}_{\mathrm{m}(k)}\right) \cdot p\left(\boldsymbol{y}_{k+1} \,|\, \boldsymbol{x}_{\mathrm{m}(k+1)}\right). \tag{3}$$

This follows from an application of Bayes' theorem, the Markov property of the state, the conditional independence of the observation errors, and a second application of Bayes' theorem. Using the convention that $\mathrm{m}(0) = 0$, and $\boldsymbol{y}_{1:0} = \emptyset$, Eq. (3) applies if $k = 0$ as well.

The model error $\boldsymbol{E}_m$ and observation error $\boldsymbol{D}_n$ need not be Gaussian in general. However, this paper assumes the a priori application of an anamorphosis transformation (Bertino et al., 2003; Weir et al., 2013) to the state and observation so that the corresponding errors are Gaussian random variables.

2.1 The effective dimension

The state of nearly every geophysical model is a collection of variables, e.g., velocity, pressure and temperature, evaluated at each point of a grid. Since the number of grid points is often $O(10^6)$ or greater, the state dimension is high as well. Fortunately, the effective dimension (Bickel et al., 2008) of the problem is usually much smaller. For example, the model can have a low-dimensional attractor, or states and observations separated by large distances can have negligible correlations. Dimensional reduction takes advantage of the smaller effective dimension by projecting the problem onto the effective subspace. There are a number of different techniques, including dynamically orthogonal decomposition (Sondergaard and Lermusiaux, 2013a, b), localization (many variations exist, but Ott et al., 2004, is one of the

most well known), and partial noise reduction (Morzfeld and Chorin, 2012).

The second fundamental assumption of this paper is the a priori application of any possible dimensional reduction, and hence that the eigenvalues of the model and observation error covariance matrices are bounded away from zero. Although the reduced model can have many possible forms, the lowest frequency mode of a climate model is quite often a double-well, i.e., a nonlinear model with two stable fixed points (Majda et al., 2003; Kravtsov et al., 2005). Under the influence of stochastic perturbations, its solutions transition periodically between these two points. It is perhaps the simplest energy balance model capable of reproducing the transitions characteristic of global temperature records (Sutera, 1981). The models considered below combine double-wells and linear maps to extend this scenario to problems in multiple dimensions with multiple attracting states.

3 The mixture-based implicit particle smoother

The assimilation technique to follow is a modification of the implicit particle filter (IPF) introduced by Chorin and Tu (2009) and extended to parameter estimation by Weir et al. (2013). In the latter, the method is continued sequentially by constructing a kernel density estimate (Silverman, 1986) of the background distribution of the model parameters. In this paper, the previous state $\boldsymbol{x}_{\mathrm{m}(k)}$ plays the role of the model parameters. Although it is successful in examples where EnKF fails, the kernel-based implicit approach requires a relatively large number of particles, $O(1000)$, to avoid collapse in a $O(10)$ dimensional sample space. One possibility is that this requirement is primarily due to the deficiencies of the kernel density estimate.

As an alternative to kernel density estimates, Sondergaard and Lermusiaux (2013a, b), referred to as SL13 from now on, suggest using a mixture model (McLachlan and Peel, 2001). The mixture-based implicit particle smoother (MIPS) does exactly that. It differs from the approach of SL13 in two ways: it constructs a mixture approximation of the background of the previous state $\boldsymbol{x}_{\mathrm{m}(k)}$ rather than the next state $\boldsymbol{x}_{\mathrm{m}(k+1)}$, and it uses optimization to find probable model solutions rather than using an analysis step based on the Kalman filter.

Figure 1 is a graphical comparison of a one-component mixture model and kernel density estimate of a standard normal density. Even in two dimensions, the kernel density estimate requires $O(1000)$ samples to have any visual similarity with the true density. The mixture model approximation with $O(100)$ samples, on the other hand, is comparable to the true density. Both estimates are quite poor with 10 samples, and their errors only increase as the dimension grows (McLachlan and Peel, 2001; Silverman, 1986). Given just a handful of samples in very many dimensions, it is thus unlikely that any representation of the true density is very accurate. In this case, it is often appropriate to use a one-component mixture model, since the statistical evidence against the Gaussianity of the true distribution is minimal (e.g., the multivariate normality test of Mardia, 1974).

3.1 Gaussian mixture models

The assimilation of the $(k+1)$-th observation begins with an ensemble of N_p particles resulting from the k-th assimilation:

$$\left\{\boldsymbol{x}^{(i)}_{\mathrm{m}(k)} \sim p\left(\boldsymbol{x}_{\mathrm{m}(k)} \,|\, \boldsymbol{y}_{1:k}\right) : i = 1, \ldots, N_p\right\}.$$

Given these samples, one may compute an approximation that is a mixture of N_m Gaussian components,

$$\begin{aligned}\hat{p}\left(\boldsymbol{x}_{\mathrm{m}(k)} \,|\, \boldsymbol{y}_{1:k}\right) &= \sum_{j=1}^{N_m} \alpha_j \mathcal{N}\left(\boldsymbol{x}_{\mathrm{m}(k)}; \boldsymbol{\mu}_j, \mathbf{B}_j\right), \\ &\approx p\left(\boldsymbol{x}_{\mathrm{m}(k)} \,|\, \boldsymbol{y}_{1:k}\right),\end{aligned} \tag{4}$$

where the weight α_j, mean $\boldsymbol{\mu}_j$ and covariance $\mathbf{B}_j$ of the components are all estimated from the samples.

Here, the only assumption on $\mathbf{B}_j$ is that it is symmetric and positive-semidefinite. In the case that $\mathbf{B}_j$ is not positive definite, its inverse is taken as the Moore–Penrose pseudoinverse (Moore, 1920; Penrose, 1951) and its determinant as the product of its non-zero eigenvalues. If $\mathbf{B}_j$ is an $N_x \times N_x$ matrix of all zeros, then $\mathcal{N}(\boldsymbol{\mu}_j, \mathbf{B}_j)$ denotes the Dirac delta function at $\boldsymbol{\mu}_j$.

Following SL13, MIPS uses the expectation-maximization (EM) algorithm (Dempster et al., 1977; McLachlan and Krishnan, 2008) to find the maximum likelihood estimate (ML/MLE) of α_j, $\boldsymbol{\mu}_j$, and $\mathbf{B}_j$. At iteration n, the EM update is computed in two steps:

1. Expectation,

$$\begin{aligned}\tau^{(i)}_{j,n} &= \frac{\alpha_{j,n}\mathcal{N}\left(\boldsymbol{x}^{(i)}_{\mathrm{m}(k)}; \boldsymbol{\mu}_{j,n}, \mathbf{B}_{j,n}\right)}{\sum_l \alpha_{l,n}\mathcal{N}\left(\boldsymbol{x}^{(i)}_{\mathrm{m}(k)}; \boldsymbol{\mu}_{l,n}, \mathbf{B}_{l,n}\right)}, \\ T_{j,n} &= \sum_i \tau^{(i)}_{j,n},\end{aligned}$$

2. Maximization,

$$\begin{aligned}\alpha_{j,n+1} &= \frac{1}{N_p} T_{j,n}, \\ \boldsymbol{\mu}_{j,n+1} &= T^{-1}_{j,n} \sum_i \tau^{(i)}_{j,n} \boldsymbol{x}^{(i)}_{\mathrm{m}(k)}, \\ \mathbf{B}_{j,n+1} &= T^{-1}_{j,n} \sum_i \left[\tau^{(i)}_{j,n}\left(\boldsymbol{x}^{(i)}_{\mathrm{m}(k)} - \boldsymbol{\mu}_{j,n+1}\right)\right. \\ &\qquad \left.\left(\boldsymbol{x}^{(i)}_{\mathrm{m}(k)} - \boldsymbol{\mu}_{j,n+1}\right)^T\right].\end{aligned}$$

This is one of many possible choices of clustering methods (see Frei and Künsch, 2013, for an example of an alternative).

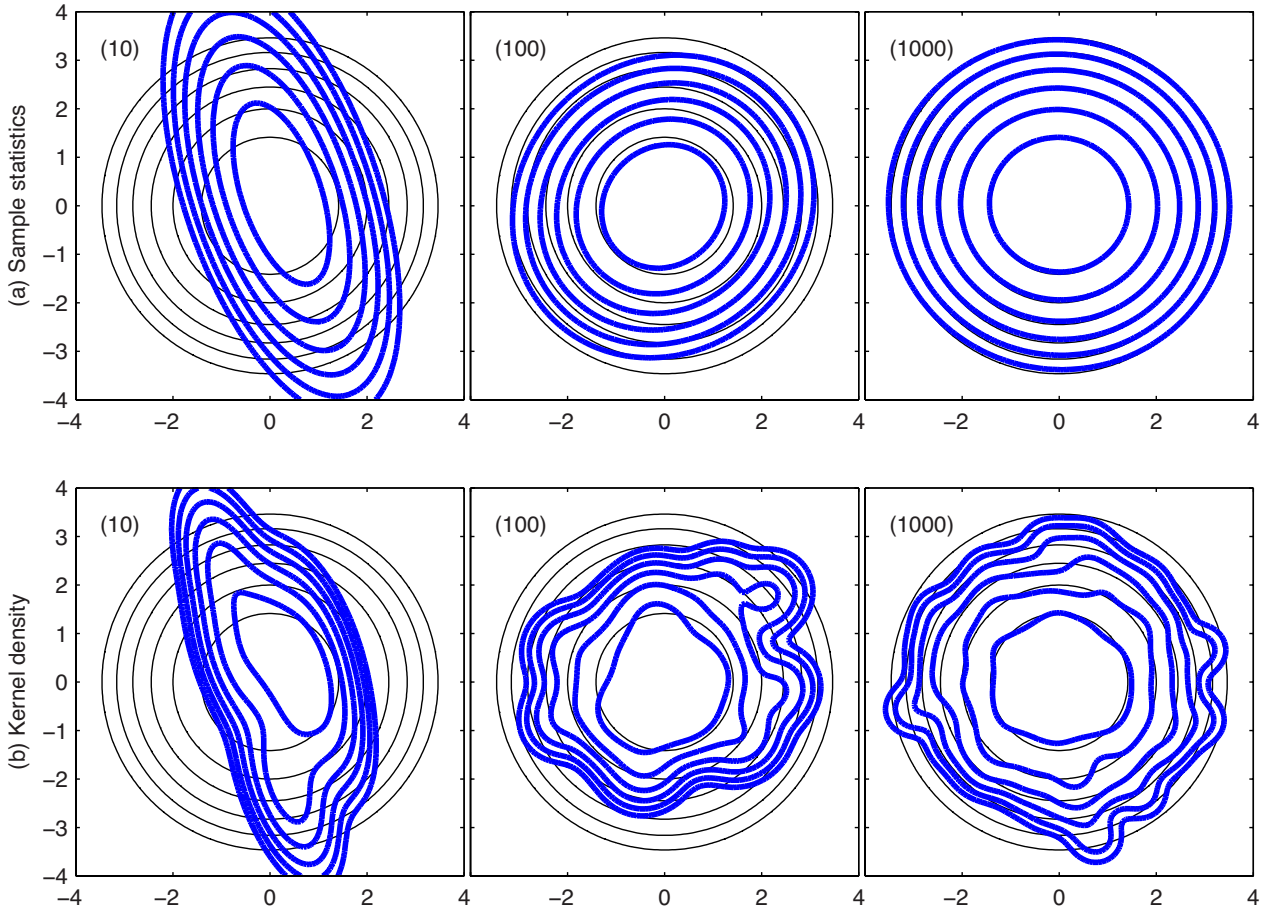

Fig. 1. Continuous estimates of the density of a two-dimensional standard normal random variable. Contours are plotted at equal intervals of the logarithm of the (thin circles) true density and (thick curves) estimated densities. The estimates are the **(a)** Gaussian with sample mean and covariance and **(b)** kernel density estimate from the same samples. The number in parentheses is the sample size. The kernel density estimates are computed using the optimal bandwidth and have the same first two moments as the samples (Silverman, 1986).

3.2 The number of components

What remains is an approach for specifying the number N_m of components in the mixture. At one extreme,

$$\left.\begin{aligned} \boldsymbol{\mu} &= \frac{1}{N_p}\sum_{i=1}^{N_p} \boldsymbol{x}_k^{(i)}, \\ \mathbf{B} &= \frac{1}{N_p}\sum_{i=1}^{N_p} \left(\boldsymbol{x}_k^{(i)} - \boldsymbol{\mu}\right)\left(\boldsymbol{x}_k^{(i)} - \boldsymbol{\mu}\right)^T, \end{aligned}\right\} \quad \text{if } N_m = 1,$$

where the covariance $\mathbf{B}$ has the normalization $1/N_p$ because the EM algorithm finds the MLE. This is a completely parametric representation of the background distribution typical of ensemble implementations of the Kalman filter. At the other extreme,

$$\left.\begin{aligned} \alpha_j &= 1/N_p, \\ \boldsymbol{\mu}_j &= \boldsymbol{x}_k^{(j)}, \qquad \mathbf{B}_j = 0, \end{aligned}\right\} \quad \text{if } N_m = N_p,$$

and the densities $\mathcal{N}(\boldsymbol{\mu}_j, \mathbf{B}_j)$ are thus Dirac delta functions. This is the traditional particle filter approach, which represents the background as a sum of delta functions, making no parametric assumptions.

For simplicity, this paper assumes that $N_m = 1$ until N_p is so large that there is no ensemble collapse. After this point, it is safe to take $N_m = N_p$. Given a small number of particles in very high dimensions, the method therefore relies upon simplifying parametric assumptions. On the other hand, in the limit as $N_p \to \infty$, it maintains the convergence properties of particle filters. An approximation for the number of particles N_p^* at which to make the transition is derived for linear and Gaussian problems in the following section. The appropriate value of N_p^* for nonlinear and non-Gaussian problems is not immediately obvious, but a reasonable choice usually can be determined from numerical experimentation. While it is not considered here, picking N_m based on the Akaike or Bayes information criteria (SL13; Konishi and Kitagawa, 2008) may have the same convergence properties and be more efficient for finite values of N_p.

3.3 The target density

Given a mixture model approximation to the background, there is a corresponding approximation to the target. Marginalizing over $\boldsymbol{x}_{0:\mathrm{m}(k)-1}$ and using the conditional independence of the state and observation, the target is

$$p\left(\boldsymbol{x}_{\mathrm{m}(k):\mathrm{m}(k+1)} \,|\, \boldsymbol{y}_{1:k+1}\right) = p\left(\boldsymbol{x}_{\mathrm{m}(k)} \,|\, \boldsymbol{y}_{1:k}\right) \cdot \prod_{m=\mathrm{m}(k)+1}^{\mathrm{m}(k+1)} p\left(\boldsymbol{x}_{m+1} \,|\, \boldsymbol{x}_m\right) \cdot p\left(\boldsymbol{y}_n \,|\, \boldsymbol{x}_{\mathrm{m}(n)}\right). \tag{5}$$

By definition of the model (1) and observation (2),

$$\boldsymbol{e}_{m+1} = \mathbf{Q}^{-1/2}\left[\boldsymbol{x}_{m+1} - \mathcal{M}\left(\boldsymbol{x}_m, \boldsymbol{x}_{m+1}\right)\right],$$
$$\boldsymbol{d}_n = \mathbf{R}^{-1/2}\left[\boldsymbol{y}_n - \mathcal{H}\left(\boldsymbol{x}_{\mathrm{m}(n)}\right)\right],$$

and expressions for the conditional pdfs $p(\boldsymbol{x}_{m+1} \,|\, \boldsymbol{x}_m)$ and $p(\boldsymbol{y}_n \,|\, \boldsymbol{x}_{\mathrm{m}(n)})$ follow from the change of variables formula. If the time discretization of the model is explicit: $\mathcal{M}(\boldsymbol{x}_m, \boldsymbol{x}_{m+1}) = \mathcal{M}(\boldsymbol{x}_m)$, then

$$p\left(\boldsymbol{x}_{m+1} \,|\, \boldsymbol{x}_m\right) \propto \exp\left(-\frac{1}{2}\boldsymbol{e}_{m+1}^T \boldsymbol{e}_{m+1}\right), \tag{6}$$

$$p\left(\boldsymbol{y}_n \,|\, \boldsymbol{x}_{\mathrm{m}(n)}\right) \propto \exp\left(-\frac{1}{2}\boldsymbol{d}_n^T \boldsymbol{d}_n\right). \tag{7}$$

Otherwise, Eq. (6) is more complex, but the algorithm below remains the same.

Substituting the mixture approximation (4) and the expressions (6) and (7) into Eq. (5) gives the approximation of the target,

$$\hat{p}\left(\boldsymbol{x}_{\mathrm{m}(k):\mathrm{m}(k+1)} \,|\, \boldsymbol{y}_{1:k+1}\right) \propto \sum_j \alpha_j \beta_j^{-1} \exp(-\varphi_j),$$

where $\beta_j = \sqrt{(2\pi)^{N_x} \det\left(\mathbf{B}_j\right)}$, and the component cost functions φ_j are defined such that

$$\varphi_j = \frac{1}{2}\left(\boldsymbol{x}_{\mathrm{m}(k)} - \boldsymbol{\mu}_j\right)^T \mathbf{B}_j^{-1}\left(\boldsymbol{x}_{\mathrm{m}(k)} - \boldsymbol{\mu}_j\right) + \frac{1}{2}\sum_{m=\mathrm{m}(k)+1}^{\mathrm{m}(k+1)} \boldsymbol{e}_m^T \boldsymbol{e}_m + \frac{1}{2}\boldsymbol{d}_{k+1}^T \boldsymbol{d}_{k+1}.$$

Finally, to simplify notation, let $\boldsymbol{v}$ denote the vector of independent variables that determine the cost φ_j, i.e.,

$$\boldsymbol{v} = \begin{bmatrix} \boldsymbol{x}_{\mathrm{m}(k)} \\ \vdots \\ \boldsymbol{x}_{\mathrm{m}(k+1)} \end{bmatrix}, \qquad \text{and} \qquad \varphi_j = \varphi_j(\boldsymbol{v}).$$

3.4 Implicit sampling

In general, sampling directly from the component density $\beta_j^{-1}\exp(-\varphi_j)$ is impossible unless φ_j has one of a limited number of parametric forms. Rather than make this restrictive assumption, implicit techniques use importance sampling (Geweke, 1989), which draws samples from an alternate density, called the importance, then weighs the samples to account for the difference between the actual and sampled densities.

The "optimal" importance is a Gaussian approximation of the component density with the same mode $\boldsymbol{v}_j^*$. In general, the mode, which is also the global minimizer of φ_j, must be found using numerical optimization, a task that is by no means trivial. A byproduct of any quasi-Newton optimization method is an approximation $\boldsymbol{\Phi}_j$ of the Hessian of φ_j at $\boldsymbol{v}_j^*$, and hence a quadratic approximation ψ_j of φ_j such that

$$\psi_j(\boldsymbol{v}) = \varphi_j^* + \frac{1}{2}\left(\boldsymbol{v} - \boldsymbol{v}_j^*\right)^T \mathbf{S}_j^{-1}\left(\boldsymbol{v} - \boldsymbol{v}_j^*\right),$$

where $\varphi_j^* = \varphi_j(\boldsymbol{v}_j^*)$ and $\mathbf{S}_j = \boldsymbol{\Phi}_j^{-1}$. The cost function ψ_j is the basis of the "optimal" importance density $\gamma_j^{-1}\exp(-\psi_j)$, where

$$\gamma_j = \exp\left(-\varphi_j^*\right)\sqrt{(2\pi)^{N_v}\det\left(\mathbf{S}_j\right)},$$

and N_v is the dimension of $\boldsymbol{v}$. The resulting Gaussian mixture approximation of the target density $\hat{p}$ is

$$q\left(\boldsymbol{x}_{\mathrm{m}(k):\mathrm{m}(k+1)} \,|\, \boldsymbol{y}_{1:k+1}\right) = \sum_j \alpha_j \gamma_j^{-1} \exp\left(-\psi_j\right).$$

The "optimal" importance is actually optimal if the component cost function is a quadratic function of $\boldsymbol{v}$ (Doucet et al., 2000). In this case, the component distribution is Gaussian and, given the exact values of $\boldsymbol{v}_j^*$ and $\boldsymbol{\Phi}_j$, is identical to the "optimal" importance. In other cases, including an example below, there are better choices for the importance.

3.5 The algorithm

MIPS begins by determining the importance density for each component in three steps:

1. Compute the mixture model approximation to the background of $\boldsymbol{x}_{\mathrm{m}(k)}$.
2. Find the mode $\boldsymbol{v}_j^*$ of $\beta_j^{-1}\exp(-\varphi_j)$ and Hessian $\boldsymbol{\Phi}_j$ of φ_j at the mode.
3. Define the covariance matrix $\mathbf{S}_j$ such that $\mathbf{S}_j = \boldsymbol{\Phi}_j^{-1}$.

It then proceeds as follows for each particle $\boldsymbol{x}_{\mathrm{m}(k)}^{(i)}$:

4. Generate a uniform random number $r \in (0, 1]$, and find the component index l that satisfies

$$\sum_{j=1}^{l-1}\alpha_j < r \le \sum_{j=1}^{l}\alpha_j,$$

using the convention that the sum from $j = 1$ to $j = 0$ is zero.

5. Draw a sample $\boldsymbol{v}^{(i)}$ from the Gaussian importance $\gamma_l^{-1}\exp(-\psi_l)$.

6. Give the sample the weight

$$w^{(i)} = \frac{\beta_l^{-1}\exp\left(-\varphi_l^{(i)}\right)}{\gamma_l^{-1}\exp\left(-\psi_l^{(i)}\right)}, \propto \gamma_l\beta_l^{-1}\exp\left(\psi_l^{(i)}-\varphi_l^{(i)}\right), \quad (8)$$

where $\varphi_l^{(i)}$ and $\psi_l^{(i)}$ denote the values of the functions evaluated at $\boldsymbol{v}^{(i)}$.

Afterward, the weighted samples can be transformed into a uniformly weighted ensemble by resampling with replacement, also known as the bootstrap (Efron, 1979). Since this step adds noise to the estimates, it is best to resample only when the effective sample size (Kong et al., 1994; Liu, 1996; Doucet et al., 2000) falls below a given fraction of N_p. There are a variety of improvements to the bootstrap that reduce the added noise as well. This reduction, however, is likely dominated by the error in the representation of the target (Kitagawa, 1996). Although it is not presented here, the generalization of the EM algorithm to weighted samples is straightforward.

In two special cases, MIPS is equivalent to other assimilation techniques. First, if there is a single component and everything is linear and Gaussian, MIPS is an ensemble Kalman smoother. If the model or observation functions are nonlinear or if $N_m \neq 1$, it is not an ensemble Kalman smoother (for two methods that are, see van Leeuwen and Evensen, 1996; Evensen and van Leeuwen, 2000). Second, if $N_m = N_p$, MIPS is equivalent to IPF. This is because the covariance of each mixture component is the degenerate form $\mathbf{B}_j = 0$ (see Sect. 3.2), which fixes the value of $\boldsymbol{x}_{\mathrm{m}(k)}$ for each particle.

Another variation to the above algorithm that can reduce the variance of the weights is to perform importance sampling on the full mixture distribution rather than its individual components. If $N_m = N_p$, this approach is equivalent to an implementation of the marginal particle filter (Klaas et al., 2005). Its biggest drawback is that it must evaluate every component density, rather than just one, at every sample to compute the weights.

4 A linear and Gaussian example

As a simple demonstration of ensemble collapse in high dimensions, Bengtsson et al. (2008), Bickel et al. (2008), and Snyder et al. (2008) propose an example where there are observations every time step, i.e., $\mathrm{m}(n) = n$, the model and observation are linear functions such that

$$\mathcal{M}(\boldsymbol{x}_m) = \mathbf{A}\boldsymbol{x}_m, \qquad \mathcal{H}(\boldsymbol{x}_m) = \mathbf{H}\boldsymbol{x}_m$$

and the distribution of the initial condition is Gaussian.

By construction, the difference between components in the mixture model is the only source of variability in the weights. This is because the component cost functions,

$$\varphi_j = \frac{1}{2}\left(\boldsymbol{x}_k-\boldsymbol{\mu}_j\right)^T\mathbf{B}_j^{-1}\left(\boldsymbol{x}_k-\boldsymbol{\mu}_j\right) + \frac{1}{2}\left(\boldsymbol{x}_{k+1}-\mathbf{A}\boldsymbol{x}_k\right)^T\mathbf{Q}^{-1}\left(\boldsymbol{x}_{k+1}-\mathbf{A}\boldsymbol{x}_k\right) + \frac{1}{2}\left(\boldsymbol{y}_{k+1}-\mathbf{H}\boldsymbol{x}_{k+1}\right)^T\mathbf{R}^{-1}\left(\boldsymbol{y}_{k+1}-\mathbf{H}\boldsymbol{x}_{k+1}\right),$$

are quadratic and the component densities are Gaussian. It is thus possible to sample the component densities exactly. The resulting samples have weights (8) that satisfy

$$w^{(i)} \propto \gamma_l\beta_l^{-1},$$

which depends only on the component of the mixture (recall that l is a random function of i).

Expressions for the mean and covariance of the component densities, and hence the weights, follow from similar algebra to the Kalman smoother (Rauch, 1963; Jazwinski, 1970). Dropping the component index l to simplify notation, the mode is the solution of the linear equations

$$\mathbf{B}^{-1}\left(\boldsymbol{x}_k^*-\boldsymbol{\mu}\right)-\mathbf{A}^T\mathbf{Q}^{-1}\left(\boldsymbol{x}_{k+1}^*-\mathbf{A}\boldsymbol{x}_k^*\right)=0,$$
$$\mathbf{Q}^{-1}\left(\boldsymbol{x}_{k+1}^*-\mathbf{A}\boldsymbol{x}_k^*\right)-\mathbf{H}^T\mathbf{R}^{-1}\left(\boldsymbol{y}_{k+1}-\mathbf{H}\boldsymbol{x}_{k+1}^*\right)=0.$$

It can be expressed as the backward recursion,

$$\boldsymbol{x}_{k+1}^* = \mathbf{A}\boldsymbol{\mu}+\mathbf{K}\left(\boldsymbol{y}_{k+1}-\mathbf{H}\mathbf{A}\boldsymbol{\mu}\right),$$
$$\boldsymbol{x}_k^* = \boldsymbol{\mu}+\mathbf{C}\left(\boldsymbol{x}_{k+1}^*-\mathbf{A}\boldsymbol{\mu}\right),$$

where

$$\mathbf{P}^f = \mathbf{A}\mathbf{B}\mathbf{A}^T+\mathbf{Q},$$
$$\mathbf{K} = \mathbf{P}^f\mathbf{H}^T\left(\mathbf{H}\mathbf{P}^f\mathbf{H}^T+\mathbf{R}\right)^{-1},$$
$$\mathbf{C} = \mathbf{B}\mathbf{A}^T\left(\mathbf{P}^f\right)^{-1}.$$

At every point, the Hessian of the cost function is the block matrix

$$\mathbf{S}^{-1} = \begin{bmatrix}\mathbf{B}^{-1}+\mathbf{A}^T\mathbf{Q}^{-1}\mathbf{A} & -\mathbf{A}^T\mathbf{Q}^{-1}\\ -\mathbf{Q}^{-1}\mathbf{A} & \mathbf{Q}^{-1}+\mathbf{H}^T\mathbf{R}^{-1}\mathbf{H}\end{bmatrix}.$$

Its inverse, the covariance matrix is

$$\mathbf{S} = \begin{bmatrix}\mathbf{P}^s & \mathbf{C}\mathbf{P}^a\\ \mathbf{P}^a\mathbf{C}^T & \mathbf{P}^a\end{bmatrix},$$

where the matrices are again the same as in the Kalman smoother:

$$\mathbf{P}^a = \left(\mathbf{I}_x-\mathbf{K}\mathbf{H}\right)\mathbf{P}^f,$$

$$\mathbf{P}^s = \mathbf{B} + \mathbf{C}\left(\mathbf{P}^a - \mathbf{P}^f\right)\mathbf{C}^T,$$

and $\mathbf{I}_x$ is the $N_x \times N_x$ identity matrix.

Substituting the expressions back into Eq. (8) and restoring the index l shows that

$$w^{(i)} = p\left(\boldsymbol{y}_{k+1} \,|\, \boldsymbol{\mu}_l, \mathbf{B}_l\right), \\ = \frac{1}{\sqrt{(2\pi)^{N_y}\det\left(\boldsymbol{\Omega}_l\right)}} \exp\left[-\frac{1}{2}\tilde{\boldsymbol{y}}_{k+1}^T \boldsymbol{\Omega}_l^{-1} \tilde{\boldsymbol{y}}_{k+1}\right],$$

where

$$\tilde{\boldsymbol{y}}_l = \boldsymbol{y}_{k+1} - \mathbf{H}\mathbf{A}\boldsymbol{\mu}_l, \qquad \text{and} \qquad \boldsymbol{\Omega}_l = \mathbf{H}\mathbf{P}_l^f\mathbf{H}^T + \mathbf{R}.$$

In other words, the weights are the values of the pdf of the innovations $\tilde{\boldsymbol{y}}_l$, whose covariance is $\boldsymbol{\Omega}_l$.

4.1 Ensemble collapse

A fundamental result of Bengtsson et al. (2008), Bickel et al. (2008), and Snyder et al. (2008) is that for importance sampling methods

$$\mathbb{E}\left[1/\max_i w^{(i)}\right] = 1 + \sqrt{2\ln N_p}/\sigma + O\left(\ln N_p/\sigma^2\right), \quad (9)$$

where

$$\sigma^2 = \mathbb{E}\left[\left(\log w^{(i)}\right)^2\right] - \mathbb{E}\left[\log w^{(i)}\right]^2.$$

To simplify the analysis, Snyder (2012) takes

$$\mathbf{A} = a\mathbf{I}_x, \qquad \mathbf{Q} = q^2\mathbf{I}_x,$$
$$\mathbf{H} = \mathbf{I}_x, \qquad \mathbf{R} = \mathbf{I}_x,$$

and a standard normal prior on the initial condition. He then shows that, provided $N_x \gg 1$,

$$\sigma^2 \approx \begin{cases} N_x\left(a^2+q^2\right)\left(\frac{3}{2}a^2+\frac{3}{2}q^2+1\right) & \text{for BPF,} \\ N_x a^2\left(\frac{3}{2}a^2+q^2+1\right)\Big/\left(q^2+1\right)^2 & \text{for IPF.} \end{cases}$$

These two cases are depicted in Fig. 2 for an example where $a^2 = q^2 = 0.5$. The line for MIPS with a one component mixture (MIPS1) is identically zero because all of the terms in Eq. (8) are constant for this example: the first two terms are constant because there is only one component in the mixture, and the final term is always 1 because the component cost φ and its quadratic approximation ψ are identical.

It is also evident from Fig. 2 how to choose N_m in a simple way to avoid collapse. If N_p is less than the value of the IPF line, let $N_m = 1$. Otherwise, let $N_m = N_p$ to revert to using IPF. The threshold N_p^* at which to make this change occurs when

$$\mathbb{E}\left[1\Big/\max_i w^{(i)}\right] = 1/0.9.$$

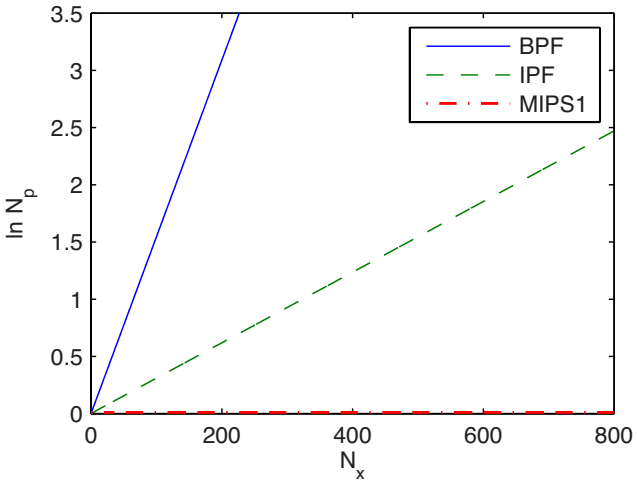

Fig. 2. Theoretical lines such that $\mathbb{E}[1/\max_i w^{(i)}] = 1/0.9$ for the (solid) bootstrap particle filter, (dashed) implicit particle filter and (dot-dashed) mixture-based implicit smoother with $N_m = 1$. The dot-dashed curve is identically zero. Similar to Fig. 3 of Snyder (2012) with a corrected y axis label.

Substituting this into Eq. (9), neglecting the higher-order terms, and using the IPF expression for σ^2 gives

$$N_p^* = \exp\left(\sigma^2/162\right), \\ = \exp\left[N_x a^2\left(3a^2/2 + q^2 + 1\right)\Big/ 162\left(q^2+1\right)^2\right].$$

In practice, computational resources may necessitate taking $N_m = 1$ even well beyond this threshold, maintaining a transition to $N_m = N_p$ only to preserve the theoretical properties of the particle filter. Often, even for nonlinear models, the choice $N_m = 1$ performs quite well, as is shown below.

Table 1 provides a numerical comparison to the theoretical lines in Fig. 2. In general, the number of particles necessary to avoid collapse is greater than the transition value

$$N_p^* = \exp(N_x/324),$$

yet by a factor less than 2. Furthermore, the difference between the two goes to zero as $N_x \to \infty$. The reason for the discrepancy is that the convergence in Eq. (9) is quite slow (David and Nagaraja, 2003, Ex. 10.5.3). While more accurate thresholds are possible, in realistic applications, the model and observation nonlinearities are likely to have a significant effect on the weights and thus the choice of N_p^*.

5 Multiple-well problems

This section considers three examples: a standard, one-dimensional double-well and two generalizations of it to multiple dimensions. Although it is simple, the nonlinearity of the double-well is significant enough to cause difficulties

Table 1. The dependence of $\mathbb{E}[1/\max_i w^{(i)}]$ on N_p and N_x for the implicit particle filter. The number of particles N_p varies along the rows and the state dimension N_x along the columns. Estimates are computed from 1000 trials, and typical sampling errors are $O(0.01)$. The italic values are the closest points to the dashed line in Fig. 2.

$N_p \backslash N_x$	100	200	400	800
2	*1.08*	1.05	1.04	1.03
4	1.15	*1.11*	1.07	1.05
8	1.24	1.16	*1.11*	1.08
16	1.34	1.22	1.14	1.10
32	1.42	1.26	1.17	*1.11*

with data assimilation methods that rely on parametric assumptions about the target, notably the extended Kalman filter (Miller et al., 1994) and EnKF (Miller et al., 1999; Evensen and van Leeuwen, 2000). For every example, the discrete model $\mathcal{M}$ is the result of the Euler–Maruyama method (Kloeden and Platen, 1999) applied to a continuous model $\boldsymbol{f}$, i.e.,

$$\mathcal{M}(\boldsymbol{x}_m, \boldsymbol{x}_{m+1}) = \boldsymbol{x}_m + \tau \boldsymbol{f}(\boldsymbol{x}_m),$$

the time step τ is 0.02, observations occur every 200 time steps, $\mathbf{Q} = 0.5\tau\mathbf{I}_x$, the observation operator is the identity, $\mathbf{R} = 0.1\mathbf{I}_x$ and the initial condition for every component of the state has mean 1 and standard deviation 0.1. These values are roughly equivalent to those used by SL13. In the double-well problem, $N_x = 1$, and

$$\boldsymbol{f}(\boldsymbol{x}) = 4x - 4x^3.$$

The multiple-well examples have one of two possible forms:

$$\boldsymbol{f}(\boldsymbol{x}) = \begin{bmatrix} 4x_1 - 4x_1^3 \\ 4x_2 - 4x_2^3 \\ 4x_3 - 4x_3^3 \\ 4 - 4x_4 \\ \vdots \\ 4 - 4x_{N_x} \end{bmatrix}, \tag{10}$$

or

$$\boldsymbol{f}(\boldsymbol{x}) = \begin{bmatrix} 4x_1 - 4x_1^3 \\ \left(-\tau x_2 - x_3\right)/\left(1+\tau^2\right) \\ \left(-\tau x_3 + x_2\right)/\left(1+\tau^2\right) \\ 4 - 4x_4 \\ \vdots \\ 4 - 4x_{N_x} \end{bmatrix}, \tag{11}$$

where x_i denotes the i-th element of the vector $\boldsymbol{x}$.

The models for the two multiple-well examples (10) and (11) are intended to illustrate two different possibilities for the asymptotic statistics of the sample solutions. Projections into three dimensions of both types of sample solutions are depicted in Fig. 3. In the first example (10), the deterministic equations

$$\frac{\mathrm{d}\boldsymbol{x}}{\mathrm{d}t} = \boldsymbol{f}(\boldsymbol{x})$$

have 8 stable fixed points. The corresponding stochastic equations have an invariant measure (climatology) that is approximately the sum of 8 Gaussians centered at these points. The pdf of this measure is the limit as $t \to \infty$ of the solution of the Fokker–Planck equation (also known as the Kolmogorov forward equation; Øksendal, 2003). In the second multiple-well example (11), the deterministic equations have two stable, two-dimensional invariant sets in the limit $\tau \to 0$. The solution of the corresponding Fokker–Planck equation, like that of a two-dimensional Wiener process, is a measure whose variances in these two subsets go to infinity as time increases. Consequently, the stochastic equations have no invariant measure.

The non-standard form of the continuous model $\boldsymbol{f}$ in the second multiple-well example (11) is meant to aid comparison with BPF and EnKF, which typically only apply to explicit time discretizations (1). It follows from applying backward/implicit Euler (Kloeden and Platen, 1999) to a rotation map. This ensures the stability of the discrete model. Implicit sampling methods, on the other hand, are straightforward to implement with implicit time discretizations. The only difference is in the form of the cost function φ_j. This is a notable advantage of the implicit approach.

5.1 Transition detection

Figure 4 compares the results of EnKF, BPF, the implicit particle filter (IPF) and the mixture-based implicit particle smoother with a 1 component mixture model (MIPS1) using 10 particles for the double-well problem. Both implicit methods perform well, yet MIPS1 requires just over a tenth of the number of floating point operations of IPF. This is because it solves a single optimization problem instead of 10, while including the previous state $\boldsymbol{x}_{\mathrm{m}(k)}$ in the optimization problem only increases the dimension of the problem from 200 to 201.

Both BPF and EnKF miss the transition even after repeated observations because their forecasting step rarely generates a particle in the correct well. This is evident in the BPF and EnKF estimates in Fig. 4, which display noticeably less variability while in the correct well than the implicit estimates. As a result, artificial inflation of the model error covariance (Anderson and Anderson, 1999) could improve the performance of the explicit methods in these examples.

The estimates of every method for the multiple-well problems are qualitatively similar to a combination of the double-well problem with the linear problem. In the dimensions where the model $\boldsymbol{f}$ is nonlinear, the problem resembles the double-well, and in the dimensions where the model $\boldsymbol{f}$ is linear, the problem resembles the linear example.

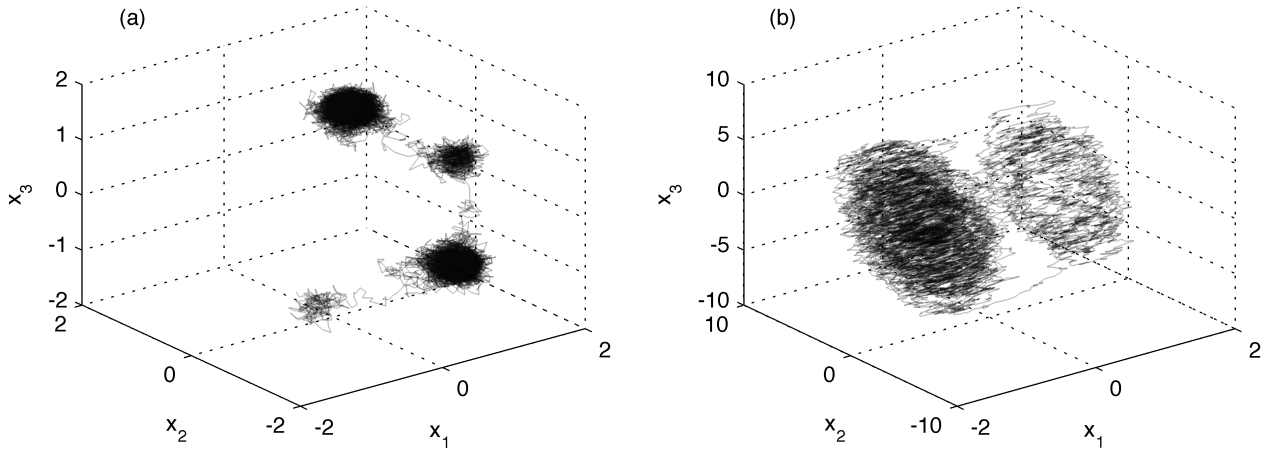

Fig. 3. Examples of three-dimensional projections of the twin solution for the two multiple-well examples: **(a)** 8 wells with three transitions and **(b)** two wells with one transition. The number of transitions in the left panel is very rare and is used for visualization purposes.

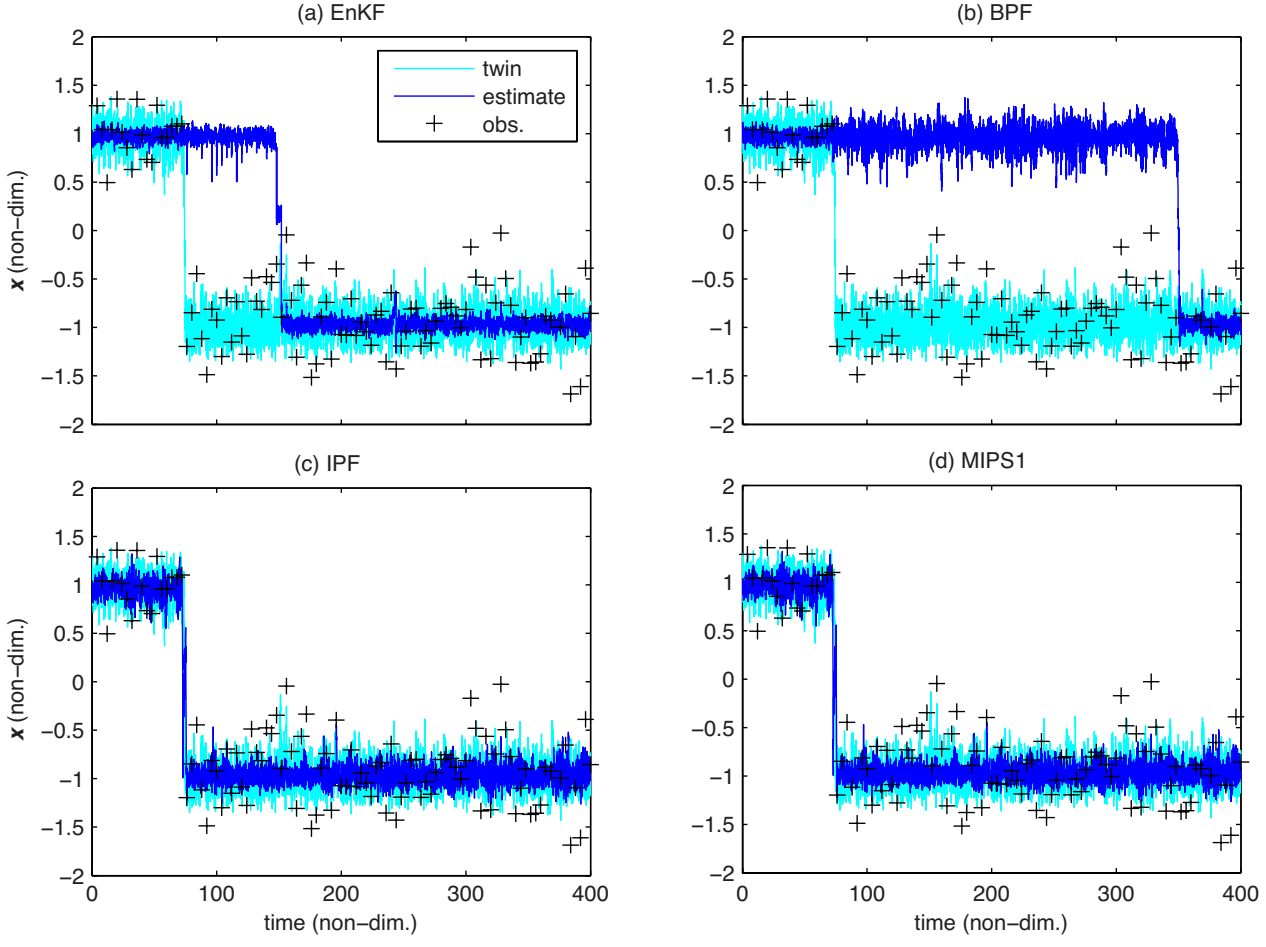

Fig. 4. Comparison of the estimates with 10 particles of the **(a)** ensemble Kalman filter, **(b)** bootstrap particle filter, **(c)** implicit particle filter and **(d)** mixture-based implicit particle smoother with $N_m = 1$.

Table 2. Percentage of trials in which an estimate computed with 10 particles is in the same well as the twin solution at the final assimilation time. The well is determined by the sign of the elements of the state vector. Results are computed from 100 trials each with at least one transition.

N_x	EnKF	BPF	IPF/MIPS
1	100 %	85 %	100 %
4	100 %	80 %	100 %
16	96 %	89 %	100 %
64	70 %	79 %	100 %
256	39 %	49 %	100 %

As the state dimension N_x increases, EnKF and BPF are increasingly less likely to detect the transitions from well to well. Table 2 quantifies this likelihood for the first example (some of the variability in the results for BPF is most likely due to sampling errors). Results for the second example are comparable. Unlike the explicit methods, the implicit methods consistently detect the transition because they find an optimal solution based on the observation. In these examples, the model and observation functions and covariances have a particularly simple form, and the problem can be decomposed into a collection of decoupled problems. If this decomposition is used, the performance of the explicit methods does not degrade as the dimension increases. Nevertheless, this decomposition is only possible in very special cases.

5.2 Hessian refinement

Although the implicit methods consistently detect the transition from one well to another, their weights collapse at the transition. This happens in MIPS1 because the component cost φ, is far from its quadratic approximation ψ, resulting in significant variation in the term

$$\exp\left(\psi^{(i)} - \varphi^{(i)}\right).$$

As a simple example, suppose $\varphi(x) = \kappa x^2 + x^4$ for some small positive number κ. Then $\psi(x) = \kappa x^2$, and importance sampling based on ψ will generate very many samples far out on the tails of the target density and very few in the region of high probability. Moreover, for a fixed N_p, the weight of the sample closest to the origin approaches 1 as κ approaches 0.

In many examples like the above, it is possible to decrease the variance of the weights by finding a better approximation to the covariance of the component density than the inverse Hessian. One approach is to repeatedly sample the importance density and update the Hessian based upon the agreement of the component cost φ and importance cost ψ. This results in a variation of the MIPS algorithm with the third step (see Sect. 3.5) replaced by the refinement step,

3′. Begin with $\mathbf{\Phi}^{(1)}$ equal to the Hessian at the mode. For n from 1 to a given number N_r, draw a sample $\boldsymbol{v}^{(n)}$ from the Gaussian importance $\gamma^{-1}\exp(-\psi)$. Then compute a new Hessian $\mathbf{\Phi}^{(n+1)}$, where

$$\mathbf{\Phi}^{(n+1)} = \mathbf{\Phi}^{(n)} + \epsilon_n \mathbf{\Delta}^{(n)},$$

$$\mathbf{\Delta}^{(n)} = \left(\psi^{(n)} - \varphi^{(n)}\right) \frac{\left(\boldsymbol{v}^{(n)} - \boldsymbol{v}^*\right)\left(\boldsymbol{v}^{(n)} - \boldsymbol{v}^*\right)^T}{\left(\boldsymbol{v}^{(n)} - \boldsymbol{v}^*\right)^T\left(\boldsymbol{v}^{(n)} - \boldsymbol{v}^*\right)},$$

and update the covariance matrix $\mathbf{S}$ to be the inverse of $\mathbf{\Phi}^{(n+1)}$.

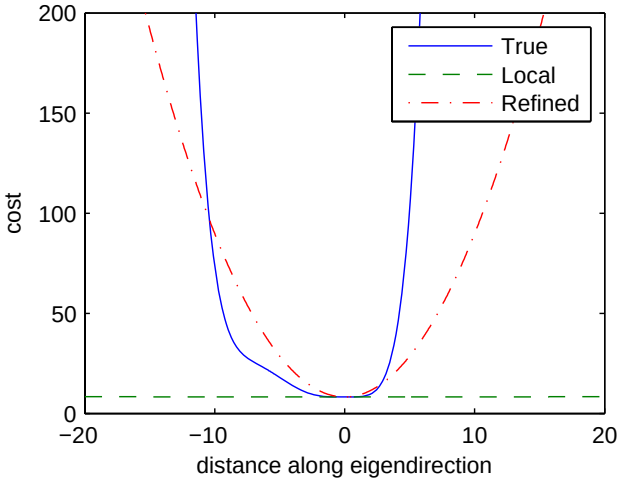

Fig. 5. Plots of the true cost function and its quadratic approximations along a line in sample space. The direction of the line is parallel to the eigenvector of the Hessian $\mathbf{\Phi}$ with the smallest eigenvalue, and the mode is translated to 0. The curvature of the local approximation (dashed) is determined by the second derivative of the cost function at its minimum, while the curvature of the refined approximation (dot-dashed) is adapted to better reflect the global properties of the cost function.

While it is not presented here, the extension of the refinement to $N_m > 1$ is straightforward.

The iteration on the Hessian follows from an application of the stochastic gradient descent algorithm of Robbins and Monro (1951). Sufficient conditions for its convergence are that ϵ_n is positive and, as $n \to \infty$,

$$\epsilon_n \to 0, \qquad \sum_n \epsilon_n \to \infty \qquad \text{and} \qquad \sum_n \epsilon_n^2 < \infty.$$

The final limit condition, however, can be weakened considerably (Kushner and Yin, 2003, Chap. 5). The result of the iteration is an approximation of Hessian whose inverse is the covariance that minimizes the variances of the weights, a claim which is made precise in Appendix A.

Figure 5 compares the true cost function φ with approximations based on its local properties and the Hessian refinement for the double-well example depicted in Fig. 4. The plot is along a line in sample space parallel to the eigenvector of

the initial Hessian with the smallest eigenvalue. In this example,

$$\epsilon_n = \frac{1}{2} n^{-3/4} \qquad \text{and} \qquad N_r = 100.$$

It is apparent that sampling from a Gaussian whose cost function is based on the refined Hessian is far more efficient than if the cost function were based on the local approximation. Finally, the similarity of the results with the example $\varphi(x) = \kappa x^2 + x^4$ is due to the fact that the transition is the point in the assimilation where the nonlinear terms of the function $\boldsymbol{f}$ matter most.

6 Conclusions

A mixture model approximation of the distribution of the background state enables particle methods to adjust the background position of the particles and is often more accurate than a kernel density estimate. When combined with an implicit assimilation method, this approach, the mixture-based particle smoother (MIPS), is a possible solution for high dimensional problems. This is true for a high-dimensional, linear, Gaussian example, where MIPS does not collapse.

With only a small number of particles, the implicit method is able to detect transitions in an example with multiple attracting states. To detect the same transitions, explicit approaches like BPF and EnKF require considerably more particles. This number increases with the system dimension, provided the problem does not admit further dimensional reduction. Moreover, with the addition of an iteration that trains the proposal covariance to the true covariance, MIPS can track transitions without weight collapse given only a handful of particles.

If MIPS is to be applied to a realistic, high-dimensional assimilation problem in the geosciences, there are a number of improvements and simplifications to consider. In particular, with a limited number of samples in very high dimensions, the analytically computed values of the mode $\boldsymbol{v}_j^*$ and covariance $\mathbf{S}_j$ may lead to better approximations of the component mean $\boldsymbol{\mu}_j$ and covariance $\mathbf{B}_j$ than the sample estimates. Perhaps most importantly, the optimization step in MIPS can require very many floating point operations, and its efficiency is vital to the applicability of the method as a whole. However, the examples in this paper show that there is reason to believe this additional computational requirement enables the implicit approach to produce accurate estimates in high dimensions even with a very small number of particles.

Appendix A

Minimization of the variance of the weights

In general, the variance of the weights with respect to the density $q(\boldsymbol{v} \mid \boldsymbol{y}_{1:k+1})$,

$$\mathrm{var}_q[w] = \mathbb{E}_q\left[w^2\right] - \mathbb{E}_q[w]^2, \tag{A1}$$

measures the success of an importance sampling method. It determines both the effective ensemble size and, to leading order, the variance of the sample mean of a general function (provided it satisfies appropriate integrability conditions; Kong et al., 1994; Liu, 1996; Doucet et al., 2000).

The goal of the Hessian refinement is to find an approximation $\boldsymbol{\Phi}$ of the matrix that minimizes the variance of the weights (A1). Since the partition function $\mathbb{E}_q[w]$ is independent of $\boldsymbol{\Phi}$,

$$\begin{aligned}
&\frac{\partial}{\partial \Phi_{ij}} \left\{ \mathbb{E}_q[w^2] - \mathbb{E}_q[w]^2 \right\} \\
&\quad = \frac{\partial}{\partial \Phi_{ij}} \mathbb{E}_q[w^2], \\
&\quad = \frac{\partial}{\partial \Phi_{ij}} \int \frac{\hat{p}(\boldsymbol{v} \mid \boldsymbol{y}_{1:k+1})^2}{q(\boldsymbol{v} \mid \boldsymbol{y}_{1:k+1})^2} q(\boldsymbol{v} \mid \boldsymbol{y}_{1:k+1}) \, \mathrm{d}\boldsymbol{v}, \\
&\quad = \int \hat{p}(\boldsymbol{v} \mid \boldsymbol{y}_{1:k+1})^2 \frac{\partial}{\partial \Phi_{ij}} \left[\frac{1}{q(\boldsymbol{v} \mid \boldsymbol{y}_{1:k+1})} \right] \mathrm{d}\boldsymbol{v}, \\
&\quad = -\int \frac{\hat{p}(\boldsymbol{v} \mid \boldsymbol{y}_{1:k+1})^2}{q(\boldsymbol{v} \mid \boldsymbol{y}_{1:k+1})^2} \frac{\partial}{\partial \Phi_{ij}} \left[q(\boldsymbol{v} \mid \boldsymbol{y}_{1:k+1}) \right] \mathrm{d}\boldsymbol{v}.
\end{aligned}$$

By definition,

$$\begin{aligned}
&\frac{\partial}{\partial \Phi_{ij}} q(\boldsymbol{v} \mid \boldsymbol{y}_{1:k+1}) \\
&\quad \propto \frac{\partial}{\partial \Phi_{ij}} \left\{ \sqrt{\det(\boldsymbol{\Phi})} \right. \\
&\qquad\qquad \left. \exp\left[\varphi^* - \frac{1}{2} \left(\boldsymbol{v} - \boldsymbol{v}^*\right)^T \boldsymbol{\Phi} \left(\boldsymbol{v} - \boldsymbol{v}^*\right) \right] \right\}, \\
&\quad = \frac{1}{2} \left[\boldsymbol{\Phi}^{-1} - \left(\boldsymbol{v} - \boldsymbol{v}^*\right)\left(\boldsymbol{v} - \boldsymbol{v}^*\right)^T \right]_{ij} q(\boldsymbol{v} \mid \boldsymbol{y}_{1:k+1}).
\end{aligned}$$

At this point, it is possible to apply the Robbins–Monro iteration (Robbins and Monro, 1951) to the integral objective equation

$$\frac{\partial}{\partial \Phi_{ij}} \left\{ \mathbb{E}_q\left[w^2\right] - \mathbb{E}_q[w]^2 \right\} = 0.$$

However, this approach performs poorly in practice because of the exponential dependence of the weights on the difference between the model cost and quadratic cost, which is very often large in magnitude. The approximation of the minimizer follows from expanding the square of the weights into

the power series

$$w^2 = \exp(2\psi - 2\varphi),$$
$$= 1 + (2\psi - 2\varphi) + \ldots.$$

Since $\psi - \varphi = O\left(\|\boldsymbol{v} - \boldsymbol{v}^*\|^2\right)$, regardless of the choice of $\boldsymbol{\Phi}$,

$$\frac{\partial}{\partial \Phi_{ij}} \mathbb{E}_q[w^2] = -\frac{1}{2}\mathbb{E}_q\left\{\left[\boldsymbol{\Phi}^{-1} - \left(\boldsymbol{v} - \boldsymbol{v}^*\right)\left(\boldsymbol{v} - \boldsymbol{v}^*\right)^T\right]_{ij}\right\} + \left[\boldsymbol{\Phi}^{-1}\right]_{ij} \mathbb{E}_q[\varphi - \psi] + O\left(\|\boldsymbol{v} - \boldsymbol{v}^*\|^4\right),$$

the first term of which is zero by definition of the covariance matrix. Hence, up to $O\left(\|\boldsymbol{v} - \boldsymbol{v}^*\|^4\right)$, the Hessian that minimizes the variance of the weights satisfies the integral objective equation

$$\mathbb{E}_q[\varphi - \psi] = 0. \tag{A2}$$

Since the objective equation (A2) is a single equation for $N_x(N_x+1)/2$ unknowns, it has infinitely many solutions. Nevertheless, each sample $\boldsymbol{v}^{(i)}$ only contains information about the model cost φ in the direction of $\boldsymbol{v}^{(i)} - \boldsymbol{v}^*$. Similar to quasi-Newton methods from deterministic optimization, the random-direction approach makes a rank-1 update to the Hessian such that

$$\boldsymbol{\Phi}^{(i+1)} = \boldsymbol{\Phi}^{(i)} + \epsilon_i\left(\psi^{(i)} - \varphi^{(i)}\right)\frac{\left(\boldsymbol{v}^{(i)} - \boldsymbol{v}^*\right)\left(\boldsymbol{v}^{(i)} - \boldsymbol{v}^*\right)^T}{\left(\boldsymbol{v}^{(i)} - \boldsymbol{v}^*\right)^T\left(\boldsymbol{v}^{(i)} - \boldsymbol{v}^*\right)}.$$

This leaves the effect of the Hessian on the null space of $\boldsymbol{v}^{(i)} - \boldsymbol{v}^*$ unchanged. However, it is important to truncate the step if it would cause one of the eigenvalues of $\boldsymbol{\Phi}^{(i+1)}$ to become negative. Although no formal proof is added, there is no evidence against the requirements for convergence (Kushner and Yin, 2003).

Acknowledgements. The authors would like to thank the organizers of the International Conference on Ensemble Methods in Geophysical Sciences held in Toulouse, France from 12 to 16 November 2012. This paper is, in many ways, an elaboration on presentations and discussions from this meeting, notably the straightforward and detailed description of the problems faced by any data assimilation method using importance sampling by Chris Snyder. Implicit sampling methods are the product of an ongoing collaboration with Alexandre Chorin, Ethan Atkins, Matthias Morzfeld and Xuemin Tu supported by the National Science Foundation, Division of Ocean Sciences, Collaboration in Mathematical Geosciences award #0934956. Their comments and suggestions, along with those of three anonymous reviewers, significantly improved the content and clarity of the manuscript. Finally, the motivation to combine Gaussian mixture models and implicit sampling is largely due to the suggestion of Pierre Lermusiaux.

Edited by: P. J. van Leeuwen
Reviewed by: three anonymous referees

References

Ades, M. and van Leeuwen, P. J.: An exploration of the equivalent weights particle filter, Q. J. Roy. Meteor. Soc., 139, 820–840, 2013.

Alspach, D. L. and Sorenson, H. W.: Nonlinear Bayesian estimation using Gaussian sum approximations, IEEE T. Automat. Contr., AC-17, 439–448, 1972.

Anderson, J. L. and Anderson, S. L.: A Monte Carlo implementation of the nonlinear filtering problem to produce ensemble assimilations and forecasts, Mon. Weather Rev., 127, 2741–2758, 1999.

Atkins, E., Morzfeld, M., and Chorin, A. J.: Implicit particle methods and their connection with variational data assimilation, Mon. Weather Rev., 141, 1786–1803, 2013.

Bellman, R. E.: Dynamic programming, Princeton University Press, Princeton, 1957.

Bengtsson, T., Snyder, C., and Nychka, D.: Toward a nonlinear ensemble filter for high-dimensional systems, J. Geophys. Res., 108, 8775, doi:10.1029/2002JD002900, 2003.

Bengtsson, T., Bickel, P., and Li, B.: Curse-of-dimensionality revisited: Collapse of the particle filter in very large scale systems, in: Probability and Statistics: Essays in Honor of David A. Freedman, Institute of Mathematical Statistics, Beachwood, Ohio, 2, 316–334, 2008.

Bertino, L., Evensen, G., and Wackernagel, H.: Sequential data assimilation techniques in oceanography, Int. Stat. Rev., 71, 223–241, 2003.

Bickel, P., Li, B., and Bengtsson, T.: Sharp failure rates for the bootstrap particle filter in high dimensions, in: Pushing the Limits of Contemporary Statistics: Contributions in Honor of Jayanta K. Ghosh, Institute of Mathematical Statistics, Beachwood, Ohio, 3, 318–329, 2008.

Bocquet, M.: Ensemble Kalman filtering without the intrinsic need for inflation, Nonlin. Processes Geophys., 18, 735–750, doi:10.5194/npg-18-735-2011, 2011.

Bocquet, M. and Sakov, P.: Combining inflation-free and iterative ensemble Kalman filters for strongly nonlinear systems, Nonlin. Processes Geophys., 19, 383–399, doi:10.5194/npg-19-383-2012, 2012.

Bocquet, M., Pires, C. A., and Wu, L.: Beyond Gaussian statistical modeling in geophysical data assimilation, Mon. Weather Rev., 138, 2997–3023, 2010.

Chen, R. and Liu, J. S.: Mixture Kalman filters, J. Roy. Stat. Soc. B, 62, 493–508, 2000.

Chorin, A. J. and Tu, X.: Implicit sampling for particle filters, P. Natl. Acad. Sci. USA, 106, 17249–17254, 2009.

Chorin, A. J., Morzfeld, M., and Tu, X.: Implicit particle filters for data assimilation, Communications in Applied Mathematics and Computational Science, 5, 221–240, 2010.

David, H. A. and Nagaraja, H. N.: Order statistics, John Wiley & Sons, 3rd Edn., Hoboken, 2003.

Dempster, A. P., Laird, N. M., and Rubin, D. B.: Maximum likelihood from incomplete data via the EM algorithm, J. Roy. Stat. Soc. B-Met., 39, 1–38, 1977.

Doucet, A., Godsill, S., and Andrieu, C.: On sequential Monte Carlo sampling methods for Bayesian filtering, Stat. Comput., 10, 197–208, 2000.

Dovera, L. and Rossa, E. D.: Multimodal ensemble Kalman filtering using Gaussian mixture models, Comput. Geosci., 15, 307–323, 2011.

Efron, B.: Bootstrap methods: another look at the jackknife, Ann. Stat., 7, 1–26, 1979.

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

Evensen, G.: Data assimilation: the ensemble Kalman filter, Springer, Dordrecht, 2nd Edn., 2009.

Evensen, G. and van Leeuwen, P. J.: An ensemble Kalman smoother for nonlinear dynamics, Mon. Weather Rev., 128, 1852–1867, 2000.

Eyink, G. L. and Kim, S.: A maximum entropy method for particle filtering, J. Stat. Phys., 123, 1071–1128, 2006.

Eyink, G. L. and Restrepo, J. M.: Most probable histories for nonlinear dynamics: tracking climate transitions, J. Stat. Phys., 101, 459–472, 2000.

Eyink, G. L., Restrepo, J. M., and Alexander, F. J.: A mean field approximation in data assimilation for nonlinear dynamics, Physica D, 195, 347–368, 2004.

Frei, M. and Künsch, H. R.: Mixture ensemble Kalman filters, Comput. Stat. Data An., 58, 127–138, 2013.

Geweke, J.: Bayesian inference in econometric models using Monte Carlo integration, Econometrica, 57, 1317–1339, 1989.

Gordon, N. J., Salmond, D. J., and Smith, A. F. M.: Novel approach to nonlinear/non-Gaussian Bayesian state estimation, IEE Proc. F, 140, 107–113, 1993.

Hoteit, I., Pham, D.-T., Triantafyllou, G., and Korres, G.: A new approximate solution of the optimal nonlinear filter for data assimilation in meteorology and oceanography, Mon. Weather Rev., 136, 317–334, 2008.

Jazwinski, A. H.: Stochastic processes and filtering theory, Academic Press, New York, 1970.

Kalnay, E. and Yang, S.-C.: Accelerating the spin-up of ensemble Kalman filtering, Q. J. Roy. Meteor. Soc., 136, 1644–1651, 2010.

Kim, S., Eyink, G. L., Restrepo, J. M., Alexander, F. J., and Johnson, G.: Ensemble filtering for nonlinear dynamics, Mon. Weather Rev., 131, 2586–2594, 2003.

Kitagawa, G.: Monte Carlo filter and smoother for non-Gaussian nonlinear state space models, J. Comput. Graph. Stat., 5, 1–25, 1996.

Klaas, M., de Freitas, N., and Doucet, A.: Toward practical N^2 Monte Carlo: the marginal particle filter, in: Proceedings of the 21st annual conference on uncertainty in artificial intelligence, AUAI Press, Arlington, 308–315, 2005.

Kloeden, P. E. and Platen, E.: Numerical solution of stochastic differential equations, Springer, Berlin, 1999.

Kong, A., Liu, J. S., and Wong, W. H.: Sequential imputations and Bayesian missing data problems, J. Am. Stat. Assoc., 89, 278–288, 1994.

Konishi, S. and Kitagawa, G.: Information criteria and statistical modeling, Springer, New York, 2008.

Kotecha, J. H. and Djurić, P. M.: Gaussian sum particle filtering, IEEE T. Signal Proces., 51, 2602–2612, 2003.

Kravtsov, S., Kondrashov, D., and Ghil, M.: Multilevel regression modeling of nonlinear processes: derivation and applications to climatic variability, J. Climate, 18, 4404–4424, 2005.

Kushner, H. J. and Yin, G. G.: Stochastic approximation and recursive algorithms and applications, Springer, New York, 2003.

Liu, J. S.: Metropolized independent sampling with comparisons to rejection sampling and importance sampling, Stat. Comput., 6, 113–119, 1996.

Majda, A. J., Timofeyev, I., and Vanden-Eijnden, E.: Systematic strategies for stochastic mode reduction in climate, J. Atmos. Sci., 60, 1705–1722, 2003.

Mardia, K. V.: Applications of some measures of multivariate skewness in testing normality and robustness studies, Sankhya Ser. B, 36, 115–128, 1974.

McLachlan, G. J. and Krishnan, T.: The EM algorithm and extensions, John Wiley & Sons, Hoboken, 2nd Edn., 2008.

McLachlan, G. J. and Peel, D.: Finite mixture models, John Wiley & Sons, New York, 2001.

Miller, R. N., Ghil, M., and Gauthiez, F.: Advanced data assimilation in strongly nonlinear dynamical systems, J. Atmos. Sci., 51, 1037–1056, 1994.

Miller, R. N., Carter Jr., E. F., and Blue, S. T.: Data assimilation into nonlinear stochastic models, Tellus A, 51, 167–194, 1999.

Moore, E. H.: On the reciprocal of the general algebraic matrix, in: The fourteenth western meeting of the American Mathematical Society, edited by: Dresden, A., Am. Math. Soc., 26, 385–396, 1920.

Morzfeld, M. and Chorin, A. J.: Implicit particle filtering for models with partial noise, and an application to geomagnetic data assimilation, Nonlin. Processes Geophys., 19, 365–382, doi:10.5194/npg-19-365-2012, 2012.

Morzfeld, M., Tu, X., Atkins, E., and Chorin, A. J.: A random map implementation of implicit filters, J. Comput. Phys., 231, 2049–2066, 2012.

Øksendal, B. K.: Stochastic differential equations: an introduction with applications, Springer, Berlin, 6th Edn., 2003.

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

Penrose, R.: A generalized inverse for matrices, Math. Proc. Cambridge, 51, 406–413, 1951.

Rauch, H. E.: Solutions to the linear smoothing problem, IEEE T. Automat. Contr., AC-8, 371–372, 1963.

Reich, S.: A Gaussian-mixture ensemble transform filter, Q. J. Roy. Meteor. Soc., 138, 222–233, 2012.

Robbins, H. and Monro, S.: A stochastic approximation method, Ann. Math. Stat., 22, 400–407, 1951.

Silverman, B. W.: Density estimation for statistics and data analysis, Chapman and Hall/CRC, London, 1986.

Smith, K. W.: Cluster ensemble Kalmn filter, Tellus A, 59, 749–757, doi:10.1111/j.1600-0870.2007.00246.x, 2007.

Snyder, C.: Particle filters, the "optimal" proposal and high-dimensional systems, Seminar on data assimilation for atmosphere and ocean, Reading, England, 6–9 September 2011, ECMWF, 2012.

Snyder, C., Bengtsson, T., Bickel, P., and Anderson, J.: Obstacles to high-dimensional particle filtering, Mon. Weather Rev., 136, 4629–4640, 2008.

Sondergaard, T. and Lermusiaux, P. F. J.: Data assimilation with Gaussian mixture models using the dynamically orthogonal field

equations. Part I: Theory and scheme, Mon. Weather Rev., 141, 1737–1760, doi:10.1175/MWR-D-11-00295.1, 2013a.

Sondergaard, T. and Lermusiaux, P. F. J.: Data assimilation with Gaussian mixture models using the dynamically orthogonal field equations. Part II: Applications, Mon. Weather Rev., 141, 1761–1785, doi:10.1175/MWR-D-11-00296.1, 2013b.

Stordal, A. S., Karlsen, H. A., Nævdal, G., Skaug, H. J., and Vallès, B.: Bridging the ensemble Kalman filter and particle filters: the adaptive Gaussian mixture filter, Comput. Geosci., 15, 293–305, 2011.

Sutera, A.: On stochastic perturbations and long-term climate behavior, Q. J. Roy. Meteor. Soc., 107, 137–151, 1981.

van Leeuwen, P. J.: Particle filtering in geophysical systems, Mon. Weather Rev., 137, 4089–4114, 2009.

van Leeuwen, P. J.: Nonlinear data assimilation in geosciences: an extremely efficient particle filter, Q. J. Roy. Meteor. Soc., 136, 1991–1999, 2010.

van Leeuwen, P. J.: Efficient nonlinear data-assimilation in geophysical fluid dynamics, Comput. Fluids, 46, 52–58, 2011.

van Leeuwen, P. J. and Evensen, G.: Data assimilation and inverse methods in terms of a probabilistic formulation, Mon. Weather Rev., 124, 2898–2913, 1996.

Weare, J.: Particle filtering with path sampling and an application to a biomodal ocean current model, J. Comput. Phys., 228, 4312–4331, 2009.

Weir, B., Miller, R. N., and Spitz, Y. H.: Implicit estimation of ecological model parameters, B. Math. Biol., 75, 223–257, doi:10.1007/s11538-012-9801-6, 2013.